Volume 87, Number 2
February 2026

ISSN 0005-1179
CODEN: AURCAT

A JOURNAL OF RUSSIAN ACADEMY OF SCIENCES

AUTOMATION AND REMOTE CONTROL

Automation and Remote Control

Editor-in-Chief
Andrey A. Galyaev

Vol. 87, No. 2, February 2026

http://ait.mtas.ru

Automation and Remote Control

ISSN 0005-1179

SCOPE

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

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

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

0005-1179/26. *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 87 (12 issues) is published in 2026.
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. 87, No. 2, 2026

Reviews

Linear Systems

Nonlinear Systems

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2026, Vol. 87, No. 2, pp. 97–140.

REVIEWS

Overview on Observer Design Principles for Systems with Unknown Parameters

A. I. Glushchenko[*,a] **and K. A. Lastochkin**[*,b]

[*] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
e-mail: [a] *aiglush@ipu.ru,* [b] *lastconst@yandex.ru*

Received September 7, 2024
Revised July 11, 2025
Accepted July 15, 2025

Abstract—The overview provides a comparative analysis of the feasibility conditions for key methods to design state observers for linear dynamical SISO systems with unknown parameters. A distinctive feature of this paper is the comparison between state observers that are based on the robust, invariant, and adaptive methods. The overview is written without excessive mathematical details and aims to familiarize a wide range of readers with the basic principles of design and functioning of state observers for linear systems with unknown parameters.

Keywords: state observers, parametric uncertainty, robustness, invariance, adaptivity

DOI: 10.7868/S1608303226020012

1. INTRODUCTION

State observers are special software implementable algorithms that reconstructs unmeasurable system states at any given time instant on the basis of available *a priori* information and online measurements of system input and output. The invention and development of state observers is a response of the mathematical community to the desire of engineers to increase reliability and simplify the design of measurement systems as an essential part of automated control systems. Observers have been most widely applied in the field of automated electric drives design, where today most control systems for synchronous and asynchronous drives are implemented by means of sensorless control [1]. Such success could not have been achieved without a deep theoretical investigation of the problem of state estimation.

The mathematical theory of observer design originated from the studies of D. Luenberger [2] and R. Kalman [3], who respectively stated deterministic and stochastic estimation problems. By now, these two parallel tasks have received an incredible number of extensions and generalizations in both continuous and discrete time. This circumstance forces us in this paper to limit ourselves to consideration and analysis of only a small part of the available results and developments. For interested readers, who would like to engage deeply into the topic and cover the problem more broadly, we recommend to refer to the reviews [4, 5] and books [6–12].

In this paper, the simplest case of a linear dynamical system with one input and one output is considered, and the task is to reconstruct the state of such a system in its original form (physical state) rather than the virtual state of its canonical forms. If the matrices of such a system are precisely known in advance, and the observability condition is met, then the Luenberger observer [2] is a solution to the estimation problem and allows one to reconstruct unmeasurable state with a given exponential convergence rate. However, how to solve the problem if the matrices of the system are unknown or not known precisely enough? The theories of robust [8–10], invariant [7, 8, 11] and adaptive [12] control are simultaneously looking for an answer to this question.[1]

Robust observers. This group mainly concerns the improvement of (and development of new) offline methods to calculate a correction gain for the basic structure of the Luenberger observer. Various design methods are being developed and improved, which (*i*) at the stage of observer parameters computation, provide guarantees of state estimation quality for the parametrically uncertain systems and additionally (*ii*) solve some optimization problems ($\mathcal{H}_2$, $\mathcal{H}_\infty$ norms of the transfer function, the quadratic functional for a linear-quadratic problem, the "size" of an invariant (or attractive) ellipsoid for problems related to the external disturbances attenuation, etc.). The general limitations of robust methods are that *i*) the amplitude of the disturbance and the range of parametric uncertainty are required to be *a priori* known, *ii*) the parametric uncertainty of the system is to satisfy special parameterizations. The approaches of this group have been developed at different times by following foreign and Russian researchers: S. Bhattacharyya, J. Doyle, K. Zhou, H. Kwakernaak, I. Petersen, B. Barmish, R. Tempo, W. Schmitendorf, F. Jabbari, D. Bernstein, M. Corless, C.H. Lien, C.E. de Souza, J.C. Geromel, D. Peaucelle, B.T. Polyak, M.V. Khlebnikov, P.S. Shcherbakov, I.G. Vladimirov, A.P. Kurdyukov, D.V. Balandin, M.M. Kogan, A.G. Alexandrov, V.N. Chestnov, S.K. Korovin, V.V. Fomichev, A.V. Ushakov, R.O. Omorov, etc.

Estimation on the basis of invariance theory. The essence of the methods from this group is the idea to ensure invariance of state reconstruction with respect to the parametric and/or additive disturbances. Invariance is achieved by algebraic elimination of a generalized disturbance from the state reconstruction error equation, or by its high-gain attenuation via choice of correction gain from the class of functions with high-gain coefficients or discontinuous signals. The feasibility conditions of observers based on the invariance theory are, firstly, knowledge of the system uncertainty range (to get to a sliding surface), and secondly, the fulfillment of strict output matching conditions (to ensure existence of such sliding surface). The application of invariance theory methods to the problem of state reconstruction is primarily related to C. Edwards, S. K. Spurgeon, B. L. Walcott, T. Floquet, J.-P. Barbot, M. Darouach, P. Kudva, S. Zak, A. Levant, L. Fridman, Y. Shtessel, K. Khalil, J. Slotine, V.I. Utkin, A.S. Poznyak, S.A. Krasnova, V.A. Utkin, A.N. Zhirabok, etc.

Adaptive observers. Observers from this group simultaneously estimate unmeasurable state and system unknown parameters. Owing to this essentially simple idea, the *a priori* known range of parametric uncertainty is not required for the observer design. Instead, classical adaptive observers require to meet the condition of persistent excitation of some function from the measured signals (control and output) for the asymptotic convergence of the state estimates to their true value. For a long period of time, this condition stymied practical interest in adaptive state reconstruction methods, since, according to various interpretations, such condition is equivalent to (*i*) the global sufficient richness of a control signal of such order that is equal to the dynamic order of the system, (*ii*) the complete observability of the extended system (states + unknown parameters). By the efforts of G. Chowdharry, R. Ortega, Y. Pan, S.B. Roy, A.A. Bobtsov, S.V. Aranovskiy and many other researchers, this condition has been relaxed to the requirement of a regressor finite excitation, which is equivalent to the sufficient richness of the control signal over some finite time interval or the local complete observability of the extended system. The development of methods to design adaptive state observers is closely related to studies by K. Narendra, G. Luders, R. Carroll, D. Lindorf, G. Kreisselmeier, A. Annaswamy, P. Ioannou, R. Marino, P. Tomei, A. Isidori, R. Ortega, S.B. Roy, D. Efimov, A.A. Bobtsov, V.O. Nikiforov, A.A. Pyrkin, S.V. Aranovskiy, N.N. Karabutov, etc.

The aim of this study is to compare the feasibility conditions of state observers designed on the basis of these three theories. In authors' opinion, the relevance of such a comparative analysis is caused by two circumstances. First, the application of observers for practical scenarios requires an

[1] Here the books are cited that summarize and systematize the main results in the mentioned fields of control theory.

engineer to choose the algorithm that best suits the conditions of the applied problem being solved. This, in turn, requires such engineer to know the basic methods to design state observers and the conditions of their applicability. This overview provides such information in a compact form, which makes it possible to familiarize oneself with the main existing approaches. Secondly, researchers interested in one theory (robust, invariant, or adaptive) do not always, but very often speak different languages with colleagues who solve the same problem, but use methods from another theory. The authors hope that this overview will help to improve the situation and achieve greater mutual understanding between these groups, which in the future will result in the enrichment of all three theories with new approaches.

Of course, it is impossible to provide an exhaustive overview, so we would like to note that this paper discusses the results that seem to be the most interesting from the subjective point of view of the authors and mainly relate to the task of physical rather than virtual state reconstruction. At the same time, not only the authors do not disclaim responsibility to *i*) the readers for subjectivity and incompleteness of information, *ii*) the researchers for the possible lack of references to their papers on the subject under consideration, but also hope to receive feedback, which will undoubtedly be of use.

The overview has the following structure. The second section show the relevance of the problem of the state vector reconstruction from the point of view of the stabilization task for a linear time-invariant system. Sections 3–5 discuss in detail the existing approaches to design observers using robust, invariant, and adaptive control methods, respectively. The sixth section is devoted to the analysis of the applicability of the state estimate obtained with the help of the overviewed observers for the purposes of feedback control. The review is wrapped up with the general conclusions presented in Section 7. The research interest of the authors is related to adaptive control theory, therefore, in the overview, more attention is paid to the methods of adaptive observers design.

Further the following notation is used: $|.|$ is the absolute value, $\|.\|$ is the suitable vector or matrix norm of $(.)$, $I_{n\times n} = I_n$ is an identity $n\times n$ matrix, $0_{n\times n}$ is a zero $n\times n$ matrix, 0_n stands for a zero vector of length n, $\det\{.\}$ stands for a matrix determinant, $\mathrm{adj}\{.\}$ represents an adjoint matrix, $\mathrm{tr}\{.\}$ is a matrix trace, $\mathrm{rank}(.)$ denotes a matrix rank, $\dim(.)$ stands for a dimension of a vector or matrix, $\sigma\{.\}$ is a matrix spectrum, $\mathrm{sgn}(.)$ denotes a sign function, $\ln(.)$ stands for a natural logarithm, $\mathrm{vec}\,(.)$ is the operation of a matrix vectorization, $\mathrm{mat}(.)$ stands for an operation that is inverse to the vectorization, $\mathrm{diag}\{.\}$ denotes a diagonal matrix. The matrix $A\in\mathbb{R}^{n\times n}$ minimum and maximum eigenvalues are denoted as $\lambda_{\min}(A) = \min\limits_i(\mathrm{Re}\{\lambda_i(A)\})$ and $\lambda_{\max}(A) = \max\limits_i(\mathrm{Re}\{\lambda_i(A)\})$, respectively, where $i = 1, 2, \ldots, n$. The Latin abbreviation exp is used to denote the exponential convergence rate. For a mapping $\mathcal{F}:\mathbb{R}^n\mapsto\mathbb{R}^n$ we denote its Jacobian by $\nabla_x\mathcal{F}(x) = \frac{\partial\mathcal{F}}{\partial x}(x)$. We also use the fact that for all (possibly singular) matrices M of dimension $n\times n$ the following holds: $\mathrm{adj}\{M\}M = \det\{M\}I_{n\times n}$, and $f\in L_q$ means $\sqrt[q]{\int_{t_0}^{t}\|f(s)\|^q ds} < \infty$ for all $t\geq t_0$.

2. STATE ESTIMATION TASK FROM PERSPECTIVE OF CONTROL PROBLEM

Many real-world technical systems can be described with sufficient accuracy by linear models with time-invariant parameters (single input–single output systems are considered hereafter):

$$\begin{gathered}\dot{x}(t) = Ax(t) + Bu(t), \quad x(t_0) = x_0,\\ y(t) = C^{\mathrm{T}}x(t),\end{gathered} \tag{2.1}$$

where $t\geq t_0\geq 0$ stands for time, $x(t)\in\mathbb{R}^n$ is a vector of unmeasurable physical states with unknown initial conditions x_0, $u(t)\in\mathbb{R}$ denotes a measurable signal or controller output, $A\in\mathbb{R}^{n\times n}$, $B\in\mathbb{R}^{n\times 1}$ are unknown matrix and vector, $C\in\mathbb{R}^n$ stands for a known vector that forms a measurable output $y(t) \in \mathbb{R}$. The pairs $\left(C^{\mathrm{T}}, A\right)$ and (A, B) are completely observable and controllable, respectively (definitions and criteria of controllability and observability can be found in [6, 23]).

One of the basic problems of control theory is to solve a stabilization task, that is, to ensure convergence of the linear system states from arbitrary bounded initial condition x_0 to the origin. If the pair of matrices (A, B) is known and controllable, then this task can be solved using a state-feedback control law:

$$u(t) = -Kx(t), \tag{2.2}$$

where the row-vector $K \in \mathbb{R}^{1\times n}$ is chosen so that to satisfy $\lambda_{\max}(A - BK) < 0$ and can be calculated, for example, via application of a pole placement method [13, 14]:

$$\begin{gathered} AM - MA_{ref} = Bh, \\ K = hM^{-1}, \end{gathered} \tag{2.3}$$

where the matrix A_{ref} and vector h are chosen so that $\sigma\{A\} \cap \sigma\{A_{ref}\} = \emptyset$, and the pair (h, A_{ref}) is observable.

However, since, according to the problem statement, the states of the system are unmeasurable, and the parameters of the matrices A and B are unknown, then the feedback (2.2) cannot be implemented, and the task to design dynamic feedback naturally takes the following forms:

$$u(t) = -K_r\hat{x}(t), \tag{2.4a}$$

$$u(t) = -\hat{K}(t)\hat{x}(t), \tag{2.4b}$$

where:

(2.4a) is a robust dynamic feedback (K_r ensures $\lambda_{\max}(A - BK_r) < 0$ for all matrices A and B that belong to some known sets);

(2.4b) denotes an adaptive dynamic feedback (the adaptive laws for $\hat{K}(t)$ and $\hat{x}(t)$ jointly ensure asymptotic stability of the system (2.1) origin). In a particular case, $\hat{K}(t)$ can be an estimate of the parameters K of a control law (2.2), (2.3).

In order to implement the feedback laws (2.4a) and (2.4b), the estimate of the system (2.1) state is required to satisfy the following:

$$\lim_{t\to\infty} \|\tilde{x}(t)\| = \lim_{t\to\infty} \|\hat{x}(t) - x(t)\| = 0, \tag{2.5}$$

where $\hat{x}(t)$ is a system state estimate, $\tilde{x}(t)$ is a error of state reconstruction/observation/estimation.

Further, the overview focuses on methods to solve the problem (2.5) using the results of the theories of robust, invariant and adaptive control, while the issues of adaptive law design for $\hat{K}(t)$ and equations (procedures) to calculate the parameters K_r are almost everywhere stay beyond the scope of the overview.

In this regard, in order to correctly analyze methods from literature, the following classical assumption is adopted.

Assumption 1. The control signal $u(t)$ is bounded and ensures boundedness of the states $x(t)$ for all time instants.

Remark 1. In addition to the problem of system stabilization via dynamic state-feedback control law discussed in this section, for which, as mentioned earlier, estimates of virtual states of the observer canonical form may be sufficient instead of $\hat{x}(t)$, practice is replete with other tasks that require to estimate the vector of unmeasurable physical states $x(t)$ of the system. For example, it is fault detection, technological processes monitoring and logging of unmeasurable signals, design of digital twins and other applied tasks. Therefore, the goal of state observer design is defined in the form (2.5).

3. ROBUST OBSERVERS

A conventional Luenberger observer is described as follows:

$$\dot{\hat{x}}(t) = A_n\hat{x}(t) + B_n u(t) + L\left(y(t) - C^{\mathrm{T}}\hat{x}(t)\right), \quad \hat{x}(t_0) = \hat{x}_0, \tag{3.1}$$

where $A_n \in \mathbb{R}^{n\times n}$ is a matrix of nominal system parameters, $B_n \in \mathbb{R}^{n\times 1}$ stands for a nominal input matrix, $L \in \mathbb{R}^{n\times 1}$ denotes a gain vector of Luenberger correction term, the pair (C^{T}, A_n) is observable.

If it is true for the system (2.1) that $A = A_n$, $B = B_n$, then, as it results from the following equation (in this section we redefine $e(t) = -\tilde{x}(t)$):

$$\dot{e}(t) = \left(A_n - LC^{\mathrm{T}}\right)e(t), \quad e(t_0) = x_0 - \hat{x}_0, \tag{3.2}$$

the observer (3.1) is a solution of the stated problem (2.5) in case $A_n - LC^{\mathrm{T}}$ is a Hurwitz matrix.

If $A \neq A_n$ and/or $B \neq B_n$, then, instead of (3.2), we have:

$$\begin{aligned}\dot{e}(t) &= Ax(t) + Bu(t) - A_n\hat{x}(t) - B_n u(t) - L\left(y(t) - C^{\mathrm{T}}\hat{x}(t)\right) \pm A_n x(t)\\ &= \left(A_n - LC^{\mathrm{T}}\right)e(t) + (B - B_n)u(t) + (A - A_n)x(t).\end{aligned} \tag{3.3}$$

Robust state observers are based on the idea of ensuring that the goal (2.5) is met not for a single pair of known matrices A_n, B_n, but for some predefined and known set of matrices. From the point of view of the system (3.3) stability, two different situations can be singled out.

Filtering problem. Assumption 1 is met, the signal $u(t)$ is measurable, but not allowed to be chosen by designer/engineer. Under these conditions, the observer (3.1) does not allow one to solve the problem (2.5) without additional restrictions imposed on the class of signal $u(t)$. If $u \in L_2$, then the goal (2.5) can be achieved and additionally we can state the task of $\mathcal{H}_\infty$ filtering, *i.e.*, to ensure a given $\gamma > 0$ ratio of L_2 norm of the error $e(t)$ to L_2 norm of the signal $u(t)$:

$$\sqrt{\int_{t_0}^{\infty} e^2(s)\,ds} \leq \gamma\sqrt{\int_{t_0}^{\infty} u^2(s)\,ds}. \tag{3.4}$$

If $u \in L_\infty$ and A is a Hurwitz matrix, then (3.1) is able to ensure only convergence of the estimation error into a certain bounded set of the phase space, and the best we can do is to state and solve a problem of such set minimization in a certain sense.

Linear dynamic controller design problem. Control signal is a linear function of the observer states, that is, $u(t) = -K_r\hat{x}(t)$. The feedback parameters K_r, along with the correction gain vector L can be chosen in the course of the observer design. In this situation, by joint calculation of K_r and L, the goal (2.5) can be achieved.

In order to use methods of robust theory to solve these problems, the parametric uncertainties in A and B matrices are parametrized into special forms. For example, a structured form is often used (such type of perturbations are also called norm-bounded) [15]:

$$A = A_n + F_A\Delta_A H_A, \quad B = B_n + F_B\Delta_B H_B, \tag{3.5}$$

where matrix uncertainties $\Delta_A \in \mathbb{R}^{p_A\times q_A}$ and $\Delta_B \in \mathbb{R}^{p_B\times q_B}$ satisfy the inequalities

$$\|\Delta_A\| \leq 1, \quad \|\Delta_B\| \leq 1,$$

and weight matrices $F_A \in \mathbb{R}^{n\times p_A}$, $H_A \in \mathbb{R}^{q_A\times n}$, $F_B \in \mathbb{R}^{n\times p_B}$, $H_B \in \mathbb{R}^{q_B\times n}$ are known.

An alernative parametrizaion is an affine one [16]:

$$R(q) = (A(q), B(q)) \in \mathcal{R},$$
$$\mathcal{R} := \left\{ R(q) : R(q) = \sum_{i=1}^{s} q_i R_i; \ \sum_{i=1}^{s} q_i = 1, \ q_i \geq 0 \right\}. \tag{3.6}$$

Regardless of the chosen uncertainty parameterization, it is usually additionally assumed that A_n is a Hurwitz matrix (this can always be achieved by *a priori* selection of the nominal part of the system).

In the following subsections, possible solutions are considered to the problems of filtering and linear dynamic controller design in the presence of a structured (3.5) or affine (3.6) uncertainty.

3.1. Robust Filtering Problem

As, considering the filtering problem, we do not design the input signal $u(t)$, then the success of its solution depends entirely on *a priori* assumptions about the class of this signal. There is a distinction between stochastic filtering tasks (for example, $u(t)$ is a random white noise with zero mean) and deterministic ones (for example, $u \in L_2$ or $u \in L_\infty$). In this overview, we consider some solutions to the filtering problem only for a deterministic case.

3.1.1. $\mathcal{H}_\infty$ filtering. The aim of $\mathcal{H}_\infty$ filtering is to obtain state estimate $\hat{x}(t)$ such that, if $u \in L_2$, then (2.5) holds, and, if $x_0 = 0$, then (3.4) holds for a given $\gamma > 0$ for all uncertainties (3.5) or (3.6).

Up to date, a solution to this problem has been obtained for both structured (3.5) and affine (3.6) uncertainties (see [17, 18] and [19–22], respectively). Considering structured uncertainty, the solution is obtained in terms of solutions of two related Riccati equations (2-Riccati approach). As for the case of affine uncertainty, solutions are obtained in terms of linear matrix inequalities derived on the basis of quadratic Lyapunov functions, both independent [19, 20] and dependent [21, 22] from a parameter. For a detailed comparative overview of existing solutions, we would like to address an interested reader to the treatise [9]. In this paper, we restrict our consideration to one algorithm that ensures (2.5) and (3.4) for affine uncertainty.

The solution to the $\mathcal{H}_\infty$ filtering problem is usually designed using an observer defined by the following equation:

$$\dot{\hat{x}}(t) = G\hat{x}(t) + Ly(t), \quad \hat{x}(t) = \hat{x}_0. \tag{3.7}$$

Unlike (3.1), here the input signal $u(t)$ is not used to obtain state estimate, and not only correction gain vector L is a design parameter, but also the matrix G, which is no longer fixed equal to the nominal matrix of the system A_n. The procedure to calculate these parameters is given by the following theorem.

Theorem 1. *Let $\hat{Y}_1$, $\hat{Y}_2$ and $\hat{P}_1$, $\hat{P}_2$ be solutions of the following linear matrix inequalities (LMI)*

$$\begin{bmatrix} P_1 & P_2 \\ P_2 & P_2 \end{bmatrix} > 0, \quad \begin{bmatrix} \Omega_{1,i} + \Omega_{1,i}^{\mathrm{T}} & Y_1 + \Omega_{2,i}^{\mathrm{T}} & P_1 B_i & I \\ * & Y_1 + Y_1^{\mathrm{T}} & P_2 B_i & 0 \\ * & * & -\gamma^2 I & 0 \\ * & * & * & -I \end{bmatrix} < 0, \quad \forall i = 1, \ldots s, \tag{3.8}$$

where

$$\Omega_{1,i} = P_1 A_i + Y_2 C^{\mathrm{T}}, \quad \Omega_{2,i} = P_2 A_i + Y_2 C^{\mathrm{T}},$$

with respect to $Y_1 \in \mathbb{R}^{n \times n}$, $Y_2 \in \mathbb{R}^n$ and $P_1 \in \mathbb{R}^{n \times n}$, $P_2 \in \mathbb{R}^{n \times n}$ for a given $\gamma > 0$.

Then the parameters $G = \hat{P}_2^{-1}\hat{Y}_1$, $L = \hat{P}_2^{-1}\hat{Y}_2$ ensure that the goals (2.5) *and* (3.4) *are achieved.*

Proof is given in [9].

The observer (3.7) designed via solution of LMI (3.8) has three main problems. First, the goal (2.5) can be achieved only when the uncertainty satisfies the predetermined parameterization (3.6), which is, in fact, typical of all robust solutions. Second, the excessive conservatism is imposed by the design procedure and related to the application of a common Lyapunov function for s vertices of the polytope (3.6). Third, the goal (2.5) is achieved only if the restrictive condition $u \in L_2$ is satisfied. As for the second problem, there exist procedures [21, 22] that solve the problem (3.4) based on a parameter-dependent Lyapunov function.

3.1.2. Invariant ellipsoid method. Considering a more practically common case $u \notin L_2$, $u \in L_\infty$ and when the matrix $A_n - LC^{\mathrm{T}}$ is Hurwitz one, we can only conclude from (3.3) that the norm of the estimation error converges asymptotically to some set:

$$\lim_{t\to\infty} \|e\,(t)\| \le e_{\max},$$

where $e_{\max}$ depends from both the upper bound of the signal $(B - B_n)\,u\,(t) + (A - A_n)\,x\,(t)$, and the chosen correction gain vector L.

For this situation, a more precise description of the asymptotic behavior of the error can be obtained using the notion of an invariant ellipsoid of the system.

Definition 1. An ellipsoid centered at the origin

$$\mathcal{E}_e = \left\{ e \in \mathbb{R}^n \colon e^{\mathrm{T}} P_e^{-1} e \le 1 \right\} \quad P_e > 0$$

is said to be positively invariant for dynamical system (3.3) + (3.5) if for all uncertainties $\|\Delta_A\| \le 1$, $\|\Delta_B\| \le 1$ and all input signals $|u\,(t)| \le 1$ it holds that:

1) $e_0 \in \mathcal{E}_e \Rightarrow e\,(t) \in \mathcal{E}_e$ for all $t \ge t_0$;
2) $e_0 \notin \mathcal{E}_e \Rightarrow e\,(t) \to \mathcal{E}_e$ for $t \to \infty$.

The matrix $P_e \in \mathbb{R}^{n\times n}$ is a configuration matrix of the ellipsoid $\mathcal{E}_e$.

Obviously, there exists an infinite number of invariant ellipsoids for the considered system (3.3). Since, in case $u \in L_\infty$, $A \ne A_n$ and/or $B \ne B_n$, the goal (2.5) for the filtering problem cannot be achieved by means of robust control, it remains only to try to provide convergence of the error $e\,(t)$ into the minimal ellipsoid. Minimal ellipsoid can be understood in different ways. Most commonly, in the literature, the minimality criterion is chosen as the minimization of the sum of squared semi-axes of the ellipsoid or, equivalently, the minimization of the trace of the ellipsoid matrix P_e [23]. Strictly, the problem of minimization of the sum of squared semi-axes of the ellipsoid is stated as follows.

Assume that the signal $u\,(t)$ is bounded by $|u\,(t)| \le 1$ (this constraint can always be satisfied by normalization). Then we need to compute the correction gain vector $L \in \mathbb{R}^{n\times 1}$ that ensures that the system (3.3) trajectories meet the conditions (1) and (2) from Definition 1, where the ellipsoid $\mathcal{E}_e$ matrix has minimal trace $\operatorname{tr} P_e$.

Such problem solution is designed in the following way. First of all, an extended system is considered:

$$\begin{bmatrix} \dot{x}\,(t) \\ \dot{e}\,(t) \end{bmatrix} = \underbrace{\begin{bmatrix} A_n + F_A\Delta_A H_A & 0 \\ F_A\Delta_A H_A & A_n - LC^{\mathrm{T}} \end{bmatrix}}_{\tilde{A}} \underbrace{\begin{bmatrix} x\,(t) \\ e\,(t) \end{bmatrix}}_{g(t)} + \underbrace{\begin{bmatrix} B_n + F_B\Delta_B H_B \\ F_B\Delta_B H_B \end{bmatrix}}_{\tilde{B}} u\,(t)\,.$$

Secondarily, its state $g\,(t)$ is put into an ellipsiod $\mathcal{E}_g$, which is defined by the matrix

$$P = \begin{bmatrix} P_x & 0 \\ 0 & P_e \end{bmatrix} \in \mathbb{R}^{2n\times 2n}, \quad P > 0$$

and the ellipsoid defined by the matrix P_e is minimized. As a result, using techniques based on the Lyapunov second method, Petersen lemma [24], S-procedure and Schur lemma, the following result is obtained.

Theorem 2. *Let $\hat{Q}_e$ and $\hat{Y}$ be the solutions of the minimization problem:*

$$\mathrm{tr}\ H \to \min$$

subjected to the constraints

$$\begin{bmatrix} \Omega_1 & 0 & Q_x B_n & Q_x F_A & Q_x F_B \\ * & \Omega_2 & 0 & Q_e F_A & Q_e F_B \\ * & * & -\alpha I + \varepsilon_2 H_B^{\mathrm{T}} H_B & 0 & 0 \\ * & * & * & -\varepsilon_1 I & 0 \\ * & * & * & * & -\varepsilon_2 I \end{bmatrix} \leq 0, \quad \begin{bmatrix} H & I \\ I & Q_e \end{bmatrix} \geq 0,$$

where

$$\Omega_1 = A_n^{\mathrm{T}} Q_x + Q_x A_n + \alpha Q_x + \varepsilon_1 H_A^{\mathrm{T}} H_A,$$
$$\Omega_2 = A_n^{\mathrm{T}} Q_e + Q_e A_n - Y C^{\mathrm{T}} - C Y^{\mathrm{T}} + \alpha Q_e,$$

with respect to the variables $Q_x = Q_x^{\mathrm{T}} \in \mathbb{R}^{n\times n}$, $Q_e = Q_e^{\mathrm{T}} \in \mathbb{R}^{n\times n}$, $Y \in \mathbb{R}^n$, ε_1, $\varepsilon_2 \in \mathbb{R}$ and scalar parameter $\alpha > 0$.

Then the minimal invariant ellipsoid is defined by the matrix $\hat{P}_e = \hat{Q}_e^{-1}$, and the respective correction gain of the observer is obtained as $L = \hat{Q}_e^{-1}\hat{Y}$.

Proof is given in [25].

According to Theorem 2, the computation of the correction gain vector L for fixed $\alpha > 0$ is reduced to a semi-definite programming problem, which is solved numerically using various software packages.

The open problems of the observers designed by the invariant ellipsoid method include the fundamental inability to meet the goal (2.5) without additional assumptions about the control signal $u(t)$. This problem is caused by the fact that regardless of the choice of the correction gain vector L, in error equation (3.3) there exists a non-vanishing component $(B - B_n)\, u(t) + (A - A_n)\, x(t)$ considered as a perturbation. The second problem is that, in order to minimize the ellipsoid trace, we need to solve optimization problem w.r.t. the parameter $\alpha > 0$ (convexity of the objective function w.r.t. this parameter has not been proved). The third and more significant problem is that the stated problem is solved only when the real uncertainty of the parameters A and B satisfies the chosen parameterization (3.5). If the real uncertainty does not satisfy this parameterization (*i.e.*, the chosen weight matrices F_A, H_A and F_B, H_B do not match the reality), then the observer correction gain no longer provides optimality in the sense of the trace of the invariant ellipsoid. The conservative choice of weight matrices, of course, can solve this problem, but leads to excessive conservatism of the obtained state estimate.

3.2. Dynamic Feedback Based Control

As it was mentioned earlier, the feedback design is almost always out of the scope of this overview. However, in this case the choice of feedback allows one to improve the properties of the solution of state reconstruction problem.

Using dynamic feedback control, it is possible to overcome the first drawback of the observer (3.1), roughly speaking, by ensuring that the summand $(B - B_n)\, u(t) + (A - A_n)\, x(t)$ converges to zero by choosing a feedback control law as $u(t) = -K_r \hat{x}(t)$ and stabilizing the system (2.1). In the presence of parametric uncertainty, the separation principle commonly used for linear systems is no longer valid [6]. This means that, in order to stabilize the system (2.1) via feedback based on the observer state, a joint calculation of the parameters of the observer and the controller is necessary to ensure the stability of the extended system composed of the equations of both the system and the observer. Let us briefly review the application of the Riccati approach and the linear matrix inequalities technique for the joint calculation of the parameters L and K_r.

3.2.1. Riccati approach. The control law for the system (2.1) with uncertainty (3.5) is chosen as a linear dynamic controller (2.4a) that uses the states of the observer (3.1). Then we need to calculate the parameters $L \in \mathbb{R}^{n\times 1}$ and $K_r \in \mathbb{R}^{1\times n}$ in such a way that

$$\lim_{t\to\infty} \|e(t)\| = \lim_{t\to\infty} \|x(t) - \hat{x}(t)\| = 0, \quad \lim_{t\to\infty} \|x(t)\| = 0. \tag{3.9}$$

The solution of this problem on the basis of Riccati approach is obtained in [26] and given below in the form of the following theorem.

Theorem 3. *Let there exist constants $\varepsilon_1 > 0$, $\varepsilon_2 > 0$ such that the following conditions hold:*

i) there exists $P_c = P_c^{\mathrm{T}} > 0$, which satisfies Riccati equation

$$\begin{gathered} A_n^{\mathrm{T}} P_c + P_c A_n + 2H_A^{\mathrm{T}} H_A + \varepsilon_1 Q_1 \\ - P_c \left[\frac{1}{\varepsilon_1} \left(B_n \left(R_1^{-1} - 2R_1^{-1} H_B^{\mathrm{T}} H_B R_1^{-1} \right) B_n^{\mathrm{T}} - 2F_B F_B^{\mathrm{T}} \right) - F_A F_A^{\mathrm{T}} \right] P_c = 0, \end{gathered} \tag{3.10a}$$

where $R_1 = R_1^{\mathrm{T}} > 0$, $Q_1 = Q_1^{\mathrm{T}} > 0$.

ii) there exists $P_o = P_o^{\mathrm{T}} > 0$, which satisfies Riccati equation

$$\begin{gathered} A_n^{\mathrm{T}} P_o + P_o A_n - \frac{1}{\varepsilon_2} C R_2^{-1} C^{\mathrm{T}} + P_o \left(F_A F_A^{\mathrm{T}} + \frac{2}{\varepsilon_1} F_B F_B^{\mathrm{T}} + \varepsilon_2 Q_2 \right) P_o \\ + \frac{2}{\varepsilon_2} P_c B_n R_1^{-1} H_B^{\mathrm{T}} H_B R_1^{-1} B_n^{\mathrm{T}} P_c = 0, \end{gathered} \tag{3.10b}$$

where $R_2 = R_2^{\mathrm{T}} > 0$, $Q_2 = Q_2^{\mathrm{T}} > 0$.

iii) the matrices P_c and P_o satisfy matrix inequality

$$\begin{gathered} \varepsilon_1 \left(Q_1 + \frac{1}{\varepsilon_1^2} P_c B_n R_1^{-1} B_n^{\mathrm{T}} P_c \right) \\ - \frac{1}{\varepsilon_1^2 \varepsilon_2} P_c B_n R_1^{-1} B_n^{\mathrm{T}} P_c \left[P_o Q_2 P_o + \frac{1}{\varepsilon_2^2} C R_2^{-1} C^{\mathrm{T}} \right]^{-1} P_c B_n R_1^{-1} B_n^{\mathrm{T}} P_c > 0. \end{gathered} \tag{3.10c}$$

Then the following choice of $K_r = \frac{1}{\varepsilon_1} R_1^{-1} B_n^{\mathrm{T}} P_c$ and $L = \frac{1}{\varepsilon_2} P_o^{-1} C R_2^{-1}$ ensures that the stated goal (3.9) *is achieved.*

Proof is given in [26].

Equations (3.10a) and (3.10b) are obtained via finding majorants of the uncertainty terms Δ_A, Δ_B, which occur in the derivative of the function $V = x^{\mathrm{T}} P_c x + e^{\mathrm{T}} P_o e$, with the help of the following inequality

$$a^{\mathrm{T}} M F N b \le a^{\mathrm{T}} M M^{\mathrm{T}} a + b^{\mathrm{T}} N^{\mathrm{T}} N b,$$

which holds for such F that $F^{\mathrm{T}} F \le I$, and any $a, b \in \mathbb{R}^n$, $M \in \mathbb{R}^{n\times p}$, $N \in \mathbb{R}^{p\times n}$.

Inequality (3.10c) is obtained via application of Schur lemma and means positive definiteness of the matrix Ω that defines the derivative $\dot{V} \le -\begin{bmatrix} x \\ e \end{bmatrix}^{\mathrm{T}} \Omega \begin{bmatrix} x & e \end{bmatrix}$.

3.2.2. Linear matrix inequalities technique. Application of the LMI [27] technique to solve the dynamic feedback control problem requires, in addition to L and K_r calculation, to compute the observer state matrix. Therefore, equation (3.1) is rewritten in the form of

$$\dot{\hat{x}}(t) = G\hat{x}(t) + B_n u(t) + L\left(y(t) - C^{\mathrm{T}}\hat{x}(t)\right), \quad \hat{x}(t) = \hat{x}_0.$$

It is required to calculate the parameters $G \in \mathbb{R}^{n\times n}$, $L \in \mathbb{R}^{n\times 1}$ and $K_r \in \mathbb{R}^{1\times n}$ such that the stated goal (3.9) is achieved. Using techniques based on the Lyapunov second method, Petersen and Schur lemmas, the following result is presented in [28].

Theorem 4. *Let $\hat{Q}_{\hat{x}}$, $\hat{Y}_1$, $\hat{Q}_e$, $\hat{Y}_2$ and $\hat{\alpha}$, $\hat{Y}_3$ be a solution of the following LMI:*

$$\begin{bmatrix} \Omega_{11} & \Omega_{12} & Q_{\hat{x}}H_A^{\mathrm{T}} & -Y_1^{\mathrm{T}}H_B^{\mathrm{T}} \\ * & \Omega_{22} & Q_e H_A^{\mathrm{T}} & 0 \\ * & * & -\varepsilon_1 I & 0 \\ * & * & * & -\varepsilon_2 I \end{bmatrix} < 0$$

under the condition

$$C^{\mathrm{T}}Q_e = \alpha C^{\mathrm{T},}$$

where

$$\Omega_{11} = Y_3 + Y_3^{\mathrm{T}} - B_n Y_1 - Y_1^{\mathrm{T}}B_n^{\mathrm{T}},$$
$$\Omega_{12} = Y_2 C^{\mathrm{T}} + Q_{\hat{x}}A_n^{\mathrm{T}} - Y_3^{\mathrm{T}},$$
$$\Omega_{22} = A_n Q_e + Q_e A_n^{\mathrm{T}} - Y_2 C^{\mathrm{T}} - C Y_2^{\mathrm{T}} + \varepsilon_1 F_A F_A^{\mathrm{T}} + \varepsilon_2 F_B F_B^{\mathrm{T}}$$

with respect to $Q_{\hat{x}} = Q_{\hat{x}}^{\mathrm{T}} \in \mathbb{R}^{n\times n}$, $Q_e = Q_e^{\mathrm{T}} \in \mathbb{R}^{n\times n}$, $Y_1 \in R^{1\times n}$, $Y_2 \in \mathbb{R}^n$, $Y_3 \in \mathbb{R}^{n\times n}$, ε_1, ε_2, $\alpha > 0$.

Then the parameters $K_r = \hat{Y}_1\hat{Q}_{\hat{x}}^{-1}$, $G = \hat{Y}_3\hat{Q}_{\hat{x}}^{-1}$ and $L = \hat{Y}_2\hat{\alpha}^{-1}$ ensure that the stated goal (3.9) *is achieved.*

Proof is given in [28].

The common open problems of the Riccati and LMI approaches include the necessity to choose a control law in the form of $u(t) = -K_r\hat{x}(t)$ and the need of joint calculation of the observer and controller parameters.

3.3. Conclusions on Overview of Robust Observers

A distinctive feature of all considered robust approaches to the state observers design is the assumption that the parametric uncertainty of the system meets the parameterizations (3.5) or (3.6). In practice, this assumption requires the range of variation of all unknown parameters to be known. Excessive conservatism in such ranges choice can lead to high values of the feedback parameters, and, as a consequence, to non-implementable solutions.

4. STATE ESTIMATION BASED ON INVARIANCE THEORY

Invariant observers are designed on the basis of idea of elimination (whether algebraic or high-gain one) of the parametric uncertainty effect on the state estimate [7, 29–31]. In the first step of such solutions design, equation (2.1) is rewritten in the following form:

$$\begin{gathered} \dot{x}(t) = (A_n + \Delta A)\,x(t) + (B_n + \Delta B)\,u(t) = A_n x(t) + B_n u(t) + Dw(x,u), \\ Dw(x,u) = \Delta A x(t) + \Delta B u(t), \end{gathered} \tag{4.1}$$

where ΔA, ΔB are the aggregated parametric uncertainties of the corresponding dimensions, $D \in \mathbb{R}^{n\times m}$ is a known vector that assigns gains to the uncertainties and allocates them to the equations, $w(x,\ u) \in \mathbb{R}^m$ is the unknown bounded (by Assumption 1) perturbation $\|w(x,\ u)\| \le w_{\max}$.

The stated goal (2.5) is now interpreted as the reconstruction of the state of the system (2.1) that has an unmeasured (unknown) input $w(x,\ u)$. Let us consider the solutions of this problem based on the Luenberger observer for systems with unknown input and various sliding mode observers.

4.1. Luenberger Observer for Systems with Unknown Input

In accordance with the results [32, 33], it is proposed to obtain the estimate of unmeasurable states by means of the following set of equations:

$$\begin{aligned}\dot{z}(t) &= Nz(t) + Gu(t) + Ly(t),\\ \hat{x}(t) &= z(t) - Ey(t),\end{aligned} \tag{4.2}$$

where $z(t) \in \mathbb{R}^n$ and matrices of corresponding dimensions are chosen as follows:

$$\begin{gathered}\left(I + EC^{\mathrm{T}}\right)D = 0,\\ N = \left(I + EC^{\mathrm{T}}\right)A_n - (L + NE)C^{\mathrm{T}},\\ \left(I + EC^{\mathrm{T}}\right)B_n = G,\end{gathered} \tag{4.3}$$

and the vector L makes the matrix N be Hurwitz one.

Owing to (4.2), the following holds for the state reconstruction error $\tilde{x}(t) = \hat{x}(t) - x(t)$:

$$\tilde{x}(t) = z(t) - EC^{\mathrm{T}}x(t) - x(t) = z(t) - \left(I + EC^{\mathrm{T}}\right)x(t),$$

and then the error equation is written as:

$$\begin{gathered}\dot{\tilde{x}}(t) = \dot{z}(t) - \left(I + EC^{\mathrm{T}}\right)\dot{x}(t)\\ = Nz(t) + Gu(t) + Ly(t) - \left(I + EC^{\mathrm{T}}\right)\left(A_n x(t) + B_n u(t) + Dw(x,u)\right)\\ = Nz(t) + Ly(t) - \left(I + EC^{\mathrm{T}}\right)A_n x(t)\\ = Nz(t) + LC^{\mathrm{T}}x(t) - Nx(t) - (L + NE)C^{\mathrm{T}}x(t)\\ = Nz(t) - Nx(t) - NEC^{\mathrm{T}}x(t) = N\tilde{x}(t).\end{gathered} \tag{4.4}$$

Considering equation (4.4), it is easy to conclude that the invariance of the estimation error $\tilde{x}(t)$ with respect to the parametric uncertainty $w(x,\ u)$ is achieved because of its algebraic elimination from the error equation. Necessary and sufficient conditions to make equations (4.3) solvable and the matrix N be Hurwitz one have been derived and revised in [32] and [33], respectively. For the considered case of the class of single-input-single-output systems, these conditions are formulated as follows.

Theorem 5. *Set of equations* (4.3) *has a solution if and only if the following conditions hold:*
1) $m = \dim(y(t)) = 1$,
2) $\operatorname{rank}\left(C^{\mathrm{T}}D\right) = \operatorname{rank}(D)$,
3) *the invariant zeros of the triple* $\left(A_n,\, D,\, C^{\mathrm{T}}\right)$ *are stable (have a negative real part).*

Proof is given in [32, 33].

The first two premises of Theorem 5 are restrictive for applications. In particular, the first condition is necessary (but not sufficient [34]) to meet the second one and does not allow two functionally different uncertainties to be in different equations of the system. For example, the observer (4.2) cannot be applied to the following second-order system:

$$\begin{aligned}
\dot{x}(t) &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t) + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 x_1(t) \\ a_2 x_1(t) + a_3 x_2(t) \end{bmatrix}, \\
y(t) &= \begin{bmatrix} 1 & 0 \end{bmatrix} x(t),
\end{aligned} \tag{4.5}$$

where $m = 2 > \dim(y(t)) = 1$, which contradicts the first condition.

In its turn, the second condition or output matching condition requires a perturbation to affect the equation, which state is measurable. For example, the observer (4.2) is not applicable to the following second-order system:

$$\begin{aligned}
\dot{x}(t) &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} (a_2 x_1(t) + a_3 x(t) + b_1 u), \\
y(t) &= \begin{bmatrix} 1 & 0 \end{bmatrix} x(t),
\end{aligned} \tag{4.6}$$

where $\operatorname{rank}\left(C^{\mathrm{T}} D\right) = 0 \neq \operatorname{rank}(D) = 1$, which contradicts the second condition.

The independence of the estimation error from an arbitrary external perturbation is the main competitive advantage of the considered Luenberger observer in comparison with the robust alternatives (3.1) and (3.7). However, strict structural constraints do not allow one to use the considered observer for state reconstruction of a wide class of real technical systems with output unmatched uncertainty.

4.2. Sliding Mode Observers. Output Matching Conditions

The invariance of the state reconstruction error with respect to parametric uncertainty can be achieved not only by its algebraic elimination, but also using special sliding modes [7, 29–31]. The first observer based on sliding modes was proposed by V.I. Utkin [35], later this solution was improved in [36] with the help of additional Luenberger correction term. In [37], an alternative way to choose the sliding surface was proposed, and necessary and sufficient conditions for the existence of all three types of sliding mode observers [35–37] were obtained.

According to an algorithm from [37], the following observer is designed for a system represented as (4.1):

$$\begin{aligned}
\dot{\hat{x}}(t) &= A_n \hat{x}(t) + B_n u(t) - G_l(\hat{y}(t) - y(t)) + G_n v(t), \\
\hat{y}(t) &= C^{\mathrm{T}} \hat{x}(t),
\end{aligned} \tag{4.7}$$

where

$$G_l = T^{-1} \begin{bmatrix} F_{12} \\ F_{22} - F_m \end{bmatrix} \in \mathbb{R}^n, \quad G_n = T^{-1} \begin{bmatrix} 0_{n-1} \\ 1 \end{bmatrix} \in \mathbb{R}^n,$$

$$v(t) := \begin{cases} -\rho \dfrac{P(\hat{y}(t) - y(t))}{\|P(\hat{y}(t) - y(t))\|}, & \text{if} \quad \hat{y}(t) - y(t) \neq 0, \\ 0 & \text{otherwise}, \end{cases}$$

and F_m is a Hurwitz matrix, P stands for a solution of Luapunov equation for F_m, $\rho > 0$ denotes a sufficiently large scalar, $T \in \mathbb{R}^{n\times n}$ is full-rank similarity transformation matrix, F_{12}, F_{22} stand for known matrices, which definition algorithm will become clear from further explanations.

According to [37, 38], the observer (4.7) is implementable and ensures convergence of the state reconstruction error (2.5) when the existence conditions of the following structural transformations are satisfied.

