Volume 86, Number 6,
June 2025

ISSN 0005-1179
CODEN: AURCAT

AUTOMATION AND REMOTE CONTROL

Automation and Remote Control

Vol. 86, No. 6, June 2025

Editor-in-Chief
Andrey A. Galyaev

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

0005-1179/25. *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 86 (12 issues) is published in 2025.
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. 86, No. 6, 2025

Linear Systems

Nonlinear Systems

Stochastic Systems

Control in Technical Systems

Control in Social Economic Systems

Optimization, System Analysis, and Operations Research

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 6, pp. 499–515.

Generalization of Gershgorin Circle Theorem with Application to Analysis and Design of Control Systems

I. B. Furtat[*,**]

[*] *Institute for Problems in Mechanical Engineering,
Russian Academy of Sciences, St. Petersburg, Russia*
[**] *St. Petersburg State University, St. Petersburg, Russia*
e-mail: cainenash@mail.ru

Received October 3, 2024
Revised February 28, 2025
Accepted March 14, 2025

Abstract—The application of the Gershgorin circle theorem and some of its derivatives to estimating matrix eigenvalues is considered. The obtained results are developed to design a localization region for matrix eigenvalues with interval-indefinite constant and non-stationary elements. The concept of *e-circles* is introduced to obtain more accurate estimates of these regions than when using Gershgorin circles. The obtained results are applied to the stability analysis of network systems, where it is shown that the proposed methods allows one to analyze a network with a much larger number of agents than when using methods for solving linear matrix inequalities in CVX and Yalmip/SeDuMi, as well as the eig (for calculating matrix eigenvalues) and lyap (for solving the Lyapunov equation) algorithms in MatLab. It is shown that if the developed methods are applied not to the system itself, but to the result obtained using the Lyapunov function method, then it is possible to study systems with matrices without diagonal dominance. This allowed us to consider the modification of the Demidovich condition for systems with non-stationary parameters and the design of the control law for non-stationary systems with matrices without diagonal dominance. All the obtained results are illustrated by numerical modeling.

Keywords: Gershgorin circle theorem, matrix eigenvalue localization domain, stability, control

DOI: 10.31857/S0005117925060012

1. INTRODUCTION

When analyzing the properties of dynamic systems and design of the control law, one of the key questions is whether the system is stable. Currently, various methods and approaches are used to determine stability: calculating the eigenvalues of the matrix [1], various algebraic and frequency stability criteria [1], the Lyapunov function method [1], divergent methods for studying stability [2], etc.

This paper focuses on design the localization domain of the eigenvalues of matrices with the application of the obtained result to the analysis and design of control laws. To construct the localization domain of the eigenvalues, the Gershgorin circle theorem [3–5] (hereinafter simply the Gershgorin theorem) and some of its consequences will be considered, and new results will be obtained on the generalization of this theorem to the case of parametrically indefinite matrices and matrices with non-stationary parameters.

Gershgorin theorem and its various modifications have been repeatedly considered in the literature. The interest in this theorem is associated with a simple method for determining the domain of localization of eigenvalues. Gershgorin theorem often leads to the study of systems containing matrices with diagonal dominance. In particular, such systems were studied in [6–8] and were called *super-stable* (if all Gershgorin circles are entirely in the left half-plane of the complex plane). It is shown that the analysis and design of control laws leads to linear programming problems. In [9–13], refinement of localization regions is obtained in the form of averaged estimates, using l_1 vector norms, etc., and in [14, 15], a design of static linear control laws using Gershgorin theorem is proposed. In [16–18], the application of Gershgorin theorem to the study of the stability of models in the chemical industry, models of electrical networks with three-phase generators, and biological models of epidemics is considered.

An analysis of the literature showed that when determining a localization region for the eigenvalues of matrices, Gershgorin theorem has advantages in the simplicity of its application, a convex procedure for finding the localization region, and low computational costs. However, limitations in the application of this theorem are associated with overestimated estimates of the localization region and consideration of matrices with diagonal dominance (or reduced to them using a diagonal matrix for transforming the basis). The requirement of diagonal dominance is also significantly restrictive in the design of the control law.

This paper will consider the solution of the following problems:

(1) estimates and localization regions of the eigenvalues of a constant matrix will be considered;

(2) localization regions of the eigenvalues of a matrix with interval parametric uncertainty will be obtained;

(3) the following will be considered as examples of application of the obtained results:

  (a) the problem of synchronization of network systems with a large number of scalar agents, where it will be shown that the proposed results can be applied to the stability analysis of a much larger number of agents than when using the methods for solving linear matrix inequalities in CVX and Yalmip/SeDuMi, as well as the eig (for calculating the eigenvalues of a matrix) and lyap (for solving the Lyapunov equation) algorithms in MatLab;

  (b) modification of the Demidovich condition (on the stability of linear systems with nonstationary parameters [19](Theorem 6.1), [23]) to systems with interval-uncertain nonstationary parameters and with the matrix of the original system without diagonal dominance;

  (c) the problem of finding a matrix in a linear control law using linear matrix inequalities for objects with a matrix without diagonal dominance.

The following *notations are used in the paper:* $\mathbb{C}$ is the set of complex numbers, $\mathbb{R}^n$ is the n-dimensional Euclidean space with the vector norm $|\cdot|$, $\mathbb{R}^{n\times m}$ is the set of all real matrices of dimension $n \times m$ with the induced matrix norm $\|Q\|$, $\lambda_i\{Q\}$ is the ith eigenvalue of the square matrix Q, $\Re\{\lambda_i\{Q\}\}$ is the real part of the ith eigenvalue of the square matrix Q, $\Im\{\lambda_i\{Q\}\}$ is the imaginary part of the ith eigenvalue of the square matrix Q, I is the identity matrix of the corresponding order, the matrix $Q \in \mathbb{R}^{n\times m}$ will also be denoted as (q_{ij}), $i = 1,\dots,n$, $j = 1,\dots,m$.

2. ESTIMATES OF THE LOCALIZATION DOMAIN OF EIGENVALUES

2.1. Constant Matrices

In this section, estimates will be obtained for the localization region of the eigenvalues of the matrix $Q = (q_{ij}) \in \mathbb{R}^{n\times n}$ with constant elements. To clarify these estimates, the diagonal matrices $D = diag\{d_1,\dots,d_n\}$ and $H = diag\{h_1,\dots,h_n\}$ will be additionally considered. Let us introduce the notation of sums over rows and columns of absolute values of elements of matrices Q, $D^{-1}QD$

and $H^{-1}QH$ without diagonal elements in the form

$$
\begin{aligned}
R_i(Q) &= \sum_{j=1,j\neq i}^{n} |q_{ij}|, \quad C_j(Q) = \sum_{i=1,i\neq j}^{n} |q_{ij}|,\\
R_i^D(Q) &= \sum_{j=1,j\neq i}^{n} \frac{d_j}{d_i}|q_{ij}|, \quad C_j^D(Q) = \sum_{i=1,i\neq j}^{n} \frac{d_i}{d_j}|a_{ij}|,\\
R_i^H(Q) &= \sum_{j=1,j\neq i}^{n} \frac{h_j}{h_i}|q_{ij}|, \quad C_j^H(Q) = \sum_{i=1,i\neq j}^{n} \frac{h_i}{h_j}|a_{ij}|.
\end{aligned}
$$

Below are two lemmas that allow one to obtain lower and upper bounds on the real parts of the eigenvalues of the matrix Q.

Lemma 1. *Consider the matrix $Q \in \mathbb{R}^{n\times n}$. There exist $d_i > 0$, $h_i > 0$, $i = 1, \ldots, n$ such that the following estimates hold:*

$$
\begin{aligned}
\max_i\{\Re\{\lambda_i\{Q\}\}\} &\leqslant \sigma^D_{\max}\{Q\} \leqslant \sigma_{\max}\{Q\},\\
\min_i\{\Re\{\lambda_i\{Q\}\}\} &\geqslant \sigma^H_{\min}\{Q\} \geqslant \sigma_{\min}\{Q\},
\end{aligned}
\tag{1}
$$

where

$$
\begin{aligned}
\sigma_{\max}(Q) &= \min\Big\{\max_i\{q_{ii} + R_i(Q)\}, \max_j\{q_{jj} + C_j(Q)\}\Big\},\\
\sigma_{\min}(Q) &= \max\Big\{\min_i\{q_{ii} - R_i(Q)\}, \min_j\{q_{jj} - C_j(Q)\}\Big\},\\
\sigma^D_{\max}(Q) &= \min_D\Big\{\max_i\{q_{ii} + R^D_i(Q)\}, \max_j\{q_{jj} + C^D_j(Q)\}\Big\},\\
\sigma^H_{\min}(Q) &= \max_H\Big\{\min_i\{q_{ii} - R^H_i(Q)\}, \min_j\{q_{jj} - C^H_j(Q)\}\Big\}.
\end{aligned}
\tag{2}
$$

Proof. By Gershgorin theorem [4] all eigenvalues of Q are contained in the union of n circles $\cup_{i=1}^n\left\{z \in \mathbb{C} : |z - q_{ii}| \leqslant \sum\limits_{j=1,j\neq i}^{n} |q_{ij}|\right\}$. Since Q^{T} has the same eigenvalues as Q, all eigenvalues of Q are also contained in the union of n circles $\cup_{i=1}^n\left\{z \in \mathbb{C} : |z - q_{ii}| \leqslant \sum\limits_{j=1,j\neq i}^{n} |q_{ji}|\right\}$. Therefore, $\sigma_{\min}\{Q\} \leqslant \min\limits_i\{\Re\{\lambda_i\{Q\}\}\}$ and $\sigma_{\max}\{Q\} \geqslant \max\limits_i\{\Re\{\lambda_i\{Q\}\}\}$.

Now consider the diagonal matrix $D = diag\{d_1, \ldots, d_n\}$. It is known that the eigenvalues of the matrices $D^{-1}QD$ and Q do not change. However, by varying the coefficients d_i, the radii of the Gershgorin circles for the matrix Q can be changed in the form $\cup_{i=1}^n\left\{z \in \mathbb{C} : |z - q_{ii}| \leqslant \sum\limits_{j=1,j\neq i}^{n} \frac{d_j}{d_i}|q_{ij}|\right\}$ and for the matrix Q^{T} in the form $\cup_{i=1}^n\left\{z \in \mathbb{C} : |z - q_{ii}| \leqslant \sum\limits_{j=1,j\neq i}^{n} \frac{d_i}{d_j}|q_{ji}|\right\}$. Therefore, there exist $d_i > 0$, $i = 1, \ldots, n$ such that $\sigma^D_{\max}\{Q\} \leqslant \sigma_{\max}\{Q\}$. However, it is impossible to simultaneously decrease the radii of all circles by varying d_i, i.e. when the radii of some circles decrease, the radii of others increase. Therefore, to obtain the estimate $\sigma^H_{\min}\{Q\} \geqslant \sigma_{\min}\{Q\}$, the matrix H is used instead of D.

Lemma 2. *Consider the matrix $Q \in \mathbb{R}^{n\times n}$. There exist $d_i > 0$, $h_i > 0$, $i = 1, \ldots, n$ and $\alpha, \beta \in [0, 1]$ such that the following estimates hold:*

$$\begin{aligned} \max_i\{\Re\{\lambda_i\{Q\}\}\} &\leqslant \sigma^{D,\alpha}_{\max}\{Q\} \leqslant \sigma^{\alpha}_{\max}\{Q\}, \\ \min_i\{\Re\{\lambda_i\{Q\}\}\} &\geqslant \sigma^{H,\beta}_{\min}\{Q\} \geqslant \sigma^{\beta}_{\min}\{Q\}, \end{aligned} \tag{3}$$

where

$$\begin{aligned} \sigma^{\alpha}_{\max}(Q) &= \min_{\alpha}\left\{\max_i\{q_{ii} + [R_i(Q)]^{\alpha}[C_i(Q)]^{1-\alpha}\}\right\}, \\ \sigma^{\beta}_{\min}(Q) &= \max_{\beta}\left\{\min_i\{q_{ii} - [R_i(Q)]^{\beta}[C_i(Q)]^{1-\beta}\}\right\}, \\ \sigma^{D,\alpha}_{\max}(Q) &= \min_{D,\alpha}\left\{\max_i\{q_{ii} + [R^D_i(Q)]^{\alpha}[C^D_i(Q)]^{1-\alpha}\}\right\}, \\ \sigma^{H,\beta}_{\min}(Q) &= \max_{H,\beta}\left\{\min_i\{q_{ii} - [R^H_i(Q)]^{\beta}[C^H_i(Q)]^{1-\beta}\}\right\}. \end{aligned} \tag{4}$$

Proof. According to Ostrovsky theorem [4], all eigenvalues of the matrix Q are contained in the union of n circles $\cup_{i=1}^n \left\{ z \in \mathbb{C} : |z - q_{ii}| \leqslant \left[\sum\limits_{j=1, j\neq i}^n |q_{ij}|\right]^{\alpha} \left[\sum\limits_{j=1, j\neq i}^n |q_{ji}|\right]^{1-\alpha} \right\}$. Therefore, there exists α such that $\sigma^{\alpha}_{\max}\{Q\} \geqslant \max\limits_i\{\Re\{\lambda_i\{Q\}\}\}$. Analogously, we obtain that there exists β such that $\sigma^{\beta}_{\min}\{Q\} \leqslant \min\limits_i\{\Re\{\lambda_i\{Q\}\}\}$.

By varying the coefficients d_i, the radii of the circles can be changed $\cup_{i=1}^n \left\{ z \in \mathbb{C} : |z - q_{ii}| \leqslant \left[\sum\limits_{j=1, j\neq i}^n \frac{d_j}{d_i}|q_{ij}|\right]^{\alpha} \left[\sum\limits_{j=1, j\neq i}^n \frac{d_j}{d_i}|q_{ji}|\right]^{1-\alpha} \right\}$. Consequently, there exist d_i, $i = 1, \ldots, n$ and α such that $\sigma^{D,\alpha}_{\max}\{Q\} \leqslant \sigma^{\alpha}_{\max}\{Q\}$. From similar reasoning it follows that there exist h_i, $i = 1, \ldots, n$ and β such that $\sigma^{H,\beta}_{\min}\{Q\} \leqslant \sigma^{\beta}_{\max}\{Q\}$.

Corollary 1. *If any upper bound in Lemmas 1 and 2 takes a negative value, then it is an estimate of the degree of stability, the concept of which was introduced by Ya.Z. Tsypkin and P.V. Bromberg in* [20].

Corollary 2. *From the proofs of Lemmas 1 and 2 it also follows that by the intersection of the corresponding circles one can find the domain of localization of the eigenvalues of the matrix Q, from which one can find not only upper and lower bounds for the real parts of the eigenvalues, but also an upper bound for the imaginary part, which we denote as $\hat{\Im}\{Q\} \geqslant \max\limits_i\{\Im\{\lambda_i\{Q\}\}\}$. The value of $\hat{\Im}\{Q\}$ is defined as the maximum value of the intersection of the circles along the imaginary axis. If the upper bound on the real part of the eigenvalue of Q is negative, then one can obtain an estimate of the oscillation μ in the form $\mu \leqslant \hat{\mu} := \frac{\hat{\Im}\{Q\}}{|\max\limits_i\{\Re\{\lambda_i\{Q\}\}\}|}$. It is well known* [1] *that the oscillation is used to estimate the overshoot Π in the form $\Pi \leqslant e^{\pi/\mu}$. Then the new estimate of the degree of overshoot is defined as $\Pi \leqslant e^{\pi/\hat{\mu}}$.*

Knowing the estimate of the degree of stability, the estimate of oscillation and the lower estimate of the real part of the eigenvalue, we can construct a majorant and a minorant for the transient process of a linear system under a single step action, which is a development of the results of S.A. Chaplygin, N.N. Luzin, A.A. Feldbaum and A.M. Rubinchik [21, 22].

The Corollaries 1 and 2 will also be true in further generalizations of the obtained results to perturbed matrices. Let us demonstrate what was noted in the lemmas and corollaries using the following example.

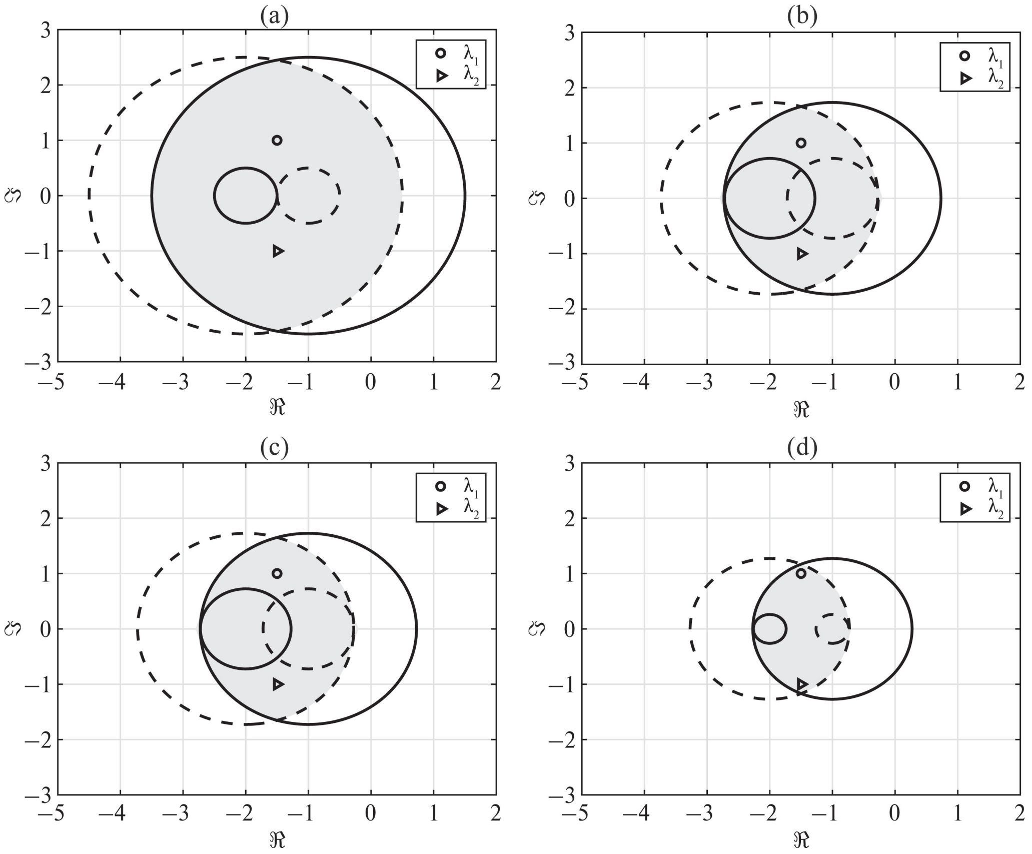

Fig. 1. Localization regions (gray region) of eigenvalues of matrix Q using estimates (1) and (3).

Example 1. Consider the matrix $Q = \begin{bmatrix} -1 & -2.5 \\ -0.5 & -2 \end{bmatrix}$, which eigenvalues are $-1.5 \pm i$. Figure 1 shows the localization regions (highlighted in gray) using:

- Gershgorin circle theorem (Fig. 1a);
- Lemma 1 with $H = diag\{1; 0.48\}$ and $D = diag\{1; 2.08\}$ (Fig. 1b);
- Lemma 2 with $\alpha = \beta = 0.23$ and $D = H = I$ (Fig. 1c);
- Lemma 2 with $\alpha = \beta = 0.01$, $H = diag\{1; 0.51\}$ and $D = diag\{1; 1.96\}$ (Fig. 1d);

The solid circles correspond to the circles which radii are calculated from the rows of the matrix, the dashed circles correspond to the columns of the matrix.

Table 1 contains upper and lower bounds for the real part of the eigenvalue of the matrix Q. The accuracy is calculated as the relative error between the corresponding estimate and $\Re\{\lambda\{Q\}\} = -1.5$ (e.g. $\frac{|-3.5+1.5|}{1.5}100\% = 133.3\%$).

Table 1. Estimates of the real part of the eigenvalues of the matrix Q, obtained using (1) and (3)

Figure	Estimate of $\Re\{\lambda\{Q\}\}$	Accuracy, %
Fig. 1a	$\sigma_{\min}(Q) = -3.5$; $\sigma_{\max}(Q) = 0.5$	133.3
Fig. 1b	$\sigma^H_{\min}(Q) = -2.72$; $\sigma^D_{\max}(Q) = -0.27$	82
Fig. 1c	$\sigma^\beta_{\min}(Q) = -2.72$; $\sigma^\alpha_{\max}(Q) = -0.27$	82
Fig. 1d	$\sigma^{H,\beta}_{\min}(Q) = -2.26$; $\sigma^{D;\alpha}_{\max}(Q) = -0.73$	51.3

From Fig. 1 one can find estimates of the imaginary part, which are reflected in Table 2. The accuracy is calculated as the relative error between the corresponding estimate and $\Im\{\lambda\{Q\}\} = 1$.

Table 2. Estimates of the imaginary part of the eigenvalues obtained using Lemmas 1 and 2

Figure	Estimate of $\Im\{\lambda\{Q\}\}$	Accuracy, %
Fig. 1a	2.3	130
Fig. 1b	1.8	80
Fig. 1c	1.7	70
Fig. 1d	1.2	20

The best estimates of the real and imaginary parts are guaranteed by the result of Lemma 2, where the variable parameters D, H, α and β are used simultaneously.

2.2. Perturbed Matrices

In this section, we consider the search for the regions of localization of eigenvalues for matrices with interval-indefinite parameters:

$$\begin{aligned}
&Q(t) = Q_0 + \Delta Q(t) \in \mathbb{R}^{n\times n}, \\
&Q_0 = (q_{ij}^0), && \Delta Q(t) = (\Delta q_{ij}(t)), \\
&\Delta \underline{q}_{ii} \leqslant \Delta q_{ii}(t) \leqslant \Delta \overline{q}_{ii}, && |\Delta q_{ij}(t)| \leqslant m_{ij} \text{ for } i \neq j.
\end{aligned} \tag{5}$$

Since the matrix elements can take any values from the admissible intervals, instead of the circles of localization of the eigenvalues considered in the proofs of Lemmas 1 and 2, we introduce the following figure into consideration.

Definition 1. The figure formed by the union of the circles $\mathcal{EC} = \cup_{q\in[\underline{q};\overline{q}]}\Big\{z \in \mathbb{C} : |z - q| \leqslant R\Big\}$ is called the e-circle.

We introduce notations for upper bounds of the sums over rows and columns of the absolute values of the elements of the matrices $Q(t)$, $D^{-1}Q(t)D$ and $H^{-1}Q(t)H$, excluding the diagonal elements, in the form

$$\begin{aligned}
\hat{R}_i(Q) &= \sum_{j=1,j\neq i}^{n} (|q_{ij}^0| + m_{ij}), & \hat{C}_j(Q) &= \sum_{i=1,i\neq j}^{n} (|q_{ij}^0| + m_{ij}), \\
\hat{R}_i^D(Q) &= \sum_{j=1,j\neq i}^{n} \frac{d_j}{d_i}(|q_{ij}^0| + m_{ij}), & \hat{C}_j^D(Q) &= \sum_{i=1,i\neq j}^{n} \frac{d_i}{d_j}(|q_{ij}^0| + m_{ij}), \\
\hat{R}_i^H(Q) &= \sum_{j=1,j\neq i}^{n} \frac{h_j}{h_i}(|q_{ij}^0| + m_{ij}), & \hat{C}_j^H(Q) &= \sum_{i=1,i\neq j}^{n} \frac{h_i}{h_j}(|q_{ij}^0| + m_{ij}).
\end{aligned}$$

Now we will consider the generalization of Lemmas 1 and 2 to the case of matrices with interval-indefinite elements. Below, in the formulations of the lemmas, we will omit the dependence of matrices and parameters on t for the sake of simplifying the expressions.

Lemma 3. *The eigenvalues of the matrix Q from* (5) *are in the intersection region of the e-circles*

$$\mathcal{EC}_{\text{row}} \cap \mathcal{EC}_{\text{col}}, \tag{6}$$

where

$$\begin{aligned}&\mathcal{EC}_{\text{row}} = \cup_{i=1}^{n}\mathcal{EC}_{\text{row},i},\\&\mathcal{EC}_{\text{row},i} = \cup_{\Delta q_{ii}\in[\Delta\underline{q}_{ii};\Delta\overline{q}_{ii}]}\Big\{\lambda\in\mathbb{C}: |\lambda - q_{ii}^{0} - \Delta q_{ii}| \leqslant \hat{R}_i(Q)\Big\},\end{aligned}\tag{7}$$

$$\begin{aligned}&\mathcal{EC}_{\text{col}} = \cup_{j=1}^{n}\mathcal{EC}_{\text{col},j},\\&\mathcal{EC}_{\text{col},j} = \cup_{\Delta q_{jj}\in[\Delta\underline{q}_{jj};\Delta\overline{q}_{jj}]}\Big\{\lambda\in\mathbb{C}: |\lambda - q_{jj}^{0} - \Delta q_{jj}| \leqslant \hat{C}_j(Q)\Big\}.\end{aligned}\tag{8}$$

Proof. Let $\lambda(t)$ be an eigenvalue of the matrix $Q(t)$ and $s(t) = col\{s_1(t),\dots,s_n(t)\}$ be the eigenvector corresponding to this eigenvalue. Choose the ith component of the vector $s(t)$ such that $\sup\{s_i(t)\} \geqslant \max\{\sup\{s_1(t)\},\dots,\sup\{s_{i-1}(t)\},\sup\{s_{i+1}(t)\},\dots,\sup\{s_n(t)\}\}$. Denote $\bar{s}_i = \sup\{s_i(t)\}$. From the relation $\lambda(t)s(t) = Q(t)s(t)$ we write out the expression for the ith coordinate in the form $\lambda(t)s_i(t) = \sum\limits_{j=1}^{n} q_{ij}(t)s(t)$ or $(\lambda(t) - q_{ii}(t))s_i(t) = \sum\limits_{j=1,j\neq i}^{n} q_{ij}(t)s(t)$. Using the triangle inequality, we consider the estimate

$$\begin{aligned}|\lambda(t) - q_{ii}(t)||s_i(t)| &= \left|\sum_{j=1,j\neq i}^{n} q_{ij}(t)s_j(t)\right|\\ \leqslant \sum_{j=1,j\neq i}^{n}|q_{ij}(t)s_j(t)| &\leqslant \sum_{j=1,j\neq i}^{n}|q_{ij}(t)||s_j(t)| \leqslant \bar{s}_i\sum_{j=1,j\neq i}^{n}|q_{ij}(t)|.\end{aligned}\tag{9}$$

Let us rewrite the expression (9) as $|\lambda(t) - q_{ii}(t)||s_i(t)| - \bar{s}_i\sum\limits_{j=1,j\neq i}^{n}|q_{ij}(t)| \leqslant 0$ or in the form

$$\bar{s}_i\left(|\lambda(t) - q_{ii}(t)|\frac{|s_i(t)|}{\bar{s}_i} - \sum_{j=1,j\neq i}^{n}|q_{ij}(t)|\right) \leqslant 0.\tag{10}$$

Since $\frac{|s_i(t)|}{\bar{s}_i} \leqslant 1$, then the expression (10) will be satisfied if inequality holds

$$|\lambda(t) - q_{ii}(t)| \leqslant \sum_{j=1,j\neq i}^{n}|q_{ij}(t)|.\tag{11}$$

Since $\Delta\underline{q}_{ii} \leqslant \Delta q_{ii}(t) \leqslant \Delta\overline{q}_{ii}$ and $|\Delta q_{ij}(t)| \leqslant m_{ij}$ for $i\neq j$, then we rewrite the inequality (11) as an e-circle $\mathcal{EC}_{\text{row},i}$ from (7).

The relation (7) is satisfied for some i. Since it is unknown which i corresponds to a given $\lambda(t)$, we can only say that $\lambda(t)$ belongs to the union of e-circles $\mathcal{EC}_{\text{row}} = \cup_{i=1}^{n}\mathcal{EC}_{\text{row},i}$. This means that all eigenvalues of the matrix $Q(t)$ are contained in the union of e-circles $\mathcal{EC}_{\text{row}}$.

Since the matrix $Q^{\mathrm{T}}(t)$ has the same eigenvalues as the matrix $Q(t)$, then all eigenvalues of the matrix $Q(t)$ are contained in the union of e-circles $\mathcal{EC}_{\text{col}} = \cup_{j=1}^{n}\mathcal{EC}_{\text{col},j}$, see (8). Further reasoning for the matrix $Q^{\mathrm{T}}(t)$ is similar to that for the matrix $Q(t)$. Since the eigenvalues of the matrix $Q(t)$ are simultaneously in $\mathcal{EC}_{\text{row}}$ and $\mathcal{EC}_{\text{col}}$, they are in the domain (6).

Lemma 4. *Let $d_i > 0$, $h_i > 0$, $i = 1,\dots,n$ be given. The eigenvalues of the matrix Q from* (5) *are in the intersection region of the e-circles*

$$\mathcal{EC}_{\text{row}}^{D} \cap \mathcal{EC}_{\text{col}}^{H},$$

where

$$\begin{aligned}
&\mathcal{EC}^D_{\text{row}} = \cup_{i=1}^n \mathcal{EC}^D_{\text{row},i},\\
&\mathcal{EC}^D_{\text{row},i} = \cup_{\Delta q_{ii}\in[\Delta \underline{q}_{ii};\Delta \overline{q}_{ii}]}\Big\{\lambda\in\mathbb{C} : |\lambda - q^0_{ii} - \Delta q_{ii}| \leqslant \hat{R}^D_i(Q)\Big\},
\end{aligned}$$

$$\begin{aligned}
&\mathcal{EC}^H_{\text{col}} = \cup_{j=1}^n \mathcal{EC}^H_{\text{col},j},\\
&\mathcal{EC}^H_{\text{col},j} = \cup_{\Delta q_{jj}\in[\Delta \underline{q}_{jj};\Delta \overline{q}_{jj}]}\Big\{\lambda\in\mathbb{C} : |\lambda - q^0_{jj} - \Delta q_{jj}| \leqslant \hat{C}^H_j(Q)\Big\}.
\end{aligned}$$

Proof. The results of Lemma 4 follow from Lemma 3 and the fact that the eigenvalues of the matrices $D^{-1}Q(t)D$, $H^{-1}Q(t)H$, and $Q(t)$ are the same.

Lemma 5. *Let $d_i > 0$, $i = 1,\dots,n$ and $\alpha\in[0;1]$ be given. The eigenvalues of the matrix Q from* (5) *are in the intersection region of the e-circles*

$$\mathcal{EC}^{D,\alpha} = \cup_{i=1}^n \mathcal{EC}^{D,\alpha}_i,$$

where

$$\mathcal{EC}^{D,\alpha}_i = \cup_{\Delta q_{ii}\in[\Delta \underline{q}_{ii};\Delta \overline{q}_{ii}]}\Big\{\lambda\in\mathbb{C} : |\lambda - q_{ii} - \Delta q_{ii}| \leqslant [\hat{R}^D_i(Q)]^\alpha[\hat{C}^D_i(Q)]^{1-\alpha}\Big\}.$$

Proof. The proof of Lemma 5 follows from the proofs of Lemmas 2 and 3 and the fact that the eigenvalues of the matrices $D^{-1}Q(t)D$ and $Q(t)$ are the same.

Corollary 3. *Similarly to Lemmas 1 and 2, one can write out estimates for the maximum and minimum values of the eigenvalues of the matrix Q using the results of Lemmas 3–5, i.e. there exist numbers $d_i > 0$, $h_i > 0$, $i = 1,\dots,n$ and $\alpha,\beta\in[0;1]$ such that the following estimates are valid:*

$$\begin{aligned}
&\max_i\Big\{\sup_t\{\Re\{\lambda_i\{Q(t)\}\}\}\Big\} \leqslant \sigma^D_{\max}\{Q(t)\} \leqslant \sigma_{\max}\{Q(t)\},\\
&\min_i\Big\{\sup_t\{\Re\{\lambda_i\{Q(t)\}\}\}\Big\} \geqslant \sigma^H_{\min}\{Q(t)\} \geqslant \sigma_{\min}\{Q(t)\},\\
&\max_i\Big\{\sup_t\{\Re\{\lambda_i\{Q(t)\}\}\}\Big\} \leqslant \sigma^{D,\alpha}_{\max}\{Q(t)\} \leqslant \sigma^\alpha_{\max}\{Q(t)\},\\
&\min_i\Big\{\sup_t\{\Re\{\lambda_i\{Q(t)\}\}\}\Big\} \geqslant \sigma^{H,\beta}_{\min}\{Q(t)\} \geqslant \sigma^\beta_{\min}\{Q(t)\},
\end{aligned}\tag{12}$$

where

$$\begin{aligned}
&\sigma_{\max}(Q(t)) = \min\Big\{\max_i\{q^0_{ii} + \Delta\overline{q}_{ii} + \hat{R}_i(Q)\}, \max_j\{q^0_{jj} + \Delta\overline{q}_{jj} + \hat{C}_j(Q)\}\Big\},\\
&\sigma_{\min}(Q(t)) = \max\Big\{\min_i\{q^0_{ii} - \Delta\overline{q}_{ii} - \hat{R}_i(Q)\}, \min_j\{q^0_{jj} - \Delta\overline{q}_{jj} - \hat{C}_j(Q)\}\Big\},\\
&\sigma^D_{\max}(Q(t)) = \min_D\Big\{\max_i\{q^0_{ii} + \Delta\overline{q}_{ii} + \hat{R}^D_i(Q)\}, \max_j\{q^0_{jj} + \Delta\overline{q}_{jj} + \hat{C}^D_j(Q)\}\Big\},\\
&\sigma^H_{\min}(Q(t)) = \max_H\Big\{\min_i\{q^0_{ii} - \Delta\overline{q}_{ii} - \hat{R}^H_i(Q)\}, \min_j\{q^0_{jj} - \Delta\overline{q}_{jj} - \hat{C}^H_j(Q)\}\Big\},
\end{aligned}$$

$$\begin{aligned}
&\sigma^\alpha_{\max}(Q(t)) = \min_\alpha\left\{\max_i\{q^0_{ii} + \Delta\overline{q}_{ii} + [\hat{R}_i(Q)]^\alpha[\hat{C}_i(Q)]^{1-\alpha}\}\right\},\\
&\sigma^\beta_{\min}(Q(t)) = \max_\beta\left\{\min_i\{q^0_{ii} - \Delta\overline{q}_{ii} - [\hat{R}_i(Q)]^\beta[\hat{C}_i(Q)]^{1-\beta}\}\right\},\\
&\sigma^{D,\alpha}_{\max}(Q(t)) = \min_{D,\alpha}\left\{\max_i\{q^0_{ii} + \Delta\overline{q}_{ii} + [\hat{R}^D_i(Q)]^\alpha[\hat{C}^D_i(Q)]^{1-\alpha}\}\right\},\\
&\sigma^{H,\beta}_{\min}(Q(t)) = \max_{H,\beta}\left\{\min_i\{q^0_{ii} - \Delta\overline{q}_{ii} - [\hat{R}^H_i(Q)]^\beta[\hat{C}^H_i(Q)]^{1-\beta}\}\right\}.
\end{aligned}$$

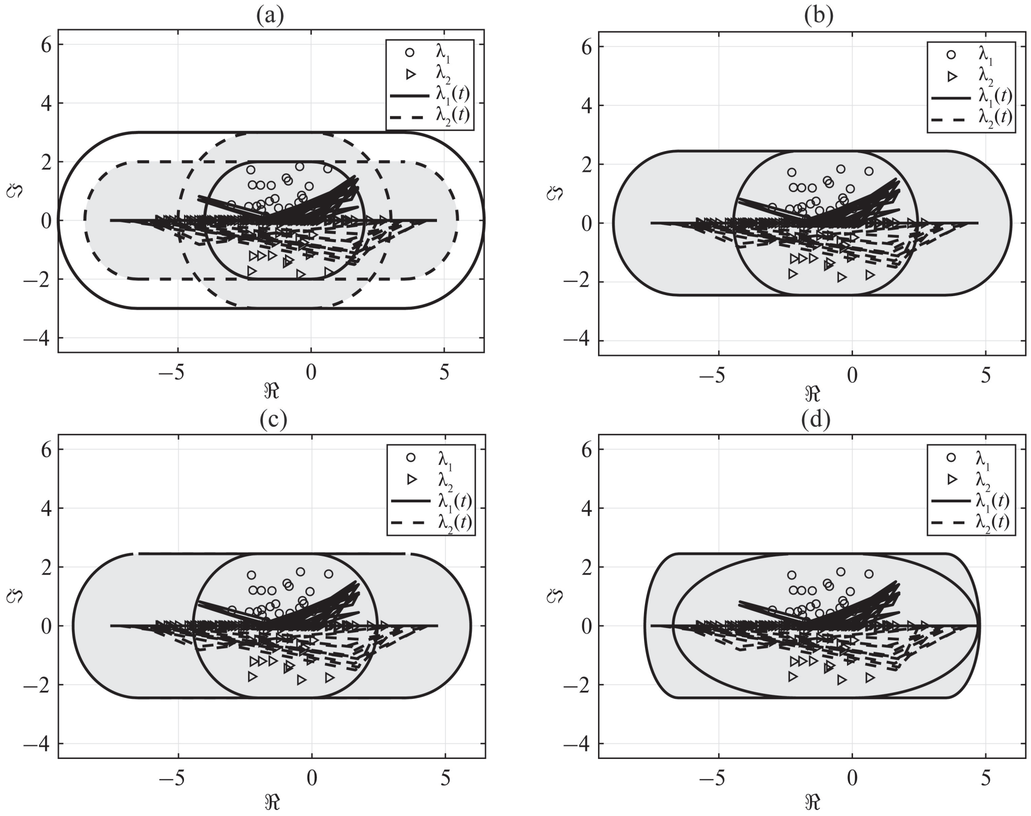

Fig. 2. Localization regions (gray region) of eigenvalues of perturbed matrices Q and $Q(t)$.

Example 2. Consider two parametrically indefinite matrices Q with constant and variable parameters in the forms

$$Q = \begin{bmatrix} -1 & 0 \\ 0 & -1.5 \end{bmatrix} + \underbrace{\begin{bmatrix} r_{11} & 2r_{12} \\ 3r_{21} & 4r_{22} \end{bmatrix}}_{\Delta Q},$$

$$Q(t) = \begin{bmatrix} -1 & 0 \\ 0 & -1.5 \end{bmatrix} + \underbrace{\begin{bmatrix} \sin(t) & 2\cos(1.5t) \\ 3\mathrm{sgn}(\sin(2t)) & 4\mathrm{sgn}(\cos(1.7t)) \end{bmatrix}}_{\Delta Q(t)},$$

where r_{ij}, $i, j = 1, 2$ are pseudorandom numbers uniformly distributed over the interval $(-1; 1)$. Let us consider 100 realizations for each r_{ij}. The matrices ΔQ and $\Delta Q(t)$ have the same m_{ij}, so the estimates of the localization region will be the same.

Figure 2 shows the localization region of the eigenvalues of Q and $Q(t)$ using the results of Lemmas 3–5 (gray regions), where small circles and triangles represent the eigenvalues of the matrix Q with constant parameters, and continuous and dashed lines (inside the gray regions) represent the eigenvalues of the matrix Q with non-stationary parameters. In three out of four figures, the pairs of e-circles coincided due to the variation of d_i, h_i, α, and β, so only two e-circles are shown in three figures. The corresponding figures were obtained using:

- Lemma 3 (Fig. 2a);
- Lemma 4 with $D = diag\{1; 1.23\}$ and $H = diag\{1; 0.81\}$ (Fig. 2b);
- Lemma 5 with $\alpha = \beta = 0.5$ and $D = H = I$ (Fig. 2c);
- Lemma 5 with $\alpha = 0.5$ and $D = diag\{1; 0.52\}$ (Fig. 2d).

The solid boundary of the e-circles corresponds to the figures composed along the rows of the matrix, and the dotted boundary corresponds to the figures composed along the columns of the matrix.

3. CONTROL SYSTEM STABILITY ANALYSIS

This section will consider several applications of the results of the previous section to the analysis and design of control systems.

3.1. Synchronization of Network Systems

Consider a network system consisting of n interconnected agents of the form

$$\dot{x}_i = \sum_{j=1}^{n} q_{ij} x_j + u_i, \quad i = 1, \ldots, n, \tag{13}$$

where $x_i \in \mathbb{R}$, $u_i \in \mathbb{R}$ is the control signal, $|q_{ij}| \leqslant m_{ij}$. It is required to ensure that the condition $\lim\limits_{t\to\infty} x_i(t) = 0$ is satisfied for all x_i by choosing u_i, $i = 1, \ldots, n$.

Let us define the control laws

$$u_i = -q x_i, \quad i = 1, \ldots, n, \tag{14}$$

where $q > 0$.

Use the following notations: $x = col\{x_1, \ldots, x_n\}$, $Q_0 = -qI$, $\Delta Q = (q_{ij})$ and $Q = Q_0 + \Delta Q$. Then (13) and (14) can be rewritten as

$$\dot{x} = Qx. \tag{15}$$

As a result, checking the condition $\lim\limits_{t\to\infty} x_i(t) = 0$ is reduced to checking the stability of the matrix $Q_0 + \Delta Q$, which can be ensured by an appropriate choice of q in (14).

Let $q = -10^3$ and q_{ij} be pseudorandom numbers uniformly distributed over the interval $(-1; 1)$.

To analyze the stability of the matrix $Q_0 + \Delta Q$, we use:

- functions eig (calculating the eigenvalues of a matrix) and lyap (solving the Lyapunov equation) in MatLab, assuming that q_{ij} are known;
- applications to solving the linear matrix inequalities CVX and Yalmip/SeDuMi, assuming q_{ij} to be known;
- Lemma 1 (calculating $\sigma_{\max}\{Q\}$) and Lemma 2 (calculating $\sigma^{\alpha}_{\max}\{Q\}$ with an exhaustive search of α from 0 to 1 with a step of 0.1), assuming q_{ij} to be known;
- Corollary 3 (calculating $\sigma^{\alpha}_{\max}\{Q\}$ with an exhaustive search of α from 0 to 1 with a step of 0.1), assuming q_{ij} to be unknown, but with known m_{ij}.

Figure 3 shows the graphs of the time spent on the operation to determine the stability of $Q_0 + \Delta Q$ depending on the dimension of the matrix (the number of agents in the network) and using the corresponding method. Regardless of whether the corresponding method indicated that the matrix Q is stable or unstable, the corresponding time was recorded to clarify this issue. The calculations were performed in Matlab R2021b on a PC with an AMD Ryzen 5 PRO 4650U processor with Radeon Graphics 2.10 GHz and 8 GB of RAM. The results for CVX and Yalmip/SeDuMi, as well as for Lemma 2 and Corollary 3 were almost identical, so their graphs in Fig. 3 matched in pairs. We also note that when analyzing the proposed results, the maximum calculation time was not reached due to the fact that Matlab R2021b did not generate a matrix with a dimension greater than 25 000.

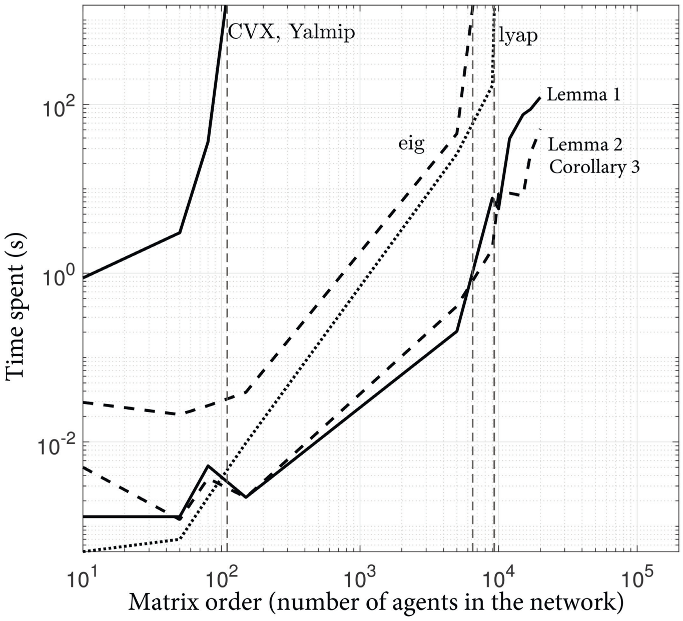

Fig. 3. Dependence of time spent on determining the stability of the system (matrix Q) on the number of agents in the network (n).

Conclusions:

- eig, lyap, CVX and Yalmip/SeDuMi algorithms provide a more accurate result in determining stability (they provide a smaller error in the deviation of the obtained solution from the true value) compared to the proposed estimates;
- the time spent on clarifying the stability issue when using CVX and Yalmip/SeDuMi increases significantly (to a lesser extent when using eig and lyap) with an increase in the number of agents in the network, while the proposed results are the least time-consuming in terms of calculation.

The closed-loop system (15) contains a matrix with diagonal dominance. In [6–18], where matrices with diagonal dominance were also used, it was noted that this is a rather narrow class of systems under study. In the following sections, we will show that the proposed results can be applied to systems with matrices without diagonal dominance. Diagonal dominance will be presented to expressions obtained using the apparatus of Lyapunov functions.

