Volume 85, Number 7,
July 2024

ISSN 0005-1179
CODEN: AURCAT

A JOURNAL OF RUSSIAN ACADEMY OF SCIENCES

# AUTOMATION AND REMOTE CONTROL

Automation and Remote Control

Editor-in-Chief
Andrey A. Galyaev

Vol. 85, No. 7, July 2024

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

0005-1179/24. *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 85 (12 issues) is published in 2024.
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. 85, No. 7, 2024**

## Linear Systems

## Nonlinear Systems

## Stochastic Systems

## Control in Social Economic Systems

## Optimization, System Analysis, and Operations Research

## Selected Articles from *Avtomatizatsiya v Promyshlennosti*

*ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 7, pp. 655–664.*
*© The Author(s), 2024 published by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 2024.*
*Russian Text © The Author(s), 2024, published in Avtomatika i Telemekhanika, 2024, No. 7, pp. 3–14.*

# A Frequency-Domain Criterion for the Quadratic Stability of Discrete-Time Systems with Switching between Three Linear Subsystems

**V. A. Kamenetskiy**

*Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
*e-mail: vlakam@ipu.ru*

Received March 13, 2024
Revised May 21, 2024
Accepted May 30, 2024

**Abstract**—Connected systems with switching between three linear discrete-time subsystems are considered, and a new frequency-domain criterion for the existence of a quadratic Lyapunov function ensuring the stability of such systems under arbitrary switching is proposed. The application of this criterion is demonstrated on an example of a third-order system.

*Keywords*: discrete-time switched systems, stability, Lyapunov functions, matrix inequalities

**DOI:** 10.31857/S0005117924070014

## 1. INTRODUCTION

The theory of discrete-time systems has been actively developing lately. Various aspects of this theory have been discussed in relatively recent publications [1–7]; also, see the bibliography therein. This paper is devoted to the quadratic stability problem of connected discrete-time systems [3] with switching between three linear stationary subsystems under any switching laws. The term "connected system" will be explained below. By quadratic stability we mean the stability of a system that can be established using a Lyapunov function from the class of quadratic forms or quadratic Lyapunov functions (QLFs). For a connected system with switching between two subsystems, this problem is equivalent to the absolute stability problem of a discrete-time system with a single nonlinearity [3], and a quadratic stability criterion for such a system is the well-known Tsypkin's criterion [8]. In the case of switching between two subsystems, connectedness means that the rank of the difference of the matrices determining the switched subsystems is one.

For connected discrete-time systems with switching between three linear subsystems, a frequency-domain criterion for the existence of a QLF was established in [3]. The disadvantages are an excessively cumbersome procedure for obtaining this criterion and an excessively cumbersome form of the final result. They can be explained as follows. The quadratic stability of a switched system ensues from the existence of a common quadratic Lyapunov function (CQLF). In the case under consideration, the existence of a CQLF is determined by the feasibility of a system of three Lyapunov linear matrix inequalities (LMIs) for discrete-time systems. This system of LMIs is connected, and one resulting matrix inequality equivalent to it was derived in [3]. However, (a) this matrix inequality is not an LMI and (b) the frequency-domain conditions of its feasibility cannot be obtained based on the generalized Kalman–Szegö–Popov lemma [9, 10], as it was done in [3] in the case of Tsypkin's criterion. To overcome inconvenience (b), a fractional linear transformation was used in [3] to pass from the system of LMIs for discrete-time systems to the equivalent system of Lyapunov LMIs for continuous-time systems. The resulting matrix inequality for this system is again not an LMI, but its feasibility conditions were established in [3] in the form of a frequency-domain criterion based on the frequency theorem [11, p. 54] (the Kalman–Yakubovich–Popov (KYP) lemma). The conditions of this criterion are expressed through the elements of a "transfer matrix" for the continuous-time system obtained by the transformation. Finally, using a rather cumbersome procedure, these elements are expressed through the elements of the "transfer matrix" of the original discrete-time system.

In this paper, we apply a new result (Theorem 2 of [12]) to the original system of three Lyapunov LMIs for discrete-time systems to obtain an equivalent resulting matrix inequality that is an LMI. Next, we demonstrate that the feasibility of this LMI can be established by the generalized Kalman–Szegö–Popov lemma in the frequency-domain criterion form. This yields a new frequency-domain criterion for the quadratic stability of the systems under consideration, the main aim of the paper.

Section 2 describes the system of three Lyapunov LMIs for discrete-time systems, whose feasibility is equivalent to the quadratic stability of the systems under consideration. The main result of this paper—the frequency-domain criterion for quadratic stability—is presented in Section 3. A numerical example of a third-order system is provided in Section 4; for this system, the proposed criterion is applied to analytically find the entire quadratic stability domain on the parameter.

## 2. PROBLEM STATEMENT

Consider a linear discrete-time switched system of the form

$$x(t+1) = A(t)x(t), \quad A(t) \in \overline{A} = \{A_1, A_2, A_3\}, \tag{1}$$

where $A_s \in \mathbb{R}^{n\times n}$ and $A(t) : \mathrm{Z}_+ \longrightarrow \overline{A}$ is a mapping from the set $\mathrm{Z}_+$ of nonnegative integers into $\overline{A}$. By assumption, the matrices $A_s$ are stable (Schur, see [13]), i.e., $r(A_s) = \max\limits_{\nu} |\mu_\nu(A_s)| < 1$ for $s = \overline{1,3}$, where $\mu_\nu$ denote the eigenvalues of the matrix $A_s$. The stability of the switched system (1) will be analyzed using QLFs of the form

$$v(x) = x^\top Lx, \quad L = L^\top = \|l_{ij}\|_{i,j=1}^{n}, \tag{2}$$

where the symbol $\{\cdot\}^\top$ means transpose.

According to [3], the existence of a QLF (2) is determined by the feasibility of the system of LMIs

$$I_s = A_s^\top LA_s - L < 0, \quad s = \overline{1,3}. \tag{3}$$

System (1) is connected [3] if the matrices $\{A_1, A_2, A_3\}$ can be represented as

$$\begin{aligned} A_1 &= A, \\ A_2 &= A + b_1c_1^\top, \qquad b_i, c_i \in \mathbb{R}^n. \\ A_3 &= A + b_2c_2^\top, \end{aligned} \tag{4}$$

In this case, system (3) can be written in the form

$$\begin{aligned} I_1 &= A^\top LA - L < 0, \\ I_2 &= (A + b_1c_1^\top)^\top L(A + b_1c_1^\top) - L < 0, \\ I_3 &= (A + b_2c_2^\top)^\top L(A + b_2c_2^\top) - L < 0. \end{aligned} \tag{5}$$

The problem under consideration is to obtain a frequency-domain criterion for the feasibility of the system of LMIs (5).

## 3. SYSTEMS WITH SWITCHING BETWEEN THREE LINEAR DISCRETE-TIME SUBSYSTEMS

To investigate the feasibility of system (5) we use Theorem 2 of [12]. In the formulas below, the symbols “•” denote the elements below the principal diagonal of an appropriate symmetric matrix that coincide with the corresponding elements above this diagonal.

**Theorem 1.** *Let the inequalities in the system*

$$I_1 < 0, \quad I_2 = I_1 + Q_1 < 0, \quad I_3 = I_1 + Q_2 < 0 \tag{6}$$

*be LMIs with respect to the unknown variable* $\nu$, *i.e.,* $I_s = I_s(\nu)$, $s = \overline{1,3}$, *and* $Q_j(\nu) = p_j(\nu)q_j^\top + q_j p_j^\top(\nu)$, *where* $p_j = p_j(\nu)$ *linearly depends on* $\nu$ *and* $q_j$ *is independent of* $\nu$, $j = 1, 2$. *Then system* (6) *is equivalent to the single matrix inequality*

$$\widehat{\widehat{I}} = \begin{pmatrix} I_1(\nu) & p_1(\nu) + \dfrac{\tau_1}{2}q_1 & p_2(\nu) - p_1(\nu) + \dfrac{\tau_2}{2}q_2 - \dfrac{\tau_1}{2}q_1 \\ (\bullet)^\top & -\tau_1 & \dfrac{\tau_1 - \tau_2 + \tau_3}{2} \\ (\bullet)^\top & \bullet & -\tau_3 \end{pmatrix} < 0, \tag{7}$$

*which is an LMI with respect to* $(\nu, \tau_1, \tau_2, \tau_3)$.

With Theorem 1 applied to system (5), the feasibility of system (5) becomes equivalent to the feasibility of the single matrix inequality with respect to the elements of the matrix $L$ and the three additional parameters $\tau_1, \tau_2, \tau_3$. The applicability of Theorem 1 to system (5) and the resulting matrix inequality follow from the relations below. Let the matrix $I_1(\nu)$ be the matrix $(A^\top LA - L)$ of system (5), i.e., $I_1(\nu) = I_1(L) = A^\top LA - L$. (The role of the parameter $\nu$ is played by the matrix $L$.) The difference of the matrices $(I_2 - I_1)$ from (5) can be represented as $p_1 q_1^\top + q_1 p_1^\top$ :

$$\begin{aligned} I_2 - I_1 = A_2^\top L A_2 - A_1^\top L A_1 &= (A + b_1 c_1^\top)^\top L (A + b_1 c_1^\top) - A^\top LA \\ &= (A^\top L + c_1 b_1^\top L)(A + b_1 c_1^\top) - A^\top LA \\ &= A^\top L b_1 c_1^\top + c_1 b_1^\top LA + c_1 b_1^\top L b_1 c_1^\top. \end{aligned} \tag{8}$$

With the notations $p_1^0 = p_1^0(L) = A^\top L b_1$ and $\delta_{11} = \delta_{11}(L) = b_1^\top L b_1$, we have

$$I_2 - I_1 = p_1^0 c_1^\top + c_1 (p_1^0)^\top + \delta_{11} c_1 c_1^\top = p_1 q_1^\top + q_1 p_1^\top, \tag{9}$$

where $p_1 = p_1(L) = A^\top L b_1 + \left(\frac{\delta_{11}(L)}{2}\right) c_1$ and $q_1 = c_1$.

Similarly, let $p_2^0 = p_2^0(L) = A^\top L b_2$ and $\delta_{22} = \delta_{22}(L) = b_2^\top L b_2$; then

$$I_3 - I_1 = p_2^0 c_2^\top + c_2 (p_2^0)^\top + \delta_{22} c_2 c_2^\top = p_2 q_2^\top + q_2 p_2^\top, \tag{10}$$

where $p_2 = p_2(L) = A^\top L b_2 + \left(\frac{\delta_{22}(L)}{2}\right) c_2$ and $q_2 = c_2$.

Thus, by Theorem 1, system (5) is equivalent to the single matrix inequality

$$\widehat{\widehat{I}} = \begin{pmatrix} A^\top LA - L & p_1(L) + \dfrac{\tau_1}{2}c_1 & p_2(L) - p_1(L) + \dfrac{\tau_2}{2}c_2 - \dfrac{\tau_1}{2}c_1 \\ (\bullet)^\top & -\tau_1 & \dfrac{\tau_1 - \tau_2 + \tau_3}{2} \\ (\bullet)^\top & \bullet & -\tau_3 \end{pmatrix} < 0, \tag{11}$$

which is an LMI with respect to $(L, \tau_1, \tau_2, \tau_3)$.

Now we demonstrate that the feasibility of the LMI (11) is determined based on the generalized Kalman–Szegö–Popov lemma [10].

**Lemma 1.** *The LMI* (11) *is equivalent to the LMI*

$$\begin{pmatrix} A^\top LA - L & A^\top L\widehat{B} + \dfrac{\widehat{C}\tau}{2} \\ \widehat{B}^\top LA + \dfrac{\tau\widehat{C}^\top}{2} & \widehat{B}^\top L\widehat{B} - \Gamma \end{pmatrix} < 0, \tag{12}$$

*where*

$$\widehat{B} = \begin{pmatrix}\widehat{B}_1 & \widehat{B}_2\end{pmatrix} = \begin{pmatrix}b_1 & b_2 - b_1\end{pmatrix}, \ \widehat{C} = \begin{pmatrix}\widehat{C}_1 & \widehat{C}_2\end{pmatrix} = \begin{pmatrix}c_1 & c_2 - \dfrac{\widehat{\tau}_1}{\widehat{\tau}_2}c_1\end{pmatrix},$$

$$\tau = \begin{pmatrix}\widehat{\tau}_1 & 0 \\ 0 & \widehat{\tau}_2\end{pmatrix}, \ \Gamma = \begin{pmatrix}\widehat{\tau}_1 & \dfrac{-\widehat{\tau}_1 + \widehat{\tau}_2 - \widehat{\tau}_3}{2} \\ \bullet & \widehat{\tau}_3\end{pmatrix}.$$

The proof of Lemma 1 is given in the Appendix.

Necessary and sufficient conditions for the feasibility of the LMI (12) are determined in the form of a frequency-domain inequality from the generalized Kalman–Szegö–Popov lemma [9, 10]. As a result, we arrive at the following quadratic stability criterion for system (1).

**Theorem 2.** *Let the matrix $A$ be Schur* ($r(A) < 1$), *and let there exist numbers $\widehat{\tau}_s > 0$, $s = \overline{1,3}$, such that $\Gamma > 0$ and the frequency-domain inequality*

$$D(\lambda) = \Gamma + \operatorname{Re}\,[\tau\widehat{C}^\top(A - \lambda E_n)^{-1}\widehat{B}] > 0 \tag{13}$$

*holds for all $\lambda \in \mathbb{C}$, $|\lambda| = 1$, where $E_n$ is an identity matrix of dimensions $(n \times n)$. (In this inequality, $\operatorname{Re} W = (W + W^*)/2$, $W^* = \overline{W}^\top$ is the Hermitian conjugate to $W$; from this point onwards, the symbol $\{\overline{\cdot}\}$ means complex conjugation and the inequality sign is interpreted as the positive definiteness of an appropriate Hermitian form.) Then the connected system* (1) *has a CQLF* (*system* (5) *is feasible, and system* (1) *is stable*). *If system* (5) *feasible, then such a set of numbers $\widehat{\tau}_s > 0$, $s = \overline{1,3}$, exists.*

Let us write the frequency-domain condition (13) in detail. It seems logical to treat $W(p) = C^\top(A - pE_n)^{-1}B$, $p \in \mathbb{C}$, as an analog of the transfer matrix for system (1), where $C = \begin{pmatrix}c_1 & c_2\end{pmatrix}$ and $B = \begin{pmatrix}b_1 & b_2\end{pmatrix}$. With the notation $\Delta(p) = (A - pE_n)^{-1}$, we have

$$W(p) = C^\top\Delta(p)B = \begin{pmatrix}w_{11} & w_{12} \\ w_{21} & w_{22}\end{pmatrix}, \quad \text{where} \quad w_{ij}(p) = c_i^\top\Delta(p)b_j. \tag{14}$$

For the sake of simplicity, we eliminate the hats, using $\tau_s$ instead of $\widehat{\tau}_s$. From (13) it follows that

$$D(\lambda) = \Gamma + \operatorname{Re}\tau\widehat{W}(\lambda) = \Gamma + 1/2\left[\tau\widehat{W}(\lambda) + \widehat{W}^*(\lambda)\tau^\top\right],$$

where

$$\widehat{W}(\lambda) = \widehat{C}^\top\Delta(\lambda)\widehat{B}\begin{pmatrix}c_1 & c_2 - \dfrac{\tau_1}{\tau_2}c_1\end{pmatrix}^\top\Delta(\lambda)\begin{pmatrix}b_1 & b_2 - b_1\end{pmatrix}$$

$$= \begin{pmatrix} w_{11}(\lambda) & w_{12}(\lambda) - w_{11}(\lambda) \\ w_{21}(\lambda) - \dfrac{\tau_1}{\tau_2}w_{11}(\lambda) & w_{22}(\lambda) - \dfrac{\tau_1}{\tau_2}w_{12}(\lambda) - w_{21}(\lambda) + \dfrac{\tau_1}{\tau_2}w_{11}(\lambda) \end{pmatrix}.$$

Finally, we write the inequality $D(\lambda) > 0$ from (13) as

$$D(\lambda) = \Gamma + \frac{1}{2}\begin{pmatrix} 2\tau_1 \mathrm{Re}\, w_{11} & \tau_1 w_{12} + \tau_2 \overline{w_{21}} - 2\tau_1 \mathrm{Re}\, w_{11} \\ \overline{(\bullet)} & 2\tau_1 \mathrm{Re}\,(w_{11} - w_{12}) + 2\tau_2 \mathrm{Re}\ (w_{22} - w_{21}) \end{pmatrix} > 0. \tag{15}$$

(For the sake of brevity, $w_{ij}$ is taken instead of $w_{ij}(\lambda)$.)

*Remark 1.* Theorem 2 remains valid when replacing inequality (13) with inequality (15), where $w_{ij} = w_{ij}(\lambda)= c_i^\top \Delta(\lambda) b_j$, $i, j = 1, 2$.

If system (1) is a triangular switched system [3], i.e., $c_1 = c_2 \triangleq c$, then $w_{11} = w_{21} \triangleq W_1 = c^\top \Delta(\lambda) b_1$ and $w_{22} = w_{12} \triangleq W_2 = c^\top \Delta(\lambda) b_2$. In this case, inequality (15) can be written as

$$D(\lambda) = \begin{pmatrix} \tau_1(1 + \mathrm{Re}\, W_1) & \dfrac{-\tau_1 + \tau_2 - \tau_3 + \tau_1 W_2 + \tau_2 \overline{W_1}}{2} - \tau_1 \mathrm{Re}\, W_1 \\ \overline{(\bullet)} & \tau_3 + (\tau_2 - \tau_1)\,(\mathrm{Re}\, W_2 - \mathrm{Re}\, W_1) \end{pmatrix} > 0. \tag{16}$$

*Remark 2.* For the triangular system (1) ($c_1 = c_2 = c$), Theorem 2 remains valid when replacing inequality (13) with inequality (16), where $W_j = W_j(\lambda) = c^\top \Delta(\lambda) b_j$, $j = 1, 2$.

Compare conditions (15) and (16) of the criterion in Theorem 2 for connected switched systems and triangular switched systems with those of Theorem 2 from [3] and their modification for triangular systems (formulas (6.3)–(6.5) from [3]). Significant progress is evident.

*Remark 3.* Inequalities (13), (15), and (16) are linear in the parameter $\mathcal{T}$; therefore, without losing generality, let $\tau_3 = 1$ in these inequalities. Thus, the inequalities under consideration will contain only two additional parameters each: $\tau_1 > 0$ and $\tau_2 > 0$.

The well-known Tsypkin's criterion [8] is a quadratic stability criterion under switching between two subsystems. The criterion of Theorem 2 can be considered an analog of Tsypkin's criterion under switching between three subsystems.

## 4. NUMERICAL SOLUTION

The quadratic stability problem for system (1) is numerically solved by applying standard software tools for checking the feasibility of the system of LMIs (5) of dimension $3n$ with respect to $n(n+1)/2$ unknowns. Due to Lemma 1, it is possible to check the feasibility of the single LMI (12) of dimension $(n + 2)$ with respect to $n(n + 1)/2 + 3$ unknowns instead of the system of LMIs (5). This transition allows significantly simplifying the problem, especially for large $n$.

## 5. AN EXAMPLE

Consider a connected switched system of the form (1) from the example presented in [3]. In this example, the matrices $A_s$ in (1) are given by (4) with

$$A_1 = A = \begin{pmatrix} 0 & 0 & -0.5 \\ 0.5 & 0 & -1.5 \\ 0 & 0.5 & -1.5 \end{pmatrix}, \quad b_1 = k_1 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad b_2 = k_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \tag{17}$$
$$c_1 = c_2 = c = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

where $k_i \geqslant 0$ are the parameters determining the stability domain of the switched system. Then the matrices $A_2$ and $A_3$ take the form

$$A_2 = \begin{pmatrix} 0 & 0 & -0.5 \\ 0.5 & 0 & -1.5 \\ 0 & 0.5 & -1.5 + k_1 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0 & 0 & -0.5 \\ 0.5 & 0 & -1.5 + k_2 \\ 0 & 0.5 & -1.5 \end{pmatrix}. \tag{18}$$

In the sequel, system (1) with the matrices $A_s$ (17), (18) will be referred to as system (1;17).

Re-examining the example from [3] can be explained as follows. In the example from [3], given $k_1 = k_2 = k$, the entire quadratic stability domain on the parameter $k$ was found. This result was obtained using the necessary (separately) and sufficient (separately) conditions for the feasibility of the system of LMIs (5). As it turned out, the estimates under these conditions coincide; hence, the resulting quadratic stability domain is entire. Note that the conditions from [3] essentially rest on the triangular property of the system, i.e., $c_1 = c_2 = c$.

This section aims to repeat the result from [3] based on the criterion of Theorem 2. Although the considerations below use a variant of Theorem 2 from Remark 2, this theorem does not include the triangularity requirement.

The presentation here involves the auxiliary calculations from [3]. Obviously, the matrix $A_1$ is Schur, $|\mu_i(A_1)| < 1$, since $\mu_i(A_1) = -0.5$, $i = \overline{1,3}$. The matrix $A_2$ is Schur for $k_1 \in [0,\, 3.375)$, whereas the matrix $A_3$ is Schur for $k_2 \in [0,\, 0.25)$.

The functions $W_j(\lambda) = c^\top (A - \lambda E_n)^{-1} b_j$ from (16) have the form

$$W_1(\lambda) = -8k_1\lambda^2/(2\lambda+1)^3 \quad\text{and}\quad W_2(\lambda) = -4k_2\lambda/(2\lambda+1)^3,$$

$\det(A - \lambda E) = -(0.5+\lambda)^3$. Inequality (16) should be checked for all $\lambda \in \mathbb{C}$ such that $|\lambda| = 1$. For the set $|\lambda| = 1$, we use the parameterization $\lambda = \frac{1-i\omega}{1+i\omega}$ for all $\omega \in [-\infty, \infty]$. Let us calculate $W_1(\lambda)$ and $W_2(\lambda)$ for $\lambda = \frac{1-i\omega}{1+i\omega}$. We write the real and imaginary parts of $W_j\left(\frac{1-i\omega}{1+i\omega}\right)$, simultaneously adopting the simplified notations $\operatorname{Re} W_j\left(\frac{1-i\omega}{1+i\omega}\right) = R_j(\omega) = R_j$ and $\operatorname{Im} W_j\left(\frac{1-i\omega}{1+i\omega}\right) = I_j(\omega) = I_j$ (see [3]):

$$\begin{aligned} R_1 = R_1(\omega) &= \operatorname{Re} W_1\left(\frac{1-i\omega}{1+i\omega}\right) = \frac{-8k_1(1+\omega^2)(27+18\omega^2-\omega^4)}{(9+\omega^2)^3}, \\ I_1 = I_1(\omega) &= \operatorname{Im} W_1\left(\frac{1-i\omega}{1+i\omega}\right) = \frac{-64k_1\omega^3(1+\omega^2)}{(9+\omega^2)^3}, \\ R_2 = R_2(\omega) &= \operatorname{Re} W_2\left(\frac{1-i\omega}{1+i\omega}\right) = \frac{-4k_2(1+\omega^2)(27-36\omega^2+\omega^4)}{(9+\omega^2)^3}, \\ I_2 = I_2(\omega) &= \operatorname{Im} W_2\left(\frac{1-i\omega}{1+i\omega}\right) = \frac{-4k_2(1+\omega^2)(54\omega-10\omega^3)}{(9+\omega^2)^3}. \end{aligned} \tag{19}$$

In terms of (19), inequality (16) takes the form

$$D(\omega) = \begin{pmatrix} \tau_1(1+R_1) & \dfrac{-\tau_1+\tau_2-\tau_3+\tau_1R_2+\tau_2R_1-2\tau_1R_1}{2} + i\dfrac{\tau_1I_2-\tau_2I_1}{2} \\ \overline{(\bullet)} & \tau_3 + (\tau_2-\tau_1)(R_2-R_1) \end{pmatrix} > 0.$$

Letting $k_2 = k_1 = k$, we make the change $\omega^2 = y \geqslant 0$. It is required to find the largest domain $[0, k^*)$ for which there exists a set of parameters $\tau_i > 0$, $j = 1,2,3$, such that $D(\omega) \cong D(y) > 0$ for $k \in [0, k^*)$ and all $y \geqslant 0$. Checking the inequality $D(y) > 0$ reduces to checking the inequalities

$$\text{(A) } D_{11} = \tau_1(1+R_1) > 0, \quad \text{(B) } D_{22} = \tau_3 + (\tau_2-\tau_1)(R_2-R_1) > 0, \quad \text{(C) } \det D(y) > 0,$$

where $D_{ij} = D_{ij}(y)$, $i, j = 1, 2$, are the elements of the matrix $D(y)$. (In fact, it suffices to check (A) and (C).) Inequality (A) is equivalent to

$$\begin{aligned} P_1(y) &= (9+y)^3 D_{11}(y) = \tau_1(1+R_1) \\ &= \tau_1(9+y)^3 - 8\tau_1 k(1+y)(27+18y-y^2) \\ &= \tau_1(1+8k)y^3 + \tau_1(27-136k)y^2 + \tau_1(243-360k)y + \tau_1 27(27-8k) > 0. \end{aligned}$$

The check of inequality (A) coincides with that of inequalities (7.4) and (7.5) from [3]. As was shown in [3], for $k < 0.44$, the inequality $P_1(y) > 0$ holds for all $y \geqslant 0$.

In view of Remark 3, we assume that $\tau_3 = 1$ and, for brevity, $\tau_2 - \tau_1 \triangleq \delta$. Then checking inequality (B) reduces to checking the inequality

$$\begin{aligned} P_2(y) &= (9+y)^3 D_{22}(y) = (9+y)^3 + 4k\delta(1+y)(27+72y-3y^2) \\ &= (1-12k\delta)y^3 + (27+276k\delta)y^2 + (243+396k\delta)y + 729 + 108k\delta > 0. \end{aligned}$$

Consider inequality (C):

$$\det D = D_{11}D_{22} - D_{12}\overline{D_{12}} = D_{11}D_{22} - (\mathrm{Re} D_{12})^2 - (\mathrm{Im} D_{12})^2 > 0.$$

With the notations $P_3(y) \triangleq 2(9+y)^3 \operatorname{Re} D_{12}$ and $P_4(y) \triangleq 2(9+y)^3 \operatorname{Im} D_{12}$, we have

$$\begin{aligned} P_3(y) &= 2(9+y)^3 \operatorname{Re} D_{12}(y) = 2\big(\tau_1(R_2-R_1) + \delta R_1 + \delta - 1\big)(9+y)^3, \\ P_4(y) &= 2(9+y)^3 \operatorname{Im} D_{12}(y) = 2\big(\tau_1 I_2 - \tau_2 I_1\big)(9+y)^3. \end{aligned}$$

Using the expressions from (19) gives

$$\begin{aligned} P_3(y) &= y^3[4k(2\delta - 3\tau_1) + \delta - 1] + y^2[27(\delta-1) + 4k(69\tau_1 - 34\delta)] \\ &\quad + y[9(27(\delta-1) - 40k\delta + 44k\tau_1)] + [27(27(\delta-1) - 4k(2\delta - \tau_1))], \\ P_4(y) &= 4k\tau_1\sqrt{y}(1+y)(10y-54) + 64k\tau_2 y\sqrt{y}(1+y) \\ &= 4k\sqrt{y}(1+y)\big(\tau_1(10y-54) + 16\tau_2 y\big). \end{aligned}$$

Inequality (C) is equivalent to

$$P(y) \triangleq (9+y)^6 \det D(y) = P_1(y)P_2(y) - \frac{1}{4}P_3(y)^2 - \frac{1}{4}P_4(y)^2 > 0. \tag{20}$$

The polynomial $P(y)$ is of degree 6 in the variable $y$. Its coefficients $f_s = f_s(k)$ for $y^s$ are functions of $k$ that depend on the additional parameters $\tau_1$ and $\tau_2$. The coefficient $f_6(k)$ of this polynomial at $y^6$ is

$$f_6(k) = \tau_1(1+8k)(1-12k\delta) - (1/4)[4k(2\delta - 3\tau_1) + \delta - 1]^2.$$

The condition $f_6(k) \geqslant 0$ is necessary for fulfilling $P(y) > 0$ for all $y \geqslant 0$. The function $f_6(k)$ represents a polynomial of degree 2 in the variable $k$. Its coefficient at $k^2$ is $a_6 = -96\tau_1\delta - 4(2\delta - 3\tau_1)^2 = -4(2\tau_2 + \tau_1)^2$, i.e., $a_6 < 0$ since $\tau_j > 0$. It follows that $f_6(k)$ is a concave function. The desired domain $[0, k^*)$ can be estimated from above by the half-interval $[0,\ 0.25)$ (the Schur domain of the matrix $A_3$). We check the values $f_6(0)$ and $f_6(0.25)$ :

$$\begin{aligned} f_6(0) &= \tau_1 - (1/4)(\delta - 1)^2, \\ f_6(0.25) &= \tau_1(1+2)(1-3\delta) - (1/4)\big((2\delta - 3\tau_1) + \delta - 1\big)^2. \end{aligned}$$

The condition $f_6(0) > 0$ gives the parameter estimate $4\tau_1 > (\delta - 1)^2$. Let us transform the expression for $f_6(0.25)$ :

$$f_6(0.25) = 3\tau_1(1-3\delta) - \frac{1}{4}(3\delta - 3\tau_1 - 1)^2 = -\frac{1}{4}(3\delta + 3\tau_1 - 1)^2 = -\frac{1}{4}(3\tau_2 - 1)^2.$$

As a result, $f_6(0.25) < 0$ for all parameter values except for $\tau_2 = 1/3$. Thus, letting $\tau_2 = 1/3$ is the single possibility to obtain the largest domain $[0, k^*)$ in which $f_6(0.25) > 0$. If we take $\tau_2 = 1/3$ and define $\tau_1$ so that $f_6(0) > 0$, the concavity of $f_6(k)$ will imply $f_6(k) > 0$ for all $k \in [0,\ 0.25)$. Partly by chance, partly to obtain $\delta = 0$, we set $\tau_1 = \tau_2 = 1/3$. In this case, $f_6(0) = 1/12 > 0$.

As it turns out, for $\tau_1 = \tau_2 = 1/3$, the other coefficients $f_s(k)$, $s = 0, \ldots, 5$, of the polynomial $P(y) = \sum\limits_{s=0}^{6} f_s(k) y^s$ from (20) are concave functions in the variable $k$. In addition, the inequalities $f_s(0) > 0$ and $f_s(0.25) > 0$, $s = 0, \ldots, 5$, hold for the values of these functions at the limit points of the half-interval $[0,\ 0.25)$. The tedious verification of this fact by elementary algebra techniques is omitted here. Thus, we have $f_s(k) > 0$ for all $k \in [0,\ 0.25)$, $s = 0, \ldots, 6$. Hence, inequality (20) is valid for all $y \geqslant 0$. According to Theorem 2, the quadratic stability domain of system (1;17) is exhausted by the set $[0,\ 0.25)$. Due to its coincidence with the Schur domain of the matrices $\{A_1, A_2, A_3\}$ defining system (1;17) (on the parameter $k_1 = k_2 = k$), this domain is the entire stability domain of system (1;17) under arbitrary switching.

## 6. CONCLUSIONS

A connected system with switching between three linear discrete-time subsystems has been considered. An existence criterion for a QLF of such systems has been established, both as a frequency-domain condition and as feasibility conditions of a single LMI. As an illustrative example, the frequency-domain criterion has been applied to a third-order system, analytically yielding its entire quadratic stability domain on the parameter $k$. In the case under study, this domain coincides with the entire stability domain of system (1;17) under arbitrary switching.

## *APPENDIX*

**Proof of Lemma 1.** We define the new parameters

$$\widehat{\tau}_1 \triangleq \delta_{11} + \tau_1, \quad \widehat{\tau}_2 \triangleq \delta_{22} + \tau_2.$$

Then

$$\begin{aligned}
&p_1(L) + \frac{\tau_1}{2}c_1 = A^\top L b_1 + \frac{\delta_{11}}{2}c_1 + \frac{\tau_1}{2}c_1 = A^\top L \widehat{B}_1 + \frac{\widehat{\tau}_1}{2}\widehat{C}_1,\\
&p_2(L) - p_1(L) + \frac{\tau_2}{2}c_2 - \frac{\tau_1}{2}c_1\\
&= A^\top L b_2 - A^\top L b_1 + \frac{\delta_{22} + \tau_2}{2}c_2 - \frac{\delta_{11} + \tau_1}{2}c_1\\
&= A^\top L(b_2 - b_1) + \frac{\widehat{\tau}_2}{2}c_2 - \frac{\widehat{\tau}_1}{2}c_1 = A^\top L \widehat{B}_2 + \frac{\widehat{\tau}_2}{2}\widehat{C}_2.
\end{aligned} \tag{A.1}$$

It suffices to represent the matrix $\begin{pmatrix} -\tau_1 & \frac{\tau_1 - \tau_2 + \tau_3}{2} \\ \bullet & -\tau_3 \end{pmatrix}$ in the form $\left(\widehat{B}^\top L \widehat{B} - \Gamma\right)$.

Considering $b_1^\top L b_2 \triangleq \delta_{12}$ and $b_2^\top L b_1 \triangleq \delta_{21}$, we write the matrix $\widehat{B}^\top L \widehat{B}$ as

$$\begin{aligned}\widehat{B}^\top L \widehat{B} = \begin{pmatrix} b_1^\top \\ b_2^\top - b_1^\top \end{pmatrix} L \begin{pmatrix} b_1 & b_2 - b_1 \end{pmatrix} &= \begin{pmatrix} b_1^\top L \\ b_2^\top L - b_1^\top L \end{pmatrix} \begin{pmatrix} b_1 & b_2 - b_1 \end{pmatrix} \\ &= \begin{pmatrix} \delta_{11} & \delta_{12} - \delta_{11} \\ \delta_{21} - \delta_{11} & \delta_{22} - 2\delta_{12} + \delta_{11} \end{pmatrix}.\end{aligned} \tag{A.2}$$

Thus, it is required to find the elements of the matrix $\Gamma = \|\gamma_{ij}\|_{i,j=1}^n$ so that

$$\begin{pmatrix} -\tau_1 & \dfrac{\tau_1 - \tau_2 + \tau_3}{2} \\ \bullet & -\tau_3 \end{pmatrix} = \begin{pmatrix} \delta_{11} - \gamma_{11} & \delta_{12} - \delta_{11} - \gamma_{12} \\ \delta_{21} - \delta_{11} - \gamma_{21} & \delta_{22} - 2\delta_{12} + \delta_{11} - \gamma_{22} \end{pmatrix}. \tag{A.3}$$

Since $-\tau_1 = \delta_{11} - \widehat{\tau}_1$, the equality of the elements $\{\cdot\}_{11}$ of the matrices from (A.3) gives $\gamma_{11} = \widehat{\tau}_1$. In view of $-\tau_2 = \delta_{22} - \widehat{\tau}_2$, the equality of the elements $\{\cdot\}_{12}$ leads to

$$\frac{\tau_1 - \tau_2 + \tau_3}{2} = \frac{-\delta_{11} + \widehat{\tau}_1 + \delta_{22} - \widehat{\tau}_2 + \tau_3}{2} = \delta_{12} - \delta_{11} - \gamma_{12}.$$

Consequently,

$$\delta_{11} + \delta_{22} + \widehat{\tau}_1 - \widehat{\tau}_2 + \tau_3 = 2\delta_{12} - 2\gamma_{12}.$$

By the equality of the elements $\{\cdot\}_{22}$, we have

$$-\tau_3 = \delta_{22} - 2\delta_{12} + \delta_{11} - \gamma_{22}.$$

Summing the last two equalities yields

$$\widehat{\tau}_1 - \widehat{\tau}_2 = -2\gamma_{12} - \gamma_{22}.$$

Letting $\gamma_{22} = \widehat{\tau}_3$, we obtain

$$\gamma_{12} = (-\widehat{\tau}_1 + \widehat{\tau}_2 - \widehat{\tau}_3)/2.$$

Thus,

$$\begin{pmatrix} -\tau_1 & \dfrac{\tau_1 - \tau_2 + \tau_3}{2} \\ \bullet & -\tau_3 \end{pmatrix} = \left(\widehat{B}^\top L \widehat{B} - \Gamma\right),$$

where

$$\Gamma = \begin{pmatrix} \widehat{\tau}_1 & \dfrac{-\widehat{\tau}_1 + \widehat{\tau}_2 - \widehat{\tau}_3}{2} \\ \bullet & \widehat{\tau}_3 \end{pmatrix}.$$

The proof of Lemma 1 is complete.

## REFERENCES

1. Aleksandrov, A. and Mason, O., Diagonal Stability of a Class of Discrete-Time Positive Switched Systems with Delay, *IET Control Theory Appl.*, 2018, vol. 12, no. 6, pp. 812–818.

2. Proskurnikov, A.V. and Matveev, A.S., Tsypkin and Jury–Lee Criteria for Synchronization and Stability of Discrete-Time Multiagent Systems, *Autom. Remote Control*, 2018, vol. 79, no. 6, pp. 1057–1073.

3. Kamenetskiy, V.A., Frequency-Domain Stability Conditions for Discrete-Time Switched Systems, *Autom. Remote Control*, 2018, vol. 79, no. 8, pp. 1371–1389.

4. Malikov, A.I., State Estimation and Stabilization of Discrete-Time Systems with Uncertain Nonlinearities and Disturbances, *Autom. Remote Control*, 2019, vol. 80, no. 11, pp. 1976–1995.

5. Aleksandrov, A.Y., Semenov, A.D., and Fradkov, A.L., Discrete-Time Deployment of Agents on a Line Segment: Delays and Switches Do Not Matter, *Autom. Remote Control*, 2020, vol. 81, no. 4, pp. 637–648.

6. Pakshin, P. and Emelianova, J., Iterative Learning Control Design for Discrete-Time Stochastic Switched Systems, *Autom. Remote Control*, 2020, vol. 81, no. 11, pp. 2011–2025.

7. Kamenetskiy, V.A., Discrete-Time Pairwise Connected Switched Systems and Lur'e Systems. Tsypkin's Criterion for Systems with Two Nonlinearities, *Autom. Remote Control*, 2022, vol. 83, no. 9, pp. 1371–1392.

8. Tsypkin, Ya.Z., On the Global Stability of Nonlinear Automatic Sampled-Data Systems, *Dokl. Akad. Nauk SSSR*, 1962, vol. 145, no. 1, pp. 52–55.

9. Yakubovich, V.A., Absolute Stability of Pulsed Systems with Several Nonlinear or Linear but Nonstationary Blocks. I, II, *Autom. Remote Control*, 1967, vol. 28, no. 9, pp. 1301–1313; 1968, vol. 29, no. 2, pp. 244–263.

10. Shepelyavyi, A.I., Absolute Instability of Nonlinear Pulse-Amplitude Control Systems. Frequency Criteria, *Autom. Remote Control*, 1972, vol. 33, no. 6, pp. 929–935.

11. Gelig, A.Kh., Leonov, G.A., and Yakubovich, V.A., *Ustoichivost' nelineinykh sistem s needinstvennym sostoyaniem ravnovesiya* (Stability of Nonlinear Systems with Nonunique Equilibrium), Moscow: Nauka, 1978.

12. Kamenetskiy, V.A., Matrix Inequalities in the Stability Theory: New Results Based on the Convolution Theorem, *Autom. Remote Control*, 2023, vol. 84, no. 3, pp. 270–284.

13. Polyak, B.T., Khlebnikov, M.V., and Shcherbakov, P.S., *Upravlenie lineinymi sistemami pri vneshnikh vozmushcheniyakh: tekhnika lineinykh matrichnykh neravenstv* (Control of Linear Systems under Exogenous Disturbances: the Technique of Linear Matrix Inequalities), Moscow: LENAND, 2014.

*This paper was recommended for publication by P.V. Pakshin, a member of the Editorial Board*

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 7, pp. 665–677.

# A Fault Tolerance Method for Control Systems with Full or Partial Fault Decoupling

**A. N. Zhirabok**[*,**,a], **V. F. Filaretov**[***,b], **A. V. Zuev**[**,c], **and A. E. Shumsky**[*,d]

[*] *Far Eastern Federal University, Vladivostok, Russia*
[**] *Institute of Marine Technology Problems, Far Eastern Branch,*
*Russian Academy of Sciences, Vladivostok, Russia*
[***] *Institute of Automation and Control Processes, Far Eastern Branch,*
*Russian Academy of Sciences, Vladivostok, Russia*
*e-mail:* [a] *zhirabok@mail.ru,* [b] *filaretov@inbox.ru,* [c] *alvzuev@yandex.ru,* [d] *a.e.shumsky@yandex.com*

Received November 15, 2023
Revised April 15, 2024
Accepted April 20, 2024

**Abstract**—This paper considers technical systems described by nonlinear dynamic models. The fault tolerance property of such systems is ensured by introducing feedback with full or partial fault decoupling. The solution is based on separating a subsystem insensitive or minimally sensitive to faults and its subsequent analysis. For this purpose, a logical-dynamic approach is used, which operates only linear algebra methods. An illustrative practical example is provided.

*Keywords*: nonlinear systems, faults, isolation, feedback, singular value decomposition

**DOI:** 10.31857/S0005117924070023

## 1. INTRODUCTION

Modern technical systems (robots, control systems) are subjected to various faults in their elements. Redundancy is one way to eliminate the effect of such faults [1]; however, it requires excessive resources and is not always implementable in practice. The use of fault diagnosis methods is a more promising approach to improving the reliability, safety, and efficiency of such systems. In real time, these methods have to detect emerging faults and determine the values of the changed system parameters and errors in the readings of their sensors. After that, all identified changes with undesirable consequences are promptly parried.

Various fault diagnosis methods were thoroughly described in [2], including the basic terminology in this area. According to [2], a fault is understood as an unacceptable deviation of at least one of the characteristic properties or variables of a system from its standard (nominal) behavior. In this paper, such a deviation is represented by an unknown bounded time-varying function $d(t)$ added to certain components of the system state vector depending on the fault location.

As is known [3], adaptive systems designed to parry the consequences of faults and changes in the parameters of control objects can be divided into two large groups: systems with self-adjusting structure (self-organizing systems) and systems with self-tuning parameters (self-tuning systems). In the former case, certain structural changes are made to the system being diagnosed, i.e., it is reconfigured to remove failed elements and use redundant ones. In the latter case, depending on the changes in the parameters of the control object, emerging faults, or external influences, only certain parameters of the used controller are tuned according to some algorithm embedded in the self-tuning device. The system with faults and changed parameters should continue functioning, preserving its most important characteristics within the admissible limits.