Lemma 1. *Let the following conditions be met:*
1) $m = \dim(y(t)) = 1$,
2) $\operatorname{rank}\left(C^{\mathrm{T}}D\right) = \operatorname{rank}(D)$,
3) *the invariant zeros of the triple* $\left(A_n,\, D,\, C^{\mathrm{T}}\right)$ *are stable.*

Then there exists a nonsingular coordinate transformation $\begin{pmatrix} z_1(t) \\ y(t) \end{pmatrix} = Tx(t)$, *which represents the system* (4.1) *and observer* (4.7) *in a block form:*

$$\begin{aligned}
\dot{z}_1(t) &= F_{11}z_1(t) + F_{12}y(t) + G_1u(t), \\
\dot{y}(t) &= F_{21}z_1(t) + F_{22}y(t) + G_2u(t) + D_2w(x,u), \\
\dot{\hat{z}}_1(t) &= F_{11}\hat{z}_1(t) + F_{12}\hat{y}(t) + G_1u(t) - F_{12}e_y(t), \\
\dot{\hat{y}}(t) &= F_{21}\hat{z}_1(t) + F_{22}\hat{y}(t) + G_2u(t) - (F_{22} - F_m)\,e_y(t) + v(t),
\end{aligned} \tag{4.8}$$

where $D_2 > 0$, F_{11} *is a Hurwitz matrix according to the design,* $e_y(t) = \hat{y}(t) - y(t)$ *is a error of system output observation.*

Proof is given in [37, 38].

The detailed algorithm to obtain the transformation matrix T is given in [7, 29–31, 35, 37] and is reduced to elimination of the external perturbation from the equation related to the unmeasured variable. The error equation between the system and observer dynamics represented via new coordinates (4.8) has the form:

$$\begin{aligned}
\dot{\tilde{z}}_1(t) &= F_{11}\tilde{z}_1(t), \\
\dot{e}_y(t) &= F_{21}\tilde{z}_1(t) + F_m e_y(t) + v(t) - D_2w(x,u).
\end{aligned} \tag{4.9}$$

The fact that F_{11} is a Hurwitz matrix leads to the exponential convergence of the error $\tilde{z}_1(t)$ to zero. Using the Lyapunov second method, the error $e_y(t)$ is shown to reach the line $e_y(t) = 0$ in finite time. By the method of equivalent control $\mu^{-1}\dot{v}_{eq}(t) = v(t) - v_{eq}(t)$, $\mu > 0$, it is also shown that $D_2^{-1}v_{eq}(t) \to w(x,\, u)$. Since T is a nonsingular transformation, then, as the errors (4.9) are exponentially stable, we immediately have that the goal (2.5) is achieved.

Thus, the sliding mode observer (4.7) requires to meet the same strict structural constraints as the perturbation invariant Luenberger observer (4.2). However, in contrast to the solution (4.2), the observer (4.7) uses another technique to eliminate the external perturbation, which is based on the robust properties of the motion on the sliding surface. A more detailed comparison of the properties of these two types of invariant observers are given in [39, 40].

4.3. Sliding Mode Observers. Extended Matching Conditions

The output matching conditions $\operatorname{rank}\left(C^{\mathrm{T}}D\right) = \operatorname{rank}(D)$ significantly restrict the class of systems, which state can be reconstructed with the help of (4.2) and (4.7) observers. It is possible to relax this requirement by increasing the dimensionality of the output $y(t)$ via addition of new physical or virtual measurements to the existing ones. New physical measurements requires to install additional sensors, which contradicts the theoretical formulation of the problem, and, considering practical scenarios, is often impractical or impossible. So, the number of measurements can be increased only by means of virtual states. In [41], it is proposed to introduce into consideration a virtual output variable

$$\xi(t) = \mathcal{O}^{-1}x(t) = \begin{bmatrix} C^{\mathrm{T}} \\ \vdots \\ C^{\mathrm{T}}A_n^{n-1} \end{bmatrix} x(t) \tag{4.10}$$

and design the following state observer:

$$\begin{aligned} \dot{\hat{x}}(t) &= A_n\hat{x}(t) + B_n u(t) - G_l\left(\hat{y}_a(t) - \hat{\xi}(t)\right) + G_n v(t), \\ \hat{y}_a(t) &= \mathcal{O}^{-1}\hat{x}(t), \end{aligned} \tag{4.11}$$

where $\hat{y}_a(t)$ is an estimate of the output (4.10) with the help of (4.11), $\hat{\xi}(t)$ denotes the estimate of the output (4.10) with the help of an observer, which will be defined further, and the matrices G_l, G_n and function $v(t)$ are chosen like in (4.7).

According to the results of [41, 42], the observer (4.11) is implementable and ensures convergence of the state reconstruction error in case the following conditions for the system that is reducible to block and triangular forms are satisfied.

Lemma 2. *Let the following conditions be met:*
1) $m = \dim(y(t)) = 1$,
2) $\operatorname{rank}(\mathcal{O}^{-1}D) = \operatorname{rank}(D)$,
3) $\begin{bmatrix} C^{\mathrm{T}} \\ \vdots \\ C^{\mathrm{T}}A_n^{n-2} \end{bmatrix} D = 0,$
4) *the invariant zeros of the triple* $\left(A_n, D, C^{\mathrm{T}}\right)$ *are stable.*

Then:

a) *there exists* [41] *a nonsingular transformation* $\begin{pmatrix} z_1(t) \\ y_a(t) \end{pmatrix} = Tx(t)$, *which represents the system* (4.1) *with the output* $\xi(t)$ *and the observer* (4.11) *in a block form* (4.8) *up to a change of* $y(t)$ *and* $\hat{y}(t)$ *by* $\xi(t)$ *and* $\hat{y}_a(t)$, *respectively;*

b) *there exists* [42] *a nonsingular transformation* (4.10), *which represents the system* (4.1) *in a triangular form:*

$$\begin{aligned} \dot{\xi}(t) &= A_0\xi(t) + B_0\delta(t) + B_e u(t), \\ y(t) &= C_0^{\mathrm{T}}\xi(t), \end{aligned} \tag{4.12}$$

where

$$A_0 = \begin{bmatrix} 0_n & \begin{matrix} I_{n-1} \\ 0_{1\times(n-1)} \end{matrix} \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0_{n-1} \\ 1 \end{bmatrix}, \quad B_e = \begin{bmatrix} C^{\mathrm{T}}B_n \\ \vdots \\ C^{\mathrm{T}}A_n^{n-2}B_n \\ C^{\mathrm{T}}A_n^{n-1}B_n \end{bmatrix},$$

$$\delta(t) = C^{\mathrm{T}}A_n^n x(t) + C^{\mathrm{T}}A_n^{n-1}Dw(t), \quad C_0^{\mathrm{T}} = \begin{bmatrix} 1 & 0_{n-1}^{\mathrm{T}} \end{bmatrix}.$$

Proof is given in [41, 42].

The implementation of the observer (4.11) requires to obtain an estimate $\hat{\xi}(t)$ of the output (4.10), which is simultaneously the state of the triangular system (4.12). In [41], a cascade observer is proposed to estimate the states of the system (4.12) in finite time:

$$\begin{cases} \dot{\hat{\zeta}}_1(t) = v\left(y(t) - \hat{\zeta}_1(t)\right) \\ \dot{\hat{\zeta}}_2(t) = E_1 v\left(\tilde{\zeta}_2(t) - \hat{\zeta}_2(t)\right) \\ \vdots \\ \dot{\hat{\zeta}}_{n-1}(t) = E_{n-2} v\left(\tilde{\zeta}_{n-1}(t) - \hat{\zeta}_{n-1}(t)\right) \\ \dot{\hat{\zeta}}_n(t) = E_{n-1} v\left(\tilde{\zeta}_n(t) - \hat{\zeta}_n(t)\right) \end{cases} + B_e u(t), \tag{4.13}$$

$$\hat{\xi}(t) = \begin{bmatrix} y(t) \\ v\left(\tilde{\zeta}_2(t) - \hat{\zeta}_2(t)\right) \\ \vdots \\ v\left(\tilde{\zeta}_{n-1}(t) - \hat{\zeta}_{n-1}(t)\right) \end{bmatrix},$$

where $\tilde{\zeta}_1(t) = y(t)$ and

$$\tilde{\zeta}_i(t) = v\left(\tilde{\zeta}_{i-1}(t) - \hat{\zeta}_{i-1}(t)\right), \quad 2 \le i \le n,$$
$$E_j = \begin{cases} 1, & \text{if } \left|\tilde{\zeta}_i(t) - \hat{\zeta}_i(t)\right| \le \varepsilon \\ 0 & \text{otherwise,} \end{cases} \qquad j \le i,$$
$$v(.) = \varphi(t) + \lambda_s |.|^{\frac{1}{2}} \operatorname{sgn}(.), \quad \lambda_s > 0,$$
$$\dot{\varphi}(t) = \alpha_s \operatorname{sgn}(.), \quad \alpha_s > 0.$$

The error equation between the systems (4.12) and (4.13) is written as:

$$\begin{cases} \dot{\bar{\zeta}}_1(t) = \xi_2(t) - v\left(y(t) - \hat{\zeta}_1(t)\right) \\ \dot{\bar{\zeta}}_2(t) = \xi_3(t) - E_1 v\left(\tilde{\zeta}_2(t) - \hat{\zeta}_2(t)\right) \\ \vdots \\ \dot{\bar{\zeta}}_{n-1}(t) = \xi_n(t) - E_{n-2} v\left(\tilde{\zeta}_{n-1}(t) - \hat{\zeta}_{n-1}(t)\right) \\ \dot{\bar{\zeta}}_n(t) = \delta(t) - E_{n-1} v\left(\tilde{\zeta}_n(t) - \hat{\zeta}_n(t)\right), \end{cases} \tag{4.14}$$

and, in case of sufficiently large values of λ_s, α_s, the sliding modes emerge one by one in (4.14) at $\tilde{\zeta}_i(t) - \hat{\zeta}_i(t) = 0$, $i = \overline{1, n}$, which results in convergence of the error $\hat{\xi}(t) - \xi(t)$ in finite time T_ξ. When the estimation process of the output (4.10) of the system (4.1) is over, a sliding mode occurs at $\hat{y}_a(t) - \hat{\xi}(t) = 0$ in finite time $T_x \ge T_\xi$, which, in accordance with the analysis (4.9), ensures that the stated goal (2.5) is achieved.

According to the results of [41], when the new output (4.10) is measured, even if the matching condition $\operatorname{rank}\left(C^{\mathrm{T}} D\right) = \operatorname{rank}(D)$ is not satisfied, the extended matching condition 2)–3) of Lemma 2 can be met, and the observer (4.11) is implementable and ensures the convergence of the state reconstruction error.

For example, we have for (4.6) that:

$$\begin{bmatrix} C^{\mathrm{T}} \\ C^{\mathrm{T}}A \end{bmatrix} D = D \Rightarrow \operatorname{rank}\left(C_a^{\mathrm{T}} D\right) = \operatorname{rank}(D),$$

which allows one to reconstruct the state of the system (4.6) with the help of (4.11) + (4.13).[2]

Alternative ways to reconstruct the states of triangular systems were discussed in detail in [43, 44] and, by analogy with (4.13), they can be used to obtain an estimate of the output variable $\xi(t)$ under the same structural constraints from Lemma 2. Also it should be noted that, as the pair $\left(A, C^{\mathrm{T}}\right)$ is observable, the states $x(t)$ can be estimated in finite time without implementation of (4.11) but using $\hat{x}(t) = \mathcal{O}\hat{\xi}(t)$.

If the extended matching conditions 2)-3) from Lemma 2 are satisfied, the system is reduced to a triangular form just after the first choice of the virtual output (4.10), which allows one to use observers of the type (4.11) + (4.13) or (4.11) + $\hat{x}(t) = \mathcal{O}\hat{\xi}(t)$ to reconstruct the states. For high-dimensional systems with multiple inputs/outputs, an introduction of a single virtual output may be not enough to reduce the whole system to a triangular form [11, 45]. In this case, the transformations are continued until the whole system is divided into triangularly shaped blocks. The system equations split obtained as a result of such iterative procedure is called the quasi block triangular observable form (QBTOF), about which, strictly speaking, two facts are known:

(i) states and perturbations of the system are observable by measurements if and only if the system is reducible to QBTOF [11],

(ii) the system is reducible to QBTOF in case the matching conditions 2) from Theorem 5 and 3) from Lemma 2 are not satisfied [45].

Therefore, the sliding mode observers from [11, 45], designed on the basis of the QBTOF, have the weakest structural constraints in comparison with other invariant state observers (e.g., (4.7), (4.11) + (4.13)). However, reduction of the system to the QBTOF is a rather laborious iterative process, and some of the structural constraints of this form of system representation seem not to be relaxable either. For example, following step 1(a) of the constructive observability criterion [11, Section 3.2], let us write down the previously considered system (4.5) in the following form (using the authors' notation):

$$\begin{aligned} \dot{y}(t) &= A_{11} y(t) + D_1 x_1(t) + Q_1 \psi(t) + B_1 u(t), \\ \dot{x}_1(t) &= A_{x_{11}} y(t) + A_{x_1} x_1(t) + Q_{x_1} \psi(t) + B_{x_1} u(t), \end{aligned} \tag{4.15}$$

where in terms of the system (4.5) we adopt the following re-definitions:

$$y_1(t) := y(t) = x_1(t), \quad x_1(t) := x_2(t), \quad \psi(t) := \begin{bmatrix} \psi_1(t) \\ \psi_2(t) \end{bmatrix} = w(x, u),$$
$$A_{11} = 0, \quad D_1 = 1, \quad Q_1 = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad B_1 = 0,$$
$$A_{x_{11}} = A_{x_1} = 0, \quad B_{x_1} = 1, \quad Q_{x_1} = \begin{bmatrix} 0 & 1 \end{bmatrix}.$$

In order to reduce the system (4.5), (4.15) into QBTOF, the following condition (equation (**) from [11]) is to be met:

$$\operatorname{rank}(Q_1) - \operatorname{rank}\left(\begin{bmatrix} Q_1 & D_1 \end{bmatrix}\right) \neq 0.$$

As for the considered case (4.5), we have that $\operatorname{rank}(Q_1) - \operatorname{rank}\left(\begin{bmatrix} Q_1 & D_1 \end{bmatrix}\right) = 0$, which does not allow one to reduce the system to QBTOF, and hence, owing to the design criterion from [11],

[2] It holds for (4.6) that $\xi(t) = x(t)$, so the implementation of (4.11) is not really required.

the states and perturbations of the system (4.5) are unobservable. In order to reduce the system to QBTOF, the inequality $m \le \dim(y(t))$ is required to be satisfied, which for the class of linear systems (2.1) with one output that is considered in this paper is almost always not met.

Thus, all existing invariant state observers based on sliding mode technique [11, 35–41, 43–45], including recent results [75, 76], which allow one to improve the quality of transients for the state estimates, impose strict structural requirements on a dynamical system with a single output. Moreover, as shown by the constructive observability criterion [11, 45], some of these conditions for single-output systems cannot even potentially be relaxed for a class of sliding mode observers. Indeed, let us introduce an observer of the following form for the system (4.5), (4.15):

$$\dot{\hat{y}}_1(t) = A_{11}\hat{y}_1(t) + D_1\hat{x}_1 + B_1u + L(\hat{y}_1(t) - y_1(t)) - \rho\,\mathrm{sgn}(\hat{y}_1(t) - y_1(t)),$$
$$\dot{\hat{x}}_1(t) = A_{x_{11}}\hat{y}_1(t) + A_{x_1}\hat{x}_1 + B_{x_1}u - K_{x_1}\mathrm{sgn}(\hat{y}_1(t) - y_1(t)),$$

which allows one to obtain the following error equation:

$$\dot{\tilde{y}}_1(t) = (A_{11} + L)\tilde{y}_1(t) + D_1\tilde{x}_1(t) - Q_1\psi(t) - v(t),$$
$$\dot{\tilde{x}}_1(t) = A_{x_{11}}\tilde{y}_1(t) + A_{x_1}\tilde{x}_1(t) - Q_{x_1}\psi(t) - \frac{K_{x_1}}{\rho}v(t),$$
$$v(t) = \rho\,\mathrm{sgn}(\tilde{y}_1(t)), \quad \tilde{y}_1(t) = \hat{y}_1(t) - y_1(t).$$

In case $\rho > 0$ is sufficiently large, a sliding mode occurs in the first subsystem at $\tilde{y}_1(t) = 0$, which, using the method of equivalent control, allows one to obtain:

$$0 = D_1\tilde{x}_1(t) - Q_1\psi(t) - v_{eq}(t),$$
$$\begin{aligned}\dot{\tilde{x}}_1(t) &= \left(A_{x_1} - \frac{K_{x_1}}{\rho}D_1\right)\tilde{x}_1(t) - Q_{x_1}\psi(t) + \frac{K_{x_1}}{\rho}Q_1\psi(t)\\ &= \left(A_{x_1} - \frac{K_{x_1}}{\rho}D_1\right)\tilde{x}_1(t) - \psi_2(t) + \frac{K_{x_1}}{\rho}\psi_1(t),\end{aligned}$$

from which, in general case, we have only boundedness of the error $\tilde{x}_1(t)$, while the stated goal (2.5) can be achieved only if the condition $\psi_2(t) - \frac{K_{x1}}{\rho}\psi_1(t) = 0$ holds, which is not satisfied for the example under consideration (4.5), (4.15), as for all $K_{x1} \in \mathbb{R}$ it holds that $a_2x_1(t) + a_3x_2(t) - \frac{K_{x1}}{\rho}a_1x_1(t) \ne 0$.

Considering real technical systems, different parametric perturbations can potentially affect equations of the system, *i.e.*, the inequality $m > 1$ is often satisfied for the parameterization (4.1), which restricts the potential of sliding mode observers application to the problem of state reconstruction of linear systems with parametric uncertainty.

4.4. Conclusions on Overview of Observers Based on Invariance Theory

Owing to algebraic uncertainty elimination or sliding modes, the invariant observers provide insensitivity of the state estimation error to the parametric uncertainty of the system. This property allows one to achieve the goal (2.5) without additional assumptions regarding the parameterization of uncertainty and the properties of the signal $u(t)$.

The open problems for invariant observers are the need to use discontinuous correction signals that are sensitive to measurement noise and to fulfill strict output matching conditions.

5. ADAPTIVE OBSERVERS

In contrast to robust state reconstruction methods, adaptive algorithms theoretically do not require any *a priori* information about the system (2.1) parameters, because they simultaneously estimate unknown parameters and reconstruct state vector. Compared to observers that are invariant to perturbations, adaptive algorithms, by means of special parameterizations, are able to take into account the structure of the perturbation that affects the system, which, in many cases, makes it possible not to require to meet output matching conditions.

It is logical (but naive) to choose the structure of an adaptive state observer as follows:

$$\dot{\hat{x}}(t) = \hat{A}(t)\hat{x}(t) + \hat{B}(t)u(t) + \hat{L}(t)(\hat{y}(t) - y(t)), \quad \hat{x}(t_0) = \hat{x}_0, \tag{5.1}$$

where all unknown matrices of the system are substituted by their dynamic estimates.

Then the error equation between (2.1) and (5.1) is written as:

$$\begin{gathered}\begin{aligned}\dot{\tilde{x}}(t) &= \hat{A}(t)\hat{x}(t) + \hat{B}(t)u(t) + \hat{L}(\hat{y}(t) - y(t)) - Ax(t) - Bu(t) \pm L(\hat{y}(t) - y(t)) \pm A\hat{x}(t) \\ &= A\hat{x}(t) + \tilde{B}(t)u(t) + \tilde{L}\tilde{y}(t) + \tilde{A}(t)\hat{x}(t) - Ax(t) + LC^{\mathrm{T}}\tilde{x}(t) \\ &= \left(A + LC^{\mathrm{T}}\right)\tilde{x}(t) + \tilde{B}(t)u(t) + \tilde{L}(t)\tilde{y}(t) + \tilde{A}(t)\hat{x}(t),\end{aligned} \\ \tilde{y}(t) = C^{\mathrm{T}}\tilde{x}(t),\end{gathered} \tag{5.2}$$

where $\tilde{A}(t)$, $\tilde{B}(t)$ are the estimation errors of the system (2.1) parameters, $\tilde{L}(t)$ stands for the estimation error of L, which ensures that $\lambda_{\max}\left(A + LC^{\mathrm{T}}\right) < 0$.

As it follows from the differential equation (5.2), there are two principal mechanisms to achieve the goal (2.5). First, one can try to obtain estimates in such a way that the sum $\tilde{B}(t)u(t) + \tilde{L}(t)\tilde{y}(t) + \tilde{A}(t)\hat{x}(t)$ converges to zero. The second way is to ensure convergence to zero of the errors $\tilde{A}(t)$, $\tilde{B}(t)$ and $\tilde{L}(t)$.

Traditionally, choosing the first mechanism, adaptive laws are designed using the Lyapunov second method. According to this approach, a quadratic form is introduced into consideration:

$$V = \tilde{x}^{\mathrm{T}}P\tilde{x} + tr\left(\tilde{A}^{\mathrm{T}}\tilde{A}\right) + \tilde{B}^{\mathrm{T}}\tilde{B} + \tilde{L}^{\mathrm{T}}\tilde{L}, \tag{5.3}$$

where $P = P^{\mathrm{T}} > 0$ is a solution of the Lyapunov equation for $Q = Q^{\mathrm{T}} > 0$.

By easy but lengthy transformations, the derivative of the function (5.3) is obtained as:

$$\dot{V} = -\tilde{x}^{\mathrm{T}}Q\tilde{x} + 2\mathrm{tr}\left(\tilde{A}^{\mathrm{T}}P\tilde{x}\hat{x}^{\mathrm{T}} + \tilde{A}^{\mathrm{T}}\dot{\hat{A}}\right) + 2\tilde{B}^{\mathrm{T}}P\tilde{x}u + 2\tilde{B}^{\mathrm{T}}\dot{\hat{B}} + 2\tilde{L}^{\mathrm{T}}P\tilde{x}\tilde{y} + 2\tilde{L}^{\mathrm{T}}\dot{\hat{L}}. \tag{5.4}$$

Then the following adaptive laws:

$$\dot{\hat{L}}(t) = -P\tilde{x}(t)\tilde{y}(t), \quad \dot{\hat{B}}(t) = -P\tilde{x}(t)u(t), \quad \dot{\hat{A}}(t) = -P\tilde{x}(t)\hat{x}^{\mathrm{T}}(t) \tag{5.5}$$

ensure $\dot{V} = -\tilde{x}^{\mathrm{T}}Q\tilde{x}$, which, with the help of Barbalat lemma, allows one to prove the asymptotic convergence of the estimation error $\tilde{x}(t)$. However, the adaptive laws (5.5) use an unmeasurable estimation error and, hence, cannot be implemented.

The adaptive laws that use the second mechanism of error equation stabilization are designed with the help of the gradient method, and therefore, they require parameterization of the regression equation based on the measured signals $u(t)$ and $y(t)$ with respect to $n^2 + n$ parameters of the matrices A, B. Let us show that such a parameterization cannot be obtained in the general case.

A transfer function from control $u(t)$ to system output $y(t)$ has the following form ($m \leq n - 1$):

$$W(s) = C^{\mathrm{T}}(sI_n - A)^{-1}B = \frac{b_m s^m + b_{m-1}s^{m-1} + \ldots + b_0}{s^n + a_{n-1}s^{n-1} + \ldots + a_0}, \tag{5.6}$$

that is, only $2n \leq n^2 + n$ parameters are associated with the measured signals $u(t)$ and $y(t)$ (and their derivatives).

Let us rewrite equation (5.6) in the following form (here s is the differentiation operator):

$$\left(b_m s^m + b_{m-1} s^{m-1} + \ldots + b_0\right) u(t) = \left(s^n + a_{n-1} s^{n-1} + \ldots + a_0\right) y(t). \tag{5.7}$$

Having applied the operator $\frac{1}{\Lambda(s)}$ ($\Lambda(s)$ is a monic Hurwitz polynomial of order n) to the left- and right-hand sides of equation (5.7) and expressed $\frac{s^n}{\Lambda(s)} y(t)$ from the obtained equation, the following parameterization is written:

$$\begin{gathered} z(t) = \frac{s^n}{\Lambda(s)} y(t) = \varphi^{\mathrm{T}}(t)\,\psi, \\ \varphi(t) = \left[-\frac{\alpha_{n-1}^{\mathrm{T}}(s)}{\Lambda(s)} y(t) \quad \frac{\alpha_{n-1}^{\mathrm{T}}(s)}{\Lambda(s)} u(t)\right]^{\mathrm{T}}, \quad \alpha_{n-1}^{\mathrm{T}}(s) = \left[s^{n-1} \ \cdots \ s \ \ 1\right], \\ \psi = \left[a_{n-1} \ \ a_{n-2} \ \cdots \ b_0\right], \end{gathered} \tag{5.8}$$

which relates the signals $u(t)$ and $y(t)$ to the parameters of the transfer function (5.6).

The signals $\varphi(t)$ and $z(t)$ are known (can be calculated via $u(t)$ and $y(t)$), and thus equation (5.8) can be used to design the estimation laws for the parameters ψ using the gradient method. However, since in case $n > 1$ an infinite number of triplets (A, B, C) is associated with one transfer function (5.6), then in general case the parameters of matrices (A, B) of arbitrary state space representation (2.1) cannot be uniquely estimated using information only about $u(t)$ and $y(t)$. We can obtain only parameters of polynomials of the transfer function (5.8) numerator and denominator, which are directly related to the measured signals and their derivatives. Therefore, in the general case, it is also impossible to design adaptive laws using the second mechanism of stabilization of the reconstruction error $\tilde{x}(t)$ equation.

Thus, regardless of the chosen mechanism to stabilize the error equation (5.2), in general case the adaptive observer (5.1) cannot be implemented, and it is required to put forward some additional conditions that allow one to consider narrower classes of systems, for which the problem of state estimation under parametric uncertainty can be solved.

Two types of conditions are distinguished in the existing literature. The first one describes the class of systems, for which the application of the Lyapunov second method (similar to (5.3)–(5.5)) allows obtaining implementable adaptive laws. The second type of conditions describes the class of systems, for which estimation of the transfer function (5.2) parameters is sufficient to reconstruct the states of the system (2.1). Let us proceed to a detailed analysis of these requirements and the adaptive observers that can be designed when they are satisfied.

5.1. Adaptive Observers for Systems with Strictly Positive Real Transfer Function from Perturbation to Output

The idea of this approach is to obtain the conditions, under which the output estimation error $\tilde{y}(t)$ can be used in the adaptive laws instead of the unmeasured state estimation error $\tilde{x}(t)$. Namely, it turns out that such a substitution is valid for systems with strictly positive real (SPR) transfer function from perturbation to output.

The main steps of observer design for systems with strictly positive real transfer function from perturbation to output are presented in Fig. 1.

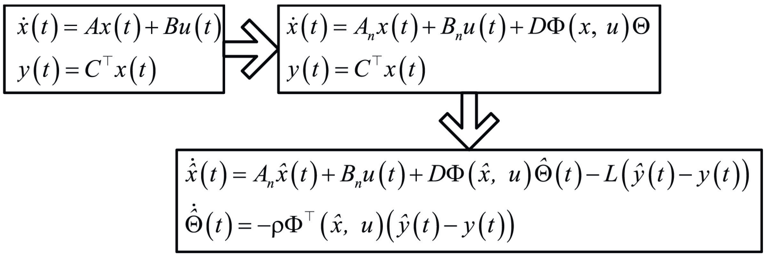

Fig. 1. Procedure of adaptive observers design for systems with strictly positive real transfer function.

As follows from Fig. 1, in the first step, the parametric perturbations of the system are represented as a new unknown input:

$$\begin{gathered}\dot{x}(t) = (A_n + \Delta A)\, x(t) + (B_n + \Delta B)\, u(t) = A_n x(t) + B_n u(t) + D\Phi(x, u)\,\Theta, \\ D\Phi(x, u)\,\Theta = \Delta A x(t) + \Delta B u(t),\end{gathered} \tag{5.9}$$

where $\Phi\colon \mathbb{R}^n \times \mathbb{R}^m \mapsto \mathbb{R}^{m\times p}$ is a linear function, $\Theta \in \mathbb{R}^p$ stands for a vector of unknown parameters.

In contrast to the parameterization (4.1) used for the invariant observers design, here the generalized perturbation is parametrized as a linear regression with unknown parameters and an unmeasured regressor.

The adaptive observer for the system (5.9) is proposed [46] as follows:

$$\begin{aligned}&\dot{\hat{x}}(t) = A_n\hat{x}(t) + B_n u(t) + D\Phi(\hat{x},\ u)\,\hat{\Theta}(t) - L(\hat{y}(t) - y(t)), \\ &\dot{\hat{\Theta}}(t) = -\rho\Phi^{\mathrm{T}}(\hat{x},\ u)(\hat{y}(t) - y(t)),\end{aligned} \tag{5.10}$$

where $\rho > 0$ is an adaptive gain, $L \in \mathbb{R}^n$ stands for a known vector that is chosen in such a way that $A_n + LC^{\mathrm{T}}$ is a Hurwitz matrix, and the adaptive law for $\hat{\Theta}(t) \in \mathbb{R}^p$ is derived using the Lyapunov second method.

The conditions of state estimates convergence using (5.10) are described in the following lemma:

Lemma 3. *Let the following conditions hold:*

1) $\|\Phi(x,\ u) - \Phi(\hat{x},\ u)\| \le \gamma\|x - \hat{x}\|,\ \gamma > 0$;
2) *there exists a solution of the following euqations for* $Q = Q^{\mathrm{T}} > 0$:

$$\begin{aligned}&\left(A_n + LC^{\mathrm{T}}\right)^{\mathrm{T}} P + P\left(A_n + LC^{\mathrm{T}}\right) = -Q, \\ &PD = C,\end{aligned}$$

and it holds that $\frac{\lambda_{\min}(Q)}{2\lambda_{\max}(P)} > \gamma$.

Then the adaptive observer (5.10) *ensures that:*

$$\lim_{t\to\infty}\|\tilde{x}(t)\| = 0, \quad \lim_{t\to\infty}\left\|D\Phi(\hat{x},\ u)\,\hat{\Theta}(t) - D\Phi(x,\ u)\,\Theta\right\| = 0.$$

Proof is given in [46].

The fact that the second premise of Lemma 3 is met allows one to use substitution of $\tilde{x}^{\mathrm{T}}(t)\,PD$ by $\tilde{y}(t) = C^{\mathrm{T}}\tilde{x}(t)$ in the estimation law and obtain an implementable adaptive law from (5.10), which is easy to check by consideration $V = \tilde{x}^{\mathrm{T}} P\tilde{x} + \rho^{-1}\tilde{\Theta}^{\mathrm{T}}\tilde{\Theta}$.

The first premise of Lemma 3 is always satisfied when a control law is linear w.r.t. states, because, owing to (5.9), the function $\Phi(x, u)$ is linear w.r.t. both inputs. According to the Kalman–Yakubovich–Popov lemma, the set of equations from the second premise has a solution if and only if the transfer function from the perturbation $\Phi(x,\ u)\,\Theta$ to the output $y(t)$ is strictly positively real. Moreover, the equation $PD = C^{\mathrm{T}}$ is solvable if PD lies in the linear span of C^{T}, which requires the output matching condition $\operatorname{rank}\left(C^{\mathrm{T}}D\right) = \operatorname{rank}(D)$ to be met, and hence also the equality $m = \dim(y(t)) = 1$.

Thus, the convergence conditions of state estimates obtained by the adaptive observer (5.10) coincide with the ones of the basic invariant Luenberger observer (4.2) and the sliding mode observer (4.7). In [47], an approach to relax the output matching condition is developed. By analogy with the solution [41], in [47], it is proposed to first obtain a new output (4.10), for which the premises of Lemma 3 are assumed to be satisfied, and then, using the new output (4.10), to implement the adaptive observer (5.5). The disadvantages of the solution [47] coincide with the ones of [41], *i.e.* the observer [47] is applicable only to the systems, for which the extended matching condition is met.

The extended and standard matching conditions are restrictive and do not allow one to design adaptive observers for systems, in which parametric uncertainties affect several equations of the system (e.g., for (4.5)).

5.2. Adaptive Observers Based on System Transformation into Observer Canonical Form

As noted earlier, the numerator and denominator parameters of the transfer function (5.3) are related to the control and output signals. This observation motivates the transformation of the system (2.1) into the observer canonical form of state space, the unknown parameters of which are precisely such transfer function parameters. For this purpose, a nonsingular (for completely observable systems) transformation $\xi(t) = Tx(t)$ is introduced as follows:

$$\begin{gathered} T^{-1} = \begin{bmatrix} A^{n-1}\mathcal{O}_n & A^{n-2}\mathcal{O}_n & \cdots & \mathcal{O}_n \end{bmatrix}, \quad \mathcal{O}_n = \mathcal{O}\begin{bmatrix} 0_{1\times(n-1)} & 1 \end{bmatrix}^{\mathrm{T}}, \\ \mathcal{O}^{-1} = \begin{bmatrix} C & A^{\mathrm{T}}C & \cdots & \left(A^{n-1}\right)^{\mathrm{T}}C \end{bmatrix}^{\mathrm{T}}, \end{gathered} \tag{5.11}$$

which, in accordance with [48, 49], allows one to rewrite equations of the system (2.1):

$$\dot{\xi}(t) = A_0\xi(t) + \psi_a y(t) + \psi_b u(t), \tag{5.12}$$

$$y(t) = C_0^{\mathrm{T}}\xi(t), \quad \xi(t_0) = Tx_0, \tag{5.13}$$

where

$$\begin{gathered} \psi_a = TAT^{-1}C_0 = \begin{bmatrix} a_{n-1} & a_{n-2} & \dots & a_0 \end{bmatrix}^{\mathrm{T}}, \\ \psi_b = TB = \begin{bmatrix} b_{n-1} & b_{n-2} & \dots & b_0 \end{bmatrix}^{\mathrm{T}}, \\ A_0 = \begin{bmatrix} 0_n & \begin{matrix} I_{n-1} \\ 0_{1\times(n-1)} \end{matrix} \end{bmatrix}, \quad C_0^{\mathrm{T}} = C^{\mathrm{T}}T^{-1} = \begin{bmatrix} 1 & 0_{n-1}^{\mathrm{T}} \end{bmatrix}. \end{gathered}$$

Then the problem of physical states estimation (2.5) is transformed into the one for the virtual states:

$$\lim_{t\to\infty}\left\|\tilde{\xi}(t)\right\| = \lim_{t\to\infty}\left\|\hat{\xi}(t) - \xi(t)\right\| = 0 \ (\exp), \tag{5.14}$$

where $\hat{\xi}(t)$ is a virtual state estimate, $\tilde{\xi}(t)$ denotes error of reconstruction/observation/estimation.

Unlike (2.5), the goal (5.14) is achievable using adaptive control techniques. For example, the estimation laws for the unknown parameters ψ_a and ψ_b can be designed on the basis of the regression equation (5.8). However, a contradiction arises between the objective (5.14) and the practical need

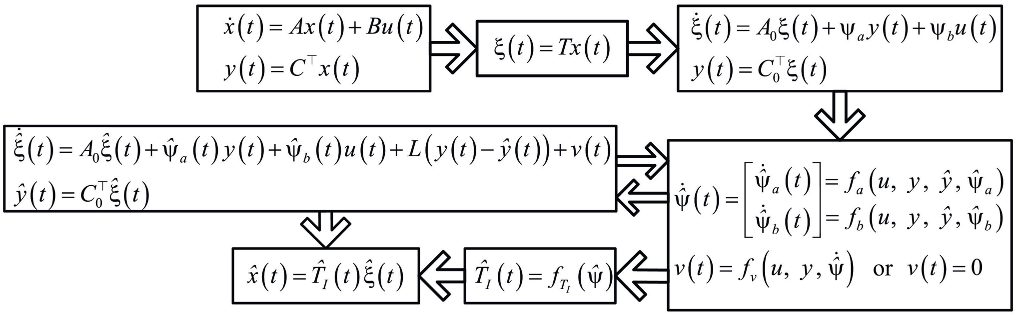

Fig. 2. Procedure of adaptive observers design on basis of system transformation into observer canonical form.

to estimate physical states to be used in the dynamic feedback (2.4a) and (2.4b). In other words, the state estimates $\hat{\xi}(t)$ are useless for *a priori* chosen feedback (2.4a) and (2.4b). Moreover, in practice, estimates of physical states $\hat{x}(t)$ are required not only for control problems, but also for other purposes (e.g., fault detection, predictive control, process monitoring, online tuning of digital twins, etc.). In order to resolve this contradiction, the physical state estimate is to be obtained as follows:

$$\hat{x}(t) = \hat{T}_I(t)\,\hat{\xi}(t), \tag{5.15}$$

where, for convenience of notation, we use the re-definition $T_I := T^{-1}$, and the estimate $\hat{T}_I(t)$ of the transformation matrix T_I is obtained on the basis of the parameter estimates of the system (5.12) via the following equation:

$$\hat{T}_I(t) = f_{T_I}\left(\hat{\psi}\right), \tag{5.16}$$

where $f_{T_I}\colon \mathbb{R}^{2n} \mapsto \mathbb{R}^n \times \mathbb{R}^n$ is a function to recalculate the transfer function parameters into the transformation matrix, $\hat{\psi}(t) = \left[\hat{\psi}_a^{\mathrm{T}}(t) \quad \hat{\psi}_b^{\mathrm{T}}(t)\right]^{\mathrm{T}} \in \mathbb{R}^{2n}$, and $\hat{\psi}_a(t)$, $\hat{\psi}_b(t)$ stand for the estimates of the parameters of the denominator and numerator of the transfer function (5.3), respectively.

Then the original goal (2.5) is equivalent to the following equalities:

$$\begin{gathered}\lim_{t\to\infty}\left\|\tilde{\xi}(t)\right\| = \lim_{t\to\infty}\left\|\hat{\xi}(t) - \xi(t)\right\| = 0 \ (\exp), \\ \lim_{t\to\infty}\left\|\tilde{\psi}(t)\right\| = \lim_{t\to\infty}\left\|\hat{\psi}(t) - \psi\right\| = 0 \ (\exp).\end{gathered} \tag{5.17}$$

In order to achieve (5.17) using the observer (5.15) + (5.16), it is required, first, the existence of a Lipschitz continuous recalculation function $f_{T_I}\colon \mathbb{R}^{2n} \mapsto \mathbb{R}^n{\times}\mathbb{R}^n \times \mathbb{R}^n$, and secondly, to design estimation laws for unknown parameters and unmeasured states, which ensure achievement of (5.17).

Block diagram of the observer (5.15) design is depicted in Fig. 2.

We first review the existing procedures to design estimation laws (functions $f_a(.)$ and $f_b(.)$ in Fig. 2). Then we show the necessity to satisfy the Lipschitz condition to achieve (2.5) when (5.17) is met and (5.16) is used.

5.2.1. Adaptive Luenberger observer. Lyapunov approach. The problem of adaptive state estimation of linear systems represented in the observer canonical form was first considered in the studies by R. Carroll and D. Lindorff [49] and G. Luders and K. Narendra [50]. It is noteworthy that these two papers, which are very close in terms of problem statement and content, were published 10 years after D. Luenberger's seminal paper [2] in the same issue of IEEE Transactions on Automatic Control with a difference of 61 pages. Later in [51] the early results of [49, 50] were revised and derived from unified perspective. Using modern notation, the observers of type [49–51] are given in [52]. In accordance with the approach under consideration, the structure of the state observer for the system (5.5) is defined in the following form:

$$\begin{aligned} \dot{\hat{\xi}}(t) &= A_0\hat{\xi}(t) + \hat{\psi}_a(t)\,y(t) + \hat{\psi}_b(t)\,u(t) + L\left(\hat{y}(t) - y(t)\right) + v(t), \\ \hat{y}(t) &= C_0^{\mathrm{T}}\hat{\xi}(t), \end{aligned} \tag{5.18}$$

where $v(t) \in \mathbb{R}^n$ is an additional signal, and the correction gain L is chosen so that to meet the condition $\lambda_{\max}\left(A_0 + LC_0^{\mathrm{T}}\right) < 0$.

The error equation between (5.18) and (5.12) is written as:

$$\begin{aligned} \dot{\tilde{\xi}}(t) &= A_m\tilde{\xi}(t) + \tilde{\psi}_a(t)\,y(t) + \tilde{\psi}_b(t)\,u(t) + v(t), \\ \tilde{y}(t) &= C_0^{\mathrm{T}}\tilde{\xi}(t), \end{aligned} \tag{5.19}$$

where $A_m = A_0 + LC_0^{\mathrm{T}}$.

The basic idea of approach from [49–51] is to transform, by means of a special choice of signal $v(t)$, the error equation (5.19) into a form such that the so-called adaptive control Lyapunov function exists.

Lemma 4. *There exists a signal vector $v(t) \in \mathbb{R}^n$ generated from measurements of known signals, for which the system* (5.19) *becomes:*

$$\begin{aligned} \dot{e}(t) &= A_m e(t) + B_c\varphi^{\mathrm{T}}(t)\,\tilde{\psi}(t), \\ \tilde{y}(t) &= C_0^{\mathrm{T}} e(t), \end{aligned}$$

where $C_0^{\mathrm{T}}(sI_n - A_m)^{-1}B_c$ is a strictly positive real (SPR) transfer function, $\varphi(t) \in \mathbb{R}^{2n}$ stands for a regressor generated on the basis of the control and output signals, $e(t) \in \mathbb{R}^n$ denotes a new state vector.

Proof is given in [52, p. 280].

By the Kalman–Yakubovich–Popov Lemma [52], the following set of equations is solvable for systems with SPR transfer function:

$$\begin{aligned} &\left(A_0 + LC_0^{\mathrm{T}}\right)^{\mathrm{T}} P + P\left(A_0 + LC_0^{\mathrm{T}}\right) = -Q, \quad Q = Q^{\mathrm{T}} > 0, \\ &PB_c = C_0, \end{aligned}$$

and, therefore, consideration of the following quadratic control Lyapunov function

$$V = e^{\mathrm{T}}Pe + \tilde{\psi}^{\mathrm{T}}\Gamma^{-1}\tilde{\psi}$$

allows one to obtain the below-given result.

Theorem 6. *The adaptive observer* (5.18) *with an adaptive law:*

$$\dot{\tilde{\psi}}(t) = \dot{\hat{\psi}}(t) = -\Gamma\tilde{y}(t)\,\varphi(t), \quad \Gamma = \Gamma^{\mathrm{T}} > 0, \tag{5.20}$$

in case $\xi \in L_\infty$ and $u \in L_\infty$, ensures that:

(i) *all signals are bounded,*

(ii) $\lim\limits_{t\to\infty} |\tilde{y}(t)| = 0$,

(iii) $\dot{\hat{\psi}} \in L_2 \cap L_\infty$ *and* $\lim\limits_{t\to\infty} \left\|\dot{\hat{\psi}}(t)\right\| = 0$.

In addition, if for all $t \geq t_0$ there exist $T > 0$ and $\alpha > 0$, such that[3]

$$\int\limits_{t}^{t+T} \varphi(\tau)\,\varphi^{\mathrm{T}}(\tau)\,d\tau \geq \alpha > 0,$$

then the following equalities hold:

$$\lim_{t\to\infty}\left\|\tilde{\psi}(t)\right\| = 0 \ (\exp), \quad \lim_{t\to\infty}\left\|\tilde{\xi}(t)\right\| = 0 \ (\exp).$$

Proof is given in [52, p. 282].

As for the considered solution, the equation of the observer (5.18) dynamics differs in the course of transients from the dynamics of the system (5.12) due to the additional signal $v(t)$, which is asymptotically decaying but imposes significant distortion into the estimates $\hat{\xi}(t)$. In addition, the adaptive law (5.18) for the observer parameters cannot be chosen arbitrarily but is derived using the control Lyapunov function, and therefore, it does not provide any guarantees of the transient quality for the parametric error $\tilde{\psi}(t)$ (e.g., it does not guarantee monotonicity of the error vector by norm or elements).

5.2.2. Adaptive Luenberger observer. Estimation-based approach. To eliminate the coupling between the dynamics of the estimator (5.20) and the observer (5.18) through $\tilde{y}(t)$, and to improve the transient quality of error $\tilde{\psi}(t)$, an estimation-based approach to design adaptive state observers [52] was proposed. The observer structure is chosen as (5.18), but the additional signal is eliminated from the equations by choosing $v(t) \overset{\Delta}{=} 0$, and the estimation laws are derived without Lyapunov functions. The basis for the design of the estimation law is the regression equation (5.8) obtained using dynamic filters.

The regression equation (5.8) allows one to use a wide class of methods from the parameter estimation theory (gradient method, variations of the least squares method, etc. [52]) to estimate the parameters of the transfer function (5.6). In particular, in case the gradient method is applied, the following theorem holds.

Theorem 7. *The adaptive observer* (5.18) *with* $v(t) \overset{\Delta}{=} 0$ *and the adaptive law:*

$$\begin{aligned}\dot{\tilde{\psi}}(t) = \dot{\hat{\psi}}(t) &= -\Gamma\varphi(t)\left(\varphi^{\mathrm{T}}(t)\,\hat{\psi}(t) - z(t)\right)\\ &= -\Gamma\varphi(t)\left(\varphi^{\mathrm{T}}(t)\,\hat{\psi}(t) - \varphi^{\mathrm{T}}(t)\,\psi\right), \quad \Gamma = \Gamma^{\mathrm{T}} > 0\end{aligned} \tag{5.21}$$

in case $\xi \in L_\infty$ and $u \in L_\infty$, ensures that:

(i) *all signals are bounded,*

(ii) $\lim\limits_{t\to\infty} |\tilde{y}(t)| = 0$,

(iii) $\dot{\tilde{\psi}} \in L_2 \cap L_\infty$ *and* $\lim\limits_{t\to\infty}\left\|\dot{\tilde{\psi}}(t)\right\| = 0$.

In addition, if the condition $\varphi \in \mathrm{PE}$ is met, then it holds that

$$\lim_{t\to\infty}\left\|\tilde{\psi}(t)\right\| = 0 \ (\exp), \quad \lim_{t\to\infty}\left\|\tilde{\xi}(t)\right\| = 0 \ (\exp).$$

Proof is given in [52, p. 271].

As it follows from comparison of the premises and statements of Theorems 6 and 7, the properties of the observer (5.18) + (5.20) derived on the basis of adaptive control Lyapunov function, and the one (5.18) + (5.21) based on estimation approach coincide to each other. However, estimation laws

[3] This requirement is a regressor persistent excitation condition, which is further denoted as $\varphi \in \mathrm{PE}$.

for $\hat{\psi}(t)$ and $\hat{\xi}(t)$ are decoulped in (5.18) + (5.21), and additional signal $v(t)$, which distorts the estimate $\hat{\xi}(t)$, is not used. Moreover, as it can be easily checked by consideration of the quadratic form $V = \tilde{\psi}^{\mathrm{T}}\Gamma^{-1}\tilde{\psi}$, unlike (5.20), in case we choose $\Gamma = \gamma I_{2n} > 0$, the estimation law (5.21) ensures monotonicity of the norm of the error $\tilde{\psi}(t)$:

$$\left\|\tilde{\psi}(t_a)\right\| \leq \left\|\tilde{\psi}(t_b)\right\| \ \forall t_a \geq t_b,$$

which, in comparison with (5.20), allows one to improve the transient quality of the estimates $\hat{\psi}(t)$ and $\hat{\xi}(t)$.