3.2. Stability Analysis of Linear Non-Stationary Systems with Interval-Uncertain Parameters and Matrices without Diagonal Dominance

In this section, we will consider a modification of the Demidovich theorem [19, Theorem 6.1; 23] (the term "Demidovich condition" is also used in the literature) on the study of the stability of linear systems with known non-stationary parameters in the case of interval uncertainty and the presence of external disturbances. Let the system be represented by the equation

$$\dot{x}(t) = A(t)x(t) + F(t)f(t), \tag{16}$$

where $t \geqslant 0$, $x \in \mathbb{R}^n$ is the state vector, $f \in \mathbb{R}^l$ is an external signal such that $\sup\{|f(t)|\} \leqslant \bar{f}$, $F(t) \in \mathbb{R}^{n\times l}$ and $A(t) = (a_{ij}(t)) \in \mathbb{R}^{n\times n}$ are such that $\sup\{\|F(t)\|\} \leqslant \bar{F}$, $A(t) = A_0 + \Delta A(t)$, $A_0 = (a_{ij}^0)$, $\Delta A(t) = (\Delta a_{ij}(t))$, $\Delta\underline{a}_{ii} \leqslant \Delta a_{ii}(t) \leqslant \Delta\overline{a}_{ii}$ and $|\Delta a_{ij}(t)| \leqslant m_{ij}$ for $i \neq j$ and for all t.

Let us introduce the matrix $\bar{A}(t)$, where

$$\begin{gathered}\bar{A}(t)=A(t)+A^{\mathrm{T}}(t)=\bar{A}_0+\Delta\bar{A}(t),\\ \bar{A}_0=(\bar{a}^0_{ij})=(a^0_{ij}+a^0_{ji}),\\ \Delta\bar{A}(t)=(\Delta\bar{a}_{ij}(t))=(\Delta a_{ij}(t)+\Delta a_{ji}(t)),\\ 2\Delta\underline{a}_{ii}\leqslant\Delta\bar{a}_{ii}(t)\leqslant 2\Delta\overline{a}_{ii},\\ |\Delta\bar{a}_{ij}(t)|\leqslant m_{ij}+m_{ji}\quad\text{at}\quad i\neq j.\end{gathered}\tag{17}$$

Note that the system (16) contains a matrix $A(t)$ without diagonal dominance. As will be shown in the theorem below, diagonal dominance will be needed in the matrix $\bar{A}(t)$.

According to Demidovich theorem [19, Theorem 6.1; 23], the system (16) is asymptotically stable for $f(t)\equiv 0$ and with a known matrix $A(t)$ if the eigenvalues of the matrix $A(t)+A^{\mathrm{T}}(t)$ take negative values for all t. Next, we consider a generalization of this theorem to interval indefinite matrices, taking into account the Corollary 3.

Theorem 1. *Denote by σ any upper bound calculated using* (12) *for the eigenvalues of the matrix $\bar{A}(t)$ in* (17). *If $\sigma<0$, then the following bound holds:*

$$|x(t)|\leqslant-\frac{2\|\bar{F}\|\bar{f}}{\sigma}+\mathcal{C}e^{0.5\sigma t},\tag{18}$$

where $\mathcal{C}=\max\left\{0,|x(0)|+\frac{2\|\bar{F}\|\bar{f}}{\sigma}\right\}$.

Proof. We choose the Lyapunov function

$$V=x^{\mathrm{T}}x\tag{19}$$

and find its derivative along the solutions (16) in the form

$$\dot{V}=x^{\mathrm{T}}\bar{A}(t)x+2x^{\mathrm{T}}F(t)f.$$

Let us find the upper estimate

$$\dot{V}\leqslant\sigma x^{\mathrm{T}}x+2|x|\|F(t)\||f|\leqslant\sigma V+2\sqrt{V}\|\bar{F}\|\bar{f}.\tag{20}$$

We solve the inequality (20) in the form

$$\sqrt{V}\leqslant-\frac{2\|\bar{F}\|\bar{f}}{\sigma}+\left(\sqrt{V(0)}+\frac{2\|\bar{F}\|\bar{f}}{\sigma}\right)e^{0.5\sigma t}.\tag{21}$$

Taking into account (19), we obtain

$$|x(t)|\leqslant-\frac{2\|\bar{F}\|\bar{f}}{\sigma}+\left(|x(0)|+\frac{2\|\bar{F}\|\bar{f}}{\sigma}\right)e^{0.5\sigma t}.\tag{22}$$

The expression (22) yields the result (18).

Example 3. A system with constant parameters and a matrix without diagonal dominance. Consider the system (16) with parameters $A=\begin{bmatrix}-1 & 3\\ -2.5 & -2\end{bmatrix}$, $B=[0\ 0.05]^{\mathrm{T}}$ and $u=\sin(t)$. The matrix A is not superstable [6–8] or diagonally dominant [4, 9–18] either in rows or in columns. There are also no $d_1>0$ and $d_2>0$ for the conditions (2) to be satisfied, since the inequalities

$d_1 - 3d_2 > 0$ and $-2.5d_1 + 2d_2 > 0$, composed for the matrix A, and the inequalities $d_1 - 2.5d_2 > 0$ and $-3d_1 + 2d_2 > 0$, composed for the matrix A^{T}, have no solution.

Consider the matrix $\bar{A} = A + A^{\mathrm{T}} = \begin{bmatrix} -2 & 0.5 \\ 0.5 & -4 \end{bmatrix}$. The condition (2) will be satisfied for $\bar{A}$, where $\sigma = \sigma_{\max}(\bar{A}) = -1.5$. The largest eigenvalue of the matrix $A + A^{\mathrm{T}}$ is -1.88. If we use another condition in (2) with $d_1 = 1$ and $d_2 = 0.711$, then the eigenvalue estimate can be improved to $\sigma = \sigma^D_{\max}(\bar{A}) = -1.6445$.

Example 4. A system with non-stationary parameters with a matrix without diagonal dominance. Consider the system (16) with parameters with $A(t) = A_0 + \Delta A(t)$, where $A_0 = A$ from the previous example, $\Delta A(t) = 0.1 \begin{bmatrix} \sin(t) & \cos(t) \\ \sin(2t) & \sin(4t) \end{bmatrix}$. The upper bounds (17) give negative values, therefore, the system (16) is stable.

3.3. Control Law Design for Linear Systems with Matrices without Diagonal Dominance

Consider the system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + F(t)f(t), \tag{23}$$

where $u \in \mathbb{R}^m$ is the control signal, $B(t) \in \mathbb{R}^{n\times m}$, $B(t) = b(t)B_0$, $\underline{b} \leqslant b(t) \leqslant \overline{b} \in \mathbb{R}$, B_0 is a known matrix, the pair $(A(t), B(t))$ is controllable for all t. The remaining notations are as in (16). Assume that the parameters $\Delta A(t)$, $b(t)$, $F(t)$ and $f(t)$ are unknown.

Introduce the control law

$$u = Kx, \tag{24}$$

where $K \in \mathbb{R}^{m\times n}$. Below we formulate theorems that allow us to calculate the matrix K that ensures the stability of the closed-loop system

$$\dot{x}(t) = (A(t) + B(t)K)x(t) + F(t)f(t). \tag{25}$$

Note that neither the matrix $A(t)$ nor the matrix $A(t) + B(t)K$ requires the diagonal dominance property to be satisfied.

Theorem 2. *Let the matrices A, B, and F in (23) be known and constant, and let for the given $\alpha > 0$ there exist the matrix $Q = Q^{\mathrm{T}}$ and the coefficient $\beta > 0$ such that the following conditions are satisfied:*

$$\begin{gathered} \Psi_{ii} < 0, \\ \Psi_{ij} \geqslant 0 \text{ for } i \neq j,\ i, j = 1, \ldots, n, \\ \sigma(Q) > 0, \end{gathered} \tag{26}$$

where

$$\Psi = (\Psi_{ij}) := QA^{\mathrm{T}} + AQ + Y^{\mathrm{T}}B^{\mathrm{T}} + BY + \alpha Q + \beta F^{\mathrm{T}}F, \tag{27}$$

$\sigma(Q)$ is one of the lower bounds for the eigenvalues of matrix Q, obtained using (2), as well as $K = YQ^{-1}$ and $P = Q^{-1}$. Then for the solutions of the system (25) the following estimate will be valid

$$|x(t)| \leqslant \left[\frac{1}{\lambda_{\min}(P)}\left(\frac{\bar{f}^2}{\alpha\beta} + \mathcal{K}e^{-\alpha t}\right)\right]^{0.5}, \tag{28}$$

where $\mathcal{K} = \max\left\{0, x(0)^{\mathrm{T}}Px(0) - \frac{\bar{f}^2}{\alpha\beta}\right\}$.

Proof. We choose the Lyapunov function

$$V = x^{\mathrm{T}} P x, \tag{29}$$

where $P = Q^{-1}$, and find its time derivative along the solutions (25):

$$\dot{V} = x^{\mathrm{T}}[(A + BK)^{\mathrm{T}} P + P(A + BK)]x + 2x^{\mathrm{T}} F f. \tag{30}$$

Denoting $z = col\{x, f\}$ and substituting (29) and (30) into the exponential stability condition $\dot{V} + \alpha V + \gamma f^{\mathrm{T}} f < 0$, $\gamma = 1/\beta$, we obtain

$$z^{\mathrm{T}} \begin{bmatrix} (A + BK)^{\mathrm{T}} P + P(A + BK) + \alpha P & PF \\ \star & -\gamma I \end{bmatrix} z < 0. \tag{31}$$

According to [24], the inequality (31) will be satisfied if the following condition is satisfied:

$$\begin{bmatrix} (A + BK)^{\mathrm{T}} P + P(A + BK) + \alpha P & PF \\ \star & -\gamma I \end{bmatrix} < 0. \tag{32}$$

Using Schur lemma [24] and the fact that $\gamma = 1/\beta$, we rewrite (32) as

$$(A + BK)^{\mathrm{T}} P + P(A + BK) + \alpha P + \beta P F^{\mathrm{T}} F P < 0. \tag{33}$$

Multiplying (33) on the left and right by Q^{-1} and replacing $Y = KQ$, we get

$$\Psi := QA^{\mathrm{T}} + AQ + Y^{\mathrm{T}} B^{\mathrm{T}} + BY + \alpha Q + \beta F^{\mathrm{T}} F < 0. \tag{34}$$

According to Lemmas 1 and 2, the eigenvalues of the symmetric matrices Ψ and Q will be negative and positive, respectively, if the inequalities (26) are satisfied. On the other hand, according to [4] (Theorem 7.2.1), a Hermitian matrix is positive (negative) definite if and only if all its eigenvalues are positive (negative). Therefore, the conditions $\Psi < 0$ and $Q > 0$ will be satisfied if the inequalities (26) are satisfied. The estimate (28) follows from the solution of the inequality $\dot{V} + \alpha V + \gamma f^{\mathrm{T}} f < 0$ taking into account (29) and the estimate $\lambda_{\min}\{P\}|x|^2 \leqslant x^{\mathrm{T}} P x$.

Using the results of Theorem 2, we formulate the following theorem for systems with unknown non-stationary parameters.

Theorem 3. *Consider the system* (23) *with non-stationary parameters. Let there exist the matrix* $Q = Q^{\mathrm{T}}$ *and the coefficient* $\beta > 0$ *such that the conditions hold*

$$\begin{gathered} \Phi_{ii} < 0, \\ \Phi_{ij} \geqslant 0 \text{ for } i \neq j, \\ \sigma(Q) > 0, \end{gathered} \tag{35}$$

at the vertices $|\Delta a_{ij}(t)| \leqslant m_{ij}$ *and* $\underline{b} \leqslant b(t) \leqslant \overline{b}$, *where*

$$\Phi = (\Phi_{ij}) := QA_0^{\mathrm{T}} + A_0 Q + Q\Delta A^{\mathrm{T}}(t) + \Delta A(t) Q + b(t) Y^{\mathrm{T}} B_0^{\mathrm{T}} + b(t) B_0 Y + \alpha Q + \beta \bar{F}^2 I,$$

$\sigma(\Psi)$ *is one of the upper bounds of the matrix* Ψ, *obtained using* (12), *and also* $K = YQ^{-1}$ *and* $P = Q^{-1}$. *Then for solutions of the system* (25) *the estimate* (28) *will be valid.*

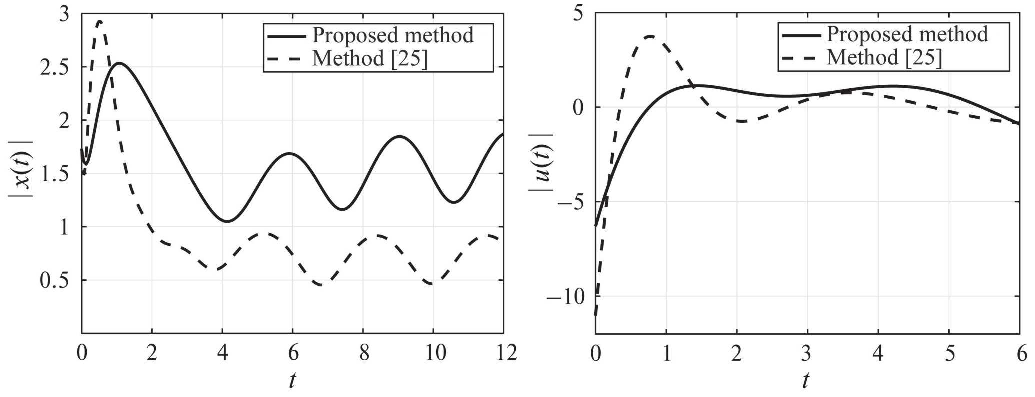

Fig. 4. The transients of $|x(t)|$ and $u(t)$ for the proposed algorithm (solid curves) and the algorithm [25] (dashed curves).

Proof. We will use the results (29)–(34) from the proof of Theorem 2 taking into account non-stationary parameters. Since $A(t) = A_0 + \Delta A(t)$, $B(t) = b(t)B_0$ and $\|F(t)\| \leqslant \bar{F}$, then we rewrite (34) as

$$\Phi := QA_0^{\mathrm{T}} + A_0Q + Q\Delta A^{\mathrm{T}}(t) + \Delta A(t)Q + b(t)Y^{\mathrm{T}}B_0^{\mathrm{T}} + b(t)B_0Y + \alpha Q + \beta \bar{F}^2 I < 0.$$

If the conditions (35) are satisfied at the vertices $|\Delta a_{ij}(t)| \leqslant m_{ij}$ and $\underline{b} \leqslant b(t) \leqslant \overline{b}$, then according to [24] the condition (35) will be satisfied for any $\Delta A(t)$ and $b(t)$ inside the polytope with vertices $|\Delta a_{ij}(t)| \leqslant m_{ij}$ and $\underline{b} \leqslant b(t) \leqslant \overline{b}$. The estimate (28) is obtained similarly to the proof of Theorem 2.

Example 5. Consider the system (23) with parameters $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 2 & 3 \end{bmatrix}$, $B = col\{0; 0; 1\}$, $F = col\{0.1; 0.5; 1\}$ and $f(t) = \sin(t)$.

Obviously, the matrix A is without diagonal dominance, and the structure of the matrix B does not allow the control law $u = Kx$ with $K \in \mathbb{R}^{1\times 3}$ to lead to the closed-loop system with a matrix with diagonal dominance. Therefore, we use Theorem 2 to analyze the localization region of the eigenvalues of the matrix Φ obtained as a result of applying the Lyapunov function method. Using Theorem 2, we obtain $K = col\{-1.3671; -2.3619; -2.5724\}$ and $trace(P) = 25.5858$ for $\alpha = 1$. Using [25], we obtain $K = col\{-2.8862; -4.9244; -3.2136\}$ and $trace(P) = 40.631$ for $\alpha = 1$. In both cases, the goal was $trace(P) \to \min$ for calculating K.

From Fig. 4 it is evident that in the steady state the value of $|x(t)|$ of the proposed algorithm is greater. However, the spike of $|x(t)|$ and the amplitude of the control signal $u(t)$ at the initial moment of time are smaller, and the value of $trace(P)$ is also smaller.

4. CONCLUSION

The paper has described the application of the Gershgorin theorem and theorems derived from it for estimating the localization domain of the eigenvalues of a matrix with constant and known parameters. These results are generalized to estimate the localization domain for matrices with parametric interval uncertainty. The concept of an e-circle is proposed, which allows obtaining more accurate estimates of the localization domain than a direct application of the Gershgorin theorem. The obtained results are applied to the control of network systems, where it is shown that for large-dimensional problems, the proposed results are the least time-consuming in terms of execution time compared to the eig and lyap procedures (commands in MatLab for finding the eigenvalues of a matrix and solving the Lyapunov equation), as well as CVX and Yalmip/SeDuMi for solving linear matrix inequalities. A generalization of the Demidovich condition is proposed for determining the stability of a non-stationary matrix. An approach has been developed for calculating the matrix in a linear control law for control of linear systems where the property of diagonal dominance for matrices in the closed-loop system is not fulfilled.

FUNDING

The work was carried out at the Institute for Problems in Mechanical Engineering of the Russian Academy of Sciences with the support of Goszadanie no. FFNF-2024-0008 (no. 124041100006-1 in EGISU NIOKTR).

REFERENCES

1. *Theory of automatic control. Part 1.*, Edited by A.A., Voronov. Moscow: Vishaya Shkola, 1986.

2. Furtat, I.B. and Gushchin, P.A., *Methods of analysis and design of linear and nonlinear control systems in the presence of disturbances and delays*, Izhevsk: Publishing office “IKI,” 2021.

3. Bellman, R., *Introduction to matrix analysis*, New York: McGraw-Hill, 1970.

4. Horn, R.A. and Johnson, C.R., *Matrix Analysis*, Cambridge: University Press, 1985.

5. Gantmakher, F.R., *Theory of matrices*, New York: Chelsea Pub. Co, 1959.

6. Polyak, B.T. and Shcherbakov, P.S., Superstable Linear Control Systems. I. Analysis, *Automation and Remote Control*, 2002, vol. 63, no. 8, pp. 1239–1254.

7. Polyak, B.T. and Shcherbakov, P.S., Superstable Linear Control Systems. II. Design, *Automation and Remote Control*, 2002, vol. 63, no. 11, pp. 1745–1763.

8. Polyak, B.T., Extended Superstability in Control Theory, *Automation and Remote Control*, 2004, vol. 65, no. 4, pp. 567–576.

9. Uronen, P. and Jutila, E.A.A., Stability via the theorem of Gershgorin, *Int. J. Control*, 1972, vol. 16, no. 6, pp. 1057–1061.

10. Solov'ev, V.N., A generalization of Gershgorin's theorem, *Mathematics of the USSR-Izvestiya*, 1984, vol. 23, no. 3, pp. 545–560.

11. Curran, P.F., On a variation of the Gershgorin circle theorem with applications to stability theory, *IET Irish Signals and Systems Conference (ISSC 2009)*, Dublin, 2009, pp. 1–5.

12. Vijay Hote and Amar Nath Jha, New approach of Gerschgorin theorem in model order reduction, *Int. J. Model. Simulat.*, 2015, vol. 35, pp. 143–149.

13. Li, C.-K. and Zhang, F., Eigenvalue continuity and Gersgorin's theorem, *Electron. J. Linear Algebra*, 2019, vol. 35, pp. 619–625.

14. Kazakova-Frehse, N. and Frick, K., The estimation of a robust domain of attraction using Gersgorin theorem, *Int. J. Robust and Nonlinear Control*, 1998, vol. 8, pp. 295–303.

15. Vijay Hote and Amar Nath Jha, Reduced order state feedback controller design, *2014 Int. Conference on Advances in Engineering & Technology Research (ICAETR – 2014)*, 2014, pp. 1–6.

16. Pachauri, N. and Rani, A., Gerschgorin theorem based stability analysis of chemical process, *2014 Int. Conference on Advances in Engineering & Technology Research (ICAETR – 2014)*, 2014, pp. 1–5.

17. Xie, L., Huang, J., Tan, E., He, F., and Liu, Z., The Stability Criterion and Stability Analysis of Three-Phase Grid-Connected Rectifier System Based on Gerschgorin Circle Theorem, *Electronics*, 2022, vol. 11, no. 20, 3270.

18. Adom-Konaduy, A., Albert Lanor Sackiteyz, and Anokyex M., Local Stability Analysis Of Epidemic Models Using A Corollary Of Gershgorin's Circle Theorem, *Applied Mathematics E-Notes*, 2023, vol. 23, pp. 159–174.

19. Afanasyev, V.N., Kolmanovsky, V.B., and Nosov, V.R., *Mathematical theory of control systems design*, Moscow: Vishaya Shkola, 2003.

20. Tsypkin, Ya.Z. and Bromberg, P.V., On the degree of stability of linear systems, *Bulletin of the USSR Academy of Sciences, OTN*, 1945, no. 12.

21. Rubinchik, A.M., *Approximate method for assessing the quality of regulation in linear systems / Devices and elements of the theory of automation and telemechanics*, Moscow: Mashgiz, 1952.

22. Feldbaum, A.A., On the distribution of roots of the characteristic equation of automatic control systems, *Automation and Remote Control*, 1948, vol. 9, no. 4, pp. 253–279.

23. Khalil, H.K., *Nonlinear System*, Pearson Prentice-Hall, 2002.

24. Boyd, S., El Ghaoui, L., Feron, E., and Balakrishnan, V., *Linear matrix inequalities in system and control theory*, Philadelphia: SIAM, 1994.

25. Nazin, S.A., Polyak, B.T., and Topunov, M.V., Rejection of bounded exogenous disturbances by the method of invariant ellipsoids, *Automation and Remote Control*, 2007, vol. 68, no. 3, pp. 467–486.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 6, pp. 516–530.

Incomplete Measurements-Based Exponential Stabilization and Asymptotic Estimation of Solutions of Linear Neutral Systems

V. E. Khartovskii[*,a], **A. V. Metelskii**[**,b], **and V. V. Karpuk**[***,c]

[*] *Yanka Kupala State University of Grodno, Grodno, Belarus*
[**] *Minsk, Belarus*
[***] *Belarusian National Technical University, Minsk, Belarus*
e-mail: [a]*hartovskij@grsu.by,* [b]*ametelskii@gmail.com,* [c]*vasvaskarpuk@gmail.com*

Received January 28, 2025
Revised March 24, 2025
Accepted March 28, 2025

Abstract—This paper is devoted to a linear autonomous differential-difference system of neutral type with lumped delays. For such systems, we propose existence criteria for output-feedback controllers based on incomplete measurements that ensure a given spectrum of the closed-loop system or its exponential stabilization. In addition, we prove existence criteria for observers forming asymptotic estimates with errors described by linear homogeneous systems with a predetermined characteristic quasipolynomial or exponential stability. All the considerations are constructive and contain a method for designing a corresponding controller or observer.

Keywords: differential-difference system, neutral type, delay, exponential stabilization, modal controllability, controller, observer

DOI: 10.31857/S0005117925060025

1. INTRODUCTION

The delay effect is inherent in almost all control processes. Therefore, it should be taken into account when building engineering, economic, and other models [1–4]. The general theory of delayed systems, as well as their applications, was studied in rather many works (for example, see the introduction in [3, 4]). In this paper, we investigate the stabilization problem for neutral delay systems. Such systems describe the behavior of plants and processes whose rate of evolution depends on both their previous states and their velocities, e.g., the motion of a pendulum with a viscous filler [2], the plunge grinding model, and plants whose dynamics are described by systems with distributed delays (in particular, telegraph equations). Let us provide other particular examples of stabilization problems for linear systems of neutral type. When studying the oscillations of the current collector of a moving locomotive far from the support (placed behind the current collector), it is necessary to consider the effect of the reflected waves of the contact wire from the strings supporting this wire and from the support placed in front of the moving current collector. For such a mechanical system, one naturally encounters the stabilization problem [5]. Another example is the stabilization problem of a system arising during the translational and rectilinear motion of some mass under the action of a linear restoring force proportional to the coordinate and some nonconservative force [6, p. 235]. Some time is needed to trigger the system's sensitive elements detecting the displacement, velocity, and acceleration of the mass, as well as the relay and servomotor; therefore, one obtains a model in the form of a linear autonomous system of neutral type [6, p. 235].

Research into the stabilization problem of delayed systems was initiated in [7, 8] and then picked up by many scientists [9–16] (see also the bibliography therein). However, despite a rather large flow of publications in this direction, the stabilization problem has not been fully studied to date.

In general, the spectrum of linear systems with aftereffect is infinite, so the analysis and subsequent elimination of unstable eigenvalues from the spectrum requires some computational effort [15]. In this regard, a more universal approach to stabilize the system is to assign a finite spectrum, usually consisting of numbers with negative real parts [17–19]. A significant disadvantage of this system stabilization approach is the solvability conditions of the corresponding (spectrum assignment) problem, which are more stringent compared to the stabilization conditions.

Modal controllability is a more general problem than finite spectrum assignment: it is required to tune the coefficients of the characteristic quasipolynomial of a system [20–22].

The Lyapunov–Krasovskii and Lyapunov–Razumikhin methods are effective for analyzing the stability of delayed systems. They allow formulating the solvability conditions of the control problem in terms of matrix inequalities [23, Chap. 3–7]. This approach to controller analysis and design provides constructive finite-dimensional conditions for its existence and can be extended to other problems. For example, the control law designed in [24] limits the influence of disturbances and measurement noise; the stability conditions of input data presented therein were described in terms of matrix inequalities.

In contrast to the above method, based to a greater extent on the differential properties of the control system, the approach proposed in this article is of a purely algebraic nature. A polynomial $\det W(p,\lambda)$, where $W(p,e^{-ph})$ is the characteristic matrix of a closed-loop system (in the case of a dynamic controller), is treated as an element of an ideal $\mathfrak{J}$ generated by a system of polynomials, i.e., algebraic complements to the elements of the last row of the matrix $W(p,\lambda)$. Therefore, the class of possible characteristic quasipolynomials $\det W(p,e^{-ph})$ can be described by computing the Gröbner basis of the ideal $\mathfrak{J}$. This circumstance reduces all controller/observer design computations to operations in the ring of polynomials. This idea was utilized in [19, 25, 26] to construct a feedback controller ensuring, after a finite time, zero values for all components of the original open-loop system, i.e., a finite stabilization controller [27]. (In other words, such a controller completely damps/calms the original open-loop system.) Such a problem is solved by constructing appropriate feedback so that the closed-loop system becomes a finite-spectrum system pointwise degenerate in the directions corresponding to the components of the solution vector of the original system [19, 25]. These ideas were extended to systems of neutral type in [26] and systematized in the monograph [4]. The next step in exploring the finite stabilization problem was the development of output-feedback controllers based on available output observations. For delayed single-input single-output (SISO) systems, such a problem was considered in [27]; for multi-input systems of neutral type, in [28, 29].

In this paper, utilizing the spectrum control methods for neutral systems [15] and the block diagrams of feedback controllers with incomplete measurements [28, 29], we prove existence criteria for output-feedback controllers based on available output observations that solve the problems of modal controllability and stabilization. In addition, we propose methods for designing two types of asymptotic observers and establish criteria for their existence.

2. NOTATION

Consider a linear autonomous differential-difference system of neutral type with commensurate delays:

$$\dot{x}(t) = A_0 x(t) + \sum_{j=1}^{m} \left(A_j x(t-jh) + D_j \dot{x}(t-jh)\right) + \sum_{j=0}^{m} b_j u(t-jh), \quad t > 0, \tag{1}$$

$$y(t) = \sum_{j=0}^{m} c_j' x(t-jh), \quad t \geqslant 0, \tag{2}$$

$$x(t) = \eta(t), \quad t \in [-mh, 0]. \tag{3}$$

Here, $x \in \mathbb{R}^n$ is the column vector of the solution of system (1) ($n \geqslant 2$); $0 < h$ is a constant delay; $A_0, A_j, D_j \in \mathbb{R}^{n\times n}$ and $b_j \in \mathbb{R}^n$, $c_j' \in \mathbb{R}^n$; the dash symbol ($'$) indicates the transpose; u is the control input (a scalar piecewise continuous function); y is the observed output (a scalar signal). By assumption, the initial function η is continuous with a piecewise continuous derivative. In this case, there exists a unique continuous solution with a piecewise continuous derivative. Throughout this paper, the initial function η is supposed to be unknown.

This study pursues the following objective: based on available observations of the output of (2), design output-feedback controllers that ensure a given characteristic quasipolynomial of the closed-loop system or its exponential stabilization. The remainder of this paper is organized as follows. First (see Section 3), two types of asymptotic observers are constructed using the controller design methods from [15]. Then (Section 4), in order to obtain feedback controllers based on available output observations, additional loops are incorporated into the controller structure from [15] in the form of asymptotic observers according to the principle developed in [28, 29]. Finally, an illustrative example is given in Section 5.

Let p, $\lambda \in \mathbb{C}$, where $\mathbb{C}$ is the set of complex numbers. Also, we introduce the following notation:

$$A(p,\lambda) = A_0 + \sum_{j=1}^{m}(A_j + pD_j)\lambda^j; \tag{4}$$

$W(p, e^{-ph}) = pI_n - A(p, e^{-ph})$ is the characteristic matrix ($I_n \in \mathbb{R}^{n\times n}$ means an identity matrix of order n); $w(p, e^{-ph}) = |W(p, e^{-ph})|$ is the characteristic quasipolynomial of the homogeneous ($u = 0$) system (1). From this point onwards, $|W|$ means the determinant of an arbitrary square matrix W.

Let $\phi \in \mathbb{N}$ be an arbitrary number. A quasipolynomial $d(p, e^{-ph})$, where

$$d(p,\lambda) = \sum_{i=0}^{\phi} \theta_i(\lambda)p^i$$

and $\theta_i(\lambda)$ are some polynomials with $\theta_\phi(0) = 1$, will be called a quasipolynomial of neutral type. If $\theta_\phi(\lambda) = 1$, we have a quasipolynomial $d(p, e^{-ph})$ of delayed type as a special case. The characteristic quasipolynomial $w(p, e^{-ph})$ of the homogeneous system (1) is in general a quasipolynomial of neutral type and $\deg_p w(p,\lambda) = n$.

Let $\mathbb{R}^{r\times m}[\lambda]$ and $\mathbb{C}^{r\times m}[\lambda]$ be the sets of matrices of dimensions $r \times m$ whose elements are polynomials of the variable λ with real and complex coefficients, respectively. (If $r = m = 1$, the superscript will be omitted.) In addition, let λ_h and p_D be the shift and differentiation operators, respectively, i.e., $p_D^i \lambda_h^j f(t) = f^{(i)}(t - jh)$ for a function f and integers $i, j \geqslant 0$.

To write the equations of the controllers and observers compactly, we introduce the set $\mathfrak{Q}^{r\times m}$ $\left(\mathfrak{Q}^{1\times 1} = \mathfrak{Q}\right)$, consisting of all mappings $\mathcal{Q} : f \mapsto \mathcal{Q}[f]$, where $f(t)$, $t \in \mathbb{R}$ is an arbitrary continuous (scalar or vector) function with a piecewise continuous derivative. (Square brackets are used to distinguish mappings and functions.) Each mapping $\mathcal{Q} \in \mathfrak{Q}^{r\times m}$ is given by the following elements: 1) $q_i(\lambda) \in \mathbb{R}^{r\times m}[\lambda]$, $i = 0, 1$; 2) $P = \{\alpha_k \pm \mathbf{i}\beta_k,\ \alpha_k,\ \beta_k \in \mathbb{R},\ k = \overline{1, n_1}\}$, representing the set of real and complex conjugate numbers ($\mathbf{i}$ denotes the imaginary unit); 3) $\widehat{q}_{ki}(\lambda) \in \mathbb{C}^{r\times m}[\lambda]$, $k = \overline{1, n_1}$, $i = \overline{1, n_2}$ ($n_1 \geqslant 1$, $n_2 \geqslant 0$ are integers). Each mapping of this kind acts according to the rule

$$\mathcal{Q}[f(t)] = q_0(\lambda_h)f(t) + q_1(\lambda_h)\dot{f}(t-h) + \sum_{k=1}^{n_1}\sum_{i=0}^{n_2}\int_0^h \widehat{q}_{ki}(\lambda_h)f(t-s)e^{p_k s}\frac{s^i}{i!}\,ds, \quad t > 0, \tag{5}$$

where $p_k \in P$. The matrices $\widehat{q}_{ki}(\lambda)$ in (5) and the set P possess the following property: with the Euler formula applied ($e^{\mathbf{i}\varphi} = \cos\varphi + \mathbf{i}\sin\varphi$), the expression (5) becomes

$$\mathcal{Q}[f(t)] = q_0(\lambda_h)f(t) + q_1(\lambda_h)\dot{f}(t-h) + \sum_{j=0}^{\widehat{n}_1} \int_0^h r_j(s)f(t-jh-s)\,ds, \tag{6}$$

where $\widehat{n}_1 = \max_{k,i}\{\deg_\lambda \widehat{q}_{ki}(\lambda)\}$, $r_j(s) = \sum_{k=1}^{n_1} e^{\alpha_k s}(\cos(\beta_k s)\nu_{jk}(s) + \sin(\beta_k s)\mu_{jk}(s))$, $(\alpha_k + \mathbf{i}\beta_k) \in P$, and $\nu_{jk}(s), \mu_{jk}(s) \in \mathbb{R}^{r\times m}[s]$ ($\deg_s \nu_{kj} \leqslant n_2$, $\deg_s \nu_{kj} \leqslant n_2$). Thus, all expressions in the relation (6) are real numbers.

When the original system is closed by controllers containing the terms (5) (equivalently, the terms (6)), distributed delays described by integral terms in (6) may appear in the closed-loop system. In this case, the distributed delay terms (see the expression (5)) are associated with the expressions $\widehat{q}_{ki}(e^{-ph})\int_0^h e^{-(p-p_k)s}s^i/i!\,ds$ in the characteristic matrix of the closed-loop system. Calculating the integrals of these expressions and then letting $\lambda = e^{-ph}$ yield the integer fractional rational functions [19]

$$\int_0^h e^{-(p-p_k)s}s^i/i!ds\bigg|_{e^{-ph}=\lambda} = \frac{(-1)^{i+1}}{i!}\frac{d^i}{dp^i}\left(\frac{\lambda - e^{-p_k h}}{e^{-p_k h}(p-p_k)}\right), \quad i = 0, 1, \ldots. \tag{7}$$

The expression (5) (or (6)) is associated with the matrix

$$\mathcal{Q}[e^{pt}]e^{-pt}\bigg|_{e^{-ph}=\lambda} = Q(p,\lambda)$$

in the characteristic matrix of the closed-loop system, where

$$Q(p,\lambda) = q_0(\lambda) + p\lambda q_1(\lambda) + q(p,\lambda) \tag{8}$$

and $q(p,\lambda)$ is a matrix of fractional rational functions of the form $\frac{\omega_1(p,\lambda)}{\omega_2(p)}$, proper in the variable p ($\omega_1(p,\lambda)$ and $\omega_2(p)$ are polynomials with complex coefficients such that $\deg_p \omega_1(p,\lambda) < \deg_p \omega_2(p)$). We further suppose that if $\widehat{W}(p, e^{-ph})$ is the characteristic matrix of a neutral system with a distributed delay given by (6), then the matrix $\widehat{W}(p,\lambda)$ is obtained by first calculating the integrals (7) and then letting $e^{-ph} = \lambda$ in the resulting expression.

For a given mapping $\mathcal{Q}$ (5), the transposed mapping $\mathcal{Q}'$ is the one obtained from (5) by replacing $q_0(\lambda)$, $q_1(\lambda)$, and $\widehat{q}_{ki}(\lambda)$ with $q_0'(\lambda)$, $q_1'(\lambda)$, and $\widehat{q}_{ki}'(\lambda)$, respectively.

3. ASYMPTOTIC ESTIMATION OF THE SOLUTION

In this section, we construct observers forming asymptotic estimates of the solution of the original system (2) from the measurements (1) with errors vanishing at a given or exponential rate, determined by the roots of the characteristic quasipolynomial. Further, these results will be needed to design a stabilizing output-feedback controller based on available output observations.

We define the following linear system of neutral type:

$$\begin{aligned} \dot{z}_1(t) &= A(p_{_D}, \lambda_h)z_1(t) + \mathcal{L}_1[z_2(t)] + b(\lambda_h)u(t), \\ \dot{z}_2(t) &= \beta_0(p_{_D})c'(\lambda_h)z_1(t) + \mathcal{L}_2[z_2(t)] - \beta_0(p_{_D})y(t), \quad t > 0, \end{aligned} \tag{9}$$

where the matrix $A(p,\lambda)$ is given by (4), $\mathcal{L}_1 \in \mathfrak{Q}^{n\times 1}$, $\mathcal{L}_2 \in \mathfrak{Q}^{1\times 1}$, $\beta_0(p) \in \mathbb{R}_0[p]$, and $\mathbb{R}_0[p] = \{1, p + \widehat{\alpha} : \widehat{\alpha} \in \mathbb{R}\}$ is the set of polynomials that have the form $p + \widehat{\alpha}$ or are equal to 1. For system (9), we choose any initial condition

$$z(t) = \varphi(t), \quad t \in [-h_0, 0], \tag{10}$$

where φ is a continuous function with a piecewise continuous derivative and h_0 is the delay of system (9).

We take the component z_1 of the solution vector $z = \mathrm{col}[z_1, z_2]$ of system (9) as an estimate of the solution x of system (1), (2) given the control input u. Obviously, the function $\zeta = z_1 - x$, representing the error of the estimate z_1 of the solution x, is a component of the solution of the homogeneous system

$$\begin{aligned} \dot{\zeta}(t) &= A(p_D, \lambda_h)\zeta(t) + \mathcal{L}_1[z_2(t)], \\ \dot{z}_2(t) &= \beta_0(p_D)c'(\lambda_h)\zeta(t) + \mathcal{L}_2[z_2(t)], \quad t > 0. \end{aligned} \tag{11}$$

Consider the characteristic matrix $W_z(p, \lambda)$ of system (11) (the homogeneous ($u = 0$) system (9)):

$$W_z(p, \lambda) = \begin{bmatrix} pI_n - A(p, \lambda) & -L_1(p, \lambda) \\ -\beta_0(p)c'(\lambda) & p - L_2(p, \lambda) \end{bmatrix}, \tag{12}$$

where $L_i(p, \lambda) = \mathcal{L}_i[e^{pt}]e^{-pt}$. Let us introduce the polynomial

$$g(p, \lambda) = \sum_{i=0}^{n+1} p^i g_i(\lambda), \quad g_i(\lambda) \in \mathbb{R}[\lambda], \quad g_{n+1}(0) = 1. \tag{13}$$

Generally speaking, the quasipolynomial $d(p, e^{-ph})$ is of neutral type.

Definition 1. System (1), (2) is said to have an observer (9) with a given characteristic polynomial if, for any polynomial (13), there exist $\mathcal{L}_1 \in \mathfrak{Q}^{n\times 1}$, $\mathcal{L}_2 \in \mathfrak{Q}^{1\times 1}$, and $\beta_0(p) \in \mathbb{R}_0[\lambda]$ such that

$$\left|W_z(p, \lambda)\right| = g(p, \lambda). \tag{14}$$

Remark 1. The main goal of observer design is to obtain an estimate for the solution of an original system. Therefore, when designing an observer with a given characteristic quasipolynomial, the quasipolynomial (13) should be chosen so that system (11) be asymptotically or exponentially stable. Regarding the computational complexity of solving system (11), the most convenient choice is a polynomial (13) that does not depend on the variable λ and has roots with negative real parts.

Definition 2. System (1), (2) is said to have an exponentially stable observer (9) if there exist $\mathcal{L}_1 \in \mathfrak{Q}^{n\times 1}$, $\mathcal{L}_2 \in \mathfrak{Q}^{1\times 1}$, and $\beta_0(p) \in \mathbb{R}_0[p]$ such that system (11) is exponentially stable.

Remark 2. A linear homogeneous autonomous system of neutral type is exponentially stable if and only if [14] its characteristic quasipolynomial possesses exponential stability (i.e., the roots p_i of the characteristic equation satisfy the inequality $\mathbf{Re}\, p_i < \varepsilon\ \exists \varepsilon < 0$). In this case, the difference equation describing the jump behavior of the first derivatives of the solution is exponentially stable [14]. We illustrate the above on an example of the system

$$\dot{x}(t) = \mathcal{Q}[x(t)], \tag{15}$$

where the mapping $\mathcal{Q}$ is given by (6). (All matrices in (6) have dimensions $n \times n$.) Let $W_0(p, e^{-ph})$ be the characteristic matrix of system (15), $W_0(p, \lambda) = p(I_n - \lambda q_1(\lambda)) - q_0(\lambda) - q(p, \lambda)$ (see (8)). We introduce the sets

$$\Delta_0 = \left\{p \in \mathbb{C} : \left|W_0(p, e^{-ph})\right| = 0\right\}, \quad \Delta_1 = \left\{\lambda \in \mathbb{C} : \left|I_n - \lambda q_1(\lambda)\right| = 0\right\}. \tag{16}$$

For the exponential stability of system (15), it is necessary and sufficient that

$$\mathbf{Re}\, p < -\varepsilon \quad \exists \varepsilon > 0, \quad p \in \Delta_0. \tag{17}$$

In this case, the exponential stability of the difference equation implies

$$|\lambda| > 1, \quad \lambda \in \Delta_1. \tag{18}$$

Consider system (1). Denoting $D(\lambda) = \sum_{j=1}^{m} \lambda^j D_j$, we formulate existence criteria for an observer with a given characteristic quasipolynomial.

Theorem 1. *For system* (1), (2) *to have an observer* (9) *with a given characteristic quasipolynomial, it is necessary and sufficient that*

$$\operatorname{rank}\begin{bmatrix} W(p,e^{-ph}) \\ c'(e^{-ph}) \end{bmatrix} = n \quad \forall p \in \mathbb{C}, \quad \operatorname{rank}\begin{bmatrix} I_n - D(\lambda) \\ c'(\lambda) \end{bmatrix} = n \quad \forall \lambda \in \mathbb{C}. \tag{19}$$

The proof of this result is provided in the Appendix.

The following theorem represents an existence criterion for an exponentially stable observer.

Theorem 2. *For system* (1), (2) *to have an exponentially stable observer* (9), *it is necessary and sufficient that*

$$\begin{gathered}\operatorname{rank}\begin{bmatrix} W(p,e^{-ph}) \\ c'(e^{-ph}) \end{bmatrix} = n \quad \forall p \in \mathbb{C}, \quad \mathbf{Re}\, p \geqslant \varepsilon_1, \quad \exists \varepsilon_1 < 0, \\ \operatorname{rank}\begin{bmatrix} I_n - D(\lambda) \\ c'(\lambda) \end{bmatrix} = n \quad \forall \lambda \in \mathbb{C}, \quad |\lambda| \leqslant 1. \end{gathered} \tag{20}$$

See the proof in the Appendix.

4. MODAL CONTROLLABILITY AND EXPONENTIAL STABILIZATION

We define a dynamic output-feedback controller based on available output measurements:

$$\begin{gathered} u(t) = \alpha_0(p_D)x_1(t), \\ \dot{x}_1(t) = \mathcal{Q}_{11}[x_1(t)] + \mathcal{Q}_{12}[x_2(t)], \\ \dot{x}_2(t) = b(\lambda_h)\alpha_0(p_D)x_1(t) + A(p_D,\lambda_h)x_2(t) + \mathcal{Q}_{23}[x_3(t)], \\ \dot{x}_3(t) = \alpha_1(p_D)c'(\lambda_h)x_2(t) + \mathcal{Q}_{33}[x_3(t)] - \alpha_1(p_D)y(t), \quad t > 0, \end{gathered} \tag{21}$$

where x_1, $x_3 \in \mathbb{R}$ and $x_2 \in \mathbb{R}^n$ are auxiliary variables; $\mathcal{Q}_{11} \in \mathfrak{Q}$, $\mathcal{Q}_{12} \in \mathfrak{Q}^{1\times n}$, $\mathcal{Q}_{23} \in \mathfrak{Q}^{n\times 1}$, $\mathcal{Q}_{33} \in \mathfrak{Q}$, and $\alpha_i(p) \in \mathbb{R}_0[p]$, $i = 0, 1$.

Let us close system (1), (2) with the controller (21). Obviously, system (1), (2), (21) is linear inhomogeneous autonomous of neutral type with commensurable lumped and distributed delays, and its inhomogeneous part depends on the output $y(t)$. Following (2), we replace the function $y(t)$ in the inhomogeneous part with $c'(\lambda_h)x(t)$ to obtain the homogeneous one. The characteristic matrix $\overline{W}(p,\lambda)$ of this homogeneous system is given by

$$\overline{W}(p,\lambda) = \begin{bmatrix} pI_n - A(p,\lambda) & -\alpha_0(p)b(\lambda) & 0_{n\times n} & 0_{n\times 1} \\ 0_{1\times n} & p - Q_{11}(p,\lambda) & -Q_{12}(p,\lambda) & 0 \\ 0_{n\times n} & -\alpha_0(p)b(\lambda) & pI_n - A(p,\lambda) & -Q_{23}(p,\lambda) \\ \alpha_1(p)c'(\lambda) & 0 & -\alpha_1(p)c'(\lambda) & p - Q_{33}(p,\lambda) \end{bmatrix}, \tag{22}$$

where $Q_{ij}(p,\lambda) = \mathcal{Q}_{ij}[e^{pt}]e^{-pt}$ and $0_{i\times j} \in \mathbb{R}^{i\times j}$ $(i,\, j > 1)$ is a zero matrix of appropriate dimensions.