Each of the above approaches has peculiarities, which somewhat restrict the scope of their practical application. In particular, the possibility of involving redundant elements is limited by the maximal design-achievable and operational (mass and size, energy, etc.) characteristics of specific robots.

Examples of implementing such an approach were described in [4, 5]. The cited authors solved the problem of fault-tolerant control of underwater robots in case of failure of one thruster (the first work mentioned) and in case of faulty electric actuators installed in the manipulator joints (the second one). In both cases, it was proposed to disconnect the faulty actuator and then distribute its control actions between the others with additional connection of the redundant ones. The disadvantage of such systems is the need for extra actuators in robots, which complicates the design and appreciably increases the cost of robots. In addition, the feasibility of using redundancy must be justified by additional calculations of reliability indicators. As a rule, redundancy elements have the same reliability as the replaced ones; as a result, the possibility of increasing the reliability of robots through redundancy is significantly limited. Fault adaptation methods based on self-tuning allow avoiding additional hardware costs, but their use admits degradation of some (usually minor) performance indicators of robots, possibly affecting the tactical and technical characteristics of robots and, in some cases, even requiring correction of the mission.

Fault-tolerant self-tuning systems with a reference model are known; their design principles were presented in [6, 7]. The main peculiarity of this class of systems is the availability of an explicit technical device (model) with given dynamic properties. In this case, the dynamics of the entire system are reduced to the desired dynamics of the model. Such adaptation systems to faults and variable parameters have found application in both ground and underwater robotics [8–10], providing high-quality control of robots with rather simple means without identifying the parameter deviations caused by faults or other external factors during their operation. As the main drawback of such systems, we note the presence of high-frequency oscillations in the self-tuning loop, which in some cases may significantly reduce the quality of adaptation to emerging faults and variable parameters. In addition, during the operation of such systems, the deviations of parameters from their nominal values are not determined; therefore, in the case of critical faults (e.g., short-circuit in some winding turns of the anchor chain of electric motors, the appearance of significant external torques on motor shafts), the robots will not be promptly stopped, and their further breakdown will not be prevented. The systems under consideration also neglect errors in the readings of robot sensors.

Optimal and robust principles of adaptive systems design are often used in engineering to compensate for the consequences of emerging faults and parameter deviations from nominal values [11–13]. The advantage of such systems is a sufficiently high level of robustness to the uncertain parameters of robots, but they are built based on a linearized model, which restricts their application to fault-tolerant control of the spatial motion of complex dynamic objects.

Currently, variable-structure systems operating in sliding mode are a common type of robust control systems. Examples of their use for fault-tolerant robot control were described in [14–17]. Control systems with adaptation to emerging faults and parameter deviations based on variable-structure systems have several considerable benefits compared to other types of fault-tolerant systems. Despite this fact, they also suffer from the disadvantage that, in order to ensure the performance of a variable-structure system within the entire range of changes in robot parameters, such systems are designed in the worst case (when these parameters correspond to the lowest system performance). As a result, even in the absence of faults, additional control signals are generated, which will increase their amplitude and energy consumption and, consequently, reduce the autonomous operation time. That is, fault-tolerant control systems of this class have a deliberately underestimated performance.

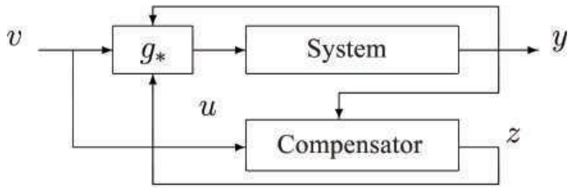


**Fig. 1.** The implementation scheme of the proposed solution.

The approaches and methods discussed above are illustrated mainly by examples of robots, but their peculiarities also apply to many modern technical systems.

A certain alternative to the considered methods is approaches based on full or partial fault decoupling: a fault is detected, but the values of the changed system parameters are not determined, and the control action on the system is corrected by using a specially built compensator and a new control. As a result, the system will execute its main operations with the previous or admissibly reduced quality. By assumption, the execution of these operations depends not on all components of the system state vector but only on some part of them, defined by a known function, and these components have to be fully or partially decoupled from possible system faults.

Figure 1 shows the implementation scheme of the proposed solution, where $u(t)$ and $y(t)$ are the control vector and output of the system, respectively, $z(t)$ is the state vector of the compensator, $v(t)$ is the new control, and $g_*$ is a function defined below. The control $u(t)$ was constructed to execute certain operations by the system, and the new control $v(t)$ must be constructed to execute the same operations by the system with the compensator, with the same or admissibly reduced quality.

This approach has certain limitations: figuratively speaking, it can be implemented if there exists a control signal between the fault location and the system variables that need to be decoupled from this fault; the control signal is used for fault decoupling.

For systems described by nonlinear difference equations, such an approach was implemented in [18, 19] based on full decoupling using a rather complex mathematical apparatus of function algebra. In distinction, this paper considers systems given by nonlinear differential equations subject to faults. For such a system, it is required to find a description of the compensator and a function $g_*$ to decouple from faults, fully or partially, given components of the system state vector.

The problem of determining the new control $v(t)$ is not considered below since this control depends on the tasks solved by the system and can be determined when specifying these tasks. After the compensator is built, the new control can be determined by known methods [20]; the compensator depends on given components of the system state vector and the fault location and is independent of the tasks solved by the system.

Note that for affine systems, such a problem was solved in [21] based on full decoupling by rather complicated methods of differential geometry. The novelty of this paper is that the systems under consideration may contain unsmooth nonlinearities; the problem is solved using the logical-dynamic approach [22], which allows analyzing nonlinear systems by linear algebra methods under definite restrictions on the class of solutions. Moreover, partial fault decoupling is studied in addition to full decoupling.

The remainder of this paper is organized as follows. Section 2 presents the main models: descriptions of the given nonlinear system and its submodel used to build the compensator. In Section 3, a fault-insensitive submodel is constructed; in Section 4, a submodel minimally sensitive to faults. Section 5 is devoted to the compensator design. An illustrative example is provided in Section 6, and Section 7 concludes the paper.

## 2. MAIN MODELS

Consider systems described by the nonlinear model

$$\begin{aligned}\dot{x}(t) &= Fx(t) + Gu(t) + C\Psi(x(t), u(t)) + Dd(t),\\ y(t) &= Hx(t),\end{aligned} \tag{2.1}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and $y \in \mathbb{R}^l$ are the state vector, control input, and output, respectively; $F$ and $G$ are known constant matrices that describe linear dynamics; $H$, $C$, and $D$ are known constant matrices; $d(t)$ is a scalar function that describes faults (if there are no faults, $d(t) = 0$; when faults occur, $d(t)$ becomes an unknown bounded time-varying function); $\Psi(x, u)$ is the nonlinear part represented as

$$\Psi(x, u) = \begin{pmatrix} \varphi_1(A_1x, u) \\ \dots \\ \varphi_q(A_qx, u) \end{pmatrix},$$

where $A_1, \dots, A_q$ are known constant row matrices, and $\varphi_1, \dots, \varphi_q$ are arbitrary nonlinear functions.

*Remark 1.* If the system may have several faults, then (generally speaking) it is necessary to build a bank of several compensators for fault decoupling. The method under consideration cannot be applied to decouple from sensor faults; if the value of such a fault is unknown, it is necessary to exclude the readings of the corresponding sensor from the control system or use a virtual sensor instead [23].

Note that the nonlinear system (2.1) can be obtained from the general nonlinear system

$$\begin{aligned}\dot{x}(t) &= f(x(t), u(t), d(t)),\\ y(t) &= h(x(t))\end{aligned} \tag{2.2}$$

by several transformations [22].

By assumption, faults in the system change the value of some system parameter. As a result, $d(t)$ represents the product of this change by some component of the vector $x(t)$ or $u(t)$ and is an unknown bounded time-varying function; the matrix $D$ indicates the fault location. The fault can be detected and isolated by known fault diagnosis methods (e.g., see [2]), but the function $d(t)$ still remains unknown.

By another assumption, the function of the components of the system state vector $x(t)$ for full or partial fault decoupling is given by a known matrix $H_0$ defining the variable $y_0(t) = H_0x(t)$. Such decoupling is ensured by introducing dynamic feedback into the system, being implemented through a compensator generally described by the nonlinear equations

$$\begin{aligned}\dot{z}(t) &= \varphi(z(t), v(t), y(t)),\\ u(t) &= g_*(z(t), v(t), y(t)),\end{aligned} \tag{2.3}$$

where $z(t) \in \mathbb{R}^k$ denotes the state vector of a compensator of dimension $k < n$, $v(t)$ is the new control, and the functions $\varphi$ and $g_*$ have to be determined. Note that the variable $y_0(t)$ must be expressed through the state vector $z(t)$.

For the discrete-time analog of system (2.2), the problem of insensitivity to (or full decoupling from) disturbances and faults via feedback was solved in a general form in [18, 19] based on a rather complex mathematical apparatus of function algebra. In this paper, we solve the problem of full or partial decoupling with insensitivity (or minimal sensitivity) to faults for system (2.1) within the logical-dynamic approach [22], which operates only linear algebra methods.

The problem solution involves a submodel of system (2.1) insensitive or minimally sensitive to faults and a compensator built on its basis. Note that interval observers in [24] were designed by building a minimal-dimension submodel. In contrast, the compensator supplying the feedback is built based on a submodel of a maximal dimension $k < n$, which provides the best conditions for satisfying the equality $y_0(t) = H_0 x(t)$. This submodel is described by the equation

$$\dot{x}_*(t) = F_* x_*(t) + G_* u(t) + J_* y(t) + C_* \Psi_*(x_*(t), y(t), u(t)), \tag{2.4}$$

where $x_*(t) \in \mathbb{R}^k$ stands for the state vector of the submodel of dimension $k < n$; $F_*$, $G_*$, $J_*$, and $C_*$ are the matrices to be determined;

$$C_* \Psi_*(x_*, y, u) = \begin{pmatrix} \varphi_{i_1}(A_{*1,i_1} x_* + A_{*2,i_1} y, u) \\ \dots \\ \varphi_{i_k}(A_{*1,i_k} x_* + A_{*2,i_k} y, u) \end{pmatrix}, \tag{2.5}$$

where $A_{*1,i_1}$, $A_{*2,i_1}$, …, $A_{*1,i_k}$, and $A_{*2,i_k}$ are the matrices to be determined; $C_* \Psi_*$ denotes the function $C_* \Psi$ in which the vector $x$ is replaced by $x_*$ and $y$ through the relation $A_i x = A_{*1,i} x_* + A_{*2,i} y$, where $i = i_1, \dots, i_k$ are the numbers of the nonzero columns of the matrix $C_*$.

## 3. BUILDING THE FAULT-INSENSITIVE SUBMODEL

We clarify that submodel (2.4) for building the compensator is a virtual object. In fact, it represents part of system (2.1) whose dynamics are determined by the state vector $x_*$ related to the vector $x$ by $x_*(t) = \Phi x(t)$, where $\Phi$ is some constant matrix. Generally speaking, these vectors can be related by a nonlinear function, and the assumption of its linearity restricts the class of solutions; it is characteristic of the logical-dynamic approach used here.

According to [22, 24], this matrix satisfies the equations

$$\begin{gathered} \Phi F = F_* \Phi + J_* H, \quad \Phi G = G_*, \quad \Phi C = C_*, \quad \Phi D = D_* \\ A_i = (A_{*1,i} \;\; A_{*2,i}) \begin{pmatrix} \Phi \\ H \end{pmatrix}, \quad i = i_1, \dots, i_k. \end{gathered} \tag{3.1}$$

The last equality in (3.1) is valid if

$$\operatorname{rank} \begin{pmatrix} \Phi \\ H \end{pmatrix} = \operatorname{rank} \begin{pmatrix} \Phi \\ H \\ A' \end{pmatrix}, \tag{3.2}$$

where the matrix $A'$ consists of the rows $A_{i_1}$, …, $A_{i_k}$.

To solve the problem, we introduce the additional condition $y_0(t) = H_* x_*(t)$ for some matrix $H_*$, i.e., the variable $y_0(t) = H_0 x(t)$ must be expressed through the compensator state vector. In view of $x_*(t) = \Phi x(t)$, it follows that

$$\operatorname{rank} \begin{pmatrix} \Phi \end{pmatrix} = \operatorname{rank} \begin{pmatrix} \Phi \\ H_0 \end{pmatrix}. \tag{3.3}$$

If this condition fails, the problem is unsolvable. Under this condition, the matrix $H_*$ is found from the equation $H_* \Phi = H_0$.

To ensure the fault-insensitivity condition $\Phi D = D_* = 0$, we introduce a matrix $D_0$ of maximal rank such that $D_0 D = 0$. Then $\Phi D = 0$ implies $\Phi = N D_0$ for some matrix $N$. Let us replace the matrix $\Phi$ in $\Phi F = F_*\Phi + J_*H$ with $ND_0$, i.e., $ND_0F = F_*ND_0 + J_*H$. After the separation of the unknown and known matrices, the resulting expression can be written as

$$\begin{pmatrix} N & -F_*N & -J_* \end{pmatrix} \begin{pmatrix} D_0F \\ D_0 \\ H \end{pmatrix} = 0. \tag{3.4}$$

Solving equation (3.4) yields the matrices $F_*$, $J_*$, and $N$, which are, in turn, allow finding the matrix $\Phi$. Let the compound matrix $\begin{pmatrix} X & Y & Z \end{pmatrix}$ contain all linearly independent solutions of equation (3.4), i.e.,

$$\begin{pmatrix} X & Y & Z \end{pmatrix} \begin{pmatrix} D_0F \\ D_0 \\ H \end{pmatrix} = 0. \tag{3.5}$$

Comparing equations (3.4) and (3.5), we obtain the equality $Y = -F_*X$. Therefore, the matrices $Y$ and $X$ cannot be arbitrary: the rows of $Y$ must be linearly expressed through the rows of $X$. To consider this fact, the rows of $Y$ that are linearly independent of the rows of $X$ must be removed. This procedure is implemented using Algorithm 1, where $Y_j$ denotes the $j$th row of the matrix $Y$, $j = 1, \ldots, p$, and $p$ is the number of rows in the matrix $Y$.

**Algorithm 1.**

```
(1) Set $j = 1$.
(2) If $\operatorname{rank}(X) = \operatorname{rank}\begin{pmatrix} X \\ Y_j \end{pmatrix}$, pass to Step 4; otherwise, to Step 3.
(3) Remove the $j$th row from the matrix $\begin{pmatrix} X & Y & Z \end{pmatrix}$, set $p := p - 1$, and return to Step 1.
(4) If $j < p$, set $j := j + 1$ and return to Step 2; otherwise, complete the procedure.
```

Let $\begin{pmatrix} X_0 & Y_0 & Z_0 \end{pmatrix}$ denote the matrix outputted by the algorithm. For this matrix, the rows of the matrix $Y_0$ are linearly expressed through the rows of the matrix $X_0$. Letting $\Phi := X_0D_0$ and $C_* := \Phi C$, we construct the matrix $A'$; if the matrix $\Phi$ satisfies condition (3.2), a nonlinear fault-insensitive compensator can be built. Otherwise, full fault decoupling is unreachable, and robust methods should be used. If condition (3.3) fails for this matrix, the problem is unsolvable.

Letting $J_* = -Z_0$ and $G_* = \Phi G$, we find the matrix $F_*$ from the algebraic equation $Y_0 = -F_*X_0$. It surely has a solution because, according to Algorithm 1, $Y_0$ is linearly expressed through the rows of the matrix $X_0$. Thus, the matrices describing the linear part of the submodel have been obtained. To construct the nonlinear part, we take $C_* = \Phi C$ and determine the matrices $A_{*1,i}$ and $A_{*2,i}$, $i = i_1, \ldots, i_k$, from equation (3.1). This gives the nonlinear part (2.5) and, consequently, the entire submodel (2.4).

## 4. BUILDING THE ROBUST SUBMODEL

If $\begin{pmatrix} X_0 & Y_0 & Z_0 \end{pmatrix} = 0$ or the matrix $\Phi$ does not satisfy condition (3.2), the fault-insensitive compensator cannot be built. In this case, it is necessary to address robust methods to minimize the fault contribution to model (2.4). For this purpose, we write the relation $\Phi F = F_*\Phi + J_*H$ in a form similar to (3.3), removing the fault-insensitivity constraint $\Phi D = D_* = 0$ and separating the unknown matrices from the known ones:

$$\begin{pmatrix} \Phi & -F_*\Phi & -J_* \end{pmatrix} \begin{pmatrix} F \\ E \\ H \end{pmatrix} = 0, \tag{4.1}$$

where $E$ is an identity matrix of appropriate dimensions. Now equation (4.1) can have solutions admitting the model's sensitivity to faults.

As above, we consider the compound matrix $( X \ \ Y \ \ Z )$ containing all linearly independent solutions of equation (4.1), i.e.,

$$( X \ \ Y \ \ Z ) \begin{pmatrix} F \\ E \\ H \end{pmatrix} = 0.$$

Applying Algorithm 1 to the matrix $( X \ \ Y \ \ Z )$, we obtain the matrix $( X_* \ \ Y_* \ \ Z_* )$ in which $Y_* = -MX_*$ with some matrix $M$. If this equation has several solutions, they will correspond to several matrices $\Phi : \Phi^{(1)}, \ldots, \Phi^{(s)}$. By determining, for each of them, the norm $\|\Phi^{(i)}D\|$ corresponding to the fault contribution to the compensator, we can choose the variant with the smallest norm value corresponding to the minimal fault contribution to the submodel.

A better result can be obtained by setting the matrix $\Phi = \sum_{i=1}^{s} v_i \Phi^{(i)}$ and assigning the weights $v_1, \ldots, v_s$ based on minimization of the norm $\|\Phi D\|$. However, this approach is possible only if the matrix $F_*$ in the expression $\Phi F = F_* \Phi + J_* H$ remains the same for different $\Phi$. We implement this approach by choosing $F_*$ in the canonical form

$$F_* = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix}, \tag{4.2}$$

which will additionally simplify the design procedure. Due to the canonical form (4.2), equations (3.1) become [22]

$$\Phi_i F = \Phi_{i+1} + J_{*i} H, \quad i = 1, \ldots, k-1, \quad \Phi_k F = J_{*k} H, \tag{4.3}$$

where $\Phi_i$ and $J_{*i}$ are the $i$th rows of the matrices $\Phi$ and $J_*$, respectively, $i = 1, \ldots, k$. According to [22], these equations can be convolved into one:

$$( \Phi_1 \ \ -J_{*1} \ \ -J_{*2} \ \ \ldots \ \ -J_{*k} ) V^{(k)} = 0, \tag{4.4}$$

where

$$V^{(k)} = \begin{pmatrix} HF^k \\ HF^{k-1} \\ \ldots \\ H \end{pmatrix}.$$

Also, see [22], the minimization problem of the fault contribution to the submodel reduces to minimizing the norm $\|\Phi D\| = \|( \Phi_1 \ \ -J_{*1} \ \ -J_{*2} \ \ \ldots \ \ -J_{*k} ) D^{(k)}\|$ subject to condition (4.4), where

$$D^{(k)} = \begin{pmatrix} D & FD & F^2D & \ldots & F^{k-1}D \\ 0 & HD & HFD & \ldots & HF^{k-2}D \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix}.$$

When solving this problem, we find a maximal dimension $k < n$ for which equation (4.4) has several (more than one) linearly independent solutions of the form $( \Phi_1 \ \ -J_{*1} \ \ -J_{*2} \ \ \ldots \ \ -J_{*k} )$.

All these solutions, $s$ totally, are combined into a matrix $W$ so that each row represents some solution of equation (4.4):

$$W = \begin{pmatrix} \Phi_1^{(1)} & -J_{*1}^{(1)} & -J_{*2}^{(1)} & \dots & -J_{*k}^{(1)} \\ \dots & \dots & \dots & \dots & \dots \\ \Phi_1^{(s)} & -J_{*1}^{(s)} & -J_{*2}^{(s)} & \dots & -J_{*k}^{(s)} \end{pmatrix}.$$

Due to the considerations above, another solution is an arbitrary linear combination of the rows of this matrix with the vector of weights $v = (v_1, \dots, v_s)$. The problem is to determine such a vector $v$ that minimizes the norm $\|vWD^{(k)}\|$.

To solve this problem, we find the singular value decomposition of the matrix product $WD^{(k)}$ :

$$WD^{(k)} = U_D \Sigma_D V_D,$$

where $U_D$ and $V_D$ are orthogonal matrices; depending on the numbers of rows and columns in the matrix $WD^{(k)}$, the matrix $\Sigma_D$ has the form

$$\Sigma_D = (\operatorname{diag}(\sigma_1, \dots, \sigma_w) \ 0)$$

or

$$\Sigma_D = \begin{pmatrix} \operatorname{diag}(\sigma_1, \dots, \sigma_w) \\ 0 \end{pmatrix},$$

with $w = \min(s, k)$ and $0 \leqslant \sigma_1 \leqslant \dots \leqslant \sigma_w$ being the singular values of the matrix $WD^{(k)}$ [22, 25]. The first transposed column of the matrix $U_D$ is chosen as the vector of weights $v = (v_1, \dots, v_s)$. By the structure of singular value decomposition and the properties of orthogonal matrices, the norm of the matrix $vWD^{(k)}$ equals the minimal singular value $\sigma_1$ [22], and $(\ \Phi_1 \ \ -J_{*1} \ \ -J_{*2} \ \ \dots \ \ -J_{*k}\ ) = vW$. Then the rows of the matrix $\Phi$ are determined from (4.3) and the matrix $A'$ is constructed. If this matrix satisfies conditions (3.2) and (3.3), we take $G_* = \Phi G$ and $C_* = \Phi C$ and find the matrices $A_{*1,i}$ and $A_{*2,i}$, $i = i_1, \dots, i_k$, from equation (3.1); this completes the robust model design. Note that this solution will be optimal for the chosen dimension $k$; changing the dimension may yield a better solution of the problem in terms of minimizing the norm $\|(\Phi_1 - J_{*1} - J_{*2} \ \dots - J_{*k})D^{(k)}\|$. If condition (3.2) or (3.3) fails, it is necessary to choose the second or subsequent transposed columns of the matrix $U_D$.

## 5. BUILDING THE COMPENSATOR

To avoid confusion, we denote the compensator state vector by $z(t) := x_*(t)$, leaving unchanged the notations for the other elements, particularly the matrix $H_*$ and the function $f_*$.

From this point onwards, condition (3.3) is assumed valid, i.e., $y_0 = H_* z$. We denote by $X_y$ the set of components of the vector $z$ participating in the formation of $y_0$. For building the compensator, model (2.4) will be written in a compact form:

$$\dot{z}(t) = f_*(z(t), u(t), y(t)). \tag{5.1}$$

Even if this model does not explicitly contain the unknown function $d(t)$ (when full decoupling is reached), its state vector is affected by faults due to the presence of the vector $y(t)$ in (5.1). To build the compensator, this effect must be eliminated by adjusting the control vector $u(t)$ via feedback with a new control vector $v(t)$. The algorithm below performs the necessary analysis and generates the feedback if possible. Let $f_{*j}$ denote the $j$th component of the function $f_*$.

**Algorithm 2.**

(1) Divide the components of the vector $y$ into two disjoint sets, $Y_g$ (*good*) and $Y_b$ (*bad*), according to the rules: the variable $y_i$ is included in $Y_g$ if it does not appear in the function $f_*$ or can be expressed through the components of the vector $z$; otherwise, $y_i$ is included in $Y_b$. If $Y_b = \emptyset$, full or partial fault decoupling is reached without the compensator since $y_i$ in the function $f_*$ can be replaced by a function of the vector $z$.

(2) If $Y_b \neq \emptyset$, for each $y_i \in Y_b$ find a variable $z_j$ such that $f_{*j}$ depends on $y_i$ and is independent of $u$. Let $X_b$ denote the set of all such $z_j$; it consists of all components of the state vector that are affected by the fault because $f_{*j}$ includes the variable $y_i$ not compensated by the control. If $X_b = \emptyset$, pass to Step 4.

(3) For each $z_j \in X_b$ find the functions $f_{*i}$ that depend on $z_j$. If all $f_{*i}$ depend on $u$, add $z_j$ to $Y_b$ and remove it from $X_b$. If for some $i$ this condition fails, then the variable $z_i$ cannot be decoupled from faults; if $z_i \in X_y$, i.e., this variable participates in the formation of the variable $y_0$, then the problem has no solution. If $z_i \notin X_y$, add $z_i$ to $X_b$ and continue executing Step 3 until $X_b = \emptyset$ or $X_b$ stops changing. The final set $Y_b$ contains the variables that will participate in the feedback to compensate for the effect of faults.

(4) Find in the function $f_*(z, u, y)$ all terms of the form $\gamma_i(z, u, y)$, $i = 1, \ldots, r$, that depend on $u$ and elements from the set $Y_b$; by assumption, $r \leqslant m$. Form a system of equations for the new control vector $v = (v_1 \ \ldots \ v_m)^T$ :

$$\begin{aligned} v_1 &= \gamma_1(z, u, y), \\ &\cdots \\ v_r &= \gamma_r(z, u, y). \end{aligned}$$

Supposing the feasibility of this system with respect to the variables $u_1, \ldots, u_r$, find its solution:

$$\begin{aligned} u_1 &= \gamma_1(z, u, y, v), \\ &\cdots \\ u_r &= \gamma_r(z, u, y, v); \\ u_{r+1} &= v_{r+1}, \ \ldots, \ u_m = v_m. \end{aligned} \tag{5.2}$$

Replace the vector $u$ in (5.1) with the vector $v$ according to the rules (5.2), which gives the dynamic part of the compensator (2.3); its static part coincides with (5.2).

## 6. EXAMPLE

Consider the nonlinear system

$$\begin{aligned} \dot{x}_1 &= u_1/\vartheta_1 - a_1\sqrt{x_1 - x_2} - d, \\ \dot{x}_2 &= u_2/\vartheta_2 + a_1\sqrt{x_1 - x_2} - a_2\sqrt{x_2 - x_3}, \\ \dot{x}_3 &= a_2\sqrt{x_2 - x_3} - a_3\sqrt{x_3 - \vartheta_7}, \\ y &= x_1, \end{aligned} \tag{6.1}$$

where $a_1 = \vartheta_4\sqrt{2\vartheta_8}/\vartheta_1$, $a_2 = \vartheta_5\sqrt{2\vartheta_8}/\vartheta_2$, and $a_3 = \vartheta_6\sqrt{2\vartheta_8}/\vartheta_3$. These equations describe the known three-tank system (Fig. 2), where $x_1$, $x_2$, and $x_3$ are the liquid levels in the tanks [26]. The system consists of three tanks with cross sections $\vartheta_1$, $\vartheta_2$, and $\vartheta_3$, respectively; the tanks are interconnected by pipes with cross sections $\vartheta_4$ and $\vartheta_5$. The liquid flows in the first and second tanks, flowing out of the third one through a pipe of a cross section $\vartheta_6$ located at a height $\vartheta_7$; the parameter $\vartheta_8$ is the gravitational constant. The controls $u_1$ and $u_2$ correspond to the externally supplied fluid. A nonzero value $d(t) > 0$ corresponds to leakage in the first tank; the variable

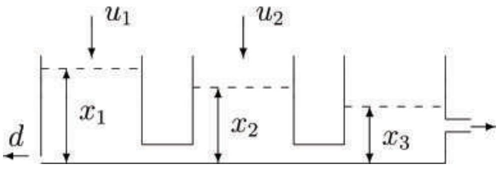


**Fig. 2.** A three-tank system.

$y_0(t) = (\ 0\ \ 0\ \ 1\ )x(t) = x_3(t)$ must be insensitive to it. The amount of leakage is assumed to be unknown, so it cannot be compensated for by increasing $u_1$ and the proposed method should be used instead.

For the sake of simplicity, let $a_1 = a_2 = a_3 = 1$ and $\vartheta_7 = 0$. The initial conditions and control are supposed to be such that $x_1(t) \geqslant x_2(t) \geqslant x_3(t) \geqslant 0$ for all $t \geqslant 0$.

Clearly, $F = 0$ for (6.1), and the considered approach cannot be applied directly. To overcome this difficulty, we transform (6.1) by introducing the formal terms $-(x_1 - x_2) + (x_1 - x_2)$, $((x_1 - x_2) - (x_2 - x_3)) - ((x_1 - x_2) - (x_2 - x_3))$, and $(x_2 - x_3 - x_3) - (x_2 - x_3 - x_3)$ into the first, second, and third equations, respectively. The term $-(x_1 - x_2)$ is added to the linear part; the term $(x_1 - x_2)$, to the nonlinear part. The remaining terms are handled similarly. As a result, the system is described by the following matrices and nonlinearities:

$$F = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix},\ G = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix},\ H = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix},\ H_0 = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix},$$

$$D = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix},\ C = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix},\ \Psi(x) = \begin{pmatrix} -\sqrt{A_1 x} + A_1 x \\ -\sqrt{A_2 x} + A_2 x \\ -\sqrt{A_3 x} + A_3 x \end{pmatrix},$$

$$A_1 = (1\ \ -1\ 0),\ \ A_2 = (0\ 1\ \ -1),\ \ A_3 = (0\ 0\ 1).$$

Since $D = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^T$, we have $D_0 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, and equation (3.5) takes the form

$$(\ X\ \ Y\ \ Z\ ) \begin{pmatrix} 1 & -2 & 1 \\ 0 & 1 & -2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = 0.$$

The solution is

$$(\ X\ \ Y\ \ Z\ ) = \begin{pmatrix} 1 & 0 & 2 & -1 & -1 \\ 0 & 1 & -1 & 2 & 0 \end{pmatrix}.$$

As is easily verified, the condition of Step 2 of Algorithm 1 holds for both rows of the matrix $Y$. Therefore,

$$(\ X_0\ \ Y_0\ \ Z_0\ ) = (\ X\ \ Y\ \ Z\ ),$$

and consequently,

$$J_* = \begin{pmatrix} 1 \\ 0 \end{pmatrix},\quad F_* = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix},\quad \Phi = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},\quad G_* = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

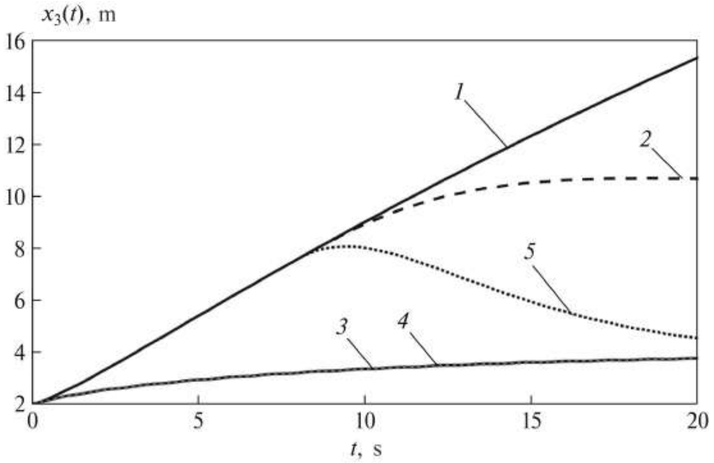


**Fig. 3.** The behavior of the variable $x_3(t) = y_0(t)$.

In view of $H_0 = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$, condition (3.3) is obviously valid; the matrix $H_*$ is found from the equation $H_0 = H_*\Phi$ and has the form $H_* = (0\ \ 1)$.

As a result, the linear part of submodel (2.4) is described by the equations

$$\begin{aligned} \dot{x}_{*1} &= u_2 - 2x_{*1} + x_{*2} + y, \\ \dot{x}_{*2} &= x_{*1} - 2x_{*2}, \end{aligned}$$

where $x_{*1} = \Phi_1 x = x_2$ and $x_{*2} = \Phi_2 x = x_3$. In addition, $y_0 = H_* x_* = x_{*2}$, i.e., $X_y = \{x_{*2}\}$.

All columns in the matrix $C_* = \Phi C = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$ are nonzero, and the matrix $A'$ hence contains three rows $A_1, A_2$, and $A_3$; condition (3.2) holds for it. Solving equation (3.1) yields

$$A_{*1,1} = (-1\ \ 0),\ A_{*2,1} = 1,\ A_{*1,2} = (1\ \ -1),\ A_{*2,2} = 0,\ A_{*1,3} = (0\ \ 1),\ A_{*2,1} = 0.$$

Therefore, the nonlinear part (2.5) takes the form

$$C_*\Psi_*(x_*, y, u) = \begin{pmatrix} \sqrt{y - x_{*1}} - (y - x_{*1}) - \sqrt{x_{*1} - x_{*2}} + (x_{*1} - x_{*2}), \\ \sqrt{x_{*1} - x_{*2}} - (x_{*1} - x_{*2}) - \sqrt{x_{*2}} + x_{*2} \end{pmatrix}.$$

Finally, adding it to the linear part gives the nonlinear submodel

$$\begin{aligned} \dot{x}_{*1} &= u_2 + \sqrt{y - x_{*1}} - \sqrt{x_{*1} - x_{*2}}, \\ \dot{x}_{*2} &= \sqrt{x_{*1} - x_{*2}} - \sqrt{x_{*2}}. \end{aligned} \tag{6.2}$$

Since $y = x_1$ is not expressed through the vector $z := x_*$, Step 1 of Algorithm 2 yields $Y_g = \emptyset$ and $Y_b = \{y\}$. Step 2 of this algorithm leads to $X_b = \emptyset$; Step 4 yields $r = 1$ and the single equation $v_2 = u_2 + \sqrt{y - z_1}$, which is obviously solvable for $u_2$ :

$$u_2 = v_2 - \sqrt{y - z_1}.$$

Setting $v_1 = u_1$ and substituting the above formula for $u_2$ into (6.2), we finally arrive at the compensator description

$$\begin{aligned} \dot{z}_1 &= v_2 - \sqrt{z_1 - z_2}, \\ \dot{z}_2 &= \sqrt{z_1 - z_2} - \sqrt{z_2}, \\ u_1 &= v_1, \\ u_2 &= v_2 - \sqrt{y - z_1}. \end{aligned} \tag{6.3}$$

For numerical simulation, we select $u_1(t) = 5$ and $u_2(t) = 2\sin(5t)$. Figure 3 shows the behavior of the variable $x_3(t) = y_0(t)$ of system (6.1) with the initial state $x(0) = 0$ for five different cases. Curve *1* corresponds to the case without the fault and decoupling; curve *2*, to the case where the fault $d = 4$ occurs at the time instant $t = 8$, but fault decoupling is not introduced (the variable changes its dynamics for $t > 8$). Curves *3* and *4* correspond to the introduction of decoupling with $v_2(t) = 2 + \sin(5t)$ at the time instant $t = 0$ in the system without the fault and with the fault, respectively; since curves *3* and *4* coincide, the fault is not manifested (has no effect on $x_3(t)$). Curve *5* corresponds to the system with the fault and decoupling with $v_2(t) = 2 + \sin(5t)$ introduced at the time instant $t = 8$; until this instant the behavior of the variable $y_0(t)$ coincides with curve *1*.

Clearly, curves *3* and *4*, where the decoupling with $v_2(t) = 2 + \sin(5t)$ is introduced at the time instant $t = 0$, do not coincide with curve *1* (the behavior of the variable without the fault). To achieve this coincidence, it is necessary to solve the control problem for the variable $v_2(t)$ in system (6.1) with the compensator (6.3). This is an independent problem solved by known methods. A similar picture is observed in case 5: when the fault occurs and the compensator is introduced, the variable $y_0(t)$, $t > 8$, changes its behavior, and the coincidence with its dynamics without the fault can be achieved by solving the control problem for the variable $v_2(t)$.

## 7. CONCLUSIONS

This paper has considered technical systems described by nonlinear dynamic models. The fault tolerance property of such systems has been ensured by introducing feedback with full or partial fault decoupling. The solution is based on the logical-dynamic approach, which operates only linear algebra methods. An illustrative practical example has been provided.

## FUNDING

This work was supported by the Ministry of Science and Higher Education of the Russian Science Foundation, project no. FZNS-2023-0011.

## REFERENCES

1. Polovko, A.M. and Gurov, S.V., *Osnovy teorii nadezhnosti* (Foundations of Reliability Theory), St. Petersburg: BKhV-Peterburg, 2006.

2. Mironovskii, L.A., *Funktsional'noe diagnostirovanie dinamicheskikh sistem* (Functional Diagnosis of Dynamic Systems), Moscow–St. Petersburg: Moscow State University, 1998.

3. Blanke, M., Kinnaert, M., Lunze, J., and Staroswiecki, M., *Diagnosis and Fault-Tolerant Control*, 3rd ed., Berlin: Springer, 2016.

4. Sarkar, N., Fault-Accommodating Thruster Force Allocation of an AUV Considering Thruster Redundancy and Saturation, *IEEE Trans. Robot. Autom.*, 2002, pp. 223–233.

5. Li, Z., Li, C., Li, S., and Cao, X., A Fault-Tolerant Method for Motion Planning of Industrial Redundant Manipulator, *IEEE Trans. Indust. Inform.*, 2020, vol. 16, pp. 7469–7478.

6. Tao, G., *Adaptive Control Design and Analysis*, Virginia: John Wiley & Sons, 2001.

7. Fradkov, A.L., Miroshnik, I.V., and Nikiforov, V.O., *Nonlinear and Adaptive Control of Complex Systems*, Springer, 1999.

8. Fan, Q.-Y., Xu, S., Deng, C., and Wang, C.-C., Event Triggered Fault Tolerant Control for Nonlinear Systems Based on Adaptive Fault Estimation, *Proc. of the 16th Int. Conf. on Control, Automation, Robotics and Vision*, Shenzhen, China, 2020, pp. 1236–1241.

9. Joshi, S. and Talange, D., Fault Tolerant Control for Autonomous Underwater Vehicle, *Proc. of the IEEE Int. Conf. on Mechatronics and Automation*, Tianjin, China, 2014, pp. 658–662.

10. Rotondo, D., Puig, V., Nejjari, F., and Romera, J., A Fault-Hiding Approach for the Switching Quasi-LPV Fault-Tolerant Control of a Four-Wheeled Omnidirectional Mobile Robot, *IEEE Trans. Indust. Electronics*, 2015, vol. 62, pp. 3932–3944.

11. Ling, Y., Sun, X., Wu, X., and Liu, J., Robust $H_\infty$ Fault-Tolerant Control for Uncertain Linear System Based on Pole Assignment, *Proc. of the 2nd IEEE Conf. Indust. Electronics and Applications*, Harbin, China, 2007, pp. 2701–2706.

12. Liu, C., Jiang, B., and Zhang, K., Adaptive Fault-Tolerant $H$-infinity Output Feedback Control for Lead-Wing Close Formation Flight, *IEEE Trans. Syst. Man. Cybernet. Syst.*, 2020, vol. 50, pp. 2804–2814.

13. Miguel, A., Puig, V., and Alenya, G., Fault-Tolerant Control of a Service Robot Using a LPV Robust Unknown Input Observer, *Proc. of the 4th Conf. Control and Fault Tolerant Systems*, Casablanca, Morocco, 2019, pp. 207–212.

14. Nair, R., Karki, H., Shukla, A., Behera, L., and Jamshidi, M., Fault-Tolerant Formation Control of Nonholonomic Robots Using Fast Adaptive Gain Nonsingular Terminal Sliding Mode Control, *IEEE Syst. J.*, 2019, vol. 13, pp. 1006–1017.

15. Van, M., Ge, S., and Ren, H., Robust Fault-Tolerant Control for a Class of Second-Order Nonlinear Systems Using an Adaptive Third-Order Sliding Mode Control, *IEEE Trans. Syst. Man. Cybernet. Syst.*, 2017, vol. 47, pp. 221–228.

16. Yin, S., Yang, H., and Kaynak, O., Sliding Mode Observer-Based FTC for Markovian Jump Systems with Actuator and Sensor Faults, *IEEE Trans. Autom. Control*, 2017, vol. 62, no. 7, pp. 3551–3558.

17. Chen, C., Xu, S., and Liang, Y., Study of Nonlinear Integral Sliding Mode Fault-Tolerant Control, *IEEE/ASME Trans. Mechatronics*, 2016, vol. 21, pp. 1160–1168.

18. Kaldmae, A., Kotta, U., Shumsky, A., and Zhirabok, A., Measurement Feedback Disturbance Decoupling in Discrete-Time Nonlinear Systems, *Automatica*, 2013, vol. 49, pp. 2887–2891.

19. Kaldmae, A., Kotta, U., Jiang, B., Shumsky, A., and Zhirabok, A., Faulty Plant Reconfiguration Based on Disturbance Decoupling Methods, *Asian J. Control*, 2016, vol. 8, no. 3, pp. 858–867.

20. *Spravochnik po teorii avtomaticheskogo upravleniya* (Handbook on Automatic Control Theory), Krasovskii, A.A., Ed., Moscow: Nauka, 1987.

21. Isidori, A., Krener, A., Gori-Giorgi, C., and Monaco, S., Nonlinear Decoupling via Feedback: A Differential Geometric Approach, *IEEE Trans. Autom. Control*, 1981, vol. AC-26, pp. 331–345.

22. Zhirabok, A., Shumsky, A., Solyanik, S., and Suvorov, A., Design of Nonlinear Robust Diagnostic Observers, *Autom. Remote Control*, 2017, vol. 78, no. 9, pp. 1572–1584.

23. Zhirabok, A.N., Zuev, A.V., and Bobko, E.Yu., Method of Virtual Sensor Design for Faulty Physical Sensor Replacement, *Mekhatronika, Avtomatizatsiya, Upravlenie*, 2023, vol. 24, no. 10, pp. 526–532.

24. Zhirabok, A.N., Zuev, A.V., and Kim, C., Interval Estimation in Discrete-Time Linear Systems with Parametric Uncertainties, *J. Comput. Syst. Sci. Int.*, 2023, vol. 62, no. 6, pp. 1037–1047.

25. Low, X., Willsky, A., and Verghese, G., Optimally Robust Redundancy Relations for Failure Detection in Uncertain Systems, *Automatica*, 1996, vol. 22, pp. 333–344.

26. Patton, R. and Chen, J., A Review of Parity Space Approach to Fault Diagnosis, *Proc. of the 1st IFAC Symp. on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS)*, Baden-Baden, Germany, 1991, pp. 239–255.