In addition, the considered estimation-based approach and parameterization (5.8) open a wide range of possibilities for the design of adaptive state observers with relaxed regressor excitation requirements and accelerated/improved convergence. These solutions will be discussed later in Section 5.2.6.

5.2.3. Extended Kalman filter. Joint estimation of the states and unknown parameters of the system (5.12) can also be achieved using the extended Kalman filter [3]. To apply it, the system is rewritten in the following form:

$$\begin{gathered}\dot{\xi}_e(t) = \begin{bmatrix} A_0 & \Psi(t) \\ 0_{2n\times n} & 0_{2n\times 2n}\end{bmatrix}\xi_e(t) = A_e(t)\,\xi_e(t),\\ y(t) = C_e^{\mathrm{T}}\xi_e(t) = \begin{bmatrix}1 & 0_{3n-1}\end{bmatrix}\xi_e(t),\end{gathered} \tag{5.22}$$

where $\xi_e(t) = \begin{bmatrix}\xi(t) & \psi\end{bmatrix}^{\mathrm{T}}$ is an extended state vector, $\Psi(t) = \begin{bmatrix}y(t)\,I_n & u(t)\,I_n\end{bmatrix} \in \mathbb{R}^{n\times 2n}$ stands for a measurable signal, $A_e(t) \in \mathbb{R}^{3n\times 3n}$ denotes a known time-varying matrix.

An extended Kalman-Bucy filter [3] for the system (5.22) is defined as follows:

$$\begin{aligned}&\dot{\hat{\xi}}_e(t) = A_e(t)\,\hat{\xi}_e(t) + L(t)\left(y(t) - C_e^{\mathrm{T}}\hat{\xi}_e(t)\right),\\ &L(t) = P(t)\,C_e R^{-1},\\ &\dot{P}(t) = A_e(t)\,P(t) + P(t)\,A_e^{\mathrm{T}}(t) + Q - P(t)\,C_e R^{-1} C_e^{\mathrm{T}} P(t),\end{aligned} \tag{5.23}$$

where $R^{-1} \in \mathbb{R}$, $Q \in \mathbb{R}^{3n\times 3n}$ stand for covariances of disturbances that affect the state and output equations, respectively.

Using the Kalman filtering algorithm (5.23), the condition for exponential convergence of the estimates $\hat{\xi}_e(t)$ to their true values is [10, 52] a uniform complete observability of the extended system (5.22).

Definition 2. A pair $(C_e, A_e(t))$ is uniformly completely observable, if for all $t \geq t_0$ there exist scalars α_1, α_2, $T > 0$ such that the following inequality holds:

$$\alpha_2 I_{3n} \geq \int\limits_t^{t+T} \Phi(\tau,\ t) C_e C_e^{\mathrm{T}} \Phi(\tau,\ t)\, d\tau \geq \alpha_1 I_{3n}, \tag{5.24}$$

where $\Phi(t, t_0)$ is a fundamental matrix of the system (5.22), which is defined as:

$$\dot{\Phi}(t,\ t_0) = A_e(t)\,\Phi(t,\ t_0), \quad \Phi(t_0,\ t_0) = I_{3n}.$$

As follows from the definition, the Kalman filter (5.23) requires uniform nonsingularity of the grammian (5.24) with dimension $3n \times 3n$, while the previously considered adaptive observers require nonsingularity of some integral of dimension $2n \times 2n$. So, in general, the convergence conditions of the adaptive observers (5.18) + (5.20) and (5.18) + (5.21) are (if not the same) weaker than the ones of the extended Kalman filter (5.23). Following the review [54], it can also be shown by means of some extensive calculations that the nonsingularity of (5.24) always requires to meet the persistent excitation condition for the regressor $\varphi(t)$ from the parameterization (5.8).

5.2.4. Algebraic observers. Kreisselmeier parametrization. The above-considered adaptive obser-vers form the state estimates using differential equations, the right-hand side of which depends on the estimates of unknown parameters and contains the Luenberger correction term. Therefore, transients of the system parameters estimates affect significantly the transient quality of the state estimates, and the correction term, which is a feedback for the observer, worsens the situation even more, causing the peak effect. In order to eliminate the noted problems, G. Kreisselmeier proposed to obtain the state estimates with the help of special parameterizations using not differential but algebraic equation, which does not require to use the correction term.

The essence of the approach is as follows. A set of differential filters for control and output signals is introduced:

$$\begin{aligned} \dot{\Omega}(t) &= A_K^{\mathrm{T}}\Omega(t) + C_0 u(t), \quad \Omega(t_0) = 0_n, \\ \dot{P}(t) &= A_K^{\mathrm{T}}P(t) + C_0 y(t), \quad P(t_0) = 0_n, \end{aligned} \tag{5.25}$$

where $A_K = A_0 + KC_0^{\mathrm{T}}$ is a Hurwitz matrix.

According to the results from [55], the following holds:

$$\begin{gathered} \xi(t) = H^{\mathrm{T}}(t)\psi + e^{A_K(t-t_0)}\xi(t_0), \\ \psi^{\mathrm{T}} = \begin{bmatrix}\psi_a^{\mathrm{T}} - K^{\mathrm{T}} & \psi_b^{\mathrm{T}}\end{bmatrix}, \quad H^{\mathrm{T}}(t) = [h_1(t)\ldots h_{2n}(t)], \\ h_i(t) = F_i P(t), \quad h_{i+n}(t) = F_i\Omega(t), \quad i = \overline{1,\ n}, \end{gathered} \tag{5.26}$$

where $H(t) \in \mathbb{R}^{2n\times n}$ is a measurable regressor, $F_i \in \mathbb{R}^{n\times n}$ stands for a transformation matrix, which is composed of the parameters of the numerator polynomial of the vector function $(sI - A_K)^{-1}e_i$, $e_i \in \mathbb{R}^n$ denotes a vector, which i^{th} element equals to one.

The parametrization (5.26) motivates to estimate the unmeasurable state of the system (5.12) with the help of an algebraic equation:

$$\hat{\xi}(t) = H^{\mathrm{T}}(t)\hat{\psi}(t), \tag{5.27}$$

where the estimates $\hat{\psi}(t)$ are obtained from the regression equation:

$$\begin{aligned} y(t) &= C_0^{\mathrm{T}}\xi(t) = \varphi(t)\psi + C_0^{\mathrm{T}}e^{A_K(t-t_0)}\xi(t_0), \\ \varphi(t) &= C_0^{\mathrm{T}}H^{\mathrm{T}}(t) \end{aligned} \tag{5.28}$$

with the help of a broad range of the parameter estimation laws.

For example, in case $\xi \in L_\infty$, $u \in L_\infty$ and the condition $\varphi \in \mathrm{PE}$ is met, a gradient estimation law

$$\dot{\tilde{\psi}}(t) = \dot{\hat{\psi}}(t) = -\Gamma\varphi(t)\left(\varphi^{\mathrm{T}}(t)\hat{\psi}(t) - y(t)\right), \quad \Gamma = \Gamma^{\mathrm{T}} > 0 \tag{5.29}$$

ensures that

$$\lim_{t\to\infty}\left\|\tilde{\psi}(t)\right\| = 0\ (\exp), \quad \lim_{t\to\infty}\left\|\tilde{\xi}(t)\right\| = 0\ (\exp).$$

Here it should be noted that, owing to the fact that A_K is a Hurwitz matrix, an exponentially decaying term in (5.26) and (5.28) does not affect the asymptotic properties of the estimation errors, but limits the maximal convergence rate [52]. In contrast to the previously considered observers, in case of the algebraic state reconstruction (5.27) + (5.29), the transient quality improvement for $\tilde{\psi}(t)$ results in uniform transient quality improvement for estimates $\hat{\xi}(t)$ (parameter estimates are not integrated in the observer equation, and there is no correction term).

5.2.5. Algebraic observers. Alternative parametrization. The computation of the regressor $H(t)$ to implement the parameterization (5.26) requires a rather laborious calculation of the matrices F_i, especially in case if n grows. A simplified parameterization is proposed in [56], [57]. Instead of two filters with vector states (5.25), it is proposed to use three filters—two with matrix state and one with vector state:

$$\begin{aligned} \dot{\chi}(t) &= A_K\chi(t) + Ky(t), \quad \chi(t_0) = 0_{n\times 1}, \\ \dot{P}(t) &= A_K P(t) + I_n y(t), \quad P(t_0) = 0_{n\times n}, \\ \dot{\Omega}(t) &= A_K\Omega(t) + I_n u(t), \quad \Omega(t_0) = 0_{n\times n}. \end{aligned} \tag{5.30}$$

Then, according to [56, 57], the parameters ψ satisfy the following linear regression equation:

$$\begin{gathered} \xi(t) = \chi(t) + H^{\mathrm{T}}(t)\psi + e^{A_K(t-t_0)}\xi(t_0), \\ H^{\mathrm{T}}(t) = \begin{bmatrix}\Omega(t) & P(t)\end{bmatrix} \in \mathbb{R}^{n\times 2n}. \end{gathered} \tag{5.31}$$

Equation (5.31) can be easily verified by direct differentiation of the error $\xi(t) - \chi(t) + H^{\mathrm{T}}(t)\psi$. Similar to (5.28), the multiplication of (5.31) by C_0^{T} yields:

$$\begin{aligned} &z(t) = y(t) - C_0^{\mathrm{T}}\chi(t) = \varphi^{\mathrm{T}}(t)\psi + C_0^{\mathrm{T}}e^{A_K(t-t_0)}\xi(t_0), \\ &\varphi^{\mathrm{T}}(t) = C_0^{\mathrm{T}}H^{\mathrm{T}}(t), \end{aligned} \tag{5.32}$$

which allows one to implement the following observer:

$$\hat{\xi}(t) = \chi(t) + H^{\mathrm{T}}(t)\hat{\psi}(t), \tag{5.33}$$

where $\hat{\psi}(t)$ can be obtained on the basis of (5.32) with the help of a broad class of parameter estimation laws.

5.2.6. Observers with accelerated/improved convergence. One of the main advantages of algebraic adaptive observers (5.27) and (5.33) over differential type adaptive observers (5.18) + (5.20), (5.18) + (5.21) is the direct dependence of the convergence rate of the observation error $\xi(t)$ on the convergence rate of the parametric error and the filter parameters (5.25) or (5.30).

Indeed, when $\xi \in L_\infty$ and $u \in L_\infty$, it holds that

$$\left\|\tilde{\xi}(t)\right\| \le H_{\max}\left\|\tilde{\psi}(t)\right\| + e^{A_K(t-t_0)}\xi_{\max},$$

which allows one to adjust the convergence rate of $\tilde{\xi}(t)$ by choosing A_K and increasing the rate of convergence of $\tilde{\psi}(t)$.

At the same time, the rate of convergence of the error from the error equation (5.19) is known to be defined not only by the rate of convergence of its input, but also by the eigenvalues of the matrix of the closed-loop system $A_0 + L_0C^{\mathrm{T}}$. In this case, the increase (in terms of absolute value) of the eigenvalues real part can lead to significant peaks in the course of transient process.

When $\Gamma = \gamma I_{2n}$ and the condition of regressor persistent excitation is met, the convergence rate of the previously discussed gradient estimation law of type (5.21) or (5.29), according to the results of [58], can be written in the following form:

$$\begin{gathered} \left\|\tilde{\psi}(t)\right\| \le \sqrt{a}\,\|\varphi\|\, e^{-\frac{1}{2}\gamma a^{-1}(t-t_0)}\left\|\tilde{\psi}(t_0)\right\|, \\ a = \gamma b^{-1}e^{2bT}, \quad b = -\frac{1}{2T}\ln\left(1 - \frac{\gamma\alpha}{1+\gamma^2T^2\|\varphi\|_\infty^4}\right). \end{gathered}$$

Thus, using conventional gradient estimation laws, the convergence rate of the parametric error cannot be arbitrarily increased by raising the gain $\gamma > 0$, since the product γa^{-1} may decrease as γ increases. Therefore, in fact, using algebraic observers (5.27) and (5.33) with conventional gradient estimation laws of the form (5.21), (5.29), as well as observers (5.18) + (5.20) or (5.18) + (5.21), it is not easy to increase the convergence rate of the unmeasured state estimation without deterioration of the transient quality. To solve this problem, modified parameter estimation laws with accelerated convergence have been proposed. Without loss of generality, and for simplicity, we illustrate the properties of these improved estimation laws using the estimation problem with parametrization (5.32). Here, we temporarily omit the term $C_0^{\mathrm{T}} e^{A_K(t-t_0)}\xi(t_0)$ but acknowledge its constraints on the achievable convergence rate.

1) Kreisselmeier's regressor extension scheme. G. Kreisselmeier proposed [55] to transform the vector regressor $\varphi(t) \in \mathbb{R}^{2n}$ from the regression equation (5.32) into a symmetric positive-semidefinite matrix $\Phi(t) \in \mathbb{R}^{2n\times 2n}$, $\Phi(t) = \Phi^{\mathrm{T}}(t) \geq 0$ with the help of the following filtering ($l > 0$):

$$\begin{aligned} \dot{Y}(t) &= -lY(t) + \varphi(t) z(t), \quad Y(t_0) = 0_{2n}, \\ \dot{\Phi}(t) &= -l\Phi(t) + \varphi(t)\varphi^{\mathrm{T}}(t), \quad \Phi(t_0) = 0_{2n\times 2n}, \end{aligned} \tag{5.34}$$

which, owing to $\psi = \text{const}$, allows one to obtain a new regression equation:

$$Y(t) = \Phi(t)\psi. \tag{5.35}$$

Based on equation (5.35), a new estimation law is designed:

$$\dot{\tilde{\psi}}(t) = \dot{\hat{\psi}}(t) = -\Gamma\left(\Phi(t)\hat{\psi}(t) - Y(t)\right), \quad \Gamma = \Gamma^{\mathrm{T}} > 0, \tag{5.36}$$

which has significantly different properties in comparison with the previously considered laws [59]:

1) $\lambda_{\min}(\Phi(t)) \notin L_1 \Leftrightarrow \lim\limits_{t\to\infty}\left\|\tilde{\psi}(t)\right\| = 0$;
   $\varphi \in \mathrm{PE} \Leftrightarrow \lim\limits_{t\to\infty}\left\|\tilde{\psi}(t)\right\| = 0 \ (\exp)$;
2) $\left\|\tilde{\psi}(t_a)\right\| \leq \left\|\tilde{\psi}(t_b)\right\| \ \forall t_a \geq t_b$,
3) if $\varphi \in \mathrm{PE}$ and $\Gamma = \gamma I_{2n}$, then the rate of exponential convergence of the parametric error $\tilde{\psi}(t)$ can be made arbitrarily fast by increasing γ.

The most important one is the third property of the law (5.36), which is explained by the property

$$\varphi \in \mathrm{PE} \Leftrightarrow \forall t \geq kT \ \lambda_{\min}(\Phi(t)) > \mu > 0,$$

proved in [60] and allowed one to obtain the following bound of the parametric error norm:

$$\frac{d}{dt}\left\|\tilde{\psi}\right\|^2 = -2\gamma\tilde{\psi}^{\mathrm{T}}\Phi\tilde{\psi} \leq -2\gamma\lambda_{\min}(\Phi)\left\|\tilde{\psi}\right\|^2$$

$$\Downarrow$$

$$\left\|\tilde{\psi}(t)\right\| \leq e^{-\gamma\int\limits_{t_0}^{t}\lambda_{\min}(\Phi)d\tau}\left\|\tilde{\psi}(t_0)\right\| \leq e^{-\gamma\mu(t-t_0)}\left\|\tilde{\psi}(t_0)\right\|,$$

from which we immediately have the third property.

Considering estimation law (5.36), the filtering operation (5.34) allows, in addition to current indirect information about system parameters, to use a certain amount of historical data (determined by parameter l). This enables the system to have indirect information (via $Y(t)$ and $\Phi(t)$) about all unknown parameters at each time instant, which in turn allows one to increase the convergence rate by means of increasing γ. Thus, when using the algebraic state observer (5.33) augmented with the accelerated convergence estimator (5.36), it becomes possible to enhance the exponential convergence rate of state reconstruction errors by increasing γ. It should also be noted that, unlike previously considered observers, the observer (5.33) + (5.36) does not require the persistent excitation condition to be satisfied to achieve asymptotic (though not exponential) convergence of the estimation error.

2) Dynamic regressor extension and mixing procedure. In [61], it is proposed to use a time-varying adaptive gain for the law (5.36):[4]

$$\Gamma = \gamma \operatorname{adj}\{\Phi(t)\}, \quad \gamma > 0, \tag{5.37}$$

which, using the property $\operatorname{adj}\{\Phi(t)\}\Phi(t) = \det\{\Phi(t)\} I_{2n\times 2n}$, allows one to obtain:

$$\begin{gathered}\dot{\hat{\psi}}(t) = \dot{\tilde{\psi}}(t) = -\gamma \operatorname{adj}\{\Phi(t)\}\Big(\Phi(t)\hat{\psi}(t) - Y(t)\Big)\\ = -\gamma\Big(\Delta(t)\hat{\psi}(t) - \mathcal{Y}(t)\Big)\\ \Downarrow\\ \dot{\hat{\psi}}_i(t) = \dot{\tilde{\psi}}_i(t) = -\gamma\Delta(t)\tilde{\psi}_i(t),\\ \Delta(t) = \det\{\Phi(t)\}, \quad \mathcal{Y}(t) = \operatorname{adj}\{\Phi(t)\}Y(t) = \Delta(t)\psi,\end{gathered} \tag{5.38}$$

as a result, the new law (5.38) has the properties 1)–3) of (5.36) and additionaly ensures element-wise monotonicity of the parametric error:

$$\left|\tilde{\psi}_i(t_a)\right| \leq \left|\tilde{\psi}_i(t_b)\right| \; \forall t_a \geq t_b, \quad i \in \{1,\ldots,n\}. \tag{5.39}$$

As it will be shown further, the property (5.39) is essential for recalculation via (5.16) of the estimates $\hat{\psi}(t) = \begin{bmatrix}\hat{\psi}_a(t) & \hat{\psi}_b(t)\end{bmatrix}^{\mathrm{T}}$ into the estimate of the transformation matrix $\hat{T}(t)$. Indeed, having (5.39) at hand, it holds that:

$$\left.\begin{array}{l}\hat{\psi}_i(t_0) \in \Xi\\ \psi_i \in \Xi\end{array}\right\} \Rightarrow \hat{\psi}_i(t) \in \Xi, \tag{5.40}$$

which, in some cases, given certain *a priori* information about the system parameters (knowledge that $\psi_i \in \Xi$), allows one to avoid discontinuities in the transformation (5.15). The role of property (5.40) will be discussed in more detail further.

3) Relaxation of Exponential Convergence Condition. Integral Dynamic Regressor Extension and Mixing Procedure. The considered estimation laws (5.20), (5.21), (5.29), (5.36), or (5.38) ensure exponential convergence of the parametric error and estimation error under the condition of regressor persistent excitation (nonsingularity of integral of dimension $2n \times 2n$), which typically requires the control input to contain at least n distinct frequencies [52, 63]. In practice, in order to meet such condition, non-decaying harmonic dither signals are required to be added to the control signal, which may contradict the original control objective.

Thus, to relax the convergence conditions for state and parameter estimates to their true values, estimation laws with weaker regressor excitation requirements have been proposed in the literature. One such condition is the finite excitation requirement ($\varphi \in \mathrm{FE}$).[5] Unlike regressor persistent

[4] It should be noted that a similar regressor scalarization procedure was proposed ten years earlier in a less cited paper [62].

[5] The condition $\varphi \in \mathrm{FE}$ means that there exist $t_e > t_r \geq t_0$ and $\alpha > 0$ such that $\int\limits_{t_r}^{t_e} \varphi(\tau)\varphi^{\mathrm{T}}(\tau)\,d\tau \geq \alpha > 0$.

excitation, the finite excitation condition only requires the nonsingularity of a certain integral over a finite time interval rather than the entire time axis. In terms of convergence conditions of the Kalman–Bucy filter (5.23), this relaxed condition is equivalent to observability of the extended system (5.22) only over a specific time interval [3]. The finite excitation condition can be satisfied by addition of an exponentially decaying harmonic signal with n distinct frequencies to the control input, which does not prevent the achievement of the asymptotic control objective. For many systems, this condition is met even in case of complete absence of dither signals.

To relax the persistent excitation requirement, numerous approaches have been proposed to handle the regressor and regressand. A detailed overview of such methods is provided in [64–67]. The core idea of most approaches is to "memorize" in a special way the indirect information about unknown parameters obtained during the finite excitation interval.

In [68], instead of (5.34), it is proposed to use the following filtering ($\sigma > 0$):

$$\begin{aligned}\dot{Y}(t) &= e^{-\sigma(t-t_0)}\varphi(t)\,z(t), \quad Y(t_0) = 0_{2n},\\ \dot{\Phi}(t) &= e^{-\sigma(t-t_0)}\varphi(t)\,\varphi^{\mathrm{T}}(t), \quad \Phi(t_0) = 0_{2n\times 2n},\end{aligned} \tag{5.41}$$

which allows one to obtain a new regression equation of the form (5.35), for which regressor, in case $\varphi \in \mathrm{FE}$, it holds that:

$$\begin{aligned}\Phi(t) &= \int_{t_0}^{t} e^{-\sigma(\tau-t_0)}\varphi(\tau)\,\varphi^{\mathrm{T}}(\tau)\,d\tau \geq \int_{t_0}^{t_e} e^{-\sigma(\tau-t_0)}\varphi(\tau)\,\varphi^{\mathrm{T}}(\tau)\,d\tau\\ &\geq e^{-\sigma(t_e-t_0)}\int_{t_0}^{t_e}\varphi(\tau)\,\varphi^{\mathrm{T}}(\tau)\,d\tau \geq \alpha e^{-\sigma(t_e-t_0)}I_{2n} > 0.\end{aligned} \tag{5.42}$$

Based on equation (5.35) obtained by filtering (5.41), an estimation law with matrix (5.36) or scalar (5.38) regressor can be implemented. The properties of these laws differ only by the type of monotonicity of the parametric error (by norm or element-wise):

1) $\varphi \in \mathrm{FE} \Leftrightarrow \lim\limits_{t\to\infty}\left\|\tilde{\psi}(t)\right\| = 0 \ (\exp)$;
2) (5.41) + (5.36)$\Rightarrow \left\|\tilde{\psi}(t_a)\right\| \leq \left\|\tilde{\psi}(t_b)\right\| \ \forall t_a \geq t_b$
   (5.41) + (5.38)$\Rightarrow \left|\tilde{\psi}_i(t_a)\right| \leq \left|\tilde{\psi}_i(t_b)\right| \ \forall t_a \geq t_b,\ i \in \{1,\ldots,n\}$,
3) if $\varphi \in \mathrm{FE}$ and $\Gamma = \gamma I_{2n}$, then the rate of exponential convergence of the parametric error $\tilde{\psi}(t)$ can be made arbitrarily fast by increasing γ.

As it follows from the comparison of properties of the laws (5.34) + (5.36), (5.34) + (5.38) and (5.41) + (5.36), (5.41) + (5.38), the second group inherits the advantages of the first one but relaxes the requirement necessary to achieve the goals (2.5) and (5.17).

Relaxation of the requirement of the regressor persistent excitation allows one to extend the estimation problem for parameterizations (5.26) and (5.31) by taking into consideration an exponentially decaying term:[6]

$$\begin{gathered}\xi(t) = H^{\mathrm{T}}(t)\,\psi_e,\\ \psi_e^{\mathrm{T}} = \begin{bmatrix}\psi_a^{\mathrm{T}} - K^{\mathrm{T}} & \psi_b^{\mathrm{T}} & \xi^{\mathrm{T}}(t_0)\end{bmatrix}, \quad H^{\mathrm{T}}(t) = \begin{bmatrix}h_1(t)\ldots h_{2n}(t) & e^{A_K(t-t_0)}\end{bmatrix},\\ \xi(t) = \chi(t) + H^{\mathrm{T}}(t)\,\psi_e,\\ H^{\mathrm{T}}(t) = \begin{bmatrix}\Omega(t) & P(t) & e^{A_K(t-t_0)}\end{bmatrix},\end{gathered}$$

[6] Extension of the estimation problem without relaxation of the persistent excitation requirement is impossible, as regressors containing a decaying term do not satisfy this condition.

which, in case of application of the corresponding parameter estimation law for ψ_e, results in the following state reconstruction error bound:

$$\left\|\tilde{\xi}\left(t\right)\right\| \leq H_{\max}\left\|\tilde{\psi}_e\left(t\right)\right\|$$

and eliminates the constraints that the eigenvalues of matrix A_K impose on the error $\tilde{\xi}\left(t\right)$ convergence rate, allowing its adjustment using only a single scalar parameter γ.

4) Relaxation of the regressor persistent excitation requirement. Estimation algorithm with finite-time convergence. An alternative estimation law [69] with relaxed regressor excitation requirements ensures finite-time convergence of parameter/state estimation errors and have the form:

$$\begin{gathered}
\hat{\psi}_i^{\mathrm{FTC}}\left(t\right) = \frac{1}{1-\phi_c\left(t\right)}\left[\hat{\psi}_i\left(t\right) - \phi_c\left(t\right)\hat{\psi}_i\left(t_0\right)\right], \\
\phi_c\left(t\right) = \begin{cases} \sigma, & \text{if} \quad \phi \geq \sigma \\ \phi\left(t\right), & \text{if} \quad \phi < \sigma, \end{cases} \qquad \dot{\phi}\left(t\right) = -\gamma\Delta^2\left(t\right)\phi\left(t\right), \quad \phi\left(t_0\right) = 1.
\end{gathered} \tag{5.43}$$

In [69], the convergence of $\hat{\psi}_i^{\mathrm{FTC}}\left(t\right)$ to ψ_i in finite time t_e is proved in case the following condition holds:

$$\int\limits_{t_0}^{t_e}\Delta^2\left(\tau\right)d\tau \geq -\frac{1}{\gamma_i}\ln\left(\sigma\right). \tag{5.44}$$

In order to explain the relaxation method, a solution of equation (5.38) is written:

$$\begin{gathered}
\tilde{\psi}_i\left(t\right) = \phi\left(t\right)\tilde{\psi}_i\left(t_0\right) \\
\Downarrow \\
\hat{\psi}_i\left(t\right) - \phi\left(t\right)\hat{\psi}_i\left(t_0\right) = \left[1-\phi\left(t\right)\right]\psi_i.
\end{gathered} \tag{5.45}$$

Since, owing to $\Delta^2\left(t\right) \geq 0$, $\phi\left(t\right)$ is a non-increasing function of time, then, when the condition (5.44) is met, the logical operator (5.43) switches at time instant t_e, and, following (5.45), the unknown parameters values are determined analytically. In parameterization (5.45), the regressor $1-\phi\left(t\right)$ and regressand $\hat{\psi}_i\left(t\right) - \phi\left(t\right)\hat{\psi}_i\left(t_0\right)$ play the roles of $\Phi\left(t\right)$ and $Y\left(t\right)$ from parametrization (5.35) + (5.41). When the condition (5.44) is met, the regressor $1-\phi\left(t\right)$ is globally bounded away from zero starting from some time instant, *i.e.*, indirect information about all unknown parameters has been obtained and preserved, enabling their analytical calculation. Consequently, the counterpart to law (5.43) in terms of parameterization (5.41) is the following one:

$$\hat{\psi}^{\mathrm{FTC}}\left(t\right) = \begin{cases} \hat{\psi}\left(t_0\right), & \text{if} \quad \det\left\{\Phi\left(t\right)\right\} < \rho \leq \det\left\{\alpha e^{-\sigma t_e}I_{2n}\right\} \\ \Phi^{-1}\left(t\right)Y\left(t\right), & \text{otherwise} \end{cases} \tag{5.46}$$

which, owing to the fact that inequality (5.42) holds, also allows one to estimate the parameters in finite-time in case of proper choice of ρ. An open problem of finite-time convergence laws (5.43) or (5.46) is their high sensitivity to parameters σ, γ_i, ρ—parametric convergence (or its absence in case of poor choice) depends entirely on the selection of these values.

5.2.7. Problem of singularities of recalculation (5.16). Having considered in detail most of the existing algorithms and procedures to estimate the virtual states $\xi\left(t\right)$ of the observer canonical form (5.12) and the parameters ψ_a, ψ_b of the numerator and denominator of the transfer function (5.6), we return to the design problem of the observer (5.15) + (5.16) of the physical states of the original system (2.1).

First, in order to obtain estimates (5.15), it is necessary to have an equation (5.16) to recalculate parameters of the transfer function (5.6) into elements of the matrix T_I. Generally speaking, complete observability of the system is a necessary, but not a sufficient condition for the existence of such an equation.

Example 1. Consider a completely observable (in case $\theta_2 \neq 0$) system:

$$A = \begin{bmatrix} \theta_1 & \theta_2 \\ \theta_3 & \theta_4 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \theta_5 \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

In accordance with the transformation (5.11), the parameters of the numerator and denominator of the transfer function (5.6) are defined as follows:

$$\psi_a = \begin{bmatrix} \theta_1 + \theta_4 \\ \theta_2\theta_3 - \theta_1\theta_4 \end{bmatrix}, \quad \psi_b = \begin{bmatrix} 0 \\ \theta_2\theta_5 \end{bmatrix}, \tag{5.47}$$

and the transformation T_I is written as:

$$T_I = \begin{bmatrix} 1 & 0 \\ \theta_4\theta_2^{-1} & \theta_2^{-1} \end{bmatrix}.$$

A set of algebraic equations (5.47) has no solution, and therefore, it is impossible to implement (5.16) despite complete observability of the system. ■

Secondly, even if the recalculation equation (5.16) exists, then exponential convergence of the physical state reconstruction error $\tilde{x}(t)$ follows from exponential convergence of $\tilde{\psi}(t)$ and $\tilde{\xi}(t)$ if the function $f_{T_I}\colon \mathbb{R}^{2n} \mapsto \mathbb{R}^n \times \mathbb{R}^n$ is Lipschitz continuous.

Lemma 5. *Let $\tilde{\psi}(t)$ and $\tilde{\xi}(t)$ converge exponentially to zero, then the limit* (2.5) *exists, if for all a, $b \in \mathbb{R}^{2n}$ there exists a scalar $\mu > 0$ such that:*

$$\left\| f_{T_I}(a) - f_{T_I}(b) \right\| \leq \mu \left\| a - b \right\|. \tag{5.48}$$

Proof. The following error equation is written:

$$\begin{aligned} \tilde{x}(t) &= \hat{T}_I(t)\hat{\xi}(t) - T_I\xi(t) \pm \hat{T}_I(t)\xi(t) = \hat{T}_I(t)\tilde{\xi}(t) + \tilde{T}_I(t)\xi(t) \pm T_I\tilde{\xi}(t) \\ &= \tilde{T}_I(t)\tilde{\xi}(t) + T_I\tilde{\xi}(t) + \tilde{T}_I\xi(t), \end{aligned}$$

Using the condition (5.48), an upper bound is obtained:

$$\begin{aligned} \left\|\tilde{x}(t)\right\| &\leq \left\|\tilde{T}_I\right\| \left\|\tilde{\xi}(t)\right\| + \left\|T_I\right\| \left\|\tilde{\xi}(t)\right\| + \left\|\tilde{T}_I(t)\right\| \left\|\xi(t)\right\| \\ &\leq \left(\left\|T_I\right\| + \mu\left\|\tilde{\psi}(t)\right\|\right)\left\|\tilde{\xi}(t)\right\| + \mu\left\|\tilde{\psi}(t)\right\|\left\|\xi(t)\right\|, \end{aligned}$$

from which we have exponential convergence of the error $\tilde{x}(t)$ in case of exponential convergence of $\tilde{\psi}(t)$ and $\tilde{\xi}(t)$. ■

For mathematical models of physical systems, recalculation equations (5.16) typically include division operation. This violates condition (5.48) and, regardless of the parameter estimation law used for $\hat{\psi}(t)$, may cause singularity of estimates $\hat{T}_I(t)$, and consequently $\hat{x}(t)$. Let us validate this claim through the following examples.

Example 2. Consider the system (2.1) with the following matrices:

$$A = \begin{bmatrix} 0 & \theta_1 \\ \theta_1 & \theta_2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \theta_3 \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

According to the transformation (5.11), the parameters of the numerator and denominator of the transfer function (5.6) are defined as follows:

$$\psi_a = \begin{bmatrix} \theta_2 \\ \theta_1^2 \end{bmatrix}, \quad \psi_b = \begin{bmatrix} 0 \\ \theta_1\theta_3 \end{bmatrix},$$

and the transformation T_I is written as:

$$T_I = \begin{bmatrix} 1 & 0 \\ \theta_2\theta_1^{-1} & \theta_1^{-1} \end{bmatrix}.$$

Then the recalculation equation (5.16) is obtained as:

$$\hat{T}_I(t) = \begin{bmatrix} 1 & 0 \\ \dfrac{\hat{\psi}_{1a}(t)}{\sqrt{\hat{\psi}_{2a}(t)}} & \dfrac{1}{\sqrt{\hat{\psi}_{2a}(t)}} \end{bmatrix}.$$

The recalculation function does not satisfy the Lipschitz condition and is singular in case $\hat{\psi}_{2a}(t) = 0$. Moreover, the negative values of $\hat{\psi}_{2a}(t)$ are out of admissible range. ■

On the basis of the dynamic properties of the parametric error, the previously discussed estimation laws of the parameters ψ_a, ψ_b can be divided into three groups:

1) do not ensure monotonicity of the error $\tilde{\psi}(t)$ neither by norm, nor by elements (5.20), (5.23);
2) guarantee monotonicity by norm (5.21), (5.29), (5.36);
3) ensure element-wise monotonicity (5.34) + (5.38), (5.41) + (5.38), (5.34) + (5.43).

Considering the estimation laws from the first two groups, it is not possible to guarantee that the estimate $\hat{\psi}_{2a}(t)$ does not enter the forbidden region $\hat{\psi}_{2a}(t) \leq 0$ in the course of the transient even if the initial conditions are chosen so that $\hat{\psi}_{2a}(t_0) > 0$. Hence, the law of physical state reconstruction (5.15) + (5.16) cannot be implemented when using observers and estimators (5.18) + (5.20), (5.18) + (5.21), (5.23), (5.26) + (5.29), (5.33) + (5.34) + (5.36) and (5.33) + (5.41) + (5.36).

The estimation laws with a scalar regressor (5.34) + (5.38), (5.41) + (5.38), (5.44) + (5.43) ensure element-wise monotonicity of the parametric error, and therefore, owing to (5.39) and (5.40), it holds for the example under consideration that

$$\left.\begin{array}{l} \hat{\psi}_2(t_0) > 0 \\ \psi_2 > 0 \end{array}\right\} \Rightarrow \hat{\psi}_2(t) > 0,$$

as it is *a priori* known from the definition of ψ_a that $\psi_2 = \psi_{2a} > 0$.

The property (5.40) allows one to apply estimators and observers (5.33) + (5.34) + (5.38), (5.33) + (5.41) + (5.38), (5.33) + (5.34) + (5.43) to implement the observer of the physical state (5.15) + (5.16). However, the properties (5.39) and (5.40) are fragile with respect to external perturbations, and more realistic practical problems may require unavailable *a priori* information to implement these estimators. Let us confirm the latter conclusion with an example.

Example 3. Consider a system (2.1) with the following matrices:

$$A = \begin{bmatrix} 0 & \theta_1 - \theta_2 \\ \theta_1 & \theta_1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \theta_3 \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{5.49}$$

Following the transformation (5.11), the parameters of the numerator and denominator of the transfer function (5.6) are defined as follows:

$$\psi_a = \begin{bmatrix} \theta_1 \\ \theta_1 \left(\theta_1 - \theta_2\right) \end{bmatrix}, \quad \psi_b = \begin{bmatrix} 0 \\ \theta_3 \left(\theta_1 - \theta_2\right) \end{bmatrix}, \tag{5.50}$$

and the transformation T_I is written as:

$$T_I = \begin{bmatrix} 1 & 0 \\ \dfrac{\theta_1}{\theta_1 - \theta_2} & \dfrac{1}{\theta_1 - \theta_2} \end{bmatrix}.$$

Then the recalculation equation (5.16) is obtained as:

$$\hat{T}_I\left(t\right) = \begin{bmatrix} 1 & 0 \\ \dfrac{\hat{\psi}_{1a}^2\left(t\right)}{\hat{\psi}_{2a}\left(t\right)} & \dfrac{\hat{\psi}_{1a}\left(t\right)}{\hat{\psi}_{2a}\left(t\right)} \end{bmatrix},$$

from which we have a singularity of the function (5.16) for $\hat{\psi}_{2a}\left(t\right) = 0$.

The estimation laws from the first two groups still cannot be used, and the implementation of the laws from the third group requires knowledge of the sign of the product $\theta_1\left(\theta_1 - \theta_2\right)$, since the following holds:

$$\begin{gathered} \operatorname{sgn}\left(\hat{\psi}_2\left(t_0\right)\right) = \operatorname{sgn}\left(\psi_2\right) \\ \Downarrow \\ \operatorname{sgn}\left(\hat{\psi}_2\left(t\right)\right) = \operatorname{sgn}\left(\psi_2\right), \end{gathered}$$

where $\psi_2 = \psi_{2a} = \theta_1\left(\theta_1 - \theta_2\right)$.

Unlike the previous example, in the considered case (5.49), (5.50), the information about the sign of ψ_2 cannot be obtained from (5.50), and hence without additional *a priori* information about the system parameters the estimation laws of the third group also cannot be used. So, for the system (5.49), the observer (5.15) + (5.16) is not implementable for any existing laws to estimate the parameters of the transfer function (5.6). ■

Examples 1–3 show that, in general case, if the system (2.1) has parametric uncertainty, then, using the observers on the basis of transformation of such system into the observer canonical form, the only solvable problem is the estimation of its virtual states $\xi\left(t\right)$, which are useless to implement the dynamic feedback of the form (2.4a) and (2.4b).

However, for few particular cases, in which the transformation equation (5.16) exists and satisfies (possibly locally) the Lipschitz condition, these solutions, unlike other adaptive and invariant observers, do not require the restrictive output matching conditions to be met. Indeed, if $\psi \in \mathcal{D}_\psi$, and the function $f_{T_I}\colon \mathbb{R}^{2n} \mapsto \mathbb{R}^n \times \mathbb{R}^n$ exists and meets the local Lipschitz condition inside $\mathcal{D}_\psi$, and, owing to (5.40), it is ensured that $\hat{\psi}\left(t\right) \in \mathcal{D}_\psi$, then, the observer (5.15) + (5.16) guarantees state reconstruction of the system, which does not meet the output matching conditions. For example, the system from Example 2 does not satisfy such condition:

$$\dot{x} = \begin{bmatrix} 0 & \theta_1 \\ \theta_1 & \theta_2 \end{bmatrix} x + \begin{bmatrix} 0 \\ \theta_3 \end{bmatrix} u = A_n x + B_n u + \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_2 & 0 & 0 & 0 \\ 0 & x_1 & x_2 & u \end{bmatrix} \begin{bmatrix} \theta_1 - 1 \\ \theta_1 \\ \theta_2 \\ \theta_3 - 1 \end{bmatrix}}_{D\Phi(x,\, u)\Theta},$$

but, nevertheless, owing to the property (5.40), for some choice of the estimates initial conditions, such system state can be reconstructed with the help of the observers (5.33) + (5.34) + (5.38), (5.33) + (5.41) + (5.38), (5.33) + (5.34) + (5.43). As it is shown in Example 3, unavailable information about the system parameters may be needed to meet the condition $\hat{\psi}\,(t) \in \mathcal{D}_{\psi}$.

Thus, the challenges of implementation of state observers based on system transformation to observer canonical form are less about to obtain estimation algorithms for $\hat{\psi}_a\,(t)\,, \hat{\psi}_b\,(t)$ and $\hat{\xi}\,(t)$ that ensure (5.17), but more about implementation of the recalculation equation (5.16), which may:

— not exist even for completely observable systems;

— exist but does not meet the Lipschitz condition (it can include operations of division, raising to a power, square root, their combinations, etc.).

Several different approaches have been proposed in the literature to overcome the problem of recalculation (5.16) singularity. Before proceeding to their consideration, let us study in detail the conditions for the existence of the recalculation equation (5.16). Note that in Examples 1–3 we considered systems with overparameterization, *i.e.*, with a number of unknown parameters that is larger than necessary for a minimal system representation. Such a class of systems in the state space is defined as follows:

$$\begin{aligned} \dot{x}\,(t) &= A\,(\theta)\,x\,(t) + B\,(\theta)\,u\,(t)\,, \\ y\,(t) &= C^{\mathrm{T}} x\,(t)\,, \quad x\,(t_0) = x_0, \end{aligned} \tag{5.51}$$

where $A : \mathbb{R}^{n_\theta} \mapsto \mathbb{R}^{n \times n}$, $B : \mathbb{R}^{n_\theta} \mapsto \mathbb{R}^{n}$ are mappings from the physical parameters $\theta \in \mathbb{R}^{n_\theta}$ to system matrices, while other notation coincides with the one from (2.1).

Then, using re-definition $\psi_a := \psi_a\,(\theta)\,, \psi_b := \psi_b\,(\theta)\,, T := T\,(\theta)$, we take into consideration the dependence of the recalculation matrix and parameters of the observer canonical form (5.12), (5.13) from the unknown parameters θ. Now the existence conditions of the recalculation equation (5.16) can be represented in terms of sufficient conditions for the existence of the inverse function that allows one to recalculate the parameters of the observer canonical form $\psi_a\,(\theta)\,,\ \psi_b\,(\theta)$ into physical parameters:

$$\begin{aligned} &\det{}^2 \left\{ \nabla_\theta \psi_{ab}\,(\theta) \right\} > 0, \\ &\psi_{ab}\,(\theta) = \mathcal{L}_{ab} \begin{bmatrix} \psi_a\,(\theta) \\ \psi_b\,(\theta) \end{bmatrix} = \mathcal{L}_{ab} \psi\,(\theta)\,, \end{aligned} \tag{5.52}$$

where $\psi\,(\theta) = \begin{bmatrix} \psi_a^\top\,(\theta) & \psi_b^\top\,(\theta) \end{bmatrix}^\top$.

If the condition (5.52) is met, then there exists an inverse function such that $\mathcal{F}\,(\psi_{ab}) = \theta$, and therefore, the mapping $f_{T_I}\colon \mathbb{R}^{2n} \mapsto \mathbb{R}^{n} \times \mathbb{R}^{n}$ from (5.16) is defined as:

$$f_{T_I}\left(\hat{\psi}\right) = \left(T_I \circ \mathcal{F}\right)\left(\mathcal{L}_{ab}\hat{\psi}\right). \tag{5.53}$$

The motivation behind the condition (5.52) is as follows. As previously shown, using the input $u\,(t)$ and the output $y\,(t)$ signals, if certain convergence conditions are met, the parameters of the observer canonical form can be estimated. The inverse coordinate transformation matrix is calculated through the physical parameters of the system. So, it is possible to calculate it only when the parameters of the observer canonical form can be recalculated into θ parameters. Considering the system with overparameterization (5.51), if the existence conditions of the inverse function are satisfied, we can describe the existing ways to avoid singularity of the recalculation (5.16).

1) Gradient projection method. Let us introduce into consideration the following set

$$\mathcal{D}_{\psi} := \left\{ \psi \in \mathbb{R}^{2n} \colon \overline{\psi_i} \leq \psi_i \leq \underline{\psi_i}, \quad i = 1, \dots 2n \right\},$$

where $\underline{\psi_i}$ and $\overline{\psi_i}$ are known scalars.

Suppose condition (5.48) holds for all $\psi \in \mathcal{D}_{\psi} \subset \mathbb{R}^{2n}$. To resolve the singularity issue, it then remains to ensure that the implication $\hat{\psi}(t_0) \in \mathcal{D}_{\psi} \Rightarrow \hat{\psi}(t) \in \mathcal{D}_{\psi}, \forall t \geq t_0$ holds. This is achieved by augmenting the estimation law with a projection operator [52, p. 788], [53, p. 133], which projects estimates onto $\mathcal{D}_{\psi}$. For example, the estimation law (5.21) with projection operator becomes:

$$\begin{gathered}
\dot{\hat{\psi}}(t) = \lambda(t) + \mu(t), \\
\lambda(t) = -\Gamma \varphi(t) \left(\varphi^{\mathrm{T}}(t) \hat{\psi}(t) - z(t) \right), \quad \hat{\psi}(t_0) \in \mathcal{D}_{\psi} \\
\mu(t) = \begin{cases} 0, & \text{if } \hat{\psi}_i(t) \in \left(\underline{\psi_i}, \overline{\psi_i} \right), \text{ or} \\ & \text{if } \hat{\psi}_i(t) = \overline{\psi_i} \text{ and } \lambda(t) \leq 0, \text{ or} \\ & \text{if } \hat{\psi}_i(t) = \underline{\psi_i} \text{ and } \lambda(t) \geq 0, \\ -\lambda(t), & \text{otherwize.} \end{cases}
\end{gathered} \tag{5.54}$$

The law (5.54) have all properties described in Theorem 7 and additionally guarantees $\hat{\psi}(t) \in \mathcal{D}_{\psi}$, $\forall t \geq t_0$, which, if Lipschitz continuity inside $\mathcal{D}_{\psi}$ holds, allows one to obtain singularity-free estimate $\hat{T}_I(t)$. Similarly, all previously discussed estimation laws can be augmented with projection operator. The limitations of the modified law (5.54) is the need to know the lower/upper bounds $\underline{\psi_i}$, $\overline{\psi_i}$ and represent the singularity-avoidance conditions via constraints for ψ_i. Generally speaking, the obtained estimates can be projected not only onto the considered set, but any convex set [52, p. 788] defined as follows:

$$\mathcal{D}_{\psi} := \left\{ \psi \in \mathbb{R}^{2n} \colon g(\psi) \leq 0 \right\}.$$

In such case, the estimation law (5.54) is written as:

$$\begin{gathered}
\dot{\hat{\psi}}(t) = \begin{cases} \lambda(t), \text{ if } \; g\left(\hat{\psi}\right) < 0 \text{ or } g\left(\hat{\psi}\right) = 0 \text{ and } \lambda^{\mathrm{T}}(t) \nabla g \leq 0, \\ \lambda(t) - \Gamma \dfrac{\nabla g \nabla g^{\mathrm{T}}}{\nabla g^{\mathrm{T}} \Gamma \nabla g} \lambda(t), \text{ otherwise,} \end{cases} \\
\lambda(t) = -\Gamma \varphi(t) \left(\varphi^{\mathrm{T}}(t) \hat{\psi}(t) - z(t) \right), \; \hat{\psi}(t_0) \in \mathcal{D}_{\psi}
\end{gathered}$$

and ensures $\hat{\psi}(t) \in \mathcal{D}_{\psi}$, $\forall t \geq t_0$. Unfortunately, considering practical scenarios, the set of $\psi(\theta)$, for which (5.16) is singular, is often non-convex, which motivated the development of other approaches.