Definition 3. System (1), (2) is said to be modally controllable (by the output) if, for any polynomial

$$\chi(p,\lambda) = \chi_1(p,\lambda)\chi_2(p,\lambda), \tag{23}$$

where $\chi_k(p,\lambda) = \sum_{i=0}^{n+1} p^i \chi_{ki}(\lambda)$, $\chi_{ki} \in \mathbb{R}[\lambda]$, $k = 1, 2$, and $\chi_{k\,n+1}(0) = 1$, there exists a controller (21) such that the characteristic matrix of the closed-loop system (1), (2), (21) satisfies

$$\left|\overline{W}(p,\lambda)\right| = \chi(p,\lambda). \tag{24}$$

Generally speaking, the quasipolynomial $\chi(p,e^{-ph})$ is of neutral type.

Definition 4. System (1), (2) is said to be exponentially stabilizable (by the output) if there exists a controller (21) such that the closed-loop system (1), (2), (21) is exponentially stable.

The following theorems are criteria for the modal controllability and exponential stabilizability of system (1), (2) in the class of controllers (21).

Theorem 3. *System* (1), (2) *is modally controllable in the class of controllers* (21) *if and only if*

$$\begin{gathered}\operatorname{rank}\bigl[W(p,e^{-ph}),b(e^{-ph})\bigr]=n\quad \forall p\in\mathbb{C},\\ \operatorname{rank}\bigl[I_n-D(\lambda),\,b(\lambda)\bigr]=n\quad \forall\lambda\in\mathbb{C},\end{gathered}\tag{25}$$

and conditions (19) *hold.*

See the proof in the Appendix.

Theorem 4. *System* (1), (2) *is exponentially stabilizable in the class of controllers* (21) *if and only if*

$$\begin{gathered}\operatorname{rank}\bigl[W(p,e^{-ph}),b(e^{-ph})\bigr]=n\quad \forall p\in\mathbb{C},\quad \mathbf{Re}\,p\geqslant\varepsilon_0,\quad \exists\varepsilon_0<0,\\ \operatorname{rank}\bigl[I_n-D(\lambda),\,b(\lambda)\bigr]=n\quad \forall\lambda\in\mathbb{C},\quad |\lambda|\leqslant 1,\end{gathered}\tag{26}$$

and conditions (20) *hold.*

See the proof in the Appendix.

5. EXAMPLE

Let system (1), (2) be of the second order and be described by the following matrices and delay:

$$\begin{gathered}A(p,\lambda)=\begin{bmatrix}-\dfrac{1}{2}p\lambda & -3+\lambda\\ -\dfrac{1}{3} & -\dfrac{5}{12}\lambda\end{bmatrix},\quad b(\lambda)=\begin{bmatrix}0\\ 2\lambda-\lambda^2\end{bmatrix},\\ c(\lambda)=[0,\,-1],\quad h=\ln 2.\end{gathered}\tag{27}$$

The original system with the matrices (27) has an infinite spectrum, and its characteristic quasipolynomial ($\lambda=e^{-ph}$) is given by

$$w(p,\lambda)=\frac{1}{2}p^2(\lambda+2)+\frac{5}{24}p\lambda(\lambda+2)+\frac{\lambda}{3}-1.$$

The quasipolynomial $w(p,e^{-ph})$ has a positive root since $w(0,1)=-\frac{2}{3}<0$; $\lim\limits_{p\to+\infty}w(p,e^{-ph})=+\infty$. Thus, the unperturbed system is not exponentially stable.

Obviously, the first condition in (25) is violated for $p=-1$ and the second for $\lambda=-1$. This means the validity of conditions (26). The first condition in (19) also holds but the second condition fails for $\lambda=-2$; therefore, conditions (20) are true. Thus, the results of [28] (the design of an incomplete measurements-based controller ensuring complete stabilization (simultaneously finite and asymptotic stabilization and finite spectrum assignment) or those of [29] (only finite stabilization) are inapplicable here. However, the conditions of Theorem 4 are satisfied, so we can construct a controller based on incomplete measurements to exponentially stabilize the closed-loop system. Looking ahead, note that the set of roots of the characteristic quasipolynomial of this closed-loop system contains the points $p=-1$ and the roots of the equation $e^{-ph}=\lambda$ with $\lambda=-2$, at which conditions (19), (25) are violated.

Now we proceed to the controller design (21).

1. Following [15], we construct a controller (A.5). Necessary calculations [15] yield

$$\begin{gathered}
u(t) = x_1(t), \\
\dot{x}_1(t) = \left[\frac{5}{6} - \frac{1}{6}\lambda_h\right]\dot{x}_1(t-h) + \left[-6 + \frac{65}{12}\lambda_h - \frac{29}{24}\lambda_h^2\right]x_1(t) \\
+ \int\limits_0^h (-12 + 6\lambda_h)x_1(t-s)e^s ds + \left[\frac{-5}{72}, \frac{5}{72}\right]\dot{x}(t-h) \\
+ \left[\frac{-223}{72} - 2\lambda_h, \frac{25}{3} + \frac{185}{288}\lambda_h\right]x(t) + \int\limits_0^h \left[1, -\frac{9}{2}\right]e^s x(t-s)ds.
\end{gathered} \tag{28}$$

In this case, the matrix (A.6) has the form

$$W_x(p,\lambda) = \begin{bmatrix} p + \dfrac{p\lambda}{2} & 3 - \lambda & 0 \\ \dfrac{1}{3} & p + \dfrac{5}{12}\lambda & (2-\lambda)\lambda \\ \nu_1(p,\lambda) & \nu_2(p,\lambda) & \nu_3(p,\lambda) \end{bmatrix}, \tag{29}$$

where

$$\begin{gathered}
\nu_1(p,\lambda) = \frac{5p\lambda}{72} - \frac{1-2\lambda}{p-1} + \frac{223}{72} + 2\lambda, \quad \nu_2(p,\lambda) = -\frac{5p\lambda}{72} + \frac{9(1-2\lambda)}{2(p-1)} - \frac{25}{3} - \frac{185\lambda}{288}, \\
\nu_3(p,\lambda) = p - \frac{5p\lambda}{6} + \frac{p\lambda^2}{6} + 6\frac{(1-2\lambda)(2-\lambda)}{p-1} + 6 - \frac{65\lambda}{12} + \frac{29\lambda^2}{24}.
\end{gathered}$$

Straightforward calculations finally lead to $|W_x(p,\lambda)| = (1 - \frac{\lambda}{3})(1 - \frac{\lambda}{2})(1 + \frac{\lambda}{2})(p+1)(p+2)(p+3)$.

2. We construct an exponentially stable observer (9). For this purpose, following the proof of Theorem 4, we construct a controller (A.2) for system (A.1) that exponentially stabilizes the closed-loop system (9), (A.2). Then, according to (A.3), we obtain an exponentially stable observer (9) with

$$\begin{gathered}
\mathcal{L}_1[z_2] = \begin{bmatrix} \dfrac{-79}{4}\lambda_h - \dfrac{31}{24}\lambda_h^2 - \dfrac{5}{24}\lambda_h^3 - 36 \\ \dfrac{25}{288}\lambda_h^3 - \dfrac{155}{144}\lambda_h^2 + \dfrac{8}{3}\lambda_h + 12 \end{bmatrix} z_2(t), \\
\mathcal{L}_2[z_2] = \frac{-1}{2}\dot{z}_2(t-h) + \left(\frac{5}{24}\lambda_h^2 - \frac{31}{12}\lambda_h - 6\right)z_2(t).
\end{gathered}$$

The characteristic matrix (12) has the form

$$W_z(p,\lambda) = \begin{bmatrix} \dfrac{p(2+\lambda)}{2} & 3-\lambda & \dfrac{79}{4}\lambda + \dfrac{31}{24}\lambda^2 + \dfrac{5}{24}\lambda^3 + 36 \\ \dfrac{1}{3} & \dfrac{5\lambda}{12} + p & -\dfrac{25}{288}\lambda^3 + \dfrac{155}{144}\lambda^2 - \dfrac{8}{3}\lambda - 12 \\ 0 & 1 & p + \dfrac{1}{2}p\lambda - \dfrac{5}{24}\lambda^2 + \dfrac{31}{12}\lambda + 6 \end{bmatrix} \tag{30}$$

and $|W_z(p,\lambda)| = \left(1 + \frac{1}{2}\lambda\right)^2 (p+3)(p+2)(p+1)$.

3. Using the parameters of the above controller and observer, we construct a controller (21):

$$
\begin{aligned}
u(t) &= x_1(t),\\
\dot{x}_1(t) &= \left[\frac{5}{6}-\frac{1}{6}\lambda_h\right]\dot{x}_1(t-h)+\left[-6+\frac{65}{12}\lambda_h-\frac{29}{24}\lambda_h^2\right]x_1(t)\\
&\quad+\int\limits_0^h(-12+6\lambda_h)x_1(t-s)e^s ds+\left[\frac{-5}{72},\frac{5}{72}\right]\dot{x}_2(t-h)\\
&\quad+\left[\frac{-223}{72}-2\lambda_h,\,\frac{25}{3}+\frac{185}{288}\lambda_h\right]x_2(t)+\int\limits_0^h\left[1,-\frac{9}{2}\right]e^s x(t-s)ds,\\
\dot{x}_2(t) &= \begin{bmatrix}0\\ 2\lambda_h-2\lambda_h^2\end{bmatrix}x_1(t)+\begin{bmatrix}\frac{1}{2} & 0\\ 0 & 0\end{bmatrix}\dot{x}_2(t-h)+\begin{bmatrix}0 & -3+\lambda_h\\ -\frac{1}{3} & -\frac{5}{12}\lambda_h\end{bmatrix}x_2(t)\\
&\quad+\begin{bmatrix}\frac{-79}{4}\lambda_h-\frac{31}{24}\lambda_h^2-\frac{5}{24}\lambda_h^3-36\\ \frac{25}{288}\lambda_h^3-\frac{155}{144}\lambda_h^2+\frac{8}{3}\lambda_h+12\end{bmatrix}x_3(t),\\
\dot{x}_3(t) &= \frac{-1}{2}\dot{x}_3(t-h)+\left(\frac{5}{24}\lambda_h^2-\frac{31}{12}\lambda_h-6\right)x_3(t)+([0,\,1]x_2(t)-y(t)).
\end{aligned}
\tag{31}
$$

The matrix of the closed-loop system (22) has the form

$$
\overline{W}(p,\lambda)=\begin{bmatrix}
\frac{p\,(2+\lambda)}{2} & 3-\lambda & 0 & 0 & 0 & 0\\
\frac{1}{3} & \frac{5\lambda}{12}+p & (2-\lambda)\,\lambda & 0 & 0 & 0\\
0 & 0 & \nu_3(p,\lambda) & \nu_1(p,\lambda) & \nu_2(p,\lambda) & 0\\
0 & 0 & 0 & \frac{p\,(2+\lambda)}{2} & 3-\lambda & \frac{79}{4}\lambda+\frac{31}{24}\lambda^2+\frac{5}{24}\lambda^3+36\\
0 & 0 & (2-\lambda)\,\lambda & \frac{1}{3} & \frac{5\lambda}{12}+p & -\frac{25}{288}\lambda^3+\frac{155}{144}\lambda^2-\frac{8}{3}\lambda-12\\
0 & -1 & 0 & 0 & 1 & p+\frac{1}{2}p\lambda-\frac{5}{24}\lambda^2+\frac{31}{12}\lambda+6
\end{bmatrix}.
$$

Straightforward calculations yield

$$
|\overline{W}(p,\lambda)|=\left(1-\frac{1}{2}\lambda\right)\left(1-\frac{1}{3}\lambda\right)\left(1+\frac{1}{2}\lambda\right)^3(p+3)^2(p+2)^2(p+1)^2.
$$

6. CONCLUSIONS

This paper has been devoted to linear autonomous differential-difference systems of neutral type with a scalar control input and an observable output. For such systems, we have derived modal controllability and exponential stabilizability criteria in the class of output-feedback controllers (function of the observed output). Modal controllability provides wider system design capabilities compared to stabilizability. In particular, it is possible to specify the rate of convergence to zero (vanishing) for the system solution by tuning the coefficients of the characteristic quasipolynomial.

Alternatively, it is possible to ensure a finite spectrum, making the system simpler from a (subsequent) control standpoint. However, the requirements for the system parameters imposed by the modal controllability criterion are more stringent than the conditions of exponential stabilizability.

Two types of asymptotic observers have been developed, namely, an observer with a given characteristic quasipolynomial and an exponentially stable observer. The behavior of the estimation errors of the observers is described by a linear homogeneous autonomous system of neutral type. Moreover, for the first type of observer, it is possible to specify a desired characteristic quasipolynomial of the system describing the error before its design (i.e., to set in advance the rate of convergence of the observer's estimate to the solution of the original system). In the case of the second type of observer, the system describing the estimation error of the solution is exponentially stable. Note that it is generally impossible to control the coefficients of the characteristic equation. However, the exponential stability of the system describing the estimation error ensures the convergence of the estimate to the solution at an exponential rate.

The methods for constructing controllers and observers developed in this study involve standard operations with polynomials and polynomial matrices and are easily implemented in modern computer algebra packages.

FUNDING

This work was supported by the grant of the President of the Republic of Belarus (order no. 15rp dated January 27, 2025).

APPENDIX

Proof of Theorem 1.

We introduce the neutral system

$$\dot{x}(t) = A'(p_D, \lambda_h)x(t) + c'(\lambda_h)u(t), \quad t > 0, \tag{A.1}$$

and define the controller

$$u(t) = \beta_1(p_D)x_1(t), \quad \dot{x}_1(t) = \mathcal{K}_1[x(t)] + \mathcal{K}_2[x_1(t)], \quad t > 0, \tag{A.2}$$

where $x_1 \in \mathbb{R}$ is an auxiliary variable, $\mathcal{K}_1 \in \mathfrak{Q}^{1\times n}$, $\mathcal{K}_2 \in \mathfrak{Q}$, and $\beta_1(p) \in \mathbb{R}_0[p]$.

For any given quasipolynomial (13) there exists a controller (A.2) such that $|W_1(p,\lambda)| = g(p,\lambda)$ if and only if conditions (19) are valid [15]. Here, $W_1(p, e^{-ph})$ is the characteristic matrix of the closed-loop system (A.1), (A.2):

$$W_1(p,\lambda) = \begin{bmatrix} pI_n - A'(p,\lambda) & -\beta_1(p)c'(\lambda) \\ -K_1(p,\lambda) & p - K_2(p,\lambda) \end{bmatrix},$$

with $K_i(p,\lambda) = \mathcal{K}_i[e^{pt}]e^{-pt}\Big|_{e^{-ph}}$. Letting

$$\beta_0(p) = \beta_1(p), \quad \mathcal{L}_1 = \mathcal{K}_1', \quad \mathcal{L}_2 = \mathcal{K}_2 \tag{A.3}$$

in equations (9) gives $\big(W_1(p,\lambda)\big)' = W_z(p,\lambda)$. Therefore, $|W_z(p,\lambda)| = g(p,\lambda)$, which implies the existence of an observer (9) with the desired characteristic quasipolynomial.

Proof of Theorem 2. For system (A.1) there exists a controller (A.2) making the closed-loop system exponentially stable if and only if conditions (20) are valid [15]. Therefore, repeating the proof of Theorem 1 with the necessary modifications, we show the existence of an exponentially stable observer (9). The proof of this theorem is complete.

Proof of Theorem 3.

Necessity. Conditions (25) [15] are necessary and sufficient for modal controllability in the class of feedback controllers based on measurements of the state vector x. Therefore, conditions (25) are necessary for modal controllability in the class of feedback controllers based on measurements of the observed output (2).

Let us prove the necessity of conditions (19). Suppose that system (1), (2) is modally controllable in the class of controllers (21). By assumption, the controller (21) ensures equality (24) for some given polynomial (23). Consider system (A.1) and define a controller of the form

$$\begin{gathered}
u(t) = -\alpha_1(p_D)x_1(t),\\
\dot{x}_1(t) = \mathcal{Q}'_{33}[x_1(t)] + \mathcal{Q}'_{23}[x_2(t)],\\
\dot{x}_2(t) = \alpha_1(p_D)c'(\lambda_h)x_1(t) + A'(p_D, \lambda_h)x_2(t) + \mathcal{Q}'_{12}[x_3(t)],\\
\dot{x}_3(t) = \alpha_0(p_D)b'(\lambda_h)(x(t) + x_2(t)) + \mathcal{Q}'_{11}[x_3(t)], \quad t > 0.
\end{gathered} \tag{A.4}$$

Let $\widehat{W}(p, e^{-ph})$ be the characteristic matrix of system (A.1),(A.4). Obviously, the matrix $\left(\overline{W}(p,\lambda)\right)'$ is obtained from the matrix $\widehat{W}(p,\lambda)$ by permuting the rows and columns of blocks with numbers 2 and 4. Therefore, we write $E_{24}\widehat{W}(p,\lambda)E_{24}^{-1} = \left(\overline{W}(p,\lambda)\right)'$, where the matrix E_{24} swaps the rows of suitable-size blocks with numbers 2 and 4 when multiplied by any matrix on the left. Hence, $|\widehat{W}(p,\lambda)| = \chi(p,\lambda)$, meaning that system (A.1) is modally controllable in the sense of [15] (i.e., by a feedback controller based on the state function x), and conditions (19) express a modal controllability criterion for system (A.1). The necessity of conditions (19) and (25) is established.

Sufficiency. Consider a given polynomial (23). We prove the sufficiency part by providing a design scheme for a controller (21) ensuring equality (24).

1. We define a state-feedback controller of the form

$$u(t) = \alpha_0(p_D)x_1(t), \quad \dot{x}_1(t) = \mathcal{Q}_{12}[x(t)] + \mathcal{Q}_{11}[x_1(t)], \quad t > 0. \tag{A.5}$$

The notation in (A.5) is the same as in (21). Due to conditions (25), for any polynomial $\chi_1(p,\lambda)$ (23) there exists [9] a controller (A.5) such that the characteristic matrix $W_x(p, e^{-ph})$ of the closed-loop system (1), (A.5),

$$W_x(p,\lambda) = \begin{bmatrix} pI_n - A(p,\lambda) & -b(\lambda)\alpha_0(p) \\ -Q_{12}(p,\lambda) & p - Q_{11}(p,\lambda) \end{bmatrix}, \tag{A.6}$$

satisfies the equality

$$|W_x(p,\lambda)| = \chi_1(p,\lambda). \tag{A.7}$$

Thus, the controller (A.5) has been constructed.

2. Under condition (19) (see Theorem 1), for any given polynomial $\chi_2(p,\lambda)$ (23) there exists an observer (9) with a given characteristic quasipolynomial such that the characteristic matrix (12) satisfies the relation

$$|W_z(p,\lambda)| = \chi_2(p,\lambda). \tag{A.8}$$

Thus, the observer (9) has been constructed.

3. Using the parameters of the controller (A.5) and observer (9), we write the controller (21) with the additional assignment

$$\mathcal{Q}_{23} = \mathcal{L}_1, \quad \mathcal{Q}_{33} = \mathcal{L}_2. \tag{A.9}$$

Let us show equality (24) for the characteristic matrix $\overline{W}(p, e^{-ph})$ of the closed-loop system (1), (2), (21). For this purpose, we introduce the matrix

$$\Gamma = \begin{bmatrix} I_n & 0_{n\times 1} & 0_{n\times n} & 0_{n\times 1} \\ 0_{1\times n} & 1 & 0_{1\times n} & 0 \\ -I_n & 0_{n\times 1} & I_n & 0_{n\times 1} \\ 0_{1\times n} & 0 & 0_{1\times n} & 1 \end{bmatrix}. \tag{A.10}$$

Direct verification yields

$$\Gamma \overline{W}(p, \lambda) \Gamma^{-1} = \widetilde{W}(p, \lambda), \tag{A.11}$$

$$\widetilde{W}(p, \lambda) = \begin{bmatrix} pI_n - A(p, \lambda) & -b(\lambda)\alpha_0(p) & 0_{n\times n} & 0_{n\times 1} \\ -\mathcal{Q}_{12}(p, \lambda) & p - \mathcal{Q}_{11}(p, \lambda) & -\mathcal{Q}_{12}(p, \lambda) & 0_{1\times 1} \\ 0_{n\times 1} & 0 & pI_n - A(p, \lambda) & -\mathcal{Q}_{23}(p, \lambda) \\ 0_{n\times 1} & 0 & -\alpha_1(p)c'(\lambda) & p - \mathcal{Q}_{33}(p, \lambda) \end{bmatrix}.$$

From equalities (A.6), (A.7), (A.8), (A.11), and (12) it follows that $\overline{W}(p, \lambda) = \Gamma \overline{W}(p, \lambda) \Gamma^{-1} = \chi(p, \lambda)$. The proof of Theorem 3 is complete.

Proof of Theorem 4. The idea of proving Theorem 4 is quite similar to that of proving Theorem 3, so we will present only its brief scheme.

Necessity. 1. Suppose that system (1), (2) closed by the controller (21) is exponentially stable. We form the sets Δ_0 and Δ_1 from Remark 2 (see (16)). If the first condition in (26) is violated, then for any $\varepsilon_0 < 0$ there exists $p_0 \in \mathbb{C}$, $\mathbf{Re}\, p_0 \geqslant \varepsilon_0$, such that $\operatorname{rank}\big[W(p_0, e^{-p_0 h}), b(e^{-p_o h})\big] < n$. In this case, for any controller of the form (21), the number p_0 remains in the spectrum of the closed-loop system (1), (2), (21), i.e., $p_0 \in \Delta_0$. Therefore, condition (17) fails and, consequently, system (1), (2), (21) cannot be exponentially stable.

If the second condition in (26) is violated, then there exists $\lambda_0 \in \mathbb{C}$, $|\lambda_0| \leqslant 1$, such that $\operatorname{rank}\big[D(\lambda_0)\big] < n$. Obviously, for any controller of the form (21), the closed-loop system (1), (2), (21) satisfies $\lambda_0 \in \Delta_1$. Hence, condition (18) is violated. The necessity of conditions (26) is established.

2. Now we prove the necessity of conditions (20). Consider system (A.1) closed by the controller (A.4). Assuming sequentially that the first or second conditions in (20) are violated, similar to (1), we show that the closed-loop system (A.1), (A.4) cannot be exponentially stable.

Sufficiency. We describe a design scheme for the controller (21) and then prove the exponential stability of the closed-loop system.

1. Following [9], we construct the controller (A.5) exponentially stabilizing the closed-loop system (1), (A.5). Conditions (26) ensure the possibility of constructing such a controller. In this case, the characteristic matrix of the closed-loop system (1), (A.5) has the form (A.6).

2. We construct the exponentially stable observer (9). Conditions (20) ensure the possibility of constructing such an observer. In this case, the characteristic matrix of the homogeneous system (9) has the form (12).

3. Using the parameters of the controller (A.5) and observer (9), we write the controller (21) with the matrices assigned by (A.9).

Let us show the exponential stability of system (1), (21). For this purpose, we apply the following nondegenerate transformation of the variables:

$$\mathrm{col}[x, x_1, x_2, x_3] = \Gamma^{-1}\mathrm{col}[\tilde{x}, \tilde{x}_1, \tilde{x}_2, \tilde{x}_3],$$

with the matrix Γ given by (A.10). This transformation yields a new system with the characteristic matrix $\widetilde{W}(p, e^{-ph})$, where the matrix $\widetilde{W}(p, \lambda)$ has the form (A.11). The resulting system will be called the system $\widetilde{\Sigma}$.

Due to the representation of the matrix $\widetilde{W}(p, \lambda)$, the components $\tilde{x}_2$ and $\tilde{x}_3$ are determined by a separate system (a subsystem of the system $\widetilde{\Sigma}$) whose characteristic matrix coincides with (12). Therefore, the system determining the components $\tilde{x}_2$ and $\tilde{x}_3$ is exponentially stable. In other words, there exist positive constants γ_1 and γ_2 such that

$$\|\tilde{x}_i(t)\| \leqslant \gamma_1 e^{-\gamma_2 t}, \quad t > 0, \quad i = 2, 3. \tag{A.12}$$

Consider the system corresponding to the first two rows of the blocks of the matrix $\widetilde{W}(p, \lambda)$. Since the components $\tilde{x}_2$ and $\tilde{x}_3$ are determined separately, they can be treated as an inhomogeneous part in the system under consideration. Then the components $\tilde{x}$ and $\tilde{x}_1$ satisfy the inhomogeneous system for which the characteristic matrix of the corresponding homogeneous system coincides with (A.6). Hence, the above homogeneous system is exponentially stable and, in view of (A.12), there exist positive constants γ_3 and γ_4 such that

$$\|\tilde{x}(t)\| \leqslant \gamma_3 e^{-\gamma_4 t}, \quad \|\tilde{x}_i(t)\| \leqslant \gamma_3 e^{-\gamma_4 t}, \quad t > 0, \quad i = \overline{1, 3}. \tag{A.13}$$

These inequalities imply the exponential stability of the system $\widetilde{\Sigma}$ and, consequently, of system (1), (21). The proof of Theorem 4 is complete.

REFERENCES

1. Dolgii, Yu.F. and Surkov, P.G., *Matematicheskie modeli dinamicheskikh sistem s zapazdyvaniem* (Mathematical Models of Dynamic Systems with Delay), Yekaterinburg: Ural University, 2012. https://elar.urfu.ru/bitstream/10995/45629/1/978-5-7996-0772-2_2012.pdf (Accessed March 17, 2025.)

2. Kolmanovskiy, V.B. and Nosov, V.R., Systems with an After-effect of the Neutral Type, *Autom. Remote Control*, 1984, vol. 45, no. 1, pp. 1–28.

3. Poloskov, I.E., *Metody analiza sistem s zapazdyvaniem* (Analysis Methods for Systems with Delay), Perm: Perm State National Research University, 2020. http://www.psu.ru/files/docs/science/books/mono/poloskov-metody-analiza-sistem.pdf

4. Khartovskii, V.E., *Upravlenie lineinymi sistemami neitral'nogo tipa: kachestvennyi analiz i realizatsiya obratnykh svyazei* (Control of Linear Systems of Neutral Type: Qualitative Analysis and Feedback Implementation), Grodno: Grodno State University, 2022.

5. Grebenshchikov, B.G., Asymptotic Properties and Stabilization of a Neutral Type System with Constant Delay, *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2021, vol. 17, no. 1, pp. 81–96. https://doi.org/10.21638/11701/spbu10.2021.108

6. Bulgakov, B.V., *Kolebaniya* (Oscillations), Moscow: Izd-vo Tekhniko-teoreticheskoi Lit-ry, 1954.

7. Krasovskii, N.N. and Osipov, Yu.S., On the Stabilization of Motions of a Plant with Delay in the Control System, *Izv. Akad. Nauk SSSR. Tekh. Kibern.*, 1963, no. 6, pp. 3–15.

8. Osipov, Yu.S., On the Stabilization of Controlled Systems with Delay, *Differ. Uravn.*, 1965, vol. 1, no. 5, pp. 606–618.

9. Pandolfi, L., Stabilization of Neutral Functional-Differential Equations, *J. Optim. Theory Appl.*, 1976, vol. 20, no. 2, pp. 191–204. https://doi.org/10.1007/BF01767451

10. Lu, W.S., Lee, E., and Zak, S., On the Stabilization of Linear Neutral Delay-Difference Systems, *IEEE Transact. Autom. Control*, 1986, vol. 31, no. 1, pp. 65–67. https://doi.org/10.1109/TAC.1986.1104115

11. Rabah, R., Sklyar, G.M., and Barkhayev, P.Y., Stability and Stabilizability of Mixed Retarded-Neutral Type Systems, *ESAIM Control, Optimization and Calculus of Variations*, 2012, vol. 18, no. 3, pp. 656–692. https://doi.org/10.1051/cocv/2011166

12. Dolgii, Yu.F. and Sesekin, A.N., Regularization Analysis of a Degenerate Problem of Impulsive Stabilization for a System with Time Delay, *Tr. Inst. Mat. Mekh. UrO RAN*, 2022, vol. 28, no. 1, pp. 74–95. https://doi.org/10.21538/0134-4889-2022-28-1-74-95

13. Hu, G.D. and Hu, R., A Frequency-Domain Method for Stabilization of Linear Neutral Delay Systems, *Syst. Control Lett.*, 2023, vol. 181, art. no. 105650. https://doi.org/10.1016/j.sysconle.2023.105650

14. Hale, J.K. and Verduyn Lunel, S.M., Strong Stabilization of Neutral Functional Differential Equations, *IMA J. Math. Control Inf.*, 2002, vol. 19, no. 1–2, pp. 5–23. https://doi.org/10.1093/imamci/19.1_and_2.5

15. Metel'skii, A.V., Spectrum Assignment for a System of Neutral Type, *Diff. Equat.*, 2024, vol. 60, no. 1, pp. 101–126. https://doi.org/10.1134/S0012266124010099

16. Minyaev, S.I. and Fursov, A.S., Topological Approach to the Simultaneous Stabilization of Plants with Delay, *Diff. Equat.*, 2013, vol. 49, no. 11, pp. 1423–1431. https://doi.org/10.1134/S0012266113110098

17. Watanabe, K., Finite Spectrum Assignment and Observer for Multivariable Systems with Commensurate Delays, *IEEE Trans. Autom. Control*, 1986, vol. AC–31, no. 6, pp. 543–550. https://doi.org/10.1109/TAC.1986.1104336

18. Wang, Q.G., Lee, T.H., and Tan, K.K., *Finite Spectrum Assignment Controllers for Time Delay Systems*, Springer-Verlag, 1999. https://doi.org/10.1007/978-1-84628-531-8

19. Metel'skii, A.V., Spectral Reduction, Complete Damping, and Stabilization of a Delay System by a Single Controller, *Diff. Equat.*, 2013, vol. 49, no. 11, pp. 1405–1422. https://doi.org/10.1134/S0012266113110086

20. Marchenko, V.M., Control of Systems with Aftereffect in Scales of Linear Controllers with Respect to the Type of Feedback, *Diff. Equat.*, 2011, vol. 47, no. 7, pp. 1014–1028. https://doi.org/10.1134/S0012266111070111

21. Metel'skii, A.V. and Khartovskii, V.E., Criteria for Modal Controllability of Linear Systems of Neutral Type, *Diff. Equat.*, 2016, vol. 52, no. 11, pp. 1453–1468. https://doi.org/10.1134/S0012266116110070

22. Khartovskii, V.E., Modal Controllability for Systems of Neutral Type in Classes of Differential-Difference Controllers, *Autom. Remote Control*, 2017, vol. 78, no. 11, pp. 1941–1954. https://doi.org/10.1134/S0005117917110017

23. Fridman, E., *Introduction to Time-Delay Systems: Analysis and Control*, Birkhäuser, 2014. https://doi.org/10.1007/978-3-319-09393-2

24. Furtat, I. and Fridman, E., Delayed Disturbance Attenuation via Measurement Noise Estimation, *IEEE Transaction on Automatic Control*, 2021, vol. 66, no. 11, pp. 5546–5553. https://doi.org/10.1109/TAC.2021.3054238

25. Karpuk, V.V. and Metel'skii, A.V., Complete Calming and Stabilization of Linear Autonomous Systems with Delay, *J. Comput. Syst. Sci. Int.*, 2009, vol. 48, no. 6, pp. 863–872. https://doi.org/10.1134/S1064230709060033

26. Metel'skii, A.V., Khartovskii, V.E., and Urban, O.I., Solution Damping Controllers for Linear Systems of the Neutral Type, *Diff. Equat.*, 2016, vol. 52, no. 3, pp. 386–399. https://doi.org/10.1134/S0012266116030125

27. Metel'skii, A.V., Complete and Finite-Time Stabilization of a Delay Differential System by Incomplete Output Feedback, *Diff. Equat.*, 2019, vol. 55, no. 12, pp. 1611–1629. https://doi.org/10.1134/S0012266119120085

28. Khartovskii, V.E., Finite Stabilization and Finite Spectrum Assignment by a Single Controller Based on Incomplete Measurements for Linear Systems of the Neutral Type, *Diff. Equat.*, 2024, vol. 60, no. 5, pp. 655–676. https://doi.org/10.1134/S0012266124050094

29. Khartovskii, V.I. and Urban, O.I., Incomplete Measurements-Based Finite Stabilization of Neutral Systems by Controllers with Lumped Commensurate Delays, *Autom. Remote Control*, 2025, vol. 86, no. 1, pp. 1–19. https://doi.org/10.31857/S000523102501

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 6, pp. 531–544.

Stabilization of Oscillations in an Autonomous Corrected Conservative System by Constructing an Attracting Cycle

V. N. Tkhai

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

Received November 27, 2024
Revised February 25, 2025
Accepted April 21, 2025

Abstract—This paper considers a conservative system admitting a family of single-frequency oscillations with a domain Ω. For the original system, an autonomous controlled (ε-corrected) system with a small gain is introduced; a given oscillation from the domain Ω is stabilized by constructing a cycle that attracts all trajectories from this domain together with its ε-neighborhood. A universal adaptive control law, acting as a nonlinear force linear in velocity, is designed to track the current value of potential energy during motion. The cycle is constructed for any system oscillation. As a result, a new class of autonomous controlled systems is obtained based on the conservative system, and the operating modes of this class are stabilized (in the large) cycles with any desired energy. Examples are provided.

Keywords: conservative system, single-frequency oscillation, universal adaptive control, feedback, potential energy tracking, attracting cycle, stabilization

DOI: 10.31857/S0005117925060037

1. INTRODUCTION

This paper is devoted to the single-frequency oscillations (periodic motions) of a conservative system. They form families by a parameter, i.e., the constant energy h. The families are divided into nondegenerate and degenerate. On a nondegenerate family, the period $T(h)$ varies monotonically with the constant h; an example is a family of pendulum oscillations. On a degenerate family, oscillations are isochronous.

Nondegenerate oscillations can always be continued [1, 2] to a global family of oscillations. In a conservative system, a global family is described by a reduced conservative system with one degree of freedom. The same result holds for a family of degenerate oscillations of a conservative system, but this issue is not considered separately in the paper. In any case, the problem of investigating a conservative system with one degree of freedom arises first.

The idea of oscillation control by constructing a limit cycle goes back to L.S. Pontryagin [3], who found necessary and sufficient conditions for isolating a limit cycle from a family of periodic solutions of a Hamiltonian system by using non-Hamiltonian perturbations. B. van der Pol [4] investigated the relaxation modes of a regerative receiver (in the absence of a perturbing force). In 1929, A.A. Andronov discovered that the stable self-oscillations constructed by him and independently by van der Pol are a physical embodiment of Poincaré curves. In 1881, A. Poincaré [5, Chap. VI] introduced the concept of a limit cycle for a system in the plane. The theory of self-oscillations was developed in Andronov's school [6].

Poincaré singled out the isolated case when periodic solutions are preserved qualitatively with varying the parameter. In the non-isolated cases, a bifurcation occurs; here, note the main results by N.N. Bogolyubov [7], I.G. Malkin [8], and V.K. Melnikov [9]. They were continued in many directions, including, e.g., nonsmooth dynamics (see the review [10]).

In an autonomous ε-perturbed system, a bifurcation gives birth to a cycle. In this case, a Jordan cell of zero characteristic exponents (CEs) splits: one zero CE is preserved whereas the other CE becomes $-\varepsilon\alpha$. In the case $\alpha > 0$, the Andronov–Witt theorem [11] on the stability of a periodic motion of an autonomous system is valid. The concept of a cycle was also applied to multidimensional systems [12]; the Andronov–Witt theorem remains valid for them as well. The formula for computing the number α was given in [13].

In a linear periodic system, a CE is a root of the characteristic equation of the Lyapunov reduced system with constant coefficients.

In an autonomous system, the solution is determined within a shift of the initial point along the trajectory. Therefore, in the Andronov–Witt theorem, when passing to the neighborhood of a periodic motion for an n-order system, the problem of asymptotic stability is posed in the Rumyantsev sense [14] with respect to $n-1$ variables (the deviations from this motion). The solution is obtained for CEs with negative real parts; the single number $-\varepsilon\alpha$ is calculated for a system in the plane. In the case of the Andronov–Witt theorem, this property is called the orbital asymptotic stability of a periodic motion or attraction [15] with a specified attraction domain of trajectories: in the small (locally) or in the large (globally).

A cycle is an isolated periodic solution of an autonomous system [12]. An attracting cycle is an orbitally asymptotically stable single-frequency oscillation. In a controlled autonomous system, an oscillation is stabilized by constructing an attracting cycle.

According to [13], in the neighborhood of a cycle, the van der Pol dissipation is universal in the sense of independence from the considered system with oscillations. For a mathematical pendulum, the result established in [13] implies that any oscillation is stabilized (in the small) by using an adaptive stabilization scheme [2]. The scheme involves a control law with a parameter chosen depending on the parameter value for the oscillation to be stabilized: the control law has adaptivity. The scheme can be applied independently or as part of a more general adaptive control system.

A cycle and an attracting cycle are achieved using a feedback loop, in which a coordinate-tracking van der Pol-type controller receives trajectory information to form dissipation at the current trajectory point without delay. Thus, the stabilization problem is solved in the neighborhood of the oscillation under consideration. For global stabilization, the feedback loop is based on potential energy tracking and is described in this paper.

Other studies on the stabilization of a desired oscillation mode differ in the use of explicit time-varying control laws. Let us mention some of these studies. A review on the example of an inverted pendulum was provided in [16]. Swinging control was proposed in [17, 18]. The orbital stabilization problem of periodic solutions of low-drive nonlinear systems was solved in [19]; the nonlinear feedback control law designed therein is time-varying. Stabilization of a desired mechanical energy by impulsive control was described in [20]; a robust stabilizing control law for oscillations was found by the implicit Lyapunov method in [21]; electrodynamic control-based stabilization was carried out in [22].

In this paper, we construct a ε-corrected conservative system possessing an attracting cycle. Its attraction domain includes the oscillation domain Ω of the conservative system and the ε-neighborhood of Ω. In addition, we solve the global stabilization problem. Note that for a reduced conservative system with one degree of freedom, the local problem was solved in [2].

2. A CONSERVATIVE SYSTEM WITH ONE DEGREE OF FREEDOM. CONTROL DESIGN

Consider a smooth conservative system with one degree of freedom admitting a family Σ of single-frequency oscillations in a parameter h, where h is the constant energy value. According to [1], such a family can always be continued to a global family, so a global family Σ will be analyzed below. It occupies an oscillation domain Ω. The period on the family can be increasing (a mathematical pendulum), constant (a harmonic oscillator), or decreasing (the equation $\ddot{x} + x^3 = 0$). Under a ε-small force (a control law with a small gain ε), we obtain an autonomous corrected (controlled) conservative system of the form

$$\ddot{x} + f(x) = \varepsilon u(x, \dot{x}). \tag{1}$$

For $\varepsilon = 0$, equation (1) admits the energy integral

$$\dot{x}^2 = 2(h - \Pi(x)), \quad \Pi = \int f(x)dx. \tag{2}$$

In the domain Ω, $0 < \Pi(x) \leqslant h$. On the family Σ, the coordinate is described by the formula $x = \varphi(h, t)$ and the period $T = T(h)$ is a function of the constant energy h.

Let us choose a smooth function u without an explicit time dependence (autonomous control). In the stabilization problem, the control law must ensure local attraction to a cycle, so the function $u(x, \dot{x})$ will be of the form $u = a(x, \dot{x})\dot{x}$ [13]. The function $a = 1 - Kx^2$ itself was found in [13]; it ensures the existence of a cycle. By choosing a constant $K = K(h^*)$ in this function, one achieves the orbital asymptotic stability of the cycle born from the oscillation of a conservative system with the energy value h^* [13]. The local result becomes global for an isochronous family of oscillations. This fact is demonstrated in the van der Pol equation. However, for a nondegenerate family of oscillations, the local result does not extend to the entire domain Ω.

When solving the local problem [13], the idea is to consider the dependence of $K(h)$ on the energy value h for a nondegenerate family of oscillations. Thereby, one constructs a cycle in the corrected system close to an oscillation of a conservative system with the desired energy $h = h^*$. For this purpose, $K = K(h^*)$ is chosen.

In this paper, we apply a control law tracking in a feedback loop the current potential energy during motion. For the controlled conservative system

$$\ddot{x} + f(x) = \varepsilon[1 - K(h^*)\Pi(x)]\dot{x}, \tag{3}$$

we study the existence of a cycle close to the oscillation of the conservative system with an energy value $h = h^* > 0$. The control law is designed below as well.

Consider the amplitude (bifurcation) equation

$$I(h) \equiv \int\limits_0^{T^*} [1 - K(h^*)\Pi(x)]\dot{x}^2 dt = 0, \tag{4}$$

in which the potential $\Pi(x)$ and kinetic $\tilde{T} = \dot{x}^2/2$ energies are calculated on the solution $x = \varphi(h, t)$. Equality (4) expresses necessary and sufficient conditions for the existence of a T^*-periodic solution in the first ε-approximation. Equality (4) takes into account the conjugate solution $\psi = -\dot{\varphi}$ of the conservative system. As it turns out, the condition $dI(h^*)/dh \neq 0$ is sufficient for the existence of a periodic solution of the perturbed equation (3); for details, see [7–9]. For the autonomous system (3), this general result means the birth of a cycle. When the derivative is negative, the cycle is stable: the formula for the CE was given in [13].

The identity

$$\int_0^T [1 - K(h)\Pi(\varphi(h,t))]\dot{\varphi}(h,t)^2 dt \equiv 0 \tag{5}$$

expresses existence conditions for a periodic solution in the first ε-approximation for all values of h. It yields the function

$$K(h) = \frac{\int\limits_0^T (h - \Pi(\varphi(h,t)))dt}{\int\limits_0^T \Pi(\varphi(h,t))(h - \Pi(\varphi(h,t)))dt}. \tag{6}$$

In (6), the denominator is nonzero in the oscillation domain Ω. The function $K(h)$ is defined uniquely through the potential energy Π, which varies with time t along the trajectory. Only one trajectory and one value of $K(h)$ corresponds to each h.

On the other hand, formula (5) with the upper limit of integration $T = T^*$ and the number $K(h) = K(h^*)$ leads to the amplitude equation (4) for finding the value of h corresponding to the cycle. The explicit form of the function $K(h)$ will be presented in Section 3.

Equation (3) defines a mapping of the phase plane onto itself: $t : 0 \to T$. In this case, the curve $\Gamma(h,0) = \{x(h,0), \dot{x}(0)\}$ is mapped into the curve $\Gamma(h,T) = \{x(h,T), \dot{x}(T)\}$. The existence of a unique root $h = h^*$ of equation (4) means the coincidence of the points $\Gamma(h^*,0)$ and $\Gamma(h^*,T)$. This fact is observed for the curves constructed using even the first ε-approximation: the necessary and sufficient conditions in the first approximation become sufficient for the existence of a fixed point of the mapping in the first ε-approximation. For the unique root, the fixed point of the mapping will be isolated, and it corresponds to an isolated periodic solution of period T. In view of $T(h^*) = T^*$, we obtain a cycle of period T^*. The condition $dI(h^*)/dh < 0$ is valid for a contracting mapping: the cycle becomes orbitally asymptotically stable.

3. ATTRACTION TO THE CYCLE IN THE SMALL

The integral (4) is taken on the interval $t \in [0, T^*]$. Here, the function $\Pi(x)$ depends only on x and is calculated via the solution $x = \varphi(h,t)$. The change of variable $\tau = (T(h)/T^*)t$ in the integral (4) makes the coordinate x directly dependent only on τ; the function $x(h,\tau)$ becomes T^*-periodic for all values of the parameter h. For an oscillation with zero initial velocity, we obtain

$$x(h,\tau) = x(h,0)e(\tau), \quad e(\tau) = e(\tau + T^*). \tag{7}$$

The function $e(\tau)$ varies on the interval $[0,1]$. The same is true for the function $\Pi(x(h,\tau))$. For $\tau = 0$, we have $\Pi(x(h,0)) = h$. Therefore, due to the expression (7), the equality $\Pi(x(h,\tau)) = hz(\tau)$ holds, where a particular function $z(\tau)$ is calculated for a given potential energy. See the Appendix for a more general analysis of the application of the variable τ; Theorem 4 provided therein essentially supplements the main results, being of independent interest.

Considering the new expression for the potential energy Π, we transform the amplitude equation (4) as follows:

$$I(h) = 0, \quad I(h) \equiv h\int_0^{T^*} (1 - K(h^*)hz)(1 - z)d\tau. \tag{8}$$

The function $I(h)$ can be represented as

$$I(h) = h(\alpha - \beta h), \quad \alpha = \int_0^{T^*} (1 - z)d\tau, \quad \beta = -K(h^*)\int_0^{T^*} z(1 - z)d\tau. \tag{9}$$

Clearly, equation (8) admits the unique nonzero root $h^* = \alpha/\beta$.

Identity (5) takes the form

$$h\int\limits_0^{T^*}(1-K(h)hz)(1-z)d\tau\equiv 0.$$

Therefore, at the point $h=h^*$,

$$\frac{d}{dh}\left(h\int\limits_0^{T^*}(1-K(h^*)hz)(1-z)d\tau\right)_{h=h^*}-\frac{dK(h^*)}{dh}(h^*)^2\int\limits_0^{T^*}z(1-z)d\tau\equiv 0.$$

The derivative at the point corresponding to the cycle is calculated by the formula

$$\frac{dI(h^*)}{dh}=\frac{dK(h^*)}{dh}(h^*)^2\int\limits_0^{T^*}z(1-z)d\tau. \tag{10}$$

The function (6) is given by

$$K(h)=\frac{\int_0^{T^*}(1-z(\tau))d\tau}{h\int_0^{T^*}z(\tau)(1-z(\tau))d\tau}=\frac{b}{h},\quad b=\mathrm{const}>0. \tag{11}$$

The integrals in (11) are independent of the energy constant h and take positive values as sums of the integrals on quarters of the period. Therefore, $K(h)=b/h$, where $b>0$ is a constant.

