Volume 85, Number 1,
January 2024

ISSN 0005-1179
CODEN: AURCAT

AUTOMATION AND REMOTE CONTROL

http://ait.mtas.ru

Automation and Remote Control

Vol. 85, No. 1, January 2024

Automation and Remote Control

ISSN 0005-1179

SCOPE

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

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

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

0005-1179/24. *Automation and Remote Control* (ISSN: 0005-1179 print version, ISSN: 1608-3032 electronic version) is published monthly by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 65 Profsoyuznaya street, Moscow 117997, Russia.

Volume 85 (12 issues) is published in 2024.

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

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

Contents

Automation and Remote Control
Vol. 85, No. 1, 2024

Linear Systems

Nonlinear Systems

Stochastic Systems

Control in Technical Systems

Optimization, System Analysis, and Operations Research

Obituary

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 1, pp. 1–14.

LINEAR SYSTEMS

Design of Generalized H_∞-Suboptimal Controllers Based on Experimental and A Priori Data

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

[*] *Lobachevsky University, Nizhny Novgorod, Russia*
[**] *Nizhny Novgorod State University of Architecture and Civil Engineering, Nizhny Novgorod, Russia*
e-mail: [a] *mkogan@nngasu.ru,*[b] *andrey8st@yahoo.com*

Received September 28, 2023
Revised November 27, 2023
Accepted December 21, 2023

Abstract—This paper considers a linear continuous- or discrete-time dynamic object in the absence of its mathematical model. As is demonstrated below, a control law that suboptimally damps initial and (or) exogenous disturbances of such objects can be implemented based on experimental and a priori data. The approach involves the methods of robust control design and duality theory as well as the technique of linear matrix inequalities.

Keywords: generalized H_∞ norm, uncertainty, robust control, experimental data, dual systems, linear matrix inequalities

DOI: 10.31857/S0005117924010014

1. INTRODUCTION

Recently, increasing attention in control theory has been paid to the design of control laws for dynamic objects with highly uncertain mathematical models, exogenous disturbances, and unknown initial conditions. Within this line of research, by assumption, a series of experiments can be conducted with an object by setting input actions and measuring output variables. The problem is to determine the feedback parameters ensuring a given quality of the closed-loop control system directly, i.e., based on available measurements and a priori data without identifying the unknown parameters of the object.

As was established in [1], a single trajectory can be used to fully characterize a linear time-invariant dynamic system under the so-called persistency of excitation. In view of this fundamental result, different direct control design schemes based on experimental data were proposed in [2] for objects with unknown state dynamics matrices and given target output matrices under the persistency of excitation. According to [3], it suffices to fulfill the data informativity condition in order to construct control laws from experimental data, which is less restrictive than the persistency of excitation. For a fully uncertain object, H_2- and H_∞-optimal control laws were constructed based on input and output measurements using a matrix version of S-lemma [5] in the publication [4] and using Petersen's lemma [7] in the publication [6]. In [8, 9], the state feedback parameters were calculated from a priori data and open-loop measurements of the input and output of a discrete-time uncertain object subjected to an unmeasured disturbance from a definite class.

In this paper, generalized H_∞-suboptimal control laws that damp initial and (or) exogenous disturbances (as a special case, linear-quadratic control laws) for continuous- or discrete-time objects with completely unknown state dynamics and target output matrices are designed from a priori and experimental data. The design procedure is based on the approach used in [9]: the uncertain

system is "immersed" into an artificial system with known equations and an additional disturbance whose influence corresponds to that of the unknown terms in the original equation. The idea of such an artificial immersion (in other words, the representation of an uncertain system as a system whose feedback loop contains a block with unknown bounded parameters or an unknown bounded operator) was actively employed in robust control based on H_∞ optimization; see the survey [10]. However, the direct application of this approach to the design of control laws based on experimental data caused difficulties. This problem is solved below by passing from the original uncertain system to a dual uncertain system immersed into the corresponding augmented system. Implementing such an approach requires establishing a connection between the generalized H_∞ norms of the primal and dual systems.

This paper is organized as follows. After the Introduction, Section 2 gives the general problem statement; in particular, two quadratic inequalities for the unknown object parameter matrices (state and target output) are derived from a priori information and experimental data. In Section 3, a necessary background is provided on the generalized H_∞ norm, and this norm is calculated in terms of the dual system; see Lemma 3.1. Section 4 describes the design procedure for the generalized H_∞-suboptimal control laws based on a priori and experimental data, including the main theorem and its proof. Several experiments with an uncertain system are presented in Section 5 to illustrate the effectiveness of this control approach. Finally, Section 6 summarizes the results and draws conclusions.

2. PROBLEM STATEMENT

Consider an uncertain system described by

$$\begin{aligned} &\partial x(t) = Ax(t) + Bu(t) + w(t), \quad x(0) = x_0, \\ &z(t) = Cx(t) + Du(t) \end{aligned} \tag{2.1}$$

with the following notations: ∂ is the differentiation operator in the continuous-time case or the shift operator in the discrete-time case; $x(t) \in \mathrm{R}^{n_x}$ is the state vector, $u(t) \in \mathrm{R}^{n_u}$ is the control vector (input), $w(t) \in \mathrm{R}^{n_w}$ is an exogenous disturbance, and $z(t) \in \mathrm{R}^{n_z}$ is the target output. By assumption, the disturbance $w(t) \in L_2(l_2)$ and the system matrices A, B, C, and D are unknown. In general, it is required to design linear state-feedback control laws based on a priori and experimental data so that the damping level of the disturbances in the closed loop system does not exceed a specified value.

The information about the unknown parameters of system (2.1) is extracted from a finite set of measurements of its trajectory. For the discrete-time system, there are available measurements of its state and target output, $x_0, x_1, \ldots, x_N$ and $z_0, \ldots, z_{N-1}$, respectively, under chosen controls $u_0, \ldots, u_{N-1}$ and some unknown disturbance $w_0, \ldots, w_{N-1}$. We compile the matrices

$$\Phi = (x_0 \cdots x_{N-1}), \quad \Phi_+ = (x_1 \cdots x_N),$$
$$U = (u_0 \cdots u_{N-1}), \quad W = (w_0 \cdots w_{N-1}), \quad Z = (z_0 \cdots z_{N-1}).$$

In the continuous-time case, there are measurements of the system state, its derivative, and the target output, $x(t_0), \ldots, x(t_{N-1})$, $\dot{x}(t_0), \ldots, \dot{x}(t_{N-1})$, and $z(t_0), \ldots, z(t_{N-1})$, respectively, under chosen controls $u(t_0), \ldots, u(t_{N-1})$ and some unknown disturbances $w(t_0), \ldots, w(t_{N-1})$ at time instants $t_0, \ldots, t_{N-1}$. By analogy, we compile the matrices

$$\Phi = (x(t_0) \cdots x(t_{N-1})), \quad \Phi_+ = (\dot{x}(t_0) \cdots \dot{x}(t_{N-1})),$$
$$U = (u(t_0) \cdots u(t_{N-1})), \quad W = (w(t_0) \cdots w(t_{N-1})), \quad Z = (z(t_0) \cdots z(t_{N-1})).$$

The experimental data matrices in both cases satisfy the relations

$$\begin{aligned} \Phi_+ &= A_{real}\Phi + B_{real}U + W, \\ Z &= C_{real}\Phi + D_{real}U, \end{aligned} \tag{2.2}$$

where A_{real}, B_{real}, C_{real}, and D_{real} are the real (unknown) system matrices. With the notations

$$\Delta_{real} = \begin{pmatrix} A_{real} & B_{real} \\ C_{real} & D_{real} \end{pmatrix}, \quad \widehat{\Phi} = \begin{pmatrix} \Phi \\ U \end{pmatrix}, \quad \widetilde{\Phi} = \begin{pmatrix} \Phi_+ \\ Z \end{pmatrix}, \quad \widehat{W} = \begin{pmatrix} W \\ 0 \end{pmatrix},$$

equations (2.2) can be written as the linear matrix regression

$$\widetilde{\Phi} = \Delta_{real}\widehat{\Phi} + \widehat{W}. \tag{2.3}$$

Assume that the disturbance in the experiment satisfies the condition

$$\sum_{i=0}^{N-1} w(t_i)w^{\mathrm{T}}(t_i) = WW^{\mathrm{T}} \leqslant \Omega. \tag{2.4}$$

In particular, if $\|w(t)\|_\infty \leqslant d_w$ for all t and a given value d_w (the damping level), then $\Omega = d_w^2 n_w N I_{n_x}$. In the case $\sum_{i=0}^{N-1} |w(t_i)|^2 \leqslant \alpha^2$ (i.e., the total energy of the disturbance is bounded during the experiment), we obtain $\Omega = \alpha^2 I$. If $w(t)$ in (2.1) has the form $w(t) = B_v v(t)$, where $v(t) \in \mathrm{R}^{n_v}$ for some matrix B_v and $\|v(t)\|_\infty \leqslant d_v$, then $\Omega = d_v^2 n_v N B_v B_v^{\mathrm{T}}$.

From (2.4) it follows that

$$\widehat{W}\widehat{W}^{\mathrm{T}} \leqslant \begin{pmatrix} \Omega & \star \\ 0 & 0 \end{pmatrix} = \widehat{\Omega}. \tag{2.5}$$

We define the set $\mathbf{\Delta_p}$ of matrices Δ of dimensions $(n_x + n_z) \times (n_x + n_u)$ that could generate the experimental matrices Φ, Φ_+, and Z under the chosen controls U and some admissible disturbances W satisfying the constraint (2.4). For these matrices, the quality $\widetilde{\Phi} = \Delta\widehat{\Phi} + \widehat{W}$ must hold with some matrix $\widehat{W}$ satisfying (2.5). Consequently,

$$\mathbf{\Delta_p} = \left\{\Delta : \ \widetilde{\Phi} = \Delta\widehat{\Phi} + \widehat{W}, \quad \widehat{W}\widehat{W}^{\mathrm{T}} \leqslant \widehat{\Omega}\right\}$$

and $\Delta \in \mathbf{\Delta_p}$ iff

$$(\widetilde{\Phi} - \Delta\widehat{\Phi})(\widetilde{\Phi} - \Delta\widehat{\Phi})^{\mathrm{T}} \leqslant \widehat{\Omega}. \tag{2.6}$$

It is obvious that $\Delta_{real} \in \mathbf{\Delta_p}$. For further use, we represent this inequality as

$$(\Delta \quad I_{n_x+n_z})\, \Psi_1\, (\Delta \quad I_{n_x+n_z})^{\mathrm{T}} \leqslant 0, \tag{2.7}$$

where the symmetric matrix Ψ_1 of order $(2n_x + n_u + n_z)$ is partitioned into appropriate blocks as follows:

$$\Psi_1 = \left(\begin{array}{cc|cc} \Phi\Phi^{\mathrm{T}} & * & * & * \\ U\Phi^{\mathrm{T}} & UU^{\mathrm{T}} & * & * \\ \hline -\Phi_+\Phi^{\mathrm{T}} & -\Phi_+U^{\mathrm{T}} & \Phi_+\Phi_+^{\mathrm{T}} - \Omega & * \\ -Z\Phi^{\mathrm{T}} & -ZU^{\mathrm{T}} & Z\Phi_+^{\mathrm{T}} & ZZ^{\mathrm{T}} \end{array}\right). \tag{2.8}$$

Thus, the set of all matrices Δ consistent with the available experimental data satisfies inequality (2.7). The lemma below formulates boundedness conditions for the set $\mathbf{\Delta_p}$. Its proof is provided in the Appendix.

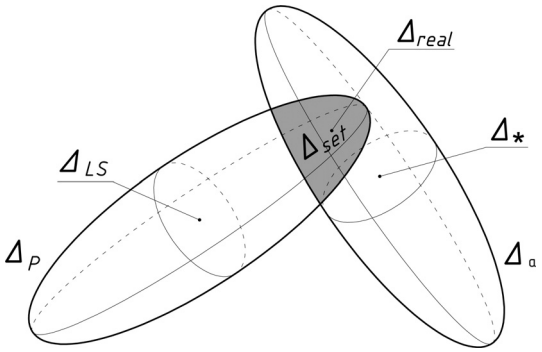


Fig. 1. The set $\mathbf{\Delta_{set}}$ of unknown parameters Δ consistent with experimental and a priori data.

Lemma 2.1. *If the information matrix $\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}}$ is nonsingular, then the set $\mathbf{\Delta_p}$ is a nondegenerate "matrix ellipsoid" centered at Δ_{LS} given by*

$$(\Delta - \Delta_{LS})(\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}})(\Delta - \Delta_{LS})^{\mathrm{T}} \leqslant \Gamma, \tag{2.9}$$

where

$$\Gamma = \widehat{\Omega} + \widetilde{\Phi}[\widehat{\Phi}^{\mathrm{T}}(\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}})^{-1}\widehat{\Phi} - I]\widetilde{\Phi}^{\mathrm{T}} \geqslant 0, \tag{2.10}$$

and $\Delta_{LS} = \widetilde{\Phi}\widehat{\Phi}^{\mathrm{T}}(\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}})^{-1}$ is the optimal least-squares estimate of the unknown matrix Δ_{real} in (2.3) *that minimizes the squared matrix norm of the residual $\|\widetilde{\Phi} - \Delta\widehat{\Phi}\|_F^2$ with respect to Δ.*

According to this lemma, given a nonsingular information matrix, the "size" of the set $\mathbf{\Delta_p}$ is determined by the regressor matrices $\widehat{\Phi}$ and ultimately depends on the real object, the controls U chosen in the experiment, and the disturbances W.

Now consider an additional information that the unknown matrix Δ_{real} satisfies the constraint

$$(\Delta - \Delta_*)(\Delta - \Delta_*)^{\mathrm{T}} \leqslant \rho^2 I, \quad \Delta_* = \begin{pmatrix} A_* & B_* \\ C_* & D_* \end{pmatrix} = \begin{pmatrix} \Delta_*^{(1)} \\ \Delta_*^{(2)} \end{pmatrix}, \tag{2.11}$$

where Δ_* and ρ are given matrix and parameter characterizing the center and size of the uncertainty domain. We write this inequality as

$$(\Delta \quad I_{n_x+n_z})\,\Psi_2\,(\Delta \quad I_{n_x+n_z})^{\mathrm{T}} \leqslant 0, \tag{2.12}$$

where

$$\Psi_2 = \left(\begin{array}{cc|cc} I_{n_x} & \star & \star & \star \\ 0_{n_u\times n_x} & I_{n_u} & \star & \star \\ \hline -A_* & -B_* & \Delta_*^{(1)}\Delta_*^{(1)\mathrm{T}} - \rho^2 I_{n_x} & \star \\ -C_* & -D_* & \Delta_*^{(2)}\Delta_*^{(1)\mathrm{T}} & \Delta_*^{(2)}\Delta_*^{(2)\mathrm{T}} - \rho^2 I_{n_z} \end{array}\right). \tag{2.13}$$

We introduce the following notations: $\mathbf{\Delta_a}$ is the set of matrices satisfying inequality (2.12), and $\mathbf{\Delta_{set}} = \mathbf{\Delta_p} \bigcap \mathbf{\Delta_a}$ is the set of matrices satisfying inequalities (2.7) and (2.12). Obviously, $\Delta_{real} \in \mathbf{\Delta_{set}}$ (see Fig. 1).

The quality of the closed loop system (2.1) with the linear state-feedback control law $u(t) = \Theta x(t)$ and a given matrix Δ will be evaluated by the damping level of the exogenous and initial disturbances, i.e., by the generalized H_∞ norm

$$\gamma_{g\infty}(\Delta,\Theta) = \sup_{x_0,\,w} \frac{\|z\|}{\left(x_0^{\mathrm{T}} R^{-1} x_0 + \|w\|^2\right)^{1/2}},$$

where $R = R^{\mathrm{T}} > 0$ is a weight matrix and $\|\xi\|^2 = \sum_{t=0}^{\infty} |\xi(t)|^2$ (in the discrete-time case) or $\|\xi\|^2 = \int_{t=0}^{\infty} |\xi(t)|^2$ (in the continuous-time case). If $w(t) \equiv 0$ (no exogenous disturbance), the generalized H_∞ norm turns into the so-called γ_0 norm given by

$$\gamma_0(\Delta,\Theta) = \sup_{x_0 \neq 0} \frac{\|z\|}{\left(x_0^{\mathrm{T}} R^{-1} x_0\right)^{1/2}}.$$

This norm characterizes the "worst" value of the quadratic functional on the system trajectories provided that the initial state is inside the ellipsoid $x^{\mathrm{T}} R^{-1} x \leqslant 1$. Under zero initial state, the generalized H_∞ norm (with $R \to 0$) turns into the conventional H_∞ norm:

$$\gamma_\infty(\Delta,\Theta) = \sup_{w \neq 0} \frac{\|z\|}{\|w\|}.$$

The quality of the closed-loop uncertain system (2.1) with the control law $u(t) = \Theta x(t)$ will be evaluated by the minimum upper bound of the damping level of the exogenous and initial disturbances, i.e., by the minimum upper bound of the generalized H_∞ norm for all object matrices consistent with experimental and a priori data:

$$\gamma_*(\Theta) = \sup_{\Delta \,\in\, \mathbf{\Delta_{set}}} \gamma_{g\infty}(\Delta,\Theta). \tag{2.14}$$

The robust generalized H_∞-optimal control law is defined as a control law with the parameter matrix Θ_* minimizing this bound, i.e., with the solution of the minimax problem

$$\inf_{\Theta} \sup_{\Delta \,\in\, \mathbf{\Delta_{set}}} \gamma_{g\infty}(\Delta,\Theta) = \inf_{\Theta} \gamma_*(\Theta) = \gamma_*(\Theta_*). \tag{2.15}$$

The problem is to design, directly from input and state measurements, a robust generalized H_∞-suboptimal control law with a parameter matrix Θ under which the generalized H_∞ norm of the closed loop system will be bounded by a given constant: $\gamma_*(\Theta) < \gamma$.

3. THE GENERALIZED H_∞ NORM IN TERMS OF THE DUAL SYSTEMS

Recall that

$$\gamma_{g\infty} = \sup_{x_0,\,v} \frac{\|z\|}{\left(x_0^{\mathrm{T}} R^{-1} x_0 + \|v\|^2\right)^{1/2}}, \tag{3.1}$$

the generalized H_∞ norm from the input v to the output z of a stable system

$$\begin{aligned} \partial x(t) &= \mathcal{A}x(t) + \mathcal{B}v(t), \\ z(t) &= \mathcal{C}x(t), \end{aligned} \tag{3.2}$$

satisfies the condition $\gamma_{g\infty} < \gamma$ iff the following LMIs are solvable in the matrix $Y = Y^{\mathrm{T}} > 0$:

$$\begin{pmatrix} Y\mathcal{A}^T + \mathcal{A}Y & \star & \star \\ \mathcal{B}^{\mathrm{T}} & -\gamma^2 I & \star \\ \mathcal{C}Y & 0 & -I \end{pmatrix} < 0, \quad \begin{pmatrix} Y & \star \\ I & \gamma^2 R^{-1} \end{pmatrix} > 0 \tag{3.3}$$

(for the continuous-time system) or

$$\begin{pmatrix} -Y & \star & \star & \star \\ Y\mathcal{A}^{\mathrm{T}} & -Y & \star & \star \\ \mathcal{B}^{\mathrm{T}} & 0 & -\gamma^2 I & \star \\ 0 & \mathcal{C}Y & 0 & -I \end{pmatrix} < 0, \quad \begin{pmatrix} Y & \star \\ I & \gamma^2 R^{-1} \end{pmatrix} > 0 \tag{3.4}$$

(for the discrete-time system). According to [11, 12], inequalities (3.3) and (3.4) mean that

$$\dot{V}(x) + |z|^2 - \gamma^2|v|^2 < 0 \text{ and } \Delta V(x) + |z|^2 - \gamma^2|v|^2 < 0 \quad \forall\, x, v, \text{ respectively,} \tag{3.5}$$

for a positive definite function $V(x) = x^{\mathrm{T}}Y^{-1}x$ with $Y > \gamma^{-2}R$ along the trajectories of system (3.2).

The next auxiliary result, proved in the Appendix, characterizes the generalized H_∞ norm of system (3.2) in terms of the dual system.

Lemma 3.1. *The generalized H_∞ norm of system* (3.2) *satisfies the condition $\gamma_{g\infty} < \gamma$ iff there exists a positive definite quadratic form $V_a(x_a) = x_a^{\mathrm{T}}Px_a$ with $P > R$ such that*

$$\begin{aligned} &\dot{V}_a(x_a(t)) + |z_a(t)|^2 - \gamma^2|v_a(t)|^2 < 0 \ \textit{or} \\ &\Delta V_a(x_a(t)) + |z_a(t)|^2 - \gamma^2|v_a(t)|^2 < 0, \ \textit{respectively,} \end{aligned} \tag{3.6}$$

along the trajectories of the dual system

$$\begin{aligned} \partial x_a(t) &= \mathcal{A}^{\mathrm{T}}x_a(t) + \mathcal{C}^{\mathrm{T}}v_a(t), \\ z_a(t) &= \mathcal{B}^{\mathrm{T}}x_a(t). \end{aligned} \tag{3.7}$$

Corollary 3.1. *For $v(t) \equiv 0$, the γ_0 norm of system* (3.2) *satisfies the condition $\gamma_0 < \gamma$ iff there exists a quadratic form $V_a(x_a) = x_a^{\mathrm{T}}Px_a$ with $P > R$ such that the corresponding inequality in* (3.6) *is valid for $z_a(t) \equiv 0$ along the trajectories of the dual system*

$$\partial x_a(t) = \mathcal{A}^{\mathrm{T}}x_a(t) + \mathcal{C}^{\mathrm{T}}v_a(t).$$

Remark 1. Formally, the dual system is described by the equations

$$\begin{aligned} \dot{\widehat{x}}_a &= -\mathcal{A}^{\mathrm{T}}\widehat{x}_a - \mathcal{C}^{\mathrm{T}}\widehat{v}_a, \\ \widehat{z}_a &= \mathcal{B}^{\mathrm{T}}\widehat{x}_a \end{aligned} \tag{3.8}$$

(in the continuous-time case) or

$$\begin{aligned} \widehat{x}_a(t) &= \mathcal{A}^{\mathrm{T}}\widehat{x}_a(t+1) + \mathcal{C}^{\mathrm{T}}\widehat{v}_a(t), \\ \widehat{z}_a(t) &= \mathcal{B}^{\mathrm{T}}\widehat{x}_a(t+1) \end{aligned} \tag{3.9}$$

(in the discrete-time case). By the proof of this lemma, from systems (3.8) and (3.9) we can pass to system (3.7), also called dual, which satisfies the corresponding inequality of (3.6).

Remark 2. The matrices of the quadratic forms $V(x) = x^{\mathrm{T}}Y^{-1}x$ and $V_a(x_a) = x_a^{\mathrm{T}}Px_a$ of the primal and dual systems have the relation $P = \gamma^2 Y$; see the proof of Lemma 3.1.

4. DESIGN OF GENERALIZED H_∞-SUBOPTIMAL CONTROLLERS

We describe the main steps for obtaining an upper bound of the generalized H_∞ norm and the corresponding parameter matrices Θ of control laws for the uncertain closed-loop system

$$\begin{aligned} \partial x(t) &= (A + B\Theta)x(t) + w(t), \\ z(t) &= (C + D\Theta)x(t). \end{aligned} \tag{4.1}$$

Assume that the closed loop system with the parameters Θ is stable. With the notations introduced above, these equations can be written as

$$\begin{aligned} \partial x(t) &= (I_{n_x}\ 0_{n_x\times n_z})\,\Delta \begin{pmatrix} I_{n_x} \\ \Theta \end{pmatrix} x(t) + w(t), \\ z(t) &= (0_{n_z\times n_x}\ I_{n_z})\,\Delta \begin{pmatrix} I_{n_x} \\ \Theta \end{pmatrix} x(t), \end{aligned} \tag{4.2}$$

where Δ is an unknown matrix of dimensions $(n_x+n_z)\times(n_x+n_u)$ and Θ is the controller's parameter matrix of dimensions $(n_u \times n_x)$. Due to Lemma 3.1, the dual continuous- and discrete-time systems are described by the equations

$$\begin{aligned} &\partial x_a(t) = \begin{pmatrix} I \\ \Theta \end{pmatrix}^{\mathrm{T}} \Delta^{\mathrm{T}} \begin{pmatrix} I \\ 0 \end{pmatrix} x_a(t) + \begin{pmatrix} I \\ \Theta \end{pmatrix}^{\mathrm{T}} \Delta^{\mathrm{T}} \begin{pmatrix} 0 \\ I \end{pmatrix} w_a(t), \\ &z_a(t) = x_a(t). \end{aligned} \tag{4.3}$$

We define an augmented system with an additional artificial input $w_\Delta(t) \in L_2(l_2)$ and an output $z_\Delta(t)$ in both cases as follows:

$$\begin{aligned} &\partial\widehat{x}(t) = \begin{pmatrix} I \\ \Theta \end{pmatrix}^{\mathrm{T}} w_\Delta(t), \\ &\widehat{z}(t) = \widehat{x}(t), \quad z_\Delta(t) = \begin{pmatrix} I \\ 0 \end{pmatrix} \widehat{x}(t) + \begin{pmatrix} 0 \\ I \end{pmatrix} \widehat{w}(t), \end{aligned} \tag{4.4}$$

where $\widehat{x}(t)$ is the state variable, $\widehat{w}(t)$ is a disturbance, and $\widehat{z}(t)$ is the target output. Note that for $w_\Delta(t) = \Delta^{\mathrm{T}} z_\Delta(t)$, equations (4.4) coincide with the equations of system (4.3). For all $t \geqslant 0$, let the additional input and output signals in system (4.4) satisfy the two inequalities

$$\begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix}^{\mathrm{T}} \Psi_1 \begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix} \leqslant 0, \quad \begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix}^{\mathrm{T}} \Psi_2 \begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix} \leqslant 0 \tag{4.5}$$

where the matrices Ψ_1 and Ψ_2 are given by (2.8) and (2.13). We denote by $\mathbf{W_\Delta}$ the set of all such signals $w_\Delta(t)$. According to (2.7) and (2.12), for $w_\Delta(t) = \Delta^{\mathrm{T}} z_\Delta(t)$ and all $\Delta \in \mathbf{\Delta_{set}}$,

$$\begin{aligned} &\begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix}^{\mathrm{T}} \Psi_1 \begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix} = z_\Delta^{\mathrm{T}}(t) \begin{pmatrix} \Delta^{\mathrm{T}} \\ I \end{pmatrix}^{\mathrm{T}} \Psi_1 \begin{pmatrix} \Delta^{\mathrm{T}} \\ I \end{pmatrix} z_\Delta(t) \leqslant 0, \\ &\begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix}^{\mathrm{T}} \Psi_2 \begin{pmatrix} w_\Delta(t) \\ z_\Delta(t) \end{pmatrix} = z_\Delta^{\mathrm{T}}(t) \begin{pmatrix} \Delta^{\mathrm{T}} \\ I \end{pmatrix}^{\mathrm{T}} \Psi_2 \begin{pmatrix} \Delta^{\mathrm{T}} \\ I \end{pmatrix} z_\Delta(t) \leqslant 0. \end{aligned}$$

Thus, $w_\Delta(t) = \Delta^{\mathrm{T}} z_\Delta(t) \in \mathbf{W_\Delta}$ and consequently, system (4.3) with $\Delta \in \mathbf{\Delta_{set}}$, dual to the original uncertain system, is immersed into the augmented system (4.4), (4.5). In view of Lemma 3.1, this fact can be used to derive an upper bound of the generalized H_∞ norm of the uncertain system through the corresponding property of the augmented system.

Theorem 4.1. *The upper bound of the generalized H_∞ norm of the uncertain system* (2.1) *with the control law $u(t) = \Theta x(t)$, $\Theta = QP^{-1}$, is less than γ if the following LMIs are solvable in $P = P^{\mathrm{T}} > 0$, Q, $\mu_1 \geqslant 0$, and $\mu_2 \geqslant 0$:*

$$\begin{pmatrix} I - \sum_{i=1}^{2} \mu_i \Xi_{11}^{(i)} & \star & \star & \star \\ -\sum_{i=1}^{2} \mu_i \Xi_{21}^{(i)} & -\sum_{i=1}^{2} \mu_i \Xi_{22}^{(i)} - \gamma^2 I & \star & \star \\ P - \sum_{i=1}^{2} \mu_i \Xi_{31}^{(i)} & -\sum_{i=1}^{2} \mu_i \Xi_{32}^{(i)} & -\sum_{i=1}^{2} \mu_i \Xi_{33}^{(i)} & \star \\ Q - \sum_{i=1}^{2} \mu_i \Xi_{41}^{(i)} & -\sum_{i=1}^{2} \mu_i \Xi_{42}^{(i)} & -\sum_{i=1}^{2} \mu_i \Xi_{43}^{(i)} & -\sum_{i=1}^{2} \mu_i \Xi_{44}^{(i)} \end{pmatrix} < 0 \tag{4.6}$$

(for the continuous-time system) or

$$\begin{pmatrix} -P & \star & \star & \star & \star \\ 0 & -P + I - \sum_{i=1}^{2} \mu_i \Xi_{11}^{(i)} & \star & \star & \star \\ 0 & -\sum_{i=1}^{2} \mu_i \Xi_{21}^{(i)} & -\sum_{i=1}^{2} \mu_i \Xi_{22}^{(i)} - \gamma^2 I & \star & \star \\ P & -\sum_{i=1}^{2} \mu_i \Xi_{31}^{(i)} & -\sum_{i=1}^{2} \mu_i \Xi_{32}^{(i)} & -\sum_{i=1}^{2} \mu_i \Xi_{33}^{(i)} & \star \\ Q & -\sum_{i=1}^{2} \mu_i \Xi_{41}^{(i)} & -\sum_{i=1}^{2} \mu_i \Xi_{42}^{(i)} & -\sum_{i=1}^{2} \mu_i \Xi_{43}^{(i)} & -\sum_{i=1}^{2} \mu_i \Xi_{44}^{(i)} \end{pmatrix} < 0 \tag{4.7}$$

(for the discrete-time system), where $P > R$,

$$\Xi_{11}^{(1)} = \Phi_+ \Phi_+^{\mathrm{T}} - \Omega, \quad \Xi_{21}^{(1)} = Z\Phi_+^{\mathrm{T}}, \quad \Xi_{22}^{(1)} = ZZ^{\mathrm{T}},$$
$$\Xi_{31}^{(1)} = -\Phi\Phi_+^{\mathrm{T}}, \quad \Xi_{32}^{(1)} = -\Phi Z^{\mathrm{T}}, \quad \Xi_{33}^{(1)} = \Phi\Phi^{\mathrm{T}},$$
$$\Xi_{41}^{(1)} = -U\Phi_+^{\mathrm{T}}, \quad \Xi_{42}^{(1)} = -UZ^{\mathrm{T}}, \quad \Xi_{43}^{(1)} = U\Phi^{\mathrm{T}}, \quad \Xi_{44}^{(1)} = UU^{\mathrm{T}},$$
$$\Xi_{11}^{(2)} = \Delta_*^{(1)}\Delta_*^{(1)\mathrm{T}} - \rho^2 I_{n_x}, \quad \Xi_{21}^{(2)} = \Delta_*^{(2)}\Delta_*^{(1)\mathrm{T}}, \quad \Xi_{22}^{(2)} = \Delta_*^{(2)}\Delta_*^{(2)\mathrm{T}} - \rho^2 I_{n_z},$$
$$\Xi_{31}^{(2)} = -A_*^{\mathrm{T}}, \quad \Xi_{32}^{(2)} = -C_*^{\mathrm{T}}, \quad \Xi_{33}^{(2)} = I_{n_x},$$
$$\Xi_{41}^{(2)} = -B_*^{\mathrm{T}}, \quad \Xi_{42}^{(2)} = -D_*^{\mathrm{T}}, \quad \Xi_{43}^{(2)} = 0_{n_u \times n_x}, \quad \Xi_{44}^{(2)} = I_{n_u}.$$

Proof of Theorem 4.1. We establish conditions for the existence of a positive definite quadratic function $\widehat{V}(\widehat{x}) = \widehat{x}^{\mathrm{T}} P \widehat{x}$ with $P > R$ that satisfies the corresponding inequality in (3.6) along the trajectories of the augmented system (4.4) for all $w_\Delta(t)$ with (4.5). By the S-procedure, a sufficient condition is the existence of a function $\widehat{V}(\widehat{x}) = \widehat{x}^{\mathrm{T}} P \widehat{x}$ with $P > R$ that satisfies the corresponding

inequality

$$\begin{aligned}\dot{\widehat{V}}(\widehat{x})+|\widehat{z}|^2-\gamma^2|\widehat{w}|^2-\sum_{i=1}^{2}\mu_i\begin{pmatrix}w_\Delta\\ z_\Delta\end{pmatrix}^{\mathrm{T}}\Psi_i\begin{pmatrix}w_\Delta\\ z_\Delta\end{pmatrix}<0,\\ \Delta\widehat{V}(\widehat{x})+|\widehat{z}|^2-\gamma^2|\widehat{w}|^2-\sum_{i=1}^{2}\mu_i\begin{pmatrix}w_\Delta\\ z_\Delta\end{pmatrix}^{\mathrm{T}}\Psi_i\begin{pmatrix}w_\Delta\\ z_\Delta\end{pmatrix}<0\end{aligned}\tag{4.8}$$

along the trajectories of system (4.4) for all $\widehat{x}$, $\widehat{w}$, w_Δ, and some $\mu_1\geqslant 0$ and $\mu_2\geqslant 0$.

These inequalities reduce to the following inequalities for the quadratic forms in the variables $\widehat{x}$, $\widehat{w}$, and w_Δ :

$$\begin{aligned}2\widehat{x}^{\mathrm{T}}P\begin{pmatrix}I\\ \Theta\end{pmatrix}^{\mathrm{T}}w_\Delta+|\widehat{z}|^2-\gamma^2|\widehat{w}|^2-\sum_{i=1}^{2}\mu_i\begin{pmatrix}w_\Delta\\ z_\Delta\end{pmatrix}^{\mathrm{T}}\Psi_i\begin{pmatrix}w_\Delta\\ z_\Delta\end{pmatrix}<0,\\ w_\Delta^{\mathrm{T}}\begin{pmatrix}I\\ \Theta\end{pmatrix}P\begin{pmatrix}I\\ \Theta\end{pmatrix}^{\mathrm{T}}w_\Delta-\widehat{x}^{\mathrm{T}}P\widehat{x}+|\widehat{z}|^2-\gamma^2|\widehat{w}|^2-\sum_{i=1}^{2}\mu_i\begin{pmatrix}w_\Delta\\ z_\Delta\end{pmatrix}^{\mathrm{T}}\Psi_i\begin{pmatrix}w_\Delta\\ z_\Delta\end{pmatrix}<0,\end{aligned}\tag{4.9}$$

where $\widehat{z}=\widehat{x}$ and $z_\Delta=\operatorname{col}(\widehat{x},\widehat{w})$. System (4.3), dual to the original one (4.2), is immersed into the augmented system, and condition (4.5) holds. Therefore, we have inequality (3.6) along the trajectories of (4.3) for all $\Delta\in\mathbf{\Delta_{set}}$. By Lemma 3.1, for any $\Delta\in\mathbf{\Delta_{set}}$, the original uncertain system satisfies $\gamma_{g\infty}(\Delta,\Theta)<\gamma$ and consequently, $\gamma_*(\Theta)<\gamma$. Finally, we write inequalities (4.9) for the quadratic forms as matrix inequalities, introduce the new matrix variable $Q=\Theta P$, and apply Schur's complement lemma to get the LMIs (4.6) and (4.7), respectively. The proof of Theorem 4.1 is complete.

Remark 3. To find the upper bound of the γ_0 norm, it is necessary to eliminate the term I from the block located in the first row and first column of inequalities (4.6) (for the continuous-time system) or from the block in the second row and second column of inequalities (4.7) (for the discrete-time system). This follows from the fact that in the case of the γ_0 norm, the term $|\widehat{z}|^2$ vanishes in inequalities (4.8) and, accordingly, in inequalities (4.9). To find the upper bound of the conventional H_∞ norm, we should use Theorem 4.1 with $R=0$.

Remark 4. According to the lossless S-procedure under two quadratic constraints (Theorem 4.1 in [13]), if $\mu_1\Psi_1+\mu_2\Psi_2>0$ for some μ_1 and μ_2 (this LMI can be directly solved with respect to μ_1 and μ_2), then the corresponding inequality (4.8) is a sufficient and also necessary condition for the existence of the above function $\widehat{V}(\widehat{x})=\widehat{x}^{\mathrm{T}}P\widehat{x}$ for the augmented system.

The minimum value of γ for which each of inequalities (4.6) or (4.7) is solvable will be denoted by $\gamma_{rob}(\Theta_{rob})$, where Θ_{rob} is the corresponding control parameter matrix. Since

$$\gamma_*(\Theta_*)\leqslant\gamma_*(\Theta_{rob})\leqslant\gamma_{rob}(\Theta_{rob}),$$

where Θ_* is the parameter matrix of the robust generalized H_∞-optimal control law (2.15), then $\gamma_{rob}(\Theta_{rob})$ is the upper bound of the minimum damping level of the disturbances in the uncertain system with the robust generalized H_∞-optimal control law under given a priori and experimental data. In addition, Theorem 4.1 can be used to find out whether the guaranteed generalized H_∞ norm of the closed-loop uncertain system (4.1) with the feedback parameter matrix $\widehat{\Theta}$ is less than a given number γ^2. For this purpose, we should let $Q=\widehat{\Theta}P$ in inequality (4.6) for the continuous-time system or inequality (4.7) for the discrete-time system and solve the resulting inequality with respect to the variables P, μ_1, and μ_2.

5. AN ILLUSTRATIVE EXAMPLE

To illustrate the approach, we consider a discrete-time object of the form (2.1) of the fifth order ($n_x = 5$) with two control actions ($n_u = 2$), a five-dimensional disturbance ($n_w = 5$), and two target outputs ($n_z = 2$) with matrices whose elements were chosen randomly on the interval $[-1, 1]$. Thus, the system contains 49 unknown parameters. In the experiment, the initial conditions and the components of the control vector were chosen randomly on the interval $[-1, 1]$, and the disturbance was chosen randomly on the interval $[-d, d]$. In total, $N = 50$ measurements were taken. The weight matrix of the initial disturbance is $R = 0.01I_5$. Figure 2 shows three typical graphs of the squared damping levels of the disturbances in the closed loop system with the control law designed from the experimental data only, depending on the disturbance level d in the experiment. The solid curve corresponds to the square $\gamma^2_{rob}(\Theta_{rob})$ of the guaranteed generalized H_∞ norm under the control law with the parameter matrix Θ_{rob} obtained by solving the LMIs (4.7) with the minimum value of γ^2. The dashed-dotted curve corresponds to the square of the damping level $\gamma_{real} = \gamma_{g\infty}(\Delta_{real}, \Theta_{rob})$ of the disturbances, i.e., the generalized H_∞ norm of the closed loop system composed of the real object with the parameter matrix Δ_{real} (if it were known) and the feedback loop with the parameters Θ_{rob}. The dotted curve corresponds to the square of the damping level $\gamma_{prob} = \gamma_{g\infty}(\Delta_{prob}, \Theta_{rob})$ of the disturbances, i.e., the generalized H_∞ norm of the closed loop system composed of the trial object with the matrix $\Delta_{prob} = \Delta_{LS} + \Gamma^{1/2}(\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}})^{-1/2}$, which lies on the boundary of the uncertainty ellipsoid $\mathbf{\Delta_{set}}$ (see Lemma 2.1), and the feedback loop with the parameters Θ_{rob}. The growth of these curves with increasing the disturbance level d in the experiment can be explained as follows: for a higher value of d, we obtain a greater ellipsoid $\mathbf{\Delta_p}$ of the unknown parameters Δ consistent with the experimental data.

According to Fig. 2, first, the curve γ^2_{rob} majorizes with some margin the damping levels of the disturbances in the closed loop system for particular objects with the matrices Δ_{real} and Δ_{prob} from the set $\mathbf{\Delta_{set}}$; second, under the control law with the parameter matrix Θ_{rob}, the generalized H_∞ norms of the closed loop systems slightly exceed (especially at small perturbation levels d) their minimum values $\gamma^2 \simeq 0.26$ for the completely known model. Note that the margin by which γ^2_{rob}

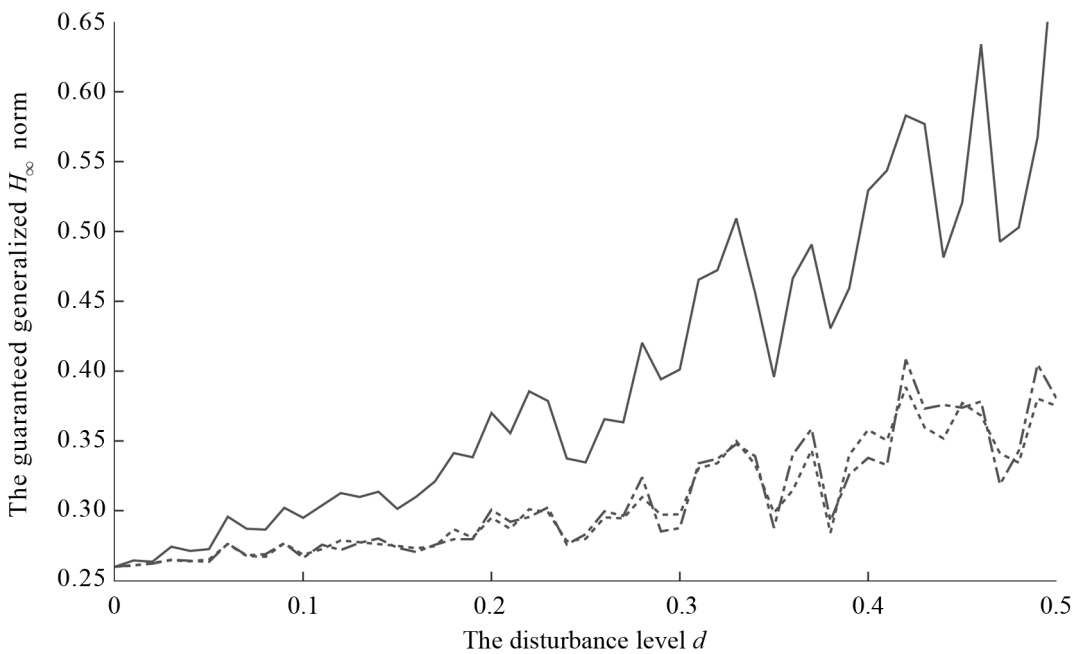


Fig. 2. The guaranteed generalized H_∞ norm and the generalized H_∞ norms for the real and trial objects as functions of the disturbance level in measurements.

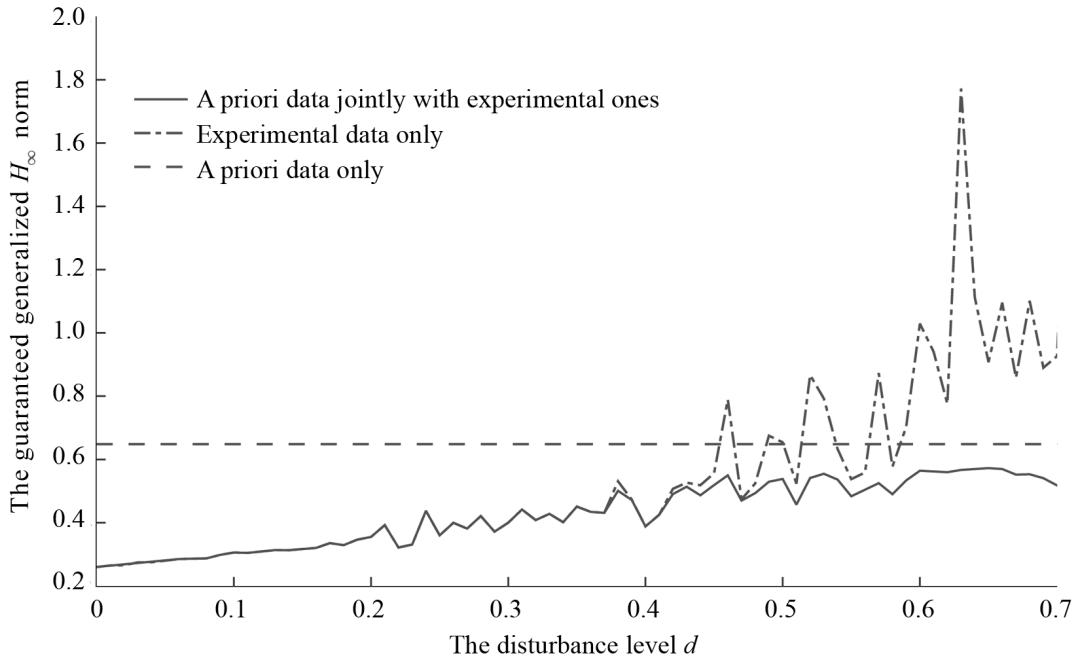


Fig. 3. The guaranteed estimates of the H_∞ norm as functions of the disturbance level in experimental data for different types of available information.

exceeds γ^2_{real} and γ^2_{prob} substantially depends on the experimental data and can be much smaller than on the graphs in Fig. 2.

Figure 3 presents the three guaranteed estimates of the generalized H_∞ norm based on different information (a priori data only, experimental data only, and a priori data jointly with experimental ones) as a function of the disturbance level d in the experiment. The a priori information was that the unknown matrices of the system satisfy condition (2.11) with $\rho = 0.1$ and $A_* = A_{real} + (\rho/2)I$, $B_* = B_{real}$, $C_* = C_{real}$, and $D_* = D_{real}$. Starting from some disturbance level in the experiment, the guaranteed estimates of the norms of the closed-loop uncertain system designed using both a priori and experimental data are much smaller than the corresponding estimates of the norms of the closed loop system with the control laws designed using only a priori or only experimental data.

Finally, we note the following aspect as well. Consider the object with the matrices A_{LS}, B_{LS}, C_{LS}, and D_{LS} constituting the parameter matrix Δ_{LS} obtained by the least squares method from the same experimental data. For this object, let us find the parameter matrix $\Theta_{LS} = QP^{-1}$ of the generalized H_∞-optimal feedback loop by solving the LMIs (3.4) with $Y = P$, $\mathcal{A}Y = A_{LS}P + B_{LS}Q$, $\mathcal{C}Y = C_{LS}P + D_{LS}Q$, and $\mathcal{B} = I$. This is essentially the so-called indirect H_∞-suboptimal adaptive control, i.e., the control law determined by estimating the unknown parameters of the object. If there are sufficiently many measurements and the information matrix is nonsingular, the generalized H_∞ norm of the closed loop system consisting of the real object and the feedback loop with $\Theta = \Theta_{LS}$ may be smaller than the corresponding guaranteed generalized H_∞ norm under the feedback loop with $\Theta = \Theta_{rob}$. In the latter case, we have an upper bound of the generalized H_∞ norm of the closed loop system for any object from the set $\mathbf{\Delta_{set}}$ consistent with the experimental data; however, for the feedback loop with $\Theta = \Theta_{LS}$, such an estimate can be obtained from inequalities (4.7) only under very small disturbance levels d (see the experimental results). In the example under consideration, for $d = 0.02$ we have $\gamma^2_{rob}(\Theta_{rob}) = 0.27$, whereas $\gamma^2_{rob}(\Theta_{LS}) = 41.77$; for $d > 0.03$, inequality (4.7) with $Q = \Theta_{LS}P$ becomes unsolvable.