2) Asymptotic cascade recalculation. An approach proposed in [70] is considered to avoid singularity of the recalculation (5.16), (5.53). Assume that the singularity is caused solely by operations of division by functions from $\psi_{ab}(\theta)$ and θ. In this case, the mappings $\mathcal{F}(\theta)$ and $T_I(\theta)$

can be decomposed into a matrix numerator and denominator as follows:

$$\mathcal{S}(\psi_{ab}) = \mathcal{G}(\psi_{ab})\,\mathcal{F}(\psi_{ab}) = \mathcal{G}(\psi_{ab})\,\theta, \tag{5.55a}$$

$$\mathcal{P}(\theta) = \mathcal{Q}(\theta)\, vec\,(T_I(\theta)), \tag{5.55b}$$

where $\mathcal{S} : \mathbb{R}^{n_\theta} \mapsto \mathbb{R}^{n_\theta}$, $\mathcal{G} : \mathbb{R}^{n_\theta} \mapsto \mathbb{R}^{n_\theta \times n_\theta}$, $\mathcal{P} : \mathbb{R}^{n_\theta} \mapsto \mathbb{R}^{n^2}$, $\mathcal{Q} : \mathbb{R}^{n_\theta} \mapsto \mathbb{R}^{n^2 \times n^2}$ stand for known mappings.

Then, having substituted $\hat{\psi}_{ab}(t) = \mathcal{L}_{ab}\hat{\psi}(t)$ into (5.55a), a regression equation is written (the estimates $\hat{\psi}(t)$ can be obtained with the help of any previously considered estimation algorithm):

$$\mathcal{Y}_\theta(t) = \mathcal{M}_\theta(t)\,\theta, \tag{5.56}$$

where

$$\mathcal{Y}_\theta(t) := \mathcal{S}\left(\hat{\psi}_{ab}\right),\ \mathcal{M}_\theta(t) := \mathcal{G}\left(\hat{\psi}_{ab}\right).$$

Based on (5.56), we can design, for example, the following estimation law ($\gamma > 0$):

$$\dot{\hat{\theta}}(t) = -\gamma\mathcal{M}_\theta^{\mathrm{T}}(t)\left(\mathcal{M}_\theta(t)\,\hat{\theta}(t) - \mathcal{Y}_\theta(t)\right), \quad \hat{\theta}(t_0) = \hat{\theta}_0. \tag{5.57}$$

Now, having the estimate $\hat{\theta}(t)$ at hand and following (5.56), (5.57), the required estimate of the recalculation matrix is obtained:

$$\dot{\hat{T}}_I(t) = -\mathrm{mat}\left[\gamma\mathcal{M}_{T_I}^{\mathrm{T}}(t)\left(\mathcal{M}_{T_I}(t)\,\mathrm{vec}\left(\hat{T}_I(t)\right) - \mathcal{Y}_{T_I}(t)\right)\right], \quad \hat{T}_I(t_0) = \hat{T}_{I0}, \tag{5.58}$$

where

$$\mathcal{M}_{T_I}(t) := \mathcal{Q}\left(\hat{\theta}\right), \quad \mathcal{Y}_{T_I}(t) := \mathcal{P}\left(\hat{\theta}\right),$$

and mat $[.]$ denotes an operation, which is inverse to vectorization.

As, owing to (5.55a), (5.55b), $\mathcal{S}(\psi_{ab})$, $\mathcal{G}(\psi_{ab})$, $\mathcal{P}(\theta)$, $\mathcal{Q}(\theta)$ have no singularities for all $\psi_{ab} \in \mathbb{R}^{n_\theta}$ and $\theta \in \mathbb{R}^{n_\theta}$, then the asymtotic cascade recalculation (5.57), (5.58) ensures $\hat{T}_I(t) \to T_I(\theta)$ in case $\hat{\psi}(t) \to \psi(\theta)$ and any of the previously considered estimation laws for $\psi(\theta)$ is used.

The disadvantages of the asymptotic cascade approach include, first, the need of additional identification of parameters θ, and, second, the slow rate of convergence of $\hat{T}_I(t)$ to $T_I(\theta)$ due to the necessity of convergence, firstly, of $\hat{\psi}(t)$ to $\psi(\theta)$, then, of $\hat{\theta}(t)$ to θ, and only after that, of $\hat{T}_I(t)$ to $T_I(\theta)$, thirdly, the cascade approach provides no protection against singularities, particularly if mappings $\mathcal{S}(\psi_{ab})$, $\mathcal{G}(\psi_{ab})$, $\mathcal{P}(\theta)$, $\mathcal{Q}(\theta)$ include, for example, arithmetical root operation. Unlike gradient projection method, the cascade procedure (5.57), (5.58) requires no prior knowledge of upper/lower bounds for elements of the vector $\psi(\theta)$ or convexity of $\mathcal{D}_\psi$.

3) Heterogeneous mappings. Another procedure to obtain the estimate $\hat{T}_I(t)$ of the recalculation matrix without singularities is based on the definiton and properties of heterogeneous mappings.

Definition 3. A mapping $\mathcal{F}\colon \mathbb{R}^{n_x} \mapsto \mathbb{R}^{n_\mathcal{F} \times m_\mathcal{F}}$ is heterogeneous of degree $\ell_\mathcal{F} \geq 1$ if there exist $\Pi_\mathcal{F}(\omega) \in \mathbb{R}^{n_\mathcal{F} \times n_\mathcal{F}}$, $\Xi_\mathcal{F}(\omega) = \overline{\Xi}_\mathcal{F}(\omega)\,\omega \in \mathbb{R}^{\Delta_\mathcal{F} \times n_x}$, and a mapping $\mathcal{T}_\mathcal{F} : \mathbb{R}^{\Delta_\mathcal{F}} \mapsto \mathbb{R}^{n_\mathcal{F} \times m_\mathcal{F}}$ such that for all $\omega \in \mathbb{R}$ and $x \in \mathbb{R}^{n_x}$ the following functional equation has a solution:

$$\Pi_\mathcal{F}(\omega)\,\mathcal{F}(x) = \mathcal{T}_\mathcal{F}(\Xi_\mathcal{F}(\omega)\,x), \tag{5.59a}$$

in such a way that:

$$\begin{gathered} \det\{\Pi_\mathcal{F}(\omega)\} \geq \omega^{\ell_\mathcal{F}}(t), \\ \Xi_{\mathcal{F}ij}(\omega) = c_{ij}\omega^{\ell_{ij}}, \quad \overline{\Xi}_{\mathcal{F}ij}(\omega) = c_{ij}\omega^{\ell_{ij}-1}, \\ c_{ij} \in \{0, 1\}, \quad \ell_{ij} \geq 1. \end{gathered} \tag{5.59b}$$

Example 4. A mapping $\mathcal{F}(\theta) = \mathrm{col}\{\theta_1\theta_2 + \theta_1, \theta_1\}$ with $\Pi_{\mathcal{F}}(\omega) = \mathrm{diag}\{\omega^2, \omega\}$, $\Xi_{\mathcal{F}}(\omega) = \mathrm{diag}\{\omega, \omega\}$ is heterogeneous of degree $\ell_{\mathcal{F}} = 3$. Thus, using a measurable signal $\mathcal{Y}_\theta(t) = \omega(t)\theta$, the main property $\Xi_{\mathcal{F}}(\omega)\theta = \overline{\Xi}_{\mathcal{F}}(\omega)\omega(t)\theta$ from Definition 3 allows one to parametrize a linear regression equation with respect to $\mathcal{F}(\theta)$ in the following way:

$$\Pi_{\mathcal{F}}(\omega)\mathcal{F}(\theta) = \mathcal{T}_{\mathcal{F}}\left(\overline{\Xi}_{\mathcal{F}}(\omega)\mathcal{Y}_\theta\right),$$
$$\begin{bmatrix}\omega^2(t) & 0\\ 0 & \omega(t)\end{bmatrix}\mathcal{F}(\theta) = \begin{bmatrix}\mathcal{Y}_{1\theta}(t)\mathcal{Y}_{2\theta}(t) + \mathcal{Y}_{1\theta}(t)\\ \mathcal{Y}_{1\theta}(t)\end{bmatrix}. \qquad \blacksquare$$

A particular case of heterogeneity conditions are homogeneity conditions, since for homogeneous mappings, in terms of equation (5.59a), it holds that

$$\Pi_{\mathcal{F}}(\omega)\mathcal{F}(x) = \mathcal{F}(\Xi_{\mathcal{F}}(\omega)x),$$
$$\Xi_{\mathcal{F}}(\omega) = \omega^\ell I_{n_x}, \quad \ell_{ij} = \ell,$$

that is, in this case, the product $\Pi_{\mathcal{F}}(\omega)\mathcal{F}(x)$ can be calculated via the function $\mathcal{F}(x)$ itself by the multiplication of its argument by $\Xi_{\mathcal{F}}(\omega)$.

In [71], it is shown that the functional equation (5.59a) has a solution for a sufficiently general algebraic function that covers all polynomials and monomials, as well as some of the irrational functions.

Proposition 1. *Let $n_{\mathcal{F}} = m_{\mathcal{F}} = 1$, then the mapping $\mathcal{F}(x) = \sum\limits_j a_j \prod\limits_{i=1}^{n_x} x_i^{k_{ji}}$, $k_{ji} \geq 0$ is heterogeneous in the sense of* (5.59a), (5.59b).

Proof of proposition is given in [71].

Let us show how the properties of heterogeneous mappings can be used to obtain an estimate of $\hat{T}_I(t)$. It should be recalled that using the dynamic regressor extension and mixing procedure (5.38), the signals $\mathcal{Y}(t)$ and $\Delta(t)$ from the following regression equation are available for measurement:

$$\mathcal{Y}(t) = \Delta(t)\psi(\theta).$$

Then, the multiplication of $\mathcal{Y}(t)$ by $\mathcal{L}_{ab}$ yields:

$$\mathcal{Y}_{ab}(t) = \Delta(t)\psi_{ab}(\theta),$$

where $\mathcal{Y}_{ab}(t) := \mathcal{L}_{ab}\mathcal{Y}(t)$.

Assume that the mappings $\mathcal{S}(\psi_{ab})$, $\mathcal{G}(\psi_{ab})$ from (5.55a) are heterogeneous in the sense of Definition 3. Then, having multiplied the left- and right-hand sides of (5.55a) by $\Pi(\Delta)$ and used the property (5.59b), it is obtained that:

$$\begin{gathered}\mathcal{T}_{\mathcal{S}}(\Xi_{\mathcal{S}}(\Delta)\psi_{ab}) = \mathcal{T}_{\mathcal{G}}(\Xi_{\mathcal{G}}(\Delta)\psi_{ab})\theta,\\ \Downarrow\\ \mathcal{T}_{\mathcal{S}}\left(\overline{\Xi}_{\mathcal{S}}(\Delta)\mathcal{Y}_{ab}\right) = \mathcal{T}_{\mathcal{G}}\left(\overline{\Xi}_{\mathcal{G}}(\Delta)\mathcal{Y}_{ab}\right)\theta,\end{gathered} \tag{5.60}$$

where the signals $\mathcal{T}_{\mathcal{S}}\left(\overline{\Xi}_{\mathcal{S}}(\Delta)\mathcal{Y}_{ab}\right)$, $\mathcal{T}_{\mathcal{G}}\left(\overline{\Xi}_{\mathcal{G}}(\Delta)\mathcal{Y}_{ab}\right)$ are measurable as $\mathcal{Y}_{ab}(t)$ and $\Delta(t)$ are measurable, and the mappings $\mathcal{T}_{\mathcal{S}}(.)$, $\mathcal{T}_{\mathcal{G}}(.)$, $\overline{\Xi}_{\mathcal{S}}(.)$, $\overline{\Xi}_{\mathcal{G}}(.)$ are known.

Having multiplied (5.60) by $\mathrm{adj}\{\mathcal{T}_{\mathcal{G}}\left(\overline{\Xi}_{\mathcal{G}}(\Delta)\mathcal{Y}_{ab}\right)\}$, a regression equation with a scalar regressor is obtained:

$$\mathcal{Y}_\theta(t) = \mathcal{M}_\theta(t)\theta, \tag{5.61}$$

where

$$\mathcal{Y}_\theta(t) := \text{adj}\left\{\mathcal{T}_\mathcal{G}\left(\overline{\Xi}_\mathcal{G}(\Delta)\mathcal{Y}_{ab}\right)\right\}\mathcal{T}_\mathcal{S}\left(\overline{\Xi}_\mathcal{S}(\Delta)\mathcal{Y}_{ab}\right), \quad \mathcal{M}_\theta(t) := \det\left\{\mathcal{T}_\mathcal{G}\left(\overline{\Xi}_\mathcal{G}(\Delta)\mathcal{Y}_{ab}\right)\right\}.$$

Then, having assumed the heterogeneity of $\mathcal{P}(\theta)$, $\mathcal{Q}(\theta)$ and used the property (5.59b), the same derivations as in (5.60), (5.61) yields

$$\mathcal{Y}_{T_I}(t) = \mathcal{M}_{T_I}(t)\,\text{vec}\left(T_I(\theta)\right), \tag{5.62}$$

where

$$\mathcal{Y}_{T_I}(t) := \text{adj}\left\{\mathcal{T}_\mathcal{Q}\left(\overline{\Xi}_\mathcal{Q}(\mathcal{M}_\theta)\mathcal{Y}_\theta\right)\right\}\mathcal{T}_\mathcal{P}\left(\overline{\Xi}_\mathcal{P}(\mathcal{M}_\theta)\mathcal{Y}_\theta\right), \quad \mathcal{M}_{T_I}(t) := \det\left\{\mathcal{T}_\mathcal{Q}\left(\overline{\Xi}_\mathcal{Q}(\mathcal{M}_\theta)\mathcal{Y}_\theta\right)\right\}.$$

Based on equation (5.62), the following estimation algorithm is introduced:

$$\dot{\hat{T}}_I(t) = -\text{mat}\left[\gamma\mathcal{M}_{T_I}(t)\left(\mathcal{M}_{T_I}(t)\,\text{vec}\left(\hat{T}_I(t)\right) - \mathcal{Y}_{T_I}(t)\right)\right], \quad \hat{T}_I(t_0) = \hat{T}_{I0}. \tag{5.63}$$

Unlike the gradient projection method, the approach based on heterogeneous mappings does not require the convexity of the set $\mathcal{D}_\psi$ or knowledge of any *a priori* information about the unknown parameters $\psi(\theta)$. Unlike the cascade recalculation (5.57), (5.56), only the inverse coordinate transformation matrix is estimated without identification of θ, which ensures faster convergence.

5.3. Conclusions on Adaptive Observers Overview

Compared to robust observers, adaptive observers *i*) do not require *a priori* information about system parameters specified by parameterization (3.5) or (3.6), *ii*) to achieve the goal (2.5), instead of the restrictive condition $u \in L_2$ or the equality $u(t) = -K_r\hat{x}(t)$, they only require the boundedness of control signal and states (for a discussion on the relaxation of this condition, see Section 6.1). In comparison with the invariant observers, the adaptive ones do not require to meet the restrictive extended and standard output matching conditions, and therefore, are applicable to a wider class of systems.

The open problems of adaptive observers are *i*) that the goal (2.5) is achieved only when the finite excitation condition $\varphi \in \text{FE}$ is satisfied, *i.e.*, the necessary and sufficient condition for the identifiability of unknown parameters [72], *ii*) the applicability of solutions only to systems with time-invariant parameters (for comparison, all observers considered earlier, both robust and invariant, generally do not assume the system unknown parameters to be time-invariant). Note that adaptive observers, in which parameter changes are described by a linear dynamic exosystems with known parameters and unknown initial conditions [73], do not allow the second problem to be overcome: in practice, system parameter changes do not usually fit such exosystem models.

6. SOME REMARKS ON STATE-ESTIMATE-BASED CONTROL

Having compared robust, invariant, and adaptive observers, it may seem that only robust observers do not require *a priori* boundedness of $u(t)$ and $x(t)$ (see Assumption 1) and allow one to use of the obtained estimates $\hat{x}(t)$ for feedback control purposes. Indeed, in most studies on adaptive and invariant observers, the problem of state estimation is considered separately from the control task. As an exception, we note the paper [70], which is devoted to the design of adaptive control based on an adaptive observer. Naturally, an interested reader may have the following questions. To satisfy Assumption 1, has the stabilizing controller already been designed? How is it designed if the system parameters are unknown and the state vector is unmeasurable?

On the one hand, a stabilizing controller can be designed based on the system output using, for example, robust control techniques. In this situation, as noted in Remark 1, the state observer is not aimed at control task solution, but rather at fault detection, process monitoring and logging of unmeasured variables, digital twin design, etc. On the other hand, following the principle of adaptive modular design, the estimation and control tasks in linear systems can be considered independently of each other. Let us focus on the latter point in more detail using the example of an adaptive observer consisting of equations (5.15), (5.18) at $v(t)=0$, *i.e.*

$$\begin{gathered}\hat{x}(t)=\hat{T}_I(t)\,\hat{\xi}(t),\\ \dot{\hat{\xi}}(t)=A_0\hat{\xi}(t)+\hat{\psi}_a(t)\,y(t)+\hat{\psi}_b(t)\,u(t)+L\left(C_0^{\mathrm{T}}\hat{\xi}(t)-y(t)\right).\end{gathered}\tag{6.1}$$

For simplicity and without loss of generality, we assume that the control signal is a robust law (2.4a). We define the parameter estimation laws $\psi_{ab}(\theta)$ and $T_I(\theta)$ by (5.38) + (5.41) and (5.63), *i.e.*:

$$\begin{aligned}\dot{\tilde{\psi}}(t)=\dot{\hat{\psi}}(t)&=-\gamma\left(\Delta(t)\,\hat{\psi}(t)-\mathcal{Y}(t)\right)=-\gamma\Delta(t)\,\tilde{\psi}(t),\\ \dot{\tilde{T}}_I(t)=\dot{\hat{T}}_I(t)&=-\mathrm{mat}\left[\gamma_{T_I}\mathcal{M}_{T_I}(t)\left(\mathcal{M}_{T_I}(t)\,vec\left(\hat{T}_I(t)\right)-\mathcal{Y}_{T_I}(t)\right)\right]\\ &=-\mathrm{mat}\left[\gamma_{T_I}\mathcal{M}^2_{T_I}(t)\,vec\left(\tilde{T}_I(t)\right)\right].\end{aligned}\tag{6.2}$$

When the condition $\varphi\in\mathrm{FE}$ is met, regardless of the boundedness of $x(t)$ and $u(t)$, it is easy to show (5.42) the existence of following lower bounds for all $t\geq t_e$:

$$|\Delta(t)|\geq\Delta_{LB}>0,\quad |\mathcal{M}_{T_I}(t)|\geq\underline{\mathcal{M}}_{T_I}>0,$$

therefore, the errors $\tilde{\psi}(t)$ and $\tilde{T}_I(t)$ converge exponentially to zero with the rate defined by parameters $\gamma>0$ and $\gamma_{T_I}>0$, *i.e.*:

$$\left\|\tilde{\psi}(t)\right\|=e^{-\gamma\Delta_{LB}(t-t_e)}\left\|\tilde{\psi}(t_e)\right\|,\quad \left\|\tilde{T}_I(t)\right\|=e^{-\gamma_{T_I}\underline{\mathcal{M}}^2_{T_I}(t-t_e)}\left\|\tilde{T}_I(t_e)\right\|.\tag{6.3}$$

The differential equations for $x(t)$ and $\tilde{\xi}(t)$ are written as follows:

$$\begin{gathered}\begin{bmatrix}\dot{x}(t)\\ \dot{\tilde{\xi}}(t)\end{bmatrix}=(\mathcal{A}+\mathcal{B}(t))\begin{bmatrix}x(t)\\ \tilde{\xi}(t)\end{bmatrix}\\ =\begin{bmatrix}A-BK_r\left(I+\tilde{T}_I(t)\,T(\theta)\right) & -BK_r\left(\tilde{T}_I(t)+T_I(\theta)\right)\\ \tilde{\psi}_a(t)C^{\mathrm{T}}-\tilde{\psi}_b(t)K_r\left(I+\tilde{T}_I(t)T(\theta)\right) & A_m-\tilde{\psi}_b(t)K_r\left(\tilde{T}_I(t)+T_I(\theta)\right)\end{bmatrix}\begin{bmatrix}x(t)\\ \tilde{\xi}(t)\end{bmatrix},\end{gathered}\tag{6.4}$$

where

$$\mathcal{A}=\begin{bmatrix}A-BK_r & -BK_rT_I(\theta)\\ 0 & A_m\end{bmatrix},$$

$$\mathcal{B}(t)=\begin{bmatrix}-BK_r\left(\tilde{T}_I(t)\,T(\theta)\right) & -BK_r\left(\tilde{T}_I(t)\right)\\ \tilde{\psi}_a(t)\,C^{\mathrm{T}}-\tilde{\psi}_b(t)\,K_r\left(I+\tilde{T}_I(t)\,T(\theta)\right) & -\tilde{\psi}_b(t)\,K_r\left(\tilde{T}_I(t)+T_I(\theta)\right)\end{bmatrix}.$$

Owing to the estimates (6.3), we have $||\mathcal{B}(t)||\to 0$ as $t\to\infty$, then, if robust control techniques were used to calculate K_r such that the matrix $A-BK_r$ is Hurwitz, then, by Lemma 9.5 from [74], we obtain global exponential convergence of states $x(t)$ and estimation errors $\tilde{\xi}(t)$, $\tilde{x}(t)$ to zero. Using adaptive feedback (2.4b), the situation is similar, except that the control scheme includes the parameter estimation law for K (e.g., for the controller (2.3)), and the matrix $\mathcal{B}(t)$ now additionally depends on the error $\tilde{K}(t)=\hat{K}(t)-K$. Thus, the adaptive observers considered in the overview, which achieve the goal (2.5), can be fully justified for application to design the control law of the type (2.4a) or (2.4b) when $\varphi\in\mathrm{FE}$.

7. CONCLUSION

Based on the stated above, the following conclusions can be made regarding the feasibility conditions of existing state observers for linear dynamical systems based on robust, invariant, and adaptive control methods:

— robust state observers require the system uncertainty to be parameterized in the form of (3.5) or (3.6). This, in turn, demands the bounds of the system parameter variation range to be known. Such observers achieve the goal if $u \in L_2$ or $u(t) = -K_r \hat{x}(t)$ and the parameter K_r is calculated jointly with the observer parameters. In the general case $u \in L_\infty$, robust observers ensure that the estimation error belongs to some invariant ellipsoid that can be minimized.

— invariant state observers solve the problem of state estimation in case of arbitrary parametric uncertainty, but they require strict structural constraints to be met—standard or extended output matching conditions. As the analysis shows, some of these conditions for systems with scalar output cannot even potentially be relaxed (in particular, the requirement that the dimension of the measured output exceed the dimension of the disturbance). Therefore, the practical application of invariant state observers is restricted to the class of systems, for which the specified strict structural constraints are satisfied.

— adaptive state observers can be divided into two groups. Methods from the first group (see Section 5.1) allow one to reconstruct the states of systems with a SPR transfer function from a parametric disturbance to a measured output, which is equivalent to the restrictive output matching conditions required for the application of invariant observers. Algorithms from the second group (see Section 5.2) reconstruct the states of linear systems with completely unknown parameters when the following requirements are met: *i*) the necessary and sufficient condition for identifiability [72] of the parameters of the observer canonical form $\psi(\theta)$, *i.e.*, $\varphi \in \mathrm{FE}$, *ii*) the sufficient condition (5.52) for the existence of the inverse function $\mathcal{F}(\psi_{ab})$, which allows the parameters $\psi(\theta)$ to be converted into physical parameters θ, *iii*) the singularity problem is avoidable using one of the techniques discussed in Section 5.2.7.

In general, research into the development of new methods to design observers for linear dynamical systems continues within all three branches of control theory discussed here, and their direction is largely driven by the existing problems mentioned in this overview, particularly in Sections 3.3, 4.4, and 5.3.

REFERENCES

1. Pacas, M., Sensorless drives in industrial applications, *IEEE Ind. Electron. Mag.*, 2011, vol. 5, no. 2, pp. 16–23.
2. Luenberger, D.G., Observing the state of a linear system, *IEEE Trans. Mil. Electron.*, 1964, vol. 8, no. 2, pp. 74–80.
3. Kalman, R.E. and Bucy, R.S., New results in linear filtering and prediction theory, *J. Basic Eng.*, 1961, vol. 83, pp. 95–108.
4. Bernard, P., Andrieu, V., and Astolfi, D., Observer design for continuous-time dynamical systems, *Annu. Rev. Control*, 2022, vol. 53, pp. 224–248.
5. Andrievsky, B. and Furtat, I., Disturbance observers: Methods and applications. I, *Autom. Remote Control*, 2020, vol. 81, no. 9, pp. 1563–1610.
6. O'Reilly, J., *Observers for Linear Systems*, Academic Press, 1983.
7. Shtessel, Y., Edwards, C., Fridman, L., and Levant, A., *Sliding Mode Control and Observation*, Springer, 2011.
8. Korovin, S.K. and Fomichev, V.V., *State Observers for Linear Systems with Uncertainty*, Walter de Gruyter, 2009.

9. Gao, H. and Li, X., *Robust Filtering for Uncertain Systems*, Springer, 2014.

10. Lewis, F.L., Xie, L., and Popa, D., *Optimal and Robust Estimation: With an Introduction to Stochastic Control Theory*, CRC Press, 2017.

11. Krasnova, S.A. and Utkin, V.A., *Kaskadnoe proektirovanie nablyudatelei sostoyaniya dinamicheskikh sistem* (Cascade Design of Dynamical Systems State Observers), Moscow: Nauka, 2006.

12. Narendra, K.S. and Annaswamy, A.M., *Stable Adaptive Systems*, Courier Corporation, 2012.

13. Bhattacharyya, S.P. and de Souza, E., Pole assignment via Sylvester's equation, *Syst. Control Lett.*, 1982, vol. 1, no. 4, pp. 261–263.

14. Dudarenko, N.A., Slita, O.V., and Ushakov, A.V., Algebraic conditions of generalized modal control, *IFAC Proc. Vol.*, 2012, vol. 45, no. 13, pp. 150–155.

15. Petersen, I.R., A stabilization algorithm for a class of uncertain linear systems, *Syst. Control Lett.*, 1987, vol. 8, no. 4, pp. 351–357.

16. Geromel, J.C., Peres, P.L.D., and Bernussou, J., On a convex parameter space method for linear control design of uncertain systems, *SIAM J. Control Optim.*, 1991, vol. 29, no. 2, pp. 381–402.

17. Fu, M., de Souza, C.E., and Xie, L., $\mathcal{H}_\infty$ estimation for uncertain systems, *Int. J. Robust Nonlinear Control*, 1992, vol. 2, no. 2, pp. 87–105.

18. Xie, L. and de Souza, C.E., On robust filtering for linear systems with parameter uncertainty, *Proc. 34th IEEE Conf. Decis. Control*, 1995, vol. 2, pp. 2087–2092.

19. Geromel, J.C. and De Oliveira, M.C., $\mathcal{H}2$ and $\mathcal{H}\infty$ robust filtering for convex bounded uncertain systems, *IEEE Trans. Autom. Control*, 2001, vol. 46, no. 1, pp. 100–107.

20. Jin, S.H. and Park, J.B., Robust $\mathcal{H}2$, $\mathcal{H}\infty$ filtering for polytopic uncertain systems via convex optimisation, *IEEE Proc., Control Theory Appl.*, 2001, vol. 148, no. 1, pp. 55–59.

21. Duan, Z., Zhang, J., Zhang, C., and Mosca, E., Robust $\mathcal{H}2$ and $\mathcal{H}\infty$ filtering for uncertain linear systems, *Automatica*, 2006, vol. 42, no. 11, pp. 1919–1926.

22. Lacerda, M.J., Oliveira, R.C.L.F., and Peres, P.L.D., Robust $\mathcal{H}2$ and $\mathcal{H}\infty$ filter design for uncertain linear systems via LMIs and polynomial matrices, *Signal Process.*, 2011, vol. 91, no. 5, pp. 1115–1122.

23. Polyak, B.T., Khlebnikov, M.V., and Rappoport, L.B., *Matematicheskaya teoriya avtomaticheskogo upravleniya* (Mathematical Theory of Automatic Control), Moscow: Lenand, 2019.

24. Petersen, I.R., A stabilization algorithm for a class of uncertain linear systems, *Syst. Control Lett.*, 1987, vol. 8, no. 4, pp. 351–357.

25. Khlebnikov, M.V., Robust filtering under nonrandom disturbances: The invariant ellipsoid approach, *Autom. Remote Control*, 2009, vol. 70, no. 1, pp. 133–146.

26. Petersen, I.R., A Riccati equation approach to the design of stabilizing controllers and observers for a class of uncertain linear systems, *Proc. Am. Control Conf.*, 1985, pp. 772–777.

27. Boyd, S., El Ghaoui, L., Feron, E., et al., *Linear Matrix Inequalities in System and Control Theory*, SIAM, 1994.

28. Lien, C.H. and Yu, K.W., LMI optimization approach on robustness and $\mathcal{H}_\infty$ control analysis for observer-based control of uncertain systems, *Chaos Solitons Fractals*, 2008, vol. 36, no. 3, pp. 617–627.

29. Perruquetti, W. and Barbot, J.P., *Sliding Mode Control in Engineering*, Marcel Dekker, 2002.

30. Alwi, H., Edwards, C., and Tan, C.P., *Fault Detection and Fault-Tolerant Control Using Sliding Modes*, Springer, 2011.

31. Utkin, V., Poznyak, A., Orlov, Y., and Polyakov, A., *Road Map for Sliding Mode Control Design*, Springer, 2020.

32. Yang, F. and Wilde, R.W., Observers for linear systems with unknown inputs, *IEEE Trans. Autom. Control*, 1988, vol. 33, no. 7, pp. 677–681.

33. Darouach, M., Zasadzinski, M., and Xu, S.J., Full-order observers for linear systems with unknown inputs, *IEEE Trans. Autom. Control*, 1994, vol. 39, no. 3, pp. 606–609.

34. Dimassi, H., Loria, A., and Belghith, S., Continuously-implemented sliding mode adaptive unknown-input observers under noisy measurements, *Syst. Control Lett.*, 2012, vol. 61, no. 12, pp. 1194–1202.

35. Utkin, V.I., *Sliding Modes in Control and Optimization*, Springer, 2013.

36. Slotine, J.J.E., Hedrick, J.K., and Misawa, E.A., On sliding observers for nonlinear systems, *J. Dyn. Syst. Meas. Control*, 1987, vol. 109, pp. 245–252.

37. Edwards, C. and Spurgeon, S.K., On the development of discontinuous observers, *Int. J. Control*, 1994, vol. 59, no. 5, pp. 1211–1229.

38. Edwards, C., Spurgeon, S.K., and Patton, R.J., Sliding mode observers for fault detection and isolation, *Automatica*, 2000, vol. 36, no. 4, pp. 541–553.

39. Edwards, C. and Tan, C.P., A comparison of sliding mode and unknown input observers for fault reconstruction, *Eur. J. Control*, 2006, vol. 12, no. 3, pp. 245–260.

40. Darouach, M., Zasadzinski, M., and Boutat-Baddas, L., Discussion on "A comparison of sliding mode and unknown input observer for fault reconstruction", *Eur. J. Control*, 2006, vol. 12, no. 3, pp. 261–266.

41. Floquet, T., Edwards, C., and Spurgeon, S.K., On sliding mode observers for systems with unknown inputs, *J. Adapt. Control Signal Process.*, 2007, vol. 21, no. 8–9, pp. 638–656.

42. Isidori, A., *Lectures in Feedback Design for Multivariable Systems*, Springer, 2017.

43. Kalsi, K., Lian, J., Hui, S., and Zak, S.H., Sliding-mode observers for systems with unknown inputs: A high-gain approach, *Automatica*, 2010, vol. 46, no. 2, pp. 347–353.

44. Zhu, F., State estimation and unknown input reconstruction via both reduced-order and high-order sliding mode observers, *J. Process Control*, 2012, vol. 22, no. 1, pp. 296–302.

45. Floquet, T. and Barbot, J.-P., A canonical form for the design of unknown input sliding mode observers, in *Advances in Variable Structure and Sliding Mode Control*, 2006, vol. 334, pp. 271–292.

46. Cho, Y.M. and Rajamani, R., A systematic approach to adaptive observer synthesis for nonlinear systems, *IEEE Trans. Autom. Control*, 1997, vol. 42, no. 4, pp. 534–537.

47. Cecilia, A., and Costa-Castell, R., Addressing the relative degree restriction in nonlinear adaptive observers: A high-gain observer approach, *J. Franklin Inst.*, 2022, vol. 359, no. 8, pp. 3857–3882.

48. Antsaklis, P.J. and Michel, A.N., *Linear Systems*, McGraw-Hill, 1997.

49. Carroll, R. and Lindorff, D., An adaptive observer for single-input single-output linear systems, *IEEE Trans. Autom. Control*, 1973, vol. 18, no. 5, pp. 428–435.

50. Luders, G. and Narendra, K.S., An adaptive observer and identifier for a linear system, *IEEE Trans. Autom. Control*, 1973, vol. 18, no. 5, pp. 496–499.

51. Narendra, K.S. and Kudva, P., Stable adaptive schemes for system identification and control-Part I, *IEEE Trans. Syst. Man Cybern.*, 1974, vol. 6, pp. 542–551.

52. Ioannou, P.A. and Sun, J., *Robust Adaptive Control*, McGraw-Hill, 1996.

53. Tao, G., *Adaptive Control Design and Analysis*, Wiley, 2003.

54. Zhang, Q., Revisiting different adaptive observers through a unified formulation, *Proc. 44th IEEE Conf. Decis. Control*, 2005, pp. 3067–3072.

55. Kreisselmeier, G., Adaptive observers with exponential rate of convergence, *IEEE Trans. Syst. Man Cybern.*, 1977, vol. 22, no. 1, pp. 2–8.

56. Prokopov, B.I., On design of adaptive observers, *Autom. Remote Control*, 1981, vol. 42, no. 5, pp. 633–637.

57. Bobtsov, A., Pyrkin, A., Vedyakov, A., Vediakova, A., and Aranovskiy, S., A modification of generalized parameter-based adaptive observer for linear systems with relaxed excitation conditions, *IFAC-PapersOnLine*, 2022, vol. 55, no. 12, pp. 324–329.

58. Efimov, D. and Fradkov, A., Design of impulsive adaptive observers for improvement of persistency of excitation, *Int. J. Adapt. Control Signal Process.*, 2015, vol. 29, no. 6, pp. 765–782.

59. Gerasimov, D.N. and Nikiforov, V.O., On key properties of the Lion's and Kreisselmeier's adaptation algorithms, *Int. J. Adapt. Control Signal Process.*, 2022, vol. 36, no. 6, pp. 1285–1304.

60. Aranovskiy, S., Ushirobira, R., Korotina, M., and Vedyakov, A., On preserving excitation properties of Kreisselmeier's regressor extension scheme, *IEEE Trans. Autom. Control*, 2023, vol. 68, no. 2, pp. 1296–1302.

61. Aranovskiy, S., Bobtsov, A., Ortega, R., and Pyrkin, A., Performance enhancement of parameter estimators via dynamic regressor extension and mixing, *IEEE Trans. Autom. Control*, 2017, vol. 62, no. 7, pp. 3546–3550.

62. Baev, S., Shkolnikov, I., Shtessel, Y., and Poznyak, A., Parameter identification of non-linear system using traditional and high order sliding modes, *Proc. Am. Control Conf.*, 2006, pp. 1–6.

63. Boyd, S. and Sastry, S.S., Necessary and sufficient conditions for parameter convergence in adaptive control, *Automatica*, 1986, vol. 22, no. 6, pp. 629–639.

64. Ortega, R., Nikiforov, V., and Gerasimov, D., On modified parameter estimators for identification and adaptive control. A unified framework and some new schemes, *Annu. Rev. Control*, 2020, vol. 50, pp. 278–293.

65. Pan, Y., Shi, T., and Ortega, R., Comparative analysis of parameter convergence for several least-squares estimation schemes, *IEEE Trans. Autom. Control*, 2024, vol. 69, no. 5, pp. 3341–3348.

66. Pan, Y. and Shi, T., Adaptive estimation and control with online data memory: A historical perspective, *IEEE Control Syst. Lett.*, 2024, vol. 8, pp. 267–278.

67. Glushchenko, A., Petrov, V., and Lastochkin, K., Regression filtration with resetting to provide exponential convergence of MRAC for plants with jump change of unknown parameters, *IEEE Trans. Autom. Control*, 2023, vol. 68, no. 8, pp. 5127–5134.

68. Glushchenko, A. and Lastochkin, K., Unknown piecewise constant parameters identification with exponential rate of convergence, *Int. J. Adapt. Control Signal Process.*, 2022, vol. 37, no. 1, pp. 315–346.

69. Ortega, R., Aranovskiy, S., Pyrkin, A.A., Astolfi, A., and Bobtsov, A.A., New results on parameter estimation via dynamic regressor extension and mixing: Continuous and discrete-time cases, *IEEE Trans. Autom. Control*, 2021, vol. 66, no. 5, pp. 2265–2272.

70. Kreisselmeier, G., Adaptive control via adaptive observation and asymptotic feedback matrix synthesis, *IEEE Trans. Autom. Control*, 1980, vol. 25, no. 4, pp. 717–722.

71. Glushchenko, A. and Lastochkin, K., Exponentially stable adaptive observation for systems parameterized by unknown physical parameters, *IEEE Trans. Autom. Control*, 2024, vol. 69, no. 8, pp. 5635–5642.

72. Wang, L., Ortega, R., Bobtsov, A., Romero, J.G., and Yi, B., Identifiability implies robust, globally exponentially convergent on-line parameter estimation, *Int. J. Control*, 2024, vol. 97, no. 9, pp. 1967–1983.

73. Pyrkin, A., Bobtsov, A., Ortega, R., and Isidori, A., An adaptive observer for uncertain linear time-varying systems with unknown additive perturbations, *Automatica*, 2023, vol. 147, p. 110677.

74. Khalil, H.K. and Grizzle, J.W., *Nonlinear Systems*, Prentice Hall, 2002.

75. Krasnova, S.A. and Utkin, A.V., Sigma-function for design of state and disturbance observers, *Probl. Control*, 2015, no. 5, pp. 27–36.

76. Krasnova, S.A. and Mysik, N.S., Cascade synthesis of a state observer with nonlinear correcting influences, *Autom. Remote Control*, 2014, vol. 75, no. 2, pp. 263–280.

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, 2026, Vol. 87, No. 2, pp. 141–158.

LINEAR SYSTEMS

Tuning Non-Fragile PI Controllers: Analysis

M. V. Khlebnikov[*,a] **and D. V. Shatov**[*,b]

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

Received July 9, 2025
Revised August 27, 2025
Accepted September 1, 2025

Abstract—This paper is devoted to the fragility analysis of PI controllers. Two different approaches are proposed to estimate the fragility of a given PI controller; they can be used independently of each other. The first approach is based on the ideas of ellipsoidal estimation and the technique of linear matrix inequalities. Within the second approach, involving the so-called *parameter loop opening*, the parameter under study is "extracted" from the plant–controller system, forming a fictitious control loop; after opening, this loop is analyzed by classical frequency-domain methods (the Nyquist criterion and D-partition). The features of both fragility analysis methods are described, and numerical examples are provided.

Keywords: linear control system, PI controller, fragility, linear matrix inequalities (LMIs), Lyapunov inequality, Nyquist criterion, D-partition

DOI: 10.7868/S1608303226020022

1. INTRODUCTION

PI and PID controllers are the most common types of automatic controllers; according to various estimates (for example, see [1, 2]), they account for over 90% of currently used controllers. Based on medium-term forecasts, this share will even remain exceptionally high, nearly 80%. The extensive application of PI and PID controllers in industry is due to several circumstances. In addition to their suitability for solving most practical problems and low cost, a high demand for such controllers is associated with their simplicity: one needs to choose correctly only two (three) coefficients—gains—to tune a PI (PID, respectively) controller.

At the same time, the procedures for their practical tuning remain much heuristic: in real plants, PI and PID controllers are often tuned manually, based on an intuitive understanding of the industrial process and the influence of separate PID/PI control components on the latter. Accordingly, the problem of developing regular approaches to their tuning still retains its meaningfulness and topicality.

The general trend of recent decades is the transition to numerical PI/PID controller design methods (including their modifications, called low-order controllers) based on solving optimization problems with a certain performance criterion: the LQR approach [3, 4], H_∞ optimization [5, 6], and others [7, 8].

In practical problems, uncertainty is inevitably introduced into a control system due to the inaccurate technical implementation of the controller itself or the need to tune its parameters during operation. In this case, an optimal controller (in one sense or another) may turn out to be "fragile," i.e., small variations in its parameters can lead to a significant deterioration in control performance or even instability of the closed-loop system. Therefore, the problem of designing non-fragile controllers, originating from the work [9], particularly PI/PID controllers, has obvious practical significance.

In this paper, the problem is to estimate the non-fragility of *given* (or *nominal*) PI controllers. We propose two approaches to solve the problem as follows.

The first approach utilizes a special Lyapunov inequality; when satisfied, it means the existence of a common quadratic Lyapunov function for a set of parameter perturbations of a given PI controller lying in some ellipse. The approach ensures stability for any (including simultaneous and time-varying) parameter perturbations and, moreover, yields ellipsoidal estimates of the system output under bounded exogenous disturbances. The technique of linear matrix inequalities (LMIs) [10, 11] is used as a technical tool here.

The second approach is based on the so-called *parameter loop opening* [12], commonly used in robust analysis problems. The core of this method is to represent a closed-loop plant–controller system in a special form: the parameter under study is "extracted" from the system, forming a *fictitious control loop*. Then this loop is investigated using a classical stability analysis method, i.e., the Nyquist criterion [13]. The approach allows analyzing static perturbations in each PI controller parameter (i.e., finding the set of stability intervals for separate parameters).

These approaches involve different ideas and can be applied both independently and jointly, complementing each other. For example, if a stabilizing PI controller is a priori unknown for a plant, then the second approach can be used to "grope" a region of stabilizing controllers, and the first approach will be naturally applied inside it.

The remainder of this paper is organized as follows. Section 2 is devoted to the formal problem statement. Sections 3 and 4 contain the main results, and Section 5 provides numerical examples.

From now on, $\|\cdot\|$ is the Euclidean norm of a vector and the spectral norm of a matrix; $^{\mathrm{T}}$ is the transpose symbol; I denotes an identity matrix of appropriate dimensions; $\lambda_i(A)$ are the eigenvalues of a matrix A; $\sigma(A) \doteq -\max_i \operatorname{Re}(\lambda_i(A)) > 0$ means the stability degree of a Hurwitz matrix A. Note that all matrix inequalities are understood in the sense of positive/negative definiteness of corresponding matrices. Finally, s is the Laplace transform symbol.

2. PROBLEM STATEMENT

Consider a linear continuous-time control system described by

$$\begin{aligned} \dot{x} &= Ax + bu + Dw, \quad x(0) = x_0, \\ y &= c^{\mathrm{T}} x, \\ z &= Cx, \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, $D \in \mathbb{R}^{n\times m}$, $c \in \mathbb{R}^n$, $C \in \mathbb{R}^{r\times n}$, with the state vector $x(t) \in \mathbb{R}^n$, the measured output $y(t) \in \mathbb{R}$, the controlled output $z(t) \in \mathbb{R}^r$, an exogenous disturbance $w(t) \in \mathbb{R}^m$ satisfying the constraint

$$\|w(t)\| \leqslant 1 \quad \text{for all } t \geqslant 0, \tag{2}$$

and the control input $u(t) \in \mathbb{R}$ in the form of a PI controller

$$u(t) = -k_1 y(t) - k_2 \int\limits_0^t y(\tau)d\tau. \tag{3}$$

Let the gains k_1 and k_2, called the *parameters* of the PI controller (and sometimes identified with the controller itself), be combined into the vector

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

Following [14] and introducing the auxiliary scalar variable ξ by

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

we arrive at the equations of the closed-loop system

$$\dot{g} = A_{\mathrm{cl}}(k)g + D_{\mathrm{cl}}w = \begin{pmatrix} A - k_1 bc^{\mathrm{T}} & -k_2 b \\ c^{\mathrm{T}} & 0 \end{pmatrix} g + \begin{pmatrix} D \\ 0 \end{pmatrix} w, \quad g(0) = \begin{pmatrix} x_0 \\ 0 \end{pmatrix},$$
$$z = C_{\mathrm{cl}} g = \begin{pmatrix} C & 0 \end{pmatrix} g$$

with respect to the augmented state vector

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

Let a given (nominal) PI controller (3) with numerical parameters

$$k = k^0 = \begin{pmatrix} k_1^0 \\ k_2^0 \end{pmatrix},$$

chosen by some considerations, stabilize the closed-loop system for (1). Two problems are of interest here.

Problem 1. What are the admissible limits to vary the parameters of the nominal PI controller with k^0 without violating the stability of the closed-loop system?

This problem is a special case of the one below.

Problem 2. It is required to find the set of stabilizing PI controllers for system (1), i.e., construct a set $\mathcal{K} \subset \mathbb{R}^2$ such that

$$\max_i \operatorname{Re} \lambda_i\big(A_{\mathrm{cl}}(k)\big) < 0 \quad \forall k \in \mathcal{K}.$$

The exact solution of Problem 2 is obtained by the D-partition method [15], which constructs the boundary $\partial\mathcal{K}$ of the set $\mathcal{K}$ analytically. However, in practice, D-partition is not as convenient for solving the same problem for a PID controller, which already includes three parameters.

Below, we propose two approaches to the approximate solution of Problem 1, which provide simple inner estimates of the desired set. In Section 3, the problem is studied from the standpoint of quadratic stability, and a simple and effective approach is presented to obtain ellipsoidal estimates of the set $\mathcal{K}$; also, a non-fragile bounding ellipsoid is constructed. The second approach (Section 4) yields the set $\mathcal{K}$ in the form of stability intervals, separately for each PI controller parameter; the application of this approach as a boundary oracle is considered.