Formula (11) is valid for any family of oscillations: that with an increasing (decreasing) period on the family or an isochronous family. The dependence (11) is an important characteristic of a family of oscillations.

We have the following result regarding the local stabilizability of a cycle.

Theorem 1. *The ε-corrected conservative system* (3) *always has an orbitally asymptotically stable (in the small) cycle close to the oscillation of the conservative system with the energy value $h=h^*$. This cycle attracts trajectories from its $O(\varepsilon)$-neighborhood.*

Proof. According to the amplitude equation (8), the corrected conservative system (3) has a cycle close to the oscillation of the conservative system with the energy value $h^*=\alpha/\beta$. At this point, the sign of the derivative (10) coincides with that of the number

$$\frac{dK(h^*)}{dh}=-\frac{b}{(h^*)^2}<0.$$

Therefore, by formula (10), the derivative is $dI(h^*)/dh<0$, and the mapping $t:0\to T^*$ is contracting: all trajectories from the neighborhood of the cycle are attracted to the cycle.

Remark 1. The derivative $dK(h^*)/dh$ is commonly used for proving local results on a cycle. However, the inequality $dI(h^*)/dh=-h^*\beta<0$ can be derived directly from the expression (9).

4. ATTRACTION TO THE CYCLE IN THE LARGE

For system (1) we define the total mechanical energy

$$E\equiv\frac{\dot{x}^2}{2}+\Pi(x). \tag{12}$$

This function takes a constant value h on the solutions of the conservative system. For the corrected conservative system,

$$\frac{dE}{dt} = \varepsilon[1 - K(h^*)\Pi(x)]\dot{x}^2, \tag{13}$$

where

$$\dot{x}^2 = 2(E - \Pi(x)).$$

For $\varepsilon = 0$ we have $E = h(\text{const}) > 0$.

The increment ΔE of the energy (12) on the period $T(h)$ is calculated for an oscillation with the energy value h. Therefore, for $\varepsilon > 0$,

$$\Pi(x) = \Pi(\varphi) + h\varepsilon\rho(\varepsilon, x)), \quad E - \Pi(x) = h - \Pi(\varphi) + h\varepsilon\sigma(\varepsilon, x), \tag{14}$$

where functions ρ and σ are of order one. Integration is performed over the variable τ on the interval $[0, T^*]$. As a result,

$$\Delta E(\varepsilon, h) = \varepsilon\frac{2T(h)}{T^*}[I(h) + \varepsilon hF(\varepsilon, h)], \tag{15}$$

where $\varepsilon hF(\varepsilon, h)$ is calculated by substituting the expressions (14) into equality (13) and matches the terms $h\varepsilon\rho$ and $h\varepsilon\sigma$. This formula remains valid in the entire oscillation domain Ω of the conservative system.

The function $E(\varepsilon, h^*, \tau)$ calculated on the cycle is T^*-periodic, so $\Delta E(\varepsilon, h^*) = 0$. For the cycle, we have the amplitude equation $I(h^*) = 0$. As a consequence, the equality $F(\varepsilon, h^*) = 0$ is true for the cycle.

Lemma 1. *There exists a h^*-independent number $\varepsilon_0 > 0$ such that, for $0 < \varepsilon < \varepsilon_0$, the equation*

$$I(h) + \varepsilon hF(\varepsilon, h) = 0 \tag{16}$$

has a unique root.

Proof. With the representation (9), equation (16) is simplified:

$$V \equiv G(h) + \varepsilon F(\varepsilon, h) = 0, \quad G \equiv \alpha - \beta h.$$

The equation $G(h) = 0$ has the root $h^* = \alpha/\beta$ corresponding to a cycle. According to Theorem 1, the cycle is locally attracting. Hence, for $h \neq h^*$, the function $G(h)$ takes values of the same sign under small ε. The function $F(\varepsilon, h)$ vanishes at the point $h = h^*$. Therefore, the function V can be transformed to

$$V = (\alpha - \beta h)(1 + \varepsilon W(\varepsilon, h)). \tag{17}$$

Under small ε, the sign of V coincides with that of G. However, as ε increases, the second factor in (17) may vanish. The corresponding value of $\varepsilon(h)$ depends on h. For the cycle with $h = h^*$, we choose the smallest value of the number $\varepsilon(h(h^*))$ and denote by ε_0 the lower bound of the set $\{\varepsilon(h^*)\}$. Then for $0 < \varepsilon < \varepsilon_0$, equation (16) has a unique root independent of the particular value of h^*. In this case, according to (9) and (17), the existence of a root in equation (16) is determined by the term $G(h)$, which has a unique root.

The proof of Lemma 1 is complete.

Next, the corrected system (3) is investigated under $0 < \varepsilon < \varepsilon_0$. The oscillation domain Ω of the conservative system is considered; in the conservative system under analysis, there may be more than one oscillation domain.

Theorem 2. *There exists a number $\varepsilon_0 > 0$ such that the corrected conservative system* (3) *with $0 < \varepsilon < \varepsilon_0$ always has a unique cycle attracting all trajectories of the oscillation domain Ω of the conservative system.*

Proof. In the oscillation domain Ω, the corrected conservative system (3) admits (Lemma 1) a unique cycle: the corresponding value of the conservative system energy is $h = h^*$. On the cycle, $\Delta E(\varepsilon, h^*) = 0$. Beyond the cycle, the sign of the function $\Delta E(\varepsilon, h)$ coincides with that of the linear function $G(h)$, $G(h^*) = 0$. The rate of change of the function is $dG(h)/dh = -\beta < 0$. Hence, the function $G(h) \to 0$, and the trajectories of the corrected system (3) tend to the cycle. This happens with any trajectory from the oscillation domain Ω of the conservative system.

The proof of Theorem 2 is complete.

5. THE MULTIDIMENSIONAL SYSTEM

By the global family theorem [1, Theorem 1], the variables are separated in a multidimensional system. The variable x describes an oscillation family Σ on the manifold Ω invariant with respect to the phase flow of the conservative system. Outside the manifold Ω, the dynamics of the conservative system are given by the vector y of dimension $n-1$. In the domain Ω, $y \equiv 0$. Therefore, outside Ω, the motion of the conservative system in the neighborhood of the trivial solution $y = 0$ is studied.

According to Poincaré, the characteristic equation of a conservative system is reciprocal: the roots of the equation are divided into pairs containing numbers with opposite signs. Therefore, in the real-root case, under the action of ε-small control, the outgoing solutions will remain outgoing. Hence, the absence of roots with real parts is a necessary condition for attracting solutions to Ω.

Together with the separation of the integral manifold Ω, the ε-corrected system (3) with the variable x is constructed. The principal coordinates are applied for the variable y in the neighborhood of the point $y = 0$. Next, we consider a controlled conservative system of the form

$$\begin{gathered}\ddot{y}_i + k_i y_i + Y_i(y) = \varepsilon[(1 - K(h^*)\Pi(x)]\dot{y}_i,\\ k_i = \text{const}, k_i \geqslant 0, \quad i = 1, \ldots, n-1.\end{gathered} \tag{18}$$

According to (18), the energy E_y of the conservative system in the variable y varies by the law

$$\frac{dE_y}{dt} = \varepsilon[1 - K(h^*)\Pi(x)]\dot{y}^2, \quad \dot{y}^2 = \sum_{i=1}^{n-1} \dot{y}_i^2, \tag{19}$$

similar to the law (13) for the variable x. In addition, $E_y = h_y = \text{const}$ for $\varepsilon = 0$.

The energy variation law of the entire conservative system

$$\frac{d(E_x + E_y)}{dt} = \varepsilon[1 - K(h^*)\Pi(x)](\dot{x}^2 + \dot{y}^2),$$

written in the variables x and y, is associated with the amplitude equation

$$\tilde{I}(\tilde{h}) \equiv -\int_0^{T^*} [1 - K(h^*)\Pi(x)](\dot{x}^2 + \dot{y}^2)dt = 0, \quad \tilde{h} = h + h_y. \tag{20}$$

The function $\tilde{I}(\tilde{h})$ admits a simple root $\tilde{h} = h^*$, $h_y^* = 0$, which follows from the existence of a simple root in equation (4). The simple root of equation (20) is associated with a cycle of the controlled system (21) (see [8, Chap. VI, §8, p. 413, §9, p. 417]). The function $\tilde{I}(\tilde{h})$ is continuous, and the cycle of the entire system coincides with the cycle on Ω.

The laws (13) and (19) imply the equality

$$\dot{y}^2 dE = \dot{x}^2 dE_y;$$

hence, the energy in the system with the variable y changes on the period (increases and decreases) in the same way as in that with the variable x. In addition, by Theorem 2, the one-period increment $\Delta E(\varepsilon, h)$ in the variable x tends to zero whereas the trajectory on the manifold Ω to a unique cycle. Then the one-period increment $\Delta E_y(\varepsilon, h)$ in the variable y tends to zero whereas the trajectory to the unique equilibrium $y = 0$ corresponding to the cycle on Ω.

Thus, all trajectories of the domain Ω and its ε-neighborhood are attracted to the cycle. This result is true regardless of the coordinates used to define the conservative system.

Next, consider the controlled system

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_s} - \frac{\partial L}{\partial q_s} = \varepsilon[1 - K(h^*)\Pi(x)]\frac{\partial \tilde{T}}{\partial \dot{q}_s}, \quad s = 1, \ldots, n, \tag{21}$$

defined by the Lagrange equations of the second kind. Here, $\tilde{T}$ and Π are the kinetic and potential energies, respectively. By assumption, for $\varepsilon = 0$ system (21) admits a family Σ of single-frequency oscillations occupying the two-dimensional domain Ω. On Ω, the system is described by the variable x; control is intended to track the potential energy on the solution. Let us choose $0 < \varepsilon < \varepsilon_0$, where ε_0 is a finite number for the oscillation family Ω; h^* is the energy value corresponding to a cycle in Ω; on Ω, the energy value $h = 0$ correspond to the equilibrium.

In a system described by Lagrange equations of the second kind, the global family of periodic motions is constructed by a continuation of its local Lyapunov family [1]. In turn, the latter is born from an equilibrium. According to the Lyapunov center theorem [23], a family of nondegenerate local nonlinear periodic solutions exists in system (18) if, in addition to pure imaginary roots, this system has non-resonant frequencies $\sqrt{k_j} \neq p\sqrt{k_s}$, $p \in \mathrm{N}$. Moreover, by the Lyapunov center theorem [23], the Lyapunov family always exists for the largest frequency. In the analysis of system (18), the above conditions have not been imposed. They arise in the Lagrangian system. Keeping this aspect in mind, we proceed to the controlled system (21).

When the principal coordinates for the system in y are not separated, the right-hand sides of system (18) will contain linear combinations of the velocities $\dot{y}_i$; they are partial derivatives of the kinetic energy $\tilde{T}$ of the system in $\dot{y}_s$. The reconstruction of system (21) is completed by returning to the initial stage of building the reduced system with one degree of freedom [1]. At this stage, the velocity $\dot{y}_i$ with a constant factor is added to the linear combination of the velocities $\dot{y}_i$.

When formulating Theorem 3, we accept the hypotheses of the Lyapunov theorem about the center adjacent to the oscillation domain Ω.

Theorem 3 (on the corrected conservative system). *Let a conservative system described by the Lagrange equation of the second kind admit an oscillation family with a domain Ω. Then there exists a number $\varepsilon_0 > 0$ such that the corrected system* (21) *with $0 < \varepsilon < \varepsilon_0$ always has a unique cycle in Ω that is ε-close to the oscillation of the conservative system with energy $h = h^*$. The cycle attracts trajectories from the oscillation domain Ω and also those ε-close to Ω. On the solutions of the corrected system, the energy variation law $E = \tilde{T} + \Pi$ of the conservative system is given by*

$$\frac{dE}{dt} = \varepsilon[1 - K(h^*)\Pi(x)]\tilde{T}.$$

Remark 2. The theorem on a corrected conservative system is formulated for a Lagrangian system. The application of Theorem 3 to a conservative system written in other variables is demonstrated in the examples below.

6. SOME EXAMPLES

First, we analyze the rate of approaching the cycle in energy terms. On the trajectories of the corrected conservative system (3), the energy variation law is given by (13). The one-period energy increment is calculated by formula (15). For a given $\varepsilon < \varepsilon_0$, it equals $2T(h)hV$. On the other hand, for an energy value h, the cycle is approached with a rate V almost linear in h. For $h > h^*$, the attraction to the cycle occurs with increasing rate V, which is negative on the above interval. For trajectories with an initial positive energy, the rate V is positive all the time while approaching the cycle. At the point $h^* = \alpha/\beta$, the rate V changes its sign from plus to minus. Let us provide several examples.

Example 1. The van der Pol equation in the adaptive stabilization scheme is described by

$$\ddot{x} + x = \varepsilon(1 - K(h^*)x^2)\dot{x}, \quad K(h^*) = \mathrm{const} > 0. \tag{22}$$

In the van der Pol equation, $K(h^*) = 1$.

For $\varepsilon = 0$, we have a harmonic oscillator with the potential energy $\Pi = x^2/2$. The generating oscillation is given by $x = A\cos t$, where A denotes an amplitude. The energy is $h = A^2/2$; therefore, $\tau = 2t$, $\Pi = hz(\tau)$, and $z = 1 + \cos 2t$. Calculations by formula (11) yield the number $b = 2$. The amplitude for the cycle is $A^* = 2/\sqrt{K(h^*)}$.

Equation (22) satisfies all the hypotheses of Theorem 2. Therefore, the corrected linear oscillator has a globally attracting cycle that passes, depending on $K(h^*)$, through any given point of the phase plane. The oscillation domain Ω in equation (22) coincides with the entire phase plane, excluding the punctured zero. Far from the cycle, as well as near zero, the rate of approaching the cycle is proportional to εh^2. Near the cycle with energy h^*, this rate equals $\varepsilon h|\Delta h|$, $|\Delta h| = h - h^*$.

In equation (22) with $K = 1$, the cycle was constructed by van der Pol [4] and, independently of him, by Andronov [11].

Note that the doubled potential energy is applied in the van der Pol oscillator. In the local ε-theory, the number ε^* is not estimated; the cycle remains attracting when increasing ε; the farther ε is from zero, the lesser the oscillator's behavior will resemble harmonic oscillations.

Example 2. Consider the corrected mathematical pendulum

$$\ddot{x} + \sin x = \varepsilon(1 - 2K(h^*)\sin^2(x/2))\dot{x}.$$

In a nonlinear system, the calculation of the function $z(\tau)$ is complicated due to the unknown function describing the oscillations. For a mathematical pendulum, $\Pi = 2\sin^2(x/2)$. The dependence (11) was numerically obtained in [24].

The peculiarity of the mathematical pendulum is that the oscillation domain has boundedness from above in h. Near the point $h = 0$ (small oscillations), the rate of approaching the cycle is proportional to εh^2.

The mathematical pendulum is investigated in the relative motion problem of a satellite in the plane of a circular orbit [25]. An adaptive scheme for stabilizing the satellite oscillation (in the small), given by

$$\ddot{x} + |\mu|\sin x = \varepsilon\sigma(1 - 2K(h^*)x^2)\dot{x}, \tag{23}$$

was proposed in [2]; also, the local attraction problem was solved for any trajectories from the neighborhood of the stabilized oscillation.

Let us introduce the following notation for the satellite [25]: μ is the inertial parameter ($|\mu| \leqslant 3$); α is the angle between the radius vector of the center of mass and the main central axis of inertia of the satellite in the orbital plane; v is the true anomaly chosen as the independent variable. In equation (23), $x = 2\alpha$ and $\mu > 0$ or $x = 2\alpha + \pi$ and $\mu < 0$, and $\sigma = 1$.

By replacing the term $2K(h^*)x^2$ in equation (23) with $2K(h^*)|\mu|\sin^2(x/2)$, we achieve a cycle attracting all trajectories from the oscillation domain (see Theorem 2). Thus, the global attraction of trajectories is ensured.

For small μ (a "flattened" satellite), it may be interesting to use the stabilized long-periodic oscillation of the satellite (in the large) near the equilibrium instead of the latter.

Example 3. The two-body problem

$$\frac{d^2x}{dt^2} = -\frac{\gamma x}{x^2+y^2}, \quad \frac{d^2y}{dt^2} = -\frac{\gamma y}{x^2+y^2}, \quad \gamma > 0,$$

has the area integral

$$x\frac{dy}{dt} - y\frac{dx}{dt} = c, \quad c = \text{const.} \tag{24}$$

On the integral manifold (24), the dynamics are described by the conservative system

$$\frac{d^2r}{dt^2} = \frac{c^2}{r^3} - \frac{\gamma}{r^2}, \quad r^2 = x^2 + y^2,$$

which possesses a family of elliptic orbits for a fixed value $c = c_*$. (The constant solution of this equation is associated with circular orbits.)

According to Theorem 2, in the corrected system

$$\frac{d^2r}{dt^2} - \frac{c_*^2}{r^3} + \frac{\gamma}{r^2} = \varepsilon((1 - K(h^*)\Pi(r))\frac{dr}{dt}, \quad \Pi = \frac{2c_*^2}{r^2} - \frac{\gamma}{r},$$

any orbit of the two-body problem close to the elliptic orbit with energy h^* is stabilized (in the large). The attraction of other orbits to the plane c_* is ensured by the equation $\dot{\Delta}c = -\Delta c$, $\Delta c = c - c_*$.

Example 4. The dynamics of a heavy solid with a fixed point are described by the classical Euler–Poisson equations

$$\begin{aligned} A\dot{p} &= (B-C)qr + P(z_0\gamma_2 - y_0\gamma_3), \qquad & \dot{\gamma}_1 &= \gamma_2 r - \gamma_3 q, \\ B\dot{q} &= (C-A)rp + P(x_0\gamma_3 - z_0\gamma_1), \qquad & \dot{\gamma}_2 &= \gamma_3 p - \gamma_1 r, \\ C\dot{r} &= (A-B)pq + P(y_0\gamma_1 - x_0\gamma_2), \qquad & \dot{\gamma}_3 &= \gamma_1 q - \gamma_2 p, \end{aligned} \tag{25}$$

written in quasi-coordinates: A, B, and C are the principal moments of inertia of the body; P is the body weight; x_0, y_0, and z_0 are the coordinates of the center of gravity; $\Omega = (p, q, r)$ is the angular velocity; finally, $\Gamma = (\gamma_1, \gamma_2, \gamma_3)$ is the unit vertical vector directed upward.

For $y_0 = 0$, system (25) admits the integral manifold

$$p = 0, \quad r = 0, \quad \gamma_2 = 0. \tag{26}$$

The Mlodzeevskii motions [26] realized on this manifold are described by

$$B\dot{q} = P(x_0\gamma_3 - z_0\gamma_1), \quad \dot{\gamma}_1 = \gamma_2 r - \gamma_3 q, \quad \dot{\gamma}_3 = \gamma_1 q - \gamma_2 p. \tag{27}$$

Using the geometric relation $\gamma_1^2 + \gamma_3^2 = 1$ and the changes $\gamma_1 = \cos\delta$ and $\gamma_2 = \sin\delta$, we reduce system (27) to the mathematical pendulum equation

$$B\ddot{\delta} + P\sqrt{x_0^2 + z_0^2}\sin(\delta - \nu) = 0, \quad \nu = \tan(z_0/x_0).$$

Thus, for any energy value chosen, an attracting cycle (in the large) is constructed for the Mlodzeevskii oscillation. As in Example 3, the attraction to the manifold (26) is ensured by linear dissipation in the variables p, r, and γ_2.

Together with the planar Mlodzeevskii oscillations, the body under study admits a second family of pendulum oscillations [27]. Being spatial, this family is described by a reduced conservative system with one degree of freedom and is separated, step by step, from system (25) with $y_0 = 0$ when constructing the reduced system. Stabilization of the oscillations (in the large) is performed according to the theory of Sections 2–4.

Remark 3. These examples have presented new results for the corresponding problems.

7. CONCLUSIONS

In a conservative system, oscillations form families. Therefore, an oscillation can be stabilized only within a controlled system. Control laws with an explicit time dependence are commonly used.

For a conservative system admitting an oscillation family, a ε-corrected system with an attracting cycle (in the large) is always constructed. An autonomous controller with a small gain is applied. It is given by nonlinear dissipation that acts without delay at the current trajectory point and tracks the potential energy of the system. The attraction domain of the cycle includes the oscillation domain Ω of the conservative system and the ε-neighborhood of Ω. The cycle is constructed for an oscillation with any desired energy of the conservative system. Stabilization is performed according to the adaptive scheme.

The main results of this study have been formulated in three theorems. Theorem 1 shows the existence of a cycle and provides a solution of the stabilization problem in the small (in the neighborhood of the oscillation considered). Next, Theorem 2 establishes the attraction of trajectories evolving from any point of the oscillation domain of the conservative system to the cycle. Theorem 3 extends the results of Theorems 1 and 2 (for a system with one degree of freedom) to the multidimensional case, including the corresponding corrected Lagrangian system. New results in classical problems have been annotated in the examples.

The paper has settled several issues in nonlinear mechanics, oscillation theory, bifurcation, and control theory. In classical mechanics, only linear dissipation is considered. This study gives an example of universal nonlinear dissipation defined by potential energy. It can also explain phenomena in nature.

To investigate the family of nonlinear oscillations of a conservative system, we have proposed to apply the theory of linear systems. The corresponding result (Theorem 4) is postponed to the Appendix, finalizing the author's efforts to present the main material in a comprehensible way. The idea of introducing *new time* brings the system of nonlinear oscillations to an isochronous family.

The conclusions on attraction in the large have become qualitatively new in bifurcation theory: a small parametric perturbation of a system leads to a global rearrangement of its phase portrait.

Concerning control theory, we have suggested the idea of utilizing the nature of an unclosed system. The control law designed is only corrective. In Examples 3 and 4, the stabilization scheme has been demonstrated on classical problems. Other applications include the problems of orbital stabilization of spacecraft in long-range missions. Here, the main constraint is the energy resource. Therefore, maneuvers using gravitational attraction (potential energy) are in demand.

APPENDIX

Transforming a family of nondegenerate (nonlinear) oscillations to an isochronous family

Consider the conservative system

$$\ddot{x} + f(x) = 0, \quad \frac{\dot{x}^2}{2} + \int f(x)dt = h(\text{const}),$$

admitting a family Σ of nondegenerate single-frequency oscillations. On the nondegenerate family, by definition, the period $T(h)$ varies monotonically with h.

Theorem 4. *The nondegenerate oscillation family $\Sigma = \{x(h,t)\}$ of a conservative system is always transformed to an isochronous family with time $\tau = (T(h)/T^*)t$, which is chosen together with the period T^* of the family oscillation, the potential energy $\tilde{\Pi}(x) = \Pi(x)(T^*/T(h))^2$ and the total energy $\tilde{h} = h(T^*/T(h))^2$. On the family Σ, the law*

$$x^2\left(h\left(\frac{T^*}{T(h)}\right)^2, 0\right) = h\left(\frac{T^*}{T(h)}\right)^2$$

relates the amplitude and energy of the oscillations.

Proof. The potential energy $\Pi(x)$ is a function of one variable x. It depends on h through the solution $x = \varphi(h,t)$. When passing to the new independent variable $\tau = (T(h)/T^*)t$, the period on the oscillation becomes T^*, so x is a T^*-periodic function of the variable τ. This occurs for all oscillations of the family, which is thus transformed to an isochronous family with period T^*. For an oscillation with zero initial velocity, we have $\Pi(x(h,0)) = h$. The initial value h is preserved in the function $x(h,\tau)$; therefore,

$$\Pi = hz(\tau), \quad 0 \leqslant z(\tau) \leqslant 1, \quad z(\tau) = z(\tau + T^*).$$

With the new independent variable τ, the energy integral on the family of oscillations takes the form

$$\left(\frac{T(h)}{T^*}\right)^2 \left(\frac{dx}{d\tau}\right)^2 = 2(h - \Pi). \tag{A.1}$$

In a more conventional representation, we obtain

$$\left(\frac{dx}{d\tau}\right)^2 = 2(\tilde{h} - \tilde{\Pi}), \quad \tilde{h} = h\left(\frac{T^*}{T(h)}\right)^2, \quad \tilde{\Pi} = \Pi(x)\left(\frac{T^*}{T(h)}\right)^2. \tag{A.2}$$

The oscillations with initial zero velocity in the transformed system are described by

$$x(\tau) = x(\tilde{h},0)e(\tau), \quad 0 \leqslant e(\tau) \leqslant 1, \quad e(\tau) = e(\tau + T^*).$$

On these oscillations, the law

$$x^2(\tilde{h},0) = \tilde{h} \tag{A.3}$$

expresses the dependence of the amplitude of isochronous oscillations on the system energy $\tilde{h}$. Hence, the amplitude of nonlinear oscillations on the family Σ depends on the energy h according to the law

$$x^2\left(h\left(\frac{T^*}{T(h)}\right)^2, 0\right) = h\left(\frac{T^*}{T(h)}\right)^2. \tag{A.4}$$

Remark 4. In the case of a linear oscillator in (A.3), the function $T(h)$ reduces to a constant and the dependence (A.3) is known.

REFERENCES

1. Tkhai, V.N., Stabilization of Oscillations of a Controlled Autonomous System, *Autom. Remote Control*, 2023, vol. 84, no. 5, pp. 534–545.

2. Tkhai, V.N., An Adaptive Stabilization Scheme for Autonomous System Oscillations, *Autom. Remote Control*, 2024, vol. 85, no. 9, pp. 894–905.

3. Pontryagin, L.S., On Dynamic Systems Close to Hamiltonian, *Zh. Eksp. Teor. Fiz.*, 1934, vol. 4, no. 9, pp. 883–885.

4. Van der Pol, B., Forced Oscillations in a Circuit with Non-linear Resistance (Reception with Reactive Triode), *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1927, ser. VII, vol. 3, no. 13, pp. 65–80.

5. Poincaré, H., Sur les courbes définies par une équation différentielle, *Journal de mathématiques pures et appliquées*, 1882, vol. 8, no. III, pp. 251–296.

6. Andronov, A.A., Vitt, A.A., and Khaikin, S.E., *Teoriya kolebanii* (Oscillation Theory), 2nd ed., edited and supplemented by Zheleztsov, N.A., Moscow: Gos. Izd-vo Fiz. Mat. Lit., 1959.

7. Bogolyubov, N.N., *O nekotorykh statisticheskikh metodakh v matematicheskoi fizike* (On Some Statistical Methods in Mathematical Physics), Kiev: Akad. Nauk Ukr. SSR, 1945.

8. Malkin, I.G., *Nekotorye zadachi teorii nelineinykh kolebanii* (Some Problems of Nonlinear Oscillation Theory), Moscow: Gostekhizdat, 1956.

9. V. K. Mel'nikov, V.K., On the Stability of a Center for Time-Periodic Perturbations, *Tr. Mosk. Mat. Obshch.*, 1963, vol. 12, pp. 3–52.

10. Makarenkov, O. and Lamb, J.S.W., Dynamics and Bifurcations of Nonsmooth Systems: A Survey, *Physica D: Nonlinear Phenomena*, 2012, vol. 241, no. 22, pp. 1826–1844. https://doi.org/10.1016/j.physd.2012.08.002

11. Andronov, A.A. and Vitt, A.A., On Lyapunov Stability, *Zh. Eksp. Teor. Fiz.*, 1933, vol. 3, no. 5, pp. 373–374.

12. Tkhai, V.N., Oscillations in the Autonomous Model Containing Coupled Subsystems, *Autom. Remote Control*, 2015, vol. 76, no. 1, pp. 64–71. https://doi.org/10.1134/S0005117915010051

13. Tkhai, V.N., Stabilizing the Oscillations of a Controlled Mechanical System, *Autom. Remote Control*, 2019, vol. 80, no. 11, pp. 1996–2004. https://doi.org/10.1134/S0005117919110043

14. Rumyantsev, V.V., On the Stability of Motion with Respect to Part of Variables, *Vest. Mosk. Gos. Univ. Ser. Mat. Mekh. Astr. Fiz. Khim.*, 1957, no. 4, pp. 9–16.

15. Rouche, N., Habets, P., and Laloy, M., *Stability Theory by Liapunov's Direct Method*, New York: Springer, 1977.

16. Boubaker, O., The Inverted Pendulum Benchmark in Nonlinear Control Theory: A Survey, *Int. J. Adv. Robot. Syst.*, 2013, vol. 10, no. 5, pp. 233–242. https://doi.org/10.5772/55058

17. Fradkov, A.L., Swinging Control of Nonlinear Oscillations, *Int. J. Control*, 1996, vol. 64, no. 6, pp. 1189–1202. https://doi.org/10.1080/00207179608921682

18. Åström, K.J. and Furuta, K., Swinging up a Pendulum by Energy Control, *Automatica*, 2000, vol. 36, no. 2, pp. 287–295. https://doi.org/10.1016/S0005-1098(99)00140-5

19. Shiriaev, A., Perram, J.W., and Canudas-de-Wit, C., Constructive Tool for Orbital Stabilization of Underactuated Nonlinear Systems: Virtual Constraints Approach, *IEEE T. Automat. Contr.*, 2005, vol. 50, no. 8, pp. 1164–1176. https://doi.org/10.1109/TAC.2005.852568

20. Kant, K., Mukherjee, R., and Khalil, H., Stabilization of Energy Level Sets of Underactuated Mechanical Systems Exploiting Impulsive Braking, *Nonlinear Dynam.*, 2021, vol. 106, pp. 279–293. https://doi.org/10.1007/s11071-021-06831-3

21. Guo, Yu., Hou, B., Xu, Sh., et al., Robust Stabilizing Control for Oscillatory Base Manipulators by Implicit Lyapunov Method, *Nonlinear Dynam.*, 2022, vol. 108, pp. 2245–2262. https://doi.org/10.1007/s11071-022-07321-w

22. Alexandrov, A.Yu. and Tikhonov, A.A., Electrodynamic Control with Distributed Delay for AES Stabilization in an Equatorial Orbit, *Cosmic Res.*, 2022, vol. 60, no. 5, pp. 366–374. https://doi.org/10.31857/S002342062204001X

23. Lyapunov, A.M., *The General Problem of the Stability of Motion*, London–Washington: Taylor & Francis, 1992.

24. Tkhai, V.N., On Stabilization of Pendulum Type Oscillations of a Rigid Body, *Proc. 2018 14th Int. Conf. on Stability and Oscillations of Nonlinear Control Systems (Pyatnitskiy's Conference) (STAB)*, IEEE Xplore: July 9, 2018. https://ieeexplore.ieee.org/document/8408408. https://doi.org/10.1109/STAB.2018.8408408

25. Beletskii, V.V., Motion of an Artificial Earth Satellite Relative to the Center of Mass, *Iskusstvennye Sputniki Zemli*, 1958, no. 1, pp. 25–43.

26. Mlodzeevskii, B.K., On the Permanent Axes in the Motion of a Heavy Rigid Body near a Fixed Point, *Tr. Otd. Fiz. Nauk O-va Lyubit. Estestv. Antropol. Etnograf.*, 1894, vol. 7, no. 1, pp. 46–48.

27. Tkhai, V.N., Spatial Oscillations of a Physical Pendulum, *Proc. 2022 16th Int. Conf. on Stability and Oscillations of Nonlinear Control Systems (Pyatnitskiy's Conference) (STAB)*, IEEE Xplore: June 29, 2022. https://ieeexplore.ieee.org/document/9807507. https://doi.org/10.1109/STAB54858.2022.9807507

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 6, pp. 545–563.

The Use of Optimal Filtering Methods for Passive Monitoring of Available Bandwidth of a Network Channel

A. V. Borisov

*Federal Research Center "Computer Science and Control",
Russian Academy of Sciences, Moscow, Russia
e-mail: ABorisov@frccsc.ru*

Received February 23, 2025
Revised April 13, 2025
Accepted April 15, 2025

Abstract—The article is devoted to the development of mathematical support for the solution to an applied problem of the available bandwidth estimation for a network data transmission channel based on the indirect observations of one of the transmitted flows. The problem is transformed to the state filtering of a Markov jump process given some indirect perfect (noiseless) and counting observations. The obtained estimates are represented as solutions to some coupled systems of ordinary differential equations and recursive relations. The performance of the proposed estimates is illustrated by a numerical example.

Keywords: available bandwidth of a channel, Markov jump process, perfect observations, martingale representation, optimal filtering equations

DOI: 10.31857/S0005117925060048

1. INTRODUCTION

The problem of real-time available bandwidth (ABW) estimation in various telecommunication channels [1–4] is highly relevant for its results to be further used in

— computer network management systems to control the efficiency of network resources utilization,

— congestion control algorithms of the transport protocols,

— multimedia information streaming systems,

— algorithms for resource allocation of software defined networks, etc.

The way ABW and related numerical indicators are understood varies in different publications and may imply

— maximum residual capacity of the given channel at the current load by external flows,

— maximum data transmission rate (throughput) through the channel ensured using some fixed protocol (UDP, TCP, etc.) at the current load by external flows,

— maximum rate of useful data transmission (goodput) through the channel at the current load ensured using the selected protocol under additional requirements to the quality of service (QoS) such as maximum permissible delay, jitter, packet loss ratio, etc.

Currently, there is a whole palette of hardware and software tools for solving this problem. In terms of the statistical information involved, they are divided into active and passive. Active ones use additional service traffic, which is a sequence of small packets, possibly of variable size. The difference between the packets sent and received, including the gaps between them, serve as the basis for calculating the current ABW value. In passive tools, this value is calculated on the basis of information about the real current traffic in the given channel collected with the help of operating systems tools (such as the *tcpdump* utility). We should separately mention the ABW estimation tools based on channel models constructed mainly on queuing systems. It applies simulation modeling rather than real statistical information.

Processing of real data for ABW estimation relies on relatively simple probabilistic models, in particular, linear stochastic observation systems. These are the ones that allow applying the classical Kalman filter [5–7]. At present, the theory of state estimation of stochastic dynamic systems is sufficiently developed, within its framework one can select a model more similar to the operation of a network channel and construct a numerically efficient algorithm for it that estimates the state of the system based on the available data.

This work deals with using the mathematical framework of Markov jump processes (MJP) to construct mathematical models of packet data transmission channels. They are designed to solve the problem of real-time estimation of channel characteristics responsible for ABW from heterogeneous statistical information. The paper has the following structure. Section 2 introduces the class of network channels and transmitted data flows under study and the structure of available observations. The section presents arguments in favor of using the MJP concept to describe the evolution of channel characteristics.

Section 3 contains the theoretical framework for solving the applied ABW monitoring problem. Section 3.1 introduces the observation system under study. Its hidden state to be estimated is a homogeneous MJP with a finite set of states. Some of the observational components are MJP functions recorded without noise while some are Cox processes whose intensity depends on the state. The problem of filtering the MJP state using the available observations is proposed to be considered as a theoretical basis for solving the applied monitoring problem. Section 3.2 deals with solving it. The sought filtering estimate is described by a system of connected ordinary differential equations and recurrence relations.

Note that the proposed optimal filtering problem solved in this work differs from the problems studied in classical monographs [8–10]. In the mentioned works, the structure of observations is such that they can be transformed to a set of Wiener and Poisson processes by a suitable change of the probability measure. In that case, the obtained equations could be interpreted to some extent as different versions of the Kallianpur–Striebel formula [11]. This transition is possible if the condition of nondegeneracy of martingales in observations is fulfilled. By contrast, in the proposed stochastic system, however, some of the observations do not contain noise at all, which makes it impossible to apply the Girsanov transformation of the measure. At the same time, the equations describing the optimal filtering estimate can be treated as a special case of the abstract formula for the optimal filtering of a semimartingale given the observation of semimartingale [12].

Section 4 contains an illustrative example of solving the ABW channel monitoring problem. The channel processes two independent packet flows. The first one is described by a Poisson process with the known intensity. The second hidden flow is described by a Cox process whose intensity varies according to some MJP. The observations include the number of packets from the first flow present in the channel and the sequence of the packets lost from this flow due to congestion. The channel itself is a simple exponential service element with the known intensity combined with a pool of packets of the known capacity. Packets from the pool are randomly selected for transmission. The current ABW depends on its occupancy rate and the intensity of packets arriving from the second flow, so these are the characteristics that are proposed to be estimated. Since no service packet flows are used to obtain statistical information, the proposed monitoring algorithm is categorized as passive. The numerical experiment presented in the section illustrates the high quality of the proposed estimates.

Section 5 presents the analysis of the obtained results and directions for further research.

2. STATEMENT OF THE APPLIED PROBLEM OF AVAILABLE BANDWIDTH MONITORING

We describe the functioning of a network channel of packet data transmission in the form of a controlled stochastic observation system. The channel ensures data transmission of several data flows described by individual characteristics such as

— the intensity of packet arrival from the flow,
— the size of individual packets,
— the total amount of transmitted data,
— the data transmission control protocol, etc.

The channel itself is a set of telecommunication equipment and transmission lines characterized by

— the number of channel hops and their characteristics,
— characteristics of individual network devices (capacity, buffer sizes, internal software characteristics), etc.

Ideally, the channel state is a “snapshot” of the location and movement of the various target and service packets in all parts that make up the given channel, as well as all input and output packet flows, including lost packets.

The channel ABW estimation problem is to determine the maximum packet data flow that could be transmitted through the channel given its current load. In this statement, the problem is unlikely to have an exhaustive solution due to the following facts.

(1) Determining the maximum data flow that can be additionally transmitted through the channel in its current state depends on a number of additional characteristics such as the type of additional data (the protocol type), reliability of data transmission, etc. The point is that the additional bandwidth must be calculated taking into account all overheads and redundancy, including the transmission of service packets, retransmission of lost data and so on. For example, the bandwidth for the subsequent use of UDP traffic will be higher than for TCP since the latter involves resending packets which are not confirmed by the reciever via a special acknowledgement flow.

(2) The channel state mentioned above must have a huge dimension that prevents it from being used in any practical estimation tasks.

(3) The channel characteristics contain uncertainties of different nature, viz.

— the parameters of the individual transmission hops, which form the channel, are usually unknown,
— characteristics of communication devices (their transmission speed,, buffer/storage size) are partially or completely unknown,
— the firmware of the communication devices is proprietary with unknown performance and implemented algorithms,
— network channelling equipment may be simultaneously used by several channels, entailing additional uncertainty of its performance.

(4) Data flows transmitted by the channel also have properties that negatively affect the quality and the very possibility of solving the ABW monitoring problem as they are nonstationary, contain a priori uncertainty in their characteristics, and are partially or completely unobservable due to information security and access sharing restrictions.

In addition, the model is bulky for solving the mentioned practical problem. For this purpose, it is sufficient to consider only the “bottle neck” of the channel, viz. the section with the lowest performance. At the same time, relatively simple queuing systems consisting of service elements, queues or temporary packet storage buffers can be used to describe its operation.

Packet flows can be described by generalized renewal processes [13] — the latter both represent random event flows and may contain some additional packet header information important for subsequent estimation of the channel characteristics. Generally speaking, statistical information available for passive ABW monitoring can include

— part of input information flows,
— part of packet loss flows arising due to various reasons,
— part of service flows such as acknowledgements,
— characteristics of buffer occupancy with packets of the observable flows,
— additional numerical characteristics of individual packets of observable flows, (individual numbers, packet sending/receiving timestamps, etc.).

The state of the communication channel should determine the pair "bottleneck state — total load of the channel". In their mathematical nature, the available observations can also be divided into two different categories, viz. counting processes with their intensity depending on the system state and some functions of the system state observed without additional noise.

As mentioned above, the ABW of a real channel depends on the type of additional load; however, in any case, it will be described by some function of the system state—the current total intensity of packet flows entering the channel and the degree of the channel occupancy. These are the ones that are proposed to be estimated using the available statistical information and then recalculated into ABW of the added flow of some type.

The additional assumption of the Markov property of the observation system under study is certainly a limitation. Nevertheless, it does not appear to be excessive. First, semi-Markov systems (Markov recovery processes) can be reduced to such systems by a suitable extension of the state vector [14–16]. Second, a wide class of non-Markov systems can be approximated using Markov systems [17]. Third, the mathematical framework of Markov processes supported by the theory of martingales allows us to solve a wide class of optimal state and parameter estimation problems. All these conclusions explain the subsequent choice of stochastic differential observation systems class describing the channel state and its filtering.

3. OPTIMAL FILTERING PROBLEM FOR THE STATE OF A MARKOV JUMP PROCESS BY A SET OF NOISELESS AND COUNTING OBSERVATIONS

In what follows, we use the following designations.

— $\mathbf{I}_{\mathcal{A}}(x)$ is the indicator function of the set $\mathcal{A}$,
— $\mathbb{S}^N = \{e_1, \dots, e_N\}$ is the set of coordinate unit vectors in $\mathbb{R}^N$,
— $\mathrm{col}(a^1, \dots, a^N)$ is the column vector composed of the components a^n, $n = \overline{1, N}$,
— $\mathrm{diag}(a)$ – is a diagonal matrix with the vector a as the diagonal,
— $a \wedge b \triangleq \min(a, b)$.

3.1. Statement of the Filtering Problem

On the complete probability space with filtration $(\Omega, \mathcal{F}, \mathsf{P}, \{\mathcal{F}_t\}_{t\geqslant 0})$ we consider the observation system

$$\theta_t = \theta_0 + \int_0^t A^\top \theta_s ds + M_t^\theta, \quad \theta_0 \sim \pi_0, \tag{1}$$

$$\xi_t = C\theta_t, \tag{2}$$

$$\eta_t = \int_0^t G\theta_s ds + M_t^\eta, \tag{3}$$

where

— $\theta_t = col(\theta_t^1, \dots, \theta_t^N) \in \mathbb{S}^N$ is an unobservable system state representing the $\mathcal{F}_t$-adapted homogeneous MJP with the values in $\mathbb{S}^N$, the transition intensity matrix (TIM) A, and the initial distribution π_0; $M_t^\theta = \mathrm{col}(M_t^{\theta,1}, \dots, M_t^{\theta,N})$ is an $\mathcal{F}_t$-adapted martingale,

— $\xi_t = col(\xi_t^1, \dots, \xi_t^M) \in \mathbb{R}^M$ is a noiseless (perfect) observation process; $C \in \mathbb{R}^{M\times N}$ is the observation plan matrix with the columns c^n, $n = \overline{1, N}$;

— $\eta_t = col(\eta_t^1, \dots, \eta_t^K) \in \mathbb{R}^K$ is an observable process with counting components: the matrix $G \in \mathbb{R}^{K\times N}$ determines conditional jump intensities of individual components η depending on the current state θ (G consists of the rows g^k, $k = \overline{1, K}$); $M_t^\eta = col(M_t^{\eta,1}, \dots, M_t^{\eta,K})$ – is an $\mathcal{F}_t$-adapted martingale.

Suppose $\mathcal{O}_t \triangleq \sigma\{\xi_s, \eta_s :\ 0 \leqslant s \leqslant t\}$ be the natural flow of σ-algebras generated by observable processes. The optimal filtering problem for the state θ_t is to calculate the conditional mathematical expectation (CME) $\widehat{\theta}_t \triangleq \mathsf{E}\{\theta_t|\mathcal{O}_t\}$, $t \in [0, T]$; $T < \infty$ is some finite deterministic instant.

We assume that the considered probability triplet with filtration and the observation system satisfy the following conditions.

A) $\mathcal{F}_t \equiv \sigma\{\theta_s, \eta_s :\ 0 \leqslant s \leqslant t\}$ for $\forall\, t \in [0, T]$.

B) The martingale components $M_t^{\eta,k}$ of the counting observations η_t^k are strongly orthogonal to each other and also orthogonal to the martingale M_t^θ in MJP θ_t:

$$\langle \eta, \eta\rangle_t = \int_0^t \mathrm{diag}(G\theta_s)ds, \qquad \langle \eta, \theta\rangle_t \equiv 0.$$

C) Let $\{\tau_j\}_{j\in\mathbb{Z}_+}$ be the instants of jumps in the block process (θ_t, η_t), and $\{\zeta_j\}_{j\in\mathbb{Z}_+}$ be the instants of observation jumps (ξ_t, η_t), $\tau_0 = \zeta_0 \triangleq 0$. We assume that

$$\lim_{j\to+\infty} \tau_j = \lim_{j\to+\infty} \zeta_j = +\infty \qquad \mathsf{P} - \text{a.s.}$$

Then, Markov points $\tau_j' \triangleq \tau_j \wedge T$ and $\zeta_j' \triangleq \zeta_j \wedge T$ will be bounded by the constant T. In what follows, the primes in the designations of the Markov points are omitted for simplicity.

The intensity matrix G of counting observations can be an arbitrary matrix of a suitable dimension consisting of non-negative elements. There are no such restrictions on the matrix of exact observations C, and it only has to have a suitable dimension. Nevertheless, in practice, the matrix C consists of 0 and 1. Often, noiseless observations ξ_t are represented by information about what some set $\mathbb{S}' \subset \mathbb{S}^N$ contains at the current instant θ_t. In this case, the respective row of C will consist of the indicators $\mathbf{I}_{\mathbb{S}'}(e_n)$, $n = \overline{1, N}$.

