Volume 84, Number 8
August 2023

ISSN 0005-1179
CODEN: AURCAT

AUTOMATION AND REMOTE CONTROL

Automation and Remote Control

Vol. 84, No. 8, August 2023

http://ait.mtas.ru

Automation and Remote Control

ISSN 0005-1179

SCOPE

Automation and Remote Control is one of the first journals on control theory. The scope of the journal is control theory problems and applications. The journal publishes reviews, original articles, and short communications (deterministic, stochastic, adaptive, and robust formulations) and its applications (computer control, components and instruments, process control, social and economy control, etc.).

Automation and Remote Control is abstracted and/or indexed in *ACM Digital Library, BFI List, CLOCKSS, CNKI, CNPIEC Current Contents/ Engineering, Computing and Technology, DBLP, Dimensions, EBSCO Academic Search, EBSCO Advanced Placement Source, EBSCO Applied Science & Technology Source, EBSCO Computer Science Index, EBSCO Computers & Applied Sciences Complete, EBSCO Discovery Service, EBSCO Engineering Source, EBSCO STM Source, EI Compendex, Google Scholar, INSPEC, Japanese Science and Technology Agency (JST), Journal Citation Reports/Science Edition, Mathematical Reviews, Naver, OCLC WorldCat Discovery Service, Portico, ProQuest Advanced Technologies & Aerospace Database, ProQuest-ExLibris Primo, ProQuest-ExLibris Summon, SCImago, SCOPUS, Science Citation Index, Science Citation Index Expanded (Sci-Search), TD Net Discovery Service, UGC-CARE List (India), WTI Frankfurt eG, zbMATH.*

Journal website: http://ait.mtas.ru

0005-1179/23. *Automation and Remote Control* (ISSN: 0005-1179 print version, ISSN: 1608-3032 electronic version) is published monthly by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 65 Profsoyuznaya street, Moscow 117997, Russia.

Volume 84 (12 issues) is published in 2023.

Publisher: Trapeznikov Institute of Control Sciences, Russian Academy of Sciences.

65 Profsoyuznaya street, Moscow 117997, Russia; e-mail: redacsia@ipu.rssi.ru; http://ait.mtas.ru, http://ait-arc.ru

Contents

Automation and Remote Control
Vol. 84, No. 8, 2023

Linear Systems

Nonlinear Systems

Stochastic Systems

Control in Technical Systems

Optimization, System Analysis, and Operations Research

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 8, pp. 901–917.

LINEAR SYSTEMS

PI Controller Design for Suppressing Exogenous Disturbances

M. V. Khlebnikov[*,**]

[*] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
[**] *Moscow Institute of Physics and Technology (National Research University), Dolgoprudny, Russia*
e-mail: khlebnik@ipu.ru

Received April 11, 2023
Revised June 6, 2023
Accepted June 9, 2023

Abstract—A novel approach is proposed to suppress bounded exogenous disturbances in linear control systems using a PI controller. The approach is based on reducing the original problem to a nonconvex matrix optimization problem. A gradient method for finding the controller's parameters is derived and its justification is provided. The corresponding recurrence procedure is rather effective and yields quite satisfactory controllers in terms of engineering performance criteria. This paper continues a series of the author's research works devoted to the design of feedback control laws from an optimization point of view.

Keywords: linear system, exogenous disturbances, PI controller, optimization, Lyapunov equation, gradient method, Newton's method, convergence

DOI: 10.25728/arcRAS.2023.66.91.001

1. INTRODUCTION

The recent paper [1] introduced a novel (optimization-based) approach to the classical problem of suppressing bounded nonrandom exogenous disturbances. This problem is posed as follows. Consider a linear control system described by

$$\begin{aligned}\dot{x} &= Ax + Bu + Dw, \quad x(0) = x_0,\\ y &= C_1 x,\\ z &= C_2 x + B_1 u\end{aligned}$$

with the state vector $x(t) \in \mathbb{R}^n$, the measured output $y(t) \in \mathbb{R}^l$, the controlled output $z(t) \in \mathbb{R}^r$, the control vector $u(t) \in \mathbb{R}^p$, and a measured disturbance $w(t) \in \mathbb{R}^m$ that is bounded at each time instant t :

$$|w(t)| \leqslant 1 \quad \text{for all } t \geqslant 0. \tag{1}$$

It is required to choose a stabilizing state-feedback $u = Kx$ or output-feedback $u = Ky$ control (if it exists) to reduce the "peak" of the output $z(t)$, i.e., the value $\max_t |z(t)|$.

Within the approach presented in [1], the original problem was reduced to a nonconvex matrix optimization problem. A gradient method for finding a static state-feedback or output-feedback control law of the system was developed, and its justification was given.

On the other hand, in [2], an optimization approach going back to [3] was applied to design a PID controller. The regular approach proposed therein involves solving a nonconvex matrix optimization problem to find the controller's parameters. The quality of this controller was evaluated by a quadratic criterion of the system output: the controller was tuned against the uncertainty in the

initial conditions to make the system output uniformly small. As it turned out, the corresponding recurrence procedure is rather effective and yields controllers that are quite satisfactory in terms of engineering performance criteria.

This paper continues both of the research lines mentioned above: we design a PI controller for suppressing bounded exogenous disturbances in linear control systems by solving an optimization problem.

From this point onwards, the following notations are adopted: $|\cdot|$ is the Euclidean norm of a vector, $\|\cdot\|$ is the spectral norm of a matrix, $\|\cdot\|_F$ is the Frobenius norm of a matrix, $^{\mathrm{T}}$ stands for the transpose operation, tr means the matrix trace, I is an identity matrix of appropriate dimensions, and $\lambda_i(A)$ are the eigenvalues of a matrix A.

2. PROBLEM STATEMENT. THE METHOD OF INVARIANT ELLIPSOIDS

Consider a linear continuous-time control system described by

$$\begin{aligned} \dot{x} &= Ax + bu + Dw, \quad x(0) = x_0, \\ y &= c^{\mathrm{T}}x, \\ z &= Cx, \end{aligned} \tag{2}$$

where $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, $D \in \mathbb{R}^{n\times m}$, and $c \in \mathbb{R}^n$, $C \in \mathbb{R}^{r\times n}$, with the state vector $x(t) \in \mathbb{R}^n$, the observed output $y(t) \in \mathbb{R}$, the controlled output $z(t) \in \mathbb{R}^r$, an exogenous disturbance $w(t) \in \mathbb{R}^m$ that satisfies the constraint (1), and the control vector $u(t) \in \mathbb{R}$ in the form of a PI controller

$$u(t) = -k_P y(t) - k_I \int_0^t y(\tau)d\tau. \tag{3}$$

The objective is to find the numerical parameters k_P and k_I of the controller (3) that stabilizes the closed loop system and suppresses the exogenous disturbances w by minimizing the bounding ellipsoid for the output z.

Let us conceptually recall the method of invariant ellipsoids; for details, see [4, 5]. Consider a linear continuous time-invariant dynamic system described by

$$\begin{aligned} \dot{x} &= Ax + Dw, \quad x(0) = x_0, \\ z &= Cx \end{aligned} \tag{4}$$

with the state vector $x(t) \in \mathbb{R}^n$, the output $z(t) \in \mathbb{R}^r$, and an exogenous disturbance $w(t) \in \mathbb{R}^l$ that satisfies the constraint (1). Assume that system (4) is stable (i.e., the matrix A is Hurwitz) and the pair (A, D) is controllable.

An ellipsoid centered at the origin is said to be *invariant* for system (4) if any of its trajectories evolving from a point inside the ellipsoid remains in this ellipsoid at any time instant under all admissible exogenous disturbances of the system.

When evaluating the effect of exogenous disturbances on the system output, it is natural to consider the minimal ellipsoids containing the system output (in a certain sense). Clearly, if an ellipsoid

$$\mathcal{E}_x = \left\{ x \in \mathbb{R}^n : \quad x^{\mathrm{T}}P^{-1}x \leqslant 1 \right\}, \quad P \succ 0, \tag{5}$$

is invariant, then the output of system (4) with $x_0 \in \mathcal{E}_x$ belongs to the so-called *bounding* ellipsoid

$$\mathcal{E}_z = \left\{ z \in \mathbb{R}^p : \quad z^{\mathrm{T}}(CPC^{\mathrm{T}})^{-1}z \leqslant 1 \right\}. \tag{6}$$

In the literature, the linear function $f(P) = \operatorname{tr} CPC^{\mathrm{T}}$ (the sum of the squares of the semi-axes of the bounding ellipsoid) is often considered a minimality criterion.

The paper [6] established an invariance criterion for ellipsoids in terms of linear matrix inequalities (LMIs). Let us formulate it as follows (see [4]).

Theorem 1. *Assume that the matrix A is Hurwitz, the pair (A, D) is controllable, and the matrix $P(\alpha) \succ 0$ satisfies the Lyapunov equation*

$$\Big(A + \frac{\alpha}{2} I\Big) P + P\Big(A + \frac{\alpha}{2} I\Big)^{\mathrm{T}} + \frac{1}{\alpha} DD^{\mathrm{T}} = 0$$

on the interval $0 < \alpha < 2\sigma(A)$.

Then the minimal bounding ellipsoid is obtained by minimizing the univariate function $f(\alpha) = \operatorname{tr} CP(\alpha)C^{\mathrm{T}}$ on the interval $0 < \alpha < 2\sigma(A)$; if α^* is the minimum point and x_0 satisfies the condition $x_0^{\mathrm{T}} P^{-1}(\alpha^*) x_0 \leqslant 1$, then the uniform estimate

$$|z(t)| \leqslant \sqrt{f(\alpha^*)}, \quad 0 \leqslant t < \infty,$$

holds.

3. SOLUTION APPROACH

Let us introduce an auxiliary scalar variable ξ as follows:

$$\dot{\xi} = y, \quad \xi(0) = 0.$$

With the extended state vector

$$g = \begin{pmatrix} x \\ \xi \end{pmatrix} \in \mathbb{R}^{n+1},$$

system (2) can be written as

$$\begin{aligned} \dot{g} &= \begin{pmatrix} A & 0 \\ c^{\mathrm{T}} & 0 \end{pmatrix} g + \begin{pmatrix} b \\ 0 \end{pmatrix} u + \begin{pmatrix} D \\ 0 \end{pmatrix} w, \quad g(0) = \begin{pmatrix} x_0 \\ 0 \end{pmatrix}, \\ y &= \begin{pmatrix} c^{\mathrm{T}} & 0 \end{pmatrix} g. \end{aligned} \tag{7}$$

According to (2) and (3), we have

$$\begin{aligned} u &= -k_P y(t) - k_I \int_0^t y(\tau) d\tau = -k_P c^{\mathrm{T}} x - k_I \xi \\ &= -k_P c^{\mathrm{T}} x - k_I \xi = -k_P \begin{pmatrix} c^{\mathrm{T}} & 0 \end{pmatrix} g - k_I \begin{pmatrix} 0 & 1 \end{pmatrix} g. \end{aligned} \tag{8}$$

The expression (8) with the more convenient notations $k_1 = k_P$ and $k_2 = k_I$ takes the form

$$u = -\begin{pmatrix} k_1 c^{\mathrm{T}} & k_2 \end{pmatrix} g. \tag{9}$$

Thus, system (7) with the feedback control law (9) is described by

$$\dot{g} = \begin{pmatrix} A - k_1 b c^{\mathrm{T}} & -k_2 b \\ c^{\mathrm{T}} & 0 \end{pmatrix} g + \begin{pmatrix} D \\ 0 \end{pmatrix} w, \quad g(0) = \begin{pmatrix} x_0 \\ 0 \end{pmatrix}.$$

It can be represented as

$$\dot g = (\mathcal{A}_0 + k_1\mathcal{A}_1 + k_2\mathcal{A}_2)g + \begin{pmatrix} D \\ 0 \end{pmatrix} w, \quad g(0) = \begin{pmatrix} x_0 \\ 0 \end{pmatrix},$$

where

$$\mathcal{A}_0 = \begin{pmatrix} A & 0 \\ c^{\mathrm{T}} & 0 \end{pmatrix}, \quad \mathcal{A}_1 = \begin{pmatrix} -bc^{\mathrm{T}} & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathcal{A}_2 = \begin{pmatrix} 0 & -b \\ 0 & 0 \end{pmatrix}.$$

Following the method of invariant ellipsoids, let the state g of system (7) belong to the invariant ellipsoid (5) generated by a matrix $P \in \mathbb{R}^{(n+1)\times(n+1)}$. We will minimize the size of the corresponding bounding ellipsoid (6) with respect to the output

$$z = Cx = \begin{pmatrix} C & 0 \end{pmatrix} g.$$

Due to Theorem 1, the associated problem is to minimize $\operatorname{tr}(C\ 0)P(C\ 0)^{\mathrm{T}}$ subject to the constraint

$$\left(\mathcal{A}_0 + k_1\mathcal{A}_1 + k_2\mathcal{A}_2 + \frac{\alpha}{2}I\right) P + P\left(\mathcal{A}_0 + k_1\mathcal{A}_1 + k_2\mathcal{A}_2 + \frac{\alpha}{2}I\right)^{\mathrm{T}} + \frac{1}{\alpha}\begin{pmatrix} D \\ 0 \end{pmatrix}\begin{pmatrix} D \\ 0 \end{pmatrix}^{\mathrm{T}} = 0 \qquad (10)$$

with respect to the matrix variables $P = P^{\mathrm{T}} \in \mathbb{R}^{n\times n}$, the scalar variables k_1 and k_2, and the scalar parameter $\alpha > 0$. Given k_1, k_2, and α, the matrix P is found from equation (10); therefore, the independent variables are k_1, k_2, and α.

Consider the vector

$$k = \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} \in \mathbb{R}^2$$

and the value

$$\operatorname{tr}\begin{pmatrix} C & 0 \end{pmatrix} P\begin{pmatrix} C & 0 \end{pmatrix}^{\mathrm{T}} + \rho|k|^2, \quad \rho \ll 1$$

as the performance criterion. Here, the second component is a control penalty (the coefficient $\rho > 0$ adjusts its significance) and ensures the coercivity of the objective function in k. (For details, see Section 5.)

Thus, the original problem (the design of a PI controller to suppress exogenous disturbances) has been reduced to the matrix optimization problem

$$\min f(k,\alpha), \quad f(k,\alpha) = \operatorname{tr} P\begin{pmatrix} C & 0 \end{pmatrix}^{\mathrm{T}}\begin{pmatrix} C & 0 \end{pmatrix} + \rho|k|^2 \qquad (11)$$

subject to the constraint (10).

4. OPTIMIZATION OF THE FUNCTION $f(\alpha)$

Consider the problem

$$\min f(\alpha), \quad f(\alpha) = \operatorname{tr} PC^{\mathrm{T}}C,$$

subject to the constraint

$$\left(A + \frac{\alpha}{2}I\right) P + P\left(A + \frac{\alpha}{2}I\right)^{\mathrm{T}} + \frac{1}{\alpha}DD^{\mathrm{T}} = 0$$

with respect to the matrix variable $P = P^{\mathrm{T}} \in \mathbb{R}^{n\times n}$ and the scalar parameter $\alpha > 0$. Assume that the matrix A is stable (Hurwitz).

As was shown in [1], minimization with respect to α can be effectively performed using Newton's method. Let us choose an initial approximation $0 < \alpha_0 < 2\sigma(A)$ and apply the iterative process

$$\alpha_{j+1} = \alpha_j - \frac{f'(\alpha_j)}{f''(\alpha_j)},$$

where

$$f'(\alpha) = \operatorname{tr} Y \left(P - \frac{1}{\alpha^2} DD^{\mathrm{T}} \right),$$
$$f''(\alpha) = 2 \operatorname{tr} Y \left(X + \frac{1}{\alpha^3} DD^{\mathrm{T}} \right),$$

and Y and X are the solutions of the Lyapunov equations

$$\left(A + \frac{\alpha}{2} I \right)^{\mathrm{T}} Y + Y \left(A + \frac{\alpha}{2} I \right) + C^{\mathrm{T}} C = 0$$

and

$$\left(A + \frac{\alpha}{2} I \right) X + X \left(A + \frac{\alpha}{2} I \right)^{\mathrm{T}} + P - \frac{1}{\alpha^2} DD^{\mathrm{T}} = 0,$$

respectively.

According to [1], the method converges globally (faster than the geometric progression with a coefficient of 1/2), with quadratic convergence in the neighborhood of the solution. It really requires at most 3–4 iterations to obtain a solution with high accuracy, unless the initial point is too close to the limits of the interval $(0, 2\sigma(A))$.

Thus, we have an efficient algorithm to perform minimization with respect to α in problem (11), (10): it suffices to replace the matrix A by $\mathcal{A}_0 + k_1\mathcal{A}_1 + k_2\mathcal{A}_2$, the matrix C by $\begin{pmatrix} C & 0 \end{pmatrix}$, and the matrix D by $\begin{pmatrix} D \\ 0 \end{pmatrix}$.

5. OPTIMIZATION OF THE FUNCTION $f(k)$

Introducing the convenient notation

$$\{\mathcal{A}, k\} = k_1\mathcal{A}_1 + k_2\mathcal{A}_2,$$

we accept the following hypothesis.

Assumption. *Let* $k_0 = \begin{pmatrix} k_1^0 \\ k_2^0 \end{pmatrix}$ *be a known stabilizing controller, i.e., the matrix* $\mathcal{A}_0 + \{\mathcal{A}, k_0\}$ *is Hurwitz.*

We will investigate the properties of the function

$$f(k) = \min_{\alpha} f(k, \alpha).$$

Lemma 1. *The function* $f(k)$ *is well-defined and positive on the set* $\mathcal{S}$ *of stabilizing controllers.*

The proofs of this and all subsequent results are given in Appendix 2.

Note that the set $\mathcal{S}$ can be nonconvex and disconnected whereas its boundaries can be nonsmooth.

Lemma 2. *The function $f(k,\alpha)$ is well-defined on the set of stabilizing feedback control laws k and for $0<\alpha<2\sigma(\mathcal{A}_0+\{\mathcal{A},k\})$. It is differentiable on this set, and the gradient is given by*

$$\frac{1}{2}\nabla_k f(k,\alpha)=\begin{pmatrix}\operatorname{tr} PY\mathcal{A}_1\\ \operatorname{tr} PY\mathcal{A}_2\end{pmatrix}+\rho k, \tag{12}$$

$$\nabla_\alpha f(k,\alpha)=\operatorname{tr} Y\left[P-\frac{1}{\alpha^2}\begin{pmatrix}D\\0\end{pmatrix}\begin{pmatrix}D\\0\end{pmatrix}^{\mathrm{T}}\right], \tag{13}$$

where the matrices P and Y are the solutions of the Lyapunov equations

$$\left(\mathcal{A}_0+\{\mathcal{A},k\}+\frac{\alpha}{2}I\right)P+P\left(\mathcal{A}_0+\{\mathcal{A},k\}+\frac{\alpha}{2}I\right)^{\mathrm{T}}+\frac{1}{\alpha}\begin{pmatrix}D\\0\end{pmatrix}\begin{pmatrix}D\\0\end{pmatrix}^{\mathrm{T}}=0$$

and

$$\left(\mathcal{A}_0+\{\mathcal{A},k\}+\frac{\alpha}{2}I\right)^{\mathrm{T}}Y+Y\left(\mathcal{A}_0+\{\mathcal{A},k\}+\frac{\alpha}{2}I\right)+\begin{pmatrix}C&0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}C&0\end{pmatrix}=0, \tag{14}$$

respectively.

The function $f(k,\alpha)$ achieves minimum at an inner point of the admissible set that is determined by the conditions

$$\nabla_k f(k,\alpha)=0,\quad \nabla_\alpha f(k,\alpha)=0.$$

In addition, $f(k,\alpha)$ as a function of α is strictly convex on $0<\alpha<2\sigma(\mathcal{A}_0+\{\mathcal{A},k\})$ and achieves minimum at an inner point of this interval.

The Hessian of the function $f(k)$ has the following properties.

Lemma 3. *The function $f(k)$ is twice differentiable, and the action of its Hessian on an arbitrary vector*[1] *$e\in\mathbb{R}^2$ is given by*

$$\frac{1}{2}\left(\nabla^2_{kk}f(k)e,e\right)=\rho(e,e)+2\operatorname{tr} P'Y\{\mathcal{A},e\}, \tag{15}$$

where P' is the solution of the Lyapunov equation

$$\left(\mathcal{A}_0+\{\mathcal{A},k\}+\frac{\alpha}{2}I\right)P'+P'\left(\mathcal{A}_0+\{\mathcal{A},k\}+\frac{\alpha}{2}I\right)^{\mathrm{T}}+\{\mathcal{A},e\}P+P\{\mathcal{A},e\}^{\mathrm{T}}=0. \tag{16}$$

Remark 1. To obtain simple quantitative estimates in Lemmas 4 and 5 below, we incorporate the regularizing terms ε_1 and ε_2 into the optimization problem (11), (10) as follows:

$$\min f(k,\alpha),\quad f(k,\alpha)=\operatorname{tr} P\left(\begin{pmatrix}C&0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}C&0\end{pmatrix}+\varepsilon_1 I\right)+\rho|k|^2,\quad \varepsilon_1\ll 1$$

subject to the constraint

$$\left(\mathcal{A}_0+\{\mathcal{A},k\}+\frac{\alpha}{2}I\right)P+P\left(\mathcal{A}_0+\{\mathcal{A},k\}+\frac{\alpha}{2}I\right)^{\mathrm{T}}+\frac{1}{\alpha}\left[\begin{pmatrix}D\\0\end{pmatrix}\begin{pmatrix}D\\0\end{pmatrix}^{\mathrm{T}}+\varepsilon_2 I\right]=0,\quad \varepsilon_2\ll 1. \tag{17}$$

The requirement of their introduction can be significantly weakened, but the current aim is to obtain the simplest and most obvious results.

[1] In the sense of the second derivative in a direction (the second directional derivative).

Lemma 4. *The function $f(k)$ is coercive on the set $\mathcal{S}$ (i.e., tends to infinity on its boundary) and, moreover,*

$$f(k) \geqslant \frac{\varepsilon_1}{4\sigma(\mathcal{A}_0+\{\mathcal{A},k\})\left(\|\mathcal{A}_0+\{\mathcal{A},k\}\|+\sigma(\mathcal{A}_0+\{\mathcal{A},k\})\right)}\|D\|_F^2, \tag{18}$$
$$f(k) \geqslant \rho|k|^2.$$

Corollary 1. *The level set*

$$\mathcal{S}_0 = \{k \in \mathcal{S}: \quad f(k) \leqslant f(k_0)\}$$

is bounded for any controller $k_0 \in \mathcal{S}$.

Corollary 2. *There exists a minimum point k_* on the set S and $\nabla f(k_*) = 0$.*

The gradient of the function $f(k)$ is not Lipschitz on the entire set $\mathcal{S}$, but it has this property on its subset $\mathcal{S}_0$. The corresponding result is presented below.

Lemma 5. *On the set $\mathcal{S}_0$, the gradient of the function $f(k)$ is Lipschitz with the constant*

$$L = \rho + \frac{8\sqrt{2n}f^2(k_0)}{\varepsilon_1\varepsilon_2^2}\left(\|\mathcal{A}_0\| + \max_i\|\mathcal{A}_i\|\sqrt{\frac{2}{\rho}f(k_0)}\right)^2\left(\frac{f^2(k_0)}{\varepsilon_1^2} + 2\max_i\|\mathcal{A}_i\|^2\right)\max\|\mathcal{A}_i\|_F. \tag{19}$$

These properties of the function $f(k)$ and its derivatives allow constructing a minimization method and justifying its convergence.

6. OPTIMIZATION ALGORITHM

We propose an iterative approach to solve problem (11). This approach is based on the application of the gradient method with respect to the variable k and Newton's method with respect to the variable α. The algorithm includes several steps as follows.

Algorithm 1 to minimize $f(k,\alpha)$:

1. Choose some values of the parameters $\varepsilon > 0$, $\gamma > 0$, $0 < \tau < 1$, and the initial stabilizing approximation k_0. Calculate $\alpha_0 = \sigma(\mathcal{A}_0 + \{\mathcal{A}, k_0\})$.
2. On the jth iteration, the values k_j and α_j are given.

   Calculate the matrix $\mathcal{A}_0 + \{\mathcal{A}, k_j\}$, solve the Lyapunov equations

$$\left(\mathcal{A}_0+\{\mathcal{A},k_j\}+\frac{\alpha_j}{2}I\right)P + P\left(\mathcal{A}_0+\{\mathcal{A},k_j\}+\frac{\alpha_j}{2}I\right)^{\mathrm{T}} + \frac{1}{\alpha_j}\begin{pmatrix}D\\0\end{pmatrix}\begin{pmatrix}D\\0\end{pmatrix}^{\mathrm{T}} = 0,$$
$$\left(\mathcal{A}_0+\{\mathcal{A},k_j\}+\frac{\alpha_j}{2}I\right)^{\mathrm{T}}Y + Y\left(\mathcal{A}_0+\{\mathcal{A},k_j\}+\frac{\alpha_j}{2}I\right) + \begin{pmatrix}C & 0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}C & 0\end{pmatrix} = 0,$$

   and find the matrices P and Y.

   Calculate the gradient

$$H_j = \nabla_k f(k_j, \alpha_j)$$

   from the relation

$$\frac{1}{2}\nabla_k f(k,\alpha) = \begin{pmatrix}\operatorname{tr} PY\mathcal{A}_1\\ \operatorname{tr} PY\mathcal{A}_2\end{pmatrix} + \rho k.$$

   If $|H_j| \leqslant \varepsilon$, then take k_j as the approximate solution.

3. Perform the gradient method step:

$$k_{j+1} = k_j - \gamma_j H_j.$$

Adjust the step length $\gamma_j > 0$ by fractionating γ until the following conditions are satisfied:

a. k_{j+1} is a stabilizing controller.

b. $f(k_{j+1}) \leqslant f(k_j) - \tau\gamma_j |H_j|^2$.

4. Minimize $f(k_{j+1}, \alpha)$ with respect to α (see Section 4) and find α_{j+1}. Revert to Step 2.

This method converges in the following sense.

Theorem 2. *In Algorithm 1, only a finite number of fractions are realized for γ_j at each iteration, the function $f(k_j)$ is monotonically decreasing, and its gradient vanishes with an exponential rate (like a geometric progression):*

$$\lim_{j\to\infty} |H_j| = 0.$$

Indeed, Algorithm 1 is well-defined at the initial point since k_0 is a stabilizing controller by the assumption. For sufficiently small γ_j, the function $f(k)$ monotonically decreases (moves in the direction of its antigradient); with this step adjustment, the values of k_j remain in the domain $\mathcal{S}_0$, where Lemma 5 ensures the Lipschitz property of the gradient. Thus, the gradient method for unconstrained minimization is convergent [7]. In particular, condition b) at Step 3 of Algorithm 1 will be satisfied after a finite number of fractions, and the gradient method will have gradient convergence with a linear rate.

Naturally, it is difficult to expect convergence to a global minimum: the domain of definition of $f(k)$ may even be disconnected.

7. EXAMPLE

Consider an illustrative example from the paper [8]. The transfer function has the form

$$G(s) = \frac{1}{(1+s)(1+\alpha s)(1+\alpha^2 s)(1+\alpha^3 s)}, \quad \alpha = 0.5.$$

MATLAB's procedure `tf2ss` gives the following matrices of system (4) in the state space:

$$A = \begin{pmatrix} -15 & -70 & -120 & -64 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad c = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 64 \end{pmatrix}.$$

Let us choose the matrix

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

and the controlled output matrix

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

We assign $\rho = 0.001$ and the stabilizing controller

$$k_0 = \begin{pmatrix} 1.7366 \\ 0.7734 \end{pmatrix}$$

as an initial one.

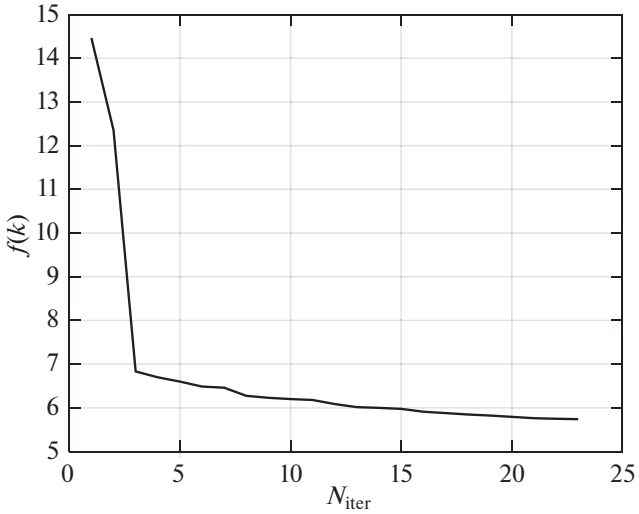


Fig. 1. Optimization procedure.

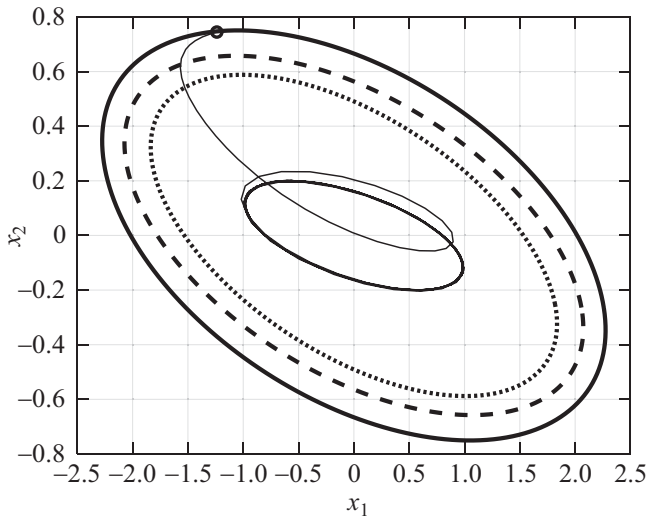


Fig. 2. Bounding ellipses.

The dynamics of the criterion $f(k)$ are demonstrated in Fig. 1. The process terminates with the PI controller with the gains

$$k_* = \begin{pmatrix} 0.2956 \\ 0.3514 \end{pmatrix}$$

and the corresponding bounding ellipse with the matrix

$$P_* = \begin{pmatrix} 5.1763 & -0.7885 \\ -0.7885 & 0.5635 \end{pmatrix}, \quad \operatorname{tr} P_* = 5.7398.$$

In Fig. 2, the solid line indicates the bounding ellipse and the trajectory of the closed loop system with the PI controller k_* under some admissible exogenous disturbance. Here, the dashed line shows the bounding ellipse for the closed loop system with the dynamic controller (see [4])

$$u = K\widehat{x},$$

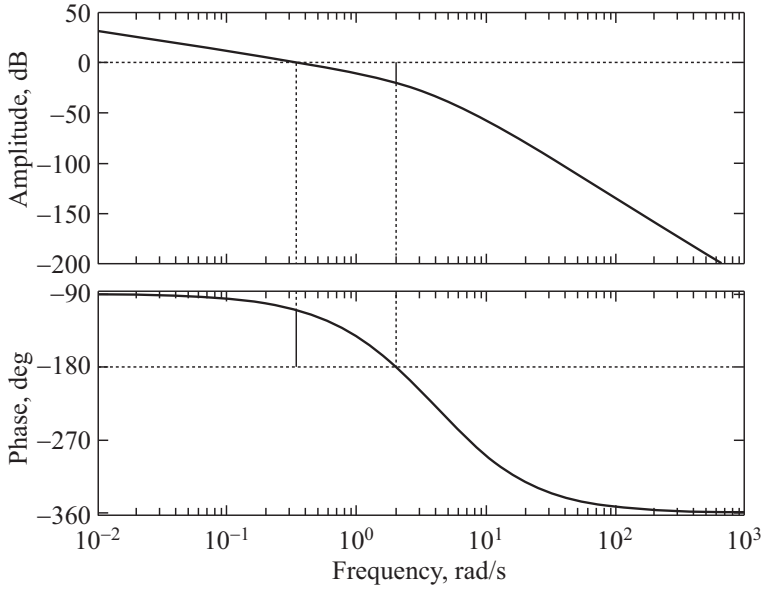


Fig. 3. The logarithmic amplitude-phase frequency response of the closed loop system.

where $\widehat{x}$ is an observer

$$\dot{\widehat{x}} = A\widehat{x} + bu + L(y - c^{\mathrm{T}}\widehat{x}), \quad \widehat{x}(0) = 0$$

with the matrices

$$K = \begin{pmatrix} -0.5154 & -2.6143 & -4.3786 & -2.4252 \end{pmatrix} \times 10^6, \qquad L = \begin{pmatrix} 0.0075 \\ -0.0225 \\ -0.0002 \\ 0.0189 \end{pmatrix}.$$

Finally, the dotted line in Fig. 2 presents the bounding ellipse for the closed loop system with the linear dynamic controller (see [4])

$$\dot{x}_r = A_r x_r + B_r y, \quad x_r(0) = 0,$$
$$u = C_r x_r + D_r y$$

with the matrices

$$A_r = \begin{pmatrix} -0.1373 & -0.6748 & -1.0932 & -0.1035 \\ 0.0140 & 0.0688 & 0.1114 & -1.7096 \\ 0.0004 & 0.0019 & 0.0031 & -0.0509 \\ 0.0000 & 0.0000 & 0.0001 & -0.0007 \end{pmatrix} \times 10^5, \qquad B_r = \begin{pmatrix} -0.7528 \\ 2.7644 \\ 0.0821 \\ 0.0011 \end{pmatrix} \times 10^3,$$

$$C_r = \begin{pmatrix} -0.1135 & -0.5579 & -0.9037 & -2.9271 \end{pmatrix} \times 10^5, \quad D_r = 3.8176 \times 10^3.$$

Clearly, the PI controller leads to quite comparable results, being advantageous by simplicity and convenience of practical implementation. In addition, the PI controller has satisfactory characteristics.

The transfer function of the PI controller with the coefficients k_* has the form

$$G_{\mathrm{PID}}(s) = 0.2956 + \frac{0.3514}{s}.$$

The closed loop system with the PI controller k_* is stable by the Nyquist criterion; its minimal gain and phase margins are 20.6 dB and 70.3°, respectively (Fig. 3).

For comparison, choosing the initial stabilizing controller

$$\widetilde{k}_0 = \begin{pmatrix} 0.8882 \\ 0.6153 \end{pmatrix},$$

we obtain the PI controller with the gains

$$\widetilde{k}_* = \begin{pmatrix} 0.3277 \\ 0.3662 \end{pmatrix}$$

and the corresponding bounding ellipse with the matrix

$$\widetilde{P}_* = \begin{pmatrix} 5.0890 & -0.7854 \\ -0.7854 & 0.5721 \end{pmatrix}, \quad \operatorname{tr} \widetilde{P}_* = 5.6611.$$

The norms of the resulting controllers differ by 6.5% only whereas the bounding ellipses by less than 1.5% (in terms of the trace criterion).

All the calculations were carried out in MATLAB using `CVX` [9], a free package.

8. DISCUSSION

This paper has proposed a novel approach to designing a PI controller that optimally suppresses bounded exogenous disturbances in a linear control system. The approach is based on reducing the original problem to a nonconvex matrix optimization problem, which is further solved by the gradient method. Its justification has been provided as well.

Note that Theorem 2 establishes the convergence of this method only in the norm of the gradient of the objective function. However, according to numerical simulations, the method yields quite satisfactory PI controllers from an engineering point of view. At the same time, it seems important to consider meaningful particular formulations of the problem where the function $f(k)$ satisfies on the level set $\mathcal{S}_0$ the Polyak–Łojasiewicz condition [7]

$$\frac{1}{2}|\nabla f(k)|^2 \geqslant \mu\,(f(k) - f(k_*))$$

with a constant $\mu > 0$ depending only on k_0 and the parameters of system (2). In this case, one could also speak of strong pointwise convergence, similar to what was shown in [3] for the linear quadratic problem with state-feedback control.

Finally, it would be interesting to extend this approach to the design of PID controllers, which will be the subject of subsequent publications.

APPENDIX A

The lemmas below contain well-known results necessary for the further presentation.

Lemma A.1 [1]. *Let X and Y be the solutions of the dual Lyapunov equations with a Hurwitz matrix A:*

$$A^{\mathrm{T}}X + XA + W = 0 \quad \textit{and} \quad AY + YA^{\mathrm{T}} + V = 0.$$

Then

$$\operatorname{tr}(XV) = \operatorname{tr}(YW).$$

Lemma A.2 [10].

1. *Matrices A and B of compatible dimensions satisfy the relations*

$$\|AB\|_F \leqslant \|A\|_F \|B\|,$$
$$|\operatorname{tr} AB| \leqslant \|A\|_F \|B\|_F,$$
$$\|A\| \leqslant \|A\|_F,$$
$$AB + B^{\mathrm{T}}A^{\mathrm{T}} \leqslant \varepsilon AA^{\mathrm{T}} + \frac{1}{\varepsilon}B^{\mathrm{T}}B \quad \text{for any } \varepsilon > 0.$$

2. *Nonnegative definite matrices A and B satisfy the relations*

$$0 \leqslant \lambda_{\min}(A)\lambda_{\max}(B) \leqslant \lambda_{\min}(A)\operatorname{tr} B \leqslant \operatorname{tr} AB \leqslant \lambda_{\max}(A)\operatorname{tr} B \leqslant \operatorname{tr} A \operatorname{tr} B.$$

Lemma A.3 [1]. *The solution P of the Lyapunov equation*

$$AP + PA^{\mathrm{T}} + Q = 0$$

with a Hurwitz matrix A and $Q \succ 0$ obey the bounds

$$\lambda_{\max}(P) \geqslant \frac{\lambda_{\min}(Q)}{2\sigma}, \quad \lambda_{\min}(P) \geqslant \frac{\lambda_{\min}(Q)}{2\|A\|},$$

where $\sigma = -\max\limits_i \operatorname{Re}\lambda_i(A)$.

If $Q = DD^{\mathrm{T}}$ and the pair (A, D) is controllable, then

$$\lambda_{\max}(P) \geqslant \frac{\|u^*D\|^2}{2\sigma} > 0,$$

where

$$u^*A = \lambda u^*, \quad \operatorname{Re}\lambda = -\sigma, \quad \|u\| = 1,$$

i.e., u is the left eigenvector of the matrix A corresponding to the eigenvalue λ of the matrix A with the greatest real part. The vector u and the number λ can be complex-valued; here, u^* denotes the Hermitian conjugate.

APPENDIX B

Proof of Lemma 1. Indeed, if the matrix $\mathcal{A}_0 + \{\mathcal{A}, k\}$ is Hurwitz, then $\sigma(\mathcal{A}_0 + \{\mathcal{A}, k\}) > 0$ and there exists the solution $P \succcurlyeq 0$ of the Lyapunov equation (10) for $0 < \alpha < 2\sigma(\mathcal{A}_0 + \{\mathcal{A}, k\})$. Thus, the function $f(k, \alpha) > 0$ is well-defined and $f(k) > 0$ by Theorem 1. The proof of Lemma 1 is complete.

Proof of Lemma 2. The optimization problem has the form

$$\min f(k, \alpha), \quad f(k, \alpha) = \operatorname{tr} P\begin{pmatrix} C & 0 \end{pmatrix}^{\mathrm{T}}\begin{pmatrix} C & 0 \end{pmatrix} + \rho|k|^2$$

subject to the constraint described by the Lyapunov equation

$$\left(\mathcal{A}_0 + \{\mathcal{A}, k\} + \frac{\alpha}{2}I\right)P + P\left(\mathcal{A}_0 + \{\mathcal{A}, k\} + \frac{\alpha}{2}I\right)^{\mathrm{T}} + \frac{1}{\alpha}\begin{pmatrix} D \\ 0 \end{pmatrix}\begin{pmatrix} D \\ 0 \end{pmatrix}^{\mathrm{T}} = 0.$$

To differentiate with respect to k, we add the increment Δk and denote the corresponding increment of P by ΔP:

$$\left(\mathcal{A}_0 + \{\mathcal{A}, k+\Delta k\} + \frac{\alpha}{2}I\right)(P+\Delta P)$$
$$+ (P+\Delta P)\left(\mathcal{A}_0 + \{\mathcal{A}, k+\Delta k\} + \frac{\alpha}{2}I\right)^{\mathrm{T}} + \frac{1}{\alpha}\begin{pmatrix}D\\0\end{pmatrix}\begin{pmatrix}D\\0\end{pmatrix}^{\mathrm{T}} = 0.$$

Let us apply linearization and subtract this and the previous equations to obtain

$$\left(\mathcal{A}_0 + \{\mathcal{A}, k\} + \frac{\alpha}{2}I\right)\Delta P + \Delta P\left(\mathcal{A}_0 + \{\mathcal{A}, k\} + \frac{\alpha}{2}I\right)^{\mathrm{T}} + \{\mathcal{A}, \Delta k\}P + P\{\mathcal{A}, \Delta k\}^{\mathrm{T}} = 0. \tag{B.1}$$

The increment of $f(k)$ is calculated by linearizing the corresponding terms:

$$\begin{aligned}\Delta f(k) =& \operatorname{tr}(P+\Delta P)\begin{pmatrix}C & 0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}C & 0\end{pmatrix} + \rho|k+\Delta k|^2\\ &- \left(\operatorname{tr} P\begin{pmatrix}C & 0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}C & 0\end{pmatrix} + \rho|k|^2\right)\\ =& \operatorname{tr}\Delta P\begin{pmatrix}C & 0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}C & 0\end{pmatrix} + 2\rho k^{\mathrm{T}}\Delta k.\end{aligned}$$

Consider equation (14), dual to (B.1). Due to Lemma A.1, from equations (B.1) and (14) it follows that

$$\Delta f(k) = 2\operatorname{tr} Y\{\mathcal{A}, \Delta k\}P + 2\rho k^{\mathrm{T}}\Delta k.$$

Thus,

$$df(k) = 2\operatorname{tr} PY\sum_{i=1}^{2}\mathcal{A}_i dk_i + 2\rho\sum_{i=1}^{2}k_i dk_i,$$

which leads to (12).

The validity of (13) is demonstrated by analogy with [1, Lemma 1]. The proof of Lemma 2 is complete.

Proof of Lemma 3. The value $(\nabla^2_{kk}f(k)e, e)$ is calculated by differentiating $\nabla_k f(k)$ in the direction $e\in\mathbb{R}^2$. For this purpose, linearizing the corresponding terms and using the convenient notation

$$[\operatorname{tr} PY\mathcal{A}] = \begin{pmatrix}\operatorname{tr} PY\mathcal{A}_1\\ \operatorname{tr} PY\mathcal{A}_2\end{pmatrix},$$

we calculate the increment of $\nabla_k f(k)$ in the direction e:

$$\begin{aligned}\frac{1}{2}\Delta\nabla_k f(k)e &= \rho(k+\delta e) + [\operatorname{tr}(P+\Delta P)(Y+\Delta Y)\mathcal{A}] - (\rho k + [\operatorname{tr} PY\mathcal{A}])\\ &= \rho(k+\delta e) + [\operatorname{tr}\left(P+\delta P'(k)e\right)\left(Y+\delta Y'(k)e\right)\mathcal{A}] - (\rho k + [\operatorname{tr} PY\mathcal{A}])\\ &= \delta\left(\rho e + [\operatorname{tr}\left(PY'(k)e + P'(k)eY\right)\mathcal{A}]\right),\end{aligned}$$

where

$$\Delta P = P(k+\delta e) - P(k) = \delta P'(k)e,$$
$$\Delta Y = Y(k+\delta e) - Y(k) = \delta Y'(k)e.$$

Thus, with $P' = P'(k)e$ and $Y' = Y'(k)e$, we have

$$\frac{1}{2}\left(\nabla^2_{kk} f(k)e, e\right) = \left(\rho e + [\operatorname{tr}(PY' + P'Y)\mathcal{A}], e\right).$$

Furthermore, $P = P(k)$ is the solution of equation (17). We write it in increments in the direction e :

$$\left(\mathcal{A}_0 + \{\mathcal{A}, k + \delta e\} + \frac{\alpha}{2} I\right)(P + \delta P')$$
$$+ (P + \delta P')\left(\mathcal{A}_0 + \{\mathcal{A}, k + \delta e\} + \frac{\alpha}{2} I\right)^{\mathrm{T}} + \frac{1}{\alpha}\begin{pmatrix} D \\ 0 \end{pmatrix}\begin{pmatrix} D \\ 0 \end{pmatrix}^{\mathrm{T}} = 0$$

or

$$\left(\mathcal{A}_0 + \{\mathcal{A}, k\} + \frac{\alpha}{2} I\right)(P + \delta P') + (P + \delta P')\left(\mathcal{A}_0 + \{\mathcal{A}, k\} + \frac{\alpha}{2} I\right)^{\mathrm{T}}$$
$$+ \delta\left(\{\mathcal{A}, e\}P + P\{\mathcal{A}, e\}^{\mathrm{T}}\right) + \frac{1}{\alpha}\begin{pmatrix} D \\ 0 \end{pmatrix}\begin{pmatrix} D \\ 0 \end{pmatrix}^{\mathrm{T}} = 0.$$

Subtracting equation (17) from this expression gives equation (16).

Similarly, $Y = Y(k)$ is the solution of the Lyapunov equation (14). We write it in increments in the direction e :

$$\left(\mathcal{A}_0 + \{\mathcal{A}, k + \delta e\} + \frac{\alpha}{2} I\right)^{\mathrm{T}}(Y + \delta Y')$$
$$+ (Y + \delta Y')\left(\mathcal{A}_0 + \{\mathcal{A}, k + \delta e\} + \frac{\alpha}{2} I\right) + \begin{pmatrix} C & 0 \end{pmatrix}^{\mathrm{T}}\begin{pmatrix} C & 0 \end{pmatrix} = 0,$$

or

$$\left(\mathcal{A}_0 + \{\mathcal{A}, k\} + \frac{\alpha}{2} I\right)^{\mathrm{T}}(Y + \delta Y') + (Y + \delta Y')\left(\mathcal{A}_0 + \{\mathcal{A}, k + \delta e\} + \frac{\alpha}{2} I\right)$$
$$+ \delta\left(\{\mathcal{A}, e\}^{\mathrm{T}} Y + Y\{\mathcal{A}, e\}\right) + \begin{pmatrix} C & 0 \end{pmatrix}^{\mathrm{T}}\begin{pmatrix} C & 0 \end{pmatrix} = 0.$$

Subtracting equation (14) from this expression yields

$$\left(\mathcal{A}_0 + \{\mathcal{A}, k\} + \frac{\alpha}{2} I\right)^{\mathrm{T}} Y' + Y'\left(\mathcal{A}_0 + \{\mathcal{A}, k\} + \frac{\alpha}{2} I\right) + \{\mathcal{A}, e\}^{\mathrm{T}} Y + Y\{\mathcal{A}, e\} = 0. \tag{B.2}$$

From (16) and (B.2) it follows that

$$\operatorname{tr} P'Y\{\mathcal{A}, e\} = \operatorname{tr} PY'\{\mathcal{A}, e\},$$

so

$$\frac{1}{2}\left(\nabla^2_{kk} f(k)e, e\right) = \rho(e, e) + \left([\operatorname{tr}(PY' + P'Y)\mathcal{A}], e\right) = \rho(e, e) + 2\operatorname{tr} P'Y\{\mathcal{A}, e\}.$$

The proof of Lemma 3 is complete.

Proof of Lemma 4. Consider a sequence of stabilizing controllers $\{k_j\} \in \mathcal{S}$ such that $k_j \to k \in \partial\mathcal{S}$, i.e., $\sigma(\mathcal{A}_0 + \{\mathcal{A}, k\}) = 0$. In other words, for any $\varepsilon > 0$ there exists a number $N = N(\varepsilon)$ such that

$$|\sigma(\mathcal{A}_0 + \{\mathcal{A}, k_j\}) - \sigma(\mathcal{A}_0 + \{\mathcal{A}, k\})| = \sigma(\mathcal{A}_0 + \{\mathcal{A}, k_j\}) < \epsilon$$

for all $j \geqslant N(\epsilon)$.

Let P_j be the solution of the Lyapunov equation (10) associated with the controller k_j :

$$\left(\mathcal{A}_0+\{\mathcal{A},k_j\}+\frac{\alpha_j}{2}I\right)P_j+P_j\left(\mathcal{A}_0+\{\mathcal{A},k_j\}+\frac{\alpha_j}{2}I\right)^{\mathrm{T}}+\frac{1}{\alpha_j}\left[\begin{pmatrix}D\\0\end{pmatrix}\begin{pmatrix}D\\0\end{pmatrix}^{\mathrm{T}}+\varepsilon_2 I\right]=0.$$

Also, let Y_j be the solution of the dual Lyapunov equation

$$\left(\mathcal{A}_0+\{\mathcal{A},k_j\}+\frac{\alpha_j}{2}I\right)^{\mathrm{T}}Y_j+Y_j\left(\mathcal{A}_0+\{\mathcal{A},k_j\}+\frac{\alpha_j}{2}I\right)+\begin{pmatrix}C&0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}C&0\end{pmatrix}+\varepsilon_1 I=0.$$

Using Lemma A.3, we have

$$\begin{aligned}f(k_j)&=\operatorname{tr}P_j\left(\begin{pmatrix}C&0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}C&0\end{pmatrix}+\varepsilon_1 I\right)+\rho|k_j|^2\geqslant\operatorname{tr}P_j\left(\begin{pmatrix}C&0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}C&0\end{pmatrix}+\varepsilon_1 I\right)\\&=\operatorname{tr}Y_j\frac{1}{\alpha_j}\left[\begin{pmatrix}D\\0\end{pmatrix}\begin{pmatrix}D\\0\end{pmatrix}^{\mathrm{T}}+\varepsilon_2 I\right]\geqslant\frac{1}{\alpha_j}\lambda_{\min}(Y_j)\operatorname{tr}\left[\begin{pmatrix}D\\0\end{pmatrix}\begin{pmatrix}D\\0\end{pmatrix}^{\mathrm{T}}+\varepsilon_2 I\right]\\&\geqslant\frac{1}{\alpha_j}\lambda_{\min}(Y_j)\left\|\begin{pmatrix}D\\0\end{pmatrix}\right\|_F^2\geqslant\frac{1}{\alpha_j}\frac{\lambda_{\min}\left(\begin{pmatrix}C&0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}C&0\end{pmatrix}+\varepsilon_1 I\right)}{2\left\|\mathcal{A}_0+\{\mathcal{A},k_j\}+\frac{\alpha_j}{2}I\right\|}\|D\|_F^2\\&\geqslant\frac{\varepsilon_1}{4\sigma\left(\mathcal{A}_0+\{\mathcal{A},k_j\}\right)\left\|\mathcal{A}_0+\{\mathcal{A},k_j\}+\frac{\alpha_j}{2}I\right\|}\|D\|_F^2\\&\geqslant\frac{\varepsilon_1}{4\epsilon\left(\|\mathcal{A}_0+\{\mathcal{A},k_j\}\|+\epsilon\right)}\|D\|_F^2\xrightarrow[\epsilon\to0]{}+\infty\end{aligned}$$

since

$$0<\alpha_j<2\sigma\left(\mathcal{A}_0+\{\mathcal{A},k_j\}\right)$$

and

$$\left\|\mathcal{A}_0+\{\mathcal{A},k_j\}+\frac{\alpha_j}{2}I\right\|\leqslant\|\mathcal{A}_0+\{\mathcal{A},k_j\}\|+\frac{\alpha_j}{2}<\|\mathcal{A}_0+\{\mathcal{A},k_j\}\|+\sigma\left(\mathcal{A}_0+\{\mathcal{A},k_j\}\right).$$

On the other hand,

$$f(k_j)=\operatorname{tr}P_j\left(\begin{pmatrix}C&0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}C&0\end{pmatrix}+\varepsilon_1 I\right)+\rho|k_j|^2\geqslant\rho|k_j|^2\xrightarrow[|k_j|\to+\infty]{}+\infty.$$

The proof of Lemma 4 is complete.

Proof of Corollary 2. The function $f(k)$ has a minimum point on the set $\mathcal{S}_0$ (as a continuous function on a compact set), but the set $\mathcal{S}_0$ shares no points with the boundary $\mathcal{S}$ due to (18). Finally, the function $f(k)$ is differentiable on $\mathcal{S}_0$ by Lemma 2, which concludes the proof of Corollary 2.