*This paper was recommended for publication by A.A. Bobtsov, a member of the Editorial Board*

*ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 7, pp. 678–685.*

# Stabilization of a Chain of Three Integrators Subject to a Phase Constraint

**A. V. Pesterev**[*,a] **and Yu. V. Morozov**[*,b]

*Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
*e-mail:* [a]*alexanderpesterev.ap@gmail.com,* [b]*tot1983@ipu.ru*

Received March 4, 2024
Revised May 7, 2024
Accepted May 30, 2024

**Abstract**—The problem of stabilizing a chain of three integrators subject to a phase constraint is studied. Continuous constrained control in the form of nested sigmoids, which guarantees the fulfillment of the phase constraint, is synthesized. A Lyapunov function is constructed, and necessary and sufficient conditions of global stability of the closed-loop system are established. The discussion is illustrated by numerical examples.

*Keywords*: stabilization of a chain of three integrators, global stability, phase constraint, nested sigmoids, Lyapunov function

**DOI:** 10.31857/S0005117924070039

## 1. INTRODUCTION

The problem of stabilizing a chain of three integrators subject to a phase constraint by means of a continuous control is studied. Stabilization of chains of integrators is one of the topical control problems, which has been widely discussed in the literature during last several decades (see, e.g., [1, 2] and references therein). The interest to this problem is due to the fact that original models in many applications are specified as chains of integrators and the controls developed for chains of integrators are easily extended to other classes of systems.

Among the variety of stabilizing controls applied to solving this problem, the class of feedbacks in the form of nested (both smooth and non-smooth) saturation functions can be distinguished [2–14]. The interest to such feedbacks is explained by the number of remarkable properties of the closed-loop system obtained: they automatically take into account boundedness of the control resource and ensure fulfillment of certain phase constraints, which is especially important far from the equilibrium state, as well as guarantee exponential rate of the deviation decrease near the equilibrium [3–7]. Note also the use of such feedbacks in the problems related to the adjustment of coefficients in the robust control laws [8].

The use of feedbacks in the form of nested saturation functions gives rise to study of quite complicated nonlinear systems (in the case of non-smooth saturation functions, these are linear switching systems), stability analysis of which is a nontrivial task. Global stability has been proved mainly for second-order systems with nested saturators [3, 5, 9] and sigmoids [3, 10]. Practically in all works studying systems of order three or higher, only local stability was proved [3, 4, 11, 12]. In rare cases of feedbacks of special form, global stability has been established for systems of order three [12] (piecewise continuous control) or four [13] (impulse control). As far as the authors know, the problem of global stability for the general case of $n$ nested saturators was considered only in the works by A. Teel [2, 14]. However, global stability has been proved only in the case where limit values of the nested saturators satisfy certain inequalities, which are seldom fulfilled in

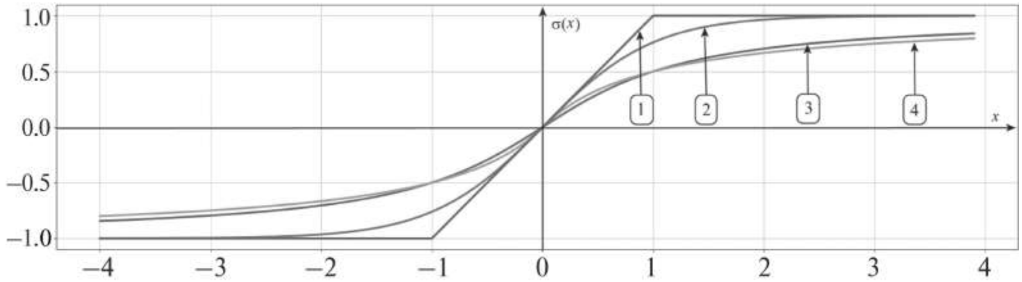


**Fig. 1.** Examples of saturation functions: $\mathrm{sat}(x)$ (1); $\tanh(x)$ (2); $2\arctan(x)/\pi$ (3); $x/(1+|x|)$ (4).

practice [2, Theorem 2.1]. The authors are not aware of works (except for abovementioned Teel's papers) where global stability were proved for a system of order three or higher stabilized by a continuous control guaranteeing fulfillment of a phase constraint.

*Saturation function* is a continuous nondecreasing function $S(x)$ of scalar variable that has finite limits when $x \to \pm\infty$. Among the saturation functions, the class of smooth strongly increasing functions called *sigmoids* can be distinguished [15]. In the literature, one can meet several slightly differing definitions of the sigmoids. We will use the following

**Definition 1.** Sigmoid is a smooth strongly increasing odd function of scalar variable $\sigma(x)$ satisfying the following conditions:

(a) $\sigma(x) \to \pm 1$ as $x \to \pm\infty$;

(b) $\max_x \sigma'(x) = \sigma'(0)$;

(c) $\sigma'(0) = 1$.

Functions satisfying the above definition but having different from ones limits at infinity and derivative at zero are referred to as *sigmoid functions.* Any sigmoid function $S(x)$ can be constructed from a sigmoid $\sigma(x)$ by specifying two coefficients: $S(x) = k_2\sigma(k_1 x)$, $k_1, k_2 > 0$. It is easy to see that, for any two sigmoid functions $S_1(x)$ and $S_2(x)$, $S(x) = S_1(S_2(x))$ is also a sigmoid function. When proving global stability, we will need the inequalities

$$S(x)x > 0 \ \forall x \neq 0, \tag{1}$$

$$[S(x+x_0) - S(x_0)]x > 0 \ \forall x \neq 0, \ \forall x_0, \tag{2}$$

which directly follow from the definition of the sigmoid.

The family of the sigmoid functions includes error function, arctangent, hyperbolic tangent, and other functions of similar form. The limit case of the sigmoid is the non-smooth saturation function called *saturators*: $\mathrm{sat}(x) = x$ when $|x| \leqslant 1$ and $\mathrm{sat}(x) = \mathrm{sgn}(x)$ when $|x| > 1$. Examples of the saturation functions are shown in Fig. 1. Other examples of the saturation functions and discussions of their properties can be found in [15]. In the control problems, the hyperbolic tangent is most often used as a sigmoid since it approximates the saturator better than other smooth saturation functions and, moreover, its derivatives are expressed in terms of the function itself. In the framework of this study, it does not matter what sigmoid is used in the feedback, since the proof of global stability is valid for any functions satisfying the above definition.

In this work, we suggest to stabilize a chain of three integrators by means of a special feedback including two nested sigmoids. The goal of the study is to prove global stability of the closed-loop system obtained under certain simple conditions on the feedback coefficients.

## 2. PROBLEM STATEMENT

We consider the problem of stabilizing a third-order integrator

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = x_3, \quad \dot{x}_3 = U(x), \quad x \equiv [x_1, x_2, x_3]^{\mathrm{T}}, \tag{3}$$

at the origin by means of a smooth feedback $U(x)$ guaranteeing the fulfillment of the phase constraint

$$|x_3(t)| \leqslant X_3. \tag{4}$$

Such a statement naturally comes to existence in many applications, for example, when stabilizing a mechanical system [11], where state variables are position, velocity, and traction (acceleration) and the system is controlled by varying the traction (e.g., by means of a step motor). A similar system with the phase constraint on the third variable, but with a discontinuous control, was considered in [16]. Since the traction in real systems is limited, the stabilizing control must not result in the violation of the phase constraint (4), where $X_3$ is the maximum possible traction.

The stabilizing control is sought in the form

$$U(x) = -k_5(x_3 + k_4\sigma_2(k_3(x_2 + k_2\sigma_1(k_1x_1)))), \tag{5}$$

where $\sigma_1$ and $\sigma_2$ are arbitrary sigmoids. The feedback of this form guarantees the fulfillment of phase constraint (4) with $X_3 = k_4$ if $|x_3(0)| \leqslant k_4$. Indeed, suppose that the phase constraint is satisfied at the initial moment. Variable $x_3(t)$ achieves local extremum on the trajectory when $U(x) = 0$; on the other hand, from formula (5), it is seen that the control equals zero when $x_3 = -k_4\sigma_2(\cdot)$. Hence, $|x_3(t)|$ cannot be greater than $k_4$; i.e., domain $|x_3| \leqslant k_4$ is an invariant set of the system. Thus, if variable $x_3$ cannot physically exceed its limit value (like, for instance, in the abovementioned example of the mechanical system), then it is sufficient to study stability of the system in this invariant set. We, however, consider a more general problem statement and will prove stability for any initial conditions in $R^3$. In so doing, if the initial point belongs to the invariant set, then the phase constraint (4) is fulfilled for any $t \geqslant 0$; otherwise, starting from some (depending on the initial conditions) finite instant.

Additional advantages of control (5) are (a) exponential rate of the deviation decrease near the equilibrium and (b) its boundedness for any deviations from the equilibrium state as long as the phase constraint is fulfilled at the initial point.

Coefficients $k_2$ and $k_4$, which set limits of sigmoid variations, are referred to as *model parameters,* since their values are determined by the model of the system under study, and, unlike the other three coefficients cannot be selected arbitrarily. Given $k_2$ and $k_4$, parameters $k_1$, $k_3$, and $k_5$ determine the character of the transition process [5, 7] and are referred to as *design parameters.* They are selected by the designer of the control system with the aim, for instance, to optimize (in one or another sense) its performance.

Without loss of generality, the model parameters can be set equal to ones, which reduces the number of system parameters to three. Indeed, let us turn to the dimensionless model by applying the same change of variables and time as in the two-dimensional case [5], i.e., $\tilde{t} = k_4t/k_2$, $\tilde{x}_1 = k_4x_1/k_2^2$, and $\tilde{x}_2 = x_2/k_2$, and define the third dimensionless variable as $\tilde{x}_3 = x_3/k_4$. Substituting the new variables into system (3), (5) and turning to differentiation with respect to the dimensionless time, we obtain the dimensionless model in which $\tilde{k}_2 = \tilde{k}_4 = 1$ and three other coefficients are given by the formulas $\tilde{k}_1 = k_1k_2^2/k_4$, $\tilde{k}_3 = k_2k_3/k_4$, and $\tilde{k}_5 = k_2k_5/k_4$. In what follows, we assume that all variables and parameters are dimensionless and will use the same notation (without tilde) for them. In the dimensionless model, feedback (5) takes the form

$$U(x) = -k_5(x_3 + \sigma_2(k_3(x_2 + \sigma_1(k_1x_1)))). \tag{6}$$

The goal of the study is to determine the conditions on the coefficients for which the proposed feedback stabilizes the system in the entire space, i.e., to establish conditions of global stability of system (3), (6). The study of stability presented in the next section is based on the construction of an integral Lyapunov function of the closed-loop system. We will prove that the necessary conditions of stability of the linearized in the neighborhood of the origin system are sufficient for its global stability. Note that the application of other known approaches to studying stability, for example, those based on the construction of the Lurie–Postnikov function or on the immersion into the class of linear nonstationary systems with subsequent application of methods of absolute stability theory allows one to prove, as a rule, only local stability (even if the system under study is stable in the whole) and construct an estimate of the invariant attraction domain.

## 3. GLOBAL STABILITY CONDITIONS

**Theorem 1.** *System* (3), (6), *where* $\sigma_1(\cdot)$ *and* $\sigma_2(\cdot)$ *are arbitrary sigmoids, is globally asymptotically stable if and only if all the feedback coefficients are positive and* $k_5 > k_1$.

**Proof.** *Necessity.* In order that the system be globally stable, it is necessary that the linearized in the neighborhood of the origin system

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = x_3, \quad \dot{x}_3 = -k_5 x_3 - k_5 k_3 x_2 - k_5 k_3 k_1 x_1$$

be stable. Applying the Hurwitz criterion to the latter system, we find that it is stable when all the coefficients are positive and the condition $k_5 > k_1$ holds. *Sufficiency.* Let coefficients $k_1$, $k_3$ and $k_5$ be positive. Let us consider the function

$$V(x) = k_5^2 \int\limits_0^{x_1} \sigma_2(k_3\sigma_1(k_1 s))ds + k_5 \int\limits_0^{x_2} \sigma_2(k_3(s + \sigma_1(k_1 x_1)))ds + \frac{1}{2}(x_3 + k_5 x_2)^2 \tag{7}$$

and prove that it is Lyapunov function of system (3), (6).

Let $\Phi_1$ and $\Phi_2$ denote the first and second, respectively, integral terms in (7). Let us prove that their sum and, hence, the entire function $V(x)$ are positive $\forall x \in R^3$.

Let us transform the second term $\Phi_2$ by changing the integration variable $\tilde{s} = s + \sigma_1(k_1 x_1)$:

$$\Phi_2 = k_5 \int\limits_{\sigma_1(k_1x_1)}^{x_2+\sigma_1(k_1x_1)} \sigma_2(k_3\tilde{s})d\tilde{s} = k_5 \int\limits_0^{x_2+\sigma_1(k_1x_1)} \sigma_2(k_3\tilde{s})d\tilde{s} - k_5 \int\limits_0^{\sigma_1(k_1x_1)} \sigma_2(k_3\tilde{s})d\tilde{s}.$$

In the second term on the right-hand side of the last formula, we perform implicit one-to-one (by virtue of monotonicity of function $\sigma_1$) change of the integration variable $\tilde{s} = \sigma_1(k_1 s)$. Taking into account that $d\tilde{s} = k_1\sigma_1'(k_1 s)ds$, where the prime denotes differentiation with respect to the argument, the sum of $\Phi_1$ and $\Phi_2$ takes the following form:

$$\Phi_1 + \Phi_2 = k_5 \int\limits_0^{x_1} \sigma_2(k_3\sigma_1(k_1 s))[k_5 - k_1\sigma_1'(k_1 s)]ds + k_5 \int\limits_0^{x_2+\sigma_1(k_1x_1)} \sigma_2(k_3\tilde{s})d\tilde{s}.$$

The second integral on the right-hand side of this formula is positive by virtue of (1). Since the derivative of the sigmoid satisfies the condition $\sigma'(s) \leqslant 1$ and, by the assumption of the theorem, $k_5 > k_1$, we have

$$k_5 - k_1\sigma_1'(k_1 s) > 0, \tag{8}$$

from which it follows that the first integral and, hence, function $V(x)$ are positive for all $x \neq 0$.

It is evident that $V(x)$ tends to infinity as $||x|| \to \infty$. Further, differentiating $V(x)$ by virtue of system (3), (6) and omitting the argument $k_1x_1$ of functions $\sigma_1$ and $\sigma_1'$ to shorten the notation, we obtain

$$\begin{aligned}\dot{V} &= k_5^2\sigma_2(k_3\sigma_1(\cdot))x_2 + k_5x_2\int\limits_0^{x_2}\sigma_2'(k_3(s+\sigma_1(\cdot)))k_3\sigma_1'(\cdot)k_1ds \\ &\quad + k_5\sigma_2(k_3(x_2+\sigma_1(\cdot)))x_3 + (x_3+k_5x_2)[-k_5(x_3+\sigma_2(k_3(x_2+\sigma_1(\cdot))) + k_5x_3] \\ &= k_5^2\sigma_2(k_3\sigma_1(\cdot))x_2 - k_5^2\sigma_2(k_3(x_2+\sigma_1(\cdot)))x_2 + k_1k_3k_5\sigma_1'(\cdot)x_2\int\limits_0^{x_2}\sigma_2'(k_3(s+\sigma_1(\cdot)))ds.\end{aligned}$$

Let us transform the integral on the right-hand side of the last expression:

$$\int\limits_0^{x_2}\sigma_2'(k_3(s+\sigma_1(\cdot)))ds = \int\limits_{\sigma_1(\cdot)}^{x_2+\sigma_1(\cdot)}\sigma_2'(k_3\tilde{s})d\tilde{s} = \frac{1}{k_3}\sigma_2(k_3(x_2+\sigma_1(\cdot))) - \frac{1}{k_3}\sigma_2(k_3\sigma_1(\cdot)).$$

Substituting the expression obtained into the formula for $\dot{V}(x)$, we get

$$\begin{aligned}\dot{V}(x) &= k_5\sigma_2(k_3\sigma_1(\cdot))x_2(k_5 - k_1\sigma_1'(\cdot)) - k_5\sigma_2(k_3(x_2+\sigma_1(\cdot)))x_2(k_5-k_1\sigma_1'(\cdot)) \\ &= -k_5(k_5-k_1\sigma_1'(\cdot))[\sigma_2(k_3(x_2+\sigma_1(\cdot))) - \sigma_2(k_3\sigma_1(\cdot))]x_2.\end{aligned}$$

The product of the expression in the square brackets and $x_2$ is positive by virtue of (2), from which, with regard to (8), it follows that the derivative is negative definite for any $x_2 \neq 0$. The derivative vanishes only on the set $x_2 = 0$, which contains no entire trajectories but $x = 0$.

Thus, function $V(x)$ satisfies all the conditions of the Barbashin–Krasovski theorem [20], and, hence, the origin is asymptotically stable equilibrium of system (3), (6) in the whole. The theorem is proved.

## 4. NUMERICAL EXAMPLES

As an illustration, we present results of numerical calculations for the feedback (6) in the form of nested hyperbolic tangents with the coefficients $k_1 = 1$, $k_3 = 3$, and $k_5 = 5$. Figure 2 shows the invariant set of the system bounded by the level surface of the Lyapunov function (7) $V(x) = k_5^2$.

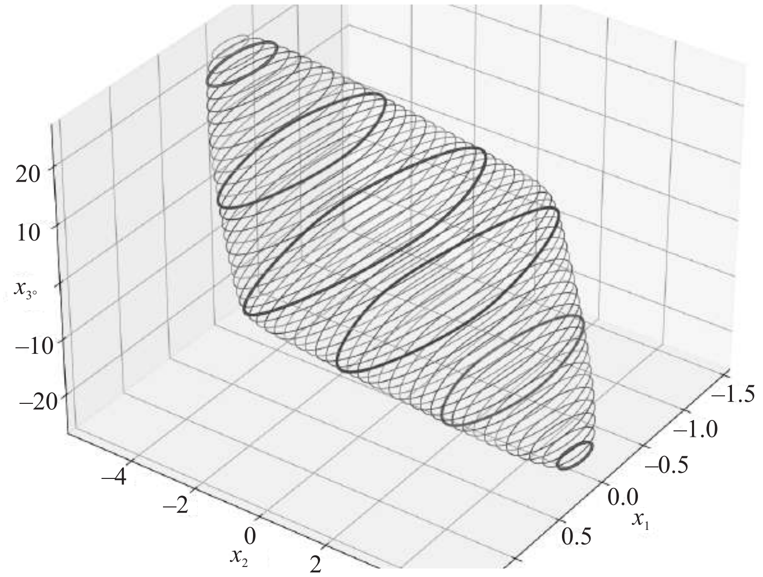


**Fig. 2.** Level surface of the Lyapunov function (7).

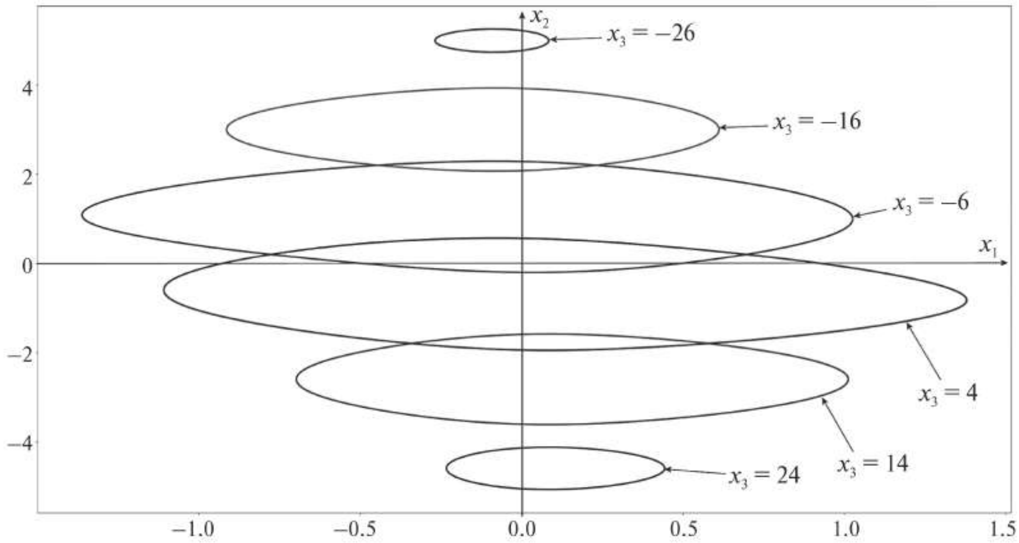


**Fig. 3.** Projections of cross-sections of the invariant set by planes $x_3 = \text{const}$ onto the plane $(x_1, x_2)$.

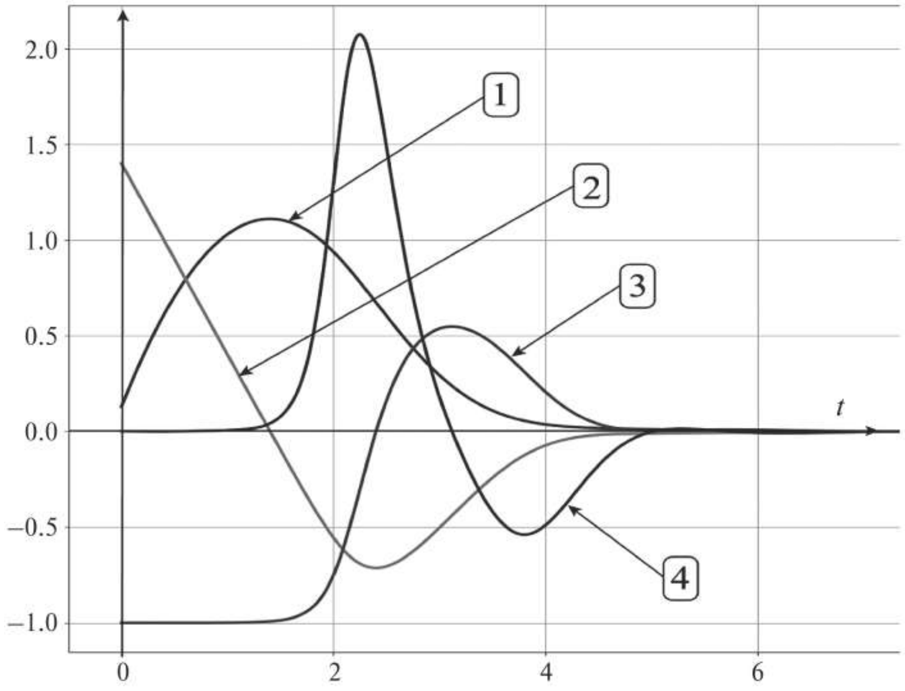


**Fig. 4.** Plots of deviation $x_1(t)$ (1), velocity $x_2(t)$ (2), acceleration $x_3(t)$ (3), and control $U(t)$ (4).

For greater clarity, Fig. 3 shows projections of six cross-sections of the level surface onto the plane $(x_1, x_2)$ (in Fig. 2, these cross-sections are depicted by bold lines) by the planes $x_3 = c_i$, $c_1 = -26$, $c_2 = -16$, $c_3 = -6$, $c_4 = 4$, $c_5 = 14$, and $c_6 = 24$.

Results of solving stabilization problem for the system with initial conditions $x_1(0) = 0.1$, $x_2(0) = 1.4$, $x_3(0) = -1$ are presented in Fig. 4, which demonstrates efficiency of the stabilization. The curves marked by 1, 2, 3, and 4 are plots of dependencies of deviation $x_1$, velocity $x_2$, acceleration $x_3$, and control $U$, respectively, on time. Although at the initial instant, the system moves in the direction opposite to the equilibrium state, the deviation, after natural growth at the initial stage, rapidly (exponentially) decreases, the phase constraint is fulfilled for any $t \geqslant 0$, control is reasonably constrained and does not result in overshooting.

## 5. CONCLUSIONS

The problem of stabilizing a chain of three integrators by a continuous control that guarantees the fulfillment of a phase constraint on the third state variable has been studied. By turning to dimensionless state variables, the original problem depending on five feedback coefficients has been reduced to study of a three-parameter system. Advantages of the proposed feedback in the form of nested sigmoids have been discussed. The basic result of the work is construction of the Lyapunov function by means of which sufficient conditions of global stability of the closed-loop system have been established. Numerical examples illustrating efficiency of stabilization by means of the proposed feedback have been presented.

## REFERENCES

1. Kurzhanski, A.B. and Varaiya, P., *Solution Examples on Ellipsoidal Methods: Computation in High Dimensions,* Cham, Switzerland: Springer, 2014.

2. Teel, A.R., Global Stabilization and Restricted Tracking for Multiple Integrators with Bounded Controls, *Sys. Cont. Lett.,* 1992, vol. 18, no. 3, pp. 165–171.

3. Olfati-Saber, R., *Nonlinear Control of Underactuated Mechanical Systems with Application to Robotics and Aerospace Vehicles*, Ph.D. dissertation, Massachusetts Inst. of Technology. Dept. of Electrical Engineering and Computer Sci., 2001.

4. Li, Y. and Lin, Z., *Stability and Performance of Control Systems with Actuator Saturation,* Basel: Birkhäuser, 2018.

5. Pesterev, A.V. and Morozov, Yu.V., Global Stabilization of a Chain of Two Integrators by a Feedback in the Form of Nested Saturators, *Autom. Remote Control,* 2024, no. 4, pp. 55–60.

6. Pesterev, A.V., Morozov, Yu.V., and Matrosov, I.V., On Optimal Selection of Coefficients of a Controller in the Point Stabilization Problem for a Robot-wheel, *Communicat. Comput. Inform. Sci.,* 2020, vol. 1340, pp. 236–249.

7. Pesterev, A.V. and Morozov, Yu.V., Optimizing Coefficients of a Controller in the Point Stabilization Problem for a Robot-Wheel, *Lect. Notes Comput. Sci.*, 2021, vol. 13078, pp. 191–202.

8. Antipov, A., Kokunko, J., and Krasnova, S., Dynamic Models Design for Processing Motion Reference Signals for Mobile Robots, *J. Intelligent Robot. Syst.,* 2022, vol. 105, pp. 1–16.

9. Hua, M.-D. and Samson, C., Time Sub-optimal Nonlinear pi and pid Controllers Applied to Longitudinal Headway Car Control, *Int. J. Control,* 2011, vol. 84, pp. 1717–1728.

10. Pesterev, A.V. and Morozov, Yu.V., Global Stabilization of a Chain of Two Integrators by a Feedback in the Form of Nested Sigmoids, *J. Comput. Syst. Sci. Int.*, 2024, vol.63, no. 3, pp. 385–389.

11. Matyukhin, V.I. and Pyatnitskii, E.S., Controllability of Mechanical Systems in the Class of Controls Bounded Together with Their Derivativesl, *Autom. Remote Control*, 2004, vol. 65, pp. 1187–1209.

12. Pesterev, A.V. and Morozov, Yu.V., The Best Ellipsoidal Estimates of Invariant Sets for a Third-Order Switched Affine System, *Lect. Notes Comput. Sci.,* 2022, vol. 13781, pp. 66–78.

13. Morozov, Yu.V. and Pesterev, A.V., Global Stability of a Fourth-Order Hybrid Affine System, *J. Comput. Syst. Sci. Int.*, 2023, vol. 62, no. 4, pp. 607–618.

14. Teel, A.R., A Nonlinear Small Gain Theorem for the Analysis of Control Systems with Saturation, *Trans. Autom. Contr. IEEE,* 1996, vol. 41, no. 9, pp. 1256–1270.

15. Mazhar, N., Malik, F.M., Raza, A., and Khan, R. Predefined-Time Control of Nonlinear Systems: A Sigmoid Function Based Sliding Manifold Design Approach, *Alexandria Engineer. J.*, 2022, vol. 61, no. 6, pp. 6831–6841.

16. Utkin, V.I. and Jingxin, Shi., Integral Sliding Mode in Systems Operating under Uncertainty Conditions, *Proc. of 35th IEEE Conf. Decision Control*, Kobe, Japan, 1996, vol. 4, pp. 4591–4596.

17. Lurie, A.I. and Postnikov, V.N., On Stability Theory of Regulated Systems, *Prikl. Matem. i Mekh.*, 1944, vol. 8, pp. 246–248.

18. Rapoport, L.B., Estimation of Attraction Domains in Wheeled Robot Control, *Autom. Remote Control*, 2006, vol. 67, no. 9, pp. 1416–1435.

19. Generalov, A., Rapoport, L., and Shavin, M., Attraction Domains in the Control Problem of a Wheeled Robot Following a Curvilinear Path over an Uneven Surface, *Lect. Notes Comput. Sci.*, 2021, vol. 13078, pp. 176–190.

20. Barbashin, E.A., *Vvedeniye v teoriyu ustoychivosti. Seriya: Fiziko-matematicheskaya biblioteka inzhenera* (Introduction to Stability Theory. Series: Physico-mathematical library of engineer), Moscow: Nauka, 1967.

*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, 2024, Vol. 85, No. 7, pp. 686–700.*

# Optimal Control of Harvesting of a Distributed Renewable Resource on the Earth's Surface

**D. V. Tunitsky**

*Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
*e-mail: dtunitsky@yahoo.com*

Received May 3, 2024
Revised May 27, 2024
Accepted May 30, 2024

**Abstract**—This paper is devoted to the optimal control of mixed (stationary and periodic impulse) harvesting of a renewable resource distributed on the Earth's surface. Examples of such a resource are biological populations, including viruses, chemical contaminants, dust particles, and the like. It is proved that on an infinite planning horizon, there exists an admissible control ensuring the maximum of time-averaged harvesting.

*Keywords*: the Kolmogorov–Petrovskii–Piskunov–Fisher equations, second-order parabolic equations, semilinear equations on a sphere, weak solutions, stabilization, optimal control

**DOI:** 10.31857/S0005117924070043

## 1. INTRODUCTION

Two-dimensional (2D) manifolds homeomorphic to a sphere are commonly used as a mathematical model of the Earth's surface. The dynamics of a renewable resource distributed on the Earth's surface can be modeled by a second-order semilinear evolutionary equation on a 2D sphere. In local coordinates, it has the form

$$\frac{\partial q}{\partial t} - \sum_{l,m=1}^{2} \frac{\partial}{\partial x^l}\left(a^{l,m}(x)\frac{\partial q}{\partial x^m}\right) = A(x)q - B(x)q^2, \quad a^{l,m}(x) = a^{m,l}(x), \tag{1}$$

where the matrix $a$ characterizes the resource diffusion, and the coefficients $A$ and $B$ are the resource renewal and saturation rates of the environment. Essentially, equation (1) combines two classical models: the Verhulst logistic model [1] and the Fourier heat propagation model [2].

Equations of the form (1) arise when modeling various reaction–diffusion processes in a distributed environment. One example is the famous model proposed by A.N. Kolmogorov, G.I. Petrovskii, and N.S. Piskunov [3] and R.A. Fischer [4]. Information about other models, the history and bibliography of the works on this topic can be found in [5]. Also, the interested reader is referred to the monograph [6], covering several applied aspects.

Second-order semilinear evolutionary equations in Euclidean space domains have been studied quite thoroughly; for example, see [7–9]. On closed manifolds, particularly spheres, they have been investigated to a lesser extent. It is appropriate to mention the papers devoted to the equations with periodically fragmented coefficients [5, 10] (in fact, equations on a torus). This case, important from an applied point of view, occurs when modeling periodic media. Of course, equations of the form (1) on a 2D sphere are also of significant interest: this is a standard model of the Earth's surface used in applications.

Note that many applied problems lead to equations of the type (1) with discontinuous coefficients. In particular, this is characteristic of optimal control problems. Therefore, it is desirable to choose a class of admissible solutions to construct a satisfactory theory of the corresponding equations with minimal regularity requirements for their coefficients. In this paper, such a class consists of weak solutions. In the class of weak solutions, it is possible to study equations of the form (1) on a 2D sphere with fairly light regularity requirements for their coefficients.

## 2. FUNCTION SPACES AND EVOLUTIONARY EQUATIONS

### *2.1. Function Spaces*

Let $\mathbb{S}^2$ be a 2D sphere of unit radius, $\{(y^1, y^2, y^3) \in \mathbb{R}^3 \mid (y^1)^2 + (y^2)^2 + (y^3)^2 = 1\}$, standardly embedded in the 3D Euclidean space $\mathbb{R}^3$. The stereographic projection

$$h : \mathbb{S}^2 \backslash (0, 0, 1) \ni (y^1, y^2, y^3) \mapsto \frac{(y^1, y^2)}{1 - y^3} \in \mathbb{R}^2$$

relative to the pole $(0, 0, 1)$ specifies a local coordinate system defined on $\mathbb{S}^2$ everywhere except the pole [11] (lecture 6). An embedding in the Euclidean space $\mathbb{R}^3$ induces on $\mathbb{S}^2$ a Riemannian metric $g$, whose inverse image relative to $h^{-1}$ has the form

$$(h^{-1})^* g = 4\frac{(dx^1)^2 + (dx^2)^2}{((x^1)^2 + (x^2)^2 + 1)^2}.$$

Here $h^{-1}$ is a mapping inverse to the stereographic projection, i.e.,

$$h^{-1} : \mathbb{R}^2 \ni (x^1, x^2) \mapsto \frac{1}{(x^1)^2 + (x^2)^2 + 1}(2x^1, 2x^2, (x^1)^2 + (x^2)^2 - 1) \in \mathbb{S}^2. \tag{2}$$

The metric $g$ defined on the tangent bundle $T\mathbb{S}^2$ admits a natural extension to tensor bundles $(T\mathbb{S}^2)^{\bigotimes^m} \bigotimes (T^*\mathbb{S}^2)^{\bigotimes^l}$, $m, l = 0, 1, 2, \ldots$, which will be denoted by the same symbol $g$. On $(T\mathbb{S}^2)^{\bigotimes^0} \bigotimes (T^*\mathbb{S}^2)^{\bigotimes^0} = \mathbb{S}^2 \times \mathbb{R}$, the metric is $g(r_1, r_2) = r_1 r_2$ for $r_1, r_2 \in \mathbb{R}$. Also, $g$ induces on $\mathbb{S}^2$ a complete metric space structure and a measure $\mu = \mu_g$, whose image relative to the stereographic projection has the form

$$d(\mu \circ h) = \frac{4dx^1 dx^2}{((x^1)^2 + (x^2)^2 + 1)^2}. \tag{3}$$

These structures are used to build the Lebesgue spaces of functions and tensor fields, $L^p(\mathbb{S}^2)$ and $L^p((T\mathbb{S}^2)^{\bigotimes^m} \bigotimes (T^*\mathbb{S}^2)^{\bigotimes^l})$, where $p \geqslant 1$ and $m, l = 0, 1, 2, \ldots$, as well as the Sobolev spaces $W^{1,p}(\mathbb{S}^2)$ and $W^{1,p}((T\mathbb{S}^2)^{\bigotimes^m} \bigotimes (T^*\mathbb{S}^2)^{\bigotimes^l})$ [12, Ch. 2] and the Hölder spaces $C^\alpha(\mathbb{S}^2)$ and $C^\alpha((T\mathbb{S}^2)^{\bigotimes^m} \bigotimes (T^*\mathbb{S}^2)^{\bigotimes^l})$, $0 < \alpha \leqslant 1$ [13, Sec. 10.2.4; 14; 15; 16, §1]. For this purpose, the stereographic coordinates (2) can be applied. For example, the function spaces $L^p(\mathbb{S}^2)$ on a sphere and $L^p(\mathbb{R}^2, \mu \circ h)$ on the plane with the measure (3) are isometric for $p \geqslant 1$.

Consider real-valued measurable functions $u$ and $v$ defined on $\mathbb{S}^2$ and let

$$ess \sup_{x \in \mathbb{S}^2} u(x) = \inf_{\substack{S \subseteq \mathbb{S}^2, \\ \mu(S)=0}} \sup_{x \in \mathbb{S}^2 \backslash S} u(x),$$

$$ess \inf_{x \in \mathbb{S}^2} u(x) = \sup_{\substack{S \subseteq \mathbb{S}^2, \\ \mu(S)=0}} \inf_{x \in \mathbb{S}^2 \backslash S} u(x), \langle u, v \rangle = \int\limits_{\mathbb{S}^2} uv d\mu.$$

If $\mathfrak{B}$ is a Banach space with a norm $\|\cdot\|_{\mathfrak{B}}$, then for fixed $T_0 \in (0,+\infty)$ and $T_1 \in (0,+\infty]$, $T_0 < T_1$, the spaces $L^p([T_0,T_1);\mathfrak{B})$ with the norms

$$\|q\|_{L^p([T_0,T_1);\mathfrak{B})} = \left(\int\limits_{T_0}^{T_1} \|q(t)\|^p_{\mathfrak{B}} dt\right)^{\frac{1}{p}}, \quad p \geqslant 1,$$

$$\|q\|_{L^\infty([T_0,T_1);\mathfrak{B})} = ess \sup_{t\in[T_0,T_1)} \|q(t)\|_{\mathfrak{B}}$$

are also Banach spaces; see [17, Ch. III, §1] and [18, Ch. II, §2]. The intersection

$$W([T_0,T_1);X) = L^2((T_0,T_1);W^{p,1}(X)) \cap L^\infty([T_0,T_1);L^2(X))$$

is also a Banach space with the norm

$$\|q\|^2_{W([T_0,T_1);X)} = ess \sup_{t\in[T_0,T_1)} \langle q(t),q(t)\rangle + \int\limits_{T_0}^{T_1} \langle g(dq(t),dq(t)),1\rangle dt.$$

For brevity, we will use the abbreviation *a.e.* whenever some properties are valid almost everywhere in the measure $\mu$ on $\mathbb{S}^2$, see (3).

### *2.2. Evolutionary Equations*

Along with $g$, let another metric $a$ be defined on the sphere $\mathbb{S}^2$. Assume that this metric is measurable and there exist $a_0, a_1 \in (0,+\infty)$ such that

$$a_0 g(\eta,\eta) \leqslant a(\eta,\eta) \leqslant a_1 g(\eta,\eta), \quad \eta \in T^*\mathbb{S}^2, \text{ a.e.} \tag{4}$$

In the stereographic coordinates $x^1$ and $x^2$ (2), the estimate (4) has the form

$$\frac{4a_0(\eta_1^2+\eta_2^2)}{((x^1)^2+(x^2)^2+1)^2} \leqslant a^{1,1}(t)\eta_1^2 + 2a^{1,2}(t)\eta_1\eta_2 + a^{2,2}(t)\eta_2^2 \leqslant \frac{4a_1(\eta_1^2+\eta_2^2)}{((x^1)^2+(x^2)^2+1)^2}.$$

Consider the differential operator $d^*_{a,g} : C^\infty(T^*\mathbb{S}^2) \ni w \mapsto d^*_{a,g}w \in C^\infty(\mathbb{S}^2)$ adjoint to the exterior differentiation operator $d$ with respect to the metrics $g$ and $a$, i.e.,

$$\langle a(du,\omega),1\rangle_g = \langle u, d^*_{a,g}\omega\rangle_g, \quad u \in C^\infty(\mathbb{S}^2), \quad \omega \in C^\infty(T^*\mathbb{S}^2);$$

for details, see [19, Ch. VIII, §1]. In the system of the stereographic coordinates $x^1$ and $x^2$ (2),

$$d^*_{a(t),g}\omega = -((x^1)^2+(x^2)^2+1)^2 \sum_{l,m=1}^{2} \frac{\partial}{\partial x^l} \frac{a(dx^l,dx^m)}{((x^1)^2+(x^2)^2+1)^2}\,\omega\left(\frac{\partial}{\partial x^m}\right).$$

Given a function $u \in C^\infty(\mathbb{S}^2)$, we define the *geometric Laplacian* (*the Laplace–de Rahm operator*), i.e., the linear second-order differential operator [20, Ch. IV, §5]

$$\triangle = \triangle_{a,g} = d^*_{a,g} \circ d. \tag{5}$$

Due to the estimate (4), the operator (5) is *uniformly elliptic* on $\mathbb{S}^2$.

Hence, the second-order evolutionary equation

$$\frac{\partial q}{\partial t} + \triangle q = (A(x) - u(x))q - B(x)q^2 \tag{6}$$

is *parabolic* on $\mathbb{S}^2$. In the stereographic coordinates $x^1$ and $x^2$ (2), it takes the form

$$\frac{\partial q}{\partial t} - ((x^1)^2 + (x^2)^2 + 1)^2 \sum_{l,m=1}^{2} \frac{\partial}{\partial x^l} \frac{a(dx^l, dx^m)}{((x^1)^2 + (x^2)^2 + 1)^2} \frac{\partial u}{\partial x^m} = (A(x) - u(x))q - B(x)q^2;$$

cf. (1). The unknown function $q = q(t, x)$ corresponds to the density of the renewable resource under consideration at a point $x$ of the sphere $\mathbb{S}^2$ at a time instant $t$, the metric $a$ characterizes the resource diffusion, the function $u$ is the control of its stationary (permanent) harvesting, and the coefficients $A$ and $B$ are the resource renewal and saturation rates of the environment.

Weak solutions, subsolutions, and supersolutions are defined in a conventional way [8, Ch. VI, §1, 5] and [9, §1.5]. In particular, a *weak solution* of equation (6) on the half-open interval $[T_0, T_1)$ is a function $q \in W([T_0, T_1); \mathbb{S}^2)$ such that $q^2 \in L^2([T_0, T_1) \times \mathbb{S}^2)$ and

$$\langle q, p\rangle(t) + \int_{T_0}^{t} \Big(\langle dq, dp\rangle_{L^2(T^*\mathbb{S}^2)} - \langle q, p'\rangle\Big)(\tau)d\tau = \langle q, p\rangle(0) + \int_{T_0}^{t} \langle (A - u)q - Bq^2, p\rangle(\tau)d\tau$$

for each $p \in C^\infty([T_0, T_1); \mathbb{S}^2)$ and $t \in [T_0, T_1)$. A weak solution $q$ of equation (1.5) that takes a given initial value of the resource density,

$$q(T_0) = q_0, \quad q_0 \in L^\infty(\mathbb{S}^2), \quad q_0 \geqslant 0 \text{ a.e.,} \tag{7}$$

is called a *weak solution of the Cauchy problem* (6), (7) *on* $[T_0, T_1)$.

In the presentation below, all solutions, subsolutions, and supersolutions are assumed to be weak, and the adjective "weak" is omitted for brevity.