3.2. Solving the Filtering Problem

Let $\mathcal{C}$ be the set of different columns of the matrix C. We construct the mapping $\Xi : \mathcal{C} \to \mathbb{R}^{1\times N}$ as follows:

$$\Xi(c) \triangleq \sum_{n:\ Ce_n = c} e_n^\top.$$

$\Xi(\cdot)$ characterizes the complete preimage of the mapping $e \to Ce$ in the following sense:

$$\mathrm{diag}(\Xi(c))e = \begin{cases} e, & \text{if } Ce = c,\\ 0 & \text{otherwise.}\end{cases}$$

We denote: $\overline{\theta}_\ell \triangleq \theta_{\zeta_\ell}$, $\overline{\xi}_\ell \triangleq \xi_{\zeta_\ell}$, $\overline{\eta}_\ell \triangleq \eta_{\zeta_\ell}$. We consider a non-decreasing sequence of σ-algebras $\mathfrak{O}_j \triangleq \sigma\{\zeta_\ell, \overline{\xi}_\ell, \overline{\eta}_\ell :\ 0 \leqslant \ell \leqslant j\}$. It is known [19] that $\mathfrak{O}_j \equiv \mathcal{O}_{\zeta_j}$ for all $j \in \mathbb{Z}_+$.

We also construct families of σ-algebras

$$\mathbf{O}_{j,t} \triangleq \sigma\{A \in \mathfrak{D}_j,\ \{\omega:\ t \in [\zeta_j(\omega), \zeta_{j+1}(\omega))\}\}.$$

Obviously, the σ-algebras $\mathbf{O}_{j,t}$ are richer than $\mathfrak{D}_j$, as they are augmented with random events of the form $\{\omega \in \Omega:\ \zeta_j(\omega) \leqslant t < \zeta_{j+1}(\omega)\}$, which carry the following meaning: there have been exactly j jumps of observations by the instant t.

To derive optimal filtering equations, the following auxiliary propositions are required.

Lemma 1. *Suppose* $\widehat{\pi}_j \triangleq \mathsf{E}\left\{\theta_{\zeta_j}|\mathfrak{D}_j\right\}$. *Then* P-*a.s. the following equalities are true*

$$\mathbf{I}_{[\zeta_j,+\infty)}(t)\mathsf{E}\left\{\theta_t \mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)|\mathfrak{D}_j\right\} = \mathbf{I}_{[\zeta_j,+\infty)}(t) m_t, \tag{4}$$

$$\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)\mathsf{E}\left\{\theta_t|\mathbf{O}_{j,t}\right\} = \mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)\mu_t, \tag{5}$$

where the functions m_t *and*

$$\mu_t = (\mathbf{1}m_t)^{-1} m_t \tag{6}$$

are the solutions to the following systems of ordinary differential equations:

$$\begin{cases} \dot{m}_t = \left[\operatorname{diag}(\Xi(\overline{\xi}_j))A^\top - \displaystyle\sum_{k=1}^{K} \operatorname{diag}(g^k)\right] m_t, \quad t > \zeta_j, \\ m_{\zeta_j} = \widehat{\pi}_j, \end{cases} \tag{7}$$

$$\begin{cases} \dot{\mu}_t = \left[\operatorname{diag}(\Xi(\overline{\xi}_j))A^\top - \displaystyle\sum_{k=1}^{K} \operatorname{diag}(g^k)\right] \mu_t - \mu_t \left[\Xi(\overline{\xi}_j)A^\top - \displaystyle\sum_{k=1}^{K} g^k\right] \mu_t, \quad t > \zeta_j, \\ \mu_{\zeta_j} = \widehat{\pi}_j. \end{cases} \tag{8}$$

Proof of Lemma 1 is given in Appendix.

Lemma 2. *The estimate* $\widehat{\pi}_{j+1} \triangleq \mathsf{E}\left\{\theta_{\zeta_{j+1}}|\mathfrak{D}_{j+1}\right\}$ *is specified by the formula*

$$\begin{aligned} \widehat{\pi}_{j+1} = \sum_{k=1}^{K} \left(g^k \mu_{\zeta_{j+1}}\right)^{-1} \operatorname{diag}(g^k)\mu_{\zeta_{j+1}}(\overline{\eta}^k_{j+1} - \overline{\eta}^k_j) \\ + \left(\Xi(\overline{\xi}_{j+1})A^\top \mu_{\zeta_{j+1}}\right)^{-1} \operatorname{diag}\left(\Xi(\overline{\xi}_{j+1})\right)\left(I - \operatorname{diag}\left(\Xi(\overline{\xi}_j)\right)\right) A^\top \mu_{\zeta_{j+1}}, \end{aligned} \tag{9}$$

where the vector $\mu_{\zeta_{j+1}}$ *is the solution to* (8) *taken at the instant* ζ_{j+1}.

The proof of Lemma 2 is given in the Appendix.

Lemmas 1 and 2 allow us to prove the main proposition of this work.

Theorem 1. *The optimal filtering estimate* $\widehat{\theta}_t$ *can be represented as*

$$\widehat{\theta}_t = \mathsf{E}\left\{\theta_t|\mathcal{O}_t\right\} = \sum_{j \geqslant 0} \mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)\mu_t, \tag{10}$$

where the functions μ_t *are specified by the solution of* (8) *on the intervals* $[\zeta_j, \zeta_{j+1})$. *At the instants* ζ_{j+1} *of jumps of observations* (ξ_t, η_t), *the estimate* $\widehat{\theta}_{j+1} = \widehat{\pi}_{j+1}$ *is calculated using recurrence relation* (9); *the filtering estimate at the initial instant is*

$$\widehat{\theta}_0 = (\Xi(\xi_0)\pi_0)^{-1} \operatorname{diag}\left(\Xi(\xi_0)\right)\pi_0. \tag{11}$$

The estimate $\widehat{\theta}_t$ *is the solution to the stochastic system*

$$\begin{aligned}\widehat{\theta}_t &= (\Xi(\xi_0)\pi_0)^{-1}\operatorname{diag}\left(\Xi(\xi_0)\right)\pi_0 \\ &+\int\limits_{\zeta_j}^{t}\left[\left(\operatorname{diag}(\Xi(\xi_s))A^\top-\sum_{k=1}^{K}\operatorname{diag}(g^k)\right)\widehat{\theta}_s-\widehat{\theta}_s\left(\Xi(\xi_s)A^\top-\sum_{k=1}^{K}g^k\right)\widehat{\theta}_s\right]ds \\ &+\sum_{\zeta_j:\zeta_j\leqslant t}\left[\sum_{k=1}^{K}\left(g^k\widehat{\theta}_{\zeta_j-}\right)^{-1}\operatorname{diag}(g^k)\widehat{\theta}_{\zeta_j-}\Delta\eta^k_{\zeta_j}\right. \\ &\left.+\left(\Xi(\xi_{\zeta_j})A^\top\widehat{\theta}_{\zeta_j-}\right)^{-1}\operatorname{diag}\left(\Xi(\xi_{\zeta_j})\right)\left(I-\operatorname{diag}\left(\Xi(\xi_{\zeta_j-})\right)\right)A^\top\widehat{\theta}_{\zeta_j-}-\widehat{\theta}_{\zeta_j-}\right].\end{aligned}\tag{12}$$

Theorem 1 is proved in the Appendix.

Remark 1. Although the integral part of final equation (12) is nonlinear and corresponds to (8), linear system (7) also plays an important role in the numerical implementation of the filtering algorithm. Note that (8) represents a system of Riccati differential equations, numerical solution of which may be difficult for some sets of parameters. The point is that the exact solution μ satisfies the conditions of non-negativity and normalization, and the approximate solution must satisfy the same conditions. Otherwise it loses the probabilistic sense of the conditional distribution, and the approximation itself diverges. To neutralize this disadvantage, more complicated numerical solution algorithms can be used or the time step can be reduced. In contrast to direct numerical solution of (8), μ_t can be computed using (7) *exactly* for any value of the time step h. For this purpose, it is sufficient to compute once the exponential $Q = \exp\left[h\left(\operatorname{diag}(\Xi(\overline{\xi}_j))A^\top - \sum_{k=1}^{K}\operatorname{diag}(g^k)\right)\right]$ and the sum of its rows $q = \mathbf{1}Q$. Then the exact values of the conditional distribution μ_{ζ_j+ih} on a uniform time grid with step h, starting at ζ_j, can be calculated by the simple recurrence

$$\mu_{\zeta_j+(i+1)h} = \frac{1}{q\mu_{\zeta_j+ih}}Q\mu_{\zeta_j+ih}, \quad i \in \mathbb{N}.$$

4. NUMERICAL EXAMPLE OF AVAILABLE BANDWIDTH ESTIMATION

We present the channel structure and the structure of information transmitted through it in more detail. Figure 1 shows the channel operating scheme.

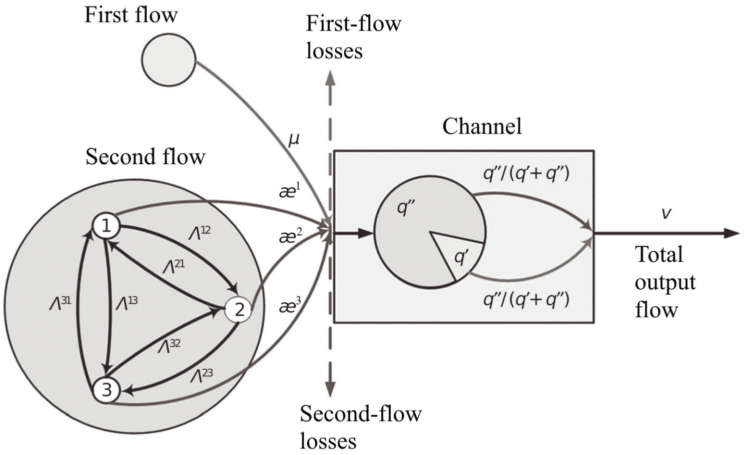

Fig. 1. Scheme of operation of the network channel.

Data in the form of packets arrive in the channel from two independent flows. The first flow—the simplest one with the intensity μ—is partially observable. The second, completely unobservable, is described by a Cox process with the intensity $\varkappa_t$ taking values from the set $\{\varkappa^s\}_{s=\overline{1,S}}$ and varies according to a hidden homogeneous MJP with the known transition intensity matrix $\Lambda = \|\Lambda^{ij}\|_{ij=\overline{1,S}}$. In fact, the second flow is an external integral non-stationary load.

A transmission channel is a service element that can simultaneously contain no more than N^p transmitted packets. Packets arriving in a completely occupied channel are lost. When the channel is not empty, it transmits a packet, spending a random time on it that has an exponential distribution with the constant parameter ν. The transmitted packet is chosen randomly from packets of both flows—if there are q' packets of the first flow and q'' packets of the second flow in the channel at the given instant, then the probabilities that a packet from the first or the second flow will be transmitted are $\frac{q'}{q'+q''}$ and $\frac{q''}{q'+q''}$, respectively. Thus, the considered model implements the *Active Queueing Management* mechanism [19], which provides different flows with fair access to resources in proportion to the number of packets of each flow that are in the channel.

Obviously, the current channel bandwidth is determined by two variables hidden from direct observation, viz. The amount of packets in the server $q_t^\Sigma \triangleq q_t' + q_t''$ and the total packet arrival intensity from the two flows $\varkappa_t^\Sigma \triangleq \mu + \varkappa_t$. These two processes are being monitored.

The arrival processes of packets of both flows into the channel and their processing are described by a unified MJP with a finite set of states $\theta_t = (s_t, q_t', q_t'')$, where s_t is the current state of the second flow ($s = \overline{1,S}$), q_t' is the number of packets of the first flow in the server, and q_t'' is the number of packets of the second flow in the server ($0 \leqslant q', q'' : q_t' + q_t'' \leqslant N^p$). One can easily check that the total number of possible MJP states is $N = \frac{S(N^p+1)(N^p+2)}{2}$.

The matrix A of MJP transition intensities X_t is defined element by element as follows:

— $(i, q', q'') \xrightarrow{\Lambda^{ij}} (j, q', q'')$, $(i, j = \overline{1,S},\ i \neq j,\ q', q'' \geqslant 0 :\ q' + q'' \leqslant N^p)$ — change of intensity of the second flow from $\varkappa^i$ to $\varkappa^i$;

— $(s, q', q'') \xrightarrow{\mu} (s, q'+1, q'')$, $(s = \overline{1,S}, q', q'' \geqslant 0 :\ q' + q'' \leqslant N^p - 1)$ — arrival of a new packet of the first flow into the channel;

— $(s, q', q'') \xrightarrow{\varkappa^s} (s, q', q''+1)$, $(s = \overline{1,S}, q', q'' \geqslant 0 :\ q' + q'' \leqslant N^p - 1)$ — arrival of a new packet of the second flow into the channel;

— $(s, q', q'') \xrightarrow{\frac{q'}{q'+q''}\nu} (s, q'-1, q'')$, $(s = \overline{1,S},\ q' > 0,\ q'' \geqslant 0 :\ q' + q'' \leqslant N^p)$ — transmission of the first flow packet through the channel;

— $(s, q', q'') \xrightarrow{\frac{q''}{q'+q''}\nu} (s, q', q''-1)$, $(s = \overline{1,S},\ q' \geqslant 0,\ q'' > 0 :\ q' + q'' \leqslant N^p)$ — transmission of the second flow packet through the channel.

To estimate the characteristics q_t^Σ and $\varkappa_t^\Sigma$, one can use the following statistical information:

— continuous observations of the number of packets of the first flow currently in the channel: $\xi_t = q_t'$,

— the process that counts packet losses of the first flow caused by channel overflow: $\eta_t = \int_0^t \mathbf{I}_{\{N^p\}}(q_u^\Sigma)\mu du + M_t^\eta$.

We performed numerical experiments for the following parameter values: $N^p = 32$, $S = 3$, $N = 1683$, $\mu = 1$, $\nu = 13$, $T = 2000$,

$$\Lambda = \begin{bmatrix} -0.002 & 0.001 & 0.001 \\ 0.001 & -0.002 & 0.001 \\ 0.001 & 0.001 & -0.002 \end{bmatrix}, \qquad \varkappa = \begin{bmatrix} 1 \\ 5 \\ 11 \end{bmatrix}.$$

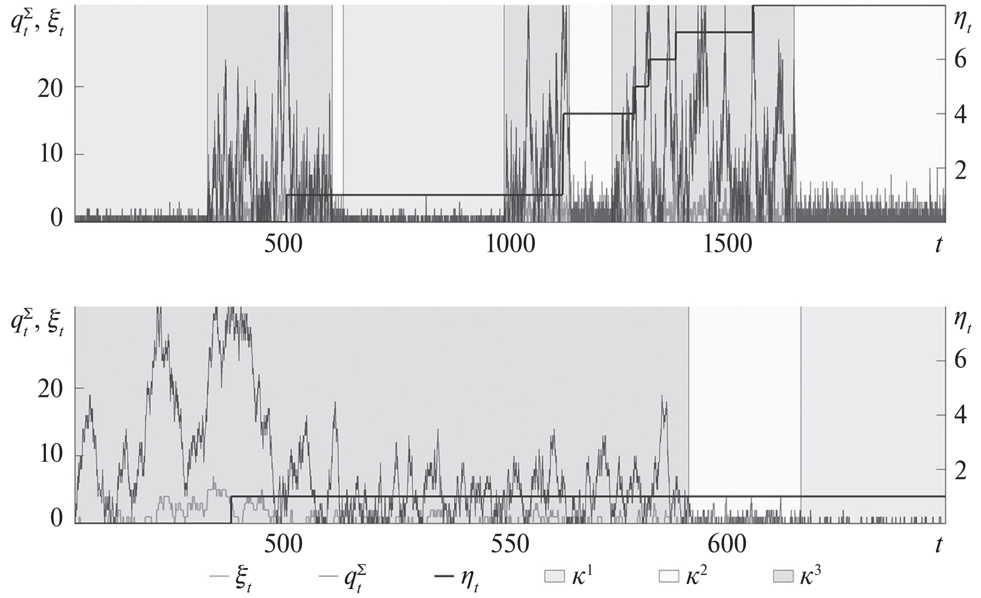

Fig. 2. Evolution of channel load and available observations.

The initial distribution of the MJP describing packet transmission coincides with a stationary one. Simulation of all processes and search for the numerical solution to the optimal filtering problem was performed with the time step $h = 0.01$.

Figure 2 gives information about the hidden state of the channel and available observations:

— the hidden intensity state of the second flow $\varkappa_t$ (displayed as the background filling),

— the hidden channel load q_t^Σ,

— the observed number of packets of the first flow ξ_t that are in the channel,

— the observable counting process of packet losses of the first flow η_t (the values are displayed on the right ordinate axis).

The MJP state filtering estimate $\widehat{\theta}_t$ obtained by solving (12) is a vector whose components are conditional probabilities $\mathsf{P}\{s_t = S, q'_t = Q', q''_t = Q''|\mathcal{O}_t\}$. Using the vector $\widehat{\theta}_t$, we can calculate the estimates of the current total channel load $\widehat{q}_t^\Sigma$:

$$\widehat{q}_t^\Sigma = \sum_{s,q',q''} (q' + q'')\mathsf{P}\{s_t = s, q'_t = q', q''_t = q''|\mathcal{O}_t\}, \tag{13}$$

and the estimates $\widehat{\varkappa}_t^\Sigma$ of the current total intensity of the packets arriving in the channel:

$$\widehat{\varkappa}_t^\Sigma = \sum_{s,q',q''} \varkappa^s \mathsf{P}\{s_t = s, q'_t = q', q''_t = q''|\mathcal{O}_t\}. \tag{14}$$

These characteristics, in turn, allow for real-time ABW estimation under different QoS conditions. We consider the channel operating with the assumption that the second flow is also simple with the constant intensity $\varkappa$. Depending on this parameter, we calculate the average number of packets in the channel $\mathsf{E}\left\{q^\Sigma\right\} = E(\varkappa)$ and the probability $\mathsf{P}\left\{q^\Sigma = N^p\right\} = P_\ell(\varkappa)$ of the packet loss in the stationary mode. Figure 3 shows the dependences $E(\varkappa)$ and $P_\ell(\varkappa)$ (on the auxiliary ordinate axis).

Suppose that the QoS requirement is fixed in the form of an upper bound for the packet loss probability $\overline{\mathsf{P}}_\ell$. We assume that the maximum bandwidth of this channel $\overline{B}$ equals the total intensity

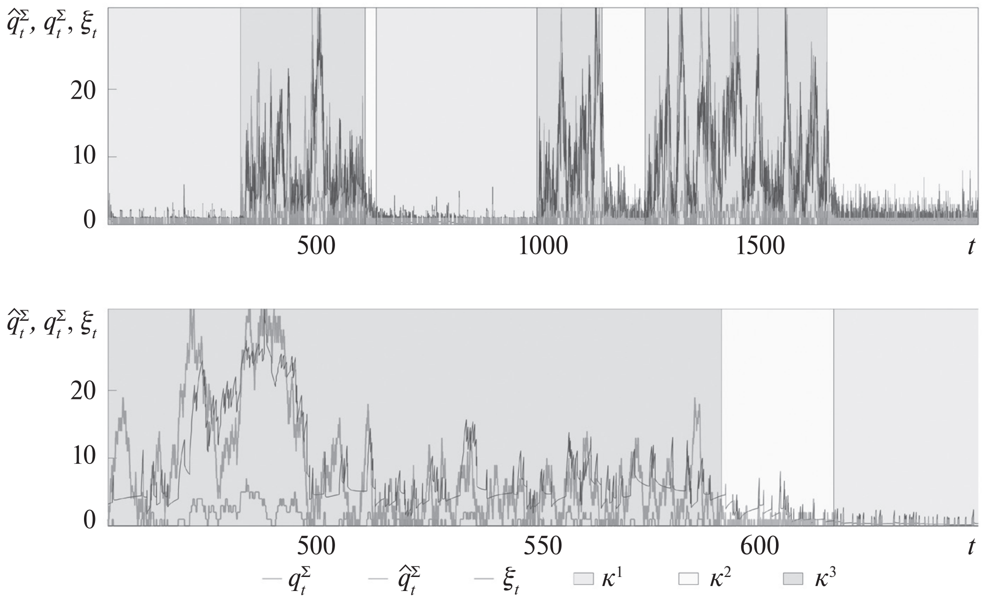

Fig. 3. Average number $E(\varkappa)$ of the packets in the channel and probability of packet loss $P(\varkappa)$.

of both flows, provided that the packet loss probability does not exceed $\overline{\mathsf{P}}_\ell : \overline{B} \triangleq P^{-1}(\overline{\mathsf{P}}_\ell) + \mu$. For example, if we choose $\overline{\mathsf{P}} = 0.05$ the respective maximum bandwidth is $\overline{B} = 12.45$. Then, we propose to take the difference $B_t^a \triangleq \overline{B} - \varkappa_t^\Sigma$, i.e., such a maximum addition to the current intensity of the second flow that does not violate the packet loss probability constraint, as the ABW at the instant t. However, the variable $\varkappa_t^\Sigma$ cannot be observed directly, so we propose to use the variable $\widehat{B}_t^a \triangleq \max(\overline{B} - \widehat{\varkappa}_t^\Sigma, 0)$, which is a function of the obtained estimate $\widehat{\varkappa}_t^\Sigma$, as an ABW estimate.

We consider another type of QoS requirement in the form of an upper bound $\overline{T}$ for the average packet transmission time. If a packet is currently on the server with the current total number of packets q_t^Σ and the channel is in the stationary mode, the average transmission time can be characterized by $\frac{q_t^\Sigma}{\nu}$. Thus, for this QoS requirement to be met, the maximum allowable number of packets being simultaneously at the server should not exceed $\overline{Q} = \overline{T}\nu$. For example, if we choose $\overline{T} = 1$ the upper value $\overline{Q} = 13$ and the respective maximum bandwidth is $\overline{B} = E^{-1}(13) = 11.55$. As an ABW estimate, we propose to use the variable $\widehat{B}_t^a \triangleq \max(\overline{B} - E^{-1}(\widehat{q}_t^\Sigma), 0)$ which is a function of the obtained estimate $\widehat{q}_t^\Sigma$.

Figure 4 shows the evolution of the channel load and its estimate:

— the hidden intensity state of the second flow $\varkappa_t$ (displayed as the background filling),

— the total hidden channel load q_t^Σ,

— the estimate of the total channel load $\widehat{q}_t^\Sigma$,

— the observed number of packets of the first flow ξ_t that are in the channel.

The upper graph shows the trajectories over the entire estimation interval [0; 2000], the lower one shows the interval [450; 650]. Note that a more detailed graph shows the piecewise continuous nature of the estimate: a continuous trajectory on the intervals of no jumps in observations and its jump change corresponding to a jump in observations. The registration of packet loss of the first flow unambiguously signals that the channel is full at the moment, i.e., $q_t^\Sigma = N^p$. The presented filtering estimate behaves in full accordance with this conclusion—at the instant $t = 485.91$, there is a packet loss, and the estimate $\widehat{q}_t^\Sigma$ coincides with the real channel load q_t^Σ, which is N^p, at this instant.

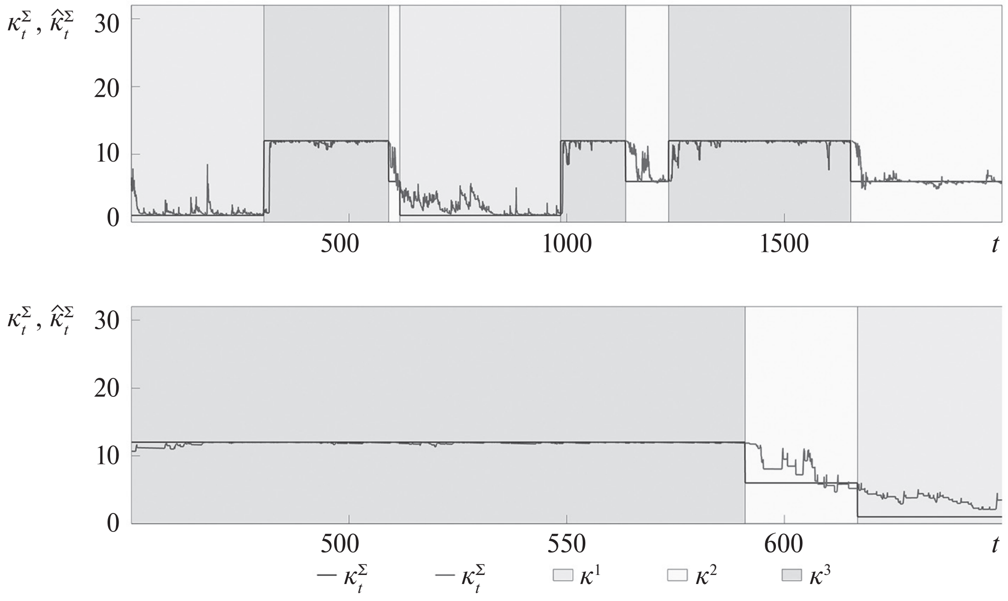

Fig. 4. Channel load and its estimate.

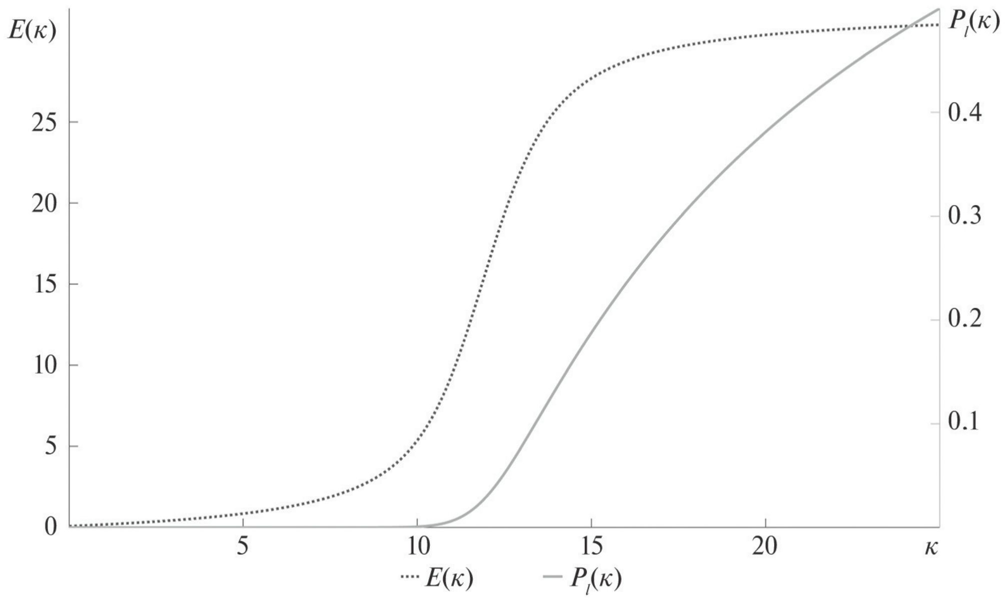

Fig. 5. Total packet arrival intensity and its estimate.

Figure 5 shows the evolution of the total packet arrival intensity in the channel and its estimate:

— the hidden intensity state of the second flow $\varkappa_t$ (displayed as the background filling),

— the packet arrival intensity $\varkappa_t^\Sigma$,

— the intensity estimate $\widehat{\varkappa}_t^\Sigma$.

The upper graph shows the trajectories over the entire estimation interval [0; 2000], the lower one shows the interval [450; 650]. Note that the more detailed graph also shows the piecewise continuous nature of the estimate.

Analyzing the graphs, we can conclude that the proposed estimates of the current characteristics of the channel bandwidth have high accuracy. We compare it with the accuracy of the trivial estimation, viz. unconditional mathematical expectation of the processes q_t^Σ and $\varkappa_t^\Sigma$ calculated for the MJP stationary distribution X. The accuracy of the trivial estimates $\mathsf{E}\left\{q^\Sigma\right\}$ and $\mathsf{E}\left\{\varkappa^\Sigma\right\}$ is characterized by the variances $\mathsf{D}\{q^\Sigma\}$ and $\mathsf{D}\{\varkappa^\Sigma\}$. As accuracy metrics for the proposed estimates we employ the following indices

$$\varepsilon^q = 1 - \frac{\int\limits_0^T \mathsf{E}\left\{(\widehat{q}_t^\Sigma - q_t^\Sigma)^2 dt\right\}}{T\mathsf{D}\{q^\Sigma\}} \quad \text{and} \quad \varepsilon^\varkappa = 1 - \frac{\int\limits_0^T \mathsf{E}\left\{(\widehat{\varkappa}_t^\Sigma - \varkappa_t^\Sigma)^2 dt\right\}}{T\mathsf{D}\{\varkappa^\Sigma\}},$$

which can be considered as analogues of the determination coefficients accepted in mathematical statistics [20]. In this example, the numerators of both indicators are obtained by the Monte Carlo method using a bundle of trajectories $N^{MC} = 10\,000$: $\varepsilon^q = 0.76$ and $\varepsilon^\varkappa = 0.94$.

5. CONCLUSIONS

In this work, we study the applied problem of real-time ABW estimation for a packet transmission channel using observations of one of the data flows served. The available observations include information on the number of packets of the flow currently in it, as well as the counting process of the packet losses. Since the proposed estimation procedure does not need to generate additional service flows through the channel that drain its resources, the proposed monitoring algorithm belongs to the passive class.

The principal idea that allowed us to construct an efficient numerical estimation algorithm is to use a partially observable MJP to describe the channel operating and incoming flows. The statistical information includes a set of some state functions observed without noise and counting processes whose intensity depends on the estimated state. The obtained filtering estimate is given by a sequence of recurrently connected ordinary differential equations calculated in the intervals between jumps of observations and discrete transformations that update the estimates at instants of changes in the observations. The work gives numerical experimental results illustrating the high quality of the presented estimates.

The research in the field of constructing efficient algorithms for ABW channel estimation can be continued in the following directions. First, it is of practical interest to solve the ABW estimation problem for an exponential element with a limited queue for the case of the non-stationary flow of incoming packets described here.

Second, it is important for telecommunication applications to complicate the model of channel and incoming flow operating by switching from Markov to semi-Markov processes.

Third, the ABW estimation problem was solved under conditions of full a priori information about the channel and flows transmitted through it. The construction of procedures for adaptive estimation of probabilistic parameters of the "channel-flows" pair and robust upgrading of the proposed monitoring algorithm also seems promising.

Fourth, the available statistical information in real data transmission networks is much richer than that used in this paper. For example, there are data linking the packet flows at the channel input and output, there is information about the individual transmission time of each packet, and so on. All this information included in the observation system may cause the extended stochastic observation system to cease to be Markov, which radically complicates the ABW estimation algorithms. Therefore, it seems promising to extend the class of observation systems in a way that, on the one hand, preserves the Markov property of the suitably extended system state and, on the other hand, allows us to use some of the additional statistical information similarly to [21, 22].

Fifth, using MJP with a finite set of states to solve applied problems involves very serious complexity. It consists of the rapid growth of the MJP dimension. Indeed, even in the considered numerical example with the channel capacity $K = 3$ and three possible variants of external load, the total number of MJP states is 48. We should also take into account that different states are described by vectors of dimension 3 rather than by scalar values, which leads to an additional increase in the amount of RAM required to implement the filtering algorithm. These circumstances make it topical to develop special efficient software that implements the estimation algorithms in stochastic observation systems with MJP.

FUNDING

The research was supported by the Ministry of Science and Higher Education of the Russian Federation, project No. 075-15-2024-544.

The research was carried out using the infrastructure of the Shared Research Facilities "High Performance Computing and Big Data" (CKP "Informatics") of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences.

APPENDIX

Proof of Lemma 1. To derive systems (7) and (8), we use the method of moments—we construct closed linear stochastic differential systems describing the evolution of the state up to the next observation jump and average them.

If $\{\zeta_\ell^{\eta,k}\}_{\ell\in\mathbb{Z}_+,\, k=\overline{1,K}}$ are the jump instants of the components of the counting observations η, and $\{\zeta_\ell^{\xi}\}_{\ell\in\mathbb{Z}_+}$ are the jump instants of the perfect observations ξ, the instant ζ_{j+1} following ζ_j is determined using an obvious recurrence

$$\zeta_{j+1} = \min_{\zeta_\ell^{\eta,k}>\zeta_j,\, \zeta_{\ell'}^{\xi}>\zeta_j} (\zeta_\ell^{\eta,k}, \zeta_{\ell'}^{\xi}).$$

On the interval $[\zeta_j, +\infty)$ we study the process

$$U_t \triangleq \mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t) = \underbrace{\mathbf{I}_{[\zeta_j,\zeta_{\ell'}^{\xi})}(t)}_{\triangleq V_t} \prod_{k=1}^{K} \underbrace{\mathbf{I}_{[\zeta_j,\zeta_\ell^{\eta,k})}(t)}_{\triangleq W_t^k}.$$

By construction, on any interval $[\zeta_j, t)$ processes V_t and W_t^k experience no more than one jump, and the relations hold

$$\operatorname{diag}(\Xi(\overline{\xi}_j))\theta_t \equiv \theta_{\zeta_j} \text{ for } \forall\, t \in [\zeta_j, \zeta_{\ell'}^{\xi}), \qquad \operatorname{diag}(\Xi(\overline{\xi}_j))\theta_{\zeta_{\ell'}^{\xi}} = 0.$$

By the Doleans formula [23] the processes V_t and W_t^k can be represented as solutions of the equations

$$V_t = \mathbf{I}_{[\zeta_j,+\infty)}(t) \left(1 + \int_{\zeta_j}^{t} V_{s-}\Xi(\overline{\xi}_j)d\theta_s\right), \tag{A.1}$$

$$W_t^k = \mathbf{I}_{[\zeta_j,+\infty)}(t) \left(1 - \int_{\zeta_j}^{t} W_{s-}^k d\eta_s^k\right). \tag{A.2}$$

Indeed, the process $\int_{\zeta_j}^t \Xi(\overline{\xi}_j)d\theta_s$ is a purely discontinuous semimartingale, and the solution of equation (A.1) by the Doleans formula has the form

$$V_t = \mathbf{I}_{[\zeta_j,+\infty)}(t)\exp\left[\Xi(\overline{\xi}_j)(\theta_t - \theta_{\zeta_j})\right]\prod_{s:\, \zeta_j < s \leqslant t}\left(1 + \Xi(\overline{\xi}_j)\Delta\theta_s\right). \tag{A.3}$$

If the process θ did not have any jumps prior to the instant t, then $V_t = V_{\zeta_j} = 1$. If at time $s > \zeta_j$ the first jump of θ occurred that did not lead to any jump of observations ξ, i.e., $\Delta\xi_s = 0$, then

$$\Xi(\overline{\xi}_j)(\theta_s - \theta_{s-}) = \Xi(\overline{\xi}_j)(\theta_s - \theta_{\zeta_j}) = 0,$$

and according to (A.3) $V_s = 1$. The process V will preserve the same value during subsequent jumps of θ that do not lead to jumps of observations ξ. If at time $s > \zeta_j$ the first jump of θ occurred that led to a jump of observations ξ, i.e., $\xi_s \neq \xi_{s-} = \overline{\xi}_j$ and $s = \min\limits_{\zeta^\xi_{\ell'} > \zeta_j} \zeta^\xi_{\ell'}$, then

$$\Xi(\overline{\xi}_j)(\theta_s - \theta_{s-}) = \Xi(\overline{\xi}_j)\theta_s - \Xi(\overline{\xi}_j)\theta_{s-} = 0 - 1 = -1,$$

and according to (A.3) $V_s = 0$. The process V_t will further preserve the same value. Thus, we showed that the solution of equation (A.1)—process (A.3)—coincides with the process $\mathbf{I}_{[\zeta_j,\zeta^\xi_{\ell'})}(t)$ on the ray $[\zeta_j, +\infty)$. We can similarly prove that the processes $W^k_t = \mathbf{I}_{[\zeta_j,\zeta^{\eta,k}_\ell)}(t)$ can be represented as the solution to Eq. (A.2).

Further, from (1)–(3) it follows that V_t and W^k_t can be expanded as follows:

$$V_t = \mathbf{I}_{[\zeta_j,+\infty)}(t)\left(1 + \int\limits_{\zeta_j}^t \Xi(\overline{\xi}_j)A^\top \underbrace{\theta_s V_s}_{\triangleq v_s} ds + M^1_t\right), \tag{A.4}$$

$$W^k_t = \mathbf{I}_{[\zeta_j,+\infty)}(t)\left(1 - \int\limits_{\zeta_j}^t g^k \underbrace{\theta_s W^k_s}_{\triangleq w^k_s} ds + M^{2,k}_t\right), \tag{A.5}$$

where $\mathbf{I}_{[\zeta_j,+\infty)}(t)M^1_t$ and $\mathbf{I}_{[\zeta_j,+\infty)}(t)M^{2,k}_t$ are some martingales. Note that (A.4) and (A.5) can be interpreted as linear stochastic differential equations with martingales in their right-hand sides. Nevertheless these equations are not closed: the right-hand side of the equation for V_t contains the process v_t, and the right-hand side of W^k_t includes w^k_t. From (A.4) and (A.5) we obtain a closed system of linear stochastic differential equations for the vector process $u_t \triangleq \theta_t U_t$.

By Ito's rule and condition (B) the process U_t admits the expansion

$$\begin{aligned} U_t &= \mathbf{I}_{[\zeta_j,+\infty)}(t)\left[1 + \int\limits_{\zeta_j}^t \left(dV_s \prod_{k=1}^K W^k_{s-} + V_{s-}\sum_{k=1}^K \prod_{i:\, i\neq k} W^i_{s-} dW^k_s\right)\right] \\ &= \mathbf{I}_{[\zeta_j,+\infty)}(t)\left[1 + \int\limits_{\zeta_j}^t \left(\Xi(\overline{\xi}_j)A^\top - \sum_{k=1}^K g^k\right) u_s ds + M^3_t\right], \end{aligned}$$

where $\mathbf{I}_{[\zeta_j,+\infty)}(t)M^3_t$ is some martingale. From the definition of processes θ and U it follows that

$$\sum_{\zeta:\, \zeta_j < \zeta \leqslant t} \Delta\theta_\zeta \Delta U_\zeta = -\mathbf{I}_{[\zeta_j,+\infty)}(t)\int\limits_{\zeta_j}^t \left[\theta_{s-} dV_s \prod_{k=1}^K W^k_{s-} - (I - \mathrm{diag}(\Xi(\overline{\xi}_j)))d\theta_s V_{s-} \prod_{k=1}^K W^k_{s-}\right],$$

therefore

$$\begin{aligned} u_t &= \mathbf{I}_{[\zeta_j,+\infty)}(t)\left[\theta_{\zeta_j} + \int\limits_{\zeta_j}^{t} (d\theta_s U_{s-} + \theta_{s-} dU_s) + \sum_{\zeta\colon\, \zeta_j<\zeta\leqslant t} \Delta\theta_\zeta \Delta U_\zeta\right] \\ &= \mathbf{I}_{[\zeta_j,+\infty)}(t)\left[\theta_{\zeta_j} + \int\limits_{\zeta_j}^{t}\left(\operatorname{diag}(\Xi(\overline{\xi}_j))A^\top - \sum_{k=1}^{K}\operatorname{diag}(g^k)\right) u_s ds + M_t^4\right], \end{aligned} \tag{A.6}$$

where $\mathbf{I}_{[\zeta_j,+\infty)}(t)M_t^4$ is some martingale. Calculating CME of both parts of (A.6) with respect to $\mathfrak{O}_j$ and using the fact that

$$\mathsf{E}\left\{\mathbf{I}_{[\zeta_j,+\infty)}(t)M_t^4|\mathfrak{O}_j\right\} = \mathsf{E}\left\{\mathsf{E}\left\{\mathbf{I}_{[\zeta_j,+\infty)}(t)M_t^4|\mathcal{F}_{\zeta_j}\right\}|\mathfrak{O}_j\right\} = 0,$$

we obtain a system of equations equivalent to (7):

$$m_t = \widehat{\pi}_j + \int\limits_{\zeta_j}^{t}\left(\operatorname{diag}(\Xi(\overline{\xi}_j))A^\top - \sum_{k=1}^{K}\operatorname{diag}(g^k)\right) m_s ds.$$

The fact that the function can be represented as a solution to (8) follows from the chain rule of function (6) and system (7).

Suppose $\mathcal{A} \in \mathfrak{O}_j$ is an arbitrary set and $\mathcal{A}' = \mathcal{A} \cap \{\omega:\ \zeta_{j+1} > t\}$. The CME properties lead to the following sequence of equalities being true

$$\begin{aligned} &\mathsf{E}\left\{\mathbf{I}_{[\zeta_j,+\infty)}(t)\left(\theta_t\mathbf{I}_{\mathcal{A}'}(\omega) - \mu_t\mathbf{I}_{\mathcal{A}'}(\omega)\right)\right\} \\ &= \mathsf{E}\left\{\theta_t\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)\mathbf{I}_{\mathcal{A}}(\omega) - \mu_t\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)\mathbf{I}_{\mathcal{A}}(\omega)\right\} \\ &= \mathsf{E}\left\{\mathsf{E}\left\{\theta_t\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)\mathbf{I}_{\mathcal{A}}(\omega) - \mu_t\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)\mathbf{I}_{\mathcal{A}}(\omega)\right\}|\mathfrak{O}_j\right\} \\ &= \mathsf{E}\left\{\left(\mathsf{E}\left\{\theta_t\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)|\mathfrak{O}_j\right\} - \mu_t\mathsf{E}\left\{\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)|\mathfrak{O}_j\right\}\right)\mathbf{I}_{\mathcal{A}}(\omega)\right\} \\ &= \mathsf{E}\left\{(m_t - \mathbf{1}m_t\mu_t)\mathbf{I}_{\mathcal{A}}(\omega)\right\} = 0, \end{aligned}$$

as well as equality (5). Lemma 1 is proved.

Proof of Lemma 2. The sequence $\{(\zeta_j,\overline{\theta}_j,\overline{\xi}_j,\overline{\eta}_j)\}_{j\in\mathbb{Z}_+}$ is Markov. We construct the elements of its transition kernel.