Proof of Lemma 5. Applying Lemma A.2 to (15) gives

$$\begin{aligned}\frac{1}{2}\|\nabla^2_{kk}f(k)\|&=\frac{1}{2}\sup_{|e|=1}|\left(\nabla^2_{kk}f(k)e,e\right)|\leqslant\sup_{|e|=1}\rho(e,e)+2\sup_{|e|=1}|\operatorname{tr}P'Y\{\mathcal{A},e\}|\\&=\rho+2\sup_{|e|=1}\|P'\|_F\|Y\{\mathcal{A},e\}\|_F\leqslant\rho+2\|P'\|_F\sup_{|e|=1}\|Y\|\|\{\mathcal{A},e\}\|_F\\&\leqslant\rho+2\sqrt{2}\|P'\|_F\|Y\|\max_i\|\mathcal{A}_i\|_F\end{aligned}$$

since

$$\|\{\mathcal{A},e\}\|_F=\left\|\sum_i\mathcal{A}_ie_i\right\|_F\leqslant\sum_i\|\mathcal{A}_i\|_F|e_i|\leqslant\max_i\|\mathcal{A}_i\|_F|e|_1\leqslant\sqrt{2}\max_i\|\mathcal{A}_i\|_F|e|.$$

Thus, it is necessary to estimate from above the value

$$\rho + 2\sqrt{2}\max_i \|\mathcal{A}_i\|_F \|P'\|_F \|Y\|.$$

For $\|Y\|$ we have the upper bound

$$\begin{aligned}\frac{\varepsilon_2}{\alpha}\|Y\| &\leqslant \frac{1}{\alpha}\lambda_{\min}\left[\begin{pmatrix}D\\0\end{pmatrix}\begin{pmatrix}D\\0\end{pmatrix}^{\mathrm{T}} + \varepsilon_2 I\right] \operatorname{tr} Y \leqslant \operatorname{tr} Y \frac{1}{\alpha}\left[\begin{pmatrix}D\\0\end{pmatrix}\begin{pmatrix}D\\0\end{pmatrix}^{\mathrm{T}} + \varepsilon_2 I\right] \\ &= \operatorname{tr} P\left(\begin{pmatrix}C & 0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}C & 0\end{pmatrix} + \varepsilon_1 I\right) = f(k) - \rho|k|^2 \leqslant f(k) \leqslant f(k_0),\end{aligned}$$

and consequently,

$$\|Y\| \leqslant \frac{\alpha}{\varepsilon_2} f(k_0). \tag{B.3}$$

An upper bound for α is established as follows:

$$\begin{aligned}&\alpha < 2\sigma\left(\mathcal{A}_0 + \{\mathcal{A}, k\}\right) \leqslant 2\left\|\mathcal{A}_0 + \{\mathcal{A}, k\}\right\| \\ &\leqslant 2\left(\|\mathcal{A}_0\| + \sum_i \|\mathcal{A}_i\||k_i|\right) \leqslant 2\left(\|\mathcal{A}_0\| + \max_i \|\mathcal{A}_i\||k|_1\right) \\ &\leqslant 2\left(\|\mathcal{A}_0\| + \max_i \|\mathcal{A}_i\|\sqrt{2}|k|\right) \leqslant 2\left(\|\mathcal{A}_0\| + \max_i \|\mathcal{A}_i\|\sqrt{\frac{2}{\rho}f(k)}\right) \\ &\leqslant 2\left(\|\mathcal{A}_0\| + \max_i \|\mathcal{A}_i\|\sqrt{\frac{2}{\rho}f(k_0)}\right),\end{aligned}$$

so

$$\|Y\| \leqslant \frac{2}{\varepsilon_2}\left(\|\mathcal{A}_0\| + \max_i \|\mathcal{A}_i\|\sqrt{\frac{2}{\rho}f(k_0)}\right) f(k_0).$$

Now, let us estimate $\|P\|$ from above:

$$\begin{aligned}&\varepsilon_1\|P\| \leqslant \lambda_{\min}\left(\begin{pmatrix}C & 0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}C & 0\end{pmatrix} + \varepsilon_1 I\right)\|P\| \\ &\leqslant \operatorname{tr} P\left(\begin{pmatrix}C & 0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}C & 0\end{pmatrix} + \varepsilon_1 I\right) = f(k) - \rho|k|^2 \leqslant f(k) \leqslant f(k_0),\end{aligned}$$

which yields

$$\|P\| \leqslant \frac{f(k_0)}{\varepsilon_1}.$$

It remains to estimate from above the value $\|P'\|_F$. In view of Lemma A.2,

$$\begin{aligned}&\lambda_{\max}\left(\{\mathcal{A}, e\}P + P\{\mathcal{A}, e\}^{\mathrm{T}}\right) = \left\|\{\mathcal{A}, e\}P + P\{\mathcal{A}, e\}^{\mathrm{T}}\right\| \leqslant \left\|P^2 + \{\mathcal{A}, e\}\{\mathcal{A}, e\}^{\mathrm{T}}\right\| \\ &\leqslant \|P\|^2 + \|\{\mathcal{A}, e\}\|^2 \leqslant \frac{f^2(k_0)}{\varepsilon_1^2} + 2\max_i \|\mathcal{A}_i\|^2 \leqslant \xi\frac{\varepsilon_2}{\alpha} \leqslant \xi\frac{1}{\alpha}\lambda_{\min}\left[\begin{pmatrix}D\\0\end{pmatrix}\begin{pmatrix}D\\0\end{pmatrix}^{\mathrm{T}} + \varepsilon_2 I\right]\end{aligned}$$

for

$$\xi = \frac{\alpha}{\varepsilon_2}\left(\frac{f^2(k_0)}{\varepsilon_1^2} + 2\max_i \|\mathcal{A}_i\|^2\right).$$

Therefore, the solution P' of the Lyapunov equation (16) satisfies the inequality

$$P' \preccurlyeq \xi P \preccurlyeq \frac{\alpha}{\varepsilon_2}\left(\frac{f^2(k_0)}{\varepsilon_1^2} + 2\max_i \|\mathcal{A}_i\|^2\right)\frac{f(k_0)}{\varepsilon_1} I$$
$$\preccurlyeq \frac{2f(k_0)}{\varepsilon_1\varepsilon_2}\left(\|\mathcal{A}_0\| + \max_i \|\mathcal{A}_i\|\sqrt{\frac{2}{\rho}f(k_0)}\right)\left(\frac{f^2(k_0)}{\varepsilon_1^2} + 2\max_i \|\mathcal{A}_i\|^2\right) I.$$

Hence, it follows that

$$\|P'\|_F \leqslant \frac{2\sqrt{n}f(k_0)}{\varepsilon_1\varepsilon_2}\left(\|\mathcal{A}_0\| + \max_i \|\mathcal{A}_i\|\sqrt{\frac{2}{\rho}f(k_0)}\right)\left(\frac{f^2(k_0)}{\varepsilon_1^2} + 2\max_i \|\mathcal{A}_i\|^2\right). \tag{B.4}$$

Considering the bounds (B.3) and (B.4), we arrive at the relation (19). The proof of Lemma 5 is complete.

FUNDING

This work was supported in part by the Russian Science Foundation, project no. 21-71-30005, `https://rscf.ru/en/project/21-71-30005/`.

REFERENCES

1. Polyak, B.T. and Khlebnikov, M.V., Static Controller Synthesis for Peak-to-Peak Gain Minimization as an Optimization Problem, *Autom. Remote Control*, 2021, vol. 82, no. 9, pp. 1530–1553.
2. Polyak, B.T. and Khlebnikov, M.V., New Criteria for Tuning PID Controllers, *Autom. Remote Control*, 2022, vol. 83, no. 11, pp. 1724–1741.
3. Fatkhullin, I. and Polyak, B., Optimizing Static Linear Feedback: Gradient Method, *SIAM J. Control Optim.*, 2021, vol. 59, no. 5, pp. 3887–3911.
4. Polyak, B.T., Khlebnikov, M.V., and Shcherbakov, P.S., *Upravlenie lineinymi sistemami pri vneshnikh vozmushcheniyakh: Tekhnika lineinykh matrichnykh neravenstv* (Control of Linear Systems Subjected to Exogenous Disturbances: The Technique of Linear Matrix Inequalities), Moscow: LENAND, 2014.
5. Polyak, B.T., Khlebnikov, M.V., and Shcherbakov, P.S., Linear Matrix Inequalities in Control Systems with Uncertainty, *Autom. Remote Control*, 2021, vol. 82, no. 1, pp. 1–40.
6. Boyd, S., El Ghaoui, L., Feron, E., and Balakrishnan, V., *Linear Matrix Inequalities in System and Control Theory*, Philadelphia: SIAM, 1994.
7. Polyak, B., *Introduction to Optimization*, Optimization Software, 1987.
8. Åström, K.J. and Hägglund, T., Benchmark Systems for PID Control, *IFAC Proceedings Volumes*, 2000, vol. 33, iss. 4, pp. 165–166.
9. Grant, M. and Boyd, S., CVX: Matlab Software for Disciplined Convex Programming, version 2.1. URL `http://cvxr.com/cvx`
10. Horn, R.A. and Johnson, Ch.R., *Matrix Analysis*, Cambridge University Press, 2012.

This paper was recommended for publication by L.B. Rapoport, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 8, pp. 918–932.

Design of Suboptimal Robust Controllers Based on A Priori and Experimental Data

M. M. Kogan[*,a] **and A. V. Stepanov**[*,b]

[*] *Nizhny Novgorod State University of Architecture and Civil Engineering, Nizhny Novgorod, Russia*

e-mail: [a] *mkogan@nngasu.ru,* [b] *andrey8st@yahoo.com*

Received March 21, 2023

Revised May 2, 2023

Accepted June 9, 2023

Abstract—This paper develops a novel unified approach to designing suboptimal robust control laws for uncertain objects with different criteria based on a priori information and experimental data. The guaranteed estimates of the γ_0, generalized H_2, and H_∞ norms of a closed loop system and the corresponding suboptimal robust control laws are expressed in terms of solutions of linear matrix inequalities considering a priori knowledge and object modeling data. A numerical example demonstrates the improved quality of control systems when a priori and experimental data are used together.

Keywords: robust control, a priori data, experimental data, γ_0 norm, generalized H_2 norm, H_∞ norm, linear matrix inequalities

DOI: 10.25728/arcRAS.2023.29.47.001

1. INTRODUCTION

In a rich variety of control design approaches for objects with an incomplete mathematical model, there exist two main ones as follows. Within one approach, the controller's parameters are found from a priori information about the possible ranges of the object's uncertain parameters. Following the other approach, the controller's parameters are tuned recursively using current information or are calculated based on experimental data. Traditionally, the former approach is associated with robust control (see [1] and the survey [2]); the latter approach, with adaptive control (see the surveys [3, 4]).

In recent years, researchers have been actively developing the data-driven design of control systems without any explicit mathematical model of the object [5–9]. The paper [10] was pioneering in this area: it was discovered that a single trajectory can be used to fully characterize a linear time-invariant dynamic system under the so-called persistency of excitation. If this condition holds, linear quadratic control of objects without disturbances and without measurement noises can be implemented without knowledge of the object's mathematical model directly from input and output measurement data [5]. According to [6], it suffices to fulfill the less restrictive condition of data informativity for the property of interest in order to construct control laws from experimental data. (Examples of such properties are stabilizability by linear state feedback control or linear quadratic control with a given performance criterion). In [7], the state feedback parameters were derived from open-loop measurements of the input and output of an uncertain object subjected to an unmeasured disturbance from a definite class. For a fully uncertain object, H_2- and H_∞-optimal control laws were constructed based on input and output measurements using a matrix version of S-lemma [11] in the publication [8] and using Petersen's lemma [12] in the publication [9].

This paper develops a novel robust control design approach for uncertain dynamic objects based on the joint use of a priori information about the structure of the uncertain object parameter matrix and the upper bound of its norm (on the one hand) and experimental data obtained by observing the object on some time interval (on the other hand). The quality of robust control is evaluated by upper bounds for one of the three performance indices: the γ_0 norm (the damping level of stochastic disturbances in the closed-loop uncertain system or the maximum value of the quadratic functional of the target output under a pulse disturbance), the generalized H_2 norm (the time-maximum deviation of the Euclidean norm of the system's target output under all deterministic disturbances bounded in the l_2 norm), and the H_∞ norm (the maximum value of the ratio of the l_2 norms of the target output and the exogenous disturbance).

The design procedure includes several basic steps. First, the set of unknown matrices consistent with a priori information is characterized by a quadratic inequality. Then, an experiment is conducted to measure the system trajectory under given initial conditions and controls and an unknown exogenous disturbance with known bounds of its components. This step yields another quadratic inequality satisfied by all unknown matrices consistent with the experimental results. Next, an extended and completely defined system with additional artificial input and output satisfying the two quadratic inequalities is determined; this system "incorporates" the original uncertain system. Finally, upper bounds are found for the damping levels of the disturbances of the original uncertain system as those of the disturbances of the extended system under all additional inputs satisfying the two quadratic inequalities.

This paper is organized as follows. After the Introduction, Section 2 gives the general problem statement; in particular, two quadratic inequalities for the unknown object parameter matrix are derived from a priori information and experimental data. In Section 3, necessary background is provided on the γ_0, generalized H_2, and H_∞ norms as well as their relations in the primal and dual systems. Section 4 describes the robust control design procedure, including the main theorem and its proof. Several experiments with an uncertain third-order system are presented in Section 5; they show the advantages of robust control laws based on a priori information and experimental data over the counterparts designed using a priori information or experimental data only. Section 6 summarizes the results and draws conclusions.

2. ROBUST CONTROL BASED ON A PRIORI AND EXPERIMENTAL DATA: PROBLEM STATEMENT

Consider an uncertain system described by

$$\begin{aligned} x(t+1) &= (A + B_\Delta \Delta C_\Delta)x(t) + (B_u + B_\Delta \Delta D_\Delta)u(t) + Bw(t), \\ z(t) &= Cx(t) + Du(t) \end{aligned} \tag{2.1}$$

with the following notations: $x(t) \in \mathrm{R}^{n_x}$ is the state vector, $z(t) \in \mathrm{R}^{n_z}$ is the target output, $w(t) \in \mathrm{R}^{n_w}$ is an exogenous disturbance, and $u(t) \in \mathrm{R}^{n_u}$ is the control vector (input). All matrices except the unknown parameter matrix Δ are given. In general, it is required to design linear state-feedback control laws based on information about the unknown parameters of the system so that the damping levels of the exogenous disturbances from different classes in the closed loop system do not exceed specified values.

The information about the unknown matrix Δ is divided into a priori one and the one obtained by a preliminary experiment. Assume that the matrix Δ has a block-diagonal structure and

$$\Delta = \operatorname{diag}(\Delta_1, \ldots, \Delta_l) = \sum_{i=1}^{l} L_i \Delta_i R_i^{\mathrm{T}}, \quad \Delta_i \Delta_i^{\mathrm{T}} \leqslant \eta_i^2 I, \tag{2.2}$$

where $\Delta_i \in \mathrm{R}^{m_i \times n_i}$ is a complete matrix block or a diagonal square matrix block $\Delta_i = \delta_i I_{n_i}$; L_i and R_i are matrices composed of unit column vectors corresponding to the location of the ith matrix block such that $L_i^{\mathrm{T}} L_j = 0$ and $R_i^{\mathrm{T}} R_j = 0$, $i \neq j$, and η_i are given values. In accordance with the structure of the matrix Δ, the matrix B_Δ can be written as $B_\Delta = (B_1 \ \dots \ B_l)$, where $B_i = B_\Delta L_i$. With the notation $\widehat{\Delta} = B_\Delta \Delta$, we have

$$\widehat{\Delta} = B_\Delta \sum_{i=1}^{l} L_i \Delta_i R_i^{\mathrm{T}} = \sum_{i=1}^{l} B_i \Delta_i R_i^{\mathrm{T}}. \tag{2.3}$$

Since $\widehat{\Delta} R_j = B_j \Delta_j$, $j = 1, \dots, l$, it follows that $\widehat{\Delta} = \left(\widehat{\Delta}_1 \ \widehat{\Delta}_2 \ \cdots \ \widehat{\Delta}_l\right)$, where $\widehat{\Delta}_i = B_i \Delta_i$.

In particular, if the state and control matrices in the object's equation are completely unknown, then

$$A = 0, \quad B_u = 0, \quad B_\Delta = I, \quad C_\Delta = (I \quad 0)^{\mathrm{T}}, \quad D_\Delta = (0 \quad I)^{\mathrm{T}} \tag{2.4}$$

in (2.1); in this case, $\widehat{\Delta} = \Delta = (A^{(real)} \ B_u^{(real)})$, where $A^{(real)}$ and $B_u^{(real)}$ are the unknown state and control matrices, respectively. This case without using a priori information was studied in the papers [5, 6, 8, 9]).

Next we express the a priori information about the matrix Δ in terms of the matrix $\widehat{\Delta}$. Following the well-known robust control design approach under structured uncertainty [13, 14], let us define the set $\mathbf{\Lambda} = \mathrm{diag}\,(\mathbf{\Lambda_1}, \dots, \mathbf{\Lambda_l})$ consisting of all $\Lambda = \mathrm{diag}\,(\Lambda_1, \dots, \Lambda_l)$ for which $\Lambda_i = \lambda_i I_{n_i}$, $\lambda_i \geqslant 0$, if the matrix block Δ_i is complete and all symmetric nonnegative definite matrices $\Lambda_i \in \mathrm{R}^{n_i \times n_i}$ if $\Delta_i = \delta_i I_{n_i}$. Due to (2.2), the inequality $\lambda_i \Delta_i \Delta_i^{\mathrm{T}} \leqslant \lambda_i \eta_i^2 I$ holds for a complete matrix block $\Delta_i \in \mathrm{R}^{n_i \times n_i}$ for all $\lambda_i \geqslant 0$ and the inequality $\Delta_i \Lambda_i \Delta_i^{\mathrm{T}} \leqslant \eta_i^2 \Lambda_i$ holds for a block $\Delta_i = \delta_i I_{n_i}$ for all symmetric nonnegative definite matrices $\Lambda_i \in \mathrm{R}^{n_i \times n_i}$. Hence, as is easily verified,

$$\Delta \Lambda \Delta^{\mathrm{T}} - \eta \Lambda \eta^{\mathrm{T}} \leqslant 0 \quad \forall\, \Lambda \in \mathbf{\Lambda} \tag{2.5}$$

with $\eta = \mathrm{diag}\,(\eta_1 I_{n_1}, \dots, \eta_l I_{n_l})$ for all the matrices Δ satisfying (2.2).

Multiplying this inequality by the matrices B_Δ and B_Δ^{T} on the left and right, respectively, yields

$$\widehat{\Delta} \Lambda \widehat{\Delta}^{\mathrm{T}} - B_\Delta \eta \Lambda \eta^{\mathrm{T}} B_\Delta^{\mathrm{T}} \leqslant 0 \quad \forall\, \Lambda \in \mathbf{\Lambda}. \tag{2.6}$$

This condition can be written as

$$\left(\widehat{\Delta} \quad I\right) \Upsilon \left(\widehat{\Delta} \quad I\right)^{\mathrm{T}} \leqslant 0 \quad \forall\, \Lambda \in \mathbf{\Lambda}, \tag{2.7}$$

where $\Upsilon = \mathrm{diag}(\Lambda, -B_\Delta \eta \Lambda \eta^{\mathrm{T}} B_\Delta^{\mathrm{T}})$. Let $\mathbf{\Delta}$ denote the set of given-structure matrices Δ satisfying (2.5) and $\widehat{\mathbf{\Delta}}_{\mathbf{a}}$ denote the set of matrices $\widehat{\Delta} = (\widehat{\Delta}_1, \dots, \widehat{\Delta}_l)$ satisfying inequality (2.6). Clearly, for any $\Delta \in \mathbf{\Delta}$ there exists $\widehat{\Delta} = B_\Delta \Delta \in \widehat{\mathbf{\Delta}}_{\mathbf{a}}$. The converse is also true as follows.

Lemma 2.1. *If matrices $B_i = B_\Delta L_i$, $i = 1, \dots, l$, have full column rank, then for any $\widehat{\Delta} \in \widehat{\mathbf{\Delta}}_{\mathbf{a}}$ there exists $\Delta \in \mathbf{\Delta}$ such that $\widehat{\Delta} = B_\Delta \Delta$.*

Proof of Lemma. Assume that $\widehat{\Delta} \in \widehat{\mathbf{\Delta}}_{\mathbf{a}}$. Due to (2.6), we have $\widehat{\Delta}^{\mathrm{T}} a = 0$ for any vector $a \neq 0$ with $B_\Delta^{\mathrm{T}} a = 0$. This means that the columns of the matrix $\widehat{\Delta}$ belong to the image of the matrix B_Δ. Hence, the linear matrix equation $B_\Delta \Delta = \widehat{\Delta}$ is solvable in the matrix Δ. It remains to show inequality (2.5) for this solution. From (2.6) it follows that

$$B_i (\Delta_i \Lambda_i \Delta_i^{\mathrm{T}} - \eta_i^2 \Lambda_i) B_i^{\mathrm{T}} \leqslant 0$$

for each block. Since the matrices B_i have full column rank, $\Delta_i \Lambda_i \Delta_i^{\mathrm{T}} - \eta_i^2 \Lambda_i \leqslant 0$ holds for all i, i.e., $\Delta \in \mathbf{\Delta}$, and the desired result is established.

According to this lemma, there is no loss of information when passing from the matrix Δ that satisfies inequality (2.5) to the matrix $\widehat{\Delta}$ that satisfies inequality (2.7). In view of this fact, we write the original uncertain system (2.1) as

$$\begin{aligned} x(t+1) &= (A+\widehat{\Delta}C_\Delta)x(t)+(B_u+\widehat{\Delta}D_\Delta)u(t)+Bw(t),\\ z(t) &= Cx(t)+Du(t), \end{aligned} \tag{2.8}$$

where the unknown parameter matrix $\widehat{\Delta}=(\widehat{\Delta}_1,\dots,\widehat{\Delta}_l)$ of the corresponding structure satisfies inequality (2.7).

Additional information about the unknown parameters of system (2.8) is extracted from a finite set of its trajectory measurements. More precisely put, it is possible to measure the system states $x_0,x_1,\dots,x_N$ under given controls $u_0,\dots,u_{N-1}$ and some unknown disturbance $w(t)$ whose components satisfy the constraint

$$|w_i(t)|\leqslant d,\quad t=0,\dots,N-1,\quad i=1,\dots,n_w \tag{2.9}$$

for some given d (the disturbance level), i.e., $\max_{0\leqslant t\leqslant N-1}\|w(t)\|_\infty\leqslant d$. Following conventional notations (e.g., see [6]), we compile the matrices

$$\begin{aligned} \Phi &= (x_0 \quad x_1 \quad \cdots \quad x_{N-1}), \quad \Phi_+ = (x_1 \quad x_2 \quad \cdots \quad x_N),\\ W &= (w_0 \quad w_1 \quad \cdots \quad w_{N-1}), \quad U = (u_0 \quad u_1 \quad \cdots \quad u_{N-1}) \end{aligned}$$

and introduce

$$C_\Delta\Phi+D_\Delta U=\widehat{\Phi}.$$

Due to the object's equation,

$$\widetilde{\Phi}=\widehat{\Delta}^{(real)}\widehat{\Phi}+BW, \tag{2.10}$$

where $\widetilde{\Phi}=\Phi_+-A\Phi-B_uU$ and $\widehat{\Delta}^{(real)}$ is the real unknown parameter matrix of the object (2.8). According to (2.9) and (2.10),

$$(\widetilde{\Phi}-\widehat{\Delta}\widehat{\Phi})(\widetilde{\Phi}-\widehat{\Delta}\widehat{\Phi})^{\mathrm{T}}=BWW^{\mathrm{T}}B^{\mathrm{T}}\leqslant d^2n_wNBB^{\mathrm{T}}$$

for $\widehat{\Delta}=\widehat{\Delta}^{(real)}$.

Let $\widehat{\mathbf{\Delta}}_{\mathbf{p}}$ denote the set of given-structure matrices $\widehat{\Delta}$ satisfying this inequality. Obviously, $\widehat{\Delta}^{(real)}\in\widehat{\mathbf{\Delta}}_{\mathbf{p}}$. Introducing the matrix

$$\Psi=\begin{pmatrix}\Psi_{11} & *\\ \Psi_{12}^{\mathrm{T}} & \Psi_{22}\end{pmatrix}=\begin{pmatrix}\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}} & *\\ -\widetilde{\Phi}\widehat{\Phi}^{\mathrm{T}} & \widetilde{\Phi}\widetilde{\Phi}^{\mathrm{T}}-d^2n_wNBB^{\mathrm{T}}\end{pmatrix}, \tag{2.11}$$

we write this inequality as

$$\begin{pmatrix}\widehat{\Delta} & I\end{pmatrix}\Psi\begin{pmatrix}\widehat{\Delta} & I\end{pmatrix}^{\mathrm{T}}\leqslant 0. \tag{2.12}$$

Let $\widehat{\mathbf{\Delta}}=\widehat{\mathbf{\Delta}}_{\mathbf{a}}\bigcap\widehat{\mathbf{\Delta}}_{\mathbf{p}}$ denote the set of matrices $\widehat{\Delta}$ satisfying the constraints (2.7) and (2.12).

The quality of the closed-loop uncertain system (2.8) with a linear state-feedback control law will be evaluated by its response to stochastic and deterministic disturbances under zero initial state, measured by three performance indices: the guaranteed estimates of the γ_0, generalized H_2, and H_∞ norms. The guaranteed estimate of the γ_0 norm is defined as the damping level of a stochastic disturbance from the class $\mathcal{G}_{n_w}$ of vector Gaussian white noises of dimension n_w, equal

to the maximum value of the square root of the ratio of the steady-state time-averaged variances of the output z and input w under all nonzero covariance matrices K_w of the input [15]:

$$\gamma_0 = \sup_{\widehat{\Delta} \in \widehat{\mathbf{\Delta}}} \gamma_0(\widehat{\Delta}), \quad \gamma_0(\widehat{\Delta}) = \operatorname{ess} \sup_{w \in \mathcal{G}_{n_w}} \frac{\|z\|_{\mathcal{P}}}{\|w\|_{\mathcal{P}}},$$

where $\|s\|_{\mathcal{P}}^2 = \lim\limits_{N\to\infty} (1/N) \sum_{t=0}^{N-1} |s(t)|^2$ and ess stands for essential supremum (the least upper bound with probability 1). The guaranteed estimates of the generalized H_2 and H_∞ norms characterize, respectively, the relative maximum values of the time-maximal deviation and the quadratic functional of the target output under deterministic disturbances from the class l_2. They are defined as

$$\gamma_{g2} = \sup_{\widehat{\Delta} \in \widehat{\mathbf{\Delta}}} \gamma_{g2}(\widehat{\Delta}), \quad \gamma_{g2}(\widehat{\Delta}) = \sup_{w(t)\neq 0} \frac{\sup_{t\geqslant 0} |z(t)|}{\|w\|},$$

$$\gamma_\infty = \sup_{\widehat{\Delta} \in \widehat{\mathbf{\Delta}}} \gamma_\infty(\widehat{\Delta}), \quad \gamma_\infty(\widehat{\Delta}) = \sup_{w(t)\neq 0} \frac{\|z\|}{\|w\|},$$

where $\|s\|^2 = \sum_{t=0}^{\infty} |s(t)|^2$. The problem is to obtain upper bounds for these norms and finally design control laws ensuring the required system quality estimates.

3. NECESSARY BACKGROUND ON THE γ_0, GENERALIZED H_2, AND H_∞ NORMS

Before deriving the guaranteed estimates of the above norms, we clarify the calculation of the norms $\gamma_0(\widehat{\Delta})$, $\gamma_{g2}(\widehat{\Delta})$, and $\gamma_\infty(\widehat{\Delta})$ for the closed loop system (2.8), $u(t) = \Theta x(t)$ under a fixed matrix $\widehat{\Delta}$ given by the equations

$$\begin{aligned} x(t+1) &= [A + B_u\Theta + \widehat{\Delta}(C_\Delta + D_\Delta\Theta)]x(t) + Bw(t), \\ z(t) &= (C + D\Theta)x(t). \end{aligned} \tag{3.1}$$

With the notations

$$A_\Theta = A + B_u\Theta, \quad C_{\Delta\Theta} = C_\Delta + D_\Delta\Theta, \quad A_\Delta = A_\Theta + \widehat{\Delta}C_{\Delta\Theta}, \quad C_\Theta = C + D\Theta,$$

these equations can be written as

$$\begin{aligned} x(t+1) &= A_\Delta x(t) + Bw(t), \\ z(t) &= C_\Theta x(t). \end{aligned} \tag{3.2}$$

The damping level of the stochastic disturbance, i.e., the γ_0 norm of this system, is found by solving a semidefinite programming problem in the covariance matrices $K_w = K_w^{\mathrm{T}} \geqslant 0$ (the disturbance) and $K_x = K_x^{\mathrm{T}} \geqslant 0$ (the state) [15]:

$$\gamma_0^2(\widehat{\Delta}) = \max \operatorname{tr} C_\Theta K_x C_\Theta^{\mathrm{T}} : \quad A_\Delta K_x A_\Delta^{\mathrm{T}} - K_x + BK_wB^{\mathrm{T}} = 0, \quad \operatorname{tr} K_w \leqslant 1. \tag{3.3}$$

Here, we need the following auxiliary result, proved in the Appendix.

Lemma 3.1. *Problem* (3.3) *is Lagrange dual to the problem*

$$\gamma_0^2(\widehat{\Delta}) = \min \gamma^2 : \quad A_\Delta^{\mathrm{T}} P A_\Delta - P + C_\Theta^{\mathrm{T}} C_\Theta \leqslant 0, \quad B^{\mathrm{T}} P B \leqslant \gamma^2 I. \tag{3.4}$$

According to problem (3.4), the increment of the function $V(x) = x^{\mathrm{T}}Px$ along the trajectories of (3.2) with the initial disturbance $w(0) = w_0$, $w(t) \equiv 0$, $t > 0$, and zero initial conditions satisfies the inequalities

$$\triangle V + |z|^2 \leqslant 0,\ t \geqslant 1, \quad V(x_1) = w_0^{\mathrm{T}} B^{\mathrm{T}} P B w_0 \leqslant \gamma^2 |w_0|^2 \quad \forall x \in \mathrm{R}^{n_x}, \forall\, w_0 \in \mathrm{R}^{n_w}. \tag{3.5}$$

In other words, the damping level of stochastic disturbances coincides with that of a deterministic initial disturbance, understood as the maximum value of the ratio of the l_2 norm of the output under the "pulse" disturbance $w(0) = w_0$, $w(t) \equiv 0$, $t \geqslant 1$, and zero initial conditions to the Euclidean norm of the disturbance:

$$\gamma_0^2(\widehat{\Delta}) = \max_{w_0 \neq 0} \frac{\|z\|^2}{|w_0|^2}.$$

The next characteristic—the maximum deviation of the output (the generalized H_2 norm [16, 17])—is found by solving the problem

$$\gamma_{g2}^2(\widehat{\Delta}) = \min \gamma^2 : \quad A_\Delta Q A_\Delta^{\mathrm{T}} - Q + BB^{\mathrm{T}} \leqslant 0, \quad C_\Theta Q C_\Theta^{\mathrm{T}} \leqslant \gamma^2 I. \tag{3.6}$$

With the change of variables $P = Q^{-1}$, it can be written as

$$\gamma_{g2}^2(\widehat{\Delta}) = \min \gamma^2 : \quad \begin{pmatrix} A_\Delta^{\mathrm{T}} P A_\Delta - P & * \\ B^{\mathrm{T}} P A_\Delta & B^{\mathrm{T}} P B - I \end{pmatrix} \leqslant 0, \quad \begin{pmatrix} P & * \\ C_\Theta & \gamma^2 I \end{pmatrix} \geqslant 0.$$

This means that the increment of the function $V(x) = x^{\mathrm{T}}Px$ along the trajectories of (3.2) with zero initial conditions satisfies the inequality

$$\triangle V - |w|^2 \leqslant 0, \quad \forall\, x \in \mathrm{R}^{n_x}, \quad \forall\, w \in \mathrm{R}^{n_w}, \quad P \geqslant \gamma^{-2} C_\Theta^{\mathrm{T}} C_\Theta. \tag{3.7}$$

As is well known, the system under consideration has the H_∞ norm below γ if and only if the linear matrix inequality (LMI)

$$\begin{pmatrix} A_\Delta^{\mathrm{T}} P A_\Delta - P & * & * \\ B^{\mathrm{T}} P A_\Delta & B^{\mathrm{T}} P B - \gamma^2 I & * \\ C_\Theta & 0 & -I \end{pmatrix} < 0 \tag{3.8}$$

is solvable in the matrix $P = P^{\mathrm{T}} > 0$. According to this LMI, the increment of the positive definite function $V(x) = x^{\mathrm{T}}Px$ along the trajectories of (3.2) satisfies the inequality

$$\triangle V + |z|^2 - \gamma^2 |w|^2 < 0 \tag{3.9}$$

for all x and w.

Direct comparison of problems (3.4) and (3.6) shows that the γ_0 and generalized H_2 norms of system (3.1), respectively, coincide with the generalized H_2 and γ_0 norms of the dual system

$$\begin{aligned} &\widehat{x}(t+1) = (A_\Theta + \widehat{\Delta} C_{\Delta\Theta})^{\mathrm{T}} \widehat{x}(t) + C_\Theta^{\mathrm{T}} \widehat{w}(t), \quad \widehat{x}(0) = 0, \\ &\widehat{z}(t) = B^{\mathrm{T}} \widehat{x}(t). \end{aligned} \tag{3.10}$$

In addition, the dual systems (3.1) and (3.10) obviously have the same H_∞ norm.

4. ROBUST γ_0-, GENERALIZED H_2-, AND H_∞-SUBOPTIMAL CONTROL LAWS

Now we present the main steps to obtain the guaranteed estimates of the γ_0, γ_{g2}, and γ_∞ norms of the uncertain system (3.1) and to find the parameters of the corresponding suboptimal robust control laws. Let $\widehat{\gamma}_0$, $\widehat{\gamma}_{g2}$, and $\widehat{\gamma}_\infty$ denote the corresponding guaranteed estimates of the norms of the dual system (3.10). According to the previous section,

$$\gamma_0 = \widehat{\gamma}_{g2}, \quad \gamma_{g2} = \widehat{\gamma}_0, \quad \gamma_\infty = \widehat{\gamma}_\infty.$$

Consider an extended system with the additional artificial input $w_\Delta(t)$ and output $z_\Delta(t)$ described by

$$\begin{aligned} &x_a(t+1) = A_\Theta^{\mathrm{T}} x_a(t) + C_{\Delta\Theta}^{\mathrm{T}} w_\Delta(t) + C_\Theta^{\mathrm{T}} w_a(t), \quad x_a(0) = 0, \\ &z_a(t) = B^{\mathrm{T}} x_a(t), \quad z_\Delta(t) = x_a(t), \end{aligned} \tag{4.1}$$

where $x_a(t)$, $w_a(t)$, and $z_a(t)$ are the state, disturbance, and target output, respectively. Suppose that for all $t \geqslant 0$, the additional input signal $w_\Delta(t)$ in system (4.1) satisfies the two inequalities

$$\begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix}^{\mathrm{T}} \Psi \begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix} \leqslant 0, \quad \begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix}^{\mathrm{T}} \Upsilon \begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix} \leqslant 0, \tag{4.2}$$

where the matrices Ψ and Υ are given by (2.11) and (2.7). The set of all such signals will be denoted by $\mathbf{W}_{\boldsymbol{\Delta}}$. System (3.10) is "immersed" in system (4.1), (4.2): for $w_\Delta(t) = \widehat{\Delta}^{\mathrm{T}} z_\Delta(t)$, equations (4.1) turn into equations (3.10); as follows from (2.12) and (2.7), for all $\widehat{\Delta} \in \widehat{\boldsymbol{\Delta}}$ we have

$$\begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix}^{\mathrm{T}} \Psi \begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix} = z_\Delta^{\mathrm{T}}(t) \begin{pmatrix} \widehat{\Delta}^{\mathrm{T}} \\ I \end{pmatrix}^{\mathrm{T}} \Psi \begin{pmatrix} \widehat{\Delta}^{\mathrm{T}} \\ I \end{pmatrix} z_\Delta(t) \leqslant 0,$$

$$\begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix}^{\mathrm{T}} \Upsilon \begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix} = z_\Delta^{\mathrm{T}}(t) \begin{pmatrix} \widehat{\Delta}^{\mathrm{T}} \\ I \end{pmatrix}^{\mathrm{T}} \Upsilon \begin{pmatrix} \widehat{\Delta}^{\mathrm{T}} \\ I \end{pmatrix} z_\Delta(t) \leqslant 0,$$

i.e., $w_\Delta(t) = \widehat{\Delta}^{\mathrm{T}} z_\Delta(t) \in \mathbf{W}_{\boldsymbol{\Delta}}$.

For the extended system (4.1), (4.2), we define the γ_0, generalized H_2, and H_∞ norms with respect to the input w_a and output z_a under all admissible inputs w_Δ as

$$\begin{aligned} \widetilde{\gamma}_0 &= \sup_{w_\Delta(t) \,\in\, \mathbf{W}_{\boldsymbol{\Delta}}} \operatorname{ess} \sup_{w_a \,\in\, \mathcal{G}_{n_w}} \frac{\|z_a\|_{\mathcal{P}}}{\|w_a\|_{\mathcal{P}}}, \\ \widetilde{\gamma}_{g2} &= \sup_{w_\Delta(t) \,\in\, \mathbf{W}_{\boldsymbol{\Delta}}} \sup_{w_a(t) \neq 0} \frac{\sup_{t \geqslant 0} |z_a(t)|}{\|w_a\|}, \\ \widetilde{\gamma}_\infty &= \sup_{w_\Delta(t) \,\in\, \mathbf{W}_{\boldsymbol{\Delta}}} \sup_{w_a(t) \neq 0} \frac{\|z_a\|}{\|w_a\|}. \end{aligned} \tag{4.3}$$

They obviously restrict from above the guaranteed estimates of the corresponding norms of system (3.10). In view of the relations between the norms of dual systems (see above), the guaranteed estimates of the norms of the original uncertain system (3.1) satisfy the inequalities

$$\gamma_0 \leqslant \widetilde{\gamma}_{g2}, \quad \gamma_{g2} \leqslant \widetilde{\gamma}_0, \quad \gamma_\infty \leqslant \widetilde{\gamma}_\infty.$$

The performance indices (4.3) will be below a given value γ if there exists a positive definite quadratic function $V(x_a) = x_a^{\mathrm{T}} P x_a$ whose increment along the trajectories of (4.1) satisfies the

following conditions for each norm (similar to conditions (3.5), (3.7), and (3.9) for system (3.2)):

$$(A_\Theta^{\mathrm{T}}x_a + C_{\Delta\Theta}^{\mathrm{T}}w_\Delta)^{\mathrm{T}}P(A_\Theta^{\mathrm{T}}x_a + C_{\Delta\Theta}^{\mathrm{T}}w_\Delta) - x_a^{\mathrm{T}}Px_a + |z_a|^2 \leqslant 0, \quad C_\Theta PC_\Theta^{\mathrm{T}} < \gamma^2 I;$$

$$(A_\Theta^{\mathrm{T}}x_a + C_{\Delta\Theta}^{\mathrm{T}}w_\Delta + C_\Theta^{\mathrm{T}}w_a)^{\mathrm{T}}P(A_\Theta^{\mathrm{T}}x_a + C_{\Delta\Theta}^{\mathrm{T}}w_\Delta + C_\Theta^{\mathrm{T}}w_a) - x_a^{\mathrm{T}}Px_a - |w_a|^2 \leqslant 0,$$

$$\begin{pmatrix} P & * \\ B^{\mathrm{T}} & \gamma^2 I \end{pmatrix} > 0;$$

$$(A_\Theta^{\mathrm{T}}x_a + C_{\Delta\Theta}^{\mathrm{T}}w_\Delta + C_\Theta^{\mathrm{T}}w_a)^{\mathrm{T}}P(A_\Theta^{\mathrm{T}}x_a + C_{\Delta\Theta}^{\mathrm{T}}w_\Delta + C_\Theta^{\mathrm{T}}w_a) - x_a^{\mathrm{T}}Px_a + |z_a|^2 - \gamma^2|w_a|^2 < 0$$

for all x_a, w_a, and all $w_\Delta \in \mathbf{W_\Delta}$, i.e., those obeying the constraints (4.2). A sufficient condition for this is the existence of a matrix $P = P^{\mathrm{T}} > 0$ and nonnegative numbers $\mu \geqslant 0$ and $\nu \geqslant 0$ such that, for all x_a, w_a, and w_Δ,

$$\triangle V + |z_a|^2 - \begin{pmatrix} w_\Delta \\ z_\Delta \end{pmatrix}^{\mathrm{T}} (\mu\Psi + \nu\Upsilon) \begin{pmatrix} w_\Delta \\ z_\Delta \end{pmatrix} \leqslant 0, \quad C_\Theta PC_\Theta^{\mathrm{T}} < \gamma^2 I;$$

$$\triangle V - |w_a|^2 - \begin{pmatrix} w_\Delta \\ z_\Delta \end{pmatrix}^{\mathrm{T}} (\mu\Psi + \nu\Upsilon) \begin{pmatrix} w_\Delta \\ z_\Delta \end{pmatrix} \leqslant 0, \quad \begin{pmatrix} P & * \\ B^{\mathrm{T}} & \gamma^2 I \end{pmatrix} > 0;$$

$$\triangle V + |z_a|^2 - \gamma^2|w_a|^2 - \begin{pmatrix} w_\Delta \\ z_\Delta \end{pmatrix}^{\mathrm{T}} (\mu\Psi + \nu\Upsilon) \begin{pmatrix} w_\Delta \\ z_\Delta \end{pmatrix} < 0,$$

where the increment of the function $V(x)$ in the first inequality is taken along the trajectory of system (4.1) with $w_a(t) \equiv 0$. We write these inequalities in matrix form and introduce the new matrix variable $Z = \Theta P$. Then replacing the matrix $\nu\Lambda$ with the matrix Λ without notational change and applying Schur's complement lemma lead to an important result.

Theorem 4.1. *The guaranteed estimates of the γ_0, generalized H_2, and H_∞ norms of the uncertain system* (2.1), (2.2) *with the control law $u(t) = \Theta x(t)$, where $\Theta = ZP^{-1}$, are below γ if the following LMIs are solvable in $P > 0$, Z, $\Lambda \in \mathbf{\Lambda}$, and $\mu \geqslant 0$:*

$$\begin{pmatrix} -P & * & * & * & * \\ \mathcal{F}_A & -P - \mu\Psi_{22} & * & * & * \\ \mathcal{F}_{C_\Delta} & -\mu\Psi_{12} & -\mu\Psi_{11} - \Lambda & * & * \\ \mathcal{F}_C & 0 & 0 & -I & * \\ 0 & \Lambda\eta^{\mathrm{T}}B_\Delta^{\mathrm{T}} & 0 & 0 & -\Lambda \end{pmatrix} \leqslant 0, \qquad \begin{pmatrix} P & * \\ B^{\mathrm{T}} & \gamma^2 I \end{pmatrix} > 0; \tag{4.4}$$

$$\begin{pmatrix} -P & * & * & * & * \\ \mathcal{F}_A & -P - \mu\Psi_{22} & * & * & * \\ 0 & B^{\mathrm{T}} & -I & * & \\ \mathcal{F}_{C_\Delta} & -\mu\Psi_{12} & 0 & -\mu\Psi_{11} - \Lambda & * \\ 0 & \Lambda\eta^{\mathrm{T}}B_\Delta^{\mathrm{T}} & 0 & 0 & -\Lambda \end{pmatrix} \leqslant 0, \qquad \begin{pmatrix} P & * \\ \mathcal{F}_C & \gamma^2 I \end{pmatrix} > 0 \tag{4.5}$$

and

$$\begin{pmatrix} -P & * & * & * & * & * \\ \mathcal{F}_A & -P-\mu\Psi_{22} & * & * & * & * \\ 0 & B^{\mathrm{T}} & -I & * & * & * \\ \mathcal{F}_{C_\Delta} & -\mu\Psi_{12} & 0 & -\mu\Psi_{11}-\Lambda & * & * \\ \mathcal{F}_C & 0 & 0 & 0 & -\gamma^2 I & * \\ 0 & \Lambda\eta^{\mathrm{T}}B_\Delta^{\mathrm{T}} & 0 & 0 & 0 & -\Lambda \end{pmatrix} < 0, \tag{4.6}$$

where $\mathcal{F}_A = AP + B_u Z$, $\mathcal{F}_C = CP + DZ$, $\mathcal{F}_{C_\Delta} = C_\Delta P + D_\Delta Z$, *the elements of the matrices* Ψ *are given by* (2.11), *and the matrix* $\eta = \operatorname{diag}(\eta_1 I_{n_1}, \ldots, \eta_l I_{n_l})$ *is given by* (2.2).

The minimum values of γ^2 obtained using this theorem will be denoted by $\gamma^2(\widehat{\Delta}, \Theta)$, where the arguments are the corresponding system parameter matrix ($\widehat{\Delta}$ for the uncertain system and $\widehat{\Delta}^{(real)}$ for the real system) and the corresponding feedback parameter matrix ($\Theta^{(ab)}$ for the robust control law based on a priori and experimental data, $\Theta^{(a)}$ for the robust control law based on a priori data only, and $\Theta^{(b)}$ for the robust control law based on experimental data only). If only a priori data are used, then the guaranteed estimates of the norms $\gamma^2(\widehat{\Delta}, \Theta^{(a)})$ are found by solving these inequalities with $\mu = 0$; if only experimental data are used, then $\gamma^2(\widehat{\Delta}, \Theta^{(b)})$ are found by solving these inequalities with $\Lambda = 0$. It is clear that $\gamma^2(\widehat{\Delta}, \Theta^{(ab)}) \leqslant \min\{\gamma^2(\widehat{\Delta}, \Theta^{(a)}), \gamma^2(\widehat{\Delta}, \Theta^{(b)})\}$.

In the case of completely unknown state and control matrices of the system with the matrices of equation (2.1) given by (2.4) and $\Delta\Delta^{\mathrm{T}} \leqslant \eta^2 I$, Theorem 4.1 provides the guaranteed estimates of the corresponding norms for $\mathbf{\Lambda} = \{\lambda I : \lambda \geqslant 0\}$.

The inequalities in Theorem 4.1 serve for calculating the parameters of control laws and the norms in different scenarios by choosing appropriate blocks $\mathcal{F}_A$, $\mathcal{F}_C$, and $\mathcal{F}_{C_\Delta}$ and variables Λ and μ. In the next section, some of these scenarios will be implemented for an illustrative example. Also, the corresponding blocks $\mathcal{F}_A$, $\mathcal{F}_C$, and $\mathcal{F}_{C_\Delta}$ and variables Λ and μ in inequalities (4.4)–(4.6) will be presented.

5. ILLUSTRATIVE EXAMPLE

Consider the results of several experiments with one system of the form (2.1):

$$x(t+1) = \begin{pmatrix} 0.3 & 0.8 & -0.3 \\ -0.2+\delta & 0.6+\Delta_{11} & -0.1+\Delta_{12} \\ 0.5 & -0.2+\Delta_{21} & 0.9+\Delta_{22} \end{pmatrix} x(t) + \begin{pmatrix} 0.2 \\ 1+\delta \\ 0.5 \end{pmatrix} u(t) + w(t),$$

$$z(t) = \begin{pmatrix} I_3 \\ 0 \end{pmatrix} x(t) + \begin{pmatrix} 0_{3\times 1} \\ 0.2 \end{pmatrix} u(t),$$

where

$$B_\Delta = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad C_\Delta = I_3, \quad D_\Delta = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \Delta = \begin{pmatrix} \Delta_1 & 0 \\ 0 & \Delta_2 \end{pmatrix},$$

$$\Delta_1 = \delta, \quad \Delta_2 = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix}, \quad |\delta| \leqslant 0.12; \quad \Delta_2\Delta_2^{\mathrm{T}} \leqslant 0.19.$$

1. Based on a priori information only, we calculate the guaranteed estimates of the norms and parameter matrix of the corresponding robust control laws using the formula $\Theta^{(a)} = ZP^{-1}$ by

solving inequalities (4.4)–(4.6) with $\mathcal{F}_A = AP + B_u Z$, $\mathcal{F}_C = CP + DZ$, $\mathcal{F}_{C_\Delta} = C_\Delta P + D_\Delta Z$, $\eta = \operatorname{diag}(0.12; 0.19 I_2)$, $\mu = 0$, and the unknown variable $\Lambda \geqslant 0$:

$$\gamma_0^2(\widehat{\Delta}, \Theta_0^{(a)}) = 12.8095; \quad \Theta_0^{(a)} = (-0.4356;\ -0.6420;\ -0.3125),$$
$$\gamma_{g2}^2(\widehat{\Delta}, \Theta_{g2}^{(a)}) = 10.5935; \quad \Theta_{g2}^{(a)} = (-0.8498;\ -0.7996;\ -0.6503),$$
$$\gamma_\infty^2(\widehat{\Delta}, \Theta_\infty^{(a)}) = 49.2653; \quad \Theta_\infty^{(a)} = (-1.2373;\ -0.8204;\ -0.9710).$$

Suppose that the real system is described by the uncertain parameters $\delta^{(real)} = -0.05$, $\Delta_{11}^{(real)} = 0.2$, $\Delta_{12}^{(real)} = \Delta_{21}^{(real)} = 0$, and $\Delta_{22}^{(real)} = -0.1$ so that

$$\widehat{\Delta} = \widehat{\Delta}^{(real)} = \begin{pmatrix} 0 & 0 & 0 \\ -0.05 & 0.2 & 0 \\ 0 & 0 & -0.1 \end{pmatrix}, \tag{5.1}$$

whereas the state and control matrices of the real object are

$$A^{(real)} = A + \widehat{\Delta}^{(real)} C_\Delta, \quad B_u^{(real)} = B_u + \widehat{\Delta}^{(real)} D_\Delta.$$

Let us calculate the three norms of the closed loop system (the real object with the robust feedback control with the parameter matrix $\Theta^{(a)}$) by solving inequalities (4.4)–(4.6) with

$$\mathcal{F}_A = (A^{(real)} + B_u^{(real)} \Theta^{(a)}) P, \quad \mathcal{F}_C = (C + D\Theta^{(a)}) P, \quad \mathcal{F}_{C_\Delta} = 0,$$

$\Lambda = 0$, and $\mu = 0$:

$$\gamma_0^2(\widehat{\Delta}^{(real)}, \Theta_0^{(a)}) = 4.8319;$$
$$\gamma_{g2}^2(\widehat{\Delta}^{(real)}, \Theta_{g2}^{(a)}) = 5.1373;$$
$$\gamma_\infty^2(\widehat{\Delta}^{(real)}, \Theta_\infty^{(a)}) = 23.5459.$$

For comparison, here are the optimal values of these norms and parameter matrices of the optimal feedback control laws for the real system (if it were known), calculated using the formula $\Theta^{(real)} = ZP^{-1}$ by solving inequalities (4.4)–(4.6) with $\mathcal{F}_A = A^{(real)} P + B_u^{(real)} Z$, $\mathcal{F}_C = CP + DZ$, $\mathcal{F}_{C_\Delta} = 0$, $\Lambda = 0$, and $\mu = 0$:

$$\gamma_0^2(\widehat{\Delta}^{(real)}, \Theta_0^{(real)}) = 3.9569; \quad \Theta_0^{(real)} = (-0.0765;\ -0.9379;\ 0.0064),$$
$$\gamma_{g2}^2(\widehat{\Delta}^{(real)}, \Theta_{g2}^{(real)}) = 4.4024; \quad \Theta_{g2}^{(real)} = (-0.1369;\ -0.9249;\ -0.0741),$$
$$\gamma_\infty^2(\widehat{\Delta}^{(real)}, \Theta_\infty^{(real)}) = 10.4651; \quad \Theta_\infty^{(real)} = (-1.2547;\ -1.3605;\ -0.3919).$$

2. Consider the case of no a priori information about the possible range of unknown parameters of the object: experimental data are used instead. We calculate the guaranteed estimates of the norms and find the parameter matrices of the suboptimal robust feedback control laws using the formula $\Theta^{(b)} = ZP^{-1}$ by solving inequalities (4.4)–(4.6) with $\mathcal{F}_A = AP + B_u Z$, $\mathcal{F}_C = CP + DZ$, $\mathcal{F}_{C_\Delta} = C_\Delta P + D_\Delta Z$, $\Lambda = 0$, and the unknown variable $\mu \geqslant 0$. To obtain experimental data, we model equation (2.8) with the initial conditions $x_0 = (9; 5; -7)^{\mathrm{T}}$ under the uncertainties $\delta^{(real)} = -0.05$, $\Delta_{11}^{(real)} = 0.2$, $\Delta_{12}^{(real)} = \Delta_{21}^{(real)} = 0$, and $\Delta_{22}^{(real)} = -0.1$ so that $\widehat{\Delta} = \widehat{\Delta}^{(real)}$. The components of the control $u(t)$ and disturbance $w(t)$ vectors in the experiment are chosen as random variables with the uniform distribution on the intervals $[-1, 1]$

and $[-d, d]$, respectively, from a random-number generator. For $d = 0.1$ and $N = 100$, the results were as follows:

$$\gamma_0^2(\widehat{\Delta}, \Theta_0^{(b)}) = 9.2104; \quad \Theta_0^{(b)} = (-0.1087;\ -0.8626;\ -0.0074),$$
$$\gamma_{g2}^2(\widehat{\Delta}, \Theta_{g2}^{(b)}) = 11.0614; \quad \Theta_{g2}^{(b)} = (-0.1745;\ -1.0321;\ -0.0257),$$
$$\gamma_\infty^2(\widehat{\Delta}, \Theta_\infty^{(b)}) = 56.6811; \quad \Theta_\infty^{(b)} = (-0.6556;\ -1.3677;\ -0.0644).$$

For the real system with robust feedback control laws with the corresponding parameter matrices $\Theta^{(b)}$, solving inequalities (4.4)–(4.6) with

$$\mathcal{F}_A = (A^{(real)} + B_u^{(real)}\Theta^{(b)})P, \quad \mathcal{F}_C = (C + D\Theta^{(b)})P, \quad \mathcal{F}_{C_\Delta} = 0,$$

$\Lambda = 0$, and $\mu = 0$ yielded the following values of the norms:

$$\gamma_0^2(\widehat{\Delta}^{(real)}, \Theta_0^{(b)}) = 3.9640;$$
$$\gamma_{g2}^2(\widehat{\Delta}^{(real)}, \Theta_{g2}^{(b)}) = 4.4416;$$
$$\gamma_\infty^2(\widehat{\Delta}^{(real)}, \Theta_\infty^{(b)}) = 12.2661.$$

3. We design the suboptimal robust control law based on a priori information and the same experimental data for the real system (see above). For this purpose, we calculate the guaranteed estimates of the norms and find the parameter matrices of the robust feedback control laws using the formula $\Theta^{(ab)} = ZP^{-1}$ by solving inequalities (4.4)–(4.6) with $\mathcal{F}_A = AP + B_uZ$, $\mathcal{F}_C = CP + DZ$, $\mathcal{F}_{C_\Delta} = C_\Delta P + D_\Delta Z$, and the unknown variables $\Lambda \geqslant 0$ and $\mu \geqslant 0$:

$$\gamma_0^2(\widehat{\Delta},\Theta_0^{(ab)}) = 8.2265; \quad \Theta_0^{(ab)} = (-0.1613;\ -0.7716;\ -0.0661),$$
$$\gamma_{g2}^2(\widehat{\Delta},\Theta_{g2}^{(ab)}) = 8.9113; \quad \Theta_{g2}^{(ab)} = (-0.4617;\ -0.8835;\ -0.2449),$$
$$\gamma_\infty^2(\widehat{\Delta},\Theta_\infty^{(ab)}) = 35.2885; \quad \Theta_\infty^{(ab)} = (-0.9790;\ -1.0324;\ -0.5212).$$

For the real system with robust feedback control laws with the parameter matrices $\Theta^{(ab)}$, the three norms calculated by solving inequalities (4.4)–(4.6) with

$$\mathcal{F}_A = (A^{(real)} + B_u^{(real)}\Theta^{(ab)})P, \quad \mathcal{F}_C = (C + D\Theta^{(ab)})P, \quad \mathcal{F}_{C_\Delta} = 0,$$

$\Lambda = 0$, and $\mu = 0$ took the following values:

$$\gamma_0^2(\widehat{\Delta}^{(real)}, \Theta_0^{(ab)}) = 4.0280;$$
$$\gamma_{g2}^2(\widehat{\Delta}^{(real)}, \Theta_{g2}^{(ab)}) = 4.5248;$$
$$\gamma_\infty^2(\widehat{\Delta}^{(real)}, \Theta_\infty^{(ab)}) = 14.3512.$$

Figures 1–3 show the guaranteed estimates of the γ_0, generalized H_2, and H_∞ norms, respectively, based on a priori information only, experimental data only, and a priori information together with experimental data, depending on the disturbance level d in the experiment; the lower horizontal lines correspond to the values of these norms for the real object whereas the upper ones to their values under robust control laws designed from a priori information only. Figure 4 plots the guaranteed estimate of the H_∞ norm obtained by the joint use of a priori information and experimental data with the disturbance level $d = 0.05$ as a function of the number of measurements N; the horizontal lines correspond to the H_∞ norm of the real object and the guaranteed estimate of the H_∞ norm obtained using a priori information only.