6. CONCLUSIONS

This paper has been devoted to constructing generalized H_∞-suboptimal (as a special case, linear-quadratic) control laws for linear continuous- and discrete-time dynamic objects without precise mathematical models. As has been demonstrated above, for dynamic objects whose equations contain unknown parameters in some bounded sets, classical robust control methods based on a priori data can be applied, after an appropriate modification, to control design from a priori and experimental data. These methods consist in immersing an uncertain system into some enlarged system with additional input and output satisfying a quadratic inequality, applying the S-procedure, and reducing the problem to the design of H_∞-optimal control for the enlarged system. The modification is to characterize the control criterion (the generalized H_∞ norm of the system under initial and exogenous disturbances or the value of the quadratic functional under the initial disturbance only) in terms of the dual system, immerse the dual uncertain system into some enlarged system, and apply the technique of LMIs. As a result, the parameters of linear suboptimal feedback loops are expressed in terms of solutions of LMIs containing only a priori and experimental data. An illustrative example with a randomly generated fifth-order object has been provided to demonstrate that when a priori and experimental data are applied together, the quality of the control system is improved significantly.

FUNDING

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

APPENDIX

Proof of Lemma 2.1. For the unknown matrix Δ_{real} in (2.3), we define the least-squares estimate Δ_{LS} minimizing the squared matrix norm of the residual with respect to Δ, i.e., the function $\|\widetilde{\Phi} - \Delta\widehat{\Phi}\|_F^2 = \operatorname{tr}(\widetilde{\Phi} - \Delta\widehat{\Phi})^{\mathrm{T}}(\widetilde{\Phi} - \Delta\widehat{\Phi})$. Equating the gradient of this function with respect to Δ to zero, $-2\widetilde{\Phi}\widehat{\Phi}^{\mathrm{T}} + 2\Delta\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}} = 0$, yields the optimal estimate $\Delta_{LS} = \widetilde{\Phi}\widehat{\Phi}^{\mathrm{T}}(\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}})^{-1}$ under the assumption that the information matrix $\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}}$ is nonsingular. Next, we transform inequality (2.6) to

$$\Delta\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}}\Delta^{\mathrm{T}} - \widetilde{\Phi}\widehat{\Phi}^{\mathrm{T}}\Delta^{\mathrm{T}} - \Delta\widehat{\Phi}\widetilde{\Phi}^{\mathrm{T}} + \widetilde{\Phi}\widetilde{\Phi}^{\mathrm{T}} - \widehat{\Omega} \leqslant 0,$$

writing the result as

$$[\Delta - \widetilde{\Phi}\widehat{\Phi}^{\mathrm{T}}(\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}})^{-1}](\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}})[\Delta - \widetilde{\Phi}\widehat{\Phi}^{\mathrm{T}}(\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}})^{-1}]^{\mathrm{T}} \leqslant \Gamma,$$

where $\Gamma = \widehat{\Omega} + \widetilde{\Phi}[\widehat{\Phi}^{\mathrm{T}}(\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}})^{-1}\widehat{\Phi} - I]\widetilde{\Phi}^{\mathrm{T}}$. Substituting the expression for $\widetilde{\Phi}$ (2.3) into Γ and using (2.5) finally give

$$\Gamma = \widehat{\Omega} + \widehat{W}[\widehat{\Phi}^{\mathrm{T}}(\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}})^{-1}\widehat{\Phi} - I]\widehat{W}^{\mathrm{T}} \geqslant \widehat{W}\widehat{\Phi}^{\mathrm{T}}(\widehat{\Phi}\widehat{\Phi}^{\mathrm{T}})^{-1}\widehat{\Phi}\widehat{W}^{\mathrm{T}} \geqslant 0.$$

Proof of Lemma 3.1. We define a linear operator Γ mapping the pair $(x(0), v(t)) \in \mathrm{R}^n \times L_2(l_2) = \Xi$ (the initial state of the system and the input disturbance) into the target output $z(t) \in L_2(l_2) = \Upsilon$, i.e.,

$$\Gamma : \Xi = \mathrm{R}^{n_x} \times L_2(l_2) \to \Upsilon = L_2(l_2) : (x(0), v) \to z.$$

The inner products in these spaces are given by

$$\langle \cdot, \cdot \rangle_\Xi = x_1^{\mathrm{T}}(0)R^{-1}x_2(0) + \langle v_1(t), v_2(t) \rangle_{L_2(l_2)}, \quad \langle \cdot, \cdot \rangle_\Upsilon = \langle z_1(t), z_2(t) \rangle_{L_2(l_2)}.$$

Moreover, the generalized H_∞ norm coincides with the induced norm of this operator since

$$\|\Gamma\| = \sup_{(x_0,\,v)\neq 0} \frac{\|\Gamma(x_0,\,v)\|}{\|(x_0,\,v)\|} = \sup_{x_0,\,v\neq 0} \frac{\|z\|}{\left(x_0^{\mathrm{T}} R^{-1} x_0 + \|v\|^2\right)^{1/2}} = \gamma_{g\infty}.$$

We show that the adjoint operator Γ^* is given by

$$\Gamma^* : \Upsilon \to \Xi : \widehat{v}_a(t) \to (R\widehat{x}_a(0), \widehat{z}_a(t)),$$

where $\widehat{x}_a(t)$ and $\widehat{z}_a(t)$ satisfy equations (3.8) and (3.9) in the continuous- and discrete-time cases, respectively.

Indeed, for the continuous-time system, from equations (3.8) it follows that

$$\frac{d(x^{\mathrm{T}}\widehat{x}_a)}{dt} = v^{\mathrm{T}}\widehat{z}_a - z^{\mathrm{T}}\widehat{v}_a;$$

for the discrete-time system (see equations (3.9)),

$$x^{\mathrm{T}}(t+1)\widehat{x}_a(t+1) - x^{\mathrm{T}}(t)\widehat{x}_a(t) = v^{\mathrm{T}}(t)\widehat{z}_a(t) - z^{\mathrm{T}}(t)\widehat{v}_a(t).$$

Integrating in the former case or summing in the latter one, we obtain

$$< z, \widehat{v}_a >= x^{\mathrm{T}}(0)R^{-1}[R\widehat{x}_a(0)] + < v, \widehat{z}_a > .$$

Thus,

$$< \Gamma(x(0), v), \widehat{v}_a >_{\Upsilon} = < (x(0), v), \Gamma^*(\widehat{v}_a) >_{\Xi} .$$

Because the norms of the adjoint operators are equal,

$$\|\Gamma\| = \|\Gamma^*\| = \sup_{\widehat{v}_a\neq 0} \frac{\left(\|\widehat{z}_a\|^2 + \widehat{x}_a^{\mathrm{T}}(0)R\widehat{x}_a(0)\right)^{1/2}}{\|\widehat{v}_a\|}.$$

Next, we establish that $\|\Gamma^*\| < \gamma$ iff there exists a function $V(\widehat{x}_a) = \widehat{x}_a^{\mathrm{T}} P \widehat{x}_a$ with $P > R$ such that

$$\begin{gathered} \dot{V}(\widehat{x}_a(t)) - |\widehat{z}_a(t)|^2 + \gamma^2 |\widehat{v}_a(t)|^2 > 0 \text{ or} \\ \Delta V(\widehat{x}_a(t)) - |\widehat{z}_a(t)|^2 + \gamma^2 |\widehat{v}_a(t)|^2 > 0 \end{gathered} \tag{A.1}$$

along the trajectories of the continuous-time system (3.8) or along the trajectories of the discrete-time system (3.9), respectively.

Indeed, integrating the former inequality or summing the latter one with $P > R$, we arrive at $\|\widehat{z}_a\|^2 + \widehat{x}_a^{\mathrm{T}}(0)R\widehat{x}_a(0) < \gamma^2\|\widehat{v}_a\|^2$ for all $\widehat{v}_a(t)$, i.e., $\|\Gamma^*\| < \gamma$. Conversely, let $\|\Gamma^*\| < \gamma$, which implies $\|\Gamma\| < \gamma$. According to [11, 12], this means the existence of a function $V(x) = x^{\mathrm{T}} Y^{-1} x$ with a matrix Y satisfying inequalities (3.3) in the continuous-time case or inequalities (3.4) in the discrete-time case. Here, we consider the former case only: the proof for the discrete-time system is analogous. Using Schur's complement lemma, the first inequality in (3.3) can be transformed to

$$\begin{pmatrix} Y\mathcal{A}^T + \mathcal{A}Y + \gamma^{-2}\mathcal{B}\mathcal{B}^{\mathrm{T}} & \star \\ \mathcal{C}Y & -I \end{pmatrix} < 0.$$

With the change of variables $Y = \gamma^{-2}P$, this condition is equivalently written as the following inequality for the quadratic form in the abstract variables $\widehat{x}_a$ and $\widehat{v}_a$:

$$2\widehat{x}_a^{\mathrm{T}} P(-\mathcal{A}^T\widehat{x}_a - \mathcal{C}^{\mathrm{T}}\widehat{v}_a) - \widehat{x}_a^{\mathrm{T}}\mathcal{B}\mathcal{B}^{\mathrm{T}}\widehat{x}_a + \gamma^2\widehat{v}_a^{\mathrm{T}}\widehat{v}_a > 0.$$

It obviously coincides with the first inequality in (A.1). Due to $Y > \gamma^{-2}R$, the function $V_a(\widehat{x}_a) = \widehat{x}_a^{\mathrm{T}} P \widehat{x}_a$ with $P > R$ satisfies the first inequality in (A.1) along the trajectories of system (3.8). Thus, $\|\Gamma^*\| < \gamma$ and consequently, $\|\Gamma\| = \gamma_{g\infty} < \gamma$ iff the corresponding inequality in (A.1) holds along the trajectories of system (3.8) or (3.9). Reverting the time, we finally pass from equations (3.8) or (3.9) to system (3.7), along whose trajectories the function $V(x_a) = x_a^{\mathrm{T}} P x_a$ will satisfy the corresponding inequality in (3.6).

REFERENCES

1. Willems, J.C., Rapisarda, P., Markovsky, I., and De Moor, B., A Note on Persistency of Excitation, *Syst. Control Lett.*, 2005, vol. 54, pp. 325–329.
2. De Persis, C. and Tesi, P., Formulas for Data-Driven Control: Stabilization, Optimality and Robustness, *IEEE Trans. Automat. Control*, 2020, vol. 65, no. 3, pp. 909–924.
3. Waarde, H.J., Eising, J., Trentelman, H.L., and Camlibel, M.K., Data Informativity: a New Perspective on Data-Driven Analysis and Control, *IEEE Trans. Automat. Control*, 2020, vol. 65, no. 11, pp. 4753–4768.
4. Waarde, H.J., Camlibel, M.K., and Mesbahi, M., From Noisy Data to Feedback Controllers: Nonconservative Design via a Matrix S-Lemma, *IEEE Trans. Automat. Control*, 2022, vol. 67, no. 1, pp. 162–175.
5. Yakubovich, V.A., S-procedure in Nonlinear Control Theory, *Vestn. Leningrad. Univ. Mat.*, 1977, vol. 4, pp. 73–93.
6. Bisoffi, A., De Persis, C., and Tesi, P., Data-Driven Control via Petersen's Lemma, *Automatica*, 2022, vol. 145, art. no. 110537.
7. Petersen, I.R., A Stabilization Algorithm for a Class of Uncertain Linear Systems, *Syst. Control Lett.*, 1987, vol. 8, pp. 351–357.
8. Berberich, J., Scherer, C.W., and Allgower, F., Combining Prior Knowledge and Data for Robust Controller Design, *IEEE Trans. Automat. Control*, 2023, vol. 68, no. 8, pp. 4618–4633.
9. Kogan, M.M. and Stepanov, A.V., Design of Suboptimal Robust Controllers Based on A Priori and Experimental Data, *Autom. Remote Control*, 2023, vol. 84, no. 8, pp. 918–932.
10. Petersen, I.R. and Tempo, R., Robust Control of Uncertain Systems: Classical Results and Recent Developments, *Automatica*, 2014, vol. 50, no. 5, pp. 1315–1335.
11. Balandin, D.V. and Kogan, M.M., Generalized H_∞-optimal Control as a Trade-off between the H_∞-optimal and γ-optimal Controls, *Autom. Remote Control*, 2010, vol. 71, no. 6, pp. 993–1010.
12. Balandin, D.V., Biryukov, R.S., and Kogan, M.M., Pareto Suboptimal H_∞ Controls with Transients, *Proc. Eur. Control Conf.*, 2021, Rotterdam, pp. 542–547.
13. Polyak, B.T., Convexity of Quadratic Transformations and Its Use in Control and Optimization, *J. Optim. Theory Appl.*, 1998, vol. 99, no. 3, pp. 553–583.

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, 2024, Vol. 85, No. 1, pp. 15–34.

NONLINEAR SYSTEMS

Inserting a Maximum-Mass Spacecraft into a Target Orbit Using a Limited-Thrust Engine with Releasing the Separable Part of Its Launch Vehicle into the Earth's Atmosphere

I. S. Grigoriev[*,a] **and A. I. Proskuryakov**[**,b]

[*] *Moscow State University, Moscow, Russia*
[**] *Moscow State University, Baku Branch, Baku, Azerbaijan*
e-mail: [a] *iliagri@yandex.ru,* [b] *ap_91@mail.ru*

Received August 24, 2023
Revised November 6, 2023
Accepted December 21, 2023

Abstract—Space debris is an urgent problem of our time. This paper considers the idea of reducing near-Earth space debris by releasing the spent additional fuel tank (AFT) and the booster's central block (CB) into the Earth's atmosphere. The spacecraft transfer from a reference circular orbit of an artificial Earth satellite to a target elliptical orbit is optimized. The transition maneuvers are carried out using a booster with a high limited-thrust engine and the AFT. The second zonal harmonic of the Earth's gravitational field is taken into account. The optimal control problem is solved based on Pontryagin's maximum principle. Bulky derivatives are calculated using a specially developed numerical-analytical differentiation technique. The Pontryagin extremals obtained below are the next step in implementing the problem hierarchy methodology.

Keywords: spacecraft, space debris, additional fuel tank, booster, limited-thrust problem, transfer optimization, release into the Earth's atmosphere, numerical-analytical differentiation

DOI: 10.31857/S0005117924010027

1. INTRODUCTION

Space debris is an urgent problem of our time. The approaches to solving this problem can be divided into two large groups: prevention and cleaning. A detailed literature review on the topic was presented in [1]. In addition, we mention the works [2, 3], covering the monitoring issues of man-made space debris, and [4–7], describing different means of capture and removal of large-size space debris.

This paper considers the idea of reducing near-Earth space debris (an approach from the prevention group) by releasing the spent parts of spacecraft launch vehicles in orbits touching the conditional boundary of the atmosphere at the transition maneuvers stage of insertion into the target orbit. The problem under study is to optimize the spacecraft transfer from a reference circular orbit of an artificial Earth satellite of a given radius and inclination to a target elliptical orbit using a booster with a high limited-thrust engine and an additional fuel tank (AFT), with releasing the AFT and the booster's central block (CB) into the Earth's atmosphere. The final ascent maneuver from the target orbit to the geostationary orbit (GEO) is considered within the simplified apsidal pulse scheme and performed using the satellite engine.

This work is the next step in implementing the problem hierarchy methodology, which consists in the sequential formalization and solution of a series of problems, where each previously obtained

solution is used as an initial approximation in the next one. Initially, the simplest problem was solved in the apsidal pulse statement [8]: according to the results, for the optimal transfer trajectory with separation of the first impulse action and a limit of 1.5 km/s on the characteristic velocity of the final ascent maneuver, the cost of releasing the CB's spent parts (stages) turn out to be small. If the characteristic velocity of the final ascent maneuver is less than 1.47 km/s, the trajectory structure changes and the cost of releasing the AFT and CB into the atmosphere becomes significant. In the next step [9, 10], the problem was solved without assuming the apsidality of impulse actions. It was established that in the problem with a phase constraint on the maximum possible distance between the spacecraft and the Earth and an unlimited transfer time, the solution is apsidal and coincides with that obtained in the previous step. The need to solve the problem in a modified pulse statement (considering the release of the AFT and CB) [1], representing the third step of the problem hierarchy methodology, was due to the difficulty of a direct transition to the problem with a high limited thrust: the modified Newton's method did not converge when using the solution of the second-step problem as an initial approximation.

Well, this paper aims at constructing Pontryagin extremals in the problem with high limited thrust. The structure of this extremal (the sequence and approximate location of active segments on the trajectory) is known from the previous studies conducted in the pulse statement. The first series of transition maneuvers of a spacecraft to the target orbit is performed using fuel from the AFT. After exhausting this fuel, the spacecraft is in an orbit touching the conditional boundary of the Earth's atmosphere (with a perigee altitude of 100 km). On the passive flight segment, lasting 120 s, the AFT is released. The spacecraft returns to a safe orbit (with a perigee altitude of 200 km) by an additional activation of the spacecraft engine. This activation, as well as the subsequent ones, are performed using fuel from the CB's main tank.

After performing the second series of maneuvers, the spacecraft is in a target orbit from which the characteristic velocity of final ascent maneuvers to the GEO is bounded by a given value. According to the earlier studies [1, 8–10], the cost of releasing is small for the bi-elliptical final ascent scheme. In the target orbit, the satellite is separated from the CB. Due to the last engine activation on the residual fuel from the main tank in the neighborhood of the target orbit apogee, the CB is transferred to an orbit touching the conditional boundary of the Earth's atmosphere, and the satellite is transferred to the GEO using its engines.

In this paper, we consider two different but similar problem statements. Within the first one, by assumption, the tanks contain exactly as much fuel as is necessary to perform the corresponding maneuvers, the dry mass of the AFT and the mass of the CB's main tank are proportional to the mass of their fuel with a coefficient α, and the engine mass is proportional to the thrust-to-weight ratio with a coefficient β [11]. Within the second statement, the following mass characteristics of the booster are given: the dry masses of the AFT and the CB's main tank as well as limits on the masses of fuel in the AFT and the CB's main tank.

The objective functional in the problems below is the payload mass, i.e., the mass of the spacecraft remaining in the target orbit after undocking the CB.

The problems under consideration are formalized as optimal control problems for a set of dynamic systems. Based on the corresponding Pontryagin's maximum principle [12], they are reduced to multipoint boundary-value problems. The boundary-value problems of the maximum principle are solved numerically by the shooting method [13, 14]. Using the previous studies, we choose the computational schemes of the shooting method and good initial approximations of the required shooting parameters. The Cauchy problem is solved by the 8(7)th order Dorman–Prince method with automatic step selection [15]; the system of nonlinear equations, by Newton's method in the Isaev–Sonin modification [16] with the Fedorenko normalization [17] used in convergence conditions; the system of linear equations arising therein, by the Gaussian elimination technique with

selection of the leading element by column and recalculation [18]. The bulky derivatives in the transversality conditions are considered through numerical-analytical differentiation [19].

2. PROBLEM STATEMENT

The transfer is considered in the rectangular Cartesian frame related to the Earth's center. The axis z of this frame is perpendicular to the equatorial plane and has south-to-north direction; the axis x lies in the equatorial plane and is directed along the node line of the initial circular orbit from the descending node to the ascending one; the axis y completes the frame to the right-hand triple.

The motion of the spacecraft's center of mass in the central Newtonian gravitational field in a vacuum is described by the system of differential equations

$$\begin{gathered}\dot{x}(t)=v_x(t),\quad \dot{y}(t)=v_y(t),\quad \dot{z}(t)=v_z(t),\\ \dot{v}_x(t)=-\frac{\mu x(t)}{r^3(t)}+\frac{P_x(t)}{m(t)},\quad \dot{v}_y(t)=-\frac{\mu y(t)}{r^3(t)}+\frac{P_y(t)}{m(t)},\\ \dot{v}_z(t)=-\frac{\mu z(t)}{r^3(t)}+\frac{P_z(t)}{m(t)},\quad \dot{m}(t)=-\frac{P(t)}{c}\end{gathered}\tag{1}$$

with the following notations: $x(t)$, $y(t)$, and $z(t)$ are the coordinates of the spacecraft's center of mass at a time instant t; $r=\sqrt{x^2(t)+y^2(t)+z^2(t)}$ is the distance between the spacecraft and the Earth's center at a time instant t; $v_x(t)$, $v_y(t)$, and $v_z(t)$ are the velocity vector components of the spacecraft's center of mass at a time instant t; $M(0)$ is the spacecraft mass at the initial time instant; $M(t)$ is the spacecraft mass at a time instant t; $m(t)=M(t)/M(0)$ is the dimensionless mass of the spacecraft (used in calculations); $\vec{F}(t)=(F_x(t),F_y(t),F_z(t))$ is the jet thrust vector at a time instant t; $F(t)=|\vec{F}(t)|=\sqrt{F_x^2(t)+F_y^2(t)+F_z^2(t)}$ is the magnitude of the jet thrust vector; $\vec{P}(t)=(P_x(t),P_y(t),P_z(t))=(F_x(t)/M(0),F_y(t)/M(0),F_z(t)/M(0))$ is the dimensionless jet thrust vector; $n=F_{\max}/(M(0)g_{\mathrm{Ear}})$ is the initial thrust-to-weight ratio; $P(t)=\sqrt{P_x^2(t)+P_y^2(t)+P_z^2(t)}$ is the magnitude of the dimensionless jet thrust vector at a time instant t; $\mu=398\,601.19\ \mathrm{km}^3/\mathrm{s}^2$ is the gravitational parameter of the Earth; $c=P_{\mathrm{spe}}g_{\mathrm{Ear}}$ is the jet velocity; P_{spe} is the specific thrust; finally, $g_{\mathrm{Ear}}=9.80665\ \mathrm{m/s}^2$ is the gravitational acceleration at the Earth surface.

In addition to the central Newtonian gravitational field, we consider the motion of the spacecraft's center of mass in the gravitational field with the second zonal harmonic:

$$\begin{gathered}\dot{x}(t)=v_x(t),\quad \dot{y}(t)=v_y(t),\quad \dot{z}(t)=v_z(t),\\ \dot{v}_x(t)=-\frac{\mu x(t)}{r^3(t)}+\frac{3}{2}J_2\mu\frac{R_0^2}{r^5(t)}\left(\frac{5x(t)z^2(t)}{r^2(t)}-x(t)\right)+\frac{P_x(t)}{m(t)},\\ \dot{v}_y(t)=-\frac{\mu y(t)}{r^3(t)}+\frac{3}{2}J_2\mu\frac{R_0^2}{r^5(t)}\left(\frac{5y(t)z^2(t)}{r^2(t)}-y(t)\right)+\frac{P_y(t)}{m(t)},\\ \dot{v}_z(t)=-\frac{\mu z(t)}{r^3(t)}+\frac{3}{2}J_2\mu\frac{R_0^2}{r^5(t)}\left(\frac{5z^3(t)}{r^2(t)}-3z(t)\right)+\frac{P_z(t)}{m(t)},\\ \dot{m}(t)=-\frac{P(t)}{c},\end{gathered}$$

where $J_2=1082.636023\times10^{-6}$ is the coefficient of the second zonal harmonic.

Controls are supposed to be piecewise continuous functions:

$$P(t) = \sqrt{(P_x(t))^2 + (P_y(t))^2 + (P_z(t))^2} \leqslant P_{\max},$$

where $P_{\max} = g_{\mathrm{Ear}} n m_0$ is the limit on the magnitude of the control thrust vector, n is the initial thrust-to-weight ratio of the spacecraft, and m_0 is the initial weight of the spacecraft.

At the initial time instant ($t = 0$) the spacecraft is in the circular reference orbit of a given radius R_0. Due to the chosen frame, the ascending node has the longitude $\Omega_0 = 0$, and the spacecraft's motion in the initial circular orbit can be formalized by the conditions

$$\begin{gathered} x(0)^2 + y(0)^2 + z(0)^2 = R_0^2, \quad x(0)C_{0x} + y(0)C_{0y} + z(0)C_{0z} = 0, \\ v_x(0) + \frac{v_0}{R_0}\left(y(0)\cos i_0 + z(0)\sin i_0\right) = 0, \;\; v_y(0) - \frac{v_0}{R_0}x(0)\cos i_0 = 0, \\ v_z(0) - \frac{v_0}{R_0}x(0)\sin i_0 = 0, \end{gathered} \tag{2}$$

where

$$\begin{gathered} C_{0x} = 0, \quad C_{0y} = -C_0 \sin i_0, \quad C_{0z} = C_0 \cos i_0, \quad C_0 = \sqrt{\mu R_0}, \\ v_0 = \sqrt{\frac{\mu}{R_0}}, \quad R_0 = R_{\mathrm{Ear}} + h_0. \end{gathered}$$

In these formulas, C_{0x}, C_{0y}, and C_{0z} are the components of the kinetic momentum vector of the spacecraft relative to the Earth's center; C_0 is the magnitude of this vector; v_0 is the magnitude of the velocity vector in the reference orbit; R_0 is the radius of the reference orbit; $R_{\mathrm{Ear}} = 6378.25$ km is the Earth's radius; finally, $h_0 = 200$ km is the altitude of the reference orbit above the Earth's surface.

The mass of the spacecraft is considered dimensionless and therefore equals 1 at the initial time instant:

$$m(0) = m_0 = 1. \tag{3}$$

The perigee radius $r_{\mathrm{p}}(x, y, z, v_x, v_y, v_z)$ of an instantaneous elliptical orbit is a function of spacecraft coordinates and velocities and is calculated using the following formulas [20]:

$$\begin{gathered} r = \sqrt{x^2 + y^2 + z^2}, \quad V = \sqrt{v_x^2 + v_y^2 + v_z^2}, \\ \cos\varphi = \frac{x v_x + y v_y + z v_z}{rV}, \quad V_{\mathrm{cir}}^2 = \frac{\mu}{r}, \\ e = \sqrt{\left[\left(\frac{V}{V_{\mathrm{cir}}}\right)^2 - 1\right]^2 + \frac{r}{a}\left(\frac{V}{V_{\mathrm{cir}}}\right)^2 \cos^2\varphi}, \\ a = \frac{r}{2 - \left(\dfrac{V}{V_{\mathrm{cir}}}\right)^2}, \quad r_{\mathrm{p}} = a(1 - e), \end{gathered} \tag{4}$$

where a is the semi-major axis; e is eccentricity; V_{cir} is the circular velocity at the distance r from the Earth's center; finally, φ is the angle between the radius vector $\vec{r} = (x,\, y,\, z)$ and the velocity vector $\overrightarrow{V} = (v_x,\, v_y,\, v_z)$.

In what follows, the perigee radius of the orbit is denoted by

$$r_{\mathrm{p}}(\tau) := r_{\mathrm{p}}\left(x(\tau), y(\tau), z(\tau), v_x(\tau), v_y(\tau), v_z(\tau)\right),$$

where τ is an arbitrary time instant.

After performing the first series of maneuvers at the time instant $\tau_{\text{rel1}}^{\text{AFT}}$, the spacecraft must be in the instantaneous Keplerian orbit touching the conditional boundary of the atmosphere. (In fact, the conditions of touching the atmosphere of the instantaneous Keplerian orbit do not ensure touching the orbit's atmosphere in the Earth's real field; by assumption in this paper, such conditions are sufficient for the rapid elimination of space debris.) The orbit perigee altitude is lowered to 100 km (the conditional boundary of the atmosphere) by activating the CB engine on the residual fuel from the AFT. At the time instant $\tau_{\text{rel1}}^{\text{AFT}}$ we have the conditions

$$r_{\text{p}}(\tau_{\text{rel1-}}^{\text{AFT}}) = R_{\text{Ear}} + 100\,\text{km}, \tag{5}$$

$$x(\tau_{\text{rel1+}}^{\text{AFT}}) - x(\tau_{\text{rel1-}}^{\text{AFT}}) = 0, \quad y(\tau_{\text{rel1+}}^{\text{AFT}}) - y(\tau_{\text{rel1-}}^{\text{AFT}}) = 0,$$

$$z(\tau_{\text{rel1+}}^{\text{AFT}}) - z(\tau_{\text{rel1-}}^{\text{AFT}}) = 0, \quad v_x(\tau_{\text{rel1+}}^{\text{AFT}}) - v_x(\tau_{\text{rel1-}}^{\text{AFT}}) = 0,$$

$$v_y(\tau_{\text{rel1+}}^{\text{AFT}}) - v_y(\tau_{\text{rel1-}}^{\text{AFT}}) = 0, \quad v_z(\tau_{\text{rel1+}}^{\text{AFT}}) - v_z(\tau_{\text{rel1-}}^{\text{AFT}}) = 0,$$

$$\tau_{\text{rel1+}}^{\text{AFT}} - \tau_{\text{rel1-}}^{\text{AFT}} = 0.$$

After the spacecraft reaches the AFT release orbit, the passive segment $[\tau_{\text{rel1}}^{\text{AFT}}, \tau_{\text{rel2}}^{\text{AFT}}]$ begins (AFT release). On this segment, the mass is neglected in the system of differential equations. By assumption, undocking the AFT takes a given time:

$$\tau_{\text{rel2+}}^{\text{AFT}} - \tau_{\text{rel1-}}^{\text{AFT}} = 120\ \text{s}.$$

We consider two different but similar problem statements. Within the first one, by assumption, the tanks contain exactly as much fuel as is necessary to perform the corresponding maneuvers, the dry mass of the AFT and the mass of the CB's main tank are proportional to the mass of their fuel with a coefficient α, and the engine mass (including the additional CB structures) is proportional to the thrust-to-weight ratio with a coefficient β. Within the second statement, the mass characteristics of the booster are given.

The mass of the spacecraft after AFT release is calculated as follows:

$$m\left(\tau_{\text{rel2+}}^{\text{AFT}}\right) = m\left(\tau_{\text{rel1-}}^{\text{AFT}}\right) - \alpha\left(m_0 - m\left(\tau_{\text{rel1-}}^{\text{AFT}}\right)\right) \tag{6}$$

(the first problem statement) and

$$m\left(\tau_{\text{rel2}}^{\text{AFT}}\right) = m\left(\tau_{\text{rel1}}^{\text{AFT}}\right) - m^{\text{AFT}} \tag{7}$$

(the second problem statement), where m^{AFT} is the given dry dimensionless mass of the AFT. (This value can be supposed to include the mass of unreduced fuel residue.)

The constraint on the AFT fuel mass in the second problem statement has the form

$$m_0 - m\left(\tau_{\text{rel1}}^{\text{AFT}}\right) \leqslant m_{\text{fuel}}^{\text{AFT}}. \tag{8}$$

After releasing the AFT, the spacecraft performs a transition maneuver to the safe orbit. This maneuver ends at the time instant τ_{safe}. Fuel from the main tank is used to perform it. As before, the pericenter radius $r_{\text{p}}(\cdot)$ is a function of the coordinates and components of the spacecraft velocity vector (4). At the time instant τ_{safe} we have the conditions

$$r_{\text{p}}\left(\tau_{\text{safe-}}\right) = R_{\text{Ear}} + 200\,\text{km}, \tag{9}$$

$$x(\tau_{\text{safe+}}) - x(\tau_{\text{safe-}}) = 0,\ y(\tau_{\text{safe+}}) - y(\tau_{\text{safe-}}) = 0,\ z(\tau_{\text{safe+}}) - z(\tau_{\text{safe-}}) = 0,$$

$$v_x(\tau_{\text{safe+}}) - v_x(\tau_{\text{safe-}}) = 0,\ v_y(\tau_{\text{safe+}}) - v_y(\tau_{\text{safe-}}) = 0,\ v_z(\tau_{\text{safe+}}) - v_z(\tau_{\text{safe-}}) = 0,$$

$$\tau_{\text{safe+}} - \tau_{\text{safe-}} = 0.$$

After reaching the safe orbit, the second series of maneuvers begins to transfer the spacecraft to the target orbit. In the target orbit, the satellite is undocked from the CB. The payload mass of the satellite remaining in the target orbit has to be optimized:

$$m_{\mathrm{p}} = m\left(\tau_{\text{tar-}}\right) - m\left(\tau_{\text{tar+}}\right) \to \max,$$

where $m\left(\tau_{\text{tar-}}\right)$ is the mass of the spacecraft in the target orbit before undocking the satellite; $m\left(\tau_{\text{tar+}}\right)$ is the CB mass in the target orbit after undocking the satellite. The satellite moves to the GEO using its engines. By assumption, the characteristic velocity of the final ascent maneuver from the target orbit to the GEO is limited by a given value Δv^* and the apsidal line of the target orbit lies in the equatorial plane, i.e., the z-component of the Laplace vector is zero. At the time instant τ_{tar} we have the conditions

$$\begin{gathered}
\Delta v_{\text{fa}}(\tau_{\text{tar-}}) := \Delta v_{\text{fa}}(x(\tau_{\text{tar-}}), y(\tau_{\text{tar-}}), z(\tau_{\text{tar-}}), v_x(\tau_{\text{tar-}}), v_y(\tau_{\text{tar-}}), v_z(\tau_{\text{tar-}})) \leqslant \Delta v^*, \\
\mathcal{A}\left(\tau_{\text{tar-}}\right) := C_y\left(\tau_{\text{tar-}}\right) v_x\left(\tau_{\text{tar-}}\right) - C_x\left(\tau_{\text{tar-}}\right) v_y\left(\tau_{\text{tar-}}\right) - \frac{\mu z\left(\tau_{\text{tar-}}\right)}{r\left(\tau_{\text{tar-}}\right)} = 0, \\
x(\tau_{\text{tar+}}) - x(\tau_{\text{tar-}}) = 0,\ y(\tau_{\text{tar+}}) - y(\tau_{\text{tar-}}) = 0,\ z(\tau_{\text{tar+}}) - z(\tau_{\text{tar-}}) = 0, \\
v_x(\tau_{\text{tar+}}) - v_x(\tau_{\text{tar-}}) = 0,\ v_y(\tau_{\text{tar+}}) - v_y(\tau_{\text{tar-}}) = 0,\ v_z(\tau_{\text{tar+}}) - v_z(\tau_{\text{tar-}}) = 0, \\
\tau_{\text{tar+}} - \tau_{\text{tar-}} = 0,
\end{gathered} \tag{10}$$

where τ_{tar} is the time instant of reaching the target orbit; $C_x\left(\tau_{\text{tar-}}\right)$, $C_y\left(\tau_{\text{tar-}}\right)$, and $C_z\left(\tau_{\text{tar-}}\right)$ are the components of the kinetic momentum vector of the spacecraft orbital motion at the time instant $\tau_{\text{tar-}}$.

Note that the characteristic velocity of the final ascent maneuvers of the satellite from the target orbit to the GEO is considered by the simplified scheme. (It is considered within the central Newtonian field, with apsidal impulse actions for maneuvering only, orbit rotation by the second impulse action only, and the first accelerating and the last setting impulse actions not changing the orbit plane.) Used together, these conditions simplify the problem statement very significantly. The value $R_{\max}$ (the distance to the Earth) is chosen, on the one hand, to be large enough, and on the other hand, to be appropriate for neglecting the influence of other bodies of the Solar System. First of all, the matter concerns the influence of the Moon and the Sun; for example, consideration of the Moon will even avoid activation of the engine at the remote point [21]. Of course, this influence can be modeled in the next steps of the problem hierarchy methodology. In this paper, the influence of other bodies is omitted. The final ascent of the satellite is implemented using three impulse actions:

$$\Delta v_{\text{fa}}\left(\tau_{\text{tar}}\right) = \Delta v_{\text{fa1}}\left(\tau_{\text{tar}}\right) + \Delta v_{\text{fa2}}\left(\tau_{\text{tar}}\right) + \Delta v_{\text{fa3}}\left(\tau_{\text{tar}}\right).$$

The first impulse action $\Delta v_{\text{fa1}}\left(\tau_{\text{tar}}\right)$ is applied at the perigee of the target orbit; without changing the inclination, it raises the apogee to the maximum possible distance $R_{\max}$ of the spacecraft from the Earth:

$$\begin{gathered}
\Delta v_{\text{fa1}}\left(\tau_{\text{tar}}\right) = \sqrt{V_{\text{tar p}}^2 + V_{\text{1p}}^2 - 2V_{\text{tar p}}V_{\text{1p}}}, \\
V_{\text{tar p}} = \sqrt{\frac{2\mu R_{\text{tar a}}}{R_{\text{tar p}}(R_{\text{tar a}} + R_{\text{tar p}})}}, \quad V_{\text{1p}} = \sqrt{\frac{2\mu R_{\max}}{R_{\text{tar p}}(R_{\max} + R_{\text{tar p}})}},
\end{gathered} \tag{11}$$

where $R_{\text{tar p}}$ is the perigee radius of the target orbit and $V_{\text{tar p}}$ is the velocity at the perigee of the target orbit.

The second impulse action $\Delta v_{\mathrm{fa2}}(\tau_{\mathrm{tar}})$ is applied at the apogee; it increases the perigee to the GEO radius R_{GEO} and decreases the inclination to zero:

$$\begin{gathered}\Delta v_{\mathrm{fa2}}(\tau_{\mathrm{tar}}) = \sqrt{V_{1\mathrm{a}}^2 + V_{2\mathrm{a}}^2 - 2V_{1\mathrm{a}}V_{2\mathrm{a}}\cos i_{\mathrm{tar}}},\\ V_{1\mathrm{a}} = \sqrt{\frac{2\mu R_{\mathrm{tar\ p}}}{R_{\max}(R_{\max}+R_{\mathrm{tar\ p}})}},\quad V_{2\mathrm{a}} = \sqrt{\frac{2\mu R_{\mathrm{GEO}}}{R_{\max}(R_{\max}+R_{\mathrm{GEO}})}},\end{gathered} \tag{12}$$

where i_{tar} is the inclination angle of the target orbit to the equatorial plane. At the time instant of passing the apogee, this value can be calculated as

$$\cos i_{\mathrm{tar}} = \frac{\sqrt{v_x^2(\tau_{\mathrm{tar\ a}}) + v_y^2(\tau_{\mathrm{tar\ a}})}}{\sqrt{v_x^2(\tau_{\mathrm{tar\ a}}) + v_y^2(\tau_{\mathrm{tar\ a}}) + v_z^2(\tau_{\mathrm{tar\ a}})}}. \tag{13}$$

The third impulse action $\Delta v_{\mathrm{fa3}}(\tau_{\mathrm{tar}})$ is applied at the perigee; without changing the inclination, it reduces the apogee to the GEO radius, thus moving the satellite to a non-predetermined point in the GEO:

$$\begin{gathered}\Delta v_{\mathrm{fa3}}(\cdot) = V_{2\mathrm{p}} - v_{\mathrm{GEO}},\\ V_{2\mathrm{p}} = \sqrt{\frac{2\mu R_{\max}}{R_{\mathrm{GEO}}(R_{\max}+R_{\mathrm{GEO}})}},\quad v_{\mathrm{GEO}} = \sqrt{\frac{\mu}{R_{\mathrm{GEO}}}}.\end{gathered} \tag{14}$$

Note that Δv_{fa3} is actually a constant (depends on the given value R_{GEO} and the problem parameter $R_{\max}$).

After undocking the satellite, the CB maneuver continues. Due to additional activation of the engine, the perigee altitude of the CB's orbit is lowered to 100 km (the conditional boundary of the atmosphere):

$$r_{\mathrm{p}}(T) = R_{\mathrm{Ear}} + 100\,\mathrm{km}. \tag{15}$$

At the final time instant T all the fuel contained in the CB's main tank is exhausted. Within the first problem statement, the tanks are filled with exactly as much fuel as is necessary to perform the corresponding maneuvers, the dry mass of the CB's main tank is proportional to the mass of fuel contained in it with the coefficient α, and the engine mass is proportional to the thrust-to-weight ratio with the coefficient β [11]. Therefore, we obtain

$$\begin{gathered}m(T) - \alpha m_{\mathrm{fuel}} - \beta n = 0,\\ m_{\mathrm{fuel}} = \left(m\left(\tau_{\mathrm{rel2+}}^{\mathrm{AFT}}\right) - m(\tau_{\mathrm{tar\text{-}}})\right) + \left(m(\tau_{\mathrm{tar+}}) - m(T)\right).\end{gathered} \tag{16}$$

Within the second problem statement, the CB's dry mass and the fuel constraint in the CB's main tank are given. In this case,

$$\begin{gathered}m(T) - m^{\mathrm{CB}} = 0,\\ \left(m\left(\tau_{\mathrm{rel2+}}^{\mathrm{AFT}}\right) - m(\tau_{\mathrm{tar\text{-}}})\right) + \left(m(\tau_{\mathrm{tar+}}) - m(T)\right) \leqslant m_{\mathrm{fuel}}^{\mathrm{CB}},\end{gathered} \tag{17}$$

where m^{CB} is the given dry dimensionless mass of the CB (the engine and additional structures) and $m_{\mathrm{fuel}}^{\mathrm{CB}}$ is the maximum dimensionless mass of fuel that can be filled into the CB's main tank.

Note that the spacecraft coordinates and velocities are continuous at all time instants. In addition, the problem under consideration has another peculiarity: its objective functional is a function of the phase variables at an intermediate time instant.

3. PONTRYAGIN'S MAXIMUM PRINCIPLE

The problem under consideration is an optimal control problem with intermediate conditions. It can be solved using Pontryagin's maximum principle [12].

In the case of the central Newtonian gravitational field, the Pontryagin function has the form

$$H = p_x v_x + p_y v_y + p_z v_z + p_m \left(-\frac{P}{c}\right)$$
$$+ p_{vx}\left(-\frac{\mu x}{r^3} + \frac{P_x}{m}\right) + p_{vy}\left(-\frac{\mu y}{r^3} + \frac{P_y}{m}\right) + p_{vz}\left(-\frac{\mu z}{r^3} + \frac{P_z}{m}\right);$$

in the problems with the second zonal harmonic, the form

$$H = p_x v_x + p_y v_y + p_z v_z + p_m\left(-\frac{P}{c}\right) + p_{vx}\left(-\frac{\mu x}{r^3} + \frac{3}{2}J_2\mu\frac{R_0^2}{r^5}\left(\frac{5xz^2}{r^2} - x\right) + \frac{P_x}{m}\right)$$
$$+ p_{vy}\left(-\frac{\mu y(t)}{r^3(t)} + \frac{3}{2}J_2\mu\frac{R_0^2}{r^5(t)}\left(\frac{5yz^2}{r^2} - y\right) + \frac{P_y}{m}\right) + p_{vz}\left(-\frac{\mu z}{r^3} + \frac{3}{2}J_2\mu\frac{R_0^2}{r^5}\left(\frac{5z^3}{r^2} - 3z\right) + \frac{P_z}{m}\right).$$

For the first problem statement, the terminant is given by

$$l = l_0 + l_{\text{rel1}} + l_{\text{rel2}} + l_{\text{safe}} + l_{\text{tar}} + l_T - \lambda_0\left(m\left(\tau_{\text{tar}}-\right) - m\left(\tau_{\text{tar}}+\right)\right),$$

where

$$l_0 = \lambda_{R0}\left(x(0)^2 + y(0)^2 + z(0)^2 - R_0^2\right) + \lambda_{C0}\left(x(0)C_{0x} + y(0)C_{0y} + z(0)C_{0z}\right)$$
$$+\lambda_{vx0}\left(v_x(0) + \frac{v_0}{R_0}\left(y(0)\cos i_0 + z(0)\sin i_0\right)\right) + \lambda_{vy0}\left(v_y(0) - \frac{v_0}{R_0}x(0)\cos i_0\right)$$
$$+\lambda_{vz0}\left(v_z(0) - \frac{v_0}{R_0}x(0)\sin i_0\right) + \lambda_{m0}\left(m(0) - m_0\right),$$

$$l_{\text{rel1}} = \sum_{\xi=(x,y,z,v_x,v_y,v_z)} \lambda_{\xi\text{rel1}}\left(\xi\left(\tau^{\text{AFT}}_{\text{rel1+}}\right) - \xi\left(\tau^{\text{AFT}}_{\text{rel1-}}\right)\right) + \lambda_{\tau\text{rel1}}\left(\tau^{\text{AFT}}_{\text{rel1+}} - \tau^{\text{AFT}}_{\text{rel1-}}\right)$$
$$+ \lambda_{\text{rel1}}\left(r_{\text{p}}\left(\tau^{\text{AFT}}_{\text{rel1-}}\right) - R_{\text{Ear}} - 100\right),$$

$$l_{\text{rel2}} = \lambda_{m\tau}\left(m\left(\tau^{\text{AFT}}_{\text{rel2+}}\right) - m\left(\tau^{\text{AFT}}_{\text{rel1-}}\right) + \alpha\left(m_0 - m\left(\tau^{\text{AFT}}_{\text{rel1-}}\right)\right)\right)$$
$$+ \lambda_\tau\left(\tau^{\text{AFT}}_{\text{rel2+}} - \tau^{\text{AFT}}_{\text{rel1-}} - 120\right) + \lambda_{\tau\text{rel2}}\left(\tau^{\text{AFT}}_{\text{rel2+}} - \tau^{\text{AFT}}_{\text{rel2-}}\right),$$

$$l_{\text{safe}} = \sum_{\xi=(x,y,z,v_x,v_y,v_z)} \lambda_{\xi\text{safe}}\left(\xi\left(\tau_{\text{safe+}}\right) - \xi\left(\tau_{\text{safe-}}\right)\right) + \lambda_{\tau\text{safe}}\left(\tau_{\text{safe+}} - \tau_{\text{safe-}}\right)$$
$$+ \lambda_{\text{safe}}\left(r_{\text{p}}\left(\tau_{\text{safe-}}\right) - R_{\text{Ear}} - 200\right),$$

$$l_{\text{tar}} = \sum_{\xi=(x,y,z,v_x,v_y,v_z)} \lambda_{\xi\text{tar}}\left(\xi\left(\tau_{\text{tar+}}\right) - \xi\left(\tau_{\text{tar-}}\right)\right) + \lambda_{\tau\text{tar}}\left(\tau_{\text{tar+}} - \tau_{\text{tar-}}\right)$$
$$+ \lambda_{\text{tar}}\left(C_y(\tau_{\text{tar-}})v_x(\tau_{\text{tar-}}) - C_x(\tau_{\text{tar-}})v_y(\tau_{\text{tar-}}) - \frac{\mu z(\tau_{\text{tar-}})}{r(\tau_{\text{tar-}})}\right)$$
$$+ \lambda_{\text{fa}}(\Delta v_{\text{fa}}(x(\tau_{\text{tar-}}), y(\tau_{\text{tar-}}), z(\tau_{\text{tar-}}), v_x(\tau_{\text{tar-}}), v_y(\tau_{\text{tar-}}), v_z(\tau_{\text{tar-}})) - \Delta v^*),$$

$$l_T = \lambda_T\left(r_{\text{p}}\left(T\right) - R_{\text{Ear}} - 100\right)$$
$$+ \lambda_{mT}\left(m(T) - \alpha\left(\left(m\left(\tau^{\text{AFT}}_{\text{rel2+}}\right) - m\left(\tau_{\text{tar-}}\right)\right) + \left(m\left(\tau_{\text{tar+}}\right) - m(T)\right)\right) - \beta n\right).$$

For the second problem statement, l_{rel2} and l_T are given by

$$l_{\text{rel2}} = \lambda_{m\tau 1}\left(m\left(\tau_{\text{rel2+}}^{\text{AFT}}\right) - m\left(\tau_{\text{rel1-}}^{\text{AFT}}\right) + m^{\text{AFT}}\right) + \lambda_{m\tau 2}\left(m_0 - m\left(\tau_{\text{rel1-}}^{\text{AFT}}\right) - m_{\text{fuel}}^{\text{AFT}}\right)$$
$$+ \lambda_\tau\left(\tau_{\text{rel2+}}^{\text{AFT}} - \tau_{\text{rel1-}}^{\text{AFT}} - 120\right) + \lambda_{\tau\text{rel2}}\left(\tau_{\text{rel2+}}^{\text{AFT}} - \tau_{\text{rel2-}}^{\text{AFT}}\right),$$

$$l_T = \lambda_T\left(r_{\text{p}}\left(T\right) - R_{\text{Ear}} - 100\right) + \lambda_{mT1}\left(m(T) - m^{\text{CB}}\right)$$
$$+ \lambda_{mT2}\left(\left(m\left(\tau_{\text{rel2+}}^{\text{AFT}}\right) - m\left(\tau_{\text{tar-}}\right)\right) + \left(m\left(\tau_{\text{tar+}}\right) - m(T)\right) - m_{\text{fuel}}^{\text{CB}}\right).$$

Here, $p_x(\cdot)$, $p_y(\cdot)$, $p_z(\cdot)$, $p_{v_x}(\cdot)$, $p_{v_y}(\cdot)$, $p_{v_z}(\cdot)$, and $p_m(\cdot)$ are the conjugate variables (the functional Lagrange multipliers) at each of the trajectory segments; λ_0, λ_{R0}, λ_{C0}, λ_{vx0}, λ_{vy0}, λ_{vz0}, λ_{m0}, $\lambda_{\xi\text{rel1}}$, $\lambda_{\xi\text{safe}}$, $\lambda_{\xi\text{tar}}$ ($\xi = x, y, z, v_x, v_y, v_z$), $\lambda_{\tau\text{rel1}}$, λ_{rel1}, $\lambda_{m\tau}$, λ_τ, $\lambda_{\tau\text{rel2}}$, $\lambda_{\tau\text{safe}}$, λ_{safe}, $\lambda_{\tau\text{tar}}$, λ_{tar}, λ_{fa}, λ_T, λ_{mT}, $\lambda_{m\tau 1}$, $\lambda_{m\tau 2}$, λ_{mT1}, and λ_{mT2} are the numerical Lagrange multipliers.