## 3. PERIODIC IMPULSE HARVESTING AND CONTROLLED SOLUTIONS

### *3.1. Periodic Impulse Harvesting*

The periodic impulse harvesting of a renewable resource is mathematically modeled by the solution $q$ of the Cauchy problem (6), (7) with the additionally imposed conditions

$$q(kT) = sq(kT-), \ \ k = 1, 2, \ldots. \tag{8}$$

Here $T \in (0, +\infty)$ is a given period, and the measurable factor $s$, $0 \leqslant s \leqslant 1$ *a.e.*, characterizes the impulse harvesting rate. The solution of problem (6), (8) is a function $q \in L^\infty([0, +\infty) \times \mathbb{S}^2)$ that resolves equation (6) on $[kT, (k + 1)T)$, has the left-hand limit values $q(kT-)$, and satisfies *a.e.* equalities (8). If for $T_0 = 0$ this solution takes *a.e.* the initial value (7), then it represents the solution of problem (6), (7), (8). The solution of problem (6), (8) is said to be *periodic* if

$$q(t + T) = q(t), \ \ t \in [0, +\infty). \tag{9}$$

We define the *admissible sets* $\mathfrak{U}$ *and* $\mathfrak{S}$ *of stationary and impulse controls*

$$\begin{gathered}\mathfrak{U} = \{u \in L^\infty(\mathbb{S}^2) \mid U_1 \leqslant u \leqslant U_2\},\\ \mathfrak{S} = \{e^{-\beta v} \mid v \in L^\infty(\mathbb{S}^2), V_1 \leqslant v \leqslant V_2, \langle 1, v\rangle \leqslant E\},\end{gathered} \tag{10}$$

where $U_1, U_2, V_1, V_2, \beta \in L^\infty(\mathbb{S}^2)$ and $E \in [0, +\infty)$. Here $U_1$ and $U_2$ characterize the constraints on the possible density of stationary resource harvesting, $E$ is the admissible harvesting effort, and the limits $V_1$ and $V_2$ describe the minimum technically feasible density of impulse harvesting and its maximum possible density given the available physical capacity of the environment and ecological constraints. In essence, $V_1(x)$ and $V_2(x)$ are the minimum and maximum efforts that can be applied at a point $x$ to achieve the goals. The impulse factor form $s = e^{-\beta(x)v(x)}$ in (8) stems from the search theory [21–23]. The factor $\beta(x)$ in the exponent characterizes the complexity of detecting and extracting the resource at a point $x \in \mathbb{S}^2$, and $v(x)$ is the effort applied.

*Remark 1.* As is easily checked, the sets of admissible stationary $\mathfrak{U}$ and impulse $\mathfrak{S}$ controls (10) are convex, closed in $L^2(\mathbb{S}^2)$, and bounded in $L^\infty(\mathbb{S}^2)$. Since the space $L^2(\mathbb{S}^2)$ is reflexive, by the Eberlein–Šmulyan theorem, the sets bounded in the norm $\|\cdot\|_{L^2(\mathbb{S}^2)}$ are sequentially weakly precompact [24, App. to Ch. V, §4]. In addition, each convex and closed subset of $L^2(\mathbb{S}^2)$ is weakly closed [25, Sec. 2.9]. Therefore, the sets of admissible controls $\mathfrak{U}$ and $\mathfrak{S}$ are sequentially weakly compact. A subset in $L^2(\mathbb{S}^2)$ is weakly compact if and only if it is sequentially weakly compact, and sequentially weakly precompact sets are norm-bounded [25, Sec. 2.9]. Therefore, the sets $\mathfrak{U}$ and $\mathfrak{S}$ are weakly sequentially compact in $L^2(\mathbb{S}^2)$ [24, App. to Ch. V, §4 ].

*Remark 2.* Obviously, $q = 0$ is a periodic subsolution of problem (6), (7), (8). Due to the constraints imposed above on equation (6) and the admissible controls, $B \geqslant B_0 > 0$ and $0 \leqslant s \leqslant 1$ *a.e.* Hence, the constant function $q = c$ is a periodic supersolution of problem (6), (7), (8) for $c \geqslant Q(\| q_0 \|_{L^\infty(\mathbb{S}^2)})$, where

$$Q : \mathbb{R} \ni r \mapsto \max\left\{r, \frac{1}{B_0}(\|A\|_{L^\infty(\mathbb{S}^2)} + \max\{\|U_1\|_{L^\infty(\mathbb{S}^2)}, \|U_2\|_{L^\infty(\mathbb{S}^2)}\})\right\} \in \mathbb{R}. \tag{11}$$

### *3.2. Controlled Solutions*

The solutions $q = q(t; q_0, u, s)$ of problem (6), (7), (8) and the periodic solutions $q = q(t; u, s)$ of problem (6), (8) with admissible controls $u \in \mathfrak{U}$ and $s \in \mathfrak{S}$ will be called *controlled solutions.* They possess the following properties.

**Theorem 1.** *Assume that the metric a is measurable and satisfies the estimate* (4) *and the coefficients $A, B \in L^\infty(\mathbb{S}^2)$ and $B \geqslant B_0$ a.e. for some $B_0 \in (0, +\infty)$. Then:*

(a) *For any $u \in \mathfrak{U}$, $s \in \mathfrak{S}$, there exists a unique controlled solution $q = q(t; q_0, u, s)$. In addition, $q \in C([(k-1)T, kT); L^2(\mathbb{S}^2))$, $k = 0, 1, \dots$, and*

$$0 \leqslant q(t; q_0, u, s) \leqslant Q\left(\| q_0\|_{L^\infty(\mathbb{S}^2)}\right), \ \ t \in [0, +\infty), \tag{12}$$

*where $Q$ is the function* (11), *and for any $\varepsilon \in (0, T)$ there exists a number $\alpha$, $0 < \alpha \leqslant 1$, such that*

$$q \in C^\alpha\left(\bigcup_{k=1}^{\infty}[(k-1)T + \varepsilon, kT) \times \mathbb{S}^2\right).$$

(b) *If sequences $\{q_m\} \subseteq L^\infty(\mathbb{S}^2)$, $\{u_m\} \subseteq \mathfrak{U}$, and $\{s_m\} \subseteq \mathfrak{S}$ weakly converge in $L^2(\mathbb{S}^2)$, i.e., $q_m \rightharpoonup q_0$, $u_m \rightharpoonup u_0$, and $s_m \rightharpoonup s_0$, and $q_m \geqslant 0$ and $q_m \neq 0$ a.e., then the weak convergence*

$$\lim_{m\to+\infty} q\,(\cdot\ ; q_m, u_m, s_m) = q\,(\cdot\ ; q_0, u_0, s_0)$$

*holds in the spaces $L^2([0, NT); W^{1,2}(\mathbb{S}^2))$ for any $N = 1, 2, \dots$ and in the norms $\|\cdot\|_{C(\cup_{k=1}^N[(k-1)T+\varepsilon, kT)\times\mathbb{S}^2)}$ for any $\varepsilon \in (0, T)$.*

(c) *For any $u \in \mathfrak{U}$ and $s \in \mathfrak{S}$, there exists a unique controlled periodic solution $q = q_\infty(t; u, s)$ such that*

$$\lim_{t\to+\infty} \|q(t; q_0, u, s) - q_\infty(t; u, s)\|_{L^\infty(\mathbb{S}^2)} = 0, \ \ \|q_0\|_{L^\infty(\mathbb{S}^2)} > 0.$$

(d) *If sequences* $\{u_m\} \subseteq \mathfrak{U}$ *and* $\{s_m\} \subseteq \mathfrak{S}$ *weakly converge in* $L^2(\mathbb{S}^2)$, *i.e.,* $u_m \rightharpoonup u_0$ *and* $s_m \rightharpoonup s_0$, *then the periodic solutions from item* (c) *have the weak convergence*

$$\lim_{m\to+\infty} q_\infty(\cdot\ ; u_m, s_m) = q_\infty(\cdot\ ; u_0, s_0)$$

*in the space* $L^2((0,T); W^{1,2}(\mathbb{S}^2))$ *and in the norms* $\|\cdot\|_{C([\varepsilon,T)\times\mathbb{S}^2)}$ *for any* $\varepsilon \in (0,T)$.

The proof is given in subsection 4.5.

*Remark 3.* There exist at most two periodic solutions $q$ of problem (6), (8). According to Remark 2, one of them is the trivial solution $q = 0$. If $q_\infty = 0$, then by Theorem 1 the other disappears; if $q_\infty \neq 0$, then the third does the same.

## 4. PROBLEM STATEMENT, THE MAIN RESULT, AND FINDINGS

### *4.1. Problem Statement*

According to assertion (a) of Theorem 1, the following functional is well-defined for the admissible sets of stationary $\mathfrak{U}$ and impulse $\mathfrak{S}$ controls (10):

$$\begin{gathered} F : \{q_0 \in L^\infty(\mathbb{S}^2) | q_0 \geqslant 0 \text{ a.e.}\} \times \mathfrak{U} \times \mathfrak{S} \ni (q_0, u, s) \\ \longmapsto \overline{\lim_{t\to+\infty}} \frac{1}{t}\left(\int\limits_0^t \langle q(\tau; q_0, u, s), u\rangle d\tau + \sum_{0<kT\leqslant t} \langle q(kT-; q_0, u, s), 1 - s\rangle\right) \in \mathbb{R}, \end{gathered} \tag{13}$$

where $q = q(t; q_0, u, s)$ is a controlled solution. Its value is the time-averaged sum of the stationary (first term) and impulse (second term) resource harvestings.

Let us pose the following problem: *It is required to establish the existence of stationary* $u_0 \in \mathfrak{U}$ *and impulse* $s_0 \in \mathfrak{S}$ *controls that maximize the functional* $F$ (13), *and investigate the impact of the initial value* $q_0$ (7) *on* $F(q_0, u_0, s_0)$, *cf.* [26] *and* [27].

### *4.2. The Main Result*

Using Theorem 1, we provide a comprehensive solution of this problem. Namely, the following result is true; cf. [28].

**Theorem 2.** *Assume that all conditions of Theorem 1 are satisfied. Then:*

(a) *For any initial values* $q_0$ (7), $\|q_0\|_{L^\infty(\mathbb{S}^2)} > 0$, *and admissible controls* $u \in \mathfrak{U}$, $s \in \mathfrak{S}$, *we have the equality*

$$F(q_0, u, s) = F(q_\infty(0; u, s), u, s) = \frac{1}{T}\left(\int\limits_0^T \langle q_\infty(\tau; u, s), u\rangle d\tau + \langle q_\infty(T-; u, s), 1 - s\rangle\right). \tag{14}$$

(b) *If sequences* $\{u_m\} \subseteq \mathfrak{U}$ *and* $\{s_m\} \subseteq \mathfrak{S}$ *weakly converge in* $L^2(\mathbb{S}^2)$, *i.e.,* $u_m \rightharpoonup u_0$ *and* $s_m \rightharpoonup s_0$, *and a sequence* $\{q_m\} \subseteq L^\infty(\mathbb{S}^2)$ *is such that* $q_m \geqslant 0$ *and* $q_m \neq 0$ *a.e., then*

$$\lim_{m\to+\infty} F(q_m, u_m, s_m) = F(q_\infty(0; u_0, s_0), u_0, s_0).$$

(c) *The functional* $F$ (13) *is bounded and its supremum is achieved at admissible controls* $u_0 \in \mathfrak{U}$ *and* $s_0 \in \mathfrak{S}$ *so that*

$$\sup F(q_0, u, s) = F(q_\infty(0;\ u_0, s_0), u_0, s_0).$$

**Proof.**

(a) Clearly, the value of the functional $F(q_0, u, s)$ will not change when replacing the zero lower limits of integration and summation in its definition (13) by any $T_0 \in [0, +\infty)$. Next, for controlled solutions $q = q(t; q_0, u, s)$ of problem (6), (7), (8) and a periodic solution $q = q_\infty(t; u, s)$ of problem (6), (8), we have

$$\left| \int\limits_{T_0}^{t} \langle q(\tau) - q_\infty(\tau), u\rangle d\tau + \sum_{T_0 < kT \leqslant t} \langle q(kT-) - q_\infty(kT-), 1 - s\rangle \right|$$

$$\leqslant \left( t \parallel u \parallel_{L^\infty(\mathbb{S}^2)} \sup_{\tau \geqslant T_0} \|q(\tau) - q_\infty(\tau)\|_{L^\infty(\mathbb{S}^2)} + [t]\|1 - s\|_{L^\infty(\mathbb{S}^2)} \sup_{kT \geqslant T_0} \|q(kT-) - q_\infty(kT-)\|_{L^\infty(\mathbb{S}^2)} \right)$$

$$\leqslant t\Big( \max\{\|U_1\|_{L^\infty(\mathbb{S}^2)}, \|U_2\|_{L^\infty(\mathbb{S}^2)}\} + 1 \Big) \sup_{\tau \geqslant T_0} \|q(\tau) - q_\infty(\tau)\|_{L^\infty(\mathbb{S}^2)}, \qquad t \in [T_0, +\infty).$$

By assertion (c) of Theorem 1, it follows that $|F(q_0, u, s) - F(q_\infty(0; u, s), u, s)| = 0$. Due to definition (13), $F(q_\infty(0; u, s), u, s)$ equals the right-hand side of equality (14).

(b) According to assertion (a), we have

$$F(q_0, u_m, s_m) = \frac{1}{T} \left( \int\limits_0^T \langle q_\infty(\tau; u_m, s_m), u_m\rangle d\tau + \langle q_\infty(T-; u_m, s_m), 1 - s_m\rangle \right).$$

By assertion (d) of Theorem 1, it is possible to pass to the limit on the right-hand side of this expression as $m \to +\infty$ [29, Ch. 1, §5]. As a result, in view of (14), we arrive at the desired conclusion.

(c) There exist sequences of initial values $\{q_m\} \subseteq L^\infty(\mathbb{S}^2)$ and admissible controls $\{u_m\} \subseteq \mathfrak{U}$ and $\{s_m\} \subseteq \mathfrak{S}$ such that

$$\sup F(q_0, u, s) = \lim_{m \to +\infty} F(q_m, u_m, s_m).$$

Due to Remark 1, the sets of admissible controls $\mathfrak{U}$ and $\mathfrak{S}$ are sequentially weakly compact in $L^2(\mathbb{S}^2)$. Hence, there exist subsequences $\{u_{m_l}\}$ and $\{s_{m_l}\}$ that weakly converge in $L^2(\mathbb{S}^2)$, i.e., $u_{m_l} \rightharpoonup u_0 \in \mathfrak{U}$ and $s_{m_l} \rightharpoonup s_0 \in \mathfrak{S}$. By assertion (b), we obtain

$$\sup F(q_0, u, s) = \lim_{m \to +\infty} F(q_m, u_m, s_m) = F(q_\infty(0; u_0, s_0), u_0, s_0).$$

The proof of Theorem 2 is complete.

### *4.3. Findings*

According to assertion (c) of Theorem 1, after choosing admissible stationary and impulse controls, the renewable resource density will uniformly tend to a unique limit state for any nonzero initial values. According to assertion (c) of Theorem 2, admissible controls can be chosen so that for each exploitation cycle, the amount of resource harvesting coincides with the maximum possible time-averaged amount of resource harvesting. In other words, with the optimal control of renewable resource exploitation, any nonzero initial resource density will tend to a limiting state ensuring the maximum of resource harvesting in one exploitation cycle.

## 5. PROOF OF THEOREM 1

### *5.1. Auxiliary Assertions*

According to Remark 1, the subsolution of problem (6), (7), (8) is the zero function $q = 0$, and the supersolution is the constant function $q = Q(\|q_0\|_{L^\infty(\mathbb{S}^2)})$. Therefore, the known results for second-order semilinear parabolic equations on a sphere [30–32] imply the following.

**Lemma 1.** *Assume that all conditions of Theorem 1 are satisfied. Then for each $u \in L^\infty(\mathbb{S}^2)$ there exists a unique solution $q = q(\,\cdot\,; q_0, u)$ of problem* (6), (7) *on the half-open interval $[T_0, +\infty)$. Moreover, $q \in C([T_0, +\infty); L^2(\mathbb{S}^2))$, $0 \leqslant q(t) \leqslant Q(\|q_0\|_{L^\infty(\mathbb{S}^2)})$ a.e. for $t \in [T_0, +\infty)$, and for each $\varepsilon > 0$ it is possible to find $\alpha = \alpha(\varepsilon, \|q\|_{L^\infty([T_0,+\infty))\times(\mathbb{S}^2)})$, $0 < \alpha \leqslant 1$, and $C = C(\varepsilon, \|q\|_{L^\infty([T_0,+\infty))\times\mathbb{S}^2}) \geqslant 0$ such that $q \in C^\alpha([T_0 + \varepsilon, +\infty)) \times \mathbb{S}^2$ and $\|q\|_{C^\alpha([T_0+\varepsilon,+\infty))\times(\mathbb{S}^2)} \leqslant C$.*

In addition, we have the following fact.

**Lemma 2.** *Assume that all conditions of Theorem 1 are satisfied. If sequences $\{q_m\} \subseteq L^\infty(\mathbb{S}^2)$ and $\{u_m\} \subseteq \mathfrak{U}$ weakly converge in $L^2(X)$, i.e., $q_m \rightharpoonup q_0$ and $u_m \rightharpoonup u_0$, then the solutions $q = q(t; q_m, u_m)$ of the Cauchy problem* (6), (7) *have the weak convergence*

$$\lim_{m\to+\infty} q(\cdot\,;\ q_m, u_m) = q(\cdot\,;\ q_0, u_0)$$

*in $L^2([T_0, T_1); W^{1,2}(X))$ and in the norms $\|\cdot\|_{C([T_0+\varepsilon,T_1)\times X)}$ for any $\varepsilon \in (0, T_1 - T_0)$.*

**Proof.** By assertion (a) of Theorem 1, for $m = 1, 2, \ldots$ there exists a unique controlled solution $q(t; q_m, u_m)$ on the half-open interval $[T_0, T_1)$. Since the sequence $\{\|q_m\|_{L^\infty(\mathbb{S}^2)}\}$ is bounded (see Remark 1), we obtain

$$0 \leqslant q(t; q_m, u_m) \leqslant Q\left(\sup_{(m=0,1,\ldots} \|q_m\|_{L^\infty(\mathbb{S}^2)}\right), \quad t \in [T_0, T_1), \quad m = 1, 2, \ldots.$$

Therefore, based on the a priori estimates for the solutions of linear second-order parabolic equations [8, ch. VI, §1] and [9, §1.5], there exists a constant $C_1$ such that

$$\|q(\cdot\,;\ q_m, u_m)\|_{L^2\left((T_0,T_1);W^{1,2}(X)\right)} \leqslant C_1, \quad m = 1, 2, \ldots; \tag{15}$$

by assertion (a) of Theorem 1, for $\varepsilon \in (0, T_1 - T_0)$ it is possible to find $C_2$ and $0 < \alpha \leqslant 1$ such that

$$\|q(\cdot\,;\ q_m, u_m)_{|[T_0+\varepsilon,T_1)\times X}\ \|_{C^\alpha([T_0+\varepsilon,T_1)\times X)} \leqslant C_2, \quad m = 1, 2, \ldots. \tag{16}$$

Due to (15) and the Eberlein–Šmulyan theorem (see [24, App. to Ch. V, §4], the sequence $\{q(\cdot\,; q_m, u_m)\}$ is sequentially weakly precompact in $L^2((T_0, T_1); W^{1,2}(X))$ because this space is reflexive [17, Ch. III, §1]. In turn, due to (16) and the Arzelá–Ascoli theorem, the sequence $\{q(\cdot\,; q_m, u_m)_{|[T_0+\varepsilon,T_1)\times X}\}$ is sequentially precompact in the norms $\|\cdot\|_{C([T_0+\varepsilon,T_1)\times X)}$. Hence, from $\{q(\cdot\,; q_m, u_m)\}$ it is possible to select a subsequence $\{q(\cdot\,; q_{m_l}, u_{m_l})\}$ that weakly converges to the limit function

$$\tilde{q}(t) = \lim_{m\to+\infty} q(t; q_{m_l}, u_{m_l}) \in L^\infty([T_0, T_1); L^\infty(X))$$

in $L^2((T_0, T_1); W^{1,2}(X))$ and in the norms $\|\cdot\|_{C([T_0+\varepsilon,T_1)\times X)}$ for any $\varepsilon \in (0, T_1 - T_0)$.

For $q_0 = q_{m_l}$ and $u = u_{m_l}$, the solution of problem (6), (7) is defined by

$$\langle q(\cdot\,; q_{m_l}, u_{m_l}), p\rangle(t) + \int_{T_0}^{t} \Big(\langle dq(\cdot\,; q_{m_l}, u_{m_l}), dp\rangle_{L^2(T^*\mathbb{S}^2)} - \langle q(\cdot\,; q_{m_l}, u_{m_l}), p'\rangle\Big)(\tau)d\tau$$

$$= \langle q_{m_l}, p\rangle(0) + \int_{T_0}^{t} \langle (A - u_{m_l})q(\cdot\,; q_{m_l}, u_{m_l}) - Bq^2(\cdot\,; q_{m_l}, u_{m_l}), p\rangle(\tau)d\tau.$$

Passing to the limit as $l \to +\infty$ [29, Ch. 1, §5] yields

$$\langle \tilde{q}, p\rangle(t) + \int\limits_{T_0}^{t} \left(\langle d\tilde{q}, dp\rangle_{L^2(T^*\mathbb{S}^2)} - \langle \tilde{q}, p'\rangle\right)(\tau)d\tau = \langle q_0, p(0)\rangle + \int\limits_{T_0}^{t} \langle (A - u_0)\tilde{q} - B\tilde{q}^2, p\rangle(\tau)d\tau,$$

i.e., the limit function $\tilde{q}$ is the solution of problem (6), (7) on $[T_0, T_1)$ with the initial value $q_0$ and the stationary control $u_0$. By assertion (a) of Theorem 1, the solution of problem (6), (7) is unique and, consequently, $\tilde{q}(t) = q(t; q_0, u_0)$. The proof of Lemma 2 is complete.

### *5.2. Proof of Assertions* (a) *and* (b)

Assertion (a) is a corollary of Lemma 1.

Assertion (b) is established by induction on $N = 1, 2, \ldots$. For $N = 1$, the desired result follows from Lemma 2 for $[T_0, T_1) = [0, T)$. Assume that it is true for $N \geqslant 1$. Then the sequence $\{q(NT-; q_m, u_m, s_m)\}$ converges to $q(NT-; q_0, u_0, s_0)$ in $\|\cdot\|_{C(X)}$; therefore, $\{s_m q(NT-; q_m, u_m, s_m)\}$ weakly converges to $s_0 q(NT-; q_0, u_0, s_0)$ in $L^2(X)$ [29, Ch. 1, §5]. By Lemma 2, for $[T_0, T_1) = [NT, (N+1)T)$, we arrive at the weak convergence

$$\lim_{m\to+\infty} q(\cdot\,;\, s_m q(NT-; q_m, u_m, s_m), u_m, s_m) = q(\cdot\,; s_0 q(NT-; q_0, u_0, s_0), u_0, s_0)$$

in $L^2((0,T); W^{1,2}(X))$ and in the norms $\|\cdot\|_{C([\varepsilon,T)\times X)}$ for any $\varepsilon \in (0, T)$. Thus, the desired result holds for $(N+1)$ as well, and the proof is complete.

### *5.3. Proof of Assertion* (c)

We choose an arbitrary number $r \in (0, +\infty)$ and consider the closed *function interval*

$$[0, Q(r)]_{L^\infty(X)} = \{w \in L^\infty(X) | 0 \leqslant w \leqslant Q(r) \text{a.e.}\},$$

where $Q$ is the function (11). By Lemma 1, the *Poincaré operator*

$$P^u_{[T_0,T_1)} : [0, Q(r)]_{L^\infty(X)} \ni w \mapsto q(T_1; w, u) \in C(X)$$

is well defined, where $q = q(t; w, u)$ is the solution of problem (6), (7) on the half-open interval $[T_0, +\infty)$ with the initial value $q_0 = w$ and the admissible control $u \in \mathfrak{U}$ (10); cf. [33, Ch. III, §21]. In addition,

$$0 = P^u_{[0,\frac{T}{2})}0, \quad P^u_{[0,\frac{T}{2})}Q(r) \leqslant Q(r), \quad 0 = P^u_{[\frac{T}{2},T)}0, \quad P^u_{[\frac{T}{2},T)}Q(r) \leqslant Q(r)$$

due to Remark 2 and the comparison principle for weak solutions [9, Sec. 2.1.2], and, consequently,

$$\begin{aligned} &P^u_{[0,\frac{T}{2})}([0, Q(r)]_{L^\infty(X)}) \subseteq [0, Q(r)]_{L^\infty(X)}, \\ &P^u_{[\frac{T}{2},T)}([0, Q(r)]_{L^\infty(X)}) \subseteq [0, Q(r)]_{L^\infty(X)}. \end{aligned} \tag{17}$$

For the admissible controls $s \in \mathfrak{S}$ (10), we have $0 \leqslant s \leqslant 1$ *a.e.*, therefore

$$s[0, Q(r)]_{L^\infty(X)} \subseteq [0, Q(r)]_{L^\infty(X)}. \tag{18}$$

Thus, the composition of the Poincaré operator and multiplication by $s$ is well-defined:

$$S : [0, Q(r)]_{L^\infty(X)} \ni v \mapsto P^u_{[0,\frac{T}{2})} s P^u_{[\frac{T}{2},T)} v \in [0, Q(r)]_{L^\infty(X)}. \tag{19}$$

Obviously, 0 is an *equilibrium* for $S$, i.e., $S(0) = 0$, whereas $Q(r)$ a *super-equilibrium*, i.e., $S(Q(r)) \leqslant Q(r)$ [33, Ch. I, §1]. According to assertion (a) and the Arzelá–Ascoli theorem, the operator $S$ is continuous and has a precompact image. By the comparison principle, $S$ strongly preserves order on $[0, Q(r)]_{L^\infty(X)}$ [33, Ch. I, §1]. Due to the strict concavity of the right-hand side of equation (6), the operator $S$ is strictly sublinear, i.e., $\beta S(v) < S(\beta v)$ for $v \in [0, Q(r)]_{L^\infty(X)} \backslash 0$ and $0 < \beta < 1$. Hence, for any $r \in (0, +\infty)$, $S$ has a unique fixed point $v_0 = Sv_0$ on the closed interval $[0, Q(r)]_{L^\infty(X)}$ such that

$$\lim_{k\to\infty} \|S^k(v) - v_0\|_{L^\infty(X)} = 0 \tag{20}$$

for any $v \in [0, Q(r)]_{L^\infty(X)} \backslash 0$ [33, Ch. I, §5]. In view of the inclusions (17) and (18), the function

$$q_{\infty,0} = sP^u_{[\frac{T}{2},T)} v_0 \in [0, Q(r)]_{L^\infty(X)}; \tag{21}$$

since equation (6) is *autonomous* (all its coefficients do not depend on $t$), it follows that

$$\begin{aligned} sP^u_{[0,T)} q_{\infty,0} &= sP^u_{[0,T)} \left( sP^u_{[\frac{T}{2},T)} v_0 \right) = sP^u_{[\frac{T}{2},T)} \left( P^u_{[0,\frac{T}{2})} sP^u_{[\frac{T}{2},T)} v_0 \right) \\ &= sP^u_{[\frac{T}{2},T)} S v_0 = sP^u_{[\frac{T}{2},T)} v_0 = q_{\infty,0}. \end{aligned}$$

Thus, $q_{\infty,0}$ is a fixed point of the operator $sP^u_{[0,T)}$. As $q_\infty(t; u, s)$ we choose the solution $q(t; q_{\infty,0}, u, s)$ of problem (6), (7), (8) with the initial value $q_{\infty,0}$ (21). By assertion (a), this solution exists, is unique, and satisfies the estimate $0 \leqslant q_\infty(t; u, s) \leqslant Q(r)$ on the half-open interval $[0, +\infty)$; moreover, it satisfies the periodicity condition (9) because $q_{\infty,0}$ is a fixed point with respect to the operator $sP^u_{[0,T)}$.

Let $q(t; q_0, u, s)$ be the solution of problem (6), (7), (8) with $q_0 \in [0, Q(r)]_{L^\infty(X)} \backslash 0$. Then

$$w(t) = \pm(q(t; q_0, u, s) - q_\infty(t; u, s))$$

satisfies the weak maximum principle on the half-open intervals $[kT, k(T + 1))$, $k = 1, 2, \ldots$ [8, Ch. VI, §7] and, consequently,

$$|q(t; q_0, u, s) - q_\infty(t; u, s)| \leqslant |q(kT; q_0, u, s) - q_\infty(kT; u, s)|, \quad t \in [kT, k(T + 1)).$$

Since $q(kT; q_0, u, s) = (sP^u_{[0,T)})^k q_0$ and $q_\infty(kT; u, s) = (sP^u_{[0,T)})^k q_{\infty,0}$, it follows that

$$\|q(t; q_0, u, s) - q_\infty(t; u, s)\|_{C(X)} \leqslant \left\| \left( sP^u_{[0,T)} \right)^k q_0 - \left( sP^u_{[0,T)} \right)^k q_{\infty,0} \right\|_{L^\infty(X)};$$

by the construction of $S$ (19) and the fixedness of $q_{\infty,0}$ (21) with respect to $sP^u_{[0,T)}$, we obtain

$$\|q(t; q_0, u, s) - q_\infty(t; u, s)\|_{C(X)} \leqslant \left\| sP^u_{[\frac{T}{2},T)} S^{k-1} P^u_{[0,\frac{T}{2})} q_0 - sP^u_{[\frac{T}{2},T)} v_0 \right\|_{L^\infty(X)}, \tag{22}$$

$t \in [kT, k(T + 1))$, because $(sP^u_{[0,T)})^k = sP^u_{[\frac{T}{2},T)} \left( P^u_{[0,\frac{T}{2})} sP^u_{[\frac{T}{2},T)} \right)^{k-1} P^u_{[0,\frac{T}{2})}$. According to (17),

$$P^u_{[0,\frac{T}{2})} q_0 \in P^u_{[0,\frac{T}{2})}([0, Q(r)]_{L^\infty(X)} \backslash 0) \subseteq [0, Q(r)]_{L^\infty(X)} \backslash 0;$$

due to (22), (20), the continuous operator $sP^u_{[\frac{T}{2},T)}$, and the arbitrary choice of $r \in (0, +\infty)$, we finally arrive at assertion (a).

### *5.4. Proof of Assertion* (d)

By Lemma 1, there exist constants $C$ and $\alpha$, $0 < \alpha \leqslant 1$, such that

$$\|q_\infty(T-; u_m, s_m)\|_{C^\alpha(X)} \leqslant C$$

uniformly in $m = 1, 2, \ldots$. According to the Arzelá–Ascoli theorem, it is therefore possible to select a subsequence $\{q_\infty(T-; u_{m_l}, s_{m_l})\}$ from $\{q_\infty(T-; u_m, s_m)\}$ that will converge in the norm $\|\cdot\|_{C(X)}$ to the limit function

$$q_T = \lim_{l\to+\infty} q_\infty(T-; u_{m_l}, s_{m_l}). \tag{23}$$

It suffices to establish the equality

$$q_T = q_\infty(T-; u_0, s_0) \tag{24}$$

regardless of the choice of $\{q_\infty(T-; u_{m_l}, s_{m_l})\}$. In this case, the entire sequence $\{q_\infty(T-; u_m, s_m)\}$ will converge to $q_\infty(T-; u_0, s_0)$ in the norm $\|\cdot\|_{C(X)}$ and $\{s_m q_\infty(T-; u_m, s_m)\}$ will weakly converge to $s_0 q_\infty(T-; u_0, s_0)$ in $L^2(X)$ [29, Ch. 1, §5]; in the final analysis, Lemma 2 will imply assertion (d) since $q_\infty$ satisfies conditions (8) and (9).

By conditions (8) and (9), $q_\infty(0; u_{m_l}, s_{m_l}) = s_{m_l} q_\infty(T-; u_{m_l}, s_{m_l})$; hence,

$$q_\infty(t; u_{m_l}, s_{m_l}) = q(t; s_{m_l} q_\infty(T-; u_{m_l}, s_{m_l}), u_{m_l}, s_{m_l}), \quad t \in [0, T),$$

as the solution of problem (6), (7) is unique (Lemma 1). Due to (23), the subsequence $\{s_{m_l} q_\infty(T-; u_{m_l}, s_{m_l})\}$ weakly converges in $L^2(X)$ to $s_0 q_T$. According to Lemma 2, passing to the limit on the right-hand side of this equality yields

$$q_T = q(T-; s_0 q_T, u_0, s_0).$$

Thus, the solution $q = q(t; s_0 q_T, u_0, s_0)$ satisfies the periodicity condition

$$q(0; s_0 q_T, u_0, s_0) = s_0 q(T-; s_0 q_T, u_0, s_0),$$

i.e., is a periodic solution of problem (6), (8). Hence, considering Remark 3, either $q_\infty(T-; u_0, s_0) = q(T-; s_0 q_T, u_0, s_0)$ (making (24) valid) or $q_\infty(T-; u_0, s_0) > 0$, $q(T-; s_0 q_T, u_0, s_0) = 0$, which is equivalent to the conditions

$$\|q_\infty(T-; u_0, s_0)\|_{C(X)} > 0, \quad q_T = 0. \tag{25}$$

We proceed by contradiction, showing that under conditions (25), the assertion

$$\lim_{\substack{k\to+\infty\\ l\to+\infty}} q(kT-; q_0, u_{m_l}, s_{m_l}) = 0, \quad q_0 > 0, \tag{26}$$

and its negation are simultaneously false; see items 1) and 2) below.

1) Assume that conditions (25) hold and assertion (26) is true.

By assertion (a), for $\varepsilon > 0$ there exists a natural number $k_0 = k_0(\varepsilon)$ such that

$$\|q(kT-; q_0, u_0, s_0) - q_\infty(T-; u_0, s_0)\|_{C(X)} < \varepsilon, \quad k = k_0, k_0 + 1, \ldots.$$

By assertion (b), for $\varepsilon > 0$ and $k = 1, 2, \ldots$ there exists a natural number $l_0 = l_0(\varepsilon, k)$ such that

$$\|q(kT-; q_0, u_{m_l}, s_{m_l}) - q(kT-; q_0, u_0, s_0)\|_{C(X)} < \varepsilon, \quad l_0, l_0 + 1, \ldots.$$

Consequently, based on

$$\begin{gathered}\|q(kT-;q_0,u_{m_l},s_{m_l})\|_{C(X)} \geqslant \|q_\infty(T-;u_0,s_0)\|_{C(X)}\\ -\|q(kT-;q_0,u_0,s_0)-q_\infty(T-;u_0,s_0)\|_{C(X)}\\ -\|q(kT-;q_0,u_{m_l},s_{m_l})-q(kT-;q_0,u_0,s_0)\|_{C(X)},\end{gathered}$$

for any $\varepsilon > 0$, $k = k_0(\varepsilon), k_0(\varepsilon)+1, \ldots$ and $l = l_0(\varepsilon,k), l_0(\varepsilon,k)+1, \ldots$, we have

$$\|q(kT-;q_0,u_{m_l},s_{m_l})\|_{C(X)} \geqslant \|q_\infty(T-;u_0,s_0)\|_{C(X)} - 2\varepsilon.$$

Choosing $\varepsilon = \frac{\|q_\infty(T-;u_0,s_0)\|_{C(X)}}{4} > 0$ in accordance with (25), we derive the estimate

$$\|q(kT-;q_0,u_{m_l},s_{m_l})\|_{C(X)} \geqslant \frac{\|q_\infty(T-;u_0,s_0)\|_{C(X)}}{2},$$

which contradicts assertion (26). Thus, conditions (25) and assertion (26) lead to a contradiction.

2) Assume that conditions (25) hold and assertion (26) is false. Then for some initial $q_0 > 0$ (7), there exists a number $\delta_0 > 0$ such that, for $N = 1, 2, \ldots,$ it is possible to find numbers $k_0 = k_0(N) \geqslant N$ and $l_0 = l_0(N) \geqslant N$ for which

$$\left\|q(k_0(N)T-;q_0,u_{m_{l_0(N)}},s_{m_{l_0(N)}})\right\|_{C(X)} \geqslant \delta_0. \tag{27}$$

Due to (23) and (25), for any $\varepsilon > 0$ there exists a number $l_1 = l_1(\varepsilon)$ such that

$$\|q_\infty(T-;u_{m_l},s_{m_l})\|_{C(X)} < \varepsilon, \quad l = l_1, l_1+1, \ldots.$$

By assertion (a), for $\varepsilon > 0$ and $l = 1, 2, \ldots$ there exists a natural number $k_1 = k_1(\varepsilon, l)$ such that

$$\|q(kT-;q_0,u_{m_l},s_{m_l}) - q_\infty(T-;u_{m_l},s_{m_l})\|_{C(X)} < \varepsilon, \quad k = k_1, k_1+1, \ldots.$$

Therefore, for $0 < \delta \leqslant \delta_0$ and $l = l_1(\frac{\delta}{2}), l_1(\frac{\delta}{2})+1, \ldots$ and $k = k_1(\frac{\delta}{2}, l), k_1(\frac{\delta}{2}, l)+1, \ldots,$ we have

$$\begin{gathered}\left\|q\left(k_0(N)T-;q_0,u_{m_{l_0(N)}},s_{m_{l_0(N)}}\right)\right\|_{C(X)} \|q(kT-;q_0,u_{m_l},s_{m_l})\|_{C(X)}\\ \leqslant \|q(kT-;q_0,u_{m_l},s_{m_l}) - q_\infty(T-;q_0,u_{m_l},s_{m_l})\|_{C(X)}\\ + \|q_\infty(T-;q_0,u_{m_l},s_{m_l})\|_{C(X)} < \delta.\end{gathered} \tag{28}$$

From (27) and (28) it follows that, for $N = l_1(\frac{\delta}{2}), l_1(\frac{\delta}{2})+1, \ldots$ there exists a number

$$k_2 = k_2(\delta, k_0(N), l_0(N)) \in \left\{k_0(N), \ldots, k_1\left(\frac{\delta}{2}, l_0(N)\right) - 1\right\}$$

for which

$$\begin{gathered}\left\|q\left(k_2(\delta,k_0(N),l_0(N))T-;q_0,u_{m_{l_0(N)}},s_{m_{l_0(N)}}\right)\right\|_{C(X)} \geqslant \delta,\\ \left\|q\left((k_2(\delta,k_0(N),l_0(N))+k)T-;q_0,u_{m_{l_0(N)}},s_{m_{l_0(N)}}\right)\right\|_{C(X)} < \delta, \quad k = 1, 2, \ldots.\end{gathered} \tag{29}$$

Consider the sequence $\{q_N\}$ composed of

$$q_N = q(k_2(\delta,k_0(N),l_0(N))T-;q_0,u_{m_{l_0(N)}},s_{m_{l_0(N)}}), \quad N = l_1\left(\frac{\delta}{2}\right), l_1\left(\frac{\delta}{2}\right)+1, \ldots.$$

By assertion (a), there exist constants $C$ and $\alpha$, $0 < \alpha \leqslant 1$, such that $\|q_N\|_{C^\alpha(X)} \leqslant C$ uniformly in $N$.

Based on the Arzelá–Ascoli theorem, we select a subsequence $\{q_{N_\beta}\}$ from the sequence $\{q_N\}$ that converges in the norm $\|\cdot\|_{C(X)}$ to the limit function

$$q_{0,\infty} = \lim_{\beta\to+\infty} q_{N_\beta}.$$

In view of the first inequality in (29), $\|q_{0,\infty}\|_{C(X)} \geqslant \delta$. Since the sequence $\{s_{m_{l_0(N_\beta)}} q_{N_\beta}\}$ weakly converges in $L^2(X)$ to $s_0 q_{0,\infty}$ [29, Ch. 1, §5], by assertion (b), for an arbitrary number $\varepsilon > 0$ and $k = 1, 2, \ldots$ there exists $\beta_0 = \beta_0(\varepsilon, k)$ such that

$$\left\| q\Big(kT-; s_{m_{l_0(N_\beta)}} q_{N_\beta}, u_{m_{l_0(N_\beta)}} q_{N_\beta}, s_{m_{l_0(N_\beta)}}\Big) - q(kT-; s_0 q_{0,\infty}, u_0, s_0) \right\|_{C(X)} < \varepsilon \tag{30}$$

for $\beta = \beta_0, \beta_0 + 1, \ldots$. By assertion (c), for $\varepsilon > 0$ there exists $k_3 = k_3(\varepsilon)$ such that

$$\|q(kT-; s_0 q_{0,\infty}, u_0, s_0) - q_\infty(T-; u_0, s_0)\|_{C(X)} < \varepsilon, \quad k = k_3, k_3 + 1, \ldots. \tag{31}$$

Hence, the following results are the case. First, since

$$\begin{gathered}
\left\| q\left(kT-; s_{m_{l_0(N_\beta)}} q_{N_\beta}, u_{m_{l_0(N_\beta)}}, s_{m_{l_0(N_\beta)}}\right) \right\|_{C(X)} \geqslant \|q_\infty(T-; u_0, s_0)\|_{C(X)} \\
- \|q(kT-; s_0 q_{0,\infty}, u_0, s_0) - q_\infty(T-; u_0, s_0)\|_{C(X)} \\
- \left\| q\Big(kT-; s_{m_{l_0(N_\beta)}} q_{N_\beta}, u_{m_{l_0(N_\beta)}}, s_{m_{l_0(N_\beta)}}\Big) - q(kT-; s_0 q_{0,\infty}, u_0, s_0) \right\|_{C(X)},
\end{gathered}$$

considering (31) and (30), for $k \geqslant k_3(\varepsilon)$ and $\beta \geqslant \beta_0(\varepsilon, k)$ we have

$$\left\| q\left(kT-; s_{m_{l_0(N_\beta)}} q_{N_\beta}, u_{m_{l_0(N_\beta)}}, s_{m_{l_0(N_\beta)}}\right) \right\|_{C(X)} \geqslant \|q_\infty(T-; u_0, s_0)\|_{C(X)} - 2\varepsilon. \tag{32}$$

Second, by the construction of $q_N$ and the autonomous property of equation (6),

$$q\left(kT-; s_{m_{l_0(N_\beta)}} q_{N_\beta}, u_{m_{l_0(N_\beta)}}, s_{m_{l_0(N_\beta)}}\right) = q\left((k_2(\delta, k_0(N), l_0(N)) + k)T-; q_0, u_{m_{l_0(N)}}, s_{m_{l_0(N)}}\right),$$

and the second inequality in (29) gives

$$\left\| q\left(kT-; s_{m_{l_0(N_\beta)}} q_{N_\beta}, u_{m_{l_0(N_\beta)}}, s_{m_{l_0(N_\beta)}}\right) \right\|_{C(X)} < \delta, \quad k = 1, 2, \ldots. \tag{33}$$

From (32) and (33), for $k = k_3(\varepsilon)$ and $\beta = \beta_0(\varepsilon, k_3(\varepsilon))$, we derive the inequality

$$\|q_\infty(T-; u_0, s_0)\|_{C(X)} < \delta + 2\varepsilon.$$

With

$$\varepsilon = \frac{\|q_\infty(T-; u_0, s_0)\|_{C(X)}}{4} \quad \text{and} \quad \delta = \min\left\{ \frac{\|q_\infty(T-; u_0, s_0)\|_{C(X)}}{4}, \delta_0 \right\},$$

the first condition in (25) leads to the contradictory estimate $\|q_\infty(T-; u_0, s_0)\|_{C(X)} < 0$. Thus, conditions (25) and the negation of assertion (26) bring to a contradiction as well.