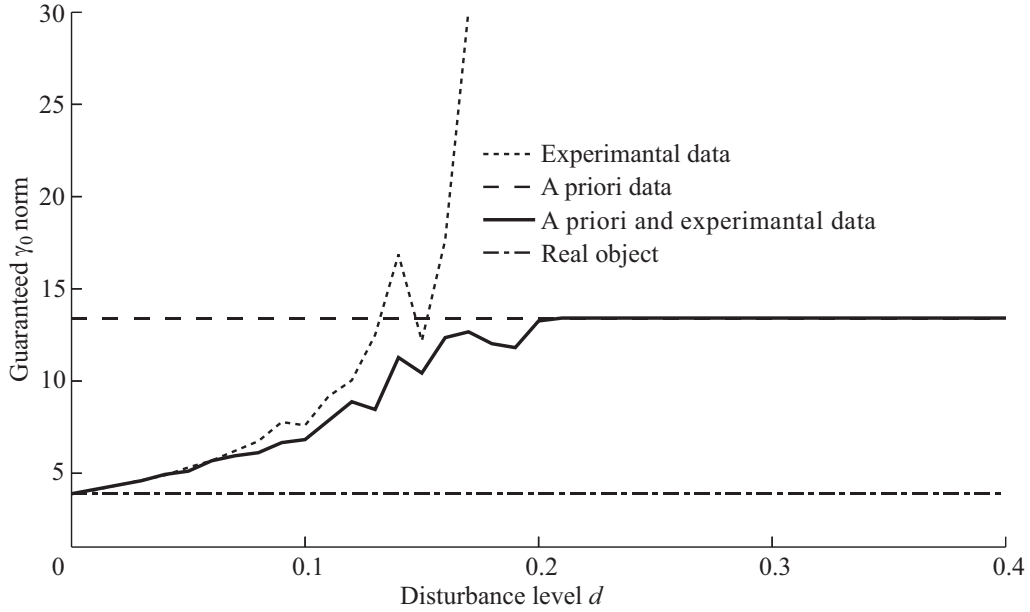


Fig. 1. The guaranteed estimates of the γ_0 norm as a function of the disturbance level in experimental data for different types of information used.

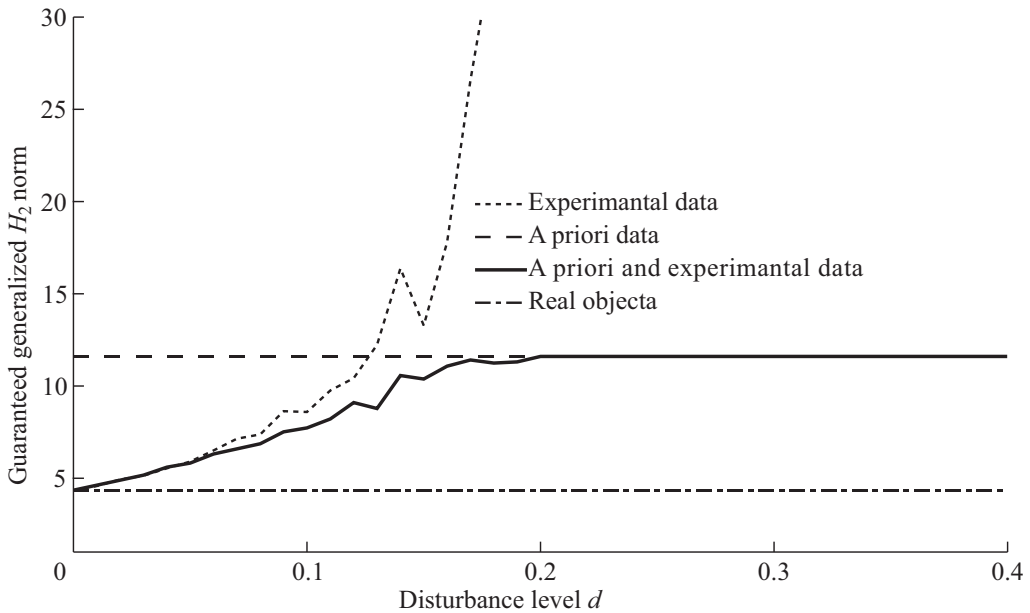


Fig. 2. The guaranteed estimates of the generalized H_2 norm as a function of the disturbance level in experimental data for different types of information used.

According to these results, if the disturbance level in the experiment is relatively small, then the guaranteed estimates of the norms of the closed-loop uncertain system designed using both a priori and experimental data are much smaller than their counterparts under the robust control laws designed using only a priori data or only experimental data. For example, the guaranteed estimates of the H_∞ norms of the closed-loop system with control laws designed using only a priori or only experimental data with the disturbance level $d = 0.1$ are $\gamma^2_\infty(\widehat{\Delta}, \Theta^{(a)}_\infty) = 49.2653$ and

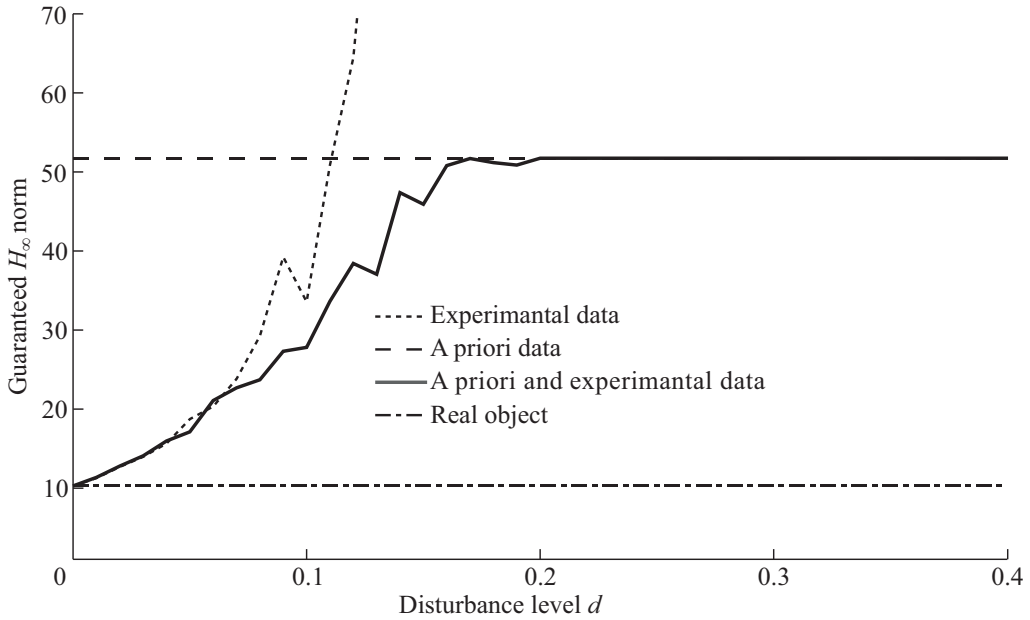


Fig. 3. The guaranteed estimates of the H_∞ norm as a function of the disturbance level in experimental data for different types of information used.

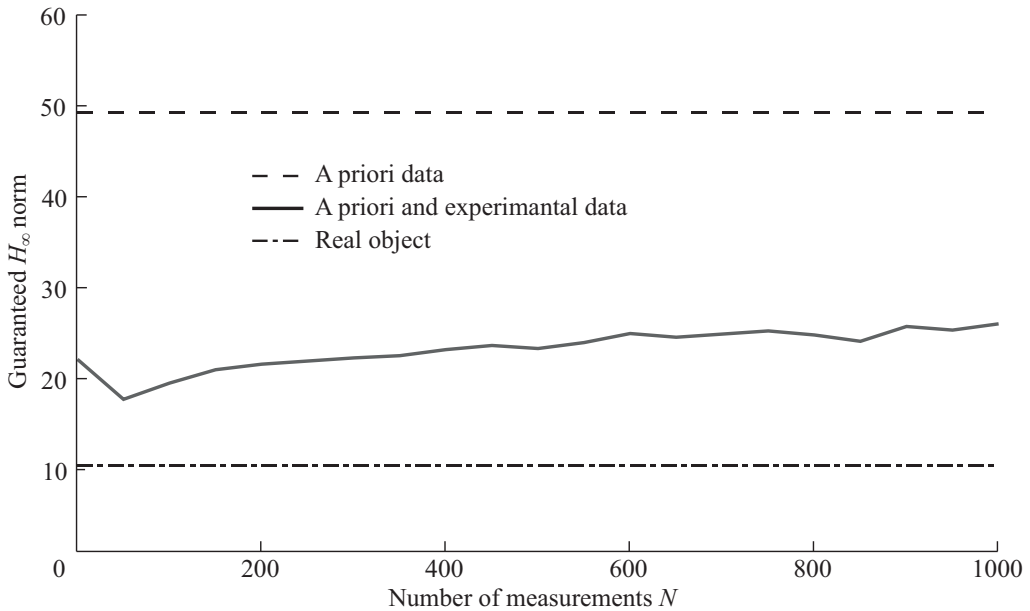


Fig. 4. The guaranteed estimate of the H_∞ norm with a given disturbance level in experimental data depending on the number of measurements.

$\gamma^2_\infty(\widehat{\Delta}, \Theta^{(b)}_\infty) = 56.6811$, respectively; when these a priori and experimental data are used together, the guaranteed estimate of the H_∞ norm equals $\gamma^2_\infty(\widehat{\Delta}, \Theta^{(ab)}_\infty) = 35.2885$. Note the effect of increasing the guaranteed estimates of the norms obtained from experimental data only. This effect can be explained as follows: the set of admissible models of the object consistent with the experimental data expands as the disturbance level increases, and the maximum value of the norm on this set grows accordingly. We emphasize another important feature: the range of disturbance

levels in which the guaranteed estimate of the norm when using a priori and experimental data together is smaller than that when using a priori data only depends on the initial conditions and the chosen controls in the experiment; therefore, this range can be varied and even (apparently) planned. Furthermore, a large number of measurements are not required to obtain acceptable results (see Fig. 4).

6. CONCLUSIONS

This paper has proposed a novel design method for suboptimal robust control laws considering a priori information about the mathematical model of the object and, moreover, experimental data of modeling the object over a small time interval. When obtaining experimental data, neither the persistency of excitation (which ensures the identifiability of unknown parameters) nor data informativity for the corresponding control law is required. In this method, the use of additional information about the unknown parameters of the object obtained from experimental data significantly reduces the guaranteed estimates of the γ_0, generalized H_2, and H_∞ norms of the closed loop system.

APPENDIX

Proof of Lemma 3.1. We write the Lagrange function for this problem and express the optimal value of its dual function as

$$\min_{P_0\geqslant 0,\gamma^2\geqslant 0}\ \max_{K_x\geqslant 0,K_w\geqslant 0}\Big[\operatorname{tr} C_\Theta K_x C_\Theta^{\mathrm{T}}+\operatorname{tr} P_0(A_\Delta K_x A_\Delta^{\mathrm{T}}-K_x+BK_wB^{\mathrm{T}})+\gamma^2(1-\operatorname{tr} K_w)\Big]$$
$$=\min_{P_0\geqslant 0,\gamma^2\geqslant 0}\ \max_{K_x\geqslant 0,K_w\geqslant 0}\Big[\gamma^2+\operatorname{tr} K_x(A_\Delta^{\mathrm{T}}P_0A_\Delta-P_0+C_\Theta^{\mathrm{T}}C_\Theta)+\operatorname{tr} K_w(B^{\mathrm{T}}P_0B-\gamma^2I)\Big].$$

This value is finite under inequalities (3.4); then the maximum is reached at $K_x=0$ and $K_w=0$. In this case, the optimal value of the dual problem coincides with $\lambda_{\max}(B^{\mathrm{T}}P_0B)$. Since the function is convex and there exists an interior point satisfying the constraint, the primal and dual problems have the same optimal value [18].

FUNDING

This work was supported by the Scientific and Educational Mathematical Center "Mathematics of Future Technologies," agreement no. 075-02-2023-945.

REFERENCES

1. Polyak, B.T. and Shcherbakov, P.S., *Robastnaya ustoichivost' i upravlenie* (Robust Stability and Control), Moscow: Nauka, 2002.
2. Petersen, I.R. and Tempo, R., Robust Control of Uncertain Systems: Classical Results and Recent Developments, *Automatica*, 2014, vol. 50, no. 5, pp. 1315–1335.
3. Andrievsky, B.R. and Fradkov, A.L., Speed Gradient Method and Its Applications, *Autom. Remote Control*, 2021, vol. 82, no. 9, pp. 1463–1518.
4. Annaswamy, A.A. and Fradkov, A.L., A Historical Perspective of Adaptive Control and Learning, *Annual Reviews in Control*, 2021, vol. 52, pp. 18–41.
5. De Persis, C. and Tesi, P., Formulas for Data-Driven Control: Stabilization, Optimality and Robustness, *IEEE Trans. Automat. Control*, 2020, vol. 65, no. 3, pp. 909–924.

6. Waarde, H.J., Eising, J., Trentelman, H.L., and Camlibel, M.K., Data Informativity: A New Perspective on Data-Driven Analysis and Control, *IEEE Trans. Automat. Control*, 2020, vol. 65, no. 11, pp. 4753–4768.

7. Berberich, J., Koch, A., Scherer, C.W., and Allgower, F., Robust Data-Driven State-Feedback Design, *Proc. Amer. Control Conf.*, 2020, pp. 1532–1538.

8. Waarde, H.J., Camlibel, M.K., and Mesbahi, M., From Noisy Data to Feedback Controllers: Nonconservative Design via a Matrix S-Lemma, *IEEE Trans. Automat. Control*, 2022, vol. 67, no. 1, pp. 162–175.

9. Bisoffi, A., De Persis, C., and Tesi, P., Data-Driven Control via Petersen's Lemma, *Automatica*, 2022, vol. 145, art. no. 110537.

10. Willems, J.C., Rapisarda, P., Markovsky, I., and De Moor, B., A Note on Persistency of Excitation, *Syst. Control Lett.*, 2005, vol. 54, pp. 325–329.

11. Yakubovich, V.A., *S*-procedure in Nonlinear Control Theory, *Vestn. Leningrad. Univ. Mat.*, 1977, vol. 4, pp. 73–93.

12. Petersen, I.R., A Stabilization Algorithm for a Class of Uncertain Linear Systems, *Syst. Control Lett.*, 1987, vol. 8, pp. 351–357.

13. Doyle, J.C., Analysis of Feedback Systems with Structured Uncertainties, *IEE Proc.*, 1982, vol. 129, part D(6), pp. 242–250.

14. Safonov, M.G., Stability Margins of Diagonally Perturbed Multivariable Feedback Systems, *IEE Proc.*, 1982, vol. 129, part D(6), pp. 251–256.

15. Kogan, M.M., Optimal Discrete-time H_∞/γ_0 Filtering and Control under Unknown Covariances, *Int. J. Control*, 2016, vol. 89, no. 4, pp. 691–700.

16. Wilson, D.A., Convolution and Hankel Operator Norms for Linear Systems, *IEEE Trans. Autom. Control*, 1989, vol. 34, no. 1, pp. 94–97.

17. Balandin, D.V., Biryukov, R.S., and Kogan, M.M., Minimax Control of Deviations for the Outputs of a Linear Discrete Time-Varying System, *Autom. Remote Control*, 2019, vol. 80, no. 12, pp. 345–359.

18. Boyd, S. and Vandenberghe, L., *Convex Optimization*, Cambridge: University Press, 2004.

This paper was recommended for publication by M.V. Khlebnikov, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 8, pp. 933–946.

NONLINEAR SYSTEMS

Continuous Processes with Fuzzy States and Their Applications

V. L. Khatskevich

Military Training and Research Center of the Air Force,
Air Force Academy named after N.E. Zhukovsky and Yu.A. Gagarin, Voronezh, Russia
e-mail: vlkhats@mail.ru

February 25, 2022
April 3, 2023
June 9, 2023

Abstract—Scalar characteristics of continuous processes with fuzzy states—mean and correlation functions—are introduced and studied. Their algebraic properties as well as some properties related to the differentiation and integration of fuzzy functions of a real argument are established. The dependence between the characteristics of a fuzzy signal at the input and output of a dynamic system described by a high-order differential equation with constant coefficients is shown.

Keywords: continuous fuzzy processes, means, correlation functions, fuzzy dynamic systems

DOI: 10.25728/arcRAS.2023.75.63.001

1. INTRODUCTION

When studying dynamic processes under limited initial information, a possible approach is to treat their parameters as realizations of some random processes [1]. However, the distribution of random variables at the time instants under consideration often has a weakly formalizable law. In this case, it is convenient to treat such processes as those with fuzzy states (fuzzy processes). In particular, an important class of fuzzy dynamic processes consists of automatic and optimal control systems.

Thus, continuous fuzzy processes represent an alternative model for automatic control problems in addition to continuous random processes. A fuzzy process is understood as a parametric system of fuzzy numbers that continuously depends on the parameter (time). At present, the theory of fuzzy sets is used in various applications [2, 3]. In particular, different fuzzy models of controlled objects have been investigated [4].

In this paper, numerical characteristics of continuous processes with fuzzy states and continuous time, namely, mean and correlation functions, are introduced and studied; see Sections 3 and 4. Their properties similar to those of the corresponding characteristics of continuous random processes are established. Section 3 considers the algebraic properties of the mean and correlation functions of continuous fuzzy processes. Section 4 is devoted to the properties of these characteristics with respect to integrals and derivatives of fuzzy processes. Integrals of fuzzy functions are understood as a special case of Aumann integrals [5] of multivalued functions (as integrals of α-cutoffs). They were studied in [6, 7] and other publications. Various definitions of derivatives of fuzzy functions were presented, e.g., in [6–8]. Here, we employ the definition in terms of Hukuhara's difference of sets (H-difference) [9]. The results of Sections 3 and 4 rest on the definition and covariance properties of fuzzy numbers discussed in the author's paper [10]; see Section 2.

Nowadays, researchers are actively investigating fuzzy differential equations and their applications; for example, see [3, Chapters 7 and 8; 7, 8, 11–13]. Among the recent works, we mention [14, 15]. Section 5 of this paper considers fuzzy dynamic systems described by nth-order linear differential equations with constant coefficients. The dependence between the numerical characteristics of a fuzzy signal at the output of a fuzzy dynamic system and the corresponding characteristics of its input fuzzy signal is obtained. In contrast to the well-known frameworks [12–15], the approach below develops the Green function method, widely used in the theory of ordinary differential equations [16, Chapter II; 17, Chapter 1], to the class of fuzzy differential equations.

2. THE MEAN, QUASI-SCALAR PRODUCT, AND COVARIANCE OF FUZZY NUMBERS

A fuzzy number is understood as a fuzzy subset of the universal set of real numbers that has a compact support and a normal, convex, and upper semicontinuous membership function; for details, e.g., see [1]. Let J denote the set of all such fuzzy numbers.

The interval representation of fuzzy numbers will be used below.

As is known, the α-level intervals (α-levels) of a fuzzy number $\tilde{z}$ with a membership function $\mu_{\tilde{z}}(x)$ are defined as

$$z_\alpha = \{x|\mu_{\tilde{z}}(x) \geqslant \alpha\}, \quad (\alpha \in (0,1]),\ z_0 = cl\{x|\mu_{\tilde{z}}(x) > 0\},$$

where cl indicates the closure of an appropriate set. Assume that all α-levels of a fuzzy number are closed and bounded intervals on the entire real axis. Let $z^-(\alpha)$ and $z^+(\alpha)$ denote the left and right bounds of an α-interval: $z_\alpha = [z^-(\alpha), z^+(\alpha)]$. The values $z^-(\alpha)$ and $z^+(\alpha)$ are called the left and right α-indices of a fuzzy number, respectively. A real number $x \in R$ is treated as a fuzzy number with the left and right α-indices equal to x.

The sum of fuzzy numbers with indices $z^-(\alpha)$, $z^+(\alpha)$ and $u^-(\alpha)$, $u^+(\alpha)$ is understood as a fuzzy number with the α-level intervals $[z^-(\alpha) + u^-(\alpha), z^+(\alpha) + u^+(\alpha)]$.

Multiplication by a positive real number c is characterized by the α-level intervals $[cz^-(\alpha), cz^+(\alpha)]$. Multiplication by a negative real number c is characterized by the α-level intervals $[cz^+(\alpha), cz^-(\alpha)]$. Equality for fuzzy numbers is understood as equality for all the corresponding α-indices $\forall\, \alpha \in [0,1]$.

According to [18], the mean value of a fuzzy number $\tilde{z}$ can be defined through the interval representation as follows:

$$m(\tilde{z}) = \frac{1}{2}\int_0^1 (z^-(\alpha) + z^+(\alpha))\, d\alpha. \tag{1}$$

Note that the mean (1) is linear.

Example 1. Consider a fuzzy triangular number $\tilde{z}$ characterized by a real-valued triple (a, b, c) with $a < b < c$ defining the membership function

$$\mu_{\tilde{z}}(x) = \begin{cases} \dfrac{x-a}{b-a} & \text{if } x \in [a,b] \\ \dfrac{x-c}{b-c} & \text{if } x \in [b,c] \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the lower and upper bounds of the α-interval have the form

$$z^-(\alpha) = (b-a)\alpha + a, \quad z^+(\alpha) = -(c-b)\alpha + c.$$

As is easily verified, the mean (1) of the fuzzy triangular number (a, b, c) is $m(\tilde{z}) = \frac{1}{4}(a+2b+c)$.

The distances between fuzzy numbers can be defined on the set of such numbers in different ways. The interval approach often involves the Hausdorff distances between the α-level sets of fuzzy numbers: for fuzzy numbers $\tilde{z}$ and $\tilde{u}$ with α-levels z_α and u_α, respectively, the corresponding metric [19] is given by

$$\rho(\tilde{z},\tilde{u}) = \sup_{0<\alpha\leqslant 1} \max\left\{\sup_{z\in z_\alpha}\inf_{u\in u_\alpha}|z-u|,\ \sup_{u\in u_\alpha}\inf_{z\in z_\alpha}|z-u|\right\}. \tag{2}$$

Definition (2) induces the equality

$$\rho(\tilde{z},\tilde{u}) = \sup_{0<\alpha\leqslant 1} \max\left\{|z^-(\alpha)-u^-(\alpha)|,\ |z^+(\alpha)-u^+(\alpha)|\right\}. \tag{3}$$

Here, $[z^-(\alpha), z^+(\alpha)]$ and $[u^-(\alpha), u^+(\alpha)]$ are the α-level intervals of the fuzzy numbers $\tilde{z}$ and $\tilde{u}$.

Note that, due to (3), the condition $\rho(\tilde{z},\tilde{u}) = 0$ matches the definition of equality for fuzzy numbers $\tilde{z}$ and $\tilde{u}$ given above.

Consider a fuzzy number $\tilde{z}$ with α-levels $z_\alpha = [z^-(\alpha), z^+(\alpha)]$. Following interval analysis, let

$$mid\, z_\alpha = \frac{1}{2}(z^+(\alpha)+z^-(\alpha)), \quad rad\, z_\alpha = \frac{1}{2}(z^+(\alpha)-z^-(\alpha)).$$

Here, $mid\, z_\alpha$ characterizes the midpoint for each $\alpha\in[0,1]$ and $rad\, z_\alpha$ the range. For fuzzy numbers $\tilde{z}$ and $\tilde{u}$ from J, we define the quasi-scalar product [10]

$$\begin{aligned}\langle\tilde{z},\tilde{u}\rangle &= \int_0^1 (mid\, z_\alpha mid\, u_\alpha + rad\, z_\alpha rad\, u_\alpha)\, d\alpha \\ &= 0.5\int_0^1 (z^+(\alpha)u^+(\alpha)+z^-(\alpha)u^-(\alpha))\, d\alpha.\end{aligned} \tag{4}$$

The quasi-norm is $\|\tilde{z}\| = \langle\tilde{z},\tilde{z}\rangle^{1/2}$.

Example 2. Consider two triangular numbers $\tilde{z}_1$ and $\tilde{z}_2$ characterized by real-valued triples a_i, b_i, c_i with $a_i < b_i < c_i$ $(i = 1, 2)$. According to the definition of their right and left indices (see Example 1) and (4), the quasi-scalar product $\langle\tilde{z}_1,\tilde{z}_2\rangle$ is given by

$$\langle\tilde{z}_1,\tilde{z}_2\rangle = \frac{2}{3}b_1b_2 + \frac{1}{3}(a_1a_2+c_1c_2) + \frac{1}{6}(a_1b_2+b_1a_2+b_1c_2+b_2c_1).$$

Proposition 1 [10]. *The quasi-scalar product* (4) *possesses the following properties:*

1) $\langle\tilde{z},\tilde{u}\rangle = \langle\tilde{u},\tilde{z}\rangle\ \forall\,\tilde{u},\tilde{z}\in J$.

2) $\langle c_1\tilde{z}, c_2\tilde{u}\rangle = c_1c_2\langle\tilde{z},\tilde{u}\rangle$ *provided that* $c_1c_2 > 0$.

3) $\langle\tilde{z}_1+\tilde{z}_2,\tilde{u}\rangle = \langle\tilde{z}_1,\tilde{u}\rangle + \langle\tilde{z}_2,\tilde{u}\rangle\ \forall\,\tilde{u},\tilde{z}_1,\tilde{z}_2\in J$.

4) $\langle\tilde{z},\tilde{z}\rangle \geqslant 0$, *and the condition* $\langle\tilde{z},\tilde{z}\rangle = 0$ *is equivalent to the zero left and right indices of* $\tilde{z}$.

5) *The generalized Cauchy–Bunyakovsky–Schwarz inequality* $|\langle\tilde{z},\tilde{u}\rangle| \leqslant \langle\tilde{z},\tilde{z}\rangle^{1/2}\langle\tilde{u},\tilde{u}\rangle^{1/2}$ *holds* $\forall\,\tilde{u},\tilde{z}\in J$.

For fuzzy numbers $\tilde{z}_1$ and $\tilde{z}_2$ with means m_1 and m_2, respectively, we define their covariance by the formula [10]

$$\begin{aligned}cov[\tilde{z}_1,\tilde{z}_2] &= \langle\tilde{z}_1-m_1,\tilde{z}_2-m_2\rangle \\ &= 0.5\int_0^1 \left((z_1^+-m_1)(z_2^+-m_2)+(z_1^--m_1)(z_2^--m_2)\right) d\alpha.\end{aligned} \tag{5}$$

The variance is denoted by $D(\tilde{z}) = cov|\tilde{z},\tilde{z}|$.

Proposition 2 [10]. *The covariance* (5) *possesses the following properties*:

1) $cov[\tilde{z}_1 + \tilde{z}_2, \tilde{u}] = cov[\tilde{z}_1, \tilde{u}] + cov[\tilde{z}_2, \tilde{u}]$ $(\forall\, \tilde{u}, \tilde{z}_1, \tilde{z}_2 \in J)$.

2) $cov[c_1\tilde{z}, c_2\tilde{u}] = c_1c_2cov[\tilde{z}, \tilde{u}]$ $(\forall\, \tilde{u}, \tilde{z} \in J)$ *for any real numbers* c_1 *and* c_2 *such that* $c_1c_2 > 0$.

3) $cov\,[\tilde{z}_1, \tilde{z}_2] = \langle \tilde{z}_1, \tilde{z}_2 \rangle - m_1m_2$, $(\forall\, \tilde{z}_1, \tilde{z}_2 \in J)$, *where* m_1 *and* m_2 *are the mean values of fuzzy numbers* $\tilde{z}_1$ *and* $\tilde{z}_2$, *respectively (the specific covariance property).*

Proposition 3 [10]. *The variance possesses the following properties*:

1) $D(c\tilde{z}) = c^2D(\tilde{z})$ *for any real number* c.

2) $D(\tilde{z} + \tilde{u}) = D(\tilde{z}) + D(\tilde{u}) + 2cov[\tilde{z}, \tilde{u}]$ $\forall\, \tilde{u}, \tilde{z} \in J$.

In several works (e.g., see [20]), the covariance of fuzzy numbers $\tilde{z}_1$ and $\tilde{z}_2$ was defined as

$$cov_1[\tilde{z}_1, \tilde{z}_2] = \frac{1}{4}\int_0^1 (z_1^+(\alpha) - z_1^-(\alpha))(z_2^+(\alpha) - z_2^-(\alpha))d\alpha.$$

With this definition, covariance is always nonnegative, which disagrees with standard covariance properties (for random variables).

3. CONTINUOUS FUZZY PROCESSES

Consider a fixed segment $[t_0, T]$ of the real axis, where $t_0 \geqslant 0$. A mapping $\tilde{z} : [t_0, T] \to J$ is called a process with fuzzy states (or a fuzzy process) and continuous time.

Let a fuzzy process $\tilde{z}(t)$, $t \in [t_0, T]$, be characterized a membership function $\mu_{\tilde{z}}(x, t)$. For a fixed number $\alpha \in (0, 1]$, consider the α-interval $z_\alpha(t) = \{x \in R : \mu_{\tilde{z}}(x, t) \geqslant \alpha\}$ and $z_0(\alpha) = cl\{x \in R : \mu_{\tilde{z}}(x, t) > 0\}$. We denote by $z_\alpha^-(t) = z^-(t, \alpha)$ and $z_\alpha^+(t) = z^+(t, \alpha)$ the left and right bounds of the α-interval, respectively: $z_\alpha(t) = [z^-(t, \alpha), z^+(t, \alpha)]$.

Assume that the indices $z^-(t, \alpha)$ and $z^+(t, \alpha)$ are square summable in α for each $t \in [t_0, T]$ and continuous in t for any $\alpha \in [0, 1]$.

For each $t \in [t_0, T]$, let the mean of $\tilde{z}(t)$ be defined as

$$m_{\tilde{z}}(t) = m(\tilde{z}(t)) = \frac{1}{2}\int_0^1 (z^-(t, \alpha) + z^+(t, \alpha))d\alpha. \tag{6}$$

Theorem 1. *The mean of a continuous fuzzy process given by* (6) *possesses the following properties:*

1. *If* $\tilde{z}_1(t)$ *and* $\tilde{z}_2(t)$ *are continuous fuzzy processes, then* $m(\tilde{z}_1(t) + \tilde{z}_2(t)) = m(\tilde{z}_1(t)) + m(\tilde{z}_2(t))$ *(additivity).*

2. *If* $\tilde{z}(t)$ *is a continuous fuzzy process and* $\varphi(t)$ *is a real-valued function, then* $m(\varphi(t)\tilde{z}(t)) = \varphi(t)m(\tilde{z}(t))$ *(homogeneity).*

Indeed, property 1 follows from the definition of interval summation and and the additivity of Lebesgue integrals.

It remains to show property 2. For a fixed number $t \in [t_0, T]$, consider the fuzzy number $\tilde{w}(t) = \varphi(t)\tilde{z}(t)$. Note that its left $w^-(t, \alpha)$ and right $w^+(t, \alpha)$ indices coincide with the expressions $\varphi(t)z^-(t, \alpha)$ and $\varphi(t)z^+(t, \alpha)$, respectively, in the case $\varphi(t) \geqslant 0$ or with the expressions $\varphi(t)z^+(t, \alpha)$ and $\varphi(t)z^-(t, \alpha)$, respectively, in the case $\varphi(t) < 0$. However, their sum $w^-(t, \alpha) + w^+(t, \alpha)$ coincides with the expression $\varphi(t)(z^-(t, \alpha) + z^+(t, \alpha))$, which is independent of the sign of $\varphi(t)$. According to (1), this fact implies property 2.

Corollary 1. *If* $f(t)$ *is a real-valued function, then* $m(\tilde{z}(t) + f(t)) = m(\tilde{z}(t)) + f(t)$.

Suppose that $f^-(t) = f^+(t) = f(t)$ for a real number $f(t)\ \forall\, t \in [t_0, T]$.

Let the correlation function of a continuous fuzzy process $\tilde{z}(t)$ be defined as

$$K_{\tilde{z}}(t_1, t_2) = \frac{1}{2}\int_0^1 \left(z^+(t_1,\alpha) - m(\tilde{z}(t_1))\right)\left(z^+(t_2,\alpha) - m(\tilde{z}(t_2))\right) + \left(z^-(t_1,\alpha) - m(\tilde{z}(t_1))\right)\left(z^-(t_2,\alpha) - m(\tilde{z}(t_2))\right) d\alpha. \tag{7}$$

The variance of a continuous fuzzy process is the value $D_{\tilde{z}}(t) = K_{\tilde{z}}(t,t)$. By definition, $D_{\tilde{z}}(t) \geqslant 0$.

Theorem 2. *The correlation function* (7) *of a continuous fuzzy process possesses the following properties.*

1. *For a continuous fuzzy process* $\tilde{z}(t)$*, the equality*

$$K_{\tilde{z}}(t_1, t_2) = K_{\tilde{z}}(t_2, t_1)$$

holds $\forall\, t_1, t_2 \in [t_0, T]$ *(symmetry).*

2. *If* $\tilde{z}(t)$ *is a continuous fuzzy process and* $\varphi(t)$ *is a real-valued function, then the correlation function* $K_{\tilde{w}}(t_1, t_2)$ *of a continuous fuzzy process* $\tilde{w}(t) = \varphi(t)\tilde{z}(t)$ *has the form* $K_{\tilde{w}}(t_1, t_2) = \varphi(t_1)\varphi(t_2)K_{\tilde{z}}(t_1, t_2)\ \forall\, t_1, t_2 \in [t_0, T]$ *such that* $\varphi(t_1)\varphi(t_2) \geqslant 0$.

3. *If* $\tilde{w}(t) = \tilde{z}(t) + \varphi(t)$*, then* $K_{\tilde{w}}(t_1, t_2) = K_{\tilde{z}}(t_1, t_2)$.

4. $|K_{\tilde{z}_1}(t_1, t_2)| \leqslant \sqrt{D_{\tilde{z}}(t_1) D_{\tilde{z}}(t_2)}$.

Theorem 2 is based on the properties of the covariance (5) of fuzzy numbers presented in Section 2.

For continuous fuzzy processes $\tilde{z}_1(t)$ and $\tilde{z}_2(t)$, consider the mutual correlation function

$$K_{\tilde{z}_1\tilde{z}_2}(t, s) = \int_0^1 \left(z_1^+(t,\alpha) - m(\tilde{z}_1(t))\right)\left(z_2^+(s,\alpha) - m(\tilde{z}_2(s))\right) + \left(z_1^-(t,\alpha) - m(\tilde{z}_1(t))\right)\left(z_2^-(s,\alpha) - m(\tilde{z}_2(s))\right) d\alpha.$$

Theorem 3. *Let* $\tilde{z}_1(t)$ *and* $\tilde{z}_2(t)$ *be continuous fuzzy processes. The correlation function of their sum* $\tilde{w}(t) = \tilde{z}_1(t) + \tilde{z}_2(t)$ *has the form*

$$K_{\tilde{w}}(t, s) = K_{\tilde{z}_1}(t, s) + K_{\tilde{z}_2}(t, s) + K_{\tilde{z}_1,\tilde{z}_2}(t, s) + K_{\tilde{z}_1,\tilde{z}_2}(s, t).$$

Continuous fuzzy processes $\tilde{z}_1(t)$ and $\tilde{z}_2(t)$ are said to be uncorrelated on a segment $[t_0, T]$ if

$$K_{\tilde{z}_1\tilde{z}_2}(t, s) = 0 \quad (\forall\, t, s \in [t_0, T]).$$

Corollary 2. *If continuous fuzzy processes* $\tilde{z}_1(t)$*,* $\tilde{z}_2(t)$ *are uncorrelated and* $\tilde{w}(t) = \tilde{z}_1(t) + \tilde{z}_2(t)$*, then*

$$K_{\tilde{w}}(t, s) = K_{\tilde{z}_1}(t, s) + K_{\tilde{z}_2}(t, s) \quad (\forall\, t, s \in [t_0, T]).$$

4. THE INTEGRATION AND DIFFERENTIATION OF CONTINUOUS FUZZY PROCESSES

The integral of a continuous fuzzy process $\tilde{z}(t)$ between the limits of a segment $[t_0, T]$ is a fuzzy number $\tilde{g}$ with the α-level intervals $g_\alpha = \int_{t_0}^T z_\alpha(t)\, dt$ for any $\alpha \in [0, 1]$; for details, see [7]. The integral is denoted by $\int_{t_0}^T \tilde{z}(t)\, dt$.

In fact, this is the Aumann integral [5] of a multi-valued mapping $z_\alpha(t)$.

If the integral $\int_{t_0}^T \tilde{z}(t)\, dt$ exists, then the process $\tilde{z}(t)$ is said to be integrable on $[t_0, T]$.

The mean of the integral possesses the following property.

Theorem 4. *Let $\tilde{z}(t)$ be an integrable fuzzy process on $[t_0, T]$. Then* $m\left(\int\limits_{t_0}^{T} \tilde{z}(\tau)d\tau\right) = \int\limits_{t_0}^{T} m(\tilde{z}(\tau))\, d\tau$.

By the definition of the integral, its indices satisfy the relation

$$\left(\int_{t_0}^{T} \tilde{z}(\tau)\, d\tau\right)_{\alpha}^{\pm} = \int_{t_0}^{T} z^{\pm}(\tau, \alpha)\, d\tau.$$

Consequently,

$$m\left(\int_{t_0}^{T} \tilde{z}(\tau)\, d\tau\right) = \frac{1}{2}\int_{t_0}^{1}\left(\int_{t_0}^{T}(z^{-}(\tau,\alpha)) + (z^{+}(\tau,\alpha))\, d\tau\right) d\alpha = \int_{t_0}^{T} m(\tilde{z}(\tau))d\tau.$$

For a continuous fuzzy process $\tilde{z}(t)\ \forall\, t \in [t_0, T]$, we define the continuous fuzzy process $\tilde{g}(t) = \int_{t_0}^{t} \tilde{z}(\tau)\, d\tau$.

Theorem 5. *The integral $\tilde{g}(t)$ of a continuous fuzzy process $\tilde{z}(t)$ has the correlation function $K_{\tilde{g}}(t_1, t_2) = \int_{t_0}^{t_1}\int_{t_0}^{t_2} K_{\tilde{z}}(\tau_1, \tau_2)\, d\tau_1 d\tau_2$.*

Proof. By definition,

$$\begin{aligned}
K_{\tilde{g}}(t_1, t_2) = \frac{1}{2}\int_0^1 &\left(\int_{t_0}^{t_1} z^{+}(\tau,\alpha)\, d\tau - \int_{t_0}^{t_1} m(\tilde{z}(\tau,\alpha))\, d\tau\right)\left(\int_{t_0}^{t_2} z^{+}(\tau,\alpha)\, d\tau - \int_{t_0}^{t_2} m(\tilde{z}(\tau))d\tau\right)\\
&+\left(\int_{t_0}^{t_1} z^{-}(\tau,\alpha)\, d\tau - \int_{t_0}^{t_1} m(\tilde{z}(\tau,\alpha))\, d\tau\right)\left(\int_{t_0}^{t_2} z^{-}(\tau,\alpha)\, d\tau - \int_{t_0}^{t_2} m(\tilde{z}(\tau))\, d\tau\right) d\alpha\\
&= \frac{1}{2}\int_0^1 \left(\int_{t_0}^{t_1}(z^{+}(\tau,\alpha) - m(\tilde{z}(\tau))\, d\tau)\left(\int_{t_0}^{t_2} z^{+}(\tau,\alpha) - m(\tilde{z}(\tau))\, d\tau\right)\right) d\alpha\\
&\quad + \frac{1}{2}\int_0^1 \left(\int_{t_0}^{t_1} z^{-}(\tau,\alpha) - m(\tilde{z}(\tau))\, d\tau \int_{t_0}^{t_2} z^{-}(\tau,\alpha) - m(\tilde{z}(\tau))\, d\tau\right) d\alpha.
\end{aligned}$$

Consider the first integral in this expression. Since the integral's value is independent of the integration variable, it can be written as

$$\begin{aligned}
&\frac{1}{2}\int_0^1 \left(\int_{t_0}^{t_1} z^{+}(\tau_1,\alpha) - m(\tilde{z}(\tau_1))\, d\tau_1\right)\left(\int_{t_0}^{t_2} z^{+}(\tau_2,\alpha) - m(\tilde{z}(\tau_2))\, d\tau_2\right) d\alpha\\
&= \frac{1}{2}\int_0^1 \int_{t_0}^{t_1}\int_{t_0}^{t_2} \left(z^{+}(\tau_1,\alpha) - m(\tilde{z}(\tau_1))\right)\left(z^{+}(\tau_2,\alpha) - m(\tilde{z}(\tau_2))\right) d\tau_1\, d\tau_2\, d\alpha.
\end{aligned}$$

The same line of reasoning applies to the indices with a minus sign. Thus,

$$\begin{aligned}
K_{\tilde{g}}(t_1, t_2) = \frac{1}{2}\int_0^1 \Bigg(&\int_{t_0}^{t_1}\int_{t_0}^{t_2} \left(z^{+}(\tau_1,\alpha) - m(\tilde{z}(\tau_1))\right)\left(z^{+}(\tau_2,\alpha) - m(\tilde{z}(\tau_2))\right)\\
&+ \left(z^{-}(\tau_1,\alpha) - m(\tilde{z}(\tau_1))\right)\left(z^{-}(\tau_2,\alpha) - m(\tilde{z}(\tau_2))\right) d\tau_1\, d\tau_2\Bigg) d\alpha.
\end{aligned}$$

Interchanging the order of integration finally gives the desired result.

Consider now the derivatives of fuzzy functions. Different definitions are introduced in the literature. A common one involves the concept of Hukuhara's difference (H-difference) [9]. For sets A and B, a set C is called their H-difference if $A = B + C$ and is denoted by $A \overset{h}{-} B$.

A mapping $\tilde{z} : [t_0, T] \to J$ is said to be differentiable at a point $t \in [t_0, T]$ [7] if $\forall\, \alpha[0,1]$ the multivalued mapping $z_\alpha(t)$ is Hukuhara differentiable at the point t with the derivative $D_H z_\alpha(t)$ and the family $\{D_H z_\alpha(t) : \alpha \in [0,1]\}$ defines a certain element $\tilde{z}\,'(t)$ belonging to J. The element $\tilde{z}\,'(t)$ is called the fuzzy derivative of $\tilde{z}(t)$ at the point t.

By definition, the fuzzy derivative $\tilde{z}\,'(t)$ satisfies the relation

$$\lim_{\Delta t \to 0} \rho\left(\frac{1}{\Delta t}\left(\tilde{z}(t+\Delta t) \overset{h}{-} \tilde{z}(t)\right), \tilde{z}\,'(t)\right) = 0,$$

where the distance ρ is given by (3).

Proposition 4 [7]. *Let a mapping $\tilde{z} : [t_0, T] \to J$ be differentiable and its fuzzy derivative $\tilde{z}\,'(t)$ be integrable on $[t_0, T]$. Then*

$$\tilde{z}(t) = \tilde{z}(t_0) + \int_{t_0}^{t} \tilde{z}\,'(s) ds. \tag{8}$$

Proposition 5 [11]. *Let a fuzzy process $\tilde{z}(t)$ be differentiable and $z_\alpha(t) = [z_\alpha^-(t), z_\alpha^+(t)]$ be its α-interval for any $\alpha \in [0,1]$. Then the functions $z_\alpha^-(t)$ and $z_\alpha^+(t)$ are differentiable with respect to t and the α-interval of the derivative $\tilde{z}\,'(t)$ has the form $[\tilde{z}\,'(t)]_\alpha = [(z_\alpha^-)'(t), (z_\alpha^+)'(t)]$.*

Proposition 5 shows the connection between the derivative introduced above and the Seikkala derivative [8].

Theorem 6. *Let $\tilde{z}(t)$ be a differentiable fuzzy process with the integrable derivative $\tilde{z}\,'(t)$. Then the mean of its derivative coincides with the derivative of its mean: $m(\tilde{z}\,'(t)) = \frac{d}{dt} m(\tilde{z}(t))$.*

Proof. Taking the mean of the left- and right-hand sides of formula (8) yields

$$m(\tilde{z}(t)) = m(\tilde{z}(t_0)) + \int_{t_0}^{t} m(\tilde{z}\,'(s))\, ds.$$

This equality is based on the additivity of means and Theorem 4. Let us differentiate its sides. In view of the properties of the integral with a variable upper limit, we obtain $\frac{d}{dt} m(\tilde{z}(t)) = m(\tilde{z}\,'(t))$, and the conclusion follows. The proof of Theorem 6 is complete.

Theorem 7. *The derivative $\tilde{z}\,'(t)$ of a differentiable fuzzy process $\tilde{z}(t)$ has the correlation function*

$$K_{\tilde{z}\,'}(t_1, t_2) = \frac{\partial^2 (K_{\tilde{z}}(t_1, t_2))}{\partial t_1 \partial t_2}.$$

Proof. Denoting $\tilde{z}\,'(t) = \tilde{w}(t)$, we consider $\tilde{g}(t) = \int_{t_0}^{t} \tilde{w}(s)\, ds$. Due to Theorem 5, the correlation function $K_{\tilde{g}}(t_1, t_2)$ is given by

$$K_{\tilde{g}}(t_1, t_2) = \frac{1}{2} \int_0^1 \left(\int_{t_0}^{t_1} w^+(\tau_1, \alpha) d\tau_1 - m(\tilde{w}(\tau_1)) \right) \left(\int_{t_0}^{t_2} w^+(\tau_2, \alpha) d\tau_2 - m(\tilde{w}(\tau_2)) \right) d\alpha$$
$$+ \frac{1}{2} \int_0^1 \left(\int_{t_0}^{t_1} w^-(\tau_1, \alpha) d\tau_1 - m(\tilde{w}(\tau_1)) \right) \left(\int_{t_0}^{t_2} w^-(\tau_2, \alpha) d\tau_2 - m(\tilde{w}(\tau_2)) \right) d\alpha.$$

Differentiating this equality, first with respect to t_1 and then with respect to t_2, yields

$$\frac{\partial^2 K_{\tilde{g}}(t_1,t_2)}{\partial t_1 \partial t_2} = \frac{1}{2}\int\limits_0^1 \left(w^+(\tau_1,\alpha) - m(\tilde{w}(\tau_1))\right)\left(w^+(\tau_2,\alpha) - m(\tilde{w}(\tau_2))\right) d\alpha$$

$$+ \frac{1}{2}\int\limits_0^1 \left(w^-(\tau_1,\alpha) - m(\tilde{w}(\tau_1))\right)\left(w^-(\tau_2,\alpha) - m(\tilde{w}(\tau_2))\right) d\alpha.$$

As a result,

$$\frac{\partial^2 K_{\tilde{g}}(t_1,t_2)}{\partial t_1 \partial t_2} = K_{\tilde{w}}(t_1,t_2). \tag{9}$$

Using formula (8) with $\tilde{z}(t_0) = \tilde{\xi}$, we write

$$\tilde{z}(t) = \tilde{\xi} + \int\limits_{t_0}^{t} \tilde{w}(s)\, ds = \tilde{\xi} + \tilde{g}(t).$$

Letting $\tilde{\eta}(t) = \tilde{\xi} + \tilde{g}(t)$ and calculating the correlation function of the sum of fuzzy processes, we obtain

$$K_{\tilde{z}}(t_1,t_2) = K_{\tilde{\eta}}(t_1,t_2) = K_{\tilde{\xi}}(t_1,t_2) + K_{\tilde{g}}(t_1,t_2) + K_{\xi g}(t_1,t_2) + K_{\xi g}(t_2,t_1).$$

By analogy, differentiating this equality, first with respect to t_1 and then with respect to t_2, yields $\frac{\partial^2 K_{\tilde{z}}(t_1,t_2)}{\partial t_1 \partial t_2} = \frac{\partial^2 K_{\tilde{g}}(t_1,t_2)}{\partial t_1 \partial t_2}$. The other terms on the right-hand side vanish since $K_{\tilde{\xi}}$ is independent of t_1 and t_2 by definition whereas $K_{\xi g}(t_1,t_2)$ and $K_{\xi g}(t_2,t_1)$ depend only on t_2 and t_1, respectively. Considering formula (9), we finally arrive at the equality

$$\frac{\partial^2 K_{\tilde{z}}(t_1,t_2)}{\partial t_1 \partial t_2} = K_{\tilde{w}}(t_1,t_2) = K_{\tilde{z}'}(t_1,t_2),$$

and the proof of Theorem 7 is complete.

5. TRANSFORMATION OF A CONTINUOUS FUZZY PROCESS BY A LINEAR DYNAMIC SYSTEM

Consider some device A with continuous fuzzy signals $\tilde{y}(t)$ and $\tilde{z}(t)$ at its input and output, respectively.

Device A is called a linear dynamic system if the relationship between the input and output signals is described by an nth-order differential equation with constant coefficients. With fuzzy input $\tilde{y}(t)$ and output $\tilde{z}(t)$ signals, the linear dynamic system is described by the fuzzy differential equation

$$\begin{aligned} &a_n \tilde{z}^{(n)}(t) + a_{n-1}\tilde{z}^{(n-1)}(t) + \cdots + a_1 \tilde{z}\,'(t) + a_0 \tilde{z}(t) \\ &= b_k \tilde{y}^{(k)}(t) + b_{k-1}\tilde{y}^{(k-1)}(t) + \cdots + b_1 \tilde{y}\,'(t) + b_0 \tilde{y}(t) \equiv \tilde{f}(t). \end{aligned} \tag{10}$$

Here, the coefficients a_i $(i = 0,\ldots,n)$ and b_i $(i = 0,\ldots,k)$ are constant numbers, the second-order derivatives of the fuzzy function are understood as $\tilde{z}\,''(t) = (\tilde{z}\,'(t))'$ (and so on for higher-order derivatives).

The next result characterizes the connection between the mean values of the input and output fuzzy signals.

Lemma 1. *The mean value $z_{mean}(t) = m(\tilde{z}(t))$ of the output fuzzy signal $\tilde{z}(t)$ of the dynamic system* (10) *satisfies the scalar differential equation*

$$a_n x^{(n)} + a_{n-1} x^{(n-1)} + \cdots + a_1 x' + a_0 x = f(t), \tag{11}$$

where f stands for the mean of the right-hand side of (10)*: $f(t) = m\tilde{f}(t)$.*

Indeed, consider the mean of the left- and right-hand sides of equality (10). Using the additivity and homogeneity of means as well as Theorem 6, we obtain

$$\begin{aligned} &a_n(m\tilde{z}(t))^{(n)} + a_{n-1}(m\tilde{z}(t))^{(n-1)} + \cdots + a_1(m\tilde{z}(t))' + a_0 m\tilde{z}(t) \\ &= b_k(m\tilde{y}(t))^{(k)} + b_{k-1}(m\tilde{y}(t))^{(k-1)} + \cdots + b_1(m\tilde{y}(t))' + b_0(m\tilde{y}(t)) \equiv m\tilde{f}(t). \end{aligned}$$

Then the scalar function $z_{mean}(t) = m(\tilde{z}(t))$ satisfies equation (11).

Proposition 6 [16, Chapter II]. *Let the roots of the characteristic equation $a_n\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0 = 0$ contain no points on the imaginary axis. Then for any continuous function $f(t)$ bounded on the entire real axis, there exists a unique solution of equation* (11) *that is bounded on the entire real axis. This solution has the form*

$$x(t) = \int_{-\infty}^{\infty} G(t-s)f(s)\,ds, \tag{12}$$

where $G(t)$ is the Green function of the problem on bounded solutions of equation (11)*.*

Note that the Green function of the problem on bounded solutions of equation (11) is known; for example, see [17, Chapter 1, § 8].

Remark 1. Assume that under the hypotheses of Proposition 6, all roots of the characteristic equation belong to the left half-plane: ($Re\lambda_i < 0$, $i = 1, \ldots, n$). Then the bounded solution of equation (11) is asymptotically Lyapunov stable. In addition, the Green function of the problem on bounded solutions of equation (11) has the form

$$G(t) = \begin{cases} k(t) & \text{for } t \geqslant 0 \\ 0 & \text{for } t < 0, \end{cases}$$

where $k(t)$ is the Cauchy function of the homogeneous equation corresponding to (11).

Theorem 8. *Let the input fuzzy process $\tilde{y}(t)$ be continuous and bounded on the entire real axis together with its derivatives $\tilde{y}^i(t)$ ($i = 1, 2, \ldots, k$). Let the roots of the characteristic equation $a_n\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0 = 0$ contain no points on the imaginary axis. Then the mean value $m(\tilde{z}(t))$ at the output of the dynamic system* (10) *can be represented as*

$$m(\tilde{z}(t)) = \int_{-\infty}^{\infty} G(t-s)m(\tilde{f}(s))\,ds, \tag{13}$$

where G is the Green function of the problem on bounded solutions of equation (11)*.*

Indeed, under the hypotheses of Theorem 8, the right-hand side of equation (11) is a bounded function on the entire real axis. Then, according to Lemma 1, the function $z_{mean}(t) = m(\tilde{z}(t))$ is the solution of equation (11) bounded on the entire real axis. Hence, Theorem 8 follows from Proposition 6.

Note that the boundedness of the fuzzy signal $\tilde{y}(t)$ (in Theorem 8 and below) is understood as the boundedness of all the corresponding α-indices $y_\alpha^{\pm}(t)$ in $t\ \forall\,\alpha \in [0, 1]$.