Note that the term corresponding to the differential equation $\dot m = -\frac{P}{c}$ is absent in the Pontryagin function on the segment $[\tau_{\text{rel1}}^{\text{AFT}}, \tau_{\text{rel2}}^{\text{AFT}}]$.

The stationarity conditions with respect to the phase variables (the Euler–Lagrange equations) have the form

$$\dot p_x = \frac{\mu}{r^3}\left[p_{vx} - \frac{3x}{r^2}\left(xp_{vx} + yp_{vy} + zp_{vz}\right)\right],$$
$$\dot p_y = \frac{\mu}{r^3}\left[p_{vy} - \frac{3y}{r^2}\left(xp_{vx} + yp_{vy} + zp_{vz}\right)\right],$$
$$\dot p_z = \frac{\mu}{r^3}\left[p_{vz} - \frac{3z}{r^2}\left(xp_{vx} + yp_{vy} + zp_{vz}\right)\right],$$
$$\dot p_{vx} = -p_x, \quad \dot p_{vy} = -p_y, \quad \dot p_{vz} = -p_z,$$
$$\dot p_m = \frac{P_x p_{vx} + P_y p_{vy} + P_z p_{vz}}{m^2}.$$

In the case of a transfer in a gravitational field with the second zonal harmonic, the Euler–Lagrange equations are not presented explicitly here. Their right-hand sides were calculated using numerical-analytical differentiation [19].

Due to their bulkiness, we formally write the transversality conditions as

$$p_\xi(0) = \frac{\partial l}{\partial \xi(0)}, \quad p_\xi(T) = -\frac{\partial l}{\partial \xi(T)},$$
$$p_\xi\left(\beta_+\right) = \frac{\partial l}{\partial \xi\left(\beta_+\right)}, \quad p_\xi\left(\beta_-\right) = -\frac{\partial l}{\partial \xi\left(\beta_-\right)},$$
$$\xi = x, y, z, v_x, v_y, v_z, \quad \beta = \tau_{\text{rel1}}^{\text{AFT}}, \tau_{\text{rel2}}^{\text{AFT}}, \tau_{\text{safe}}, \tau_{\text{tar}}.$$

The transversality conditions at the initial time instant imply

$$p_x(0) = 2\lambda_{R0}x(0) + \lambda_{C0}C_{0x} - \frac{v_0}{R_0}\left(p_{vy}(0)\cos i_0 + p_{vz}(0)\sin i_0\right),$$
$$p_y(0) = 2\lambda_{R0}y(0) + \lambda_{C0}C_{0y} + \frac{v_0}{R_0}p_{vx}(0)\cos i_0, \tag{18}$$
$$p_z(0) = 2\lambda_{R0}z(0) + \lambda_{C0}C_{0z} + \frac{v_0}{R_0}p_{vy}(0)\sin i_0.$$

At the time instants $\tau_{\text{rel1}}^{\text{AFT}}$ and τ_{safe}, these conditions yield

$$\begin{gathered} p_\xi(\gamma_-) - p_\xi(\gamma_+) + \lambda_i \frac{\partial r_{\text{p}}(\gamma_-)}{\partial \xi(\gamma_-)} = 0, \\ \xi = x, y, z, v_x, v_y, v_z, \quad \gamma = \tau_{\text{rel1}}^{\text{AFT}}, \tau_{\text{safe}}, \quad i = \text{rel1}, \text{safe}. \end{gathered} \tag{19}$$

Finally, at the time instant τ_{tar}, from the transversality conditions it follows that

$$\begin{gathered} p_\xi(\tau_{\text{tar-}}) - p_\xi(\tau_{\text{tar+}}) + \lambda_{\text{fa}} \frac{\Delta v_{\text{fa}}(\tau_{\text{tar-}})}{\partial \xi(\tau_{\text{tar-}})} + \lambda_{\text{tar}} \frac{\partial \mathcal{A}(\tau_{\text{tar-}})}{\partial \xi(\tau_{\text{tar-}})} = 0, \\ \xi = x, y, z, v_x, v_y, v_z. \end{gathered} \tag{20}$$

The derivatives of the functions $r_{\text{p}}(\cdot)$, $\Delta v_{\text{fa}}(\cdot)$, and $\mathcal{A}(\cdot)$ (see (19), (20), and the transversality conditions at the final time instant T) are calculated using numerical-analytical differentiation.

In the first problem statement, the transversality conditions with respect to the variable m at the time instants $\tau_{\text{rel1}}^{\text{AFT}}$, $\tau_{\text{rel2}}^{\text{AFT}}$, and T imply the equality

$$(1+\alpha) p_m(\tau_{\text{rel2+}}^{\text{AFT}}) - p_m(\tau_{\text{rel1-}}^{\text{AFT}}) - \alpha p_m(T) = 0. \tag{21}$$

Let us prove this equality. The transversality conditions with respect to the variable m at the time instants $\tau_{\text{rel1}}^{\text{AFT}}$, $\tau_{\text{rel2}}^{\text{AFT}}$, and T have the form

$$\begin{gathered} p_m(\tau_{\text{rel1-}}^{\text{AFT}}) = -\frac{\partial l}{\partial m(\tau_{\text{rel1-}}^{\text{AFT}})} = \lambda_{m\tau}(1+\alpha), \\ p_m(\tau_{\text{rel2+}}^{\text{AFT}}) = \frac{\partial l}{\partial m(\tau_{\text{rel2+}}^{\text{AFT}})} = \lambda_{m\tau} - \alpha\lambda_{mT}, \\ p_m(T) = -\frac{\partial l}{\partial m(T)} = -\lambda_{mT}(1+\alpha). \end{gathered}$$

We obtain the following chain of equalities:

$$\begin{gathered} p_m(\tau_{\text{rel2+}}^{\text{AFT}}) - \lambda_{m\tau} + \alpha\lambda_{mT} = 0, \quad \lambda_{m\tau} = \frac{p_m(\tau_{\text{rel1-}}^{\text{AFT}})}{1+\alpha}, \quad \lambda_{mT} = -\frac{p_m(T)}{1+\alpha} \\ \Rightarrow p_m(\tau_{\text{rel2+}}^{\text{AFT}}) - \frac{p_m(\tau_{\text{rel1-}}^{\text{AFT}})}{1+\alpha} - \alpha\frac{p_m(T)}{1+\alpha} = 0 \\ \Rightarrow (1+\alpha) p_m(\tau_{\text{rel2+}}^{\text{AFT}}) - p_m(\tau_{\text{rel1-}}^{\text{AFT}}) - \alpha p_m(T) = 0. \end{gathered}$$

In the second problem statement, the transversality conditions with respect to the variable m at the time instants $\tau_{\text{rel1}}^{\text{AFT}}$ and $\tau_{\text{rel2}}^{\text{AFT}}$ imply the equality

$$p_m(\tau_{\text{rel2+}}^{\text{AFT}}) - p_m(\tau_{\text{rel1-}}^{\text{AFT}}) = \lambda_{mT2} - \lambda_{m\tau 2}. \tag{22}$$

Let us prove this equality. The transversality conditions with respect to the variable m at the time instants $\tau_{\text{rel1}}^{\text{AFT}}$ and $\tau_{\text{rel2}}^{\text{AFT}}$ have the form

$$p_m(\tau_{\text{rel2+}}^{\text{AFT}}) = \lambda_{m\tau 1} + \lambda_{mT2}, \quad p_m(\tau_{\text{rel1-}}^{\text{AFT}}) = \lambda_{m\tau 1} + \lambda_{m\tau 2}.$$

Subtracting the second equality from the first one yields (22).

The transversality conditions with respect to the variable m at the time instant τ_{tar} imply the continuity of the conjugate variable:

$$p_m(\tau_{\text{tar+}}) = p_m(\tau_{\text{tar-}}). \tag{23}$$

Indeed, $p_m(\tau_{\text{tar-}}) = -\alpha\lambda_{mT}$ and $p_m(\tau_{\text{tar+}}) = -\alpha\lambda_{mT}$ (the first problem statement) and $p_m(\tau_{\text{tar-}}) = -\alpha\lambda_{mT2}$ and $p_m(\tau_{\text{tar+}}) = -\alpha\lambda_{mT2}$ (the second problem statement).

At the initial time instant the stationarity condition is absent. The stationarity conditions at the time instants $\tau_{\text{rel1}}^{\text{AFT}}$ and $\tau_{\text{rel2}}^{\text{AFT}}$ imply $H(\tau_{\text{rel2+}}^{\text{AFT}}) = H(\tau_{\text{rel1-}}^{\text{AFT}})$. At the time instants τ_{safe} and τ_{tar}, the Pontryagin function is continuous: $H(\tau_{\text{safe+}}) = H(\tau_{\text{safe-}})$ and $H(\tau_{\text{tar+}}) = H(\tau_{\text{tar-}})$. The stationarity condition at a time instant T (unknown in advance) has the form $H(T) = 0$.

Let $\vec{e}$ be a unit vector. Then the optimality conditions with respect to the control actions P_x, P_y, and P_z have the form

$$\overrightarrow{P} = P\vec{e}, \quad \vec{e} = (\cos\alpha, \cos\beta, \cos\gamma),$$
$$P_x = P\cos\alpha, \quad P_y = P\cos\beta, \quad P_z = P\cos\gamma,$$

$$\overrightarrow{P}_{opt} = \underset{0\leqslant P\leqslant P_{\max}}{arg\,abs\,max}\left[\frac{p_{vx}P_x + p_{vy}P_y + p_{vz}P_z}{m} - \frac{p_m}{c}P\right]$$
$$= \underset{0\leqslant P\leqslant P_{\max}}{arg\,abs\,max}\left[\frac{p_{vx}P\cos\alpha + p_{vy}P\cos\beta + p_{vz}P\cos\gamma}{m} - \frac{p_m}{c}P\right]$$
$$= \underset{0\leqslant P\leqslant P_{\max}}{arg\,abs\,max}\left[P\left(\frac{p_{vx}\cos\alpha + p_{vy}\cos\beta + p_{vz}\cos\gamma}{m} - \frac{p_m}{c}\right)\right],$$

where $\cos\alpha$, $\cos\beta$, and $\cos\gamma$ are the direction cosines.

If $\left(\dfrac{p_{vx}\cos\alpha + p_{vy}\cos\beta + p_{vz}\cos\gamma}{m} - \dfrac{p_m}{c}\right) > 0$, then $P_{opt} = P_{\max}$; in the case $\left(\dfrac{p_{vx}\cos\alpha + p_{vy}\cos\beta + p_{vz}\cos\gamma}{m} - \dfrac{p_m}{c}\right) < 0$, $P_{opt} = 0$. Thus,

$$P_{opt} = \begin{cases} P_{\max}, & \chi > 0 \\ 0, & \chi < 0, \end{cases}$$

where $\chi \equiv \dfrac{\rho}{m} - \dfrac{p_m}{c}$ is the switching function.

Note that $p_{vx}\cos\alpha + p_{vy}\cos\beta + p_{vz}\cos\gamma$ is the inner product of the vectors $\vec{p}_v = (p_{vx}, p_{vy}, p_{vz})$ and $\vec{e} = (\cos\alpha, \cos\beta, \cos\gamma)$. It achieves maximum for the codirectional vectors $\vec{p}_v$ and $\vec{e}$:

$$\cos\alpha_{opt} = \frac{p_{vx}}{\rho}, \quad \cos\beta_{opt} = \frac{p_{vy}}{\rho}, \quad \cos\gamma_{opt} = \frac{p_{vz}}{\rho},$$

where $\rho = \sqrt{p_{vx}^2 + p_{vy}^2 + p_{vz}^2}$. Thus, due to the optimal direction of the thrust vector, we find

$$(P_x)_{opt} = P_{opt}\frac{p_{vx}}{\rho}, \quad (P_y)_{opt} = P_{opt}\frac{p_{vy}}{\rho}, \quad (P_z)_{opt} = P_{opt}\frac{p_{vz}}{\rho}.$$

Special control regimes, potentially possible in the problems under study, are not considered in this paper.

For the first problem statement, the complementary slackness and nonnegativity conditions have the form

$$\lambda_{\text{fa}}(\Delta v_{\text{fa}}(\tau_{\text{tar-}}) - \Delta v^*) = 0,$$
$$\lambda_0 \geqslant 0, \quad \lambda_{\text{fa}} \geqslant 0. \tag{24}$$

For the second problem statement, in addition to (24), we get the following complementary slackness and nonnegativity conditions:

$$\lambda_{m\tau 2}\left(m_0 - m\left(\tau_{\text{rel1-}}^{\text{AFT}}\right) - m_{\text{fuel}}^{\text{AFT}}\right) = 0,$$
$$\lambda_{mT2}\left(\left(m\left(\tau_{\text{rel2+}}^{\text{AFT}}\right) - m\left(\tau_{\text{tar-}}\right)\right) + \left(m\left(\tau_{\text{tar+}}\right) - m(T)\right) - m_{\text{fuel}}^{\text{CB}}\right) = 0,$$
$$\lambda_{m\tau 2} \geqslant 0, \quad \lambda_{mT2} \geqslant 0. \quad (25)$$

The normalization condition is

$$p_{vx}^2(0) + p_{vy}^2(0) + p_{vz}^2(0) = 1. \quad (26)$$

4. TRAJECTORY STRUCTURE AND NUMERICAL RESULTS

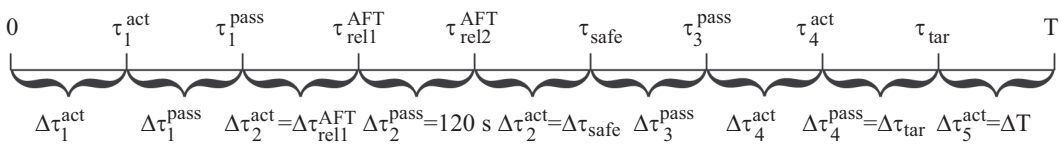
The trajectory structure is determined based on the previous studies [1, 8–10]. The main advantage of the given-structure trajectory approach is that it yields Pontryagin extremals: under a "good" computational scheme for the shooting method and a "good" initial approximation, the modified Newton's method converges in a few iterations.

Other possible methods for solving the corresponding boundary-value problems were tested as well, but without success (the Pontryagin extremals were not constructed).

The initial approximation to the values of the phase and conjugate variables in the shooting parameter vector is chosen using the solution obtained previously in the modified pulse statement [1] in accordance with [12]: at the thrust activation time instants they correspond to the values of the phase and conjugate variables before the impulse action; at the thrust deactivation time instants, to the values of the phase and conjugate variables after the impulse action. The duration of active segments is estimated based on fuel consumption for a given engine activation; the duration of passive segments is equal to the corresponding duration of passive segments between the impulse actions. First, the problem with high limited thrust is solved in the first statement with $n = 10$. Then the parameter continuation method is applied for the thrust-to-weight ratio to obtain the solution for $n = 0.1$. The transition from the first problem statement to the second one is also carried out using the parameter continuation method: the corresponding equations from the first and second problem statements are multiplied by $(1 - \gamma)$ and γ, respectively, where $\gamma \in [0, 1]$.

Let us describe the computational scheme of the shooting method (see the figure). The shooting parameter vector consists of the following components:

- the numerical Lagrange multipliers λ_{R0}, λ_{C0}, λ_{rel1}, λ_{safe}, λ_{tar}, λ_{fa}, and λ_T; in the second problem statement, also the numerical Lagrange multiplier λ_{mT2};
- the angular position of the spacecraft in the reference circular orbit, φ_0, and the values of the four conjugate variables at the initial time instant, $p_{vx}(0)$, $p_{vy}(0)$, $p_{vz}(0)$, and $p_m(0)$; (The coordinates and velocities of the spacecraft at the initial time instant are calculated by the angular position; the values $p_x(0)$, $p_y(0)$, and $p_z(0)$ of the conjugate variables are calculated using (18). By the condition, $m(0) = 1$ and, therefore, $m(0)$ is not included in the shooting parameter vector. Thus, we obtain the starting point for solving the Cauchy problem.)
- the duration of the first active segment, $\Delta\tau_1^{\text{act}}$;



The computational scheme of the shooting method.

- the coordinates and velocities as well as the values of the conjugate variables after engine deactivation, $x(\tau_{1+}^{\text{act}})$, $y(\tau_{1+}^{\text{act}})$, $z(\tau_{1+}^{\text{act}})$, $v_x(\tau_{1+}^{\text{act}})$, $v_y(\tau_{1+}^{\text{act}})$, $v_z(\tau_{1+}^{\text{act}})$, $p_x(\tau_{1+}^{\text{act}})$, $p_y(\tau_{1+}^{\text{act}})$, $p_z(\tau_{1+}^{\text{act}})$, $p_{vx}(\tau_{1+}^{\text{act}})$, $p_{vy}(\tau_{1+}^{\text{act}})$, and $p_{vz}(\tau_{1+}^{\text{act}})$;
- the duration of the first passive segment, $\Delta\tau_1^{\text{pass}}$;
- the coordinates and velocities as well as the values of the conjugate variables after engine activation, $x(\tau_{1+}^{\text{pass}})$, $y(\tau_{1+}^{\text{pass}})$, $z(\tau_{1+}^{\text{pass}})$, $v_x(\tau_{1+}^{\text{pass}})$, $v_y(\tau_{1+}^{\text{pass}})$, $v_z(\tau_{1+}^{\text{pass}})$, $p_x(\tau_{1+}^{\text{pass}})$, $p_y(\tau_{1+}^{\text{pass}})$, $p_z(\tau_{1+}^{\text{pass}})$, $p_{vx}(\tau_{1+}^{\text{pass}})$, $p_{vy}(\tau_{1+}^{\text{pass}})$, and $p_{vz}(\tau_{1+}^{\text{pass}})$;
- the duration of the second active segment, $\Delta\tau_2^{\text{act}} = \Delta\tau_{\text{rel1}}^{\text{AFT}}$, where the perigee altitude of the spacecraft orbit is lowered to the conditional boundary of the atmosphere;
- the coordinates and velocities as well as the values of the conjugate variables after engine deactivation, $x(\tau_{\text{rel1}+}^{\text{AFT}})$, $y(\tau_{\text{rel1}+}^{\text{AFT}})$, $z(\tau_{\text{rel1}+}^{\text{AFT}})$, $v_x(\tau_{\text{rel1}+}^{\text{AFT}})$, $v_y(\tau_{\text{rel1}+}^{\text{AFT}})$, $v_z(\tau_{\text{rel1}+}^{\text{AFT}})$, $p_x(\tau_{\text{rel1}+}^{\text{AFT}})$, $p_y(\tau_{\text{rel1}+}^{\text{AFT}})$, $p_z(\tau_{\text{rel1}+}^{\text{AFT}})$, $p_{vx}(\tau_{\text{rel1}+}^{\text{AFT}})$, $p_{vy}(\tau_{\text{rel1}+}^{\text{AFT}})$, and $p_{vz}(\tau_{\text{rel1}+}^{\text{AFT}})$; (The duration of the second passive segment (AFT release), $\Delta\tau_2^{\text{pass}}$, is a parameter of the problem (120 s), being therefore not included in the shooting parameter vector.)
- the coordinates and velocities as well as the values of the conjugate variables after engine activation, $x(\tau_{\text{rel2}+}^{\text{AFT}})$, $y(\tau_{\text{rel2}+}^{\text{AFT}})$, $z(\tau_{\text{rel2}+}^{\text{AFT}})$, $v_x(\tau_{\text{rel2}+}^{\text{AFT}})$, $v_y(\tau_{\text{rel2}+}^{\text{AFT}})$, $v_z(\tau_{\text{rel2}+}^{\text{AFT}})$, $p_x(\tau_{\text{rel2}+}^{\text{AFT}})$, $p_y(\tau_{\text{rel2}+}^{\text{AFT}})$, $p_z(\tau_{\text{rel2}+}^{\text{AFT}})$, $p_{vx}(\tau_{\text{rel2}+}^{\text{AFT}})$, $p_{vy}(\tau_{\text{rel2}+}^{\text{AFT}})$, and $p_{vz}(\tau_{\text{rel2}+}^{\text{AFT}})$, including the conjugate variable corresponding to the mass, $p_m(\tau_{\text{rel2}}^{\text{AFT}}) = p_m(\tau_{\text{rel2}+}^{\text{AFT}})$; (The mass of the spacecraft after AFT release is not included in the shooting parameter vector and is calculated by formulas (6) (the first problem statement) and (7) (the second problem statement).)
- the duration of the third active segment, $\Delta\tau_3^{\text{act}} = \Delta\tau_{\text{safe}}$, where the perigee altitude of the spacecraft orbit is increased to 200 km;
- the coordinates and velocities as well as the values of the conjugate variables after engine deactivation, $x(\tau_{\text{safe}+})$, $y(\tau_{\text{safe}+})$, $z(\tau_{\text{safe}+})$, $v_x(\tau_{\text{safe}+})$, $v_y(\tau_{\text{safe}+})$, $v_z(\tau_{\text{safe}+})$, $p_x(\tau_{\text{safe}+})$, $p_y(\tau_{\text{safe}+})$, $p_z(\tau_{\text{safe}+})$, $p_{vx}(\tau_{\text{safe}+})$, $p_{vy}(\tau_{\text{safe}+})$, and $p_{vz}(\tau_{\text{safe}+})$;
- the duration of the third passive segment, $\Delta\tau_3^{\text{pass}}$;
- the coordinates and velocities as well as the values of the conjugate variables after engine activation, $x(\tau_{3+}^{\text{pass}})$, $y(\tau_{3+}^{\text{pass}})$, $z(\tau_{3+}^{\text{pass}})$, $v_x(\tau_{3+}^{\text{pass}})$, $v_y(\tau_{3+}^{\text{pass}})$, $v_z(\tau_{3+}^{\text{pass}})$, $p_x(\tau_{3+}^{\text{pass}})$, $p_y(\tau_{3+}^{\text{pass}})$, $p_z(\tau_{3+}^{\text{pass}})$, $p_{vx}(\tau_{3+}^{\text{pass}})$, $p_{vy}(\tau_{3+}^{\text{pass}})$, and $p_{vz}(\tau_{3+}^{\text{pass}})$;
- the duration of the fourth active segment, $\Delta\tau_4^{\text{act}}$, at the end of which the spacecraft reaches the target orbit;
- the coordinates and velocities as well as the values of the conjugate variables after engine deactivation, $x(\tau_{4+}^{\text{act}})$, $y(\tau_{4+}^{\text{act}})$, $z(\tau_{4+}^{\text{act}})$, $v_x(\tau_{4+}^{\text{act}})$, $v_y(\tau_{4+}^{\text{act}})$, $v_z(\tau_{4+}^{\text{act}})$, $p_x(\tau_{4+}^{\text{act}})$, $p_y(\tau_{4+}^{\text{act}})$, $p_z(\tau_{4+}^{\text{act}})$, $p_{vx}(\tau_{4+}^{\text{act}})$, $p_{vy}(\tau_{4+}^{\text{act}})$, and $p_{vz}(\tau_{4+}^{\text{act}})$;
- the duration of the fourth passive segment, $\Delta\tau_4^{\text{pass}} = \Delta\tau_{\text{tar}}$, where the spacecraft moves in the target orbit; (For convenience of calculations, τ_{tar} is the last engine activation point for the CB release instead of the first point in the target orbit of the spacecraft; this is possible because the points are connected by the passive segment.)
- the coordinates and velocities as well as the values of the conjugate variables after engine activation, $x(\tau_{\text{tar}+})$, $y(\tau_{\text{tar}+})$, $z(\tau_{\text{tar}+})$, $v_x(\tau_{\text{tar}+})$, $v_y(\tau_{\text{tar}+})$, $v_z(\tau_{\text{tar}+})$, $p_x(\tau_{\text{tar}+})$, $p_y(\tau_{\text{tar}+})$, $p_z(\tau_{\text{tar}+})$, $p_{vx}(\tau_{\text{tar}+})$, $p_{vy}(\tau_{\text{tar}+})$, and $p_{vz}(\tau_{\text{tar}+})$, and the CB mass $m(\tau_{\text{tar}+})$ after undocking the satellite; (The conjugate variable $p_m(\tau_{\text{tar}+})$ is not included in the shooting parameter vector since p_m is continuous at the point τ_{tar} by (23).)
- the duration of the fifth active segment, $\Delta\tau_5^{\text{act}} = \Delta T$, where the CB perigee is lowered to 100 km (the conditional boundary of the atmosphere).

The residual vector function includes the following elements:

- the twelve continuity conditions of the phase and conjugate variables at the time instant τ_1^{act};

- the twelve continuity conditions of the phase and conjugate variables at the time instant τ_1^{pass};
- the six continuity conditions of the phase variables and the six implications of the transversality conditions at the time instant $\tau_{\text{rel1}}^{\text{AFT}}$ (19);
- the twelve continuity conditions of the phase and conjugate variables at the time instant $\tau_{\text{rel2}}^{\text{AFT}}$;
- the six continuity conditions of the phase variables and the six implications of the transversality conditions at the time instant τ_{safe} (19);
- the twelve continuity conditions of the phase and conjugate variables at the time instant τ_3^{pass};
- the twelve continuity conditions of the phase and conjugate variables at the time instant τ_4^{act};
- the six continuity conditions of the phase variables and the six implications of the transversality conditions at the time instant τ_{tar} (20);
- the zero value of the z-component of the Laplace vector (the second condition from (10));
- the condition of exhausting all fuel from the CB's main tank at the final time instant: formulas (16) and (17) in the first and second problem statements, respectively;
- the complementary slackness condition: a given value of the final ascent impulse from the target orbit to the geostationary orbit (the first condition from (24));
- the three conditions on the perigee of the spacecraft orbit at the time instants $\tau_{\text{rel1}}^{\text{AFT}}$, τ_{safe}, and T: formulas (5), (9), and (15), respectively;
- the four conditions on the switching function: $\chi(\tau_{1-}^{\text{act}}) = 0$, $\chi(\tau_{1+}^{\text{pass}}) = 0$, $\chi(\tau_{3+}^{\text{pass}}) = 0$, and $\chi(\tau_{4-}^{\text{act}}) = 0$;
- the six transversality conditions at the final time instant T;
- the implication of the transversality conditions with respect to the variable m: formulas (21) and (22) in the first and second problem statements, respectively (in the latter case, with $\lambda_{m\tau 2} = 0$);
- the three implications of the stationarity conditions, $H(\tau_{\text{rel2}+}^{\text{AFT}}) = H(\tau_{\text{rel1-}}^{\text{AFT}})$, $H(\tau_{\text{safe}+}) = H(\tau_{\text{safe-}})$, and $H(\tau_{\text{tar}+}) = H(\tau_{\text{tar-}})$;
- the stationarity condition at the final time instant, $H(T) = 0$;
- the normalization condition (26);
- in the second problem statement, also the second complementary slackness condition from (25). (The first complementary slackness condition from (25) is not included in the residual vector function, and the corresponding inequality (8) is verified after solving the problem: the strict inequality holds on the Pontryagin extremal, which matches the case $\lambda_{m\tau 2} = 0$.)

Thus, the first problem statement has one hundred and eighteen shooting parameters and one hundred and eighteen residuals; the second problem statement, one hundred and nineteen shooting parameters and one hundred and nineteen residuals. In other words, in both statements, the number of unknown parameters coincides with the number of equations for their determination.

In the Appendix, we present the Pontryagin extremal in the second problem statement with the second zonal harmonic and $n = 0.1$, $P_{\text{spe}} = 350$ s, $i_0 = 0.9$ rad, $\Delta v^* = 1.5$ km/s, $m(0) = 1$ ($M(0) = 22\,500$ kg), the AFT's dry mass $m^{\text{AFT}} = 0.052$ (which corresponds to the mass 1170 kg), the CB's dry mass $m^{\text{CB}} = 0.0635556$ (which corresponds to the mass 1430 kg), the maximum AFT fuel mass $m_{\text{fuel}}^{\text{AFT}} = 0.6488889$ (which corresponds to the mass 14 600 kg), the maximum CB fuel mass $m_{\text{fuel}}^{\text{CB}} = 0.2266667$ (which corresponds to the mass 5100 kg), and $R_{\max} = 280\,000$ km.

5. CONCLUSIONS

One result of the previous studies in the pulse statement was the possibility of releasing the additional fuel tank and the booster of the central block into the Earth's atmosphere at low cost. The same result has been confirmed above in the case of spacecraft with a high limited-thrust engine.

As it has turned out, the solution of the spacecraft transfer problem with a high limited-thrust engine is, to some extent, close to that obtained in the pulse statement. In the case under consideration, the method for passing from the latter solution to the former one [12] is effective: the Pontryagin extremal has been constructed.

The problems in the first and second statements (with optimizable design and fixed mass characteristics) have been successfully included in the parametric family. The transition from the solution of the first problem (with the chosen constants $\alpha = 0.08$ and $\beta = 0.01$) to the second one (see the extremal in the Appendix) has been effectively implemented the parameter continuation method.

The difference between the extremal considering the second zonal harmonic and the one without such consideration is small in the sense of convergence of Newton's method. (This method converges in 11 iterations.)

The numerical-analytical differentiation technique has demonstrated its effectiveness and advantages (the simplified program code and the reduced probability of programming errors).

The problem hierarchy methodology has been adopted to solve the original problem, choose an appropriate computational scheme and a good initial approximation, and cope with the difficulties of numerical solution due to the complexity and bulkiness of the problem statement, thus demonstrating its effectiveness. The candidate's dissertation by A.S. Samokhin [22] is another complete example of the effective application of this methodology. Note that when the maximum possible distance $R_{\max}$ of the spacecraft to the Earth increases, the problem statement changes: it will be necessary to consider the influence of the Moon's gravitational field [21]; moreover, the resulting problem will be in the next steps of the problem hierarchy methodology.

APPENDIX

The Pontryagin Extremal in the Second Problem Statement with the Second Zonal Harmonic

Here, we present the Pontryagin extremal in the second problem statement with the second zonal harmonic and $n = 0.1$, $P_{\text{spe}} = 350$ s, $i_0 = 0.9$ rad, $\Delta v^* = 1.5$ km/s, $m(0) = 1$ ($M(0) = 22\,500$ kg), the AFT's dry mass $m^{\text{AFT}} = 0.052$ (which corresponds to the mass 1170 kg), the CB's dry mass $m^{\text{CB}} = 0.0635556$ (which corresponds to the mass 1430 kg), the maximum AFT fuel mass $m^{\text{AFT}}_{\text{fuel}} = 0.6488889$ (which corresponds to the mass 14 600 kg), the maximum CB fuel mass $m^{\text{CB}}_{\text{fuel}} = 0.2266667$ (which corresponds to the mass 5100 kg), and $R_{\max} = 280\,000$ km.

The numerical Lagrange multipliers are:

$$\lambda_{R0} = 0.000143381, \quad \lambda_{C0} = 0.000591830,$$
$$\lambda_{vx0} = 0.454024806, \quad \lambda_{vy0} = 0.573821987, \quad \lambda_{vz0} = 0.681608247,$$
$$\lambda_{m0} = 0.003461557, \quad \lambda_{x\text{rel}1} = -0.000341292, \quad \lambda_{y\text{rel}1} = 7.242061031 \times 10^{-7},$$
$$\lambda_{z\text{rel}1} = -4.016946803 \times 10^{-7}, \quad \lambda_{vx\text{rel}1} = -0.006462504, \quad \lambda_{vy\text{rel}1} = -0.749532226,$$
$$\lambda_{vz\text{rel}1} = -0.815268562, \quad \lambda_{x\text{safe}} = -4.192288919 \times 10^{-5}, \quad \lambda_{y\text{safe}} = 1.862663610 \times 10^{-6},$$
$$\lambda_{z\text{safe}} = 3.375542260 \times 10^{-6}, \quad \lambda_{vx\text{safe}} = 0.005756096, \quad \lambda_{vy\text{safe}} = 0.122589304,$$
$$\lambda_{vz\text{safe}} = 0.271310189, \quad \lambda_{x\text{tar}} = 2.712182075 \times 10^{-7}, \quad \lambda_{y\text{tar}} = -1.138471581 \times 10^{-9},$$
$$\lambda_{z\text{tar}} = -6.622593433 \times 10^{-13}, \quad \lambda_{vx\text{tar}} = 0.000524475, \quad \lambda_{vy\text{tar}} = 0.124996420,$$
$$\lambda_{vz\text{tar}} = 0.154483787, \quad \lambda_{\tau\text{rel}1} = 1.017515559 \times 10^{-10}, \quad \lambda_{\text{rel}1} = 0.000457417,$$
$$\lambda_{m\tau 1} = 0.005674790, \quad \lambda_{m\tau 2} = 0, \quad \lambda_{\tau} = -7.385740053 \times 10^{-11},$$
$$\lambda_{\tau\text{rel}2} = 8.771225602 \times 10^{-11}, \quad \lambda_{\tau\text{safe}} = 1.237010821 \times 10^{-11}, \quad \lambda_{\text{safe}} = -0.000291125,$$
$$\lambda_{\tau\text{tar}} = -1.948360860 \times 10^{-19}, \quad \lambda_{\text{tar}} = -0.020742502, \quad \lambda_{\text{fa}} = 1.164436713,$$
$$\lambda_T = 4.669488715 \times 10^{-6}, \quad \lambda_{mT1} = -0.009927899, \quad \lambda_{mT2} = 0.000804002, \quad \lambda_0 = 0.009915863.$$

The spacecraft engine is activated at the initial time instant $t = 0$ under the angular position $\varphi_0 = -0.7817944$ rad in the reference orbit:

$$x(0) = 4668.258 \text{ km}, \quad y(0) = -2880.996 \text{ km}, \quad z(0) = -3630.510 \text{ km},$$
$$v_x(0) = 5.484390 \text{ km/s}, \quad v_y(0) = 3.433812 \text{ km/s}, \quad v_z(0) = 4.327147 \text{ km/s},$$
$$p_x(0) = 0.000284790, \quad p_y(0) = -0.000515932, \quad p_z(0) = -0.000601403,$$
$$p_{vx}(0) = 0.454024806, \quad p_{vy}(0) = 0.573821987, \quad p_{vz}(0) = 0.681608247,$$
$$m(0) = 1, \quad p_m(0) = 0.003461557.$$

The duration of the first active segment is $\Delta\tau_1^{\text{act}} = 1234.190$ s. The spacecraft moves to the elliptic orbit with the apogee $r_{\text{a1}} = 15\,500.572$ km, the perigee $r_{\text{p1}} = 6702.795$ km, and the inclination angle $i_1 = 0.8956402$ rad. The coordinates, velocities, and mass of the spacecraft at the engine activation time instant τ_1^{act} are

$$x(\tau_{1-}^{\text{act}}) = x(\tau_{1+}^{\text{act}}) = 5360.198 \text{ km}, \quad y(\tau_{1-}^{\text{act}}) = y(\tau_{1+}^{\text{act}}) = 3045.731 \text{ km},$$
$$z(\tau_{1-}^{\text{act}}) = z(\tau_{1+}^{\text{act}}) = 3807.202 \text{ km}, \quad v_x(\tau_{1-}^{\text{act}}) = v_x(\tau_{1+}^{\text{act}}) = -4.376498 \text{ km/s},$$
$$v_y(\tau_{1-}^{\text{act}}) = v_y(\tau_{1+}^{\text{act}}) = 4.635010 \text{ km/s}, \quad v_z(\tau_{1-}^{\text{act}}) = v_z(\tau_{1+}^{\text{act}}) = 5.786158 \text{ km/s},$$
$$p_x(\tau_{1-}^{\text{act}}) = p_x(\tau_{1+}^{\text{act}}) = 0.000331052, \quad p_y(\tau_{1-}^{\text{act}}) = p_y(\tau_{1+}^{\text{act}}) = 0.000386951,$$
$$p_z(\tau_{1-}^{\text{act}}) = p_z(\tau_{1+}^{\text{act}}) = 0.000452488, \quad p_{vx}(\tau_{1-}^{\text{act}}) = p_{vx}(\tau_{1+}^{\text{act}}) = -0.371919208,$$
$$p_{vy}(\tau_{1-}^{\text{act}}) = p_{vy}(\tau_{1+}^{\text{act}}) = 0.637424282, \quad p_{vz}(\tau_{1-}^{\text{act}}) = p_{vz}(\tau_{1+}^{\text{act}}) = 0.754898302,$$
$$m(\tau_1^{\text{act}}) = 0.6473743, \quad p_m(\tau_1^{\text{act}}) = 0.005597245.$$

The duration of the first passive segment is $\Delta\tau_1^{\text{pass}} = 5219.504$ s. At the end of this segment, the spacecraft is in the orbit with the apogee $r_{\text{a2}} = 15\,497.241$ km, the perigee $r_{\text{p2}} = 6704.141$ km, and the inclination angle $i_2 = 0.8956703$ rad. The coordinates, velocities, and mass of the spacecraft at the engine activation time instant τ_1^{pass} are

$$x(\tau_{1-}^{\text{pass}}) = x(\tau_{1+}^{\text{pass}}) = -15495.958 \text{ km}, \quad y(\tau_{1-}^{\text{pass}}) = y(\tau_{1+}^{\text{pass}}) = 133.386 \text{ km},$$
$$z(\tau_{1-}^{\text{pass}}) = z(\tau_{1+}^{\text{pass}}) = 131.434 \text{ km}, \quad v_x(\tau_{1-}^{\text{pass}}) = v_x(\tau_{1+}^{\text{pass}}) = -0.061410 \text{ km/s},$$
$$v_y(\tau_{1-}^{\text{pass}}) = v_y(\tau_{1+}^{\text{pass}}) = -2.462966 \text{ km/s}, \quad v_z(\tau_{1-}^{\text{pass}}) = v_z(\tau_{1+}^{\text{pass}}) = -3.076413 \text{ km/s},$$
$$p_x(\tau_{1-}^{\text{pass}}) = p_x(\tau_{1+}^{\text{pass}}) = 0.000121015, \quad p_y(\tau_{1-}^{\text{pass}}) = p_y(\tau_{1+}^{\text{pass}}) = -2.300740522 \times 10^{-6},$$
$$p_z(\tau_{1-}^{\text{pass}}) = p_z(\tau_{1+}^{\text{pass}}) = -3.527557800 \times 10^{-6}, \quad p_{vx}(\tau_{1-}^{\text{pass}}) = p_{vx}(\tau_{1+}^{\text{pass}}) = 0.007053866,$$
$$p_{vy}(\tau_{1-}^{\text{pass}}) = p_{vy}(\tau_{1+}^{\text{pass}}) = 0.600617145, \quad p_{vz}(\tau_{1-}^{\text{pass}}) = p_{vz}(\tau_{1+}^{\text{pass}}) = 0.868167234,$$
$$m(\tau_1^{\text{pass}}) = 0.6473743, \quad p_m(\tau_1^{\text{pass}}) = 0.005597245.$$

The duration of the second active segment is $\Delta\tau_2^{\text{act}} = \Delta\tau_{\text{rel1}}^{\text{AFT}} = 30.961$ s. The spacecraft moves to the elliptic orbit with the apogee $r_{\text{a3}} = 15\,497.241$ km, the perigee $r_{\text{p3}} = 6478.25$ km, and the inclination angle $i_3 = 0.8948234$ rad. This orbit touches the conditional boundary of the atmosphere. The coordinates, velocities, and mass of the spacecraft at the engine deactivation time instant $\tau_{\text{rel1}}^{\text{AFT}}$ are

$$x(\tau_{\text{rel1-}}^{\text{AFT}}) = x(\tau_{\text{rel1+}}^{\text{AFT}}) = -15\,497.060 \text{ km}, \quad y(\tau_{\text{rel1-}}^{\text{AFT}}) = y(\tau_{\text{rel1+}}^{\text{AFT}}) = 57.540 \text{ km},$$
$$z(\tau_{\text{rel1-}}^{\text{AFT}}) = z(\tau_{\text{rel1+}}^{\text{AFT}}) = 36.780 \text{ km}, \quad v_x(\tau_{\text{rel1-}}^{\text{AFT}}) = v_x(\tau_{\text{rel1+}}^{\text{AFT}}) = -0.009780 \text{ km/s},$$
$$v_y(\tau_{\text{rel1-}}^{\text{AFT}}) = v_y(\tau_{\text{rel1+}}^{\text{AFT}}) = -2.436415 \text{ km/s}, \quad v_z(\tau_{\text{rel1-}}^{\text{AFT}}) = v_z(\tau_{\text{rel1+}}^{\text{AFT}}) = -3.037856 \text{ km/s},$$
$$p_x(\tau_{\text{rel1-}}^{\text{AFT}}) = 0.000121065, \quad p_y(\tau_{\text{rel1-}}^{\text{AFT}}) = -3.087854169 \times 10^{-7}, \quad p_z(\tau_{\text{rel1-}}^{\text{AFT}}) = -6.465598107 \times 10^{-7},$$
$$p_{vx}(\tau_{\text{rel1-}}^{\text{AFT}}) = 0.003306216, \quad p_{vy}(\tau_{\text{rel1-}}^{\text{AFT}}) = 0.600657543, \quad p_{vz}(\tau_{\text{rel1-}}^{\text{AFT}}) = 0.868231852,$$
$$p_x(\tau_{\text{rel1+}}^{\text{AFT}}) = -0.000341292, \quad p_y(\tau_{\text{rel1+}}^{\text{AFT}}) = 7.242061031 \times 10^{-7}, \quad p_z(\tau_{\text{rel1+}}^{\text{AFT}}) = -4.016946803 \times 10^{-7},$$
$$p_{vx}(\tau_{\text{rel1+}}^{\text{AFT}}) = -0.006462504, \quad p_{vy}(\tau_{\text{rel1+}}^{\text{AFT}}) = -0.749532226, \quad p_{vz}(\tau_{\text{rel1+}}^{\text{AFT}}) = -0.815268562,$$
$$m(\tau_{\text{rel1-}}^{\text{AFT}}) = 0.6385284, \quad p_m(\tau_{\text{rel1-}}^{\text{AFT}}) = 0.005674790.$$

The duration of the second passive segment is $\Delta\tau_2^{\text{pass}} = 120$ s. On this segment, the AFT is undocked from the spacecraft. At the end of the second passive segment, the spacecraft is in the orbit with the apogee $r_{\text{a}4} = 15\,497.245$ km, the perigee $r_{\text{p}4} = 6478.246$ km, and the inclination angle $i_4 = 0.8948232$ rad. The coordinates, velocities, and mass of the spacecraft at the engine activation time instant $\tau_{\text{rel}2}^{\text{AFT}}$ are

$$x(\tau_{\text{rel}2-}^{\text{AFT}}) = x(\tau_{\text{rel}2+}^{\text{AFT}}) = -15486.279 \text{ km}, \quad y(\tau_{\text{rel}2-}^{\text{AFT}}) = y(\tau_{\text{rel}2+}^{\text{AFT}}) = -234.799 \text{ km},$$
$$z(\tau_{\text{rel}2-}^{\text{AFT}}) = z(\tau_{\text{rel}2+}^{\text{AFT}}) = -327.697 \text{ km}, \quad v_x(\tau_{\text{rel}2-}^{\text{AFT}}) = v_x(\tau_{\text{rel}2+}^{\text{AFT}}) = 0.189472 \text{ km/s},$$
$$v_y(\tau_{\text{rel}2-}^{\text{AFT}}) = v_y(\tau_{\text{rel}2+}^{\text{AFT}}) = -2.435275 \text{ km/s}, \quad v_z(\tau_{\text{rel}2-}^{\text{AFT}}) = v_z(\tau_{\text{rel}2+}^{\text{AFT}}) = -3.035984 \text{ km/s},$$
$$p_x(\tau_{\text{rel}2-}^{\text{AFT}}) = p_x(\tau_{\text{rel}2+}^{\text{AFT}}) = -0.000341191, \quad p_y(\tau_{\text{rel}2-}^{\text{AFT}}) = p_y(\tau_{\text{rel}2+}^{\text{AFT}}) = -8.913952605 \times 10^{-6},$$
$$p_z(\tau_{\text{rel}2-}^{\text{AFT}}) = p_z(\tau_{\text{rel}2+}^{\text{AFT}}) = -1.089032490 \times 10^{-5}, \quad p_{vx}(\tau_{\text{rel}2-}^{\text{AFT}}) = p_{vx}(\tau_{\text{rel}2+}^{\text{AFT}}) = 0.034488810,$$
$$p_{vy}(\tau_{\text{rel}2-}^{\text{AFT}}) = p_{vy}(\tau_{\text{rel}2+}^{\text{AFT}}) = -0.749040916, \quad p_{vz}(\tau_{\text{rel}2-}^{\text{AFT}}) = p_{vz}(\tau_{\text{rel}2+}^{\text{AFT}}) = -0.814591115,$$
$$m(\tau_{\text{rel}2+}^{\text{AFT}}) = 0.5865284, \quad p_m(\tau_{\text{rel}2+}^{\text{AFT}}) = 0.006478792.$$