3. FRAGILITY ANALYSIS OF A PI CONTROLLER BASED ON ELLIPSOIDAL ESTIMATION

3.1. Estimation of the Non-Fragility Radius of a PI Controller

We introduce the notation

$$A_0 = \begin{pmatrix} A & 0 \\ c^{\mathrm{T}} & 0 \end{pmatrix}, \quad F = \begin{pmatrix} -b \\ 0 \end{pmatrix}, \quad H = \begin{pmatrix} c^{\mathrm{T}} & 0 \\ 0 & 1 \end{pmatrix},$$

with

$$A_{\mathrm{cl}}(k) = A_0 + Fk^{\mathrm{T}}H.$$

Perturbing the parameters k^0 of the stabilizing PI controller (3) in the form

$$k_1^0 + \delta_1, \quad k_2^0 + \delta_2,$$

we obtain the vector $k^0 + \delta$ of the perturbed controller parameters. By assumption, the perturbations belong to some ellipse

$$\delta = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} \in \mathcal{E}(R) = \{\delta \in \mathbb{R}^2 \colon \quad \delta^{\mathrm{T}} R^{-1} \delta \leqslant 1\}, \quad R \succ 0. \tag{4}$$

It is necessary to find the largest ellipse (4) (by some criterion) such that system (1) with the stabilizing PI controller with the parameters k^0 will remain stable under all perturbations

$$\delta = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} \in \operatorname{int} \mathcal{E}(R)$$

of its parameters. In particular, an important issue is to calculate the largest *non-fragility radius* $\rho(k)$ such that system (1) will remain stable under all perturbations $\delta \colon \|\delta\| < \rho$ of the PI controller parameters.

Consider the matrix of the so-called *perturbed* system, i.e., the closed-loop one with the controller with the "perturbed" parameters:

$$A_{\mathrm{cl}}(k^0 + \delta) = A_0 + F(k^0 + \delta)^{\mathrm{T}} H = A_{\mathrm{cl}}(k^0) + F\delta^{\mathrm{T}} H. \tag{5}$$

As is well known (e.g., see [11]), if the Lyapunov inequality

$$\left(A_{\mathrm{cl}}(k^0) + F\delta^{\mathrm{T}} H\right)^{\mathrm{T}} Q + Q\left(A_{\mathrm{cl}}(k^0) + F\delta^{\mathrm{T}} H\right) \prec 0$$

with some matrix $0 \prec Q \in \mathbb{S}^{n+1}$ holds under all *admissible* uncertainties δ, then the family (5) has a common quadratic Lyapunov function $V(x) = x^{\mathrm{T}} Q x$ and is robustly quadratically stable.

The last inequality can be written as

$$A_{\mathrm{cl}}^{\mathrm{T}}(k^0) Q + Q A_{\mathrm{cl}}(k^0) + H^{\mathrm{T}} \delta F^{\mathrm{T}} Q + Q F \delta^{\mathrm{T}} H \prec 0$$

or

$$A_{\mathrm{cl}}^{\mathrm{T}}(k^0) Q + Q A_{\mathrm{cl}}(k^0) + H^{\mathrm{T}} R^{1/2} \bar{\delta} F^{\mathrm{T}} Q + Q F \bar{\delta}^{\mathrm{T}} R^{1/2} H \prec 0$$

with respect to the uncertainty $\bar{\delta} \doteq R^{-1/2}\delta$, $\|\bar{\delta}\| \leqslant 1$.

Using Petersen's lemma [16], we arrive at the matrix inequality

$$\begin{pmatrix} A_{\mathrm{cl}}^{\mathrm{T}}(k^0) Q + Q A_{\mathrm{cl}}(k^0) + \varepsilon H^{\mathrm{T}} R H & QF \\ F^{\mathrm{T}} Q & -\varepsilon I \end{pmatrix} \prec 0$$

with respect to the scalar variable ε and the matrix variables $Q, R \succ 0$. Due to the homogeneity of this LMI in the aggregate of variables, we set $\varepsilon = 1$. Thus, the following result is true.

Lemma 1. *The feasibility of the LMI*

$$\begin{pmatrix} A_{\mathrm{cl}}^{\mathrm{T}}(k^0) Q + Q A_{\mathrm{cl}}(k^0) + H^{\mathrm{T}} R H & QF \\ F^{\mathrm{T}} Q & -I \end{pmatrix} \prec 0$$

with respect to the matrix variables $0 \prec Q \in \mathbb{S}^{n+1}$ and $0 \prec R \in \mathbb{S}^2$ is equivalent to the quadratic stability of the family $A_{\mathrm{cl}}(k^0 + \delta)$ under all admissible uncertainties δ.

Now, simple inner estimates can be proposed for the admissibility region of the uncertainty δ by minimizing a certain criterion under the LMI constraint established in Lemma 1.

As is known (for more details, see [11]), an ellipsoid of the form

$$\mathcal{E} = \{x \in \mathbb{R}^n\colon \quad x^{\mathrm{T}} P^{-1} x \leqslant 1\}, \quad P \succ 0, \tag{6}$$

has volume $c_n\sqrt{\det P}$, where c_n is the volume of the unit ball in the n-dimensional space; in addition, the function $f(X) = -\operatorname{logdet} X$ of a matrix argument $X \succ 0$ is convex. Accordingly, the volume of an ellipsoid (in the two-dimensional case under consideration, the area of an ellipse) is maximized as follows.

Lemma 2. *Let $\widehat{R}_{\text{area}}$ be the solution of the problem*

$$\min(-\operatorname{logdet} R)$$

subject to the constraints

$$\begin{pmatrix} A_{\text{cl}}^{\mathrm{T}}(k^0)Q + QA_{\text{cl}}(k^0) + H^{\mathrm{T}}RH & QF \\ F^{\mathrm{T}}Q & -I \end{pmatrix} \prec 0, \quad Q \succ 0, \quad R \succ 0,$$

where optimization is performed with respect to the matrix variables $Q \in \mathbb{S}^{n+1}$ and $R \in \mathbb{S}^2$. Then system (1) *closed by the stabilizing PI controller* (3) *with the parameters k^0 remains stable under all perturbations $\delta = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}$ of its parameters such that*

$$\delta^{\mathrm{T}} \widehat{R}_{\text{area}}^{-1} \delta < 1.$$

The uncertainty ellipse can be optimized using other criteria, e.g., by maximizing the radius of the circle contained in it. (This also gives the maximum magnitude of the uncertainty δ.) As is known, if

$$P \succcurlyeq \gamma I,$$

the ellipsoid (6) will contain a ball of radius $\sqrt{\delta}$ centered at the origin (details can be found in [11, Remark 2.2.3]). Therefore, we arrive at the following result.

Lemma 3. *Let $\widehat{\gamma}$, $\widehat{R}_{\text{axis}}$ be the solution of the problem*

$$\max \gamma$$

subject to the constraints

$$\begin{pmatrix} A_{\text{cl}}^{\mathrm{T}}(k^0)Q + QA_{\text{cl}}(k^0) + H^{\mathrm{T}}RH & QF \\ F^{\mathrm{T}}Q & -I \end{pmatrix} \prec 0, \quad R \succcurlyeq \gamma I, \quad Q \succ 0,$$

where optimization is performed with respect to the matrix variables $Q \in \mathbb{S}^{n+1}$ and $R \in \mathbb{S}^2$ and the scalar variable γ. Then system (1) *closed by the stabilizing PI controller* (3) *with the parameters k^0 remains stable under all perturbations $\delta = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}$ of its parameters such that*

$$\delta^{\mathrm{T}} \widehat{R}_{\text{axis}}^{-1} \delta < 1,$$

and, in particular, those satisfying

$$\|\delta\| < \sqrt{\widehat{\gamma}}.$$

Thus, the non-fragility radius $\rho(k^0)$ of a given PI controller with the parameters k^0 is defined by the value $\sqrt{\widehat{\gamma}}$.

Remark 1. The last assertion of Lemma 3 can be obtained more simply: it suffices to set $R = rI$, which leads to the following result. Let $\widehat{r}$ be the solution of the problem

$$\max r \ \text{ subject to the constraints } \begin{pmatrix} A_{\rm cl}^{\rm T}(k^0)Q + QA_{\rm cl}(k^0) + rH^{\rm T}H & QF \\ F^{\rm T}Q & -I \end{pmatrix} \prec 0, \quad Q \succ 0,$$

where optimization is performed with respect to the matrix variable $Q \in \mathbb{S}^{n+1}$ and the scalar variable r. Then system (1) closed by the stabilizing PI controller (3) with the parameters k^0 remains stable under all perturbations $\delta = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}$ of its parameters such that

$$\|\delta\| < \sqrt{\widehat{r}}.$$

3.2. Non-Fragile Bounding Ellipsoid

Now it is natural to estimate the influence of uncertainties in the controller parameters on the performance criterion in a certain control problem associated with system (1). In this paper, we consider the classical problem of suppressing nonrandom bounded exogenous disturbances (2). An effective solution approach is to find the minimal bounding ellipsoid containing the controlled output of system (1); for details, see [11].

Within the analysis problem, we still assume a given stabilizing PI controller (3) with parameters k^0. It is required to find the minimal bounding ellipsoid for the output z under all perturbations δ of the controller parameters such that $\|\delta\| \leqslant \gamma < \rho$. It will be called the *non-fragile bounding ellipsoid* corresponding to the non-fragility level γ.

According to [11], the minimal bounding ellipsoid for the controlled output of system (1) closed by a PI controller with parameters k^0 is given by the matrix $C_{\rm cl}\widetilde{P}C_{\rm cl}^{\rm T}$, where $\widetilde{P}$ is the solution of the parameterized semidefinite programming problem

$$\min \operatorname{tr} C_{\rm cl}PC_{\rm cl}^{\rm T}$$

subject to the constraint

$$A_{\rm cl}(k^0)P + PA_{\rm cl}^{\rm T}(k^0) + \alpha P + \frac{1}{\alpha}D_{\rm cl}D_{\rm cl}^{\rm T} \preccurlyeq 0,$$

where optimization is performed with respect to the matrix variable $0 \prec P \in \mathbb{S}^{n+1}$ and the scalar parameter $\alpha > 0$. (The minimal ellipsoid is understood in terms of the so-called trace criterion, i.e., the one obtained by minimizing the sum of its squared semi-axes.)

By utilizing this result and passing to the matrix (5) of the perturbed system, we have the condition

$$\big(A_{\rm cl}(k^0) + F\delta^{\rm T}H\big)P + P\big(A_{\rm cl}(k^0) + F\delta^{\rm T}H\big)^{\rm T} + \alpha(\delta)P + \frac{1}{\alpha(\delta)}D_{\rm cl}D_{\rm cl}^{\rm T} \preccurlyeq 0,$$

holding for all $\delta\colon \|\delta\| \leqslant \gamma$. Suppose the existence of a number α such that the above matrix inequality is valid under all admissible values of δ. We represent it in the form

$$A_{\rm cl}(k^0)P + PA_{\rm cl}^{\rm T}(k^0) + \alpha P + \frac{1}{\alpha}D_{\rm cl}D_{\rm cl}^{\rm T} + PH^{\rm T}\delta F^{\rm T} + F\delta^{\rm T}HP \preccurlyeq 0$$

and use Petersen's lemma [16], taking the non-fragility level γ into account by appropriately scaling the matrix F. Consequently,

$$A(k^0)P + PA^{\mathrm{T}}(k^0) + \alpha P + \frac{1}{\alpha} D_{\mathrm{cl}} D_{\mathrm{cl}}^{\mathrm{T}} + \varepsilon \gamma^2 F F^{\mathrm{T}} + \frac{1}{\varepsilon} P H^{\mathrm{T}} H P \preccurlyeq 0$$

which is equivalent to

$$\begin{pmatrix} A_{\mathrm{cl}}(k^0)P + PA_{\mathrm{cl}}^{\mathrm{T}}(k^0) + \alpha P + \frac{1}{\alpha} D_{\mathrm{cl}} D_{\mathrm{cl}}^{\mathrm{T}} + \varepsilon \gamma^2 F F^{\mathrm{T}} & PH^{\mathrm{T}} \\ HP & -\varepsilon I \end{pmatrix} \preccurlyeq 0,$$

holding for some scalar $\varepsilon > 0$. Thus, the following fact has been established.

Theorem 1. *Let $\widehat{P}$ be the solution of the problem*

$$\min \operatorname{tr} C_{\mathrm{cl}} P C_{\mathrm{cl}}^{\mathrm{T}}$$

subject to the constraints

$$\begin{pmatrix} A_{\mathrm{cl}}(k^0)P + PA_{\mathrm{cl}}^{\mathrm{T}}(k^0) + \alpha P + \frac{1}{\alpha} D_{\mathrm{cl}} D_{\mathrm{cl}}^{\mathrm{T}} + \varepsilon \gamma^2 F F^{\mathrm{T}} & PH^{\mathrm{T}} \\ HP & -\varepsilon I \end{pmatrix} \preccurlyeq 0, \quad P \succ 0,$$

where optimization is performed with respect to the matrix variable $P \in \mathbb{S}^{n+1}$, the scalar variable ε, and the scalar parameter $\alpha > 0$.

Then the ellipsoid with the matrix $C_{\mathrm{cl}} \widehat{P} C_{\mathrm{cl}}^{\mathrm{T}}$ is a non-fragile bounding ellipsoid for the output z of system (1) *with $x_0 = 0$ closed by the stabilizing PI controller* (3) *with the parameters k^0, corresponding to the admissible non-fragility level γ.*

Remark 2. Within the approach discussed, the case of a nonzero initial condition can also be considered easily. In this case, the LMI

$$\begin{pmatrix} 1 & x_0^{\mathrm{T}} \\ x_0 & P \end{pmatrix} \succcurlyeq 0$$

is added to the constraints of the theorem, meaning that the point x_0 belongs to the invariant ellipsoid with the matrix P; see details in [11].

4. FRAGILITY ANALYSIS OF A PI CONTROLLER BY PARAMETER LOOP OPENING

4.1. Idea of the Approach

This section describes a frequency-domain approach to the fragility analysis of a PI controller based on the so-called *parameter loop opening* [12]. The approach estimates the set $\mathcal{K}$ of Problem 2 as follows: for a known stabilizing controller with parameters k^0, with one fixed parameter $k_i^0 = \mathrm{const}$, $i = 1, 2$, the stability interval (or union of intervals) is determined for the second parameter k_j, $j = 1, 2$, $i \neq j$.

The plant (1) is written in the "input–output" form as

$$a(s)y = b(s)u, \tag{7}$$

where $b(s)$ and $a(s)$ denote the numerator and denominator polynomials of its transfer function $b(s)/a(s) = c^{\mathrm{T}}(sI - A)^{-1}b$.

We apply the Laplace transform to the PI controller equation (3):

$$u = -\frac{k_1 s + k_2}{s} y. \tag{8}$$

In view of the negative feedback in the Nyquist criterion, this yields the open-loop transfer function

$$w_o(s) = \frac{b(s)}{a(s)} \frac{k_1 s + k_2}{s},$$

which is associated with the characteristic polynomial

$$D(s) = a(s)s + b(s)(k_1 s + k_2)$$

of the closed-loop system.

The Nyquist criterion [13] can be used to analyze the stability of the polynomial $D(s)$ from the Nyquist plot of the frequency response $w_o(j\omega)$. In classical automatic control theory, the robustness measure of a control system is the gain and phase margins (the frequency-domain segment stability criteria in [13]).

These classical tools for stability and robustness analysis are used in parameter loop opening [12]. Let us demonstrate the use of this method for analyzing the fragility of a nominal stabilizing PI controller k^0 with respect to the parameter k_1.

We introduce the variables

$$\tilde{u} \doteq -k_1^0 \tilde{y}, \quad \tilde{y} \doteq y,$$

where $\tilde{u}$ and $\tilde{y}$ are the fictitious control and output of the plant, respectively; then the plant equations (7) can be written as

$$a(s)y = b(s)\Big(\tilde{u} - \frac{k_2^0}{s} y\Big) \quad \Rightarrow \quad \Big(a(s) + \frac{k_2^0 b(s)}{s}\Big)\tilde{y} = b(s)\tilde{u}$$
$$\Rightarrow \quad \tilde{y} = \frac{sb(s)}{sa(s) + k_2^0 b(s)} \tilde{u} \quad \Rightarrow \quad \tilde{y} = \tilde{w}_1(s)\tilde{u}.$$

The resulting fictitious control loop consists of the transfer function $\tilde{w}_1(s)$ of a fictitious plant closed by the static controller $-k_1^0$. Performing a similar opening with respect to the parameter k_2 (setting $\tilde{u} = -k_2^0 \tilde{y}$), we get the following transfer function of the fictitious plant:

$$\tilde{w}_2(s) = \frac{b(s)}{s\big(a(s) + k_1^0 b(s)\big)}.$$

The next result is easily proved.

Lemma 4. *The characteristic polynomials obtained by closing the transfer functions of the open fictitious loops coincide with the characteristic polynomial $D(s)$ of the original closed-loop system* (7), (8).

Indeed, these open-loop transfer functions are written as

$$\tilde{w}_{oi}(s) = k_i^0 \tilde{w}_i(s), \quad i = 1, 2, \tag{9}$$

yielding $D(s)$ after closing, where the sign of the feedback in the Nyquist criterion is taken into account similarly to $w_o(s)$. In other words, the functions (9) can be employed to judge the stability of system (7), (8).

Note another important fact: both transfer functions (9) contain the PI controller parameters as cofactors. Then stability intervals of the closed-loop system with respect to the parameters k_i, $i = 1, 2$, are found by studying the intersection points of the real axis with the corresponding Nyquist plots $\mathrm{Im}\{\tilde{w}_{oi}(j\omega)\} = 0$, $i = 1, 2$. As shown in [15], such an approach is equivalent to the one-dimensional D-partition of the polynomial $D(s) = \tilde{a}_i(s) + k_i\tilde{b}_i(s)$, where $\tilde{b}_i(s)$ and $\tilde{a}_i(s)$ are the numerator and denominator of the corresponding transfer function $\tilde{w}_{oi}(s)$.

4.2. Fragility Analysis Procedure

Using the parameter k_1 as an example, we describe the procedure for finding stability intervals with respect to $\tilde{w}_{o1}(s)$ provided that all roots of $b(s)$ are nonzero and there are no zero or pure imaginary numbers among the roots of $\tilde{a}(s)$.

1. The Nyquist plot $\tilde{w}_{o1}(j\omega)$ starts and ends at the point $(0,\ j0)$ when varying ω from 0 to ∞, i.e., it forms a closed contour. Solve the polynomial equation $\mathrm{Im}\{\tilde{w}_{o1}(j\omega)\} = 0$, $\omega \in [0\ \infty)$, denoting its solutions by ω_i, $i = 1, \dots, r$.

Compute $g_i = \mathrm{Re}\,\{\tilde{w}_{o1}(j\omega_i)\}$, $i = 1, \dots, r$, and associate with each number g_i an index $f_i \in \{0,\ \pm 1/2,\ \pm 1\}$, $i = 1, \dots, r$, defined by the following rule [13, p. 141]:

$$f_i = \begin{cases} 1 & \text{if } \mathrm{Im}\{\tilde{w}_{o1}(j(\omega_i - \epsilon))\} < 0 \text{ and } \mathrm{Im}\{\tilde{w}_{o1}(j(\omega_i + \epsilon))\} > 0, \\ 1/2 & \text{if } \tilde{w}_{o1}(\omega_i) \text{ is the start of the plot and } \mathrm{Im}\{\tilde{w}_{o1}(j(\omega_i + \epsilon))\} > 0, \\ 0 & \text{if } \mathrm{Im}\{\tilde{w}_{o1}(j(\omega_i - \epsilon))\} \cdot \mathrm{Im}\{\tilde{w}_{o1}(j(\omega_i + \epsilon))\} > 0,\ i = 1, \dots, r, \\ -1/2 & \text{if } \tilde{w}_{o1}(\omega_i) \text{ is the start of the plot and } \mathrm{Im}\{\tilde{w}_{o1}(j(\omega_i + \epsilon))\} < 0, \\ -1 & \text{if } \mathrm{Im}\{\tilde{w}_{o1}(j(\omega_i - \epsilon))\} > 0 \text{ and } \mathrm{Im}\{\tilde{w}_{o1}(j(\omega_i + \epsilon))\} < 0, \end{cases}$$

where $\epsilon > 0$ is a sufficiently small number such that

$$\epsilon \ll \min_i |\omega_i - \omega_{i-1}|, \quad i = 2, 3, \dots, r.$$

2. According to the Nyquist criterion [13], the closed-loop system is stable if and only if the frequency response $\tilde{w}_{o1}(j\omega)$ has $s_{\tilde{a}}/2$ counterclockwise encirclements of the critical point $-1/k_1$, where $s_{\tilde{a}}$ is the number of unstable roots of the polynomial $\tilde{a}(s)$. For a stable $\tilde{w}_{o1}(j\omega)$, one obtains $s_{\tilde{a}} = 0$; therefore, for closed-loop stability, its frequency response shall not encircle the critical point.

The number of encirclements of the critical point can change only at the real axis points g_i, $i = 1, \dots, r$. An encirclement occurs clockwise for $f_i > 0$ and counterclockwise for $f_i < 0$; if $f_i = 0$, the number of encirclements remains invariable.

Sort the numbers g_i, $i = 1, \dots, r$, in ascending order and rearrange the subscripts of g_i and f_i accordingly.

3. First consider the case where $g_i < 0$, $i = 1, \dots, r-1$, and $g_r = 0$.

Split the real axis into the open intervals

$$(-\infty,\ g_1), \quad (g_1,\ g_2), \quad \dots, \quad (g_{r-1},\ g_r), \quad (g_r,\ \infty).$$

For the first interval $(-\infty,\ g_1)$, the number of encirclements of the critical point by the frequency response is zero, and the closed-loop system will be stable only if $\tilde{w}_1(s)$ is stable; then the parameter k_1 varies within the interval

$$k_1^{[-\infty]} \in \left[0,\ -\frac{k_1^0}{g_1}\right).$$

The left bound is included in this interval since the open-loop system (7) is stable.

For the remaining intervals, starting from the second, check the equality

$$\sum_{i=1}^{q} f_i = -\frac{s_{\tilde{a}}}{2}, \quad q = 1, \ldots, r-1, \tag{10}$$

where q is the subscript of the left bound of the interval $(g_q,\ g_{q+1})$.

For all intervals where condition (10) holds, the closed-loop system is stable; then for a particular qth interval, one has the following interval of the parameter k_1 :

$$k_1^{[-q]} \in \left(-\frac{k_1^0}{g_q},\ -\frac{k_1^0}{g_{q+1}}\right).$$

The complete stability region is the union of all stability intervals: $\bigcup k_1^{[-q]}$.

4. In the general case, g_i, $i = 1, \ldots, r$, can have different signs or be zero.

For all $g_i < 0$, perform the actions of Step 3. For $g_i = 0$, the stability of the closed-loop system is determined by the open-loop transfer function (9).

For the remaining $g_i > 0$, apply the following technique: simultaneously change the signs of $g_i > 0$ and their corresponding f_i, which means changing the sign of the feedback in the closed-loop system, and then perform the actions of Step 3, albeit with calculating the bounds of the qth stability interval as

$$k_1^{[+q]} \in \left(\frac{k_1^0}{g_q},\ \frac{k_1^0}{g_{q+1}}\right);$$

for the interval $(g_r,\ \infty)$, denote the corresponding stability interval by

$$k_1^{[+\infty]} \in \left(\frac{k_1^0}{g_r},\ 0\right).$$

The complete stability region is determined by analogy with Step 3.

Note that if stability holds in the intervals $k_1^{[-\infty]}$ and $k_1^{[+\infty]}$ simultaneously, they can be united into one common interval $k_1^{[\infty]}$. However, for the points g_i with $f_i = 0$ (the number of encirclements remains invariable!), such a union cannot be performed for stable intervals $k_1^{[-(i-1)]}$ and $k_1^{[-i]}$.

Remark 3. When analyzing a general-form transfer function $\tilde{w}_{oi}$ containing zero or pure imaginary poles, the above procedure is preserved, but it becomes more complicated to count encirclements. (See the analysis of the transfer function $\tilde{w}_{o2}$ with one zero pole in Example 2.)

The approach developed gives a complete description of stability intervals, even if the controller k^0 is not stabilizing. In this case, the approach also detects stability intervals (if they exist) with respect to one controller parameter with the second fixed.

4.3. Fragility Analysis Procedure as a Boundary Oracle

The fragility analysis procedure proposed in the previous section can be effectively used as a *boundary oracle* for an approximate description of the region of stabilizing PI (and also PID) controllers. It is based on a modification of *Hit-and-Run* [15, 17], a variant of the Monte Carlo method to generate points with the uniform distribution inside some region $\mathcal{D} \subset \mathbb{R}^m$, generally neither convex nor connected.

Hit-and-Run includes three sequential steps:

1) Through a current point $x_i \in \mathcal{D}$ in a random direction $d \in \mathbb{R}^m$, a line $x_i + \tau d$, $\tau \in (-\infty,\ \infty)$, is drawn.

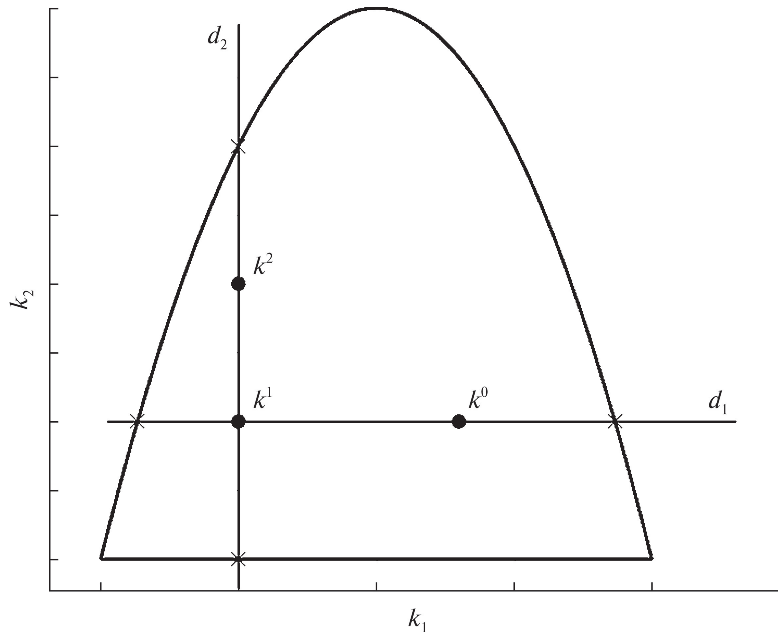

Fig. 1. The general scheme of the coordinate-wise modification of Hit-and-Run.

2) The so-called boundary oracle is used to find the intersection points of this line with the boundary of the region $\mathcal{D}$.

3) The next point x_{i+1} is chosen randomly on the resulting segment (or segments in the multiply connected case), and the process is repeated.

According to the results of Section 4.2, we implement the boundary oracle by providing the intersection points of the boundary of the region $\mathcal{K}$ (all stabilizing PI controllers) with the line $k + \tau d$, where k is some stabilizing PI controller and $d \in \mathbb{R}^2$ is one of the two directions

$$\tilde{d}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \tilde{d}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

coinciding with those of the coordinate axes in the plane (k_1, k_2). Therefore, in this (coordinate-wise) modification of the Hit-and-Run method, the direction $d^i \in \mathbb{R}^2$ at the current point k^i is chosen not randomly but as

$$d^i = \begin{cases} \tilde{d}_1 & \text{if } i \bmod 2 = 1, \\ \tilde{d}_2 & \text{if } i \bmod 2 = 0. \end{cases}$$

In other words, the directions $\tilde{d}_1$ and $\tilde{d}_2$ are alternated.

Figure 1 shows the first two iterations of this modification of the Hit-and-Run method in the parameter plane $(k_1,\ k_2)$. Since the region $\mathcal{K}$ of all stabilizing PI controllers has dimension 2 (for PID controllers, dimension 3), this simplification is not overly restrictive; the examples below confirm the effectiveness of the corresponding procedure.

5. EXAMPLES

Example 1. Consider the linearized model of a first-order plant with delay, widely used in practice [18]:

$$\frac{\kappa}{Ts+1} e^{-\tau s} \to \frac{\kappa(-\tau s/2 + 1)}{(Ts+1)(\tau s/2+1)}, \tag{11}$$

where κ is the gain, T is the time constant, and τ is the delay.

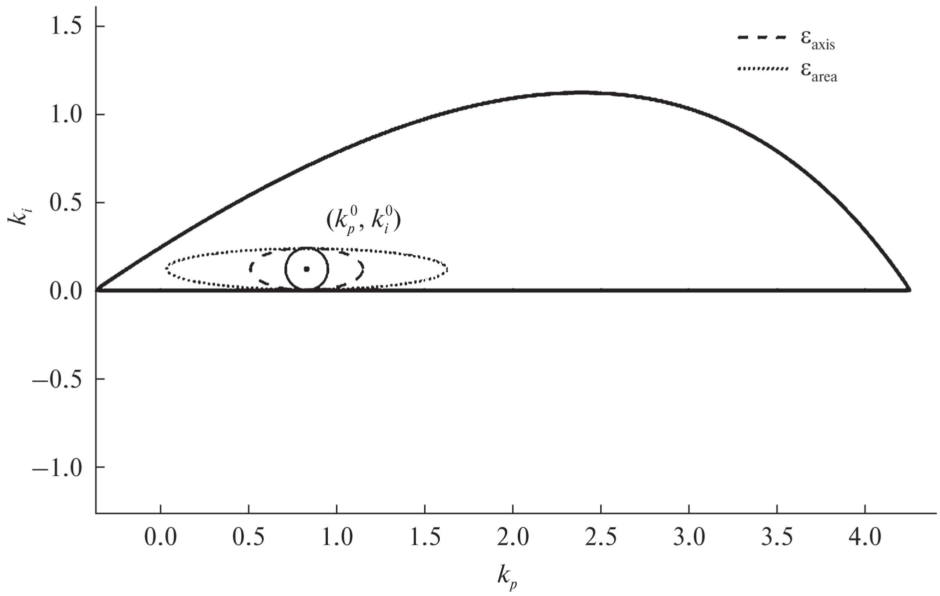

Fig. 2. The ellipses of admissible perturbations.

In the state space, this model is described by a system of the form (1) with the matrices

$$A = \begin{pmatrix} -a_1 & 1 \\ -a_2 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad c = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

where

$$a_1 = \frac{2T+\tau}{T\tau}, \quad a_2 = \frac{2}{T\tau}, \quad b_1 = -\frac{\kappa}{T}, \quad b_2 = \frac{2\kappa}{T\tau}. \tag{12}$$

Taking

$$\kappa = 2.7, \quad T = 8.4, \quad \tau = 1.6 \tag{13}$$

as the parameters of the original model, we have

$$A = \begin{pmatrix} -1.3690 & 1 \\ -0.1488 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} -0.3214 \\ 0.4018 \end{pmatrix}.$$

Let a stabilizing controller have the parameters [19]

$$k^0 = \begin{pmatrix} 0.832 \\ 0.120 \end{pmatrix}.$$

Lemmas 2 and 3 give the ellipses of admissible perturbations with the matrices

$$\widehat{R}_{\rm area} = \begin{pmatrix} 0.6343 & 0 \\ 0 & 0.0135 \end{pmatrix} \quad \text{and} \quad \widehat{R}_{\rm axis} = \begin{pmatrix} 0.1021 & 0 \\ 0 & 0.0144 \end{pmatrix},$$

respectively; and the non-fragility radius is $\rho(k^0) = 0.12$.

In Fig. 2, the bold lines show the set $\mathcal{K}$ of all stabilizing PI controllers yielded by D-partition (for details, see [15]), as well as the ellipses of admissible perturbations: the dotted line corresponds to the ellipse $\mathcal{E}_{\rm area}$, obtained by maximizing the area; the dashed line, to the ellipse $\mathcal{E}_{\rm axis}$, obtained by

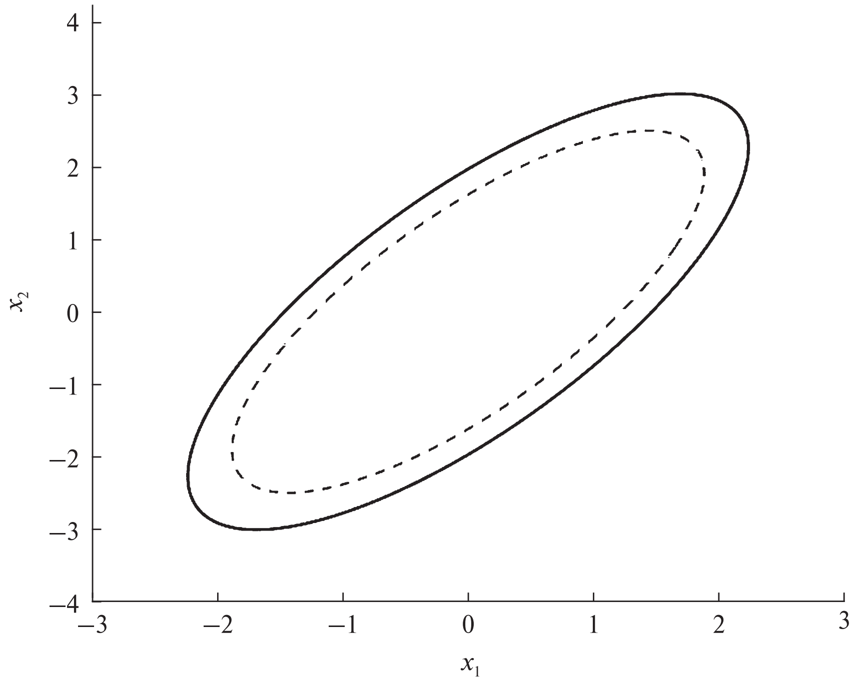

Fig. 3. The non-fragile bounding ellipsoid.

maximizing the length of the smallest semi-axis. The thin solid line is a circle with the non-fragility radius $\rho(k^0)$. The point indicates the PI controller with the parameters k^0.

We choose the controlled output $z = x$ (the full state of the system), i.e., $C = I$, and set $D = b$. The minimal bounding ellipsoid for the system closed by the PI controller with the parameters k^0 is given by the matrix

$$C_{\rm cl}\widetilde{P}C_{\rm cl}^{\rm T} = \begin{pmatrix} 3.5473 & 3.5892 \\ 3.5892 & 6.2562 \end{pmatrix}.$$

Let the non-fragility level be $\gamma = 0.02$. Theorem 1 provides the matrix

$$C_{\rm cl}\widehat{P}C_{\rm cl}^{\rm T} = \begin{pmatrix} 5.0026 & 5.0881 \\ 5.0881 & 9.0666 \end{pmatrix}$$

of the non-fragile bounding ellipsoid corresponding to this non-fragility level.

In Fig. 3, the dotted line shows the minimal bounding ellipsoid constructed without the non-fragility requirement, and the solid line is the non-fragile bounding ellipsoid found; it exceeds the minimal one by about 40% in terms of the trace criterion.

In all examples, numerical simulation was carried out in MATLAB using the `cvx` package [20].

Example 2. We return to the plant from Example 1. The transfer function (11), written in the form (7), is

$$(s^2 + a_1 s + a_2)y = (b_1 s + b_2)u;$$

in view of (12)–(13), it becomes

$$(s^2 + 1.369s + 0.149)y = (-0.321s + 0.402)u.$$

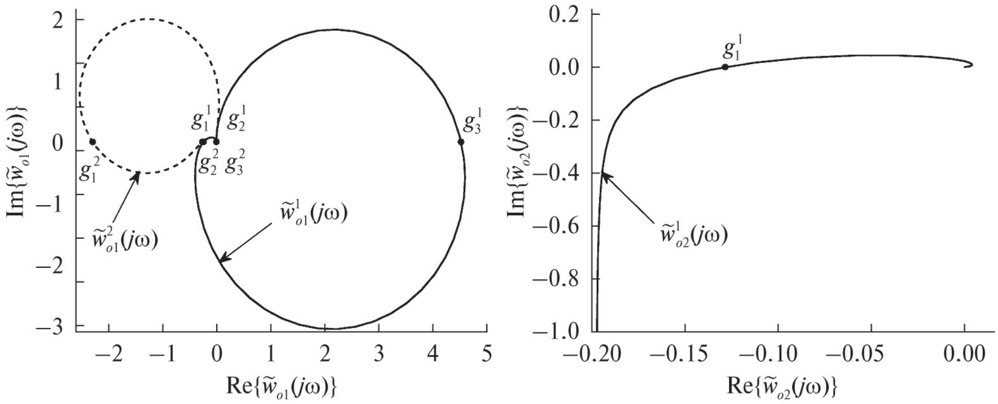

Fig. 4. The frequency response of the open-loop system: $\tilde{w}_{o1}^i(j\omega)$, $i = 1, 2$ (left) and $\tilde{w}_{o2}^1(j\omega)$ (right).

Using the procedure described in Section 4, we analyze the fragility of two stabilizing PI controllers with the parameters

$$k^1 = \begin{pmatrix} k_1^1 \\ k_2^1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.1 \end{pmatrix}, \quad k^2 = \begin{pmatrix} k_1^2 \\ k_2^2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}.$$

We start with the parameter k_1. In the corresponding transfer functions (9), the superscript indicates which of the PI controllers, k^1 or k^2, is being analyzed. (Similar notation will be employed for all parameters of the fragility analysis procedure.) As a result,

$$\tilde{w}_{o1}^1(s) = \frac{-0.321s^2 + 0.402s}{s^3 + 1.369s^2 + 0.117s + 0.0402},$$
$$\tilde{w}_{o1}^2(s) = \frac{-0.321s^2 + 0.402s}{s^3 + 1.369s^2 - 0.012s + 0.201}.$$

Their frequency responses are demonstrated in Fig. 4 on the left, namely, $\tilde{w}_{o1}^1(j\omega)$ (the solid line) and $\tilde{w}_{o1}^2(j\omega)$ (the dotted line).

The polynomial equations $\mathrm{Im}\{\tilde{w}_{o1}^i(j\omega)\} = 0$, $i = 1, 2$, have the solutions

$$\omega_1^1 = 0, \quad \omega_2^1 = 0.1670, \quad \omega_3^1 = 1.3417,$$
$$\omega_1^2 = 0, \quad \omega_2^2 = 0.4043, \quad \omega_3^2 = 1.2393,$$

and the corresponding pairs of g_i^j and f_i^j, $j = 1, 2$, $i = 1, 2, 3$ (sorted in ascending order of g_i^j) are

$$\{g_1^1 = -0.2387,\ f_1^1 = 1\}, \quad \{g_2^1 = 0,\ f_2^1 = 1/2\}, \quad \{g_3^1 = 4.5263,\ f_3^1 = -1\},$$
$$\{g_1^2 = -2.2907,\ f_1^2 = -1\}, \quad \{g_2^2 = -0.2596,\ f_2^2 = 1\}, \quad \{g_3^2 = 0,\ f_3^1 = 1/2\}.$$

They are presented in Fig. 4 on the left, with the values g_i^1 above the real axis and g_i^2 below it.

We begin with the frequency response $\tilde{w}_{o1}^2(j\omega)$ since the corresponding arrangement of the points g_i^2, $i = 1, 2, 3$, coincides with the assumption of Step 3 of the fragility analysis procedure. The transfer function $\tilde{w}_{o1}^2(s)$ has a pair of unstable complex conjugate poles; therefore, $s_{\tilde{a}}^2 = 1$, and condition (10) holds only for $q = 1$ on the interval $(g_1^2,\ g_2^2)$. As a result, we arrive at the following interval:

$$k_1^{[-1]} \in (0.4366,\ 3.8523).$$

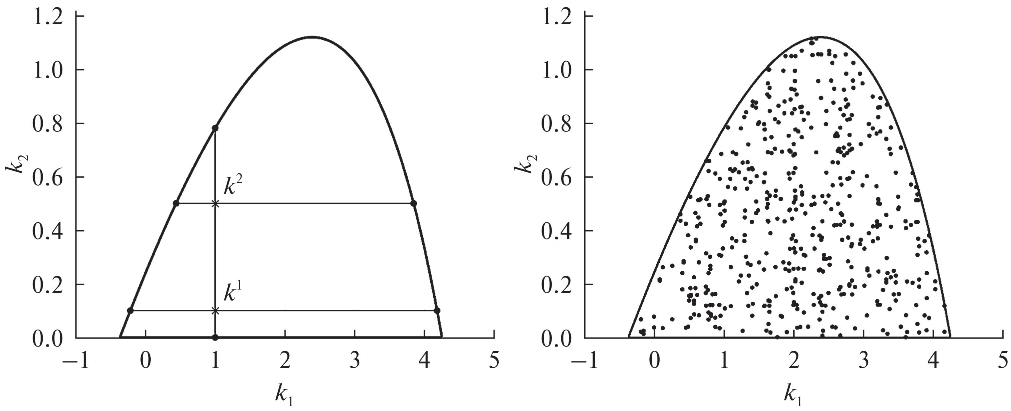

Fig. 5. Comparison with D-partition: fragility analysis of k^1 and k^2 (left) and Hit-and-Run implementation (right).

The transfer function $\tilde{w}_{o1}^1(s)$ is stable ($s_{\bar{a}}^1 = 0$), but g_i^1, $i = 1, 2, 3$, have different signs; hence, it is necessary to use Step 4 of the fragility analysis procedure of PI controllers. Considering g_i^1 shows that $k_1^{[-\infty]}$ and $k_1^{[+\infty]}$ are stable and can be united, finally yielding

$$k_1^{[\infty]} \in (-0.2209,\ 4.1898).$$

When analyzing the parameter k_2, any of the controllers k^1 and k^2 can be chosen because $k_1^1 = k_1^2 = 1$. We will use k^1, the corresponding transfer function (9) is

$$\tilde{w}_{o2}^1(s) = \frac{-0.0321s + 0.0402}{s(s^2 + 1.048s + 0.551)}.$$

This function is neutral: it has one zero pole and the others are stable ($s_{\bar{a}}^1 = 0$).

The frequency response $\tilde{w}_{o2}^1(j\omega)$ is shown in Fig. 4 on the right. The polynomial equation $\operatorname{Im}\{\tilde{w}_{o2}^1(j\omega)\} = 0$ has the solution $\omega_1 = 0.547$, giving the pair

$$\{g_1^2 = -0.128,\ f_1^2 = 1\}.$$

Stability holds on the interval $(-\infty,\ g_1^2)$ or, recalculated into the parameter k_2,

$$k_2^{[-\infty]} \in (0,\ 0.781).$$

Here, the left bound does not include zero since $\tilde{w}_{o2}^1(s)$ is not asymptotically stable.

In Fig. 5 on the left, in the parameter plane $(k_1,\ k_2)$, the bold lines present the set of stabilizing PI controllers found by D-partition [15], the crosses indicate the controllers with the parameters k^1 and k^2, and the thin lines demonstrate the stability intervals for each parameter.

In Fig. 5 on the right, the results of the coordinate-wise modification of the Hit-and-Run method are shown (see Section 4.3). A PI controller with the parameters $k^2 = (1,\ 0.5)^{\mathrm{T}}$ was chosen as the starting point, and 500 points were generated.

Example 3. Consider a plant (7) of the form

$$(s^3 + s^2 + s + 1)y = (s^2 + 2s + 5)u,$$

which is regulated by a PI controller with the parameters $k = (2,\ 0.01)^{\mathrm{T}}$.

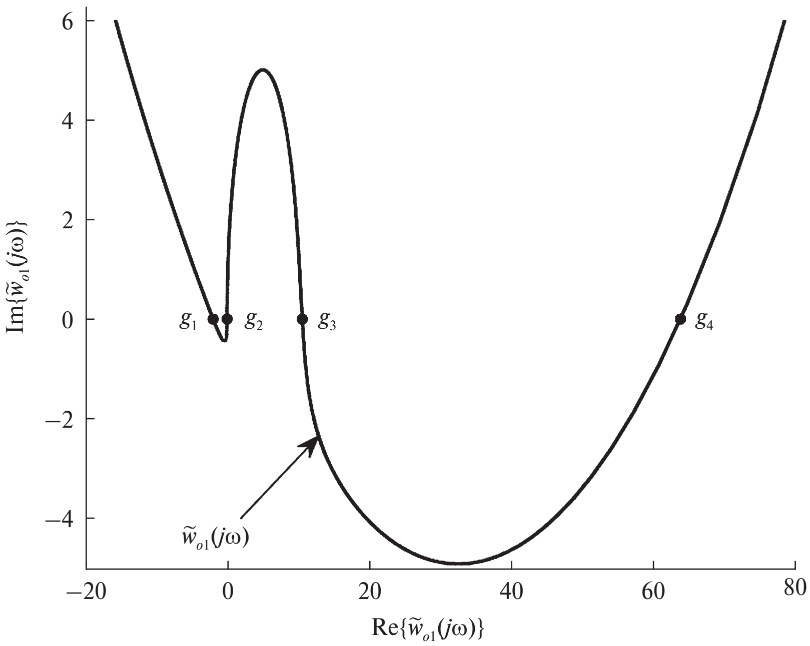

Fig. 6. The frequency response $\tilde{w}_{o1}(j\omega)$ of the open-loop system.

We analyze fragility with respect to the parameter $k_1 = 2$ using the procedure from Section 4. The transfer function (9) is written as

$$\tilde{w}_{o1}(s) = \frac{2s^3 + 4s^2 + 10s}{s^4 + s^3 + 1.01s^2 + 1.02s + 0.05};$$

it has one pair of unstable complex conjugate poles and $s_{\tilde{a}} = 1$ for it. The frequency response $\tilde{w}_{o1}(j\omega)$ is shown in Fig. 6.

The polynomial equation $\mathrm{Im}\{\tilde{w}_{o1}(j\omega)\} = 0$ has the solution

$$\omega_1 = 0, \quad \omega_2 = 0.3045, \quad \omega_3 = 0.944, \quad \omega_4 = 1.7396,$$

yielding the corresponding pairs of g_i and f_i, $i = 1, 2, 3, 4$:

$$\{g_1 = -1.9679,\ f_1 = -1\}, \quad \{g_2 = 0,\ f_2 = 1/2\},$$
$$\{g_3 = 10.5841,\ f_3 = -1\}, \quad \{g_4 = 63.8024,\ f_4 = 1\}$$

(see Fig. 6).

Condition (10) holds on the intervals $(g_1,\ g_2)$ and $(g_3,\ g_4)$; therefore, for $k_2 = 0.01$, the closed-loop system will be stable on the intervals

$$\bigcup_q k_1^{[q]}, \quad q \in \{-1,\ +3\},$$

where

$$k_1^{[-1]} \in (1.0163,\ \infty), \quad k_1^{[+3]} \in (-0.189,\ -0.0313).$$

6. CONCLUSIONS

This paper has proposed two approaches to the fragility analysis of PI controllers. The first one is based on the technique of linear matrix inequalities; for a nominal stabilizing PI controller, it effectively constructs ellipsoidal estimates for admissible uncertainties in the PI controller parameters under which the system remains quadratically stable. Also, a method for constructing a non-fragile bounding ellipsoid has been presented.

The second approach, involving the Nyquist criterion, provides intervals of static variations for each PI controller parameter under which the closed-loop system is still stable. The use of this approach as a boundary oracle in a modification of the Hit-and-Run method has been described.

Both approaches are simply implementable and employ standard functions available in most modern software tools for engineers.