The processes $\theta_t(\eta_t^k - \overline{\eta}_j^k)\mathbf{I}_{[\zeta_j,+\infty)}(t)$ can be expanded as

$$\theta_t(\eta_t^k - \overline{\eta}_j^k)\mathbf{I}_{[\zeta_j,+\infty)}(t) = \mathbf{I}_{[\zeta_j,+\infty)}(t)\left[\int\limits_{\zeta_j}^{t}\left(A^\top\theta_s(\eta_s^k - \overline{\eta}_j^k) + \operatorname{diag}(g^k)\theta_s\right)ds + M_t^5\right],$$

where $\mathbf{I}_{[\zeta_j,+\infty)}(t)M_t^5$ is some martingale. On the other hand,

$$\theta_t(\eta_t^k - \overline{\eta}_j^k)\mathbf{I}_{[\zeta_j,+\infty)}(t) = \theta_t\underbrace{(\eta_t^k - \overline{\eta}_j^k)\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)}_{=0} + \theta_t(\eta_t^k - \overline{\eta}_j^k)\mathbf{I}_{[\zeta_{j+1},+\infty)}(t).$$

From the last two equalities it follows that

$$\begin{aligned}
&\theta_{t\wedge\zeta_{j+1}}(\eta^k_{t\wedge\zeta_{j+1}}-\overline{\eta}^k_j)\mathbf{I}_{[\zeta_j,+\infty)}(t\wedge\zeta_{j+1})\\
&=\theta_{t\wedge\zeta_{j+1}}(\eta^k_{t\wedge\zeta_{j+1}}-\overline{\eta}^k_j)\mathbf{I}_{[\zeta_{j+1},+\infty)}(t\wedge\zeta_{j+1})\\
&=\overline{\theta}_{j+1}(\overline{\eta}^k_{j+1}-\overline{\eta}^k_j)\mathbf{I}_{[\zeta_{j+1},+\infty)}(t)\\
&=\mathbf{I}_{[\zeta_j,+\infty)}(t)\left[\int_{\zeta_j}^{t}\left(A^\top\underbrace{u_s(\eta^k_s-\overline{\eta}^k_j)}_{=0}+\mathrm{diag}(g^k)u_s\right)ds+M^5_{t\wedge\zeta_{j+1}}\right].
\end{aligned}$$

Calculating the CME with respect to $\mathfrak{O}_j$ of the left and right parts of the last equality and using the optional stopping theorem of the right-continuous martingale, we obtain that

$$\begin{aligned}
&\mathsf{E}\left\{\overline{\theta}_{j+1}(\overline{\eta}^k_{j+1}-\overline{\eta}^k_j)\mathbf{I}_{[\zeta_{j+1},+\infty)}(t)|\mathfrak{O}_j\right\}\\
&=\mathsf{E}\left\{\overline{\theta}_{j+1}\mathbf{I}_{\{1\}}(\overline{\eta}^k_{j+1}-\overline{\eta}^k_j)\mathbf{I}_{[\zeta_{j+1},+\infty)}(t)|\mathfrak{O}_j\right\}\\
&=\mathbf{I}_{[\zeta_j,+\infty)}(t)\int_{\zeta_j}^{t}\mathrm{diag}(g^k)m_s ds=\mathbf{I}_{[\zeta_j,+\infty)}(t)\int_{\zeta_j}^{t}\mathrm{diag}(g^k)\mu_s(\mathbf{1}m_s)ds.
\end{aligned}$$

The considered transition corresponds to a jump of the component η^k, i.e. $\zeta_{j+1}=\zeta^{\eta,k}_{\ell}$. Now consider the case when the transition is generated by a jump of observations ξ, i.e. when $\overline{\xi}_{j+1}\neq\overline{\xi}_j$ and $\zeta_{j+1}=\zeta^{\xi}_{\ell}$. Let $c\in\mathcal{C}$ (one of the possible values of observation ξ) be some column of matrix C. Note that

$$\mathrm{diag}(c)\left(I-\mathrm{diag}(\Xi(\overline{\xi}_j))\right)\overline{\theta}_{j+1}=\begin{cases}0, \text{ if } \overline{\xi}_{j+1}=\overline{\xi}_j,\\ \overline{\theta}_{j+1}, \text{ if } \overline{\xi}_{j+1}\neq\overline{\xi}_j.\end{cases}$$

The process $\mathrm{diag}(c)\left(I-\mathrm{diag}(\Xi(\overline{\xi}_j))\right)\theta_t\mathbf{I}_{[\zeta_j,+\infty)}(t)$ can be represented as

$$\begin{aligned}
&\mathrm{diag}(c)\left(I-\mathrm{diag}(\Xi(\overline{\xi}_j))\right)\theta_t\mathbf{I}_{[\zeta_j,+\infty)}(t)\\
&=\mathrm{diag}(c)\left(I-\mathrm{diag}(\Xi(\overline{\xi}_j))\right)\left[\int_{\zeta_j}^{t}A^\top\theta_s ds+M^6_t\right]\mathbf{I}_{[\zeta_j,+\infty)}(t),
\end{aligned}$$

where $\mathbf{I}_{[\zeta_j,+\infty)}(t)M^6_t$ is some martingale. On the other hand,

$$\begin{aligned}
&\mathrm{diag}(c)\left(I-\mathrm{diag}(\Xi(\overline{\xi}_j))\right)\theta_t\mathbf{I}_{[\zeta_j,+\infty)}(t)\\
&=\underbrace{\mathrm{diag}(c)\left(I-\mathrm{diag}(\Xi(\overline{\xi}_j))\right)\theta_t\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)}_{=0}\\
&+\mathrm{diag}(c)\left(I-\mathrm{diag}(\Xi(\overline{\xi}_j))\right)\theta_t\mathbf{I}_{[\zeta_{j+1},+\infty)}(t).
\end{aligned}$$

From the last two equalities it follows that

$$\begin{aligned}
&\mathrm{diag}(c)\left(I-\mathrm{diag}(\Xi(\overline{\xi}_j))\right)\theta_{t\wedge\zeta_{j+1}}\mathbf{I}_{[\zeta_j,+\infty)}(t\wedge\zeta_{j+1})\\
&=\mathrm{diag}(c)\left(I-\mathrm{diag}(\Xi(\overline{\xi}_j))\right)\theta_{t\wedge\zeta_{j+1}}\mathbf{I}_{[\zeta_{j+1},+\infty)}(t\wedge\zeta_{j+1})
\end{aligned}$$

$$= \mathrm{diag}(c)\left(I - \mathrm{diag}(\Xi(\overline{\xi}_j))\right)\overline{\theta}_{j+1}\mathbf{I}_{[\zeta_{j+1},+\infty)}(t)$$

$$= \mathrm{diag}(c)\left(I - \mathrm{diag}(\Xi(\overline{\xi}_j))\right)\left[\int\limits_{\zeta_j}^{t} A^\top u_s ds + M^6_{t\wedge\zeta_{j+1}}\right]\mathbf{I}_{[\zeta_j,+\infty)}(t).$$

Again calculating the CME with respect to $\mathfrak{O}_j$ of the left and right parts of the equality and using the optional stopping theorem of the martingale, we obtain

$$\begin{aligned}
&\mathsf{E}\left\{\mathrm{diag}(c)\left(I - \mathrm{diag}(\Xi(\overline{\xi}_j))\right)\overline{\theta}_{j+1}\mathbf{I}_{[\zeta_{j+1},+\infty)}(t)|\mathfrak{O}_j\right\}\\
&= \mathsf{E}\left\{\overline{\theta}_{j+1}\mathbf{I}_{\{c\}}(\overline{\xi}_{j+1})\left(1 - \mathbf{I}_{\{\overline{\xi}_j\}}(\overline{\xi}_{j+1})\right)\mathbf{I}_{[\zeta_{j+1},+\infty)}(t)|\mathfrak{O}_j\right\}\\
&= \mathbf{I}_{[\zeta_j,+\infty)}(t)\int\limits_{\zeta_j}^{t}\mathrm{diag}(c)\left(I - \mathrm{diag}(\Xi(\overline{\xi}_j))\right)A^\top m_s ds\\
&= \mathbf{I}_{[\zeta_j,+\infty)}(t)\int\limits_{\zeta_j}^{t}\mathrm{diag}(c)\left(I - \mathrm{diag}(\Xi(\overline{\xi}_j))\right)A^\top \mu_s(\mathbf{1}m_s)ds.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathsf{P}\left\{\overline{\theta}_{j+1} = e_i,\ \overline{\xi}_{j+1} = c,\ \overline{\xi}_{j+1} \neq \overline{\xi}_j,\ \zeta_{j+1} \in [t, t+dt)|\mathfrak{O}_j\right\}\\
&\quad = e_i^\top \mathrm{diag}(c)\left(I - \mathrm{diag}(\Xi(\overline{\xi}_j))\right)A^\top \mu_t(\mathbf{1}m_t)dt
\end{aligned}\tag{A.7}$$

and

$$\begin{aligned}
&\mathsf{P}\left\{\overline{\theta}_{j+1} = e_i,\ \overline{\xi}_{j+1} = \overline{\xi}_j,\ \overline{\eta}^k_{j+1} - \overline{\eta}^k_j = 1,\ \zeta_{j+1} \in [t, t+dt)|\mathfrak{O}_j\right\}\\
&\quad = e_i^\top \mathrm{diag}(g^k)\mu_t(\mathbf{1}m_t)dt.
\end{aligned}\tag{A.8}$$

Further, we use a technique standard for deriving the equations of optimal state filtering of Markov observation systems with discrete time [24, 25]. Let (α, β, γ) be a block random vector, $P(\mathcal{A}, \mathcal{B}|\gamma)$ be the conditional distribution of the pair (α, β) with respect to γ, i.e.

$$\mathsf{P}\{\alpha \in \mathcal{A}, \beta \in \mathcal{B}|\gamma\} = P(\mathcal{A}, \mathcal{B}|\gamma) \qquad \mathsf{P} - \text{a.s.}$$

Let there also exist a measure $\chi(a, b|\gamma)$ such that $P \ll \chi$ and $\rho(a, b|\gamma) = = \frac{dP}{d\chi}(a, b|\gamma)$ be the corresponding Radon-Nikodym derivative. Then the CME $\mathsf{E}\{\alpha|\beta, \gamma\}$ can be computed using the following variant of Bayes formula:

$$\mathsf{E}\{\alpha|\beta, \gamma\} = \left(\int \rho(a', \beta|\gamma)d\chi(a', \beta|\gamma)\right)^{-1}\int a\rho(a, \beta|\gamma)d\chi(a, \beta|\gamma).\tag{A.9}$$

Formula (9) is a special case of (A.9) obtained by substituting (A.7) and (A.8) into it. Lemma 2 is proved.

Proof of Theorem 1. By direct substitution, we can check that the estimate $\widehat{\theta}_t$, "glued" from solutions of systems (8) with jumps described by (9) and initial condition (11), is a solution of (12). Therefore, to prove the theorem it is sufficient to check the truth of equality (10).

The observable process (ξ_t, η_t) represents a multivariate point process (MPP) with state space $\mathbf{B} \triangleq \mathcal{C} \times \mathbb{Z}^K_+$, which can be represented in the equivalent form of stochastic measure ϕ [18], defined on the measurable space $([0, T] \times \mathbf{B}, \mathcal{B}([0, T]) \times 2^{\mathbf{B}})$:

$$\phi(\omega, dt, dy_1, dy_2) = \sum_{j\in\mathbb{Z}_+}\delta_{(\zeta_j(\omega), \overline{\xi}_j(\omega), \overline{\eta}_j(\omega))}(dt, dy_1, dy_2).$$

In [18] it was proved that the natural flow of σ-algebras generated by observations coincides with the one generated by the stochastic measure, i.e.

$$\sigma\left\{\phi([a,b)\times\{c\}\times\{\mathbf{z}\}):\ [a,b)\in\mathcal{B}([0,T]),\ c\in\mathcal{C},\ \mathbf{z}\in\mathbb{Z}_+^K\right\}\equiv\mathcal{O}_t,\quad t\in[0,T].$$

The base of the σ-algebra $\mathcal{B}([0,T])\times 2^{\mathbf{B}}$ consists of sets of the form $[a,b)\times\{c\}\times\{\mathbf{z}\}$, so by virtue of the theorem on monotone classes [23] to prove the truth of equality (10) it is sufficient to check the validity of equality

$$\mathsf{E}\left\{\left(\sum_{j\geqslant 0}\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)\mu_t-\theta_t\right)\phi([a,b)\times\{c\}\times\{\mathbf{z}\})\right\}\equiv 0$$

for all sets $[a,b)\times\{c\}\times\{\mathbf{z}\}$ of the base.

From the properties of CME and (4)–(6) follows the sequence of equalities

$$\begin{aligned}
&\mathsf{E}\left\{\left(\sum_{j\geqslant 0}\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)\mu_t-\theta_t\right)\phi([a,b)\times\{c\}\times\{\mathbf{z}\})\right\}\\
&=\mathsf{E}\left\{\sum_{j\geqslant 0}\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)\left(\mu_t-\theta_t\right)\mathbf{I}_{[a,b)}(t)\sum_{\ell\geqslant 0}\mathbf{I}_{[\zeta_\ell,\zeta_{\ell+1})}(t)\mathbf{I}_{\{c\}}(\overline{\xi}_\ell)\mathbf{I}_{\{\mathbf{z}\}}(\overline{\eta}_\ell)\right\}\\
&=\mathbf{I}_{[a,b)}(t)\sum_{j\geqslant 0}\mathsf{E}\left\{\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)\left(\mu_t-\theta_t\right)\mathbf{I}_{\{c\}}(\overline{\xi}_j)\mathbf{I}_{\{\mathbf{z}\}}(\overline{\eta}_j)\right\}\\
&=\mathbf{I}_{[a,b)}(t)\sum_{j\geqslant 0}\mathsf{E}\left\{\mathsf{E}\left\{\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)\left(\mu_t-\theta_t\right)\mathbf{I}_{\{c\}}(\overline{\xi}_j)\mathbf{I}_{\{\mathbf{z}\}}(\overline{\eta}_j)|\mathfrak{O}_j\right\}\right\}\\
&=\mathbf{I}_{[a,b)}(t)\sum_{j\geqslant 0}\mathsf{E}\left\{\left(\underbrace{\mathbf{I}_{[\zeta_j,+\infty)}(t)\mathsf{E}\left\{\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)|\mathfrak{O}_j\right\}}_{=\mathbf{I}_{[\zeta_j,+\infty)}(t)\mathbf{1}m_t}\mu_t\right.\right.\\
&\qquad\left.\left.-\underbrace{\mathbf{I}_{[\zeta_j,+\infty)}(t)\mathsf{E}\left\{\theta_t\mathbf{I}_{[\zeta_j,\zeta_{j+1})}(t)|\mathfrak{O}_j\right\}}_{=\mathbf{I}_{[\zeta_j,+\infty)}(t)m_t}\right)\mathbf{I}_{\{c\}}(\overline{\xi}_j)\mathbf{I}_{\{\mathbf{z}\}}(\overline{\eta}_j)\right\}=0.
\end{aligned}$$

Theorem 1 is proved.

REFERENCES

1. Guerrero, C., *Available Bandwidth Estimation: A Hidden Markov Model Approach*, Saarbrücken: Lambert Academic Publishing, 2010.

2. Chaudhari, S. and Biradar, R., Survey of Bandwidth Estimation Techniques in Communication Networks, *Wireless Pers. Commun.*, 2015, vol. 83, pp. 1425–1476.

3. Airon, M. and Gupta, N., Bandwidth Estimation Tools and Techniques: A Review, *International Journal of Research*, 2017, vol. 4, pp. 1250–1265.

4. Salcedo, D., Cesar, D. Guerrero, C., and Martinez, R., Available Bandwidth Estimation Tools: Metrics, Approach and Performance, *Int. J. Commun. Networks Inform. Security*, 2018, vol. 10, no. 3, pp. 580–587.

5. Kalman, R., A New Approach to Linear Filtering and Prediction Problems, *J. Basic Engineer*, 1960, vol. 82, no. 1, pp. 35–45.

6. Bergfeldt, E., Ekelin, S., and Karlsson, J., Real-time Available-Bandwidth Estimation Using Filtering and Change Detection, *Computer Networks*, 2009, vol. 53, no. 15, pp. 2617–2645.

7. Bozakov, Z. and Bredel, M., Online Estimation of Available Bandwidth and Fair Share Using Kalman Filtering, *Proc. of 8th International IFIP-TC 6 Networking Conference*, 2009, LNCS, vol. 5550, Springer, Berlin, Heidelberg, pp. 548–561.

8. Liptser, R. and Shiryaev, A., *Statistika sluchainykh protsessov*, Moscow: Nauka, 1974; *Statistics of Random Processes*, Berlin: Springer-Verlag, 2001.

9. Wong, E. and Hajek, B., *Stochastic Processes in Engineering Systems*, New York: Springer, 1984.

10. Elliott, R., Aggoun, L., and Moore, J., *Hidden Markov Models: Estimation and Control*, New York: Springer, 2008.

11. Kallianpur, G. and Striebel, C., Stochastic Differential Equations Occurring in the Estimation of Continuous Parameter Stochastic Processes, TVP, 1969, vol. 14, no. 4, pp. 597–622.

12. Liptser, R. and Shiryaev, A., *Teoriya martingalov*, Moscow: Fizmatlit, 1986; *Theory of Martingales*, Dordrecht: Kluwer Academic Publishers, 1989.

13. Brémaud, P., *Point Process Calculus in Time and Space*, New York: Springer, 2021.

14. Limnios, N. and Oprisan, G., *Semi-Markov Processes and Reliability*, New York: Springer-Science+Business Media, LLC, 2001.

15. Grabski, F., *Semi-Markov Processes: Applications in System Reliability and Maintenance*, Amsterdam: Elsevier, 2015.

16. Cocozza-Thivent, C., *Markov Renewal and Piecewise Deterministic Processes*, Cham: Springer Nature Switzerland AG, 2021.

17. Kalashnikov, V. and Rachev, S., *Matematicheskie metody postroeniya stokhasticheskikh modelei obsluzhivaniya*. Moscow: Nauka, 1988; *Mathematical Methods for Construction of Queueing Models*, Belmont: Thomson Brooks/Cole, 1990.

18. Jacod, J., Multivariate Point Processes: Predictable Projection, Radon-Nikodym Derivatives, Representation of Martingales, *Z. Wahrsch. Verw. Geb.*, 1975, vol. 31, pp. 235–253.

19. Floyd, S. and Jacobson, V., Random Early Detection Gateways for Congestion Avoidance, *IEEE/ACM Trans Netw.*, 1993, vol. 1, no. 4, pp. 397–413.

20. Draper, N. and Smith, H., *Applied Regression Analysis*, Hoboken: John Wiley & Sons, 1998.

21. Borisov, A, Kurinov, Yu., and Smelyansky, R., Probabilistic Analysis of a Class of Markov Jump Processes, *Inform. i ee primen.*, 2024, vol. 18, no. 3, pp. 30–37.

22. Borisov, A., Filtering of States and Parameters of Special Markov Jump Processes via Indirect Perfect Observations, *Inform. i ee primen.*, 2025, vol. 19, no. 1 (in print).

23. Elliott, R., *Stochastic Calculus and Applications*, New York: Springer, 1982.

24. Sorenson, H. and Stubberud, A., Non-linear Filtering by Approximation of the A Posteriori Density, *Int. J. Contr.*, 1968, vol. 8, no. 1, pp. 33–51.

25. Bertsekas, D. and Shreve, S., *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, 1978.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 6, pp. 564–576.

Kalman Filter in the Strapdown Airborne Gravimetry Problem Based on the Refined Model of GNSS Data Errors

A. S. Arkhipova[*,a] **and V. S. Vyazmin**[**,b]

[*] *Robotics Center, Sberbank, Moscow, Russia*
[**] *Lomonosov Moscow State University, Moscow, Russia*
e-mail: [a] *alsearkhipova@sberbank.ru, vadim.vyazmin@math.msu.ru*

Received December 30, 2024
Revised February 19, 2025
Accepted March 28, 2025

Abstract—The paper considers the problem of gravity disturbance determination on the aircraft flight trajectory using measurements from a strapdown airborne gravimeter. The gravimeter measurements include raw data from the inertial sensors and global navigation satellite system (GNSS) receivers. The problem is reduced to optimal stochastic estimation given an a priori model of gravity disturbance in the time domain and stochastic models of the inertial sensor measurement errors and the GNSS data errors (the errors of kinematic accelerations derived from carrier phase measurements). The estimation algorithm is the Kalman filter and smoothing. We show that the accuracy of gravity estimation can be improved when using a refined model of the kinematic acceleration errors instead of using the traditional model (a white noise process). The refined model is given as the second difference of a discrete-time white noise.

Keywords: airborne gravimetry, strapdown gravimeter, GNSS, kinematic acceleration errors, optimal estimation, Kalman filter

DOI: 10.31857/S0005117925060057

1. INTRODUCTION

Strapdown airborne gravimetry aims to determine gravity disturbance from measurements of a strapdown gravimeter on the flight trajectory of an aircraft (fixed-wing aircraft, helicopter or drone). The gravity disturbance is the difference between the magnitudes of the real gravity vector and normal gravity vector. The normal gravity is defined for the ellipsoid model of the Earth [1].

A strapdown airborne gravimeter consists of a strapdown inertial navigation system or inertial measurement unit (IMU), which includes high-precision inertial sensors (accelerometers and gyroscopes) and geodetic-grade global navigation satellite system (GNSS) receivers (onboard and ground-based). Raw gravimeter data includes measurements from the IMU inertial sensors and GNSS receivers and are processed following the steps of a postprocessing scheme [2] (see also, e.g., [3–5]):

(1) computing GNSS solutions (determining position, velocity, and accelerations of the aircraft from raw GNSS measurements);

(2) computing integrated IMU/GNSS solutions (estimating the attitude of the gravimeter's IMU, the instrumental errors of the IMU inertial sensors, GNSS antenna offsets, etc.);

(3) computing gravimetric solution (determining the gravity disturbance along the flight trajectory).

In the first stage of the postprocessing scheme, GNSS solutions are calculated using code pseudorange, Doppler pseudorange rate and carrier-phase measurements from dual-frequency receivers accessing signals from one or several satellite constellations (GPS, GLONASS, GALILEO, BeiDou, etc.) [4, 5]. The carrier-phase differential mode of data processing is typically used to reduce GNSS measurement errors caused by ionospheric and tropospheric effects. Alternatively, the Precise Point Positioning (PPP) technology, which does not require ground-based receivers (base stations), may be used to obtain high-accuracy GNSS solutions [6]. The purpose of this postprocessing stage is to calculate the aircraft's velocity and/or accelerations with high accuracy, primarily using carrier phase measurements and, less commonly, Doppler pseudorange rate measurements (see, e.g., [5, 7]). Another approach is to use these two types of measurements simultaneously, with Doppler pseudorange rate measurements serving as auxiliary data in case of carrier phase measurement failures [4].

In the second stage of postprocessing (IMU/GNSS integration), the position, velocity and attitude of the gravimeter's IMU computed from the inertial sensor measurements are refined using the GNSS solutions. The mathematical foundation of this task is the IMU error dynamics equations expressed in the axes of the navigation (geodetic) frame (see, e.g., Section 5.2 in [1]). The vertical channel is excluded from the equations, which removes gravity disturbance (as an unknown unknwn variable) from the equations. The influence of gravity disturbance on the horizontal channels, expressed as the product of gravity disturbance and the horizontal attitude errors, is commonly neglected as a second-order small quantity. Next, an optimal stochastic estimation problem is formulated and solved using Kalman filtering and smoothing, which results in the state vector estimates obtained at each point of the aircraft's flight trajectory. In particular, the estimates of the IMU attitude errors and the inertial sensor errors are obtained at this stage (for details, please see [2]).

In this work, we focus on the third stage of postprocessing, which is solving the gravimetry problem (determining gravity on the flight trajectory). The basic idea of solving the problem is to form the difference between the IMU vertical channel data and GNSS data (kinematic acceleration in the vertical direction). The problem is formulated as an optimal stochastic estimation problem given a priori stochastic models of gravity disturbance, sensor measurement errors and GNSS data errors. The solution to the estimation problem is provided by the Kalman filter and smoothing. Traditionally, the GNSS acceleration errors are modeled as white noise [8, 9]. However, these errors typically have a more complex structure and are correlated in time.

In this work, a refined stochastic model of the GNSS acceleration errors is introduced (for the first time to our knowledge) under the assumption that GNSS accelerations are computed using carrier phase measurements (Doppler pseudorange rates are not considered in the paper). Namely, we assume that the kinematic accelerations are computed based on double differences of GNSS carrier phase measurements (involving three successive time epochs) [10]. The refined model of GNSS acceleration errors is defined in the time domain as the second central difference of a discrete-time white noise process. The airborne gravimetry problem is then reduced to a linear optimal stochastic estimation problem and solved in the Kalman filter framework. The results of testing the proposed approach using real airborne gravimeter data are presented and discussed. The results demonstrate a higher accuracy of gravity determination comparing to the traditional approach based on the simplified GNSS acceleration error model (white noise).

2. MATHEMATICAL MODELS

2.1. Basic Equations

The following notation is used in this study:

- M is the proof mass of the accelerometer triad of the gravimeter's IMU;

- Mx is the geodetic frame with the origin at the point M and the axes pointing east, north, and up along the normal to the reference ellipsoid (denoted as E, N, Up) [1];
- Mz is the IMU body frame with the axes z_1, z_2, z_3 mutually perpendicular and aligned with the sensitivity axes of the calibrated IMU accelerometers.

The mathematical foundation of the strapdown airborne gravimetry problem is the equation of motion of the point M expressed in the geodetic frame Mx (for the expressions in other reference frames see, e.g., [1]):

$$\mathbf{a}_x = -(\boldsymbol{\Omega}_x + 2\mathbf{u}_x) \times \mathbf{v}_x + \mathbf{g}_x^0 + \delta\mathbf{g}_x + L_{zx}^{\mathrm{T}}\mathbf{f}_z, \tag{1}$$

where $\mathbf{v}_x, \mathbf{a}_x$ are the velocity and acceleration vectors (relative to the Earth) of the point M, respectively, expressed in the geodetic frame Mx; $\boldsymbol{\Omega}_x$, $\mathbf{u}_x$ are the angular velocity of Mx relative to the Earth and the angular velocity of the Earth relative to the inertial space, respectively; $\mathbf{g}_x^0 = (0, 0, -g_0)^{\mathrm{T}}$ is the normal gravity vector at the point M [11]; $\delta\mathbf{g}_x$ is the gravity disturbance vector [1]; $\mathbf{f}_z$ is the specific force at the point M expressed in the IMU body frame Mz; L_{zx} is the transformation matrix from the geodetic frame Mx to the body frame Mz (an orthogonal matrix).

The vertical projection of (1) is given by the formula (the fundamental equation of airborne scalar gravimetry):

$$a_{up} = g_{etv} - g_0 - \delta g + L_3^{\mathrm{T}}\mathbf{f}_z, \tag{2}$$

where a_{up} is the relative vertical acceleration of the point M, g_{etv} is the Eötvös correction term (the vertical projection of the inertial forces), g_0 is the magnitude of the normal gravity vector at the point M, δg is the magnitude of $\delta\mathbf{g}_x$ (gravity disturbance), L_3 is the third column of the transformation matrix L_{zx}.

The gravimetric problem, as noted earlier, is solved at the final stage of the postprocessing strategy and involves determining the gravity disturbance δg on the flight trajectory based on (2) using raw accelerometer measurements, GNSS solutions (positions, velocities and accelerations), and integrated IMU/GNSS solutions (estimates of the IMU attitude).

2.2. Measurement Models

In (2), only the specific force $\mathbf{f}_z$ is measured directly (by the IMU accelerometers). The measurement model is

$$\mathbf{f}_z' = \mathbf{f}_z + \mathbf{q}_f, \tag{3}$$

where $\mathbf{f}_z'$ is the vector of three accelerometer measurements and $\mathbf{q}_f$ is the vector of measurement errors.

The vertical kinematic acceleration of the aircraft is derived from raw (carrier phase) GNSS measurements and can be expressed as

$$a_{up}^{gps} = a_{up} + e_a, \tag{4}$$

where a_{up}^{gps} is the vertical acceleration computed from raw GNSS data and e_a is the acceleration error.

The Eötvös correction term g_{etv} is determined using the position and the horizontal components of the velocity vector at the point M [1]. The normal gravity g_0 at the point M is determined using a theoretical model for the normal gravity at the ellipsoid (e.g., Helmert's formula or Somigliana's formula) and the height correction term [11]. For computing the Eötvös correction term and normal gravity, the GNSS position and velocity solutions (or the IMU/GNSS solutions) are used [4]. The lever-arm effect caused by the distance between the onboard GNSS antenna and the IMU should be taken into account.

The estimate $\tilde{L}_3$ of the third column L_3 is also assumed to be known and computed from the estimates of the IMU attitude angles (heading, roll and pitch) [12].

The computed vertical acceleration a'_{up} is defined as

$$a'_{up} = g_{etv} - g_0 + \tilde{L}_3^{\mathrm{T}} \mathbf{f}'_z. \tag{5}$$

2.3. Fundamental Equation of Airborne Strapdown Gravimetry

Define the difference between the computed vertical acceleration and the true value as $\Delta a_{up} = a_{up} - a'_{up}$, which can be expressed using (2) and (5) in the following form:

$$\Delta a_{up} = -\delta g + L_3^{\mathrm{T}} \mathbf{f}_z - \tilde{L}_3^{\mathrm{T}} \mathbf{f}'_z. \tag{6}$$

The transformation matrix $\tilde{L}_{zx}$ is computed from the integrated IMU/GNSS solution, that is, from the estimated (refined) IMU attitude angles (heading, roll, pitch), and is an orthogonal matrix. The relationship between the true transformation matrix L_{zx} and the computed matrix $\tilde{L}_{zx}$ can be written as

$$L_{zx} = (I + \hat{\kappa}) \tilde{L}_{zx},$$

where $\hat{\kappa}$ is a skew-symmetric matrix composed of components of the small rotation vector $\kappa = (k_E, k_N, k_{Up})^{\mathrm{T}}$. The vector κ characterizes the residual attitude error, that is, the errors in the IMU attitude estimates obtained at the IMU/GNSS integration stage. Here, k_E, k_N are the residual attitude errors in the east and north directions, respectively, and k_{Up} is the error of the azimuthal error estimate.

The right-hand side of (6) can be rewritten using (3) as:

$$L_3^{\mathrm{T}} \mathbf{f}_z - \tilde{L}_3^{\mathrm{T}} \mathbf{f}'_z = (L_3^{\mathrm{T}} - \tilde{L}_3^{\mathrm{T}}) \mathbf{f}'_z - \tilde{L}_3^{\mathrm{T}} \mathbf{q}_f = -k_E f'_N + k_N f'_E - \tilde{L}_3^{\mathrm{T}} \mathbf{q}_f, \tag{7}$$

where $f'_E = \tilde{L}_1^{\mathrm{T}} \mathbf{f}'_z$, $f'_N = \tilde{L}_2^{\mathrm{T}} \mathbf{f}'_z$ are the horizontal projections of the accelerometer measurements in the east and north directions, respectively, and $\tilde{L}_1, \tilde{L}_2$ are the first two columns of the transformation matrix $\tilde{L}_{zx}$.

Using the GNSS-derived vertical acceleration (4), the measurement of Δa_{up} can be formed as

$$y := a_{up}^{gps} - a'_{up} = \Delta a_{up} + e_a, \tag{8}$$

where e_a is the error of the GNSS-derived acceleration.

Substituting (7) and (6) into (8), we finally obtain the basic equation of airborne strapdown gravimetry in the form containing measurements and measurement errors:

$$y = -\delta g - k_E f'_N + k_N f'_E - \tilde{L}_3^{\mathrm{T}} \mathbf{q}_f + e_a. \tag{9}$$

Other systematic errors, such as GNSS antenna offsets or time-synchronization errors between the IMU and GNSS data [2], may also be included in (9). However, for simplicity, these errors are not considered here.

Equation (9) is considered over the flight time interval $[t_0, t_n]$. All measurements in (9) (y, f'_E, f'_N, and $\tilde{L}_3$) are assumed to be resampled at the GNSS data rate. Let t_i be a time stamp of GNSS data ($i = 0, \ldots, n$) and Δt the time step. The remaining variables in (9) (δg, k_E, k_N, $\mathbf{q}_f$, and e_a) are treated as unknown functions of time, for which a priori models are introduced below.

3. FORMULATION OF THE OPTIMAL ESTIMATION PROBLEM

Problem (9) is reduced to optimal stochastic estimation. A priori stochastic models are introduced for gravity disturbance δg, residual attitude errors k_E, k_N and GNSS vertical acceleration error e_a. The measurement errors of three accelerometers $\mathbf{q}_f$ are modeled as a modeled as stochastic processes, namely, as discrete-time white noises with zero mean and variance σ_f^2 (it is assumed here that all three accelerometers have the same accuracy).

3.1. Stochastic Models of Gravity Disturbance and IMU Systematic Errors

In airborne gravimetry, it is usually assumed that gravity (as a function of flight time) is a slowly-varying function (with the spectrum mostly concentrated at low frequencies) [13]. An a priori stochastic model of gravity as a stationary random process in time is typically introduced. The models commonly used in airborne gravimetry algorithms are Gauss–Markov models (typically of order two or three) [14, 15], integrals of a white noise [3, 4, 13], and Jordan's model [16]. Deterministic spatial models are also occasionally used in airborne gravimetry [17, 18]. A detailed comparison of all these models is beyond the scope of this paper, but a partial comparison can be found in [15] (Section 5.2) and [14, 15, 17, 18]. For instance, [15] notes that using the Gauss–Markov models of different orders yields similar results in airborne gravimetry as when using the integrals of a white noise.

In this work, gravity disturbance is modeled as the second integral of a white-noise process. The model takes into account the long-wavelength nature of gravity, agrees well with real gravimetric data in many areas [13] and is defined by a simple equation in the time domain: $\delta\ddot{g} = q_g$, where q_g is a white noise. The power spectral density (PSD) of the gravity model is:

$$S_g(\omega) = \frac{\sigma^2}{2\pi\omega^4}, \tag{10}$$

where ω is the angular frequency and σ^2 is the intensity of white noise.

Let us write the equations of the gravity model in discrete time denoting with the subscript i the value at the time moment t_i:

$$\begin{cases} \delta g_{i+1} = \delta g_i + \Delta t\, p_i, \\ p_{i+1} = p_i + q_{g,i}, \end{cases} \tag{11}$$

where $q_{g,i}$ is a discrete-time white noise with zero mean and variance σ_g^2.

Further, we introduce the stochastic models for the residual attitude errors k_E, k_N in east and north directions. Recall that the IMU attitude errors are estimated at the IMU/GNSS integration stage. The horizontal attitude errors contain the so-called Schueler oscillations [12] and in absolute value typically do not exceed 0.5 arcmin when using a state-of-the-art IMU [2]. The residual attitude errors k_E, k_N do not contain the Schueler oscillations, are typically less than 10 arcsec in absolute value and can be modeled as slowly varying functions of flight time [2, 4].

Based on the above, we introduce the models of k_E, k_N as integrals of white noise: $\dot{k}_E = q_E$, $\dot{k}_N = q_N$, or in discrete time as:

$$\begin{cases} k_{E,i+1} = k_{E,i} + q_{E,i}, \\ k_{N,i+1} = k_{N,i} + q_{N,i}, \end{cases} \tag{12}$$

where $q_{E,i}$, $q_{N,i}$ are discrete-time white noises with zero mean and variances σ_E^2, σ_N^2, respectively.

3.2. Refined Model of Kinematic Acceleration Error

Traditionally, in airborne gravimetry algorithms the error in GNSS-derived accelerations is assumed to be a white noise. In this work, a refined model of the acceleration error is introduced taking into account the specifics of the acceleration computation method. Namely, we assume that the kinematic accelerations are computed based on numerical differentiation of carrier phase measurements (by forming double differences of measurements using three successive epochs t_{i-1}, t_i, t_{i+1}) [10]. Taking this into account, we introduce the refined model of GNSS acceleration error in the following form:

$$e_{a,i} = \frac{q_{a,i+1} - 2q_{a,i} + q_{a,i-1}}{\Delta t^2}, \tag{13}$$

where $q_{a,i}$ is a discrete-time white noise with zero mean and variance σ_a^2.

The autocorrelation function of the process $e_{a,i}$ denoted by $K_e(m)$ (m is an integer) takes the following values:

$$K_e(0) = \frac{6\sigma_a^2}{\Delta t^4}, \quad K_e(\pm 1) = -\frac{4\sigma_a^2}{\Delta t^4}, \quad K_e(\pm 2) = \frac{\sigma_a^2}{\Delta t^4},$$

and zero for other m.

The PSD of the process $e_{a,i}$ is given by (assuming $\Delta t = 1$ for simplicity):

$$S_e(e^{j\omega}) = \frac{1}{2\pi} \sum_{m=-\infty}^{\infty} K_e(m)\, e^{-j\omega m} = \frac{2\sigma_a^2}{\pi}(1 - \cos\omega)^2, \tag{14}$$

where j is the imaginary unit.

Figure 1 shows the PSD of the refined error model (13)–(14) and PSD of real kinematic accelerations from a static test (the GNSS receiver from JAVAD with the sampling rate of 10 Hz was used). The recording was made while the aircraft was at the parking position at the airport (see

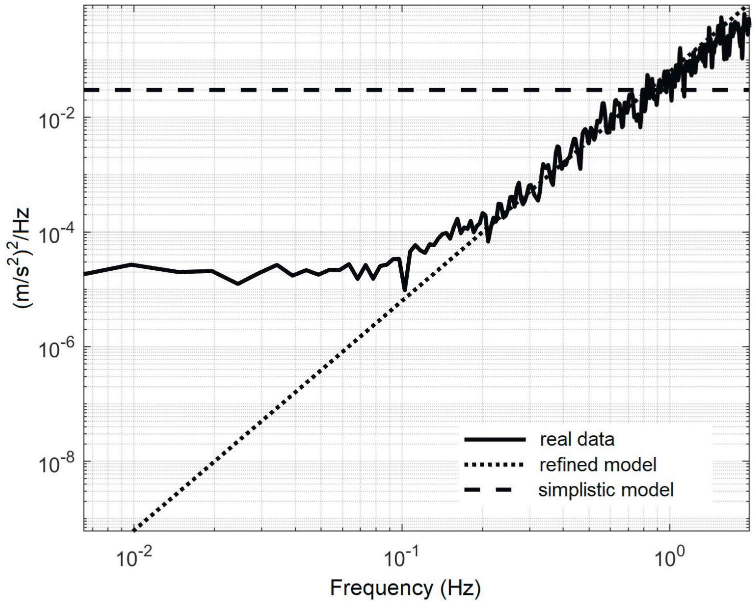

Fig. 1. Power spectral densities of GNSS carrier-phase acceleration errors (real data; solid line) and of theoretical models of GNSS acceleration errors: refined model (dotted line) and traditional simplistic model (dashed line), $(\mathrm{m}^2/\mathrm{s}^4)$/Hz.

Section 4 for details). The raw GNSS data processing (acceleration computation) was performed using the software developed by Lomonosov Moscow State University [10]. Since the PSD of the refined model (14) is proportional to ω^4 at low frequencies, its plot is shown in Fig. 1 as a straight line (in a logarithmic scale).

In Fig. 1, the PSD of the simplified GNSS acceleration error model (white noise), which is traditionally used in airborne gravimetry algorithms, is also shown. From Fig. 1, it follows that the refined model matches the real data significantly better, both in the high-frequency domain and near the cutoff frequency of the gravimetric filter [1], that is, in the range of 0.01–0.1 Hz, than the simplified model.

Further, a state-space representation of the refined model (13) will be required. Let us introduce auxiliary variables ξ_i, η_i:

$$\begin{cases} \eta_{i+1} = \xi_i, \\ \xi_{i+1} = q_{\xi,i}, \end{cases} \tag{15}$$

where $q_{\xi,i} := q_{a,i+1}$. Rewriting the model (13) using the auxiliary variables, we obtain:

$$e_{a,i} = \frac{\eta_i - 2\xi_i + q_{\xi,i}}{\Delta t^2}. \tag{16}$$

3.3. Problem Statement and Estimation Algorithm

Let us combine the basic equation of airborne gravimetry (9), the stochastic models of the gravity disturbance (11), the residual attitude errors (12) and the refined model of GNSS acceleration error (15)–(16) into one state-space system:

$$\begin{cases} k_{E,i+1} = k_{E,i} + q_{E,i}, \\ k_{N,i+1} = k_{N,i} + q_{N,i}, \\ \delta g_{i+1} = \delta g_i + \Delta t\, p_i, \\ p_{i+1} = p_i + q_{g,i}, \\ \eta_{i+1} = \xi_i, \\ \xi_{i+1} = q_{\xi,i}, \\ y_i = -\delta g_i - k_{E,i} f'_{N,i} + k_{N,i} f'_{E,i} + \dfrac{1}{\Delta t^2}(\eta_i - 2\xi_i + q_{\xi,i}) - \tilde{L}^{\mathrm{T}}_{3,i}\mathbf{q}_{f,i}. \end{cases} \tag{17}$$

The state-space system can be written in matrix form:

$$\begin{cases} \mathbf{x}_{i+1} = A_i\mathbf{x}_i + B_i\mathbf{q}_i, \\ y_i = C_i\mathbf{x}_i + r_i, \end{cases} \tag{18}$$

where the state vector $\mathbf{x}_i$ has the following form

$$\mathbf{x}_i = (k_{E,i}, k_{N,i}, \delta g_i, p_i, \eta_i, \xi_i)^{\mathrm{T}} \in \mathbb{R}^6. \tag{19}$$

The vector $\mathbf{q}_i$ is the system noise vector:

$$\mathbf{q}_i = (q_{E,i}, q_{N,i}, q_{g,i}, q_{\xi,i})^{\mathrm{T}} \in \mathbb{R}^4. \tag{20}$$

The scalar r_i is the measurement noise:

$$r_i = \frac{1}{\Delta t^2} q_{\xi,i} - \tilde{L}^{\mathrm{T}}_{3,i}\mathbf{q}_{f,i}. \tag{21}$$

The matrices A_i, B_i, C_i in (18) consist of the coefficients at the unknown variables and noises in (17) and have dimensions 6×6, 6×4 and 1×6, respectively.

The covariance matrix of the system noise vector (20) $E[\mathbf{q}_i\mathbf{q}_i^{\mathrm{T}}]$ is a diagonal 4×4-matrix whose diagonal elements are the variances of the components of $\mathbf{q}_i$. The variance of the measurement noise r_i (white noise) is easily computed from (21) and is given by the formula

$$E[r_i^2] = \frac{\sigma_\xi^2}{\Delta t^4} + \sigma_f^2,$$

where we use the fact that $\tilde{L}_{3,i}$ is a column of an orthogonal matrix.

The processes r_i and $\mathbf{q}_i$ are cross-correlated and their cross-covariance matrix is given by the expression:

$$E[r_i\mathbf{q}_i^{\mathrm{T}}] = \frac{1}{\Delta t^2}(0, 0, 0, \sigma_\xi^2).$$

Under the above assumptions, we now can formulate the optimal estimation problem (in the mean-square error sense) for the state vector $\mathbf{x}_i$ at each time instant given the state-space system (18) and measurements y_i, $i = 0, \ldots, n$. We assume that the estimate of the state vector at the initial time moment $\mathbf{x}_0$ is 0 and the initial covariance matrix $E[\mathbf{x}_0\mathbf{x}_0^{\mathrm{T}}]$ is given. The optimal estimation algorithm is the Kalman filter and smoothing [13].

3.4. Theoretical Analysis of Gravity Estimation Accuracy

Let us determine the gravity estimation accuracy when using the proposed approach (based on the refined model of kinematic acceleration error). For this, we derive an approximate expression for the transfer function of the optimal filter, which maps the measurement y (9) to the gravity disturbance estimate (the so-called gravimetric filter).

First, we reduce equation (9) to stationary form by neglecting the systematic errors k_E, k_N and accelerometer measurement noise $\mathbf{q}_f$. The gravity disturbance δg and GNSS acceleration error e_a are assumed here to be continuous-time stationary processes with given PSDs (10) and (14), respectively. Then the optimal (in the mean-square error sense) linear estimate of gravity is determined by a smoothing filter that is given in the frequency domain by the expression (the Wiener filter) [19]

$$W_1(\omega) = S_g(\omega)\,(S_g(\omega) + S_e(\omega))^{-1} = \left(1 + \left(\frac{\sigma_a}{\sigma_g}\right)^2 \omega^8\right)^{-1}. \tag{22}$$

In deriving (22), we used expression (14) in an approximate form as $\frac{\sigma_a^2}{2\pi}\omega^4$ for small ω.

Thus, the gravimetric filter based on the refined GNSS acceleration error model (14) approximately corresponds to the two-pass Butterworth filter of order 4 (Fig. 2a). In Fig. 2a, also shown is the gravimetric filter $W_2(\omega)$ constructed using the simplified GNSS acceleration error model (white noise with intensity σ_q^2) and the same model for gravity:

$$W_2(\omega) = \left(1 + \left(\frac{\sigma_q}{\sigma_g}\right)^2 \omega^4\right)^{-1}. \tag{23}$$

The transfer function (23) corresponds to the two-pass Butterworth filter of order 2.

Let us determine the accuracy of gravity estimation using the constructed filters as the PSD of the estimate error, provided that the true PSD of gravity coincides with the a priori model (10)

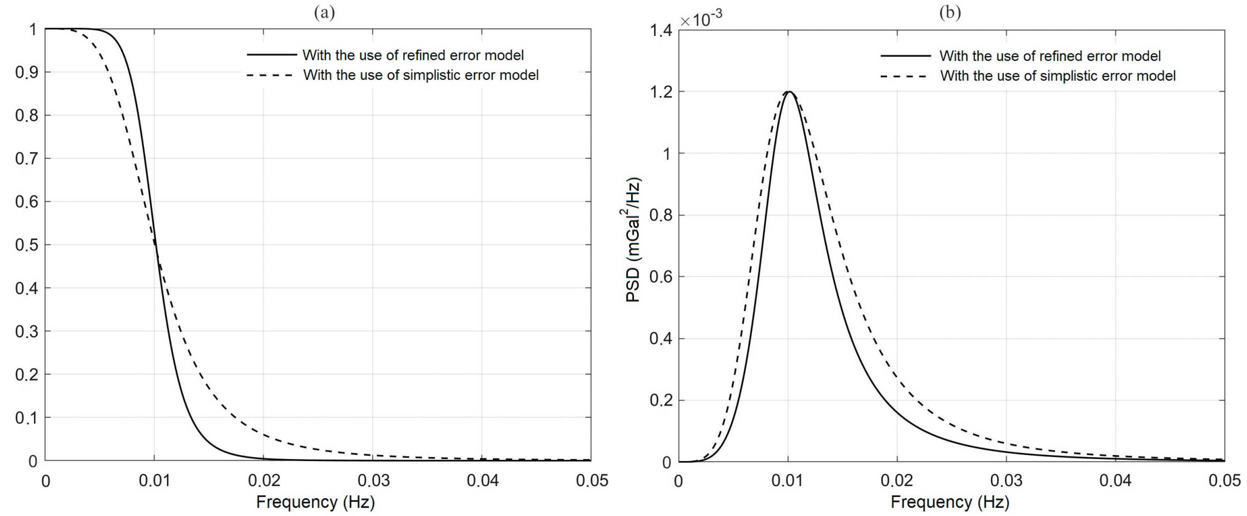

Fig. 2. (a) Transfer functions of the gravimetric filters in the new approach (based on the refined model of GNSS acceleration error) and in the common approach (based on a simplistic model of the GNSS acceleration error). (b) Power spectral densities of the errors in gravity estimates provided by the gravimetric filters, mGal^2/Hz (1 mGal $= 10^{-5}$ m/s^2).

and the true PSD of the GNSS acceleration error coincides with the model (14). Then the PSD of the error of the gravity estimate provided by the gravimetric filter $W_1(\omega)$ is determined as

$$S_{\delta g}(\omega) = S_g(\omega)S_e(\omega)\left(S_g(\omega) + S_e(\omega)\right)^{-1} = \frac{\sigma_a^2\omega^4}{2\pi}\left(1 + \left(\frac{\sigma_a}{\sigma_g}\right)^2\omega^8\right)^{-1}. \tag{24}$$

The PSD of the error of gravity estimate obtained using the gravimetric filter $W_2(\omega)$ is determined by:

$$\begin{aligned} S_{\delta g}(\omega) &= |1 - W_2(\omega)|^2 S_g(\omega) + |W_2(\omega)|^2 S_e(\omega) \\ &= \frac{\sigma_a^2\omega^4}{2\pi}\left(1 + \frac{\sigma_q^4}{\sigma_g^2\sigma_a^2}\right)\left(1 + \left(\frac{\sigma_q}{\sigma_g}\right)^2\omega^4\right)^{-2}. \end{aligned} \tag{25}$$

The plots of PSDs (24)–(25) are shown in Fig. 2b. The figure shows that when using the filter based on the refined GNSS acceleration error model, the PSD of the gravity estimation error is smaller than when using the filter based on the simplified acceleration error model.