The duration of the third active segment is $\Delta\tau_3^{\text{act}} = \Delta\tau_{\text{safe}} = 12.584$ s. The spacecraft moves to the elliptic orbit with the apogee $r_{\text{a}5} = 15\,497.362$ km, the perigee $r_{\text{p}5} = 6578.25$ km, and the inclination angle $i_5 = 0.8944602$ rad. This is the safe orbit. The coordinates, velocities, and mass of the spacecraft at the engine deactivation time instant τ_{safe} are

$$x(\tau_{\text{safe-}}) = x(\tau_{\text{safe+}}) = -15\,483.759 \text{ km}, \quad y(\tau_{\text{safe-}}) = y(\tau_{\text{safe+}}) = -265.532 \text{ km},$$
$$z(\tau_{\text{safe-}}) = z(\tau_{\text{safe+}}) = -365.996 \text{ km}, \quad v_x(\tau_{\text{safe-}}) = v_x(\tau_{\text{safe+}}) = 0.211071 \text{ km/s},$$
$$v_y(\tau_{\text{safe-}}) = v_y(\tau_{\text{safe+}}) = -2.449215 \text{ km/s}, \; v_z(\tau_{\text{safe-}}) = v_z(\tau_{\text{safe+}}) = -3.051042 \text{ km/s},$$
$$p_x(\tau_{\text{safe-}}) = -0.000341167, \quad p_y(\tau_{\text{safe-}}) = -9.925271032 \times 10^{-6},$$
$$p_z(\tau_{\text{safe-}}) = -1.199086427 \times 10^{-5}, \quad p_{vx}(\tau_{\text{safe-}}) = 0.038782187,$$
$$p_{vy}(\tau_{\text{safe-}}) = -0.748922381, \quad p_{vz}(\tau_{\text{safe-}}) = -0.814447148,$$
$$p_x(\tau_{\text{safe+}}) = -4.192288919 \times 10^{-5}, \quad p_y(\tau_{\text{safe+}}) = 1.862663610 \times 10^{-6},$$
$$p_z(\tau_{\text{safe+}}) = 3.375542260 \times 10^{-6}, \quad p_{vx}(\tau_{\text{safe+}}) = 0.005756096,$$
$$p_{vy}(\tau_{\text{safe+}}) = 0.122589304, \quad p_{vz}(\tau_{\text{safe+}}) = 0.271310189,$$
$$m(\tau_{\text{safe}}) = 0.5829330, \quad p_m(\tau_{\text{safe}}) = 0.006518753.$$

The duration of the third passive segment is $\Delta\tau_3^{\text{pass}} = 5213.308$ s. At the end of this segment, the spacecraft is in the orbit with the apogee $r_{\text{a}6} = 15\,511.458$ km, the perigee $r_{\text{p}6} = 6576.991$ km, and the inclination angle $i_6 = 0.8945149$ rad. The coordinates, velocities, and mass of the spacecraft at the engine activation time instant τ_3^{pass} are

$$x(\tau_{3-}^{\text{pass}}) = x(\tau_{3+}^{\text{pass}}) = 5800.915 \text{ km}, \quad y(\tau_{3-}^{\text{pass}}) = y(\tau_{3+}^{\text{pass}}) = -2325.058 \text{ km},$$
$$z(\tau_{3-}^{\text{pass}}) = z(\tau_{3+}^{\text{pass}}) = -2873.476 \text{ km}, \quad v_x(\tau_{3-}^{\text{pass}}) = v_x(\tau_{3+}^{\text{pass}}) = 3.552179 \text{ km/s},$$
$$v_y(\tau_{3-}^{\text{pass}}) = v_y(\tau_{3+}^{\text{pass}}) = 5.123342 \text{ km/s}, \quad v_z(\tau_{3-}^{\text{pass}}) = v_z(\tau_{3+}^{\text{pass}}) = 6.398453 \text{ km/s},$$
$$p_x(\tau_{3-}^{\text{pass}}) = p_x(\tau_{3+}^{\text{pass}}) = 0.000705371, \quad p_y(\tau_{3-}^{\text{pass}}) = p_y(\tau_{3+}^{\text{pass}}) = -0.000362590,$$
$$p_z(\tau_{3-}^{\text{pass}}) = p_z(\tau_{3+}^{\text{pass}}) = -0.000422128, \quad p_{vx}(\tau_{3-}^{\text{pass}}) = p_{vx}(\tau_{3+}^{\text{pass}}) = 0.375643830,$$
$$p_{vy}(\tau_{3-}^{\text{pass}}) = p_{vy}(\tau_{3+}^{\text{pass}}) = 0.672228633, \quad p_{vz}(\tau_{3-}^{\text{pass}}) = p_{vz}(\tau_{3+}^{\text{pass}}) = 0.795432792,$$
$$m(\tau_3^{\text{pass}}) = 0.5829330, \quad p_m(\tau_3^{\text{pass}}) = 0.006518753.$$

The duration of the fourth active segment is $\Delta\tau_4^{\text{act}} = 780.500$ s. The spacecraft moves to the target orbit with the apogee $r_{\text{a}7} = 227\,835.611$ km, the perigee $r_{\text{p}7} = 6644.321$ km, and the inclination angle $i_7 = 0.8906535$ rad. The coordinates, velocities, and mass of the spacecraft at the engine

deactivation time instant τ_4^{act} are

$$x(\tau_{4-}^{\text{act}}) = x(\tau_{4+}^{\text{act}}) = 6084.753 \text{ km}, \quad y(\tau_{4-}^{\text{act}}) = y(\tau_{4+}^{\text{act}}) = 2384.542 \text{ km},$$
$$z(\tau_{4-}^{\text{act}}) = z(\tau_{4+}^{\text{act}}) = 2973.927 \text{ km}, \quad v_x(\tau_{4-}^{\text{act}}) = v_x(\tau_{4+}^{\text{act}}) = -2.938015 \text{ km/s},$$
$$v_y(\tau_{4-}^{\text{act}}) = v_y(\tau_{4+}^{\text{act}}) = 6.263591 \text{ km/s}, \quad v_z(\tau_{4-}^{\text{act}}) = v_z(\tau_{4+}^{\text{act}}) = 7.730739 \text{ km/s},$$
$$p_x(\tau_{4-}^{\text{act}}) = p_x(\tau_{4+}^{\text{act}}) = 0.000708160, \quad p_y(\tau_{4-}^{\text{act}}) = p_y(\tau_{4+}^{\text{act}}) = 0.000286427,$$
$$p_z(\tau_{4-}^{\text{act}}) = p_z(\tau_{4+}^{\text{act}}) = 0.000340122, \quad p_{vx}(\tau_{4-}^{\text{act}}) = p_{vx}(\tau_{4+}^{\text{act}}) = -0.318610391,$$
$$p_{vy}(\tau_{4-}^{\text{act}}) = p_{vy}(\tau_{4+}^{\text{act}}) = 0.697696303, \quad p_{vz}(\tau_{4-}^{\text{act}}) = p_{vz}(\tau_{4+}^{\text{act}}) = 0.821832874,$$
$$m(\tau_4^{\text{act}}) = 0.3599331, \quad p_m(\tau_4^{\text{act}}) = 0.010719865.$$

In the target orbit, the satellite is separated from the CB. The satellite mass in the target orbit (payload mass) is $m_{\text{p}} = 0.2963061$ (6666.888 kg). At the time instant τ_{tar} the last engine activation occurs to lower the perigee altitude of the CB orbit to the conditional boundary of the atmosphere. For convenience of calculations, the mass jump (after undocking the satellite) is considered at the last engine activation instant. The duration of the fourth passive section (passive flight of the CB in the target orbit) is $\Delta\tau_4^{\text{pass}} = \Delta\tau_{\text{tar}} = 197\,376.995$ s. At the end of this passive segment, the spacecraft is in the orbit with the apogee $r_{\text{a8}} = 226\,259.913$ km, the perigee $r_{\text{p8}} = 6643.293$ km, and the inclination angle $i_8 = 0.8905128$ rad. The coordinates and mass of the CB at the engine activation time instant τ_{tar} are

$$x(\tau_{\text{tar-}}) = x(\tau_{\text{tar+}}) = -226\,257.921 \text{ km}, \quad y(\tau_{\text{tar-}}) = y(\tau_{\text{tar+}}) = 949.323 \text{ km},$$
$$z(\tau_{\text{tar-}}) = z(\tau_{\text{tar+}}) = 0.031 \text{ km}, \quad v_x(\tau_{\text{tar-}}) = v_x(\tau_{\text{tar+}}) = -0.000838 \text{ km/s},$$
$$v_y(\tau_{\text{tar-}}) = v_y(\tau_{\text{tar+}}) = -0.199407 \text{ km/s}, \quad v_z(\tau_{\text{tar-}}) = v_z(\tau_{\text{tar+}}) = -0.246448 \text{ km/s},$$
$$p_x(\tau_{\text{tar-}}) = -7.173006000 \times 10^{-7}, \quad p_y(\tau_{\text{tar-}}) = 1.993046144 \times 10^{-9},$$
$$p_z(\tau_{\text{tar-}}) = -3.571982769 \times 10^{-8}, \quad p_{vx}(\tau_{\text{tar-}}) = -0.003979552,$$
$$p_{vy}(\tau_{\text{tar-}}) = -0.672797915, \quad p_{vz}(\tau_{\text{tar-}}) = 0.617436221,$$
$$p_x(\tau_{\text{tar+}}) = 2.712182075 \times 10^{-7}, \quad p_y(\tau_{\text{tar+}}) = -1.138471581 \times 10^{-9},$$
$$p_z(\tau_{\text{tar+}}) = -6.622593433 \times 10^{-13}, \quad p_{vx}(\tau_{\text{tar+}}) = 0.000524475,$$
$$p_{vy}(\tau_{\text{tar+}}) = 0.124996420, \quad p_{vz}(\tau_{\text{tar+}}) = 0.154483787,$$
$$m(\tau_{\text{tar-}}) = 0.3599331, \quad m(\tau_{\text{tar+}}) = 0.0636269,$$
$$p_m(\tau_{\text{tar-}}) = p_m(\tau_{\text{tar+}}) = 0.010719865.$$

The duration of the fifth (last) active segment is $\Delta\tau_5^{\text{act}} = \Delta T = 0.250$ s. The spacecraft moves to the orbit touching the conditional boundary of the atmosphere with the apogee $r_{\text{a9}} = 226\,259.913$ km, the perigee $r_{\text{p9}} = 6478.25$ km, and the angle of inclination $i_9 = 0.8905128$ rad. The coordinates, velocities, and mass of the CB at the engine deactivation time instant T are

$$x(T) = -226\,257.922 \text{ km}, \quad y(T) = 949.274 \text{ km},$$
$$z(T) = -0.030 \text{ km}, \quad v_x(T) = -0.000826 \text{ km/s},$$
$$v_y(T) = -0.196984 \text{ km/s}, \quad v_z(T) = -0.243454 \text{ km/s},$$
$$p_x(T) = 2.712182120 \times 10^{-7}, \quad p_y(T) = -1.137397235 \times 10^{-9},$$
$$p_z(T) = 6.655346255 \times 10^{-13}, \quad p_{vx}(T) = 0.000524407,$$
$$p_{vy}(T) = 0.124996420, \quad p_{vz}(T) = 0.154483787,$$
$$m(T) = 0.0635556, \quad p_m(T) = 0.010731902.$$

The final ascent impulses to transfer the satellite from the target orbit to the GEO are

$$\Delta v_{\text{fa1}} = 0.029677 \text{ km/s}, \quad \Delta v_{\text{fa2}} = 0.491271 \text{ km/s}, \quad \Delta v_{\text{fa3}} = 0.979052 \text{ km/s}.$$

The fuel consumed to lower the perigee altitude to 100 km (to release the AFT) is 199.034 kg (0.0088460). The fuel consumed to raise the perigee altitude to 200 km (to reach the safe orbit) is 80.897 kg (0.0035954). The fuel consumed to lower the perigee altitude to 100 km (to release the CB) is 1.606 kg (7.1361228×10^{-5}). The total fuel consumption for releasing the AFT and CB constitutes 281.536 kg (0.0125127).

The correspondence of the phase and conjugate variables at the starts and ends of the passive segments can be verified by numerical integration. The conditions of Pontryagin's maximum principle can be verified by substituting the phase and conjugate variables and the numerical Lagrange multipliers into the corresponding formulas, and numerical-analytical differentiation can be used to verify the transversality conditions. The basic dimensional units in the calculations are 1000 km and 1 s. When passing to other dimensional units, the conjugate variables must be recalculated by appropriate formulas. The 8(7)th order Dorman–Prince method was used for numerical integration.

REFERENCES

1. Grigoriev, I.S. and Proskuryakov, A.I., Spacecraft Transfer Optimization with Releasing the Additional Fuel Tank and the Booster to the Earth Atmosphere, *Autom. Remote Control*, 2023, vol. 84, no. 3, pp. 211–225.
2. Makarov, Yu.N., Space Debris Monitoring. Problems and Solutions, *Nanoindustry*, 2019, no. 1(87), pp. 6–14.
3. Vedeshin, L.A., Concept of a System for Monitoring and Control of Near Space Environmental Condition, *Transactions of IAA RAS*, 2019, no. 51, pp. 26–31.
4. Pikalov, R.S. and Yudintsev, V.V., Bulky Space Debris Removal Means Review and Selection, *Tr. MAI*, 2018, no. 100.
5. Shan, M., Guo, J., and Gill, E., Review and Comparison of Active Space Debris Capturing and Removal Methods, *Progr. Aerospac. Sci.*, 2015, vol. 80, pp. 18–32.
6. Trushlyakov, V.I. and Yutkin, E.A., A Review of Docking and Capture Means for Large-Size Space Debris Objects, *Omskii Mauchnyi Vest.*, 2013, no. 2, pp. 56–61.
7. Pelton, J.N., *New Solutions for the Space Debris Problem*, Springer, 2015.
8. Grigoryev, I.S. and Proskuryakov, A.I., Optimization of the Spacecraft Final Orbit and the Trajectory of the Apsidal Impulse Launch, with due Regard to Spent Stage Jettisons into the Atmosphere, *Engineering Journal: Science and Innovation*, 2019, no. 4(88). https://doi.org/10.18698/2308-6033-2019-4-1869
9. Grigoriev, I.S. and Proskuryakov, A.I., Spacecraft Pulsed Flights Trajectories with the Stages Jettison into the Atmosphere and Phase Restriction (Part I), *Engineering Journal: Science and Innovation*, 2019, no. 9(93). https://doi.org/10.18698/2308-6033-2019-9-1917
10. Grigoriev, I.S. and Proskuryakov, A.I., Spacecraft Pulsed Flights Trajectories with the Stages Jettison into the Atmosphere and Phase Restriction (Part II), *Engineering Journal: Science and Innovation*, 2019, no. 10(94). https://doi.org/10.18698/2308-6033-2019-9-1925
11. Grodzovskii, G.L., Ivanov, Yu.N., and Tokarev, V.V., *Mekhanika kosmicheskogo poleta. Problemy optimizatsii* (Mechanics of Space Flight. Optimization Problems), Moscow: Nauka, 1975.
12. Grigoriev, I.S. and Grigoriev, K.G., The Use of Solutions to Problems of Spacecraft Trajectory Optimization in Impulse Formulation When Solving the Problems of Optimal Control of Trajectories of a Spacecraft with Limited Thrust Engine: I, *Cosmic Res.*, 2007, vol. 45, pp. 339–347.
13. Grigoriev, I.S., *Metodicheskoe posobie po chislennym metodam resheniya kraevykh zadach printsipa maksimuma v zadachakh optimal'nogo upravleniya* (Manual in Numerical Methods of Solving the Boundary Value Problems of the Maximum Principle in Optimal Control Problems), Moscow: Mosk. Gos. Univ., 2005.

14. Grigoriev, K.G., Grigoriev, I.S., and Zapletin, M.P., *Praktikum po chislennym metodam v zadachakh optimal'nogo upravleniya* (Tutorial on Numerical Methods in Optimal Control Problems), Moscow: Center for Applied Research, the Faculty of Mechanics and Mathematics, Moscow State University, 2007.

15. Hairer, E., Norsett, S.P., and Wanner, G., *Solving Ordinary Differential Equations: I. Nonstiff Problems*, Berlin: Springer, 1987. Translated under the title *Resheniye obyknovennykh differentsial'nykh uravnenii*, Moscow: Mir, 1990.

16. Isaev, V.K. and Sonin, V.V., On a Modification of Newton's Methods for the Numerical Solution of Boundary Problems, *USSR Comput. Math. Math. Phys.*, 1963, vol. 3, no. 6, pp. 1525–1528.

17. Fedorenko, R.P., *Vvedenie v vychislitel'nuyu fiziku* (Introduction to Computational Physics), Moscow: Mosk. Fiz.-Tekhn. Inst., 1994.

18. McCracken, D.D. and Dorn, W.S., *Numerical Methods and FORTRAN Programming*, New York: Wiley, 1964. Translated under the title *Chislennye metody i programmirovanie na FORTRANe*, Moscow: Mir, 1977.

19. Numerical-analytical Differentiation ext_value. http://mech.math.msu.su/ iliagri/ext_value.htm

20. Duboshin, G.N., *Spravochnoe rukovodstvo po nebesnoi mekhanike i astrodinamike* (Handbook of Celestial Mechanics and Astrodynamics), Moscow: Nauka, 1976.

21. Ivashkin, V.V. and Tupitsyn, N.N., Using the Moon's Gravitational Field for Inserting a Spacecraft into a Stationary Earth Satellite Orbit, *Kosm. Issled.*, 1971, vol. 9, no. 2, pp. 163–172.

22. Samokhin, A.S., A Method for Constructing Pontryagin Extremals in End-to-End Trajectory Optimization of Interplanetary Transfers Considering Planetocentric Sections, *Cand. Sci. (Phys.–Math.) Dissertation*, Moscow, 2021.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 1, pp. 35–48.

NONLINEAR SYSTEMS

Output Stabilization of Lurie-Type Nonlinear Systems in a Given Set

B.H. Nguyen[*,**]

[*] *Institute for Problems in Mechanical Engineering, Russian Academy of Sciences, St. Petersburg, Russia*
[**] *ITMO University, St. Petersburg, Russia*
e-mail: leningrat206@gmail.com

Received March 18, 2023
Revised October 10, 2023
Accepted December 21, 2023

Abstract—This paper considers the problem of stabilizing the output variables of a Lurie-type nonlinear system in a given set at any time instant. A special output transformation is used to reduce the original constrained problem to that of analyzing the input-to-state stability of a new extended system without constraints. For this system, nonlinear control laws are obtained using the technique of linear matrix inequalities. Examples are given to illustrate the effectiveness of the method proposed and confirm the theoretical conclusions.

Keywords: Lurie-type nonlinear system, stabilization, nonlinear control, coordinate transformation, stability, linear matrix inequalities

DOI: 10.31857/S0005117924010038

1. INTRODUCTION

Guaranteeing the desired quality of transients is a key criterion in the design of automatic control systems. Classical control methods, such as modal control [1], adaptive robust control [2, 3], etc., ensure control performance only in the steady-state mode. The transient mode remains uncontrollable.

Control problems for linear plants with a guarantee for the controlled variable to stay in a given set at any time instant were presented in [4–6]. Within this approach, control performance is ensured not only in the steady-state mode but also in the transient mode. Such problems often arise in practice, e.g., when controlling electric power systems to maintain the frequency and voltage of electric generators in specified ranges [7, 8], when stabilizing the formation pressure of oil production, where the pressure at the wellhead must strictly belong to a given band [9], etc. To solve such problems, the authors [4, 5] proposed a method based on a special output transformation that reduces the original control problem with output constraints to a new control problem without constraints on the auxiliary variable. For the class of linear plants, the corresponding control problems were well studied and solved in [4]. However, they remain open for Lurie-type nonlinear systems.

Below we consider Lurie-type systems with an unstable linear part and unknown bounded disturbances and pose the problem of stabilizing such systems in a given set of output variables. The remainder of this paper is organized as follows. Section 2 formulates the problem of stabilizing the controlled variables of Lurie-type nonlinear systems in given sets. In section 3, control design methods are proposed. Finally, section 4 provides some numerical examples in MATLAB to illustrate the theoretical results.

The following notations are used in the presentation: $\mathbb{R}^n$ is the n-dimensional Euclidean space with the Euclidean norm $|\cdot|$; $\mathbb{R}^{n\times m}$ is the set of all real matrices of dimensions $n\times m$ with the Euclidean norm $\|\cdot\|$; for $A\in\mathbb{R}^{n\times n}$, the relation $A\succ 0$ ($A\prec 0$) means that A is a positive (negative, respectively) definite matrix, whereas the relation $A\succeq 0$ ($A\preceq 0$) means that A is a nonnegative (nonpositive, respectively) definite matrix; $I, 0$, and $diag\{\cdot\}$ are identity, zero, and diagonal, respectively, matrices of appropriate dimensions; $\mathbf{1_m}\in\mathbb{R}^m$ is an m dimensional vector composed of unit elements; $col\{\cdot\}\in\mathbb{R}^m$ is a column vector in the space $\mathbb{R}^m$; finally, the symbol "$\star$" indicates a symmetric block in a symmetric matrix.

2. PROBLEM STATEMENT

Consider a Lurie-type nonlinear system of the form

$$\begin{gathered}\dot{x}(t)=Ax(t)+Bu(t)+G\phi(z(t))+Df(t),\\ y(t)=Lx(t),\quad z(t)=Cx(t),\end{gathered}\tag{1}$$

where $t\geqslant 0$; $x(t)\in\mathbb{R}^n$ is the vector of measured states; $u(t)\in\mathbb{R}^m$ is the control variable (input); $y(t)=col\{y_1(t),\dots,y_m(t)\}\in\mathbb{R}^m$ is the controlled output; $f(t)\in\mathbb{R}^l$ is an unknown disturbance such that $|f(t)|\leqslant\bar{f}$; $z(t)\in\mathbb{R}^q$ is the argument of the nonlinearity ϕ; the matrices $A\in\mathbb{R}^{n\times n}$, $B\in\mathbb{R}^{n\times m}$, $G\in\mathbb{R}^{n\times q}$, $D\in\mathbb{R}^{n\times l}$, $L\in\mathbb{R}^{m\times n}$, and $C\in\mathbb{R}^{q\times n}$ are known. The pair of matrices (A,B) is controllable, and the pair of matrices (A,L) is observable. System (1) has a relative degree of $\mathbf{1_m}$, i.e., $det(LB)\neq 0$ [10, 11]. The unknown nonlinearity $\phi(\cdot):\mathbb{R}^q\to\mathbb{R}^q$ satisfies sector constraints: for all $z, \phi(z)=col\{\phi_1(z_1),\dots,\phi_q(z_q)\}\in\mathbb{R}^q$,

$$k_{1i}\leqslant\frac{\phi_i(z_i)}{z_i}\leqslant k_{2i},\quad \forall z_i\neq 0,\ i=1,\dots,q,\tag{2}$$

where k_{1i} and k_{2i} are some known constants.

This paper aims at designing a control law that stabilizes the output $y(t)$ of the plant (1) in a given set at any time instant:

$$\mathcal{Y}=\left\{y(t)\in\mathbb{R}^m:\underline{g}_i(t)<y_i(t)<\overline{g}_i(t),\ i=1,\dots,m\right\},\quad\forall t\geqslant 0,\tag{3}$$

where $\underline{g}_i(t)$ and $\overline{g}_i(t)$ are bounded differentiable functions with bounded first derivatives. These functions can be selected by the designers according to system performance requirements. To illustrate the objective of control, Fig. 1 shows a given pipe where the output must be at any time instant.

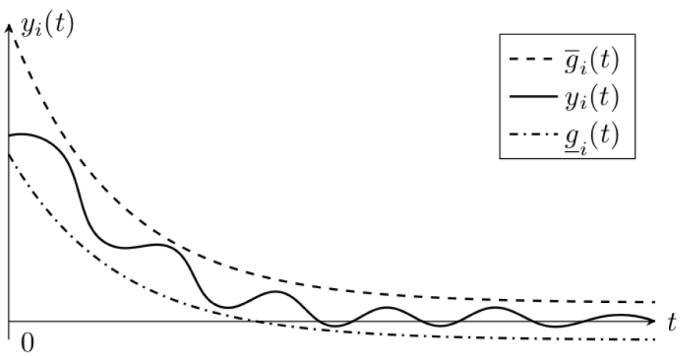


Fig. 1. The objective of control: one illustration.

3. SOLUTION METHOD

Following [4, 5], we introduce the output transformation

$$\varepsilon(t) = \Phi(y(t), t), \tag{4}$$

where $\varepsilon(t) = col\{\varepsilon_i(t), i = 1, \dots, m\} \in \mathbb{R}^m$ and $\Phi : \mathcal{Y} \times [0, \infty) \to \mathbb{R}^m$ is a differentiable function (with respect to all arguments) in the diagonal form that satisfies several conditions:

(a) There exists the inverse mapping

$$y = \Phi^{-1}(\varepsilon, t), \forall \varepsilon \in \mathbb{R}^m, \quad t \geqslant 0. \tag{5}$$

(b) The function $\Phi^{-1}(\varepsilon, t)$ is differentiable with respect to ε and t, and $\frac{\partial \Phi^{-1}(\varepsilon,t)}{\partial \varepsilon} \succ 0 \forall \varepsilon \in \mathbb{R}^m$ and $t \geqslant 0$.

(c) $\underline{g}_i(t) < \Phi_i^{-1}(\varepsilon_i, t) < \overline{g}_i(t),\ i = 1, \dots, m,\ \forall \varepsilon_i \in \mathbb{R}$ and $t \geqslant 0$.

(d) $\left|\frac{\partial \Phi^{-1}(\varepsilon,t)}{\partial t}\right| < \gamma$ for ε and $t \geqslant 0$, where $\gamma > 0$ is some constant defined by the transformation (4).

In this paper, the functions $\Phi_i^{-1}(\varepsilon_i, t)$ depend on $\varepsilon_i \in \mathbb{R}$ and t; therefore, the matrix $\frac{\partial \Phi^{-1}(\varepsilon,t)}{\partial \varepsilon}$ has the diagonal form. Some knowledge regarding the dynamics of the variable $\varepsilon(t)$ is needed to construct a control law. For this purpose, we take the total time derivative of the function $y(t)$ considering (5):

$$\dot{y} = \frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon}\dot{\varepsilon} + \frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial t}. \tag{6}$$

Due to (1) and $det\left(\frac{\partial \Phi^{-1}(\varepsilon,t)}{\partial \varepsilon}\right) \neq 0$, the expression (6) can be written as

$$\dot{\varepsilon} = \left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon}\right)^{-1} \left[LAx + LBu + LG\phi + LDf - \frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial t}\right]. \tag{7}$$

In (7), $LDf(t)$ and $\frac{\partial \Phi^{-1}(\varepsilon,t)}{\partial t}$ are bounded values. Applying the change $\psi(t) = LDf(t) - \frac{\partial \Phi^{-1}(\varepsilon,t)}{\partial t}$ yields $|\psi(t)| \leqslant \kappa$, where $\kappa = \|LD\|\bar{f} + \gamma$. In view of this change, (7) reduces to

$$\dot{\varepsilon} = \left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon}\right)^{-1} \Big[LAx + LBu + LG\phi + \psi\Big]. \tag{8}$$

We recall the main result of [4] to solve the problem.

Theorem 1. *Let conditions* (a)–(d) *hold for the transformation* (4). *If there exists a control law* $u(t)$ *under which the solutions of* (8) *and* (1) *are bounded, then* $y(t) \in \mathcal{Y}$.

Remark 1. Conditions (a)–(d) affect the choice of the transformation function (4) only: they do not guarantee condition (3). For example, if the trajectories $\varepsilon(t)$ tend to infinity in a finite time, $y(t)$ will converge to a boundary of the pipe defined by (3). Therefore, after choosing the output transformation function based on conditions (a)–(d), it is required to obtain a control law that would ensure the boundedness of $\varepsilon(t)$. Theorem 1 reduces the original control problem (1) with the constraints (3) on the output $y(t)$ to an auxiliary control problem without any constraints on the variable $\varepsilon(t)$.

Now we find a control law $u(t)$ ensuring the boundedness of $\varepsilon(t)$. Consider a Lyapunov function of the form $V = \frac{1}{2}\varepsilon^{\mathrm{T}}\varepsilon$. According to (8),

$$\dot{V} = \varepsilon^{\mathrm{T}}\dot{\varepsilon} = \varepsilon^{\mathrm{T}}\left(\frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon}\right)^{-1}\Big[LAx + LBu + LG\phi + \psi\Big]. \tag{9}$$

Let the set Ω be an open Euclidean ball in the space $\mathbb{R}^m$, i.e.,

$$\Omega = \Big\{\varepsilon\in\mathbb{R}^m : |\varepsilon| < \sqrt{2c},\ c>0\Big\}, \tag{10}$$

where c is a given positive number. The idea is to stabilize the trajectory $\varepsilon(t)$ in the set Ω. For $\varepsilon(t)$ to stay in Ω, it suffices to guarantee the negative derivative of the Lyapunov function for all ε outside the set Ω, i.e., $\dot{V} < 0\ \forall\varepsilon\notin\Omega$. (See the concept of input-to-state stability in the book [12].) The derivative of the Lyapunov function in (9) contains the matrix $\left(\frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon}\right)^{-1}$, which is positive definite. In particular, if the plant (1) is one-dimensional, then $\left(\frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon}\right)^{-1}$ is a positive scalar not affecting the sign of $\dot{V} < 0$. The control law can be constructed using the technique of linear matrix inequalities (LMIs) as described in [6]. Below, we present control design procedures for a particular case of one-dimensional systems and extend them to the general case of multidimensional ones.

Remark 2. The sector nonlinearity condition (2) can be written as the norm constraint $|\frac{\phi_i(z_i)}{z_i}| \leqslant \overline{k}_i = \max\{|k_{1i}|,\ |k_{2i}|\}$. Hence, it follows that $|\phi(z)| \leqslant \mu|z|$, where $\mu = \sqrt{q}\max_i\{\overline{k}_i\},\ i = 1,\dots,q$. When passing from the original nonlinearity sector to the new one, the nonlinearity range will be expanded: $[k_{1i}, k_{2i}] \subset [-\mu,\mu]$. Then a control law will be designed for any nonlinearity in the new sector. In other words, this law can handle any nonlinearity in the original sector as well (i.e., it has higher "robustness" to the nonlinearity).

3.1. One-Dimensional Systems

We define a piecewise continuous control law of the form

$$u = -(LB)^{-1}[K\varepsilon + LAx + \mu\mathrm{sgn}(\varepsilon)\|LG\|\|C\||x|], \tag{11}$$

where $K\in\mathbb{R}$ is the desired gain and

$$\mathrm{sgn}(\varepsilon) = \begin{cases} 1, & \varepsilon\geqslant 0,\\ -1, & \varepsilon < 0.\end{cases}$$

The following result is true for the one-dimensional system.

Theorem 2. *Let conditions* (a)–(d) *hold for the transformation* (4)*, and let* $\frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon} > 0$ *for any* $\varepsilon\in\mathbb{R}$ *and* $t\geqslant 0$*. For given numbers* c, $\alpha > 0$, *assume the existence of a positive number* K *and positive coefficients* τ_i, $i = 1, 2$, *such that*

$$\begin{bmatrix} -K+\alpha+0.5\tau_1 & 0.5\\ \star & -\tau_2\end{bmatrix} \leqslant 0,$$
$$-c\tau_1 + \kappa^2\tau_2 \leqslant 0. \tag{12}$$

Then the control law (11) *ensures the target condition* (3).

The proof of Theorem 2 is postponed to the Appendix.

3.2. Multidimensional Systems

Proposition 1. *Consider given block matrices*

$$M = \begin{bmatrix} Q & 0 \\ \star & Q \end{bmatrix} \succ 0, \quad N = \begin{bmatrix} N_{11} & N_{12} \\ \star & N_{22} \end{bmatrix} \prec 0,$$

where $Q, N_{11}, N_{12}, N_{21}, N_{22} \in \mathbb{R}^{n\times n}$ *are diagonal matrices. Then the matrix*

$$MN = \begin{bmatrix} QN_{11} & QN_{12} \\ \star & QN_{22} \end{bmatrix}$$

is negative definite.

Proposition 1 is used to prove the main result of this subsection, see below. The proof of Proposition 1 is provided in the Appendix.

We define a piecewise continuous control law of the form

$$u = -(LB)^{-1}\left[K\varepsilon + LAx + \bar{\sigma}\mu\mathrm{Sign}(\varepsilon)\left\|\left(\frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon}\right)^{-1}\right\|\|LG\|\|C\||x|\right], \tag{13}$$

where $K \in \mathbb{R}^{m\times m}$ is the gain matrix and $\bar{\sigma}$ is a constant determined by the transformation (4). In addition, $\mathrm{Sign}(\varepsilon) = col\{\mathrm{sgn}(\varepsilon_i), i = 1, \ldots, m\}$.

Substituting the control law (13) into (8) gives the closed loop system

$$\dot{\varepsilon} = \left(\frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon}\right)^{-1}\Big[-K\varepsilon - \bar{\sigma}\mu\mathrm{Sign}(\varepsilon)\left\|\left(\frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon}\right)^{-1}\right\|\|LG\|\|C\||x| + LG\phi + \psi\Big]. \tag{14}$$

We arrive at the following result.

Theorem 3. *Let conditions* (a)–(d) *hold for the transformation* (4)*, and let* $0 \prec \frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon} \preceq \overline{\sigma}I$ *for any* $\varepsilon \in \mathbb{R}^m$ *and* $t \geqslant 0$*. For a given number* $c > 0$*, assume the existence of a diagonal matrix* $K \in \mathbb{R}^{m\times m}$ *and positive coefficients* $\tau_i,\ i = 1, 2$*, such that*

$$\begin{aligned} &\begin{bmatrix} -K + [(0.5\tau_1 - \alpha)\sigma + \beta]I & 0.5I \\ \star & -\tau_2\sigma I \end{bmatrix} \preceq 0, \\ &-c\tau_1 + \kappa^2\tau_2 \leqslant 0 \end{aligned} \tag{15}$$

for any $\sigma \in (0, \overline{\sigma}]$ *and* $\alpha > 0,\ \beta > 0$*.*

Then the control law (13) *ensures the target condition* (3)*.*

The proof of Theorem 3 can be found in the Appendix.

Remark 3. The technique of LMIs and the S-procedure allow analyzing the input-to-state stability of the closed loop system under unknown bounded disturbances. Moreover, the gain for ε in (11), (13) can be obtained by finding an admissible solution of (12), (15), which is easy to do using widespread solvers for semidefinite programming problems (e.g., SEDUMI [14], SDPT3 [15], CSDP [16], and others.)

Remark 4. Obviously, the parameter c in (12), (15) is related to the radius of the open balls Ω attracting the system trajectories $\varepsilon(t)$: this radius equals $\sqrt{2c}$. Decreasing the value of c will reduce the radius of the ball and, in turn, the limit value of $\varepsilon(t)$. Therefore, a decrease in the limit value of $\varepsilon(t)$ will also reduce the fluctuation of the variable $y(t)$ in the set $\mathcal{Y}$ due to the exogenous disturbance $f(t)$.

4. NUMERICAL EXAMPLES

4.1. Example 1. One-Dimensional System

Consider an unstable plant of the form (1) with the following parameters:

$$A=\begin{bmatrix}0 & 1\\ 2 & -3\end{bmatrix},\ B=\begin{bmatrix}0\\ 1\end{bmatrix},\ G=\begin{bmatrix}0\\ 0.1\end{bmatrix},\ D=\begin{bmatrix}1\\ 1\end{bmatrix},\ L=\begin{bmatrix}2 & 1\end{bmatrix},\ C=\begin{bmatrix}1 & 2\end{bmatrix},$$
$$f(t)=0.1+\sin(3t)+0.5sat(d(t)),\ \phi(z)=\sin(z).$$

where $sat\{\cdot\}$ is the saturation function and $d(t)$ is a white noise with an intensity and sampling time of 0.1. Then $\bar{f}=1.6$ and $\mu=1$.

Let the function $\varepsilon(t)$ be specified as

$$\varepsilon(t)=\ln\left(\frac{y(t)-\underline{g}(t)}{\overline{g}(t)-y(t)}\right).$$

Consequently, the inverse function $\Phi^{-1}(\varepsilon(t),t)$ is given by

$$\Phi^{-1}(\varepsilon,t)=\frac{\overline{g}(t)e^{\varepsilon}+\underline{g}(t)}{e^{\varepsilon}+1}.$$

For all $\varepsilon\in\mathbb{R}$ and $t\geqslant 0$, we have

$$\frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon}=\frac{e^{\varepsilon}(\overline{g}(t)-\underline{g}(t))}{(e^{\varepsilon}+1)^2}>0$$

and

$$\left|\frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial t}\right|=\left|\frac{\dot{\overline{g}}(t)e^{\varepsilon}+\dot{\underline{g}}(t)}{e^{\varepsilon}+1}\right|\leqslant\max\left\{\sup_{t\geqslant 0}|\dot{\overline{g}}(t)|,\sup_{t\geqslant 0}|\dot{\underline{g}}(t)|\right\}=\gamma. \tag{16}$$

Let the functions $\underline{g}(t)$ and $\overline{g}(t)$ be specified as

$$\overline{g}(t)=\begin{cases}-3\cos(t)+0.2, & t<2\pi,\\ \cos(t)+2.2, & t\geqslant 2\pi,\end{cases}$$
$$\underline{g}(t)=\begin{cases}3\cos(t)-0.2, & t<2\pi,\\ \cos(t)+1.8, & t\geqslant 2\pi.\end{cases}$$

The control law (11) can be written as

$$u=-(LB)^{-1}\left[K\ln\left(\frac{y-\underline{g}}{\overline{g}-y}\right)+LAx+\mu\mathrm{sgn}\left(\ln\left(\frac{y-\underline{g}}{\overline{g}-y}\right)\right)\|LG\|\|C\||x|\right].$$

In view of (16), we find $\gamma=3$ and $\kappa=8.6$. Inequality (12) was solved using YALMIP [17] with SEDUMI. For $c=100$ and $\alpha=2$, the result is $\tau_1=3.10$, $\tau_2=4.14$, and $K=6.54$; for $c=0.1$ and $\alpha=2$, the result is $\tau_1=45.47$, $\tau_2=0.04$, and $K=35.36$.

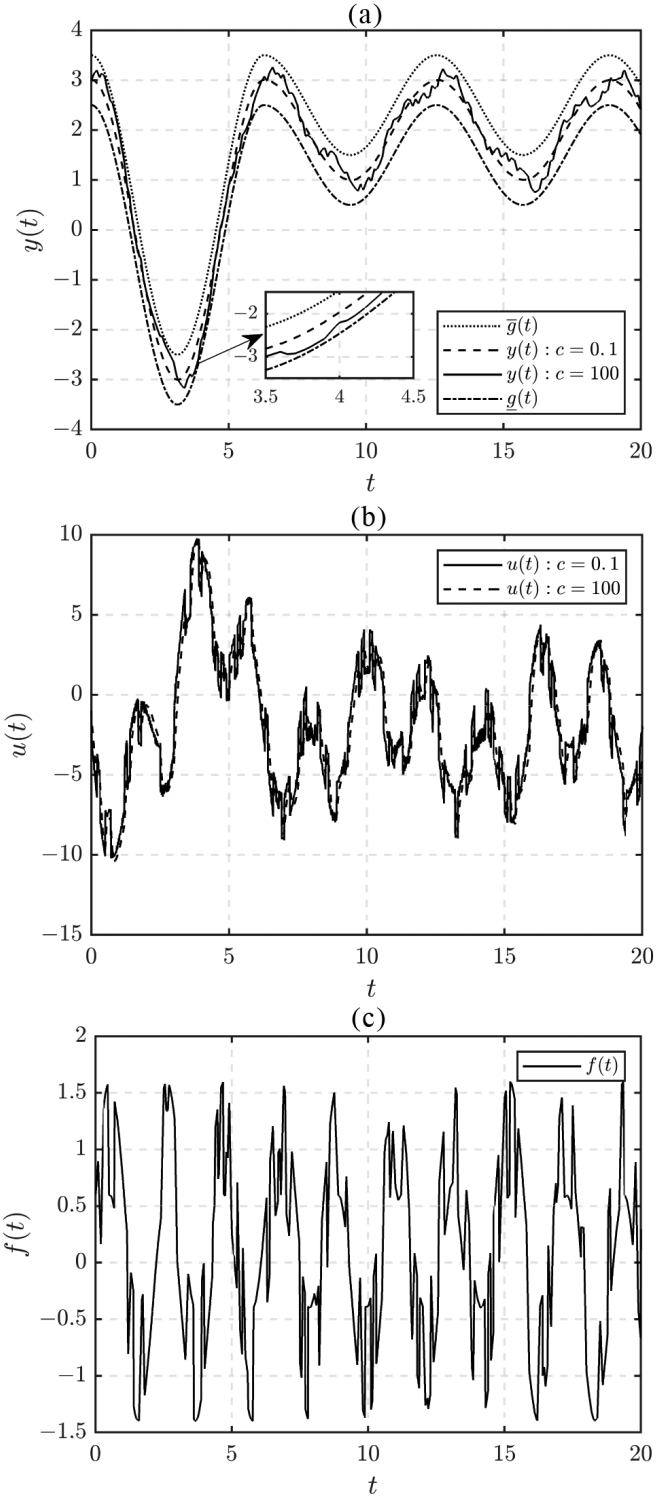


Fig. 2. Transients in the closed loop system for $c = 0.01$ and $c = 100$: (a) output $y(t)$, (b) control variable $u(t)$, and (c) disturbance $f(t)$.

The transients in $y(t)$, $u(t)$, and $f(t)$ for $x(0) = col\{1, 1\}$ are shown in Fig. 2. According to Fig. 2a, the output $y(t)$ never reaches the boundaries of the given set. Note also that the smaller the parameter c is, the better the effect of exogenous disturbances will be suppressed. The fluctuations of the control variable in Fig. 2b are explained by the disturbance $f(t)$ present in the system.

4.2. Example 2. Multidimensional System

We demonstrate control performance for an unstable double-input double-output plant with the following parameters:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.1 & 2 & -3 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 3 \end{bmatrix}, \quad G = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \\ 0,1 & 0.1 \end{bmatrix},$$

$$D = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad L = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix},$$

$$\phi(z) = col\{z_1 + \sin(z_1), \sin(z_2)\},$$

where $f(t)$ is the same as in Example 1. Then $\bar{f} = 1.6$ and $\mu = 2$.

Let $\Phi(y(t), t) = diag\{\Phi_1(y_1(t), t), \Phi_2(y_2(t), t)\}$, where Φ_i, $i = 1, 2$, are the same as in Example 1, i.e., $\Phi(y_i(t), t) = \ln\left(\frac{y_i(t) - \underline{g}_i(t)}{\overline{g}_i(t) - y_i(t)}\right)$. Consequently, $\Phi^{-1}(\varepsilon_i, t) = \frac{\overline{g}_i(t)e^{\varepsilon_i} + \underline{g}_i(t)}{e^{\varepsilon_i} + 1}$. For $\varepsilon \notin \Omega$, we have $0 \prec \frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \preceq \overline{\sigma} I$, where $\overline{\sigma} = \frac{1}{4} \max_i \left[\sup_{t \geqslant 0}(\overline{g}_i(t) - \underline{g}_i(t))\right]$, $i = 1, 2$.

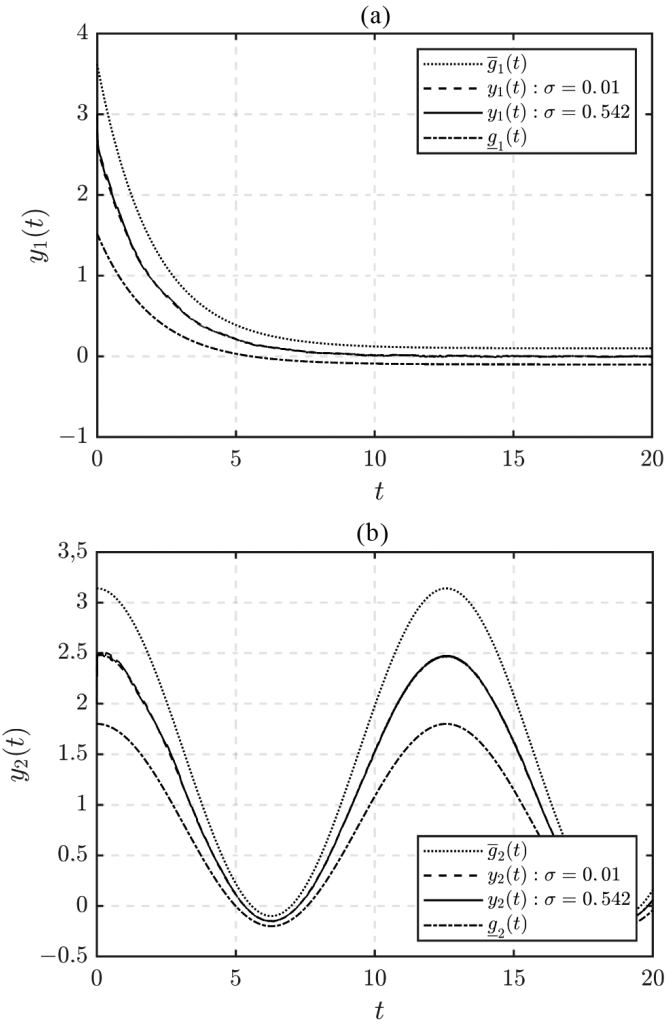


Fig. 3. Transients in the closed loop system for $c = 0.1$, $\sigma = 0.01$, and $\sigma = 0.542$: (a) $y_1(t)$ and (b) $y_2(t)$.

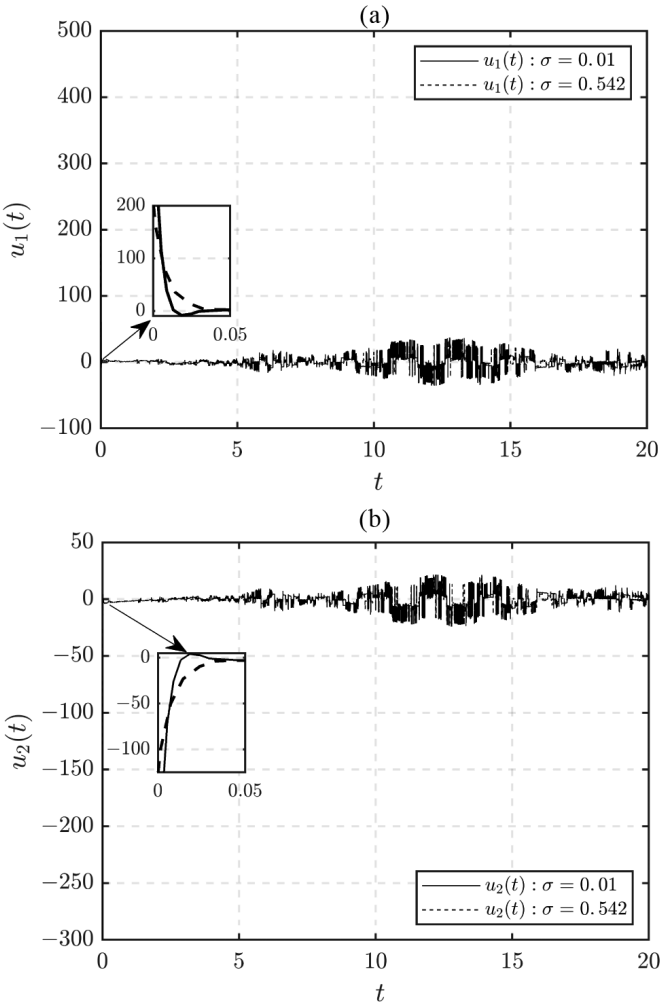


Fig. 4. Control variables in the closed loop system for $c = 0.1$, $\sigma = 0.01$, and $\sigma = 0.542$: (a) $u_1(t)$ and (b) $u_2(t)$.