Corollary 3. *Assume that under the hypotheses of Theorem 8, the input signal is quasi-stationary:* $m(y(t)) = m_{\tilde{y}} = \text{const}$. *Then the output signal is quasi-stationary as well, and its mean value is* $m(\tilde{z}(t)) = m_{\tilde{z}} = \frac{b_0}{a_0} m_{\tilde{y}}$.

Indeed, the arbitrary-order derivative of a constant is zero and, in this case, the right-hand side of equation (11) is $b_0 m_{\tilde{y}}$. Then $m_{\tilde{z}}$ is the solution of the corresponding equation (11): $a_0 m_{\tilde{z}} = b_0 m_{\tilde{y}}$. Equation (11) has no other bounded solutions under the hypotheses of Theorem 8.

The same conclusion can be drawn for the mean value of the fuzzy input signal that stabilizes over time, i.e., $m(y(t)) \to m_{\tilde{y}}$ as $t \to \infty$.

In some cases, the indices of the output fuzzy signal of the dynamic system (10) can be written explicitly.

Theorem 9. *Assume that under the hypotheses of Theorem 8, all coefficients of the dynamic system* (10) *are positive* ($a_i > 0$, $i = 0, \dots, n$). *Then the indices of the output fuzzy signal* $\tilde{z}(t)$ *of the dynamic system* (10) *have the form*

$$z_\alpha^-(t) = \int\limits_{-\infty}^{\infty} G(t-s) f_\alpha^-(s)\, ds, \quad z_\alpha^+(t) = \int\limits_{-\infty}^{\infty} G(t-s) f_\alpha^+(s)\, ds, \tag{14}$$

where $f_\alpha^\pm(s)$ *are the indices of the function* $\tilde{f}(s)$.

Indeed, equality for fuzzy numbers means equality for all the corresponding α-intervals. Due to the positivity of the coefficients a_i and Theorem (6), by the rules of interval arithmetic, equation (10) $\forall\, \alpha \in [0,1]$ implies

$$a_n (z_\alpha^-)^{(n)}(t) + a_{n-1}(z_\alpha^-)^{(n-1)}(t) + \dots + a_1 (z_\alpha^-)'(t) + a_0 z_\alpha^-(t) = f_\alpha^-(t); \tag{15}$$

by analogy, for the indices with a plus sign,

$$a_n (z_\alpha^+)^{(n)}(t) + a_{n-1}(z_\alpha^+)^{(n-1)}(t) + \dots + a_1 (z_\alpha^+)'(t) + a_0 z_\alpha^+(t) = f_\alpha^+(t). \tag{16}$$

According to (15) and (16), equalities (14) hold by Proposition 6.

Proposition 7. *Assume that under the hypotheses of Theorem 9, the Green function* G *of problem* (10) *is nonnegative. Then the bounded fuzzy signal at the output of the dynamic system* (10) *can be represented as*

$$\tilde{z}(t) = \int\limits_{-\infty}^{\infty} G(t-s) \tilde{f}(s)\, ds. \tag{17}$$

Indeed, by the definition of the integral of a fuzzy function, we have the index relations

$$\left(\int\limits_{-\infty}^{\infty} G(t-s) \tilde{f}(s)\, ds \right)_\alpha^- = \int\limits_{-\infty}^{\infty} G(t-s) f_\alpha^-(s)\, ds,$$

$$\left(\int\limits_{-\infty}^{\infty} G(t-s) \tilde{f}(s)\, ds \right)_\alpha^+ = \int\limits_{-\infty}^{\infty} G(t-s) \tilde{f}_\alpha^+(s)\, ds.$$

In view of (14), they imply the representation (17).

Theorem 10. *Assume that under the hypotheses of Theorem 9, all roots of the characteristic equation have negative real parts* ($Re\lambda_i < 0$, $i = 1, \dots, n$). *Then the output fuzzy signal* $\tilde{z}(t)$ *of the*

dynamic system (10) *has the correlation function*

$$K_{\tilde z}(t_1,t_2)=\int\limits_{-\infty}^{t_1}\int\limits_{-\infty}^{t_2}G(t_1-\tau_1)G(t_2-\tau_2)K_{\tilde f}(\tau_1,\tau_2)d\tau_1 d\tau_2, \tag{18}$$

where $K_{\tilde f}(\tau_1,\tau_2)$ *is the correlation function of the input signal* $\tilde f=\sum_{i=1}^n b_i\tilde y^{(i)}$ *and* G *is the Green function of the problem on bounded solutions of equation* (11).

Proof. Considering Remark 1, by definition (7) and formulas (13) and (14), we write

$$\begin{aligned}K_{\tilde z}(t_1,t_2)&=\frac12\int\limits_0^1\left[\left(\int\limits_{-\infty}^{t_1}G(t_1-s)(f_\alpha^+(s)-m(\tilde f(s)))ds\right)\left(\int\limits_{-\infty}^{t_2}G(t_2-s)(f_\alpha^+(s)-m(\tilde f(s)))ds\right)\right.\\&\left.+\left(\int\limits_{-\infty}^{t_1}G(t_1-s)(f_\alpha^-(s)-m(\tilde f(s)))ds\right)\left(\int\limits_{-\infty}^{t_2}G(t_2-s)(f_\alpha^-(s)-m(\tilde f(s)))ds\right)\right]d\alpha\\&=\frac12\int\limits_0^1\int\limits_{-\infty}^{t_1}\int\limits_{-\infty}^{t_2}G(t_1-\tau_1)G(t_2-\tau_2)\Big[(f_\alpha^+(\tau_1)-m(\tilde f(\tau_1)))(f_\alpha^+(\tau_2)-m(\tilde f(\tau_2)))\\&\qquad+(f_\alpha^-(\tau_1)-m(\tilde f(\tau_1)))(f_\alpha^-(\tau_2)-m(\tilde f(\tau_2)))\Big]d\tau_1 d\tau_2 d\alpha.\end{aligned}$$

Interchanging the order of integration gives

$$\begin{aligned}K_{\tilde z}(t_1,t_2)=\int\limits_{-\infty}^{t_1}\int\limits_{-\infty}^{t_2}G(t_1-\tau_1)G(t_2-\tau_2)&\left(\frac12\int\limits_0^1(f_\alpha^-(\tau_1)-m(\tilde f(\tau_1)))(f_\alpha^-(\tau_2)-m(\tilde f(\tau_2)))\right.\\&\left.+(f_\alpha^+(\tau_1)-m(\tilde f(\tau_1)))(f_\alpha^+(\tau_2)-m(\hat f(\tau_2)))d\alpha\right)d\tau_1 d\tau_2,\end{aligned}$$

directly leading to (18).

Note that the assumption $Re\lambda_i<0$, $i=1,\dots,n$, in Theorem 10 serves only for clarity when comparing with Theorem 5. Without this assumption, formula (18) becomes

$$K_{\tilde z}(t_1,t_2)=\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}G(t_1-\tau_1)G(t_2-\tau_2)K_{\tilde f}(\tau_1,\tau_2)d\tau_1 d\tau_2.$$

Example 3. Consider a linear dynamic system described by the first-order differential equation with constant coefficients

$$\tilde z\,'(t)+\beta\tilde z(t)=\tilde y\,'(t),\quad \beta>0.$$

Let a fuzzy signal $\tilde y\,'(t)$ bounded on the entire real axis be supplied to the input of this system. It is required to find the numerical characteristics of the bounded output fuzzy signal $\tilde z(t)$.

Note that the Green function of the problem on bounded solutions of the scalar equation $x'+\beta x=y(t)$ is represented as

$$G_1(t)=\begin{cases}e^{-\beta t} & \text{for } t\geqslant 0\\ 0 & \text{for } t<0.\end{cases}$$

Then, according to Theorem 8, the mean value at the system output has the form

$$m(\tilde{z}(t)) = \int\limits_{-\infty}^{t} e^{-\beta(t-s)} m(\tilde{y}'(s))ds = e^{-\beta t} \int\limits_{-\infty}^{t} e^{\beta s} m(\tilde{y}(s))'ds.$$

Integrating by parts the right-hand side gives

$$m(\tilde{z}(t)) = m(\tilde{y}(t)) - \beta e^{-\beta t} \int\limits_{-\infty}^{t} e^{\beta s} m(\tilde{y}(s))ds.$$

Using Theorem 10 and property 2 from Theorem 2, we write the correlation function at the output as follows:

$$K_{\tilde{z}}(t_1, t_2) = e^{-\beta(t_1+t_2)} \int\limits_{-\infty}^{t_1} \int\limits_{-\infty}^{t_2} e^{\beta(\tau_1+\tau_2)} \frac{\partial^2 K_{\tilde{y}}(\tau_1, \tau_2)}{\partial\tau_1 \partial\tau_2} d\tau_1 d\tau_2,$$

where $K_{\tilde{y}}(\tau_1, \tau_2)$ is the correlation function of the input signal.

Example 4. Consider a linear dynamic system described by the second-order differential equation with constant coefficients

$$\tilde{z}''(t) + a_1\tilde{z}'(t) + a_0\tilde{z}(t) = \tilde{y}(t).$$

Let a continuous fuzzy signal $\tilde{y}(t)$ bounded on the entire real axis be supplied to the input of this system. It is required to find the numerical characteristics of the bounded output fuzzy signal $\tilde{z}(t)$.

Suppose that the coefficients of this equation satisfy the conditions $a_1, a_0 > 0$ and $a_1^2 - 4a_0 > 0$. Then the roots λ_1 and λ_2 of the characteristic equation $\lambda^2 + a_1\lambda + a_0 = 0$ are real and $\lambda_1 < \lambda_2 < 0$. In this case, the Green function G_2 of the problem on bounded solutions of the equation $a_2x'' + a_1x' + a_0x = f(t)$ has the form

$$G_2(t) = \begin{cases} (e^{\lambda_2 t} - e^{\lambda_1 t})(\lambda_2 - \lambda_1)^{-1} & \text{for } t \geqslant 0 \\ 0 & \text{for } t < 0. \end{cases}$$

Then, according to Theorems 8 and 10, the output fuzzy signal $\tilde{z}(t)$ satisfies the relations

$$m(\tilde{z}(t)) = \int\limits_{-\infty}^{t} G_2(t-s) m(\tilde{y}(s))ds,$$

$$K_{\tilde{z}}(t_1, t_2) = \int\limits_{-\infty}^{t_1} \int\limits_{-\infty}^{t_2} G_2(t_1-\tau_1) G_2(t_2-\tau_2) K_{\tilde{y}}(\tau_1, \tau_2) d\tau_1 d\tau_2.$$

Note that the Green functions G_1 and G_2 in Examples 3 and 4 are nonnegative. Hence, the representation (17) holds in these examples.

Example 5. Consider a linear dynamic system described by the third-order differential equation with constant coefficients

$$\tilde{z}'''(t) + a_2\tilde{z}''(t) + a_1\tilde{z}''(t) + a_0\tilde{z}(t) = \tilde{y}(t).$$

Let a continuous fuzzy signal $\tilde{y}(t)$ bounded on the entire real axis be supplied to the input of this system. It is required to find the numerical characteristics of the output fuzzy signal $\tilde{z}(t)$.

Suppose that $a_2 > 0$, $a_1 > 0$, $a_0 > 0$, and $a_2 a_1 - a_0 > 0$. Then, by the Hurwitz criterion, the equation $\lambda^3 + a_2\lambda^2 + a_1 + a_0 = 0$ has the roots with $Re\lambda_i < 0$ $(i = 1, 2, 3)$. Therefore, according to Theorems 8 and 10, the output fuzzy signal $\tilde{z}(t)$ satisfies the relations

$$m(\tilde{z}(t)) = \int_{-\infty}^{t} G_3(t-s)m(\tilde{y}(s))ds,$$

$$K_{\tilde{z}}(t_1, t_2) = \int_{-\infty}^{t_1}\int_{-\infty}^{t_2} G_3(t_1-\tau_1)G_3(t_2-\tau_2)K_{\tilde{y}}(\tau_1, \tau_2)d\tau_1 d\tau_2.$$

Here, $G_3(t)$ is the Green function of the problem on bounded solutions of the equation

$$x'''(t) + a_2x''(t) + a_1x'(t) + a_0x(t) = f(t),$$

which has the form $G_3(t) = \begin{cases} k(t) & \text{for } t \geqslant 0 \\ 0 & \text{for } t < 0, \end{cases}$ where $k(t)$ is the Cauchy function representing the solution of the homogeneous equation

$$k'''(t) + a_2k''(t) + a_1k'(t) + a_0k(t) = 0$$

with the initial conditions

$$k(0) = k'(0) = 0, \quad k''(0) = 1.$$

(For details, see [17, Chapter 2, § 8].)

For example, if the characteristic equation has different roots, the Cauchy function is given by

$$k(t) = C_1e^{\lambda_1 t} + C_2e^{\lambda_2 t} + C_3e^{\lambda_3 t},$$

where

$$C_1 = \frac{1}{(\lambda_1-\lambda_2)(\lambda_1-\lambda_3)}, \quad C_2 = \frac{1}{(\lambda_2-\lambda_3)(\lambda_2-\lambda_1)}, \quad C_3 = \frac{1}{(\lambda_3-\lambda_1)(\lambda_3-\lambda_2)}.$$

6. CONCLUSIONS

The results of Sections 3 and 4 of this paper—the properties of numerical characteristics of fuzzy processes—are similar to the well-known counterparts for continuous random processes. However, despite their significance, they have not been established before.

The main results of this paper concern fuzzy dynamic systems described by nth-order linear differential equations with bounded input fuzzy signals (Section 5). They are based on the new properties of the mean and correlation functions of continuous fuzzy processes (Sections 3 and 4) as well as on the development of the Green function method to the class of fuzzy differential equations.

The approach outlined here is an alternative to the conventional one used to study linear dynamic systems with constant coefficients in terms of frequency response and direct and inverse Fourier transform. Unlike the known approaches, it does not assume stationarity (in any sense) for the processes under consideration. Note that this approach can be extended to continuous processes with fuzzy random states.

REFERENCES

1. Venttsel', E.S. and Ovcharov, L.A., *Teoriya sluchainykh protsessov i ikh inzhenernye prilozheniya* (Theory of Random Processes and Their Engineering Applications), Moscow: Knorus, 2016.
2. Averkin, A.N., *Nechetkie mnozhestva v modelyakh upravleniya i iskusstvennogo intellekta* (Fuzzy Sets in Models of Control and Artificial Intelligence), Moscow: Nauka, 1986.
3. Buckley, J.J., Eslami, E., and Feuring, T., *Fuzzy Mathematics in Economics and Engineering*, Heidelberg–New York: Physica-Verl., 2002.
4. Pegat, A., *Nechetkoe modelirovanie i upravlenie* (Fuzzy Modeling and Control), Moscow: BINOM. Laboratoriya Znanii, 2015.
5. Aumann, R.J., Integrals of Set-Valued Functions, *J. Math. Anal. Appl.*, 1965, no. 12, pp. 1–12.
6. Puri, M.L. and Ralescu, D.A., Differential of Fuzzy Functions, *J. Math. Anal. Appl.*, 1983, vol. 91, pp. 552–558.
7. Kaleva, O., Fuzzy Differential Equations, *Fuzzy Sets and Syst.*, 1987, vol. 24, no. 3, pp. 301–317.
8. Seikkala, S., On the Fuzzy Initial Value Problem, *Fuzzy Sets Syst.*, 1987, vol. 24, no. 3, pp. 319–330.
9. Hukuhara, M., Integration des applications mesurables dont la valeur est un compact convexe, *Func. Ekvacioj.*, 1967, no. 11, pp. 205–223.
10. Khatskevich, V.L., Means, Quasi-scalar Product and Covariance of Fuzzy Numbers, *Journal of Physics: Conference Series*, 2021, vol. 1902(1), art. no. 012136.
11. Park, J.Y. and Han, H., Existence and Uniqueness Theorem for a Solution of Fuzzy Differential Equations, *Int. J. Math. Mathem. Sci.*, 1999, no. 22(2), pp. 271–280.
12. Ahmad, L., Farooq, M., and Abdullah, S., Solving nth Order Fuzzy Differential Equation by Fuzzy Laplace Transform, *Ind. J. Pure Appl. Math.*, 2014, no. 2, pp. 1–20.
13. Mochalov, I.A., Khrisat, M.S., and Shihab Eddin, M.Ya., Fuzzy Differential Equations in Control. Part II, *Information Technologies*, 2015, vol. 21, no. 4, pp. 243–250.
14. Demenkov, N.P., Mikrin, E.A., and Mochalov, I.A., Fuzzy Optimal Control of Linear Systems. Part 1. Positional Control, *Information Technologies*, 2019, vol. 25, no. 5, pp. 259–270.
15. Esmi, E., Sanchez, D.E., Wasques, V.F., and de Barros, L.C., Solutions of Higher Order Linear Fuzzy Differential Equations with Interactive Fuzzy Values, *Fuzzy Sets and Systems*, 2021, vol. 419, pp. 122–140.
16. Daletskii, Yu.L. and Krein, M.G., *Ustoichivost' reshenii differentsial'nykh uravnenii v banakhovom prostranstve* (Stability of Solutions of Differential Equations in a Banach Space), Moscow: Nauka, 1970.
17. Krasnosel'skii, M.A., Burd, V.Sh., and Kolesov, Yu.S., *Nonlinear Almost Periodic Oscillations*, New York: J. Wiley, 1973.
18. Dubois, D. and Prade, H., The Mean Value of Fuzzy Number, *Fuzzy Sets and Syst.*, 1987, vol. 24, no. 3, pp. 279–300.
19. Kaleva, O. and Seikkala, S., On Fuzzy Metric Spaces, *Fuzzy Sets and Systems*, 1984, vol. 12, pp. 215–229.
20. Fuller, R. and Majlender, P., On Weighted Possibilistic Mean Value and Variance of Fuzzy Numbers, *Fuzzy Sets and Systems*, 2003, vol. 136, pp. 363–374.

This paper was recommended for publication by D.V. Vinogradov, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 8, pp. 947–955.

NONLINEAR SYSTEMS

Generalization of the Carathéodory Theorem and the Maximum Principle in Averaged Problems of Non-Linear Programming

A. M. Tsirlin

Ailamazyan Program Systems Institute, Russian Academy of Sciences, Pereslavl-Zalessky, Russia
e-mail: tsirlin@sarc.botik.ru

Received June 9, 2022
Revised March 7, 2023
Accepted June 9, 2023

Abstract—The relationship between the averaging of functions over time and its averaging over the set of values of the required variables is considered. Optimization problems are studied, the criterion and constraints of which include the averaging of functions or functions of the average values of variables. It is shown that the optimality conditions for these problems have the form of the maximum principle, and their optimal solution in the time domain is a piecewise constant function. A generalization of Carathéodory's theorem on convex hulls of a function is proved. Optimality conditions are obtained for non-linear programming problems with averaging over a part of the variables and functions depending on the average values of the variables.

Keywords: averaged constraints, sliding modes, convex hulls of functions, reachability function, maximum principle in averaged problems

DOI: 10.25728/arcRAS.2023.17.50.001

1. INTRODUCTION

For a wide class of problems, the optimality criterion and all or part of the constraints averagely depend on all or part of the variables. Such problems arise when, in technological processes, some variables to be selected must be unchanged (design parameters), while others may change over time, and the presence of devices that smooth out fluctuations, e.g. capacitances, leads to the average influence of these changes [1]. Such problems arise in the optimal control of macrosystems (systems consisting of a set of individually uncontrollable elements), in which it is possible to control only the average parameters of the set of these elements. All such problems are called averaged optimization problems.

In systems, whose set of admissible controls is non-convex (e.g. relay systems), the optimal solution is often a sliding mode, in which the change of the object state depends averagely on any frequently switching control [2–5]. Averaged problems also arise as auxiliary estimation problems in the optimization of cyclic modes, when the introduction of averaging expands the set of admissible solutions and simplifies the solution, allowing to obtain an estimate of the efficiency of the cyclic mode without finding the form of optimal cycles. The value of such an estimation problem is known to be "not worse" than the value of the initial one, and its optimal solution contains useful information about the nature of the optimal solution of the initial one. For definiteness, we will consider problems for the maximum of the optimality criterion.

In the first section of this paper, we will discuss the relationship between the averaging of functions whose argument varies in time over a set of values of that argument and over time, and define what is sought as a solution to the averaged problem and how this solution can be implemented. In the second section, we will formulate the theorem on the optimality conditions of the non-linear programming problem with averaging of the optimality criterion and constraints and

give its proof based on Carathéodory's theorem on convex hulls of functions. In the third section, we will consider possible generalizations of the proved theorem.

2. ON THE RELATIONSHIP BETWEEN TIME AVERAGING AND SET AVERAGING

The mean value of the continuous scalar function $f(x(t))$, $t \in [0, \tau]$, $x \in V \subset R^n$ can be calculated on time as

$$\overline{f_t(x)} = \frac{1}{\tau} \int_0^\tau f(x(t))dt \tag{1}$$

or on set as

$$\overline{f_p(x)} = \int_V f(x)p(x)dx. \tag{2}$$

The function $p(x)$ is called the distribution density. When $x(t)$ is a random function, $p(x)$ is the distribution density of the random variable. It is non-negative and its integral on V is equal to one. In particular, the set V can be a parallelepiped in R^n. In our case, $x(t)$ is a determined function, so let us focus more on the properties of $p(x)$ such that the results of averaging by formulas (1) and (2) are the same.

Let us consider the variable x as scalar, the set V here and below as bounded and closed, and introduce the function $\theta(x_0)$, $x_0 \in V$, equal to the total duration of those time intervals t, for which $x(t) \leqslant x_0$. It is obvious that this function does not exceed τ. Through $P(x_0)$, let us denote the ratio $\frac{\theta(x_0)}{\tau}$, i.e., the fraction of the interval $[0, \tau]$, for which $x(t) \leqslant x_0$. This function grows monotonically as x_0 increases, varying from zero to one. It is similar to the distribution function of a random variable.

The distribution density is equal to

$$p(x_0) = \frac{dP(x_0)}{dx_0} = \frac{1}{\tau}\frac{d\theta(x_0)}{dx_0} = \frac{1}{\tau}\frac{1}{\sum\limits_\nu \left|\frac{dx_\nu}{dt}\right|_{x_\nu = x_0}}. \tag{3}$$

The interval θ increases as x_0 increases for any sign of the derivative at those values x_ν of the function $x(t)$, in which it is equal to x_0.

If at some value of x_0 the function $x(t)$ is constant over a fraction γ of the interval $[0, \tau]$, then the function $P(x_0)$ experiences a jump of magnitude γ at that point, and the distribution density at that point is equal to $\gamma\delta(x - x_0)$.

Examples

1. **Linear functions.** Let $x(t) = \frac{ht}{\tau}$. Then, according to formula (3), we get $p(x) = \frac{1}{h} = \text{const}$. The same distribution density corresponds to all triangles with base $[0, \tau]$ and height h.

2. **Piecewise constant functions.** These functions take discrete values of x_i, each within a fraction γ_i of the interval $[0, \tau]$. Any such function, according to formula (3), corresponds to the distribution density function (3)

$$p(x_0) = \sum_i \gamma_i \delta(x - x_i), \quad \gamma_i > 0, \quad \sum_i \gamma_i = 1. \tag{4}$$

The order, in which the piecewise constant function takes one or another of the possible values, does not matter.

From these examples we see that *every function $x(t)$ corresponds to the distribution density of its values $p(x)$ defined on V, and every distribution density corresponds to any number of functions $x(t)$, for which $\overline{f_p(x)} = \overline{f_t(x)}$.* An exception is the distribution density of the form

$p(x) = \delta(x - x_1)$. In this case, the corresponding function is $x(t) = x_1 = \text{const}$ over the entire interval $[0, \tau]$, and it is unique.

Let us consider the case when the function f depends on several variables (e.g., for the sake of simplicity, on two variables, $x_1(t)$ and $x_2(t)$). In this case, the distribution function $P(x^0)$ of the values of vector x represents the fraction of the interval $[0, \tau]$, for which the two following inequalities are satisfied: $x_1(t) \leqslant x_1^0$ and $x_2(t) \leqslant x_2^0$. This function grows monotonically with the growth of each of the arguments. When the first of the components of the vector x^0 is at the maximum ($p_1(x_1) = 1$), it is equal to and its derivative is equal to the distribution density $p(x_1^{\max}, x_2) = p_2(x_2)$. Similarly, when $x_2 = x_2^{\max}$, $p(x_2^{\max}, x_1) = p_1(x_1)$. The functions $x_1(t)$ and $x_2(t)$ are independent of each other, so $p(x_1, x_2) = p_1(x_1)p_2(x_2)$.

The sought solution to the averaged optimization problem is the distribution density $p^*(x)$ of the vector x on the set V of its admissible values. To implement this solution over time, we need to find one of the possible functions $x(t)$ having the distribution $p^*(x)$. The solution of this last problem is greatly facilitated by the peculiarities of optimal solutions of $p^*(x)$ proved in the next section.

3. ON THE OPTIMAL SOLUTION OF AVERAGED OPTIMIZATION PROBLEMS

We will denote the averaging operation by a line drawn over the function or vector to be averaged. Thus,

$$\overline{x} = \int\limits_V xp(x)dx, \quad \overline{f(x)} = \int\limits_V f(x)p(x)dx.$$

The simplest problem of averaged optimization is the problem of maximizing the average value of a scalar function f(x) at a given average value of its argument:

$$\overline{f(x)} \to \max \Big/ \overline{x} = x_0, \quad x \in V \subset R^n. \tag{5}$$

Or in a more detailed form

$$\int\limits_V f(x)p(x)dx \to \max \Big/ \int\limits_V xp(x)dx = x_0, \quad p(x) \geqslant 0, \quad \int\limits_V p(x)dx = 1. \tag{6}$$

The sought function in this problem is $p(x)$ (the distribution density of the vector of sought variables). This function is non-negative and its integral on the set V is equal to one.

4. CARATHÉODORY'S THEOREM ON CONVEX HULLS OF FUNCTIONS

Carathéodory's theorem [3, 4, 6] on convex hulls of sets states that any element of a convex hull CoD of a compact set D in Euclidean space of dimension n can be represented as an element of a simplex having at most $n + 1$ vertices (base points), each of which belongs to D.

In particular, a subgraph of function $f(x)$ can be the set D. The convex hull of a function is the convex hull of a subgraph. A function depending on n variables is the boundary of a set in R^{n+1} space of dimension n. The basis points are known to lie on this boundary, and hence their number does not exceed $n + 1$. Below we will call Carathéodory's theorem the theorem on convex hulls of functions.

The ordinate of the convex hull of the function $f_0(x)$ at the point x_0 belonging to the convex hull of the set of the function definition is the value of the problem

$$\overline{f_0(x)} \to \max_{p(x)} \Big/ \begin{array}{l} \overline{x_i} = x_{i0}, \quad i = \overline{1, n}, \\ x \in V \subset R^n, \end{array} \tag{7}$$

where V is the compact space.

According to Carathéodory's theorem, the optimal solution of this problem is

$$p^*(x) = \sum_{j=0}^{n} \gamma_j \delta(x - x^j), \quad \gamma_j \geqslant 0, \quad \sum_{i=0}^{n} \gamma_j = 1.$$

That is, the optimal allocation is concentrated in at most $(n+1)$ base points.

This fact allows us to rewrite the problem (7) as a non-linear programming problem

$$\sum_{j=0}^{n} \gamma_j f_0(x^j) \to \max \Bigg/ \begin{array}{l} \sum\limits_{j=0}^{n} \gamma_j x^j = x_0, \\ x^j \in V \subset R^m, \quad \sum\limits_{j=0}^{n} \gamma_j = 1, \quad \gamma_j \geqslant 0, \end{array} \tag{8}$$

whose variables are the basis vectors x^j and the vector of weight coefficients γ, and use the Kuhn–Tucker theorem [7] to solve it:

If y^* is a solution to a non-linear programming problem

$$f(y) \to \max \Big/ \varphi_i(y) \leqslant 0, \quad y_j \geqslant 0, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n, \tag{9}$$

then there is such a non-zero vector of multipliers

$$\lambda = \lambda_0, \ldots, \lambda_m \ \ (\lambda_0 \ \textit{equal to}\ 0 \ \textit{or}\ 1, \ \ \lambda_i \leqslant 0 \ \ \textit{when} \ \ i > 0),$$

that for the Lagrangian function

$$R = \lambda_0 f(y) + \sum_{i=1}^{m} \lambda_i \varphi_i(y)$$

the following conditions are true:

$$\left(\frac{\partial R}{\partial y_j}\right)_{y=y^*} = 0, \ \text{if } y_j^* > 0; \quad \left(\frac{\partial R}{\partial y_j}\right)_{y=y^*} \leqslant 0, \ \text{if } y_j^* = 0; \tag{10}$$

$$\lambda_i = 0, \ \text{if } \varphi_i(y^*) < 0; \quad \lambda_i \leqslant 0, \ \text{if } \varphi_i(y^*) = 0. \tag{11}$$

For problem (8) the Lagrangian function takes the form

$$R = \sum_{j=0}^{n} \gamma_j \left[f_0(x^j) + \sum_{i=1}^{n} \lambda_i x_i^j - \Lambda \right], \tag{12}$$

where Λ is the Lagrange multiplier corresponding to the condition of equality of the sum of weight coefficients to one.

Kuhn–Tucker conditions on weighting factors lead to requirements:

$$R^0(x_j, \lambda) = f_0(x^j) + \sum_{i=1}^{n} \lambda_i x_i^j < \Lambda, \ \text{if } \gamma_j = 0, \tag{13}$$

$$R^0(x_j, \lambda) = f_0(x^j) + \sum_{i=1}^{n} \lambda_i x_i^j = \Lambda, \ \text{if } \gamma_j > 0, \quad j = 0, \ldots, n+1.$$

Here, R^0 is the Lagrangian function of problem (8) without averaging. Hereafter such a problem will be called the *initial* one.

Thus, *for all base values of x included in the optimal solution of the convex hull problem of the function f_0 with non-zero weight, the Lagrangian function of the original problem is maximal. The number of such points does not exceed $n+1$.*

5. PROBLEM WITH BOND AVERAGING, GENERALIZATION OF CARATHÉODORY'S THEOREM

In the non-linear programming problem with averaging functions defining relations between variables, it is required to maximize the average value of the function $f_0(x)$ on the set V of admissible values of x, provided that the average value of the vector function $f(x) = (f_1(x), \dots, f_i(x), \dots, f_m(x))$ is equal to zero. Formally,

$$\overline{f_0(x)} \to \max \Big/ \overline{f_i(x)} = 0, \quad i = 1, \dots, m, \quad x \in V \in R^n. \tag{14}$$

Theorem 1. 1. *The optimal distribution density in problem* (14) *has the form*

$$p^*(x) = \sum_{j=0}^{m} \gamma_j \delta(x - x^j), \quad \gamma_j \geqslant 0, \ \sum_{j=0}^{m} \gamma_j = 1. \tag{15}$$

2. *There is such a non-zero vector*

$$\lambda = \lambda_0, \dots, \lambda_i, \dots, \lambda_m, \quad \lambda_0 = (0; 1),$$

that, at each base point x^j, the Lagrangian function of the original problem

$$R = \sum_{i=0}^{m} \lambda_i f_i(x) \tag{16}$$

is maximal over $x \in V$.

Proof. To prove this statement, we will introduce the concept of the *reachability function* of the problem (14):

$$f_0^*(C) = \max f_0(x) \Big/ f_k(x) = C_k, \quad k = 1, \dots, m, \quad x \in V. \tag{17}$$

This function is defined algorithmically on the set

$$V_c = \{C \in R^m : f(x) = C, x \in V \subset R^n\}.$$

It may be non-smooth and semi-continuous on top.

The following statement is true.

Statement. *For those values of x, for which $f(x) = C$, $p^*(x)$ is deliberately equal to zero if $f_0(x) \neq f_0^*(C)$.*

Thus, only those values $x = x^*(C)$, for which the value of $f_0(x)$ coincides with the ordinate of the reachability function, can be included in the solution of the averaged problem with non-zero weight. If this statement were not true, it would be possible to change the density of the distribution so that the average value of $f_0(x)$ would increase.

Since for each C the value of f_0 coincides with the ordinate of the reachability function, the problem (14) can be rewritten as

$$\overline{f_0^*(C)} \to \max \Big/ \overline{C_k} = 0, \quad k = 1, \dots, m, \quad C \in V_c \subset R^m. \tag{18}$$

This is the problem on the ordinate of the convex hull of the reachability function at zero. According to Carathéodory's theorem, its optimal solution is equal to

$$p^*(C) = \sum_{j=0}^{m} \gamma_j \delta(C - C^j), \quad \gamma_j \geqslant 0, \ \sum_{j=0}^{m} \gamma_j = 1. \tag{19}$$

Since each base value of C^j corresponds to the value of $x^{j*}(C^j)$, the optimal distribution density in problem (14) is of the form (15). The first statement of Theorem 1 is proved.

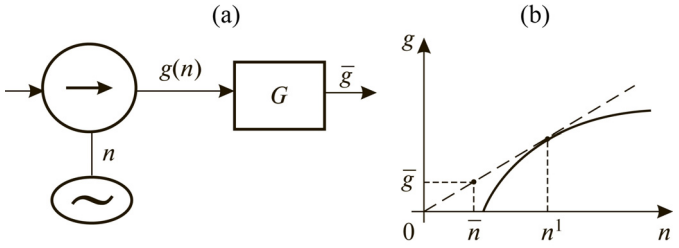


Fig. 1. System consisting of a pump and a smoothing tank (a); relation between flow rate and power input (b).

The proof of the second statement completely repeats the analogous proof for the problem on the ordinate of the convex hull of a function with the difference that the Lagrangian function of the non-averaged problem has the form (16). We emphasize that the number of base points does not depend on the dimensionality of the vector x, but is determined by the dimensionality m of the vector function f.

Note that here and below conditions in the form of the maximum principle do not require the functions defining the averaged problem to be smooth on x, the set V can be non-contiguous [8–10].

Example 1. Let us consider the system consisting of an electric motor, a pump rotated by it and a vessel in Fig. 1a. The motor consumes power n, on which depends the pump capacity g. The dependence of $g(n)$ is shown in Fig. 1b. It is required to find the mode for which, for a given average power input $\overline{n}$, the average pumping capacity $\overline{g}$ is maximized. This is the problem of the ordinate of the convex hull of the function $g(n)$ at the point $\overline{n}$. The number of base points is two, one of them is the origin of coordinates, and the second one, n^1, is defined by the condition that the Lagrangian function $R = g(n) + \lambda n$ reaches the maximum in it (the same as at $n = 0$). Excluding λ from the conditions for the maximum of the Lagrangian function and the requirement that this maximum be zero, we reach the equation for n^1:

$$\frac{g(n)}{n} = \frac{dg(n)}{dn}.$$

There are many optimal implementations of this solution over time, and for each of them the pump power takes values zero and n^1, and the fraction of the interval τ, for which $n = n^1$, is equal to $1 - \frac{\overline{n}}{n^1}$. The maximum value of the interval τ is determined by the value of capacitance G, it is equal to

$$\tau_{\max} = \frac{2G}{g(n^1)}.$$

The value of the problem is equal to

$$\overline{g}^* = g(n^1)\left(1 - \frac{\overline{n}}{n^1}\right).$$

It does not depend on G, and the sliding mode is the optimal solution when the capacitance goes down to zero.

6. GENERALIZATIONS OF THE AVERAGED NON-LINEAR PROGRAMMING PROBLEM

6.1. Averaged Problem with Deterministic Variables

As mentioned in the introduction, there can be two types of variables in averaged problems: randomized and deterministic. There is no averaging for variables of the second type. Let us consider a non-linear programming problem, in which some variables are not averaged.

The problem with averaging over a part of variables will take the form:

$$\overline{f_0(x,y)} \to \max \Big/ \overline{f_j(x,y)} = 0, \quad x \in V \subset R^n, \quad y \in V_y \subset R^K, \quad j = 1, \ldots, m, \tag{20}$$

functions $f_0, \ldots, f_m$ are continuous and continuously differentiable over the set of arguments, the line corresponds to averaging over $x \in V$, the sets V and V_y are closed and bounded.

For any y, this problem is an averaged non-linear programming problem (14), and hence, due to the theorem, the optimal distribution density x is concentrated in at most $(m+1)$ base points, so that $p^*(x) = \sum_0^m \gamma_j \delta(x - x^j)$ and there exists such a non-zero vector λ that, at each of these points, the Lagrangian function of the original problem

$$R = \sum_{j=0}^{m} \lambda_j f_j(x,y), \quad x \in V \subset R^n, \quad y \in V_y \subset R^K \tag{21}$$

is maximal on x.

The Lagrangian function of the problem (20), in which the distribution density x is equal to $p^*(x)$, has the form

$$\overline{R^*} = \sum_{j=0}^{m} \lambda_j \sum_{i=0}^{m} \gamma_i f_j(x^i, y), \quad x^i \in V \subset R^n, \quad y \in V_y \subset R^K. \tag{22}$$

For any distribution density of randomized variables $p(x)$, the problem (20) is a non-linear programming problem and, according to Kuhn–Tucker theorem, there is such a non-zero vector λ with components $\lambda_0 = (0;1)$, λ_j, $j = 1, \ldots, m$, that the conditions of local non-improvability on y are satisfied for the function (21) at the optimal solution

$$\frac{\partial \overline{R^*}}{\partial y_l} \delta y_l \leqslant 0, \quad l = 1, \ldots, K. \tag{23}$$

Here, δy_l is the acceptable variation of y_l.

The non-linear programming problem with averaging over a part of variables has much in common with the optimal control problem with links in the form of differential equations. There, the control actions enter the problem in such a way that their fast changes are averaged in the neighborhood of each time instant, which cannot be said about the phase coordinates. That is why the conditions in the form of Pontryagin's maximum principle are valid for the control actions.

6.2. Problem Containing Functions of Mean Values of Variables

This problem has the form

$$\overline{f_0(x, \overline{x_l})} \to \max \Big/ \overline{f_j(x, \overline{x_l})} = 0, \quad x \in V \subset R^n, \quad l = 1, \ldots, \quad K \leqslant n. \tag{24}$$

Let us introduce the notation: $y_l = \overline{x_l}$. The variable y_l belongs to the convex hull CoV_{x_l} of the set of admissible values x_l. Given the introduced notations, the problem (24) can be rewritten as

$$\overline{f_0(x,y)} \to \max \Big/ \overline{f_j(x,y)} = 0, \quad \overline{x_l} - y_l = 0, \quad x \in V \subset R^n, \quad y_l \in CoV_{x_l} \subset R^K. \tag{25}$$

When written in this form, problem (25) differs from problem (20) only by additional averaged conditions $\overline{x_l} - y_l = 0$. The Lagrangian function of the original problem will take the form

$$R = \sum_{j=0}^{m} \lambda_j f_j(x,y) + \sum_{l=1}^{K} \lambda_l (x_l - y_l), \quad x \in V \subset R^n, \quad y_l \in CoV_{x_l}. \tag{26}$$

From the optimality conditions (21), (23) it follows that the maximum number of base values of x in problem (24) is $m + K + 1$ and that there is such a non-zero vector λ that at each of the base points the function R appearing in (26) reaches a maximum on the optimal solution on x, while, on y, the function (22) is locally non-improvable.

When solving averaged problems, the Lagrange multipliers are expressed through the base values x^j and y from the condition of maximum of the Lagrangian function on x and equality of these maxima to each other, as well as the conditions of non-improvability on y are written down. After that, from the averaged conditions, the weight coefficients for each of the base points are found, given that the sum of these weight coefficients is equal to one.

Example 2. As an illustrative example, let us consider the following problem

$$\overline{(x-\overline{x})^2} \to \min \Big/ \overline{\left(\frac{1}{x+\overline{x}}\right)} = 1, \quad x = -1; 0; 1. \tag{27}$$

The Lagrangian function for this problem is equal to

$$L = (x-y)^2 + \lambda\left(\frac{1}{x+y-1}\right) + \mu(y-x). \tag{28}$$

The number of averaged conditions is two, hence all three admissible values of x are basic, and the uncertain multipliers must be chosen so that the maximum of the function L^* is the same at these points, which leads to the following conditions:

$$\begin{aligned} L^* &= (1+y)^2 + \lambda\left(\frac{1}{y-1} - 1\right) - \mu(1+y) = y^2 + \lambda\left(\frac{1}{y} - 1\right) - \mu y \\ &= (1-y)^2 + \lambda\left(\frac{1}{y+1} - 1\right) + \mu(1-y). \end{aligned} \tag{29}$$

Thus,

$$\lambda = \frac{2y}{2+y}(1-y-2y^2), \quad \mu = \frac{y(2y-1)}{2+y}. \tag{30}$$

After substituting these expressions into L^* and differentiating the resulting expression by y, we come to the equation

$$3y^3 + 17y^2 + 20y - 10 = 0. \tag{31}$$

To the second decimal place, $y = 0.37$. Weight multipliers $\gamma_1, \gamma_2, \gamma_3$ for $x = -1$, $x = 0$, $x = 1$, respectively, can now be found from the following conditions

$$\sum_{i=1}^{3}\gamma_i = 1, \quad \sum_{i=1}^{3}\gamma_i x_i = 0.37, \quad \sum_{i=1}^{3}\gamma_i \frac{1}{x_i + 0.37} = 1. \tag{32}$$

Then we get $\gamma_1 = 0.155$, $\gamma_2 = 0.320$, $\gamma_3 = 0.525$.

7. CONCLUSION

Various formulations of non-linear programming problems with averaging were considered. It is shown that, with the introduction of the concept of reachability function of non-linear programming problems, the problems containing averaging of functions from a vector of randomized variables x can be reduced to extremal problems on convex hulls of sets and functions. The optimal distribution in all these problems is centered in discrete "base" points of the compact set V of admissible values

of x. The maximum principle for such problems is proved. It is shown that the number of base points does not exceed the number of averaged conditions in the problem by more than one. The criterion and constraints of averaged non-linear programming problems may depend on time. If these dependencies are continuous, then the above optimality conditions are valid for each moment of time and determine the time variation of the coordinates of the base points and their weights. The vector of Lagrange indeterminate multipliers corresponding to the optimal solution delivers the minimum to the maximum value of the maximized function with respect to the sought variables, which serves as a basis for computational algorithms.

FUNDING

This work was supported by the Russian Science Foundation, project no. 20-61-46013.

REFERENCES

1. Tsirlin, A.M., *Optimal'nye tsikly i tsiklicheskie rezhimy* (Optimal Cycles and Cycling Modes), Moscow: Energoatomizdat, 1983.
2. Rozonoer, L.I., Pontryagin Maximum Principle in the Theory of Optimal Systems, *Autom. Remote Control*, 1959, no. 10, pp. 1320–1334; no. 11, pp. 1441–1458; no. 12, pp. 1561–1578.
3. Yudin, D.B., *Matematicheskie metody upravleniya v usloviyakh nepolnoi informatsii* (Mathematical Methods of Management under Conditions of Incomplete Information), Moscow: Sovetskoe Radio, 1974.
4. Yang, L., *Lectures on the Calculus of Variations and Optimal Control Theory*, London: Saunders, 1969. Translated under the title *Lektsii po variatsionnomu ischisleniyu i optimal'nomu upravleniyu*, Moscow: Mir, 1974.
5. Tsirlin, A.M., Optimization in Mean and Sliding Modes in Optimal Control Problems, *Izv. Akad. Nauk SSSR. Tekhn. Kibernetika*, 1974, no. 2, pp. 143–151.
6. Polovinkin, E.S. and Balashov, M.V., *Elementy vypuklogo i sil'no vypuklogo analiza* (Elements of Convex and Strongly Convex Analysis), Moscow: Fizmatlit, 2004. ISBN 5-9221-0499-3.
7. Himmelblau, D.M., *Applied Nonlinear Programming*, New York: McGraw-Hill, 1972. Translated under the title *Prikladnoe nelineinoe programmirovanie*, Moscow: Mir, 1975.
8. Afanas'ev, A.P., Dikusar, V.V., Milyutin, A.A., and Chyukanov, S.A., *Neobkhodimoe uslovie v optimal'nom upravlenii* (Necessary Condition in Optimal Control), Moscow: Nauka, 1990.
9. Dubovitskii, A.Ya. and Milyutin, A.A., *Teoriya printsipa maksimuma. Metody teorii ekstremal'nykh zadach v ekonomike* (Theory of the Maximum Principle. Methods of the Theory of Extreme Problems in Economics), Moscow: Nauka, 1981.
10. Tsirlin, A.M., Optimality Conditions of Sliding Modes and the Maximum Principle for Control Problems with the Scalar Argument, *Autom. Remote Control*, 2009, no. 5, pp. 839–854.

This paper was recommended for publication by P.S. Shcherbakov, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 8, pp. 956–966.

STOCHASTIC SYSTEMS

Parametric Algorithm for Finding a Guaranteed Solution to a Quantile Optimization Problem

S. V. Ivanov[*,a], **A. I. Kibzun**[*,b], **and V. N. Akmaeva**[*,c]

[*] *Moscow Aviation Institute (National Research University), Moscow, Russia*
e-mail: [a] *sergeyivanov89@mail.ru,* [b] *kibzun@mail.ru,* [c] *akmaeva@mai.ru*

Received January 30, 2023
Revised May 15, 2023
Accepted June 9, 2023

Abstract—The problem of stochastic programming with a quantile criterion for a normal distribution is studied in the case of a loss function that is piecewise linear in random parameters and convex in strategy. Using the confidence method, the original problem is approximated by a deterministic minimax problem parameterized by the radius of a ball inscribed in a confidence polyhedral set. The approximating problem is reduced to a convex programming problem. The properties of the measure of the confidence set are investigated when the radius of the ball changes. An algorithm is proposed for finding the radius of a ball that provides a guaranteeing solution to the problem. A method for obtaining a lower estimate of the optimal value of the criterion function is described. The theorems are proved on the convergence of the algorithm with any predetermined probability and on the accuracy of the resulting solution.

Keywords: stochastic programming, quantile criterion, confidence method, quantile optimization, guaranteeing solution

DOI: 10.25728/arcRAS.2023.74.92.001

1. INTRODUCTION

Stochastic programming problems with a quantile criterion are optimization problems in which the minimum point of the quantile of the loss function is sought, depending on the optimization strategy and random parameters. Similar problems arise when modeling technical and economic systems, in which the requirements for the reliability of the decision being made play an important role. The quantile function describes the level of loss that cannot be exceeded with a given fixed probability, usually close to one. The monographs [1, 2] are devoted to problems of this class.

An effective way to solve the problem of minimizing the quantile function is the confidence method [1, 2]. The essence of this method is that the original quantile optimization problem is approximated by a minimax problem. In this problem, we first consider the maximum of the objective function on a certain set of values of random parameters (confidence set) as a function of the confidence set and the optimization strategy. Then, the minimum of the obtained maximum function is searched for by the optimization strategy and the confidence set. The choice of the optimal confidence set is not an easy task. However, with a properly chosen fixed confidence set, one can obtain a fairly accurate upper estimation of the quantile function. In particular, it is shown [2], that for a Gaussian distribution of random factors, the choice of a confidence set in the form of a ball for large values of the reliability level ensures high accuracy of the resulting estimate. This article discusses the loss functions that are presented as the maximum of a finite number of linear (with respect to random parameters) functions. For this class of loss functions, the optimal confidence set is a polyhedron. In this regard, the estimate on the ball can be improved by performing an additional optimization over the class of confidence sets in the form of polyhedra,

parametrized by the radius of the inscribed ball. This idea was implemented for the Gaussian distribution in [3]. In [4], this algorithm was extended to the case of an arbitrary distribution of random factors, and an algorithm was proposed for further improving the guaranteeing solution by moving the faces of a convex polyhedral confidence set while maintaining its probability measure. It should be noted that in [3, 4] the loss function was assumed to be linear in the optimization strategy. This allowed the approximating minimax problem to be reduced to a linear programming problem.

A feature of the approximating problem obtained by using the algorithms [3, 4] is the fact that in the case of a Gaussian distribution, it can be used to obtain not only the upper, but also the lower estimate of the optimal value of the quantile function. To do this, in the approximating problem, instead of the confidence set, take the kernel of the probability measure [2], which, in the case of a standard Gaussian distribution, is a ball of radius calculated as a quantile of the standard normal distribution of the same level as the quantile function. It should be noted that the kernel of a probability measure is not a confidence set.

Of special interest is the case of a loss function that is linear in random parameters. In [1] it is proved that, under the condition of regularity of the kernel, the quantile function can be calculated as a maximum by random parameters of the loss function on the core. Later, the regularity conditions for the kernel were loosen in [5]. The said kernel property was used in [6] to construct an algorithm for solving a stochastic programming problem with a quantile criterion and a bilinear loss function, as well as in [7] for approximating probabilistic constraints.

Stochastic programming problems with a quantile criterion are a special case of problems with probabilistic constraints [8, 9]. The review of methods for solving problems with probabilistic constraints can be found in [10]. In particular, we should note the approach based on the use of p-efficient points [11, 12]. However, problems with a quantile criterion have a number of properties that are not characteristic of problems with arbitrary probabilistic constraints, which makes it possible to use special methods of analysis, in particular, the confidence method. Problems with a quantile criterion and additional probabilistic constraints were studied in detail in [1].

This article considers a stochastic programming problem with a loss function that is piecewise linear in random parameters and convex in terms of the optimization strategy, which makes it possible to approximate the problem under study by a convex programming problem. For this problem, an algorithm is developed based on the ideas of constructing algorithms in [3, 4] for piecewise linear problems. Estimates are given for the accuracy of the proposed algorithm.

2. FORMULATION OF THE PROBLEM

Let X be the random vector (column) with realizations $x \in \mathbb{R}^m$, given on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. It is assumed that the distribution X is standard normal. We assume that the loss function Φ is piecewise linear in random parameters:

$$\Phi(u,x) \triangleq \max_{i=\overline{1,k_1}} \{B_{1i}(u)x + b_{1i}(u)\}.$$

The constraints in the problem are described by the function

$$Q(u,x) \triangleq \max_{j=\overline{1,k_2}} \{B_{2j}(u)x + b_{2j}(u)\},$$

where $u \in U \subset \mathbb{R}^n$ is the strategy; $B_{1i}(u)$, $B_{2j}(u)$ are rows of matrices $B_1(u)$, $B_2(u)$ respectively, $b_{1j}(u)$, $i = \overline{1,k_1}$, and $b_{2j}(u)$, $j = \overline{1,k_2}$, are elements of vectors (columns) $b_1(u)$ and $b_2(u)$ respectively. This article assumes that the functions $u \mapsto B_1(u)$, $u \mapsto B_2(u)$ are linear (i.e., $B_l(u) = D_l u + a_l$, where D_l is a matrix, a_l is a vector, $l \in \{1,2\}$), and functions $u \mapsto b_1(u)$, $u \mapsto b_2(u)$ are convex and

continuous on a convex closed set U. Note that the linear transformation of the random vector X does not change the structure of the functions Φ and Q. Moreover, any normal vector can be obtained by a linear transformation of the vector X of suitable dimension. For these reasons, the case of an arbitrary normal distribution of the vector X reduces to the case under consideration.

Define the probability function as

$$P_\varphi(u) \triangleq \mathbf{P}\{\Phi(u,X) \leqslant \varphi, \quad Q(u,X) \leqslant 0\},$$

where $\varphi \in \mathbb{R}$ is a given value of the loss function, and the quantile function as

$$\Phi_\alpha(u) \triangleq \min\{\varphi \mid P_\varphi(u) \geqslant \alpha\}, \quad \alpha \in (0, P^*),$$

where

$$P^* \triangleq \sup_{u \in U} \mathbf{P}\{Q(u,X) \leqslant 0\}.$$

The article considers the problem of quantile optimization

$$U_\alpha \triangleq \operatorname{Arg} \min_{u \in U} \Phi_\alpha(u). \tag{1}$$

Since the functions Φ and Q are continuous and measurable, according to the result of [13, Theorem 6], which is a generalization of a similar result in [1], the function $u \mapsto \Phi_\alpha(u)$ is lower semicontinuous. Therefore, a solution to the problem (1) exists if the set U is compact. Let us determine the optimal value of the criterion function as

$$\varphi_\alpha \triangleq \Phi_\alpha(u_\alpha),$$

where $u_\alpha \in U_\alpha$. In what follows, we will assume that a solution to the problem (1) exists. In this case, the boundedness of the set U, generally speaking, is not required.

3. CONSTRUCTION OF SOLUTION ESTIMATES

According to the confidence method, [1] the problem (1) is equivalent to

$$\varphi_\alpha = \min_{S \in \mathcal{E}_\alpha, u \in U} \left\{ \sup_{x \in S} \Phi(u,x) \mid \sup_{x \in S} Q(u,x) \leqslant 0 \right\}, \tag{2}$$

where $\mathcal{E}_\alpha$ is the family of all confidence sets $S \subset \mathbb{R}^m$ of level α, i.e. Borel sets such that $\mathbf{P}\{X \in S\} \geqslant \alpha$.

Denote by B_r the ball of radius r:

$$B_r \triangleq \{x \in \mathbb{R}^m \mid \|x\| \leqslant r\},$$

where $\|x\| \triangleq \sqrt{x^\top x}$ is the Euclidean norm of the vector x.

Let us consider a problem similar to the problem (2), in which the set $S = B_r$ is fixed:

$$\psi(r) \triangleq \min_{u \in U} \left\{ \max_{x \in B_r} \Phi(u,x) \mid \max_{x \in B_r} Q(u,x) \leqslant 0 \right\}. \tag{3}$$

We will assume that the minimum in u in problem (3) is reached, which is true, for example, in the case of compact set U. In the problem (3) the supremum is replaced by the maximum, because

$$\max_{x \in B_r} \Phi(u,x) = \max_{x \in B_r} \max_{i=\overline{1,k_1}} \{B_{1i}(u)x + b_{1i}(u)\}$$

$$= \max_{i=\overline{1,k_1}} \max_{x \in B_r} \{B_{1i}(u)x + b_{1i}(u)\} = \max_{i=\overline{1,k_1}} \{b_{1i}(u) + \|B_{1i}(u)\| r\}.$$

In a similar way is $\max\limits_{x\in B_r} Q(u,x)$. Thus, the problem (3) can be rewritten as

$$\psi(r)=\min_{u\in U}\left\{\max_{i=\overline{1,k_1}}\{b_{1i}(u)+\|B_{1i}(u)\|r\}\mid \max_{j=\overline{1,k_2}}\{b_{2j}(u)+\|B_{2j}(u)\|r\}\leqslant 0\right\}. \tag{4}$$

If the constraints of this problem are inconsistent, we will assume that $\psi(r)=+\infty$. From the monotonic nondecreasing of the objective function and the narrowing of the set of admissible strategies as r increases it follows that the function ψ is non-decreasing. Problem (4) is equivalent to the convex programming problem

$$\varphi\to\min_{u\in U,\,\varphi\in\mathbb{R}} \tag{5}$$

under constraints

$$\begin{aligned} &b_{1i}(u)+\|B_{1i}(u)\|r\leqslant\varphi,\quad i=\overline{1,k_1},\\ &b_{2j}(u)+\|B_{2j}(u)\|r\leqslant 0,\quad j=\overline{1,k_2}. \end{aligned}$$

Equivalence is understood here in the sense that the optimal value of the variable φ coincides with $\psi(r)$, and the sets of admissible values u coincide. The total number of constraints in this problem will be denoted by $k=k_1+k_2$. Problem (5) can be solved with high accuracy using convex optimization methods [14].