## FUNDING

This work was supported by the Russian Science Foundation, project no. 19-11-00223.

## ACKNOWLEDGMENTS

The author is grateful to A.A. Davydov for the problem statement and useful discussions.

## REFERENCES

1. Verhulst, P.F., Notice sur la loi que la population poursuit dans son accroissement, *Correspondance Math. Phys.*, 1838, no. 10, pp. 113–121.
2. Fourier, J.B.J., *Theorie Analytique de la Chaleur*, Paris: F. Didot, 1822.
3. Kolmogorov, A.N., Petrovskii, I.G., and Piskunov, N.S., The Investigation of a Diffusion Equation Connected with an Increasing Amount of Substance and Its Application to a Biological Problem, *Byull. Mosk. Gos. Univ. Mat. Mekh.*, 1937, vol. 1, no. 6, pp. 1–26. (In Russian.) Kolmogoroff, A., Petrovsky, I., Piscounoff, N., Étude de l'éequation de la diffusion avec croissance de la quantite de matière et son application à un problème biologique, *Moscou Univ. Bull.Math.*, 1937, vol. 1(6), pp. 1–25.
4. Fisher, R.A., The Advance of Advantageous Genes, *Ann. Eugenics*, 1937, vol. 7, pp. 335–369.
5. Berestycki, H., Francois, H., and Roques, L., Analysis of the Periodically Fragmented Environment Model: I – Species Persistence, *J. Math. Biol.*, 2005, vol. 51, pp. 75–113.
6. Pethame, B., *Parabolic Equations in Biology*, Heidelberg: Springer, 2015.
7. Ladyzhenskaya, O.A., Solonnikov, V.A., and Ural'tseva, N.N., *Linear and Quasilinear Equations of Parabolic Type*, Providence: American Mathematical Society, 1968.
8. Lieberman, G.M., *Second Order Parabolic Differential Equations*, New Jersey: World Scientific, 2005.
9. Wang, M., *Nonlinear Second Order Parabolic Equations*, Boca Raton: CRC Press, 2021.
10. Berestycki, H., Francois, H., and Roques, L., Analysis of the Periodically Fragmented Environment Model: II – Biological Invasions and Pulsating Travelling Fronts, *J. Math. Pures Appl.*, 2005, vol. 84, pp. 1101–1146.
11. Postnikov, M.M., *Smooth Manifolds*, Moscow: Mir Publishers, 1989.
12. Hebbey, E., *Sobolev Spaces on Riemannian Manifolds*, Berlin: Springer, 1996.
13. Nicolaescu, L.I., *Lectures on the Geometry of Manifolds*, New Jersey: World Scientific, 2021.
14. Tunitsky, D.V., On the Construction of Solutions of Semilinear Second-Order Elliptic Equations on Closed Manifolds, *Trudy 14-oi Mezhdunarodnoi konferentsii "Upravlenie razvitiem krupnomasshtabnykh sistem"* (Proceedings of the 14th International Conference on Management of Large-Scale System Development (MLSD'2021)), September 27–29, 2021, Moscow: Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 2021, pp. 717–723.
15. Tunitsky, D.V., On Solvability of Second-Order Semilinear Elliptic Equations on Spheres, *Proceedings of the 14th International Conference on Management of Large-Scale System Development (MLSD)*, September 27–29, 2021, Moscow, Russia. IEEE Explore, November 22, 2021, pp. 1–4. https://ieeexplore.ieee.org/document/9600203
16. Tunitsky, D.V., On Solvability of Semilinear Second-Order Elliptic Equations on Closed Manifolds, *Izv. Math.*, 2022, vol. 86, no. 5, pp. 925–942.
17. Showalter, R.E., *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, Providence: American Mathematical Society, 1997.
18. Lions, J.L., *Equations differentielles operationnelles et problemes aux limites*, Berlin: Springer-Verlag, 1961.
19. Palais, R.S., *Seminar on the Atiyah–Singer Index Theorem*, Princeton, NJ: Princeton Univ. Press, 1965.
20. Wells, R.O., *Differential Analysis on Complex Manifolds*, New York: Springer, 2008.

21. Davydov, A.A. and Melnik, D.A., Optimal States of Distributed Exploited Populations with Periodic Impulse Harvesting, *Proc. Steklov Inst. Math.*, 2021, vol. 315, suppl. 1, pp. S1–S8.

22. Koopman, B.O., The Theory of Search. III. The Optimum Distribution of Search Effort, *Oper. Res.*, 1957, no. 5, pp. 613–626.

23. Zhikov, V.V., Mathematical Problems of Search Theory, in *Proceedings of the Vladimir Polytechnic Institute*, Moscow: Vysshaya Shkola, 1968, pp. 263–270.

24. Yosida, K., *Functional Analysis*, 6th ed., Berlin–Heidelberg–New York: Springer-Verlag, 1980.

25. Hille, E. and Phillips, R.S., *Functional Analysis and Semi-Groups*, 2nd ed., American Mathematical Society, 1957.

26. Arnold, V.I., Optimization in Mean and Phase Transitions in Controlled Dynamical Systems, *Functional Analysis and Its Applications*, 2002, vol. 36, pp. 83–92.

27. Davydov, A. and Vinnikov, E., Optimal Cyclic Dynamic of Distributed Population under Permanent and Impulse Harvesting, in *Springer Proceedings in Mathematics & Statistics*, 2023, vol. 407, pp. 101–112.

28. Vinnikov, E.V., Davydov, A.A., and Tunitsky, D.V., Existence of a Maximum of Time-Averaged Harvesting in the KPP Model on Sphere with Permanent and Impulse Harvesting, *Dokl. Math.*, 2023, vol. 108, pp. 472–476.

29. Gajewski, H., Greger, K., and Zacharias, K., *Nichtlineare Operator Gleichungen und Operator Differential Gleichungen*, Berlin: Akademie-Verlag, 1974.

30. Tunitsky, D.V., On Weak Solutions of Semilinear Second-Order Parabolic Equations on Closed Manifolds, *Trudy 15-oi Mezhdunarodnoi konferentsii "Upravlenie razvitiem krupnomasshtabnykh sistem"* (Proceedings of the 15th International Conference on Management of Large-Scale System Development (MLSD'2022)), September 26–28, 2022, Moscow: Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 2022, pp. 613–619.

31. Tunitsky, D.V., On Initial Value Problem for Semilinear Second Order Parabolic Equations on Spheres, *Proceedings of the 15th International Conference on Management of Large-Scale System Development (MLSD)*, September 26–28, 2022, Moscow, Russia. IEEE Explore, November 9, 2022, pp. 1–4. https://ieeexplore.ieee.org/document/9934193

32. Tunitsky, D.V., On Stabilization of Solutions of Second-Order Semilinear Parabolic Equations on Closed Manifolds, *Izv. Math.*, 2023, vol. 87, no. 4, pp. 817–834.

33. Hess, P., *Periodic-Parabolic Boundary Value Problems and Positivity*, New York: Pitman Research Notes in Math. Series, 1991.

*This paper was recommended for publication by A.G. Kushner, a member of the Editorial Board*

*ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 7, pp. 701–710.*
*© The Author(s), 2024 published by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 2024.*
*Russian Text © The Author(s), 2024, published in Avtomatika i Telemekhanika, 2024, No. 7, pp. 61–72.*

# Synthesis of Itô Equations for a Shaping Filter with a Given Spectrum

**M. M. Khrustalev*† and D. S. Rumyantsev*,[a]**

*Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
*e-mail: [a] n3030@mail.ru*

Received August 6, 2023
Revised January 15, 2024
Accepted January 20, 2024

**Abstract**—The analytical method for the synthesis of a generator of a random process with a given spectrum in the form of a linear system of Ito's equations is proposed. The stationarity of a random process is assumed, the spectral and corresponding transfer functions of which are defined in the form of rational fractions. The coefficients of the system of Ito's equations of the generator are found from recurrent algebraic relations. The method is focused on working with mathematical models of nature random processes, such as the Dryden's wind model. The transformation of the spectra of the wind gust model in three directions is presented in detail and the corresponding stochastic equations are given.

*Keywords*: Ito's stochastic differential equation, spectral density, transfer function, shaping filter, random disturbance generator, Dryden wind turbulence model

**DOI:** 10.31857/S0005117924070059

## 1. INTRODUCTION

The shaping filter allows you to generate a random signal with a given spectral density from a white noise signal [1, Section 6.6; 2, Section 10.1; 3, Section 5.1.5]. The shaping filter and the analyzed system form some extended system, the input of which is affected by white noise (Fig. 1). This shows a way to move from representing a system in terms of transfer functions (shown in the diagram) to stochastic differential equations. The results of the article will be useful to researchers for adding random factors to a dynamic model and simulating natural phenomena (movement of air masses, water flow, etc.).

There are many known models of wind gusts [4], but in the article only the Dryden [5] turbulence model is considered in detail, which at the output gives a stochastic process determined by velocity spectra. The spectral density of the signal is an even fractional-rational function of frequency and can be represented in the form of two complex conjugate factors, from which the transfer function of the shaping filter is found [1, Section 6.6; 3, Section 5.1.5].

Trying to directly write a high-order differential equation whose output has a given spectrum usually results in high-order white noise derivatives. The representation of these derivatives in the form of generalized functions [6] and the generalization of the Itô equations in the form of Leontief-type equations [6] are known, but such equations are complicated and little studied. In [3, Section 3.3.3] the transition from a linear stochastic differential equation of higher order to a linear system of stochastic differential equations of first order is considered, but to find the coefficients of the system it is necessary to differentiate the coefficients of the original equation (if it is not stationary). The proposed method allows us to describe a natural random process using the well-studied Ito

† Deceased.

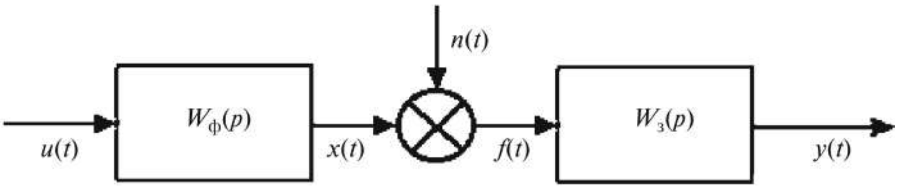


**Fig. 1.** Connection diagram for the shaping filter.

equations [3]. The resulting equations, for example, can be used in conjunction with the equations of the mathematical model of the aircraft [7].

A method is proposed for obtaining relatively simple stochastic differential equations for synthesizing the output signal using a known transfer function. Next, we will consider transfer functions under the assumption that the corresponding spectra are known.

Two ways of transforming any fractional-rational transfer function, leading to the same result, are presented below. The function is decomposed into a sum of fractions, the numerators of which are real numbers, and the denominators of which are polynomials. In this case, all operations are arithmetic. And the process flow diagram can be depicted as a sum of integrating links. Based on the new notation, it is possible to construct a system of linear stochastic differential Ito equations.

There is a slightly more complex way of similar transformation of the transfer function [8, Section 2.3]. If the transfer function $W(p)$ is given, then for the corresponding system of linear equations of the form

$$\begin{gathered} \dot{x} = Ax + bu \\ y = cx \end{gathered} \tag{1}$$

it is necessary to find the matrix $A$ and the vector $b$. The output vector $c$ is given. The first Frobenius form of the state equation matrix $A$ is selected so that its characteristic polynomial coincides with the denominator of the transfer function. The elements of the vector $b$ are found from solving the system of equations

$$W(p) = c(Ep - A)^{-1}b$$

by the method of indefinite coefficients by equating factors with equal powers of the variable $p$ of polynomials of numerators on the left and right [8, Example 2.7].

In the proposed approach, the output vector $c$ is in the process of being solved and is not known in advance. As a result, matrix inversion is not required and only the coefficients of the Ito equation are calculated using arithmetic operations.

The transformation to obtain the equation (1) is not unique [8]. Therefore, it is not always possible to achieve "minimal implementation" (1), i.e., obtain the minimum possible number of variables in the Ito equation.

## 2. MATHEMATICAL PROBLEM STATEMENT

Let the spectral density of the disturbance under study be defined as $\Phi(\omega) = |W(i\omega)|^2$, here

$$W(p) = \frac{P_m(p)}{Q_n(p)} = \frac{a_0p^m + a_1p^{m-1} + \ldots + a_{m-1}p + a_m}{b_0p^n + b_1p^{n-1} + \ldots + b_{n-1}p + b_n}, \tag{2}$$

and $a_i$ $(i = \overline{0,m})$, $b_j$ $(j = \overline{0,n})$ are constant real coefficients. The poles and zeros of the function $W(p)$ are located in the left half-plane. $W(p)$ is the transfer function of the linear differential

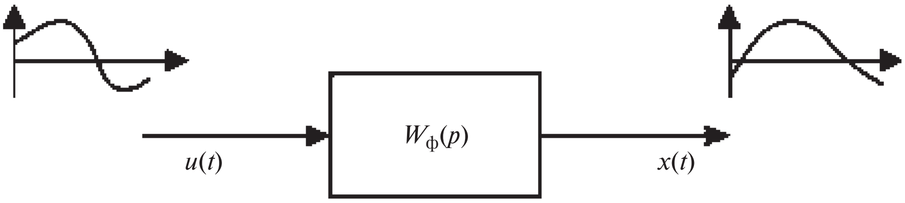


**Fig. 2.** Block diagram of a shaping filter.

equation

$$b_0x^{(n)} + b_1x^{(n-1)} + \ldots + b_{n-1}x' + b_nx \\ = a_0u^{(m)} + a_1u^{(m-1)} + \ldots + a_mu. \tag{3}$$

The block diagram of the shaping filter is shown in Fig. 2. If standard white noise (the derivative of the standard Wiener process) $u(t)$ is supplied to the input, then the equation (3) becomes stochastic, but in the general case contains higher derivatives of white noise.

The goal is to replace the equation with a system of linear differential equations that satisfies two conditions: a) the system does not contain derivatives of the input signal, b) its output, linearly dependent on the state, coincides with the output of the equation. Such a system of equations, as shown below, can easily be converted into the Ito system of equations.

It turns out that for the transformation it is enough to represent the transfer function (2) as a sum of rational fractions, the numerators of which are real coefficients (zero-order polynomials). In the case when all the zeros of the denominator of a rational fraction are real, such a transformation is known [9], but requires finding the zeros of the denominator, which in the general case is only possible numerically. The proposed transformation does not require finding zeros and for any proper rational fraction with both real and complex zeros of the denominator, it can be performed analytically. The coefficients of the modified transfer function are found sequentially from a chain of linear equations.

A method for transforming a $n$ order linear stochastic equation (3) to an equivalent first order linear system of equations not containing white noise derivatives is shown in [3, Section 3.3.3]. In this case, the equation is not stationary and the coefficients $a_i$ $(i = \overline{0,m})$, $b_j$ $(j = \overline{0,n})$ depend on time $t$, and to find the coefficients of an equivalent system, it is necessary to differentiate the functions $a_i$, $b_j$. For a stationary system, new coefficients are found from recurrent arithmetic relations.

## 3. TRANSFER FUNCTION CONVERSION

The main idea is to represent the original function $W(p)$ as the sum of several functions $W_i(p)$, $i = \overline{1, m+n}$, see Fig. 3. The input and output signals will not change as a result of this conversion.

The proposed representation of the transfer function (2) has the form

$$W(p) = \frac{P_m(p)}{Q_n(p)} = \frac{\alpha_1}{p^{n-m}} + \frac{\alpha_2}{p^{n-m+1}} + \ldots + \frac{\alpha_m}{p^{n-1}} \\ + \frac{1}{Q_n(p)}\left(\frac{\beta_{n-1}}{p^{n-1}} + \frac{\beta_{n-2}}{p^{n-2}} + \ldots + \frac{\beta_1}{p} + \beta_0\right). \tag{4}$$

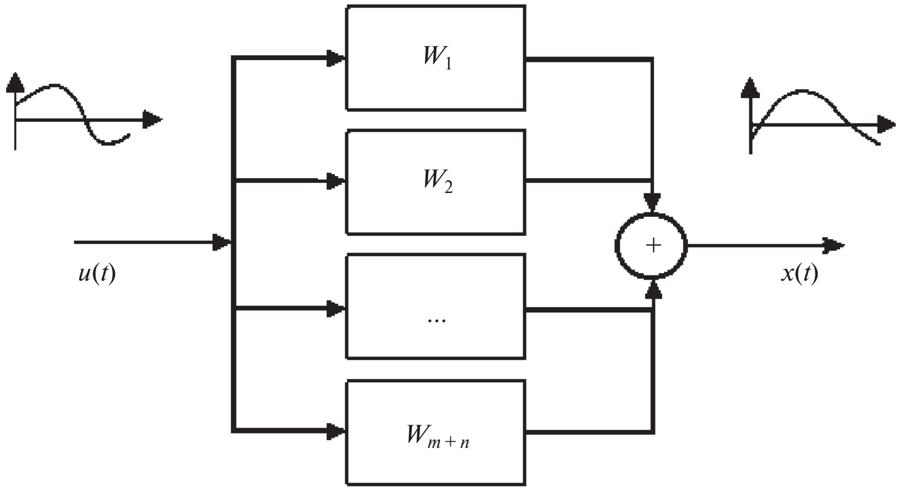


**Fig. 3.** Block diagram of the sum of several shaping filters.

The number of coefficients $\alpha_i$ $(i = \overline{1, m})$, $\beta_j$ $(j = \overline{0, n-1})$ is $m + n$. They can be obtained by equating the left and right sides of the equations (2) and (4).

$$\begin{array}{lll}
\beta_{n-1} & + \alpha_m b_n & = 0, \\
\beta_{n-2} & + \alpha_m b_{n-1} + \alpha_{m-1} b_n & = 0, \\
\beta_{n-3} & + \alpha_m b_{n-2} + \alpha_{m-1} b_{n-1} + \alpha_{m-2} b_n & = 0, \\
\ldots & & \\
\beta_{n-m+1} & + \alpha_m b_{n-m+2} + \ldots \quad + \alpha_2 b_n & = 0, \\
\beta_{n-m} & + \alpha_m b_{n-m+1} + \ldots \quad + \alpha_2 b_{n-1} + \alpha_1 b_n & = 0, \\
\beta_{n-m-1} & + \alpha_m b_{n-m} \quad + \ldots \quad + \alpha_2 b_{n-2} + \alpha_1 b_{n-1} & = 0, \\
\ldots & & \\
\beta_2 & + \alpha_m b_3 + \alpha_{m-1} b_4 + \ldots + \alpha_2 b_{m+1} + \alpha_1 b_{m+2} & = 0, \\
\beta_1 & + \alpha_m b_2 + \alpha_{m-1} b_3 + \ldots \quad + \alpha_1 b_{m+1} & = 0, \\
\beta_0 & + \alpha_m b_1 + \alpha_{m-1} b_2 + \ldots \quad + \alpha_1 b_m & = a_m, \\
 & \alpha_m b_0 + \alpha_{m-1} b_1 + \ldots \quad + \alpha_1 b_{m-1} & = a_{m-1}, \\
 & \alpha_{m-1} b_0 + \ldots \quad + \alpha_1 b_{m-2} & = a_{m-2}, \\
\ldots & & \\
 & \alpha_2 b_0 \quad + \alpha_1 b_1 & = a_1, \\
 & \alpha_1 b_0 & = a_0.
\end{array}$$

Let us write a short form, which is a system of recurrent equations, with the help of which the coefficients can be calculated sequentially:

$$\alpha_1 = \frac{a_0}{b_0}, \quad \alpha_k = \frac{1}{b_0}\left[a_{k-1} - \sum_{s=1}^{k-1} \alpha_s b_{k-s}\right], \quad k = \overline{2, m},$$

$$\beta_0 = a_m - \sum_{s=1}^{m} \alpha_s b_{m-s+1},$$

$$\beta_k = -\sum_{s=1}^{m} \alpha_s b_{m+k-s+1}, \quad k = \overline{1, n-m},$$

$$\beta_k = -\sum_{s=1}^{n-k} \alpha_{-n+m+k+s} b_{n-s+1}, \quad k = \overline{n-m+1, n-1}.$$

There is another way to converse the transfer function. Let the transfer function have the form

$$W^{(r)}(p) = \frac{a_0^{(r)}p^{n-r} + a_1^{(r)}p^{n-r-1} + \ldots + a_{n-r-1}^{(r)}p + a_{n-r}^{(r)}}{b_0p^n + b_1p^{n-1} + \ldots + b_{n-1}p + b_n}, \tag{5}$$

$1 \leqslant r \leqslant n$. The superscript $(r)$ indicates the function number and the degree of the numerator polynomial. The degree of the numerator polynomial $W^{(r+1)}(p)$ is less than the degree of the numerator polynomial $W^{(r)}(p)$, since $n-(r+1) < n-r$. Let us denote $B(p) = b_0p^n + b_1p^{n-1} + \ldots + b_{n-1}p + b_n$ and perform a series of conversions of the function $W^{(r)}(p)$, consisting of successively reducing the degree of the numerator polynomial to zero.

$$\begin{aligned}
W^{(r)}(p) &= \frac{a_0^{(r)}p^n + a_1^{(r)}p^{n-1} + \ldots + a_{n-r}^{(r)}p^r}{p^rB(p)} \\
&= \frac{1}{p^rB(p)}\left[\frac{a_0^{(r)}}{b_0}B(p) - \frac{a_0^{(r)}}{b_0}\left(b_1p^{n-1} + \ldots + b_{n-1}p + b_n\right)\right. \\
&\qquad \left. + \left(a_1^{(r)}p^{n-1} + a_2^{(r)}p^{n-2} + \ldots + a_{n-r}^{(r)}p^r\right)\right] \\
&= \frac{a_0^{(r)}}{b_0}\frac{1}{p^r} + \frac{1}{B(p)}\left[\left(a_1^{(r)} - \frac{a_0^{(r)}}{b_0}b_1\right)p^{n-r-1} + \ldots\right. \\
&\qquad \left. + \left(a_{n-r-1}^{(r)} - \frac{a_0^{(r)}}{b_0}b_{n-r-1}\right)p + \left(a_{n-r}^{(r)} - \frac{a_0^{(r)}}{b_0}b_{n-r}\right)\right] \\
&\qquad - \frac{a_0^{(r)}}{b_0}\frac{1}{B(p)}\left[\frac{b_{n-r+1}}{p} + \frac{b_{n-r+2}}{p^2} + \ldots + \frac{b_n}{p^r}\right].
\end{aligned}$$

Let's determine the coefficients $a_\alpha^{(r+1)} = a_{1+\alpha}^{(r)} - (a_0^{(r)}/b_0)b_{1+\alpha}$, $\alpha = \overline{0, n-r-1}$, for the new function $W^{(r+1)}$. Then

$$W^{(r)}(p) = \frac{a_0^{(r)}}{b_0}\frac{1}{p^r} + W^{(r+1)}(p) - \frac{a_0^{(r)}}{b_0}\frac{1}{B(p)}\sum_{k=1}^{r}\frac{b_{n-r+k}}{p^k}, \tag{6}$$

$$\text{and} \quad W^{(r+1)}(p) = W^{(r)}(p) - \frac{a_0^{(r)}}{b_0}\frac{1}{p^r} + \frac{a_0^{(r)}}{b_0}\frac{1}{B(p)}\sum_{k=1}^{r}\frac{b_{n-r+k}}{p^k}.$$

The maximum number of steps is $n-r$. The function $W^{(r)}(p)$ is defined at the $r$th step, it is necessary to find $W^{(r+1)}(p)$, $W^{(r+2)}(p)$, .... The numerator of the last function $W^{(r+s)}(p)$ is a zeroth order polynomial, and then the calculation will be completed. Each next found function $W^{(r+k+1)}(p)$ must be substituted into the current function $W^{(r+k)}(p)$.

## 4. CREATING A RANDOM DISTURBANCE GENERATOR

Let's consider the transfer function (4), which is the sum of integrating links with their own gain factors [10]. For link $1/p^{n-m}$ the corresponding equation will be $x_1 = u/p^{n-m}$, or $x_1^{(n-m)} = u$. The link $1/p^{n-m+1}$ will give the equation $x_2 = u/p^{n-m+1} = u/(p^{n-m}p) = u/p^{n-m} \times 1/p = x_1 \times 1/p$, or $x_2' = x_1$. This is how differential equations for the first $m$ outputs are successively found. In the same way, using the denominator $Q_n(p)$, we obtain the output equation $x_{m+1}$, which then needs to be integrated another $n-1$ times using the terms in brackets from (4). The last step will be the summation of all outputs with the corresponding coefficients $\alpha$, $\beta$. Let us write down the system of differential equations and the output equation corresponding to the transfer function (4) and (3):

$$
\begin{gathered}
x_1^{(n-m)} = u, \quad x_2' = x_1, \quad x_3' = x_2, \quad \ldots \quad x_m' = x_{m-1}, \\
b_0 x_{m+1}^{(n)} + b_1 x_{m+1}^{(n-1)} + \ldots + b_{n-1} x_{m+1}' + b_n x_{m+1} = u, \\
x_{m+2}' = x_{m+1}, \quad \ldots \quad x_{m+n}' = x_{m+n-1}, \\
x = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_m x_m + \beta_0 x_{m+1} + \beta_1 x_{m+2} + \ldots + \beta_{n-1} x_{m+n}.
\end{gathered} \tag{7}
$$

The order of the resulting system is $N \leqslant 3n - 2$.

The system (3) can be easily written as a system of first-order equations (7) and, assuming that $u(t)$ is standard white noise, transformed into a system of Ito equations $N$th order.

The transformation does not change the transfer function $W(p)$ for the output $x = \alpha_1 x_1 + \alpha_2 x_2 + \ldots$ (see (7)). Nevertheless, the transfer functions of the outputs $x_1$, $x_2$, $\ldots$ have a certain number of zero poles. Thus, $W_k(p)$ for $x_k$, $k = \overline{1, m}$ has the form $W_k(p) = 1/p^{n-m+k-1}$. In practice, this will lead to instability in process modeling (due to calculation errors). However, this situation can be corrected by making the replacement $p = (q - \Delta)/\lambda$ $(q = \lambda p + \Delta)$, $\Delta/\lambda > 0$ in the original transfer function $W(p)$. Then

$$
\begin{gathered}
W^*(q) = \frac{P_m^*(q)}{Q_n^*(q)} = \frac{P_m\left(\frac{q-\Delta}{\lambda}\right)}{Q_n\left(\frac{q-\Delta}{\lambda}\right)} = \frac{\alpha_1^*}{q^{n-m}} + \frac{\alpha_2^*}{q^{n-m+1}} + \ldots \\
+ \frac{\alpha_m^*}{q^{n-1}} + \frac{1}{Q_n^*(q)} \left( \frac{\beta_{n-1}^*}{q^{n-1}} + \frac{\beta_{n-2}^*}{q^{n-2}} + \ldots + \frac{\beta_1^*}{q} + \beta_0^* \right).
\end{gathered}
$$

By inverse transformation $q = \lambda p + \Delta$ we get

$$
W(p) = \frac{P_m(p)}{Q_n(p)} = \frac{\alpha_1^*}{(\lambda p + \Delta)^{n-m}} + \frac{\alpha_2^*}{(\lambda p + \Delta)^{n-m+1}} + \ldots.
$$

The transfer functions of the output components $x_1$, $x_2$, $\ldots$ of such an expansion will have poles in the left half-plane.

## 5. EXAMPLE

Let the transfer function be given

$$
W(p) = \frac{p^2 + 2p + 1}{p^3 + 3p^2 + 2p + 2}.
$$

It is required to write it in the form of a sum of fractions with zero-order polynomials in the numerators.

As a result of the transformation we get

$$
W^*(q) = W(q-1) = \frac{(q-1)^2 + 2(q-1) + 1}{(q-1)^3 + 3(q-1)^2 + 2(q-1) + 2} = \frac{q^2}{q^3 - q + 2}.
$$

Here $n = 3$, $m = 2$, $a_0^* = 1$, $a_1^* = 0$, $a_2^* = 0$, $b_0^* = 1$, $b_1^* = 0$, $b_2^* = -1$, $b_3^* = 2$, $\alpha_1^* = \frac{a_0^*}{b_0^*} = 1$, $\alpha_2^* = \frac{1}{b_0^*}[a_1^* - \alpha_1^* b_1^*] = 0$, $\beta_0^* = a_2^* - [\alpha_1^* b_2^* + \alpha_2^* b_1^*] = 1$, $\beta_1^* = -[\alpha_1^* b_3^* + \alpha_2^* b_2^*] = -2$, $\beta_2^* = -[\alpha_2^* b_3^*] = 0$.

$$
\begin{gathered}
W^*(q) = \frac{1}{q} + \left(1 - \frac{2}{q}\right) \frac{1}{q^3 - q + 2}, \\
W(p) = \frac{1}{p+1} + \left(1 - \frac{2}{p+1}\right) \frac{1}{p^3 + 3p^2 + 2p + 2}.
\end{gathered}
$$

Let's write the solution in the second way

$$W^{(1)}(q) = \frac{q^2}{q^3 - q + 2} = \frac{q^2}{B(q)}, \quad B(q) = q^3 - q + 2,$$
$$W^{(2)}(q) = \frac{q^2}{B(q)} - \frac{1}{q} + \frac{1}{B(q)} \times \frac{2}{q} = \frac{q^3 - B(q) + 2}{qB(q)} = \frac{1}{B(q)}.$$

We use (6)

$$W^{(1)}(q) = \frac{1}{q} + \frac{1}{B(q)} - \frac{1}{B(q)}\frac{2}{q} = \frac{1}{q} + \left(1 - \frac{2}{q}\right)\frac{1}{B(q)} = W^*(q).$$

The results are the same.

Let us write down the derivation of the Ito equations for $W(p)$ in accordance with (7). The first term gives the equation $u = (p+1)x_1$. The second equation will be $u = (p^3 + 3p^2 + 2p + 2)x_2 = ((p+1)^3 - (p+1) + 2)x_2$. Let us denote $(p+1)x_2 = x_4$, $(p+1)x_4 = x_5$, then $u = (p+1)x_5 - x_4 + 2x_2$. The third equation would be $u = (p+1)(p^3 + 3p^2 + 2p + 2)x_3$, or $(p+1)x_3 = x_2$.

The required system of Ito equations and the output equation have the form

$$dx_1 + x_1 dt = dw, \quad dx_2 + (x_2 - x_4)dt = 0,$$
$$dx_3 + (x_3 - x_2)dt = 0, \quad dx_4 + (x_4 - x_5)dt = 0,$$
$$dx_5 + (x_5 - x_4 + 2x_2)dt = dw,$$
$$x = x_1 + x_2 - 2x_3.$$

Based on the transfer function, a linear system of Ito differential equations was obtained that does not contain derivatives of the input signal. Of course, the choice of a new variable was made so that it would be easy to isolate the cube of the sum in the denominator of the transfer function, and then obtain first-order linear equations.

## 6. DRYDEN WIND TURBULENCE MODEL

The US Department of Defense uses the Dryden gust model in some aircraft design and simulation applications. This mathematical model considers the speed components of continuous gusts of wind as random processes [5, 11]. The MATLAB documentation provides an implementation of the transfer function for wind gusts [12]. Twelve transfer functions are defined for gust models in the longitudinal, horizontal and vertical directions. However, only three types of different functions can be distinguished, differing from the model functions only by constant coefficients $A$, $B$, $C$, $D$ (see [12]):

$$G_1(p) = A\frac{1}{1 + Cp}, \quad G_2(p) = A\frac{1 + Bp}{(1 + Cp)^2}, \quad G_3(p) = \frac{Ap}{1 + Cp} \times \frac{1 + Bp}{(1 + Dp)^2}.$$

The first type of function $G_1(p)$ is a simple integrator and does not require any transformation. The required system of Ito equations for $G_1(p)$ has the form

$$dx + \frac{1}{C}xdt = \frac{A}{C}dw.$$

Let's look at the second one. It is required to obtain a system of Ito equations for the transfer function $G_2(p)$. Then

$$G_2^*(q) = G_2\left(\frac{q-1}{C}\right) = \frac{A}{C}\frac{Bq + C - B}{q^2} = \frac{A}{C}\left(\frac{B}{q} + \frac{C - B}{q^2}\right).$$
$$G_2(p) = \frac{A}{C}\left(\frac{B}{1 + Cp} + \frac{C - B}{(1 + Cp)^2}\right).$$

The desired system of Ito equations and the output equation for $G_2(p)$ have the form

$$dx_1 + \frac{1}{C}x_1 dt = \frac{1}{C}dw, \quad dx_2 + \frac{1}{C}(x_2 - x_1)dt = 0,$$
$$x = \frac{A}{C}\Big(Bx_1 + (C - B)x_2\Big).$$

If we make another change of variables $q = p + 1$, we will get a rather cumbersome system of 5th order equations. We invite readers to see this for themselves.

Let's consider the third function $G_3(p)$ and replace the variable: $p = (q - 1)/D$. Then

$$G_3^*(q) = G_3\left(\frac{q-1}{D}\right) = A\frac{Bq^2 + (D - 2B)q + B - D}{DCq^3 + (D^2 - DC)q^2}.$$

Let's represent the last expression using (4):

$$G_3^*(q) = A\left[\frac{\alpha_1}{q} + \frac{\alpha_2}{q^2} + \frac{1}{b_0 q^3 + b_1 q^2}\left(\frac{\beta_2}{q^2} + \frac{\beta_1}{q} + \beta_0\right)\right],$$

then $b_0 = DC$, $b_1 = D^2 - DC$, $\alpha_1 = B/(DC)$, $\alpha_2 = \left[D - 2B - B(D^2 - DC)/(DC)\right]/(DC)$, $\beta_0 = B - D - (D^2 - DC)\left[D - 2B - B(D^2 - DC)/(DC)\right]/(DC)$, $\beta_1 = 0$, $\beta_2 = 0$.

Let's do the reverse change of variables and get the function

$$G_3(p) = A\left[\frac{\alpha_1}{1 + Dp} + \frac{\alpha_2}{(1 + Dp)^2} + \frac{\beta_0}{b_0(1 + Dp)^3 + b_1(1 + Dp)^2}\right].$$

Let us write down the derivation of the Ito equations for $G_3(p)$ in more detail. The first and second terms give $x_1 = \frac{u}{1+Dp}$, $x_2 = \frac{u}{(1+Dp)^2} = \frac{u}{1+Dp} \times \frac{1}{1+Dp} = \frac{x_1}{1+Dp}$. Let's consider the third term: $x_3 = \frac{u}{b_0(1+Dp)^3 + b_1(1+Dp)^2} = \frac{u}{(1+Dp)^2} \times \frac{1}{b_0(1+Dp)+b_1} = \frac{x_2}{b_0(1+Dp)+b_1}$. Then the desired system of Ito equations and the output equation for $G_3(p)$ have the form

$$dx_1 + \frac{1}{D}x_1 dt = \frac{1}{D}dw, \quad dx_2 + \frac{1}{D}(x_2 - x_1)dt = 0,$$
$$dx_3 - \frac{1}{Db_0}x_2 dt + \frac{b_0 + b_1}{Db_0}x_3 dt = 0,$$
$$x = A\big[\alpha_1 x_1 + \alpha_2 x_2 + \beta_0 x_3\big].$$

Dryden's wind turbulence model is not the only one. For example, the von Karman model [13] has other transfer functions such as

$$G(p) = A\frac{1 + Bp}{1 + Cp + Dp^2}.$$

The corresponding system of Ito equations for this function will contain 6 variables. We do not present the transformed function here because it turned out to be too cumbersome. Perhaps a not very successful variable replacement was chosen. Therefore, the researchers themselves, depending on the coefficients $C$ and $D$ of the denominator polynomial, must choose a manner for replacing the variable.

The discussion about the choice of turbulence model continues [14]. It can be seen that the number of variables in the Ito equation for the Dryden model is no more than three, and in the von Karman model no less than six. Accordingly, the computational complexity of the wind gust modeling algorithm increases.

## 7. CONCLUSION

The proposed method for conversing the transfer function allows us to bring it to such a form that the system of differential equations equivalent to the differential equation (3) does not contain derivatives of the input signal $u(t)$. Assuming that $u(t)$ is white noise, the system can easily be transformed into a system of Ito equations.

The results can be used not only for stochastic differential equations, but also for ordinary differential equations with constant coefficients of the form (3) with scalar input and output signals [3, Section 1.3.4].

In the presented method it is impossible to influence the number of variables, but in the method [8] it is possible to influence the number of output variables for the output signal $y = cx$ (see (1)) and, accordingly, the type shaping filter shown in Fig. 3. Therefore, in the approach discussed above, the structure of the output signal becomes known only as a result of solving the problem. And in the method described in [8], the type of the output signal is known in advance. But the proposed solution, unlike [8], does not require matrix inversion, but uses only recursive arithmetic operations to find the coefficients of polynomials.

The results of calculations and numerical modeling of dynamic processes for the systems considered in the article are not specifically presented here. In the opinion of the authors, a fairly complete study with various modeling results was carried out in [14].

## FUNDING

The work was carried out in part with the financial support of the Russian Foundation for Basic Research (project no. 20-08-00400).

## REFERENCES

1. Konovalov, G.F., *Radioavtomatika: Ucheb. dlya vuzov po spets. "Radiotekhnika"* (Radioautomation: Textbook for universities specializing in "Radio Engineering"), Moskow: Vysshaya shkola, 1990.

2. Kim, D.P., *Sbornik zadach po teorii avtomaticheskogo upravleniya. Mnogomernye, nelineinye, optimal'nye i adaptivnye sistemy* (Collection of problems on the theory of automatic control. Multidimensional, nonlinear, optimal and adaptive systems), Moskow: FIZMATLIT, 2008.

3. Pugachev, V.S. and Sinitsin, I.N., *Stokhasticheskie differentsial'nye sistemy* (Stochastic differential systems), Moscow: Nauka, 1985.

4. Hoblit, F.M., *Gust Loads on Aircraft: Concepts and Applications*, Washington, DC: American Institute of Aeronautics and Astronautics, Inc. ISBN 0930403452, 1988.

5. McLean, D., *Automatic Flight Control Systems*, Prentice Hall Inc., Englewood Cliffs, 1990, 593 p.

6. Gliklikh, Yu.E., Study of Leontief-type equations with white noise using methods of average derivatives of random processess, *Ser. Matem. modelirovanie i programmirovanie. Vestn. YuUrGU*, 2012, no. 27 (286), pp. 24–34.

7. Agapova, A.S. and Khrustalev, M.M., System shape optimization and stabilization of controlled quasilinear stochastic systems that operate on an infinite time interval, *J. Comput. Syst. Sci. Int.*, 2017, vol. 56, no. 1, pp. 64–86. https://doi.org/10.1134/S1064230717010099

8. Veremey, E.I., *Lineinye sistemy s obratnoi svyaz'yu* (Linear systems with feedback), St. Petersburg: Lan', 2013, vol. 1.

9. Kudryavtsev, L.D., *Matematicheskii analiz* (Mathematical analysis), Moskow: Vysshaya shkola, 1970, vol. 1, pp. 369-370.

10. Panteleev, A.V. and Bortakovsky, A.S., *Teoriya upravleniya v primerakh i zadachakh* (Control theory in examples and tasks), Moskow: Vysshaya Shkola, 2003.

11. Liepmann, H.W., On the Application of Statistical Concepts to the Buffeting Problem, *J. Aeronaut. Sci.*, 1952, vol. 19, no.12, pp. 793–800. https://doi.org/10.2514/8.2491

12. MATLAB Reference Pages. The MathWorks, Inc. 2010. Retrieved Jan 31, 2024. https://www.mathworks.com/help/aeroblks/ drydenwindturbulencemodelcontinuous.html

13. MATLAB Reference Pages. The MathWorks, Inc. 2010. Retrieved Jan 31, 2024. https://www.mathworks.com/help/aeroblks/ vonkarmanwindturbulencemodelcontinuous.html

14. U.S. Military Handbook MIL-HDBK-1797, December 19, 1997.

*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, 2024, Vol. 85, No. 7, pp. 711–724.*
*© The Author(s), 2024 published by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 2024.*
*Russian Text © The Author(s), 2024, published in Avtomatika i Telemekhanika, 2024, No. 7, pp. 73–90.*

# Quantity Conjectural Variations in Oligopoly Games under Different Demand and Cost Functions and Multilevel Leadership

**M. I. Geraskin**

*Samara University, Samara, Russia*
*e-mail: innovation@ssau.ru*

Received October 31, 2023
Revised February 27, 2024
Accepted April 30, 2024

**Abstract**—This paper considers a noncooperative game of quantity competition among firms in an oligopoly market under general demand and cost functions. Each firm's optimal response to the strategies of other firms is assessed by the magnitude and sign of its conjectural variation, expressing the firm's expectation regarding the counterparty's supply quantity change in response to the firm's unit change in its supply quantity. A game of $n$ firms with the sum of conjectural variations (SCV) regarding all counterparties as the generalized response characteristic is studied. The existence of a bifurcation of the players' response is revealed; a bifurcation is a strategy profile of the game in which both positive and negative responses are possible with an infinite-magnitude SCV value. Methods are developed for calculating the SCV value under different types of inverse demand functions (linear and power) and cost functions (linear, power, and quadratic), and the impact of these characteristics of firms on the bifurcation state is comparatively analyzed.

*Keywords*: oligopoly, conjectural variation, bifurcation, Stackelberg leadership

**DOI:** 10.31857/S0005117924070066

## 1. INTRODUCTION

In an oligopoly game, players (firms) make assumptions about the strategies of other players (the environment) underlying their optimal response to these strategies. In the case of quantity competition, the assumptions of firms are formalized by conjectural variations [1]. This case is often considered by researchers [2] due to the preferability of quantity competition: it results in a smaller output, higher prices, and higher profits than price competition [3]. A conjectural variation (hereinafter, meaning a quantity conjectural variation) is the firm's expectation regarding the counterparty's supply quantity change in response to the firm's unit change in its supply quantity. In oligopoly theory, it is conventional to consider the optimal (consistent [4]) conjectural variation calculated from the player's necessary condition of optimality, i.e., the one corresponding to the player's best response. In other words, the player's strategy choice model (utility function) being unknown, the awareness of the conjectural variation allows predicting the player's behavior.