Let the parameters of the constraint functions $\underline{g}(t)$ and $\overline{g}(t)$ be specified as

$$\overline{g}_1(t) = 3.52e^{-0.5t} + 0.1,$$
$$\underline{g}_1(t) = 1.62e^{-0.5t} - 0.1,$$
$$\overline{g}_2(t) = 1.62\cos(0.5t) + 1.52,$$
$$\underline{g}_2(t) = \cos(0.5t) + 0.8.$$

In this case,

$$\gamma = \sqrt{2}\max_i \left\{ \sup_{t\geqslant 0} |\dot{\overline{g}}_i(t)|,\ \sup_{t\geqslant 0} |\dot{\underline{g}}_i(t)| \right\} = 2.49, \quad i = 1, 2$$

and

$$\kappa = 11.54, \quad \overline{\sigma} = 0.542.$$

For $c = 0.1$, we solve inequality (15) under some values of $\sigma \in (0,\ 0.542]$:

$\tau_1 = 527.72$, $\tau_2 = 0.3$, and $K = diag\{108.42,\ 108.42\}$ (for $\sigma = 0.01$) and

$\tau_1 = 92.33$, $\tau_2 = 0.05$, and $K = diag\{39.26,\ 39.26\}$ (for $\sigma = 0.542$).

Figure 3 presents the transients of $y_1(t)$ and $y_2(t)$ for $x(0) = col\left\{\frac{5}{3}, \frac{2}{3}, -1\right\}$ whereas Fig. 4 the control variables $u_1(t)$ and $u_2(t)$. According to Fig. 3, the outputs always belong to the given pipes.

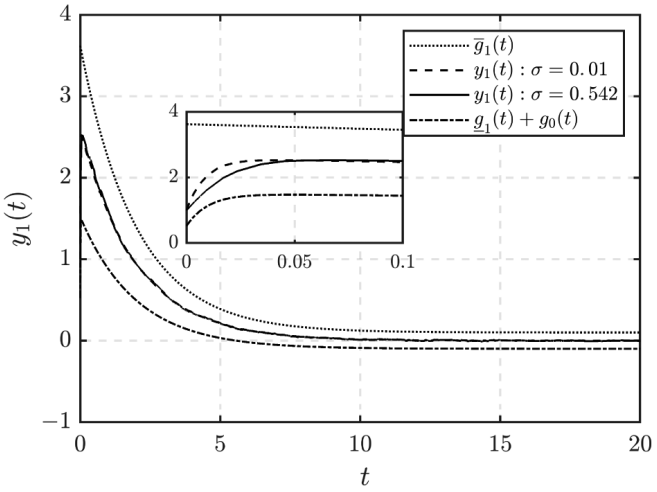


Fig. 5. Transients of $y_1(t)$ for $x(0) = col\{-\frac{1}{3}, \frac{2}{3}, 1\}$.

Remark 5. In the above examples, the initial values of the outputs are supposed to belong to a given set. However, if they are outside it, the control design method will fail: by the transformation (3), the outputs must be specified inside this set. This drawback can be eliminated by adding a fast exponentially decaying function to the limit functions, so the new limits will cover the initial conditions. Figure 5 shows the transients of $y_1(t)$ for $x(0) = col\left\{-\frac{1}{3}, \frac{2}{3}, 1\right\}$, i.e., $y_1(0) = 1$ falls beyond the initial set $\mathcal{Y}$. The function $g_0(t) = -e^{-100t}$ is added to the function $\underline{g}_1(t)$ so that the initial condition $y_1(0)$ is bounded from below by the new constraint function.

5. CONCLUSIONS

This paper has proposed a new method for stabilizing the output variables of nonlinear Lurie-type systems in given sets at any time instant. The method is based on a special output transformation and the technique of LMIs. With this transformation, the original problem with a constraint on the output variables is reduced to a problem without any constraints on the auxiliary variable. The control law for the new perturbing closed-loop system is designed using the Lyapunov function method in combination with the technique of LMIs. Simulation results in MATLAB/Simulink have illustrated the effectiveness of the method and confirmed the theoretical conclusions.

FUNDING

This work was performed in the Institute for Problems in Mechanical Engineering, the Russian Academy of Sciences, under the support of state order no. 121112500298-6 (The Unified State Information System for Recording Research, Development, Design, and Technological Work for Civilian Purposes).

APPENDIX

Proof of Theorem 2. Substituting (11) into (8) yields the closed loop system

$$\dot{\varepsilon} = \left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon}\right)^{-1} \Big[- K\varepsilon - \mu \operatorname{sgn}(\varepsilon) \|LG\| \|C\| |x| + LG\phi + \psi \Big]. \tag{A.1}$$

We choose a Lyapunov function of the form $V = \frac{1}{2}\varepsilon^2$. Its total time derivative along the solutions of (A.1) is given by

$$\dot V = \varepsilon\dot\varepsilon = \varepsilon\left(\frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon}\right)^{-1}\Big[-K\varepsilon-\mu\,\mathrm{sgn}(\varepsilon)\|LG\|\|C\||x|+LG\phi+\psi\Big]. \tag{A.2}$$

For $V \geqslant c$, we require the condition $\dot V \leqslant -2\alpha V$, where α is any known positive number, i.e., $\dot V < 0\ \forall\varepsilon \notin \Omega$. Due to $LG\phi \leqslant |LG\phi| \leqslant \mu\|LG\|\|C\||x|$ and the constraint $|\psi| \leqslant \kappa$, the above conditions can be written as

$$\begin{aligned}&(-K+\alpha)\varepsilon^2+\varepsilon\psi \leqslant 0\ \forall\,(\varepsilon,\psi):\\ &0.5\varepsilon^2 \geqslant c,\ \ \psi^2 \leqslant \kappa^2.\end{aligned} \tag{A.3}$$

Denoting $z = col\{\varepsilon,\psi\}$, we represent (A.3) in the matrix form

$$\begin{gathered}z^{\mathrm T}\begin{bmatrix}-K+\alpha & 0.5\\ \star & 0\end{bmatrix}z \leqslant 0,\\ z^{\mathrm T}\begin{bmatrix}-0.5 & 0\\ \star & 0\end{bmatrix}z \leqslant -c,\ z^{\mathrm T}\begin{bmatrix}0 & 0\\ \star & 1\end{bmatrix}z \leqslant \kappa^2.\end{gathered} \tag{A.4}$$

By the S-procedure [13], inequalities (A.4) hold under conditions (12). Hence, system (A.1) is input-to-state stable, and the variable $\varepsilon(t)$ is bounded. Owing to the transformation (4), the output $y(t)$ is also bounded, and the state vector $x(t)$ of system (1) possesses the same property accordingly. Therefore, the control variable $u(t)$ in (11) is bounded as well. Due to Theorem 1, the target condition (3) holds.

The proof of Theorem 2 is complete.

Proof of Proposition 1. Obviously, the matrix MN is symmetric. Let $\lambda_i, x_i,\ i = 1,\dots,2n$, be the eigenvalues and eigenvectors of the matrix MN, respectively. Then

$$x_i^{\mathrm T}NMNx_i = \lambda_i x_i^{\mathrm T}Nx_i.$$

Hence, the values λ_i can be expressed as

$$\lambda_i = \frac{x_i^{\mathrm T}NMNx_i}{x_i^{\mathrm T}Nx_i}.$$

Since $M \succ 0$ and $N = N^{\mathrm T} \prec 0$, we obtain $NMN \succ 0$, i.e., $x^{\mathrm T}NMNx > 0\ \forall x \neq 0$. In view of $x^{\mathrm T}Nx < 0\ \forall x \neq 0$, it follows that $\lambda_i < 0,\ i = 1,\dots,2n$. All eigenvalues of the symmetric matrix MN are negative, so the matrix MN is negative definite.

The proof of Proposition 1 is complete.

Proof of Theorem 3. We choose a Lyapunov function of the form $V = \frac{1}{2}\varepsilon^{\mathrm T}\varepsilon$. Its total time derivative along the solutions of (14) is given by

$$\begin{aligned}\dot V = \varepsilon^{\mathrm T}\dot\varepsilon = \varepsilon^{\mathrm T}&\left(\frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon}\right)^{-1}\Big[-K\varepsilon\\ &-\bar\sigma\mu\,\mathrm{Sign}(\varepsilon)\Big\|\left(\frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon}\right)^{-1}\Big\|\|LG\|\|C\||x|+LG\phi+\psi\Big].\end{aligned} \tag{A.5}$$

Formula (A.5) can be written as

$$\dot{V} = \dot{V}_1 + \dot{V}_2, \tag{A.6}$$

where

$$\dot{V}_1 = -\varepsilon^{\mathrm{T}} \left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right)^{-1} K\varepsilon + \varepsilon^{\mathrm{T}} \left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right)^{-1} \psi,$$

$$\dot{V}_2 = -\varepsilon^{\mathrm{T}} \left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right)^{-1} \mathrm{Sign}(\varepsilon) \bar{\sigma}\mu \left\| \left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right)^{-1} \right\| \|LG\| \|C\| |x|$$
$$+ \varepsilon^{\mathrm{T}} \left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right)^{-1} LG\phi.$$

Considering $\left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right) \leqslant \bar{\sigma} I$, we estimate $\dot{V}_2$ as

$$\dot{V}_2 \leqslant -\left(\sum_{i=1}^{v} |\varepsilon_i| \right) \bar{\sigma}^{-1} \bar{\sigma}\mu \left\| \left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right)^{-1} \right\| \|LG\| \|C\| |x|$$
$$+ \mu |\varepsilon| \left\| \left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right)^{-1} \right\| \|LG\| \|C\| |x| \leqslant 0.$$

Based on this inequality, the condition $\dot{V} \leqslant 0$ is equivalent to $\dot{V}_1 \leqslant 0$. In the case under study, $\left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right)^{-1}$ is a matrix and cannot be neglected when analyzing the sign definiteness of $\dot{V}$, in contrast to the previous section. For $V \geqslant c$, we require the condition $\dot{V} \leqslant -2\alpha V$, where α is any known positive number. Due to the constraints $|\psi| \leqslant \kappa$, the above conditions can be written as

$$\begin{gathered} -\varepsilon^{\mathrm{T}} \left[\left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right)^{-1} K + \alpha I \right] \varepsilon + \varepsilon^{\mathrm{T}} \left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right)^{-1} \psi \leqslant 0 \\ \forall\, (\varepsilon, \psi) : 0.5\varepsilon^{\mathrm{T}}\varepsilon \geqslant c, \psi^{\mathrm{T}}\psi \leqslant \kappa^2. \end{gathered} \tag{A.7}$$

Denoting $z = col\{\varepsilon, \psi\},\ z \in \mathbb{R}^{2m}$, we represent (A.7) in the matrix form

$$\begin{gathered} z^{\mathrm{T}} \begin{bmatrix} -\left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right)^{-1} K - \alpha I & 0.5 \left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right)^{-1} \\ \star & 0 \end{bmatrix} z \leqslant 0, \\ z^{\mathrm{T}} \begin{bmatrix} -0.5I & 0 \\ \star & 0 \end{bmatrix} z \leqslant -c, \quad z^{\mathrm{T}} \begin{bmatrix} 0 & 0 \\ \star & I \end{bmatrix} z \leqslant \kappa^2. \end{gathered} \tag{A.8}$$

By the S-procedure, inequalities (A.8) hold under the conditions

$$\begin{gathered} \begin{bmatrix} -\left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right)^{-1} K - \alpha I + 0.5\tau_1 I & 0.5 \left(\frac{\partial \Phi^{-1}(\varepsilon, t)}{\partial \varepsilon} \right)^{-1} \\ \star & -\tau_2 I \end{bmatrix} \prec 0, \\ -c\tau_1 + \kappa^2 \tau_2 \leqslant 0. \end{gathered} \tag{A.9}$$

The first inequality in (A.9) is equivalent to

$$\begin{bmatrix} \left(\dfrac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon}\right)^{-1} & 0 \\ \star & \left(\dfrac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon}\right)^{-1} \end{bmatrix} \times \begin{bmatrix} -K+(0.5\tau_1-\alpha)\dfrac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon} & 0.5I \\ \star & -\tau_2\dfrac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon} \end{bmatrix} \prec 0. \tag{A.10}$$

Since $\left(\frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon}\right)^{-1} \succ 0$, by Proposition 1, the latter inequality holds if, for any $\beta > 0$,

$$\begin{bmatrix} -K+(0.5\tau_1-\alpha)\dfrac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon} & 0.5I \\ \star & -\tau_2\dfrac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon} \end{bmatrix} \preceq -\beta I \prec 0. \tag{A.11}$$

Due to condition (A.7), it is required to ensure $\dot V < 0$ for all ε from the set $\{\varepsilon\in\mathbb{R}^m : |\varepsilon| \geqslant \sqrt{2c},\ c>0\}$. In addition, for all ε from this set, we have an interval uncertainty in (A.11) with $0 \prec \frac{\partial\Phi^{-1}(\varepsilon,t)}{\partial\varepsilon} \preceq \overline{\sigma} I$. Conditions (A.7) will be valid if the LMIs (15) are feasible for any $\sigma\in(0,\ \overline{\sigma}]$. Moreover, obviously, there always exist a matrix K and $\tau_1,\ \tau_2 > 0$ such that (15) are feasible. Indeed, using Schur's complement lemma [13], we write (15) as

$$\begin{gathered} -\tau_2\sigma+\beta<0, \\ -K+[(0.5\tau_1-\alpha)\sigma+\beta]I+\frac{1}{\tau_2\sigma-\beta}I\preceq 0, \\ -c\tau_1+\kappa^2\tau_2\leqslant 0, \\ \tau_1>0,\ \tau_2>0, \\ \alpha>0,\ \beta>0,\ 0<\sigma\leqslant\overline{\sigma}. \end{gathered} \tag{A.12}$$

For a given number $c > 0$ and fixed numbers σ, α, and β, inequalities (A.12) always have finite solutions (K, τ_1, τ_2). Thus, according to Theorem 1, the control law (13) with the gain matrix K satisfying (15) ensures the target condition (3).

The proof of Theorem 3 is complete.

REFERENCES

1. Grigor'ev, V.V., Zhuravleva, N.V., Luk'yanova, G.V., and Sergeev, K.A., *Sintez sistem metodom modal'nogo upravleniya* (Systems Design by the Modal Control Method), St. Petersburg: ITMO University, 2007.
2. Ioannou, P.A. and Sun, J., *Robust Adaptive Control*, Courier Corporation, 2012.
3. Narendra, K.S. and Annaswamy, A.M., *Stable Adaptive Systems*, Courier Corporation, 2012.
4. Furtat, I.B. and Gushchin, P.A., Control of Dynamical Plants with a Guarantee for the Controlled Signal to Stay in a Given Set, *Autom. Remote Control*, 2021, vol. 82, no. 4, pp. 654–669.
5. Furtat, I. and Gushchin, P., Nonlinear Feedback Control Providing Plant Output in Given Set, *Int. J. Control*, 2022, vol. 95, no. 6, pp. 1533–1542. https://doi.org/10.1080/00207179.2020.1861336

6. Nguyen, B.H., Furtat, I.B., and Nguyen, Q.C., Observer-Based Control of Linear Plants with the Guarantee for the Controlled Signal to Stay in a Given Set, *Diff. Eqs. Control Processes*, 2022, no. 4, pp. 95–104.

7. Furtat, I., Nekhoroshikh, A., and Gushchin, P., Synchronization of Multi-Machine Power Systems under Disturbances and Measurement Errors, *Int. J. Adaptive Control Signal Proc.*, 2022, vol. 36, no. 6, pp. 1272–1284. https://doi.org/10.1002/acs.3372

8. Pavlov, G.M. and Merkur'ev, G.V., Automation of Power Systems, St. Petersburg: Personnel Training Center of the Unified Energy Systems of Russia, 2001.

9. Verevkin, A.P. and Kiryushin, O.V., Control of a Reservoir Pressure Maintenance System Using Finite-State Machines, *Territoriya Neftegaz*, 2008, no. 10, pp. 14–19.

10. Miroshnik, I.V., Nikiforov, V.O., and Fradkov, A.L., *Nonlinear and Adaptive Control of Complex Systems*, Dordrecht–Boston–London: Kluwer Academic Publishers, 1999.

11. Isidori, A., *Nonlinear Control Systems*, Springer, 1995.

12. Khalil, H.K., *Nonlinear Systems*, 3rd ed., Pearson, 2001.

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

14. Sturm, J.F., Using SeDuMi 1.02, a MATLAB Toolbox for Optimization over Symmetric Cones, *Optim. Method. Softwar.*, 1999, vol. 11, no. 1, pp. 625–653. https://doi.org/10.1080/10556789908805766

15. Toh, K.C., Todd, M.J., and Tutuncu, R.H., SDPT3-a MATLAB Software Package for Semidefinite Programming, ver. 1.3, *Optim. Method. Softwar.*, 1999, vol. 11, pp. 545–581. https://doi.org/10.1080/10556789908805762

16. Borchers, B., A C Library for Semidefinite Programming, *Optim. Method. Softwar.*, 1999, vol. 11, pp. 613–623.

17. Lofberg, J., YALMIP: A Toolbox for Modeling and Optimization in MATLAB, *Proc. of the IEEE International Conference on Robotics and Automation*, 2004, pp. 284–289. IEEE Cat. No. 04CH37508. https://doi.org/10.1109/CACSD.2004.1393890

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 1, pp. 49–63.

STOCHASTIC SYSTEMS

Transient Behavior of a Two-Phase Queuing System with a Limitation on the Total Queue Size

V. M. Vishnevsky[*,a], **K. A. Vytovtov**[*,b], **and E. A. Barabanova**[*,c]

[*] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
e-mail: [a] *vishn@inbox.ru,* [b] *vytovtov_konstan@mail.ru,* [c] *elizavetaalexb@yandex.ru*

Received April 16, 2023
Revised November 15, 2023
Accepted December 21, 2023

Abstract—The transient mode of a two-phase queuing system with a Poisson input flow, exponential distribution of service time in each phase, and a limitation on the total buffer size of the two phases is considered. Nonstationary probabilities of system states are found using the Laplace transform. A numerical calculation and analysis of the system performance characteristics in transient mode with parameters corresponding to new-generation optical networks were carried out.

Keywords: two-phase queuing system, transient mode, Laplace transform

DOI: 10.31857/S0005117924010041

1. INTRODUCTION

Multiphase queuing systems (QSs), or so-called tandem networks, are widely used to describe the operation of telecommunication systems, in which the process of processing requests consists of several stages [1]. This class of systems includes, for example, multi-stage switching systems or a network of linear topology base stations. Moreover, the stationary operation mode of such systems is well studied both for the case of Poisson and correlated input flow. Let us note only some recent works on this topic [1–7].

In recent years, in addition to the study of the stationary mode of QSs, the study of the transient mode of their operation has continued. For example, an important problem in designing optical telecommunication networks with high information transfer rates is to study changes in the system performance characteristics over time and estimate the transition time in stationary mode after the system reboot process or a failure of service devices [8]. A similar situation arises when studying new-generation 5G and 6G networks [9]. Due to the relevance of this problem, in recent years the number of works devoted to the study of the transient operating mode of QSs and their non-stationary Markov models has increased [10–19]. One of the first works where such a problem for a two-phase QS with a Poisson input flow, an infinite buffer in the first phase, and a zero buffer in the second phase was considered is the paper of 1967 [20]. In further works, more complex systems are studied, such as systems with phase-type service time [11, 12] and various types of tandem networks [14, 15]. It should be noted that in most of these works, the authors do not provide final expressions that allow analyzing the performance characteristics of the QS in the transient mode, but use ready-made numerical methods of existing software packages. An analysis of the stability of non-stationary Markov processes with continuous time, describing the functioning of the main classes of QSs with non-stationary input flows, including those varying according to a sinusoidal law, was carried out in [10, 13, 18, 19]. In [21–23], the main performance characteristics of single-line and multi-line QSs with Poisson and correlated flows in transient mode are analyzed.

This work studies the non-stationary performance characteristics of a two-phase QS with a limitation on the total buffer size of the two phases. One example of a real system, the model of which is represented by this QS, is a car service station with two stages of service: diagnostics and repair. Cars queued for service at each stage are placed in a common parking lot with a certain number of parking spaces, which determines the limit on the total number of cars simultaneously located at the service station. The stationary mode of this QS was studied in [24]. There is no study of the non-stationary mode of such a system in the world literature, which determines the novelty of this article.

The structure of the article is as follows. Section 3 presents differential equations that describe the functioning of a two-phase QS, for the convenience of writing which new functions are introduced. Section 4 presents an expression for finding the probabilities of states of a two-phase QS, containing an auxiliary matrix, the elements of which are found using the Laplace transform apparatus. Section 5 provides expressions for finding the main performance indicators of a two-phase QS in transient mode. The numerical results of the study are presented in Section 6.

2. STATEMENT OF THE PROBLEM

A two-phase QS with one single-line service device on each phase is considered. The input flow is Poisson with intensity λ, and the time for servicing requests by devices of the first and second phases has an exponential distribution with intensities μ_1 and μ_2, respectively.

After completing the servicing of a request in the first phase, each request moves to the second phase. The number of requests in the first and second phases can take the values $n_1 = \overline{0, N}$, $n_2 = \overline{0, N}$, respectively, where N is the maximum number of requests in the system. In this case, a limitation is imposed on the total buffer size of the two phases of the system, such that $n_1 + n_2 \leqslant N$ at any time. A new request can enter the system only under the condition $n_1 + n_2 < N$ (Fig. 1).

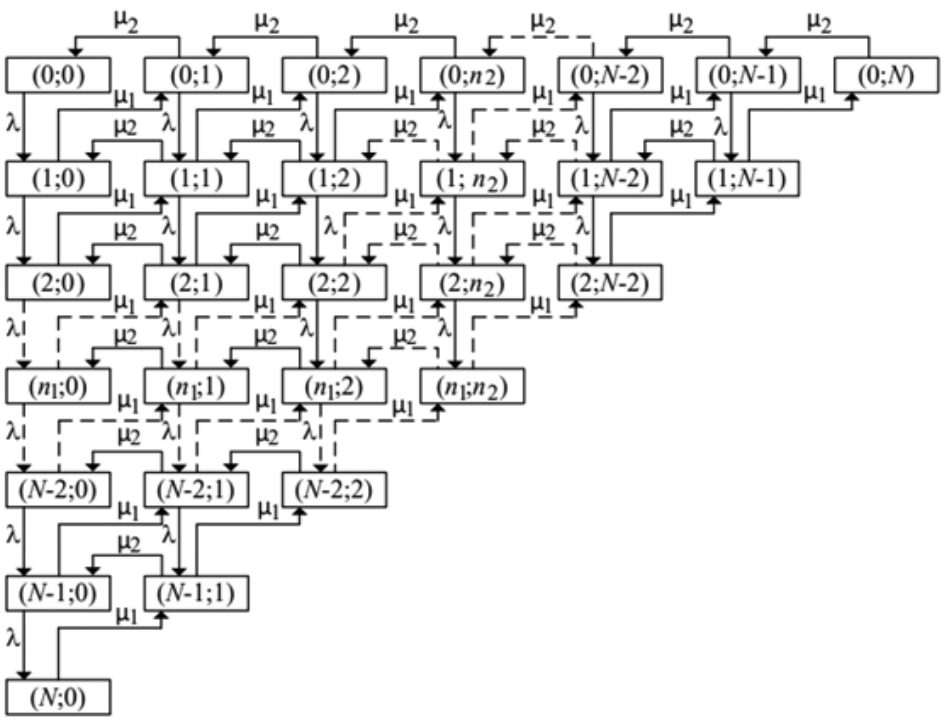


Fig. 1. System state graph.

The purpose of this work is to analyze the performance characteristics of the system described above in transient mode, such as transition time, loss probability, throughput, and the average number of requests in the system.

3. CONSTRUCTION OF DIFFERENTIAL EQUATIONS DESCRIBING THE FUNCTIONING OF A TWO-PHASE QS WITH A LIMITATION ON THE TOTAL BUFFER SIZE

The Markov process describing the operation of the considered QS consists of $R=\frac{1}{2}\left(N^2+3N+2\right)$ states of the system $S\left(n_1,n_2,t\right)$, when n_1 requests are served in the first phase, and n_2 requests are served in the second phase at time t, where $n_1+n_2\leqslant N$ (Fig. 1). The system of differential equations for such a QS has the form:

$$\begin{aligned}
\frac{dP\left(0,0,t\right)}{dt} &= -\lambda P\left(0,0,t\right)+\mu_2 P(0,1,t),(n_1,n_2=0);\\
\frac{dP\left(0,n_2,t\right)}{dt} &= -(\lambda+\mu_2)P\left(0,n_2,t\right)+\mu_2 P(0,n_2+1,t)+\mu_1 P(0,n_2-1,t),\\
&\qquad (n_1=0,n_2=\overline{1,N-1});\\
\frac{dP\left(0,N,t\right)}{dt} &= -\mu_2 P\left(0,N,t\right)+\mu_1 P(1,N-1,t),(n_1=0,n_2=N);\\
\frac{dP\left(n_1,0,t\right)}{dt} &= -\left(\lambda+\mu_2\right)P\left(n_1,0,t\right)+\mu_2 P(n_1,1,t)+\lambda P\left(n_1-1,0,t\right),\\
&\qquad (n_1=\overline{1,N-1},n_2=0);\\
\frac{dP\left(N,0,t\right)}{dt} &= -\mu_1 P\left(N,0,t\right)+\lambda P(N-1,0,t),(n_1=N,n_2=0);\\
\frac{dP\left(n_1,n_2,t\right)}{dt} &= -\left(\lambda+\mu_1+\mu_2\right)P\left(n_1,n_2,t\right)+\mu_2 P(n_1,n_2+1,t)\\
&+\mu_1 P(n_1+1,n_2-1,t)+\lambda P(n_1-1,n_2,t),(n_1,n_2>0,n_1+n_2<N);\\
\frac{dP\left(n_1,n_2,t\right)}{dt} &= -\left(\mu_1+\mu_2\right)P\left(n_1,n_2,t\right)+\mu_1 P(n_1+1,n_2-1,t)\\
&\qquad +\lambda P(n_1-1,n_2,t),(n_1,n_2>0,n_1+n_2=N).
\end{aligned} \tag{1}$$

It should be noted that the well-known approach to constructing a system of differential equations, which involves the use of various forms of writing the equations for different permissible values of n_1 and n_2, is very inconvenient for calculating and analyzing the characteristics of the QS in the transient mode. For the convenience of further solution and analysis of system (1), we introduce the functions

$$\upsilon_1\left(x,M\right)=\frac{|x-M+0.5|+x-M+0.5}{2\,|x-M+0.5|}, \tag{2}$$

$$\upsilon_2\left(x,K\right)=\frac{|K-x-0.5|+K-x-0.5}{2\,|K-x-0.5|}, \tag{3}$$

where $M=\overline{0,N}$, $K=\overline{0,N}$. Then system (1) can be written in the form

$$\begin{aligned}
\frac{dP\left(n_1,n_2,t\right)}{dt} &= -\left[\lambda\upsilon_2\left(n_1+n_2,N-1\right)+\mu_1\upsilon_1\left(n_1,1\right)+\mu_2\upsilon_1\left(n_2,1\right)\right]P\left(n_1,n_2,t\right)\\
&+\mu_1\upsilon_1\left(n_2,1\right)\upsilon_2\left(n_1+n_2,N\right)P\left(n_1+1,n_2-1,t\right)+\mu_2\upsilon_2\left(n_1+n_2,N-1\right)\\
&\times P\left(n_1,n_2+1,t\right)+\lambda\upsilon_1\left(n_1,1\right)\upsilon_2\left(n_1+n_2,N\right)P\left(n_1-1,n_2,t\right),
\end{aligned} \tag{4}$$

where $n_1 = \overline{0,N}$, $n_2 = \overline{0,N}$, $n_1 + n_2 \leqslant N$. The described system can be represented in matrix form:

$$\frac{d\vec{P}\,(t)}{dt} = \mathbf{A}\vec{P}\,(t)\,, \tag{5}$$

where $\mathbf{A}$ is the matrix of coefficients of the system of differential Eqs. (4), $\vec{P}(t) = \{P(n_1, n_2, t)\}^{\mathrm{T}}$ is the column vector of system state probabilities. To construct the matrix $\mathbf{A}$, we additionally introduce the function

$$\vartheta\,(n_k, n_l) = (N+1)\,n_k + n_l - \frac{n_k\,(n_k-1)}{2} + 1, \tag{6}$$

transforming the number of requests n_k, n_l in the first and second buffer, respectively, into the number of a column or row of this matrix. A brief description of functions (2), (3) and (6) is given in the Appendix. Then the elements of the matrix of system (5), located on the main diagonal, are written in the form

$$A_{\vartheta(n_1,n_2),\vartheta(n_1,n_2)} = -\left[\lambda\upsilon_2\,(n_1+n_2,N) + \mu_1\upsilon_1\,(n_1,1) + \mu_2\upsilon_1\,(n_2,1)\right]. \tag{7}$$

The remaining non-zero elements are determined by the relations

$$\begin{gathered} A_{\vartheta(n_1,n_2),\vartheta(n_3,n_4)} = \mu_1\upsilon_1\,(n_2,1)\,\upsilon_2\,(n_1+n_2,N+1)\,; \\ A_{\vartheta(n_1,n_2),\vartheta(n_1,n_5)} = \mu_2\upsilon_2\,(n_1+n_2,N)\,; \\ A_{\vartheta(n_1,n_2),\vartheta(n_6,n_2)} = \lambda\upsilon_1\,(n_2,1)\,\upsilon_2\,(n_1+n_2,N+1)\,. \end{gathered} \tag{8}$$

Here $n_1 = \overline{0,N}$, $n_2 = \overline{0,N}$, $n_3 = n_1 + 1$, $n_4 = n_2 - 1$, $n_5 = n_2 + 1$, $n_6 = n_1 - 1$. The remaining elements $A_{i,j}$ of the matrix $\mathbf{A}$ in (5) are equal to zero. The new function (6) is also necessary for the ordered construction of a column vector of system state probabilities at time t in equation (5). Indeed, in terms of n_k and n_l it has the form

$$\begin{gathered} \vec{P}\,(t) = \{p\,(0,0,t)\,,\ldots,p\,(0,N,t)\,,p\,(1,0,t)\,,\ldots, \\ p\,(1,N-1,t)\,,\ldots\ldots,p\,(N-1,0,t)\,,p\,(N-1,1,t)\,,p(N,0,t)\}^{\mathrm{T}}\,, \end{gathered} \tag{9}$$

where T is the transposition operator. However, to correctly solve (5), it is necessary to use not a two-dimensional array of numbers n_k and n_l when indicating the state of the system, but a sequence number from 1 to $R = \frac{1}{2}\,(N^2+3N+2)$. To do this, using function (6), we finally obtain

$$\vec{P}\,(t) = \{P\,(1,t)\,,P\,(2,t)\,,P\,(3,t)\,,P\,(4,t)\,,\ldots,P\,(\vartheta\,(n_k,n_l)\,,t)\,,\ldots,P(\vartheta\,(N,0)\,,t)\}^{\mathrm{T}}\,, \tag{10}$$

where $P\,(\vartheta\,(n_k,n_l)\,,t)$ corresponds to $p\,(n_k,n_l,t)$ in (9).

Thus, using the functions $\vartheta\,(n_k,n_l)$, $\upsilon_1\,(x,M)$, $\upsilon_2\,(x,K)$ makes it possible to construct a matrix of coefficients in (5) in general form for any number of requests N.

4. STATE PROBABILITIES OF A TWO-PHASE QS IN A TRANSIENT MODE

To connect the system state probabilities at time t with the probabilities of system states at some initial time t_0, we introduce the matrix $\mathbf{L}$, the order of which is one greater than the order of the fundamental matrix of the system of equations (4) and such that

$$\vec{P}\,(t) = \mathbf{L}\,(t-t_0)\,\vec{P}\,(t_0)\,, \tag{11}$$

where $\vec{P}\,(t) = \{P\,(\vartheta\,(n_1,n_2)\,,t)\}^{\mathrm{T}}$ is the vector column of system state probabilities at time t.

Let us apply the direct Laplace transform to the system of equations (5):

$$\int_0^\infty e^{-st}\frac{d\vec{P}\,(t)}{dt}dt = \int_0^\infty e^{-st}\mathbf{A}\vec{P}\,(t)dt. \tag{12}$$

Then the elements of the matrix $\mathbf{L}\,(t-t_0)$ are determined by the following theorem.

Theorem 1. *The elements of the matrix $\mathbf{L}\,(t-t_0)$ of a two-phase QS with a limitation on the total buffer size of two phases N, described by the system of equations (4), have the form*

$$L_{l,j}\,(t-t_0) = \sum_{k=1}^{R}(-1)^{l+j}\,\frac{\Delta_{j,l}(s_k)}{\left.\frac{d\Delta(s)}{ds}\right|_{s=s_k}}\exp\left(s_k\,(t-t_0)\right), \tag{13}$$

where $\Delta(s)$ is the determinant of the matrix $\mathbf{B}=\mathbf{A}-s\mathbf{I}$, $\mathbf{A}$ is the coefficient matrix in (5), $\mathbf{I}$ is the unit diagonal matrix, $s=\alpha+i\beta$ is the independent variable in the complex domain, $i=\sqrt{-1}$, $\Delta_{li}(s)$ is the determinant of the minor element B_{li} of the matrix $\mathbf{B}$, s_k is the kth root of the polynomial $\Delta\,(s)$ in the case when all its roots are simple, $R=(N^2+3N+2)/2$ is the number of roots of the polynomial $\Delta\,(s)$, equal to the number of differential equations in system (4).

Proof. Considering that $\int_0^\infty e^{-st}\left(\frac{d\vec{P}(t)}{dt}\right)dt = s\vec{P}\,(s)-\vec{P}\,(0)$, where $\vec{P}\,(0)$ is the column vector of initial conditions, and also that in this case $\mathbf{A}$ is a constant matrix, let us carry out the transformations

$$s\vec{P}\,(s)-\vec{P}\,(0)=\mathbf{A}\vec{P}\,(s)\;\Rightarrow\;\mathbf{A}\vec{P}\,(s)-s\vec{P}\,(s)=-\vec{P}\,(0)\;\Rightarrow\;(\mathbf{A}-s\mathbf{I})\,\vec{P}\,(s)=-\vec{P}\,(0)\,. \tag{14}$$

As a result, we obtain a system of linear inhomogeneous algebraic equations

$$\mathbf{B}\vec{P}\,(s)=-\vec{P}\,(0) \tag{15}$$

with constant coefficients. Taking into account (4), system (15) can be written in the form

$$\begin{aligned}-\left[\lambda\upsilon_2\,(n_1+n_2,N-1)+\mu_1\upsilon_1\,(n_1,1)+\mu_2\upsilon_1\,(n_2,1)+s\right]P\,(n_1,n_2,s)\\ +\mu_1\upsilon_1\,(n_2,1)\,\upsilon_2\,(n_1+n_2,N)\,P\,(n_1+1,n_2-1,s)+\mu_2\upsilon_2\,(n_1+n_2,N-1)\,P\,(n_1,n_2+1,s)\\ +\lambda\upsilon_1\,(n_1,1)\,\upsilon_2\,(n_1+n_2,N)\,P\,(n_1-1,n_2,s)=P\,(n_1,n_2,0)\,,\end{aligned} \tag{16}$$

where $n_1=\overline{0,N}$, $n_2=\overline{0,N}$, $n_1+n_2\leqslant N$. Then, in accordance with (7) and (8), the non-zero elements of the matrix $\mathbf{B}$ are written as

$$\begin{aligned}B_{\vartheta(n_1,n_2),\vartheta(n_1,n_2)}(s)&=-\left[\lambda\upsilon_2\,(n_1+n_2,N)+\mu_1\upsilon_1\,(n_1,1)+\mu_2\upsilon_1\,(n_2,1)+s\right];\\ B_{\vartheta(n_1,n_2),\vartheta(n_3,n_4)}(s)&=\mu_1\upsilon_1\,(n_2,1)\,\upsilon_2\,(n_1+n_2,N+1)\,;\\ B_{\vartheta(n_1,n_2),\vartheta(n_1,n_5)}&=\mu_2\upsilon_2\,(n_1+n_2,N)\,;\\ B_{\vartheta(n_1,n_2),\vartheta(n_6,n_2)}(s)&=\lambda\upsilon_1\,(n_2,1)\,\upsilon_2\,(n_1+n_2,N+1)\,.\end{aligned} \tag{17}$$

To find images of elements of the matrix $\mathbf{L}$ it is necessary to use linearly independent initial conditions. These conditions are:

$$P\,(N_1,N_2,0)=1(N_1=\overline{0,N},N_2=\overline{0,N},N_1+N_2\leqslant N);P\,(n_1,n_2,0)=0(n_1\neq N_1,n_2\neq N_2). \tag{18}$$

Solutions to system (16) for $P\,(n_1,n_2,0)=1$ $(n_1=n_2=0)$, $P\,(n_1,n_2,0)=0$ $(n_1=\overline{1,N}$, $n_2=\overline{1,N}$, $n_1+n_2\leqslant N)$ give the first column of images of the elements of the transformation matrix, the

solution to system (15) for $P(0,1,0)=1$ and the rest $P(n_1,n_2,0)=0$ give the second column of images of the elements of the transformation matrix. Similarly, all columns of images of elements of the transformation matrix $\mathbf{L}(s-s_0)$ are found. To obtain an image of the matrix $\mathbf{L}$, it is advisable to use the Cramer method. In accordance with this method, the elements of the matrix, which are linearly independent solutions (16), are fractions of the form

$$L_{l,j}(s-s_0)=(-1)^{l+j}\frac{\Delta_{j,l}(s)}{\Delta(s)}, \tag{19}$$

where $\Delta(s)$ is the determinant of the matrix $\mathbf{B}$, $\Delta_{jl}(s)$ is the minor of the element B_{jl} of the matrix $\mathbf{B}$. Now consider the inverse Laplace transform. First of all, we note that the image of the element of the probability transformation matrix (19) is a proper fraction

$$L_{l,j}(s-s_0)=(-1)^{l+j}\frac{\Delta_{j,l}(s)}{\Delta(s)}=(-1)^{l+j}\frac{a_ns^n+a_{n-1}s^{n-1}+\ldots+a_2s^2+a_1s+a_0}{b_ms^m+b_{m-1}s^{m-1}+\ldots+b_2s^2+b_1s+b_0}. \tag{20}$$

Moreover, $n<m$, since the numerator is the determinant of the algebraic complement of the matrix element whose determinant is in the denominator. Then the fraction in (20) can be factorized

$$L(s)=\frac{\Delta_{j,l}(s)}{\Delta(s)}=A_1\frac{1}{s-s_1}+A_2\frac{1}{s-s_2}+\ldots+A_m\frac{1}{s-s_m}=\sum_{k=1}^{m}A_k\frac{1}{s-s_k}. \tag{21}$$

To find the coefficients A_k, multiply both sides of (21) by $(s-s_1)$ and get

$$L(s)=\frac{\Delta_{j,l}(s)}{\Delta(s)}(s-s_1)=A_1+(s-s_1)\sum_{k=2}^{m}A_k\frac{1}{s-s_k}. \tag{22}$$

The right-hand side of (22) at $s\to s_1$ is equal to A_1, since $s-s_1\to 0$. The left side represents the uncertainty $0/0$, since the factor $s-s_1$ is present in both the numerator and the denominator. Let us reveal this uncertainty using L'Hopital's rule and obtain the left-hand side in the form

$$\lim_{x\to x_1}\frac{\Delta_{j,l}(s)}{\Delta(s)}(s-s_1)=\lim_{x\to x_1}\frac{\Delta_{j,l}(s)+(s-s_1)\frac{d\Delta_{j,l}(s)}{ds}}{\frac{d\Delta(s)}{ds}}=\frac{\Delta_{j,l}(s_1)}{\left[\frac{d\Delta(s)}{ds}\right]|_{s=s_1}}. \tag{23}$$

Taking into account (22) and (23), we obtain

$$A_1=\frac{\Delta_{j,l}(s_1)}{\left[\frac{d\Delta(s)}{ds}\right]|_{s=s_1}}. \tag{24}$$

Similarly, we find the kth coefficient in (22) as

$$A_k=\frac{\Delta_{j,l}(s_k)}{\left[\frac{d\Delta(s)}{ds}\right]|_{s=s_k}}. \tag{25}$$

Thus, expression (21) takes the form

$$L(s)=\sum_{k=1}^{m}\frac{\Delta_{j,l}(s_k)}{\left[\frac{d\Delta(s)}{ds}\right]|_{s=s_k}}\cdot\frac{1}{s-s_k}. \tag{26}$$

Applying the inverse Laplace transform to (26) and carrying out mathematical transformations

$$L(t) = \frac{1}{2\pi i}\int\limits_{\sigma-i\infty}^{\sigma+i\infty}\sum_{k=1}^{m}\frac{\Delta_{j,l}(s_k)}{\left[\frac{d\Delta(s)}{ds}\right]|_{s=s_k}}\frac{exp(st)ds}{s-s_k} = \frac{1}{2\pi i}\sum_{k=1}^{m}\frac{\Delta_{j,l}(s_k)}{\left[\frac{d\Delta(s)}{ds}\right]|_{s=s_k}}\int\limits_{\sigma-i\infty}^{\sigma+i\infty}\frac{exp(st)ds}{s-s_k}$$

$$= \sum_{k=2}^{m}\frac{\Delta_{j,l}(s_k)}{\left[\frac{d\Delta(s)}{ds}\right]|_{s=s_k}}exp(s_k t), \quad (27)$$

we obtain an expression for the original from image (26) in the form (12). The theorem has been proven.

By substituting (27) and expression (11), we can find the probabilities of states of a two-phase QS in the transition mode under given initial conditions. These expressions make it possible to calculate and analyze the performance indicators of the system under consideration at an arbitrary moment of time t in both transient and stationary modes: the time before the system enters stationary mode, the probability of losses, throughput, and the number of requests served in each phase.

5. PERFORMANCE INDICATORS OF A TWO-PHASE QS IN TRANSIENT MODE

5.1. Transition Time

The transition time is the time during which the QS goes into stationary mode. In accordance with [22], the time of the transition mode is determined by the smallest absolute value of the real part of the pole of the state probability images:

$$\tau_{tr} = \frac{k}{\alpha_{\min}}. \quad (28)$$

Here $\forall\alpha_j \in \Gamma$: $(\Gamma = \alpha_j, \alpha_j \geqslant \alpha_{\min} \Longrightarrow \alpha_j = \alpha_{\min})$ and $k > 0$, $k \in R$. The value of k is selected based on the formulation of a specific problem. It was shown in [22] that the transition mode can be considered completed when $k = (3 \div 5)$.

5.2. Probability of Losses

Since the maximum number of requests in the system is $n_1 + n_2 = N$, all requests in the states $(i, N-i)$, $i = \overline{0, N}$ will be lost. Considering that the presence of requests in the specified states are independent events, the sum of the probabilities of these states at time t

$$P_{loss}(t) = \sum_{i=0}^{N} P(i, N-i, t) = \sum_{i=0}^{N} P(\vartheta(i, N-i), t) \quad (29)$$

determines the probability of loss of requests.

5.3. Throughput

Since expression (29) determines the resulting probability of requests being lost in the system, it is obvious that requests entering the system in any other states will be serviced. Then the throughput at time t in the transition mode is equal to

$$A(t) = [1 - P_{loss}(t)]\lambda = \left[1 - \sum_{i=0}^{N} P(i, N-i, t)\right]\lambda = \left[1 - \sum_{i=0}^{N} P(\vartheta(i, N-i), t)\right]\lambda. \quad (30)$$

Since the throughput actually represents the intensity of the system servicing the requests received by it, then in the time dt the system services $A(t)\,dt$ requests. Consequently, during the transition mode the number of requests served is equal to

$$Z_{service\ tr} = \int_{t_0}^{t_0+\tau_{tr}} \lambda \left[1 - \sum_{i=0}^{N} P\left(\vartheta\left(i, N-i\right), t\right)\right] dt, \tag{31}$$

and the number of lost requests is

$$Z_{loss\ tr} = \int_{t_0}^{t_0+\tau_{tr}} \lambda \sum_{i=0}^{N} P\left(\vartheta\left(i, N-i\right), t\right) dt. \tag{32}$$

Thus, the sum of (31) and (32) gives the number of requests received during the transition mode $\lambda\tau_{tr}$, which confirms the correctness of the obtained relationships.

5.4. Number of Requests Served at Each Phase in Transition Mode

Let $P(n_1, n_2, t)$ be the probability of finding n_1 requests in the first phase and n_2 requests in the second phase at time t, then the number of requests in the first phase, provided that the system is in the state (n_1, n_2), is equal to $n_1 P(n_1, n_2, t)$. Summing $n_1 P(n_1, n_2, t)$ over all possible states, we obtain the average number of requests in the first phase at time t as

$$Z_{phase1}(t) = \sum_{n_1=0}^{N} \sum_{n_2=0}^{N} \left[n_1 P(n_1, n_2, t)\right] = \sum_{n_1=0}^{N} \sum_{n_2=0}^{N} \left[n_1 P\left(\vartheta\left(n_1, n_2\right), t\right)\right], \tag{33}$$

where $n_1 + n_2 \leqslant N$. Similarly, the average number of requests in the second phase at time t in the transition mode is equal to

$$Z_{phase2}(t) = \sum_{n_1=0}^{N} \sum_{n_2=0}^{N} \left[n_2 P(n_1, n_2, t)\right] = \sum_{n_1=0}^{N} \sum_{n_2=0}^{N} \left[n_2 P\left(\vartheta\left(n_1, n_2\right), t\right)\right], \tag{34}$$

where $n_1 + n_2 \leqslant N$. Then the average number of requests in the system in the transition mode will be

$$Z(t) = \sum_{n_1=0}^{N} \sum_{n_2=0}^{N} \left[\left(n_1 + n_2\right) P(n_1, n_2, t)\right] = \sum_{n_1=0}^{N} \sum_{n_2=0}^{N} \left[\left(n_1 + n_2\right) P\left(\vartheta\left(n_1, n_2\right), t\right)\right], \tag{35}$$

where $n_1 + n_2 \leqslant N$.

6. NUMERICAL STUDY OF A TWO-PHASE QS TRANSIENT MODE

Let us consider the transient mode of a two-phase QS operation, which adequately describes the operation of a switch in an all-optical network. In the presented numerical experiment, the values $\lambda = 8 \cdot 10^6$ packets/s, $\mu_1 = 15 \cdot 10^6$ packets/s, $\mu_2 = 10 \cdot 10^6$ packets/s ($\lambda < \mu_2 < \mu_1$) correspond to the actual characteristics of modern optical networks [25]. Here n_1 is the number of packets in the first servicing phase, n_2 is the number of packets in the second servicing phase, $N = n_1 + n_2 = 4$ is the maximum number of packets in the system. The small buffer size in this numerical example is determined by the technical limitations of modern optical devices.