Let R_α be the ball of probabilistic measure α, i.e. the solution of the equation

$$\mathbf{P}\{X\in B_{R_\alpha}\}=\alpha.$$

Let us fix in the problem (2) a confidence set S in the form of a ball B_{R_α}. Thus, an upper estimate of the quantile function can be found.

To search for a lower estimate, the kernel of the probability measure can be used, defined as the intersection of all closed half-spaces A such that $\mathbf{P}\{X\in A\}=\alpha$. It is known that for $\alpha>\frac{1}{2}$ the kernel of the distribution of the standard normal Gaussian vector is a ρ_α-radius ball centered at zero, where ρ_α is the quantile of the standard normal distribution of the α level. In [1, Section 3.4.3, Corollary 2] it is shown that $\psi(\rho_\alpha)\leqslant\varphi_\alpha$, when X distributed normally.

Thus, we have obtained the estimate

$$\psi(\rho_\alpha)\leqslant\varphi_\alpha\leqslant\psi(R_\alpha). \tag{6}$$

The upper estimate for $\psi(R_\alpha)$ can be improved. Let $(u(r),\psi(r))$ be some solution to the problem (5). Let us define the set

$$\begin{aligned} C_r&\triangleq\left\{x\in\mathbb{R}^m\mid \Phi(u(r),x)\leqslant\psi(r),\ Q(u(r),x)\leqslant 0\right\}\\ &=\left\{x\in\mathbb{R}^m\mid B_{1i}(u(r))x+b_{1i}(u(r))\leqslant\psi(r),\ B_{2j}(u(r))x+b_{2j}(u(r))\leqslant 0,\ i=\overline{1,k_1},\ j=\overline{1,k_2}\right\}. \end{aligned} \tag{7}$$

We introduce the notation $h(r)\triangleq\mathbf{P}\{X\in C_r\}$ for the probability measure of the set C_r. Note that $h(r)$ and C_r depend on the choice of $u(r)$. Therefore, in what follows, the choice of $u(r)$ is assumed to be fixed.

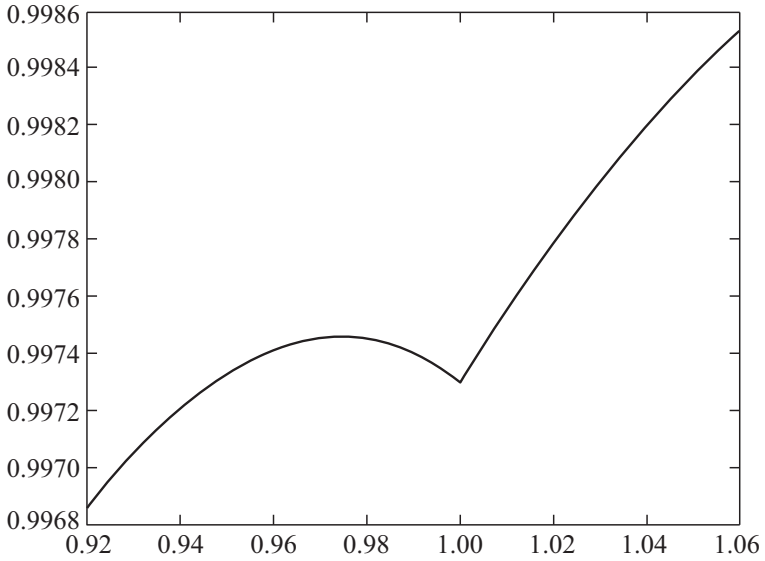
Because

$$\max_{x\in B_r}\Phi(u(r),x)=\psi(r),\quad \max_{x\in B_r}Q(u(r),x)\leqslant 0, \tag{8}$$

the inclusion $B_r\subset C_r$ is valid. Besides,

$$\max_{x\in B_r}\Phi(u(r),x)=\max_{x\in C_r}\Phi(u(r),x),\quad \max_{x\in C_r}Q(u(r),x)\leqslant 0. \tag{9}$$

It follows from (8) and (9) that if $h(r)\geqslant\alpha$, then C_r is a confidence set and $\psi(r)\geqslant\varphi_\alpha$.



Dependency graph for $h(r) = \mathbf{P}\{X \in C_r\}$ of r.

It follows from the monotonicity of ψ that the upper bound for the quantile function can be improved by finding r close to $r^* \triangleq \inf\{r \mid h(r) \geqslant \alpha\}$, such that $h(r) \geqslant \alpha$. If the function $r \mapsto h(r)$ is monotonic, then the dichotomy method can be used to find r^* . Unfortunately, the function h can be non-monotone, as the following example demonstrates.

Example 1. Let the loss function be

$$\Phi(u, x) = \max\{u + 4x, -u + 2x + 2, -11u - 4x\},$$

$u \in \mathbb{R}$, x is a realization of a random variable $X \sim \mathcal{N}(0, \sigma^2)$, $\sigma^2 = \frac{1}{9}$.

It is easy to check that the problem (5) has a solution

$$u(r) = 1 - r, \quad \psi(r) = 1 + 3r \quad \text{if} \quad r \in [0, 1];$$
$$u(r) = 0, \quad \psi(r) = 4r \quad \text{if} \quad r \in [1, +\infty).$$

Therefore,

$$C_r = \{x \mid \Phi(u(r), x) \leqslant \psi(r)\} = \begin{cases} [-3 + 2r, r], & \text{if } r \in [0, 1], \\ [-r, r], & \text{if } r \in [1, +\infty). \end{cases}$$

Let us calculate the measure of the set C_r if $r \in [0, 1]$:

$$h(r) = \mathbf{P}\{X \in C_r\} = \int\limits_{-3+2r}^{r} \frac{3}{\sqrt{2\pi}} e^{-\frac{3x^2}{2}} dx.$$

Let us calculate the derivative of the obtained function:

$$\frac{dh}{dr}(r) = \frac{3}{\sqrt{2\pi}} e^{-\frac{3r^2}{2}} - 2\frac{3}{\sqrt{2\pi}} e^{-\frac{3(2r-3)^2}{2}}.$$

Let us calculate the left-hand limit

$$\lim_{r \to 1-} \frac{dh}{dr}(r) = -\frac{3}{\sqrt{2\pi}} e^{-\frac{3}{2}} < 0.$$

This means that on some interval $(1 - \varepsilon, 1)$, where $\varepsilon > 0$, function h is decreasing. Moreover, $h(1) \approx 0.9973$. The dependence graph for $h(r)$ is shown in the figure.

Table 1. Dependence of R_α on m

$\alpha \setminus m$	1	2	3	4	5	6	7	8	9	10	50
0.95	1.96	2.45	2.80	3.08	3.32	3.55	3.75	3.94	4.11	4.28	8.22
0.99	2.58	3.03	3.37	3.64	3.88	4.10	4.30	4.48	4.65	4.82	8.73

Table 2. Dependence of ρ_β on k

$\alpha \setminus k$	1	2	3	4	5	6	7	8	9	10	50
0.95	1.64	1.96	2.13	2.24	2.33	2.39	2.45	2.50	2.54	2.58	3.09
0.99	2.33	2.58	2.71	2.81	2.88	2.93	2.98	3.02	3.06	3.09	3.54

As can be seen from the example above, the function h may turn out to be nonmonotone. In this connection, we propose sufficient conditions that ensure the monotonicity of the function h.

Theorem 1. *Let $U = \mathbb{R}^n$ and the conditions are fulfilled:*

1) *$b_{1i}(u) = A_{1i}u + c_{1i}$, A_{1i} are the rows of the matrix A_1, $b_{2j}(u) = A_{2j}u + c_{2j}$, A_{2j} are the rows of the matrix A_2, matrices $B_1(u)$ and $B_2(u)$ do not depend on u;*

2) *the rows of the block matrix*

$$\begin{pmatrix} A_1 & e_{k_1} \\ A_2 & 0_{k_2} \end{pmatrix}$$

are linearly independent, where $e_{k_1}, 0_{k_2}$ are the columns of ones and zeros respectively (if $Q(u,x) \equiv 0$, then there are no rows corresponding to A_2 in the above matrix);

3) *for some $r = R$ the solution to the problem (5) exists.*

Then the function h is non-decreasing on the interval $[0, R]$.

The proof of the 1 and all subsequent theorems are in the Appendix.

Note that in Theorem 1 the set U is not compact. Unfortunately, it is difficult to propose more general conditions for the monotonicity of the function h since the monotonicity measures can only be guaranteed under the assumption that the set C_r expands as r increases. However, only the distance from the origin of the faces touching the ball B_r can be guaranteed. The remaining faces can both move away and approach the origin of coordinates.

In connection with the nonmonotonicity of the function h it is necessary to indicate as accurately as possible the interval in which it is necessary to look for r^*. For this we get the following result.

Theorem 2. *Let $k = k_1 + k_2$. The inequality $h(r) \geqslant \alpha$ holds if $r \geqslant \rho_\beta$ and the set C_r is defined, where ρ_β is the quantile of standard normal distribution of the level $\beta = 1 - \frac{1-\alpha}{k}$.*

From the Theorem 2 and the inequality (6) it follows that

$$\psi(\rho_\alpha) \leqslant \varphi_\alpha \leqslant \min\{\psi(R_\alpha), \psi(\rho_\beta)\} = \psi\left(\min\{R_\alpha, \rho_\beta\}\right). \tag{10}$$

It follows from the definition of a confidence ball that $R_\alpha = \sqrt{\chi^2_\alpha(m)}$, where $\chi^2_\alpha(m)$ is the chi-square distribution quantile with m degrees of freedom. In contrast to R_α the value of ρ_β does not depend on the dimension of the random vector, but depends only on the number of constraints k. It is known [2], that $R_\alpha - \rho_\alpha \to 0$ for $\alpha \to 1$, but the rate of convergence depends on dimension n. It is easy to see that $\rho_\beta \to +\infty$ for $k \to 1$. However, it turns out that for small values k the inequality $\rho_\beta < R_\alpha$ can be satisfied. The dependence of R_α on m is given in Table 1, and the dependence of ρ_β on k is given in Table 2. The levels $\alpha = 0.95$ and $\alpha = 0.99$ are considered. Let, for example, $m = 8$, $\alpha = 0.95$. Then $R_\alpha = 3.94$, and $\rho_\beta < R_\alpha$ even for $k = 50$.

Note that for $k = 1$ we have the equality $\rho_\beta = \rho_\alpha$. Therefore, $\varphi_\alpha = \psi(\rho_\alpha)$, and the optimal strategy u_α can be found from problem (5) for $r = \rho_\alpha$, which agrees with the known result [5].

4. ALGORITHM FOR SEARCHING FOR A GUARANTEEING SOLUTION

A strategy $u \in U$, satisfying the relation $\varphi_\alpha(u) \leqslant \psi(\min\{R_\alpha, \rho_\beta\})$, will be called a guaranteeing solution. Thus, a guaranteeing solution can be found from the problem (5) for $r = \bar{R}_\alpha$, where $\bar{R}_\alpha \triangleq \min\{R_\alpha, \rho_\beta\}$. Denote this guaranteeing solution by u^0. In this section, we propose an algorithm for improving the guaranteeing solution u^0, i.e. providing a smaller value of the criterion function $\varphi_\alpha(u)$ than $\varphi_\alpha(u^0)$.

As noted in the previous section, the dichotomy method can be used to find the radius of the ball r^*, inscribed in the confidence polyhedron C_r. In this case, the following difficulties arise: first, the continuity and monotonicity of $h(r) = \mathbf{P}\{X \in C_r\}$ is not guaranteed in the general case, secondly, the calculation of the probability of X falling into the polyhedron C_r requires the use of approximate methods. Nevertheless, we will use the dichotomy method to find an improved guaranteeing solution. Due to the fact that $h(r)$ will be calculated approximately using the Monte Carlo procedure, we will look for a value of r, such that $h(r) \geqslant \alpha + \varepsilon$, where ε is a small positive constant ($\varepsilon < 1 - \alpha$). Approximate calculation of the measure can lead to the fact that an unacceptable solution of the problem will be found, therefore it is necessary to specify the probability p of finding an acceptable solution. Since the quantile setting implies finding a solution that guarantees a given level of objective function value with a probability α, it is recommended to choose $p \geqslant \alpha$.

Algorithm 1.

1. Set algorithm parameters $\varepsilon \in (0, 1 - \alpha)$ (accuracy parameter for measure calculation), $\delta > 0$ (accuracy parameter for radius calculation) and $p \in [\alpha, 1)$ (probability of finding an acceptable solution).

2. Calculate ρ_α being the α level quantile of the standard normal distribution and $\bar{R}_\alpha \triangleq \min\{R_\alpha, \rho_\beta\}$, where $R_\alpha = \sqrt{\chi^2_\alpha(m)}$, $\chi^2_\alpha(m)$ being chi-square distribution quantile with m degrees of freedom, $\beta = 1 - \frac{1-\alpha}{k}$.

3. Calculate sample size

$$N = \left\lceil \frac{\ln(1/(1 - \sqrt[K]{p}))}{2\varepsilon^2} \right\rceil,$$

where $K = \left\lceil \log_2 \frac{|\bar{R}_\alpha - \rho_\alpha|}{\delta} \right\rceil$, $\lceil a \rceil$ is the rounding of a up to the nearest integer.

4. Set $r_1 := \rho_\alpha$, $r_2 := \bar{R}_\alpha$.

5. Find the lower estimate for the solution $\psi(r_1)$ and the upper estimate $\psi(r_2)$ of the optimal value of the criterion function, as well as the initial guaranteeing solution $u(r_2)$, by solving the problem (5) for $r = r_1$ and $r = r_2$.

6. While $|r_1 - r_2| > \delta$ repeat the following steps:

6.1. Assign $r := \frac{r_1 + r_2}{2}$.

6.2. Calculate $u(r)$ and $\psi(r)$, by solving the problem (5).

6.3. Simulate N independent realizations of a random vector X.

6.4. Calculate $\mu(r) \triangleq \mathbf{P}\{X \in B_r\} = F_{\chi^2(m)}(r^2)$, where $F_{\chi^2(m)}(r^2)$ is the value of the distribution function of the chi-square law with m degrees of freedom at point r^2.

6.5. Find $\hat{h}(r)$ being an estimate of the measure of the set C_r, defined by the formula (7):

$$\hat{h}(r) = \mu(r) + \frac{s(r)}{N},$$

where $s(r)$ is the number of sample elements included in the set $C_r \setminus B_r$.

6.6. If $\hat{h}(r) \geqslant \alpha + \varepsilon$, then $r_2 := r$. Otherwise $r_1 := r$.

7. As a guaranteeing solution, take $u(r_2)$.

Note that to improve the accuracy of the algorithm, one can use not the dichotomy method, but divide the segment of the search for a solution into several equal parts. In this case, at step 6.1 of the algorithm, it will be necessary to take several values of r in the segment $[r_1, r_2]$. It should also be noted that in the case of a nonmonotonic dependence of $r \mapsto h(r)$ the algorithm may not find the root of the equation $h(r) = \alpha + \varepsilon$, but some guaranteeing solution will be found.

Let us formulate a theorem on the convergence of the algorithm.

Theorem 3. *Let the problem* (5) *have a solution for* $r \in [\rho_\alpha, \bar{R}_\alpha]$. *Then application of the algorithm ensures finding a guaranteeing solution with a probability not less than* p.

The following theorem characterizes the accuracy of the solution found using the proposed algorithm 1. This result is a refinement of [2, Theorem 3.13] for optimization problems of the class under consideration.

Theorem 4. *Let the function* ψ *be defined and takes finite values on the segment* $[\rho, R]$, *and let the loss function be Lipschitz with constant* L, *i.e.*

$$|\Phi(u, x) - \Phi(u, y)| \leqslant L\|x - y\|.$$

Also suppose that

$$\max_{j=\overline{1,k_2}} \{b_{2j}(u(\rho)) + \|B_{2j}(u(\rho))\|R\} \leqslant 0. \tag{11}$$

Then $0 \leqslant \psi(R) - \psi(\rho) \leqslant (R - \rho)L$.

This inequality indicates the closeness of the found upper estimate of the criterion function to its optimal value. Theorem 4 gives an estimate of the bounds in these inequalities, which can be obtained even before applying the Algorithm 1. According to this estimation

$$0 \leqslant \psi(\bar{R}_\alpha) - \psi(\rho_\alpha) \leqslant L|\bar{R}_\alpha - \rho_\alpha|,$$

if the conditions of Theorem 4 are satisfied. Note that these conditions are satisfied for a Lipschitz loss function, for example, for $Q(u, x) \equiv 0$.

5. NUMERICAL EXPERIMENT

Example 2. Let us find a guaranteeing solution to the problem (1) for

$$\begin{aligned}\Phi(u,x) = \max\{&u_1 + 3u_3 + 2u_5 + x_1 + 2x_3 + 4,\\ &-u_1 + 2u_2 - u_3 + 3u_4 + 2u_5 + 2x_1 - x_2 + 2x_3,\\ &2u_1 + u_2 + 2u_3 - 2u_4 - u_5 + 3x_1 + x_2 + 2x_3 + 2,\\ &3u_1 - 2u_2 + u_3 + 3u_4 - 3u_5 - 2x_1 + 3x_2 - 3x_3 + 5,\\ &0.1u_1^2 - 0.02u_1u_2 - 0.03u_1u_3 + 0.2u_2^2 + 0.05u_3^2 + 0.3u_4^2 +\\ &+ 0.1u_5^2 - 0.2u_1 - 0.3u_2 - 0.1u_3 - 0.2u_5 - 3x_1 - 2x_2 + x_3 + 6\},\\ Q(u,x) = &\,3u_2 + u_1 + 4u_3 - 2u_5 - x_1 - 3x_2 - 4x_3 - 10,\end{aligned}$$

$U = \{u \in \mathbb{R}^5 \mid u_i \in [0; 10], i = \overline{1,5}\}$, $\alpha = 0.95$. For this level α, $\rho_\alpha = 1.645$, $R_\alpha = 2.796$, $\beta = 0.992$, $\rho_\beta = 2.394$, $\bar{R}_\alpha = 2.394$. Therefore, the function h must be considered on the segment $[1.645; 2.394]$. Solving problem (5) for $r = \rho_\alpha$ and $r = \bar{R}_\alpha$, find an estimate

$$\varphi_\alpha \in [\psi(\rho_\alpha), \psi(\bar{R}_\alpha)] = [11.813; 14.754].$$

Table 3. Application of the Algorithm 1

Iteration	r	$\hat{h}(r)$	$\psi(r)$
1	2.019	0.949	13.267
2	2.207	0.970	14.007
3	2.113	0.961	13.635
4	2.066	0.956	13.451
5	2.043	0.952	13.359
6	2.031	0.950	13.313
7	2.037	0.9507	13.336

The initial guaranteeing solution has the form

$$u(\bar{R}_\alpha) = (0.139; 0.602; 0.000; 0.004; 1.613)^\top.$$

Let us set the algorithm parameters: $\varepsilon = 0.001$, $\delta = 0.01$, $p = 0.99$. These parameters require a sample size of $N = 3\,273\,389$. The application of Algorithm 1 is shown in Table 3. Improved guaranteeing solution complies with $r = r^* \triangleq 2.043$ and it has the form

$$u(r^*) = (0.536; 0.688; 0.000; 0.003; 1.356)^\top.$$

At the same time

$$\varphi_\alpha \in [\psi(\rho_\alpha), \psi(r^*)] = [11.813; 13.359].$$

Thus, the use of Algorithm 1 made it possible to reduce the length of the uncertainty interval of the optimal value of the criterion function on $(1 - \frac{13.359-11.813}{14.754-11.813})100\% = 47\%$, which indicates the efficiency of the proposed algorithm.

All calculations were carried out on a computer with Intel(R) Core(TM) i5-6300U CPU, 2.40 GHz, RAM 8 GB RAM in Matlab system using program for solving quadratic Gurobi optimization problems. The counting time was 1035 s. The bulk of the calculation was the calculation of the measure of the polyhedron C_r using the Monte Carlo method.

6. CONCLUSION

The paper proposes an algorithm for solving a stochastic programming problem with a quantile criterion in the case of a loss function that is piecewise linear in random parameters and convex in strategy. The advantage of the proposed algorithm is the ease of constructing approximating problems, which can later be solved using convex optimization methods. The main computational difficulty in its application is the need to estimate the measure using the Monte Carlo method. The proposed algorithm for choosing a confidence set parameterized by the radius of the inscribed ball, as the example showed, can be successfully applied to solve stochastic optimization problems with a quantile criterion in the case of a convex piecewise quadratic linear loss function. It can be seen that this algorithm can also be applied to the case of discrete optimization strategies. The form of Algorithm 1 will not change, but in the course of applying the algorithm, it will be necessary to solve not a convex continuous optimization problem, but a discrete optimization problem. Algorithms for solving such problems may be the subject of further research.

FUNDING

The work was supported by the Russian Science Foundation (project no. 22-21-00213, https://rscf.ru/project/22-21-00213/).

APPENDIX

Proof of Theorem 1. Conditions 2 and 3 ensure that all constraints in the problem (5) are active. This means that all faces of the set C_r touch the ball B_r. As r increases on the segment $[0, R]$ the faces of the set C_r are transferred in parallel, touching the ball B_r. This means that the set C_r expands as r increases. Therefore, the function h, defined as the measure C_r, is non-decreasing. Theorem 1 is proved.

Proof of Theorem 2. Let $\gamma \in (0, 1)$. The set C_{ρ_γ} is defined as the intersection of k half-planes of measure no less than γ. Denote these half-planes by L_i, $i = \overline{1, k}$. Then

$$h(\rho_\gamma) = \mathbf{P}\left\{X \in \bigcap_{i=1}^{k} L_i\right\} = 1 - \mathbf{P}\left\{X \in \bigcup_{i=1}^{k} (\mathbb{R}^m \setminus L_i)\right\} \geqslant 1 - \sum_{i=1}^{k} \mathbf{P}\{X \notin L_i\} = 1 - (1-\gamma)k.$$

Thus, $h(\rho_\gamma) \geqslant \alpha$ for $\alpha \leqslant 1 - (1-\gamma)k$, which is equivalent to $\gamma \geqslant \beta = 1 - \frac{1-\alpha}{k}$. Theorem 2 is proved.

Proof of Theorem 3. Since at each iteration the segment of the search for a solution narrows two times, the number of iterations K of the algorithm can be found as the minimum natural number K, that satisfies the inequality

$$\frac{|\bar{R}_\alpha - \rho_\alpha|}{2^K} \leqslant \delta.$$

It follows from this inequality that $K = \left\lceil \log_2 \frac{|\bar{R}_\alpha - \rho_\alpha|}{\delta} \right\rceil$. The algorithm can make an error in its work only if at some iteration it turns out that $\hat{h}(r) \geqslant \alpha + \varepsilon$, although in fact $h(r) < \alpha$. It is easy to see that the random variable $s(r)$ is distributed according to the binomial law with the success probability $h(r) - \mu(r)$. The inequality is known [15, ch. 1, § 6]:

$$\mathbf{P}\{\hat{h}(r) - h(r) \geqslant \varepsilon\} = \mathbf{P}\left\{\frac{s(r)}{N} - (h(r) - \mu(r)) \geqslant \varepsilon\right\} \leqslant e^{-2N\varepsilon^2}.$$

Therefore, if we assume that $h(r) < \alpha$, then $\mathbf{P}\{\hat{h}(r) \geqslant \alpha + \varepsilon\} \leqslant e^{-2N\varepsilon^2}$. Since the samples used to evaluate the measure are independent, the probability that the algorithm will work correctly is at least $(1 - e^{-2N\varepsilon^2})^K$. Hence it follows that in order to ensure the probability p of successful operation of Algorithm he inequality

$$p \leqslant (1 - e^{-2N\varepsilon^2})^K \iff N \geqslant \frac{\ln(1/(1 - \sqrt[K]{p}))}{2\varepsilon^2}.$$

must be satisfied.

Theorem 3 is proved.

Proof of Theorem 4. Let $\Psi(u, r) \triangleq \max_{x \in B_r} \Phi(u, x) = \Phi(u, x^0(r))$, where x^0 is the point on the boundary of the ball B_r, where the specified maximum is reached. Since $B_\rho \subset B_R$, $\Psi(u, \rho) \leqslant \Psi(u, R)$ holds. Since the point $y = \frac{\rho}{R} x^0(R)$ lies on the boundary of the ball B_ρ, $\Phi(u, y) \leqslant \Psi(u, \rho)$. That's why

$$0 \leqslant \Psi(u, R) - \Psi(u, \rho) \leqslant \Phi(u, x^0(R)) - \Phi(u, y) \leqslant L\|x^0(R) - y\| = (R - \rho)L.$$

Thus, the inequalities

$$\Psi(u, \rho) \leqslant \Psi(u, R) \leqslant \Psi(u, \rho) + (R - \rho)L \tag{A.1}$$

are true. Minimizing the left and right parts of the first inequality in (A.1) with respect to $u \in U$ so that $\max_{j=\overline{1,k_2}} \{b_{2j}(u) + \|B_{2j}(u)\|R\} \leqslant 0$ (constraints of the problem (4) for $r = R$), we obtain

the first inequality to be proved $\psi(\rho) \leqslant \psi(R)$ (here we take into account that $\psi(\rho)$ is defined at least on a wider set). From (11) and the second inequality in (A.1) it follows that

$$\psi(R) \leqslant \Psi(u(\rho), R) \leqslant \Psi(u(\rho), \rho) + (R - \rho)L = \psi(\rho) + (R - \rho)L.$$

This estimate implies the second inequality to be proved. Theorem 4 is proved.

REFERENCES

1. Kibzun, A.I. and Kan, Y.S., *Stochastic Programming Problems with Probability and Quantile Functions*, Chichester, New York, Brisbane, Toronto, Singapore: John Wiley & Sons, 1996.
2. Kibzun, A.I. and Kan, Yu.S., *Zadachi stokhasticheskogo programmirovaniya s veroyatnostnymi kriteriyami* (Stochastic Programming Problems with Probabilistic Criteria), Moscow: Fizmatlit, 2009.
3. Kibzun, A.I. and Naumov, A.V., A Guaranteeing Algorithm for Quantile Optimization, *Kosm. Issled.*, 1995, vol. 33, no. 2, pp. 160–165.
4. Naumov, A.V. and Ivanov, S.V., On Stochastic Linear Programming Problems with the Quantile Criterion, *Autom. Remote Control*, 2011, vol. 72, no. 2, pp. 353–369.
5. Kan, Yu.S., An Extension of the Quantile Optimization Problem with a Loss Function Linear in Random Parameters, *Autom. Remote Control*, 2020, vol. 81, no. 12, pp. 2194–2205.
6. Vasil'eva, S.N. and Kan, Yu.S., A Method for Solving Quantile Optimization Problems with a Bilinear Loss Function, *Autom. Remote Control*, 2015, vol. 76, no. 9, pp. 1582–1597.
7. Vasil'eva, S.N. and Kan, Yu.S., Approximation of Probabilistic Constraints in Stochastic Programming Problems with a Probability Measure Kernel, *Autom. Remote Control*, 2019, vol. 80, no. 11, pp. 2005–2016.
8. Prékopa, A., *Stochastic Programming*, Dordrecht–Boston: Kluwer, 1995.
9. Shapiro, A., Dentcheva, D., and Ruszczyński, A., *Lectures on Stochastic Programming. Modeling and Theory*, Philadelphia: Society for Industrial and Applied Mathematics (SIAM), 2014.
10. Lejeune, M.A. and Prékopa, A., Relaxations for Probabilistically Constrained Stochastic Programming Problems: Review and Extensions, *Ann. Oper. Res.*, 2018. https://doi.org/10.1007/s10479-018-2934-8
11. Dentcheva, D., Prékopa, A., and Ruszczyński, A., On Convex Probabilistic Programming with Discrete Distributions *Nonlinear Anal.-Theor.*, 2001, vol. 47, no. 3, pp. 1997–2009.
12. Van Ackooij, W., Berge, V., de Oliveira, W., and Sagastizábal, C., Probabilistic Optimization via Approximate p-Efficient Points and Bundle Methods, *Comput. Oper. Res.*, 2017, vol. 77, pp. 177–193.
13. Ivanov, S.V. and Kibzun, A.I., General Properties of Two-Stage Stochastic Programming Problems with Probabilistic Criteria, *Autom. Remote Control*, 2019, vol. 80, no. 6, pp. 1041–1057.
14. Boyd, S. and Vandenberghe, L., *Convex Optimization*, Cambridge: University Press, 2009.
15. Shiryaev, A.N., *Probability*, New York: Springer, 1996.

This paper was recommended for publication by E.Ya. Rubinovich, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 8, pp. 967–980.

STOCHASTIC SYSTEMS

Resolvents of the Ito Differential Equations Multiplicative with Respect to the State Vector

M. E. Shaikin

Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
e-mail: shaikin@ipu.ru

Received September 1, 2022
Revised May 25, 2023
Accepted June 9, 2023

Abstract—Integral representations of solutions of linear multiplicatively perturbed differential equations are obtained, the diffusion part of which is bilinear on the state vector and the vector of independent Wiener processes. Equations of such class serve as models of stochastic systems with control functioning under conditions of parametric uncertainty or undesirable influence of external disturbances. The concepts and analytical apparatus of the theory of Lie algebras are used to find integral representations and fundamental matrices of the equations.

Keywords: multiplicative stochastic system, fundamental matrix, Fisk–Stratonovich differential, group-theoretic method, matrix Lie algebra, Wei–Norman theorem, stochastic resolvent

DOI: 10.25728/arcRAS.2023.32.54.001

1. INTRODUCTION

In the theory of optimization of dynamical systems an important place is given to the control problems of objects functioning under conditions of parametric uncertainty or undesirable influence of external disturbances. The simplest models of such systems in the stochastic section of the theory are linear, called multiplicative, Ito equations, the diffusion components of which are linear on vectors of state, control and external or parametric perturbation. Multiplicative equations are simple enough mathematical objects, and it is hoped to obtain in closed analytic form their solutions or integral representations for them.

Consider a stochastic Ito system (1.1), (1.2), whose *dynamics* is given by the multiplicative Markov equation

$$dx_t = a(t, x_t)dt + b(t)(x_t; dw(t)), \quad x_t \in R^d, \quad w(t) \in R^r, \quad x_0 = \text{const} \tag{1.1}$$

(coefficients depend on t), *driving force* is determined by a random function f with a differential

$$df(t) = (B_1(t)u_t + B_2(t)\upsilon_t)dt + B_{01}(t)u_t dw_1(t) + B_{02}(t)\upsilon_t dw_2(t), \tag{1.2}$$

where u_t and υ_t are vector signals of control and external perturbation respectively; $w(t)$ with or without indices denotes the vector Wiener process. Equation (1.1) is assumed to be linear in the state vector x_t such that $a(t, x) = A(t)x$, where $A(t) \in R^{d\times d}$ is the matrix $d \times d$ at each t, the diffusion component is defined by the function $b(t)(\cdot\,;\cdot)$ of two variables $(x, h) \in R^d \times R^r$ taking values in R^d, and the mapping $R^d \times R^r \to R^d$ is bilinear. The operator $B(t)h$ defined by the relation $(B(t)h)x = b(t)(x; h)$ is linear $R^d \to R^d$ at fixed h. All matrix functions in (1.1), (1.2) are assumed to be continuous on each finite interval of values of the parameter t. The system (1.1), (1.2) is called below (x, u, υ)-*multiplicative*; in particular, the system (1.1)—(x)-multiplicative. Multiplicative

models of the type (1.1), (1.2) are used, in particular, in the theory of H_2/H_∞—optimization of stochastic systems [1].

The purpose of the paper is to obtain in integral form the solution of the linear (x, u, υ)-*multiplicative* equation or the stochastic analog of its fundamental matrix. Let's make it clear what kind of fundamental matrix and which solution in integral form is talking about. The solution in the deterministic case of the linear differential equation $\dot{x} = A(t)x + B(t)$ has the following form

$$x(t) = R(t, t_0)x_0 + \int_{t_0}^{t} R(t, \tau)B(\tau)d\tau, \tag{1.3}$$

where $R(t, t_0)$ is the *resolvent* (or fundamental matrix) of the homogeneous at $B = 0$ equation [2, p. 144]. The function $R(t, t_0)x_0$ is a general solution of the homogeneous equation taking the value x_0 at $t = t_0$, and the integral in (1.3) is the solution of the perturbed equation going to zero at $t = t_0$. The fundamental matrix of equation (1.1) in the *stochastic* case is a matrix *random* function $\Phi(t, \tau)$, and the general solution of the perturbed equation, following the analogy with (1.3), should be given by the formula

$$x(t) = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, \tau) \circ df(\tau), \tag{1.4}$$

where integral is stochastic; $\circ df$ is denoted the Stratonovich differential [3, p. 105–109]. The integral is chosen stochastic in the Stratonovich sense for the reason that the differentiation rule of a complex function $t \mapsto f(\xi^1(t), \ldots, \xi^d(t))$ is represented in the in the same form as in the classical calculus, that is, as $df = \sum_{i=1}^{d} \frac{\partial f}{\partial x^i} \circ d\xi^i$ [3]. This integral in Stratonovich form makes it possible to extend some group-theoretic methods to the stochastic case. In the deterministic case, the group-theoretical concepts allow to overcome the difficulties of studying multidimensional systems caused by the *non-commutativity* of the matrix coefficients defining the dynamics of the system [4]. Perhaps, the same concepts can be useful in the problem of multiplicativity.

Some examples of the application of group-theoretic methods to statistical research are known in the literature. Here is a small list of publications thematically close to the problem of analyzing multiplicative systems [5–11]. In [5] the problem of numerical approximation of the solution of the stochastic equation is considered in the following form

$$dx_t = (Ax_t + f(x_t))dt + \sum_{i=1}^{n}(B_i x_t + g_i(x_t))dw_i, \quad x(0) = x_0 \in R^d$$

with nonlinear functions $f, g_i : R^d \to R^d$ and matrices $A, B_i \in R^{d\times d}$ satisfying the following conditions: A, B_i take values in the matrix Lie algebra $\mathfrak{g}$ with commutator relations $[A, B_i] = 0$, $[B_i, B_j] = 0$ for all i, j. On the background of works on the group-theoretic analysis of *deterministic* equations, the number of which has clearly decreased recently [6], the analysis of solution properties and numerical algorithms for finding solutions (so-called exponential integrators) for *stochastic* equations remains an active field of research on multiplicative and additive noise equations [7, 8]. The question of the mean-square stability of numerical methods for the calculation of exponential integrators is investigated in [9]. As shown in [10], group-theoretic methods are also effective for the numerical integration of partial equations. Among the works of Russian authors we note the research of multiplicative stochastic differential-operator equation with operators A, B acting in a separable Hilbert space [11]. In this paper, it is assumed that the operator A gives rise

to a a semigroup of operators $S(t)$, $t > 0$ of class C_0; it guarantees the correctness of the Cauchy problem for the unperturbed equation $\dot{X}(t) = AX(t)$.

The problem solved in this paper considers a finite-dimensional multiplicative equation, for the computation of its resolvent analog the group-theoretic method is applied, which is a generalization to the stochastic case of the deterministic Wei–Norman method [12] of finding the resolvents of linear differential equations. Wei–Norman method: if in the matrix equation $\dot{\Phi}(t) = B(t)\Phi(t)$, $\Phi(0) = E$ (E is a unit matrix), the non-random function $B(t)$ takes values in the matrix Lie algebra $\mathfrak{g}$, then the solution $\Phi(t)$ belongs to the corresponding Lie group $\mathcal{G}$. In this case, one way to construct the solution of $\Phi(t)$ is to represent by a finite product of matrix exponentials

$$\Phi(t) = \exp(s_1(t)A_1) \dots \exp(s_m(t)A_m), \tag{1.5}$$

where $\{A_1, \dots, A_m\}$ is the basis of the minimal Lie algebra $\mathfrak{g}$ generated by matrices $A(t)$ for all t, and $s_i(t)$, $i = 1, \dots, m$ are some real functions. Finding the desired $s_i(t)$ is reduced to the solution of some system of nonlinear differential equations [12]. The basis of the Wei–Norman method proposed here for the case of the multiplicative Ito equation is to write the latter in the form of the Fisk–Stratonovich equation and to find a solution of the latter in the form of the product of matrix exponents $\exp(A_i s_i(t))$ with the needed *semimartingales* (in the terminology adopted in the [3]) $s_i(t)$. Regarding the matrices A_i, $i = 1, \dots, m$, it is assumed, as in the deterministic case, that they form the basis of some matrix Lie algebra.

Applications of group theory to the problems of analyzing and finding solutions of deterministic differential equations are widely known from the monographic literature [4, 13, 14]. Applications to the theory of stochastic differential equations are much more modest; from the textbook literature we mention [3, 15, 16]. An exposition of the group-theoretic method of Wei–Norman to the problem of computing the of the resolvents of multiplicative Ito equations has not been found in the literature.

2. PROBLEM FORMULATION

By characterizing the stochastic system in the previous section as being given by the (x, u, υ)-multiplicative Ito equation was separation of the equation into its dynamical part and the forcing force, which does not depend on the state vector of the system. This is dictated by the character of the problem to compute the fundamental matrix (resolvent) of the stochastic Ito equation, which is defined by its homogeneous x_t-dependent part. Having calculated the resolvent, it is not difficult to obtain then an integral representation of the solution of the equation. Following this consideration, it is possible to pass from the general (x, u, υ)-multiplicative system to its dynamic part, i.e., to equation (1.1), which is multiplicative only on the state.

Let us list the tasks solved in the paper. The first problem is determination of the Wiener and martingale species of the diffusion component $b(t)(x_t; dw(t))$ of equation (1.1). The second problem in Sections 3, 4 is to write the multiplicative equation (1.1) in the symmetrized Fisk–Stratonovich form. The third task is to obtain the *integral* representation of the solution of the multiplicative equation (1.1). The more general case of the diffusion equation with the matrix $\sigma(t, x)$, depending *affinely* (not simply linearly) on x, see Section 5, the interesting phenomenon of the appearance of an additional forcing force in the integral representation for the solution of the equation. When solving the following two problems in Sections 6 and 7, there arise group-theoretic aspects of solving a multiplicative equation written in a symmetrized form, with *solvable* (in Section 6) Lie algebra and with arbitrary Lie algebra for the matrix coefficients of the diffusion component of the equation in Section 7. The equation in Section 7 is given in the unsymmetrized martingale form instead of Wiener processes. In a separate section we give *example* of finding the resolvent of the equation by the group-theoretic method. Concluding remarks and a list of cited references conclude the paper.

3. WIENER AND MARTINGALE REPRESENTATIONS OF THE DIFFUSION COMPONENT

Both the Wiener process and martingale representations of the differential equation for perturbing forces are quite interesting in multiplicative theory. The martingale equation is discussed in more detail in Section 7.

Proposition 1. *The diffusion component* $b(t)(x_t; dw(t))$ *of the homogeneous equation* (1.1) *admits the following equivalent representations:*

(a) $b(t)(x_t; dw(t)) = (B_1(t)x_t, \ldots, B_r(t)x_t)dw(t)$, *where* $B_j(t)$, $j = 1, \ldots, r$ *is a matrix with size* $d \times d$;

(b) $b(t)(x_t; dw(t)) = \sum\limits_{i=1}^{m} A_i x_t d\zeta^i(t)$, *where* A_i, $i = 1, \ldots, m$ *are matrices* $d \times d$ *and* $d\zeta^i(t) = \sum\limits_{j=1}^{r} b_j^i(t)dw^j(t)$, $b_j^i(t) \in R$, *where* $\zeta^i(t)$ *are martingales.*

Proof. As noted in Section 1, the diffusion part $b(t)(x_t; dw(t))$ of the linear equation at each t is given by the bilinear mapping $b(t)$ of the product $V \times H$, where $V = R^d$, $H = R^r$, of vector spaces into the space V. When $h \in H$ is fixed, the operator $B(t)h$, defined by the equality $(B(t)h)x = b(t)(x; h)$, is an element of the space $End\, V$ of linear operators from V to V. Let $\{h_j, j = 1, \ldots, r\}$ be a basis in H such that in the decomposition $w(t) = \sum_j w^j(t)h_j$ the Wiener processes $w^j(t)$ are mutually independent. There is

$$b(t)(x; dw(t)) = b(t)\left(x; \sum_{j=1}^{r} dw^j(t)(b(t)h_j)\right) = \sum_{j=1}^{r} dw^j(t)(B(t)h_j)x,$$

where $B(t)h_j \in End\, V$.

Denoting $B_j(t) := B(t)h_j$, we obtain statement (a) $b(t)(x_t; dw(t)) = (B_1(t)x_t, \ldots, B_r(t)x_t)dw(t)$ Proposition 1. Thus, the dependence of $b(t)$ on x is given by a set of r arbitrary square $d \times d$ matrices $B_j(t)$, not necessarily linearly independent [1, 17].

Further, let $\{A_i, i = 1, \ldots, m\}$ be the basis of a linear subspace $L \subset End\, V$, generated by the operators $B(t)h_j$. Assuming $B(t)h_j = \sum_{i=1}^{m} b_j^i(t)A_i$, $j = 1, \ldots, m$, where $b_j^i(t) \in R$, and introducing the notations $d\zeta^i(t) := \sum_{j=1}^{r} b_j^i(t)dw^j(t)$, $i = 1, \ldots, m$, we get $b(t)(x; dw(t)) = \sum_{i=1}^{m} d\zeta^i(t)A_i x$, which finishes the the proof Proposition 1. Below, without loss of generality, we assume $\dim L = m = r$.

In the proof of Proposition 1, the drift $a(t, x_t)dt$ in the in equation (1.1) was not taken into account. Implicitly, it was assumed to be zero. It can indeed be converted to zero by the well-known transformation (of course, in this case (1.1) will be replaced by an equation with another bilinear mapping $b(t)$). Indeed, let $y_t = \Lambda_t^{-1}x_t$, where Λ_t is a matrix exponent satisfying, as is known, the integral equation $\Lambda_t = E + \int_0^t a(s)\Lambda_s ds$ with initial condition $\Lambda_0 = E$. Since $dy_t = (d\Lambda_t^{-1})x_t + \Lambda_t^{-1}dx_t$ and $d\Lambda_t^{-1} = -\Lambda_t^{-1}a(t)dt$, then

$$dy_t = \Lambda_t^{-1}a(t)x_t dt + \Lambda_t^{-1}b_t(x_t; dw(t)) - \Lambda_t^{-1}a(t)x_t dt$$

(note that the matrices $a(t)$ and Λ_t commute), thus we obtain the equation $dy_t = \Lambda_t^{-1}b_t(\Lambda_t y_t; dw(t))$ with zero drift. See that the matrices defined above in Proposition 1 $B_j(t)$ are replaced by the matrices $\tilde{\Lambda}_t = \Lambda_t^{-1}B_j(t)\,\Lambda_t$, $j = 1, \ldots, r$.

Let us now find out how to transform the multiplicative equation (1.1) to the symmetrized Fisk–Stratonovich form. It has been noted above that such transformation is a necessary requirement of the methodology proposed here.

Proposition 2. *In the symmetrized Fisk–Stratonovich form the equation of state* (1.1) *written in the form*

$$dx_t = a(t)x_t dt + (B_1(t)x_t, \dots, B_r(t)x_t)dw(t) \tag{3.1}$$

(*Proposition* 1,(a)) *takes the form*

$$dx_t = a(t)x_t dt + B_0(t)x_t dt + \sum_{j=1}^{r} B_j(t)x_t \circ dw^j(t), \tag{3.2}$$

where $B_0(t) := -1/2\sum_{j=1}^{r} B_j^2(t)$.

Proof. Starting from the theory of Markov type equations

$$dx_t = a(t, x_t)dt + \sigma(t, x_t)dw(t), \tag{3.3}$$

which does not even assume linearity on x_t of the functions $a(t, x_t)$ and $\sigma(t, x_t)$ [3], let us write (3.3) in coordinate form

$$dx_t^i = a^i(t, x_t)dt + \sum_{j=1}^{r} b_j^i(t, x_t)dw^j(t), \;\; i = 1, \dots, d.$$

According to the general theory, equation (3.3), using the Fisk–Stratonovich differential, is represented as

$$dx_t = \bar{a}(t, x_t)dt + \sigma(t, x_t) \circ dw(t), \tag{3.4}$$

where the vector $\bar{a}(t, x)$ has components

$$\bar{a}^i(t, x) = a^i(t, x) - 1/2\sum_{j=1}^{d}\sum_{k=1}^{r}\left(\frac{\partial}{\partial x^j}b_k^i(t, x)\right)b_k^j(t, x). \tag{3.5}$$

Recall that the stochastic Ito differential $dw^j(t)$ and the differential $\circ dw^j(t)$ are related by the formula

$$x_t dw^q(t) = x_t \circ dw^q(t) - 1/2 dx_t dw^q(t). \tag{3.6}$$

Consider equation (3.1) in the form

$$dx_t = a(t)x_t dt + \sum_{j=1}^{r} B_j(t)x_t dw^j(t) \tag{3.7}$$

and refer to formula (3.6). Since $x_t dw^j(t) = x_t \circ dw^j(t) - 1/2 dx_t dw^j(t)$, we have, ignoring for now the drift in (3.7), the equation $dx_t = \sum_j B_j(t)x_t \circ dw^j(t) - 1/2\sum_j B_q(t)dx_t dw^j(t)$. Noting that

$$dx_t dw^j(t) = \sum_k B_k(t)x_t dw^k(t)dw^j(t) = \sum_k B_k(t)x_t\delta_{jk}dt = B_j(t)x_t dt,$$

equation (3.1) in the transformed form can be written as

$$dx_t = a(t)x_t dt + \sum_{j=1}^{r} B_j(t)x_t \circ dw^j(t) - 1/2\sum_{j=1}^{r} B_j^2(t)x_t\, dt, \tag{3.8}$$

which is what was required. The drift in this equation is determined by the matrix $A(t) := a(t) - 1/2\sum_{j=1}^{r} B_j^2(t)$; it can be converted to zero by passing to the state vector $y_t = \Lambda_t^{-1}x_t$, where $\Lambda_t = E + \int_0^t A(s)\Lambda_s ds$.

In particular, if the matrices $B_j(t)$, $j = 1, \ldots, r$, commute, then the solution of the last equation is written as products

$$x_t = \prod_{q=1}^{r} \exp\left\{\int_0^t B_q(s)dw^q(s) - 1/2\int_0^t B_q^2(s)ds\right\} x_0.$$

It is also clear that the matrices $B_q(t)$ and $B_q^2(t)$ commute, so that the the multipliers in the product can be represented as

$$\exp\left\{\int_0^t B_q(s)dw^q(s)\right\} \exp\left\{-1/2\int_0^t B_q^2(s)ds\right\}, \quad q = 1, \ldots, r.$$

The solution of the equation $dx_t = B_q(t)x_t \circ dw^q(t)$ with zero drift, with initial condition x_0 is the function $U_q(t)x_0 = \exp\left\{\int_0^t B_q(s)dw^q(s)\right\} x_0$. The mapping $t \mapsto U_q(t)$ is the stochastic resolvent of this equation.

4. STOCHASTIC RESOLVENT OF MULTIPLICATIVE EQUATION

The non-random component $a(t)x_t dt$ in a multiplicative equation of the type

$$dx_t = a(t)x_t dt + \sum_{j=1}^{r} B_j(t)x_t \circ dw^j(t) \tag{4.1}$$

can be converted to zero (Section 3) and, without loss of generality, one can consider the equation to be given in the form $dx_t = \sum_{j=1}^{r} B_j(t)x_t \circ dw^j(t)$ with new matrix coefficients. To do this, let us put $y_t = \Lambda_t^{-1}x_t$ and then $dy_t = \Lambda_t^{-1}b(t)(x_t; dw(t))$. Since it is realized that $b(t)(x_t; dw(t)) = \sum_{j=1}^{r} B_j(t)x_t dw^j(t)$, then

$$dy_t = \sum_{j=1}^{r} B_j(t)\Lambda_t y_t \circ dw^j(t). \tag{4.2}$$

Within the group-theoretic formalism, it is the matrices $\tilde{B}_j(t) = \Lambda_t^{-1}B_j(t)\Lambda_t$, not the original matrices B_j, $j = 1, \ldots, r$, must give rise to the Lie algebra basic to the Wei–Norman method, where $\Lambda_t = \exp\int_0^t a(s)ds$ is the exponent of the matrix function $t \mapsto \int_0^t a(s)ds$.

Let us now define the (stochastic) resolvent of an equation linear in x_t. If Ψ_t is the solution of

$$d\Psi_t = \sum_{j=1}^{r} \Lambda_t^{-1}B_j(t)\Lambda_t\Psi_t \circ dw^j(t), \quad \Psi_t|_{t=0} = E \tag{4.3}$$

with zero drift, then the the fundamental matrix (resolvent) $\Phi(t)$ of the initial equation (4.1) is defined by the formula $\Phi(t) = \Lambda_t\Psi_t$. But since equation (4.1) is equivalent to Ito's equation $dx_t = \sum_{j=1}^{r} dw^j(t)B_j(t)x_t$, hence, $\Phi(t)$ satisfies the Stratonovich equation

$$d\Phi_t = a(t)\Phi_t dt + \sum_{j=1}^{r} B_j(t)\Phi_t \circ dw^j(t), \quad \Phi_t|_{t=0} = E. \tag{4.4}$$

If $f(t)$ is the driving force in an inhomogeneous stochastic equation, then the solution of the latter must have an integral representation (1.4).

5. INTEGRAL CAUCHY REPRESENTATION OF THE SOLUTION OF THE MULTIPLICATIVE EQUATION WITH AFFINE COEFFICIENTS

Equations with affine coefficients are necessary in the theory of such linear controllable systems that are multiplicative not only on the state vector, but also on the vectors of control and external perturbation. The stochastic resolvent theory of the previous section dealt with a multiplicative equation with linear but not *affine* coefficients. Now consider a vector equation (with a vector Wiener process) of the form

$$dx_t = a(t,x_t)dt + \sigma(t,x_t)d\,w(t),\ \ x_0 \in R^d, \tag{5.1}$$

where $a(t,x) = a(t)x + a_0(t)$, $\sigma(t,x) = B(t,x) + b_0(t)$, $a(t) \in R^{d\times d}$, $a_0(t) \in R^d$, $B(t,x)$, $b_0(t) \in R^{d\times r}$, $w(t) \in R^r$. If (5.1) is a multiplicative system, then $B(t,x) = (B_1(t)x, \dots, B_r(t)x)$ is a matrix with columns $B_j(t)x$, where $B_j(t) \in R^{d\times d}$, $j = 1,\dots,r$, which is established in (4.1). Then, $b_0(t)$ is a matrix with columns $b_{0j}(t)$, $j = 1,\dots,r$. A special case of a one-dimensional ($x_t \in R$) system with affine coefficients and scalar $w(t)$ is considered by in [15]. In the vector case $x_t \in R^d$, let us find the fundamental matrix of the of equation (5.1).

Proposition 3. *The solution of the multiplicative equation with coefficients affine with respect to x_t has the following integral representation:*

$$x_t = \Phi_t\left(x_0 + \int\limits_0^t \Phi_s^{-1}\left(a_0(s) - \sum_{j=1}^r B_j(s)b_{0j}(s)\right)ds + \int\limits_0^t \Phi_s^{-1}\sum_{j=1}^r b_{0j}(s)dw^j(s)\right). \tag{5.2}$$

Here Φ_t defined in the in the previous section, the resolvent (4.1) *written for* (5.1) *with the conditions $a_0(t) = 0$, $b_{0j}(t) = 0$, $j = 1,\dots,r$.*

Note that the appearance under the integrals in (5.2), in addition to $a_0(s)ds + \sum_{j=1}^r b_{0j}(s)dw^j(s)$, *additional* driving force $B(s)b_0(s) := -\sum_{j=1}^r B_j(s)b_{0j}(s)ds$ (it is caused by the "affine" additive $b_0(s)$) could not be foreseen in advance, but a direct check shows that the function (5.2) indeed satisfies equation (5.1).

Proof. Let us again turn to equation (5.1). In stochastic case, let us apply a method analogous to the deterministic method of constant variation. Let's put $x_t = \Phi_t\eta_t$ and consider η_t as the new unknown instead of x_t. Differentiating $x_t = \Phi_t\eta_t$ stochastically, we obtain

$$dx_t = (d\Phi_t)\eta_t + \Phi_t d\eta_t + d\Phi_t d\eta_t,$$

or by the definition of Φ_t as the equation (4.4)

$$dx_t = \left(a(t)dt + \sum_{j=1}^r dw^j B_j(t)\right)x_t + \Phi_t d\eta_t + \left(a(t)dt + \sum_{j=1}^r dw^j B_j(t)\right)\Phi_t d\eta_t.$$

Equating the right-hand sides of this equation and the original equation (5.1)

$$dx_t = (a(t)x_t + a_0(t))dt + \sum_{j=1}^r dw^j\Big(B_j(t)x_t + b_{0j}(t)\Big),$$

we obtain after abbreviations

$$\left(E + a(t)dt + \sum_{j=1}^r dw^j(t)B_j(t)\right)\Phi_t d\eta_t = a_0(t)dt + \sum_{j=1}^r b_{0j}(t)dw^j(t). \tag{5.3}$$

Because, if we're being formal,

$$\left(E + a(t)dt + \sum_{j=1}^{r} dw^j B_j(t)\right)^{-1} = E - \left(a(t)dt + \sum_{j=1}^{r} dw^j B_j(t)\right) + \left(a(t)dt + \sum_{j=1}^{r} dw^j B_j(t)\right)^2 + \dots$$

and

$$\left(a(t)dt + \sum_{j=1}^{r} dw^j B_j(t)\right)^2 = \sum_{j=1}^{r} B_j^2 d(t),$$

we obtain from (5.3)

$$d\eta_t = \Phi_t^{-1}\left(a_0(t)dt - \sum_{j=1}^{r} B_j(t)b_{0j}(t)dt + \sum_{j=1}^{r} dw^j(t)b_{0j}(t)\right). \tag{5.4}$$

This equation expresses the fact that η_t is primitive for the right-hand side in (5.3), so that integrating (5.4) gives exactly the formula (5.2). Being expressed in terms of resolvent $\mathcal{R}(t,s) = \Phi_t\Phi_s^{-1}$, the same formula gives the integral Cauchy representation of the solution of the multiplicative equation (4.1) with affine coefficients. This was required to prove.