In addition, when assessing the conjectural variations, a player can assume that the counterparty is also assessing him, i.e., suppose the former's optimal behavior. In this case, the counterparty is called a Stackelberg leader whereas the given player a follower. However, the counterparty may argue by analogy, treating the given player as a Stackelberg leader and calculating the conjectural variation from the leader's optimal response (thereby becoming a second-level leader for the given player). This sequence of players' reasoning is called strategic reflexion. Thus, an analysis of conjectural variations inevitably leads to the problem of multilevel leadership [5]. Consequently, the vector of the conjectural variations of all players is a complex characteristic of the strategy profile of the game with these mental profiles of the firms, as conjectural variations are functions of the role of each player in the hierarchy of multilevel leadership.

In an oligopoly game of $n > 2$ players, the firm's behavior is determined by the sum of its conjectural variations (SCV) regarding all environment players. If the player's SCV is negative, then its optimal strategy is to increase the supply quantity, and vice versa. Therefore, in the $n$-player game, the awareness of all components of the vector of conjectural variations of all players is not necessary for predicting the game outcome: it suffices to know the components of the SCV vector of all players. No doubt, the awareness of the players' utility functions is required to determine the SCV vector; nevertheless, given available limits for typical utility functions and the nature of SCV changes, one receives an information base for predicting game outcome limits.

Typical utility functions are defined by a set of demand functions and cost functions [5–19]. In the studies of oligopoly, the most common inverse demand functions are the linear [5, 6, 9–15] and power [5, 16–19] ones. The set of functions describing the costs of oligopolists is somewhat wider: the linear function [10, 12–14, 16, 18], the power function [6, 17], and the quadratic function [5, 7–9, 11, 15, 19]. Obviously, in the vast majority of publications, researchers consider the linear models of demand and costs: in this case, it is easy to calculate conjectural variations from the best response functions (reaction functions) in explicit form. The power cost function can be either convex or concave for different degrees; a concave cost function corresponds to the positive scale effect whereas a convex cost function to the negative scale effect. The quadratic cost function is used only to describe the negative scale effect in the convex case: otherwise, the transition to a decreasing dependence of costs on output may occur, which disagrees with the economic realities.

Thus, when assessing the behavior of firms in an oligopoly game, a topical problem is to analyze the nature and limits of SCV changes due to the changes in the profile of their reflexive beliefs under different utility functions of the players.

## 2. FORMULATION OF THE OLIGOPOLY GAME MODEL

Consider quantity mono-product competition in an oligopoly of $n > 2$ firms. Let all firms have a common inverse demand function $P(Q)$ decreasing in the total supply quantity $Q$, and let the cost function $C_i(Q_i)$ of each firm $i$ be nondecreasing in its supply quantity $Q_i$.

We suppose the possibility of reflexion for each player (firm), given by a reflexion rank $r$. The player's reflexive behavior consists in putting forward some beliefs about the strategies of its environment (the other players), which leads to the appearance of phantom players in the game [20]. In this case, reflexion rank is a numerical characteristic of such beliefs, and the sequence of reflexion ranks defines the following hierarchy of phantom players:

—At rank $r = 1$, the player is aware that the environment does not know its strategy, i.e., the other players are followers and this player becomes a first-level Stackelberg leader.

—At rank $r = 2$, according to the player's information, it is surrounded by first-level Stackelberg leaders; hence, this player becomes a second-level Stackelberg leader.

—At an arbitrary rank $r$, the player knows that the environment players are $(r-1)$th-level Stackelberg leaders; therefore, this player becomes an $r$th-level Stackelberg leader.

Thus, the real game of firms in an oligopoly market will be treated below as an information game of phantom players, each having different leadership levels depending on the degree of its awareness. Such a situation is commonly called multilevel leadership (a multiple leader–follower game) [5], and leadership levels are given by the reflexion rank $r$.

A multiple leader–follower game is a tuple of the form

$$\Gamma = \langle N, \{Q_i, i \in N\}, \{\Pi_i, i \in N\}, \{r_i, i \in N\}\rangle,$$

where $N = \{1, \ldots, n\}$ denotes the set of players, $\{Q_i, i \in N\}$ is their action vector (the strategy profile of the game), $\{\Pi_i, i \in N\}$ is the vector of their utility functions, and $\{r_i, i \in N\}$ is the vector of their ranks.

The utility function of player $i$ has the form

$$\Pi_i(Q, Q_i) = P(Q)Q_i - C_i(Q_i).$$

Differentiating the utility functions of the players, we define the system of necessary conditions for Nash equilibrium:

$$P(Q) + (1 + S_i^r)Q_i P_Q^{/} - C_{iQ_i}^{/} = 0, \quad i \in N, \tag{1}$$

where $S_i^r = \sum_{j \in N \setminus i} Q_{j(r)Q_i}^{/}$ is the sum of the conjectural variations of player $i$ at a reflexion rank $r$ (each component $Q_{jQ_i}^{r/}$ is the conjectural variation of player $i$, i.e., the expected change in the quantity of player $j$ in response to the unit quantity increase of player $i$); the value $Q_{jQ_i}^{r/} = \rho_{ij}^r$ is calculated by differentiating equation (1) for player $j$, which confirms its optimality.

An equilibrium in this game, i.e., a solution of system (1) that maximizes the utility functions $\Pi_i(Q, Q_i)$ of the players, exists under the condition established by W. Novshek [21]:

$$P_Q^{/} + P_{QQ}^{//} Q < 0.$$

This condition depends on the type of demand functions: for linear and exponential demand functions, it is satisfied; for the power demand function, it fails, and the existence of an equilibrium requires the nondecreasing property of the cost functions, $C_{iQ_i}^{/} \geqslant 0$.

The solution of system (1) can be found if the SCV values $S_i^r$ are known for all players. They are calculated using the following recurrent formula [6] at an arbitrary reflexion rank:

$$S_i^r = \left( \frac{1}{\sum\limits_{j \in N \setminus i} \dfrac{1}{u_j - S_j^{r-1} + 1}} - 1 \right)^{-1}. \tag{2}$$

Due to (2), the player's SCV depends on two characteristics of the environment players:

—the mental types of players, defined by their SCV values $S_j^{r-1}$ at the previous reflexion rank;

—the technological types related to the type of the cost functions of the environment players, defined by the parameters $u_j$; for some types of demand functions (if $P_{QQQ_i}^{//} \neq 0$, see below), this parameter also describes the player's mental type.

Note that formula (2) is presented for the conjectural variations independent of the actions of players, i.e., under the condition $\rho_{ijQ_i}^{/} = 0$; the more general case $\rho_{ijQ_i}^{/} \neq 0$ was described in [6]. It was also demonstrated therein that conjectural variations weakly depend on the supply quantities of players, i.e., $\rho_{ijQ_i}^{/} \approx 0$. Below, we will justify this premise for the demand and cost functions under considerations, showing that the SCV values and types of the demand and cost functions of the environment players have the greatest impact on the player's SCV value.

**Proposition 1.** *The parameter $u_i$ in* (2) *is given by*

$$u_i = -1 + \frac{P_{Q_i}^{/} + (1 + S_i^{r-1})Q_i P_{QQ_i}^{//} - C_{iQ_iQ_i}^{//}}{\mid P_Q^{/} \mid}. \tag{2a}$$

We will call $u_i$ *the nonlinearity coefficient* since it characterizes the impact of the nonlinearity of the demand and cost functions on the type of equation (1) of player $i$ : for $u_i = -2$, the corresponding equation of system (1) is linear.

Thus, according to (1), the computation of the game equilibrium directly depends on the SCV value. In turn, this value is predetermined by the peculiarities of the functions $P(Q)$ and $C_i(Q_i)$; see formula (2). Therefore, we will study possible SCV values under different combinations of these functions.

## 3. RESULTS

### *3.1. Methods for Calculating Conjectural Variations*

Whenever the player's number $i$ does not matter, it will be omitted below, and the player's action will be denoted by $q = Q_i \forall i \in N$. Consider the inverse demand functions

$$P_1(Q) = a - bQ, \quad a > 0, \quad b > 0, \quad a \gg b, \tag{3a}$$

$$P_2(Q) = AQ^{\alpha}, \quad A > 0, \quad \alpha < 0, \quad \mid \alpha \mid < 1, \tag{3b}$$

and the cost functions

$$C_1(q) = B_0 + B_1 q, \quad B_0 \geqslant 0, \quad B_1 > 0, \tag{4a}$$

$$C_2(q) = B_0 + B_1 q^{\beta}, \quad B_0 \geqslant 0, \quad B_1 > 0, \quad \beta \in (0, 2), \tag{4b}$$

$$C_3(q) = B_0 + B_1 q + \frac{B_2}{2} q^2, \quad B_0 \geqslant 0, \quad B_1, B_2 > 0, \tag{4c}$$

where $a, b, A$, and $\alpha$ are the constant coefficients of the demand functions and $B_0, B_1, B_2$, and $\beta$ are the constant coefficients of the cost functions.

Here, we adopt the notations of function types: $P_k(Q)$, $k = 1, 2$, is the demand function of type $k$ ($k = 1$ corresponds to the linear function and $k = 2$ to the power function); $C_m(q)$, $m = 1, 2, 3$, is the cost function of type $m$ ($m = 1$ corresponds to the linear function, $m = 2$ to the power function, and $m = 3$ to the quadratic function).

Using formula (2), we derive expressions for $P^{/}_{Q}$, $P^{/}_{q}$, $P^{//}_{Qq}$, and $C^{//}_{qq}$ in the case of the functions (3) and (4):

$$P^{/}_{1_Q} = P^{/}_{1_q} = -b, \quad P^{//}_{1_{Qq}} = 0, \tag{5a}$$

$$P^{/}_{2_Q} = P^{/}_{2_q} = A\alpha Q^{\alpha-1}, \quad P^{//}_{2_{Qq}} = A\alpha(\alpha - 1)Q^{\alpha-2}, \tag{5b}$$

$$C^{//}_{1_{qq}} = 0, \quad C^{//}_{2_{qq}} = B_1\beta(\beta - 1)q^{\beta-2}, \quad C^{//}_{3_{qq}} = B_2. \tag{5c}$$

As a result, the parameters $u_{km}$, $k = 1, 2$, $m = 1, 2, 3$, of the functions (3) and (4) are given by

$$u_{11} = -2, \quad u_{21} = -2 + (1 + S^{r-1})(1 - \alpha)\frac{q}{Q}, \tag{6a}$$

$$\begin{gathered} u_{12} = -2 - \frac{B_1}{b}\beta(\beta - 1)q^{\beta-2}, \\ u_{22} = -2 + (1 + S^{r-1})(1 - \alpha)\frac{q}{Q} - \frac{B_1}{A \mid \alpha \mid Q^{\alpha-1}}\beta(\beta - 1)q^{\beta-2}, \end{gathered} \tag{6b}$$

$$u_{13} = -2 - \frac{B_2}{b}, \quad u_{23} = -2 + (1 + S^{r-1})(1 - \alpha)\frac{q}{Q} - \frac{B_2}{A \mid \alpha \mid Q^{\alpha-1}}. \tag{6c}$$

Note that the parameter $B_1$ in (6) corresponds only to the case of the power function.

### 3.2. Comparative Analysis of Conjectural Variations

With the notation $s_i^r = \sum_{j\in N\setminus i} \frac{1}{u_j - S_j^{r-1}+1}$, formula (2) is simplified to

$$S_i^r = \left(\frac{1}{s_i^r} - 1\right)^{-1}, \tag{7}$$

where $s_i^r$ expresses the *aggregate* of the cost functions and SCV values of the environment of player $i$, i.e., a generalized characteristic of the technological and mental types of the other players.

Due to formula (7), the function $S_i^r(u_j, S_j^{r-1})$ suffers from a discontinuity of the second kind (Fig. 1) under the condition

$$s_i^r = \sum_{j\in N\setminus i} \frac{1}{u_j - S_j^{r-1} + 1} = 1; \tag{7a}$$

moreover, it takes infinitely large positive and negative values as $s_i^r \to 1 - 0$ and $s_i^r \to 1 + 0$, respectively. The function $s_i^r(u_j, S_j^{r-1})$ has a discontinuity of the second kind for $S_j^{r-1} = u_j + 1 \forall j \in N\setminus i$, which does not cause a discontinuity of the function $S_i^r(u_j, S_j^{r-1})$.

The discontinuity of the second kind of the function $S_i^r(u_j, S_j^{r-1})$ means that at the point of discontinuity $(u_j^0, S_j^{0,r-1})$, $j \in N\setminus i$, player $i$ at a reflexion rank $r$ can simultaneously have two SCV values ($+\infty$ and $-\infty$). Let us consider the sequential reflexion of the players regarding each other's behavior as a dynamic process on the numerical sequence of ranks $r = 1, 2, \ldots$. Then, by analogy with the solutions of some differential equations, we can say that there is a bifurcation of the player's beliefs. In this case, the *bifurcation state* of the beliefs of player $i$ is a combination of the

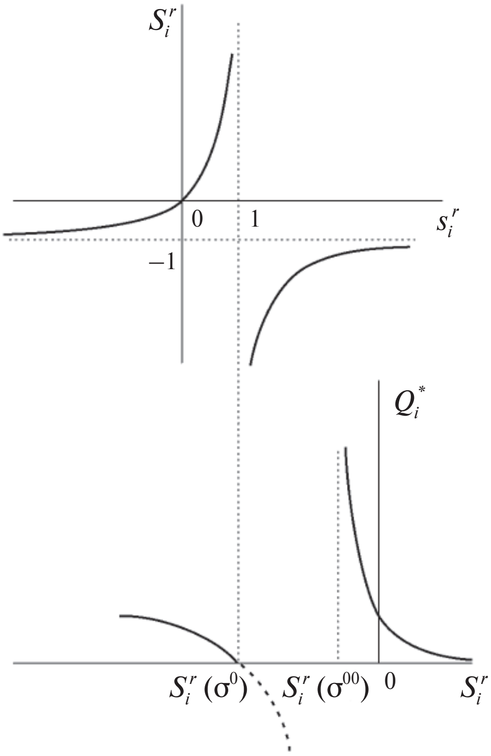


**Fig. 1.** The SCV value of player $i$ depending on the aggregate of the cost functions and SCV values of the environment (top) and the equilibrium action of player $i$ depending on SCV value (bottom).

technological types of the environment players, defined by their cost functions, and the mental types of the environment players, expressed by their leadership levels (numerically defined in the SCV form), under which player $i$ can simultaneously expect infinitely large, both positive and negative, reactions (SCV values) of the environment.

Function (7) (see the upper part of Fig. 1) allows estimating the following intervals of SCV changes:

$$S_i^r = \begin{cases} \in [-1, 0) & \text{if } s_i^r \leqslant 0 \\ \in (0, \infty) & \text{if } 0 < s_i^r < 1 \\ \in (-\infty, -1) & \text{if } s_i^r < -1. \end{cases} \tag{7b}$$

Hence, by characterizing the dependence of the aggregate $s_i^r$ on the environment's nonlinearity coefficients $u_j$, which depend on the types of the demand and cost functions and are low-sensitive to the supply quantities, and on the environment's SCV value $S_j^{r-1}$, which depends on the latter's mental type, we can estimate the impact of these parameters on the player's SCV value.

To qualitatively analyze in comparative terms the impact of the types of the functions $P_k(Q)$ and $C_m(q)$ and study the bifurcation phenomenon, we consider the case of identical players. Assume that for all environment players, the nonlinearity coefficients and SCV values are the same: $u_j = u$ $\forall j \in N$, $S_j^{r-1} = \sigma$ $\forall j \in N$. In this case,

$$S_i^r = \frac{n-1}{u + 2 - \sigma - n}, \quad s_i^r = \frac{n-1}{u - \sigma + 1}. \tag{8}$$

Formulas (7a) and (8) lead to the following result.

**Proposition 2.** *If $u_j = u \, \forall j \in N$ and $S_j^{r-1} = \sigma \, \forall j \in N$, then the SCV function $S_i^r(u, \sigma)$ has the following properties:*

i) *a discontinuity of the second kind at $\sigma = \sigma^0 = u + 2 - n$ (except for the case of the linear demand and cost functions) and*

$$\lim_{\sigma \to (u+2-n)-0} S_i^r = \infty, \quad \lim_{\sigma \to (u+2-n)+0} S_i^r = -\infty; \tag{8a}$$

ii) *values belonging to the intervals*

$$S_i^r \begin{cases} \in (0, \infty) & \textit{if } \sigma \in (-\infty, \sigma^0) \\ \in (-\infty, 0) & \textit{if } \sigma \in (\sigma^0, \infty) \end{cases} \tag{8b}$$

(*except for the case of the linear demand and cost functions, in which $S_i^r \in [-1, 0)$*).

A clear illustration of a bifurcation follows from the explicit-form solution of the system of equilibrium equations (1), known for the case of the linear demand and cost functions. However, in this case, infinite values of conjectural variations do not arise (they are bounded by the range $(-1, 0]$). For power cost functions, an explicit-form solution does not exist [6], so we consider the case of the linear demand function and the quadratic cost functions.

**Proposition 3.** *In the case of the linear demand function and the quadratic cost functions, the general solution of game* (1) *has the form*

$$Q_i^* = \frac{D_i \left[ \prod\limits_{j=1 \backslash i}^{n} \left(\gamma_j^r - 1\right) + \sum\limits_{j=1 \backslash i}^{n} \prod\limits_{\mu=1 \backslash j,i}^{n} \left(\gamma_\mu^r - 1\right) \right] - \sum\limits_{j=1 \backslash i}^{n} \left[ D_j \prod\limits_{\mu=1 \backslash i,j}^{n} \left(\gamma_\mu^r - 1\right) \right]}{\prod\limits_{j=1}^{n} \left(\gamma_j^r - 1\right) + \sum\limits_{j=1}^{n} \prod\limits_{\mu=1 \backslash j}^{n} \left(\gamma_\mu^r - 1\right)}; \tag{9}$$

*in the particular case $n = 3$, the formula reduces to*

$$Q_i^* = \frac{D_i\left(\prod\limits_{j=1\backslash i}^{3} \gamma_j^r - 1\right) - \sum\limits_{j=1\backslash i}^{3} \prod\limits_{\mu=1\backslash i}^{3} D_j \gamma_\mu^r + \sum\limits_{j=1\backslash i}^{3} D_j}{\gamma_1^r \gamma_2^r \gamma_3^r - \gamma_1^r - \gamma_2^r - \gamma_3^r + 2}; \tag{9a}$$

*in the particular case $n = 3$ with the identical types of the players, $D = D_i\ \forall i \in N$, the formula becomes*

$$Q_i^* = D\frac{\prod\limits_{j=1\backslash i}^{3} \gamma_j^r - \sum\limits_{j=1\backslash i}^{3} \gamma_j^r + 1}{\gamma_1^r \gamma_2^r \gamma_3^r - \gamma_1^r - \gamma_2^r - \gamma_3^r + 2}, \tag{9b}$$

*where*

$$D_i = \frac{a - B_{1i}}{b}, \quad \gamma_i = 2 + S_i^r + \frac{B_{2i}}{b},$$

*and the symbol "*" indicates the game equilibrium.*

Under a *belief bifurcation*, two cases are simultaneously possible. They will be described based on (9b) for the identical (same-type) environment players, see the corresponding condition in Proposition 2. Then, considering (6c), we have $\gamma_j = \gamma = -u_j + S_j^{r-1} = -u + \sigma$, $j = 2, 3$, and $\gamma < 0$ for $\sigma = \sigma^0$ since $B_2 > 0$.

The first case $S_i^r \to \infty$ means that for the environment players, the optimal strategy is infinite growth of the supply quantity, limited by the parameters of the demand function $P(Q)$ and the technological capabilities of the firms. (Recall that conjectural variations are considered as strategies.) In this case, if the environment's SCV values $S_j^r$, $j = N\backslash i$, are finite numbers, then by (9b) the player's optimal response vanishes on the right, i.e., $Q_i^{r^*} \to 0 + 0$. In other words, the player seeks to reduce the supply quantity to zero.

The second case $S_i^r \to -\infty$ implies an infinite reduction in the supply quantity by the environment players, although they can actually reduce the supply only to zero. Due to (9b), the player's optimal response vanishes on the left, i.e., $Q_i^{r^*} \to 0 - 0$. This can be interpreted as the player's largest acceptable response to the negative value of the total supply quantity predicted by this player based on $S_i^r \to -\infty$.

Interestingly, a belief bifurcation should lead to an *action bifurcation* at the subsequent reflexion ranks. This fact can also be demonstrated from the optimal SCV formula (8) and the equilibrium action (9b).

If $S_i^r \to \infty$, at the next reflexion rank $(r + 1)$, we consider the situation from the environment's viewpoint (i.e., as $\sigma \to \infty$); from (8) it follows that $S_j^{r+1} \to 0$. Returning to player $i$ at rank $(r + 2)$, for which $\sigma \to 0$, from (8) we also obtain $\lim_{\sigma\to 0} s_i^{r+2} = \frac{n-1}{u-\sigma+1} < 0$ since $u < -2$ by (6c). Consequently, $S_i^{r+2} \to -1$ as $\sigma \to \sigma^{00} = u + 1$, and when preserving the environment's responses by the type $S_j^{r+1} \to 0$, formula (9b) implies $Q_1^* \to \infty$ ($\gamma = 1$ and $Q_1^* = D\frac{\gamma^2 - 2\gamma + 1}{\gamma_1(\gamma^2-1)-2(\gamma-1)}$). Thus, the SCV-defined mental response bifurcation leads to an equilibrium bifurcation in the game. These considerations are illustrated in Fig. 1 (the lower part).

The case of identical players is the basis for comparatively analyzing the impact of the types of demand and cost functions on the SCV value. According to (8a), the bifurcation point $\sigma^0$ shifts upwards when increasing the nonlinearity coefficients $u$ of the environment players and behaves oppositely when decreasing $u$ decreases. In other words, a bifurcation state occurs under higher values of the environment's SCV value. Due to conditions (8b), if the nonlinearity coefficients are larger, the environment's SCV value should be larger so that $S_i^r$ belongs to the corresponding ranges.

Let us characterize the dependence of the SCV value of player $i$ on the nonlinearity coefficient $u_l$ of some environment player $l$ as well as on the environment's SCV value and $q, Q$.

**Proposition 4.** *The SCV value $S_i^r$ of player $i$ at a reflexion rank $r$ has the following properties:*

i) *goes down when increasing the nonlinearity coefficient $u_l$ of environment player $l$ and up when increasing $S^{r-1}$ :*

$$S^{r/}_{iul} < 0, \;\; S^{r/}_{iS^{r-1}} > 0; \tag{10a}$$

ii) *in the case of the linear demand function, is independent of $q$ under the linear and quadratic cost functions of the environment, goes down (up) when increasing $q$ for $\beta > 1$ (for $\beta < 1$, respectively) under the power cost functions of the environment, and is independent of $Q$ under any cost functions of the environment:*

$$S^{r/}_{i\;q}\Big|_{\substack{k=1\\ m=1,3}} = S^{r/}_{i\;Q}\Big|_{\substack{k=1\\ m=1,2,3}} = 0, \quad S^{r/}_{i\;q}\Big|_{\substack{k=1\\ m=2}} \begin{cases} < 0 & \textit{for } \beta > 1, \\ > 0 & \textit{for } \beta < 1; \end{cases} \tag{10b}$$

iii) *in the case of the power demand function, goes down (up) when increasing $q$ if $S^{r-1} > -1$ (if $S^{r-1} < -1$, respectively) under the linear and quadratic cost functions of the environment, down if $S^{r-1} > -1$ under the convex power cost functions ($\beta > 1$) and under the concave power cost functions ($\beta < 1$) provided that $\varphi < 1$, and up (down) if $S^{r-1} < -1$ under the concave power cost functions (under the convex power cost functions provided that $\varphi < -1$, respectively);*

*goes up (down) when increasing $Q$ if $S^{r-1} > -1$ (if $S^{r-1} < -1$, respectively) under the linear cost functions and the quadratic cost functions (in the latter case, goes down provided that $\zeta < 1$), up (down) if $S^{r-1} > -1$ under the convex power cost functions ($\beta > 1$) (under the concave power cost functions ($\beta < 1$) provided that $\varphi < 1$, respectively), and down (up) if $S^{r-1} < -1$ under the concave power cost functions (under the convex power cost functions provided that $\psi > -1$, respectively):*

$$S^{r/}_{i\;q}\Big|_{\substack{k=2\\ m=1,3}} \begin{cases} < 0 & \textit{for } S^{r-1} > -1, \\ > 0 & \textit{for } S^{r-1} < -1, \end{cases}$$

$$S^{r/}_{i\;Q}\Big|_{\substack{k=2\\ m=1}} \begin{cases} > 0 & \textit{for } S^{r-1} > -1, \\ < 0 & \textit{for } S^{r-1} < -1, \end{cases}$$

$$S^{r/}_{i\;q}\Big|_{\substack{k=2\\ m=2}} \begin{cases} < 0 & \textit{if } \varphi < 1 \textit{ for } t = 1, \\ < 0 & \textit{for } t = 2, \\ > 0 & \textit{for } t = 3, \\ < 0 & \textit{if } \varphi < -1 \textit{ for } t = 4, \end{cases} \tag{10c}$$

$$S^{r/}_{i\;Q}\Big|_{\substack{k=2\\ m=2}} \begin{cases} < 0 & \textit{if } \psi > 1 \textit{ for } t = 1, \\ > 0 & \textit{for } t = 2, \\ < 0 & \textit{for } t = 3, \\ < 0 & \textit{if } \psi > -1 \textit{ for } t = 4, \end{cases}$$

$$S^{r/}_{i\;Q}\Big|_{\substack{k=2\\ m=3}} \begin{cases} > 0 & \textit{for } S^{r-1} > -1, \\ < 0 & \textit{if } \zeta < 1 \textit{ for } S^{r-1} < -1; \end{cases}$$

iv) *weakly depends on the supply quantities of the players compared to the impact of the environment's SCV value:*

$$S^{r/}_{i\;q} \ll S^{r/}_{i\;S^{r-1}}, \tag{10d}$$

*where*

$$\varphi = \frac{B_1\beta(1-\beta)(2-\beta)q^{\beta-3}}{A|\alpha||1+S^{r-1}|(1-\alpha)Q^{\alpha-2}}, \quad \psi = \varphi\frac{1-\alpha}{2-\beta}, \quad \zeta = \frac{B_2Q^{2-\alpha}}{A|\alpha||1+S^{r-1}|q},$$

*and the additional notations are*

$$\begin{aligned} &t=1: \ S^{r-1} > -1 \wedge \beta < 1, \qquad t=2: \ S^{r-1} > -1 \wedge \beta > 1,\\ &t=3: \ S^{r-1} < -1 \wedge \beta < 1, \qquad t=4: S^{r-1} < -1 \wedge \beta > 1. \end{aligned}$$

Now we compare the bifurcation points under different demand and cost functions.

**Proposition 5.** *Under different types of the demand and cost functions of the environment, the bifurcation point $\sigma^0$ satisfies the following relations:*

$$\sigma^0_{23} > \sigma^0_{21}, \tag{11a}$$

$$\sigma^0_{22} > \sigma^0_{21} \text{ for } \beta > 1, \tag{11b}$$

$$\sigma^0_{21} > \sigma^0_{12} \text{ for } B_1 > \bar{B}_1, \tag{11c}$$

$$\sigma^0_{21} > \sigma^0_{13} \text{ for } B_2 < \bar{B}_2, \tag{11d}$$

$$\sigma^0_{12} > \sigma^0_{22} \text{ for } B_1 < \bar{\bar{B}}_1, \text{ if } \beta > 1 \text{ or for } B_1 > \bar{\bar{B}}_1 \text{ if } \beta < 1, \tag{11e}$$

$$\sigma^0_{13} < \sigma^0_{23} \text{ for } B_2 < \bar{\bar{B}}_2, \tag{11f}$$

$$\sigma^0_{12} > \sigma^0_{13} \wedge \sigma^0_{22} > \sigma^0_{23} \text{ for } \beta > 1 \text{ and } \frac{B_1}{B_2} > \frac{1}{\lambda}, \tag{11g}$$

*where*

$$\begin{gathered} \bar{B}_1 = b\frac{\delta}{\lambda}, \quad \bar{B}_2 = b\delta, \quad \bar{\bar{B}}_1 = \frac{\delta}{\lambda}(\chi - b), \quad \bar{\bar{B}}_2 = \delta\frac{\chi b}{\chi - b},\\ \delta = (1+S^{r-1})(1-\alpha)\frac{q}{Q}, \quad \chi = A|\alpha|Q^{\alpha-1} > 0, \quad \lambda = \beta(\beta-1)q^{\beta-2}. \end{gathered}$$

The corresponding relations for the bifurcation point under the other types of the demand and cost functions of the environment are presented in the Appendix.

## 4. FINDINGS

According to Proposition 2, a belief bifurcation occurs for a player when the environment's SCV value increases from $\sigma = u + 2 - n - 0$ to $\sigma = u + 2 - n + 0$; furthermore, *a greater magnitude of the SCV value is required to destabilize the equilibrium in the case of more players* since the bifurcation point decreases with increasing $n$.

Proposition 4 reveals the following major factors affecting the player's SCV value. First, there is the factor of the market-technological conditions of the game, determined by the nonlinearity coefficient: the greater the nonlinearity coefficient of the environment players is, the smaller the SCV value will be (i.e., the greater its magnitude will be). As a rule, the SCV value is negative, and *the growing magnitude of the SCV value indicates enhancing the player's response*. Therefore, the combinations of the demand and cost functions resulting in higher values of the nonlinearity coefficient contribute to enhancing the player's response. In particular, these include games with quadratic cost functions or power cost functions with the positive scale effect ($\beta < 1$), which lead to greater values of the nonlinearity coefficient compared to the linear cost model regardless of the demand model; for details, see the Appendix.

Second, *symmetric response* is observed for the players, i.e., the greater the environment's SCV value is, the greater the player's SCV value will be. This player response consonance, qualitatively

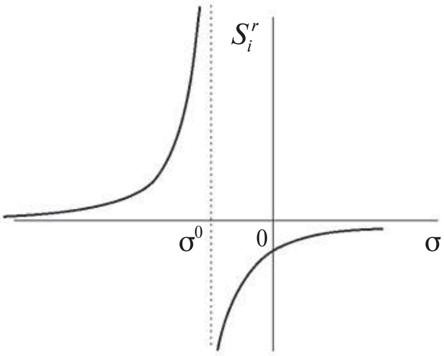


**Fig. 2.** The player's SCV value depending on the environment's SCV value.

illustrated in Fig. 2, is expressed as a two-step process. If the environment has negative SCV values, then increasing them (decreasing their magnitudes) is accompanied by a growth in the positive SCV value of the player ($S_i^r \to \infty$). In economic terms, the player expects expansion of the environment and, as a result, will reduce the supply quantity to zero according to formula (9b). Then a bifurcation point occurs, and the process changes in the opposite direction: applying formula (8) to the environment yields $S_j^{r+1} = \frac{n-1}{u+2-S_i^r-n}$ and, consequently, $S_j^{r+1} \to 0$. In other words, the environment expects the player's zero reaction. But in response to this situation (Fig. 2), the player's SCV value again increases, i.e., $S_j^{r+2} \to -1$, motivating the player to increase the supply quantity.

Third, the greatest impact on the player's SCV value is exerted by the environment's SPV values and types of the demand and cost functions. Despite that the player's SCV value depends on its supply quantity and the total supply quantity of all players through the nonlinearity coefficients of the environment (for the nonlinear functions), this impact is negligibly small compared to the impact of the mental types of the environment conditioned by its leadership levels.

Proposition 5 identifies the following properties of a bifurcation.

The case of *the power demand function* leads to a greater value of the bifurcation point under the quadratic cost function ($k = 2, m = 3$) compared to the linear cost function ($k = 2, m = 1$) since the quadratic function models the negative scale effect. Similarly, in the case of the power demand function, the bifurcation point under the power cost function ($k = 2, m = 2$) with the negative scale effect ($\beta > 1$) exceeds the corresponding point under the linear cost function ($k = 2, m = 1$); the converse situation occurs given the positive scale effect.

Compared to the linear demand function and different cost functions ($m = 2, 3$), the case of *the power demand function and the linear cost function* ($k = 2, m = 1$) leads to an increase in the bifurcation point under certain values of the coefficients $B_1$ and $B_2$ :

i) if $\bar{B}_1 < 0$ (i.e., given the positive scale effect for $S^{r-1} < -1$ and the negative scale effect for $S^{r-1} > -1$) since

$$\bar{B}_1 = \frac{b(1+S^{r-1})(1-\alpha)}{\beta(\beta-1)Qq^{\beta-3}} \begin{cases} > 0 \text{ if } (\beta > 1 \wedge S^{r-1} > -1) \vee (\beta < 1 \wedge S^{r-1} < -1) \\ < 0 \text{ if } (\beta < 1 \wedge S^{r-1} < -1) \vee (\beta > 1 \wedge S^{r-1} > -1); \end{cases}$$

and the scale effect as the technological type of players is opposite to the impact of the environment's SCV value as the mental type of players;

ii) if $\bar{B}_1 > 0$ and $B_1 > \bar{B}_1$, i.e., for a high growth rate of the power function;

iii) if $\bar{B}_2 > 0$, i.e., for $S^{r-1} > -1$ since $\bar{B}_2 > 0 = b(1+S^{r-1})(1-\alpha)\frac{q}{Q}$.

Compared to the *quadratic cost functions* $(m = 3)$ with any demand functions $(k = 1, 2)$, the case of the *power cost functions* $(m = 2)$ gives the following relations for the bifurcation point:

i) The bifurcation point under the power cost functions is greater if the scale effect is negative $(\beta > 1)$, for $B_2 \ll B_1$, since $\frac{B_1}{B_2} > \frac{1}{\lambda}$ implies $B_1\beta(\beta-1)q^{\beta-2} > B_2$, and $\beta(\beta-1)q^{\beta-2} \ll 1$.

ii) The bifurcation point under the power cost functions is smaller if the scale effect is positive $(\beta < 1)$ because, in this case, $B_1\beta(\beta-1)q^{\beta-2} < 0$.

The case of the *quadratic cost functions* $(m = 3)$ with the linear $(k = 1)$ and power $(k = 2)$ demand functions demonstrates two possibilities in which enhancing the environment's response compensates for the nonlinearity impact of the demand function:

i) The bifurcation point under the power cost functions is greater if $S^{r-1} > -1$ and $B_2$ is sufficiently small $(B_2 < \overset{=}{\bar{B}}_2)$ since $\overset{=}{\bar{B}}_2 = \frac{b(1+S^{r-1})(1-\alpha)A|\alpha|qQ^{\alpha-2}}{A|\alpha|Q^{\alpha-1}-b} > 0$ if $A \gg b$.

ii) The bifurcation point under the power cost functions is smaller if $S^{r-1} > -1(\overset{=}{\bar{B}}_2 < 0)$.

## 5. CONCLUSIONS

The market-technological conditions of an oligopoly game are described by a combination of the market demand function and the players' cost functions, which together determine their utility functions. Based on the analysis of the variety of such combinations arising in different applied problems of oligopoly modeling, this study has demonstrated the importance of the market-technological conditions of an oligopoly game for the stability of the game equilibrium. As has been established, the reason of destabilizing equilibrium, or a bifurcation of the players' actions, is a bifurcation of their beliefs: under a definite constellation of the beliefs of environment players, the player can evaluate their optimal reaction as positive and negative simultaneously. In turn, the specified constellation of players' beliefs is predetermined by their Stackelberg leadership levels and expressed by some SCV value of the environment, which can be called a bifurcation point.

The bifurcation point depends on the number of players and the nonlinearity coefficient of their utility functions, and the nonlinearity coefficient is determined by the types of the demand and cost functions. If the bifurcation point is greater for a particular combination of the demand and cost functions of the players (i.e., the SCV value of the environment has a smaller magnitude), the game situation will be more sensitive to changes in the mental types of the players. In other words, the game equilibrium can be easier destabilized in dynamics.

It is characteristic that the equilibrium cannot be destabilized under the linear demand and cost functions. Therefore, gradual changes of the equilibrium actions will be observed in real oligopoly games with the linear dependencies of the market-technological parameters, a phenomenon often encountered in practice.

## *APPENDIX*

**Proof of Proposition 1.** The parameter $u_i$ in [6] is the component of the second-order condition for the optimum of the player's utility function[2], i.e., $\Pi^{//}_{iQ_iQ_i} = u_i - S_i^{r-1} < 0$. Based on (1), we write this condition as $P^{/}_{Q_i} + (1 + S_i^{r-1})P^{/}_{Q} + (1 + S_i^{r-1})Q_iP^{//}_{QQ_i} - C^{//}_{iQ_iQ_i} < 0$; in view of $P^{/}_{Q} < 0$, this inequality can be divided by $|P^{/}_{Q}|$ : $\frac{P^{/}_{Q_i}}{|P^{/}_{Q}|} - 1 - S_i^{r-1} + \frac{(1+S_i^{r-1})Q_iP^{//}_{QQ_i}}{|P^{/}_{Q}|} - \frac{C^{//}_{iQ_iQ_i}}{|P^{/}_{Q}|} < 0$, which finally yields (2a).

[2] In [6], this parameter has the form $u_i = -2 - \frac{C^{//}_{iQ_iQ_i}}{b}$ since it was derived under the linear demand function, for which $P^{/}_{Q} = P^{/}_{Q_i} = -b$.

**Proof of Proposition 3.** Given (3a) and (4c), equations (1) take the form

$$a - bQ - b(1 + S_i^r)Q_i - B_{2i}Q_i - B_{1i} = 0,$$

or

$$\gamma_i Q_i + \sum_{j=N\backslash i} Q_j q_{-i} = D_i, \quad i \in N;$$

solving this system by Cramer's rule gives (9).

**Proof of Proposition 4.** By denoting $z_j^{r-1} = \frac{1}{u_j - S_j^{r-1} + 1}$ and differentiating the expression (7), we obtain

$$S_{i\ u_l}^{r/} = \left(\frac{1}{S_i^r} - 1\right)(s_i^r)^{-2}\left(z_j^{r-1}\right)_{u_l}^{/} = -\left(1 - s_i^r\right)^{-2}\left(u_l - S_l^{r-1} + 1\right)^{-2} < 0, \tag{A.1}$$

$$S_{i\ q}^{r/} = S_{i\ u_l}^{r/} u_{l\ q}^{/}, \quad S_{i\ Q}^{r/} = S_{i\ u_l}^{r/} u_{l\ Q}^{/}, \quad S_{i\ S^{r-1}}^{r/} > 0, \quad S_{i\ u_l}^{r/} < 0. \tag{A.2}$$

To simplify the analysis of formulas (6), let us introduce the notations

$$\delta = (1 + S^{r-1})(1 - \alpha)\frac{q}{Q}\begin{cases} > 0 & \text{for } S^{r-1} > -1 \\ < 0 & \text{for } S^{r-1} < -1, \end{cases} \quad \chi = A|\alpha|Q^{\alpha-1} > 0, \tag{A.3}$$

$$\lambda = \beta(\beta - 1)q^{\beta-2}\begin{cases} < 0 & \text{for } \beta < 1 \\ > 0 & \text{for } \beta > 1. \end{cases}$$

In addition, with the compact notations $x, y, z, X, Y$, and $Z$ for $u_{km}, k = 1, 2, m = 1, 2, 3$, formulas (6) reduce to

$$x = u_{11} = -2, \quad X = u_{21} = -2 + \delta, \tag{A.4a}$$

$$y = u_{12} = -2 - \frac{B_1}{b}\lambda, \quad Y = u_{22} = -2 + \delta - \frac{B_1}{\chi}\lambda, \tag{A.4b}$$

$$z = u_{13} = -2 - \frac{B_2}{b}, \quad Z = u_{23} = -2 + \delta - \frac{B_2}{\chi}. \tag{A.4c}$$

Analysis of (A.3) and (A.4) shows the existence of four possible cases depending on the values of the parameters $\beta$ and $S^{r-1}$, further indicated by the symbol $t$: 1) $t = 1 : S^{r-1} > -1 \wedge \beta < 1$; in this case, $\delta > 0 \wedge \lambda < 0$; 2) $t = 2 : S^{r-1} > -1 \wedge \beta > 1$; in this case, $\delta > 0 \wedge \lambda > 0$; 3) $t = 3 : S^{r-1} < -1 \wedge \beta < 1$; in this case, $\delta < 0 \wedge \lambda < 0$; 4) $t = 4 : S^{r-1} < -1 \wedge \beta > 1$; in this case, $\delta < 0 \wedge \lambda > 0$.

Differentiating (A.4) yields

$$\begin{gathered} x_q^{/} = z_q^{/} = 0, \quad X_q^{/} = Z_q^{/} = \frac{\delta}{q}, \\ y_q^{/} = -\frac{B_1\lambda}{bq}(\beta - 2), \quad Y_q^{/} = \frac{\delta}{q} - \frac{B_1\lambda}{\chi q}(\beta - 2), \end{gathered} \tag{A.5}$$

$$\begin{gathered} x_Q^{/} = y_Q^{/} = z_Q^{/} = 0, \quad X_Q^{/} = -\frac{\delta}{Q}, \\ Y_q^{/} = -\frac{\delta}{Q} - \frac{B_1\lambda}{\chi q}(\beta - 2), \quad Z_Q^{/} = -\frac{\delta}{Q} - \frac{B_2}{\chi Q}(1 - \alpha). \end{gathered} \tag{A.6}$$

Due to (A.2) and (A.3), from these formulas we obtain the following results:

1) under the linear demand function ($k = 1$),

$$S^{r/}_{i\ q}\Big|_{\substack{k=1\\ m=1,3}} = S^{r/}_{i\ Q}\Big|_{\substack{k=1\\ m=1,3}} = 0, \quad S^{r/}_{i\ q}\Big|_{\substack{k=1\\ m=2}} \begin{cases} < 0 & \text{for } \beta > 1 \\ > 0 & \text{for } \beta < 1, \end{cases}$$

2) under the power demand function $(k = 2), Y^{/}_{q} > 0$, i.e., due to (A.1), $S^{r/}_{i\ q}\Big|_{\substack{k=2\\ m=2}} < 0$ if $\delta + \frac{B_1\lambda}{\chi}(2-\beta) > 0$; this inequality leads to the four possible cases: for $t = 1$, the inequality $1 > -\frac{B_1\lambda}{\chi\delta}(2-\beta)$ is valid, and the substitution of (A.3) gives $1 > \varphi = \frac{B_1\beta(1-\beta)(2-\beta)q^{\beta-3}}{A|\alpha||1+S^{r-1}|(1-\alpha)Q^{\alpha-2}}$; for $t = 2$, the inequality is the same and $\varphi < 0$, i.e., $Y^{/}_{q} > 0$ without additional conditions; for $t = 3$, we have $\delta + \frac{B_1\lambda}{\chi}(2-\beta) < 0$, and consequently, $Y^{/}_{q} < 0$; for $t = 4$, $Y^{/}_{q} > 0$ if $\varphi < -1$; the derivatives $Y^{/}_{Q} > 0$ and $Z^{/}_{Q} > 0$ are considered by analogy.