To analyze the performance characteristics of the considered QS, first of all, the matrix $\mathbf{A}$ is constructed in accordance with (7) and (8), and then the matrix $\mathbf{B}$ is constructed in accordance

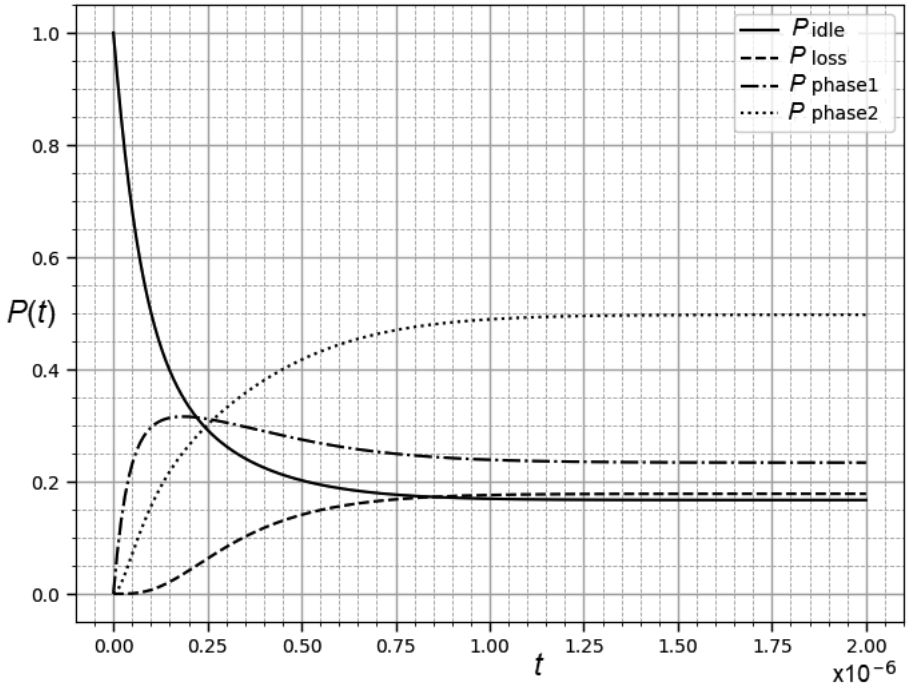


Fig. 2. Dependence of system state probabilities on time in transition mode.

with (17). Next, the elements of the matrix $\mathbf{L(t)}$ are written in accordance with (13). To do this, we find the poles of functions that describe the elements of the matrix $\mathbf{L(s - s_0)}$ in terms of the Laplace transform: $s_0 = 0$, $s_1 = -1.2 \cdot 10^7$, $s_2 = -2.8 \cdot 10^7$, $s_{3,4} = -3.9 \cdot 10^7 \pm i1.9 \cdot 10^7$, $s_{5,6} = -3.2 \cdot 10^7 \pm i9.9 \cdot 10^7$, $s_{7,8} = -2.8 \cdot 10^7 \pm i1.2 \cdot 10^7$, $s_{9,10} = -5.1 \cdot 10^7 \pm i1.4 \cdot 10^7$, $s_{11,12} = -1.5 \cdot 10^7 \pm i6.0 \cdot 10^7$, $s_{13,14} = -2.2 \cdot 10^7 \pm i4.0 \cdot 10^7$. One of these poles is zero, all others have a negative real part. This indicates the presence of a stationary mode in the system. Moreover, 12 of the 15 poles are pairwise complex conjugate, which indicates the oscillatory nature of the probabilities of states in the transition mode. Indeed, the exponent of a complex number in (12) is a combination of trigonometric functions in accordance with Euler's formula.

Studying the poles of state probability images also allows one to calculate the time constant using formula (28) $\tau = 1/|\alpha_{\min}| = 1/5\,138\,202.473908113 = 1.9462 \cdot 10^{-7}$ s and transition time $\tau_{tr} = 5\tau = 9,731 \cdot 10^{-7}$ s.

The dependence of state probabilities on time for the case under consideration is presented in Fig. 2. The figure shows: $P_{idle}(t)$ is the probability that the system is free; $P_{phase1}(t)$, $P_{phase2}(t)$ are dependencies of the probabilities of the states of finding requests only in the first and only in the second phases of service, respectively; $P_{loss}(t)$ is the probability of losses calculated in accordance with (29).

From Fig. 2 it can be seen that the time of the transition mode, calculated from (28), corresponds to the time of reaching the stationary mode according to the state probability graphs. The oscillatory nature of the transition mode is clearly visible from the dependence of the probability of finding requests in the first phase of service $P_{phase1}(t)$ (Fig. 3). Note that the probabilities of states in a stationary mode, obtained by the authors using the proposed approach, are equal to the stationary probabilities calculated using a well-known technique [24]. Indeed, from Fig. 2, it is clear that $\pi_{idle} = 0.167$, $\pi_{loss} = 0.172$, $\pi_{phase1} = 0.24$, $\pi_{phase2} = 0.49$, which corresponds to the stationary probabilities calculated using formulas (6) and (7) presented in [24].

Next, the performance indicators of the considered QS are calculated in accordance with Section 5 of this work.

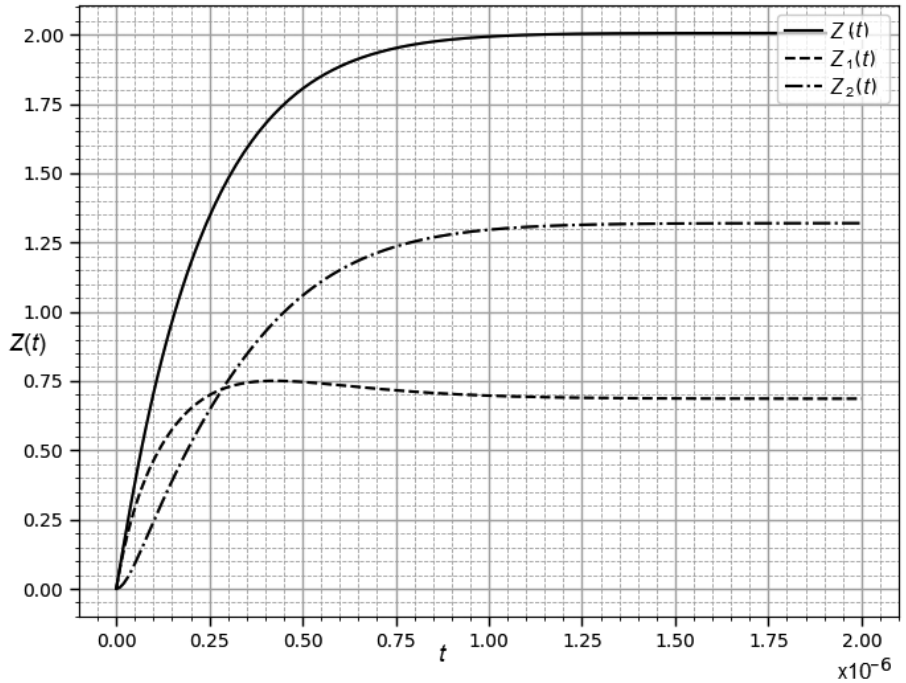


Fig. 3. Dependence of the average number of requests in each phase on time.

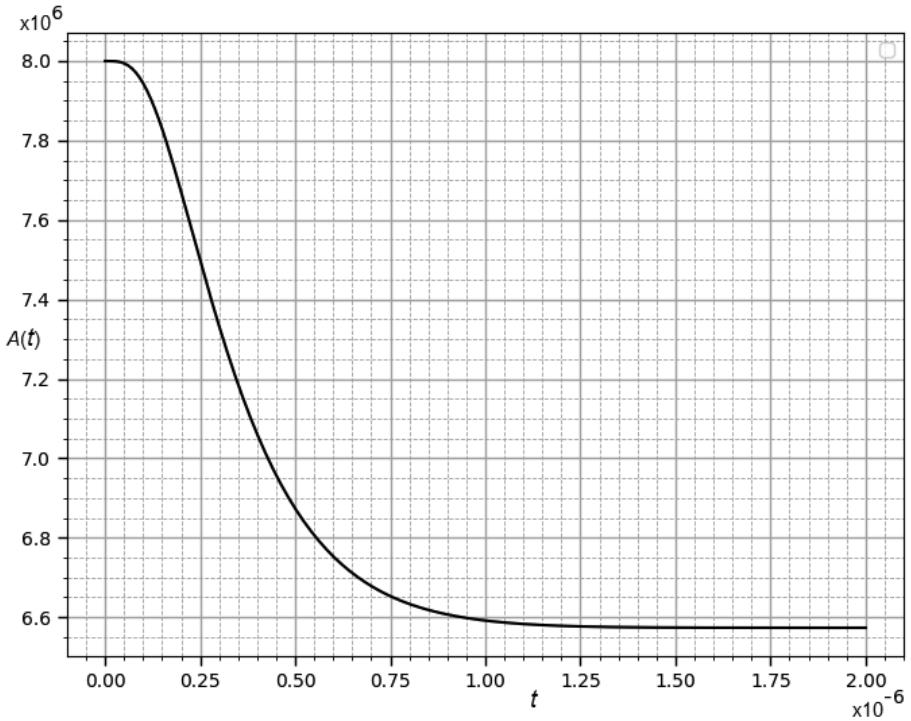


Fig. 4. Dependence of system throughput on time in transition mode.

Figure 4 shows the dependence of the system throughput on time in the transient mode, calculated in accordance with (30). The system throughput at the initial time is equal to $8 \cdot 10^6$ packets/s and decreases to a stationary value of $6.62 \cdot 10^6$ packets/s. Studying changes in the throughput of an all-optical switch in transient mode makes it possible to obtain more accurate estimates of its

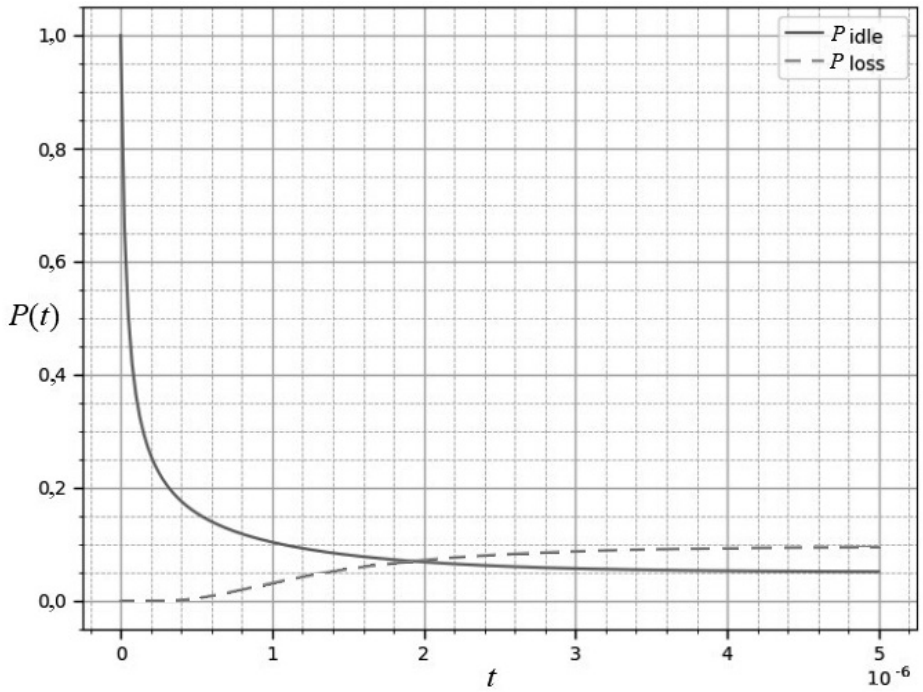


Fig. 5. Dependence of system state probabilities on time in transition mode.

performance, taking into account possible switch reboots when changing information transmission routes in all-optical networks.

Figure 5 shows the time dependence of the number of packets in the first and second phases, as well as the total number of packets in the system in the transient mode, calculated in accordance with (33)–(35). It can be seen that until the moment $t = 0.28 \cdot 10^{-7}$ s the number of packets in the first phase exceeds the number of packets in the second phase. At the same time, in stationary mode, the average number of packets in the first phase is less than in the second phase of service, which is obvious, since $\mu_1 > \mu_2$. Considering that the number of requests in the first and second phases of the QS under study corresponds to the number of packets processed in the first and second stages of the all-optical switch [8], the results obtained make it possible to estimate the degree of filling of the switch buffers during the transition mode.

As the buffer size increases, the size of the matrices in (12) increases, which requires additional computational resources. Figure 4 shows the calculation of a two-phase system with a buffer volume of $N = 15$: the probability of losses $P_{loss}(t)$ and the probability that the system is empty, $P_{idle}(t)$. The graph shows that with an increase in the buffer size, the probability of losses in the stationary mode decreased — $\pi_{loss} = 0.1$, and the time of the transition mode increased — $\tau_{tr} = 4 \cdot 10^{-7}$ s.

7. CONCLUSION

In this paper, the transient mode of a two-phase QS with a Poisson input flow, an exponential law of distribution of service time in each phase and a limitation on the total buffer size of two phases is considered and analyzed. Previously, the non-stationary mode of such a system was not considered in the world literature. However, it is of interest for various applications, in particular in the design of all-optical network switches. It should be noted that the study of the non-stationary mode of an all-optical switch allows a more accurate assessment of its performance metrics, which differ significantly from stationary values due to the high information transfer rate of all-optical networks [8].

A system of differential equations describing the functioning of this QS is presented, the solution of which is written using the Laplace transform. The characteristics of system performance in transient mode, such as the probability of losses, throughput, the average number of serviced requests, and the transition time, were obtained. Obviously, as the buffer size increases, obtaining numerical solutions to the characteristics of a two-phase QS with a limited buffer is a computationally intensive task and requires the use of high-performance computing systems or the use of approaches based on simulation modeling and machine learning [26].

FUNDING

This work was supported by the Russian Science Foundation, project no. 23-29-00795.
https://rscf.ru/en/project/23-29-00795/

APPENDIX

Formally, eliminating certain terms in equations (4) and preserving the remaining ones can be done using the Heaviside function. However, this function is essentially logical, not analytical, and, therefore, does not allow one to write down an expression for the probabilities of system states in a general form. In particular, when using it in program code, it is necessary to organize additional loops. Therefore, to enable a compact analytical representation of the system of equations (4), the analytical function was introduced

$$\sigma_1(x, x_0) = \frac{|x - x_0| + x - x_0}{2|x - x_0|}. \tag{A.1}$$

Thus, the function limiting from below the permissible states of the system has the form

$$\upsilon_1(x, M) = \frac{|x - M + 0.5| + x - M + 0.5}{2|x - M + 0.5|}. \tag{A.2}$$

For example, for $M = 0$ the function $\upsilon_1(x, M)$ has the form shown in Fig. 6.

A shift of 0.5 along the time axis was chosen due to the fact that otherwise, in the state $x = M$ of the system, this function would be indefinite, and its derivative would tend to infinity at this point. Similarly with (A.1), we introduce the function

$$\sigma_2(x, x_0) = \frac{|x_0 - x| + x_0 - x}{2|x_0 - x|}. \tag{A.3}$$

Thus, the function that limits from above the permissible states of the system can be written in the form

$$\upsilon_2(x, K) = \frac{|K - x - 0.5| + K - x - 0.5}{2|K - x - 0.5|}, \tag{A.4}$$

where $K = \overline{0, N}$ is the state of the system. For $K = 4$, the function $\upsilon_2(x, M)$ has the form shown in Fig. 7.

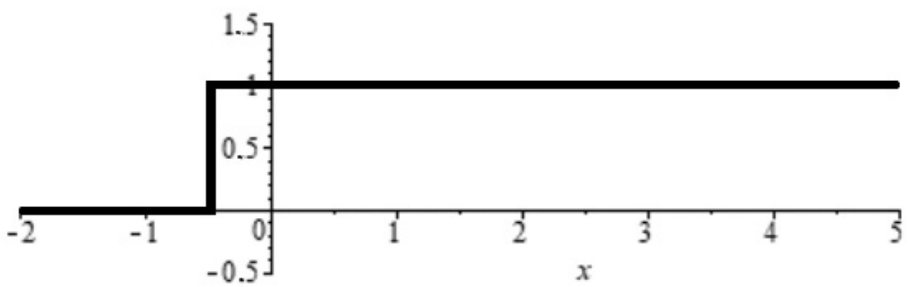


Fig. 6. Function $\upsilon_1(x, 0)$.

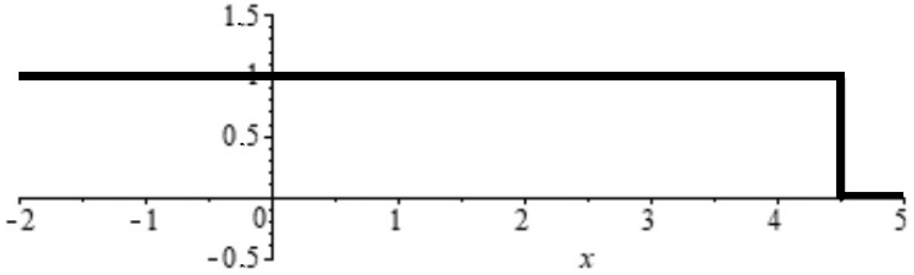


Fig. 7. Function $\upsilon_2(x, 4)$.

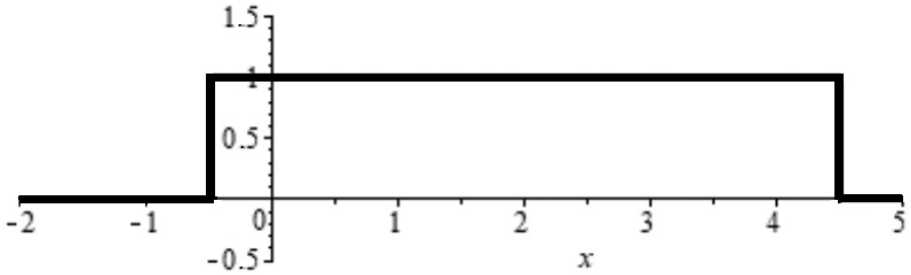


Fig. 8. Function $\upsilon(x, 0, 4)$.

Obviously, the function limiting the permissible range of values from the smallest M to the largest K takes the form

$$\upsilon(x, M, N) = \upsilon_1(x, M)\,\upsilon_2(x, N) = \frac{(|x - M + 0.5| + x - M + 0.5)\,(|K - x - 0.5| + K - x - 0.5)}{4\,|x - M + 0.5|\,|K - x - 0.5|}. \quad \text{(A.5)}$$

For example, with $M = 0$ and $K = 4$ it has the form shown in Fig. 8.

In relation to the problem being solved, x can take the values n_1, n_2, $n_1 + n_2$, etc. The advantage of functions (A.2), (A.4) and (A.5) is the absence of conditions. However, it should be noted that such conditions still exist when the module is expanded. However, despite the fact that these functions do not speed up the calculation process, they allow for analytical study of the resulting expressions and simplify the program code.

To find the function $\vartheta(n_k, n_l)$ (see (6)), which transforms a pair of numbers n_k, n_l, characterizing the state of the system, into the column number of the matrix $\mathbf{A}$, let us analyze the following pattern for $N = 4$: for $n_k = 0$ the values of n_l change from 0 to N, and the values of $\vartheta(n_k, n_l)$ change from 1 up to $N + 1$; for $n_k = 1$ the values of n_l change from 0 to $N - 1$, and the values of $\vartheta(n_k, n_l)$ change from $N + 2$ to $2N + 1$; for $n_k = 2$ the values of n_l change from 0 to $N - 2$, and the values of $\vartheta(n_k, n_l)$ change from $2N + 2$ to $3N$; for $n_k = 3$ the values of n_l change from 0 to $N - 3$, and the values of $\vartheta(n_k, n_l)$ change from $3N + 1$ to $4N - 2$; for $n_k = 4$ we have $n_l = 0$ and $\vartheta(n_k, n_l) = 4N - 1$. Therefore, the expression for $\vartheta(n_k, n_l)$ must contain the term $n_k(N + 1)$, as well as the term n_l. Thus, for $n_k = 0$:

$$\vartheta(0, n_l) = (N + 1)\,n_k + n_l + 1 = (N + 1)\,n_k + n_l + 0 \cdot (-0.5) + 1, \quad \text{(A.6)}$$

for $n_k = 1$:

$$\vartheta(1, n_l) = (N + 1)\,n_k + n_l + 1 = (N + 1)\,n_k + n_l - 1 \cdot 0 + 1, \quad \text{(A.7)}$$

for $n_k = 2$:

$$\vartheta(2, n_l) = (N+1)\, n_k + n_l + 0 = (N+1)\, n_k + n_l - 2 \cdot 0.5 + 1, \tag{A.8}$$

for $n_k = 3$:

$$\vartheta(3, n_l) = (N+1)\, n_k + n_l - 2 = (N+1)\, n_k + n_l - 3 \cdot 1 + 1, \tag{A.9}$$

for $n_k = 4$:

$$\vartheta(4, n_l) = (N+1)\, n_k + n_l - 5 = (N+1)\, n_k + n_l - 4 \cdot 1.5 + 1, \tag{A.10}$$

for $n_k = m$:

$$\vartheta(m, n_l) = (N+1)\, m + n_l - (m+1) = (N+1)\, m + n_l - m\frac{m-1}{2} + 1. \tag{A.11}$$

Thus, expressions (A.6)–(A.11) connect a pair of numbers to the corresponding ordinal number of the element in the row (column) of the coefficient matrix. It is easy to check that relation (6) is valid for any N.

REFERENCES

1. Dudin, A.N., Klimenok, V.I., and Vishnevsky, V.M., *Methods to Study Queuing Systems with Correlated Arrivals*, Berlin/Heidelberg: Springer, 2020.
2. Rohit Singh Tomar and Dr.R.K. Shrivastav, Three Phases of Service For A Single Server Queueing System Subject To Server Breakdown And Bernoulli Vacation, *Int. J. Math. Trend. Techn. (IJMTT)*, 2020, vol. 66 (5), pp. 124–136.
3. Murat Sagir and Vedat Saglam, Optimization and analysis of a tandem queueing system with parallel channel at second station, *Communications in Statistics — Theory and Methods*, 2022, vol. 51, no. 21, pp. 1–14.
4. Sudhesh, R. and Vaithiyanathan, A., Stationary analysis of infinite queueing system with two — stage network server, *RAIRO-Oper. Res.*, 55, 2021, pp. 2349–2357.
5. Al-Rawi, Z.R. and Al Shboul, K.M.S., A Single Server Queue with Coxian-2 Service and One-Phase Vacation (M/C-2/M/1 Queue), *Open J. Appl. Sci.*, 2021, vol. 11, no. 6, pp. 766–774.
6. Serite Ozkar, Two-commodity queueing-inventory system with phase-type distribution of service times, *Annals of Operations Research*, 2022. https://link.springer.com/article/10.1007/s10479-022-04865-3
7. Anastasia Galileyskaya, Ekaterina Lisovskaya, Michele Pagano, and Svetlana Moiseeva, Two-Phase Resource Queueing System with Requests Duplication and Renewal Arrival Process, *LNCS*, 2020, 12563, pp. 350–364.
8. Barabanova, E.A., Vytovtov, K.A., Vishnevsky, V.M., and Podlazov, V.S., High-capacity strictly non-blocking optical switches based on new dual principle, *J. Phys.: Conf. Ser.*, 2021, vol. 2091, no. 1. https://iopscience.iop.org/article/10.1088/1742-6596/2091/1/012040
9. Ivanova, D., Adou, Y., Markova, E., Gaidamaka, Y., and Samouylov, K., Mathematical Framework for Mixed Reservation- and Priority-Based Traffic Coexistence in 5G NR Systems, *Mathematics*, 2023, vol. 11, no. 4. https://doi.org/10.3390/math11041046
10. Zeifman, A.I., Razumchik, R.V., Satin, Y.A., and Kovalev, I.A., Ergodicity bounds for the markovian queue with time-varying transition intensities, batch arrivals and one queue skipping policy, *Appl. Math. Comput.*, 2021, vol. 395, p. 125846.
11. Kempa Wojciech, M. and Paprocka Iwona, Transient behavior of a queueing model with hyper-exponentially distributed processing times and finite buffer capacity, *Sensors*, 2022, vol. 22, no. 24. https://doi.org/10.3390/s22249909

12. Rubino, G., *Transient analysis of Markovian queueing systems: a survey with focus on closed forms and uniformization / Queueing Theory 2: Advanced Trends*, Wiley-ISTE: Hoboken, NJ,. USA, 2021, pp. 269–307.

13. Zeifman, A., Korolev, V., and Satin, Y., Review Two Approaches to the Construction of Perturbation Bounds for Continuous-Time Markov Chains, *Mathematics*, 2020, vol. 8. https://doi.org/10.3390/math8020253

14. Sita Rama Murthy, M., Srinivasa Rao, K., Ravindranath, V., and Srinivasa Rao, P., Transient Analysis of K-node Tandem Queuing Model with Load Dependent Service Rates, *Int. J. Engin. Techno.*, 2018, vol. 7, no. 3.31, pp. 141–149.

15. Suhasini, A.V.S., Rao K. Srinivasa, and Reddy, P.R.S., Transient analysis of tandem queueing model with nonhomogenous poisson bulk arrivals having statedependent service rates, *Int. J. Advanc. Comput. Math. Sci.*, 2012, vol. 3, no. 3, pp. 272–289.

16. Neelam Singla and Garg P.C., Transient and Numerical Solution of a Feedback Queueing System with Correlated Departures, *Amer. J. Numer. Anal.*, 2014, vol. 2, no. 1, pp. 20–28.

17. Shyam Sundar Sah and Ram Prasad Ghimire, Transient Analysis of Queueing Model, *J. Inst. Engin.*, 2015, vol. 11, no. 1, pp. 165–171.

18. Zeifman, A.I., On the Nonstationary Erlang Loss Model, *Autom. Remote Control*, 2009, vol. 70, no. 12, pp. 2003–2012.

19. Kovalev, I.A., Satin, Ya.A., Sinitsina, A.V., and Zeifman, A.I., On one approach to estimating the convergence rate of non-stationary Markov models of queuing systems, *Inform. and its application*, 2022, vol 16, no. 3, pp. 75–82.

20. Prabhu, N.U., Transient Behaviour of a Tandem Queue, *Management Science*, 1967, vol. 13, no. 9, pp. 631–639. https://doi.org/10.1287/mnsc.13.9.631

21. Vyshnevsky, V.M., Vytovtov, K.A., Barabanova, E.A., and Semenova, O.V., Analysis of an $MAP/M/1/N$ queue with periodic and non-periodic piecewise constant input rate, *Mathematics*, 2022, vol. 10, no. 10. https://www.mdpi.com/2227-7390/10/10/1684

22. Vishnevsky, V., Vytovtov, K., Barabanova, E., and Semenova, O., Transient behavior of the $MAP/M/1/N$ queuing system, *Mathematics*, 2021, vol. 9, no. 2559. https://doi.org/10.3390/math9202559

23. Vytovtov, K.A., Barabanova, E.A., and Vishnevsky, V.M., Modeling and Analysis of Multi-channel Queuing System Transient Behavior for Piecewise-Constant Rates, *LNCS*, 2023, vol. 13766, pp. 397–409.

24. Jackson, R.R.P., Queueing Systems with Phase Type Service, *Oper. Res. Soc.*, 1954, vol. 5, no. 4, pp. 109–120.

25. Zhuravlev, A.P., Ryumshin, K.Yu., Atakischev, O.I., Titenko, E.A., and Titenko, M.A., Modulation parameters of modern communication systems, *T-Comm: telecommunications and transport.*, 2023, vol. 17, no. 7, pp. 13–20.

26. Vishnevsky, V.M. and Semyonova, O.V., *Metody mashinnogo obucheniya dlya issledovaniya stokhasticheskikh modelei tsiklicheskogo oprosa v shirokopolosnykh besprovodnykh setyakh* (Machine learning methods for studying stochastic models of cyclic polling in broadband wireless networks), Moscow: IPU, RAS, 2023.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 1, pp. 64–72.

STOCHASTIC SYSTEMS

On the Problem of Maximizing the Probability of Successful Passing of a Time-Limited Test

A. V. Naumov[*,a], **A. E. Stepanov**[*,b], **and A. E. Ustinov**[*,c]

[*] *Moscow Aviation Institute (National Research University), Moscow, Russia*
e-mail: [a] *naumovav@mail.ru,* [b] *Rus.fta@yandex.ru,* [c] *entro1122@gmail.com*

Received October 18, 2023
Revised November 30, 2023
Accepted December 21, 2023

Abstract—The problem of finding the optimal sequence of performing a set of tasks in a time-limited test is considered. That is, a task group is allocated for mandatory initial execution in the test, the remaining tasks are performed during the remaining time until the end of the test. For each correctly completed task of the test, the subject is awarded a certain number of points. The proposed criterion is the probability that the total number of points scored for the test exceeds a certain level, which is a fixed parameter, while simultaneously fulfilling the time limit of the test. Random parameters are the user's response time to each test task. The correctness of the user's answer to the task is modeled by a random variable with a Bernoulli distribution. The resulting stochastic bilinear programming problem boils down to a deterministic integer problem of mathematical programming.

Keywords: time-constrained test, maximum likelihood estimation, integer mathematical programming

DOI: 10.31857/S0005117924010051

1. INTRODUCTION

In the paradigm of computerized adaptive testing [1–6] the problem of constructing an optimal test passing strategy is related to the formation of an individual learning trajectory. This problem seems to be relevant in various fields of the educational process: preparation for passing the Unified State Exam (USE) by applicants, passing regular tests in the learning management system (LMS), checking the residual knowledge of students, etc. As a rule, such forms of testing are limited in time, and the structure of the test is known in advance with an accuracy of the types of tasks, or sections of the course, to test the knowledge of which these tasks are aimed. At the same time, there are considerable statistics of the performance of such tasks by the subjects during the training both by the type of tasks and by the individual user. Often, for example, in the context of using LMS in the educational process, the collection and storage of this information are automated, as is done in the CLASS.NET LMS of the Moscow Aviation Institute [7, 8]. This allows us to reasonably use in the problem under consideration mathematical models of random parameters taken into account, for example, the time spent by the subject on solving a task of a certain type. Models of the response time of the user to the test task are widely represented in the literature. Van der Linden proposed a lognormal time model [1], and in [9, 10] gamma distribution and discrete distribution were used as models. The frequency of the correct solution of the problems of a certain type by the subject, obtained on the basis of data on the work of the subject in the learning process, can serve as a good estimate of the parameter of the Bernoulli distribution, modeling the correct solution of the corresponding task in the test by the subject. All tasks of the test, as a rule, are characterized by a certain number of points that the subject scores, having solved them correctly. The total

number of points scored during testing characterizes the quality of the subject's preparation and is the basis for his assessment. The achievement by the subject of a certain total score for the test can serve as a certain target indicator for him. As a strategy of the subject in the presence of the above-mentioned random parameters in the problem, a set of test tasks can be used, which should be performed first.

The literature presents quite widely the problems of constructing various tests in order to check the level of knowledge of the subjects, including in probabilistic or quantile statement [3, 4, 6, 9, 10]. However, the authors are not aware of publications in which the problems of constructing an optimal strategy for the subject to pass the test would be considered.

The paper formulates the problem of finding the optimal strategy of the subject (in the above sense) according to the criterion of maximizing the probability of gaining a total score for the test above a certain level chosen by the subject. This takes into account the probabilistic constraint associated with the fact that the time spent by the subject on completing the test should not exceed the total fixed testing time.

This problem of stochastic bilinear programming based on the generalized minimax approach [11] according to the technique proposed in [12] is reduced to an integer mathematical programming problem. The initial data for estimating the parameters of the probabilistic models used in the problem are taken from the statistics of the work of users of the MAI CLASS.NET LMS. The paper also discusses the results of a numerical experiment and the dependence of the optimal value of the criterion on the total score, which the subject seeks to exceed.

2. DISTRIBUTIONS OF RANDOM PARAMETERS USED IN THE WORK

Two vector random parameters are used in the problem under consideration. One of them is the vector $X = \mathrm{col}\,(X_1, \ldots, X_n)$, the ith coordinate of which model the correctness of the solution of the ith test task. It is assumed that X_i are independent random variables having a Bernoulli distribution, with parameters $p_i, i = 1, \ldots, n$, estimated by the frequency of correct answers of the subject to similar tasks of the ith type during preparation for testing or in the learning process. Equality $X_i = 1$ models the correctness of the solution of the ith task, and equality to zero — the opposite event. Another random parameter is the vector $T = \mathrm{col}(T_1, \ldots, T_n)$, the coordinates of which characterize the time spent by the subject on solving the task of the ith type. Random variables $T_i,\ i = 1, \ldots, n$ are also assumed to be independent. However, it would be reckless to assume independence between the values of X and T, therefore for each value of X_i (0 or 1) its own distribution of the random variable T_i is estimated also on the basis of statistics of solving tasks of a similar type by the subject. Continuous distributions of the user's response time to the task (Van der Linden [1], Gamma distributions [9]) do not allow finding an exact solution to the problem in a probabilistic statement, therefore a discretized response time model with three values is used in the work, modeling situations of quick solution, standard solution and solution with difficulties. The technique for constructing a discrete distribution law of the response time of the subject to the test tasks can be different: from discretization in various ways of continuous distribution models (for example, the Van der Linden model [1]), the parameters of which are determined based on statistical data on the time of solving the problems of the corresponding class by the subject, to using the initial distribution histogram, built according to the same statistical data. In this work, for each task of the test, based on the available statistical data obtained from the CLASS.NET system [8], a variation series of the response time of the subject to similar tasks is constructed, which is divided into three equal parts at a distance between the maximum and minimum elements. For each part, the sample mean is calculated, which is used as the corresponding possible value of the random variable of interest to us. The probabilities of the obtained three possible values are assumed to be equal to the frequencies of the elements of the sample used falling into the corresponding selected

ranges. Thus, the general vector of random variables has a discrete distribution with the number of implementations $D = 2^n 3^n$. The probabilities of each implementation can be found using the formula for multiplying probabilities, and using the conditional distribution of the response time of the subject to the test tasks under the conditions of its correct or incorrect solution.

3. STATEMENT OF THE PROBLEM AND METHOD OF ITS SOLUTION

It is required to define the strategy of the testee while performing a time-limited test consisting of n tasks. The strategy is defined by a vector of Boolean variables $u \in \{0,1\}^n$, where

$$u_i \triangleq \begin{cases} 1, & \text{if the testee tries to solve the } i\text{th task of the test,} \\ 0 & \text{otherwise,} \end{cases} \qquad i = \overline{1,n}.$$

For each ith task of the test, b_i points are awarded. To successfully pass the test, it is necessary to score at least φ points. The time for completing the test is limited to $\overline{T}$. The probability of successfully passing the test while simultaneously fulfilling the time limit for its completion was chosen as the optimization criterion.

Let us consider the probability function

$$P_{\varphi,\overline{T}}(u) \triangleq P\left\{\sum_{i=1}^{n} u_i X_i b_i \geqslant \varphi,\ \sum_{i=1}^{n} u_i T_i \leqslant \overline{T}\right\}.$$

In it, the values φ and $\overline{T}$ play the role of parameters. To ensure the reasonableness of the problem statement, we impose restrictions on the specified parameters: $0 < \varphi \leqslant \sum_{i=1}^{n} b_i$ and $\sum_{i=1}^{n} T_i^{\min} \leqslant \overline{T}$, where $T_i^{\min}$ is the minimum time for the testee to solve the ith task. Then the problem of finding the optimal strategy for the testee can be formulated as follows:

$$P_{\varphi,\overline{T}}(u) \to \max_{u \in \{0,1\}^n}. \tag{1}$$

This problem is a stochastic programming problem with Boolean variables.

As mentioned above, the number of all possible realizations of the random parameter vector $\mathrm{col}(X^\top, T^\top)$ is $D = 2^n 3^n$. Let us consider the vector $\delta \in \{0,1\}^D$, each νth coordinate of which corresponds to one of the realizations $\mathrm{col}((x^\nu)^\top, (t^\nu)^\top)$ of the vector $(X^\top, T^\top)^\top$ and can take values 0 or 1. Let $\Upsilon \triangleq e^T b$, where $e = \mathrm{col}(1, \ldots, 1) \in R^n$, i.e. $\Upsilon = \sum_{i=1}^{n} b_i$ — the maximum number of points that can be scored for the test. Let $p_\nu = P(\mathrm{col}(X^\top, T^\top) = \mathrm{col}((x^\nu)^\top, (t^\nu)^\top))$, $\nu = \overline{1,D}$. Then, similarly to the method proposed in [12] for solving the problem of minimizing the quantile function based on the confidence method [11], the stochastic programming problem (1) can be reduced to a deterministic optimization problem with Boolean variables:

$$\sum_{\nu=1}^{D} p_\nu \delta_\nu \to \max_{\substack{u \in \{0,1\}^n \\ \delta \in \{0,1\}^D}} \tag{2}$$

under the constraints

$$\varphi - \Upsilon - \delta_\nu \left(\sum_{i=1}^{n} u_i x_i^\nu\ b_i - \Upsilon\right) \leqslant 0, \quad \nu = \overline{1,D}, \tag{3}$$

$$\delta_\nu u^T t^\nu - \overline{T} \leqslant 0, \quad \nu = \overline{1,D}. \tag{4}$$

In the problem considered above, the optimal value of the vector δ determines the type of the optimal confidence set (in terms of the confidence method [11]) as the implementations of the random

parameter vector corresponding to ones in the optimal vector δ. The total probability measure of such implementations is maximized in the problem under consideration, and the constraints on the test execution time and the number of points scored during the test are satisfied, while for the remaining implementations corresponding to zero values of the coordinates of the optimal vector δ, these constraints in the original problem may not be satisfied, and constraints (3) and (4) are satisfied by construction.

Problem (2)–(4), constructed strictly according to the technique of [12], is a bilinear programming problem (with a bilinear system of constraints), which, together with the Boolean nature of the variables and the large dimensionality, makes it difficult to solve. However, the structure of the problem under consideration makes it possible to rewrite it as a linear programming problem (LPP), which will possibly allow the use of special methods for solving LPPs with Boolean variables implemented in modern applied optimization software packages. The form of this LPP is as follows:

$$\sum_{\nu=1}^{D} p_\nu \delta_\nu \to \max_{\substack{u\in\{0,1\}^n \\ \delta\in\{0,1\}^D}} \tag{5}$$

under the constraints

$$\delta_\nu \varphi \leqslant \sum_{i=1}^{n} u_i x_i^\nu \; b_i, \quad \nu = \overline{1,D}, \tag{6}$$

$$\sum_{i=1}^{n} u_i b_i \geqslant \varphi, \;\; \sum_{i=1}^{n} u_i T_i^{\min} \leqslant \overline{T}.$$

$$u^T t^\nu \leqslant \delta_\nu \overline{T} + (1-\delta_\nu) T^{MAX}, \;\; \nu = \overline{1,D}, \tag{7}$$

where $T^{MAX} = \sum_{i=1}^{n} T_i^{\max}$, and $T_i^{\max}$ is the maximum of all possible realizations of the random variable T_i, $i = \overline{1,n}$.

If the dimensions of the problems (2)–(4), (5)–(7) allow to solve them using standard procedures from known libraries of optimization programs, then the solution can be found with their help. However, these problems contain an additional vector of optimization variables $\delta \in \{0,1\}^D$ of large dimension, which, taking into account the large number of constraints, makes them difficult to solve by exhaustive search of all possible values of Boolean optimization variables and requires the development of special solution methods that take into account the structure of the problem. Next, we consider an algorithm for solving the original problem. The efficiency of its application in comparison with standard library procedures for solving problems (2)–(4), (5)–(7) will be discussed in the section concerning the results of a numerical experiment.

4. ALGORITHM FOR SOLVING THE ORIGINAL PROBLEM

Step 0.

Of all 2^n strategies $u \in \{0,1\}^n$ we choose N, forming the set $\underline{U}$, for the elements of which the following conditions are satisfied

$$\sum_{i=1}^{n} u_i b_i \geqslant \varphi, \;\; \sum_{i=1}^{n} u_i T_i^{\min} \leqslant \overline{T}.$$

The point is that in this way we filter out strategies that are obviously unsuitable in terms of the total time or the number of points even in the most optimistic case, when all the tasks selected for solving the problem are solved correctly and in the minimum possible time.

We renumber all elements of the set $\underline{U}$. Thus, a number from 1 to N uniquely defines an element of the set. By u^m we will mean the mth element of the set $\underline{U}$. Let $m := 1$, $P^* := 0$, and $u^* := (0, \ldots, 0)^\top$.

At this step, the external loop for enumerating all N selected optimization strategies is initialized.

Step 1.

If $m > N$, then go to Step 5. Otherwise $P_m := 0$.

The auxiliary parameter P_m is used below to calculate the probability of fulfilling the constraints at $u = u^m$.

Step 2.

Suppose that the vector u^m contains exactly K ones. Suppose that the nonzero components of the vector u^m are the components with numbers $i_1, \ldots, i_K$. Consider the subvector $\mathrm{col}(X_{i_1}, \ldots, X_{i_K})$ of the random vector X. Let $J := 2^K$, and $j := 1$.

At this step, the cycle of enumerating all possible realizations $\mathrm{col}(x^j_{i_1}, \ldots, x^j_{i_K})$, $j = \overline{1, 2^K}$ is initialized.

Step 3.

If $j > J$ and $P_m > P^*$, then we assume $P^* := P_m$, $u^* := u^m$, $m := m + 1$ and proceed to Step 1.

If $j > J$ and $P_m \leqslant P^*$, then we assume $m := m + 1$ and proceed to Step 1.

Otherwise, if for the realization $\mathrm{col}(x^j_{i_1}, \ldots, x^j_{i_K})$ the condition

$$\sum_{i \in \{i_1, i_2, \ldots, i_K\}} u^m_i x^j_i \geqslant \varphi,$$

is satisfied, then we assume $L := 3^K$, $l := 1$ and proceed to Step 4. If the specified condition is not satisfied, then we assume $j := j + 1$ and proceed to the beginning of Step 3.

At this step, the cycle of enumerating all possible realizations $\mathrm{col}(t^l_{i_1}, \ldots, t^l_{i_K})$, $l = \overline{1, L}$ of the subvector $\mathrm{col}(T_{i_1}, \ldots, T_{i_K})$ of the random vector T is initialized.

Step 4.

If $l > L$, then we assume $j := j + 1$ and proceed to Step 3. Otherwise, if for the realization $\mathrm{col}(t^l_{i_1}, \ldots, t^l_{i_K})$ the condition

$$\sum_{i \in \{i_1, i_2, \ldots, i_K\}} u^m_i t^l_i \leqslant \overline{T},$$

is satisfied, then we assume $P_m := P_m + \prod_{i \in \{i_1, i_2, \ldots, i_K\}} P(T_i = t^l_i | X_i = x^j_i) P(X_i = x^j_i)$.

We assume $l := l + 1$ and proceed to the beginning of Step 4.

Step 5. We assume the optimal value of the criterion to be equal to P^*, and the optimal value of the strategy to be equal to u^*.

Note that in all nested cycles considered in the algorithm, there is a significant reduction in the required volume of enumeration of possible values of optimization variables. The volume of full enumeration can be reduced by an order of magnitude depending on the selected values of the task parameters φ and $\overline{T}$.

5. RESULTS OF NUMERICAL EXPERIMENT

The initial distributions for solving the problem were obtained based on the analysis functioning of the MAI CLASS.NET learning management system [8]. We will assume the number of tasks in the test is $n = 10$. Estimates of the parameters of the initial distributions are given in Tables 1–6.

Table 1. Probability of correct solution of test tasks

Task number in the test	Probability of correct solution	Number of points for the task
1	0.9	1
2	0.91	1
3	0.95	1
4	0.97	1
5	0.90	1
6	0.8	2
7	0.65	3
8	0.75	2
9	0.5	4
10	0.55	3

Table 2. Conditional distribution of the test subject response time for a test task in case of its incorrect solution

Task number in the test	Conditional distribution of response time			
1	t_1^j	60	100	160
	$P(T_1 = t_1^j \mid X_1 = 0)$	0.3	0.55	0.15
2	t_2^j	70	130	250
	$P(T_2 = t_2^j \mid X_2 = 0)$	0.25	0.6	0.15
3	t_3^j	60	150	270
	$P(T_3 = t_3^j \mid X_3 = 0)$	0.2	0.45	0.35
4	t_4^j	100	200	350
	$P(T_4 = t_4^j \mid X_4 = 0)$	0.15	0.6	0.25
5	t_5^j	75	140	210
	$P(T_5 = t_5^j \mid X_5 = 0)$	0.2	0.45	0.35
6	t_6^j	190	290	400
	$P(T_6 = t_6^j \mid X_6 = 0)$	0.1	0.65	0.25
7	t_7^j	310	380	450
	$P(T_7 = t_7^j \mid X_7 = 0)$	0.2	0.4	0.4
8	t_8^j	180	250	320
	$P(T_8 = t_8^j \mid X_8 = 0)$	0.1	0.3	0.6
9	t_9^j	360	480	600
	$P(T_9 = t_9^j \mid X_9 = 0)$	0.1	0.3	0.6
10	t_{10}^j	320	400	470
	$P(T_{10} = t_{10}^j \mid X_{10} = 0)$	0.3	0.55	0.15

Consider the value $b^{\max} = \sum_{i=1}^{n} b_i$. As a result of the proposed algorithm, the dependences of optimal solutions on the values of the problem parameters φ and $\overline{T}$ were obtained, see Tables 4 and 5.