6. MULTIPLICATIVE EQUATION WITH SOLVABLE LIE ALGEBRA

In this section, an exhaustive solution to the problem of the integral of the solution of the multiplicative equation is obtained at the cost of a strong assumption that the Lie algebra associated to the equation is solvable. A result with a solvable Lie algebra is obtained by H. Kunita [18] and is given in [3] as as one of the examples.[1] The equation of state is assumed here to be given a priori in the symmetrized Fisk–Stratonovich form, the coefficients of the equation do not depend on t.

So, the equation is considered (with constant coefficients)

$$\begin{gathered} dx_t = (B_0x_t + b_0)dt + \sum_{p=1}^{r}(B_px_t + b_p)\circ dw^p(t), \\ B_p \in R^{d\times d}, \quad b_p \in R^d, \quad p = 0,1,\dots,d. \end{gathered} \tag{6.1}$$

Lie algebra generated by vector fields fields $L_p = \sum_{i=1}^{d}(B_px + b_p)^i\frac{\partial}{\partial x^i}$, $p = 0,1,\dots,r$, is solvable, which holds when $(B_p)_j^i = 0$ for $i > j$, $p = 0,1,\dots,r$. This condition means that in each of the matrices B_p, its elements under the of the main diagonal are zero. In particular, the only non-zero element of the last dth row is only the diagonal element $(B_p)_d^d$, $p = 0,1,\dots,r$ at $i = d$. It follows from equation (6.1)

$$dx_t^d = \left((B_0)_d^d x_t^d + b_0^d\right)dt + \sum_{p=1}^{r}\left((B_p)_d^d x_t^d + b_p^d\right)\circ dw^p(t) \tag{6.2}$$

when $i = d$. This (scalar) stochastic equation is similar to a deterministic equation in the sense that it is written using the Fisk–Stratonovich differential, so its solution has the form

$$x_t^d = e^{c_d(t)}\left(x_0^d + \int_0^t e^{-c_d(s)}\circ df_d(s)\right), \tag{6.3}$$

[1] A simple example of a solvable Lie algebra is generated by the group of translations of the plane R^2 and rotations about an axis perpendicular to it. The Lie algebra is three-dimensional, its commutation relations are $[X_1, X_2] = 0$, $[X_1, X_3] = X_2$, $[X_3, X_2] = X_1$ [19].

where the function $c_d(t)$ under the exponent sign and the driving force $f_d(t)$ are given respectively by the formulas

$$c_d(t) = (B_0)_d^d\, t + \sum_{p=1}^{r} (B_p)_d^d w^p(t), \quad f_d(t) = b_0^d\, t + \sum_{p=1}^{r} b_p^d w^p(t).$$

By proceeding analogously, let us consider the equation for x_t^{d-1}:

$$\begin{aligned} dx_t^{d-1} &= \left((B_0)_{d-1}^{d-1} x_t^{d-1} + (B_0)_d^{d-1} x_t^d + b_0^{d-1}\right) dt \\ &+ \sum_{p=1}^{r} \left((B_p)_{d-1}^{d-1} x_t^{d-1} + (B_p)_d^{d-1} x_t^d + b_p^{d-1}\right) \circ dw^p(t). \end{aligned} \tag{6.4}$$

In the right-hand side of equation (6.4) depends on x_t^{d-1} the sum of

$$(B_0)_{d-1}^{d-1} x_t^{d-1} dt + \sum_{p=1}^{r} (B_p)_{d-1}^{d-1} x_t^{d-1} \circ dw^p(t).$$

Let us denote the integral of the coefficient at x_t^{d-1} in this sum by

$$c_{d-1}(t) := (B_0)_{d-1}^{d-1} t + \sum_{p=1}^{r} (B_p)_{d-1}^{d-1} w^p(t).$$

The summands in the right-hand side of equation (6.4) independent of x_t^{d-1} form the sum

$$df_{d-1}(t) := \left((B_0)_d^{d-1} x_t^d + b_0^{d-1}\right) dt + \sum_{p=1}^{r} \left((B)_d^{d-1} x_t^d + b_p^{d-1}\right) \circ dw^p(t), \tag{6.5}$$

in which x_t^d is already known as the solution (6.3) of equation (6.2). The solution of equation (6.4) is written, therefore, in the form

$$x_t^{d-1} = e^{c_{d-1}(t)} \left(x_0^{d-1} + \int_0^t e^{-c_{d-1}(s)} \circ df_{d-1}(s) \right). \tag{6.6}$$

The procedure of sequential solution of scalar equations for components x_t^k, $k = n, n-1, \ldots, 1$ of the vector $x_t \in R^n$ is quite obvious from the above. However, the equivalence between this form of solution and the one given in [3] is not obvious. To establish the equivalence, let us write the original equation (6.1) by components:

$$dx_t^i = \sum_{j \geqslant i} \left((B_0)_j^i x_t^j + b_0^i\right) dt + \sum_{p=1}^{r} \sum_{j \geqslant i} \left((B_p)_j^i x_t^j + b_p^i\right) \circ dw^p(t). \tag{6.7}$$

When i is fixed, the summands in the of the right-hand side, depending on x_t^j with $j \geqslant i$, play a special role. First, the differential dx_t^i is related to the variable x_t^j by the coefficient $(B_0)_j^i dt + \sum_{p=1}^{r} (B_p)_j^i dw^p(t)$, the integral of which is denoted by c_j^i:

$$c_j^i(t) := (B_0)_j^i t + \sum_{p=1}^{r} (B_p)_j^i w^p(t), \quad j = i, i+1, \ldots, d.$$

The diagonal element $c_i^i(t)$ of this matrix coincides with the function, which was denoted above by $c_i(t)$. Second, the sum of summands on the right-hand side in (6.7) with in $\sum_{j=i+1}^{d} x_t^j \circ dc_j^i(t)$. Finally, the terms that do not depend on the components of the vector x_t at all form the sum $\sum_{p=1}^{r} b_p^i dw^p(t) + b_0^i dt$. Thus, the solution of the system of equations (6.7) is given by the formulas

$$x_t^d = e^{c_d^d(t)} \left(x_0^d + \int_0^t e^{-c_d^d(s)} \circ df_d(s) \right), \quad f_d(t) = b_0^d t + \sum_{p=1}^{r} b_p^d dw^p(t),$$

if $i = d$, and by the formulas

$$x_t^i = e^{c_i^i(t)} \left(x_0^i + \int_0^t e^{-c_i^i(s)} \circ df_i(s) \right),$$

where

$$f_i(t) := \sum_{j=i+1}^{d} \int_0^t x^j(s) \circ dc_j^i(s) + b_0^i t + \sum_{p=1}^{r} b_p^i dw^p(t),$$

if $i = d-1, d-2, \ldots, 1$ [3].

7. LIE ALGEBRA OF MULTIPLICATIVE EQUATION WITH CONTINUOUS SEMIMARTINGALES

Let us consider the slightly more general case of Ito's equation

$$dx_t = \sum_{p=1}^{m} A_p x_t d\zeta^p(t), \quad \zeta(0) = 0 \tag{7.1}$$

with continuous semimartingales $\zeta^i(t) = \int_0^t \sum_{j=1}^{m} b_j^i(t) dw^j(t)$ instead of the Wiener processes $w^j(t)$, $j = 1, \ldots, m$. The matrices A_p are assumed to be non-commutative, giving rise to an arbitrary finite-dimensional Lie algebra. It should be noted that the topic of the interaction of the *stochastic* structure of the differential equation with the *algebraic* group-theoretic structure of its coefficients remains to date insufficiently studied.

Suppose that the equation has a single solution x_t, $t > 0$, then x_t depends linearly on x_0. We will let $x_t = U(t)x_0$. The solution of the equation $dx_t = A_p x_t d\zeta^p(t)$, $\zeta(0) = 0$ with a single matrix A_p and supermartingale $\zeta^p(t)$ is an exponential supermartingale (if $\sum_j (b_j^p)^2 < \infty$)

$$\sigma_p(t) = e^{A_p \zeta^p(t) - 1/2 A_p^2 < \zeta^p > (t)} \sigma_p(0),$$

where $< \zeta^p > (t) = \sum_j \int_0^t (b_j^p)^2(s) ds$; [20, Section 2.7]. To obtain the solution, let us again use the method, already described in the introduction, namely: we write the original Ito equation (7.1) in the symmetrized Fisk–Stratonovich form and apply to the obtained equation an analogue of the deterministic Wei–Norman method [12]. After that, it will not be difficult to obtain the integral representation of the solution of equation (7.1).

Theorem. *Let*

$$dx_t = \sum_{p=1}^{m} A_p x_t d\zeta^p(t), \quad \zeta(0) = 0 \tag{7.2}$$

be a stochastic system with semimartingales

$$\zeta^i(t) = \int\limits_0^t \sum_{j=1}^m b_j^i(t) dw^j(t).$$

The functions $b_j^p(t)$ *are known. Consider the functions*

$$F_i = \prod_{k=1}^{i-1} e^{X_k s_k} X_i \prod_{k=i-1}^{1} e^{-X_k s_k},$$

where $X_1, \ldots, X_n$ *is a basis of the Lie algebra generated by the matrix coefficients* $A_p(t)$, $p = 1, \ldots, m$, *and* $s^i(t)$ *are the desired functions. Then for the differentials* $\circ ds^i(t)$ *of the unknown functions* $s^i(t)$ *are valid the system of equations* $s^i(t)$:

$$\sum_{i=1}^n F_i(t) \circ ds^i(t) = \sum_{p=1}^n \tilde{A}_p(t) \circ d\zeta^p(t). \tag{7.3}$$

Through the functions $s^i(t)$ *a solution* $x_t = U(t)x_0$ *of the original equation* (7.1) *is expressed.*

Proof. Keeping in mind the above remarks, let us write down equation (7.1) in the symmetrized form. There are $x_t \times d\zeta^p(t) = x_t \circ d\zeta^p(t) - 1/2\, dx_t \times d\zeta^p(t)$, where

$$dx_t \times d\zeta^p(t) = \sum_{q=1}^m A_q x_t\, d\zeta^q(t) \times d\zeta^p(t).$$

Since $d\zeta^q(t) \times d\zeta^p(t) = \sum_{j=1}^r b_j^q(t) b_j^p(t)\, dt =: c^{qp}(t)dt$, where $c^{qp}(t)$ are elements of the matrix $c(t) = b^*(t)b(t)$ of order $m \times m$, and matrix $b(t) = (b_j^p(t))$ of order $r \times m$, then

$$x_t \times d\zeta^p(t) = x_t \circ d\zeta^p(t) - 1/2 \sum_{q=1}^m A_q\, x_t c^{qp}(t)dt,$$

and equation (7.1) is written in the form of

$$dx_t = a(t)x_t dt + \sum_{p=1}^m A_p x_t \circ d\zeta^p(t) \tag{7.4}$$

with the drift coefficient $a(t) = -1/2 \sum_{p,q=1}^m A_p A_q\, c^{qp}(t)$. The fundamental matrix of equation (7.3), as above in Section 5 above, let us find it as a product of $\Phi_t = \Lambda_t \Psi_t$, where matrix $\Lambda_t = \exp\{\int_0^t a(s)ds\}$ satisfies the matrix equation $d\Lambda_t = a(t)\Lambda_t dt$. Given that $\Psi_t = \Lambda_t^{-1}\Phi_t$ and $d\Psi_t = (d\Lambda_t^{-1})\Phi + \Lambda_t^{-1} d\Phi_t$ and $d\Lambda_t^{-1} = -\Lambda_t^{-1}(d\Lambda_t)\Lambda_t^{-1}$ for the unknown function Ψ_t one gets the matrix differential equation

$$d\Psi_t = \sum_{p=1}^m (\Lambda_t^{-1} A_p \Lambda_t)\, \Psi_t \circ d\zeta_t^p. \tag{7.5}$$

The matrix drift coefficient turns here to zero, and the matrix coefficients A_p of the initial equation (7.1) turn into coefficients $\tilde{A}_p(t) = \Lambda_t^{-1} A_p \Lambda_t$. In such a case, from the Campbell-Baker-Hausdorff theorem [21], according to which $a, b \in L \Rightarrow e^a b e^{-a} \in L$, it follows that also $\tilde{A}_p \in L$, $p = 1, \ldots, m$. Here there arises a limitation for the application of group-theoretic methods caused by the necessity to transform the Ito equation to its symmetrized form. Probably, for this reason,

in most statistical applications, group analysis is applied to equations given immediately in the Fisk–Stratonovich form.

To continue the topic of group-theoretic analysis of equation (7.5), here let us also assume that the matrix coefficients $\tilde{A}_p(t) = \Lambda_t^{-1} A_p \Lambda_t$ in equation (7.5) are known a priori and $\tilde{L}$ is a Lie algebra generated by them for all t with some basis $\{X_1, \ldots, X_n\}$, then the Wei–Norman method can be applied to the algebra $\tilde{L}$. Below, the assumption of the existence Lie algebra $\tilde{L}$ for equation (7.5) is considered to be satisfied.

Purposing to search for the fundamental matrix Ψ_t in the form of the product $\prod_{i=1}^n e^{X_i s^i(t)}$ with unknown scalar functions $s^i(t)$, consider the matrix function

$$u(s) = u(s_1, \ldots, s_n) = e^{X_1 s_1} \cdots e^{X_n s_n}, \quad s_i \in R, \;\; i = 1, \ldots, n$$

(caution against confusing the numeric variable s_i with the function $s^i(t)$). The partial derivative $\frac{\partial}{\partial s_i} u(s)$ equals $u_{s_i} = \prod_{k=1}^{i-1} e^{X_k s_k} X_i \prod_{k=i}^{n} e^{X_k s_k}$, which can be be written in the form $u_{s_i} = F_i u(x)$, where denoted by $F_i = \prod_{k=1}^{i-1} e^{X_k s_k} X_i \prod_{k=i-1}^{1} e^{-X_k s_k}$. Therefore, the Fisk–Stratonovich differential of the function $t \mapsto \Psi_t = u(s^1(t), \ldots, s^n(t))$ is equal to

$$d\Psi_t = \sum_{i=1}^n F_i(t) \Psi_t \circ ds^i(t),$$

where $F_i(t)$ is obtained from the formula for F_i by substituting into it $s^i(t)$ instead of s_i for all $i = 1, \ldots, n$. Comparing $d\Psi_t$ with the differential for Ψ_t from equation (7.5), which (by replacing m by n) we rewrite as $d\Psi_t = \sum_{p=1}^n \tilde{A}_p(t) \Psi_t \circ d\zeta^p(t)$, we obtain, after reduction by the to the special matrix Ψ_t, the basic equation for the differentials $\circ ds^i(t)$ of the desired processes $s^i(t)$:

$$\sum_{i=1}^n F_i(t) \circ ds^i(t) = \sum_{p=1}^n \tilde{A}_p(t) \circ d\zeta^p(t). \tag{7.6}$$

Let us remind once again that there are relations

$$d\zeta^p(t) = \sum_{j=1}^r b_j^p(t) dw^j(t), \;\; p = 1, \ldots, n, \quad ds^q(t) = \sum_{j=1}^r g_j^q(t) dw^j(t), \;\; q = 1, \ldots, n,$$

where the functions $b_j^p(t)$ are known and the functions $g_j^q(t)$ are sought, where $n = \dim \tilde{L}$. If one decomposes both parts of the basic equation (7.6) by the basis $\{X_1, \ldots, X_n\}$ of the Lie algebra $\tilde{L}$, then we obtain a system of equations relating the unknowns functions g_j^p to the known b_j^p. Equation (7.6) is obtained by assumption that the drift coefficient $a(t)$ (7.4) is zero. The latter is ensured by transforming the original equation (7.4) to equation (7.5). The proof is now complete.

8. EXAMPLE

Let us consider an example of solving equation Ito of type (7.1), in which the assumption $a(t) = 0$ is violated, but still $a(t) \in \tilde{L}$. The fundamental matrix of the equation of the state is in the form of a product of exponential semimartingales. This is an example of using a modification of the Wei–Norman method (its stochastic version).

Let us find the fundamental matrix U_t of the stochastic equation $dx_t = \sum_{p=1}^3 X_p x_t d\zeta^p(t)$, $d\zeta^p(t) = \sum_{j=1}^3 b_j^p(t) dw^j(t)$, $\zeta_i(0) = 0$. Let $L = L_3$ be a Lie algebra of dimension $\dim L = 3$ with basis (X_i) and multiplication table $[X_1, X_2] = X_3$, $[X_2, X_3] = [X_1, X_3] = 0$. The algebra L_3 admits a representation of (3×3)-matrices $X_1 = E_{12}$, $X_2 = E_{23}$, $X_3 = E_{13}$ (E_{ij}—matrix canonical units), with $X_i^2 = 0$ for all i. The matrix $U_t$will be found as the product $U_t = \prod_{i=1}^3 \exp\{s^i(t) X_i\}$, where the

components $Z_k(t) = s^k(t)X_k$ are absent by virtue of $X_k^2 = 0$, and the functions $s^i(t)$ are suitably chosen random processes with differentials $ds^k(t) = \sum_{j=1}^3 g_j^k(t)dw^j(t)$, $s^k(0) = 0$. We assume

$$F_1(t) = X_1,\ F_2(t) = e^{Z_1(t)}X_2e^{-Z_1(t)},\ F_3 = e^{Z_1(t)}e^{Z_2(t)}X_3e^{-Z_2(t)}e^{-Z_1(t)}.$$

It is directly verified that

$$X_i^2 = 0 \quad \forall i,\ F_1 = X_1,\ F_2 = X_2 + s_1X_3,\ F_3 = X_3,\ F_1F_2 = X_3,$$

the remaining F_iF_j are zero. Using the modification of the method outlined in Section 3. Wei–Norman method of formulating equations for the unknown functions (in this example they are $s^i(t)$), one obtains the equation

$$\sum_{i=1}^3 F_ids^i = \sum_{i=1}^3 X_id\zeta^i.$$

Taking into account the formulas for F_i in the X_i basis decomposition, from this equation we get

$$ds^1 = d\zeta^1,\ ds^2 = d\zeta^2,\ ds^3 = d\zeta^3 - s^1ds^2 - ds^1ds^2. \tag{8.1}$$

To check the correctness of the obtained solution, let us find the the stochastic differential of the function U_t by calculating the function itself.

Since

$$\exp\{s^1X_1\} = I + s^1X_1,\ \exp\{s^2X_2\} = I + s^2X_2,\ \exp\{s^3X_3\} = I + s^3X_3,$$

then, by multiplication we find $U_t = I + s^1X_1 + s^2X_2 + (s^3 + s^1s^2)X_3$ and it follows that, $dU_t = X_1ds^1 + X_2ds^2 + X_3(ds^3 + d(s^1s^2))$, where $d(s^1s^2) = s^1ds^2 + s^2ds^1 + ds^1\,ds^2$. Substituting here the expressions for ds^i from (8.1), after the reduction we obtain $X_1d\zeta^1 + X_2d\zeta^2 + X_3d\zeta^3$, which coincides with the coefficient in the right-hand side of the original Ito equation. Thus, the solution of the stochastic Ito equation is found in the form of the product of the of stochastic semimartingales ("stochastic exponents").

9. CONCLUSION

The base of the integral representation of the solution of the linear of a stochastic equation is, as in the deterministic case, the fundamental matrix of solutions, through which the Green's function for the inhomogeneous equation is expressed. During the finding of the fundamental matrix of a multivariate equation, the main difficulty belongs to the noncommutativity of matrix coefficients of drift and diffusion components.

The non-commutativity of matrices is overcome in a known way if they are in involution. Turning to the methodology of group theory, we should assume that the coefficients of the equation belong to a certain matrix Lie algebra L closed with respect to a matrix commutator. For a linear system with diffusion components, depending only on the Wiener processes, but independent of the state vector, the Lie algebra associated to the system is organized quite simply: it is generated by the diffusion and drift coefficients. In the case of diffusion depending linearly on the state vector, it is necessary to preliminary transformation of the initial equation to the form, using the Fisk–Stratonovich differential. The drift coefficient becomes in this case depending on squares of diffusion coefficients, and diffusion coefficients, in their turn, undergoes transformations depending on the drift coefficient. And only in the commutative case (or in the case of a solvable algebra), it is possible to avoid the difficulties noted above. Thus, the situation with the application of standard group-theoretic concepts to the stochastic equation is satisfactory. Perhaps some algebraic structure other than Lie algebra, would be more appropriate in this problem, but the clarification of this question requires further study.

REFERENCES

1. Petersen, I.R., Ugrinovsky, V.A., and Savkin, A.V., *Robust Control Design using H_∞-methods*, London: Springer, 2006. ISBN 1-85233-171-2.
2. Kartan, A., *Differentsial'noe ischislenie. Differentsial'nye formy* (Differential calculus. Differential forms), Moscow: Mir, 1971.
3. Vatanabe, S. and Ikeda, N., *Stokhasticheskie differentsial'nye uravneniya i diffuzionnye protsessy* (Stochastic Differential Equations and Diffusion Processes), Moscow: Nauka, 1986.
4. Olver, P., *Prilozheniya grupp Li k differentsial'nym uravneniyam* (Applications of Lie Groups to Differential Equations), Moscow: Mir, 1989.
5. Erdogan, U. and Lord, G.J., *A New Class of Exponential Integrators for Stochastic Differential Equations with Multiplicative Noise*, 2016, arXiv:1608.07096v2.
6. Hochbruck, M. and Ostermann, A., Exponential Integrators, *Acta Numerica*, 2010, no. 19, pp. 209–286.
7. Mora, C.M., Weak Exponential Schemes for Stochastic Differential Equations with Additive Noise, *IMA J. Numer. Anal.*, 2005, vol. 25, no. 3, pp. 486–506.
8. Jimenez, J.C. and Carbonell, F., Convergence Rate of Weak Local Linearization Schemes for Stochastic Differential Equations with Additive Noise, *J. Comput. Appl. Math.*, 2015, vol. 279, pp. 106–122.
9. Komori, Y. and Burrag, K., A Stochastic Exponential Euler Scheme for Simulation of Stiff Biochemical Reaction Systems, *BIT*, 2014, vol. 54, no. 4, pp. 1067–1085.
10. Lord, G.J. and Tambue, A., Stochastic Exponential Integrators for the Finite Element Discretization of SPDEs for Multiplicative and Additive Noise, *IMECO J. Numer. Anal.*, 2012, drr059.
11. Mel'nikova, I.V. and Al'shanskii, M.A., *Stokhasticheskie uravneniya s neogranichennym operatornym koeffitsientom pri mul'tiplikativnom shume* (Stochastic Equations with Unbounded Operator Coefficient under Multiplicative Noise), *Sib. Mat. Zhurn.*, 2017, vol. 58, no. 6, pp. 1354–1371.
12. Wei, J. and Norman, E., On Global Representations of the Solutions of Linear Differential Equations as a Product of Exponentials, *Proc. Amer. Math. Soc.*, 1964, vol. 15, no. 2, pp. 327–334.
13. Ovsyannikov, L.V., *Gruppovoi analiz differentsial'nykh uravnenii* (Group Analysis of Differential Equations), Moscow: Nauka, 1978.
14. Miller, U., *Simmetriya i razdelenie peremennykh* (Symmetry and Separation of Variables), Moscow: Mir, 1981.
15. Kallianpur, G., *Stokhasticheskaya teoriya fil'tratsii* (Stochastic Filtration Theory), Moscow: Nauka, 1987.
16. Hida, T., *Brounovskoe dvizhenie* (Brownian Motion), Moscow: Nauka, 1987.
17. Shaikin, M.E., Multiplicative Stochastic Systems with Multiple External Disturbances, *Autom. Remote Control.*, 2018, vol. 79, no. 2, pp. 299–309.
18. Kunita, Kh., On the Representation of Solutions of Stochastic Differential Equations, *Seminare de Prob. XIV, Lecture Notes in Math.*, 1980, vol. 784, pp. 282–304, Berlin: Springer-Verlag.
19. Barut, A. and Ronchka, R., *Teoriya predstavlenii grupp i ee prilozheniya* (Group Representation Theory and Its Applications), vol. 1, Moscow: Mir, 1980.
20. Makkin, G., *Stokhasticheskie integraly* (Stochastic Integrals), Moscow: Mir, 1972.
21. Burbaki, N., *Gruppy i algebry Li. Chast' 1* (Lie Groups and Algebras. Part 1), Moscow: Mir, 1976.

This paper was recommended for publication by A.V. Nazin, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 8, pp. 981–992.

CONTROL IN TECHNICAL SYSTEMS

Synthesis of Test Control for Identification of Aerodynamic Characteristics of Aircraft

N. V. Grigor'ev

Public Joint-Stock Company "Gromov Flight Reseach Institute", Zukovskii, Russia
e-mail: lab76@lii.ru

Received December 7, 2022
Revised April 25, 2023
Accepted April 28, 2023

Abstract—The synthesis of a control law for tracking a target informative path as a new approach to solving the problem of planning a flight experiment for identifying the aerodynamic characteristics of automatically controlled aircraft is proposed. The mathematical statement and the method for solving the synthesis problem are obtained. In the numerical experiment, it is shown that the identification accuracy on the synthesized control can be significantly improved compared to the identification accuracy on the optimal program test signal.

Keywords: aerodynamic characteristics, planning of test signals, parametric identification, automatic control

DOI: 10.25728/arcRAS.2023.99.47.001

1. INTRODUCTION

The task of planning test signals for identifying the aerodynamic characteristics (ADC) of an aircraft is to generate a specially perturbed motion of the aircraft in order to increase the accuracy of ADC identification. The disturbed movement of the aircraft (test maneuver) is formed by applying so-called test input signals (test signals) to the aircraft's controls. As a rule, criteria adopted in the theory of optimal design of experiments, which characterize to one degree or another the expected identification accuracy, are used as criteria for selecting a test signal.

Problems of active identification of ADC of aircraft are characterized by a wide variety of mathematical formulations. Already solved problems differ in their mathematical formulations in: test signal dimension (scalar [1–9], vector [1, 4, 9–16]), class of functions in which the test signal is optimized (continuous functions [9], discrete functions [1, 2, 4, 8, 17], polyharmonic functions [2, 10, 12, 15, 16, 18], type "bang-zero-bang" controls and similar controls [2, 6, 8, 12, 14, 15], parameterized controls [2, 4, 7, 11], functions of simple form [5]), by type of restrictions (only for test signal [2, 3, 5, 9] on the components of the state vector of the aircraft in perturbed motion [1–4, 6, 7, 11, 14, 17]), criterion (Turing number [1], L-, D-criterion [1–7, 9, 11, 12, 14, 15, 17, 18], peak factor [10, 12, 15, 16]). It is usually assumed that the choice of test signal is made before the experiment, but the possibility of step-by-step optimization of the test signal during the experiment is also considered [3]. Optimization of test signals is performed most often in the time domain [1–12, 14, 16], but can also occur in the frequency domain [18] or in the time and frequency domain simultaneously [13, 17]. For further presentation, it is important to note that in the known formulations of the problem of active identification of aircraft ADCs, restrictions on the components of the aircraft state vector in perturbed motion do not take into account (except [6]) possible differences between unknown ADCs and their a priori estimates, and the choice of test

signals is made in the class of program controls, i.e. adaptive control for the purpose of active identification of aircraft ADCs is practically not considered [2, 19].

The safety conditions of the flight experiment, various physical and methodological restrictions determine the restrictions on the disturbances of the components of the aircraft state vector in the test maneuver. In a number of important applications, taking into account these restrictions is a necessary condition for performing a test maneuver [11]. If the restrictions are violated, the test maneuver is not performed (interrupted by the aircraft automatic control system). The fulfillment of the restrictions must be ensured whenever the ADC and the initial conditions of the test maneuver are known approximately when selecting a test signal.

In [6] a method for optimizing a test signal is proposed taking into account the specified restrictions in the class "bang-zero-bang" controls. In [7], a method for optimizing a test signal is proposed taking into account the specified restrictions in the class of parameterized controls, in particular, a solution was obtained in the class of piecewise constant functions with a short persistence time, which differs significantly from "bang-zero-bang" management. The program test signal obtained in [7] ensures that the specified restrictions are met for all a priori possible values of the ADC. But a consequence of this positive property of the test signal is its optimality "on average" on the set of all restrictions, determined by the set of possible values of the ADC. This means that in each specific case (in particular, with true ADC values), such a test signal will obviously be suboptimal. In the class of program test signals, it is impossible to select a test signal that will be optimal for all possible values of the ADC. However, it is possible to improve the informative properties of the selected test signal directly during the flight experiment due to the information obtained about the aircraft state vector. In [20] a method for approximate solution of this problem was proposed. Below we propose a method for finding its optimal solution.

2. FORMULATION OF THE PROBLEM

The proposed mathematical formulation of the control synthesis problem for identifying the ADC contains a model of the dynamics of the object in a test mode lasting T seconds, described by a linear (linearized with respect to the reference motion of the aircraft) differential equation

$$\frac{dx}{dt} = A(b)x + Gu,\ t[0,T], \quad x(0) = x_0, \tag{1}$$

and discrete measurement model

$$z_i = z(x(t_i)) = Hx(t_i) + v_i, \quad i = \overline{1,N}, \tag{2}$$

where: x is n-dimensional vector of the aircraft state; $u = u(t,x)$ is optimized dimension control vector m; z_i is p is dimensional vector of measurements; v_i is vector of "white" Gaussian noise of measurements, $E(v_i) = 0$, $E(v_i v_j^T) = 0$, $i \neq j$, $E(v_i v_i^T) = R$, $i = \overline{1,N}$, $j = \overline{1,N}$ (E is mathematical expectation); $A(b)$, G, H is matrices of corresponding dimensions; t_i is timepoints at which measurements are taken, $t_i = h(i-1)$, $h = T/(N-1)$; N is number of measurements. The matrix $A(b)$ depends on the identified vector of unknown parameters b (the desired ADCs) of dimension k. The true values of the b^{true} parameters b are not known. The a priori estimate b^{pri} of the vector b^{true} contains an error Δb, $b^{\text{pri}} = b^{\text{true}} + \Delta b$, with respect to which it is known that the components Δb_i of the vector Δb belong to the intervals $[-\Delta_i, \Delta_i]$: $\Delta b_i \in [-\Delta_i, \Delta_i]$, $i = \overline{1,k}$. We denote the set of possible values of b by the symbol B.

We will assume that the movement of the aircraft before the start of the test maneuver is quasi-stationary. This means that the components of the vector x_0 in (1) are close to zero, but may be different from zero. We will assume that x_0^{true} belongs to the closed bounded set X^0 containing the

zero vector. We will consider the possible values of the components of the vectors x_0 and b to be independent of each other.

It's required by choosing on the interval $[0, T]$ a vector function $u = u(t, x)$ from a certain class of functions U (defined below):

1) ensure the fulfillment of scalar linear restrictions on the state vector of the aircraft for all a priori possible values of b and x_0

$$|x_s(t, b, x_0, u)| \leqslant q_s(t), \quad b \in B, \quad x_0 \in X^0, \quad s = \overline{1, r}, \tag{3}$$

where: x_s is components of the vector x on which restrictions are imposed; $q_s(t)$ is specified functions; r is number of restrictions;

2) minimize the control u functional

$$J = \text{tr}\left(WM^{-1}(b^{\text{pri}}, x_0, u)\right), \tag{4}$$

where tr is matrix trace notation, W is non-negative definite weight matrix (usually diagonal), M is information matrix:

$$M(b, x_0, u) = \sum_{i=1}^{N} (\partial x(t_i, b, x_0, u)/\partial b)^T \; Q(\partial x(t_i, b, x_0, u)/\partial b). \tag{5}$$

In (5) matrix $Q = H^T R^{-1} H$; $x_0 = 0$; derivatives $S_j = \dfrac{\partial x(t, b, x_0, u)}{\partial b_j}$, $j = \overline{1, k}$ are determined from a system of differential equations for sensitivity functions:

$$\begin{cases} \dfrac{dS_j}{dt} = A(b)S_j + \dfrac{\partial A(b)}{\partial b_j} x(t, b, x_0, u), \\ S_j(0) = 0, \quad j = \overline{1, k}. \end{cases} \tag{6}$$

Equations (6) and (1) are solved jointly.

The mathematical formulation of the test signal planning problem in the class of program controls differs from the above formulation of the problem only in that the desired control is sought in a given class of time functions, i.e., $u = u(t)$ (usually in the class of continuous or piecewise continuous functions of time [1–18]).

The solution $u = u(t, x)$ to problem (1)–(4) is proposed to be sought among controls that ensure tracking of a certain trajectory of system (1), which has good information content about the identified parameters and satisfies restrictions (3), and precisely in the class of functions representable in the form

$$u(t, x) = \mu u^{\text{pri}}(t) + L\left(\mu x^{\text{pri}}(t) - x(t)\right), \tag{7}$$

where $x^{\text{pri}}(t) = x(t, b^{\text{pri}}, 0, u^{\text{pri}})$—trajectory of system (1) for optimal program test signal $u^{\text{pri}}(t)$ at restrictions $B = b^{\text{pri}}$, $X^0 = 0$; coefficient μ, $0 \leqslant \mu \leqslant 1$, and elements L_{ij} of matrix L

$$|L_{i,j}| \leqslant C, \quad i = \overline{1, l}, \quad j = \overline{1, n} \tag{8}$$

subject to determination from the minimum condition of criterion (4) under restrictions (3). The constant C reflects restrictions on the feedback coefficients of the automatic control system (ACS). For the convenience of further references, we will call the given problem the problem of selecting a test control, and the desired function $u(t, x(t))$ will be called a test control.

System (1) under control (7) can be written in the form

$$\frac{dx}{dt} = (A(b) - GL)x + \mu G(u^{\text{pri}}(t) + Lx^{\text{pri}}(t)), \quad x(0) = x_0, \tag{9}$$

therefore, for sufficiently small values of the coefficient μ, restrictions (3) will certainly be satisfied. In addition, from (4)–(6) and (9) it follows that for an arbitrary function $u = u(t)$ the equality $J(\mu u) = J(u)/\mu^2$ is true, therefore, to minimize functional (4), the value of μ should be chosen as maximum as possible, subject to the fulfillment of restrictions (3).

The equations for the sensitivity functions S_j, $j = \overline{1,k}$ are written in the form (6), since it is assumed that in the procedure for post-flight estimation of the vector b the technique of artificially disconnecting the system will be used (9), when a signal $u_\Sigma(t)$ known from a flight experiment is applied to the input of a customized motion model with an excluded ACS circuit $u_\Sigma(t) = \mu u^{\text{pri}}(t) + L(\mu x^{\text{pri}}(t) - x(t))$. If the customized model includes an ACS model, then the matrix A in (6) must be replaced by the matrix $A - GL$.

A fairly complete characteristic of the solution to problem (1)–(5) is the distribution density of the function values $J(b, x_0) = \operatorname{tr} M^{-1}(b, x_0, u)$. The function $J(b, x_0)$ characterizes the expected identification error (the lower bound of the sum of variances of parameter estimates) on the test control $u\,(t, x(t))$ (or on the test signal $u(t)$) if $b^{\text{true}} = b$, $x(0) = x_0$. To construct an estimate of a given distribution density (polygon), it is sufficient to calculate the values of the function $J(b, x_0)$ for a plenty large number N_P of pairs of vectors b and x_0, $b \in B$, $x_0 \in X^0$, selected randomly. If the distribution densities of the components of the vectors b and x_0 on the intervals of their possible values are unknown, then according to the recommendations [21] they should be assumed to be uniform. The number N_P is chosen so that when it increases, the position and shape of the polygon do not change. With little computational time spent, the polygon of expected identification error values represents an integral characteristic of test management quality that is convenient for analysis, allowing one to estimate the probability of obtaining certain values of the expected identification error parameters.

3. SOLUTION METHOD

In the case of $B = b^{\text{pri}}$, $X^0 = 0$, the optimal program test signal $u^{\text{pri}}(t)$ and the corresponding trajectory $x^{\text{pri}}(t) = x(t, b^{\text{pri}}, 0, u^{\text{pri}})$ can be found, for example, by one of the methods described in [2, 7]. Further, we present a method for optimizing the coefficient μ and matrix L in (7).

We combine the elements of the matrix L and the coefficient μ into one vector $v \in V$, where V is a hypercube defined by inequalities (8) and the inequality $0 \leqslant \mu \leqslant 1$. The dimension of the vector v is equal to $N_v \leqslant nl + 1$ (some elements of the matrix L can be set equal to zero to eliminate feedback on the corresponding components and reduce the number of adjustable coefficients). We set $v_{N_v} = \mu$. We denote by $x(t, b, x_0, u^v)$ the solution of system (1) on control (7) for a given vector v.

Let N_C be a positive integer. We divide the optimization interval $[0, T]$ with points $t_i = \Delta_C(i-1)$, $i = \overline{1, N_C}$ into subintervals of equal length $\Delta_C = T/(N_C - 1)$. We choose N_C so large that when the restrictions are satisfied

$$\begin{gathered} |x_s(t_i, b, x_0, u^v)| \leqslant q_s(t_i), \quad t_i = \Delta_C(i-1), \quad i = \overline{1, N_C}, \\ b \in B, \quad x_0 \in X^0, \quad s = \overline{1, r} \end{gathered} \tag{10}$$

restrictions (3) can be considered fulfilled for all $t \in [0, T]$ with sufficient accuracy. Thus, to solve the problem posed, it is sufficient to solve the problem of minimizing criterion (4) on the set S of vectors v satisfying the set of restrictions (10).

We define the following auxiliary problem. Minimize by $v \in V$ the criterion

$$J = \operatorname{tr}\left(WM^{-1}(b^{\mathrm{pri}}, 0, u^v)\right) \tag{11}$$

on some closed, bounded set $\check{S}$ of vectors v, defined by a finite number of restrictions

$$\begin{gathered} |x_s(t_i, b^j, x_0^j, u^v)| \leqslant q_s(t_i), \quad i = \overline{1, N_C}, \\ b^j \in B, \quad x_0^j \in X^0, \quad s = \overline{1, r}, \quad j = \overline{1, K}, \quad v \in V. \end{gathered} \tag{12}$$

The solution to this typical nonlinear programming problem can be found by various methods, for example, the linearization method [22]. The gradients of restrictions (12) over the components of the vector v are equal

$$S_{v_j} = \frac{\partial x(t, b, x_0, u^v)}{\partial v_j}, \quad j = \overline{1, N_v}.$$

The gradient of functional (11) can be calculated if the functions are known

$$S_{v_j}^{b_i} = S_{v_j}^{b_i}(t, b, x_0, u^v) = \frac{\partial}{\partial v_j} S_i, \quad i = \overline{1, k}, \quad j = \overline{1, N_v}.$$

The functions $S_{v_j}, S_{v_j}^{b_i}$ can be determined from solving the following systems of equations, which must be solved together with equations (1) and (6):

$$\begin{cases} \dfrac{dS_{v_j}}{dt} = (A(b) - GL)S_{v_j} - G\dfrac{\partial L}{\partial v_j}x(t, b, x_0, u) + \mu G\dfrac{\partial L}{\partial v_j}x^{\mathrm{pri}}, \text{ if } j = \overline{1, N_v - 1}, \\ \dfrac{dS_{v_{N_v}}}{dt} = (A(b) - GL)S_{v_{N_v}} + G(u^{\mathrm{pri}} + Lx^{\mathrm{pri}}), \\ S_{v_j}(0) = 0, \quad j = \overline{1, N_v}; \end{cases}$$

$$\begin{cases} \dfrac{dS_{v_j}^{b_i}}{dt} = A(b)S_{v_j}^{b_i} + \dfrac{\partial A(b)}{\partial b_i}S_{v_j}, \\ S_{v_j}^{b_i}(0) = 0, \quad j = \overline{1, N_v}, \quad i = \overline{1, k}. \end{cases}$$

The solution to the original minimization problem with respect to the vector v of criterion (11) under restrictions (8) and (10) can be obtained by the following iterative algorithm:

Step 0. We set the counter for the number of iterations: $iter = 0$. We define arbitrary $b^j \in B$, $x_0^j \in X^0$, $j = \overline{1, K}$ and define the set S^{iter} as set of vectors v satisfying inequalities and conditions (12).

Step 1. We solve an auxiliary problem in which $\check{S} = S^{iter}$. We denote the solution by v^{iter}, the corresponding test control (7)—by $u^{v^{iter}}$.

Step 2. To check the fulfillment of restrictions (10) on the found control $u^{v^{iter}}$ for each $s = \overline{1, r}$ and $i = \overline{1, N_C}$ we define $\max\limits_{b \in B, x_0 \in X^0} \left|x_s(t_i, b, x_0, u^{v^{iter}})\right|$.

Step 3. If for all $s = \overline{1, r}$, $i = \overline{1, N_C}$

$$\max_{b \in B, x_0 \in X^0} \left|x_s(t_i, b, x_0, u^{v^{iter}})\right| \leqslant q_s(t_i)$$

is true, then problem (11)–(10) is solved—a test control that satisfies restrictions (10) and minimizes functional (11) is found. Next go to Step 5.

Step 4. If for some s^*, i^*

$$\max_{b\in B, x_0\in X^0}\left|x_{s^*}(t_{i^*}, b, x_0, u^{v^{iter}})\right| = \left|x_{s^*}(t_{i^*}, b^*, x_0^*, u^{v^{iter}})\right| > q_{s^*}(t_{i^*})$$

is true, that is, restrictions (10) are violated, then we supplement the set S^{iter} with restrictions $\left|x_{s^*}(t_{i^*}, b^*, x_0^*, u^{v^{iter}})\right| \leqslant q_{s^*}(t_{i^*})$. We again denote the obtained set by S^{iter}, having previously set $iter = iter + 1$. Next go to Step 1.

Step 5. Constructing a polygon of function values $J(b, x^0) = \operatorname{tr}\left(WM^{-1}(b, x_0, u^{opt})\right)$, where $u^{opt} = u^{v^{iter}}$. The method for constructing the polygon was described in Section 2.

We explain: each subsequent set S^{i+1} of vectors v is already contained in the previous set S^i due to the fact that each restriction added at Step 4 narrows the set on which criterion (11) is minimized. Thus $S^0 \supset S^1 \supset \cdots \supset S^i \supset \cdots \supset S$, where S is the set of vectors v defined by the formulas (10). Consequently, the minimum of criterion (11) on the set S is not less than the minimum on the set S^i. Therefore, if at the ith iteration the conditions of Step 3 of the algorithm are met, then restrictions (10) are satisfied, and the minimum found on the set S^i is the minimum on the set S.

Thus, the solution to problem (4), (10) is reduced to solving a sequence of standard nonlinear programming problems that "approximate" the original problem in the vicinity of the desired minimum with the approximation accuracy increasing during iterations. This approach seems preferable to optimization of test signals using the dynamic programming method [6, 14] due to the "curse of dimensionality."

The presented method for solving the problem can be generalized to the case of dependence of the matrices G and H on the identified parameters b.

4. NUMERICAL MODELING

We consider the problem of constructing a two-component ($m = 2$) test control $u(t, x(t))$ on a time interval of eight seconds ($T = 8$) in order to identify the coefficients b_i, $i = \overline{1,5}$ models of aircraft lateral movement [9]

$$\begin{cases} \dot\beta = b_1\beta + w_y + 0.0565\gamma + 0.0289\delta_N, \\ \dot w_x = b_2\beta - 0.935w_x - 0.124w_y + 1.4\delta_N + 2.88\delta_e, \\ \dot w_y = b_3\beta + 0.119w_x + b_4w_y + b_5\delta_N, \\ \dot\gamma = w_x, \end{cases} \tag{13}$$

supplemented with the simplest models of the rudder and aileron drive:

$$\begin{cases} \dot\delta_N = \omega_N, \\ \dot\omega_N = k(\delta_N^{\text{set}} - \delta_N) - k_2\omega_N, \quad \delta_N^{\text{set}} = u_1(t, x(t)), \\ \dot\delta_e = \omega_e, \\ \dot\omega_e = k(\delta_e^{\text{set}} - \delta_e) - k_2\omega_e, \quad \delta_e^{\text{set}} = u_2(t, x(t)), \\ k = \dfrac{0.456}{\tau^2}, \quad k_2 = \dfrac{0.8}{\tau}, \quad \tau = 0.02. \end{cases} \tag{14}$$

In (13) and (14): β is gliding angle of the aircraft, w_x, w_y is angular velocities of roll and yaw, γ is roll angle, δ_N, δ_e are rudder and aileron deflection angles, ω_N, ω_e are rudder and aileron deflection angular rates, k, k_2, τ are parameters of the rudder and aileron drives, coefficients b_1, b_2, b_3, b_4, b_5 are derivatives of the lateral aerodynamic force and aerodynamic moments of roll and yaw to be identified corresponding components of the aircraft state vector: β, w_x, w_y, δ_N. The dimension of angular velocities is—degrees per second, angles are—degrees. The variables β, w_x, w_y, γ, δ_N, δ_e are measured independently at a frequency of 25 hertz.

We have the state vector of the aircraft $x = (\beta, w_x, w_y, \gamma, \delta_N, \delta_e, \omega_N, \omega_e)^T$, vector of identifiable parameters $b = (b_1, b_2, b_3, b_4, b_5)^T$, measurement vector $z_i = z(t_i) = Hx(t_i) + v_i$, $t_i = h(i-1)$,

$i = \overline{1,N}$, where H is matrix with elements $H_{ii} = 1$ when $i = \overline{1,6}$, $H_{ij} = 0$ when $i = \overline{1,6}$, $j = \overline{1,8}$, $i \neq j$; v_i is vector of "white" Gaussian noise of measurements, $E(v_i) = 0$, $E(v_i v_j^T) = 0$, $i \neq j$, $E(v_i v_i^T) = R$, $i = \overline{1,N}$, $j = \overline{1,N}$, $h = 0.04$ s, $N = 201$. The root-mean-square measurement errors ($\sqrt{R_{ii}}$, $i = \overline{1,6}$) are: for β—1°, for w_x, w_y—0.71°/s, for δ_N, δ_e—0.5°.

A priori estimate of the true values b^{true} of the vector b:

$$b^{\text{pri}} = (-0.119,\ -4.43,\ -2.99,\ 0.178,\ 1.55)^T.$$

The boundaries of the tolerance intervals $[-\Delta_i, \Delta_i]$, such that $\Delta b_i \in [-\Delta_i, \Delta_i]$, have the form: $\Delta_i = \pm 0.5 \left| b_i^{\text{pri}} \right|$, $i = \overline{1,4}$, $\Delta_5 = \pm 0.2 \left| b_5^{\text{pri}} \right|$. Thus, the a priori uncertainty of the first four components of the vector b is ±50% of the nominal values. The set of possible values of the vector b defines a parallelepiped with center at the point b^{pri}—set B. The test maneuver should start from a quasi-stationary state:

$$\begin{gathered} |\omega_x(0)| \leqslant 0.25^{\circ/\text{s}}, \quad |\beta(0)| \leqslant 0.5^\circ, \quad |\omega_N(0)| \leqslant 0.025^{\circ/\text{s}}, \quad |\delta_N(0)| \leqslant 0.25^\circ, \\ |\omega_y(0)| \leqslant 0.25^{\circ/\text{s}}, \quad |\gamma(0)| \leqslant 0.25^\circ, \quad |\omega_e(0)| \leqslant 0.025^{\circ/\text{s}}, \quad |\delta_e(0)| \leqslant 0.25^\circ. \end{gathered} \tag{15}$$

The set of possible values of the initial conditions of the test maneuver $x_0 = x(0)$ defines the polyhedron—set X^0. Intervals $I_6 = \pm 0.25^{\circ/\text{s}}$, $I_7 = \pm 0.5^\circ$, $I_8 = \pm 0.025^{\circ/\text{s}}$, $I_9 = \pm 0.25^\circ$, $I_{10} = \pm 0.25^{\circ/\text{s}}$, $I_{11} = \pm 0.25^\circ$, $I_{12} = \pm 0.025^{\circ/\text{s}}$, $I_{13} = \pm 0.25^\circ$, defining possible values x_0, as well as tolerance intervals $I_i = [-\Delta_i, \Delta_i]$, $i = \overline{1,5}$ will be further called intervals of a priori uncertainty.

When constructing polygons of values $J(b, x_0)$, we will assume that the components of the a priori estimate of the vector b and the components of the vector x_0 are uniformly distributed over the intervals of a priori uncertainty $I_i = [-\Delta_i, \Delta_i]$, $i = \overline{1,13}$ and are independent of each other. The matrix W in (4) was assumed to be unit.

We will impose restrictions on the permissible disturbances of each of the components of the vector x in the test maneuver:

$$\begin{gathered} |\omega_N(t,b,x_0,u)| \leqslant 30^{\circ/\text{s}}, \quad |\omega_e(t,b,x_0,u)| \leqslant 30^{\circ/\text{s}}, \quad |\beta(t,b,x_0,u)| \leqslant 3^\circ, \\ |w_x(t,b,x_0,u)| \leqslant 5^{\circ/\text{s}}, \quad |w_y(t,b,x_0,u)| \leqslant 5^{\circ/\text{s}}, \quad |\gamma(t,b,x_0,u)| \leqslant 5^\circ, \\ b \in B, \quad x_0 \in X^0, \quad t \in [0,8]. \end{gathered} \tag{16}$$

The first two restrictions in (16) reflect physical restrictions on the speed of movement of the drives, and the remaining restrictions are intended to ensure the safety of the test maneuver. Time discretization (see (10)) of restrictions (16) was carried out with the parameter $\Delta_C = h$.

The task is to determine such a test control $u^A(t, x(t))$:

$$u_i^A(t,x(t)) = \mu u_i^{\text{pri}}(t) + \sum_{j=1}^{4} L_{i,j}(\mu x_i^{\text{pri}}(t) - x_i(t)), \quad i = 1,2, \tag{17}$$

on which the functional (4) reaches its minimum value. The restrictions on the elements of the matrix $L_{i,j}$ in test control (17) were taken in the form (8) with $C = 0.5, 1, 2$. The optimal test control $u^A(t, x(t))$ was determined in accordance with the algorithm of Section 3. Optimization of the program test signal $u^{\text{apr}}(t)$ with $B = b^{\text{apr}}$, $x_0 = 0$ and replacing restrictions (16) with restrictions

$$\begin{gathered} |\omega_N(t,b^{\text{pri}},0,u)| \leqslant 30^{\circ/\text{s}}, \quad |\omega_e(t,b^{\text{pri}},0,u)| \leqslant 30^{\circ/\text{s}}, \quad |\beta(t,b^{\text{pri}},0,u)| \leqslant 3^\circ, \\ |w_x(t,b^{\text{pri}},0,u)| \leqslant 5^{\circ/\text{s}}, \quad |w_y(t,b^{\text{pri}},0,u)| \leqslant 5^{\circ/\text{s}}, \quad |\gamma(t,b^{\text{pri}},0,u)| \leqslant 5^\circ, \quad t \in [0,8] \end{gathered}$$

was performed by the method described in [7], in the class of parameterized controls, presented in the form

$$u_j^{\text{pri}}(t) = \sum_{i=1}^{50} d_{i+50(j-1)} \sin(\pi i t/T), \quad j = 1,2,$$

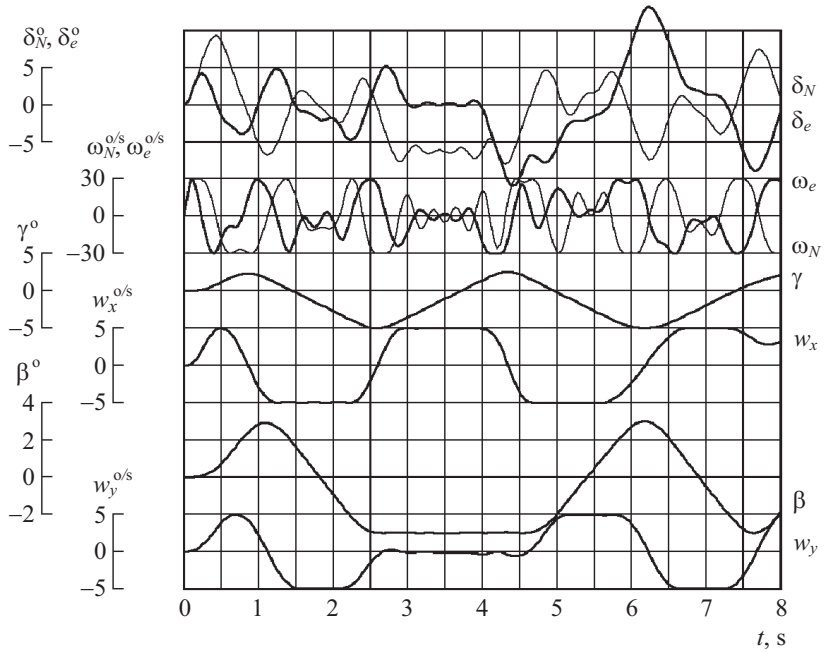


Fig. 1. Optimal solution to the problem in the class of program controls for $B = b^{\text{pri}},\ X^0 = 0$.