Let us compare $S^{r/}_{i\ S^{r-1}}$ and $S^{r/}_{i\ q}$ by magnitude, observing that

$$S^{r/}_{i\ S^{r-1}} = (1-s^r_i)^{-2}(u_l - S^{r-1}_l + 1)^{-2}, \quad S^{r/}_{i\ q} = (1-s^r_i)^{-2}(u_l - S^{r-1}_l + 1)^{-2}u^{/}_{l\ q}.$$

From (A.5) it follows that

$$u^{/}_{11q} = u^{/}_{13q} = 0, \quad u^{/}_{21q} = u^{/}_{23q} = (1+S^{r-1})(1-\alpha)\frac{1}{Q},$$
$$u^{/}_{12q} = \frac{B_1}{b}(\beta-2)\beta(\beta-1)q^{\beta-3},$$
$$u^{/}_{22q} = (1+S^{r-1})(1-\alpha)\frac{1}{Q} - \frac{B_1}{A|\alpha|Q^{\alpha-1}}(\beta-2)\beta(\beta-1)q^{\beta-3}.$$

Obviously, $\lim_{q\to\infty} u^{/}_{l\ q} \to 0$, and therefore, $S^{r/}_{i\ q} \ll S^{r/}_{i\ S^{r-1}}$.

**Proof of Proposition 5.** For the linear and quadratic cost functions with any parameter values, we have the relations

$$x > z, \quad X > Z.$$

In the other cases, the nonlinearity coefficients satisfy the following relations: $x < X$ for $S^{r-1} > -1$; $x < y$ for $\beta < 1$; $x < Y$ for $B_1 < \bar{\bar{\bar{B}}}_1$; $x < Z$ for $B_2 < \bar{\bar{\bar{B}}}_2$; $X < Y$ for $\beta > 1$; $X < y$ for $B_1 > \bar{B}_1$; $X < z$ for $B_2 < \bar{B}_2$; $y < Y$ for $B_1 < \bar{\bar{B}}_1$ if $\beta > 1$, or for $B_1 > \bar{\bar{B}}_1$ if $\beta < 1$; $z < Z$ for $B_2 < \bar{\bar{B}}_2$; $y < z$ for $\beta > 1$ and $\frac{B_1}{B_2} > \frac{1}{\lambda}$; $Y < Z$ for $\beta > 1$ and $\frac{B_1}{B_2} > \frac{1}{\lambda}$; $Y < Z$ for $\beta > 1$ and $\frac{B_1}{B_2} > \frac{1}{\lambda}$, where $\bar{\bar{\bar{B}}}_1 = \delta\frac{\chi}{\lambda}$, $\bar{B}_1 = b\frac{\delta}{\lambda}$, $\bar{\bar{\bar{B}}}_2 = \delta\chi$, $\bar{B}_2 = b\delta$, $\bar{\bar{B}}_1 = \frac{\delta}{\lambda}(\chi - b)$, and $\bar{\bar{B}}_2 = \delta\frac{\chi b}{\chi - b}$. Due to (10a), greater values of $u_{km}$ lead to smaller values of $S^r_{ikm}$. Therefore, these relations yield the desired inequalities for $\sigma^0_{km}$.

## REFERENCES

1. Bowley, A.L., *The Mathematical Groundwork of Economics*, Oxford: Oxford Univ. Press, 1924.

2. Jehle, G.A. and Reny, Ph.J., *Advanced Microeconomic Theory*, 2nd. ed., Pearson, 2001.

3. Singh, N. and Vives, X., Price and Quantity Competition in a Differential Duopoly, *Rand J. Econ.*, 1984, vol. 15, pp. 546–554.

4. Daughety, A., Reconsidering Cournot: The Cournot Equilibrium Is Consistent, *Rand J. Econ.*, 1985, vol. 16, pp. 368–380.

5. Julien, L.A., On Noncooperative Oligopoly Equilibrium in the Multiple Leader–Follower Game, *Eur. J. Oper. Res.*, 2017, vol. 256, no. 2, pp. 650–662.

6. Geraskin, M.I., The Properties of Conjectural Variations in the Nonlinear Stackelberg Oligopoly Model, *Autom. Remote Control*, 2020, vol. 81, no. 6, pp. 1051–1072.

7. Kalashnikov, V.V., Bulavsky, V.A., and Kalashnykova, N.I., Existence of the Nash-Optimal Strategies in the Meta-game, *Stud. Syst. Decis. Control*, 2018, no. 100, pp. 95–100.

8. Kalashnykova, N., Kalashnikov, V., Watada, J., Anwar, T., and Lin, P., Consistent Conjectural Variations Equilibrium in a Mixed Oligopoly Model with a Labor-Managed Company and a Discontinuous Demand Function, *IEEE Access*, 2022, vol. 10, pp. 107799–107808.

9. Aizenberg, N.I., Zorkaltsev, V.I., and Mokryi, I.V., A Study into Unsteady Oligopolistic Markets, *J. Appl. Industr. Math.*, 2017, vol. 11, no. 1, pp. 8–16.

10. Algazin, G.I. and Algazina, Y.G., To the Analytical Investigation of the Convergence Conditions of the Processes of Reflexive Collective Behavior in Oligopoly Models, *Autom. Remote Control*, 2022, vol. 83, no. 3, pp. 367–388.

11. Fedyanin, D.N., Monotonicity of Equilibriums in Cournot Competition with Mixed Interactions of Agents and Epistemic Models of Uncertain Market, *Proc. Comp. Sci.*, 2021, vol. 186, pp. 411–417.

12. Lo, C.F. and Yeung, C.F., Quantum Stackelberg Oligopoly, *Quant. Inform. Proc.*, 2022, vol. 21, no. 3, p. 85.

13. Ougolnitsky, G. and Gorbaneva, O., Sustainability of Intertwined Supply Networks: A Game-Theoretic Approach, *Games*, 2022, vol. 13, no. 3, p. 35.

14. Ougolnitsky, G.A. and Usov, A.B., The Interaction of Economic Agents in Cournot Duopoly Models under Ecological Conditions: A Comparison of Organizational Modes, *Autom. Remote Control*, 2023, vol. 84, no. 2, pp. 175–189.

15. Filatov, A.Yu., The Heterogeneity of Firms Behavior at Oligopolistic Market: Price-Makers and Price-Takers, *Bullet. Irkutsk State Univ. Ser. Math.*, 2015, vol. 13, pp. 72–83.

16. Cornes, R., Fiorini, L.C., and Maldonado, W.L., Expectational Stability in Aggregative Games, *J. Evolut. Econom.*, 2021, vol. 31, no. 1, pp. 235–249.

17. Geras'kin, M.I. and Chkhartishvili, A.G., Structural Modeling of Oligopoly Market under the Nonlinear Functions of Demand and Agents' Costs, *Autom. Remote Control*, 2017, vol. 78, no. 2, pp. 332–348.

18. Kanieski da Silva, B., Tanger, S., Marufuzzaman, M., and Cubbage, F., Perfect Assumptions in an Imperfect World: Managing Timberland in an Oligopoly Market, *Forest Policy Econ.*, 2022, vol. 137, p. 102691.

19. Zhou, X., Pei, Z., and Qin, B., Assessing Market Competition in the Chinese Banking Industry Based on a Conjectural Variation Model, *China and World Economy*, 2021, vol. 29, no. 2, pp. 73–98.

20. Novikov, D.A. and Chkhartishvili, A.G., *Reflexion and Control: Mathematical Models*, Leiden: CRC Press, 2014.

21. Novshek, W., On the Existence of Cournot Equilibrium, *Rev. Econ. Stud.*, 1985, vol. 52, pp. 85–98.

*This paper was recommended for publication by D.A. Novikov, a member of the Editorial Board*

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 7, pp. 725–737.
© The Author(s), 2024 published by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 2024.
Russian Text © The Author(s), 2024, published in Avtomatika i Telemekhanika, 2024, No. 7, pp. 113–130.

# "Pitfalls" of Bio-Inspired Models on the Example of Ant Trails

**I. P. Karpova**[*,a] **and V. E. Karpov**[**,b]

[*]*HSE University, Moscow, Russia*
[**]*National Research Centre "Kurchatov Institute", Moscow, Russia*
*e-mail:* [a]*karpova_ip@mail.ru,* [b]*karpov-ve@yandex.ru*

Received November 30, 2023
Revised April 18, 2024
Accepted April 30, 2024

**Abstract**—This paper explores the problem of influencing the environment by a group of autonomous robots through the creation and use of road infrastructure. The model object is ant roads (trails). We identify the main aspects of the behavior of different ant species in the process of collective foraging, and actions that together lead to the appearance of a phenomenon that the observer perceives as an ant road. We develop and describe an animat behavior model in the process of arranging a route. We define a list of mechanisms, a set of sensory capabilities, and effectors that are necessary for the robot to implement options for arranging the route. The results of simulation modeling for solving the foraging problem with route clearing are consistent with theoretical models. The simulation results confirm our assumption that the route arrangement can be carried out by individual efforts of animats (robots) and without the need to organize joint actions.

*Keywords*: social behavior models, collective robotics, autonomous mobile robots, bio-inspired models, foraging task

**DOI:** 10.31857/S0005117924070074

## 1. INTRODUCTION

Usually, the task of moving autonomous agents (robots) is solved by methods of constructing an optimal or suboptimal route. If the robot has an environmental map and the starting and ending points of the route are determined, then various optimization methods are used to solve the problem [1]. If there is no map, the robot either pre-builds this map (SLAM methods, Simultaneous Localization and Mapping), or the map is constructed with the target point search [2]. Some of the methods used take into account changes in the environment, but practically nowhere is the problem that a robot can change this environment itself considered. For example, if there is an obstacle that prevents the robot from moving in the right direction, it can bypass the obstacle or remove it if it is movable and the robot has manipulators to move it. Here the problem arises of finding a balance between the costs of bypassing the obstacle and clearing the road.

There are two aspects to this work. On the one hand, the behavior of an agent moving along the route will be considered from the point of view of the bio-inspired methods application. On the other hand, this problem is used to raise the question of the technical inexpediency of modeling the external, phenomenological side of animal behavior instead of identifying and implementing the basic mechanisms of their behavior.

The use of social behavior models (SBM) is one of the approaches to solving complex tasks of group robots control in difficult non-deterministic environments. SBM consider the socio-inspired organization of robot interaction as one of the adaptive mechanisms that allows solving group tasks in complex dynamic environments where the use of centralized control methods is difficult or impossible [3, 4]. Here the term "society" is considered as an exclusively biological concept.

Ants are a prime example of social animals. They form so-called *eusocial communities* — the most complex form of social organization. This community is characterized by such signs as the presence of a territory assigned to the group, a permanent composition of the group, cohesion (the desire of group members to stick together), individuals specialization, etc. [5, p. 109].

The use of SBM to solve group robotics tasks is as follows: formalization of various behavioral patterns of social animals; development of mechanisms and algorithms that implement these models; creation of software and technical solutions based on them, allowing to perform applied tasks of group robots control. Previous research in the field of SBM has focused on the interaction of agents with the environment and with each other. Now the issue of agents influencing the environment is becoming relevant.

SBM belong to bio-inspired approaches, therefore, the issues of methodology for creating bio-inspired models are important.

## 2. METHODOLOGICAL ASPECTS

There are two extremes in the field of bio-inspired models related to the modeling of social behavior of animals. The first is to create artificial models "inspired by nature." These are Ant Colony Optimization algorithms [6], Grey Wolf Optimizer [7], Butterfly Optimization Algorithm [8] and similar stochastic algorithms [9], optimizing the search in the solution space. For example, the Ant Colony Optimization algorithm is based on the concept of a pheromone trace [6]. Some ant's species leave an odorous trail: a pheromone that evaporates over time. Foraging ants move in the direction of increasing the intensity of this odor when searching for resources. The pheromone analogue serves as the basis for gradient search in a solution space. Another example is the Gray Wolf Optimizer algorithm (GWO). The GWO mimics the hierarchy of individuals and hunting mechanism of grey wolves in nature [7]. This algorithm contains many biological terms, but in fact "individuals" are solutions in a space that is described by some non–monotonic function. At each step, the algorithm evaluates three best solutions, then the "individuals" are shifted in space according to certain rules, the solutions are evaluated again under the assumption that the best solution is located in the geometric center of the "dominant" individuals. Thus, this algorithm is a completely artificial mechanism, for which only the general principle of dividing individuals by weight in decision-making is taken from nature. Such methods solve optimization problems well, but have little to do with the actual behavior of living organisms.

The other extreme is an attempt to simulate natural mechanisms in the form in which they are observed in nature (a phenomenological approach). The authors take certain natural phenomena, more precisely, a description of human observations of these phenomena, and model their external effects. This approach leads to the emergence of numerous realizations of these phenomena. For example, in [10, 11], the authors describe an implementation of ant foraging, which is a complex behavior and involves the search and transportation of food. Other authors even propose "generalized approaches" to such modeling [12]. But all these solutions are particular ways of solving specific problems.

One of the most striking examples is aggression or agonistic behavior. Quite a long time ago, biologists proposed to consider aggression as an external manifestation of certain social behavior types, such as parental, nutritional, etc. [13]. But so far, aggression is explicitly or implicitly declared a basic mechanism or a separate behavior type (see [14, 15]). However, this phenomenon can be realized with the help of other basic elements [16].

In contrast to the above approaches, the SBM paradigm assumes that any complex social behavior or phenomenon consists of a small number of basic mechanisms. This corresponds to the approach of M.L. Tsetlin school in the field of collective behavior [17]. To model behavior, it is necessary to understand what basic elements it consists of and how the observed effects arise, not to introduce unnecessary entities, but to use a combination of basic mechanisms to implement any behavior. This has both a technical and a pragmatic rationale.

As an example, let's give the phenomenon of leadership. Leadership is just some observable phenomenon. The individual does not have a special "leader" block, and insects do not have specific tasks related to dominance. It's just that the behavior of an individual depends on the presence or absence of other individuals nearby. If there is someone stronger (larger, well-fed, etc.) nearby, then the individual begins to behave like a subordinate — it follows the leader. Following a leader is understood in a broad sense: both movement and imitation of the leader's actions in the end. This is one of the basic models of social behavior [5, 18]. If there is no one more successful or stronger nearby, then the individual itself becomes someone whom others perceive as a leader. The individual itself does not perform any "leadership functions", but continues to do its own actions — construction, foraging, etc. And we observe the effect of self-organization: individuals next to the "leader" begin to perform similar work.

Another good example is the phenomenon of ant roads. This is a well-established term often used in literature [19–21]. In this paper, we will try to show that the term "ant trails" appeared as a result of human interpretation of the observed actions of individual ants and their groups, which are usually performed during foraging.

Trail construction is considered one of the most interesting examples of ants working together, which has an impact on the habitat. There are many descriptions of how roads arise, how they are maintained in working order for a long time, etc. Therefore, the desire to model this mechanism is justified, in particular, in order to efficiently use resources or minimize energy consumption during movement. However, an attempt at formalization leads to a rethinking of the road phenomenon in determining the mechanisms underlying it.

The first question is: what are we really observing? The answer to this question determines which models and mechanisms need to be implemented. Strictly speaking, a trail is not only an element of infrastructure. A trail or road is a concept included in the agent's knowledge base about his environment, part of the so-called world model. From the point of view of semiotics as the science of sign systems, which determines the applied aspect of the knowledge representation form, this concept should include the image (perception of the sign), ways of using it (meaning of the sign) and influence goal setting (personal meaning of the sign) [22]. Formally, the sign $S$ describes the entities or concepts of the agent's world. It can be represented as follows: $S =< n, p, a, m >$, where $n$ is the sign name, $p$ is the perception (description or set of characteristics), $m$ is the sign meaning (procedures related to the concept), $a$ is the personal meaning (the component responsible for goal setting).

On the other hand, there is the concept of *a route*. This is a fundamentally different entity. This is an observable external; it does not have to be part of the agent's world model. As will be seen later, the "road" activity of social insects can be reduced to the arrangement of routes. We will understand the route arrangement as a set of actions performed on the area through which the route runs, and aimed at changing its physical characteristics in order to reduce the energy costs of passing the route.

## 3. TRAILS AND ROUTES

The road aspect is very interesting for group robotics (GR). GR solves practical tasks such as monitoring, reconnaissance, patrolling, etc.; therefore, moving along certain routes plays an essential role. Important mechanisms are not only cooperation and coordination of actions, but also the creation of road infrastructure by robots themselves during self-organization. Researchers do not pay enough attention to the latter issue, although the benefits of clearing a route to increase movement speed, for example, are obvious.

Biologists consider roads to be one of the main structure elements of the protected area of many ant species [20], and the roads construction is a vivid example of collective behavior [23]. But an attempt to find a definition of the ant road did not yield results. Biologists often use terms without giving any definitions at all. For example, one of the prominent Russian researchers A.A. Zakharov gives a classification of ant roads, but does not define the road [19]. Biologists from other countries use similar terms: trails, trail construction [24], infrastructure construction [23]. But they also do not give definitions, at best they give a brief description, for example: “physical trails, i.e. pathways that are cleared of obstacles” [24]. And it gives the impression of “well-established terminology” from the category of “everyone knows what ant roads (trails) are.”

Let’s try to approach this issue from a constructive point of view. We will understand by a road **a strip of land equipped or adapted and used for movement, or the surface of an artificial structure**, i.e. something that is the result of purposeful activity. This is similar to the definition that can be found in official documents.

Such definitions are not constructive for SBM. This is an exclusively external phenomenological description. To implement behavioral models, it is necessary that the road become the essence of the agent’s world model. It should be a kind of sign that has at least two components: a number of signs recognized by the agent (perception) and many related behavioral procedures (meaning). But besides the term “road” there is the concept of a path or route. A route is a way from one point of interest to another. The route is recorded or evaluated by an external observer, i.e. he does not have to be represented in the agent’s world model.

In this work, it is assumed that in the ant world there is no road as a specially created structure, but there is a route, for example, from the nest to the foraging area. There are many confirmations that the concept of “ant road” is a human interpretation of the movement of ants along a certain route. For example, in [21, p. 38] it is said about determining “the direction of forage roads or *forager flows* in cases where there are no permanent roads.” Or in [20] when describing roads deep into the soil: “In the rest of the territory, ants used *ordinary roads which represent a stream of foragers* while actively using trunks and branches of fallen trees to move.”

Let’s further consider the two main *road* phenomena described in insect biology, replacing the term road (trail) with the term **route**.

### *3.1. Route Formation*

Let’s describe the process of forming the ant’s route based on data from [21, p. 10]. First, the scout ants explore the area in search of food resources (for example, aphid habitats). If scouts discover new food source, they return to the nest, mobilizes its nestmates and take them to this place. If it is a renewable resource, foragers begin to visit this place regularly. The route usually does not run in a straight line, but where it is more convenient to walk: partially along fallen branches and tree trunks. But if the route passes over the ground then there are irregularities, small debris, vegetation that interfere with walking. Then the ants begin to arrange the route to make it more “convenient,” reducing energy costs. One removes debris, the other pushes aside soil particles, the third destroys small vegetation. The surface is leveled, and the route sometimes becomes more direct and shorter. As a result of the individual, over time, the same “cleared trail” is formed [24], which the observer sees. People perceive this usually narrow cleared strip along which ants move as a road or trail (Fig. 1). In places, it passes through fallen branches or tree trunks (Fig. 1a) [20–21, 23], then it can only be detected by the ants flow. But on the ground, this “trail” is clearly visible even in the absence of ants on it (Fig. 1b), and the observer may have

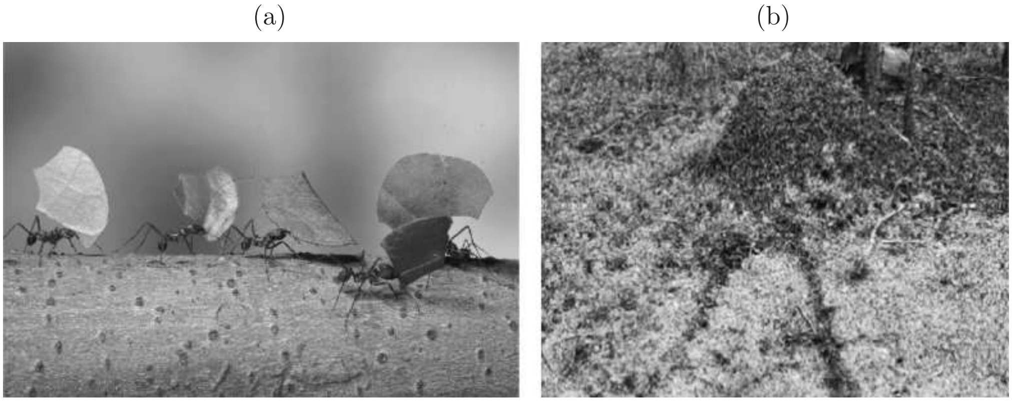


**Fig. 1.** Examples of “ant roads”: (a) a stream of leaf-cutting ants passing along the trunk of a tree; (b) an anthill of a red forest ant with two “roads.”

the impression that he sees a certain construction that appeared as a result of coordinated actions (construction).

It is convenient to call this “construction” a road by analogy with human roads. But for a person, a road is a construction with a certain set of signs that determine human behavior in relation to it. Ants do not have such unambiguous recognition. The experience indirectly confirming this is described by A.A. Zakharov [21, p. 11]. On one of the roads, the researchers seized all foragers and observer ants from the side of the nest adjacent to the road. Thus, there are no ants left in the nest that know this road. The ant family regained possession of the lost part of the territory a few days later, but the original road network and many aphid colonies in the experimental sector were lost. Consequently, other ants who do not know the area could not recognize the roads that exist on it and reuse them.

### *3.2. Clearing the Route*

On the one hand, the ant’s activity in “road maintenance” is energetically beneficial. For example, loaded leafcutter ants travel 4–10 times faster on cleared trails than on uncleaned ones [25], and on average a colony of such ants spends only a few days per season clearing trails [26]. On the other hand, there are studies [23] with the same leafcutter ants, confirming the hypothesis that there are no feedback mechanisms between individuals, nor recruitment mechanisms specifically for clearing the trail. The mathematical model [23] shows that the results of trail clearing experiments can be explained by a fixed probability that the forager will eliminate the interfering obstacle, and this does not require the mobilization of other ants.

So, using the concept of a route, we can explain all the observed phenomena of the appearance, use and support of the “road ants infrastructure,” more precisely, the rational use of the territory.

This statement may seem unnecessarily categorical. We will not delve into terminology issues, we will talk about tunnels [27], roads deep into the soil [20], etc. Of course, environmental change affects the nature of the agent’s behavior: he will *preferably* move along a convenient, well-trodden area. In this sense, the tunnel is the ultimate case of such a “preference”, because in the tunnel the ant does not have the opportunity to choose another path. Let us repeat that if we consider the road not as the essence of the world model of an external observer, but as the essence (sign) of the agent’s world model, then behavioral procedures (the sign meaning), its image (perception), and meaning (explicit meaning from the point of view of the objective function) should be associated with such an entity-sign. But exactly all this is not observed in the agent’s behavior.

## 4. THE ROUTE ARRANGEMENT

Based on the above, we will assume that ants do not have a separate type of activity for the construction and maintenance of road infrastructure. Individual foraging ants, moving along a familiar route, perform some actions. The result of these actions is perceived by an outside observer as a road. These actions are auxiliary. Therefore, the solution of the task of arranging the route should, if possible, be performed using methods and mechanisms that have previously been implemented and have already been used for other tasks, and not introduce entities beyond what is necessary.

The agent must move from one point to another to solve various tasks: patrolling, foraging, etc. Here, the task of movement is solved not at the level of route planning and building an optimal trajectory, but at the behavioral level (like ants). Let's take foraging as an example. Foragers regularly go to the food source and move it to the "base," and the route connects the base and the source location. The agent does not have an environmental map (like the ant [28]): he remembers the route by visual landmarks and compass [29, 30]. During the movement, the agent remembers landmarks and approximate compass direction, so his route is a set of relatively straight segments from one landmark to another (Fig. 2).

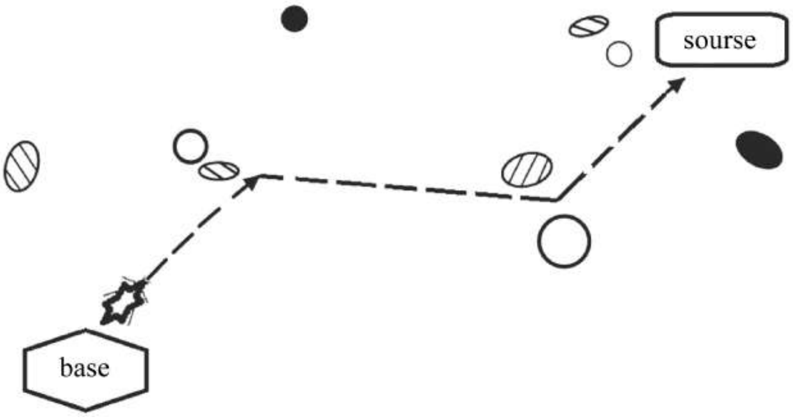


**Fig. 2.** Example of a route from a "base" to a source.

**Movement characteristics.** The direction of the agent's movement is determined by the local goal and context at every time. A local goal for ants can be a visible landmark, a pheromone trail, polarized light data. The current context is the state of surface, obstacles, etc. The context determines the characteristics of movement, obstacle avoidance, local preferences, etc. In this sense, movement is the "resultant" of tendencies to go in the right direction, as well as to go in such a way that it is "more convenient." This is precisely the effect of the "trodden" path on the agent's movement. The more intense the flow, the more the path is "trampled," the more preferable it will be to move along it. But such a "well–equipped" route is not a road, it's just "more convenient" to go that way.

The efficiency of movement is determined by energy consumption, which depends on the time of movement along the route. The time depends on the total length and curves of the route: the agent moves in a straight line faster than when turning, in particular, when bypassing an obstacle. Thus, removing obstacles and straightening the route will increase energy efficiency. In addition, the number of landmarks that the agent can recognize should be sufficient for steady movement along the route. This increases the probability of successful passage from the base to the source and back.

**The route arrangement** for ants in general may include different actions:

1. Trampling (compaction of the soil without additional effort of ants).
2. Clearing the route (cleaning up debris and vegetation).
3. The use of improvised material to increase the convenience of passing along the route, for example, laying plant debris on a swampy route section.
4. Surface leveling, including digging into the soil.

**The procedure for clearing the route.** When an agent goes along a known route, he knows in which direction he needs to move. If there is an obstacle in front of him that prevents him from walking straight, he can bypass it or move it to the side. To do this, it must recognize obstacles and distinguish between movable and non-movable ones. He can perceive these same obstacles as landmarks.

The clearing procedure determines where and how the obstacles are removed. The main question is not in which direction or how far the interfering objects are moving. Difficulties arise when obstacles shift into piles or shafts, forming new landmarks. In fact, the creation of a new landmark means the agent's explicit impact on the environment.

Note that using displaced obstacles as new landmarks is not the same as reacting to a pheromone trail. On the one hand, if an agent detects such a landmark, it cannot identify it as a landmark along the path. On the other hand, the pheromone in ants is a label that is perceived by ants as a sign that another ant's route is passing here. The odorous trail left by the scout can be used by foragers for self-mobilization: they begin to move along this trail in the direction from the nest (they know its location) [31]. In the case of moving an obstacle, only the agent who walks along the route known to him can remember this obstacle as a landmark.

**Participants in the clearing.** Any forager can become a participant in the clearing, since there are no special clearing ants [32]. The probability that an ant will start clearing depends on its condition. Foragers carrying cargo are never engaged in clearing (for leafcutter ants, see [23]). There is also an estimate of the probability that an ant, when faced with an obstacle, will eliminate it.

So, clearing a route leads to its opening, a reduction in its length and to the creation of additional landmarks along the route.

## 5. THE AGENT'S BEHAVIOR MODEL FOR ROUTE ARRANGEMENT

Biologists' research confirms that the route arrangement in ants is energetically efficient [25, 26]. Consequently, the assessment of the agent's actions during foraging can also be based on changes in energy costs. Here is a model describing this process and allowing evaluation of the effectiveness of the agent. This is a simplified qualitative model; it does not aim to describe all the details of the route clearing process.

Suppose there is an agent solving the problem of transporting some resource ("food") from the foraging area to the base. The amount of "food" determines the positive contribution to its energy balance. The agent expends energy on traversing the route from the base to the source location, as well as clearing the route from obstacles. Let's assume that the agent functions in such a discrete time, where each clock cycle can determine a certain period of its existence. The environment in which the agent operates is determined by a limited amount of non-reproducible resource in the foraging area, as well as many obstacles along the route. The agent can remove these obstacles with some probability, reducing the route length, but at the same time spending some of its energy on cleaning.

The task is to evaluate energy efficiency as a function of the agent energy consumption, which depends on the properties of the medium and the probability that the agent will clear the route.

Let $f(t)$ is the delivered resource, and $C(t)$ is the cost of delivering the resource. The delivered resource is determined by the agent's load capacity and in the simple case $f(t) = f(0) = \text{const}$. All the values used are dimensionless and are defined as the energy received or spent in conventional units, and the time $t$ is discrete. Then the effectiveness of agent $E(t)$ at time $t$ can be determined as follows:

$$E(t) = f(t) - C(t). \tag{1}$$

The cost of delivering the resource $C(t)$ (1) consists of the cost of completing the route $L(t)$, clearing work $W(t)$ and searching for food on the area $C_f(t)$:

$$C(t) = L(t) + W(t) + C_f(t). \tag{2}$$

The cost of completing the route $L(t)$ (2) depends on the distance $L_0$ between the "base" and the "foraging area" and on the saturation of obstacles:

$$L(t) = L_0 + k_L\rho(t). \tag{3}$$

Here $L_0$ is an approximately direct route, $L_0 = \text{const}$; $\rho(t)$ is the saturation of obstacles; $k_L$ is the coefficient determining the cost of bypassing the obstacle.

The cost of clearing work $W(t)$ (2) depends on the saturation of obstacles $\rho(t)$ and on the $p_w$ is probability that the agent will remove the obstacle, $p_w = \text{const}$:

$$W(t) = p_w\rho(t). \tag{4}$$

The cost of searching for food $C_f(t)$ (2) is inversely proportional to the amount of food in the area:

$$C_f(t) = k_F/(F(t) + \epsilon). \tag{5}$$

Here $k_F$ is coefficient of the searching cost, $k_F \in \mathbb{R}$, and $\epsilon$ is introduced so that when $F(t) = 0$ the costs would be finite, $\epsilon \in \mathcal{R}$, $\epsilon > 0$. The amount of food in the area $F(t)$ (5) decreases with time:

$$F(t) = F_0 - ft. \tag{6}$$

Here $F_0$ is the initial amount of food on the area. The saturation of obstacles $\rho(t)$ (4) also decreases as the clearing progresses:

$$\rho(t+1) = \rho(t) - k_wW(t) = \rho(t) - k_wp_w\rho(t) = \rho(t)(1 - k_wp_w) \tag{7}$$

or, in the end:

$$\rho(t) = \rho_0(1 - k_wp_w)^t. \tag{8}$$

Here $k_w$ is the coefficient of actions' effectiveness to remove an obstacle. As a result, we get an expression for the agent effectiveness:

$$E(t) = \frac{F_0f - f^2t + f\epsilon - k_F}{F(t) + \epsilon} - L_0 - \rho_0(1 - k_wp_w)^t(k_L + p_w). \tag{9}$$

Obviously, the function determining the amount of resource in the area $F(t)$ must be redefined so that it is bounded from below by zero. If $F(t) = 0$, then the value of $f(t)$ is reset (the agent does not bring anything):

$$F(t) = \max(F_0 - ft, 0), \quad f(t) = \begin{cases} f_0, & \text{if } F(t) - f_0 > 0, \\ 0 & \text{else.} \end{cases} \tag{10}$$

This model allows us to evaluate the effectiveness of the agent's actions during foraging and build a qualitative graph for $E(t)$ (10). Figure 3 shows graphs of $E(t)$ and obstacle saturation $\rho(t)$ (7)

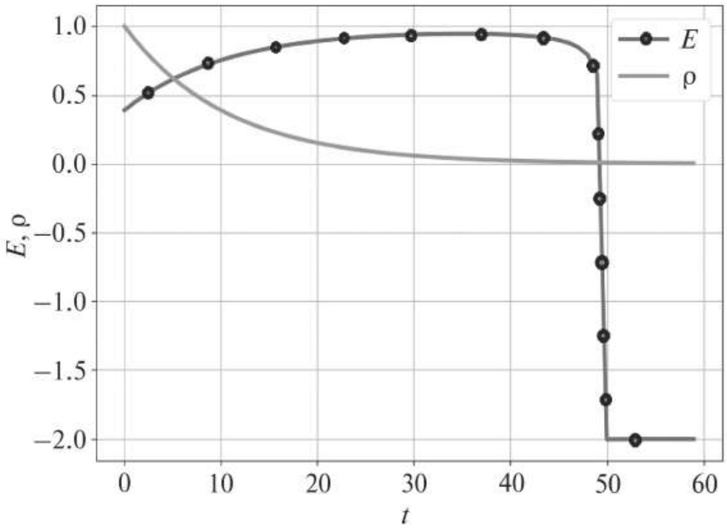


**Fig. 3.** Graphs of the effectiveness of the agent's actions $E(t)$ and the saturation of obstacles $\rho(t)$ in numerical modeling.

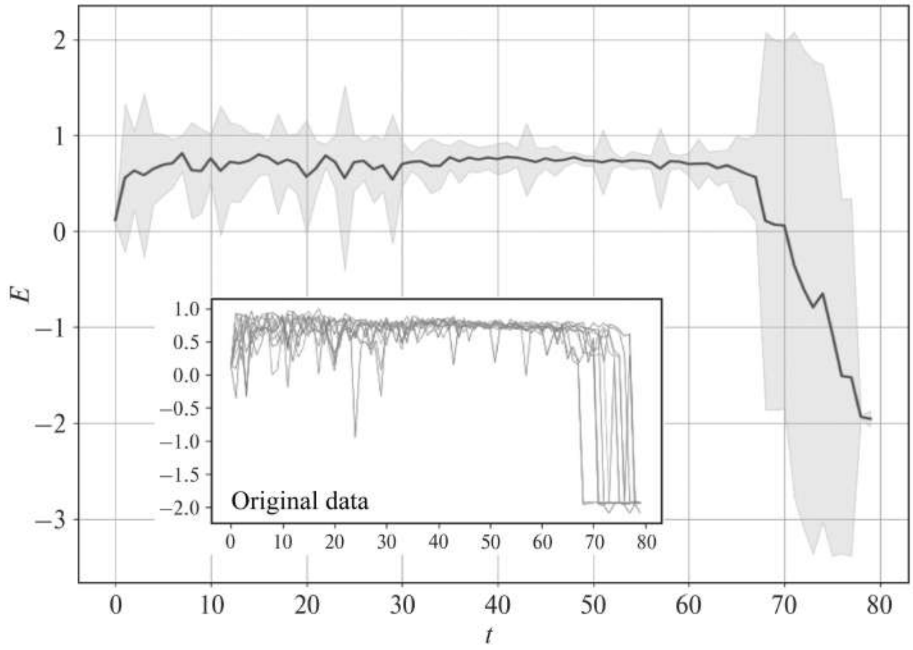


**Fig. 4.** Graphs of the effectiveness of the agent's actions $E(t)$ in simulation modeling. The average of 10 experiments and the standard deviation.

for the following parameter values: the cost of passing a direct route $L_0 = 1$; the probability that the agent will remove the obstacle $p_w = 0.1$; the cost factor for bypassing the obstacle $k_L = 0.5$; the initial saturation of obstacles $\rho(0) = 1$; initial amount of food in the area $F(0) = 100$; load capacity of the agent $f = 2$; clearing efficiency coefficient $k_w = 0.9$; cost factor for food search $k_F = 1$, $\epsilon = 1$.

Graph $E(t)$ in Fig. 3 shows that the agent's efficiency increases as the route is cleared, reaches a maximum when the number of obstacles decreases below a certain threshold, but then decreases due to a decrease in the amount of food.

Negative values of $E(t)$ on the graph mean that when $F(t) = 0$, the agent only spends its resources while traveling along the route, not replenishing them (works at a loss).

## 6. SIMULATION MODELING

Simulation modeling was performed using the Kvorum multi-agent modeling system created at the Kurchatov Institute Research Center [33]. The agent moved between two points: from the "nest" to the "foraging area." The time spent on the road was calculated, taking into account the avoidance of obstacles and/or the cost of removing them, and the time spent on the site searching for food. After finding the resource, the agent instantly returned to the "nest" and went back to the foraging area. The experiment ended when the time to search for food exceeded 2000 cycles: then it was believed that the food on the site was over. Figure 4 shows a graph of the $E(t)$ efficiency for a number of simulation experiments. It can be seen from the graph in Fig. 4 that the effectiveness of the agent's actions varies in a similar way to what numerical modeling shows. At first, the effectiveness of the agent's actions increases, since clearing the route leads to a decrease in the time to complete the path. Then comes the stabilization period (35–55 passes, Fig. 4). As the area is depleted, the time to search for food increases, and efficiency decreases.

## 7. MECHANISMS FOR THE IMPLEMENTATION OF ROUTE ARRANGEMENT OPTIONS

**Trampling** (natural compaction of the soil). This process is difficult to implement in the laboratory conditions, but in a natural environment it happens naturally when many robots move in the same way. The assessment of the sufficient number of robots is based on an estimate of the number of active foraging ants: for the Formica family of 500 individuals, it ranges from 15 to 45 individuals [34]. Minimum requirements are imposed on the robot: it must be able to navigate by visual landmarks and compass, memorize the route, return to "base" and repeat the route.

**Clearing the route** (removing obstacles from it). To do this, in addition to the previously listed mechanisms, the robot must be able to: (1) identify obstacles, (2) distinguish movable obstacles from stationary ones, (3) shift or transfer obstacles to the side, (4) return to movement along the route, (5) refine the memorized route, because shifting obstacles changes the configuration of landmarks along the route. In papers [35, 36], a way of navigating by visual landmarks and compass is described, in which the route is remembered approximately. Returning to the route can be implemented as a continuation of the movement "in the same direction," taking into account the landmarks. Therefore, it is necessary to move the obstacle a short distance, sufficient to clear the way and comparable to the size of the robot.

To clear the route, the robot must determine how and when it makes a choice between bypassing and moving an obstacle. After moving the obstacle, the robot must determine further actions: will he follow the route or continue clearing the way.

**Surface alignment** (horizontal alignment) To do this, the robot must have effectors capable of cutting off the top layer of soil or "laying trenches." This is too strong a requirement, but you can limit yourself to movable elements (obstacles) that you can either drive over or go around them. The robot can shift such elements with an effector in the form of a blade: in this case, clearing the route will also lead to alignment.

**The use of improvised material.** This is a more difficult option. First, the robot must be able to determine that there is an area in front of it that is inconvenient for movement, for example, a recess. Secondly, he must find an element nearby that can align this area. But this option can be considered as a continuation of the previous one, by analogy with ants that *shift the soil* to level the surface of the trail [23]. And the robot can move obstacles that interfere with the passage to these inconvenient areas. A general list of mechanisms is given in Table. In it, all the previous mechanisms are needed for each next option.

Mechanisms for implementing options for a route arranging by a robot

| Options for arranging a route | Mechanisms |
|---|---|
| 1. Trampling | Localization by visual landmarks and compass<br>Memorizing the route<br>Returning to the "base"<br>Repeating the route |
| 2. Clearing the route | Obstacle identification<br>Recognition of movable and stationary obstacles<br>Shifting obstacles to the side |
| 3. Surface alignment | Identification of an obstacle that can be skirted from the side or run over |
| 4. The use of improvised material | Identification of an inconvenient area where an obstacle can be moved |

## 8. CONCLUSION

This paper has two aspects — technical and methodological. The first aspect is concerning the behavior modeling and is as follows. The reasoning and simulation results above confirm the assumption that in fact there are no roads in the world of ants in the sense that man puts into this concept. There are only routes and directions of movement, and the route arrangement in order to reduce energy costs in a simple case can be carried out without organizing joint purposeful activities of many agents (robots). Thus, we have shown that without creating new entities, without involving any artificial structures, we can get the same result and observe the same phenomenon, which biologists call "ant roads." Formally, this means the absence of the "road" entity in the agent's world model, which entails the absence of the need to create behavioral procedures corresponding to this sign, representation/recognition, etc. This greatly simplifies the solution of the problem.

The second aspect is methodological. It emphasizes that the attitude towards bio-inspired models should be critical and constructive. Let's go back to the title of the paper and summarize what the "pitfalls" of bio-inspired models are.

1. Superficial analogies in bio-inspired behaviors are a dangerous thing. "Nature-inspired" models often have nothing to do with what is available in nature. The danger lies in the fact that such a superficial view ignores the mechanisms underlying a particular behavior. As a result, specific models are obtained that reflect only the external, phenomenological aspects of natural phenomena.
2. The identification and implementation of basic behavioral mechanisms has purely practical aspects. This saves effort when developing systems, makes it possible to combine these basic mechanisms and provides flexibility. An example of this approach is the paradigm of social behavior models.
3. Real biological models and descriptions of phenomena also require a critical attitude. The point is that biologists and technical specialists use different concepts. The latter should consider important the essence of the phenomenon, as well as its constituent elements and the causal relationships between them. Without this, it is unclear what needs to be modeled. An example of this phenomenon is ant roads, which were discussed in this paper.

## FUNDING

The simulation modelling was performed at the expense of the state assignment of the Kurchatov Institute Research Center.