A natural further development is to extend the above results to the fragility analysis of PID controllers, as well as to apply these approaches to design non-fragile PI and PID controllers.

FUNDING

The research presented in Sections 3 and 5 was supported by the Russian Science Foundation, project no. 25-29-20062, https://rscf.ru/en/project/25-29-20062/.

REFERENCES

1. Alexandrov, A.G. and Palenov, M.V., Adaptive PID Controllers: State of the Art and Development Prospects, *Autom. Remote Control*, 2014, vol. 75, no. 2, pp. 188–199.
2. The IFAC Newsletter. No. 2. April, 2019. URL https://www.ifac-control.org/newsletter_archive/IFAC_Newsletter_2019_2_April.pdf
3. He, J.-B., Wang, Q.-G., and Lee, T.-H., PI/PID Controller Tuning via LQR Approach, *Chemical Engineering Science*, 2000, vol. 55, no. 13, pp. 2429–2439.
4. Polyak, B.T., and Khlebnikov, M.V., New Criteria for Tuning PID Controllers, *Autom. Remote Control*, 2022, vol. 83, no. 11, pp. 1724–1741.
5. Gryazina, E.N., Polyak, B.T., and Tremba, A.A., Design of the Low-Order Controllers by the H_∞ Criterion: A Parametric Approach, *Autom. Remote Control*, 2007, vol. 68, no. 3, pp. 456–466.
6. Han, S., Keel, L.H., and Bhattacharyya, S.P., PID Controller Design with an H_∞ Criterion, *IFAC-PapersOnLine*, 2018, vol. 51, no. 4, pp. 400–405.
7. Jin, L., and Kim, Y.C., Fixed, Low-Order Controller Design with Time Response Specifications Using Non-Convex Optimization, *ISA Transactions*, 2008, vol. 47, no. 4, pp. 429–438.
8. Hast, M., Astrom, K.J., Bernhardsson, B., and Boyd, S., PID Design by Convex-Concave Optimization, *Proc. 2013 European Control Conference (ECC)*, Zurich, Switzerland, 2013, pp. 4460–4465.
9. Keel, L.H., and Bhattacharyya, S.P., Robust, Fragile, or Optimal?, *IEEE Trans. Autom. Control*, 1997, vol. 42, no. 8, pp. 1098–1105.
10. Boyd, S., El Ghaoui, L., Feron, E., et al., *Linear Matrix Inequalities in System and Control Theory*, Philadelphia: SIAM, 1994.
11. Polyak, B.T., Khlebnikov, M.V., and Shcherbakov, P.S., *Upravlenie lineinymi sistemami pri vneshnikh vozmushcheniyakh: Tekhnika lineinykh matrichnykh neravenstv* (Control of Linear Systems Subjected to Exogenous Disturbances: The Technique of Linear Matrix Inequalities), Moscow: LENAND, 2014.
12. Chestnov, V.N., An Approach to the Problem of Tolerance Design for Parameters of Linear Multivariable Systems, *Izv. Ross. Akad. Nauk. Teor. Sist. Upravlen.*, 1995, vol. 33, no. 6, pp. 43–50.

13. Tsypkin, Ya.Z., *Osnovy teorii avtomaticheskogo upravleniya* (Fundamentals of the Theory of Automatic Control), Moscow: Nauka, 1977.

14. Khlebnikov, M.V., PI Controller Design for Suppressing Exogenous Disturbances, *Autom. Remote Control*, 2023, vol. 84, no. 8, pp. 901–917.

15. Gryazina, E.N., Polyak, B.T., and Tremba, A.A., D-Decomposition Technique State-of-the-Art, *Autom. Remote Control*, 2008, vol. 69, no. 12, pp. 1991–2026.

16. Khlebnikov, M.V. and Shcherbakov, P.S., Petersen's Lemma on Matrix Uncertainty and Its Generalization, *Autom. Remote Control*, 2008, vol. 69, no. 11, pp. 1932–1945.

17. Smith, R.L., Efficient Monte Carlo Procedures for Generating Points Uniformly Distributed over Bounded Regions, *Oper. Res.*, 1984, vol. 32, no. 6, pp. 1296–1308.

18. Åström, K.J. and Hägglund, T., Benchmark Systems for PID Control, *IFAC Proceedings Volumes*, 2000, vol. 33, no. 4, pp. 165–166.

19. Shatov, D.V., PI and PID Controllers Design for Tracking Systems via LQ Criterion, *Proc. 2023 5th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*, Lipetsk, 2023, pp. 551–556.

20. Grant, M. and Boyd, S., CVX: Matlab Software for Disciplined Convex Programming, ver. 2.1. URL http://cvxr.com/cvx

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2026, Vol. 87, No. 2, pp. 159–169.

NONLINEAR SYSTEMS

Delay Estimation to Ensure the Maximum Degree of Convergence in Consensus Problems

R. P. Agaev[*,a] **and D. K. Khomutov**[*,b]

[*] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
e-mail: [a]*agaraf3@gmail.com*, [b]*homutov_dk@mail.ru*

Received May 24, 2025
Revised October 29, 2025
Accepted November 6, 2025

Abstract—The Lambert W function is used to study a linear consensus model in multi-agent systems with delay. In particular, the case where all nonzero eigenvalues of the Laplacian matrix are real is considered. An explicit expression is obtained for the delay ensuring the maximum degree of convergence. A formula for the maximum degree of convergence is derived. As proved, the maximum degree of convergence depends only on the maximum and minimum nonzero eigenvalues, while the other eigenvalues have no influence on this characteristic. The results presented are a basis for one still unsolved problem, i.e., the direct estimation of the convergence rate in multi-agent systems with a directed structure.

Keywords: Lambert W function, degree of convergence, degree of stability, control with delay, Laplacian matrix, multi-agent system, consensus

DOI: 10.7868/S1608303226020032

1. INTRODUCTION

Several results on the theory of multi-agent systems with information links and delay were obtained in [1–9]. However, the dependence of the convergence rate of such systems with delay on their spectral properties remains a relevant topic for research. The degree of convergence conceptualized in this paper and its estimate can serve as a basis for estimating the convergence rate. This is valid for models of multi-agent systems with a Laplacian matrix and, moreover, for systems with an arbitrary stable matrix (when estimating the rate of reaching consensus).

In this paper, we apply the Lambert W function to obtain the main results. Let us recall some key results on this topic.

Over the past 20 years, after the famous publication [10], a series of works on control of systems with delay using the Lambert W function have appeared [11–16]. The *Lambert W function $W(z)$* is defined by $z = W(z)e^{W(z)}$, i.e., $W(z)$ is the inverse function of $f(z) = ze^z$. The application of this function has proven very effective and, in some cases, allows establishing necessary statements in a shorter way.

The system

$$\dot{y}(t) = Ay(t) + By(t-\tau) \tag{1}$$

with simultaneously triangularizable matrices A and B was studied in [12]; due to this property, the matrix equation can be reduced to a scalar one. A necessary and sufficient condition for robust stability of the corresponding system was presented under certain inequalities with undetermined coefficients. Similar to the results in [12], an expression for the spectrum of system (1) with simultaneously triangularizable or commuting matrices A and B was also derived in [13]. (The overlapping statements in [12] and [13] were obtained simultaneously and independently.)

Consider in detail the method for computing the primary matrix Lambert W function described in [14]. The iterative method (4.4) proposed in [14] was called by the authors a stable variant of the simplified Newton method, namely, that of the iterations $y_{k+1} = ({y_k}^2 + ae^{-y_k})/(y_k + 1)$. Since this formula directly follows from $y_{k+1} = y_k - f(y_k)/f'(y_k)$, it was termed the Newton–Raphson method in [14]. To determine the primary Lambert W function satisfying the equation $We^W - A = 0$, expressed as a polynomial in a matrix A, the Schur decomposition for A is used: the matrix A is replaced by its Schur form $T = Q^{\mathrm{T}}AQ$, where T is a block-diagonal matrix whose diagonal blocks are of order 1 (order 2) if they correspond to a real eigenvalue (to a pair of complex eigenvalues, respectively). Also, the Schur matrix T is represented as a block-triangular matrix with two blocks on the diagonal:

$$T = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}.$$

For each diagonal block T_{11} and T_{22}, the proposed Newton method with chosen initial matrices leads to the matrices $X_{11} = W_k(T_{11})$ and $X_{22} = W_k(T_{22})$. The missing block X_{12} is found from the Sylvester equation (see [14]). Thus, one obtains

$$W_k(T) = \begin{pmatrix} X_{11} & X_{12} \\ 0 & X_{22} \end{pmatrix}.$$

The inverse transformation $A = QTQ^T$ is applied to compute the primary matrix $W_k(A)$ satisfying the equation $W_k(A)e^{W_k(A)} - A = 0$.

In the scalar case, the stability region of a system with a delay link was also constructed by using the Lambert W function in [16].

The reader can find other applications of the Lambert W function in physical problems in the book [17], where, in particular, results on conformal mappings with the Lambert W function acting on rectangular domains in the z-plane were presented. For a system governed by the equation $x^{(\alpha)}(t) = Bx(t-1)$ with unit delay and the derivative of an integer or fractional order α, a stability analysis using the Lambert W function was described by the same authors in [18].

2. NECESSARY BACKGROUND AND AUXILIARY RESULTS

In this section, when presenting some properties of the Lambert W function, we utilize the results from the famous paper [10] (also, see [19]). The properties of the Lambert W function were also considered in [17].

As in the case of a complex variable, the real Lambert W function $W(x)$ is defined as a solution of the functional equation $W(x)e^{W(x)} = x$.

The function $W(x)$ is defined on the interval $-e^{-1} \leqslant x < +\infty$, where it takes values from $-\infty$ to $+\infty$. For negative arguments, $W(x)$ is negative and two-valued. At the point $(-e^{-1}, -1)$ the graph splits into two branches: the upper one is the principal (or zeroth) branch $W_0(x)$, and the lower one is the minus-first branch $W_{-1}(x)$. In Fig. 1, the zeroth branch is drawn with a solid line and the -1st branch with a dashed line.

In the complex domain, the Lambert W function is multivalued. The complex plane is mapped to a value set (range) $w = W_k(z)$, $k = 0, \pm 1, \pm 2, \ldots$, $z = x + iy$, $w = u + iv$. The definitional domain of $W_0(z)$ is the entire complex plane with the cut $(-\infty, -e^{-1}]$, i.e., with a cut along the negative real semi-axis: $\{z : -\infty < \mathrm{Re}(z) \leqslant -e^{-1}, \Im(z) = 0\}$. For other branches ($k \neq 0$) the definitional domain is the entire plane with the cut $(-\infty, 0]$, i.e., $\{z : -\infty < \mathrm{Re}(z) \leqslant 0, \Im(z) = 0\}$. Note that the above cuts up to the branching points allow defining single-valued branches of the Lambert

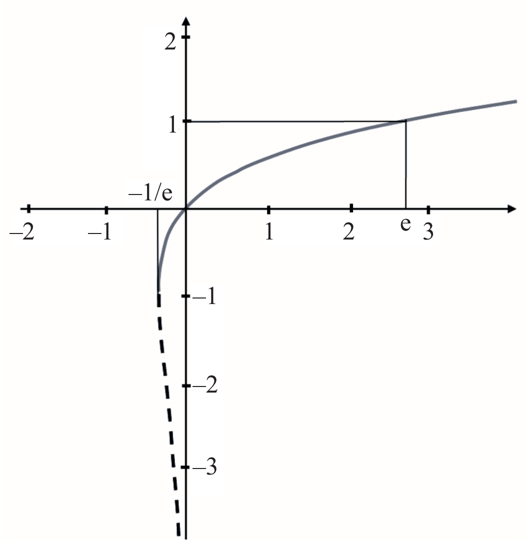

Fig. 1. Two branches of the Lambert function for the real case.

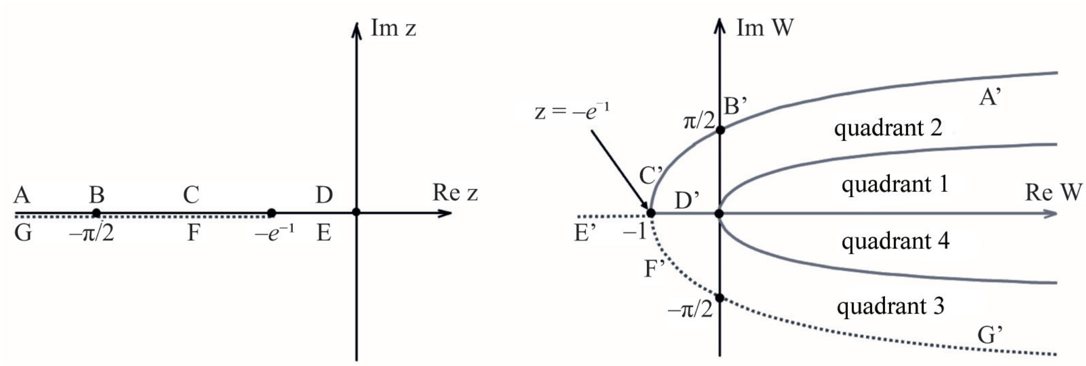

Fig. 2. The solid line shows the mapping of the imaginary and real axes by the function $W_0(z)$. The dashed line, defining the upper boundary of the branch $W_{-1}(z)$, indicates the mappings of the cut and the interval $(-e^{-1}, 0)$ by the function $W_{-1}(z)$.

function that do not transform into each other. Due to $W(z)e^{W(z)} = z$,

$$\begin{cases} x = e^u(u\cos v - v\sin v), \\ y = e^u(v\cos v + u\sin v). \end{cases} \tag{2}$$

From the condition $y = 0$ in (2) we obtain the expression $u = -v\cot v$, $-\pi < v < \pi$, defining the boundary between the range of the branch W_0 and those of the branches W_{-1} and W_1 (see Fig. 2).

Remark 1. For the curve $u = -v\cot v$, the point $v = 0$ is a removable discontinuity point. However, as $W_0(-e^{-1}) = -1$ and $W_{-1}(-e^{-1}) = -1$, the boundary between the ranges will have no discontinuity.

The curves separating the ranges of other branches are defined by the following set of points [10]:

$$\{(u, v) : u = -v\cot v,\ 2\pi k < \pm v < (2k+1)\pi\},\ k = 1, 2, \ldots.$$

Note that for one value of k, two curves are obtained, defining the boundaries between the branches W_k, W_{k+1} and W_{-k}, W_{-k-1}.

To prove the main results, the following fact [12, Lemma 3] will be needed.

Lemma 1. *For any* $z \in \mathbb{C}$,

$$\max\{\mathrm{Re}(W_k(z)) : k = 0, \pm 1, \pm 2, \ldots\} = \mathrm{Re}(W_0(z)).$$

A multi-agent system with a set of agents $V = \{1, \ldots, n\}$ is conveniently represented as a dependency digraph $\Gamma = (V, E)$, where $E \subset V \times V$ is a set of weighted arcs. If agent j influences agent i with a weight a_{ij}, then there exists an arc from vertex j to vertex i with the same weight.

Definition 1. The Laplacian matrix $L = (l_{ij})$ corresponding to a digraph Γ is constructed as follows:

$$l_{ij} = \begin{cases} -a_{ij}, & j \neq i, \\ \sum\limits_{k \neq i} a_{ik}, & j = i. \end{cases}$$

Consider the basic consensus protocol in a first-order multi-agent system with delay

$$\begin{cases} \dot{x}(t) = -Lx(t - \tau), & t > 0, \\ x(\theta) = \phi(\theta), & \theta \in [-\tau, 0], \end{cases} \tag{3}$$

where $x_i(t)$ denotes the characteristic of agent i at a time instant t and $x(t) = (x_1(t), \ldots, x_n(t))^T$ is the column vector of agents' characteristics. The time delay τ in the system can arise from data transmission, measurement, control, etc.

Definition 2. The protocol (3) converges to consensus if there exists the finite limit $\lim_{t \to \infty} x(t) = c\mathbf{1}$, where $x(t)$ is the solution of system (3), $\mathbf{1}$ is the column vector of ones, and c is some constant (the consensus value). In such a case, we say that system (3) reaches asymptotic consensus (or simply, reaches consensus).

For each initial function in Definition 2, only the existence of the constant c is required. However, as proved in [21], for the protocol (3) the value of c depends on the initial function.

For L, zero is always an eigenvalue, and its algebraic multiplicity coincides with the geometric one. This paper considers a system for which L has only a real spectrum with a simple zero eigenvalue and a simple structure. For a graph, the condition of a simple zero eigenvalue is equivalent to connectivity; for a digraph, to the presence of a spanning tree. This condition is necessary for reaching consensus for any initial value vector. (And for $\tau = 0$, it is also sufficient.) The class of Laplacian matrices L satisfying these two conditions is the minimal extension of the class of connected symmetric matrices L. The nonzero eigenvalues of an arbitrary matrix L have positive real parts, which can be verified, e.g., using Gershgorin's circle theorem.

For the protocol to converge, it is necessary that the real parts of all nonzero roots of the characteristic function be negative. The roots of the characteristic function depend on τ. The boundary value τ_0 is a delay value such that consensus is reached in the system for any initial function $\phi(\theta)$ if $\tau < \tau_0$ and is not achieved if $\tau > \tau_0$. The dependence of the boundary delay on the spectral properties of the corresponding matrix has been studied by many researchers. For example, for system (3) with a symmetric matrix, the boundary delay $\tau_0 = \frac{\pi}{2\lambda_{\max}}$ was obtained in [20]. However, a linear system with an arbitrary stable matrix A (and a spectrum possibly containing complex numbers) was considered earlier in [22]; the authors gave a rather complicated proof of a theorem providing an expression for the boundary value τ_0. This result for the consensus problem is presented below as Theorem 1, including a short proof using the Lambert W function.

The characteristic function of system (3) is

$$F(s) = \det(sI + e^{-\tau s}L) = s \prod_{\lambda \in \sigma(L) \setminus \{0\}} (s + \lambda e^{-\tau s}) = s \prod_{\lambda \in \sigma(L) \setminus \{0\}} f_\lambda(s), \tag{4}$$

where $\sigma(L)$ denotes the spectrum of the matrix L.

We write the equation

$$s + \lambda e^{-\tau s} = 0 \tag{5}$$

as $\tau s e^{\tau s} = -\tau\lambda$.

Since $W(z)e^{W(z)} = z$, the number τs (s being the root of equation (5)) is the value of the Lambert W function at $-\tau\lambda$:

$$W(-\tau\lambda) = \tau s. \tag{6}$$

Analogously to the *degree of stability*, a concept introduced for stable matrices, we define the degree of convergence to consensus for system (3).

Definition 3. The number $\zeta_0 = -\max(\mathrm{Re}(s))$, where s is the zero of the function $G(s) = \prod_{\lambda\in\sigma(L)\setminus\{0\}} f_\lambda(s)$, is called the degree of convergence to consensus for system (3).

The aim of this paper is to find the delay value τ^* maximizing the degree of convergence. However, even in the case of a matrix with a real spectrum, the degree of convergence for systems with delay does not always characterize the convergence rate of the corresponding protocol. (For symmetric networks without delay, the degree of convergence coincides with the convergence rate and is called the *Fiedler number*.) Therefore, the concept of the degree of convergence characterizes, to a greater extent, the "margin" of convergence to consensus (like stability margins in control theory). In [23], $\gamma^* = -\zeta_0$ was called the spectral abscissa function, and the corresponding root of the characteristic quasipolynomial was called the dominant root. Based on dominant roots, the authors claimed that for some $\gamma \in \mathbb{R}$, the value $\gamma^* = -\gamma$ approximates the exponential decay rate of the agents' states via the rightmost roots [23].

If consensus is achieved in the system, then $\zeta_0 > 0$. By Lemma 1, to study the degree of convergence, as well as to find the boundary delay, it suffices to consider the root corresponding to the zeroth branch of the Lambert W function.

As mentioned above, for any stable matrix A, the theorem below was established earlier [22, Theorem 3.4]. Here, we provide its short proof for system (3). Note that Theorem 3.4 from [22] can be proved by analogy.

Theorem 1. *If* 0 *is a simple eigenvalue of* L, *then the consensus protocol* (3) *converges for any* $\tau < \tau_0 = \min\limits_{\lambda\in\sigma(L)\setminus\{0\}} \frac{1}{\rho}\left(\frac{\pi}{2} - \varphi\right)$, *where* $\rho = |\lambda|$ *and* $\varphi = |\arg(\lambda)|$.

3. MAIN RESULTS

Consider system (3) in which the matrix L has a real spectrum, i.e., $\sigma(L) = \{0 = \lambda_1 < \lambda_2 \leqslant \lambda_3 \leqslant \ldots \leqslant \lambda_n\}$. For such a system, we pose the following problem: find the value of τ^* ensuring the maximum degree of convergence. For this purpose, under a fixed $\lambda \neq 0$, it is necessary to analyze the behavior of the function $s = \frac{1}{\tau}W_0(-\tau\lambda)$ on the interval $-\pi/2 < -\tau\lambda < 0$. (For $-\tau\lambda < -\pi/2$ the real parts of $W_0(-\tau\lambda)$ are positive.)

Proposition 1. 1) *The real part of the function* $s = \frac{1}{\tau}W_0(-\tau\lambda)$ *increases on the interval* $-\pi/2 < -\tau\lambda < -e^{-1}$. 2) *On the interval* $-e^{-1} < -\tau\lambda < 0$, *this function decreases.*

Proposition 2. *Let* s_0 *be the solution of the equation* $f_\lambda(s) = s + \lambda e^{-\tau s} = 0$ $(\lambda > 0)$ *with the maximum real part. Then* $\arg\min\limits_{\tau} \mathrm{Re}(s_0) = \tau^*_\lambda = \frac{1}{e\lambda}$.

Proposition 3. *For each* $\lambda_i > 0$, $i = 2, \ldots, n$, *let the minimum real part of the solution of equation* (5) *be achieved at* $\tau^*_{\lambda_i}$. *Then*:

1) *If* $\lambda_2 < \lambda_n$, *the maximum degree of convergence of system* (3) *will be achieved at* $\tau^* \in (\tau^*_{\lambda_n}, \tau^*_{\lambda_2})$. 2) *If* $\lambda_2 = \lambda_n = \lambda$, *then* $\tau^* = \tau^*_\lambda$.

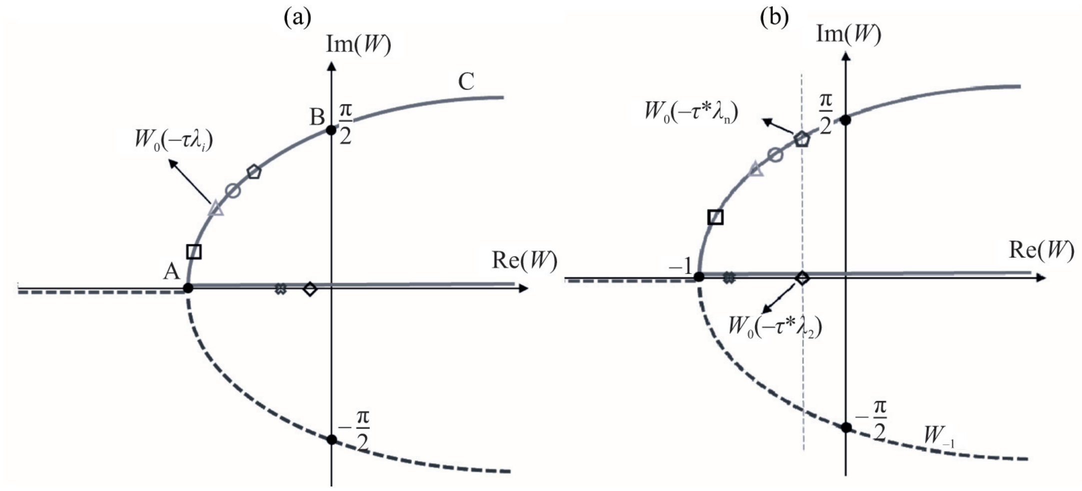

Fig. 3. (a) The values of $W_0(-\tau\lambda)$ and (b) the values of $W_0(-\tau^*\lambda_2)$ and $W_0(-\tau^*\lambda_n)$ for $\tau = \tau^*$. The dashed line indicates the mapping of $W_{-1}(z)$ whereas the solid line the mapping of $W_0(z)$.

Thus, for any $\tau \in (\tau^*_{\lambda_n}, \tau^*_{\lambda_2})$, the value $W_0(-\tau\lambda_n)$ will lie on the arc AC, and $W_0(-\tau\lambda_2)$ will belong to the interval $(-1, 0)$ (see Fig. 3a).

If the real part of $W_0(-\tau\lambda_n)$ is smaller than $W_0(-\tau\lambda_2)$, then by increasing τ, we can achieve the case $\mathrm{Re}(W_0(-\tau\lambda_n)) = W_0(-\tau\lambda_2)$. And if $\mathrm{Re}(W_0(-\tau\lambda_n)) > W_0(-\tau\lambda_2)$, then by decreasing τ, we can achieve the equality here (see Fig. 3b).

The result below provides an expression for τ ensuring the maximum degree of convergence ζ_0.

Theorem 2. *Let* $q = \frac{\lambda_2}{\lambda_n}$ *and* $\xi = -\arccos(q) \cdot \frac{q}{\sqrt{1-q^2}}$. *Then the maximum degree of convergence* ζ_0 *is achieved at* $\tau = \tau^* = \frac{-\xi e^{\xi}}{\lambda_2}$, *and its value is* $\zeta_0 = \frac{\lambda_2}{e^{\xi}}$.

In the paper [24], for an undirected graph and the case of delay, the number ζ_0 was used to characterize the rate of reaching consensus in a second-order consensus protocol with some parameter γ. The authors reduced such a problem to an optimization problem subject to a constraint involving the parameter γ and the first nonzero eigenvalue of the matrix L. Nevertheless, for two systems with the degrees of convergence ζ_0^1 and ζ_0^2 such that $\zeta_0^1 > \zeta_0^2$, the convergence rate of the first system may be less than that of the second. This may be due to, e.g., a Jordan cell of dimension above 1 in the Laplacian matrix. Therefore, for a system with a real spectrum of L, the convergence rate problem shall be investigated considering the convergence rate of the functional series expressing the solution.

4. NUMERICAL EXAMPLE

In this section, we present an example of a multi-agent system with informational influences, delay, and a digraph corresponding to a Laplacian matrix of simple structure with a real spectrum. Figure 4 shows the dependency digraph for five agents. The weights of the agents' influences are indicated on the arcs. The Laplacian matrix of this digraph has the form

$$L = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1.8 & 0 & -1.8 & 0 \\ 0 & -2 & 2.5 & 0 & -0.5 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & -2 & 2 \end{pmatrix}.$$

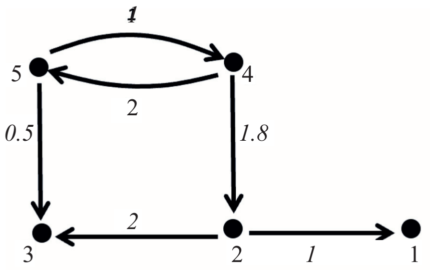

Fig. 4. The dependency digraph for five agents.

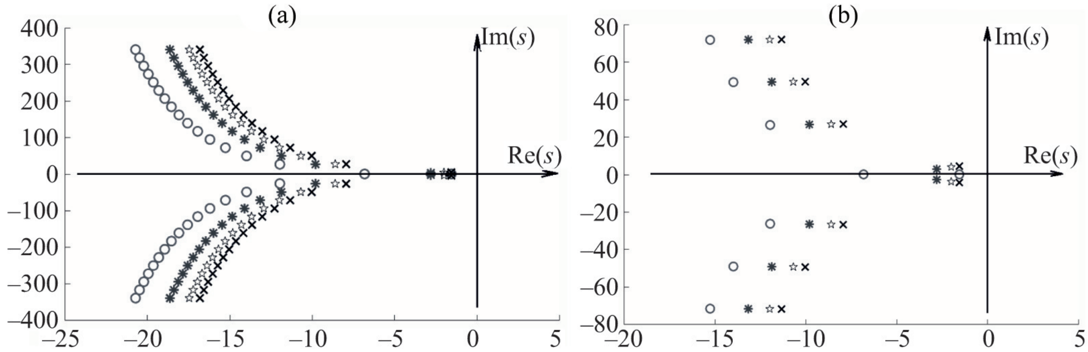

Fig. 5. The localization of roots s of the quasipolynomial computed using the Lambert W function (a) for fourteen branches and (b) for four branches. $\frac{1}{\tau^*}W_0(-\tau^*\lambda_2) = \frac{1}{\tau^*}\mathrm{Re}(W_0(-\tau^*\lambda_n)) = -1.545$.

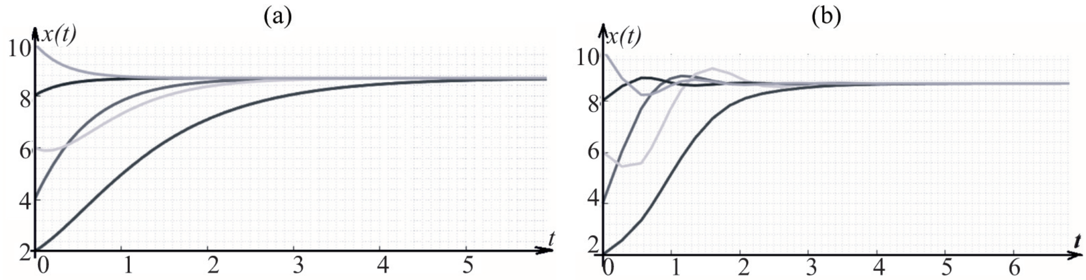

Fig. 6. Reaching consensus (a) without delay and (b) with delay $\tau^* \approx 0.2816$.

Its spectrum is $\sigma(L) = \{0;\ 1;\ 1.8;\ 2.5;\ 3\}$. According to Theorem 2, we find $\lambda_2 = 1$; $\lambda_5 = 3$; $q = 1/3$, and $-\xi = \arccos(q)\frac{q}{\sqrt{1-q^2}} = \arccos(1/3)\frac{1/3}{\sqrt{8/9}} \approx 0.4352$. Then, for the delay $\tau^* = -\frac{\xi e^{\xi}}{\lambda_2} \approx 0.2816$, the degree of convergence to consensus achieves maximum, and its value is $\zeta_0 = \frac{\lambda_2}{e^{\xi}} = 1.545$. The roots of the characteristic function (4), except the zero one, are marked in Fig. 5. Note that the boundary delay is $\tau_0 = \frac{\pi}{2\cdot 3} \approx 0.5236$.

Figure 6a demonstrates the graph of reaching consensus without delay with the initial value vector x_0. Clearly, with the delay $\tau^* \approx 0.2816$ and the initial function $\phi(\theta) = x_0$, the agents reach consensus faster than without delay (Fig. 6b). For both cases, the initial value vector is $x_0 = (2, 4, 6, 8, 10)^T$. For this initial condition, the consensus value does not depend on the delay and is given by the expression $px_0 = \frac{26}{3}$, where $p = \left(0, 0, 0, 0, \frac{2}{3}, \frac{1}{3}\right)$ is the row of the eigenprojector of rank 1 for the Laplacian matrix. Note that p is the left eigenvector of the zero eigenvalue for L.

In this example, for a nonsymmetric matrix of simple structure, an increase in the degree of convergence leads to an increase in the convergence rate. However, if a nonsymmetric matrix contains Jordan cells of dimension above 1, such a correlation may be violated. In other words, an increase in the degree of convergence does not always lead to an increase in the convergence rate.

According to the numerical example, an increase in the convergence rate causes oscillations in the transients. Unlike conventional control systems, where filters, damping devices, etc., can be used to reject oscillations, no methods for solving this problem in multi-agent systems are yet known. Indeed, in group motion, this can lead to collisions between agents. This problem is of interest and can be a subject for further study.

5. CONCLUSIONS

In this paper, the Lambert W function has been used to investigate the dependence of the degree of convergence on the delay in a first-order multi-agent system. For a linear system with delay, where the Laplacian matrix has a real spectrum, an expression for the delay τ^* ensuring the maximum degree of convergence has been derived. As has been established, the value τ^* depends only on the minimum nonzero and maximum eigenvalues of the Laplacian matrix with a real spectrum. Owing to the Lambert W function, a new (short) proof for the boundary delay has been obtained. The results of this paper can be applied to an arbitrary system with a nonsymmetric matrix whose spectrum consists of real numbers only.

APPENDIX

Proof of Theorem 1. Consider the characteristic function (4) of system (3), where $f_\lambda(s) = s + \lambda e^{-\tau s}$ and $\lambda = a + ib \in \sigma(L) \setminus \{0\}$. Recall that $a > 0$ for a Laplacian matrix. By (6) and Lemma 1, the boundary delay ensuring the negative real parts of the zeros of the function $f_\lambda(s)$ is given by the equality

$$W_0(-\tau\lambda) = iv,$$

which can be written as

$$ive^{iv} = -\lambda\tau. \tag{A.1}$$

From (A.1) it directly follows that

$$\begin{cases} v \sin v = a\tau, \\ v \cos v = -b\tau. \end{cases} \tag{A.2}$$

Then (A.2) implies $\tan v = -\frac{a}{b}$. Let $\varphi = |\arg(\lambda)|$, $\rho = |\lambda|$ and consider all possible cases of b.

a) $b = 0$. In this case, $-\tau\lambda$ belongs to the negative part of the real axis and $W_0(-\tau\lambda) = W_0(-\frac{\pi}{2})$ (see Fig. 2). Obviously, $-\tau\lambda = -\frac{\pi}{2}$ and $\tau = \frac{\pi}{2\lambda}$.

b) $b < 0$. Here, $-\tau\lambda$ belongs to the second quadrant and $0 < v < \pi/2$, as can be observed in Fig. 2. Therefore, $\tan v = \tan(\frac{\pi}{2} - \varphi) \Rightarrow v = \frac{\pi}{2} - \varphi$.

c) If $b > 0$, then $-\tau\lambda$ belongs to the third quadrant and, according to Fig. 2, $-\pi/2 < v < 0$. Therefore, $\tan v = -\tan(\frac{\pi}{2} - \varphi)$ and consequently, $v = \varphi - \frac{\pi}{2}$.

In case b), substituting $v = \frac{\pi}{2} - \varphi$ into (A.1) yields

$$i\left(\frac{\pi}{2} - \varphi\right)e^{i(\frac{\pi}{2}-\varphi)} = -\rho\tau e^{-i\varphi}.$$

Multiplying both sides by $-i$, we obtain

$$\left(\frac{\pi}{2} - \varphi\right)e^{i(\frac{\pi}{2}-\varphi)} = \rho\tau e^{i(\frac{\pi}{2}-\varphi)} \Rightarrow \frac{\pi}{2} - \varphi = \rho\tau. \tag{A.3}$$

In case c), with $v = \varphi - \frac{\pi}{2}$ substituted into (A.1),

$$i\left(\varphi - \frac{\pi}{2}\right) e^{i\left(\varphi - \frac{\pi}{2}\right)} = -\rho\tau e^{i\varphi},$$

$$\left(\frac{\pi}{2} - \varphi\right) e^{i\left(\varphi - \frac{\pi}{2}\right)} = -i\rho\tau e^{i\varphi} = \rho\tau e^{i\left(\varphi - \frac{\pi}{2}\right)} \Rightarrow \frac{\pi}{2} - \varphi = \rho\tau. \tag{A.4}$$

Thus, from (A.3) and (A.4) it follows that

$$\tau = \frac{1}{\rho}\left(\frac{\pi}{2} - \varphi\right).$$

In view of case a), the boundary delay for system (3) is $\tau_0 = \min\limits_{\lambda \in \sigma(L) \setminus \{0\}} \frac{1}{\rho}\left(\frac{\pi}{2} - \varphi\right)$. □

Proof of Proposition 1. 1) For $-\tau\lambda < -\frac{1}{e}$ (i.e., for $\tau > \frac{1}{e\lambda}$), the function $s = \frac{1}{\tau} W_0(-\tau\lambda)$ takes complex values. Since $W_0(-\frac{\pi}{2}) = i\frac{\pi}{2}$ and $\mathrm{Re}(W_0(x)) < 0$ for $x \in (-\frac{\pi}{2}, 0)$, we have $\mathrm{Re}(s) < 0$ for $-\frac{\pi}{2} < -\tau\lambda < -\frac{1}{e}$ (i.e., for $\frac{1}{e\lambda} < \tau < \frac{\pi}{2\lambda}$). On this interval, as τ increases, the value of $\frac{1}{\tau}$ decreases, like the magnitude of the real part of $W_0(-\tau\lambda)$. Thus, the real part of $s = \frac{1}{\tau} W_0(-\tau\lambda)$ takes a negative value and decreases in magnitude with increasing τ. In other words, it increases on the interval under consideration.

2) Due to

$$W'(x) = \frac{W(x)}{x(1 + W(x))},$$

we have

$$s' = \left(\frac{W_0(-\tau\lambda)}{\tau}\right)'_\tau = -\frac{W_0^2(-\tau\lambda)}{\tau^2(1 + W_0(-\tau\lambda))}.$$

On the other hand, on the interval $-\frac{1}{e} < -\tau\lambda < 0$ (i.e., for $0 < \tau < \frac{1}{e\lambda}$), the values of the function $W_0(-\tau\lambda)$ are negative and greater than -1 (see Fig. 1). Therefore, the function $s = \frac{1}{\tau} W_0(-\tau\lambda)$ decreases on the interval $0 < \tau < \frac{1}{e\lambda}$ with respect to τ. On this interval, $-e\lambda < s < -\lambda$. □

Proof of Proposition 2. If s_0 is the solution of the equation $s + \lambda e^{-\tau s} = 0$, then according to (5) and (6), τs_0 is defined by the multivalued function $W(-\tau\lambda)$. On the other hand, by Lemma 1, the maximum real part of s_0 can be defined by the function $\frac{1}{\tau} W_0(-\tau\lambda)$. Also, based on Proposition 1, the minimum real part of the function $\frac{1}{\tau} W_0(-\tau\lambda)$ is achieved at $-\tau\lambda = -\frac{1}{e}$, i.e., at $\tau = \frac{1}{e\lambda}$. □

Proof of Proposition 3. 1) Obviously, $\tau^*_{\lambda_n} < \ldots < \tau^*_{\lambda_2}$. Let τ^* be the delay value ensuring the maximum degree of convergence for system (3). We show that $\tau^* \in (\tau^*_{\lambda_n}, \tau^*_{\lambda_2})$. Indeed, if $\tau^* \leqslant \tau^*_{\lambda_n}$, then all values $W_0(-\tau^*\lambda_i)$, $i = 2, \ldots, n$, are real and belong to the interval $(-1, 0)$ of the real axis. In this case, the root corresponding to λ_2 will be closest to the imaginary axis. Then, as the value τ^* increases, the root corresponding to λ_2 will move away from the imaginary axis, thereby increasing the degree of convergence of the system. And if $\tau^* \geqslant \tau^*_{\lambda_2}$, then all values $W_0(-\tau^*\lambda_i)$, $i = 2, \ldots, n$, contain an imaginary part and belong to the arc AC (see Fig. 3a). In addition, the root corresponding to λ_n will have the largest real part. In this case, by decreasing τ^*, one can increase the degree of convergence of the system.

Thus, the maximum degree of convergence of system (3) will be achieved at $\tau = \tau^* \in (\tau^*_{\lambda_n}, \tau^*_{\lambda_2})$, for which

$$W_0(-\tau^*\lambda_2) = \mathrm{Re}(W_0(-\tau^*\lambda_n)). \tag{A.5}$$

2) If all nonzero eigenvalues λ of the Laplacian matrix are equal, then the degree of convergence is determined by λ and is equal to $e\lambda$. □

Proof of Theorem 2. According to Proposition 3, the eigenvalues λ_i, $i = 3, \ldots, n-1$, have no influence on τ^*. Therefore, τ^* can be found from equality (A.5).

Let us denote $W_0(-\tau^*\lambda_2) = u$ and $W_0(-\tau^*\lambda_n) = u + iv$.

Note that $-\tau^*\lambda_n$ belongs to the ray $(-\infty, -e^{-1}]$ (see Fig. 2). Then for $W_0(-\tau^*\lambda_n)$ we have $u = -v\mathrm{cot}v$.

Thus, from $W_0(-\tau^*\lambda_2) = \mathrm{Re}(W_0(-\tau^*\lambda_n))$ it follows that

$$\begin{cases} -\tau^*\lambda_2 = ue^u, \\ -\tau^*\lambda_n = e^u(u\cos v - v\sin v), \\ u = -v\mathrm{cot}v. \end{cases} \tag{A.6}$$

We transform the second equation of system (A.6):

$$-\tau^*\lambda_n = ue^u\left(\cos v - \frac{v}{u}\sin v\right).$$

Since $u = -v\mathrm{cot}v$,

$$\cos v - \frac{v}{u}\sin v = \cos v + \frac{\sin^2 v}{\cos v} = \frac{1}{\cos v}.$$

Then system (A.6) becomes

$$\begin{cases} -\tau^*\lambda_2 = ue^u, \\ -\tau^*\lambda_n = ue^u\dfrac{1}{\cos v}, \\ u = -v\mathrm{cot}v. \end{cases}$$

From the first and second equations we obtain

$$v = \arccos(q),$$

where $q = \frac{\lambda_2}{\lambda_n}$. Substituting this expression into the third equation yields

$$u = -\arccos(q)\cot(\arccos(q)).$$

Note that

$$\cot(\arccos(q)) = \frac{\cos(\arccos(q))}{\sin(\arccos(q))} = \frac{q}{\sqrt{1-q^2}}.$$

Hence, $u = -\frac{q}{\sqrt{1-q^2}} \cdot \arccos(q)$, we denote this value by ξ. Then τ^* can be obtained from the first equation of system (A.6):

$$\tau^* = -\frac{\xi e^\xi}{\lambda_2};$$

and the maximum degree of convergence is $\zeta_0 = -\frac{\xi}{\tau^*} = \frac{\lambda_2}{e^\xi}$. □

REFERENCES

1. Furtat, I., Fridman, E., and Fradkov, A., Disturbance Compensation with Finite Spectrum Assignment for Plants with Input Delay, *IEEE Transactions on Automatic Control*, 2017, vol. 63, no. 1, pp. 298–305.
2. Seuret, A., Dimarogonas, D.V., and Johansson, K.H., Consensus under Communication Delays, *47th IEEE Conference on Decision and Control*, Cancun, Mexico, 2008, pp. 4922–4927.
3. Amelina, N.O., Anan'evskii, M.S., Proskurnikov, A.V., et al., *Problemy setevogo upravleniya* (Problems of Network Control), Fradkov, A.L., Ed., Izhevsk: Institute of Computer Science, 2015.

4. Munz, U., Papachristodoulou, A., and Allgower, F., Delay Robustness in Consensus Problems, *Automatica*, 2010, vol. 46, no. 8, pp. 1252–1265.
5. Hou, W., Fu, M., Zhang, H., and Wu, Z., Consensus Conditions for General Second-Order Multi-Agent Systems with Communication Delay, *Automatica*, 2017, vol. 75, pp. 293–298.
6. Hara, S., Hayakawa, T., and Sugatat, H., Stability Analysis of Linear Systems with Generalized Frequency Variables and Its Applications to Formation Control, *Proc. 46th IEEE Conference on Decision and Control*, New Orleans, USA, 2007, pp. 1459–1466.
7. Agaev, R.P. and Khomutov, D.E., On the Boundary Value of the Time-Delay and the Asymptotic Behavior of a Continuous First-Order Consensus, *Autom. Remote Control*, 2024, vol. 85, no. 6, pp. 610–620.
8. Yang, W., Wang, X., and Shi, H., Fast Consensus Seeking in Multi-Agent Systems with Time Delay, *Syst. Control Lett.*, 2013, vol. 62, no. 3, pp. 269–276.
9. Yang, W., Bertozzi, A.L., and Wang, X., Stability of a Second Order Consensus Algorithm with Time Delay, *47th IEEE Conference on Decision and Control*, Cancun, Mexico, 2008, pp. 2926–2931.
10. Corless, R.M., Gonnet, G.H., Hare, D., et. al., On the Lambert W Function, *Advances in Computational Mathematics*, 1996, no. 5, pp. 329–359.
11. Asl, F.M. and Ulsoy, A.G., Analysis of a System of Linear Delay Differential Equations, *J. Dyn. Sys., Meas., Control*, 2003, vol. 125, no. 2, pp. 215–223.
12. Shinozaki, H. and Mori, T., Robust Stability Analysis of Linear Time-Delay Systems by Lambert W Function: Some Extreme Point Results, *Automatica*, 2006, vol. 42, no. 10, pp. 1791–1799.
13. Jarlebring, E. and Damm, T., The Lambert W Function and the Spectrum of Some Multidimensional Time-Delay Systems, *Automatica*, 2007, vol. 43, no. 12, pp. 2124–2128.
14. Fasi, M., Higham, N.J., and Iannazzo, B., An Algorithm for the Matrix Lambert W Function, *SIAM Journal on Matrix Analysis and Applications*, 2015, vol. 36, no. 2, pp. 669–685.
15. Yi, S., Nelson, P.W., and Ulsoy, A.G., Analysis and Control of Time Delayed Systems via the Lambert W Function, *IFAC Proceedings Volumes*, 2008, vol. 41, no. 2, pp. 13414–13419.
16. Michurin, R.A. and Nikolayev, V.T., Using Lambert W Functions for Construction Domain of Stability in Automatic Control System Including Time Delay, *Vest. Mosk. Gos. Akad. Delov. Administr. Ser. Ekon.*, 2013, vol. 24, no. 3, pp. 191–197.
17. Dubinov, A.E., Dubinova, I.D., and Saikov, S.K., *W-funktsiya Lamberta i ee primenenie v matematicheskoi fizike* (Lambert W Function and Its Application in Mathematical Physics), the Russian Federal Nuclear Center – All-Russian Scientific Research Institute of Experimental Physics (RFNC-VNIIEF), 2006.
18. Dubinov, A.E., Dubinova, I.D., and Saikov, S.K., Characteristic Roots and Stability Domains of One Dynamic Delay System, *Autom. Remote Control*, 2005, vol. 66, no. 8, pp. 1212–1213.
19. Corless, R.M., Hare, D.E., and Jeffrey, D.J., Unwinding the Branches of the Lambert W Function, *The Mathematical Scientist*, 1996, vol. 21, pp. 1–7.
20. Olfati-Saber, R.M. and Murray, R.M., Consensus Problems in Networks of Agents with Switching Topology and Time-Delays, *IEEE Trans. Autom. Control*, 2004, vol. 49, no. 9, pp. 1520–1533.
21. Khomutov, D.K., Latent Consensus Protocol with Weak Background Links and Time-Delay, *Large-Scale Systems Control*, 2025, no. 114, pp. 138–155.
22. Hara, T. and Sugie, J., Stability Region for Systems of Differential-Difference Equations, *Funkcialaj Ekvacioj*, 1996, vol. 39, no. 1, pp. 69–86.
23. Ramirez, A., Sipahi, R., Mondie, S., and Garrido, R., Fast Consensus in a Large-Scale Multi-Agent System with Directed Graphs Using Time-Delayed Measurements, *Philosophical Transactions of the Royal Society A*, 2019, vol. 377, no. 2153, art. no. 20180130.
24. Yang, W., Wang, X., and Shi, H., Fast Consensus Seeking in Multi-Agent Systems with Time Delay, *Systems & Control Letters*, 2013, vol. 62, no. 3, pp. 269–276.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2026, Vol. 87, No. 2, pp. 170–194.