The calculations were performed on a computer ASUS X550LC (Intel Core i5 2.3 GHz, 8Gb RAM). The linear programming problem was solved by the IBM Cplex package from the Python library. As can be seen, the values of the problem parameters significantly affect the speed of the

Table 3. Conditional distribution of the test subject response time for a test task in case of its correct solution

Task number in the test	Conditional distribution of response time			
1	t_i^j	60	180	300
	$P(T_1 = t_1^j \mid X_1 = 1)$	0.3	0.5	0.2
2	t_i^j	75	190	330
	$P(T_2 = t_2^j \mid X_2 = 1)$	0.15	0.6	0.25
3	t_i^j	60	120	250
	$P(T_3 = t_3^j \mid X_3 = 1)$	0.15	0.35	0.5
4	t_i^j	130	200	350
	$P(T_4 = t_4^j \mid X_4 = 1)$	0.1	0.3	0.6
5	t_i^j	75	140	210
	$P(T_5 = t_5^j \mid X_5 = 1)$	0.2	0.45	0.35
6	t_i^j	200	275	380
	$P(T_6 = t_6^j \mid X_6 = 1)$	0.2	0.35	0.45
7	t_i^j	310	380	450
	$P(T_7 = t_7^j \mid X_7 = 1)$	0.2	0.4	0.4
8	t_i^j	200	290	370
	$P(T_8 = t_8^j \mid X_8 = 1)$	0.25	0.4	0.35
9	t_i^j	380	470	650
	$P(T_9 = t_9^j \mid X_9 = 1)$	0.1	0.25	0.65
10	t_i^j	150	275	500
	$P(T_{10} = t_{10}^j \mid X_{10} = 1)$	0.3	0.55	0.15

Table 4. Dependence of the optimal solution on the parameter φ at $\overline{T} = 0.8\, T^{\max}$

φ	Optimal strategy	Opt. value of the criterion	Calculation time (sec)	Number of investigated strategies
$0.4b^{\max}$	1. 1. 1. 1. 1. 1. 1. 1. 0. 1.	0.9241	29.9	727
$0.5b^{\max}$	1. 1. 1. 0. 1. 1. 1. 1. 1. 1.	0.7874	18.2	511
$0.6b^{\max}$	1. 1. 1. 0. 1. 1. 1. 1. 1. 1.	0.5777	8.9	295
$0.7b^{\max}$	1. 1. 1. 0. 1. 1. 1. 1. 1. 1.	0.3830	3.2	131
$0.8b^{\max}$	1. 1. 1. 1. 1. 1. 1. 1. 1. 1.	0.2028	0.9	39
$0.9b^{\max}$	1. 1. 1. 1. 1. 1. 1. 1. 1. 1.	0.0816	0.1	5

author's algorithm. Thus, increasing the desired number of points for the test leads to a decrease in the probability of achieving this result and a decrease in the calculation time using the proposed algorithm due to a decrease in the volume of enumeration of admissible optimization strategies u. Comparative analysis of the running time of the algorithms shows its significant growth with an increase in the number of tasks in the test, see Table 6. All algorithms with the same problem parameters received coinciding solutions for all values of n. At $n \geqslant 7$, it was not possible to solve the LP problem due to problems with insufficient memory for storing the matrix of the constraint system. The most effective was the authors' algorithm, which for large n exceeded by an order of magnitude the running time of other considered algorithms.

Table 5. Dependence of the optimal solution on the parameter $\overline{T}$ at $\varphi = 0.6\ b^{\max}$

$\overline{T}$	Optimal strategy	Opt. value of the criterion	Calculation time (sec)	Number of investigated strategies
$0.4T^{\max}$	0. 0. 0. 0. 0. 1. 1. 0. 1. 1.	0.0633	1.8	273
$0.5T^{\max}$	0. 0. 1. 0. 1. 1. 1. 1. 0. 1.	0.1733	9.5	296
$0.6T^{\max}$	1. 1. 1. 0. 1. 1. 1. 1. 0. 1.	0.2786	8.7	296
$0.7T^{\max}$	1. 1. 1. 1. 1. 1. 1. 1. 0. 1.	0.5049	9.0	296
$0.8T^{\max}$	1. 1. 1. 0. 1. 1. 1. 1. 1. 1.	0.5777	8.9	296
$0.9T^{\max}$	1. 1. 1. 1. 1. 1. 1. 1. 1. 1.	0.7213	8.3	296

Table 6. Algorithms running time (s) for different values of n

n	LPP	Full enumeration	Authors' algorithm
1	0.0010	0.0340	0.0010
2	0.0010	0.0400	0.0016
3	0.0102	0.0580	0.0020
4	0.0202	0.0800	0.0029
5	0.0628	0.2000	0.0077
6	0.0991	0.6100	0.0117
7	–	2.1900	0.0636
8	–	9.3800	0.1600
9	–	48.1000	2.1500
10	–	3010.0000	9.1100

6. CONCLUSION

This article considers the problem of stochastic programming to search for an optimal strategy for passing a time-limited test according to the criterion of the maximum probability of the testee overcoming a certain number of points scored for the test, taking into account the time limit for completing the test. For the testee, the probability of a correct solution to each test task is considered known. The time spent by the testee on solving each task is also random.

The considered problem of stochastic programming with a probabilistic quality criterion is reduced to a deterministic problem of large dimensionality, for which standard optimization procedures from known program libraries can be used. In addition, an algorithm is proposed for a directed search possible values of a discrete optimization strategy, reducing time costs for solving the problem. The conducted numerical experiment confirmed the effectiveness of using the developed algorithm for solving the original problem in comparison with the use of standard optimization procedures for its deterministic equivalents. This efficiency, determined by the difference in the time of calculating the optimal strategy in various ways, increases with an increase in the number of tasks in the test. Numerical experiment also showed a significant dependence of the optimal solution and the time of its calculation on the parameters tasks, which justifies the relevance of further improvement her solution algorithm.

The results obtained in the work can be extended to the quantile formulation of the problem under consideration, when the testee seeks to maximize the number of points scored for the test while maintaining the selected probability level of fulfilling all the constraints of the problem. This is a separate study, the results of which the authors plan to publish.

FUNDING

This work was supported by the Russian Science Foundation, project no. 23-21-00293.

REFERENCES

1. Van der Linden, W.J., Scrams, D.J., Schnipke, D.L., et al., Using Response-Time Constraints to Control for Differential Speededness in Computerized Adaptive Testing, *Apll. Psychol. Meas.*, 1999, vol. 23, no. 3, pp. 195–210.
2. Rasch, G., *Probabilistic models for some intelligence and attainment tests*, Chicago: The University of Chicago Press, 1980.
3. Kuravsky, L.S., Marmalyuk, P.A., Alkhimov, V.I., and Yuryev, G.A., A New Approach to the Construction of Intellectual and Competency Tests, *Modelirovanie i analiz dannyh*, 2013, no. 1, pp. 4–28.
4. Kuravsky, L.S., Margolis, A.A., Marmalyuk, P.A., Panfilova, A.S., Yuryev, G.A., and Dumin, P.N., A Probabilistic Model of Adaptive Training, *Appl. Math. Sci.*, 2016, vol. 10, no. 48, pp. 2369–2380.
5. Naumov, A.V. and Martyushova, Ya.G., Adapting a Distance Learning System Based on Statistical Processing of User Performance Results, *Elektronnyi Zhurnal "Trudy MAI"*, no. 109, December, 2019.
6. Bosov, A.V., Martyushova, Ya.G., Naumov, A.V., and Sapunova, A.P., Bayesian Approach to Building an Individual Trajectory of a User in a Distance Learning System, *Informatika i ee Primeneniya*, 2020, vol. 14, no. 3, pp. 86–93.
7. Naumov, A.V., Dzhumurat, A.S., and Inozemtsev, A.O., The CLASS.NET System for Distance Learning of Mathematical Disciplines, *Vestnik Komp'yuternykh i Informatsionnykh Tekhnologii*, 2014, no. 10, pp. 36–40.
8. MAI CLASS.NET Distance Learning System [Electronic resource], URL: `https://distance.kaf804.ru/` (access date: 27.02.2023)
9. Bosov, A.V., Mkhitarian, G.A., Naumov, A.V., and Sapunova, A.P., Using the Gamma Distribution in the Task of Forming a Time-Limited Test, *Informatika i ee Primeneniya*, 2019, vol. 13, no. 4, pp. 12–18.
10. Naumov, A.V., Mkhitarian, G.A., and Cherygova, E.E., Stochastic Formulation of the Task of Forming a Test of a Given Level of Difficulty with Minimization of the Time of Completion Quantile, *Vestnik Komp'yuternykh i Informatsionnykh Tekhnologii*, 2019, no. 2, pp. 37–46.
11. Kan, Yu.S. and Kibzun, A.I., *Zadachi stokhasticheskogo programmirovaniya s veroyatnostnymi kriteriyami* (Problems of Stochastic Programming with Probabilistic Criteria), Moscow: Fizmatlit, 2009.
12. Kibzun, A.I., Naumov, A.V., and Norkin, V.I. Reduction of a Quantile Optimization Problem with Discrete Distribution to a Mixed Integer Programming Problem, *Autom. Remote Control*, 2013, vol. 74, no. 6, pp. 951–967.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 1, pp. 73–84.

CONTROL IN TECHNICAL SYSTEMS

On the Determination of the Region Border Prior to the Limit Steady Modes of Electric Power Systems by the Analysis Method of the Tropical Geometry of the Power Balance Equations

M. I. Danilov[*,a] **and I. G. Romanenko**[*,b]

[*] *North Caucasus Federal University, Stavropol, Russia*
e-mail: [a] *mdanilov@ncfu.ru,* [b] *irina_romanenko_@mail.ru*

Received September 28, 2023
Revised October 30, 2023
Accepted December 21, 2023

Abstract—The analysis of the known [8] approach in which tropical geometry over complex multifields of active power balances is used to estimate the region of existence of the electric power system mode. Its limitations are shown and a new approach is proposed, a criterion is also represented for determining the boundary that precedes the violation of the stability of the energy system due to the restructuring of the tropical set of solutions. The developed approach allows to determine the approach of the power system mode to the limit by the known parameters of the lines and the dynamics of changes of the nodes voltage modules and the nodes load.

Keywords: electric power system, static stability, regime existence subdomains, tropical geometry

DOI: 10.31857/S0005117924010066

1. INTRODUCTION

Calculation of steady-state modes (states) of electric power systems is necessary to verify the actual possibility of transmitting of the required power to consumers from existing generators [1–3]. In addition, it is important for control problems [4–6] to be able to estimate the design parameters of the power system state under study relative to the maximum possible mode [7–9]. The limiting mode is determined by increasing the energy consumption of nodes (buses) to critical values, at which the power balance in the system is still maintained. In the event of a further increase of the nodes load, the balance in power system will not be maintained due to a violation of static stability, and such a mode is impossible to actually implement. The procedure of finding limiting modes, called weighting, is performed by selecting nodes and the step of incrementing their power, which can be determined by empirical considerations based on an analysis of the network topology or other necessary criteria. There are various approaches to searching of limiting modes and the criteria used in this case. The traditional [3] method includes the Zhdanov approach and criterion; in this case, weighting is performed and the Jacobian of the linearized equations of the power system steady state is controlled to be equal zero [1, 2]. An optimization model of electrical power systems limiting modes is proposed and the importance of method of balancing node setting is shown in [3], which is also noted in [7]. It is proposed to use an approach based on tropical geometry over complex multipoles of active power balance equations in [8]. In [9], measurement data performed by PMU (Phasor Measurement Unit) devices are used to estimate the voltage safety margin. In [10], the limiting states of the power system are found by incrementing of the nodes load at each step of

the iterative calculation of the steady state. In [11, 12] it is proposed to use a modal approach to solve the problem of ensuring static stability, which consists in analyzing of the eigenvalues of the Jacobian matrix. In [13, 14], the limiting steady states of the power system are found by the Holomorphic Embedding Load-flow Method (HELM), which guarantees that the found solution, when it exists, always corresponds to the correct one, and otherwise signals about the absence of solutions. In [15], an approach to approximating of region boundaries of the regime existence is proposed, taking into account the limitation of generator reactive power. In [16], emergency modes and the using of group lines switching to control and ensure static voltage stability were considered. In [17], an algorithmic procedure of adjusting of the known (previously found) base point of the network limit mode according to the occurrence of changes in node loads is proposed. In [18, 19], methods of estimating of the voltage stability margin for power supply systems with distributed and [20] renewable energy sources are proposed. The authors of [21] proposed a method using the holomorphic load embedding approach, as well as arc length parameterization and piecewise approximations to determine the boundary of the mode existence and to track the entire PV curve of power system nodes. The shape of the boundary of the energy network limiting modes (in the multidimensional space of parameters), which has a complex topology with non-intersecting isolated sections, is studied in [22, 23]. In [24], the conditions of the mode existence of are given only for linear circuits of four-terminal networks operating on alternating current with constant power loads. It must be noted that such schemes correspond to simulated power lines when calculating critical modes.

Determination and research of the stability boundary of the power system and its power lines and associated mode calculation methods [25, 26] are also important when optimizing of the operation of both the distribution network [27–29] and power plants [30].

The presented article analyzes the results of work [8], notes its shortcomings, and proposes an original approach and criterion for determining the boundary of the region of permissible modes preceding instability.

The material of the article is structured as follows. In Section 2, the problem of determining the boundary of the region preceding the limiting steady states of a power system with an arbitrary number of buses is formulated. Section 3 discusses the problem of load power supply through a line from an infinite power bus and the application of analysis over a complex tropical multipole to this system. The main properties of their solutions are noted. Section 4 demonstrates the results obtained using two examples of power system calculations. The first example: a power system with four nodes, in one of which the active power is increased. The second is a standard IEEE 5-bus scheme. The possibility of determining of the boundary preceding the loss of the power system mode stability is shown. The appendix provides the derivation of the theoretical expressions used in the article.

2. STATEMENT OF THE PROBLEM OF DETERMINING OF THE REGION BOUNDARY PRECEDING THE LIMITING POWER SYSTEM STEADY STATES WITH AN ARBITRARY NUMBER OF BUSES

Let's consider a power system with a known electrical network topology, nominal voltage classes, and a number of buses n. Let's introduce the variable v ($v = \overline{0, n}$) to indicate the node numbers. Electricity consumption in load nodes is specified by complex power values $\dot{p}_v^{\text{load}} = {p_{\text{load}}}_v^{\text{Re}} + j{p_{\text{load}}}_v^{\text{Im}}$ ($v \in PQ$), where PQ is the set of load nodes with active ${p_{\text{load}}}_v^{\text{Re}}$ and reactive ${p_{\text{load}}}_v^{\text{Im}}$ powers. Electricity enters the system through generating units. Nodes with given complex values of powers $\dot{p}_v^{\text{gen}} = {p_{\text{gen}}}_v^{\text{Re}} + j{p_{\text{gen}}}_v^{\text{Im}}$ are called generator nodes of PQ-type ($v \in PQ$). Nodes with given values of active ${p_{\text{gen}}}_v^{\text{Re}}$ powers and adjustable voltage modules U_v are called PV-type nodes ($v \in PV$).

Also, when calculating modes in the power system, one of the nodes is assumed to be balancing ($v = 0$), it represents an infinite power bus with a given complex voltage $\dot{U}_0$.

Next, we make the following assumptions.

1. Electricity transmission is carried out using lines, which are represented by π-shaped equivalent circuits with lumped parameters.

2. The parameters of all electrical network lines are known and are represented by longitudinal active $R^{\text{line}}_{v,k}$ and reactive $X^{\text{line}}_{v,k}$ resistances, transverse capacitive $B^{\text{line}}_{v,k}$ conductivities to ground, where v and k are the numbers of nodes between which the line is connected.

The equations necessary to calculate the steady state of the power system are written in complex form according to the nodal voltage method, taking into account the presence of PV-type nodes:

$$\dot{U}^*_v \dot{\hat{I}}_v = \dot{p}^{*\,\text{gen}}_v - \dot{p}^{*\,\text{load}}_v,\ v \in PQ,\quad v = \overline{1,n}, \tag{1}$$
$$\text{Re}\left[\dot{U}^*_v \dot{\hat{I}}_v\right] = \dot{p}^{\text{Re}}_v,\quad \left|\dot{U}_v\right| = \text{const},\quad v \in PV,\quad v = \overline{1,n},$$

in which

$$\dot{\hat{I}}_v = \dot{U}_v \underline{Y}_{v,v} - \sum_{k \in A^g_v} \dot{U}_k \underline{Y}_{v,k} - \dot{U}_0 \underline{Y}_{0,v},$$

where $\dot{U}^*_v$ is the conjugate voltage complex $\dot{U}_v$; $\underline{Y}_{v,v}$ – self-conductance of all branches connected to node v; A^g_v – nodes directly connected to node v; $\dot{U}_k$ – voltage of node k; $\underline{Y}_{v,k}$ – mutual conductance between nodes v and k; $\dot{U}_0$ – the known complex voltage of the balancing ($v = 0$) node; $\underline{Y}_{0,v}$ – conductivity of all branches directly connecting node v and balancing; $\dot{p}^{*\,\text{gen}}_v$, $\dot{p}^{*\,\text{load}}_v$ – conjugate complexes $\dot{p}^{\,\text{gen}}_v$ and $\dot{p}^{\text{load}}_v$.

The conductivities $\underline{Y}_{v,v}$, $\underline{Y}_{v,k}$ and $\underline{Y}_{0,v}$ are defined as follows:

$$\underline{Y}_{v,v} = \sum_{m \in A^b_v} \frac{1}{R_m + jX_m} + \sum_{m \in A^b_v} jB_m,\quad v = \overline{1,n}, \tag{2}$$
$$\underline{Y}_{v,k} = \sum_{k \in A^g_v} \frac{1}{R_{v,k} + jX_{v,k}},\quad v \neq k,\quad \underline{Y}_{0,v} = \frac{1}{R_{0,v} + jX_{0,v}},$$

where R_m, X_m – active and inductive resistance; B_m – the capacitive conductivity of one branch m from the set of nodes A^b_v connected to node v; $R_{v,k}$, $X_{v,k}$ – active and inductive resistance of the branch between nodes v and k; $R_{0,v}$, $X_{0,v}$ – active and inductive resistance of the branch between the balancing node and node v.

The solution of nonlinear equations system (1) is the complex voltages of all nodes $\dot{U}_v$ ($v = \overline{1,n}$), which can be found by various numerical iterative methods.

The procedure of searching of limiting modes is performed by selecting nodes and incremental step of their power to critical values, beyond which leads to an imbalance of power in the system.

The problem solved in the presented article is to determine the boundary within the region of existence of the steady state of the power system that precedes the emergence of the limiting mode in terms of static stability.

3. CONSUMPTION EQUATIONS FOR A LOAD FED THROUGH A LINE FROM AN INFINITE POWER BUS AND APPLICATION OF TROPICAL MULTIPOLE ANALYSIS TO THIS SYSTEM

First let's consider consumption and the appearance of the limiting (critical) mode in the simplest scheme (Fig. 1). The load $\dot{p} = p^{\text{Re}} + jp^{\text{Im}}$, where R, X – active and inductive resistance of the line,

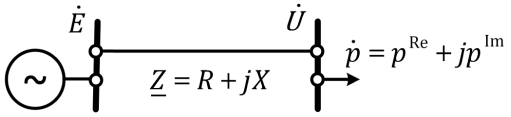


Fig. 1. Power supply circuit through the load line from the infinite power bus.

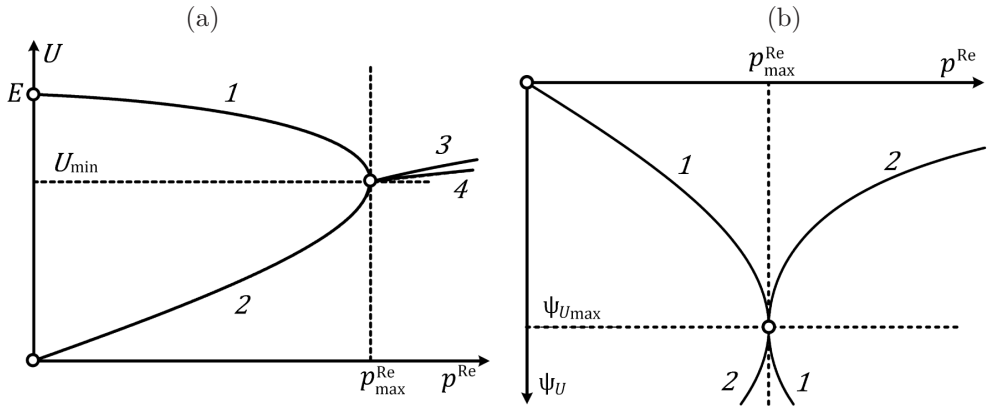


Fig. 2. Dependences of the (a) modulus and (b) angle of the load node voltage on p^{Re}: *1* and *2* – respectively, the first $U_1(p^{\mathrm{Re}})$, $\Psi_{U_1}(p^{\mathrm{Re}})$ and the second $U_2(p^{\mathrm{Re}})$, $\Psi_{U_2}(p^{\mathrm{Re}})$ roots (4) and (5); *3* – $\left|U_{1,2}(p^{\mathrm{Re}})\right|$; *4* – $\mathrm{Re}\left[U_{1,2}(p^{\mathrm{Re}})\right]$.

p^{Re}, p^{Im} – active and reactive power consumption, is connected to a generator (infinite power bus) with a known complex voltage $\dot{E} = Ee^{j\Psi_E}$ through a power line $\underline{Z} = R + jX$.

The equation for determining the steady state of the presented circuit is written in the following form:

$$\dot{U} + \frac{p^{\mathrm{Re}} - jp^{\mathrm{Im}}}{\dot{U}^*}(R + jX) = \dot{E}, \tag{3}$$

in which $\dot{U} = Ue^{j\Psi_U}$, $\dot{U}^* = Ue^{-j\Psi_U}$, where U is the module voltage of the node (bus) with load $\dot{p}$, Ψ_U is the angle of complex voltage $\dot{U}$, measured from the generator voltage vector $\dot{E}$.

Nonlinear equation (3) has an analytical solution, which roots (modules $U_{1,2}(p^{\mathrm{Re}}, p^{\mathrm{Im}})$ and angles (arguments) $\Psi_{U_{1,2}}(p^{\mathrm{Re}}, p^{\mathrm{Im}})$ voltages) are determined according to the expressions:

$$(U_{1,2})^2 = \frac{E^2}{2} - p^{\mathrm{Re}}R - p^{\mathrm{Im}}X \pm \frac{\sqrt{(2(p^{\mathrm{Re}}R + p^{\mathrm{Im}}X) - E^2)^2 - 4(R^2 + X^2)((p^{\mathrm{Re}})^2 + (p^{\mathrm{Im}})^2)}}{2}, \tag{4}$$

$$\Psi_{U_{1,2}} = -\arg\left[(U_{1,2})^2 + p^{\mathrm{Re}}R + p^{\mathrm{Im}}X + j(p^{\mathrm{Re}}X - p^{\mathrm{Im}}R)\right]. \tag{5}$$

Assuming as known and constant power factor ($\cos\phi$) of the node load with an increase in active power p^{Re}, we determine the component $p^{\mathrm{Im}} = p^{\mathrm{Re}}\tan\phi$. Let's build (Fig. 2) dependencies (*2*) and (*3*). The angle $\Psi_{U_{\max}}$ is determined by the expression

$$\Psi_{U_{\max}} = -\mathrm{atan}\left[\frac{X - R\tan\phi}{R + X\tan\phi}\right]. \tag{6}$$

At $p^{\mathrm{Re}} > p^{\mathrm{Re}}_{\max}$ the voltage modulus U takes complex values, and therefore, in the specified range the curves *1* and *2* are missing. The value $p^{\mathrm{Re}}_{\max}$ corresponds to the regime existence boundary. Dependencies *3* and *4* are the modulus and real part of the voltage U at $p^{\mathrm{Re}} > p^{\mathrm{Re}}_{\max}$, respectively.

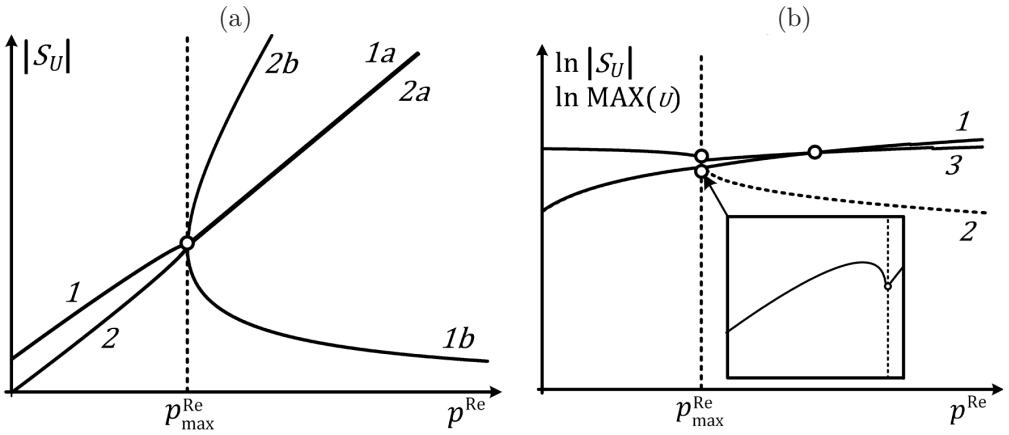


Fig. 3. Dependencies of parameters S_U, $\ln|S_U|$ and $\ln \mathrm{MAX}(U)$ from p^{Re}: on (a) *1* and *2* – $S_U(p^{\mathrm{Re}})$ for the first and second roots, respectively, while *1*, *1a* and *2*, *2a* according to the left side of (7), and *1*, *1b* and *2*, *2b* according to the right side (7); on (b) – for the first roots *1* and *2* – $\ln|S_U|$ and $\ln \mathrm{MAX}(U)$, respectively, according to the data of the left and right sides of (7), *3* – $\ln \mathrm{MAX}(U)$.

For the diagram in Fig. 1, we apply the approach proposed in [8] for determining the proximity of a regime to the region of its existence. The expression for the parameter S_U, written from the condition of the balance of nodes active power, has the form

$$\frac{U^2}{2}\left(\frac{1}{R+jX}+\frac{1}{R-jX}\right)+\mathrm{Re}(\dot{p})=S_U=\frac{1}{2}\left(\frac{\dot{E}\dot{U}^*}{R+jX}+\frac{\dot{E}^*\dot{U}}{R-jX}\right). \tag{7}$$

The dependences of S_U from zero to $p^{\mathrm{Re}}_{\max}$ are the same for the left and right expressions (7) and differ for $p^{\mathrm{Re}} > p^{\mathrm{Re}}_{\max}$ (see Fig. 3). From Fig. 3 it is clear that $\ln|S_U| < \ln\left(\frac{EU}{\sqrt{R^2+X^2}}\right)$ for p^{Re} from zero to $p^{\mathrm{Re}}_{\max}$. Tropical equations used for the diagram in Fig. 1, have the form

$$\ln|S_U| \oplus \ln\left(\frac{EU}{\sqrt{R^2+X^2}}\right). \tag{8}$$

The parameter $\mathrm{MAX}(U)$ denotes the component $\frac{EU}{\sqrt{R^2+X^2}}$. It must be noted that the curve *1* in Fig. 3b (inset) has a maximum near the value $p^{\mathrm{Re}}_{\max}$, explained by the restructuring of the tropical set of solutions, which may be an additional criterion for the proximity of the limiting regime. In this case, in the entire range of p^{Re} from zero to $p^{\mathrm{Re}}_{\max}$ the condition

$$|S_U| \leqslant \max_U\left(\frac{EU}{\sqrt{R^2+X^2}}\right), \tag{9}$$

which is used in [8] to determine the proximity of a regime to the boundary of the region of its existence is not violated. Thus, it is not possible to apply (9), as proposed in [8], to the diagram in Fig. 1 to identify the region preceding the onset of the limiting regime.

4. EXAMPLES OF POWER SYSTEM CALCULATIONS

Let's consider an example of the calculation presented by the authors of [8] to demonstrate their proposed approach. A four-node power system (Fig. 4) with two generators is being studied.

The first generator is specified by an infinite power bus with the known voltage $\dot{U}_1$, the second one – by the voltage module U_3 and active power p^{Re}_3. With this formulation of the problem, it is

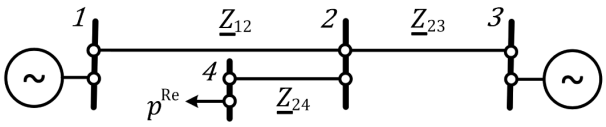


Fig. 4. Calculation scheme of a four-node electrical power system.

impossible to solve the nonlinear equations [14] of the power system state analytically:

$$\dot{U}_2\left(\frac{1}{\underline{z}_{12}}+\frac{1}{\underline{z}_{23}}+\frac{1}{\underline{z}_{24}}\right)-\dot{U}_3\left(\frac{1}{\underline{z}_{23}}\right)-\dot{U}_4\left(\frac{1}{\underline{z}_{24}}\right)=\dot{U}_1\left(\frac{1}{\underline{z}_{12}}\right), \tag{10}$$

$$\mathrm{Re}\left[\dot{U}_3^*\left(\dot{U}_3\frac{1}{\underline{z}_{23}}-\dot{U}_2\frac{1}{\underline{z}_{12}}\right)\right]=\dot{p}_3^{\mathrm{Re}},\quad \left|\dot{U}_3\right|=\mathrm{const},$$

$$\dot{U}_4\frac{1}{\underline{z}_{24}}-\dot{U}_2\frac{1}{\underline{z}_{24}}=-\frac{\dot{p}_4^*}{\dot{U}_4^*},$$

where

— branch resistances $\underline{z}_{12}=8.91+j80.91$; $\underline{z}_{23}=4.45+j40.46$; $\underline{z}_{24}=6.68+j60.68$;

— balancing node voltage $\dot{U}_1=500$;

— active power $p_3^{\mathrm{Re}}=400$ and voltage modulus $U_3=500$ in the third node;

— conjugate complex of fourth node power $\dot{p}_4^*=p_4^{\mathrm{Re}}-jp_4^{\mathrm{Im}}$ for different modes; it must be noted that the condition is accepted $p_4^{\mathrm{Im}}=0$.

The solution of system (10) was carried out in the Mathcad package by the Levenberg–Marquardt method with increasing load in the *4* node until the limiting mode was obtained (Table 1). There was no imbalance in the active power of the power system in all existing modes, i.e. it was equal to zero (with an accuracy of 10^{-307}). At $p_4^{\mathrm{Re}}>p_{4\max}^{\mathrm{Re}}$ the power balance in the power system was disturbed and it could be simply identified by the calculated value of the voltage modulus U_3, which became different from the specified value $U_3=500$ in the indicated modes. It should be noted that in [8] the limiting mode corresponded to the value of 1243.8 obtained at step 10 of Table 1, although the presented calculation results show the limiting mode at 1250.1993. In the case of $p_4^{\mathrm{Re}}=1250.1994$, the voltage U_3, taking into account fifteen decimal places, is 500.000000000000170, and the unbalance is $454.7474\cdot10^{-15}$.

The results of calculation (Table 2) of mode parameters in a logarithmic scale over a complex multipole for each step p_4^{Re} are given. The parameters S_2, S_3 and S_4 are determined from the active power balances of nodes as follows [8]:

$$\frac{1}{2}\left(\frac{\dot{U}_2^*\dot{U}_1}{\underline{z}_{12}}+\frac{\dot{U}_2\dot{U}_1^*}{\underline{z}_{12}^*}+\frac{\dot{U}_2^*\dot{U}_3}{\underline{z}_{23}}+\frac{\dot{U}_2\dot{U}_3^*}{\underline{z}_{23}^*}+\frac{\dot{U}_2^*\dot{U}_4}{\underline{z}_{24}}+\frac{\dot{U}_2\dot{U}_4^*}{\underline{z}_{24}^*}+\right)=S_2,$$

$$\frac{1}{2}\left(\frac{\dot{U}_3^*\dot{U}_2}{\underline{z}_{23}}+\frac{\dot{U}_3\dot{U}_2^*}{\underline{z}_{23}^*}\right)=S_3=\frac{(U_3)^2}{2}\left(\frac{1}{\underline{z}_{23}}+\frac{1}{\underline{z}_{23}^*}\right)-p_3^{\mathrm{Re}},$$

$$\frac{1}{2}\left(\frac{\dot{U}_4^*\dot{U}_2}{\underline{z}_{24}}+\frac{\dot{U}_4\dot{U}_2^*}{\underline{z}_{24}^*}\right)=S_4=\frac{(U_4)^2}{2}\left(\frac{1}{\underline{z}_{24}}+\frac{1}{\underline{z}_{24}^*}\right)-p_4^{\mathrm{Re}},$$

where

$$S_2=\frac{(U_2)^2}{2}\left(\frac{1}{\underline{z}_{12}}+\frac{1}{\underline{z}_{12}^*}+\frac{1}{\underline{z}_{23}}+\frac{1}{\underline{z}_{23}^*}+\frac{1}{\underline{z}_{24}}+\frac{1}{\underline{z}_{24}^*}\right).$$

Table 1. Four-node power system mode parameters

Step	p_4^{Re}	U_2	Ψ_2	Ψ_3	U_4	Ψ_4
0	500	492,5637	−2,0083	1,6966	481,5826	−9,3567
1	600	489,8857	−3,9833	−0,2924	475,4331	−12,9766
2	700	486,5099	−6,0115	−2,3386	467,9762	−16,7639
3	800	482,2942	−8,1121	−4,4620	458,9055	−20,7819
4	900	477,0080	−10,313	−6,6920	447,7206	−25,1289
5	1000	470,2383	−12,6602	−9,0775	433,5139	−29,9775
6	1100	461,0926	−15,2455	−11,7162	414,2925	−35,6971
7	1200	446,5674	−18,3529	−14,9131	383,2047	−43,5358
8	1225	440,4618	−19,3571	−15,9566	369,7892	−46,5104
9	1237,5	436,0835	−19,9756	−16,6041	359,9997	−48,5514
10	1243,8	432,9471	−20,3673	−17,0168	352,8867	−49,9711
11	1245	432,1875	−20,4557	−17,1103	351,1503	−50,3099
12	1250	426,5797	−21,0298	−17,7230	338,1517	−52,7538
13	1250,1993	425,2433	−21,1460	−17,8485	335,0037	−53,3220

Table 2. Mode parameters on a logarithmic scale

step	p_4^{Re}	$\ln \mathrm{MAX}(2)$	$\ln \lvert S_2 \rvert$	$\ln \mathrm{MAX}(3)$	$\ln \lvert S_3 \rvert$	$\ln \mathrm{MAX}(4)$	$\ln \lvert S_4 \rvert$
0	500	8.7079	7.2533	8.7079	5.6038	8.2651	6.8197
1	600	8.7025	7.2424	8.7025	5.6038	8.2468	6.9129
2	700	8.6955	7.2286	8.6955	5.6038	8.2240	6.9963
3	800	8.6868	7.2112	8.6868	5.6038	8.1958	7.0711
4	900	8.6758	7.1891	8.6758	5.6038	8.1601	7.1383
5	1000	8.6615	7.1606	8.6615	5.6038	8.1135	7.1981
6	1100	8.6419	7.1213	8.6419	5.6038	8.0485	7.2497
7	1200	8.6099	7.0573	8.6099	5.6038	7.9385	7.2884
8	1225	8.5961	7.0297	8.5961	5.6038	7.8891	7.2931
9	1237.5	8.5861	7.0097	8.5861	5.6038	7.8523	7.2929
10	1243.8	8.5789	6.9953	8.5789	5.6038	7.8251	7.2910
11	1245	8.5771	6.9918	8.5771	5.6038	7.8184	7.2903
12	1250	8.5641	6.9657	8.5641	5.6038	7.7677	7.2827
13	1250.1993	8.5609	6.9594	8.5609	5.6038	7.7552	7.2803

Parameters $\ln \mathrm{MAX}(2)$, $\ln \mathrm{MAX}(3)$ and $\ln \mathrm{MAX}(4)$ at each step p_4^{Re} corresponded to the same branches and were determined by the formulas:

$$\ln \mathrm{MAX}(2) = \ln \mathrm{MAX}(3) = \left(\frac{U_2 U_3}{|\underline{z}_{23}|} \right),$$

$$\ln \mathrm{MAX}(4) = \left(\frac{U_2 U_4}{|\underline{z}_{24}|} \right).$$

It must be noted that for all nodes v of the power system (Table 2) the condition is satisfied

$$\ln \mathrm{MAX}(v) > \ln |S_v|,$$

and accordingly, given in [8]

$$|S_U| \leqslant \max_k \left(\frac{U_v U_k}{|\underline{z}_{vk}|} \right). \tag{11}$$

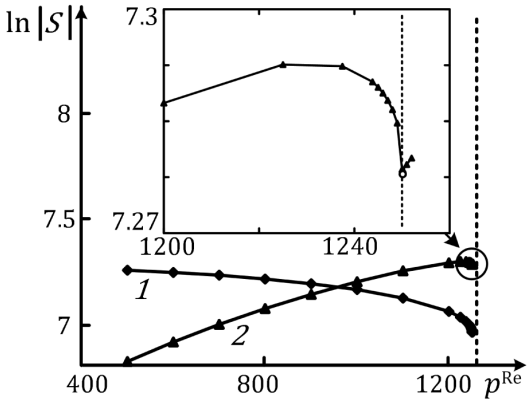


Fig. 5. Parameter dependencies $\ln|S|$ from p^{Re}: *1* – $\ln|S_2|$; *2* – $\ln|S_4|$.

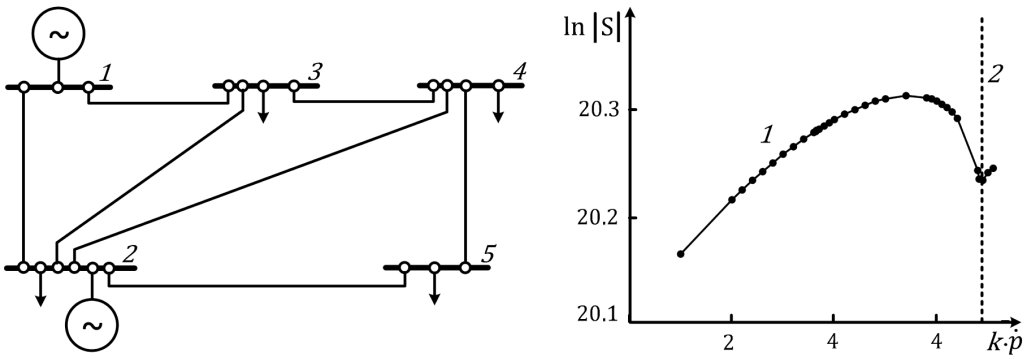


Fig. 6. Scheme *a* IEEE 5 buses and dependencies *b* *1* – $\ln|S_5|$ from $k\sum\dot{p}$; *2* – $\dot{p}_{\max}$.

At the same time, in [8] in the range from $p_4^{\mathrm{Re}} = 1200$ to $p_{4\max}^{\mathrm{Re}}$ (1243.8) condition (11) is violated, that makes it possible to identify a subregion within the region of existence of the regime, the exit of which precedes the limit regime.

It should be noted that compliance with condition (11) in the existence region of the regime is consistent with the qualitative results obtained for the circuit in Fig. 1. In addition, we can segregate the node *4*, in which the dependence $\ln|S_4|$ has a maximum (Fig. 5 and Table 2, cell with fill), similar to that shown in Fig. 3b (see box). To check the conditions (11) and the possibility of using the $\ln|S|$ dependencies when identifying the region preceding the limiting mode, the calculations were carried out on the IEEE 5-bus test circuit (Fig. 6a) with standard initial data presented in Table 3. The parameter values in the table are given in relative units at basic power values of 100 MBA and voltage of 100 kV. The circuit takes into account the capacitive conductivity B of the network lines and a PV-type generator is connected to the *2* bus.

As the power system mode became heavier, the load of all nodes $\sum\dot{p} = \dot{p}_3+\dot{p}_4+\dot{p}_5$ was increased simultaneously, by multiplying by coefficient k. As a result of the calculations, it was established that for all nodes (buses) of the power system, conditions (11) are met, and for node 5 the maximum of dependence $\ln|S_5|$ from $k\sum\dot{p}$, preceding the limiting mode $\dot{p}_{\max}$ (see Fig. 6b) was observed. It should be noted that for arbitrary values of node loads and line resistances conditions (11) are met and there is the indicated maximums.

Studies out of the scalability of the proposed method and criterion on circuits with a large number of buses (standard IEEE 30-bus circuit), in which only PQ nodes of loads and generators were specified, as well as in the presence of PV generators were carried. For all the cases considered, only some (several) nodes connected to generators exhibit growth and maxima in the dependencies $\ln|S|$,

Table 3. IEEE 5-Bus Power System Test Circuit Source Data

Information on network nodes					Information on network branches				
no.	Node type	V, o.e..	p^{Re}, o.e.	p^{Im}, o.e.	Branches		Resistance and conductivity		
							R, o.e.	X, o.e.	B, o.e.
1	Slack bus	1.20	–	–	1	2	0.02	0.06	0.06
					1	3	0.08	0.24	0.05
2	Generator (PV)	1.20	0.40	–	2	3	0.06	0.18	0.04
					2	4	0.06	0.18	0.04
3	Load	–	0.45	0.15	2	5	0.04	0.12	0.03
4	Load	–	0.60	0.10	3	4	0.01	0.03	0.02
5	Load	–	0.40	0.05	4	5	0.08	0.24	0.05

which can be used to determine the approach to the region boundary of the permissible power system modes. For other nodes and branches of the power system, these dependencies decrease with increasing load, as shown in Fig. 5, curve *1*.

Thus, studies show that as the regime becomes heavier, the maximum of dependence $\ln|S|$ for power system nodes can be used to determine the boundary beyond which precedes the regime reaching the boundary of its domain of existence.

5. CONCLUSION

1. It is possible to identify the lack of balance in the power system, and, accordingly, whether the mode under consideration belongs to the region of an unstable state, by monitoring the value of the specified voltage of the PV-type generating unit (bus) during calculations.

2. In the region where the mode exists, conditions (11) are satisfied for each node (bus) of the power system; they cannot be used to identify the subregion, the exit beyond which precedes the limit mode.

3. It is possible to identify the boundary within the existence region of the regime, the exit beyond which precedes the limiting regime, by determining the maximum of the values of $\ln|S|$ node (bus) of the power system. This will make it possible, at a low cost of computational resources, to determine the nodes that are critical in terms of weighting and to obtain additional information to enter the regime into a more stable region.

FUNDING

The work was carried out with the financial support of the Priority 2030 program (grant no. 122060300035-2).

APPENDIX

Expressions (4) and (5) are obtained from (3) as follows:

$$U^2 + \left(p^{\mathrm{Re}} - jp^{\mathrm{Im}}\right)(R + jX) = EUe^{-j\Psi_U},$$
$$U^2 + p^{\mathrm{Re}}R + p^{\mathrm{Im}}X + j\left(p^{\mathrm{Re}}X + p^{\mathrm{Im}}R\right) = EUe^{-j\Psi_U}. \quad \text{(A.1)}$$

The balance of modules of expression (A.1) is reduced to a quadratic equation for the unknown $\hat{U} = U^2$:

$$a\hat{U}^2 + b\hat{U} + c = 0, \tag{A.2}$$

where

$$a = 1, \quad b = 2\left(p^{\mathrm{Re}}R + p^{\mathrm{Im}}X\right) - E^2, \quad c = \left(R^2 + X^2\right)\left[\left(p^{\mathrm{Re}}\right)^2 + \left(p^{\mathrm{Im}}\right)^2\right].$$

The solution to (A.2) is expression (4). The angle Ψ_U in expression (5) is determined by substituting the found expression (4) for U into equation (A.1).

Expression (6) of the article is obtained from the load bus power equation:

$$\dot{p} = p^{\mathrm{Re}} + jp^{\mathrm{Im}} = \dot{U}\left(\frac{\dot{E} - \dot{U}}{R + jX}\right)^*. \tag{A.3}$$

From (A.3) we obtain

$$\frac{p^{\mathrm{Im}}}{p^{\mathrm{Re}}} = \tan\phi = \frac{EX\cos\Psi_U - UX + ER\sin\Psi_U}{ER\cos\Psi_U - UR - EX\sin\Psi_U}. \tag{A.4}$$

Let's express the voltage modulus U of the load bus from (A.4)

$$U = E\left(\cos\Psi_U + \sin\Psi_U\left(\frac{R + X\tan\phi}{X - R\tan\phi}\right)\right)$$

and put it into the expression for active power obtained from (A.3):

$$p^{\mathrm{Re}} = \frac{U}{(R^2 + X^2)}\left[R\left(E\cos\Psi_U - U\right) - EX\sin\Psi_U\right]. \tag{A.5}$$

Taking the derivative of (A.5) with respect to the angle Ψ_U and equating it to zero, we obtain the expression

$$\frac{dp^{\mathrm{Re}}}{d\Psi_U} = \frac{\sin\left(2\Psi_U\right)\left(R + X\tan\phi\right) - \cos\left(2\Psi_U\right)\left(R\tan\phi - X\right)}{\left(X - R\tan\phi\right)^2} = 0 \tag{A.6}$$

from which we determine

$$\tan\left(2\Psi_U\right) = \frac{R\tan\phi - X}{R + X\tan\phi}. \tag{A.7}$$

The resulting expression (A.7) is equivalent to equation (6).

REFERENCES

1. Venikov, V.A., Stroev, V.A., Idelchik, V.I., and Tarasov, V.I., Estimation of Electrical Power System Steady State Stability in Load Flow Calculations, *IEEE Trans. Power App. Syst.* 1975, vol. 94, no. 3, pp. 1034–1041.

2. Dobson, A. and Lu, L., New Methods for Computing a Closest Saddle Node Bifurcation and Worst Case Load Power Margin for Voltage Collapse, *IEEE Trans. Power Syst.* 1993, vol. 8, no. 3, pp. 905–911.

3. Ayuev, B.I., Davydov, V.V., and Erokhin, P.M., Optimization Model of Limit Modes of Electrical Systems, *Elektrichestvo*, 2010, no. 11, pp. 2–12.

4. Voropai, N.I., Golub, I.I., Efimov, D.N., et al., Spectral and Modal Methods for Studying Stability and Control of Electric Power Systems, *Autom. Remote Control*, 2020, vol. 81, no. 10, pp. 1751–1774.

5. Wang, Y., Lopez, J.A., and Sznaier, M., Convex Optimization Approaches to Information Structured Decentralized Control, *IEEE Trans. Autom. Control*, 2018, vol. 63, no. 10, pp. 3393–3403.

6. Matveev, A.S., MacHado, J.E., Ortega, R., et al., Tool for Analysis of Existence of Equilibria and Voltage Stability in Power Systems with Constant Power Loads, *IEEE Trans. Autom. Control*, 2020, vol. 65, no. 11, pp. 4726–4740.

7. Ghiocel, S.G. and Chow J.H., A Power Flow Method Using a New Bus Type for Computing Steady-State Voltage Stability Margins, *IEEE Trans. Power Syst.*, 2014, vol. 29, no. 2, pp. 958–965.

8. Kirshtein, B.K. and Litvinov, G.L., Analyzing Stable Regimes of Electrical Power Systems and Tropical Geometry of Power Balance Equations Over Complex Multifields, *Autom. Remote Control*, 2014, vol. 75, no. 10, pp. 1802–1813.

9. Su, H.Y. and Liu, C.W., Estimating the Voltage Stability Margin Using PMU Measurements, *IEEE Trans. Power Syst.*, 2016, vol. 31, no. 4, pp. 3221–3229.

10. Ayuev, B.I., Davydov, V.V., and Erokhin, P.M., Fast and Reliable Method of Searching Power System Marginal States, *IEEE Trans. Power Syst.*, 2016, vol. 31, no. 6, pp. 4525–4533. https://doi.org/10.1109/TPWRS.2016.2538299

11. Sharov, Ju.V., About Development of Analysis Methods Static Stability of Electric Power Systems, *Elektrichestvo*, 2017, no. 1, pp. 12–18.

12. Sharov, Ju.V., Application Modal Approach for Solving the Problem of Ensuring Power System Static Stability, *Izvestiya RAN. Energetika*, 2017, no 2, pp. 13–29.

13. Rao, S., Tylavsky, D., and Feng, Y., Estimating the Saddle-Node Bifurcation Point of Static Power Systems Using the Holomorphic Embedding Method, *Int. J. Electr. Power Energ. Syst.*, 2017, vol. 84, pp. 1–12.

14. Liu, C., Wang, B., Hu, F., Sun, K., and Bak, C.L., Online Voltage Stability Assessment for Load Areas Based On the Holomorphic Embedding Method, *IEEE Trans. Power Syst.*, 2018, vol. 33, no. 4, pp. 3720–3734.