4. NUMERICAL RESULTS

To test the developed gravity estimation algorithm (Section 3.3), we used data from a state-of-the-art strapdown airborne gravimeter (iCORUS by iMAR) recorded on December 17, 2022 during a flight of an airborne gravity campaign. The gravimeter was flown along ten repeated lines (i.e., above the same ground track) during this flight (Fig. 3). The repeated lines are oriented in the east-west and west-east directions. The length of each line is about 110 km. The flight was carried out using a Cessna 208B aircraft at a constant altitude of 760 m above the reference ellipsoid. The average aircraft speed along survey lines was 70 m/s. The total flight duration is 7 hours. The gravimetry campaign was conducted by Aerogeophysica JSC (Russia) in the Krasnoyarsk Krai.

The gravimeter data included raw measurements from the IMU inertial sensors (at the data rate of 400 Hz) and raw measurements from the onboard and ground-based GNSS (GPS) receivers

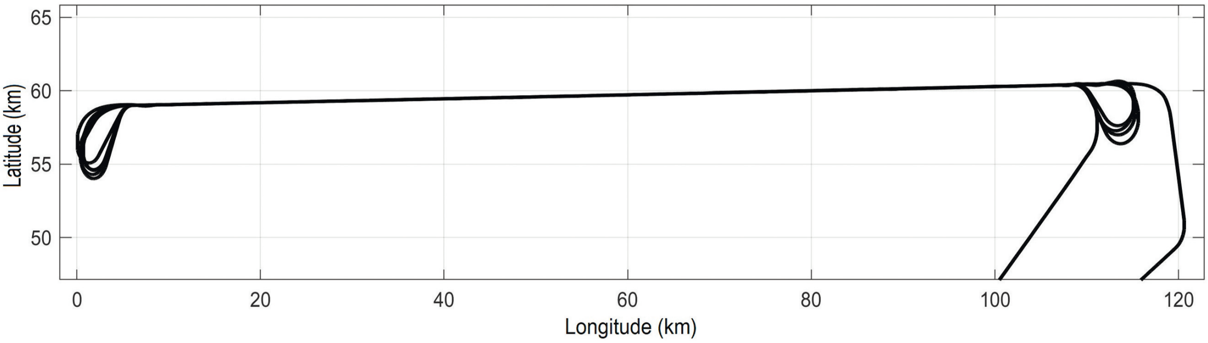

Fig. 3. Flight trajectory on the longitude-latitude plane (GNSS data).

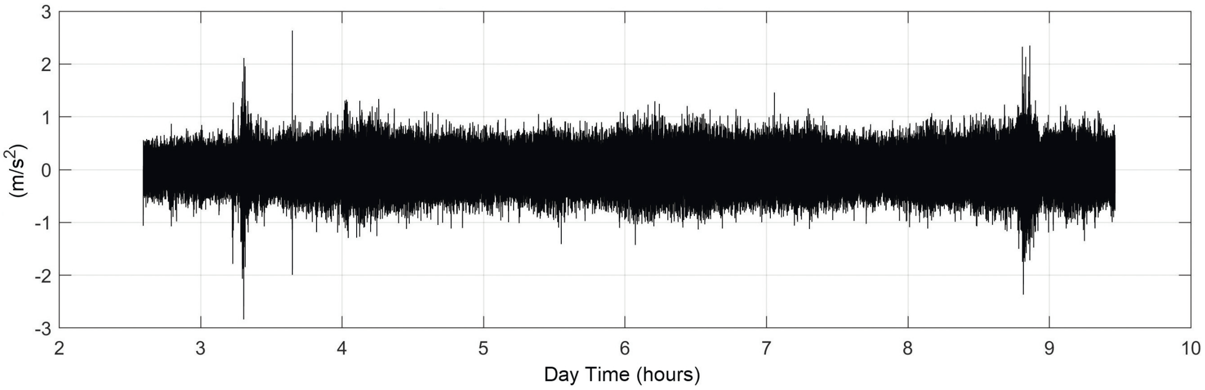

Fig. 4. Vertical kinematic accelerations from GNSS carrier-phase measurements, m/s^2.

from JAVAD (at 10 Hz). Preliminary stages of raw data postprocessing were performed using the software developed by the Faculty of Mechanics and Mathematics at Lomonosov Moscow State University [2] (software packages *INS-GNSS* and *IMU-GRAV* [20, 21]). Namely, the following tasks were solved at the initial stage:

1) computing the GNSS solutions (in the carrier phase differential mode), which included calculation of

- position (latitude, longitude and height above the reference ellipsoid) of the antenna of the onboard GNSS receiver;
- velocity (east, north and vertical components) of the onboard receiver;
- kinematic accelerations (east, north and vertical components) of the onboard receiver;

2) computing integrated IMU/GNSS solutions, which included estimation of

- IMU attitude angles (heading, roll and pitch);
- systematic errors of the IMU inertial sensors.

The vertical kinematic accelerations calculated from the carrier phase measurements using the algorithm from [10] are shown in Fig. 4.

The gravity estimate along the flight trajectory was computed using the proposed algorithm based on the refined GNSS acceleration error model. The gravity estimation accuracy was determined based on the statistics from ten repeated flight lines. The root-mean-square (RMS) value is 0.706 mGal. The gravity estimates at the repeated lines are shown in Fig. 5.

For comparison, another estimate of the gravity disturbance was obtained using the standard approach, which is based on the simplified GNSS acceleration error model (white noise). The same stochastic models were used for the residual attitude errors k_E, k_N, accelerometer measurement

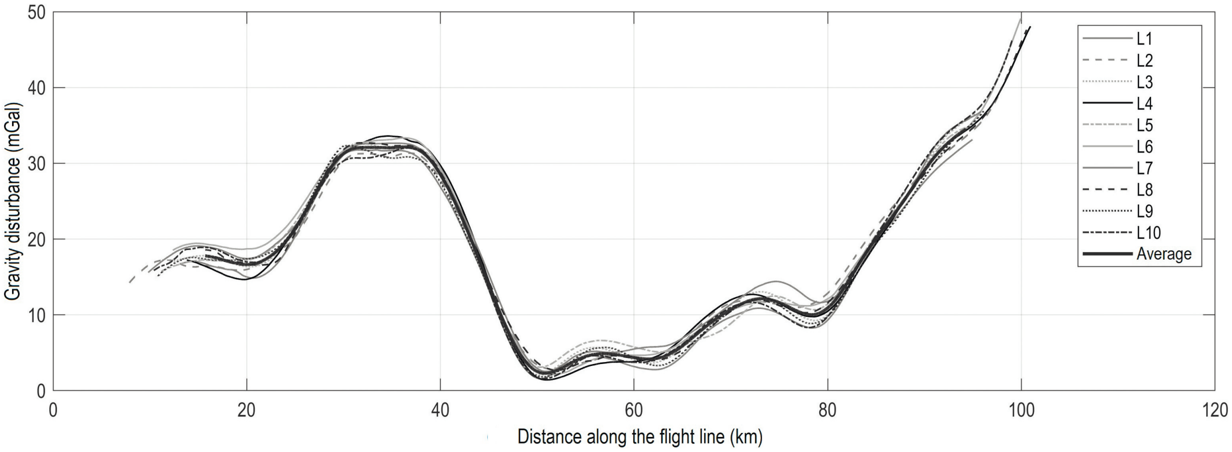

Fig. 5. Gravity disturbance estimates at the repeated lines provided by the new algorithm based on the refined model of GNSS acceleration error, mGal.

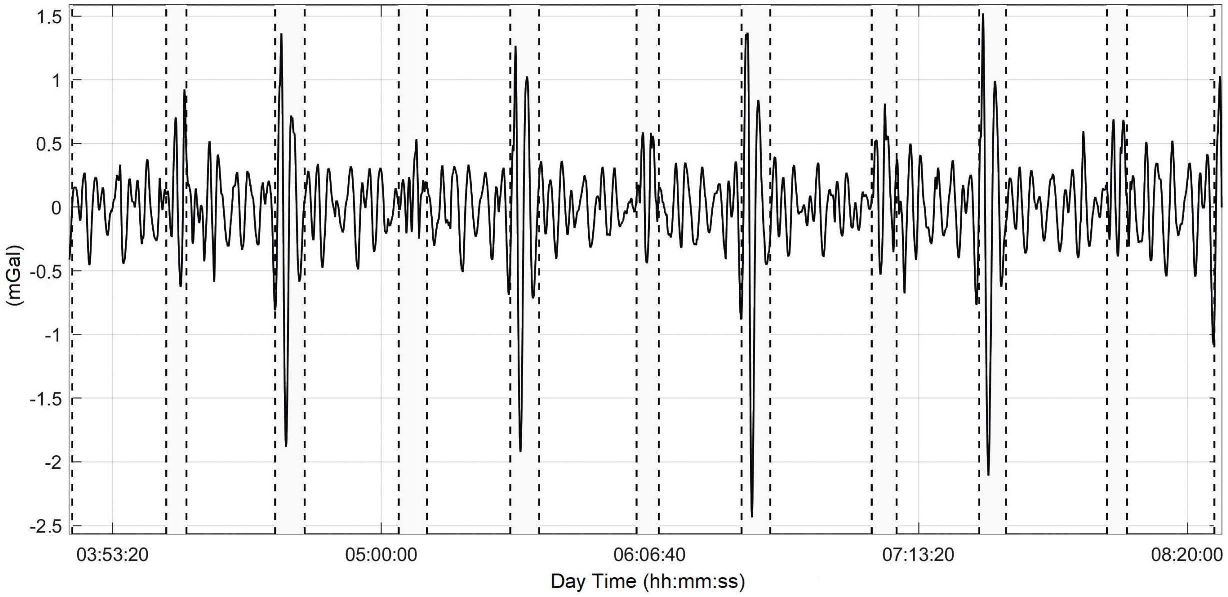

Fig. 6. Difference between the gravity disturbance estimates obtained in the new and standard approaches, mGal.

noise $\mathbf{q}_f$ and gravity disturbance in (9) as in the proposed approach. The estimation algorithm in the standard approach is the Kalman filter and smoothing. As shown above, the transfer function of the gravimetric filter in the standard approach is close to the 2nd-order two-pass Butterworth filter (Fig. 2a). The accuracy of the gravity estimates obtained in the standard approach was also determined based on the statistics from ten repeated lines and is 0.749 mGal (RMS). This demonstrates a slightly worse repeatability of the gravity estimates compared to the results from the proposed approach (based on the refined GNSS acceleration error model).

In Fig. 6, the difference between the gravity estimates computed by the proposed and standard approaches is shown. The standard deviation of the difference is 0.764 mGal, which is quite significant. We attribute this discrepancy to the gravity estimation error introduced by the standard approach as it showed worse repeatability of the gravity estimates at the repeated lines.

The absolute values of the difference between the gravity estimates provided by the two approaches reach 2.5 mGal, with maxima occurring during the aircraft turns (each lasting from 5 to 8 minutes) between the repeated lines (visible as spikes in Fig. 6). This can probably be explained by the fact that during aircraft turns, GNSS acceleration errors have a wider frequency range and are more effectively suppressed by the gravimetric filter of the proposed approach, which has a steeper roll-off near the cutoff frequency (Fig. 2a).

5. CONCLUSIONS

The gravity estimation algorithm based on the refined model of errors in the kinematic accelerations computed from the GNSS carrier phase measurements has been proposed. The refined error model takes into account the specificity of the method that was used for computing the kinematic accelerations and is defined in the time domain as the second central difference of a discrete-time white noise process.

The proposed approach was compared with the standard one, which uses a simplified GNSS acceleration error model (white noise). The approaches were compared based on processing raw data from a strapdown gravimetry flight (ten repeated lines). The results show a higher accuracy of the gravity estimates provided by the proposed approach, which is 0.71 mGal (RMS), while the standard approach showed the 0.75 mGal accuracy. The difference between the gravity estimates obtained by the two approaches reaches 2.5 mGal and is attributed to the estimation errors introduced by the standard approach.

Based on the obtained test results, we conclude that the proposed approach seems promising for implementing in postprocessing software packages of state-of-the-art strapdown airborne gravimeters.

REFERENCES

1. *Methods and Technologies for Measuring the Earth's Gravity Field Parameters*, Peshekhonov, V.G. and Stepanov, O.A. (eds.)., Springer, Cham, Switzerland, 2022.
2. Golovan, A.A. and Vyazmin, V.S., Methodology of Airborne Gravimetry Surveying and Strapdown Gravimeter Data Processing, *Gyroscopy Navig.*, 2023, vol. 14, pp. 36–47.
3. Ayres-Sampaio, D., Deurloo, R., Bos., M., et al., A Comparison Between Three IMUs for Strapdown Airborne Gravimetry, *Surv. Geophys.*, 2015, vol. 36, pp. 571–586.
4. Bolotin, Yu.V. and Golovan, A.A., Methods of inertial gravimetry, *Moscow Univer. Mech. Bull.*, 2013, vol. 68, pp. 117–125.
5. Lu, B., Barthelmes, F., Petrovic, S., et al., Airborne Gravimetry of GEOHALO Mission: Data Processing and Gravity Field Modeling, *J. Geoph. Res.*, 2017, vol. 122, pp. 586–604.
6. Li, M., Xu, T., Lu, B., et al., Multi-GNSS Precise Orbit Positioning for Airborne Gravimetry Over Antarctica, *GPS Solutions*, 2019, vol. 23, pp. 1–14.
7. He, K., Xu, T., Förste, C., et al., Integrated GNSS Doppler Velocity Determination for GEOHALO Airborne Gravimetry, *GPS Solutions*, 2021, vol. 25, pp. 1–12.
8. Jekeli, C. and Garcia, R., GPS Phase Accelerations for Moving-Base Vector Gravimetry, *J. Geod.*, 1997, vol. 71, pp. 630–639.
9. Bruton, A., Schwarz, K., Ferguson, S., et al., Deriving Acceleration from DGPS: Toward Higher Resolution Applications of Airborne Gravimetry, *GPS Solutions*, 2002, vol. 5, pp. 1–14.
10. Golovan, A.A. and Vavilova, N.B., Satellite Navigation. Raw Data Processing for Geophysical Applications, *J. Math. Sci.*, 2007, vol. 146, pp. 5920–5930.
11. Torge, W., *Gravimetry*, De Gruyter, Berlin, 1989.

12. Vavilova, N.B., Golovan, A.A., and Parusnikov, N.A., *Matematicheskie osnovy inertsial'nykh navigatsionnykh sistem* (Mathematical Foundations of Inertial Navigation Systems), Moscow: Moscow State University, 2020.

13. Forsberg, R., A New Covariance Model for Inertial Gravimetry and Gradiometry, *J. Geophys. Res.*, 1987, vol. 92, pp. 1305–1310.

14. Jekeli, C., Airborne Vector Gravimetry Using Precise, Position-Aided Inertial Measurement Units, *Bull. Géodésique*, 1994, vol. 69, pp. 1–11.

15. Becker, D., *Advanced Calibration Methods for Strapdown Airborne Gravimetry*, Ph.D. Thesis, Technische Universität Darmstadt, Darmstadt, Germany, 2016.

16. Stepanov, O.A., Koshaev, D.A., and Motorin, A.V., Identification of Gravity Anomaly Model Parameters in Airborne Gravimetry Problems Using Nonlinear Filtering Methods, *Gyroscopy Navig.*, 2015, vol. 6, pp. 318–323.

17. Vyazmin, V.S., New Algorithm for Gravity Vector Estimation from Airborne Data Using Spherical Scaling Functions, *International Association of Geodesy Symposia*, Springer, Berlin, Heidelberg, 2020, pp. 1–7.

18. Vyazmin, V.S., Bolotin, Yu.V., and Smirnov, A.O., Improving Gravity Estimation Accuracy for the GT-2A Airborne Gravimeter Using Spline-Based Gravity Models, *International Association of Geodesy Symposia*, Springer, Berlin, Heidelberg, 2020, pp. 1–8.

19. Kailath, T., Sayed, A.H., and Hassibi, B., *Linear Estimation*, Prentice Hall, Englewood Cliffs, 2000.

20. *Certificate of State Registration of Software Program no. 2023668582, Russian Federation*, Program for Calculating an Integrated Navigation Solution Based on Strapdown Airborne Gravimeter Data, Registered on 08.22.2023, Authors: Golovan, A.A. and Vyazmin, V.S.

21. *Certificate of State Registration of Software Program no. 2024680457, Russian Federation*, Program for Calculating Gravity Disturbance Estimate Based on Inertial and Satellite Navigation Data from a Strapdown Airborne Gravimeter, Registered on 08.08.2024, Author: Vyazmin, V.S.

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, 2025, Vol. 86, No. 6, pp. 577–588.

Informational Control of Strategies in an n-Player Oligopoly Game with Reflexive Behavior

M. I. Geraskin

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

Received November 29, 2024
Revised March 21, 2025
Accepted March 25, 2025

Abstract—This paper is devoted to an n-player oligopoly game with quantity competition under general demand and cost functions. Players are assumed to be reflexive: each player conjectures about the strategies of all other players. As a result, the subsets of players with different Stackelberg leadership levels are formed in this game (a game with multilevel leadership). The reflexion of players is formalized by conjectural variations, i.e., players' expectations regarding the impact of their actions on the counterparty's action. The problem of controlling the strategy of one player (the controlled player) by the other $(n-1)$ players (the Principal) is investigated, and an optimal Nash equilibrium is established in terms of the Principal's utility criterion. A hierarchical game model of players' interactions is proposed, and the dependence of the maximum of the Principal's utility function on the vector of the sums of conjectural variations (SCV) of all players is found within this model. The dependence is used to calculate the controlled player's SCV value optimizing the Principal's utility function. An informational control method is developed, enabling the Principal to induce the controlled player to choose the reaction function optimal from the former's standpoint.

Keywords: oligopoly, conjectural variation, Stackelberg leadership, hierarchical game

DOI: 10.31857/S0005117925060065

1. INTRODUCTION

The oligopoly game is an aggregative game [1], i.e., one where each player's payoff depends on the sum (aggregate) of the actions of all players. The solution of this game is the Cournot–Nash equilibrium [2, 3]. Based on a conjectural variation, H. Stackelberg was the first to define the leader's strategy in the game as opposed to the follower's strategy [4].

The prerequisite for the concept of a conjectural variation in an aggregative game was the comprehension that, when choosing their optimal actions, players will inevitably expect the optimal behavior of their rivals (i.e., they will perform reflexion). Consequently, the conjectural variation is a mathematical formalization of the mental process of reflexion [5], in this case being interpreted as a thought operation executed by some player to calculate the optimal reaction (best response) of another player to the former's action. As a rule, a quantity conjectural variation is considered, which characterizes the player's expected reciprocal change in the counterparty's action (supply quantity), optimizing the latter's utility function under the action chosen by the former. In modern research, the conjectural variation is widely used to analyze Stackelberg leadership in two directions as follows. First, an increase in the number of reflexive players leads to the emergence of multiple Stackelberg leaders in the game [6]. Second, deeper reflexion causes the emergence of higher-level leaders (multilevel leadership) [7]. The second aspect is expressed in the hierarchy of players' conjectures, bringing to the following hierarchy of their mental types: 1) a follower, who makes no conjectures regarding the strategies of the environment (its conjectural variation is therefore zero); 2) a (first-level) Stackelberg leader, who expects followers in the environment; 3) a second-level (or higher-level) Stackelberg leader, who expects first-level (or other lower-level) Stackelberg leaders in the environment. The hierarchy of mental types determines the reflexion rank of player r as the number in the described sequence of mental types. Note that the above hierarchy is constructed only in the players' beliefs (in this case, we have a game with phantom players); in fact, however, there is a nonhierarchical game with equal players, investigated in most studies of the oligopoly problem. As an exception, a hierarchical aggregative game with Principal's control was considered in [8] as an incentive problem with players (universities) institutionally dependent on the Principal (the government).

Given the above stratification of leaders, depending on the awareness of each player, players of different mental types can coexist in an oligopoly game, and a player of a definite mental type will choose a predictable action according to its conjectural variation. Therefore, it becomes possible to change, in a purposeful way, some player's action by forming a definite information field for him/her. This possibility leads to the well-known concept of informational control. The idea of informational control [9–11] is based on the formation of a purposeful sequence of opinions in a social group depending on the opinions of the so-called influence agents. Formally speaking, informational control is intended to induce purposefully the desired way of thinking, set by a control authority, for one or several players.

In the context of oligopoly games, the concept of informational control is constructed as follows. Consider a group consisting of $n-1$ players, further denoted by the symbol j. Let this group strive to achieve a favorable action of a non-group player i. To do so, the group performs actions from which player i concludes on some reflexion rank r of the group. Therefore, for player i, the optimal reflexion rank is $r+1$, which corresponds to definite values of its conjectural variations; in turn, they predetermine the desired action of this player for the group. For a particular realization of such a control process, it is necessary to find the sum of conjectural variations (SCV) of player i that are optimal (consistent) from the group's standpoint and then determine the dependences of the equilibrium actions of all players on the parameters of their mental type.

In this paper, we consider a procedure for calculating the optimal SCV value of a certain player in terms of the environment's utility criterion, a method for estimating the player's mental type corresponding to this SCV value or its reaction function, and an algorithm for calculating the group's actions inducing the required player's response.

2. THE BASIC OLIGOPOLY GAME MODEL

The game-theoretic model describes the interactions of n players representing firms in an oligopoly market. By a traditional assumption [6], these firms offer an identical product to the market with a common decreasing inverse demand function; in the case of quantity competition, they choose actions in the form of supply quantities. Players are rational, i.e., maximize individual action-concave utility functions $\pi_i(Q, Q_i) = P(Q)Q_i - C_i(Q_i)$; in addition, they are informed about the utility functions of the environment and choose their actions simultaneously, once, and independently. Then the basic model of player's action choice is described by

$$\max_{Q_i \geqslant 0} \pi_i(Q, Q_i) = \max_{Q_i \geqslant 0} \left[P(Q)Q_i - C_i(Q_i) \right], \quad i \in N = \{1, \ldots, n\}, \tag{1}$$

$$Q = \sum_{i \in N} Q_i, \tag{2}$$

where Q_i and π_i denote the action and utility function of player i; Q is the aggregate of actions (the total action of all players); N stands for the set of players; n is the number of players; $P(Q)$ is the inverse demand function, $P'_Q < 0$; finally, $C_i(Q_i)$ is the cost function of player i, $C'_{Q_i} > 0$.

For a known vector of conjectural variations, the Nash equilibrium in the game $\Gamma = \langle N, \{Q_i, i \in N\}, \{\pi_i, i \in N\}\rangle$ is determined by solving the following system of reaction equations:

$$\frac{\partial \pi_i(Q_i^*, \rho_{ij})}{\partial Q_i} = 0, \quad i, j \in N, \tag{3}$$

where $\rho_{ij} = Q'_{jQ_i}$ is the quantity conjectural variation of player i, i.e., its expectation regarding the supply quantity change of player j in response to the unit increase in the supply quantity of player i; Q_i^* is the equilibrium value.

The optimal conjectural variation (also called consistent in the literature) is calculated from equation (3) of player j, i.e., this variation corresponds to its best response. For the utility function (1), system (3) takes the form

$$P(Q) + (1 + S_i^r)Q_i P'_Q - C'_{iQ_i} = 0, \quad i \in N, \quad S_i^r = \sum_{j \in N\backslash i} \rho_{ij}^r, \tag{4}$$

where S_i^r is the SCV value of player i at a reflexion rank r.

In the case of action-independent conjectural variations ($\rho'_{ijQ_i} = 0$), SCV values at an arbitrary reflexion rank are given by the recurrent formula [7]

$$S_i^r = \left(\frac{1}{\sum_{j \in N\backslash i} \dfrac{1}{u_j - S_j^{r-1} + 1}} - 1 \right)^{-1}, \tag{5}$$

where $u_i = -1 + \frac{P'_{Q_i} + (1 + S_i^{r-1})Q_i P''_{QQ_i} - C''_{iQ_iQ_i}}{|P'_Q|}$ is a nonlinearity coefficient expressing the influence of the nonlinearity of the demand and cost functions on the unimodality of the utility function of player i.

Due to (4), the Nash equilibrium vector $\mathbf{Q}^* = \{Q_i^*, i \in N\}$ in the n-player oligopoly game depends on the SCV vector $\mathbf{S^r} = \{S_i^r, i \in N\}$ (the conjectural variations of all players). Therefore, an inverse dependence also exists: given a known action vector $\mathbf{Q}^* = \{Q_i^*, i \in N\}$, it is possible to establish a vector $\mathbf{S^r} = \{S_i^r, i \in N\}$ inducing these actions of the players. On this basis, let us consider some tools for controlling (manipulating) the player's behavior by the environment.

3. AN OPTIMAL CONTROL MODEL FOR PLAYER'S BEHAVIOR

Consider the following modification of the basic oligopoly game model in the form of a hierarchical game. Player i is the controlled object, and its environment (i.e., the other players) acts as the control subject, also called the Principal. Thus, we study a hierarchical game of the Principal–agent type (Fig. 1). For the sake of simplicity, the environment of player i will be assigned number j (i.e., $j = \{N\backslash i\}$). The environment has a common goal: induce player i to choose an optimal action $\overline{Q}_i$ in terms of the former's utility functions. Therefore, let us define the Principal's goal function as the vector of the utility functions of the environment players. Since the latter are supposed to be identical, the goal function can be represented as a single function of the form

$$\pi^{(i)} = \pi_j, \quad j \in \{1, \ldots, i-1, i+1, \ldots, n\},$$

and briefly written as

$$\pi = \pi^{(i)}.$$

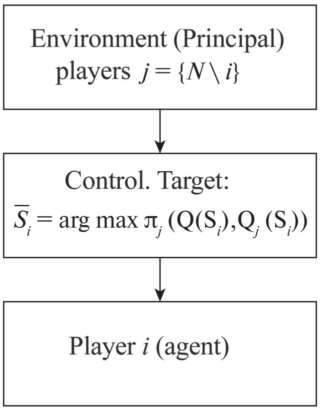
Environment (Principal)
players $j = \{N \setminus i\}$
Control. Target:
$\overline{S}_i$ = arg max π_j (Q(S$_i$),Q$_j$ (S$_i$))
Player i (agent)

Fig. 1. The diagram of a hierarchical game.

Now we describe the main assumptions adopted in the hierarchical game analysis.

1) The utility functions of all environment players are identical, i.e., these players have the same cost functions with the same coefficient:

$$C_j(Q_j) = C_k(Q_k), \quad \pi_j(Q_j) = \pi_k(Q_k)\forall j, \quad k = \{N\backslash i\}.$$

2) The awareness of the players is described by the following awareness sets I_i and I_j:

— The awareness set of the controlled player includes the set of players and the actions and utility functions of all players:

$$I_i = \{N, Q_k, \pi_k, k \in N\}.$$

— The awareness set of the environment encompasses the set of players, the actions and utility functions of all players, the SCV values of the environment players, and the Principal's goal function:

$$I_j = \{N, Q_k, \pi_k, S_j, \pi, j \in N\backslash i, k \in N\},$$

where $\pi = \pi_j$, $j \in N\backslash i$, is the goal function of the environment (Principal), identical to the utility functions of the environment players.

3) All players calculate their conjectural variations based on the sets I_i and I_j. These variations may be optimal (i.e., consistent with the utility functions of the players) or be determined by the players using the actions of the other players. If players' actions are inconsistent with their utility functions, the players are action-oriented. At each time instant of the game, either the SCV of the controlled player or the SCV of the environment has a constant value:

$$S_i = \text{const} \vee S_j = \text{const},$$

since to change the SCV, players estimate the actions of the other players at the previous time instant.

The environment controls the behavior of player i through its reflexion in the following procedure:

— The environment calculates the target SCV value $\overline{S}_i$ of player i that is optimal in terms of the former's utility functions:

$$\overline{S}_i = \arg\max_{S_i} \pi_j(Q(S_i), Q_j(S_i)).$$

— The environment performs actions $\overline{Q}_j$ that will induce player i to choose the SCV value $\overline{S}_i$ and, consequently, the optimal action $\overline{Q}_i$ from the environment's standpoint.

The second stage of the above behavior control procedure explains the meaning of Assumption 3 in the context of the players' dual approach to estimating the conjectural variations. The environment controls player i by performing the actions $\overline{Q}_j$ determined not from the maximum of its utility function but from the condition of inducing this player to choose $\overline{S}_i$; therefore, having predicted the SCV values from the environment's utility functions, player i will arrive at a contradiction. Consequently, the player in this contradiction will favor the method for estimating the SCV values by the environment's actions. Therefore, two alternatives are considered when estimating the SCV values: if the players' actions are consistent with their utility functions, the other players will estimate the optimal SCV values; if the SCV values found by the players' actions do not coincide with the SCV values based on the utility functions, the first estimation method as more realistic will be given priority.

4. METHODS FOR CALCULATING OPTIMAL CONTROL

The behavioral control of player i is based on the dependence of each player's utility function on the SCV values of all players, see system (4). Therefore, we first derive a formula for the maximum of the environment's utility function depending on the SCV vector of all players.

Proposition 1. *The maximum value of the environment's utility function is given by*

$$\pi_j^*(\mathbf{S}^r) = P\left[Q^*(\mathbf{S}^r)\right] Q_j^*(\mathbf{S}^r) - \int\limits_0^{Q_j^*(\mathbf{S}^r)} \left[P(Q^*) + (1+S_j^r)Q_j^* P'_Q\right] dQ_j + C_j(0), \tag{6}$$

where $\mathbf{S}^r = \{S_k^r, k \in N\}$ *denotes the SCV vector.*

Proof of Proposition 1. Let us express the marginal cost from equation (4), written for the environment:

$$C'_{jQ_j} = P(Q^*) + (1+S_j^r)Q_j^* P'_Q.$$

Integration over Q_j yields

$$C_j(Q_j^*) = \int\limits_0^{Q_j^*} C'_{jQ_j} dQ_j + C_j(0) = \int\limits_0^{Q_j^*} \left[P(Q^*) + (1+S_j^r)Q_j^* P'_Q\right] dQ_j + C_j(0);$$

after substitution into the environment's utility function, we obtain

$$\pi_j^* = P(Q^*)Q_j^* - C_j(Q_j^*) = P(Q^*)Q_j^* - \int\limits_0^{Q_j^*} \left[P(Q^*) + (1+S_j^r)Q_j^* P'_Q\right] dQ_j + C_j(0),$$

where $C_j(0)$ is fixed costs. Note that in this formula, the equilibrium action of player i, Q_j^*, the equilibrium price $P(Q^*)$, and the equilibrium aggregate action Q^* are all functions of the SCV $\mathbf{S}^r = \{S_i^r, i \in N\}$. Thus, the maximum utility of the environment also depends on this vector, which finally implies (6). ∎

Let us express the SCV of player i that is optimal in terms of the environment's utility criterion:

$$\overline{S}_i = \arg\max_{S_i} \pi_j^*(\mathbf{S}^r).$$

Proposition 2. *The optimal SCV value* $\overline{S}_i$ *of the controlled player, in terms of the environment's utility function, is calculated from the equation*

$$2(1+S_j^r)Q_j^* Q_{jS_i}^{*'} P'_Q + \left((1+S_j^r)P''_{QS_i} + P'_Q \frac{\partial S_j^r}{\partial S_i}\right) Q_j^{*2} = 0 \tag{7}$$

under the condition

$$2\frac{\partial S_j^r}{\partial S_i} Q^{*'}_{jS_i} + (1+S_j^r) Q^{*'}_{jS_i} < 0, \tag{7a}$$

in the case of a weak impact of the SCV change on the equilibrium shift and a relatively small value of the second derivative of the environment's SCV with respect to the player's SCV compared to the first derivative.

Proof of Proposition 2.

With the Leibniz integral rule (differentiation under the integral sign) applied to the second term in (6), where the integration limits are functions of the parameter S_i, we write the necessary first-order maximum condition for the function (6):

$$\pi^{*'}_{jS_i} = P'_{S_i} Q_j^* + P Q^{*'}_{jS_i} - \left\{ \Big[P + (1+S_j^r) Q_j^* P'_Q\Big] Q^{*'}_{jS_i} + \int\limits_0^{Q_j^*} \Big[P(Q^*) + (1+S_j^r) Q_j^* P'_Q\Big]'_{S_i} dQ_j \right\} = 0. \tag{7b}$$

The integral in this equation can be transformed as follows:

$$I = \int\limits_0^{Q_j^*} \Big[P + (1+S_j^r) Q_j^* P'_Q\Big]'_{S_i} dQ_j = \int\limits_0^{Q_j^*} \left[P'_{S_i} + (1+S_j^r)(Q^{*'}_{jS_i} P'_Q + Q_j^* P''_{QS_i}) + Q_j^* P'_Q \frac{\partial S_j^r}{\partial S_i}\right] dQ_j.$$

In the integrand, the parameters $Q_j^*, Q^*, Q^{*'}_{jS_i}, P'_{S_i}(Q^*)$, and $P'_Q(Q^*)$ characterize the equilibrium of all players, therefore being independent of the action Q_j of player j; the parameter S_j^r and hence $\frac{\partial S_j^r}{\partial S_i}$ weakly depend on the action Q_j (see the proof in [7]). Therefore, the following variables are considered to be independent of Q_j:

$$Q_j^*, Q^*, P'_{S_i}, P'_Q, \frac{\partial S_j^r}{\partial S_i}, Q^{*'}_{jS_i}, S_j^r.$$

In this case, the integral becomes

$$I = (P'_{S_i} + (1+S_j^r) Q^{*'}_{jS_i} P'_Q) Q_j^* + \left((1+S_j^r) P''_{QS_i} + P'_Q \frac{\partial S_j^r}{\partial S_i}\right) Q_j^{*2}.$$

Substituting it into (7a) gives the expression

$$\pi^{*'}_{jS_i} = P'_{S_i} Q_j^* + P Q^{*'}_{jS_i} - \Big[P + (1+S_j^r) Q_j^* P'_Q\Big] Q^{*'}_{jS_i} - (P'_{S_i} + (1+S_j^r) Q^{*'}_{jS_i} P'_Q) Q_j^*$$
$$- \left((1+S_j^r) P''_{QS_i} + P'_Q \frac{\partial S_j^r}{\partial S_i}\right) Q_j^{*2} = -2(1+S_j^r) Q_j^* Q^{*'}_{jS_i} P'_Q - \left((1+S_j^r) P''_{QS_i} + P'_Q \frac{\partial S_j^r}{\partial S_i}\right) Q_j^{*2}.$$

Then equation (7b), used to calculate the optimal SCV value of player i in terms of the environment's utility criterion, takes the form (7).

The second-order maximum condition for the function (6) is given by

$$\pi^{*''}_{jS_iS_i} = -\left\{ 2\frac{\partial S_j^r}{\partial S_i} Q_j^* Q^{*'}_{jS_i} P'_Q + 2(1+S_j^r)\Big[Q_j^* Q^{*''}_{jS_iS_i} P'_Q + Q_j^* (Q^{*''}_{jS_iS_i} P'_Q + Q^{*'}_{jS_i} P''_{QS_i})\Big] \right.$$
$$+ \frac{\partial S_j^r}{\partial S_i} P''_{QS_i} Q_j^{*2} + (1+S_j^r)(P'''_{QS_iS_i} Q_j^{*2} + 2Q_j^* Q^{*'}_{jS_i} P''_{QS_i})$$
$$\left. + P''_{QS_i} \frac{\partial S_j^r}{\partial S_i} Q_j^{*2} + P'_Q \left(\frac{\partial^2 S_j^r}{\partial S_i^2} Q_j^{*2} + 2Q_j^* Q^{*'}_{jS_i} \frac{\partial S_j^r}{\partial S_i}\right) \right\} < 0.$$

After straightforward transformations, we obtain

$$\frac{\partial S_j^r}{\partial S_i} Q_j^* (2P''_{QS_i} Q_j^* + 4Q^{*'}_{jS_i} P'_Q)$$

$$+ (1+S_j^r)\left(2(Q^{*'}_{jS_i})^2 P'_Q + 2Q_j^* Q^{*''}_{jS_iS_i} P'_Q + P'''_{QS_iS_i} Q_j^{*2} + 4Q_j^* Q^{*'}_{jS_i} P''_{QS_i}\right) + P'_Q \frac{\partial^2 S_j^r}{\partial S_i^2} Q_j^{*2} > 0.$$

Due to the assumption of a weak influence of the SCV change on the equilibrium shift,

$$P''_{QS_i} = 0, \quad P'''_{QS_iS_i} = 0, \quad Q^{*''}_{jS_iS_i} = 0.$$

Due to the assumption of a small value of the second derivative of the environment's SCV with respect to the player's SCV compared to the first derivative,

$$\left|\frac{\partial^2 S_j^r}{\partial S_i^2}\right| \ll \left|\frac{\partial S_j^r}{\partial S_i}\right| \Rightarrow \frac{\partial^2 S_j^r}{\partial S_i^2} \approx 0.$$

Under these assumptions, the above condition becomes

$$4\frac{\partial S_j^r}{\partial S_i} Q_j^* Q^{*'}_{jS_i} P'_Q + 2(1+S_j^r)(Q^{*'}_{jS_i})^2 P'_Q > 0.$$

Since $P'_Q < 0$ by the inverse demand function property and $Q^{*'}_{jS_i} > 0$ by the Stackelberg leadership property, we finally arrive at a sufficient maximum condition for the solution (7) in the form (7a). ■

Let us present a methodology for calculating the derivatives $Q^{*'}_{jS_i}$ in equation (7).

Proposition 3. *The derivatives $Q^{*'}_{jS_i}$ are the roots of the following system of linear equations:*

$$\sum_{k\in N} a_{jk} Q^{*'}_{kS_i} = b_j, \quad j \in N, \tag{8}$$

where $b_j = -\left(\dfrac{\partial S_j^r}{\partial S_i} Q_j^* P'_Q + (1+S_j^r) Q_j^* P''_{QS_i} C''_{jQ_jS_i}\right),$

$$a_{jk} = \begin{cases} \gamma_{jk} + P'_Q & \textit{for } j \neq k \\ \gamma_{jk} + P'_Q + (1+S_j^r)P'_Q & \textit{for } j = k, \end{cases}$$

$$\gamma_{jk} = P'_{Q_k}(Q^*) + (1+S_j^r)\left\{Q_j^* P''_{QQ_k} + P'_Q Q'_{jQ_jQ_k}\right\} - C''_{jQ_jQ_k}.$$

Proof of Proposition 3.

Assuming that the optimal actions of all environment players (system (4)) depend on S_i, we consider the n implicit functions

$$F_j(Q^*, S_i) = P(Q^*) + (1+S_j^r) Q_j^* P'_Q - C'_{jQ_j} = 0, \quad j \in N.$$

In this case, the derivatives $Q^{*'}_{jS_i}$ of the implicit functions with several independent variables are calculated from the following system [12]:

$$\sum_{k\in N} \frac{\partial F_j}{\partial Q_k}\frac{\partial Q_k}{\partial S_i} + \frac{\partial F_j}{\partial S_i} = 0, \quad j \in N, \tag{8a}$$

$$\text{where}\quad \frac{\partial F_j}{\partial Q_k} = P'_Q(Q^*) + (1+S_j^r)\left\{Q_j^* P''_{QQ_k} + P'_Q Q^{*'}_{jQ_k}\right\} - C''_{jQ_jQ_k} = \gamma_{jk},$$

$$\frac{\partial F_j}{\partial S_i} = P'_Q(Q^*)Q^{*'}_{S_i} + \frac{\partial S_j^r}{\partial S_i} Q_j^* P'_Q$$

$$+ (1+S_j^r)\left\{Q_j^* P''_{QS_i} + P'_Q Q^{*'}_{jS_i}\right\} - C_{jQ_jS_i}, \quad Q^{*'}_{S_i} = \sum_{k\in N} Q^{*'}_{S_i}.$$

Here, γ_{jk} stands for the component without an explicit dependence of the desired parameters $Q^{*'}_{jS_i}$, further denoted by $x_j = Q^{*'}_{jS_i}$. In this case, system (8a) has the form

$$\sum_{k\in N} \gamma_{jk} x_k + P'_Q \sum_{k\in N} x_k + (1+S_j^r)P'_Q x_j + \frac{\partial S_j^r}{\partial S_i} Q_j^* P'_Q + (1+S_j^r)Q_j^* P''_{QS_i} - C''_{jQ_jS_i} = 0.$$

With $b_j = -\left(\frac{\partial S_j^r}{\partial S_i} Q_j^* P'_Q + (1+S_j^r)Q_j^* P''_{QS_i} - C''_{jQ_jS_i}\right)$ and

$$a_{jk} = \begin{cases} \gamma_{jk} + P'_Q & \text{for } j \neq k \\ \gamma_{jk} + P'_Q + (1+S_j^r)P'_Q & \text{for } j = k, \end{cases}$$

we write the following system of linear algebraic equations for the unknowns x_k: $\sum_{k\in N} a_{jk}x_k = b_j$, $j \in N$, which matches (9). ■

System (9) allows determining the derivatives $Q^{*'}_{jS_i}$ as functions of the SCV values S_j^r of all players, including the desired value $\overline{S}_i$. Thus, we have provided a method for calculating the target SCV value of the controlled player: solve equation (7) considering the derivatives $Q^{*'}_{jS_i}$ expressed through S_i from the solution of system (8).

5. A MECHANISM FOR CALCULATING OPTIMAL CONTROL

Consider a possible method for the environment to induce the controlled player to choose the target SCV value $\overline{S}_i$. As an illustration, we will interpret the considerations by the example of duopoly. Let us start with the description of the classical principle of Stackelberg leader emergence in the game of initially equal participants (Fig. 2), i.e., from the situation of Cournot responses. (Here, the equilibrium and Cournot responses are indicated by the symbol K.) As is known [13], the optimal reaction functions of the players in the linear Cournot duopoly model have the form

$$Q_1 = \frac{\alpha_1 - Q_2}{2}, \quad Q_2 = \frac{\alpha_2 - Q_1}{2},$$

where $\alpha_1 = \frac{a-B_i}{b}$, with a and b representing the maximum price and the rate of price reduction in the inverse demand function, respectively, and B_i specifying the marginal cost of player i. However, if the reaction functions were unknown to the players, they could reconstruct these functions from observations of each other's actions. When treated as a potential leader, the first player observes in the game dynamics the response of the second player: the second player takes the action Q_2^t in response to the action M_1 and the action $Q_2^{(t+1)}$ in response to the first player's reciprocal action Q_1^t. Based on these observations, the first player (with the reaction function R_1^K) determines the second player's reaction function R_2^K and calculates from it the conjectural variation (equal to the SCV in the duopoly) as follows:

$$S_1 = Q'_{2Q_1} = -\frac{1}{2}.$$

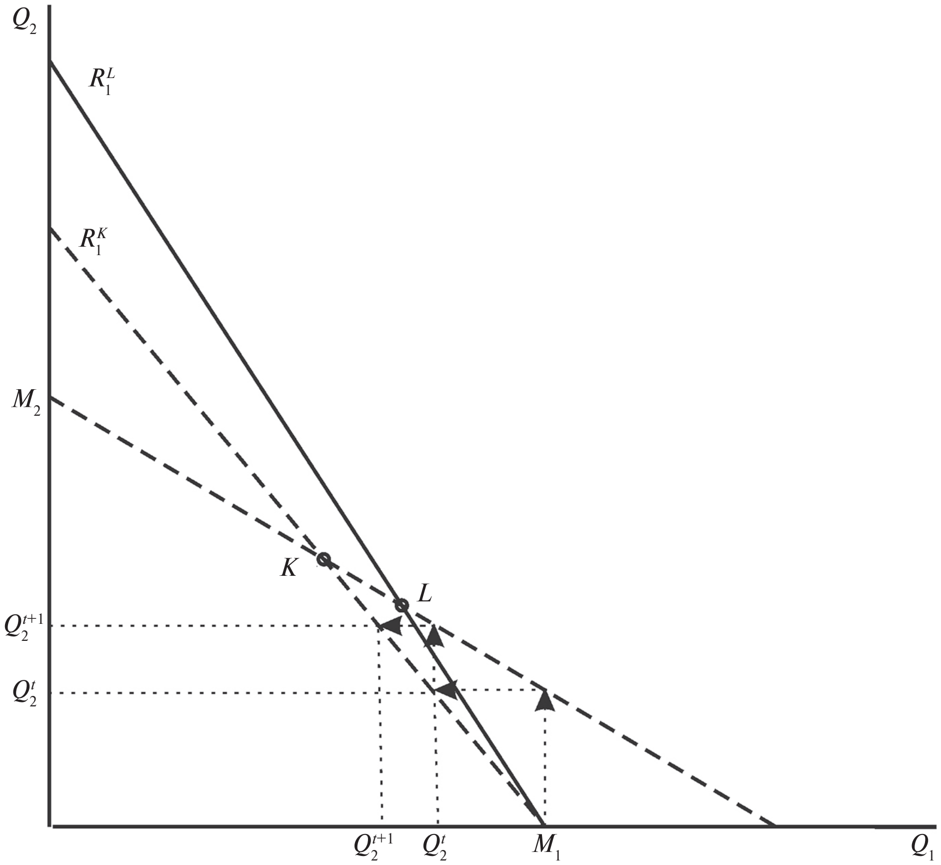

Fig. 2. The emergence of a Stackelberg leader: an illustration.