NONLINEAR SYSTEMS

Analytical Optimal Solution to the Minimum Time and Energy Problem of Programmed Control for Spacecraft Spatial Motion in the Class of Generalized Helical Conic Motions

Yu. N. Chelnokov[*,a], **A. V. Molodenkov**[*,b], **and I. A. Pankratov**[*,**,c]

[*] *Institute of Precision Mechanics and Control, Russian Academy of Sciences, Saratov, Russia*
[**] *Saratov National Research State University named after N.G. Chernyshevsky, Saratov, Russia*
e-mail: [a] *chelnokovyun@gmail.com,* [b] *molalexei@yandex.ru,* [c] *pankratovia.mechanic@gmail.com*

Received May 8, 2025
Revised July 28, 2025
Accepted July 29, 2025

Abstract—The study examines the problem of optimal programmed control for the spatial motion (particularly, maneuvering) of a spacecraft, treated as a free rigid body of arbitrary dynamic configuration, with a combined cost functional incorporating both time and energy expended during the controlled motion of the spacecraft. In the class of generalized helical conic motions, an optimal analytical solution to the problem is obtained for arbitrary boundary conditions on the angular and linear positions as well as the angular and linear velocities of the spacecraft, which is then formulated into an algorithm. To describe the spatial motion, four-dimensional dual Euler parameters (Rodrigues–Hamilton parameters) are used, which are components of the Clifford dual quaternion (parabolic biquaternion) of finite displacement. The solution is derived using Chasles' theorem on the motion of a free rigid body and the Kotelnikov–Study transfer principle. Numerical examples demonstrating the effectiveness of the proposed solution are provided.

Keywords: spacecraft, optimal programmed control, spatial motion, Clifford dual quaternion (biquaternion), dual integral combined cost functional, analytical solution, algorithm

DOI: 10.7868/S1608303226020042

1. INTRODUCTION

It is well known that the general spatial displacement of a free rigid body is equivalent to its helical displacement (Chasles' theorem in mechanics: any displacement of a free rigid body can be accomplished by a single helical displacement along some axis, called the axis of finite helical displacement). Therefore, the motion of a free rigid body represents a continuous sequence of instantaneous helical motions. Here, the orientation of the axis of instantaneous helical motion coincides with the orientation of the instantaneous rotation axis of the rigid body, and the body's rotation around this axis and its translational motion along this axis form (using Clifford's duality) the so-called dual rotation angle: a complex composition of the body's rotational and translational displacements.

Previously, the literature considered the kinematic problem of constructing a time-optimal helical displacement of a free rigid body equivalent to its general spatial displacement [1, 2]. Here, the vectors of angular and linear velocities of the body served as programmed controls. To obtain an exact (analytical) solution to the problem, the dual quaternion (biquaternion) kinematic equation of helical motion of a free rigid body, first proposed by Yu.N. Chelnokov in 1980 [3–5], was used.

Also, in [6–8], dynamic problems of analytical construction of control for the spatial motion of a rigid body (using the dual quaternion models of free rigid body motion proposed in these works) were solved within the framework of the feedback control concept.

In [6], a method for the analytical construction of control for the spatial motion of a rigid body, equivalent to the composition of angular (rotational) and translational (orbital) motions, was developed in a nonlinear dynamic formulation using Clifford dual quaternions and dual matrices. The controls ensure asymptotic stability "in the large" of any chosen programmed spatial motion in the inertial coordinate system and the desired dynamics of the controlled spatial motion of the rigid body. To construct the control laws, dual quaternion and matrix models of the spatial motion of a rigid body, the concept of solving inverse problems of dynamics, the principle of feedback control, and an approach based on reducing the nonlinear differential equations of perturbed spatial motion of the rigid body constructed in the paper to reference linear stationary differential forms of a chosen structure through the use of proposed nonlinear feedbacks in the control laws were employed. Various dual matrix (screw) control laws for the spatial motion of a rigid body are proposed, for which the nonlinear non-stationary differential equations of perturbed spatial motion of the rigid body take the form of linear stationary dual matrix differential equations of the second order (with respect to the screw part of the dual quaternion of the body's position error), invariant with respect to any chosen programmed spatial motion of the rigid body. The constant coefficients (scalar dual or matrix dual) of these equations are the gain coefficients of the nonlinear feedbacks in the proposed dual control laws, ensuring the desired quality of control transients. The determination of the gain coefficients of the nonlinear feedbacks and the properties of such controlled motion of the rigid body are discussed.

In [7], another method for the analytical construction of control for the spatial motion of a rigid body (in particular, a spacecraft considered as a free rigid body) was developed in a nonlinear dynamic formulation using dual quaternions (Clifford parabolic biquaternions). The control ensures asymptotic stability "in the whole" of any chosen programmed motion in the inertial coordinate system and the desired dynamics of the controlled body motion. To construct the control laws, new dual quaternion differential equations of perturbed spatial motion of a rigid body are proposed, which, unlike [6], use non-normalized dual quaternions of finite displacements, dual quaternions of angular and linear velocities and accelerations of the body with non-zero dual scalar parts (i.e., four-dimensional, rather than three-dimensional, dual angular and linear velocities and accelerations). This allowed for effectively solving the synthesis problem of control ensuring asymptotic stability "in the whole" of any chosen programmed motion. To construct the control laws, the concept of solving inverse problems of dynamics, the principle of feedback control, and an approach based on reducing the constructed equations of perturbed body motion to linear stationary differential forms of a chosen structure, invariant with respect to any chosen programmed motion, through the appropriate choice of dual nonlinear feedbacks in the proposed dual quaternion control laws were also used. This approach is similar to that used in [6], but is applied to more general dual quaternion differential equations of perturbed spatial motion of a rigid body. Analytical solutions of the dual quaternion differential equations describing the dynamics of the spatial motion control process using the proposed dual quaternion control laws are constructed. The properties and regularities of such control are analyzed.

This article addresses the problem of constructing interrelated time-optimal and energy-optimal programmed control for the spatial motion of a spacecraft, considered as a free rigid body of arbitrary dynamic configuration, performing an optimal helical conic motion in an inertial coordinate system, equivalent to the composition of the translational (orbital) motion of the spacecraft's center of mass and its angular motion (rotation) around the center of mass. The spacecraft is subject to the principal vector and principal moment of external forces, including the control force vector and the control torque vector. Four-dimensional dual Euler parameters (Rodrigues–Hamilton parameters) and three-dimensional dual velocities of the spacecraft, which are components of the Clifford dual quaternion of finite displacement and the dual quaternion of angular and linear velocities, respectively, i.e., dual quaternions with complex (dual) components, are used to describe the general spatial motion of the spacecraft. The Clifford duality s used to describe this motion possesses the property that its square is zero: $s^2 = 0$.

The problem formulation (for arbitrary boundary conditions on the angular and linear positions of the spacecraft in inertial space and on its absolute angular and linear velocities, with the dual control vector being unconstrained) is based on using a dynamic model of the general spatial motion of the spacecraft, equivalent to its helical motion. To solve the problem, the dual quaternion (biquaternion) kinematic equation of motion of a free rigid body [3–5] (where the dual variable is the dual quaternion (biquaternion) of the body's finite displacement in inertial space) was used, as well as analytical expressions obtained by the authors in the article as explicit functions of time for the four-dimensional dual quaternion of finite displacement and for the three-dimensional dual quaternions of angular and linear velocities and accelerations of the spacecraft, describing the generalized helical conic motion of the spacecraft in inertial space. To obtain these analytical expressions, a new analytical solution to the dual quaternion kinematic equation of spatial motion of a free rigid body for its generalized helical conic motion is constructed in the article. Using these analytical expressions (after extracting their principal and moment parts), the vectors of programmed control force and programmed control torque applied to the spacecraft are formed in accordance with the concept of solving inverse problems of dynamics.

Using the Kotelnikov–Study transfer principle, which allows extending quaternion formulas describing angular motion control to dual quaternion (biquaternion) formulas describing the control of general spatial motion of a rigid body, an algorithm for programmed control of spacecraft spatial motion, optimal in terms of time and minimum energy costs in the class of generalized helical conic motions, is obtained. This algorithm is constructed based on a generalization of the previously obtained approximate analytical solution of the quaternion problem of time-optimal and energy-optimal spacecraft (as a rigid body) reorientation under arbitrary boundary conditions on the angular position and angular velocity of the spacecraft. This solution was obtained within the framework of the classical Poinsot concept, interpreting arbitrary angular motion of a rigid body in terms of precession cones, or otherwise, generalized conic angular motion [9–11].

To find the optimal parameters of the generalized helical conic motion and the corresponding optimal angular and linear velocities and accelerations of the spacecraft, the L.S. Pontryagin maximum principle was used. The found analytical solution of the problem can be considered as a solution to the problem of time-optimal and energy-optimal programmed control of spacecraft spatial motion under arbitrary boundary conditions in the class of generalized dual (helical) conic motions. The results of modeling optimal spatial maneuvering of a spacecraft in an inertial coordinate system using the proposed dual quaternion models and dual control laws are presented, demonstrating the effectiveness of the proposed method for controlling spacecraft motion.

The proposed optimal analytical solution to the problem of programmed control of spacecraft spatial motion has not only theoretical but also applied significance, as it allows using in the spacecraft motion control system the obtained analytical expressions (as explicit functions of time) for the dual quaternion of the spacecraft's finite displacement and for the dual quaternion of its absolute angular and linear velocities, describing the optimal programmed helical conic motion of the spacecraft in the inertial coordinate system. It also allows using optimal analytical time-varying laws for the programmed control force and programmed control torque applied to the spacecraft. This is important, in particular, for spacecraft spatial maneuvering at exceptionally high rates of displacement, when the time for calculating the optimal spatial trajectory of the spacecraft maneuver and for calculating the programmed control laws enabling this trajectory is extremely limited.

The article continues the research initiated in [12].

2. INITIAL EQUATIONS OF SPATIAL MOTION OF A FREE RIGID BODY

Consider a free rigid body, e.g., a spacecraft, capable of performing arbitrary spatial motion relative to the main (inertial) coordinate system, equivalent to a spatial helical motion, and also equivalent to the composition of the translational motion of the body together with an arbitrarily chosen point of the body and the rotation of the body around that point. The body is subject to an arbitrary principal vector and principal moment of external forces. They include the control force vector and the control torque vector. We will consider the controlled spatial motion of the body relative to the inertial coordinate system $O_1\xi_1\xi_2\xi_3$ (ξ) (its origin O_1 is the Earth's center of mass, the axis $O_1\xi_3$ is directed along the Earth's rotation axis, and the axes $O_1\xi_1$ and $O_1\xi_2$ lie in the equatorial plane and do not participate in the Earth's diurnal rotation). We rigidly attach to the rigid body a coordinate system $CX_1X_2X_3$ (X) with origin at the body's center of mass C.

Let us denote: $\mathbf{r}$ and $\mathbf{v}$—the radius vector and velocity vector of the center of mass of the rigid body (spacecraft) in the inertial coordinate system; $\boldsymbol{\lambda}$—the quaternion of orientation of the body in this coordinate system, its components are the Rodrigues–Hamilton (Euler) parameters λ_j, $j=\overline{0,3}$, identical in bases ξ and X (it is assumed that in the initial position (before body rotation), the orientations of coordinate systems ξ and X coincide); $\boldsymbol{\omega}$ and $\boldsymbol{\varepsilon}$—the vectors of absolute angular velocity and absolute angular acceleration of the body; $\mathbf{F}_c$ and $\mathbf{M}_c$—the vectors of control force and control torque applied to the body; $\mathbf{F}=\mathbf{F}(t,\mathbf{r},\mathbf{v})$ and $\mathbf{M}=\mathbf{M}(t,\boldsymbol{\lambda},\boldsymbol{\omega})$—the principal vector of other external forces acting on the rigid body (gravity, drag, and other forces of interaction of the body with the external environment), and the principal moment of these forces, calculated relative to the body's center of mass, assumed to be known functions of time t and the variables $\mathbf{r}$, $\mathbf{v}$ and $\boldsymbol{\lambda}$, $\boldsymbol{\omega}$.

The initial differential equations of motion of a rigid body, written in the body-fixed coordinate system X, have the form:

$$m\left[\dot{\mathbf{v}}_x+\mathbf{K}(\boldsymbol{\omega}_x)\mathbf{v}_x\right]=\mathbf{F}_x(t,\mathbf{r}_x,\mathbf{v}_x)+\mathbf{F}_{cx},\quad \dot{\mathbf{r}}_x+\mathbf{K}(\boldsymbol{\omega}_x)\mathbf{r}_x=\mathbf{v}_x; \tag{2.1}$$

$$\dot{\boldsymbol{\omega}}_x=\boldsymbol{\varepsilon}_x=\mathbf{J}^{-1}\left[\mathbf{M}_x(t,\boldsymbol{\lambda},\boldsymbol{\omega}_x)-\mathbf{K}(\boldsymbol{\omega}_x)\mathbf{J}\boldsymbol{\omega}_x+\mathbf{M}_{cx}\right]; \tag{2.2}$$

$$2\dot{\boldsymbol{\lambda}}=\boldsymbol{\lambda}\circ\boldsymbol{\omega}_x; \tag{2.3}$$

$$\mathbf{r}_x=x_1\mathbf{i}+x_2\mathbf{j}+x_3\mathbf{k},\quad \dot{\mathbf{r}}_x=\dot{x}_1\mathbf{i}+\dot{x}_2\mathbf{j}+\dot{x}_3\mathbf{k},$$

$$\mathbf{v}_x=v_1\mathbf{i}+v_2\mathbf{j}+v_3\mathbf{k},\quad \dot{\mathbf{v}}_x=\dot{v}_1\mathbf{i}+\dot{v}_2\mathbf{j}+\dot{v}_3\mathbf{k};$$

$$\boldsymbol{\lambda}=\lambda_0+\boldsymbol{\lambda}_v=\lambda_0+\lambda_1\mathbf{i}+\lambda_2\mathbf{j}+\lambda_3\mathbf{k},\quad \dot{\boldsymbol{\lambda}}=\dot{\lambda}_0+\dot{\boldsymbol{\lambda}}_v=\dot{\lambda}_0+\dot{\lambda}_1\mathbf{i}+\dot{\lambda}_2\mathbf{j}+\dot{\lambda}_3\mathbf{k};$$

$$\boldsymbol{\omega}_x=\omega_1\mathbf{i}+\omega_2\mathbf{j}+\omega_3\mathbf{k},\quad \dot{\boldsymbol{\omega}}_x=\dot{\omega}_1\mathbf{i}+\dot{\omega}_2\mathbf{j}+\dot{\omega}_3\mathbf{k},\quad \boldsymbol{\varepsilon}_x=\varepsilon_1\mathbf{i}+\varepsilon_2\mathbf{j}+\varepsilon_3\mathbf{k};$$

$$\mathbf{J}=\begin{pmatrix} J_{11} & -J_{12} & -J_{13}\\ -J_{21} & J_{22} & -J_{23}\\ -J_{31} & -J_{32} & J_{33}\end{pmatrix},\quad \mathbf{K}(\boldsymbol{\omega}_x)=\begin{pmatrix} 0 & -\omega_3 & \omega_2\\ \omega_3 & 0 & -\omega_1\\ -\omega_2 & \omega_1 & 0\end{pmatrix}.$$

Here in equations (2.1), (2.2) $\mathbf{r}_x$, $\mathbf{v}_x$, $\boldsymbol{\omega}_x$, $\boldsymbol{\varepsilon}_x$, $\mathbf{F}_x$, $\mathbf{M}_x$, $\mathbf{F}_{cx}$, $\mathbf{M}_{cx}$—column vectors of size 3×1 or, further, quaternions with zero scalar parts, composed of the projections x_k, v_k, ω_k, ε_k, F_k, M_k, F_{ck}, M_{ck}, $k=\overline{1,3}$, of vectors $\mathbf{r}$, $\mathbf{v}$, $\boldsymbol{\omega}$, $\boldsymbol{\varepsilon}$, $\mathbf{F}$, $\mathbf{M}$, $\mathbf{F}_c$, $\mathbf{M}_c$ onto the axes of the body-fixed coordinate system X; m—body mass, $\mathbf{J}$—constant inertia matrix of the rigid body; $\mathbf{K}(\boldsymbol{\omega}_x)$—skew-symmetric matrix of angular velocities of the body, associated with the vector $\boldsymbol{\omega}$; $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$—orts of the hypercomplex space (Hamilton's vector imaginary units); $\mathbf{a}_y$—mapping of vector $\mathbf{a}$ onto basis Y

$(Y = \xi,\ X)$, defined as a quaternion $\mathbf{a}_y = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$, whose components are the projections a_k of vector $\mathbf{a}$ onto basis Y; the superscript dot denotes the time derivative t (when calculating the derivative of a quaternion, the orts $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ are considered constant), the sign "$\circ$" denotes quaternion multiplication. The first matrix equation (2.1) and the matrix equation (2.2) are dynamic, while the second matrix equation (2.1) and the quaternion equation (2.3)—are kinematic equations of spatial motion of a rigid body, representing the composition of translational (trajectory) and angular (rotational) motions. These equations constitute a system of nonlinear, non-stationary differential equations of the thirteenth order with respect to the variables x_k, v_k and λ_j, ω_k. In the case where the coordinate system $OX_1X_2X_3$ (X) with origin at another arbitrarily chosen point O of the body is rigidly attached to the rigid body, the principal vector $\mathbf{F}$ and principal moment $\mathbf{M}$ of external forces include the translational inertial force and its moment relative to the point O of the rigid body.

3. PROBLEM STATEMENT FOR SPACECRAFT CONTROL USING DUAL QUATERNIONS AND THE CONCEPT OF SOLVING THE INVERSE PROBLEM OF DYNAMICS

In accordance with the concept of solving inverse problems of dynamics, the laws for forming the column vector $\mathbf{F}_{cx}$ of the control force $\mathbf{F}_c$ and the column vector $\mathbf{M}_{cx}$ of the control torque $\mathbf{M}_c$, composed of the projections F_{ck} and M_{ck}, $k = \overline{1,3}$, of the vectors $\mathbf{F}_c$ and $\mathbf{M}_c$ onto the axes of the coordinate system X associated with the spacecraft, are obtained based on the equations of motion (2.1), (2.2) and have the form:

$$\mathbf{F}_{cx} = m\,[\dot{\mathbf{v}}_x + \mathbf{K}(\boldsymbol{\omega}_x)\mathbf{v}_x] - \mathbf{F}_x(t, \mathbf{r}_x, \mathbf{v}_x), \tag{3.1}$$

$$\mathbf{M}_{cx} = \mathbf{J}\boldsymbol{\varepsilon}_x + \mathbf{K}(\boldsymbol{\omega}_x)\mathbf{J}\boldsymbol{\omega}_x - \mathbf{M}_x(t, \boldsymbol{\lambda}, \boldsymbol{\omega}_x). \tag{3.2}$$

The vector-matrix control law (3.1) includes the column vector $\dot{\mathbf{v}}_x = \mathbf{w}_x$, composed of the time derivatives of the projections v_k of the vector $\mathbf{v}$ of the velocity of the body's center of mass C in the inertial coordinate system ξ (the vector of absolute velocity of the spacecraft's center of mass) onto the axes of the rotating coordinate system X, and the vector-matrix control law (3.2) includes the column vector $\boldsymbol{\varepsilon}_x$, composed of the projections ε_k of the vector $\boldsymbol{\varepsilon}$ of angular acceleration of the spacecraft in the inertial coordinate system (the vector of absolute angular acceleration of the body) onto the axes of the coordinate system X.

These column vectors $\dot{\mathbf{v}}_x = \mathbf{w}_x$ and $\boldsymbol{\varepsilon}_x$ will be considered as new controls. They can be constructed, for example, using the L.S. Pontryagin maximum principle based on the vector-matrix and quaternion equations:

$$\dot{\mathbf{v}}_x = \mathbf{w}_x, \quad \dot{\mathbf{r}}_x + \mathbf{K}(\boldsymbol{\omega}_x)\mathbf{r}_x = \mathbf{v}_x, \tag{3.3}$$

$$\dot{\boldsymbol{\omega}}_x = \boldsymbol{\varepsilon}_x, \quad 2\dot{\boldsymbol{\lambda}} = \boldsymbol{\lambda} \circ \boldsymbol{\omega}_x, \tag{3.4}$$

derived from the equations of motion of a rigid body (2.1)–(2.3).

We emphasize that the new controls $\mathbf{w}_x$ and $\boldsymbol{\varepsilon}_x$ have a clear mechanical meaning. $\mathbf{w}_x$—is the column vector whose components are the projections onto the axes of the body-fixed coordinate system X of the component $\dot{\mathbf{v}} = (d\mathbf{v}/dt)_{\mathrm{loc}}$ of the vector $\mathbf{W}$ of absolute linear acceleration of the spacecraft's center of mass C (projections of the local derivative $(d\mathbf{v}/dt)_{\mathrm{loc}}$ of the vector $\mathbf{v}$ of absolute velocity of point C, computed in this coordinate system). The total vector $\mathbf{W}$ of absolute linear acceleration of the spacecraft's center of mass is equal to the sum of the vectors $\dot{\mathbf{v}}$ and $\boldsymbol{\omega} \times \mathbf{v}$, defined by their projections in the coordinate system X: $\mathbf{W}_x = \dot{\mathbf{v}}_x + \boldsymbol{\omega}_x \times \mathbf{v}_x = \mathbf{w}_x + \boldsymbol{\omega}_x \times \mathbf{v}_x$ (here "$\times$" denotes the cross product).

$\boldsymbol{\varepsilon}$—is the vector of absolute angular acceleration of the rigid body (one of the main characteristics of absolute angular (rotational) motion of a rigid body), $\boldsymbol{\varepsilon}_x$—is the column vector whose components are the projections of the vector $\boldsymbol{\varepsilon}$ onto the axes of the coordinate system X.

Introducing the component $\dot{\mathbf{v}} = (d\mathbf{v}/dt)_{\text{loc}}$ of the vector of absolute linear acceleration of the spacecraft's center of mass as a control allows proposing, for describing the spatial motion of the spacecraft, a biquaternion differential model of spacecraft spatial motion, convenient for solving the problem of controlling spacecraft spatial motion using dual quaternions and the maximum principle.

After finding the new controls $\mathbf{w}_x$ and $\boldsymbol{\varepsilon}_x$ based on the differential equations (3.3) and (3.4), the phase variables $\mathbf{r}_x$, $\mathbf{v}_x$, $\boldsymbol{\lambda}$, $\boldsymbol{\omega}_x$, appearing in these equations, corresponding to these controls will also be found, and then the control force $\mathbf{F}_{cx}$ and control torque $\mathbf{M}_{cx}$ will be found in accordance with the algebraic relations (3.1), (3.2).

Thus, the problem of constructing the control force $\mathbf{F}_{cx}$ and control torque $\mathbf{M}_{cx}$ using the concept of solving inverse problems of dynamics is reduced by the authors to constructing the required component $\mathbf{w}_x$ of absolute linear acceleration and the required absolute angular acceleration $\boldsymbol{\varepsilon}_x$, which act as new controls in the differential equations (3.3), (3.4).

Obviously, equations (3.3), (3.4) are valid for any moving object (i.e., a free rigid body). Therefore, the controls $\mathbf{w}_x$ and $\boldsymbol{\varepsilon}_x$ for all such bodies are constructed in the same way. The specifics of the object (its mass-inertial and other characteristics, the acting external perturbing forces and their moments) are taken into account when constructing the control force $\mathbf{F}_{cx}$ and control torque $\mathbf{M}_{cx}$ based on the final relations (3.1), (3.2); their right-hand sides include the principal vector $\mathbf{F}_x(t,\, \mathbf{r}_x,\, \mathbf{v}_x)$ of other external forces acting on the rigid body (gravity, drag, and other forces of interaction of the body with the external environment) and the principal moment $\mathbf{M} = \mathbf{M}(t,\, \boldsymbol{\lambda},\, \boldsymbol{\omega})$ of these forces, calculated relative to the body's center of mass. The principal vector and principal moment are assumed to be known functions of time t and the variables $\mathbf{r}$, $\mathbf{v}$ and $\boldsymbol{\lambda}$, $\boldsymbol{\omega}$.

Let us set the problem of finding the new controls $\mathbf{w}_x$ and $\boldsymbol{\varepsilon}_x$ in terms of dual quaternions. Let us introduce the kinematic screw $\mathbf{U}$ of the spacecraft, whose mapping $\mathbf{U}_x$ onto the body-fixed (spacecraft) basis X is defined by the dual quaternion:

$$\mathbf{U}_x = U_1\mathbf{i} + U_2\mathbf{j} + U_3\mathbf{k} = \boldsymbol{\omega}_x + s\mathbf{v}_x, \quad U_k = \omega_k + sv_k, \tag{3.5}$$

where $\boldsymbol{\omega}_x$ and $\mathbf{v}_x$ are quaternions: $\boldsymbol{\omega}_x = \omega_1\mathbf{i} + \omega_2\mathbf{j} + \omega_3\mathbf{k}$, $\mathbf{v}_x = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$; s—is the Clifford symbol (duality) possessing the property $s^2 = 0$; $U_k = \omega_k + sv_k$, $k = \overline{1,3}$—are the dual orthogonal projections of the kinematic screw $\mathbf{U}$ onto basis X.

Then the vector-matrix (3.3) and quaternion (3.4) differential equations can be replaced by the following two dual quaternion (biquaternion) differential equations [6–8]:

$$\dot{\mathbf{U}}_x = \boldsymbol{\varepsilon}_x + s\mathbf{w}_x = \mathbf{H}_x; \tag{3.6}$$

$$2\dot{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda} \circ \mathbf{U}_x; \tag{3.7}$$

$$\begin{gathered} \mathbf{U}_x = U_1\mathbf{i} + U_2\mathbf{j} + U_3\mathbf{k}, \qquad U_k = \omega_k + sv_k, \qquad \boldsymbol{\varepsilon}_x = \varepsilon_1\mathbf{i} + \varepsilon_2\mathbf{j} + \varepsilon_3\mathbf{k}, \\ \mathbf{w}_x = w_1\mathbf{i} + w_2\mathbf{j} + w_3\mathbf{k}, \quad \mathbf{H}_x = H_1\mathbf{i} + H_2\mathbf{j} + H_3\mathbf{k}, \qquad H_k = \varepsilon_k + sw_k; \end{gathered} \tag{3.8}$$

$$\boldsymbol{\Lambda} = \Lambda_0 + \boldsymbol{\Lambda}_v = \Lambda_0 + \Lambda_1\mathbf{i} + \Lambda_2\mathbf{j} + \Lambda_3\mathbf{k} = \boldsymbol{\lambda} + s\boldsymbol{\lambda}^0, \quad \Lambda_j = \lambda_j + s\lambda_j^0. \tag{3.9}$$

Here, the sought variables are the dual quaternion $\mathbf{U}_x = \boldsymbol{\omega}_x + s\mathbf{v}_x$ (mapping of the kinematic screw $\mathbf{U}$ of the spacecraft onto the spacecraft-fixed basis X) and the dual quaternion $\boldsymbol{\Lambda} = \boldsymbol{\lambda} + s\boldsymbol{\lambda}^0$ of the spacecraft's displacement in inertial space. Its principal part (the quaternion $\boldsymbol{\lambda}$) describes the orientation of the spacecraft in the inertial coordinate system, while the moment part (the quaternion $\boldsymbol{\lambda}^0$) describes the location of the spacecraft in this coordinate system. The Cartesian coordinates x_k, $k = \overline{1,3}$, of the spacecraft's center of mass in this coordinate system can be determined by knowing the components of the quaternions $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^0$. The dual control in this problem is the dual quaternion $\mathbf{H}_x = \boldsymbol{\varepsilon}_x + s\mathbf{w}_x$, i.e., the dual composition of the required absolute angular acceleration $\boldsymbol{\varepsilon}_x$ and the required component $\mathbf{w}_x$ of absolute linear acceleration of the spacecraft (the dual composition of the new controls $\mathbf{w}_x$ and $\boldsymbol{\varepsilon}_x$).

Introducing the component of the vector of absolute linear acceleration of the spacecraft's center of mass (the local derivative of the vector of absolute velocity of the center of mass) as a control allows proposing, for describing the spatial motion of the spacecraft, the differential dual quaternion equation (3.6) for the kinematic screw, which, together with the kinematic dual quaternion equation (3.7) of spatial motion of the spacecraft, forms a complete mathematical model of its spatial motion, convenient for solving the problem of controlling spacecraft spatial motion using dual quaternions and the maximum principle.

The coordinates ξ_k of the spacecraft's center of mass in the inertial coordinate system ξ (i.e., the projections of its radius vector $\mathbf{r}$) and the projections x_k of this vector onto the axes of the spacecraft-fixed coordinate system X are related to the components of the quaternions $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^0$ by the relations [3, 5]:

$$\mathbf{r}_\xi = \xi_1\mathbf{i} + \xi_2\mathbf{j} + \xi_3\mathbf{k} = 2\boldsymbol{\lambda}^0 \circ \widetilde{\boldsymbol{\lambda}}, \quad \mathbf{r}_x = x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k} = \widetilde{\boldsymbol{\lambda}} \circ \boldsymbol{\lambda}^0, \tag{3.10}$$

where the upper tilde denotes quaternion conjugation.

Let us formulate the following problem for finding the new controls $\mathbf{w}_x$ and $\boldsymbol{\varepsilon}_x$ using dual quaternions: it is necessary to construct, in the class of optimal generalized helical conic motions, a dual quaternion programmed control $\mathbf{H}_x = \boldsymbol{\varepsilon}_x + s\mathbf{w}_x$ (the dual composition of the component of absolute linear acceleration of the spacecraft's center of mass and the vector of absolute angular acceleration of the rigid body), ensuring a programmed transfer of the spacecraft, whose motion is described by equations (3.6), (3.7), from its arbitrary given initial state

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}(0) = \boldsymbol{\lambda}(0) + s\boldsymbol{\lambda}^0(0), \;\; \mathbf{U}_x = \mathbf{U}_x(0) = \boldsymbol{\omega}_x(0) + s\mathbf{v}_x(0) \tag{3.11}$$

to its given final state

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}(t_1) = \boldsymbol{\Lambda}_{\mathrm{k}} = \boldsymbol{\lambda}_{\mathrm{k}} + s\boldsymbol{\lambda}^0_{\mathrm{k}}, \;\; \mathbf{U}_x = \mathbf{U}_x(t_1) = \mathbf{U}_x^{\mathrm{k}} = \boldsymbol{\omega}_x^{\mathrm{k}} + s\mathbf{v}_x^{\mathrm{k}} \tag{3.12}$$

in time t_1, which is to be determined.

The boundary conditions $\boldsymbol{\lambda}^0(0)$ and $\boldsymbol{\lambda}^0_{\mathrm{k}}$ are found through the given initial and final values of the quaternions $\boldsymbol{\lambda}$ and $\mathbf{r}_\xi$ using the formula:

$$\boldsymbol{\lambda}^0 = (1/2)\,\mathbf{r}_\xi \circ \boldsymbol{\lambda} = (1/2)\,(\xi_1\mathbf{i} + \xi_2\mathbf{j} + \xi_3\mathbf{k}) \circ (\lambda_0 + \lambda_1\mathbf{i} + \lambda_2\mathbf{j} + \lambda_3\mathbf{k})\,, \tag{3.13}$$

where ξ_k, $k = \overline{1,3}$ are the coordinates of the object in the inertial coordinate system. After solving the stated problem, it is necessary to extract the principal part $\boldsymbol{\varepsilon}_x$ and the moment part $\mathbf{w}_x = \dot{\mathbf{v}}_x$ from the constructed control $\mathbf{H}_x = \boldsymbol{\varepsilon}_x + s\mathbf{w}_x$. After this extraction, the laws for forming the control force $\mathbf{F}_{cx}$ and control torque $\mathbf{M}_{cx}$ are obtained in accordance with the concept of solving inverse problems of dynamics using formulas (3.1) and (3.2).

4. ANALYTICAL SOLUTION OF THE OPTIMAL PROGRAMMED CONTROL PROBLEM

Using the Kotelnikov–Study transfer principle, which allows extending quaternion formulas describing angular motion control to dual quaternion formulas describing the control of general spatial motion of a rigid body, we obtain an algorithm for programmed control of spacecraft spatial motion, optimal in the class of generalized helical conic motions, for the problem of Section 3 of the article. This algorithm is constructed using Clifford dual quaternions and the Kotelnikov–Study transfer principle by generalizing the analytical solution of the quaternion problem of time-optimal and energy-optimal spacecraft reorientation in the class of generalized conic angular motions of a rigid body under arbitrary boundary conditions on the angular position and angular velocity of the spacecraft [9–11], which, in turn, was obtained based on the exact solution of a modified problem of optimal rotation of a rigid body introduced by Ya.G. Sapunkov [13]. We present this biquaternion solution.

In the case where a free rigid body performs a spatial generalized helical conic motion, for which the kinematic screw $\mathbf{U}_x$ of the body has the form

$$\mathbf{U}_x(t) = U_1\mathbf{i} + U_2\mathbf{j} + U_3\mathbf{k} = \boldsymbol{\omega}_x + s\mathbf{v}_x = \Big(\dot{F}(t)\sin G(t)\Big)\ \mathbf{i} + \Big(\dot{F}(t)\cos G(t)\Big)\ \mathbf{j} + \dot{G}(t)\ \mathbf{k}, \tag{4.1}$$

where $F(t)$, $G(t)$, $\dot{F}(t)$, $\dot{G}(t)$—are arbitrary differentiable dual functions of time:

$$\begin{aligned} F(t) &= f(t) + sf^0(t), \quad \dot{F}(t) = \dot{f}(t) + s\dot{f}^0(t),\\ G(t) &= g(t) + sg^0(t), \quad \dot{G}(t) = \dot{g}(t) + s\dot{g}^0(t), \end{aligned}$$

the dual quaternion kinematic equation (3.7) of the general spatial motion of a rigid body has an analytical solution constructed by the authors:

$$\boldsymbol{\Lambda}(t) = \boldsymbol{\Lambda}(0) \circ \mathbf{exp}\left\{-\left(G(0)/2\right)\mathbf{k}\right\} \circ \mathbf{exp}\left\{\left(\left(F(t) - F(0)\right)/2\right)\mathbf{j}\right\} \circ \mathbf{exp}\left\{\left(G(t)/2\right)\mathbf{k}\right\} \tag{4.2}$$

—a dual analogue of the known quaternion solution [9–11] of the kinematic equation of rotational motion of a body, where "$\mathbf{exp}\{\cdot\}$" denotes the dual quaternion exponential [5].

The helical motion of a free rigid body (spacecraft) described by relations (4.1), (4.2) can be generalized by adding an arbitrary dual rotation in the inertial coordinate system by a dual constant angle around some axis. Such a dual rotation is given by an arbitrary constant dual quaternion $\mathbf{K} = \boldsymbol{\kappa} + s\boldsymbol{\kappa}^0$, $\|\mathbf{K}\| = 1$ (multiplications on the left and right by the dual quaternions $\widetilde{\mathbf{K}}$ and $\mathbf{K}$ respectively, where $\widetilde{\mathbf{K}} \circ \mathbf{K} = \mathbf{K} \circ \widetilde{\mathbf{K}} = 1$, are added to the right-hand sides of formulas (4.1), (4.2)):

$$\mathbf{U}_x(t) = \widetilde{\mathbf{K}} \circ \Big[\Big(\dot{F}(t)\sin G(t)\Big)\ \mathbf{i} + \Big(\dot{F}(t)\cos G(t)\Big)\ \mathbf{j} + \dot{G}(t)\ \mathbf{k}\Big] \circ \mathbf{K}, \tag{4.3}$$

$$\boldsymbol{\Lambda}(t) = \boldsymbol{\Lambda}(0) \circ \widetilde{\mathbf{K}} \circ \mathbf{exp}\left\{-\left(G(0)/2\right)\mathbf{k}\right\} \circ \mathbf{exp}\left\{\left(\left(F(t) - F(0)\right)/2\right)\mathbf{j}\right\} \circ \mathbf{exp}\left\{\left(G(t)/2\right)\mathbf{k}\right\} \circ \mathbf{K}. \tag{4.4}$$

Formulas (4.3), (4.4) contain (for $s = 0$) all previously found exact quaternion analytical solutions of the classical problem of optimal rotation of a dynamically symmetric rigid body in the case where the angular velocity vector maintains a constant direction or describes a circular cone in space throughout its motion [14, 15], and their dual analogues (for $s \neq 0$), constructed here in this article. The proposed structure of the kinematic screw (4.1) or (4.3), in the quaternion case—the angular velocity vector of the spacecraft, correlates well with the Poinsot concept: any arbitrary angular motion of a rigid body around a fixed point can be considered as some generalized conic motion of the rigid body.

We restrict ourselves to the case where the constant dual quaternion $\mathbf{K}$ is represented as a product of two constant dual quaternions:

$$\mathbf{K} = \mathbf{K}_2 \circ \mathbf{K}_1, \quad \mathbf{K}_1 = \mathbf{exp}\left\{\left(A_1/2\right)\mathbf{i}\right\}, \quad \mathbf{K}_2 = \mathbf{exp}\left\{\left(A_2/2\right)\mathbf{j}\right\}, \tag{4.5}$$

where $A_1 = \alpha_1 + s\alpha_1^0$, $A_2 = \alpha_2 + s\alpha_2^0$—are some dual scalar constants. Note that the dual quaternions $\mathbf{K}_1$ and $\mathbf{K}_2$ define dual rotations of the screw $\mathbf{U}_x$ around the axes $\mathbf{i}$ and $\mathbf{j}$. The dual rotation around the axis $\mathbf{k}$ is already included in (4.4), as an additive constant is included in the dual function $G(t)$.

The conjugate dual quaternion $\widetilde{\mathbf{K}}$ will be represented as:

$$\widetilde{\mathbf{K}} = \widetilde{\mathbf{K}}_1 \circ \widetilde{\mathbf{K}}_2, \quad \widetilde{\mathbf{K}}_1 = \mathbf{exp}\left\{-\left(A_1/2\right)\mathbf{i}\right\}, \quad \widetilde{\mathbf{K}}_2 = \mathbf{exp}\left\{-\left(A_2/2\right)\mathbf{j}\right\}. \tag{4.6}$$

Then the kinematic screw $\mathbf{U}_x$ of the free rigid body (spacecraft) and the analytical solution of the dual quaternion kinematic equation (3.7) for the considered spatial helical motion of the spacecraft take the form:

$$\begin{aligned}\mathbf{U}_x(t) = &\, \mathbf{exp}\left\{-\left(A_1/2\right)\mathbf{i}\right\} \circ \mathbf{exp}\left\{-\left(A_2/2\right)\mathbf{j}\right\} \\ &\circ \left[\left(\dot{F}(t)\sin G(t)\right)\mathbf{i} + \left(\dot{F}(t)\cos G(t)\right)\mathbf{j} + \dot{G}(t)\mathbf{k}\right] \\ &\circ \mathbf{exp}\left\{\left(A_2/2\right)\mathbf{j}\right\} \circ \mathbf{exp}\left\{\left(A_1/2\right)\mathbf{i}\right\},\end{aligned} \tag{4.7}$$

$$\begin{aligned}\mathbf{\Lambda}(t) = &\, \mathbf{\Lambda}(0) \circ \mathbf{exp}\left\{-\left(A_1/2\right)\mathbf{i}\right\} \circ \mathbf{exp}\left\{-\left(A_2/2\right)\ \mathbf{j}\right\} \\ &\circ \mathbf{exp}\left\{-\left(G(0)/2\right)\mathbf{k}\right\} \circ \mathbf{exp}\left\{\left(\left(F(t)-F(0)\right)/2\right)\mathbf{j}\right\} \\ &\circ \mathbf{exp}\left\{\left(G(t)/2\right)\mathbf{k}\right\} \circ \mathbf{exp}\left\{\left(A_2/2\right)\mathbf{j}\right\} \circ \mathbf{exp}\left\{\left(A_1/2\right)\mathbf{i}\right\}.\end{aligned} \tag{4.8}$$

The Cartesian coordinates ξ_k of the body's center of mass in the inertial coordinate system and the projections x_k of the radius vector of the center of mass onto the axes of the body-fixed coordinate system X are found through the components of the quaternions $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^0$ using formulas (3.10), and the projections of the velocity vector of the spacecraft's center of mass in the inertial coordinate system onto its own coordinate axes are found using the quaternion formula

$$\begin{aligned}\mathbf{v}_\xi = \dot{\mathbf{r}}_\xi = \dot{\xi}_1\mathbf{i} + \dot{\xi}_2\mathbf{j} + \dot{\xi}_3\mathbf{k} &= 2\left(\dot{\boldsymbol{\lambda}}^0 \circ \widetilde{\boldsymbol{\lambda}} + \boldsymbol{\lambda}^0 \circ \dot{\widetilde{\boldsymbol{\lambda}}}\right) \\ &= 2\dot{\boldsymbol{\lambda}}^0 \circ \widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 \circ \boldsymbol{\omega}_x \circ \widetilde{\boldsymbol{\lambda}} = \left(2\dot{\boldsymbol{\lambda}}^0 - \boldsymbol{\lambda}^0 \circ \boldsymbol{\omega}_x\right) \circ \widetilde{\boldsymbol{\lambda}}.\end{aligned} \tag{4.9}$$

Let us set the problem of finding the optimal parameters of the helical conic motion of the spacecraft (generalizing the problem of finding the optimal parameters of the generalized angular conic motion of the spacecraft [9–11]). We will consider the second derivatives of the dual functions $F(t)$ and $G(t)$ (of the dual parameters of the helical conic motion) as control parameters. Let $\dot{F} = F_1$, $\dot{G} = G_1$—be the dual phase coordinates, and U_{c1}, U_{c2}—be the dual control parameters. Then the controlled system will be described by the following system of scalar dual differential equations:

$$\dot{F} = F_1, \quad \dot{G} = G_1, \quad \dot{F}_1 = U_{c1}, \quad \dot{G}_1 = U_{c2}, \tag{4.10}$$

where $F = f + sf^0$, $G = g + sg^0$, $F_1 = f_1 + sf_1^0$, $G_1 = g_1 + sg_1^0$—are the dual phase coordinates, $U_{c1} = u_{c1} + su_{c1}^0$, $U_{c2} = u_{c2} + su_{c2}^0$—are the dual controls.

System (4.10) of four scalar dual differential equations is equivalent to a system of eight scalar real differential equations:

$$\dot{f} = f_1, \quad \dot{g} = g_1, \quad \dot{f}_1 = u_{c1}, \quad \dot{g}_1 = u_{c2}; \tag{4.11}$$

$$\dot{f}^0 = f_1^0, \quad \dot{g}^0 = g_1^0, \quad \dot{f}_1^0 = u_{c1}^0, \quad \dot{g}_1^0 = u_{c2}^0. \tag{4.12}$$

The subsystems of real differential equations (4.11) and (4.12) are independent, since the original system (4.10) of dual differential equations is linear.