15. Qiu, Y., Wu, H., Song, Y., and Wang, J., Global Approximation of Static Voltage Stability Region Boundaries Considering Generator Reactive Power Limits, *IEEE Trans. Power Syst.*, 2018, vol. 33, no. 5, pp. 5682–5691.

16. Wang, L. and Chiang, H.D., Group-Based Line Switching for Enhancing Contingency-Constrained Static Voltage Stability, *IEEE Trans. Power Syst.*, 2020, vol. 35, no. 2, pp. 1489–1498.

17. Ali, M., Gryazina, E., Khamisov, O., and Sayfutdinov, T., Online Assessment of Voltage Stability Using Newton-Corrector Algorithm, *IET Generat., Transmiss. Distribut.*, 2020, vol. 14, no. 19, pp. 4207–4216.

18. Bulatov, Yu.N., Kryukov, A.V., Suslov K.V., et al., Timely Determination of Static Stability Margins in Power Supply Systems Equipped with Distributed Generation Installations, *Vestnik Irkut. Gos. Tekh. Univ.*, 2021, vol. 25, no. 1(156), pp. 31–43. https://doi.org/10.21285/1814-3520-2021-1-31-43

19. Bulatov, Y., Kryukov, A., Suslov, K., et al., A Stochastic Model for Determining Static Stability Margins in Electric Power Systems, *Computation*, 2022, vol. 10, no. 5. https://doi.org/10.3390/computation10050067

20. Weng, Y., Yu, S., Dvijotham, K., and Nguyen, H.D., Fixed-Point Theorem-Based Voltage Stability Margin Estimation Techniques for Distribution Systems with Renewables, *IEEE Transact. Industr. Inform.*, 2022, vol. 18, no. 6, pp. 3766–3776. https://doi.org/10.1109/TII.2021.3112097

21. Zhang, W., Wang, T., and Chiang, H.D., A Novel FFHE-Inspired Method for Large Power System Static Stability Computation, *IEEE Trans. Power Syst.*, 2022, vol. 37, no. 1, pp. 726–737. https://doi.org/10.1109/TPWRS.2021.3093236

22. Ali, M., Gryazina, E., Dymarsky, A., and Vorobev, P., Calculating Voltage Feasibility Boundaries for Power System Security Assessment, *Int. J. Electr. Power Energ. Syst.*, 2023, vol. 146, 108739. https://doi.org/10.1016/j.ijepes.2022.108915

23. Ali, M., Ali, M.H., Gryazina, E., and Terzija, V., Calculating Multiple Loadability Points in the Power Flow Solution Space, *Int. J. Electr. Power Energ. Syst.*, 2023, vol. 148, 108915. https://doi.org/10.1016/j.ijepes.2022.108739

24. Machado, J.E., Grino, R., Barabanov, N., et al., On Existence of Equilibria of Multi-Port Linear AC Networks with Constant-Power Loads, *IEEE Transact. Circuits and Systems. Part 1: Regular Papers*, 2017, vol. 64, no. 10, pp. 2772–2782.https://doi.org/10.1109/TCSI.2017.2697906

25. Danilov, M.I. and Romanenko, I.G., Determination of Power Flows and Temperature of Electrical Network Wires of a Power System Steady State, *Power Technol. Engineer.*, 2023, vol. 56, no. 5, pp. 739–750. https://doi.org/10.1007/s10749-023-01583-z

26. Karimi, M., Shahriari, A., Aghamohammadi, M.R., et al., Application of Newton-Based Load Flow Methods for Determining Steady-State Condition of Well and Ill-Conditioned Power Systems: A Review, *Int. J. Electr. Power Energ. Syst.*, 2019, vol. 113, pp. 298–309.

27. Zorin, I.A. and Gryazina, E.N., An Overview of Semidefinite Relaxations for Optimal Power Flow Problem, *Autom. Remote Control*, 2019, vol. 80, no. 5, pp. 813–833. https://doi.org/10.1134/S0005231019050027

28. Danilov, M.I. and Romanenko, I.G., Identification of Unauthorized Electric-Power Consumption in the Phases of Distribution Networks with Automated Metering Systems, *Power Technol. Engineer.*, 2022, vol. 56, no. 3, pp. 414–422. https://doi.org/10.1007/s10749-023-01530-y

29. Danilov, M.I. and Romanenko, I.G., Operational Identification of Resistances of Wires of 380 V Distribution Networks by Automated Accounting Systems, *Energetika. Izv. Vuzov i energ. ob"edinenii SNG*, 2023, vol. 66, no. 2, pp. 124–140. https://doi.org/10.21122/1029-7448-2023-66-2-124-140

30. Bonchuk, I.A., Shaposhnikov, A.P., Sozinov, M.A., and Erokhin, P.M., Optimization of the Operating Modes of Power Plants in Isolated Electrical Power Systems, *Power Technol. Engineer.*, 2021, vol. 55, no. 3, pp. 445–453. https://doi.org/10.1007/s10749-021-01380-6

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, 2024, Vol. 85, No. 1, pp. 85–96.

Convex Isoquants in Dea Models with Selective Convexity

A. P. Afanasiev[*,**,a], **V. E. Krivonozhko**[***,****,b], **A. V. Lychev**[***,c], **and O. V. Sukhoroslov**[*,d]

[*] *Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia*
[**] *RUDN University, Moscow, Russia*
[***] *National University of Science and Technology "MISIS", Moscow, Russia*
[****] *Federal Research Center "Computer Science and Control", Russian Academy of Sciences, Moscow, Russia*
e-mail: [a] *apa@iitp.ru,* [b] *KrivonozhkoVE@mail.ru,* [c] *lychev@misis.ru,* [d] *sukhoroslov@iitp.ru*

Received August 31, 2023
Revised December 22, 2023
Accepted December 30, 2023

Abstract—Models with selective convexity are an important class of data envelopment analysis (DEA) models. This type of model allows managers to consider variables such as ratios, averages, percentages, etc. The paper proposes algorithms for constructing input and output isoquants using volume variables in models with selective convexity. These algorithms help investigate the relationship between any volume variables in the model. Computational experiments confirm the reliability and efficiency of the proposed methods.

Keywords: data envelopment analysis, production possibility set, selective convexity, efficient frontier, isoquant

DOI: 10.31857/S0005117924010075

1. INTRODUCTION

The data envelopment analysis (DEA) approach arose as a generalization of simple indicators of units behavior to a multidimensional case. Mathematically, this approach leads to solving a large family of optimization problems. The founders of this approach were famous American scientists A. Charnes, W. Cooper, E. Rhodes and R. Banker [1, 2]. The FDH (free disposal hull) models appeared almost simultaneously with VRS formulation of DEA in the works of D. Deprins, L. Simar and G. Tulkens [3] in the end of last century. Constraints sets of the DEA models are convex, so optimization methods are widely used for DEA models. Production possibility set of the FDH models are non-convex. For this reason, the development of visualization methods for FDH models slows down.

The notion of selective convexity was proposed in [4]. This notion considers a range of new DEA models, where DEA and FDH models are two extreme cases. Such models expands the possibilities of DEA and FDH models, since problems with selective convexity include such variables into models as ratios, percentages, averages, etc.

The DEA and FDH models aim to develop models and instruments for analyzing the behavior of complex socio-economic systems, such as regions, banks, universities, hospitals, industrial facilities, etc. For developing and applying these models it was necessary to develop new approaches.

Visualization techniques are utilized in various fields of human activity, including the study of the behavior of large-scale socio-economic systems. It enables managers to construct the trajectories

of units' development, to obtain unknown dependencies between model components, detect and correct incorrectness in models, to explore the problem of units' separation and merging, as noted in [5]. In general, visualization enhances a manager's intuition in making strategic decisions.

However, there exist a few works [5–7] in the scientific papers devoted to the visualization of multidimensional production possibility sets and dispositions of production units in such figures. In [7], the methods for multidimensional visualization of convex DEA models were presented. In [5] a review of visualization methods in DEA is presented. Visualization means the construction of intersections of multidimensional polyhedral production possibility set with two- or three-dimensional hyperplanes. This approach reduces the efficiency analysis of production units to the investigation of well-known functions in economics, such as production function, isoquant, isocost, isoprofit, etc. [8, 9].

In paper [10], visualization methods were proposed for models with selective convexity, in which some of the variables are ratios. For such models, solution and visualization methods were proposed and for any two ratio input or output variables. The new methods have shown their efficiency on real-life problems.

Moreover, it was shown in paper [10] that not taking into account specifics of the task leads to significant distortions of the result. In this paper, algorithms are considered for construction of input and output isoquants in models with selective convexity with the use of volume variables.

2. BACKGROUND

Consider a set of production units (X_j, Y_j), $j = 1, \ldots, n$, where the vector of outputs $Y_j = (y_{1j}, \ldots, y_{rj}) \geqslant 0$ is produced from the vector of inputs $X_j = (x_{1j}, \ldots, x_{mj}) \geqslant 0$. All data are assumed to be nonnegative, but at least one component of every input and output vector is positive.

Now consider the notion of selective convexity [4]. Let the input and output sets I and O have the following partition

$$I = I^C \cup I^{NC}, \quad O = O^C \cup O^{NC},$$

where the subsets I^C and I^{NC}, and O^C and O^{NC}, are mutually disjoint.

Subsets I^C and O^C are called the subsets of volume inputs and outputs (volume measures). The complementary subsets $I^{NC} = I \setminus I^C$ and $O^{NC} = O \setminus O^C$ are marked as ratio inputs and outputs (ratio measures).

Let us assume that the set I^C contains the inputs from 1 to m', at the same time the set I^{NC} contains the inputs from $(m'+1)$ to m. Then it is evident that any vector of inputs can be written in the form $X = (X^C, X^{NC})$, where X^C is the vector of the first m' components of X, and X^{NC} is the vector of the last components of X.

In the same way, let us assume that the set O^C contains the output components from 1 to r', and the set O^{NC} contains the output components from $(r'+1)$ to r. Hence any vector of outputs can be written in the form $Y = (Y^C, Y^{NC})$.

The production possibility set T of the technology with selective convexity is determined by the following postulates [4].

(A1) Feasibility of observed data. Unit $(X_j, Y_j) \in T$ for any $j = 1, \ldots, n$.

(A2) Free disposability. $(X, Y) \in T$, and $Y \geqslant Y' \geqslant 0$ and $X' \leqslant X$ implies $(X', Y') \in T$.

(A3) Selective convexity. Let $(X', Y') \in T$ and $(X'', Y'') \in T$. Assume that $(X')_i = (X'')_i$ for all $i \in I^{NC}$, and $(Y')_r = (Y'')_r$ for all $r \in O^{NC}$. Then, for any $\lambda \in [0, 1]$, the unit $\lambda(X', Y') + (1 - \lambda)(X'', Y'') \in T$.

The production possibility set T, which satisfies (A1)–(A3) can be written in the following form:

$$T = \left\{ (X^C, X^{NC}, Y^C, Y^{NC}) \geqslant 0 \;\middle|\; \sum_{j=1}^{n} X_j^C \lambda_j \leqslant X^C,\ \sum_{j=1}^{n} Y_j^C \lambda_j \geqslant Y^C,\ \text{if } \lambda_j \geqslant 0, \right.$$
$$\left. \text{then } X_j^{NC} \leqslant X^{NC} \text{ and } Y_j^{NC} \geqslant Y^{NC},\ \sum_{j=1}^{n} \lambda_j = 1,\ \lambda_j \geqslant 0,\ j = 1, \ldots, n \right\}. \quad (1)$$

The selective convexity model combines two well-known DEA models. So if $I^{NC} = O^{NC} = \emptyset$ (all variables are volume), then set (1) determines the BCC model. If set (1) contains only ratios, i.e., $I^C = O^C = \emptyset$, then the selective convexity model becomes the FDH model.

Podinovski [4] used binary variables δ_j to transform set T to a mixed integer linear constraints. However, for construction of isoquants for variables from I^C and O^C, ratio variables I^{NC} and O^{NC} do not change. So the mixed integer constraints in this case can be replaced by the equivalent constraints $(X_j^{NC} - X^{NC})\lambda_j \leqslant 0$ and $(Y_j^{NC} - Y^{NC})\lambda_j \geqslant 0$; see [11, 12] and Remark 3 in [4].

3. ALGORITHM FOR CONSTRUCTION OF THE INPUT ISOQUANT

Input two-dimensional section of set T for unit (X_o, Y_o) is determined by the following formula

$$I_1(X_o, Y_o) = \{(X, Y) \mid X = X_o + \alpha d_1 + \beta d_2,\ Y = Y_o,\ \alpha, \beta \in E^1\}, \quad (2)$$

where $d_1, d_2 \in E^m$, $(X_o, Y_o) \in T$, vectors d_1 and d_2 are directional vectors, and d_1 is perpendicular to d_2.

Next, define the input two-dimensional isoquant as the intersection of the frontier and two-dimensional plane I_1.

$$\mathrm{Sec}_I(X_o, Y_o) = \{(X, Y) \mid (X, Y) \in \mathrm{WEff}_P T \cap I_1\}, \quad (3)$$

where $\mathrm{WEff}_P T$ is a set of weakly Pareto efficient points of set T.

Output two-dimensional section of set T for unit $(X_o Y_o)$ is written as

$$I_2(X_o, Y_o) = \{(X, Y) \mid X = X_o,\ Y = Y_o + \alpha g_1 + \beta g_2,\ \alpha, \beta \in E^1\}, \quad (4)$$

where $g_1, g_2 \in E^r$, $(X_o, Y_o) \in T$, g_1 is perpendicular to g_2.

Now, define the output two-dimensional isoquant as the intersection of the frontier and two-dimensional plane I_2.

$$\mathrm{Sec}_O(X_o, Y_o) = \{(X, Y) \mid (X, Y) \in \mathrm{WEff}_P T \cap I_2\}. \quad (5)$$

Consider an optimization algorithm for construction of the input isoquant for unit $(X_o Y_o)$. The isoquant is determined by directions $e_p \in E^{m'}$ and $e_s \in E^{m'}$, where e_p and e_s are unity vectors with ones in positions p and s, correspondingly. In addition, the inputs p and s belong to the set I^C.

Algorithm 1 (construction of the input isoquant).

Step 1. Find a leftmost point on the input isoquant going through unit (X_o, Y_o) and associated with directions $e_p \in E^{m'}$ and $e_s \in E^{m'}$.

Step 1a. Solve the following optimization problem.

$$\begin{aligned}
&\max \theta_1\\
&\sum_{j=1}^{n} x_{sj}^{C}\lambda_j + \theta_1 \leqslant x_{so},\\
&\sum_{j=1}^{n} x_{pj}^{C}\lambda_j + \tau_1 \leqslant x_{po},\\
&\sum_{j=1}^{n} x_{ij}^{C}\lambda_j \leqslant x_{io},\ i \neq p, s,\\
&\sum_{j=1}^{n} Y_{j}^{C}\lambda_j \geqslant Y_o\\
&(X_j^{NC} - X_o^{NC})\lambda_j \leqslant 0,\ j = 1, \ldots, n,\\
&(Y_j^{NC} - Y_o^{NC})\lambda_j \geqslant 0,\ j = 1, \ldots, n,\\
&\sum_{j=1}^{n} \lambda_j = 1,\ \lambda_j \geqslant 0,\ j = 1, \ldots, n,
\end{aligned} \tag{6}$$

where τ_1 and θ_1 are free variables.

Step 1b. Let θ_1^* be optimal objective of (6). Solve the following problem.

$$\begin{aligned}
&\max \tau_1\\
&\sum_{j=1}^{n} x_{sj}^{C}\lambda_j + \theta_1^* \leqslant x_{so},\\
&\sum_{j=1}^{n} x_{pj}^{C}\lambda_j + \tau_1 \leqslant x_{po},\\
&\sum_{j=1}^{n} x_{ij}^{C}\lambda_j \leqslant x_{io},\ i \neq p, s,\\
&\sum_{j=1}^{n} Y_{j}^{C}\lambda_j \geqslant Y_o,\\
&(X_j^{NC} - X_o^{NC})\lambda_j \leqslant 0,\ j = 1, \ldots, n,\\
&(Y_j^{NC} - Y_o^{NC})\lambda_j \geqslant 0,\ j = 1, \ldots, n,\\
&\sum_{j=1}^{n} \lambda_j = 1,\ \lambda_j \geqslant 0,\ j = 1, \ldots, n,
\end{aligned} \tag{7}$$

where τ_1 is a free variable.

Let $\tilde{Z}_1^1 = (X_o^C - \theta_1^* e_s - \tau_1^* e_p, X_o^{NC}, Y_o^C, Y_o^{NC})$, where θ_1^* and τ_1^* are optimal objectives of problems (6) and (7), respectively.

Step 2. Find the second point on the input isoquant going through unit (X_o, Y_o) and determined by directions $e_p \in E^{m'}$ and $e_s \in E^{m'}$.

Step 2a. Solve the following optimization problem.

$$\begin{aligned}
&\max \tau_2\\
&\sum_{j=1}^{n} x_{sj}^{C}\lambda_j + \theta_2 \leqslant x_{so},\\
&\sum_{j=1}^{n} x_{pj}^{C}\lambda_j + \tau_2 \leqslant x_{po},\\
&\sum_{j=1}^{n} x_{ij}^{C}\lambda_j \leqslant x_{io},\ i \neq p, s,\\
&\sum_{j=1}^{n} Y_{j}^{C}\lambda_j \geqslant Y_o,\\
&(X_j^{NC} - X_o^{NC})\lambda_j \leqslant 0,\ j = 1, \ldots, n,\\
&(Y_j^{NC} - Y_o^{NC})\lambda_j \geqslant 0,\ j = 1, \ldots, n,\\
&\sum_{j=1}^{n} \lambda_j = 1,\ \lambda_j \geqslant 0,\ j = 1, \ldots, n,
\end{aligned} \tag{8}$$

where τ_2 and θ_2 are free variables.

Step 2b. Let τ_2^* be optimal objective of (8). Solve the following problem.

$$\begin{aligned}
&\max \theta_2\\
&\sum_{j=1}^{n} x_{sj}^{C}\lambda_j + \theta_2 \leqslant x_{so},\\
&\sum_{j=1}^{n} x_{pj}^{C}\lambda_j + \tau_2^* \leqslant x_{po},\\
&\sum_{j=1}^{n} x_{ij}^{C}\lambda_j \leqslant x_{io},\ i \neq p, s,\\
&\sum_{j=1}^{n} Y_{j}^{C}\lambda_j \geqslant Y_o,\\
&(X_j^{NC} - X_o^{NC})\lambda_j \leqslant 0,\ j = 1, \ldots, n,\\
&(Y_j^{NC} - Y_o^{NC})\lambda_j \geqslant 0,\ j = 1, \ldots, n,\\
&\sum_{j=1}^{n} \lambda_j = 1,\ \lambda_j \geqslant 0,\ j = 1, \ldots, n,
\end{aligned} \tag{9}$$

where τ_2 is a free variable.

Let $\tilde{Z}_2^1 = (X_o^C - \theta_2^* e_s - \tau_2^* e_p, X_o^{NC}, Y_o^C, Y_o^{NC})$, where θ_2^* and τ_2^* are optimal objectives of problems (8) and (9), respectively.

Step 3. Set $l := 1$, $k := 1$, $i_1 := 1$, $i_2 := 2$. Create flow F_k^l with points $Z_{i_1}^l = Z_1^1$, $Z_{i_2}^l = Z_2^1$ of production possibility set T. Define set $M = \{Z_1^1, Z_2^1\}$.

Step 4. Perform the following operations. Take any unprocessed flow F_k^l, solve optimization problem of the following type

$$\begin{aligned}
&\max \beta_1\\
&(Z_{i_1}^l + Z_{i_2}^l)/2 + \beta_1 d_1 + \tau d_2 \in T,
\end{aligned} \tag{10}$$

where β_1 and τ are scalar variables, vector d_1 is perpendicular to the vector d_2, it lies in the plane of the section, and is directed to the low left corner of a two-dimensional section, vector $d_2 = Z_{i_1}^l - Z_{i_2}^l$.

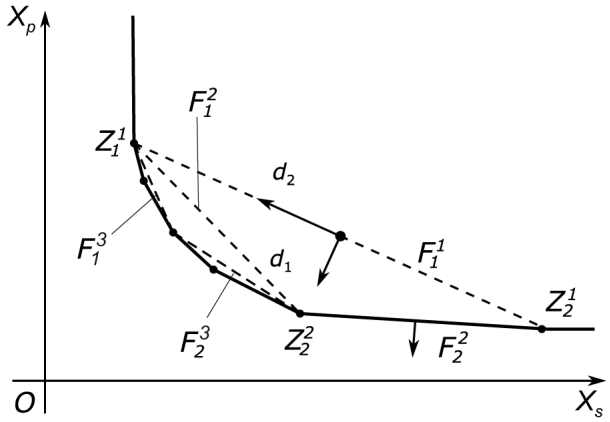


Fig. 1. Construction of input isoquant.

If optimal objective value of problem (10) $\beta_1^* > 0$, then start new flows $F_{k_1}^{l+1}$ and $F_{k_2}^{l+1}$ and solve optimization sub-problems.

Flow $F_{k_1}^{l+1}$ contains points

$$Z_{i_1}^{l+1} = Z_{i_1}^l, \quad Z_{i_2}^{l+1} = (Z_{i_1}^l + Z_{i_2}^l)/2 + \beta_1^* d_1 + \tau^* d_2,$$

where β_1^* and τ^* are optimal values of variables in (10).

Flow $F_{k_2}^{l+1}$ contains points

$$Z_{i_3}^{l+1} = Z_{i_2}^{l+1}, \quad Z_{i_4}^{l+1} = Z_{i_2}^l, \quad d_2 = Z_{i_4}^{l+1} - Z_{i_3}^{l+1},$$

where d_1 is perpendicular to the vector d_2.

If optimal solution value in (10) $\beta_1^* \leqslant 0$, then points $Z_{i_1}^l$ and $Z_{i_2}^l$ are angular points of the segment of input isoquant. Include these points to the set of corner points M. Flow F_k^l is deleted from the list of flow tasks.

Step 5. Set $l{:=}l+1$. If exist unprocessed flows F_k^l, then go to the Step 4, else go to the Step 6.

Step 6. Points of set M are angular points of input isoquant. Connect adjacent pair points by line segments. At last, add a vertical line after the first point, and a horizontal line starting from the last point. This completes construction of the input isoquant.

Figure 1 illustrates the construction of isoquant with the help of the algorithm. At first, the angular points Z_1^1 and Z_2^1 are found by solving models (6)–(9). Then the flow F_1^1 containing these two points is started. At the next steps, point Z_2^2 will be found with the help of solving problem (10); and the flow F_1^1 will be split into two flows F_1^2 and F_2^2, that will have vertices Z_1^1 and Z_2^2 for flow F_1^2 and vertices Z_2^2 and Z_2^1 for flow F_2^2.

After this, the computations are repeated until all segments of the isoquant are found.

For the algorithm presented above the following assertion is valid.

Assertion 1. *Algorithm constructs an input isoquant for production possibility set* (1) *in a finite number of steps.*

Proof. An input isoquant of the two-dimensional set (2) envelops this set or, in other words, it is a boundary of this set. At Steps 1 and 2 the algorithm founds two points Z_1^1 and Z_1^2 of the isoquant and determines segment $[Z_1^1, Z_1^2]$ belonging to the set (2). This initial approximation of the set (3) is found. After this, two optimization problems of the type (10) are solved. If $\beta_1^* > 0$ for at least one of this problem, then the algorithm starts new flows. The approximation of the set (3)

is expanded. If $\beta_1^* \leqslant 0$, then flow F_k^l is deleted from the list of flow tasks. Iterations continued if there exist unprocessed flows. However, all approximations of the set (3) belong to this set and they are expanded during the iterations. The last approximation coincides with set (3). Since the number of boundary segments is finite and the directions of the objective functions differ from each other at every iteration. This completes the proof.

4. ALGORITHM FOR CONSTRUCTION OF THE OUTPUT ISOQUANT

The algorithm for construction of the output isoquant can be written in a similar way. Next, we will focus only on the main differences. Let (X_o, Y_o) be a production unit for which the isoquant is being constructed, and let p and s be two outputs that determined that isoquant. At the first step, we find a rightmost vertex Z_1^1 of isoquant by solving the following optimization problems.

$$\begin{aligned}
&\max \theta_1\\
&\sum_{j=1}^{n} X_j^C \lambda_j \leqslant X_o,\\
&\sum_{j=1}^{n} y_{sj}^C \lambda_j - \theta_1 \geqslant y_{so},\\
&\sum_{j=1}^{n} y_{pj}^C \lambda_j - \tau_1 \geqslant y_{po},\\
&\sum_{j=1}^{n} y_{ij}^C \lambda_j \geqslant y_{io},\ i \neq p, s,\\
&(X_j^{NC} - X_o^{NC})\lambda_j \leqslant 0,\ j = 1, \ldots, n,\\
&(Y_j^{NC} - Y_o^{NC})\lambda_j \geqslant 0,\ j = 1, \ldots, n,\\
&\sum_{j=1}^{n} \lambda_j = 1,\ \lambda_j \geqslant 0,\ j = 1, \ldots, n,
\end{aligned} \tag{11}$$

where τ_1 and θ_1 are free variables.

$$\begin{aligned}
&\max \tau_1\\
&\sum_{j=1}^{n} X_j^C \lambda_j \leqslant X_o,\\
&\sum_{j=1}^{n} y_{sj}^C \lambda_j - \theta_1^* \geqslant y_{so},\\
&\sum_{j=1}^{n} y_{pj}^C \lambda_j - \tau_1 \geqslant y_{po},\\
&\sum_{j=1}^{n} y_{ij}^C \lambda_j \geqslant y_{io},\ i \neq p, s,\\
&(X_j^{NC} - X_o^{NC})\lambda_j \leqslant 0,\ j = 1, \ldots, n,\\
&(Y_j^{NC} - Y_o^{NC})\lambda_j \geqslant 0,\ j = 1, \ldots, n,\\
&\sum_{j=1}^{n} \lambda_j = 1,\ \lambda_j \geqslant 0,\ j = 1, \ldots, n,
\end{aligned} \tag{12}$$

where τ_1 is a free variable.

Point Z_1^1 is expressed as:

$$Z_1^1 = (X_o^C, X_o^{NC}, Y_o^C + \theta_1^* e_s + \tau_1^* e_p, Y_o^{NC}),$$

where $e_p \in E^{r'}$ and $e_s \in E^{r'}$ are direction vectors of isoquant, θ_1^* and τ_1^* are optimal objectives of problems (11) and (12), respectively.

Second vertex Z_2^1 of the output isoquant is determined using following problems.

$$\begin{aligned}
&\max \tau_2 \\
&\sum_{j=1}^{n} X_j^C \lambda_j \leqslant X_o, \\
&\sum_{j=1}^{n} y_{sj}^C \lambda_j - \theta_2 \geqslant y_{so}, \\
&\sum_{j=1}^{n} y_{pj}^C \lambda_j - \tau_2 \geqslant y_{po}, \\
&\sum_{j=1}^{n} y_{ij}^C \lambda_j \geqslant y_{io},\ i \neq p, s, \\
&(X_j^{NC} - X_o^{NC})\lambda_j \leqslant 0,\ j = 1, \ldots, n, \\
&(Y_j^{NC} - Y_o^{NC})\lambda_j \geqslant 0,\ j = 1, \ldots, n, \\
&\sum_{j=1}^{n} \lambda_j = 1,\ \lambda_j \geqslant 0,\ j = 1, \ldots, n,
\end{aligned} \tag{13}$$

where τ_2 and θ_2 are free variables.

$$\begin{aligned}
&\max \theta_2 \\
&\sum_{j=1}^{n} X_j^C \lambda_j \leqslant X_o, \\
&\sum_{j=1}^{n} y_{sj}^C \lambda_j - \theta_2 \geqslant y_{so}, \\
&\sum_{j=1}^{n} y_{pj}^C \lambda_j - \tau_2^* \geqslant y_{po}, \\
&\sum_{j=1}^{n} y_{ij}^C \lambda_j \geqslant y_{io},\ i \neq p, s, \\
&(X_j^{NC} - X_o^{NC})\lambda_j \leqslant 0,\ j = 1, \ldots, n, \\
&(Y_j^{NC} - Y_o^{NC})\lambda_j \geqslant 0,\ j = 1, \ldots, n, \\
&\sum_{j=1}^{n} \lambda_j = 1,\ \lambda_j \geqslant 0,\ j = 1, \ldots, n,
\end{aligned} \tag{14}$$

where τ_2 is a free variable.

Thus we have $Z_2^1 = (X_o^C, X_o^{NC}, Y_o^C + \theta_2^* e_s + \tau_2^* e_p, Y_o^{NC})$, where θ_2^* and τ_2^* are optimal objective values of problems (13) and (14), respectively.

Steps 3–6 of the algorithm for output isoquant coincide with the algorithm for the input isoquant. The only difference is that vector d_1 in model (10) must have positive p and s coordinates to secure the correct shape of the output isoquant.

Assertion 2. *Algorithm constructs an output isoquant for production possibility set* (1) *in a finite number of steps.*

The proof of this assertion is similar to the input isoquant case and hence omitted.

5. COMPUTATIONAL EXPERIMENTS

To perform the computational experiments we use a dataset with artificially generated DMUs. It contains 100 units with 6 variables (3 inputs and 3 outputs). The variables were generated randomly in a range from 5 to 95. Figure 2 shows three isoquants constructed for unit 78 (depicted by point Z_0) using three different models.

Curve *1* corresponds to the isoquant of BCC model, where all variables are from the set $I^C \cup O^C$, i.e., $I^{NC} \cup O^{NC} = \emptyset$. Curve *2* is associated with the model with selective convexity, where all variables are volume except two outputs y_2 and y_3 that are ratio variables. Third model differs from the previous only in inputs x_1 and x_2. In this model they belong to I^{NC}. The isoquant for this model is depicted as curve *3*. Input isoquant for FDH model looks exactly the same as curve *3*; it so happened that the two curves coincided. We see from Fig. 2 that BCC and FDH models are two extreme cases, and curve *2* lies between them. Points Z_1, Z_2, and Z_3 are radial projections of unit Z_0 onto the frontier of models 1, 2, and 3, respectively.

Figure 3 shows three output isoquants constructed for unit 78 (point Z_0 in the figure) using three different models. Curve *1* is associated with the output isoquant of the BCC model. Curve *2* corresponds to the model with one ratio output y_3. Curve *3* is obtained for the model with two ratio variables x_3 and y_3, and the rest are volume. Recall that the distances from the point Z_0 to the points Z_1, Z_2 and Z_3 in relative units are measures of efficiency in models 1, 2 and 3, respectively. This confirms the fact that the choice of the model significantly affects the accuracy of the analysis of the production units' behavior.

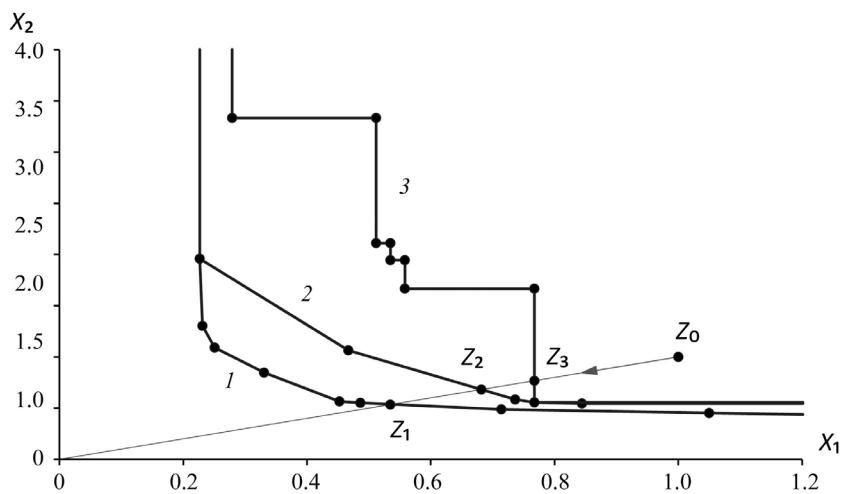


Fig. 2. Input isoquants for BCC model (curve *1*), model with selective convexity (curve *2*), and FDH model (curve *3*) for unit 78.

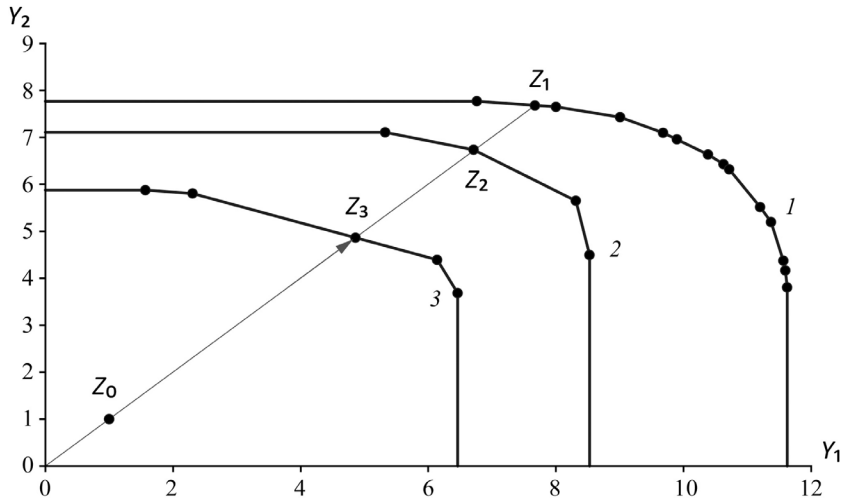


Fig. 3. Output isoquants for BCC model (curve *1*), model with one ratio variable (curve *2*), and model with two ratio variables (curve *3*) for unit 78.

6. CONCLUSION

Visualization plays a huge role in the science and practice of mankind. Indeed, the invention of the telescope by Giordano Bruno at the beginning of the 17th century allowed Newton at the end of this century to discover the laws of planetary motion and formulate as a result world-famous laws, without which it is impossible to create the modern development of science and technology. Visualization methods are used in many areas of human activity, no captain goes on a long trip without detailed maps, no doctor will start operation without a set of patient images, and no engineer will start construction without detailed drawings. However, the leaders of large-scale socio-economic systems often do not have all this instruments and rely on their intuition. However, the cost of an error may be quite huge.

The DEA and FDH technologies do not embrace all possible model cases for production units descriptions. In paper [4], the concept of selective convexity was proposed, which provides the development of a range of new DEA models [13–16], where FDH and DEA models are two extreme cases. Such modifications allow one to explain the class of model's variables and include the averages, percentages, ratios, etc. into DEA models.

In paper [10], algorithms were developed for the construction of input isoquants in DEA models with selective convexity with the use of ratio variables.

In this paper, algorithms are developed for construction of two-dimensional input and output isoquants with the use of volume input and output variables. The proposed algorithm requires considerably fewer computations than the algorithm [10] for ratio variables since it involves only linear problems, whereas the second uses mixed-integer programs.

Computational experiments documented that the proposed algorithms are reliable and efficient. The proposed algorithm allows parallel and distributed implementation similar to the approach proposed in [7]. The development of efficient parallel and distributed implementations [17–19] to speed up computations and conducting computational experiments with large-scale datasets we consider as a direction of our future research.

FUNDING

This work was supported by the Russian Science Foundation, project no. 23-11-00197. https://rscf.ru/en/project/23-11-00197/.

REFERENCES

1. Charnes, A., Cooper, W.W., and Rhodes, E., Measuring the efficiency of decision making units, *Eur. J. Oper. Res.*, 1978, vol. 2, no. 6, pp. 429–444. https://doi.org/10.1016/0377-2217(78)90138-8
2. Banker, R.D., Charnes, A., and Cooper, W.W., Some models for estimating technical and scale efficiency in data envelopment analysis, *Management Sci.*, 1984, vol. 30, no. 9, pp. 1078–1092. https://doi.org/10.1287/mnsc.30.9.1078
3. Deprins, D., Simar, L., and Tulkens, H., Measuring Labor-Efficiency in Post Offices, in *The Performance of Public Enterprises: Concepts and Measurements*, Marchand, M., Pestieau, P., and Tulkens, H., Eds., 1984, Chapter 10, pp. 243–268.
4. Podinovski, V.V., Selective convexity in DEA models, *Eur. J. Oper. Res.*, 2005, vol. 161, no. 2, pp. 552–563. https://doi.org/10.1016/j.ejor.2003.09.008
5. Afanasyev, A.P., Krivonozhko, V.E., Forsund, F.R., and Lychev, A.V., Multidimensional visualization of Data Envelopment Analysis Models, *Data Envelopment Anal. J.*, 2021, vol. 5, no. 2, pp. 339–361. https://doi.org/10.1561/103.00000040
6. Cesaroni, G., Kerstens, K., and Van de Woestyne, I., Global and local scale characteristics in convex and nonconvex nonparametric technologies: A first empirical exploration, *Eur. J. Oper. Res.*, 2017, vol. 259, no. 2, pp. 576–586. https://doi.org/10.1016/j.ejor.2016.10.030
7. Afanasiev, A.P., Krivonozhko, V.E., Lychev, A.V., and Sukhoroslov, O.V., Multidimensional frontier visualization based on optimization methods using parallel computations, *J. Global. Optim.*, 2020, vol. 76, pp. 563–574. https://doi.org/10.1007/s10898-019-00812-y
8. Krivonozhko, V.E., Utkin, O.B., Volodin, A.V., Sablin, I.A., and Patrin, M.V., Constructions of economic functions and calculations of marginal rates in DEA using parametric optimization methods, *J. Oper. Res. Soc.*, 2004, vol. 55, no. 10, pp. 1049–1058. https://doi.org/10.1057/palgrave.jors.2601759
9. Varian, H.R., *Intermediate Microeconomics, a Modern Approach*, 8th ed., New York: W.W. Norton, 2010. ISBN: 978-0-393-93424-3
10. Afanasyev, A.P., Krivonozhko, V.E., Lychev, A.V., and Sukhoroslov, O.V., Constructions of input and output isoquants in DEA models with selective convexity, *Appl. Comput. Math.*, 2022, vol. 21, no. 3, pp. 317–328. https://doi.org/10.30546/1683-6154.21.3.2022.317
11. Kuosmanen, T., DEA with efficiency classification preserving conditional convexity, *Eur. J. Oper. Res.*, 2001, vol. 132, no. 2, pp. 326–342. https://doi.org/10.1016/S0377-2217(00)00155-7
12. Dekker, D. and Post, T., A quasi-concave DEA model with an application for branch performance evaluation, *Eur. J. Oper. Res.*, 2001, vol. 132, no. 2, pp. 296–311. https://doi.org/10.1016/S0377-2217(00)00153-3
13. Olesen, O.B., Petersen, N.C., and Podinovski, V.V., Efficiency analysis with ratio measures, *Eur. J. Oper. Res.*, 2015, vol. 245, no. 2, pp. 446–462. https://doi.org/10.1016/j.ejor.2015.03.013
14. Olesen, O.B., Petersen, N.C., and Podinovski, V.V., Efficiency measures and computational approaches for data envelopment analysis models with ratio inputs and outputs, *Eur. J. Oper. Res.*, 2017, vol. 261, no. 2, pp. 640–655. https://doi.org/10.1016/j.ejor.2017.02.021
15. Olesen, O.B., Petersen, N.C., and Podinovski, V.V., The structure of production technologies with ratio inputs and outputs, *J. Prod. Anal.*, 2022, vol. 57, pp. 255–267. https://doi.org/10.1007/s11123-022-00631-6
16. Olesen, O.B., Petersen, N.C., and Podinovski, V.V., Scale characteristics of variable returns-to-scale production technologies with ratio inputs and outputs, *Annals Oper. Res.*, 2022, vol. 318, pp. 383–423. https://doi.org/10.1007/s10479-022-04862-6
17. Smirnov, S., Voloshinov, V., and Sukhosroslov, O., Distributed Optimization on the Base of AMPL Modeling Language and Everest Platform, *Procedia Comput. Sci.*, 2016, vol. 101, pp. 313–322. https://doi.org/10.1016/j.procs.2016.11.037

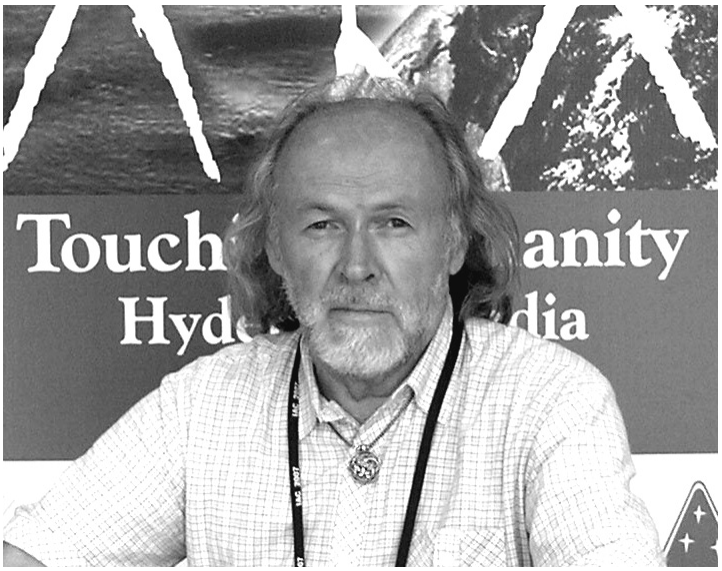
18. Sukhoroslov, O., Volkov, S., and Afanasiev, A., A web-based platform for publication and distributed execution of computing applications, *14th International Symposium on Parallel and Distributed Computing.*, 2015, pp. 175–184. https://doi.org/10.1109/ISPDC.2015.27

19. Sukhoroslov, O., Voloshinov, V., and Smirnov, S., Running Many-Task Applications Across Multiple Resources with Everest Platform, in *Supercomputing. RuSCDays 2020*, Voevodin, V. and Sobolev, S., Eds., *Commun. Comput. Inform. Sci.*, 2020, vol. 1331, pp. 634–646. https://doi.org/10.1007/978-3-030-64616-5_54

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 1, pp. 97–98.

OBITUARY



Mikhail M. Khrustalev
(1938–2023)

Doctor of Physics and Mathematics, Professor Mikhail Khrustalev passed away on August 11, 2023, at the age of 86. He was an outstanding researcher, a wise teacher, and a reliable friend.

Mikhail Khrustalev, in full Mikhail Mikhailovich Khrustalev, was born on May 2, 1938, in Vologda. In 1963, he graduated from Kazan Aviation Institute with a degree in the dynamics of aircraft. Mikhail began his career in the ballistics department of V.N. Chelomei's Design Bureau (nowadays Military-Industrial Corporation "Research and Industrial Association of Machine Building" (MIC NPO Mashinostroyenia), Reutov). In 1970, Khrustalev defended his candidate's dissertation at Moscow State University. Mikhail's doctoral dissertation was defended in a specialized council of Moscow Aviation Institute (MAI) in 1984. Three years later, he was conferred the title of professor.

A lecturer of the highest level and a scientist in the field of optimal control, Khrustalev taught MAI's students to mathematical analysis, differential equations, and the theory of functions of a complex variable as well as a course on the modern theory of optimal control. He was able to explain the most complex mathematical constructs simply and understandably, which was appreciated by both students and colleagues: Mikhail thought in terms of the world known to him and was able to explain it to all his friends. He was a lifelong learner of new things and perfectly selfless in sharing his knowledge. Many young researchers became candidates of physics and mathematics under Khrustalev's supervision. It was a good school of attitude to life.

In addition to teaching, Mikhail devoted much time to organizational and administrative work. In different years, he served as Deputy Director for Science at the Moscow branch of the Institute of Transport Problems, the Russian Academy of Sciences (RAS), and then at the Research Center for Stability and Nonlinear Dynamics, Mechanical Engineering Research Institute RAS. Simultaneously, he worked at the Institute of Applied Mechanics and Electrodynamics (MAI) to create mathematical models of working fluid flows in plasma engines. Numerous colleagues and friends have recognized Mikhail as "one of the few who equally well understands both mathematics and the way aircraft moves."

Khrustalev's scientific accomplishments are quite extensive. He proposed and rigorously justified sufficient and necessary conditions of global optimality for systems described by ordinary differential equations. In particular, these conditions are applicable to optimal control problems with state constraints. His works formulated global optimality conditions for stochastic diffusion systems with incomplete information about the state, as well as conditions of Nash equilibrium in stochastic differential n-person games.

Mikhail considered necessary and sufficient conditions of terminal invariance to be one of his most striking results. Even when working in the ballistics department, he noticed that this problem has a much higher applied significance compared to the classical invariance problem: it admits a solution much more often and, as a rule, essentially nonunique ones. Khrustalev proposed to use the available freedom of choice in terminally invariant control for the parallel solution of additional problems regularly encountered in practice. In particular, he introduced the concept of absolute invariance as the property where the terminal criterion is independent of both current disturbances and the initial state of the system. He formulated sufficient conditions for this problem.

Khrustalev always maintained close ties with the Institute of Control Sciences (ICS), especially with the Laboratory of Optimal Controlled Systems headed by V.F. Krotov, at whose invitation he came to the Institute in 2014. After Krotov's decease, Mikhail headed the Laboratory from 2015 to 2019. Led by him, employees of the Laboratory obtained necessary optimality conditions and effective numerical optimization algorithms for control processes of nonlinear stochastic diffusion systems and jump diffusion systems on finite and infinite horizons.

During his work at ICS RAS, Khrustalev returned to terminal invariance and advanced brilliantly in this area of research. In particular, he posed a new problem of terminal invariance for stochastic diffusion systems and jump diffusion systems and established sufficient conditions of terminal invariance for both classes of systems.

Khrustalev's last scientific results were connected with the development of Krotov's theory of space-time continuum, a generalization extending the well-known Einstein's general relativity theory and the Poincaré gauge theory of gravity. He investigated the elastic properties of the space continuum and proposed a space-time analog of the Hubble redshift theory and the hypothesis of the distribution of dark matter across the universe. In Khrustalev's theory, the effects attributed to dark matter and dark energy arise by themselves due to time deformation.

Mikhail actively practiced yoga and was well-versed in Eastern philosophy and religion. As destined in his well-studied Buddhism, Mikhail will stay with us, passing into the next form of existence of an enlightened person. He was interesting, non-conflicted, able to defend his point of view, persistent, and hardworking; appreciated simplicity and beauty, and understood and accepted the complexity of our unimaginable world. A pleasant, warm, and peaceful kindness always emanated from that marvelous man. He was as pure as rock crystal, *khrustal'* in Russian. The brightest feelings about him will remain in our memory.

A.S. Agapova (Cand. Sci. (Phys.–Math.), ICS RAS),
A.V. Arutyunov (Dr. Sci. (Phys.–Math.), ICS RAS),
S.N. Vassilyev (Academician of RAS, ICS RAS),
A.A. Galyaev (Corresponding Member of RAS, ICS RAS),
D.A. Novikov (Academician of RAS, Director of ICS RAS),
E.E. Onegin (Cand. Sci. (Phys.–Math.)),
D.S. Rumyantsev (Cand. Sci. (Phys.–Math.), ICS RAS),
N.B. Filimonov (Dr. Sci. (Eng.), ICS RAS),
and *K.A. Tsarkov* (Cand. Sci. (Phys.–Math.), ICS RAS).