## REFERENCES

1. Sahoo, S.K. and Choudhury, B.B., A Review of Methodologies for Path Planning and Optimization of Mobile Robots, *J. Process Manag. New Technol.*, 2023, vol. 11, no. 1–2, pp. 34–52.
2. Abaspur Kazerouni, I., Fitzgerald, L., Dooly, G., and Toal, D., A survey of state-of-the-art on visual SLAM, *Expert Syst. Appl.*, 2022, vol. 205, no. 2, p. 117734.
3. Karpov, V.E., Karpova, I.P., and Kulinich, A.A., *Sotsial'nye soobshchestva robotov* (Social communities of robots), Moscow: URSS, 2019.
4. Karpov, V.E., Social Robot Communities: from Reactive to Cognitive Agent, *Myagkie izmereniya i vychisleniya*, 2019, vol. 2, no. 15, pp. 61–78.
5. Dewsbury, D.A., *Comparative Animal Behavior*, New York: McGraw-Hill, 1978.
6. Dorigo, M., Maniezzo, V., and Colorni, A., Ant System: Optimization by a Colony of Cooperating Agents, *IEEE Trans. Syst. Man, Cybern. Part B (Cybernetics)*, 1996, vol. 26, no. 1, pp. 29–41.
7. Mirjalili, S., Mirjalili, S.M., and Lewis, A., Grey Wolf Optimizer, *Adv. Eng. Softw.*, 2014, vol. 69, pp. 46–61.
8. Arora, S. and Singh, S., Butterfly optimization algorithm: a novel approach for global optimization, *Soft Comput.*, 2019, vol. 23, no. 3, pp. 715–734.
9. Toaza, B. and Esztergár-Kiss, D., A review of metaheuristic algorithms for solving TSP-based scheduling optimization problems [Formula presented], *Appl. Soft Comput.*, 2023, vol. 148, no. 11, p. 110908.
10. Imai, K. and Okuyama, A., Research on a Multi-agent System That Mimics Ant Foraging Behavior, in *Lecture Notes in Networks and Systems. Proceedings of Eighth International Congress on Information and Communication Technology ICICT 2023*, London, 2024, vol. 696, pp. 193–203.
11. Zhang, N. and Yong, E.H., Dynamics, statistics, and task allocation of foraging ants, *Phys. Rev. E*, 2023, vol. 108, no. 5, p. 054306.
12. De Nicola, R., Di Stefano, L., Inverso, O., and Valiani, S., Intuitive Modelling and Formal Analysis of Collective Behaviour in Foraging Ants, in *Comput. Meth. in Syst. Biol.*, 2023, pp. 44–61.
13. Lorenz, K., *On Aggression*, London: Routledge, 2002.
14. Kudryavceva, N.N., Markel', A.L., and Orlov, Yu.L., Aggressive behavior: genetic and physiological mechanisms, *Vavilovskii zhurnal genetiki i selektsii*, 2014, vol. 18, no. 4/3, pp. 1133–1155.
15. Nordell, S.E. and Valone, T.J., *Habitat Selection, Territoriality, and Aggression, Animal Behaviour*, Oxford University Press, 2021, p. 476.
16. Karpova, I.P. and Karpov, V.E., Aggression in the animats world, or about some mechanisms for aggressive behavior control in group robotics, *Upravlenie Bol'shimi Sistemami*, 2018, vol. 76, pp. 173–218.
17. Tsetlin, M.L., *Issledovaniya po teorii avtomatov i modelirovaniyu biologicheskikh sistem* (Research on automata theory and modeling of biological systems), Moscow: Nauka, 1969.
18. Karpov, V.E., Models of social behaviour in the group robotics, *Upravlenie Bol'shimi Sistemami*, 2016, vol. 59, pp. 165–232.
19. Zakharov, A.A., Ant Roads (terminology issues), "Ants and forest protection": Proc. of the VI All-Union Myrmecological Symposium, Tartu, 1979, pp. 152–155.
20. Novgorodova, T.A., Use of natural trenches by ants of the Formica rufa (Hymenoptera, Formicidae), *Evroaziatskii entomologicheskii zhurnal*, 2011, vol. 10(3), no. 3, pp. 401–405.
21. Zakharov, A.A., *Muravej, Sem'ya, koloniya* (Ant, family, colony), Moscow: Nauka, 1978.
22. Osipov, G.S., Panov, A.I., Chudova, N.V., and Kuzneczova, Yu.M., *Znakovaya kartina mira sub"ekta povedeniya* (Significant picture of the world of the subject of behavior), Moscow: Fizmatlit, 2018.
23. Bochynek, T., Burd, M., Kleineidam, C., and Meyer, B., Infrastructure construction without information exchange: the trail clearing mechanism in Atta leafcutter ants, *Proc. R. Soc. B Biol. Sci.*, 2019, vol. 286, no. 1895.
24. Bouchebti, S., Travaglini, R.V., Forti, L.C., and Fourcassi, V., Dynamics of physical trail construction and of trail usage in the leaf-cutting ant Atta laevigata, *Ethol. Ecol. Evol.*, 2019, vol. 31, no. 2, pp. 105–120.

25. Rockwood, L.L. and Hubbell, S.P., Host-plant selection, diet diversity, and optimal foraging in a tropical leafcutting ant, *Oecologia*, 1987, vol. 74, pp. 55–61.

26. Howard, J.J., Costs of trail construction and maintenance in the leaf-cutting ant Atta columbica, *Behav. Ecol. Sociobiol.*, 2001, vol. 49, no. 5, pp. 348–356.

27. Viles, H.A., Goudie, A.S., and Goudie, A.M., Ants as geomorphological agents: A global assessment, *Earth-Science Rev.*, 2021, vol. 213, p. 103469.

28. Wehner, R., Hoinville, T., and Cruse, H., On the 'cognitive map debate' in insect navigation, *Stud. Hist. Philos. Sci.*, 2023, vol. 102, no. August, pp. 87–89.

29. Dall'Osto, D., Fischer, T., and Milford, M., Fast and Robust Bio-inspired Teach and Repeat Navigation, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 500–507.

30. Dupeyroux, J., Viollet, S., and Serres, J.R., An ant-inspired celestial compass applied to autonomous outdoor robot navigation, *Rob. Auton. Syst.*, 2019, vol. 117, pp. 40–56.

31. Dlussky, G.M., Behavioral mechanisms of regulation of foraging in ants, *"Ants and forest protection": Proc. of the VI All-Union Myrmecological Symposium*, Tartu, 1979, pp. 147–151.

32. Alma, A.M., Farji-Brener, A.G., and Elizalde, L., When and how obstacle size and the number of foragers affect clearing a foraging trail in leaf-cutting ants, *Insectes Soc.*, 2019, vol. 66, no. 2, pp. 305–316.

33. Karpov, V.E., Rovbo, M.A., and Ovsyannikova, E.E., A system for modeling the behavior of groups of robotic agents with elements of a social organization Quorum, *Programmnye produkty i sistemy*, 2018, vol. 31, no. 3, pp. 581–590.

34. Malyshev, A. and Burgov, E., Revisiting Parameters of Bio-inspired Behavior Models in Group Foraging Modeling, *SPIIRAS Proceedings*, 2020, vol. 19, no. 1, pp. 79–103.

35. Karpova, I.P., Animate orientation based on visual landmarks and scene recognition, *Mekhatronika, Avtomatizatsiya, Upravlenie*, 2021, vol. 22, no. 10, pp. 537–546.

36. Karpova, I.P., A Bio-inspired Approach to Robot Orientation or a Real "Ant" Algorithm, *Upravlenie Bol'shimi Sistemami*, 2022, vol. 96, pp. 69–117.

*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, 2024, Vol. 85, No. 7, pp. 738–745.*

# Optimization of the Parameters of a Model Predictive Control System for an Industrial Fractionator

**O. Yu. Snegirev**[*,a] **and A. Yu. Torgashov**[*,b]

*Institute of Automation and Control Processes, Far Eastern Branch,*
*Russian Academy of Sciences, Vladivostok, Russia*
*e-mail:* [a]*snegirevoleg@iacp.dvo.ru,* [b]*torgashov@iacp.dvo.ru*

Received September 28, 2023
Revised November 27, 2023
Accepted December 21, 2023

**Abstract**—The problem of parametric synthesis of a model predictive control (MPC) system by the chemical process of production of the kerosene fraction of an industrial fractionator under conditions of constraints and uncertainty is considered. The optimal parameters of the MPC algorithm are obtained as a result of solving the problem of multi-criteria optimization, taking into account the intervally specified parameters of the plant model.

*Keywords*: model predictive control, robust control, multi-criteria optimization

**DOI:** 10.31857/S0005117924070088

## 1. INTRODUCTION

Model predictive control (MPC) has recently developed a lot due to the fact that it has a number of significant advantages in solving the problems of control multidimensional industrial plants in the presence of constraints on control actions [1–3].

When finding the values of control actions at each time step $k$, the optimization problem is solved. The objective function using the predict of $P$ forward time steps ($\tilde{y}_{k+j}$, $j = 1, \ldots, P$) is minimized by selecting the increment values of the control variables $\Delta u$ on the control horizon $M$. The values of the control actions are determined by $M$ ssteps forward, but only the first change is used $\Delta u_k$, i.e. at the current time. After $u_k$ is executed, the measurement of the output variable comes in the next step $y_{k+1}$ and the model error is corrected, since the measured value $y_{k+1}$, as a rule, does not coincide with the forecast value. For a multidimensional system consisting of controlled variables (CV) $n_{CV}$ and manipulated variables (MV) $n_{MV}$, the matrix of the system dynamics is formed from the coefficients of the finite step response (FSR):

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_1 & 0 & \ldots & 0 \\ \mathbf{S}_2 & \mathbf{S}_1 & 0 & \vdots \\ \vdots & \vdots & \ddots & 0 \\ \mathbf{S}_M & \mathbf{S}_{M-1} & \ldots & \mathbf{S}_1 \\ \mathbf{S}_{M+1} & \mathbf{S}_M & \ldots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_P & \mathbf{S}_{P-1} & \ldots & \mathbf{S}_{P-M+1} \end{bmatrix},$$

where $\mathbf{S}_i = \begin{bmatrix} S_{11,i} & S_{12,i} & \dots & S_{1n_{MV},i} \\ S_{21,i} & \dots & \dots & S_{2n_{MV},i} \\ \vdots & \vdots & \vdots & \vdots \\ S_{n_{CV}1,i} & \dots & \dots & S_{n_{CV}n_{MV},i} \end{bmatrix}$ – matrix $n_{CV} \times n_{MV}$ of step response coefficients for the $i$th time step.

First-order transfer function with a delay are mainly used as initial data for predictive models in the form of an FSR:

$$F(s) = \frac{g}{\tau s + 1} e^{-\theta s}.$$

The parameters of the model for a multivariable system can be written as matrices:

$$\hat{\mathbf{G}} = \begin{pmatrix} \hat{g}_{1,1} & \dots & \hat{g}_{1,n_{MV}} \\ \vdots & \ddots & \vdots \\ \hat{g}_{n_{CV},1} & \dots & \hat{g}_{n_{CV},n_{MV}} \end{pmatrix}, \quad \hat{\mathbf{T}} = \begin{pmatrix} \hat{\tau}_{1,1} & \dots & \hat{\tau}_{1,n_{MV}} \\ \vdots & \ddots & \vdots \\ \hat{\tau}_{n_{CV},1} & \dots & \hat{\tau}_{n_{CV},n_{MV}} \end{pmatrix}, \quad \hat{\mathbf{\Theta}} = \begin{pmatrix} \hat{\theta}_{1,1} & \dots & \hat{\theta}_{1,n_{MV}} \\ \vdots & \ddots & \vdots \\ \hat{\theta}_{CV,1} & \dots & \theta_{n_{CV},n_{MV}} \end{pmatrix},$$

where $\hat{\mathbf{G}}$ is the matrix of the gain coefficients, $\hat{\mathbf{T}}$ is the matrix of the time constants, $\hat{\mathbf{\Theta}}$ is the matrix of the delay values.

The elements of the matrix $\mathbf{S}$ are formed on the basis of $\hat{\mathbf{G}}$, $\hat{\mathbf{T}}$ and $\hat{\mathbf{\Theta}}$:

$$\begin{cases} \mathbf{S}_i = 0, & i\Delta t \leqslant \hat{\mathbf{\Theta}} \\ \mathbf{S}_i = \hat{\mathbf{G}} \left(1 - e^{\frac{-\left(i\Delta t - \hat{\mathbf{\Theta}}\right)}{\hat{\mathbf{T}}}}\right), & i\Delta t > \hat{\mathbf{\Theta}}. \end{cases}$$

The control problem can be formulated as an optimization problem with the following objective function [4]:

$$\min_{\Delta \mathbf{U}_k} \mathbf{J} = \hat{\mathbf{E}}_{k+1}^{\mathrm{T}} \mathbf{Q} \hat{\mathbf{E}}_{k+1} + \Delta \mathbf{U}_k^{\mathrm{T}} \mathbf{R} \Delta \mathbf{U}_k$$

s.t.

$$\begin{gathered} \mathbf{y}_{k+j}^{-} \leqslant \tilde{\mathbf{y}}_{k+j} \leqslant \mathbf{y}_{k+j}^{+} \quad (j = 1, \dots, P), \\ \mathbf{u}^{-} \leqslant \mathbf{u}_{k+j} \leqslant \mathbf{u}^{+} \quad (j = 0, 1, \dots, M-1), \\ \Delta \mathbf{u}^{-} \leqslant \Delta \mathbf{u}_{k+j} \leqslant \Delta \mathbf{u}^{+} \quad (j = 0, 1, \dots, M-1), \end{gathered}$$

where $\mathbf{u}_k = \left[u_{1|k}, \dots, u_{n_{MV}|k}\right]^{\mathrm{T}}$ is a vector of MV values at time $k$; $\Delta \mathbf{U}_k = [\Delta \mathbf{u}_k, \dots, \Delta \mathbf{u}_{k+M-1}]$ is a matrix $M$ of changes in MV values at time $k$ ($\Delta \mathbf{u}_k = \left[\Delta u_{1|k}, \dots, \Delta u_{n_{MV}|k}\right]^{\mathrm{T}}$); $\tilde{\mathbf{y}}_{k+j} = \left[\tilde{y}_{1|k+j}, \dots, \tilde{y}_{n_{CV}|k+j}\right]^{\mathrm{T}}$ is a vector of corrected predicted CV values at time $k+j$; $\mathbf{Q}$ and $\mathbf{R}$ are diagonal weight matrices for prioritizing elements $\hat{\mathbf{E}}_{k+1}$ and controlling changes $\Delta \mathbf{U}_k$, respectively.

The predictable error vector $\hat{\mathbf{E}}_{k+1}$ is defined as

$$\hat{\mathbf{E}}_{k+1} = \mathbf{Y}_{k+1}^{ref} - \tilde{\mathbf{Y}}_{k+1},$$

where $\mathbf{Y}_{k+1}^{ref}$ is the vector of given CV values at time $k+1$, $\tilde{\mathbf{Y}}_{k+1}$ is the vector of corrected predicted values:

$$\tilde{\mathbf{Y}}_{k+1} = \mathbf{S} \Delta \mathbf{U}_k + \hat{\mathbf{Y}}_{k+1}^{o} + [\mathbf{y}_k - \hat{\mathbf{y}}_k],$$

where $\hat{\mathbf{Y}}_{k+1}^{o} = \sum_{i=1}^{N-2} \mathbf{S}_{i+1} \Delta \mathbf{u}_{k-i} + \mathbf{S}_N \mathbf{u}_{k-N+1}$ is the vector of forecasts of unforced responses.

This paper considers the MPC algorithm, in which the increments of control actions (MV) are determined analytically [5]:

$$\Delta\mathbf{U}(k) = \mathbf{K}_C\hat{\mathbf{E}}^o(k+1),$$

where $\hat{\mathbf{E}}^o(k+1)$ is the predicted deviations from the initial trajectory with the constancy of the values of future control actions; $\mathbf{K}_C$ is the matrix of the regulator gain, which is calculated as $\mathbf{K}_C = \left(\mathbf{S}^{\mathrm{T}}\mathbf{Q}\mathbf{S} + \mathbf{R}\right)^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{Q}$.

In spite of its advantages, MPC depends on the accuracy of the model, and transients in the control system can deteriorate in the presence of uncertainty, perturbations, and model errors (mismatch of the MPC model with the model of the controlled plant) [6]. In the existing works, in order to compensate for the uncertainty, it is proposed to introduce output predictors into the structure of the control system for various parameters of the plant (family of plants) [7, 8], which increases the complexity of the control system in the case of multidimensional plants, and the number of required computing resources of the MPC algorithm for finding a sequence of optimal increments of control actions increases significantly. In contrast to the known works, in this paper, it is proposed to take into account the uncertainty of the plant at the design stage of the MPC algorithm, i.e. to search for the optimal parameters of the regulator based on the predictive model (weight matrices $\mathbf{Q}$ and $\mathbf{R}$), when the parameters of the plant are set intervally.

## 2. DESCRIPTION OF THE PROCESS UNIT AND FORMULATION OF THE PROBLEM

A fractionation column C-2 is considered (Fig. 1), in which a multicomponent hydrocarbon crude mixture is divided into naphta, kerosene, and other fractions. Column C-2 has an additional side stripping column—a column for stripping the C-3 kerosene fraction. Column C-2 contains 44 valve trays in the rectification section and 12 valve trays in the stripping section. Excess heat in the column is removed by bottom pumparound (BPA). The top temperature (TIC1) is controlled by the supply of reflux in the upper part of the C-2 fractionator. The purpose of column C-3 is the stripping of light hydrocarbons from the kerosene fraction due to the heat of the BPA of column C-2 supplied to the reboiler E-2. Light hydrocarbon vapors from column C-3 are returned to column C-2. The temperature of the product kerosene at the outlet of column C-3 is controlled by the TIC2 loop. The plant in question has $n_{CV} = 2$ and $n_{MV} = 2$. The matrix of transfer

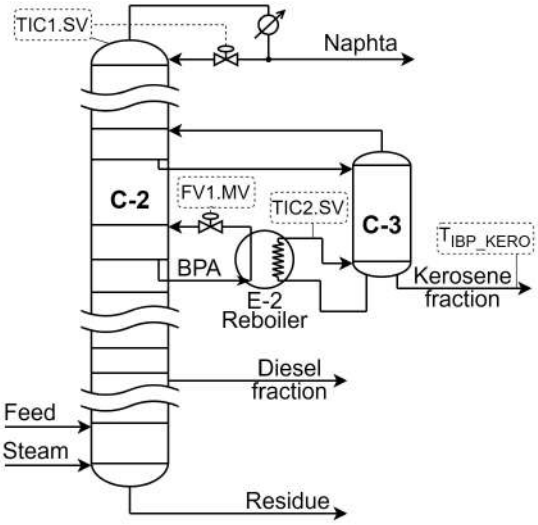


**Fig. 1.** Process unit block diagram.

**Table 1.** Transfer matrix of the plant

| | Top temperature C-2 (TIC1.SV) | Bottom temperature C-3 (TIC2.SV) |
|---|---|---|
| $\mathrm{T_{IBP\_KERO}}$ | $F_{1,1}=\dfrac{g_{1,1}}{\tau_{1,1}s+1}e^{-\theta_{1,1}s}$ | $F_{1,1}=\dfrac{g_{1,2}}{\tau_{1,2}s+1}e^{-\theta_{1,2}s}$ |
| FV1.MV % of valve opening E-2 | $F_{2,1}=\dfrac{g_{2,1}}{\tau_{2,1}s+1}e^{-\theta_{2,1}s}$ | $F_{2,2}=\dfrac{g_{2,2}}{\tau_{2,2}s+1}e^{-\theta_{2,2}s}$ |

functions of the plant is presented in Table 1. Transfer functions are aperiodic links of the 1st order with a delay.

The control task is to maintain the initial boiling point of the kerosene fraction ($\mathrm{T_{IBP\_KERO}}$) within a set range.

## 3. DETERMINATION OF THE OPTIMAL PARAMETERS OF THE REGULATOR BASED ON THE PREDICTIVE MODEL FOR QUALITY CONTROL OF THE INDUSTRIAL FRACTIONATOR PRODUCT

Let us denote the parameters of the transfer functions of the plant in the form of the following matrices:

$$\mathbf{G}=\begin{pmatrix} g_{1,1} & g_{1,2}\\ g_{2,1} & g_{2,2}\end{pmatrix},\quad \mathbf{T}=\begin{pmatrix} \tau_{1,1} & \tau_{1,2}\\ \tau_{2,1} & \tau_{2,2}\end{pmatrix},\quad \mathbf{\Theta}=\begin{pmatrix} \theta_{1,1} & \theta_{1,2}\\ \theta_{2,1} & \theta_{2,2}\end{pmatrix}.$$

Matrices of parameters of the transfer functions of the regulator (to find the FSR):

$$\hat{\mathbf{G}}=\begin{pmatrix} \hat{g}_{1,1} & \hat{g}_{1,2}\\ \hat{g}_{2,1} & \hat{g}_{2,2}\end{pmatrix},\quad \hat{\mathbf{T}}=\begin{pmatrix} \hat{\tau}_{1,1} & \hat{\tau}_{1,2}\\ \hat{\tau}_{2,1} & \hat{\tau}_{2,2}\end{pmatrix},\quad \hat{\mathbf{\Theta}}=\begin{pmatrix} \hat{\theta}_{1,1} & \hat{\theta}_{1,2}\\ \hat{\theta}_{2,1} & \hat{\theta}_{2,2}\end{pmatrix}.$$

Set the upper and lower limits of the parameter ranges:

$$\overline{\mathbf{G}}=\begin{pmatrix} 0.2 & 0.9\\ -0.5 & 0.9\end{pmatrix},\quad \overline{\mathbf{T}}=\begin{pmatrix} 18 & 18\\ 20 & 14\end{pmatrix},\quad \overline{\mathbf{\Theta}}=\begin{pmatrix} 8 & 7\\ 9 & 7\end{pmatrix},$$

$$\underline{\mathbf{G}}=\begin{pmatrix} 0.1 & 0.5\\ -1 & 0.45\end{pmatrix},\quad \underline{\mathbf{T}}=\begin{pmatrix} 6 & 6\\ 8 & 6\end{pmatrix},\quad \underline{\mathbf{\Theta}}=\begin{pmatrix} 2 & 3\\ 1 & 1\end{pmatrix}.$$

Let us assume that the actual parameters of the transfer functions of the plant lie in the middle of the specified range:

$$\mathbf{G}=\frac{\overline{\mathbf{G}}+\underline{\mathbf{G}}}{2}=\begin{pmatrix} 0.15 & 0.7\\ -0.75 & 0.675\end{pmatrix},\quad \mathbf{T}=\frac{\overline{\mathbf{T}}+\underline{\mathbf{T}}}{2}=\begin{pmatrix} 12 & 12\\ 14 & 10\end{pmatrix},\quad \mathbf{\Theta}=\frac{\overline{\mathbf{\Theta}}+\underline{\mathbf{\Theta}}}{2}=\begin{pmatrix} 5 & 4\\ 6 & 4\end{pmatrix}.$$

To study the transient processes in the control system, we set the vector of tasks by CV $\mathbf{r}=[r_1 \quad r_2]=[1 \quad 0.8]$. The adjusting parameters of the regulator are the matrix of weights by error CV $\mathbf{Q}=\operatorname{diag}\{Q_1,Q_2\}$ and the matrix of weights by increment MV $\mathbf{R}=\operatorname{diag}\{R_1,R_2\}$. In the course of the study, it was established that the quality of regulation depended not so much on the values of the weights $\mathbf{Q}$ and $\mathbf{R}$ as on the ratio of the weights relative to each other. Therefore, within the framework of this study, the weights $\mathbf{Q}=\operatorname{diag}\{Q_1,Q_2\}$ are set.

As a criterion for the accuracy of control tasks according to CV, the mean square error relative to the desired dynamics was chosen [9]:

$$J=\sum_{i=1}^{N_M}\sum_{q=1}^{n_{CV}}\left(y_{i,q}^{ref}-y_{i,q}\right)^2,$$

where $y_{i,q}^{ref}$ is the value of the desired trajectory $q$ CV at the $i$th time point, $y_{i,q}$ is the actual value of $q$ CV at the $i$th time point.

Thus, the optimization problem can be written as:

$$\min_{\mathbf{R}>0} J = \sum_{i=1}^{N_M} \sum_{q=1}^{n_{CV}} \left( y_{i,q}^{ref} - y_{i,q} \right)^2,$$

where $\mathbf{R} > 0$ means that all diagonal elements are positive. The calculation of the desired trajectory is made according to the following expression:

$$y_{i,q}^{ref} = \frac{\mathbf{r}_q \sum_{j=1}^{n_{MV}} \mathbf{G}_{q,j} \left( 1 - e^{\frac{\tilde{t}}{\mathbf{T}_{q,j}}} \right)}{\sum_{j=1}^{n_{MV}} |\mathbf{G}_{q,j}|}, \qquad \tilde{t} = \begin{cases} i - \mathbf{\Theta}_{q,j}, & i \geqslant \mathbf{\Theta}_{q,j} \\ 0, & i < \mathbf{\Theta}_{q,j}, \end{cases}$$

where $i = 1, \ldots, N_M$ are the time points, and $q = 1, \ldots, n_{CV}$ is the CV number for which the desired trajectory is calculated. The purpose of this study is to find such values of the weights $\mathbf{R} = \operatorname{diag}\{R_1, R_2\}$ that the output variables of the plant are as close as possible to the desired dynamics for various parameters $\hat{\mathbf{G}}$, $\hat{\mathbf{T}}$ and $\hat{\mathbf{\Theta}}$, lying within the specified range.

To determine the robust optimal values of $\mathbf{R}$, consider cases where one of the parameters $\hat{\mathbf{G}}$, $\hat{\mathbf{T}}$ and $\hat{\mathbf{\Theta}}$ lies at the boundary of the ranges, and the rest are in the middle. The cases under consideration are presented in Table 2.

**Table 2.** Controller model parameter variations

| Index $\tilde{\mathbf{p}}$ | Parameter value | | | Index $\tilde{\mathbf{p}}$ | Parameter value | | |
|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{G}}$ | $\hat{\mathbf{T}}$ | $\hat{\mathbf{\Theta}}$ | | $\hat{\mathbf{G}}$ | $\hat{\mathbf{T}}$ | $\hat{\mathbf{\Theta}}$ |
| 1 | $\underline{\mathbf{G}}$ | $\mathbf{T}$ | $\mathbf{\Theta}$ | 4 | $\mathbf{G}$ | $\overline{\mathbf{T}}$ | $\mathbf{\Theta}$ |
| 2 | $\overline{\mathbf{G}}$ | $\mathbf{T}$ | $\mathbf{\Theta}$ | 5 | $\mathbf{G}$ | $\mathbf{T}$ | $\underline{\mathbf{\Theta}}$ |
| 3 | $\mathbf{G}$ | $\underline{\mathbf{T}}$ | $\mathbf{\Theta}$ | 6 | $\mathbf{G}$ | $\mathbf{T}$ | $\overline{\mathbf{\Theta}}$ |

The optimization problem in a general form for each of the cases under consideration can be represented as:

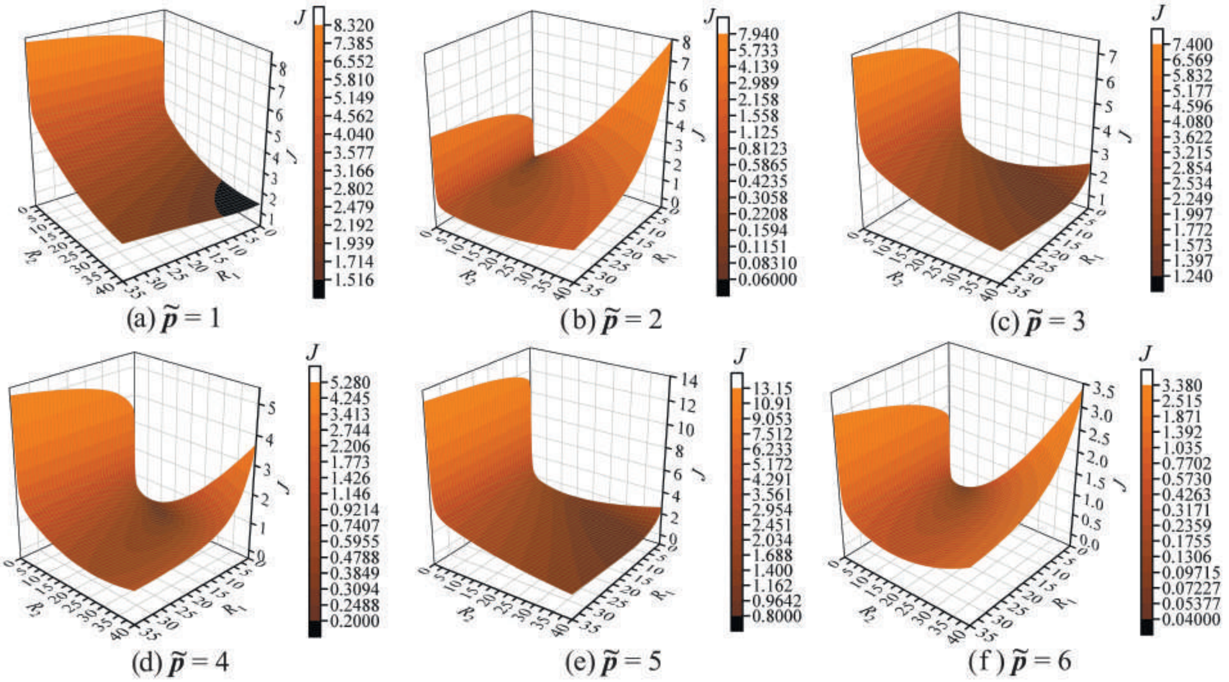
$$\min_{\mathbf{R}>0} J^{\tilde{\mathbf{p}}} = \sum_{i=1}^{N_M} \sum_{q=1}^{n_{CV}} \left( y_{i,q}^{ref} - \left( y_{i-1,q} + \mathbf{S}_i^{\tilde{\mathbf{p}}} \left( \mathbf{S}_i^{\tilde{\mathbf{p}}\mathrm{T}} \mathbf{Q} \mathbf{S}_i^{\tilde{\mathbf{p}}} + \mathbf{R} \right)^{-1} \mathbf{S}_i^{\tilde{\mathbf{p}}\mathrm{T}} \mathbf{Q} \hat{\mathbf{E}}^o (k+1) \right) \right)^2,$$

where $\begin{cases} \mathbf{S}_i^{\tilde{\mathbf{p}}} = 0, & i\Delta t \leqslant \hat{\mathbf{\Theta}}^{\tilde{\mathbf{p}}} \\ \mathbf{S}_i^{\tilde{\mathbf{p}}} = \hat{\mathbf{G}}^{\tilde{\mathbf{p}}} \left( 1 - e^{-\frac{i\Delta t - \hat{\mathbf{\Theta}}^{\tilde{\mathbf{p}}}}{\hat{\mathbf{T}}^{\tilde{\mathbf{p}}}}} \right), & i\Delta t > \hat{\mathbf{\Theta}}^{\tilde{\mathbf{p}}}. \end{cases}$

To determine the optimal (in this case, robust) parameters of the MPC algorithm, we will vary the values $\mathbf{R} = \operatorname{diag}\{R_1, R_2\}$ in the range $0.1 \leqslant R_1 \leqslant 35$ and $0.1 \leqslant R_2 \leqslant 40$ in increments of 0.2. The graphs in Fig. 2 show the surfaces of the change in the accuracy criterion. Table 3 shows the weights $\mathbf{R}$ at which the criteria values $J^{\tilde{\mathbf{p}}}$ are minimal.

**Table 3.** Optimal values of $\mathbf{R}$ for different criteria $J^{\tilde{\mathbf{p}}}$

| | $\mathbf{R}_{opt}^{\tilde{\mathbf{p}}=1}$ | $\mathbf{R}_{opt}^{\tilde{\mathbf{p}}=2}$ | $\mathbf{R}_{opt}^{\tilde{\mathbf{p}}=3}$ | $\mathbf{R}_{opt}^{\tilde{\mathbf{p}}=4}$ | $\mathbf{R}_{opt}^{\tilde{\mathbf{p}}=5}$ | $\mathbf{R}_{opt}^{\tilde{\mathbf{p}}=6}$ |
|---|---|---|---|---|---|---|
| $R_1$ | 0.1 | 0.7 | 6.9 | 2.5 | 8.9 | 0.1 |
| $R_2$ | 32.9 | 1.1 | 28.5 | 12.1 | 32.9 | 3.5 |



(a) $\tilde{p} = 1$ (b) $\tilde{p} = 2$ (c) $\tilde{p} = 3$

(d) $\tilde{p} = 4$ (e) $\tilde{p} = 5$ (f) $\tilde{p} = 6$

**Fig. 2.** Criteria values $J^{\tilde{\mathbf{p}}}$ for different $R_1$ and $R_2$.

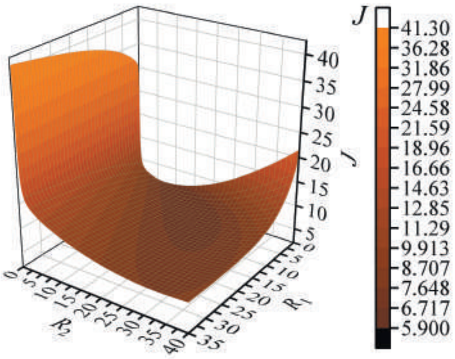


**Fig. 3.** Change of $\breve{J}$ at different values $R_1$ and $R_2$.

Due to the fact that the optimal values of the weight matrix $\mathbf{R}$ are different for the 6 cases under consideration, we will use the convolution of criteria [10] to find the robust optimal parameters:

$$\breve{J} = \sum_{\tilde{\mathbf{p}}=1}^{6} w^{\tilde{\mathbf{p}}} \times J^{\tilde{\mathbf{p}}}.$$

Since the values of the criteria for the cases under consideration have the same physical dimension, we assume that the value of $w^{\tilde{\mathbf{p}}} = 1$, $\tilde{\mathbf{p}} = 1, \ldots, 6$. Figure 3 shows the surface of change $\breve{J}$ when the values of $R_1$ and $R_2$ change. The values of the weights $\mathbf{R}$ corresponding to the minimum value of the criterion $\breve{J}$ are equal to $\mathbf{R}_{opt}^{\breve{J}} = \operatorname{diag}\{5.3, 17.1\}$.

Figure 4 presents the optimal values of $\mathbf{R}$ in the plane $R_1R_2$. Figure 5 shows the CV transients when the values of $\mathbf{R} = \mathbf{R}_{opt}^{\tilde{\mathbf{p}}=1}$ and $\mathbf{R} = \mathbf{R}_{opt}^{\breve{J}}$.

It can be concluded from the graphs in Fig. 5 that the use of weights $\mathbf{R}$ selected on the basis of $\breve{J}$, i.e. taking into account the variations in the parameters of the object, makes it possible to

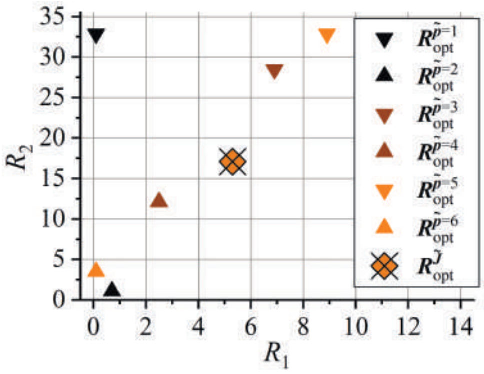
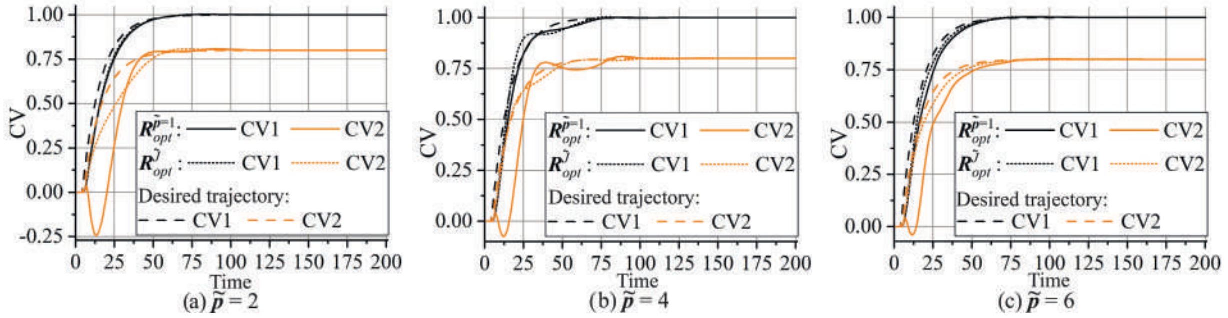


**Fig. 4.** Location of optimal values of **R**.



(a) $\tilde{p} = 2$ (b) $\tilde{p} = 4$ (c) $\tilde{p} = 6$

**Fig. 5.** CV transients at $\mathbf{R} = \mathbf{R}_{opt}^{\tilde{\mathbf{p}}=1}$ and $\mathbf{R} = \mathbf{R}_{opt}^{\breve{J}}$.

reduce the deviation from the desired trajectory in comparison with the case when the optimal weights are selected on the basis of only one of the criteria $J^{\tilde{\mathbf{p}}}$, i.e. without taking into account the uncertainty of the parameters of the plant.

**Table 4.** Values of criteria $J^{\tilde{\mathbf{p}}}$ (deviation from the desired dynamics) for different **R**

| | $R_1$ | $R_2$ | $\tilde{\mathbf{p}} = 1$ | $\tilde{\mathbf{p}} = 2$ | $\tilde{\mathbf{p}} = 3$ | $\tilde{\mathbf{p}} = 4$ | $\tilde{\mathbf{p}} = 5$ | $\tilde{\mathbf{p}} = 6$ |
|---|---|---|---|---|---|---|---|---|
| $\mathbf{R}_{opt}^{\tilde{\mathbf{p}}=1}$ | 0.1 | 32.9 | 1.34848 | 6.06283 | 2.00167 | 2.62597 | 2.03054 | 2.39714 |
| $\mathbf{R}_{opt}^{\tilde{\mathbf{p}}=2}$ | 0.7 | 1.1 | 3.7571 | 0.06194 | 2.29674 | 0.73448 | 2.6284 | 0.18394 |
| $\mathbf{R}_{opt}^{\tilde{\mathbf{p}}=3}$ | 6.9 | 28.5 | 1.59663 | 1.89024 | 1.25891 | 0.64345 | 0.82054 | 0.90192 |
| $\mathbf{R}_{opt}^{\tilde{\mathbf{p}}=4}$ | 2.5 | 12.1 | 2.38149 | 0.71668 | 1.44793 | 0.21418 | 1.2088 | 0.15819 |
| $\mathbf{R}_{opt}^{\tilde{\mathbf{p}}=5}$ | 8.9 | 32.9 | 1.54609 | 2.1035 | 1.26915 | 0.82553 | 0.80598 | 1.13606 |
| $\mathbf{R}_{opt}^{\tilde{\mathbf{p}}=6}$ | 0.1 | 3.5 | 2.97203 | 0.43404 | 1.54271 | 0.3417 | 1.48232 | 0.04001 |
| $\mathbf{R}_{opt}^{\breve{J}}$ | 5.3 | 17.1 | 1.94841 | 0.87779 | 1.38022 | 0.2607 | 1.05938 | 0.28966 |

## 4. CONCLUSION

In the framework of this work, a search was made for the robust optimal values of the weights **R** for the control system based on the predictive model, taking into account the parametric uncertainty of the parameters of the control plant. The optimal weights were found for cases where one of the parameters is at the boundary of the range, and the rest are in the middle. With the help of convolution of criteria, robustly optimal values $\mathbf{R}_{opt}^{\breve{J}} = \operatorname{diag}\{5.3, 17.1\}$ were found that ensure the best quality of control of the $\mathrm{T}_{\mathrm{IBP\_KERO}}$ of kerosene fraction of the industrial fractionator. It is shown that the use of $\mathbf{R}^{\check{J}}_{opt}$ made it possible to reduce the deviation from the desired dynamics in comparison with the use of the optimal value of $\mathbf{R}$ without taking into account the uncertainty.

## FUNDING

The research was carried out within the state assignment of IACP FEB RAS (Theme FWFW-2021-0003).

## REFERENCES

1. Qian, X., Jia, S., Huang, K., Chen, H., Yuan, Y., and Zhang, L., Model predictive control of azeotropic dividing wall distillation column for separating furfural-water mixture, Optimality and Robustness, *ISA transactions*, 2021, vol. 111, pp. 302–308.

2. Martin, P.A., Zanin, A.C., and Odloak, D., Integrating real time optimization and model predictive control of a crude distillation unit, *Brazilian Journal of Chemical Engineering*, 2019, vol. 36, pp. 1205–1222.

3. Mendis, P., Wickramasinghe, C., Narayana, M., and Bayer, C., Adaptive model predictive control with successive linearization for distillate composition control in batch distillation, *2019 Moratuwa Engineering Research Conference*, 2019, pp. 366–369.

4. Schwenzer, M., Ay, M., Bergs, T., and Abel, D., Review on model predictive control: an engineering perspective, *The International Journal of Advanced Manufacturing Technology*, 2021, vol. 117, pp. 1327–1349.

5. Seborg, D.E., Edgar, T.E., Mellichamp, D.A., and Doyle, III F.J., Process Dynamics and Control, 4nd ed., *John Wiley* & *Sons*, 2016, pp. 369–389.

6. Mayne, D.Q., Kerrigan, E.C., and Falugi, P., Robust model predictive control: advantages and disadvantages of tube-based methods, *IFAC Proceedings Volumes*, 2011, vol. 44, no. 1, pp. 191–196.

7. Hill, E., Biglarbegian, M., and Gadsden, S.A., Tube-based model predictive control of small satellite systems with uncertainty dynamics, *Proceedings of CSME Congress*, 2021.

8. Kayacan, E. and Peschel, J., Robust model predictive control of systems by modeling mismatched uncertainty, *IFAC-PapersOnLine*, 2016, vol. 49, no. 18, pp. 265–269.

9. Giraldo, S.A.C., Melo, P.A., and Secchi, A.R., Tuning of model predictive controllers based on hybrid optimization, *Processes*, 2022, vol. 10, 351.

10. Al-Jamimi, H.A., BinMakhashen, G.M., Deb, K., and Saleh, T.A., Multiobjective optimization and analysis of petroleum refinery catalytic processes: A review, *Fuel*, 2021, vol. 288.

*This paper was recommended for publication by A.A. Galyaev, a member of the Editorial Board*