As a result, the reaction function of the first player is transformed to

$$Q_1 = \frac{\alpha_1 - Q_2}{2 + S_2} = \frac{\alpha_1 - Q_2}{2 - \frac{1}{2}};$$

and this player becomes a Stackelberg leader (in Fig. 2, its response and the corresponding equilibrium are indicated by the symbol L). In other words, observing the response $Q_2 = \frac{\alpha_2 - Q_1}{2+0}$ of the second player, the first player has revised its SCV: the SCV $S_2 = 0$ of the second player has induced the first player to set the SCV value $S_1 = -\frac{1}{2}$.

Extending this procedure to multilevel leadership, we can formulate the following law: for a certain player to change its conjectural variation to some given value corresponding to a definite-level Stackelberg leader, this player must observe another player's action corresponding to the response of a previous-level Stackelberg leader. Consequently, the other player must create the so-called phantom agent acting not according to its true reaction function, so this response will be called phantom and denoted by the symbol f. Formally, this means that to induce player i to set the SCV value $\overline{S}_i$, the environment must act according to the phantom reaction function

$$Q_j^f = \frac{\alpha_j - Q_i}{2 + S_j^f}$$

under the condition

$$S_i = Q_{jQ_i}^{f'} = -\frac{1}{2 + S_j^f} = \overline{S}_i.$$

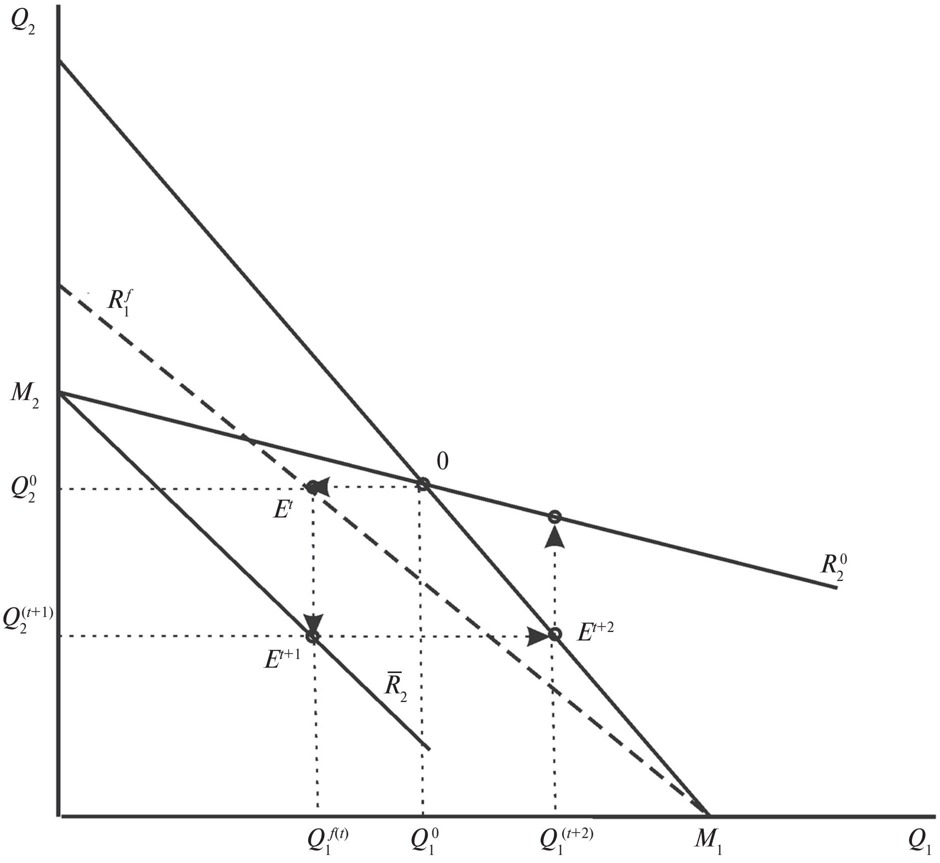

Fig. 3. The informational control process: an illustration.

Hence, the SCV of the environment for this action is given by

$$S_j^f = -\frac{1}{\overline{S}_i} - 2.$$

This general principle was proved earlier [7] as formula (5). Based on the latter, we provide a method for calculating the phantom response in the general case of nonlinear cost functions, when the reaction functions cannot be expressed explicitly. Let us summarize the considerations as follows.

Proposition 4. *The environment's phantom reaction function inducing the controlled player to set the target SCV value $\overline{S}_i$ corresponds to the environment's SCV value S_j^f calculated by solving the equation*

$$\overline{S}_i = \left(\frac{1}{\sum_{j \in N \setminus i} \dfrac{1}{u_j - S_j^f + 1}} - 1 \right)^{-1}. \tag{9}$$

Based on this principle, we describe the informational control process in the above duopoly example, assuming that the target SCV value of the controlled (second) player is $\overline{S}_2 = -\frac{3}{4}$. In other words, the first player strives to make the second player's response match a third-level Stackelberg leader. (Recall that in the linear duopoly, the leaders of the first, second, and third levels have the SCV values $-\frac{1}{2}, -\frac{2}{3}$, and $-\frac{3}{4}$, respectively.) We assign the number "0" to the initial equilibrium state, i.e., the equilibrium actions are Q_1^0 and Q_2^0, the SCV values of the players are S_1^0 and S_2^0, and the reaction functions are R_1^0 and R_2^0. The control process is illustrated in Fig. 3.

At the time instant t, the first player calculates its action using the phantom reaction function R_1^f of the second-level leader: $Q_1^{f(t)} = \frac{\alpha_1 - Q_2^0}{2-\frac{2}{3}}$. Therefore, the game state at this instant (the point E^t) is described by the action vector $(Q_1^{f(t)}, Q_2^0)$.

At the time instant $t+1$, the second player calculates $\overline{S}_2 = -\frac{3}{4}$ using this action and passes to the target reaction function $\overline{R}_2$, given by the equation $Q_2^{t+1} = \frac{\alpha_2 - Q_1^{f(t)}}{2-\frac{3}{4}}$. With this reaction function, it responds to the action $Q_1^{f(t)}$ by the action $Q_2^{t+1} = \frac{\alpha_2 - Q_1^{f(t)}}{2-\frac{3}{4}}$. At this time instant, the game state is denoted by the point E^{t+1}.

At the time instant $t+2$, the first player performs an action according to its true reaction function $Q_1^{t+2} = \frac{\alpha_1 - Q_2^{t+1}}{2+S_2^0}$ (maximizes its utility for the initial equilibrium). For the combination (Q_1^{t+2}, Q_2^{t+1}), the game state is denoted by the point E^{t+2}.

At the subsequent time instants of the game, the initial equilibrium is restored according to the above procedure (see Fig. 2). With this procedure, the first player gains an additional utility at the time instants $t+1$ and $t+2$ since the second player performs actions according to the SCV target value $\overline{S}_2$.

The game state returns to the initial equilibrium in an infinite number of steps. Therefore, the Principal's control efficiency can be assessed by the following condition:

$$\sum_{\tau=t}^{\infty} \pi^{*(\tau)} e^{-\rho\tau} - \pi^{*(0)} \sum_{\tau=t}^{\infty} e^{-\rho\tau} \geqslant 0,$$

where $\pi^{*(0)}$ is the Principal's maximum utility at the initial equilibrium; $\pi^{*(\tau)}$ is the Principal's maximum utility at a time instant τ; finally, ρ is the discount rate.

6. CONCLUSIONS

This paper has developed an informational control method for the actions of a given player in an oligopoly game model: other players perform a control action inducing this player to make an optimal response from the environment's standpoint. The foundations of this informational control are, first, the dependence of players' actions on their conjectures regarding the expected actions of counterparties and, second, the a priori unawareness of players regarding each other's conjectures due to the dual nature of their conjectural variations. On the one hand, the variations are based on the analysis of the utility functions of the environment; on the other, a player cannot ignore the nature of the responses of its environment. Therefore, the following hypothesis has been adopted above: in the case of contradiction between these two approaches, the players estimate the conjectural variations by each other's actions, which are more reliable information. Under this hypothesis, without changing its conjectural variations, the environment can perform an action as if on behalf of a phantom player that induces the controlled player to respond in a way favorable to the environment, and the latter interprets this action as a signal of a change in the environment's true reaction and performs the desired action.

The main results of this study can be summarized as follows. A hierarchical game model of players' interactions in an oligopoly has been presented, where the environment is treated as the Principal and some player as a controlled object. An explicit expression has been derived for the maximum of the environment's utility function depending on the SCV vector of all players; this expression allows finding the controlled player's SCV value optimizing the environment's utility function. A methodology for calculating the target SCV value of the controlled player from the environment's standpoint has been defined. An iterative procedure has been developed to induce the controlled player to choose a reaction function corresponding to the target SCV value; as a result, the environment maximizes its utility.

The optimal control problem has been formulated for an n-player game with general utility functions. Therefore, it is impossible to obtain explicit solutions to analyze the game results in the developed techniques and procedures. Hence, the next stage of research is to apply these general tools to particular utility functions and carry out numerical experiments.

REFERENCES

1. Anderson, S.P., Erkal, N., and Piccinin, D., Aggregative Games and Oligopoly Theory: Short-Run and Long-Run Analysis, *RAND J. Econom.*, 2020, vol. 51, no. 2, pp. 470–495.

2. Cournot, A.A., *Researches into the Mathematical Principles of the Theory of Wealth*, London: Hafner, 1960. (Original 1838.)

3. Nash, J.F.Jr., Non-cooperative Games, *Annal. Mathem.*, 1951, vol. 54, no. 2, pp. 286–295.

4. Stackelberg, H., *Market Structure and Equilibrium*, 1st ed. Translation into English, Bazin, Urch & Hill, Springer, 2011. (Original 1934.)

5. Novikov, D.A. and Chkhartishvili, A.G., *Reflexion and Control: Mathematical Models*, Leiden: CRC Press, 2014.

6. 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.

7. 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.

8. Malsagov, M., Ougolnitsky, G., and Usov, A., A Differential Stackelberg Game Theoretic Model of the Promotion of Innovations in Universities, *Advances Syst. Sci. Appl.*, 2020, vol. 20, no. 3, pp. 166–177.

9. Gubanov, D.A., Novikov, D.A., and Chkhartishvili, A.G., Informational Influence and Information Control Models in Social Networks, *Control Sciences*, 2009, no. 5, pp. 28–35.

10. Gubanov, D.A., Novikov, D.A., and Chkhartishvili, A.G., *Social Networks: Models of Information Influence, Control and Confrontation*, Cham: Springer, 2019.

11. Gubanov, D.A., Petrov, I.V., and Chkhartishvili, A.G., Multidimensional Model of Opinion Dynamics in Social Networks: Polarization Indices, *Autom. Remote Control*, 2021, vol. 82, no. 10, pp. 1802–1811.

12. Korn, G.A. and Korn, T.M., *Mathematical Handbook for Scientists and Engineers. Definitions, Theorems, and Formulas for Reference and Review*, New York: Dover Publications, 2000.

13. Algazin, G.I. and Algazina, D.G., Modeling the Dynamics of Collective Behavior in a Reflexive Game with an Arbitrary Number of Leaders, *Informatics and Automation*, 2022, vol. 21, no. 2, pp. 339–375.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 6, pp. 589–597.
© The Author(s), 2025 published by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 2025.
Russian Text © The Author(s), 2025, published in Avtomatika i Telemekhanika, 2025, No. 6, pp. 118–130.

Optimization by a Probabilistic Criterion in a Dynamic Test Passing Model

S. V. Ivanov[*,a] **and A. V. Stepanov**[*,b]

[*]*Moscow Aviation Institute (National Research University), Moscow, Russia*
e-mail: [a]*sergeyivanov89@mail.ru,* [b]*rus.fta@yandex.ru*

Received February 23, 2025
Revised April 8, 2025
Accepted April 12, 2025

Abstract—A dynamic model of passing a time-limited test is considered. The problems of finding program and positional strategies that maximize the probability of test passing are stated. A strategy is to complete or not complete a current test task. A positional strategy is defined as a function of the time spent since the test start and the sum of points scored for previous tasks. Bellman's dynamic programming method is used to design a positional strategy. An algorithm based on the branch-and-bound method is proposed to find an optimal program strategy. The results of calculations are presented and compared with those of a similar problem with a time limit (i.e., when the test is considered unpassed if the testee does not meet the limit).

Keywords: testing model, probabilistic criterion, control design

DOI: 10.31857/S0005117925060076

1. INTRODUCTION

With the modern application level of information technologies in education, it is topical to develop, in particular, the theory of computerized adaptive testing (CAT). Its foundations were laid at the end of the twentieth century, e.g., in the works of G. Rasch [1]. A good survey of the state-of-the-art results in this field can be found in [2]. Conceptually CAT involves information about the testee's performance (the results of learning) and passing of a control event in the process of forming the test content and deciding on test completion. (Below testees will be called subjects.) The processing of this information and decision-making are often implemented with contemporary machine learning and artificial intelligence techniques. The same technologies are also used to form individual learning trajectories in learning management systems (LMSs); for example, see [3, 4]. Recent works in this area include [5–7]. Such technologies are commonly applied by test organizers and LMS software developers. However, the testing problem can also be viewed from the subject's standpoint, who needs to develop an individual strategy for passing an (often, time-limited) test. The test completion time is directly related to the subject's response time to test tasks. Many publications have been devoted to probabilistic models of the response time of a subject or an LMS user to a task; for example, we refer to [8–10]. The problem of finding the subject's optimal strategy by the criterion of maximizing the probability of scoring a given sum of points under a time limit was considered in [11]. A similar problem in a quantile formulation (maximizing the sum of points scored during a test under a probabilistic time constraint) was studied in [12]. In the last two papers mentioned, the following stochastic programming problems were solved: before a test (a priori), a subject chooses as an individual strategy a group of tasks he/she will tackle first. Other test tasks are performed by the subject if he/she has enough time after completing the first group of tasks. The above formulations neglect the sequence in which a subject performs test tasks. This may be natural if a subject receives no information about the correctness of performing a current task and, consequently, about the current sum of points scored. However, very often, especially in LMS testing, e.g., [4], such information is available to a user and can be utilized by him/her during the test. Therefore, it is relevant to consider dynamic testing models with a test passing strategy corrected after each task considered by a subject.

This paper deals with a dynamic model of passing a fixed time-limited test by the criterion of maximizing the probability of the subject's scoring at least a given sum of points. The problems of finding the subject's program and positional strategies are stated. Algorithms are proposed to solve these problems based on stochastic and dynamic programming methods. The resulting solutions are compared with those obtained in [11, 12] for the same initial data.

2. DESCRIPTION OF THE MODEL AND CONTROL DESIGN PROBLEM

Consider a dynamic system describing test passing by a subject (e.g., a student). At the kth step, the subject is asked to perform a task; in case of success, he/she is awarded b_k points, $k = \overline{1, n}$. It is required to score at least φ points to pass the test successfully. The test completion time is limited by $\bar{T}$. The time for performing the kth task is described by a random variable τ_k. The correctness of performing the kth task is described by a random variable X_k : it takes value 1 if the kth task has been performed correctly and value 0 otherwise. A strategy is defined by a set of variables $u = (u_1, \dots, u_n)$, where $u_k = 1$ if the subject attempts to perform the kth task and $u_k = 0$ otherwise. Now we introduce the state variables of the dynamic system. Let T_k be the time spent to complete the first k tasks and S_k be the sum of points scored during this time. Then the system dynamics are described by the equations

$$T_k = T_{k-1} + \tau_k u_k, \tag{1}$$

$$S_k = S_{k-1} + b_k X_k I\{T_k \leqslant \bar{T}\} u_k, \tag{2}$$

$$T_0 = 0,\ S_0 = 0, \quad k = \overline{1, n}, \tag{3}$$

where $I\{\cdot\}$ denotes the indicator of the event in curly brackets, equal to 1 if the condition is satisfied and to 0 otherwise. Thus, b_k points are awarded if the subject has spent no more than $\bar{T}$ units of time on the first k tasks. Assume that all the random variables τ_k are discrete with a finite set $\mathcal{T}_k$ of realizations and let $p_k(t, x) = \mathbf{P}\{\tau_k = t, X_k = x\}$. In this paper, for all $k = \overline{2, n}$, the sigma algebra generated by the random variables X_k and τ_k is supposed to be independent of the sigma algebra generated by the random variables $X_1, \dots, X_{k-1}$ and $\tau_1, \dots, \tau_{k-1}$. Such an assumption means that the answer to the current test questions does not affect further test performance. How does the knowledge of the correctness of performing previous tasks influence the correctness of performing subsequent ones? This issue requires additional statistical analysis. Let $S(u)$ be the random sum of points scored, equal to the value S_n under the chosen control action u.

We state the problem of finding the maximum of the probability function

$$P^1_{\varphi, \bar{T}}(u) = \mathbf{P}\{S(u) \geqslant \varphi\} \to \max_{u \in \{0,1\}^n} . \tag{4}$$

This problem is to design a program strategy for choosing tasks to be performed under which the subject will maximize the probability of test passing.

Now assume that the control action is chosen in the class of positional strategies. In other words, when choosing his/her strategy at each step, the subject considers the current time spent and the current sum of points scored for the previous tasks. We denote by $\mathbf{u}_k(T_{k-1}, S_{k-1})$ the subject's strategy at the kth step, represented by the following possible values of some function $\mathbf{u}_k$ of the system's state variables: 1 if the subject tries to perform the kth task and 0 otherwise. Let $\mathbf{u}$ be a vector composed of the functions $\mathbf{u}_1, \ldots, \mathbf{u}_n$ to describe a positional strategy. We denote by $\mathbf{S}(\mathbf{u})$ the random sum S_n of points defined by equations (1) and (2) when substituting $u_k = \mathbf{u}_k(T_k, S_k)$ therein. Boldface is used here to distinguish the positional strategy $\mathbf{u}$ (a function) from the program strategy u (a set of variables) as well as the sum $\mathbf{S}(\mathbf{u})$ of points scored by the positional strategy from the sum S_n of points scored by the program strategy.

We state the problem of finding the maximum of the probability function under the positional strategy:

$$P^2_{\varphi,\bar{T}}(\mathbf{u}) = \mathbf{P}\{\mathbf{S}(\mathbf{u}) \geqslant \varphi\} \to \max_{\mathbf{u}\in\mathcal{U}}, \tag{5}$$

$$\mathbf{u}^* = (\mathbf{u}_1^*, \ldots, \mathbf{u}_n^*) \in \operatorname{Arg}\max_{\mathbf{u}\in\mathcal{U}} P^2_{\varphi,\bar{T}}(\mathbf{u}), \tag{6}$$

where $\mathcal{U}$ is the set of admissible control functions. Since the random response time for any task is a discrete random variable with a finite number of values and the correct answer is modeled by the Bernoulli distribution, the number of states (T_k, S_k) at the kth step is finite. However, for convenience of optimization, the control action $\mathbf{u}_k$ at the kth step will be chosen as a function on the wider set of states $\mathbb{T}_k \times \mathbb{S}_k$, where $\mathbb{T}_k$ and $\mathbb{S}_k$ are some finite sets containing all possible values of the states T_k and S_k, respectively. For this reason, $\mathcal{U}$ is supposed to be the set of all functions from $\bigotimes\limits_{k=1}^{n} \mathbb{T}_k \times \mathbb{S}_k$ into $\{0,1\}^n$.

Remark 1. In the class of program strategies, the problem of maximizing the probability of scoring a required sum of points under a time limit was solved in [11]. In the current notation, this problem has the form

$$\tilde{P}^1_{\varphi,\bar{T}}(u) = \mathbf{P}\{S(u) \geqslant \varphi,\ T(u) \leqslant \bar{T}\} \to \max_{u\in\{0,1\}^n}, \tag{7}$$

where $T(u)$ is the random test completion time under the program strategy u. Within problem (7), a test is considered unpassed if the subject fails to meet the time limit even when scoring the required sum of points. There is no requirement to meet the time limit in problem (4), but no points are awarded for performing tasks after the time $\bar{T}$. Meanwhile, the advantage of (7) is the independence of the optimal choice of tasks from their order in the test. In problems (4) and (5), the optimal set of tasks to be performed depends on their order. Note that the optimal value of the objective function in problem (7) is a lower bound for those of the objective functions in problems (4) and (5) for any task order in the test.

Remark 2. Problem (7) can be generalized to the case of positional strategies. As a result, we obtain the problem

$$\tilde{P}^2_{\varphi,\bar{T}}(\mathbf{u}) = \mathbf{P}\{\mathbf{S}(\mathbf{u}) \geqslant \varphi,\ \mathbf{T}(\mathbf{u}) \leqslant \bar{T}\} \to \max_{\mathbf{u}\in\mathcal{U}}, \tag{8}$$

where $\mathbf{T}(\mathbf{u})$ is the random test completion time T_n defined by equations (1) when substituting $u_k = \mathbf{u}_k(T_k, S_k)$ therein. As will be shown below, problem (5) turns out to be completely equivalent to problem (8), in contrast to the pair of problems (4) and (7).

3. POSITIONAL CONTROL DESIGN

To solve the positional control design problem (5), we apply Bellman's dynamic programming method. Let us define the Bellman function for $k = \overline{1,n}$ by the rule

$$B_k(T_{k-1}, S_{k-1}) = \max_{\mathbf{u}_k,\ldots,\mathbf{u}_n} \mathbf{P}\{\mathbf{S}(\mathbf{u}) \geqslant \varphi \mid T_{k-1}, S_{k-1}\}. \tag{9}$$

At the last step, the Bellman function is defined by the equality

$$B_{n+1}(T_n, S_n) = I\{S_n \geqslant \varphi\}. \tag{10}$$

Note the absence of dependence on T_n at the last step: it has been introduced for the convenience of writing the dynamic programming relations. Obviously,

$$\max_{\mathbf{u}\in\mathcal{U}} P^2_{\varphi,\bar{T}}(\mathbf{u}) = B_1(T_0, S_0).$$

We obtain the dynamic programming relations using the formula of total probability:

$$B_k(T_{k-1}, S_{k-1}) = \max_{u_k=\mathbf{u}_k(T_{k-1},S_{k-1})} \max_{\mathbf{u}_{k+1},\dots,\mathbf{u}_n} \sum_{t\in\mathcal{T}_k,\, x\in\{0,1\}} \mathbf{P}\{\mathbf{S}(\mathbf{u})$$

$$\geqslant \varphi \mid S_k = S_{k-1} + b_k x I\{T_{k-1} + tu_k \leqslant \bar{T}\}u_k,\ T_k = T_{k-1} + tu_k\} p_k(t,x)$$

$$= \max_{u_k\in\{0,1\}} \sum_{t\in\mathcal{T}_k,\, x\in\{0,1\}} B_{k+1}(T_{k-1} + tu_k, S_{k-1} + b_k x I\{T_{k-1} + tu_k \leqslant \bar{T}\}u_k) p_k(t,x). \tag{11}$$

The control action $\mathbf{u}^*_k(T_{k-1}, S_{k-1})$ will be determined by maximizing the Bellman function at the kth step.

Since the random variables have discrete distributions, all the functions in the dynamic programming method are measurable, which ensures the correctness of this method.

Thus, the problem can be solved using the following algorithm, which implements dynamic programming.

Algorithm 1 (positional control design).

1. Set $k := n + 1$; for all $S_n \in \mathbb{S}_n$ calculate the value $B_{n+1}(T_n, S_n)$.
2. If $k > 1$, set $k := k - 1$; otherwise, proceed to Step 4.
3. For all $T_{k-1} \in \mathbb{T}_{k-1}$ and $S_{k-1} \in \mathbb{S}_{k-1}$ calculate the value $B_k(T_{k-1}, S_{k-1})$ and determine $\mathbf{u}^*_k(T_{k-1}, S_{k-1})$.
4. Calculate the value $B_1(0, 0)$ and determine the control action $\mathbf{u}^*_1(0, 0)$ of the first step.

Note that with additionally calculating $B_1(t, s)$ at Step 4 of Algorithm 1, we will maximize the objective functional $P^2_{\varphi-s,\bar{T}-t}(\mathbf{u})$. These calculations can be carried out if the set of reachable states of the dynamic system with the initial conditions $T_0 = t$, $S_0 = s$ is a subset of $\mathbb{T}_k \times \mathbb{S}_k$ at each kth step. Thus, it is possible to solve the problem for several values of φ and $\bar{T}$ at once.

The above solution method allows proving the equivalence of problems (5) and (8).

Proposition 1. *The optimal values of the objective functionals in problems* (5) *and* (8) *are equal. There exist strategies optimal in both problems, namely, those satisfying the condition* $\mathbf{u}^*_k(T_{k-1}, S_{k-1}) = 0$ *for* $S_{k-1} \geqslant \varphi$ *or* $T_{k-1} > \bar{T}$*.*

Proof. To solve problem (8), we also apply the dynamic programming method. For this purpose, it is necessary to define the Bellman function at the last step by the rule $\tilde{B}_{n+1}(T_n, S_n) = I\{T_n \leqslant \bar{T},\ S_n \geqslant \varphi\}$. The dynamic programming relations for solving problem (8) will be the same as (11), with the functions $\tilde{B}_k$ written instead of B_k.

Let us analyze the resulting dynamic programming relations. Note that for $T_{k-1} > \bar{T}$ (the time limit is exhausted) or $S_{k-1} \geqslant \varphi$ (the required sum of points is scored), when computing $B_k(T_{k-1}, S_{k-1})$ and $\tilde{B}_k(T_{k-1}, S_{k-1})$, we can assign $\mathbf{u}^*_k(T_{k-1}, S_{k-1}) = 0$ (the condition formulated in Proposition 1).

Now let us show the following result by induction: $B_k(T_{k-1}, S_{k-1}) = \tilde{B}_k(T_{k-1}, S_{k-1})$ for $T_{k-1} \leqslant \bar{T}$, with the presence of coincident strategies among those on which the maximum is reached, and if $S_{k-1} < \varphi$, then the maximum is reached on the same strategies. For $k = n$ this assertion is true by the definition of the above Bellman functions. Assuming its truth for $B_{k+1}(T_k, S_k)$, we establish it for $B_k(T_{k-1}, S_{k-1})$. If $S_{k-1} \geqslant \varphi$ and $T_{k-1} \leqslant \bar{T}$, then $B_k(T_{k-1}, S_{k-1}) = \tilde{B}_k(T_{k-1}, S_{k-1}) = 1$ (as noted above, we can take $\mathbf{u}_k^*(T_{k-1}, S_{k-1}) = 0$). Two cases may arise when calculating $B_k(T_{k-1}, S_{k-1})$ under the conditions $S_{k-1} < \varphi$ and $T_{k-1} \leqslant \bar{T}$: 1) If $T_k > \bar{T}$, then the equality $S_k = S_{k-1}$ holds and the value $B_k(T_k, S_{k-1}) = \tilde{B}_k(T_k, S_{k-1}) = 0$ is considered. 2) If $T_k \leqslant \bar{T}$, then the value $B_k(T_k, S_{k-1}) = \tilde{B}_k(T_k, S_{k-1})$ is considered (equality is valid by the inductive hypothesis). In both cases, we have $B_k(T_{k-1}, S_{k-1}) = \tilde{B}_k(T_{k-1}, S_{k-1})$ and the maximum in the definitions of Bellman functions is achieved on the same strategies. Thus, the proof of Proposition 1 is complete.

Remark 3. Among the optimal strategies in problem (5), there may be strategies satisfying the condition $\mathbf{u}_k^*(T_{k-1}, S_{k-1}) = 1$ for $S_{k-1} \geqslant \varphi$, which are not optimal ones in problem (8). Indeed, when the required sum of points has been scored, performing additional tasks reduces the probability of not exceeding the time limit, decreasing the value of the objective functional in problem (8) compared to the value of that in problem (5).

4. PROGRAM CONTROL DESIGN

First, we describe a method for calculating the objective functional in (4) for a fixed $u \in \{0,1\}^n$. Let us define the following functions:

$$\begin{aligned} &B_{n+1}^u(T_n, S_n) = I\{S_n \geqslant \varphi\}, \\ &B_k^u(T_{k-1}, S_{k-1}) = \sum_{t \in \mathcal{T}_k,\, x \in \{0,1\}} B_{k+1}^u(T_{k-1} + tu_k, S_{k-1} + b_k x I\{T_{k-1} + tu_k \leqslant \bar{T}\} u_k) p_k(t, x), \end{aligned} \tag{12}$$

where $k = \overline{1,n}$. The value $B_1^u(0,0)$ yielded by these formulas is equal to that of the probability functional $P_{\varphi,\bar{T}}^1(u)$. This fact follows from the definition of the Bellman function (9) since the relations (12) are similar to the Bellman equations but without maximization and with the control strategy u_k at the kth step.

Owing to the use of dynamic relations, the above procedure is much more efficient than the direct calculation of probability by enumerating all possible realizations of the random variables X_k and τ_k.

The optimal program control can be found by enumerating all possible control actions $u \in \{0,1\}^n$. In this case, it is possible to eliminate the control actions for which $\sum\limits_{k=1}^{n} b_k u_k < \varphi$: such actions do not ensure test passing for any realizations of the random variables with a nonzero probability. Also, we eliminate the control actions u with $u_n = 0$ because they are no worse than the corresponding ones with $u_n = 1$. Indeed, trying to perform the last task cannot reduce the probability of test passing.

We propose a procedure for significantly reducing the number of program control actions enumerated. Assume that the values of the Bellman function $B_k(S_{k-1}, T_{k-1})$ are known. (They can be calculated using the algorithm from the previous section.) Let us introduce the notation

$$\begin{aligned} &C_{k,k}^{u_1,\ldots,u_k}(T_{k-1}, S_{k-1}) = \sum_{t \in \mathcal{T}_k,\, x \in \{0,1\}} B_{k+1}(T_{k-1} + tu_k, S_{k-1} + b_k x I\{T_{k-1} + tu_k \leqslant \bar{T}\} u_k) p_k(t, x), \\ &C_{l,k}^{u_1,\ldots,u_k}(T_{l-1}, S_{l-1}) = \sum_{t \in \mathcal{T}_l,\, x \in \{0,1\}} C_{l+1,k}^{u_1,\ldots,u_k}(T_{l-1} + tu_l, S_{l-1} + b_l x I\{T_{l-1} + tu_l \leqslant \bar{T}\} u_l) p_l(t, x), \\ &l = k-1, k-2, \ldots, 1, \quad C_k(u_1, \ldots, u_k) = C_{1,k}^{u_1,\ldots,u_k}(T_0, S_0). \end{aligned} \tag{13}$$

The value $C_{l,k}^{u_1,\ldots,u_k}(T_{k-1}, S_{k-1})$ describes the minimum loss starting from the kth step under the control actions $u_1, \ldots, u_k$ applied at the first k steps. These values are calculated from the dynamic programming relations with fixed control actions at the first k steps. The values $C_k(u_1, \ldots, u_k)$ can be used as upper bounds for the objective function.

Proposition 2. *The following relations hold for the values given by* (13)*:*

1) $C_k(u_1,\dots,u_k) \geqslant B_1^u(0,0) = P^1_{\varphi,\bar{T}}(u) = C_n(u_1,\dots,u_n)$.

2) $C_k(u_1,\dots,u_k) \geqslant C_{k+1}(u_1,\dots,u_{k+1})$ *for all* $k = \overline{1,n-1}$.

Proof. Equations (13) are the dynamic Bellman relations with the control actions $u_1,\dots,u_k$ at the first k steps. Therefore, the value $C_k(u_1,\dots,u_k)$ coincides with that of the objective functional in problem (5) with positional control, where the control vector $\mathbf{u}$ has fixed components equal to $u_1,\dots,u_k$ at the first k steps and the optimally chosen components at the subsequent steps (with respect to the state achieved in k steps). Hence, for any program control u with fixed $u_1,\dots,u_k$, we have the inequality $C_k(u_1,\dots,u_k) \geqslant B_1^u(0,0)$. In addition, the control vector with fixed components $u_1,\dots,u_{k+1}$ gives a smaller value of the performance functional in (5) than that with fixed components $u_1,\dots,u_k$. Therefore, we obtain $C_k(u_1,\dots,u_k) \geqslant C_{k+1}(u_1,\dots,u_{k+1})$ for all $k = \overline{1,n-1}$. The equality $B_1^u(0,0) = P^1_{\varphi,\bar{T}}(u)$ has been established above. The proof of Proposition 2 is complete.

Thus, if $\psi < \max\limits_{u\in\{0,1\}^n} P^1_{\varphi,\bar{T}}(u)$ (the known lower bound for the optimal value of the objective function) and $C_k(u_1,\dots,u_k) \leqslant \psi$ for some k, then by Proposition 2 we have $P^1_{\varphi,\bar{T}}(u) < \max\limits_{u\in\{0,1\}^n} P^1_{\varphi,\bar{T}}(u)$ and, consequently, any strategy u with the corresponding first k components is not optimal. Therefore, Algorithm 2 based on the branch-and-bound method (see below) can be proposed to find the program strategy. This algorithm traverses a binary tree with the following properties: the root corresponds to the variable u_1; the outgoing edges of the root, to values 1 and 0 of this variable; the adjacent vertices of the root, to the variable u_2; the outgoing edges of these vertices, to the values of the variable u_2, and so on. The vertices corresponding to the variable u_n have two outgoing edges corresponding to the values of the variable u_n and connecting them to the leaves. A depth-first search is applied to traverse the tree; the search can be terminated if its further execution does not yield an optimal strategy. To describe the termination time, we introduce a function $L(u_1,\dots,u_k)$: it takes value 0 if no further depth-first search is performed and value 1 otherwise. The variables ψ and u^* modified when executing the function $L(u_1,\dots,u_k)$ are global.

Algorithm 2 (program control design).

```
0. Define the function $L(u_1,\dots,u_k)$ by the following rule:
0a. Set $\tilde{u} = (u_1,\dots,u_k,1,\dots,1)$. If $\sum_{k=1}^{n} b_k\tilde{u}_k \geqslant \varphi$ or ($k = n$ and $u_n = 1$), calculate $C_k(u_1,\dots,u_k)$;
    otherwise, return $L(u_1,\dots,u_k) = 0$.
0b. If $C_k(u_1,\dots,u_k) \leqslant \psi$, return $L(u_1,\dots,u_k) = 0$.
    If $k = n$ and $C_n(u_1,\dots,u_n) > \psi$, assign $\psi := C_n(u_1,\dots,u_n)$ and $u^* := (u_1,\dots,u_n)$ and return $L(u_1,\dots,u_k) = 0$.
    If $k = n$ and $C_n(u_1,\dots,u_n) \leqslant \psi$, return $L(u_1,\dots,u_k) = 0$;
    otherwise return $L(u_1,\dots,u_k) = 1$.
1. Set $k := 1$, $\psi := 0$, $u^* := (0,\dots,0)$, and $u_1 := 1$.
2. Calculate $L(u_1,\dots,u_k)$.
3. If $L(u_1,\dots,u_k) = 1$, set $k := k+1$ and $u_k := 1$ and get back to Step 2.
   If $L(u_1,\dots,u_k) = 0$ and $u_k = 1$, set $u_k := 0$ and get back to Step 2.
   If $L(u_1,\dots,u_k) = 0$ and $u_l = 0$ for all $l = \overline{1,k}$, terminate execution of the algorithm;
   otherwise, assign $k := \max\{l \mid u_l = 1, l < k\}$ and $u_k := 0$ and get back to Step 2.
```

Algorithm 2 yields the optimal control strategy u^* and the corresponding optimal value of the objective function $P^1_{\varphi,\bar{T}}(u^*) = \psi$.

The depth-first search of the algorithm starts from the variable values $u_k = 1$: in this case, the first strategy to be considered is "perform all tasks," which usually corresponds to a large value of the probability of test passing.

Remark 4. Algorithm 2 can be applied to solve problem (7) as well. For this purpose, we need to set $B_{n+1}(T_n, S_n) = I\{T_n \leqslant \bar{T},\ S_n \geqslant \varphi\}$ when determining the values $C_k(u_1, \ldots, u_k)$. In addition, at Step 0a, it is necessary to eliminate the check ($k = n$ and $u_n = 1$) since the optimal strategy may have $u_n = 0$.

5. NUMERICAL RESULTS

The two problems posed were successfully solved for the data from [11]. In the test under consideration, there are 10 tasks with the points $b_1 = \ldots = b_5 = 1$, $b_6 = b_8 = 2$, $b_7 = b_{10} = 3$, and $b_9 = 4$ scored for performing them. A subject needs to score $\varphi = 11$ points. The probabilities of correct answers to each task are known. Each task is associated with three possible realizations of the time spent on a correct answer and its three realizations for an incorrect answer; their conditional probabilities are known and were given in [11].

Table 1 presents the solutions of problems (4) and (5) depending on the parameter $\bar{T}$. Here, the values of $\bar{T}$ are different fractions of $T^m = 3830$, the maximum possible time for performing all tasks. The run times of Algorithms 2 and 1 to find program (prog.) and positional (pos.) strategies, respectively, are indicated. The last column shows the time for solving problem (4) when enumerating all possible program strategies, except for the obviously suboptimal ones, using formulas (12).

Similarly, Table 2 provides the solutions of problems (7) and (8). The run times of Algorithms 2 and 1 are indicated as well. All calculations were carried out on an Acer Aspire A315-54K laptop (Intel Core i5-6300U 2.4 GHz CPU, 8Gb RAM).

Table 1. The solutions of problems (4) and (5) depending on the parameter $\bar{T}$ for $\varphi = 11$

$\bar{T}$	Optimal program strategy u^*	$P^1_{\varphi,\bar{T}}(u^*)$	$\max\limits_{\mathbf{u}\in\mathcal{U}} P^2_{\varphi,\bar{T}}(\mathbf{u})$	Run time, prog./pos. (s)	Run time, enumeration (s)
$0.4T^m$	$(0,0,0,0,1,1,1,1,0,1)$	0.1447	0.1689	0.83 / 0.08	1.69
$0.5T^m$	$(1,1,1,0,1,1,1,1,0,1)$	0.3583	0.4220	0.45 / 0.07	1.78
$0.6T^m$	$(1,0,1,1,1,1,1,1,0,1)$	0.5244	0.6130	0.53 / 0.07	1.93
$0.7T^m$	$(1,1,1,0,1,1,1,1,1,1)$	0.6906	0.7242	0.28 / 0.07	2.17
$0.8T^m$	$(1,1,1,1,1,1,1,1,1,1)$	0.7815	0.7876	0.11 / 0.08	2.34
$0.9T^m$	$(1,1,1,1,1,1,1,1,1,1)$	0.8006	0.8007	0.15 / 0.07	2.64

Table 2. The solutions of problems (7) and (8) depending on the parameter $\bar{T}$ for $\varphi = 11$

$\bar{T}$	Optimal program strategy u^*	$\tilde{P}^1_{\varphi,\bar{T}}(u^*)$	$\max\limits_{\mathbf{u}\in\mathcal{U}} \tilde{P}^2_{\varphi,\bar{T}}(\mathbf{u})$	Run time, prog./pos. (s)	Run time, enumeration (s)
$0.4T^m$	$(0,0,0,0,1,1,1,1,0,1)$	0.1447	0.1689	0.99 / 0.08	3.47
$0.5T^m$	$(1,0,1,0,1,1,1,1,0,1)$	0.2460	0.4220	3.02 / 0.07	3.45
$0.6T^m$	$(1,1,1,0,1,1,1,1,0,1)$	0.4963	0.6130	1.49 / 0.07	3.45
$0.7T^m$	$(1,1,1,1,1,1,1,1,0,1)$	0.6059	0.7242	0.64 / 0.07	3.46
$0.8T^m$	$(1,1,1,0,1,1,1,1,1,1)$	0.7216	0.7876	0.63 / 0.07	3.41
$0.9T^m$	$(1,1,1,1,1,1,1,1,1,1)$	0.7965	0.8007	0.14 / 0.07	3.42

According to these tables, positional strategies have an advantage over program ones, not only in terms of the optimal value of the objective functional but also in terms of the speed of computation. Comparing problem (4) with the one from [11], we note that, as a rule, the solution of (4) implies a larger number of subproblems and a larger value of the objective functional. This is due to the absence of a time limit in problem (4).

As mentioned earlier, the solution of (4) and (5) depends on the order of test tasks. Therefore, additional calculations of the strategies for passing the same test, but with the reverse order of tasks, were performed: the first task of the test was made the tenth, the second the ninth, and so on. The points awarded for the tasks and the probabilities of performing the tasks remained the same. The results of solving this problem are described in Table 3. For convenience of comparing the tables, the components of the optimal program strategies are given in reverse order. The run times of the algorithms are omitted here because they insignificantly differ from the ones in Table 1. Direct comparison of the solutions shows that the values of the objective functions changed slightly with the reverse order of the tasks, and the program strategy differs for the time limits $0.4T^m$ and $0.6T^m$.

Table 3. The solutions of problems (4) and (5) depending on the parameter $\bar{T}$ for $\varphi = 11$: the reverse order of tasks

$\bar{T}$	Optimal program strategy $(u_{10}^*, \ldots, u_1^*)$	$P^1_{\varphi,\bar{T}}(u^*)$	$\max\limits_{\mathbf{u}\in\mathcal{U}} P^2_{\varphi,\bar{T}}(\mathbf{u})$
$0.4T^m$	$(1,0,1,0,1,1,1,1,0,1)$	0.1639	0.1787
$0.5T^m$	$(1,1,1,0,1,1,1,1,0,1)$	0.3526	0.3834
$0.6T^m$	$(1,1,1,1,1,1,1,1,0,1)$	0.5340	0.5718
$0.7T^m$	$(1,1,1,0,1,1,1,1,1,1)$	0.6781	0.7003
$0.8T^m$	$(1,1,1,1,1,1,1,1,1,1)$	0.7824	0.7847
$0.9T^m$	$(1,1,1,1,1,1,1,1,1,1)$	0.8007	0.8007

6. CONCLUSIONS

In this paper, we have found program and positional strategies for passing a time-limited test in which a testee (subject) knows the current sum of points scored after performing each task. Algorithms for solving these problems have been proposed and their effectiveness has been demonstrated. The results have been compared with those of the problem statement considered by one of the authors previously. According to the analysis, with the current sum of points being known to a subject, there is a test passing strategy under which the subject will exceed the required level with a higher probability.

Future research may deal with various modifications of this model. In particular, the dependence of the results of performing test tasks on the current sum of points scored can be considered. Such a dependence arises due to the psychological peculiarities of some subjects. The possibility of modifying the Bellman equations (11) in this case needs to be investigated. Also, it seems interesting to describe the response time of subjects (their answers to test tasks) by continuous random variables and to study the solution accuracy in the discretization case.

REFERENCES

1. Rasch, G., *Probabilistic Models for Some Intelligence and Attainment Tests*, Chicago: The University of Chicago Press, 1980.

2. Xiao, J. and Bulut, O., Item Selection with Collaborative Filtering in On-the-fly Multistage Adaptive Testing, *Appl. Psychol. Meas.*, 2022, vol. 46, no. 8, pp. 690–704.

3. Naumov, A.V., Dzhumurat, A.S., and Inozemtsev, A.O., Distance Learning System for Mathematical Disciplines CLASS.NET, *Herald of Computer and Information Technologies*, 2014, no. 10, pp. 36–44.

4. CLASS.NET. The Distance Learning System of Moscow Aviation Institute.
URL: https://distance.kaf804.ru// (Accessed October 12, 2024.)

5. Kuravsky, L.S., Margolis, A.A., Marmalyuk, P.A., et al., A Probabilistic Model of Adaptive Training, *Appl. Math. Sci. (Ruse)*, 2016, vol. 10, no. 48, pp. 2369–2380.

6. Bosov, A.V., Martyushova, Ya.G., Naumov, A.V., and Sapunova, A.P., Bayesian Approach to the Construction of an Individual User Trajectory in the System of Distance Learning, *Informatics and Applications*, 2020, vol. 14, no. 3, pp. 86–93.

7. Bosov, A.V., Application of Self-Organizing Neural Networks to the Process of Forming an Individual Learning Path, *Informatics and Applications*, 2022, vol. 16, no. 3, pp. 7–15.

8. van der Linden, W.J., Scrams, D.J., and Schnipke, D.L., Using Response-Time Constraints to Control for Differential Speededness in Computerized Adaptive Testing, *Appl. Psych. Meast.*, 1999, vol. 23, no. 3, pp. 195–210.

9. Bosov, A.V., Mkhitaryan, G.A., Naumov, A.V., and Sapunova, A.P., Using the Model of Gamma Distribution in the Problem of Forming a Time-Limited Test in a Distance Learning System, *Informatics and Applications*, 2019, vol. 13, no. 4, pp. 11–17.

10. Naumov, A.V., Mkhitaryan, G.A., and Cherygova, E.E., Stochastic Statement of the Problem of Generating Tests with Defined Complexity with the Minimization of Quantile of Test Passing Time, *Herald of Computer and Information Technologies*, 2019, no. 2, pp. 37–46.

11. Naumov, A.V., Stepanov, A.E., and Ustinov, A.E., On the Problem of Maximizing the Probability of Successful Passing of a Time-Limited Test, *Autom. Remote Control*, 2024, vol. 85, no. 1, pp. 64–72.

12. Martyushova, Ya.G., Naumov, A.V., and Stepanov, A.E., Optimization of the Strategy of Passing the Time-Limited Test According to the Quantile Criterion, *Informatics and Applications*, 2024, vol. 18, no. 4, pp. 44–51.

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