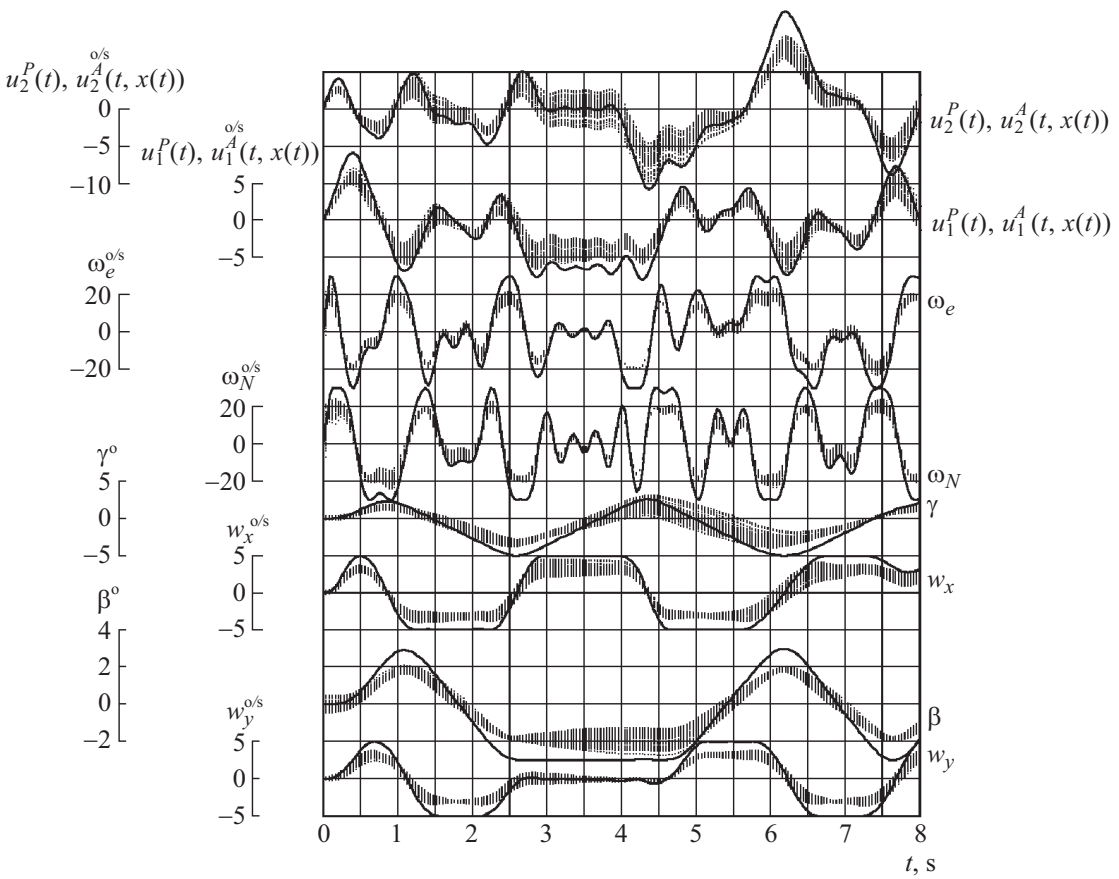


Fig. 2. Trajectory fields of system (1) under optimal test control for $C = 2$. Trajectory components $x^{\text{pri}}(t) = x(t, b^{\text{pri}}, 0, u^{\text{pri}})$ are shown by thick lines.

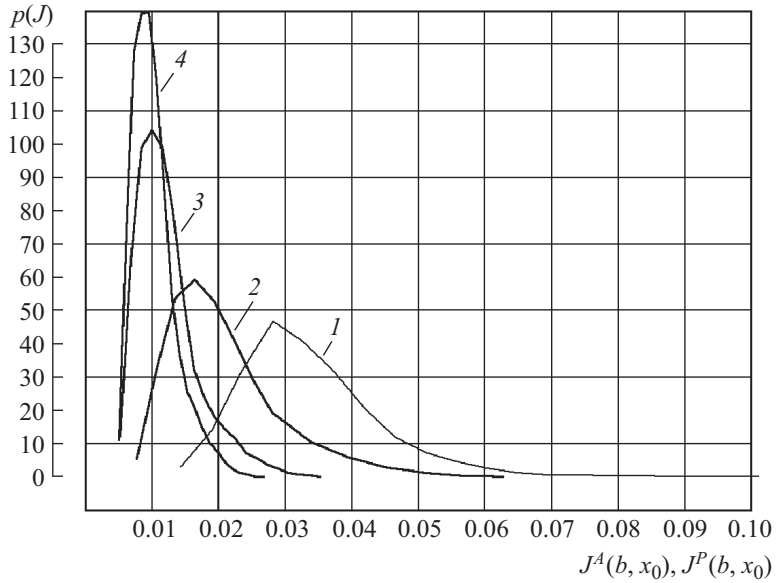


Fig. 3. Polygons of the expected identification error on (*1*) the optimal program test signal and on test controls with C are equal to: (*2*) 0.5, (*3*) 1, (*4*) 2.

where d_i, $i = \overline{1,100}$ is optimized parameters. Figure 1 shows the trajectory $x(t, b^{\text{pri}}, 0, u^{\text{pri}})$ of system (13)–(14), corresponding to the optimal program test signal $u^{\text{pri}}(t)$ for this task. The components of the optimal program test signal $u^{\text{pri}}(t)$ practically coincide with the dependences $\delta_N(t)$, $\delta_e(t)$ shown on the graph. The value of the criterion on the optimal test signal is equal to $\operatorname{tr}\left(M^{-1}(b^{\text{pri}}, 0, u^{\text{pri}})\right) = 0.0036$.

Next, in accordance with the algorithm of Section 3, we found the optimal values of μ and $L_{i,j}$, $i = 1, 2$, $j = \overline{1,4}$ in (17). All corner points of cube B were taken as the initial sample of values $b^j \in B$, $x_0^j \in X^0$, $j = \overline{1,32}$ for $x_0^j = 0$. Finding test controls for each $C = 0.5, 1, 2$ required five to eight iterations of the algorithm. The values of the criterion $\operatorname{tr} M^{-1}\left(b^{\text{pri}}, 0, u^A(t, x(t))\right)$ on optimal test controls are equal to: 0.0089 at $C = 2$; 0.011 for $C = 1$ and 0.018 for $C = 0.5$.

Figure 2 shows the fields of values of the components of the vector x, calculated on the test control $u^A(t, x(t))$ with $C = 2$ for 60 different pairs b^j, x_0^j from a priori possible ones (i.e. for 60 possible solutions of system (13)–(14)). At $C = 1$ and 0.5, the fields of the components of the vector x differed mainly in the larger width of the "tracks" values. The figure shows that all specified restrictions (16) are satisfied. Numerical verification of the fulfillment of restrictions (16) was carried out for 20 000 different pairs b^j, x_0^j for each value of $C = 0.5, 1, 2$. The optimal value of μ for $C = 2$ was equal to $\mu = 0.75$. We note that on the program test signal $u(t) = \mu u^{\text{pri}}(t)$ restrictions (15) would be violated already at $\mu = 0.1$.

At the same time, the limitations and stability of the system were tested (9). In all these cases, all eigenvalues of the matrices $A(b^j) - GL$ had negative real parts.

Figure 3 shows the polygons of expected identification errors $J^A(b, x_0) = \operatorname{tr} M^{-1}\left(b, x_0, u^A(t, x(t))\right)$ on optimal test controls in comparison with the polygon of expected identification errors $J^P(b, x_0) = \operatorname{tr} M^{-1}(b, x_0, u^P(t))$ on the optimal program test signal $u^P(t)$. The program test signal $u^P(t)$ for problem (13)–(16) was found using the method described in [7]. The value of the criterion on the optimal program test signal is $\operatorname{tr}\left(M^{-1}(b^{\text{pri}}, 0, u^P(t))\right) = 0.031$.

The expected identification errors $J^P(b, x_0)$ and $J^A(b, x_0)$ were calculated using solutions to the same systems of equations (13)–(14) and (6), differing only in input signals $u = u(t) = u^P(t)$ and $u = u_\Sigma(t) = u^A(t, x(t))$ respectively. The number of points to construct the polygon was $N_P = 20\,000$.

Figure 3 shows that the test control is significantly better than the program test signal. The polygons of expected identification errors on test controls are located to the left of the polygon of expected identification errors on the program test signal in the region of lower values of expected identification errors. The spread of possible values of the expected identification error in test controls is significantly smaller. The right "tails" of the polygons, corresponding to large values of the expected error, are noticeably shorter on the test controls than on the polygon on the program test signal. When $C = 2$ the average value (standard deviation) of the expected identification error on the test control is more than 3.2 (3.2) times less than the expected error on the program test signal, with $C = 1$—more than 2.7 (2.2) times, with $C = 0.5$—more than 1.6 (1.2) times. Out of 20 000 realizations of the values b and x_0, out of a priori possible ones, the share of realizations for which the ratio of expected identification errors on the test signal and test control was more than two was equal to: when $C = 2$—93%, when $C = 1$—78%, when $C = 0.5$—28%.

We note that, within the framework of the comparison, the formulation of the problem of optimizing the program test signal fit to conditions favorable for identification for conducting a test maneuver with an open control loop.

The optimal values of μ and $L_{i,j}$ in the problem under consideration were such that: $\max\limits_{i,j} L_{i,j} = C$; $\mu = 0.64$ for $C = 2$, $\mu = 0.58$ for $C = 1$, $\mu = 0.45$ for $C = 0.5$. We can assume that the optimal (maximum achievable) values of the parameter μ in control (7) are limited by the value of the parameter C in (8). To confirm this assumption, criterion (4) in this problem was replaced by the criterion $J = \mu$, which was maximized over μ and L under the same restrictions (16), (8) and in that class controls (17). The values of μ and $L_{i,j}$, optimal for the criterion $J = \mu$, obtained at $C = 0.5, 1, 2$ practically did not differ from the corresponding previously obtained values. We note that the problem of maximizing μ is significantly simpler than the problem of minimizing the nonlinear criterion (4).

The a priori uncertainty of the initial conditions of the test maneuver significantly affects the effectiveness of the test control. The influence of this uncertainty can be weakened if feedback is introduced gradually at the beginning of the test maneuver (see [20]). In the considered example, this technique leads to a decrease in the average expected identification error on the test control by 4.2 times (at $C = 2$) compared to the error on the program test signal $u^P(t)$.

5. CONCLUSION

The problem of planning an experiment for parametric identification of an object's motion model is considered under restrictions on permissible disturbances of the object's state vector in the experiment and a priori uncertainty regarding the initial conditions of the experiment. Methods are proposed for solving this problem in the class of feedback controls. This ensures tracking of an object's trajectory that satisfies the specified restrictions and has good informativeness about the identified parameters.

The scope of application of the methods proposed in the article is limited to the tasks of planning experiments to clarify the characteristics of automatically controlled objects, in particular the aerodynamic characteristics of automatically controlled aircraft. It should be expected that the effectiveness of the proposed methods in such problems increases with the increase in the uncertainty of the prior estimates of the identified characteristics and the tightening of restrictions on the permissible disturbances of the object's state vector in the experiment.

The control synthesized for active parametric identification in the class of controls with feedback is proposed to be called test control by analogy with test signals selected in the class of program controls.

The results of statistical modeling, carried out with a fifty percent a priori uncertainty regarding the true values of the identified parameters, confirmed that by choosing a test control, the identification error can be significantly reduced compared to the identification error on the optimal program test signal, both on average and "by probability," i.e., for most priori possible trajectories of object movement.

REFERENCES

1. Kas'yanov, V.A. and Udartsev, E.P., *Opredelenie kharakteristik vozdushnykh sudov metodami identifikatsii* (Determination of Aircraft Characteristics by Identification Methods), Moscow: Mashinostroenie, 1988.
2. Ovcharenko, V.N., *Aerodinamicheskie kharakteristiki letatel'nykh apparatov: identifikatsiya po poletnym dannym* (Aerodynamic Characteristics of Aircraft: Identification by Flight Data), Moscow: LENAD, 2019.
3. Hosseini, B., Diepolder, J., and Holzapfel, F., Online Parameter Estimation and Optimal Input Design, *MMSC*, 2020, pp. 128–139. CEUR-WS.org/vol-2783/paper-09.pdf.
4. Licitra, G., Burgerc, A., Williamsa, P., et al., Optimal Input Design for Autonomous Aircraft, *Control Engineering Practice*, 2018, vol. 77, pp. 15–27.
5. Ovcharenko, V.N., Planning of Identifying Input Signals in Linear Dynamic Systems, *Autom. Remote Control*, 2001, vol. 62, no. 2, pp. 236–247.
6. Hosseini, B., Botkin, N., Diepolder, J., and Holzapfel, F., Robust Optimal Input Design for Flight Vehicle System Identification, *AIAA Scitech 2020 Forum*, 2020. https://doi.org/10.2514/6.2020-0290
7. Grigor'ev, N.V., Test Signal Planning for Identifying the Aerodynamic Characteristics of Automatically Controlled Aircraft Taking into Account the Uncertainty of A Priori Data, *AiT*, 2022, no. 4, pp. 125–139.
8. Jayanti, E.B., Atmasari, N., Mardikasari, H., et al., Pengaruh Masukan Kendali Terhadap Hasil Identifikasi Parameter Pesawat Udara Konfigurasi Konvensional Matra Terbang Longitudinal, *J. Techn. Sist. Comput.*, 2019, no. 7(1), pp. 25–30. https://doi.org/10.14710/jtsiskom.7.1.2019.25-30
9. Gupta, N.K., Hall, W.E., Jr., Input Design for Identification of Aircraft Stability and Control Derivatives, *NASA CR-2493*, 1975.
10. Belokon', S.A., Zolotukhin, Yu.N., and Filippov, M.N., Method of Test Signal Design for Estimating the Aircraft Aerodynamic Parameters, *Avtometriya*, 2017, vol. 53, no. 4, pp. 59–65.
11. Grigor'ev, N.V. and Nesterov, V.E., Active Identification of the ADC of a Re-entry Rocket Unit in Flight Conditions on a Scalable Demonstrator, *Aviakosmicheskaya Tekhnika i Tekhnologiya*, 2014, no. 1, pp. 47–56.
12. Lichota, P., Multi-Axis Inputs for Identification of a Reconfigurable Fixed-Wing UAV, *Aerospace*, 2020, 7. https://doi.org/10.3390/aerospace7080113
13. Roeser, M.S. and Fezans, N., Method for Designing Multi-Input System Identification Signals Using a Compact Time-Frequency Representation, *J. CEAS Aeronaut.*, 2021, vol. 12, pp. 291–306. https://doi.org/10.1007/s13272-021-00499-6
14. Morelli, E.A., Flight Test of Optimal Inputs And Comparison with Conventional Inputs, *J. Aircr.*, 1999, vol. 36(2), pp. 389–397. https://doi.org/10.2514/2.2469
15. Morelli, E.A., Optimal Input Design for Aircraft Stability and Control Flight Testing, *J. Optim. Theory Appl.*, 2021, 191, pp. 415–439. https://doi.org/10.1007/s10957-021-01912-0
16. Grauer, J.A. and Boucher, M., Aircraft System Identification from Multisine Inputs and Frequency Responses, *AIAA Scitech 2020 Forum*, Orlando, FL, USA (2020). https://doi.org/10.2514/6.2020-0287
17. Hosseini, B. and Holzapfel, F., Optimal Input Design for Flight Vehicle System Identification in Frequency Domain, *AIAA Scitech 2022 Forum*, 2022. https://doi.org/10.2514/6.2022-2297

18. Berestov, L.M., Poplavskii, B.K., and Miroshnichenko, L.Ya., *Chastotnye metody identifikatsii letatel'nykh apparatov* (Frequency Methods for Aircraft Identification), Moscow: Mashinostroenie, 1985.

19. Talalay, A.M., Active Identification in the case of Adaptive Control, *Autom. Remote Control*, 1986, vol. 47, no. 2, pp. 1226–1230.

20. Grigor'ev, N.V., Active Identification of Aerodynamic Characteristics: from Test Signal to Test Control, *Polet*, 2022, no. 10, pp. 3–11.

21. Kan, Yu.S. and Kibzun, A.I., *Zadachi stokhasticheskogo programmirovaniya s veroyatnostnymi kriteriyami* (Stochastic Programming Problems with Probabilistic Criteria), Moscow: Fizmatlit, 2009.

22. Pshenichnyi, B.N. and Danilin, Yu.M., *Chislennye metody v ekstremal'nykh zadachakh* (Numerical Methods in Extremal Problems), Moscow: Nauka, 1978.

This paper was recommended for publication by A.A. Galyaev, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 8, pp. 993–1004.

CONTROL IN TECHNICAL SYSTEMS

Angular Motion Control of a Large Space Structure with Elastic Elements

V. Yu. Rutkovskii[*,a], **V. M. Glumov**[*,a], **and A. S. Ermilov**[*,b]

[*] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
e-mail: [a] *vglum@ipu.ru,* [b] *44eas@mail.ru*

Received October 12, 2022
Revised February 6, 2023
Accepted March 30, 2023

Abstract—The task of angular orientation and stabilization of a space structure during its assembly in orbit is solved. The structure includes elastic elements that are installed during the assembly process. The elastic elements of the structure have no sensors to obtain information about their deformation parameters. Control algorithms are proposed to ensure the stability of the angular motion of the structure. A nonlinear extended Kalman filter is used to obtain the necessary information. A joint estimation algorithm for the coordinates of the angular motion of the considered mechanical system and the coordinates of the elastic vibration tones, as well as an algorithm for the identification of their unobservable parameters are developed. The results of mathematical modeling of a variant of the mechanical system of a space structure are presented, which confirm the operability and efficiency of the developed algorithms for estimating coordinates and parameters.

Keywords: mathematical model, control algorithm, space structure, gyroscopic drive, vibration damping, coordinate estimation

DOI: 10.25728/arcRAS.2023.42.67.001

1. INTRODUCTION

Modern spacecraft are dynamic control objects with mechanical structures containing elastic elements. It is noted in [1] that as the size and complexity of the mechanical structure of such vehicles grows, the influence of the elastic properties of the structure on the dynamics of the orientation mode increases. In addition, there is a tendency to complexity of modern spacecraft structure itself, for example, the use of extended elastic elements. Perturbation in the dynamics of spacecraft is also brought by the transformation of elements of the design during operation [2]. With the development of space technology, large-size space structures have emerged, called "large space structures" (LSS), which can be created in space in various ways. LSS—a multidimensional multi-frequency mechanical system with varying parameters [3, 4]. Some of the first LSSs were considered to be large-sized umbrella reflectors, whose structures were envisioned to be built by assembly in space [5]. The development of space robotics makes it possible to solve LSS assembly problems using various robotic devices [6]. In [7] it is noted that the development of space robotics is characterized by two trends. On the one hand, elements of future space infrastructure such as large, multi-modular spacecraft, for example, orbital stations, are expected to be improved, with robotics as an integral component. On the other hand, more and more attention is being paid to robotic servicing, interpreted in a broad sense, which also includes robotic assembly operations for a very broad class of objects [8]. In the robotic environment, space manipulation robots [9], including free-flying [10] robots, are expected to be used extensively.

This paper considers an umbrella-type LSS assembled in space, which is a dynamic control object with variable parameters and a large and discretely time-varying number of degrees of freedom. As a mechanical system, such LSS can be considered as a sequence of intermediate mechanical structures formed in the assembly process. The structure contains elastic elements installed in the assembly process using a space manipulator or a free-flying space manipulation robot. A variant of LSS is considered, in which elastic elements have no sensors of information about coordinates and vibration parameters. One of the main LSS control tasks is orientation control and stabilization of the structure hull axes. The solution of this problem is traditionally obtained on the basis of relay or discrete algorithms [11]. The breaking character of control actions on the hull and shock effects during installation of new structural elements are the causes of elastic oscillations of LSS. When controlling the angular motion of the LSS, a contradiction arises between the main goal of controlling an elastic dynamic object as a rigid body and the need to damp the appearing elastic oscillations. Absence of atmospheric resistance forces leads to accumulation of energy of elastic vibrations in the process of controlling "rigid" motion of LSS. Exceeding the critical amplitude of elastic oscillations and the proximity of their frequencies to the frequencies of controlling the "rigid" motion lead to instability of the system [12]. The lack of accurate determination of the mathematical model (MM) variables in ground conditions leads to the necessity to solve the problem of stable and accurate control of angular motion at all stages of LSS assembly using robust or adaptive control methods of dynamic objects [13].

In [14] an algorithm for LSS orientation control is proposed with low frequencies of elastic vibrations, which significantly affect the quality of transients due to the proximity of natural frequencies of the structure to the frequency of control of its "rigid" motion. In [15] the problem of providing robust stability of elastic vibrations of spacecraft with a nonlinear orientation control system using flywheel motors is solved. The solution is based on the purposeful change of stability region boundaries in the space of object and regulator parameters to maximize the number of robustly stable elastic components of the spacecraft structure. It should be noted that the algorithms providing robust control are effective for the final assembled spacecraft structure (SS). The approach proposed in [14, 15] is limited by the need to obtain current information on the state of the system and its MM parameters. Adaptive control algorithms allow to ensure stability and damping vibration in a wide enough range of LSS elastic vibration natural frequencies with a minimum value of structural damping. In [16] three types of adaptive control strategy for SS on the sequence of stages of its change during assembly in space were defined. The first type: control using analysis and prediction of LSS elastic vibration state. The second type: control with estimation of the phase of the dominant vibrational component in the frequency spectrum of elastic vibrations at the moment of control switching. The third type: control based on fuzzy logic [17]. In [18], an adaptive control algorithm with a reference model for the angular motion of the assembled LSS is proposed. Its functioning does not depend on the intensity and spectral composition of the input influences and does not require the estimation of the elastic vibrations of the LSS. However, the algorithm provides high control accuracy at high energy costs. Currently, attention is paid to the realization of the first type of LSS adaptive control strategy, which uses methods of identification and estimation of the state of the mechanical system of the structure. In [11], active damping of elastic vibrations of the International Space Station structure by the orientation motors using identification algorithms is proposed. To obtain the necessary information for controlling the angular motion of a space structure with an elastic mechanical system, it is reasonable to use estimation algorithms based on the Kalman–Bucy [19] filtering theory. In [20], the task of estimating the coordinates of elastic vibrations of SS using a nonlinear extended Kalman filter is solved. In the present work (as a follow-up to [20]), an algorithm for joint estimation of the coordinates of angular motion of a mechanical system and unmeasured coordinates of elastic vibrations tones, as well as an algorithm for identification of their unobservable parameters, is developed. The problem of forming algorithms

for controlling angular stabilization of SS at the assembly stages is solved. It is assumed that at each stage of assembly there is a connection of a structural element causing elastic vibrations that need to be damped within a given time interval using gyroscopic power drive of the LSS angular stabilization system.

2. MATHEMATICAL MODEL OF ANGULAR MOTION OF LSS

The structure of the umbrella-type LSS mechanical system will be considered as a set of solid bodies, one of which is a carrying body. The other (carried) bodies are building elements attached in one or another order to the carrying body using the spiral scheme of the umbrella-type frame assembly. Such a mechanical system contains non-rigid elements and is characterized by a discretely varying number of degrees of freedom [21]. At the connection points of the structural elements, the rotational degree of freedom in the considered plane of motion and elastic coupling that limits the possible displacements of the elements to the area of small deviations relative to the equilibrium state are taken into account [22]. Using a gyroscopic power drive for LSS assembly containing three identical control moment gyroscope (CMG) installed in a three-beam star pattern, gyrostabilization channel interconnections arise due to inertial and gyroscopic influences [23]. A simplified MM of the spatial angular motion of the mechanical system of the considered LSS type, obtained from the full MM, is presented in detail in [23]. To solve the problem of analytical synthesis of the structure of CMG control algorithms, the model of gyro-power-driven LSS motion, neglecting the cross-effects of CMG motions, can be simplified to three single-type control and gyrostabilization channels of the following form

$$\begin{gathered} I_x\ddot{\chi} + \sum_{i=1}^{n_x} \tilde{I}_{i,x}\ddot{q}_{i,x} - H\dot{\beta} + aI_\beta\ddot{\beta}_s + F(\dot{\chi}) = M_x, \\ a_{i,x}\ddot{\chi} + \ddot{q}_{i,x} + b_{i,x}\dot{q}_{i,x} + c_{i,x}q_{i,x} = 0, \quad i = \overline{1,n_x}, \\ I_\beta\ddot{\beta} + k_d\dot{\beta} + H\dot{\chi} + aI_\beta\ddot{\beta}_s = M_u(u_x), \end{gathered} \tag{1}$$

where $\chi = (\psi, \varphi, \vartheta)^{\mathrm{T}}$ is a vector of hull orientation angles, $\beta = (\beta_\psi, \beta_\varphi, \beta_\vartheta)^{\mathrm{T}}$ is a vector of precession angles of CMG frames, $q = (q_k)^{\mathrm{T}}$ is a composite vector of coordinates characterizing the elastic vibrations of the structure elements along each of the three channels of orientation angles such that $q_k = (q_{i,k})^{\mathrm{T}}$, $i = \overline{1,n_x}$, where n_x is a number of elastic coordinates considered in channel χ_k, $(k = \overline{1,3})$; $\beta_s = [(\beta_\varphi + \beta_\vartheta), (\beta_\psi + \beta_\vartheta), (\beta_\varphi + \beta_\psi)]^{\mathrm{T}}$; $a = \cos(\pi/4 = 0.707$ (for installation of CMGs of "star" type), I_β are moments of inertia of CMG frames; $H = \mathrm{diag}(h_1, h_2, h_3)$ is a diagonal matrix of CMG kinetic moments; k_d is a damping coefficient along the CMG suspension axis; $a_{i,x}, b_{i,x}, c_{i,x}$ are parameters of the equations of vibrations of elastic elements; $I_x = \bar{I}_x + \sum\limits_{i=1}^{n_x} \tilde{I}_{i,x}$, where $\bar{I}_x$ is a diagonal matrix of axial moments of inertia of the hull, $\tilde{I}_{i,x}$ is a matrix of inertial influence of ith elastic element on the dynamics of the structure; $F(\dot{\chi})$ is a vector of nonlinear functions containing products $\chi_i\chi_j$, $i, j = \overline{1,3}$, $i \neq j$; M_x is a vector of disturbing moments of external forces acting on the hull; $M_u(u_x)$ is a vector of control moments applied with respect to the CMG frame axes; u_x is a vector of control voltages, whose components are fed to the inputs of the corresponding CMG momentum drives.

In the mode of angular orientation and stabilization of the LSS at the assembly stage, the values of velocities $\dot{\chi}_k$ are small enough to allow neglecting in $F(\dot{\chi})$ the products $\chi_i\chi_j$, $i, j = \overline{1,3}$, $i \neq j$. In the analytical study of gyro-power-driven control with three identical CMGs, it is reasonable to neglect the interchannel cross-couplings and take $aI_\beta\ddot{\beta}_s = 0$ in (1) [22]. Then the system (1) has

the form

$$\begin{gathered} I_x\ddot{\chi} + \sum_{i=1}^{n_x} \tilde{I}_{i,x}\ddot{q}_{i,x} - H\dot{\beta} = M_x, \\ a_{i,x}\ddot{\chi} + \ddot{q}_{i,x} + b_{i,x}\dot{q}_{i,x} + c_{i,x}q_{i,x} = 0, \quad i = \overline{1, n_x}, \\ I_\beta\ddot{\beta} + k_d\dot{\beta} + H\dot{\chi} + aI_\beta\ddot{\beta}_s = M_u(u_x). \end{gathered} \tag{2}$$

MM (2) is the basis for its decomposition into three subsystems, which correspond to isolated gyrostabilization channels [22].

3. LSS ANGULAR MOTION CONTROL ALGORITHMS

Synthesis of control algorithms for dynamic objects with MM of the form (1) or (2) is traditionally carried out sequentially by two steps [22]. In the first step, the type and parameters of the algorithms forming the values of the components of the vector $u_x(t)$ are determined before the start of the assembly without taking into account elastic vibrations ($q = 0$). Such algorithms are called basic algorithms, during the synthesis of which the MM (2) is transformed to the form of

$$\begin{gathered} I_x\ddot{\chi} - H\dot{\beta} = M_x, \\ I_\beta\ddot{\beta} + k_d\dot{\beta} + H\dot{\chi} + aI_\beta\ddot{\beta}_s = M_u(u_x). \end{gathered} \tag{3}$$

At the second step of synthesis for stabilization and damping of elastic vibrations it is proposed to form a control algorithm in addition to the basic algorithm, which uses information about elastic vibrations of elements and their parameters.

It is reasonable to apply basic algorithms for controlling CMG in the LSS stabilization mode at the stage of assembling PD-algorithms in each kth channel in the form of

$$u_{x,k}(t) = p_{1,k}\chi_k(t) + p_{2,k}\dot{\chi}_k(t), \quad k = \overline{1, 3},$$

where $p_{1,k}, p_{2,k}$ are coefficients, which are chosen taking into account the parameters of the equations (3) and without taking into account the elasticity of the structure from the conditions of ensuring stability and the required quality of control.

The control moments applied relative to the CMG precession axes are formed as [22]

$$M_{u,k}(u_{x,k}) = p_{0,k}(p_{1,k}\chi_k(t) + p_{2,k}\dot{\chi}_k(t)), \quad k = \overline{1, 3}, \tag{4}$$

where the coefficients $p_{0,k}$ are determined by the structure characteristics of the hull and are set depending on the moments of inertia I_x at the assembly stage. It should be noted that MM (3) with algorithms (4) describe a linear dynamic system with constant parameters at the assembly stage, whose stability condition on the angular velocity vector $\dot{\chi}$ is determined from the analysis of its characteristic equations in each kth channel in the form of [22]

$$k_d(h_k + p_{0,k}p_{2,k}) > I_\beta p_{0,k}p_{1,k}, \quad k = \overline{1, 3}. \tag{5}$$

Based on the same characteristic equations, the problem of determining the values of the coefficients $p_{1,k}, p_{2,k}$ of the algorithms (4) that provide the required regulation time $t_{r,k} \approx 3/\eta_k^*$, $k = \overline{1, 3}$ at the coordinates of the vector χ. Here η_k^* are the given values of the stability degrees of the characteristic equations of [22].

Studies of the dynamics of the umbrella-type LSS have shown that, when the number of elastic elements increases, lower frequencies of elastic vibrations appear in the frequency spectrum.

It should be noted that the gyro-power-driven system with the basic algorithm (4) provides the necessary damping of high-frequency elastic vibrations. However, in the low-frequency region, the processes of elastic vibration damping by means of basic control (4) under the condition (5) appear to be overly delayed [22]. Such dynamics of the processes of orientation and stabilization of the angular position of the LSS is unsatisfactory. In addition, the increase in the elastic vibration damping time creates known difficulties when a free-flying space manipulation robot is used in the LSS assembly process [24]. The mentioned disadvantages require complication of the initial basic control algorithm (4). A possible way to correct the basic algorithm is to organize a subsystem of additional gyro-power-driven stabilization of low-frequency elastic vibrations of the LSS, using estimates $\hat{q}_{i,x}, \hat{\dot{q}}_{i,x}$ of the corresponding elastic coordinates. The additional subsystem is connected after the reorientation maneuver is completed and the structural element is installed at the assembly stage. To accelerate the vibration damping, the subsystem generates additional influences of the following type at the CMG inputs

$$M_{d,k}(u_{q,k}) = \sum_{i=1}^{n_k} \tilde{p}_{1,k,i}\hat{q}_{k,i} + \sum_{i=1}^{n_k} \tilde{p}_{2,k,i}\hat{\dot{q}}_{k,i}, \quad k = \overline{1,3}, \tag{6}$$

where $\hat{q}_k,\ \hat{\dot{q}}_k$ are estimation vectors of elastic coordinates and their derivatives, $\tilde{p}_{1,k,i},\ \tilde{p}_{2,k,i}$ are constant coefficients at the assembly stage.

In choosing the values of the coefficients in (6), it is necessary to take into account the values of the parameter estimates in the equations of MM elastic vibrations (2). Estimates of partial frequency values $\omega_{i,x} = \sqrt{c_{i,x}}$ from the low-frequency spectrum of elastic vibrations allow us to choose the coefficients $\tilde{p}_{1,k,i}, \tilde{p}_{2,k,i}$ that ensure stability and minimum damping time of the elastic component [22]. Using the estimates $\hat{\chi}, \hat{\dot{\chi}}$ taking into account (6), the control moments are formed in the form of

$$M_{u,k}(u_{x,k}) = p_{0,k}\left[p_{1,k}\left(\hat{\bar{\chi}}_k - I_x^{-1}\sum_{i=1}^{n_x}\tilde{I}_{i,k}\hat{q}_{i,k}\right) + p_{2,k}\left(\hat{\dot{\bar{\chi}}}_k - I_x^{-1}\sum_{i=1}^{n_x}\tilde{I}_{i,k}\hat{\dot{q}}_{i,k}\right)\right], \quad k = \overline{1,3}. \tag{7}$$

The gain coefficients in (7) at estimates $\hat{q}_{i,k}, \hat{\dot{q}}_{i,k}$ depend on the values of $\tilde{I}_{i,k}$, which can be less than the values of I_x by an order of magnitude or more. For accelerated active compensation of the effect of elastic vibrations on the angular orientation of the LSS, it is reasonable to introduce reconfigurable coefficients $\tilde{p}_{1,k,i},\ \tilde{p}_{2,k,i}$ in (7). Then the algorithms (7) take the following form

$$M_{u,k}(u_{x,k}) = p_{0,k}\left[p_{1,k}\left(\hat{\bar{\chi}}_k - \sum_{i=1}^{n_x}\tilde{p}_{1,k,i}\hat{q}_{i,k}\right) + p_{2,k}\left(\hat{\dot{\bar{\chi}}}_k - \sum_{i=1}^{n_x}\tilde{p}_{2,k,i}\hat{\dot{q}}_{i,k}\right)\right], \quad k = \overline{1,3}, \tag{8}$$

where $\tilde{p}_{1,k,i} \gg p_{1,k}I_x^{-1}\tilde{I}_i,\ \tilde{p}_{2,k,i} \gg p_{2,k}I_x^{-1}\tilde{I}_i$.

If the elastic elements do not have information sensors, it is necessary to solve the problem of obtaining estimates of $\hat{q}$ and elastic vibrations parameters at each stage of LSS assembly after its completion. To solve this problem, a modified version of the Kalman filter-based estimation algorithm proposed in [20] is used.

4. SYNTHESIS OF AN ALGORITHM FOR JOINT ESTIMATION OF COORDINATES ELASTIC VIBRATIONS AND THEIR PARAMETERS

The synthesis of the algorithm for joint estimation of the coordinates of angular motion and coordinates of vibrations (tones) of elastic elements of the structure will be carried out on the

example of an isolated channel $\chi_2 = \varphi$, which is obtained from MM (2) in the form of

$$\begin{gathered} I_\varphi \ddot{\varphi} + \sum_{i=1}^{n} \tilde{I}_i \ddot{q}_i - h_2 \dot{\beta} = M_\varphi, \\ a_i \ddot{\varphi} + \ddot{q}_i + b_i \dot{q}_i + c_i q_i = 0, \quad i = \overline{1, n_x}, \\ I_\beta \ddot{\beta} + k_d \dot{\beta} + h_2 \dot{\varphi} = M_u(u_\varphi), \end{gathered} \tag{9}$$

where $M_u(u_\varphi) = p_\varphi u_\varphi$, $p_\varphi = (p_{1,\varphi}, p_{2,\varphi})$ is a vector of coefficients, $u_\varphi = (\varphi, \dot{\varphi})^{\mathrm{T}}$.

During the synthesis of the estimation algorithm, assume $M_\varphi = 0$. Then the system (9) is transformed to the form [20]:

$$\begin{gathered} \dddot{\bar{\varphi}} - I_\varphi^{-1} h_2 \dot{\beta} = 0, \\ \left(1 - I_\varphi^{-1} \sum_{i=1}^{n} a_i \tilde{I}_i\right) \ddot{q}_i + \left(1 - I_\varphi^{-1} \sum_{i=1, j \neq i}^{n} a_j \tilde{I}_j\right) (b_i \dot{q}_i + c_i q_i) \\ + a_i \sum_{i=1, j \neq i}^{n} \tilde{I}_j (b_j \dot{q}_j + c_j q_j) + a_i h_2 \dot{\beta} = 0, \\ I_\beta \ddot{\beta} + k_d \dot{\beta} + h_2 \left(\dot{\varphi} - I_\phi^{-1} \sum_{i=1}^{n} \tilde{I}_i \dot{q}_j\right) = p_\varphi u_\varphi. \end{gathered} \tag{10}$$

and the angle φ is defined by the expression

$$\varphi = \bar{\varphi} - I_\varphi^{-1} \sum_{i=1}^{n} \tilde{I}_i q_i, \tag{11}$$

where $\bar{\varphi}$ is an angle of rotation of the hull caused by the rotation of the LSS as a rigid object.

The representation of the φ coordinate in the form of (11) allows to apply filtering algorithms for joint estimation of the coordinates of angular motion of the considered mechanical system of LSS with CMG, unmeasured coordinates q_i of elastic vibration tones, and identification of elastic vibration parameters in real time. It should be noted that unlike the [20] system (10) is nonlinear because it contains unknown parameters. A nonlinear extended Kalman filter is used to obtain the estimates. During the synthesis of the estimation algorithm, let represent the MM equations (10) and (11) in the Cauchy form

$$\dot{x}(t) = f(x(t)) + d u_\varphi + C w(t), \tag{12}$$

where $x \in R^{5n+4}$ is a state vector, $x = (\bar{\varphi}, \dot{\bar{\varphi}}, \beta, \dot{\beta}, q_i, \dot{q}_i, a_i, b_i, c_i)^{\mathrm{T}}$, $i = \overline{1, n}$, $b \in R^{5n+4}$ with non-zero element $d_4 = 1$, $f(x)$ is a nonlinear vector-function defined from (10) and (11),

$$\begin{aligned} &f_1 = x_2, \quad f_2 = I_\varphi^{-1} h_2 x_{2n+4}, \quad f_{2i+1} = x_{2i+2}, \quad f_{2n+3} = x_{2n+4}, \\ &f_{2n+4} = I_\beta^{-1} \left[d_4 u_\varphi - k_d x_{2n+4} - h_2 \left(x_2 - I_\varphi^{-1} \sum_{i=1}^{n} \tilde{I}_i x_{2i+2} \right) \right], \\ &f_{2i+2} = (\cdot)^{-1} \Bigg[x_{2n+4+i} h_2 x_{2n+4} - (\cdot)_j (x_{3n+4+i} x_{2i+2} + x_{4n+4+i} x_{2i+1}) \\ &\qquad - x_{2n+4+i} \sum_{j=1, j \neq i}^{n} \tilde{I}_j (x_{3n+4+j} x_{2j+2} + x_{4n+4+j} x_{2j+1}) \Bigg], \end{aligned}$$

where

$$(\cdot) = 1 - I_\varphi^{-1}\sum_{i=1}^{n} a_i\tilde{I}_i, \quad (\cdot)_j = 1 - I_\varphi^{-1}\sum_{j=1,j\neq i}^{n} a_j\tilde{I}_j, \quad j = \overline{1,n}, \quad j \neq i,$$
$$f_{2n+4+i} = f_{3n+4+i} = f_{4n+4+i} = 0;$$

$w \in R^{4n+2}$ is a noise vector, $C = \mathrm{diag}(C_0 \cdots C_i \cdots)$ is a block-diagonal matrix of object noise, containing blocks $C_0 \in R^{4\times 2}$, $C_i \in R^{5\times 4}$. The elements of matrix C_0 are zero except $c_{21} = c_{42} = 1$, matrices C_i also have zero elements except $c_{21,i} = c_{32,i} = c_{43,i} = c_{64,i} = 1$.

It is assumed that in (10) the unknown parameters of elastic vibrations are assumed constant at the assembly stage. If necessary, any parameters can be included in the state vector χ, which leads to cumbersome mathematical expressions.

If only coordinates φ and $\dot{\varphi}$ are measured on board the LSS, the measurement equation has the form

$$z(t) = Gx(t) + v(t), \tag{13}$$

where the measurement vector $z \in R^2$ has coordinates

$$z_1 = x_1 - I_\varphi^{-1}\sum_{i=1}^{n}\tilde{I}_i x_{4+i} + v_1, \quad z_2 = x_2 - I_\varphi^{-1}\sum_{i=1}^{n}\tilde{I}_i x_{4+n+i} + v_2;$$

v is a noise vector of the meters.

The structure of the measurement matrix $G \in R^{2\times(5n+4)}$ has the form [20]

$$G = [C_1 G_2 \cdots G_{i+2}],$$

where C_1, G_2, G_{i+2} are adjoint matrices, $i = \overline{1,n}$; G_1 is a square unit matrix, G_2 is a square zero matrix; the matrix $G_{i+2} \in R^{2\times 5}$ consists of the following non-zero elements: $g_{11,i} = g_{22,i} = -I_\varphi^{-1}\tilde{I}_\varphi$.

It is assumed that the initial values of $x(t_0), w, v$ are independent of each other, w and v are Gaussian white noise with zero mathematical expectations and correlation functions:

$$\mathrm{M}\langle w(t)w^{\mathrm{T}}(\tau)\rangle = Q_w(t)\delta(t-\tau), \quad \mathrm{M}\langle v(t)v^{\mathrm{T}}(\tau)\rangle = Q_v(t)\delta(t-\tau).$$

Here δ is the Dirac delta function, the diagonal noise intensity matrices $Q_w(t)$ and $Q_v(t)$ are continuous and positively defined for $t \geqslant t_0$. Then the problem of synthesizing an algorithm for estimating the coordinates $x(t)$ from the measurements $z(t)$ reduces to a special case of a continuous nonlinear extended Kalman filter [20] with constant matrices C and G:

$$\begin{gathered}\dot{\hat{x}}(t) = f(\hat{x}) + du(t) + P(t)G^{\mathrm{T}}Q_v^{-1}[z(t) - G\hat{x}(t)],\\ \dot{P}(t) = D(\hat{x})P(t) + P(t)D^{-1}(\hat{x}) - P(t)G^{\mathrm{T}}Q_v^{-1}GP(t) + CQ_w(t)C^{\mathrm{T}},\end{gathered} \tag{14}$$

where $\hat{x}(t)$ is a vector of estimates of the coordinates of vector $x(t)$, $P(t)$ is a covariance matrix, $D(\hat{x}) = \partial f(\hat{x})/\partial\hat{x}$ is a Jacobi matrix.

5. MATHEMATICAL MODELING

The investigation of the capabilities of the control algorithm (8) for active compensation of elastic vibrations at the angular orientation of the LSS along φ coordinate was carried out by means of mathematical modeling using in (8) estimates derived from the (14) algorithm. The number of tones and the values of their parameters were assumed to be known and the same in both MM (9)

and the estimation algorithm (14), except for those parameters that are assumed to be unknown in (14). To reduce the simulation time in (9), only two tones $n = 2$ were investigated, and the $\hat{\bar{\varphi}}$, $\hat{\dot{\bar{\varphi}}}$ and $\hat{q}_i$, $\hat{\dot{q}}_i$ LSS estimates were used to form the control moment. The constant parameters c_1 and c_2 were chosen as unknowns, and their estimates $\hat{c}_1$ and $\hat{c}_2$ were used in (14).

The control signal is generated on the basis of (8) as follows

$$u_\phi = p_1\hat{\bar{\varphi}} - \sum_{i=1}^{2}\tilde{p}_{1,i}\hat{q}_{1,i} + p_2\hat{\dot{\bar{\varphi}}} - \sum_{i=1}^{2}\tilde{p}_{2,i}\hat{\dot{q}}_{2,i}, \tag{15}$$

where the coefficients $\tilde{p}_{1,i}$, $\tilde{p}_{2,i}$ have the same order as p_1 and p_2, respectively.

In the modeling of angular orientation dynamics to obtain measurements, a variant of the system (9) with the algorithm (15) was used as MM in the form of [20]

$$\dot{y} = Ay + \bar{d}u_\varphi, \tag{16}$$

where $y \in R^8$ is a state vector, $y = (\bar{\varphi}, \dot{\bar{\varphi}}, \beta, \dot{\beta}, q_1, q_2, \dot{q}_1, \dot{q}_2)^{\mathrm{T}}$, $\bar{d} \in R^8$ is a vector with one non-zero element $\bar{d}_4 = 1$.

Based on (16), a vector of measured coordinates $z = (\varphi^*, \dot{\varphi}^*)^{\mathrm{T}}$ was generated using the expression $z = \bar{G}y + v$, where the matrix $\bar{G} \in R^{2\times 8}$ has non-zero elements $\bar{g}_{1,1} = \bar{g}_{2,2} = 1$, $\bar{g}_{1,5} = \bar{g}_{2,7} = -I_\varphi^{-1}\tilde{I}_1$, $\bar{g}_{2,6} = \bar{g}_{2,8} = -I_\varphi^{-1}\tilde{I}_2$, $v = (v_1, v_2)^{\mathrm{T}}$ is a vector of measurement noise.

In (14), MM (12) was used with vector $x \in R^{10}$, which includes identifiable unknown parameters c_1 and c_2, $x = (\bar{\varphi}, \dot{\bar{\varphi}}, \beta, \dot{\beta}, q_i, \dot{q}_i, c_1, c_2)^{\mathrm{T}}$, $i = \overline{1,2}$, $d \in R^{10}$ is a vector with one non-zero element $d_4 = 1$. $C \in R^{10\times 6}$ is a noise matrix with non-zero elements $c_{2,1} = c_{4,2} = c_{6,3} = c_{8,4} = c_{9,5} = c_{10,6} = 1$. The measurement model for the algorithm (14) is formed as $\hat{z} = G\hat{x}$, where the matrix $G \in R^{2\times 10}$ differs from the matrix $\bar{G}$ by the presence of the ninth and tenth zero columns. Matrices $Q_w \in R^{6\times 6}$ and $Q_v \in R^{2\times 2}$ (14) are assumed constant.

The initial values at $t_0 = 0$ of the coordinates and parameters, as well as the vectors y, $\hat{x}$, and the elements of the diagonal covariance matrix $P(0)$ are assumed to be [20] as follows:

$$y_1(0) = 0.017;\quad y_2(0) = 0.016\ \mathrm{s}^{-1};\quad y_3(0) = 0.18\times 10^{-3};$$
$$y_4(0) = 0.7\times 10^{-4}\ \mathrm{s}^{-1};\quad y_5(0) = 0.017;\quad y_6(0) = 0.19\times 10^{-4}\ \mathrm{s}^{-1};$$
$$y_7(0) = 0.37\times 10^{-2};\quad y_8(0) = 0.13\times 10^{-4}\ \mathrm{s}^{-1};$$
$$a_1 = 1.2;\quad a_2 = 2.32;\quad b_1 = 0.24\ \mathrm{s}^{-1};\quad b_2 = 0.12\ \mathrm{s}^{-1};$$
$$c_1 = (0.34)^2\ \mathrm{s}^{-2};\quad c_2 = (0.47)^2\ \mathrm{s}^{-2};$$
$$I_\varphi = 69\,200\ \mathrm{Nms}^2;\quad \tilde{I}_1 = 1270\ \mathrm{Nms}^2;\quad \tilde{I}_2 = 2500\ \mathrm{Nms}^2;$$
$$I_\beta = 1.1\ \mathrm{Nms}^2;\quad k_d = 2.5\ \mathrm{Nms};\quad h = 240\ \mathrm{Nms};$$
$$p_{1,1} = 3.9\times 10^{-6};\quad p_{2,2} = 3.8\times 10^{-6}\ \mathrm{s}^{-2};\quad p_{3,3} = 0.49\times 10^{-2};$$
$$p_{4,4} = 6.1\times 10^{-4}\ \mathrm{s}^{-2};\quad p_{5,5} = 4.7\times 10^{-4};\quad p_{6,6} = 0.54\times 10^{-2}\ \mathrm{s}^{-2};$$
$$p_{7,7} = 0.11\times 10^{-4};\quad p_{8,8} = 0.11\times 10^{-6}\ \mathrm{s}^{-2};$$
$$p_{9,9} = 1.1\times 10^{-3}\ \mathrm{s}^{-4};\quad p_{10,10} = 1.8\times 10^{-3}\ \mathrm{s}^{-4}.$$

The initial values of the estimates were used:

$$\hat{x}_1(0) = \varphi^*,\quad \hat{x}_2(0) = \dot{\varphi}^*,\quad \hat{x}_j(0) = 0\ \forall j = \overline{3,8}.$$

Given that the parameters c_1 and c_2 can only be positive, then $\hat{x}_9(0) = 0.002\ \mathrm{s}^{-2}$; $\hat{x}_{10}(0) = 0.005\ \mathrm{s}^{-2}$.

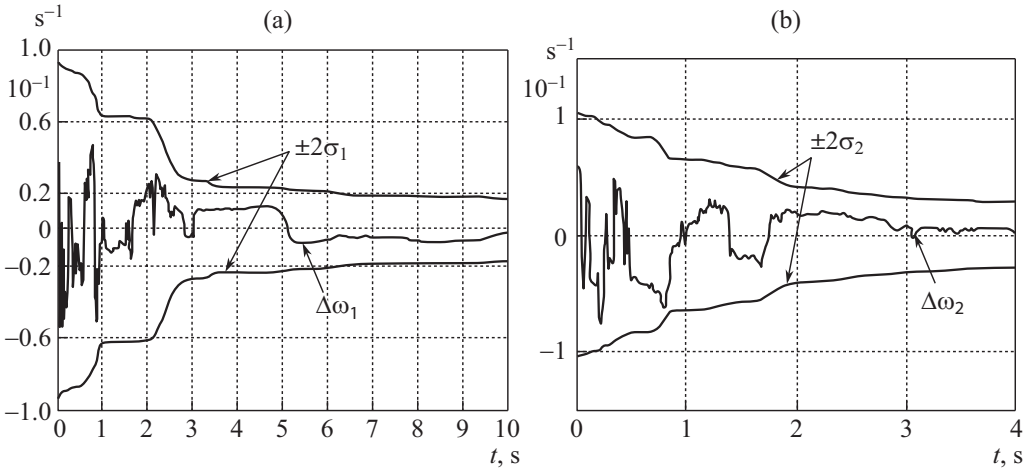


Fig. 1. Identification errors of partial frequencies.

The following standard deviations were assumed for modeling discrete analogs of the continuous white noise of the object and meters:

$$\sigma_{w,1} = 1.5 \times 10^{-5}\ \text{s}^{-2};\quad \sigma_{w,2} = 2 \times 10^{-5}\ \text{s}^{-2};\quad \sigma_{w,3} = 2.2 \times 10^{-6}\ \text{s}^{-2};$$
$$\sigma_{w,4} = 1.8 \times 10^{-6}\ \text{s}^{-2};\quad \sigma_{w,5} = 4.8 \times 10^{-2}\ \text{s}^{-1};\quad \sigma_{w,6} = 3.6 \times 10^{-2}\ \text{s}^{-1};$$
$$\sigma_{v,1} = 2.6 \times 10^{-4};\quad \sigma_{v,2} = 1.34 \times 10^{-5}\ \text{s}^{-1}.$$

The white noise intensity matrices Q_w and Q_v are assumed to be diagonal due to the lack of correlation between the object noise and the noise in the measurement channels. The elements of these matrices are calculated using the expressions

$$q_{w,kk} = 2\sigma_{w,k}^2\tau,\quad k = \overline{1,6}\quad \text{and}\quad q_{w,jj} = 2\sigma_{v,j}^2\tau,\quad j = \overline{1,2},$$

where τ is a correlation time, $\tau \leqslant \Delta t$, Δt is an integration step. The following values are adopted:

$$q_{w,11} = 2.3 \times 10^{-12}\ \text{s}^{-3};\quad q_{w,22} = 0.49 \times 10^{-14}\ \text{s}^{-3};\quad q_{w,33} = 4.8 \times 10^{-14}\ \text{s}^{-3};$$
$$q_{w,44} = 3.2 \times 10^{-14}\ \text{s}^{-3};\quad q_{w,55} = 2.3 \times 10^{-6}\ \text{s}^{-3};\quad q_{w,66} = 1.3 \times 10^{-6}\ \text{s}^{-3};$$
$$q_{v,11} = 4.7 \times 10^{-12}\ \text{s};\quad q_{v,22} = 2.9 \times 10^{-13}\ \text{s}^{-1}.$$

In statistical modeling, discretization of the equations (14) was performed using the fourth-order Runge–Kutta method with Δt, which was chosen to range from 0.002 to 0.005 s.

Figure 1 presents the identification error plots of unmeasured partial frequencies $\Delta\omega_i(t) = \sqrt{c_i} - \sqrt{\hat{c}_i(t)}$, $i = \overline{1,2}$ with doubled standard deviations calculated as the corresponding diagonal elements of the matrix $P(t)$: $\sigma_{w,1} = p_{9,9}^{-4}$, $\sigma_{w,2} = p_{10,10}^{-4}$. From the results of statistical modeling, it follows that the convergence time of the parameter estimates c_1 to 2% of the maximum value of the initial value is on average from 3 to 6 s. At the same time, the convergence time of coordinate estimates $\hat{\varphi}$ and $\dot{\hat{\varphi}}$ to 2% of their maximum values averages 20–25 s.

In order to verify the possibility of using the algorithm (15) to actively compensate for the influence of vibrations of elastic parts of the LSS on its angular dynamics, mathematical simulations in the angular stabilization mode have been carried out. The results of comparative modeling of the angular motion of the LSS along the coordinate φ with the control algorithm (7) and the algorithm (15) are presented in Fig. 2.

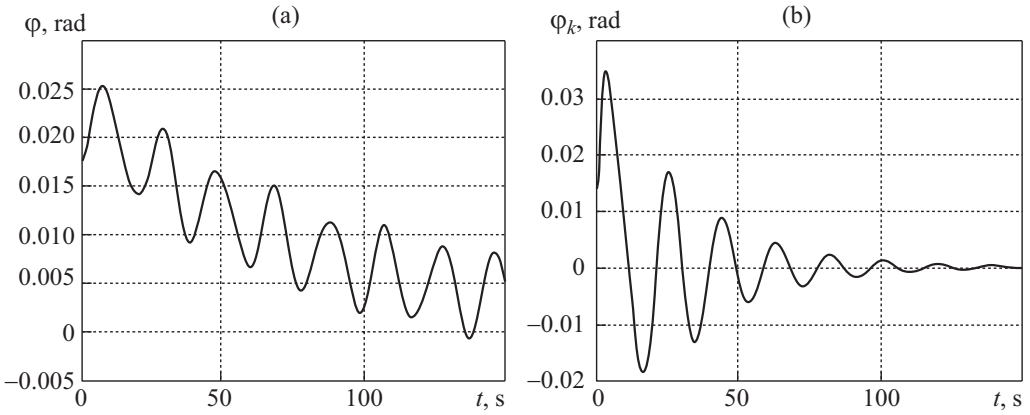


Fig. 2. Stabilization processes for the rotation angle of a structure.

Figure 2a shows plots of the real values of φ (11) obtained when using the algorithm (7) with $\tilde{p}_{1,i} = 0.017$, $\tilde{p}_{2,i} = 0.01$, $i = \overline{1,2}$, in Fig. 2b shows the graphs when using the algorithm (15) in which $\tilde{p}_{1,i} = 3.6$, $\tilde{p}_{2,i} = 2.3$. In the first case, the elastic vibrations decay to 2% of the maximum value of the initial amplitude in ~ 6000 s, while in the second case, with active compensation of the effect of elastic vibrations occurs for ~ 80 s.

In simulations of the stabilization process of the steering angle up to 150 s (see Fig. 2b), errors in the identification of partial frequencies ranged from 0.7 to 1.6% of the true values of the ω_i parameters.

6. CONCLUSION

The problem of vibration damping arises in controlling the angular motion of an LSS assembled in orbit, containing elastic elements, in the absence of information about new mechanical parameters of the assembled structure and initial characteristics arising at each stage of assembly of new elastic components. This requires ensuring not only a timely change of the estimation strategy and, accordingly, of the control during the transition of the structure from one class of mechanical systems to another, but also the application of the principles of adaptive control on the interval of the structure development within each stage of assembly above the first one. The task of optimizing the coefficients in the algorithm (8) at the coordinates of elastic vibration tones from the point of view of rapidity should be solved at the assembly stage, if necessary.

The synthesized algorithm of joint estimation of LSS angular motion coordinates, tones of elastic vibrations of the structure and their parameters allows to obtain with high accuracy estimates of their unmeasured coordinates and parameters in real time based only on the readings of LSS angular motion meters in the absence of any objective information on elastic vibrations.

It should be noted that the construction of an extended Kalman filter for estimating the motion coordinates and their parameters of such a complex mechanical system as the umbrella-type LSS considered in this paper requires using a full MM of a much higher order, in which the mutual influence of vibrational components is taken into account. It is advisable to solve such a problem when developing the system for a particular variant of the assembled structure using an appropriate amount of computational means. This paper considers the principal possibility of using the proposed approach to solve the problem of estimation of such complex dynamic objects.

The use of the synthesized algorithms (8) and (14) in LSS assembly has a number of advantages. Thus, when the first elastic element is installed, the dimensionality of the state vector in the estimation algorithm is increased by five unmeasured coordinates: two coordinates of the elastic

vibration tone and three vibration parameters. Since these parameters remain constant for a long time, after their identification they become known and further identification of them is not reasonable, then these three parameters can be excluded from the state vector. After installation of the next elastic element, the above change of dimensionality is repeated and the state vector is increased by five coordinates. After the identification procedure of the next constant parameters, the state vector is also decreased by the next three coordinates and so on. Thus, the state vector after identification of vibration parameters of all elastic elements of the LSS increases only by $2n$ coordinates, where n—the number of elastic elements installed on the LSS.

The results of statistical mathematical modeling prove the possibility of active compensation of the influence of vibrations of elastic parts of the LSS on the dynamics of angular orientation and stabilization of the LSS itself using control of the form (15). When modeling the algorithm of identification of parameters a, b, c, three variants of estimation were tested, in which one of the three parameters was chosen and the other two were considered to be known.

FUNDING

The work was partially supported by the Russian Foundation for Basic Research under grant no. 20-08-00073.

REFERENCES

1. Akulenko, L.D., Krylov, S.S., Markov, Yu.G., Win, T.T., and Filippova, A.S., Dynamics of Spacecraft with Elastic and Dissipative Elements in the Attitude Control Mode, *Izv. RAN. Teoriya i Sistemy Upravleniya*, 2014, no. 5, pp. 106–115.

2. Gecha, V.Ya., Grinevich, D.V., Chirkov, V.P., and Kanunnikova, E.A., Influence of Elastic Transformable Structural Elements on Spacecraft Stabilization Accuracy, *Spravochnik. Inzhenernyj zhurnal*, 2013, no. 5, pp. 3–6.

3. Nurre, G.S., Ryan, R.S., Scofield, H.N., and Sims, J.L., Dynamics and Control of Large Space Structures, *J. of Guidance, Control, Dynam.*, 1984, vol. 7, no. 5, pp. 514–526.

4. Rutkovsky, V.Yu. and Sukhanov, V.M., Large Space Structure: Models, Methods of Study and Control, *Autom. Remote Control*, 1996, vol. 57, no. 7, pp. 953–963.

5. Mikulas, M.M., Collins, T.J., and Hedgepeth, J.M., Preliminary Design Considerations for 10–40 Meter-Diameter Precision Truss Reflectors, *J. Spacecraft Rockets*, 1991, vol. 28, no. 4, pp. 439–447.

6. Boning, P. and Dubowsky, S., Coordinated Control of Space Robot Teams for the On-Orbit Construction of Large Flexible Space Structures, *Advanced Robotics*, 2010, vol. 24, no. 3, pp. 303–323.

7. Belonozhko, P.P., Space Robotics for Assembly and Service, *Potencial'nye zadachi, koncepcii perspektivnyh sistem Vozdushno-kosmicheskaya sfera*, 2018, no. 4, pp. 84–97.

8. Flores-Abad, A., Ma, O., Pham, K., and Ulrich, S., A Review of Space Robotics Technologies for On-Orbit Serving, *Progr. Aerospace Scie.*, 2014, no. 68, pp. 1–26.

9. Papadopoulos, E., Aghili, F., Ma, O., and Lampariello, R., Manipulation and Capture in Space: A Survey, *Front. Robot. AI.*, 2021, no. 8, pp. 1–36.

10. Ishijima, Yo., Tzeranis, D., and Dubowsky, S., The On-Orbit Maneuvering of Large Space Flexible Structures by Free-Flying Robots, *Pros. of the 8 Int. Symp. on Artificial Intelligence, Robotics and Automation in Space (SAIRAS-2005)*, Munich, 5–8 Sept., 2005, Noordwijk: ESTEC, 2005, pp. 419–426, (ESA SP ISSN 1609-042X. no. 603).

11. Timakov, S.N. and Zhirnov, A.V., Algorithm of Active Damping of Elastic Vibrations of the International Space Station Structure, *Vestnik MGTU im. N.E. Baumana. Ser. "Priborostroenie,"* 2014, no. 3, pp. 37–53.

12. Krutova, I.N. and Sukhanov, V.M., Dynamic Features of Flexible Spacecraft Control in Process of Its Transformation into a Large Space Structure, *Autom. Remote Control*, 2008, vol. 69, no. 5, pp. 774–787.

13. *Teoriya upravleniya. Dopolnitel'nye glavy* (Control Theory. Additional Chapters), Novikov, D.A., Ed., Moscow: LENAND, 2019.

14. Krutova, I.N. and Sukhanov, V.M., Design of Modified PD Algorithm to Control Angular Motion of Large Space Structure, *Autom. Remote Control*, 2009, vol. 70, no. 1, pp. 33–42.

15. Krutova, I.N. and Sukhanov, V.M., Design of Discrete Control System of Flexible Spacecraft Maintaining Robust Stability of Elastic Oscillations, *Autom. Remote Control*, 2009, vol. 70, no. 7, pp. 1109–1119.

16. Rutkovsky, V.Yu., Sukhanov, V.M., Zemlyakov, S.D., and Glumov, V.M., Models and Methods of Control of Large Space Structures with Discretely Changing Construction, *Proceedings of International Conference on Nonlinear Problems in Aviation and Aerospace (ICNPAA-2008). Cambridge Scientific publishers. Ed. by S. Sivasundaram*, 2009, Plenary paper, chapter 12, pp. 115–142.

17. Glumov, V.M., Krutova, I.N., and Sukhanov, V.M., Fuzzy Logic-based Adaptive Control System for In-orbit Assembly of Large Space Structures, *Autom. Remote Control*, 2004, vol. 65, no. 10, pp. 1618–1634.

18. Rutkovsky, V.Yu., Glumov, V.M., and Sukhanov, V.M., New Adaptive Algorithm of Flexible Spacecraft Control, *Complex Systems. Relationships between Control, Communications and Computing*, Dordrecht, The Netherlands: Springer International Publishing, 2016, pp. 313–326.

19. Kalman, R.E. and Bucy, R., New Results in Linear Filtering and Prediction Theory, *Trans. ASME. J. Basic Eng.*, 1961, vol. 83D, pp. 95–108.

20. Ermilov, A.S. and Ermilova, T.V., Estimating Nonmeasurable Coordinates of Elastic Oscillations for Large Space Constructions with a Gyroforse Engine, *Autom. Remote Control*, 2013, vol. 74, no. 9, pp. 1545–1556.

21. Glumov, V.M., Krutova, I.N., and Sukhanov, V.M., A Method of Constructing the Mathematical Model of a Discretely Developing Large Space Structure, *Autom. Remote Control*, 2003, vol. 64, no. 10, pp. 1527–1543.

22. Glumov, V.M., Krutova, I.N., and Sukhanov, V.M., Some Features of Powered Gyrostabilization of a Large Space Structure Assembled in Orbit, *Autom. Remote Control*, 2017, vol. 78, no. 12, pp. 1345–1355.

23. *Dinamika i upravlenie kosmicheskimi ob"ektami* (Dynamics and Control of Space Objects), Matrosova, V.M. and Reshetneva, M.F., Eds., Novosibirsk: Nauka. Sib. Otd. RAN, 1992.

24. Glumov, V.M., Adaptive Control of Free-Flying Space Manipulation Robot in the Working Area when Assembling a Large Space Structure in Orbit, *Proceedings of the 12th International Conference "Management of Large Space System Development" (MLSD)*, M: IEEE, 2019, pp. 1–4. https//ieeeexplore.ieee.org/document/8911089

This paper was recommended for publication by A.I. Matasov, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 8, pp. 1005–1016.

Control of Set of System Parameter Values by the Ant Colony Method

I. N. Sinitsyn[*,**,a] **and Yu. P. Titov**[*,**,b]

[*] *Moscow Aviation Institute (National Research University), Moscow, Russia*
[**] *Federal Research Center for Computer Science and Control, Russian Academy of Sciences, Moscow, Russia*
e-mail: [a] *sinitsin@dol.ru,* [b] *kalengul@mail.ru*

Received January 23, 2023
Revised March 21, 2023
Accepted June 9, 2023

Abstract—The paper considers the modification and application of the ant colony method for the problem of directed enumeration of the values of system parameters when performing calculated multiple calculations. Interaction with the user makes it possible to stop the process of exhaustive enumeration of sets of parameter values, and the application of a modification of the ant colony method will allow us to consider rational sets at early iterations. If the user does not terminate the algorithm, then the proposed modifications allow one to enumerate all solutions using the ant colony method. To modify the ant colony method, a new probabilistic formula and various algorithms of the ant colony method are proposed, allowing for each agent to find a new set of parameter values. The optimal algorithm, according to the research results, is the use of repeated endless cyclic search for a new solution. This modification allows you to consider all solutions, and at the same time, find all the optimal solutions among the first 5% of the considered solutions.

Keywords: ant colony method, parametric graph, reordering, computing cluster, hyperparameter optimization

DOI: 10.25728/arcRAS.2023.91.87.001

1. INTRODUCTION

Nowadays, due to the development of computing clusters, many computational and optimization tasks are transferred from human experts to computing machines. For such systems arise the problems of finding rational values of parameters for solving the computational problem, called hyperparameter optimization [1]. Among the algorithms of hyperparameter optimization one can mention the Bayesian optimizer, which allows finding regularities of influence of individual parameter values (and various combinations) on performance criteria on the basis of hypotheses and a posteriori information [2]. For multi-criteria and multi-extreme problems it is often necessary to consider all sets of parameter values, usually by the method of complete search. This paper considers the possibility of using the ant colony method to solve the problem of directed enumeration of sets of hyperparameters before sending them to a computing system. By interacting with the user, it is possible to stop the method before considering all sets of parameter values when a set satisfying the user's conditions is found. Directed search by the ant colony method will allow to consider rational sets of parameter values as early as possible, but in case of user's dissatisfaction with the results, it will be possible to reconsider all sets of values.

The ant colony method was originally developed for solving the traveling salesman problem [3, 4]. Modern research allows to apply the ant colony method to search for continuous optimiza-

tion. The search methods of CACO (ContinuousAntColonyOptimization) [5], ACOR (Ant Colony Optimization for continuous domain) [6] and CIAC (ContinuousInteractingAntColony) [7, 8] do not involve the use of a graph and have been actively investigated by different researchers, including researchers from Russia [9, 10]. Studies describe the possibility of using the ant colony method for solving problems on graphs: assignment problems with fuzzy execution time [11, 12], problems of finding optimal routes for a group of salesmen [13, 14] and problems of supporting spare parts supply processes [15]. Parametric problems related to finding an optimal set of parameters, classification, dependency detection, etc. are actively researched in the global community [16–19]. For such problems, a special graph structure is created. The presented methods and modifications of the ant colony method are designed to find approximate and rational solutions. Usually all agents (ants) should converge to one solution. The search for new solutions is carried out by the multistart [9].

For a directed search of parameter values, it is necessary not to converge to one solution (set of parameter values) but to consider new solutions sequentially until the user stops the method or considers all possible solutions. In this paper, modifications of the algorithm are proposed to consider all solutions instead of converging to one. The proposed approach allows to solve problems with vector optimality criterion and multimodal target functions without restarting the algorithm. At the same time, the property of optimization algorithms, the fastest finding of optimal solutions, is preserved.

2. MODELS AND METHODS

The ant colony method is based on the probabilistic search for an arc in a graph according to the formula

$$P_{ij,k}(t) = \frac{\tau_{ij}^{\alpha} \times \mu_{ij,t}^{\beta}}{\sum\limits_{z \in J_{i,k}} \left(\tau_{iz}^{\alpha} \times \mu_{iz,t}^{\beta}\right)}. \tag{1}$$

Using (1), the transition probability of an agent from the current vertex i to the vertices from the set $J_{i,k}$ at iteration t is determined. From the computational results, the transition probability to vertex j for the kth agent is determined. The (1) takes the arc length information τ_{ij}^{α} (remoteness) and some weight $\mu_{ij,t}^{\beta}$ (pheromone) into account. The value of τ_{ij}^{α} is fixed and does not depend on the iteration number t. The number of weights $\mu_{ij,t}^{\beta}$ changes between iterations, updating the state of the graph and the external environment for the agents' movement.

To determine the values of the system (solution) parameters, it is suggested to represent the sets of values in the form of a parametric graph. Each specific value of one parameter represents a vertex of the graph. All values of one parameter are combined into layers. From each vertex of one layer there is an arc to each vertex of the next layer. The layers of vertices are arranged in a certain order, which reduces the number of arcs in the graph. An example of a parametric graph is shown in Fig. 1. Similar graphs are found in [13, 15, 18–21].

The weights (pheromone) in such a graph are recorded not on arcs but on vertices. As a result, the arcs are fictitious, and such a parametric graph can be represented as a set of layers (parameters) and a set of vertices (parameter values) [22].

For a parametric graph in the probabilistic formula (1), the value τ_{ij}^{α} can only be set on the basis of a priori information from an expert, but this parameter cannot be defined in general. If a single factor is used in the probability formula, the stagnation of the solution search process increases. The algorithm converges to the first good solution and does not continue the search for the optimal one.

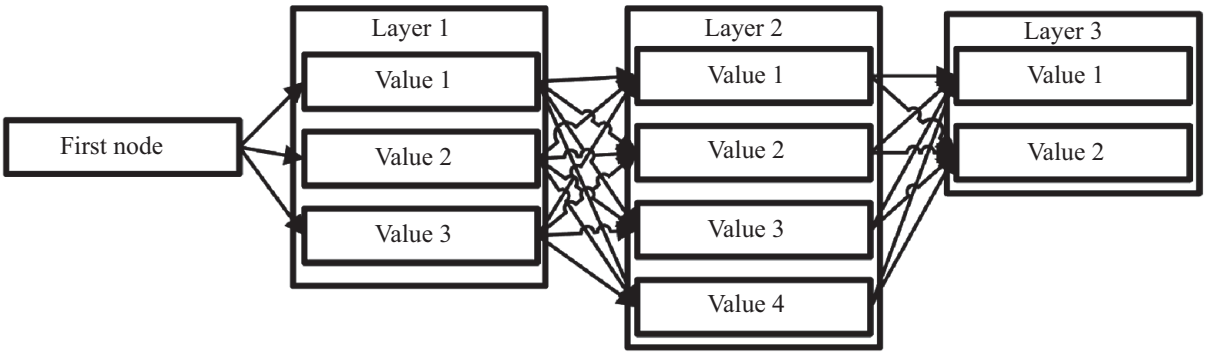


Fig. 1. Diagram of the parametric graph structure.

To solve the stagnation problem, we can "reset" the parametric graph, transfer the state of the graph vertices to the initial state, or modify the probabilistic formula

$$P_{ij,k}(t) = \frac{k1 \times \mu^{\alpha}_{norm\,ij,t} + k2 \times \left(\frac{1}{kol(t)_{ij,t}}\right)^{\beta} + k3 \times \left(\frac{kol(t)_{ij,t}}{MaxKol_{j,t}}\right)^{\gamma}}{\sum\limits_{z \in J_{i,k}} \left(k1 \times \mu^{\alpha}_{norm\,iz,t} + k2 \times \left(\frac{1}{kol(t)_{iz,t}}\right)^{\beta} + k3 \times \left(\frac{kol(t)_{iz,t}}{MaxKol_{z,t}}\right)^{\gamma}\right)}. \tag{2}$$

The formula (2) is a linear convolution (not multiplicative as in (1)) of three summands and weights. The first summand $\mu^{\alpha}_{norm\,ij,t}$ is determined by the number of weights at the ith vertex of the jth parameter (parametric graph layer) at iteration t. To apply this parameter in the weighted sum, it is necessary to use the normalized value. The second summand $(\frac{1}{kol(t)_{ij,t}})^{\beta}$ is determined by the number of visits of agents to the ith vertex of the jth parameter during the running time of the algorithm. This summand increases the probability of visiting a vertex that is rarely present in the solutions and avoids stagnation in the early iterations of the ant colony method. The third summand $(\frac{kol(t)_{ij,t}}{MaxKol_{j,t}})^{\gamma}$ considers the maximum number of possible visits to a vertex: the values of $MaxKol_{j,t}$ for the parameter j at iteration t. Since, in a parametric graph, one vertex (one parameter value) must be selected on each layer, we can calculate the total number of solutions or sets of parameter values. The total number of solutions can be calculated as the product of the number of vertices in each layer. The maximum number of solutions that can contain a particular vertex of a parametric graph is computed as the ratio of the total number of solutions and the number of vertices in a given layer, i.e., for each vertex of layer j, the value $MaxKol_{j,t}$ will be the same. The third summand at later iterations allows to increase the probability of choosing the vertex, for which the majority of solutions are considered. If all possible solutions are considered for a vertex, this vertex can be excluded from the probabilistic search. Additive convolution allows to compensate the values of summands. Vertices with a large number of weights and frequent visits (frequently considered vertices) can be compensated by vertices with a small number of visits (rare vertices) or vertices, for which almost all solutions are considered.

Another specific feature of the algorithm modification is the need to interact with an external computing system. For such modifications, it is necessary to store the state of the system computed for a particular solution. If the ant colony method repeatedly finds a solution, this solution is not sent to the calculator again, and the value of the target function is taken from the hash table [15, 21].

For the considered algorithm, it is important to find a new solution at each iteration. Therefore, if the solution already exists in the hash table, different actions are possible:

1. Using the target function values from the hash table, enter the weights as in the original algorithm.

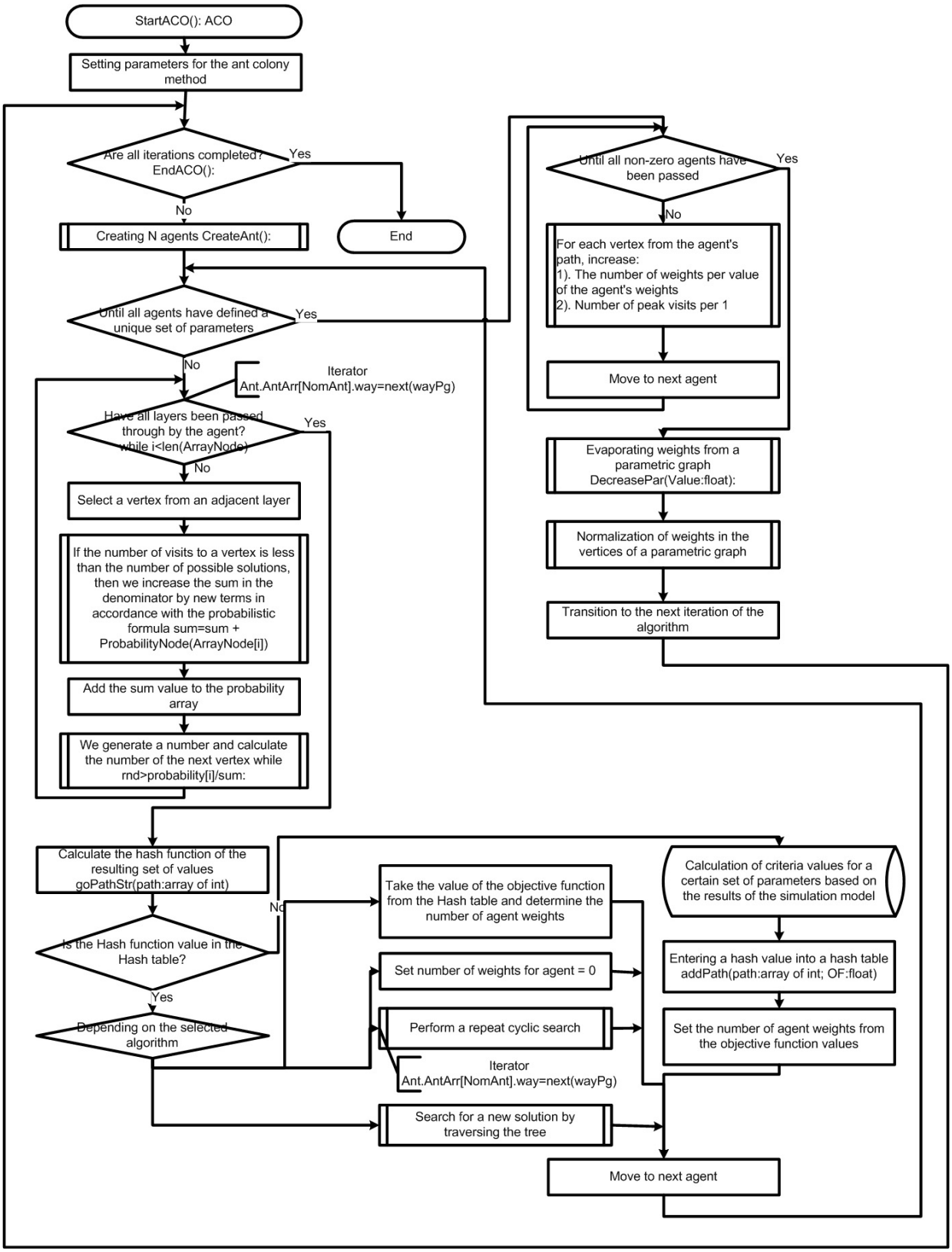


Fig. 2. Schematic diagram of the algorithm of the modified ant colony method.

2. Ignore the agent. The agent does not put weights on the parametric graph.
3. Repeated search for a new solution, not yet considered on the calculator, by the ant colony method with a limited number of iterations. If no new solution is found for the set number of iterations, the agent is ignored.
4. Repeated search for a new solution, not yet considered on the calculator, by the ant colony method with an unlimited number of iterations. The restriction in item no. 3 can solve the stagnation problem.
5. Repeat the search for a new solution by another algorithm. The possibility of traversing the parametric graph as a tree has been considered.

The algorithm flowchart of the proposed modification of the ant colony method is shown in Fig. 2.

3. EXPERIMENT

The objectives of directed enumeration of parameter values are:

—enumeration of all sets of parameter values without any exception;

—the fastest possible acquisition of the optimal set of parameter values.

It can be noticed that both problems contradict each other, because among all sets of parameter values there will always be the best one. But since the system interacts with the calculator and the user in real time, there is a possibility of stopping the program by the user if a satisfactory solution is found. It should also be noted that it is possible to use the vector criterion, which transfers the problem into the area of decision support and multi-criteria optimization [20].

The effectiveness of modifications of the ant colony method is determined by two main evaluations:

—ability of the algorithm to consider all solutions. It is determined by the value of the criterion: "estimate of the probability of finding a new solution by the agent." This estimate is calculated by the ratio of the number of solutions found during the algorithm's running time to the total number of solutions considered;

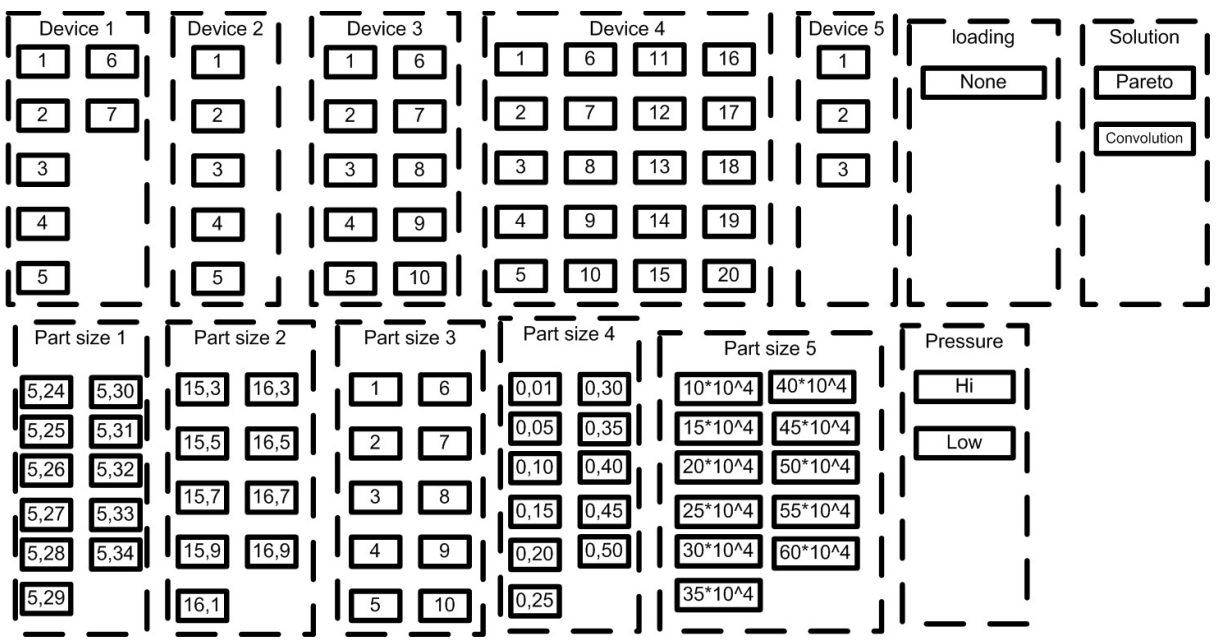


Fig. 3. Diagram of a parametric graph of high dimensionality.

—speed of finding the optimal solution. It is determined by the value of the criterion: "estimate of the expected value of the number of iterations at which this solution was found."

The experiments were performed on the ACO Cluster software written using the Python language. Most of the benchmarks taken from [22, 23] and test graphs of high dimensionality [21] were considered as test data. The structure of the high-dimensional graph is shown in Fig. 3. The presented high-dimensional graph contains more than 10^9 solutions for 13 parameters given in discrete and qualitative scales. With 25 agents per iteration and 20 000 iterations, only 0.05% of the total number of solutions can be considered at most.

To analyze the performance of the algorithm in finding the last (not yet considered) solutions, investigation of the Carrom table function are given in the results:

$$f(x) = -\frac{\left(\cos(x_1)\cos(x_2)e^{|100-(x_1^2+x_2^2)^{0.5}|}\right)^2}{30}. \tag{3}$$

The parametric graph for the function (3) contains two layers: for parameters x_1 and x_2. Each layer of the parametric graph has 201 vertices defining specific parameter values in the range $[-10, 10]$ with precision of 0.1. It takes 1936 iterations of the ant colony method to consider all solutions by 25 agents, assuming that each agent finds a new solution.

4. RESULTS

In this paper, the following modifications of the ant colony algorithm are investigated:

—ACOCC (ACO Cluster Classic)—Classic ant colony method. In (2), the coefficients $k2$ and $k3$ are equal to zero. In this case, the additive convolution formula 2 becomes the multiplicative convolution of the formula (1) with parameter $\tau_{ij}^{\alpha} = 1$. If the solution already considered by the algorithm is found, the value of the target function is determined from the hash table.

—ACOCN (ACO Cluster New)—Similar to the classic ant colony method, but in (2) the coefficients $k2$ and $k3$ are equal to one. The transition probability is determined by additive convolution.

—ACOCNI (ACO Cluster New Ignor)—In (2), the coefficients $k2$ and $k3$ are equal to one. If a solution already written to the hash table is found, this agent is ignored and it does not put weights on the vertices of the parametric graph.

—ACOCCy3 (ACO Cluster Cycle3)—In (2), the coefficients $k2$ and $k3$ are equal to one. If a solution already written to the hash table is found, the solution is searched again using the ant colony method. The search for a new solution is performed cyclically. A limit on the number of iterations of the repeated search is set to 3. If no new solution is found within the set number of iterations, the agent is ignored.

—ACOCCyI (ACO Cluster Cycle Infinity)—In (2), the coefficients $k2$ and $k3$ are equal to one. If a solution already written to the hash table is found, the solution is searched again using the ant colony method. The search for a new solution is performed cyclically with no limit on the number of iterations.

—ACOCT (ACO Cluster Tree)—In (2), the coefficients $k2$ and $k3$ are equal to one. If a solution already written to the hash table is found, the new solution is searched again with a different algorithm. The traversal of the parametric graph as a tree is considered.

Estimates of the results of the modifications for a parametric graph of high dimensionality are given in Tables 1–4. When applying the ACOCN algorithm, the number of solutions required to find the best set of parameter values is minimal, and weakly increases with the number of iterations (3rd column of Table 4). Only the ACOCC algorithm finds a solution faster, but the probability of finding an optimal solution by the ACOCC algorithm is less than 0.1 (2nd column of Table 3). The

Table 1. Estimation of expected value of time of solution search by one agent (in seconds)

Number of iterations	ACOCC	ACOCN	ACOCNI	ACOCCy3	ACOCCyI	ACOCT
2500	1.404×10^{-4}	1.547×10^{-4}	1.562×10^{-4}	1.627×10^{-4}	1.619×10^{-4}	1.887×10^{-4}
5000	1.381×10^{-4}	1.517×10^{-4}	1.560×10^{-4}	1.648×10^{-4}	1.636×10^{-4}	2.745×10^{-4}
7500	1.388×10^{-4}	1.505×10^{-4}	1.567×10^{-4}	1.665×10^{-4}	1.647×10^{-4}	4.158×10^{-4}
10 000	1.391×10^{-4}	1.501×10^{-4}	1.578×10^{-4}	1.690×10^{-4}	1.654×10^{-4}	5.645×10^{-4}
12 500	1.370×10^{-4}	1.547×10^{-4}	1.562×10^{-4}	1.706×10^{-4}	1.657×10^{-4}	6.980×10^{-4}
15 000	1.364×10^{-4}	1.526×10^{-4}	1.569×10^{-4}	1.700×10^{-4}	1.650×10^{-4}	8.593×10^{-4}
17 500	1.328×10^{-4}	1.472×10^{-4}	1.585×10^{-4}	1.695×10^{-4}	1.655×10^{-4}	9.776×10^{-4}
20 000	1.325×10^{-4}	1.469×10^{-4}	1.582×10^{-4}	1.741×10^{-4}	1.688×10^{-4}	10.549×10^{-4}

Table 2. Estimation of the probability of finding a new solution by a single agent per iteration of the ant colony method

Number of iterations	ACOCC	ACOCN	ACOCNI	ACOCCy3	ACOCCyI	ACOCT
2500	0.248	0.618	0.966	1.000	1.000	1.000
5000	0.122	0.381	0.951	1.000	1.000	1.000
7500	0.082	0.282	0.942	1.000	1.000	1.000
10 000	0.063	0.223	0.935	1.000	1.000	1.000
12 500	0.049	0.184	0.929	1.000	1.000	1.000
15 000	0.041	0.158	0.925	1.000	1.000	1.000
17 500	0.035	0.136	0.921	1.000	1.000	1.000
20 000	0.030	0.125	0.918	1.000	1.000	1.000

Table 3. Estimation of the probability of finding a new solution by a single agent per iteration of the ant colony method

Number of iterations	ACOCC	ACOCN	ACOCNI	ACOCCy3	ACOCCyI	ACOCT
2500	0.03	0.31	0.21	0.27	0.24	0.13
5000	0.03	0.49	0.26	0.26	0.33	0.25
7500	0.03	0.50	0.35	0.30	0.32	0.24
10 000	0.02	0.60	0.22	0.32	0.31	0.33
12 500	0.03	0.78	0.32	0.37	0.42	0.27
15 000	0.02	0.71	0.31	0.44	0.37	0.31
17 500	0.03	0.74	0.28	0.39	0.44	0.26
20 000	0.03	0.72	0.41	0.46	0.43	0.39

Table 4. Estimation of expected value of the solution number (in %), at which the optimal parameter values were found

Number of iterations	ACOCC	ACOCN	ACOCNI	ACOCCy3	ACOCCyI	ACOCT
2500	0.801×10^{-6}	2.956×10^{-6}	3.404×10^{-6}	2.970×10^{-6}	3.299×10^{-6}	2.863×10^{-6}
5000	1.122×10^{-6}	3.191×10^{-6}	4.501×10^{-6}	4.415×10^{-6}	5.648×10^{-6}	5.030×10^{-6}
7500	1.014×10^{-6}	3.378×10^{-6}	5.701×10^{-6}	6.137×10^{-6}	5.830×10^{-6}	6.838×10^{-6}
10 000	1.142×10^{-6}	3.667×10^{-6}	5.693×10^{-6}	5.337×10^{-6}	5.910×10^{-6}	7.704×10^{-6}
12 500	0.829×10^{-6}	3.770×10^{-6}	6.779×10^{-6}	6.320×10^{-6}	8.108×10^{-6}	5.269×10^{-6}
15 000	0.740×10^{-6}	3.741×10^{-6}	8.240×10^{-6}	10.174×10^{-6}	8.794×10^{-6}	7.834×10^{-6}
17 500	1.393×10^{-6}	3.845×10^{-6}	8.175×10^{-6}	9.221×10^{-6}	11.678×10^{-6}	15.820×10^{-6}
20 000	1.119×10^{-6}	3.936×10^{-6}	9.864×10^{-6}	11.344×10^{-6}	10.513×10^{-6}	23.592×10^{-6}

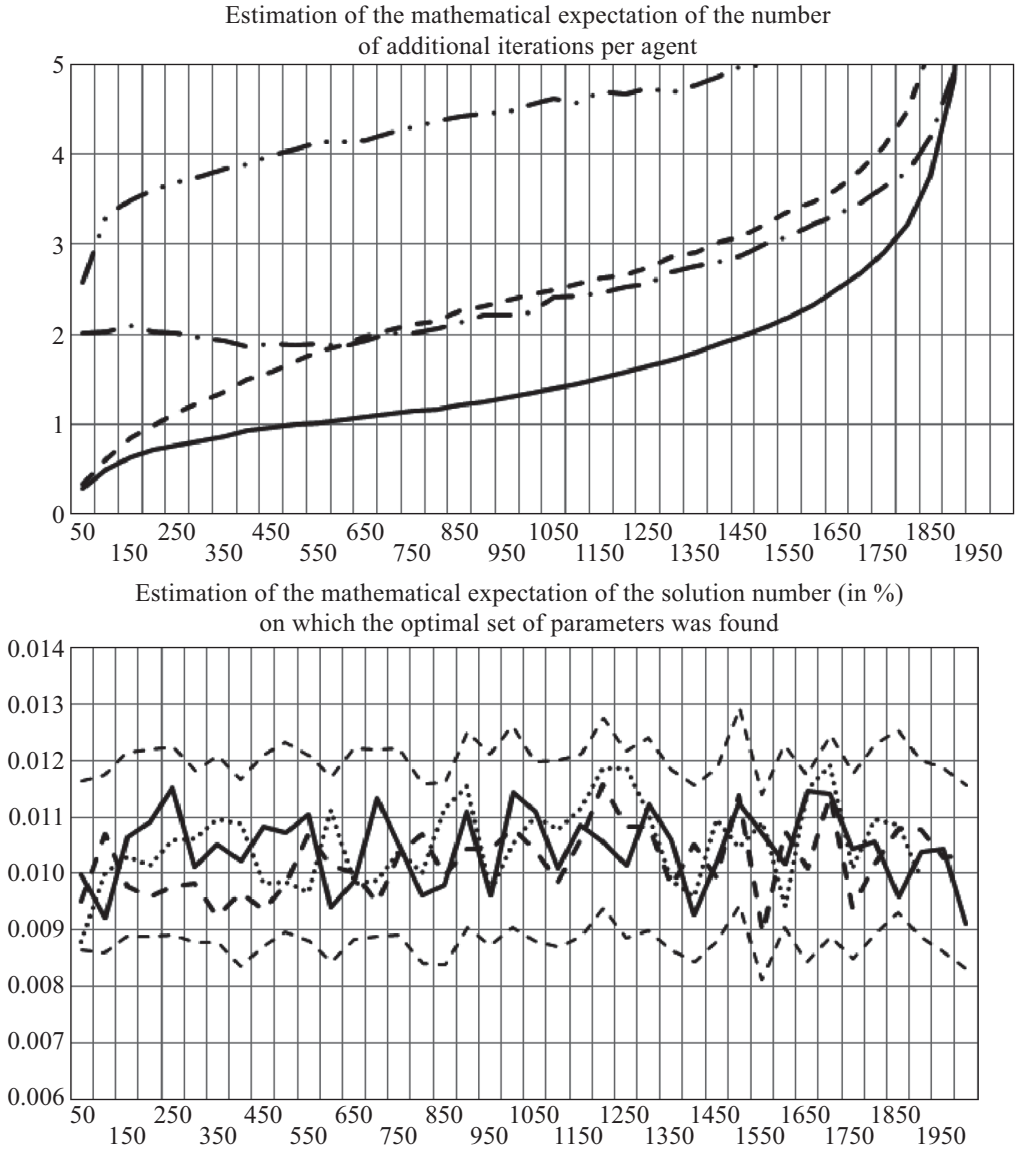


Fig. 4. Dependence of the efficiency of the algorithm on the number of iterations of the ant colony method for the parametric Carrom table function graph.

ACOCC algorithm stagnates at the nearest rational solution. The algorithms that use repeated search for zero agents (ACOCCy3, ACOCCyI and ACOCT) find a new solution for each agent (columns 5–7 of Table 2). For 25 agents, for 2×10^4 iterations, the presented algorithms will consider 5×10^5 sets of parameter values. The time taken for the agent to find a solution does not depend on the number of iterations of the algorithm for most algorithms (Table 1). As a result, it is possible to predict the required number of iterations to find all solutions. The exception is the ACOCT algorithm, which requires more time for one agent to search for a path as the number of iterations increases. The algorithms ACOCC, ACOCN, ACOCNI, ACOCCy3 and ACOCCyI can be ranked in ascending order based on the agent's solution search time. The time of search for a solution by an agent in algorithms ACOCCy3 and ACOCCyI is close, because the number of additional iterations is most often less than the limit of 3 iterations.

It should be noted that the cyclic search actually increases the number of iterations, since it works according to the rules of the ant colony method. If, among ten agents in an iteration, one

Table 5. Estimation of expected value of number of additional iterations by ACOCCyI modification

Number of iterations	1900	1910	1920	1930	1940	1950
$k2$=0; $k3$=0;	8.379	8.712	9.582	10.487	23.004	25.153
$k2$=0; $k3$=0; *ignor*;	8.399	9.028	9.880	10.719	19.658	18.996
$k2$=0; $k3$=1;	5.035	5.248	5.735	6.536	13.408	12.861
$k2$=0; $k3$=1; *ignor*;	5.054	5.291	5.711	6.558	10.658	10.299
$k2$=1; $k3$=0;	7.049	7.780	8.990	11.493	44.962	42.247
$k2$=1; $k3$=0; *ignor*;	7.113	7.744	8.698	10.263	13.658	13.533
$k2$=1; $k3$=1;	4.864	5.312	5.990	7.383	19.585	19.828
$k2$=1; $k3$=1; *ignor*;	4.874	5.314	5.977	7.268	11.172	11.292

agent did not find a new solution and needed ten additional iterations, then twenty iterations were actually performed. Unlike twenty iterations of the original algorithm, there is no updating of weights on the parametric graph for these iterations and the necessary number of iterations can be controlled. Since modifications of ACOCCy3 and ACOCCyI showed good convergence to the optimal solution and each agent in this algorithm finds a new solution (Table 2), let us investigate the possibility of finding all solutions in a parametric graph. For the purpose of the study, we will use a graph of low dimensionality (40 401 solutions) with a target Carrom table function (3). The results of applying the ACOCCyI modification are shown in Fig. 4. Unlike ACOCCyI, ACOCNI modification has considered only 35% of the solutions for 2000 iterations. During the investigation of the ACOCCyI modification, the values of the coefficients $k2$ and $k3$ in (2) were modified.

The long dashed lines with dots in Fig. 4 define the modification that has $k2 = 0$ (at $k3 = 0$ the point is a double point, at $k3 = 1$ the point is a single point). At $k2 = 0$ the optimal solution with probability 1 is determined only after 1000 iterations (more than 60% of the considered solutions). More efficient are the modifications in which $k2 = 1$ (at $k3 = 0$ the line is dashed, at $k3 = 1$ the line is solid). For any number of iterations, the optimal solution is found after examining from 0.009% ($40\,401 \times 0.009 = 363$) solutions to 0.012% (484 solutions)—a confidence interval with a confidence level of 0.99 (bottom plot of Fig. 4). Finding the 0.99% optimal solution requires 2 times less than the considered solutions, sets of parameter values. The ACOCNI modification (line of dots) requires a similar number of considered solutions to find the optimal one.

The top graph of Fig. 4 shows the number of additional iterations required per agent. On average, an agent requires less than 5 additional iterations when considering most possible solutions. As a result, it is inefficient to set a constraint as in the ACOCCy3 modification. The minimum number of additional iterations required to find a new, not yet considered, set of parameter values guarantees minimal delays in the running time of the algorithm.

The inefficient behavior of the modification occurs at the last 36 iterations, at which it is necessary to find the last 900 remaining unexamined solutions (Table 5). The even rows of the table labeled *ignor* are the results of the ACOCCyI modification with the condition of ignoring the vertices for which all solutions are considered. The modification does not consider vertices with $kol(t)_{ij,t} = MaxKol_{j,t}$ in the probability formula (2).

The results presented in Table 5 prove the efficiency of using the third summand in the formula (2). When $k3 = 1$, the number of additional iterations required is significantly smaller than when $k3 = 0$. The smaller number of additional iterations corresponds to a shorter search time for the remaining unexamined sets of parameter values. Ignoring the vertices, for which all possible solutions are considered (lines labeled "ignor"), shows high efficiency.

Both test parametric graphs have several optimal parameter values (equal to the value of the target function). Figure 5 shows the Carrom table function graph and the solution number estimation graph, where the specific values of the parameters x_1 and x_2 were evaluated for 3000 runs of

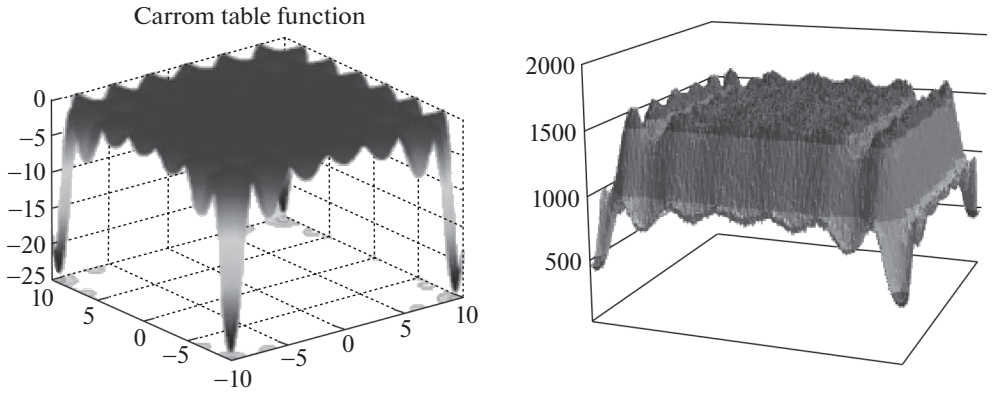


Fig. 5. Function graph and solution number estimates for Carrom table function.

the ACOCCyI modification of the ant colony method. With a large number of runs, the algorithm finds all 4 optima at the earliest iterations. The graph of iteration number estimation visually follows the graph of the function, i.e., the optimal and rational solutions proposed by the modification of the ant colony method are statistically determined at the initial runs.

5. CONCLUSION

In this paper, the problem of directed enumeration of sets of parameter values was considered. To solve the problem, a parametric graph is used, in which the set of parameters is represented as a set of vertices united in groups (layers). This graph has no arcs, and it is necessary to determine the order of the layers, which can lead to differences in the effectiveness of the proposed modifications. But the advantage of this type of parametric graph is the simplicity of its creation for the user. For the ant colony method, we considered the problem of enumerating all values of the system parameters. Optimization properties of the method allow to consider the optimal set of parameter values as early as possible. Modifications are proposed to simplify and control the search for new sets of parameter values using the ant colony method. The use of an interface for interaction with the user is assumed to provide the possibility of stopping the method when a set of system parameter values satisfying the user is found. If such a set is not found, the ant colony method will consider all combinations of discrete parameter values. This approach will also allow to solve multi-extremal and multi-criteria problems.

The considered ant colony method is a rather powerful tool for solving the given problem due to the probabilistic formula for selecting the next vertex. A simple modification of the probabilistic formula is proposed, which allows to significantly improve the performance of the algorithm. This modification defines the probability of choosing the next vertex as an additive convolution of three components: the weights in the vertex, the number of visits to the vertex by agents, and the number of remaining solutions containing this vertex.

In this paper, modifications are proposed that allow choosing different algorithms of behavior when the agent obtains the already considered solution. The verification of the found solution is performed using a hash table. The repeated cyclic search has shown good convergence to the optimal solution and the best performance when searching for the last sets of parameter values on different parametric graphs. The repeated cyclic search performs worse than the original algorithm with the new probabilistic formula.

Consideration of all possible combinations of values of system parameters is a specific task, since it is possible to search all variants by a complete search and, as a result, to find an optimal solution. Most of the algorithms of search for system variants are aimed at finding a rational solution by

convergence to it, and such algorithms do not consider "bad" solutions. Despite the probabilistic convergence of the ant colony method to a single solution, the paper proves that it is possible for the algorithm to consider all solutions, even all suboptimal solutions, under any distribution of the probability formula. For multimodal functions, the multistart procedure, which introduces additional stochasticity into the algorithm, can be abandoned.

The disadvantages of the ant colony method include the presence of a large number of free parameters [9]. The parameters related to the "classical" ant colony method (number of agents per iteration, weight evaporation factor, and addition factor) are discussed in detail in [21]. When investigating the asynchronous behavior of the ant colony method in interaction with an asynchronous computational cluster, it is assumed to set the number of agents equal to the number of threads performing computations on the cluster. The lack of convergence of the algorithm to a single solution does not require the setting of boundary conditions and multistart parameters. But the presence of coefficients in the new probabilistic formula requires additional research. Various discrete values of weight and degree coefficients in formula (2) have been investigated. The optimal values are those that are equal to 1. It should be noted that the situation when the weight coefficients take real values and sum up to 1 requires further research. The dynamic change of the coefficient values also requires further research, since the coefficient k_1 is effective in the early iterations to quickly find the optimal set of parameter values, and the coefficient k_3—to find the remaining solutions in the last iterations.

Further development is expected in the following directions.

1. The proposed alternative algorithm for finding a new solution by traversing the tree has shown low efficiency and requires improvement.
2. The value of individual parameter layers of a parametric graph has not been considered. This study will allow to allocate the most and the least significant parameters of the technical system.
3. The probabilistic formula for the agent's choice of the next vertex is a very powerful tool, and it can be modified based on the value of individual layers.
4. Further research on the structure of the parametric graph, the partitioning of the layer into sub-layers, layer permutations, and methods of graph formation is needed.
5. Application of the proposed method for solving problems with vector criterion. Research of modifications for fast consideration on cluster of all solutions from the Pareto set.
6. When the ant colony method works together with an asynchronous calculator, it is assumed to obtain the values of the target function asynchronously with respect to the operation of the ant colony method. For this modification, it is proposed that each agent is computed in its own thread, and as a result, the asynchronous operation of the ant colony method is considered. The addition and evaporation of weights from the parametric graph can be performed as separate threads.

REFERENCES

1. Feurer, M., Hutter, F., and Vanschoren, J., Hyperparameter Optimization, in *The Springer Series on Challenges in Machine Learning*, Cham: Springer, 2019. https://doi.org/10.1007/978-3-030-05318-5_1
2. Koehrsen, W., *A Conceptual Explanation of Bayesian Hyperparameter Optimization for Machine Learning*, 2018. https://towardsdatascience.com/a-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f
3. Colorni, A., Dorigo, M., and Maniezzo, V., Distributed Optimization by Ant Colonies, in *Proc. First Eur. Conf. on Artific. Life*, Paris: Elsevier Publishing, 1992, pp. 134–142.
4. Dorigo, M. and Stützle, T., *Ant Colony Optimization*, Cambridge: MIT Press, 2004.
5. Socha, K. and Dorigo, M., Ant Colony Optimization for Continuous Domains, *Eur. J. Oper. Res.*, 2008, vol. 185, no. 3, pp. 1155–1173. https://doi.org/10.1016/j.ejor.2006.06.046

6. Mohamad, M., Tokhi, M., and Omar, O.M., Continuous Ant Colony Optimization for Active Vibration Control of Flexible Beam Structures, *IEEE International Conf. on Mechatronics (ICM)*, 2011, no. 4, pp. 803–808.
7. Karpenko, A.P. and Chernobrivchenko, K.A., Efficiency of Optimization by a Continuously Interacting Ant Colony Method (CIAC), *Nauka i obrazovanie: scientific edition of Bauman Moscow State Technical University*, 2011, no. 2. https://doi.org/10.7463/0211.0165551
8. Karpenko, A.P. and Chernobrivchenko, K.A., Multimemetic Modification of Hybrid Ant Algorithm for Continuous Optimization HCIAC, *Nauka i obrazovanie: scientific edition of Bauman Moscow State Technical University*, 2012, no. 9. https://doi.org/10.7463/0912.0470529
9. Karpenko, A.P., Modern Search Engine Optimization Algorithms. Algorithms Inspired by Nature, *Bauman Moscow State Technical University publishing house*, Moscow, 2017, 2nd ed.
10. Simon, D., *Algoritmy evolyutsionnoi optimizatsii: prakticheskoe rukovodstvo* (Evolutionary Optimization Algorithms: A Practical Guide), Moscow: DMK Press, 2020.
11. Sudakov, V.A. and Titov, Y.P., Modified Method of Ant Colonies Application in Search for Rational Assignment of Employees to Tasks, in *Proceedings of 4th Computational Methods in Systems and Software 2020*, vol. 2, Vsetin: Springer Nature, 2020. P. 342–348. DOI 10.1007/978-3-030-63319-6_30
12. Khakhulin, G.F. and Titov, Yu.P., Decision Support System for the Supply of Military Aircraft Spare Parts, *Izv. of Samara scientific center of RAS*, 2014, vol. 16, nos. 1–5, pp. 1619–1623.
13. Sinitsyn, I.N. and Titov, Yu.P., *Razvitie stohasticheskikh algoritmov murav'inoi organizatsii* (Development of Stochastic Algorithms for Ant Organization), in *Bionika—60 let. Results and Prospects. Collection of Articles of the First International Scientific and Practical Conference*, Karpenko, A.P., Ed., Dec. 17–19, 2021, Moscow, 2022, pp. 210–220. https://doi.org/10.53677/9785919160496_210_220
14. Titov, Y.P., Modifications of the Ant Colony Method for Aviation Routing, *Autom. Remote Control*, 2015, vol. 76, no. 3, pp. 458–471. https://doi.org/10.1134/S0005117915030091
15. Sudakov, V.A., Bat'kovskii, A.M., and Titov, Yu.P., Speedup Algorithms for Modifying Ant Colony Method for Finding Rational Assignment of Employees to Tasks with Uncertain Completion Times, *Sovremennye informatsionnye tekhnologii i IT-obrazovanie*, 2020, vol. 16, no. 2, pp. 338–350. https://doi.org/10.25559/SITITO.16.202002.338-350
16. Parpinelli, R., Lopes, H., and Freitas, A., Data Mining with an Ant Colony Optimization Algorithm, *IEEE Trans. Evol. Comput.*, 2002, vol. 6, no. 4, pp. 321–332.
17. Junior, I.C., Data Mining with Ant Colony Algorithms, *ICIC. LNCS*, 2013, vol. 7996, pp. 30–38.
18. Martens, D., De Backer, M., Haesen, R., and Vanthienen, J., Classification with Ant Colony Optimization, *IEEE Trans. Evol. Comput.*, 2007, vol. 11, no. 5, pp. 651–665.
19. Pasia, J.M., Hartl, R.F., and Doerner, K.F., Solving a Bi-objective Flowshop Scheduling Problem by Pareto-Ant Colony Optimization, *ANTS*, 2006, pp. 294–305.
20. Titov, Yu.P., Experience in Modeling Supply Planning Using Modifications of the Ant Colony Method in High Availability Systems, *Sistemy vysokoi dostupnosti*, 2018, vol. 14, no. 1, pp. 27–42.
21. Sinitsyn, I.N. and Titov, Yu.P., Optimization of Hyperparameter Ordering of Computational Cluster by Ant Colony Method, *Sistemy vysokoi dostupnosti*, 2022, vol. 18, no. 3, pp. 23–37. https://doi.org/10.18127/j20729472-202203-02
22. Mishra Sudhanshu, K., *Some New Test Functions for Global Optimization and Performance of Repulsive Particle Swarm Method*, University Library of Munich, Germany, MPRA Paper, 2006. https://doi.org/10.2139/ssrn.926132
23. Layeb Abdesslem, *New Hard Benchmark Functions for Global Optimization*, 2022. https://doi.org/10.48550/arXiv.2202.04606

This paper was recommended for publication by O.P. Kuznetsov, a member of the Editorial Board