Then, for the dual controlled system (4.10) or the real controlled system (4.11), (4.12), the following problem can be formulated: it is required to find the dual optimal controls $U_{c1}(t) = u_{c1}(t) + su_{c1}^0(t)$, $U_{c2}(t) = u_{c2}(t) + su_{c2}^0(t)$ or the real optimal controls $u_{c1}(t)$, $u_{c2}(t)$ and $u_{c1}^0(t)$, $u_{c2}^0(t)$, which transfer the controlled system (4.10) ((4.11), (4.12)) from the initial state

$$\begin{aligned}&\left.\begin{aligned}F = F(0) = f(0) + sf^0(0), \qquad G = G(0) = g(0) + sg^0(0), \\ F_1 = F_1(0) = f_1(0) + sf_1^0(0), \quad G_1 = G_1(0) = g_1(0) + sg_1^0(0)\end{aligned}\right\} \\ &\sim \left\{\begin{aligned}f = f(0), \quad f^0 = f^0(0), \quad g = g(0), \quad g^0 = g^0(0), \\ f_1 = f_1(0), \quad f_1^0 = f_1^0(0), \quad g_1 = g_1(0), \quad g_1^0 = g_1^0(0)\end{aligned}\right.\end{aligned} \tag{4.13}$$

to the final state

$$\left.\begin{aligned} F = F(t_1) = f(t_1) + sf^0(t_1), \quad & G = G(t_1) = g(t_1) + sg^0(t_1), \\ F_1 = F_1(t_1) = f_1(t_1) + sf_1^0(t_1), \quad & G_1 = G_1(t_1) = g_1(t_1) + sg_1^0(t_1) \end{aligned}\right\}$$
$$\sim \left\{\begin{aligned} & f = f(t_1), \quad f^0 = f^0(t_1), \quad g = g(t_1), \quad g^0 = g^0(t_1), \\ & f_1 = f_1(t_1), \quad f_1^0 = f_1^0(t_1), \quad g_1 = g_1(t_1), \quad g_1^0 = g_1^0(t_1) \end{aligned}\right. \tag{4.14}$$

and minimize the dual combined functional:

$$J = \int\limits_0^{t_1} \left(1 + U_{c1}^2 + U_{c2}^2\right) dt \tag{4.15}$$

and satisfy relations (4.7), (4.8) at the initial and final instants of time (the final time instant t_1 is not fixed and is subject to determination):

$$\begin{gathered} [(F_1(0)\sin G(0))\ \mathbf{i} + (F_1(0)\cos G(0))\ \mathbf{j} + G_1(0)\ \mathbf{k}] \\ = \mathbf{exp}\left\{(A_2/2)\,\mathbf{j}\right\} \circ \mathbf{exp}\left\{(A_1/2)\,\mathbf{i}\right\} \circ \mathbf{U}_x(0) \\ \circ\, \mathbf{exp}\left\{-\,(A_1/2)\,\mathbf{i}\right\} \circ \mathbf{exp}\left\{-\,(A_2/2)\,\mathbf{j}\right\}, \end{gathered} \tag{4.16}$$

$$\begin{gathered} [(F_1(t_1)\sin G(t_1))\,\mathbf{i} + (F_1(t_1)\cos G(t_1))\,\mathbf{j} + G_1(t_1)\mathbf{k}] \\ = \mathbf{exp}\left\{(A_2/2)\,\mathbf{j}\right\} \circ \mathbf{exp}\left\{(A_1/2)\,\mathbf{i}\right\} \circ \mathbf{U}_x(t_1) \\ \circ\, \mathbf{exp}\left\{-\,(A_1/2)\,\mathbf{i}\right\} \circ \mathbf{exp}\left\{-\,(A_2/2)\,\mathbf{j}\right\}; \end{gathered} \tag{4.17}$$

$$\begin{gathered} \mathbf{\Lambda}(0) \circ \mathbf{exp}\left\{-\,(A_1/2)\,\mathbf{i}\right\} \circ \mathbf{exp}\left\{-\,(A_2/2)\,\mathbf{j}\right\} \\ \circ\, [\mathbf{exp}\left\{-\,(G(0)/2)\,\mathbf{k}\right\} \circ \mathbf{exp}\left\{((F(t_1) - F(0))/2)\,\mathbf{j}\right\} \circ \mathbf{exp}\left\{(G(t_1)/2)\,\mathbf{k}\right\}] \\ \circ\, \mathbf{exp}\left\{(A_2/2)\,\mathbf{j}\right\} \circ \mathbf{exp}\left\{(A_1/2)\,\mathbf{i}\right\} = \mathbf{\Lambda}(t_1) \end{gathered} \tag{4.18}$$

or

$$\begin{gathered} \mathbf{exp}\left\{-\,(G(0)/2)\,\mathbf{k}\right\} \circ \mathbf{exp}\left\{((F(t_1) - F(0))/2)\,\mathbf{j}\right\} \circ \mathbf{exp}\left\{(G(t_1)/2)\,\mathbf{k}\right\} \\ = \mathbf{exp}\left\{(A_2/2)\,\mathbf{j}\right\} \circ \mathbf{exp}\left\{(A_1/2)\,\mathbf{i}\right\} \circ \widetilde{\mathbf{\Lambda}}(0) \circ \mathbf{\Lambda}(t_1) \\ \circ\, \mathbf{exp}\left\{-\,(A_1/2)\,\mathbf{i}\right\} \circ \mathbf{exp}\left\{-\,(A_2/2)\,\mathbf{j}\right\}. \end{gathered} \tag{4.19}$$

The laws for forming the control force $\mathbf{F}_{cx}$ and control torque $\mathbf{M}_{cx}$ are obtained in accordance with the concept of solving inverse problems of dynamics using the formulas:

$$\begin{aligned} \mathbf{F}_{cx} = F_{c1}\mathbf{i} + F_{c2}\mathbf{j} + F_{c3}\mathbf{k} &= m\left[\dot{\mathbf{v}}_x + \mathbf{K}(\boldsymbol{\omega}_x)\mathbf{v}_x\right] - \mathbf{F}_x\left(t, \mathbf{r}_x, \mathbf{v}_x\right) \\ &= m\left[\mathbf{w}_x + \mathbf{K}(\boldsymbol{\omega}_x)\mathbf{v}_x\right] - \mathbf{F}_x\left(t, \mathbf{r}_x, \mathbf{v}_x\right), \end{aligned} \tag{4.20}$$

$$\mathbf{M}_{cx} = M_{c1}\mathbf{i} + M_{c2}\mathbf{j} + M_{c3}\mathbf{k} = \mathbf{J}\boldsymbol{\varepsilon}_x + \mathbf{K}(\boldsymbol{\omega}_x)\mathbf{J}\boldsymbol{\omega}_x - \mathbf{M}_x\left(t, \boldsymbol{\lambda}, \boldsymbol{\omega}_x\right), \tag{4.21}$$

where the quaternions $\mathbf{v}_x = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$, $\boldsymbol{\omega}_x = \omega_1\mathbf{i} + \omega_2\mathbf{j} + \omega_3\mathbf{k}$, $\dot{\mathbf{v}}_x = \mathbf{w}_x = \dot{v}_1\mathbf{i} + \dot{v}_2\mathbf{j} + \dot{v}_3\mathbf{k}$, $\dot{\boldsymbol{\omega}}_x = \dot{\omega}_1\mathbf{i} + \dot{\omega}_2\mathbf{j} + \dot{\omega}_3\mathbf{k}$, $\boldsymbol{\varepsilon}_x = \dot{\boldsymbol{\omega}}_x = \varepsilon_1\mathbf{i} + \varepsilon_2\mathbf{j} + \varepsilon_3\mathbf{k}$ are obtained by extracting the principal and moment parts from the dual relation (4.7):

$$\begin{aligned} \omega_1 &= f_1 \sin g \cos\alpha_2 - g_1 \sin\alpha_2, \\ \omega_2 &= f_1(\sin g \sin\alpha_1 \sin\alpha_2 + \cos g \cos\alpha_1) + g_1 \sin\alpha_1 \cos\alpha_2, \\ \omega_3 &= f_1(\sin g \cos\alpha_1 \sin\alpha_2 - \cos g \sin\alpha_1) + g_1 \cos\alpha_1 \cos\alpha_2; \end{aligned} \tag{4.22}$$

$$
\begin{aligned}
&\dot\omega_1 = (u_{c1}\sin g + f_1 g_1\cos g)\cos\alpha_2 - u_{c2}\sin\alpha_2,\\
&\dot\omega_2 = u_{c1}(\sin g\sin\alpha_1\sin\alpha_2 + \cos g\cos\alpha_1)\\
&\qquad + f_1 g_1(\cos g\sin\alpha_1\sin\alpha_2 - \sin g\cos\alpha_1) + u_{c2}\sin\alpha_1\cos\alpha_2,\\
&\dot\omega_3 = u_{c1}(\sin g\cos\alpha_1\ \ \sin\alpha_2 - \cos g\sin\alpha_1)\\
&\qquad + f_1 g_1(\cos g\cos\alpha_1\sin\alpha_2 + \sin g\sin\alpha_1) + u_{c2}\cos\alpha_1\cos\alpha_2;
\end{aligned}
\tag{4.23}
$$

$$
\begin{aligned}
\mathbf{v}_x = e^{-\frac{\alpha_1}{2}\mathbf{i}} \circ \Big[&\boldsymbol\beta \times \left(\alpha_1^0\mathbf{i} + \alpha_2^0\mathbf{j}\right) + e^{-\frac{\alpha_2}{2}\mathbf{j}} \circ ((f_1^0\sin g + f_1 g^0\cos g)\mathbf{i}\\
&+ (f_1^0\cos g - f_1 g^0\sin g)\mathbf{j} + g_1^0\mathbf{k}) \circ e^{\frac{\alpha_2}{2}\mathbf{j}}\Big] \circ e^{\frac{\alpha_1}{2}\mathbf{i}},\\
\boldsymbol\beta = e^{-\frac{\alpha_2}{2}\mathbf{j}} \circ (&f_1\sin g\,\mathbf{i} + f_1\cos g\,\mathbf{j} + g_1\mathbf{k}) \circ e^{\frac{\alpha_2}{2}\mathbf{j}};
\end{aligned}
\tag{4.24}
$$

$$
\begin{aligned}
\dot{\mathbf{v}}_x = e^{-\frac{\alpha_1}{2}\mathbf{i}} \circ \Big[&\dot{\boldsymbol\beta} \times (\alpha_1^0\mathbf{i} + \alpha_2^0\mathbf{j}) + e^{-\frac{\alpha_2}{2}\mathbf{j}} \circ \Big(((u_{c1}^0 - f_1 g^0 g_1)\sin g\\
&+ (f_1^0 g_1 + u_{c1}g^0 + f_1 g_1^0)\cos g)\mathbf{i} + ((u_{c1}^0 - f_1 g^0 g_1)\cos g\\
&- (f_1^0 g_1 + u_{c1}g^0 + f_1 g_1^0)\sin g)\mathbf{j} + u_{c2}\mathbf{k}\Big) \circ e^{\frac{\alpha_2}{2}\mathbf{j}}\Big] \circ e^{\frac{\alpha_1}{2}\mathbf{i}},\\
\dot{\boldsymbol\beta} = e^{-\frac{\alpha_2}{2}\mathbf{j}} \circ \big(&(u_{c1}\sin g + f_1 g_1\cos g)\mathbf{i} + (u_{c1}\cos g - f_1 g_1\sin g)\mathbf{j} + u_{c2}\mathbf{k}\big) \circ e^{\frac{\alpha_2}{2}\mathbf{j}}.
\end{aligned}
\tag{4.25}
$$

The problem (4.3)–(4.25) is solved by the authors using the maximum principle. The dual Hamilton–Pontryagin function for the controlled system (4.10) or (4.11) and (4.12) and the combined functional (4.15) has the form:

$$
H = -\left(1 + U_{c1}^2 + U_{c2}^2\right) + \Psi_1 F_1 + \Psi_2 G_1 + \Psi_3 U_{c1} + \Psi_4 U_{c2} = 0, \quad \forall t \in [0, t_1], \tag{4.26}
$$

where $\Psi_1 = \psi_1 + s\psi_1^0$, $\Psi_2 = \psi_2 + s\psi_2^0$, $\Psi_3 = \psi_3 + s\psi_3^0$, $\Psi_4 = \psi_4 + s\psi_4^0$—dual adjoint variables satisfying the dual system of equations:

$$
\dot\Psi_1 = 0, \quad \dot\Psi_2 = 0, \quad \dot\Psi_3 = -\Psi_1, \quad \dot\Psi_4 = -\Psi_2, \tag{4.27}
$$

which is equivalent to two systems of real equations:

$$
\dot\psi_1 = 0, \quad \dot\psi_2 = 0, \quad \dot\psi_3 = -\psi_1, \quad \dot\psi_4 = -\psi_2; \tag{4.28}
$$

$$
\dot\psi_1^0 = 0, \quad \dot\psi_2^0 = 0, \quad \dot\psi_3^0 = -\psi_1^0, \quad \dot\psi_4^0 = -\psi_2^0. \tag{4.29}
$$

The general solutions of equations (4.27) or (4.28), (4.29), containing dual arbitrary integration constants $C_i = c_i + sc_i^0$, $i = \overline{0,4}$, or real arbitrary integration constants $c_1, \ldots, c_4$ and $c_1^0, \ldots, c_4^0$, have the form:

$$
\Psi_1 = C_1, \quad \Psi_2 = C_2, \quad \Psi_3 = -C_1 t + C_3, \quad \Psi_4 = -C_2 t + C_4; \tag{4.30}
$$

$$
\psi_1 = c_1, \quad \psi_2 = c_2, \quad \psi_3 = -c_1 t + c_3, \quad \psi_4 = -c_2 t + c_4, \tag{4.31}
$$

$$
\psi_1^0 = c_1^0, \quad \psi_2^0 = c_2^0, \quad \psi_3^0 = -c_1^0 t + c_3^0, \quad \psi_4^0 = -c_2^0 t + c_4^0. \tag{4.32}
$$

The sought optimal dual controls

$$
U_{c1} = (-C_1 t + C_3)/2, \quad U_{c2} = (-C_2 t + C_4)/2 \tag{4.33}
$$

maximize the Hamilton–Pontryagin function (4.26).

Consider the dual controls (4.33), which contain the dual quantities

$$U_{c1} = u_{c1} + su_{c1}^0, \quad U_{c2} = u_{c2} + su_{c2}^0,$$
$$\Psi_3 = \psi_3 + s\psi_3^0, \quad \Psi_4 = \psi_4 + s\psi_4^0, \quad C_i = c_i + sc_i^0, \; i = \overline{0,4}.$$

From these controls, we extract the principal and moment parts, i.e., the scalar real controls u_{c1}, u_{c2} and u_{c1}^0, u_{c2}^0 :

$$U_{c1} = u_{c1} + su_{c1}^0 = \left(\psi_3 + s\psi_3^0\right)/2 = \left[(-c_1t + c_3) + s\left(-c_1^0t + c_3^0\right)\right]/2,$$
$$U_{c2} = u_{c2} + su_{c2}^0 = \left(\psi_4 + s\psi_4^0\right)/2 = \left[(-c_2t + c_4) + s\left(-c_2^0t + c_4^0\right)\right]/2.$$

Substituting (4.33) into the system of dual differential equations (4.10), we find the general solution for the phase coordinates F, G, F_1, G_1, containing eight dual arbitrary constants $C_1 = c_1 + sc_1^0, \ldots, C_8 = c_8 + sc_8^0$:

$$\begin{aligned} F(t) &= -C_1t^3/12 + C_3t^2/4 + C_5t + C_6, \\ G(t) &= -C_2t^3/12 + C_4t^2/4 + C_7t + C_8, \\ F_1(t) &= -C_1t^2/4 + C_3t/2 + C_5, \\ G_1(t) &= -C_2t^2/4 + C_4t/2 + C_7. \end{aligned} \tag{4.34}$$

The unknown dual constants $C_1, \ldots, C_8$ are subject to determination. Since the constant C_6 enters the function F as an additive constant, from formula (4.8), containing the expression $\mathbf{exp}\left\{((F(t) - F(0))/2)\,\mathbf{j}\right\}$, it can be seen that this dual constant has no effect. For this reason, the constant C_6 can be set to zero. Thus, to determine the nine unknown dual constants of the problem $C_1, \ldots, C_5$, C_7, C_8, A_1, A_2 and the transient time t_1, we use nine dual equations from the system (4.16), (4.17), (4.19) and the scalar part of condition (4.26) at the time instant t_1 (note that in the dual quaternion equation (4.19) only three equations in dual scalar form are independent due to the normalization of the dual quaternion $\mathbf{\Lambda}$). Substituting formulas (4.33), (4.34) into (4.7), (4.8), we obtain analytical dual formulas for finding the sought optimal laws of variation of the kinematic screw $\mathbf{U}_x$ of the free rigid body and the dual quaternion $\mathbf{\Lambda}$, describing the optimal trajectories of the angular (rotational) and translational (orbital) motions of the free rigid body. These formulas will describe the optimal, in the sense of the combined functional, displacement of the body in the class of generalized helical conic motions. The laws for forming the control force $\mathbf{F}_{cx}$ and control torque $\mathbf{M}_{cx}$ are found in accordance with the concept of solving inverse problems of dynamics using formulas (4.20), (4.21).

The problem of optimal, in the sense of a combined functional incorporating time and energy, displacement of a free rigid body (spacecraft) in the class of generalized helical conic motions is thus completely solved.

The optimal algorithm for the spatial motion of a spacecraft of arbitrary dynamic configuration under arbitrary boundary conditions has the form:

1) from the given dual quantities $\mathbf{\Lambda}(0)$, $\mathbf{\Lambda}(t_1)$, $\mathbf{U}_x(0)$, $\mathbf{U}_x(t_1)$ (formulas (3.11), (3.12) for the boundary conditions of the problem) by solving the system of algebraic dual equations (4.5), (4.6), (4.16), (4.17), (4.19), supplemented by condition (4.26) for the scalar part of the Hamilton–Pontryagin function, the nine dual undetermined constants $C_1, \ldots, C_5$, C_7, C_8, A_1, A_2 and the time t_1 are calculated; then the dual functions F, G, F_1, G_1 are found;

2) using (4.5), the constant dual quaternion $\mathbf{K}$ is calculated;

3) using formula (4.7), the kinematic screw of the spacecraft $\mathbf{U}_x(t)$ is determined;

4) using formula (4.8), the dual quaternion of the spatial helical motion of the spacecraft $\mathbf{\Lambda}(t)$ is determined;

5) using formulas (4.20), (4.21), the laws for forming the control force $\mathbf{F}_{cx}$ and control torque $\mathbf{M}_{cx}$ of the spacecraft are constructed;

6) from expression (4.15), the value of the optimization criterion for the problem of optimal spatial motion (maneuvering) of the spacecraft is found.

It should be noted that in the cases of quaternion problems of optimal reorientation of a spherically symmetric spacecraft in the classes of planar Euler rotations and regular angular conic motions, the known optimal solutions of the classical problem and the analytical solutions of the problem of controlling angular motion of the spacecraft [9–11], optimal in the class of generalized angular conic motions, completely coincide. Based on the Kotelnikov–Study transfer principle, the same can be stated regarding the dual spatial analogues of these controlled motions of the spacecraft.

5. NUMERICAL EXAMPLES

This section considers examples of calculations of optimal, in the class of generalized helical conic motions, spatial motion (maneuvering) of a spacecraft according to the proposed algorithm, demonstrating the effectiveness of the obtained analytical solution of the problem for various dynamic configurations of the spacecraft.

Let us transition from dimensional variables of the problem to dimensionless ones using the formulas (the quaternion $\boldsymbol{\lambda}$ is dimensionless) [16]:

$$J^{\text{scale}} = \left(\left(J_1^2 + J_2^2 + J_3^2\right)/3\right)^{1/2}, \quad J_k^{\text{dimless}} = J_k/J^{\text{scale}};$$

$$\boldsymbol{\omega}_x^{\text{dimless}} = \left(J^{\text{scale}}\right)^{1/2} a^{1/4}\, \boldsymbol{\omega}_x, \quad t^{\text{dimless}} = \left(J^{\text{scale}}\right)^{-1/2} a^{-1/4}\, t,$$

$$\mathbf{M}^{\text{dimless}} = a^{1/2}\mathbf{M},$$

and also by the formulas:

$$\boldsymbol{\lambda}^{0\,\text{dimless}} = \boldsymbol{\lambda}^0/L, \;\; \boldsymbol{\varepsilon}_x^{\text{dimless}} = J^{\text{scale}} a^{1/2} \boldsymbol{\varepsilon}_x, \;\; \mathbf{w}_x^{\text{dimless}} = L\left(J^{\text{scale}}\right)^{-1} a^{-1/2}\, \mathbf{w}_x.$$

Here $a = 1(\mathrm{N} \times \mathrm{m})^{-2} = 1\ \mathrm{kg}^{-2} \times \mathrm{m}^{-4} \times \mathrm{s}^4$, L—is the scale factor for distance.

In this case, the phase equations of the problem and the main formulas of the proposed algorithm will not change.

We assume that the principal vector of other external forces $\mathbf{F}_x\,(t,\, \mathbf{r}_x,\, \mathbf{v}_x)$, appearing in formula (4.20) for calculating the control force, is caused solely by the central force of gravitational attraction of the spacecraft to the Earth, equal to

$$\mathbf{F}_x\,(t,\, \mathbf{r}_x,\, \mathbf{v}_x) = \mathbf{F}_x\,(\mathbf{r}_x) = -G\frac{M_0 m}{r^3}\mathbf{r}_x,$$

where G is the gravitational constant, and M_0—is the mass of the attracting body (the Earth), $r = |\mathbf{r}_x|$.

The gravitational moment in this central force field is determined by the following relation [17]:

$$\mathbf{M}_{\text{grav}} = 3\mu\frac{\boldsymbol{\gamma} \times \mathbf{J}\boldsymbol{\gamma}}{r^3}, \;\; \boldsymbol{\gamma} = \frac{\mathbf{r}}{r},$$

where $\mu = GM_0$ is the gravitational parameter of the attracting body (Earth).

Then the expressions for the control force and torque (4.20), (4.21) in dimensionless form will be written as follows:

$$\begin{gathered} \mathbf{F}_x = \mathbf{w}_x + \mathbf{K}(\boldsymbol{\omega}_x)\mathbf{v}_x + N_G\mathbf{r}_x/\left|\mathbf{r}_x\right|^3, \\ \mathbf{M}_{cx} = \mathbf{J}\boldsymbol{\varepsilon}_x + \mathbf{K}(\boldsymbol{\omega}_x)\mathbf{J}\boldsymbol{\omega}_x - 3N_G\mathbf{r}_x \times \mathbf{J}\mathbf{r}_x/\left|\mathbf{r}_x\right|^5. \end{gathered} \tag{5.1}$$

Here $N_G = GM_0T^2/L^3$—is the dimensionless parameter of the problem, $T = \left(J^{\text{scale}}\right)^{1/2} \times a^{1/4}$ – the scale factor for time.

Taking into account that $\|\boldsymbol{\lambda}\| = 1$, formulas (5.1) for finding the control force and control torque in dimensionless form take the form

$$\mathbf{F}_{cx} = \mathbf{w}_x + \mathbf{K}(\boldsymbol{\omega}_x)\mathbf{v}_x + N_G\frac{\widetilde{\boldsymbol{\lambda}} \circ \boldsymbol{\lambda}^0}{\|\boldsymbol{\lambda}^0\|^{3/2}},$$
$$\left\|\boldsymbol{\lambda}^0\right\| = \left(\lambda_0^0\right)^2 + \left(\lambda_1^0\right)^2 + \left(\lambda_2^0\right)^2 + \left(\lambda_3^0\right)^2. \tag{5.2}$$

$$\mathbf{M}_{cx} = \mathbf{J}\boldsymbol{\varepsilon}_x + \mathbf{K}\left(\boldsymbol{\omega}_x\right)\mathbf{J}\boldsymbol{\omega}_x - 3N_G\frac{\left(\widetilde{\boldsymbol{\lambda}} \circ \boldsymbol{\lambda}^0\right) \times \mathbf{J}\left(\widetilde{\boldsymbol{\lambda}} \circ \boldsymbol{\lambda}^0\right)}{\|\boldsymbol{\lambda}^0\|^{5/2}}.$$

Note that expressions (5.1) and (5.2) for the control force do not contain the inertia tensor $\mathbf{J}$, i.e., the law of variation of the control force does not depend on the type of mass distribution of the spacecraft.

The boundary conditions for the angular position in space and angular velocity of the spacecraft have the form ([16, p. 137],) (reference [16] considered the problem of optimal reorientation of a spacecraft as a rigid body of arbitrary dynamic configuration under arbitrary boundary conditions on the angular position and angular velocity of the spacecraft's reorientation in the absence of translational displacement and gravitational torque):

$$\begin{gathered}\boldsymbol{\lambda}(0) = (0.7951;\ 0.2981;\ -0.3975;\ 0.3478),\\ \boldsymbol{\omega}(0) = (0.2739;\ -0.2388;\ -0.3),\end{gathered} \tag{5.3}$$

$$\begin{gathered}\boldsymbol{\lambda}_k = (0.8443;\ 0.3984;\ -0.3260;\ 0.1485),\\ \boldsymbol{\omega}_k = (0.0;\ 0.0;\ -0.59).\end{gathered} \tag{5.4}$$

SC 1. Spherically symmetric rigid body: $J_1 = J_2 = J_3 = J^*$ kg $\times$ m^2. Here $J^{\text{scale}} = J^*$ kg $\times$ m^2, then the dimensionless moments of inertia are $J_1 = J_2 = J_3 = J^*/J^{\text{scale}} = 1.0$.

SC 2. International Space Station (ISS) [18]:

$J_1 = 4\,853\,000$ kg $\times$ m^2, $J_2 = 23\,601\,000$ kg $\times$ m^2, $J_3 = 26\,278\,000$ kg $\times$ m^2 (in dimensional form) or $J_1 = 0.2358$, $J_2 = 1.1466$, $J_3 = 1.2766$ (in dimensionless form).

SC 3. Space Shuttle spacecraft (almost axisymmetric rigid body): $J_1 = 3\,400\,648$ kg $\times$ m^2, $J_2 \approx J_3 = 21\,041\,672$ kg $\times$ m^2 or $J_1 = 0.1967$, $J_2 \approx J_3 = 1.2168$.

In order to compare the calculation results with those obtained in [16] in the absence of translational displacement of the spacecraft, the initial and final values of the spacecraft's linear velocity vector must be set to the zero vector. The values of the real constants α_1, α_2, $c_1, \ldots, c_8$, appearing in the analytical solution of the problem obtained in [16] for angular motion with boundary conditions (5.3), (5.4) are as follows ($\alpha_1^0 = \alpha_2^0 = 0$, $c_1^0 = c_2^0 = \ldots = c_8^0 = 0$):

$$\begin{array}{lllll}\alpha_1 = -0.0421; & \alpha_2 = -0.2226; & c_1 = 3.4020; & c_2 = -2.0123; & c_3 = 2.2293;\\ c_4 = -1.7026; & c_5 = -0.4156; & c_6 = 0; & c_7 = -0.2220; & c_8 = -0.9216.\end{array}$$

Furthermore, the values of the components of the radius vector of the spacecraft's center of mass (in the inertial coordinate system) at the initial and final instants of time must be the same. In the authors' calculations, the coordinates of the dimensionless radius vector were calculated from the spacecraft orbit parameters given in [19, p. 95]:

$$\xi_1^0 = \xi_1^k = 23\,399\,727.8 \text{ m},$$
$$\xi_2^0 = \xi_2^k = 23\,962\,416.6 \text{ m},$$
$$\xi_3^0 = \xi_3^k = -18\,801\,552.4 \text{ m}.$$

In [19], the effectiveness of numerical prediction of transformations regularizing and stabilizing the equations of motion was investigated using this orbit as an example.

In this problem formulation, the mass-inertial characteristics of the spacecraft appear only in the formulas for finding the control force and control torque and do not affect the laws of variation of the optimal controls (optimal angular and linear accelerations). Therefore, the calculation results in dimensionless variables correspond to the calculations presented in [16] when solving the spacecraft reorientation problem.

In the case where the initial and final positions of the spacecraft coincide, the values of the components of the dimensionless control torque vector $\mathbf{M}$ found by the authors at the beginning, middle, and end of the motion in the absence of gravitational torque are close to those given in Tables 1, 4, 5 of reference [16, pp. 176–177] for a spherically symmetric spacecraft, the ISS, and the Space Shuttle spacecraft, respectively.

The value of the integral

$$J^{\text{modif}} = \int_0^{t_1} \left(1 + \mathbf{M}^2\right) dt \tag{5.5}$$

(the dimensionless minimized functional in [16]) in the case of spherical symmetry of the spacecraft turned out to be 1.4746, which is close to the number 1.4749 indicated at the end of p. 176 of reference [16] for the optimal control problem of angular motion of a spacecraft in the class of angular conic motions. The final time of the controlled process turned out to be 0.9657, which is close to the number 0.966 given in Table 1.

At the same time, the value of the optimization criterion (4.15) for the problem of optimal spatial motion (maneuvering) of the spacecraft turned out to be $J = 1.4708 + 0 \cdot s$, i.e., the functional value is a scalar. Let us further consider the case where the scale factors correspond to the work [20]: $L = 37\,000\,000.0$ m, $T = 11\,272.855470$ s. Then with $M_0 = 5.9722 \times 10^{24}$ kg and $G = 6.67408 \times 10^{-11}$ $\text{N} \cdot \text{m}^2 \cdot \text{kg}^{-2}$, the dimensionless parameter of the problem is $N_G = 0.99997 \approx 1.0$.

Note that in the case where the initial and final positions of the spacecraft coincide and do not change during its angular motion, the specific values of the components of the vector $\mathbf{r}_x$ do not affect the values of the sought constants $C_1, \ldots, C_5$, C_7, C_8, A_1, A_2. Therefore, the laws of variation of the phase variables and optimal control also do not depend on them. Moreover, the projections of the spacecraft's radius vector onto the axes of the inertial coordinate system ξ will be constant. Also, in the coordinate system ξ, the control force will be a constant vector. The control force vector will be anti-collinear with respect to the gravitational force vector (with the magnitudes of these vectors being equal). Over time, the projections of the gravitational force onto the axes of the spacecraft-fixed coordinate system X will change (the projections of this force onto the axes of the ξ coordinate system will not change).

Figure 1 shows the time-varying laws of the components of the control torque and control force vectors (they do not depend on the spacecraft's inertia tensor) in the body-fixed coordinate system, for $N_G = 0.99997 \approx 1.0$ for a spherically symmetric spacecraft (in this case, the gravitational torque is absent). In the coordinate system ξ, the control force vector is a constant vector.

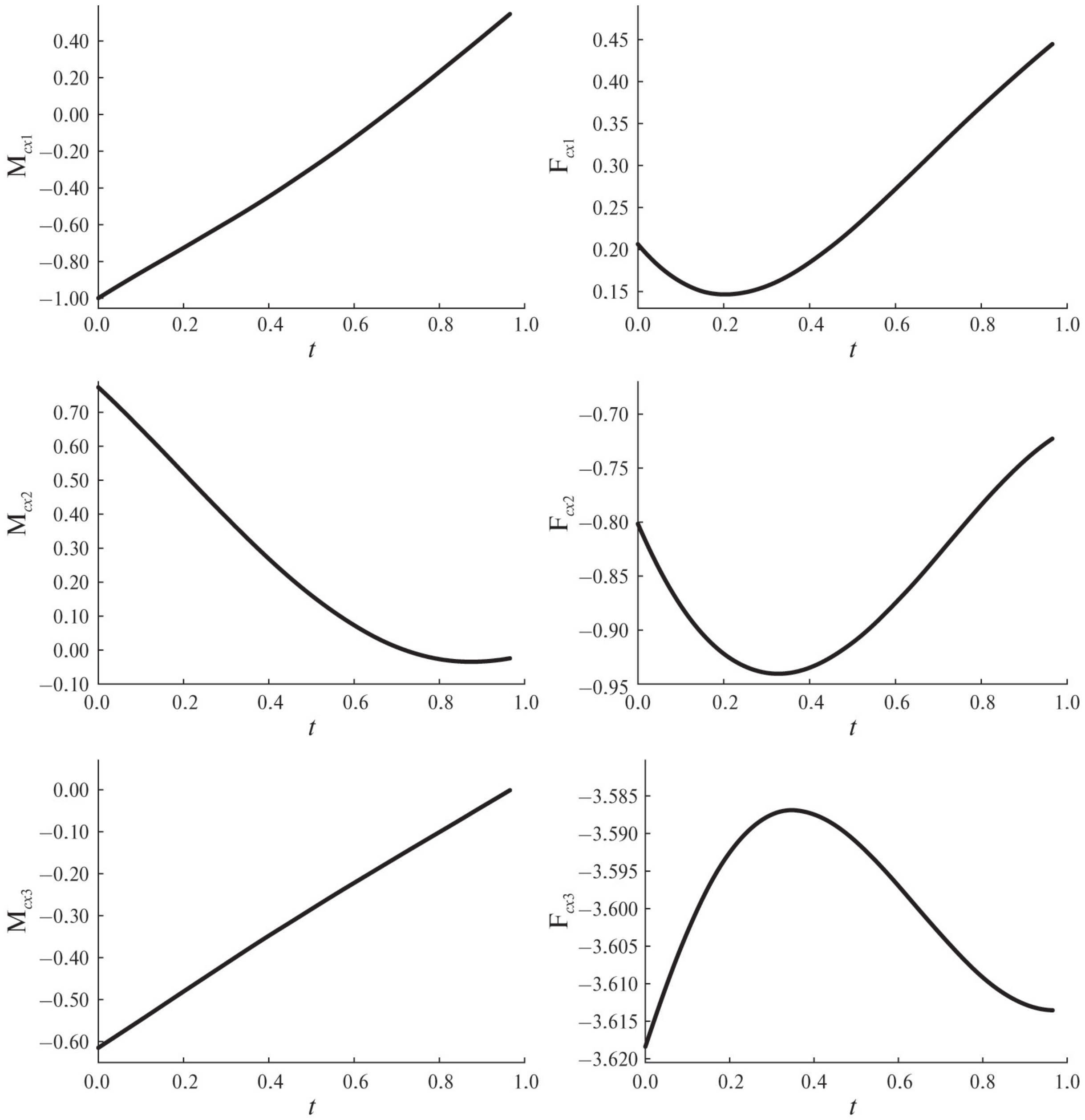

Fig. 1. Components of the control torque and control force vectors for SC 1 in the absence of translational displacement.

Laws of variation of the control torque vector components were also found for the ISS and the Space Shuttle spacecraft. In this case, the values of the integral (5.5) in the absence of gravitational torque turned out to be 1.3671 (ISS) and 1.3680 (Space Shuttle spacecraft), which is close to the numbers 1.3674 and 1.3683, respectively, indicated on p. 177 of reference [16] for the modified optimal control problem of angular motion.

Figures 2 and 3 show the dimensionless time-varying laws of the control torque vector components in the body-fixed coordinate system for the ISS and the Space Shuttle spacecraft (left column—in the absence of gravitational torque, right column—in its presence) when the coordinates of the initial and final positions of the spacecraft coincide.

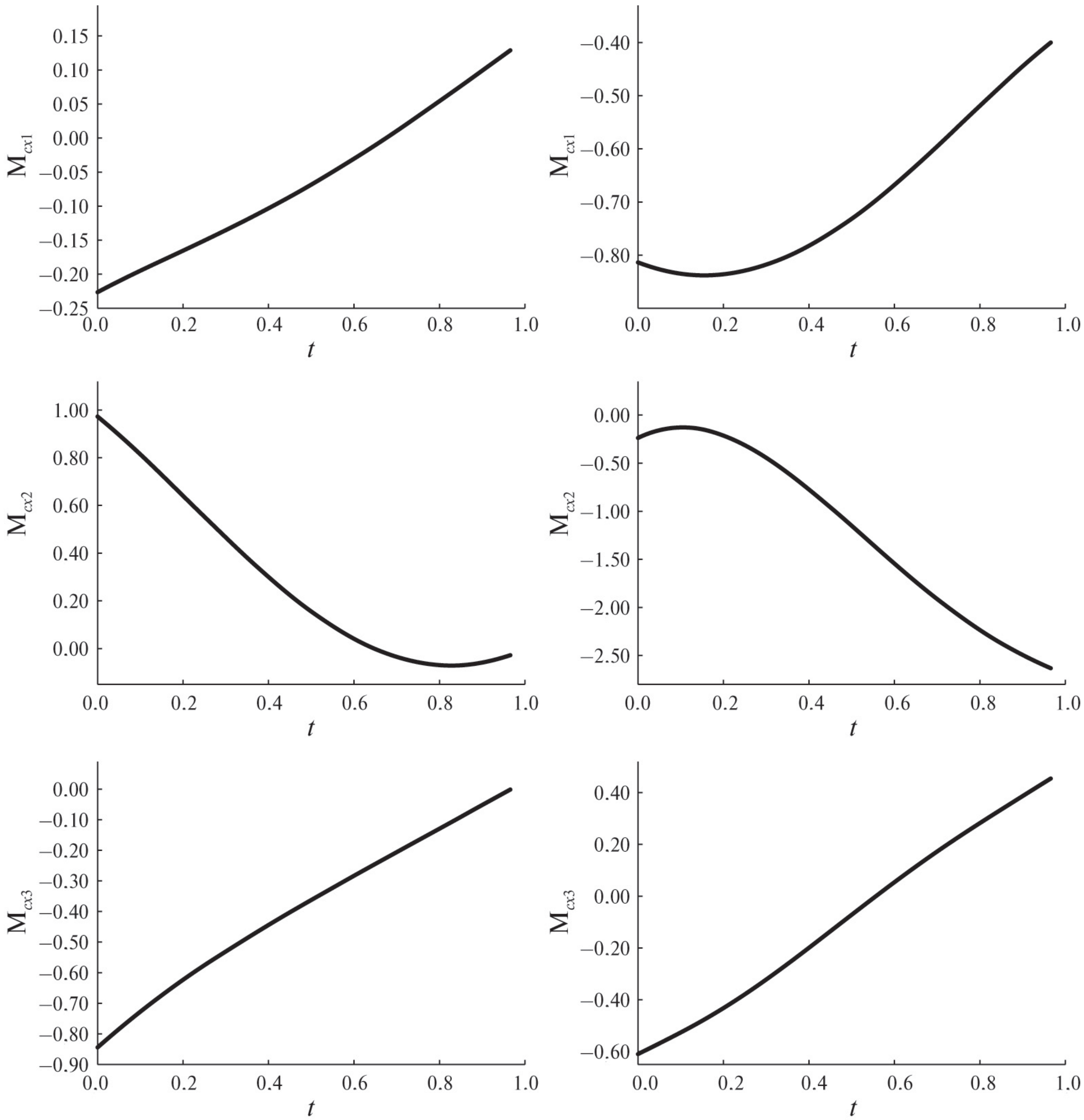

Fig. 2. Control torque vector components for SC 2 in the absence of translational displacement.

From the calculation results, it follows that for all spacecraft, the range of variation of the third component of the control force vector is an order of magnitude smaller than the ranges of variation of the other two components.

Next, calculations are presented for the case where, in addition to angular motion, the spacecraft also performs controlled translational motion. Let the position and velocity of the spacecraft in dimensionless variables correspond to the following orbit parameters [21]:

initial position of the spacecraft

$$\xi_1^0 = -12\,194\,795.0 \text{ m}, \quad \xi_2^0 = 21\,779\,195.0 \text{ m}, \quad \xi_3^0 = 8\,278\,547.0 \text{ m}.$$

$$\dot{\xi}_1^0 = -1080.750 \text{ m/s}, \quad \dot{\xi}_2^0 = -1849.256 \text{ m/s}, \quad \dot{\xi}_3^0 = 3274.225 \text{ m/s},$$

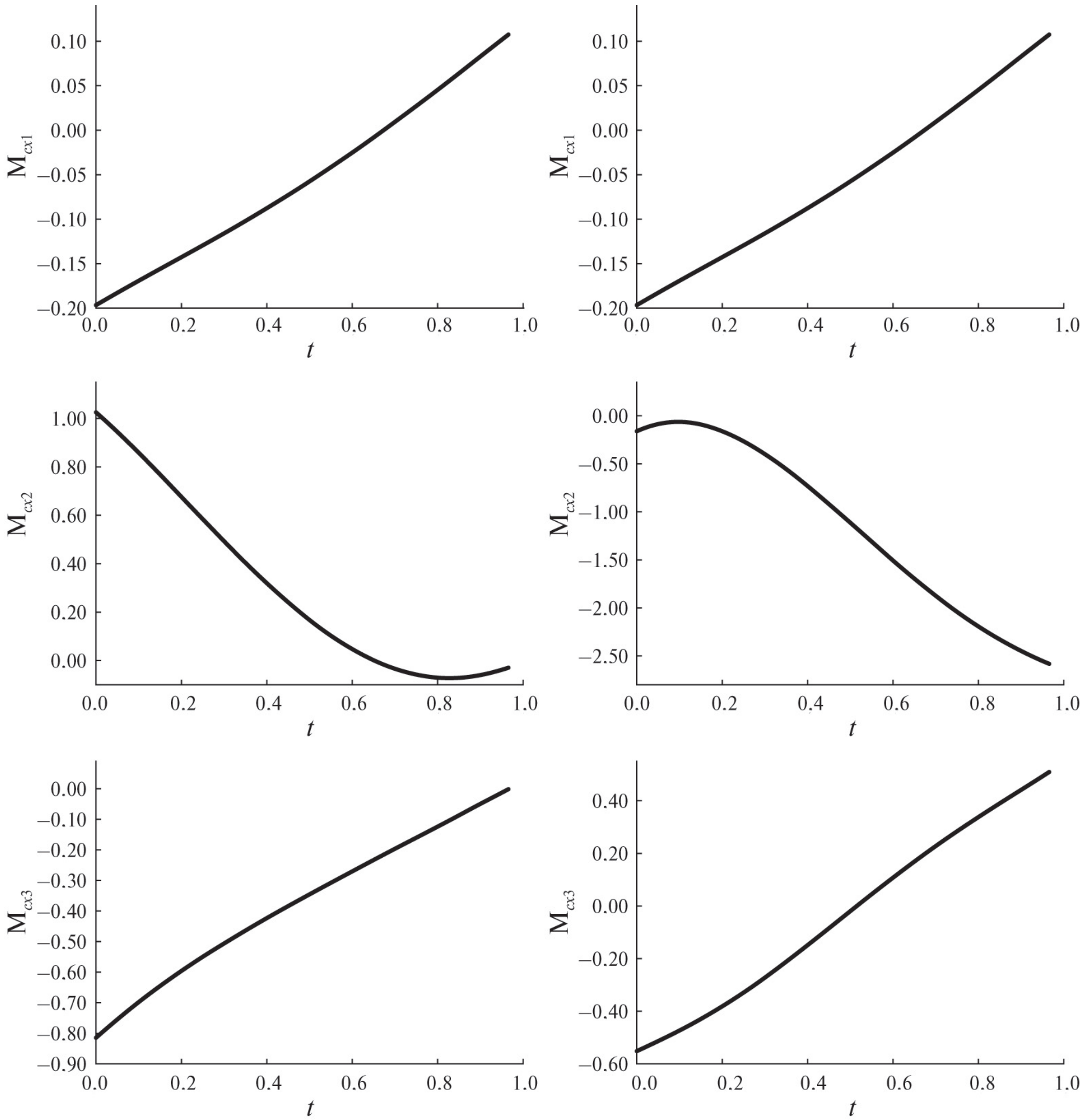

Fig. 3. Control torque vector components for SC 3 in the absence of translational displacement.

final position of the spacecraft

$$\xi_1^{\mathrm{k}} = -12\,110\,249.0 \text{ m}, \quad \xi_2^{\mathrm{k}} = 21\,643\,438.0 \text{ m}, \quad \xi_3^{\mathrm{k}} = 8\,744\,787.0 \text{ m}.$$

$$\dot{\xi}_1^{\mathrm{k}} = -1063.392 \text{ m/s}, \quad \dot{\xi}_2^{\mathrm{k}} = -1905.728 \text{ m/s}, \quad \dot{\xi}_3^{\mathrm{k}} = 3247.462 \text{ m/s}.$$

As an initial approximation for the undetermined dual constants $C_1, \ldots, C_5$, C_7, C_8, A_1, A_2 in the presence of translational displacement, the real values of these constants corresponding to a pure rotation of the spacecraft were taken. In this case, the principal parts of these dual constants (which correspond to the pure rotation of the spacecraft) and the final time of the controlled process remained practically unchanged as a result of solving the system of nonlinear algebraic equations.

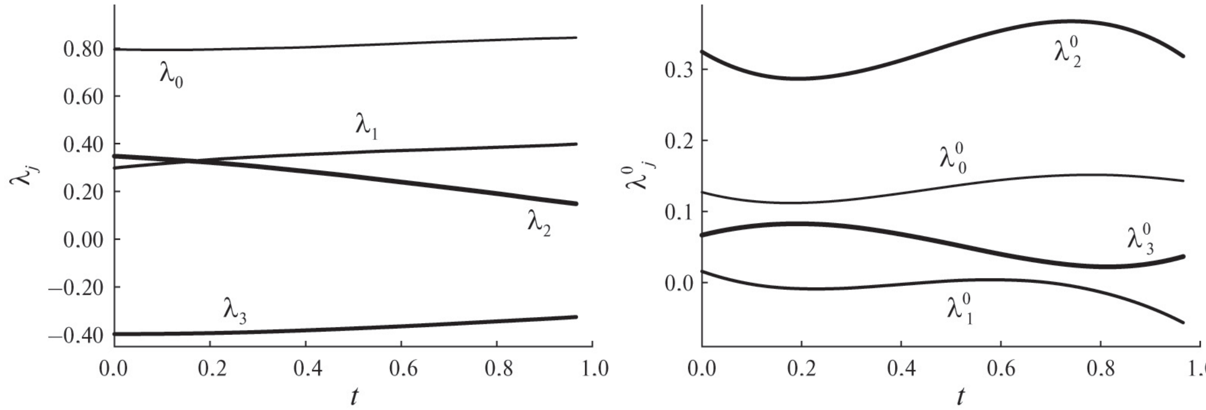

Fig. 4. Components of the dual quaternion of finite displacement for SC 1 in the presence of translational displacement.

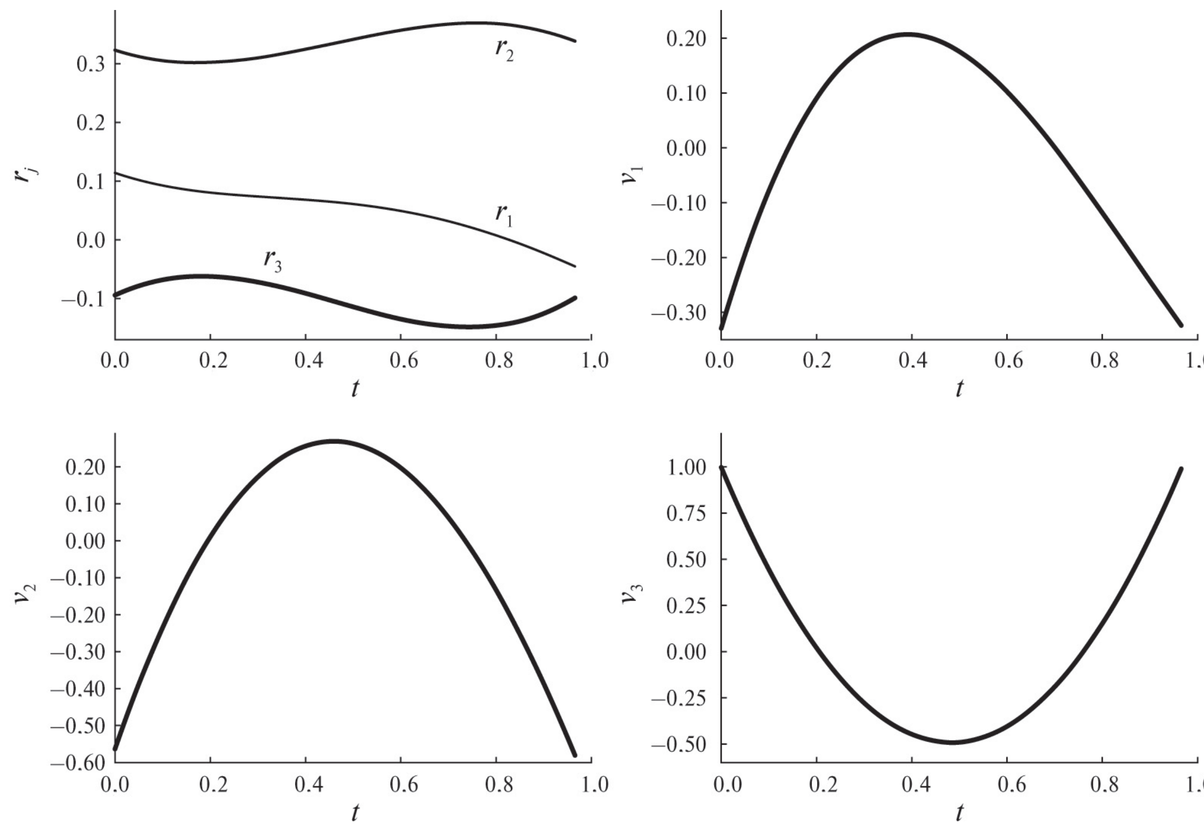

Fig. 5. Components of the radius vector and linear velocity vector for SC 1 in the presence of translational displacement.

The value of the optimization criterion (4.15) for the problem of optimal spatial motion (maneuvering) of the spacecraft turned out to be $J = 1.4708 + 1.2897 \cdot s$ (its principal part is the same as in the case of pure rotation of the spacecraft [16]).

Below are the calculation results in the presence of translational displacement of the spacecraft's center of mass. Note that the error in determining the components of the dual quaternion of the spacecraft's finite displacement at the final instant of time, both in the absence and in the presence of translational displacement, was about 10^{-16} dimensionless units. In this case, the error in

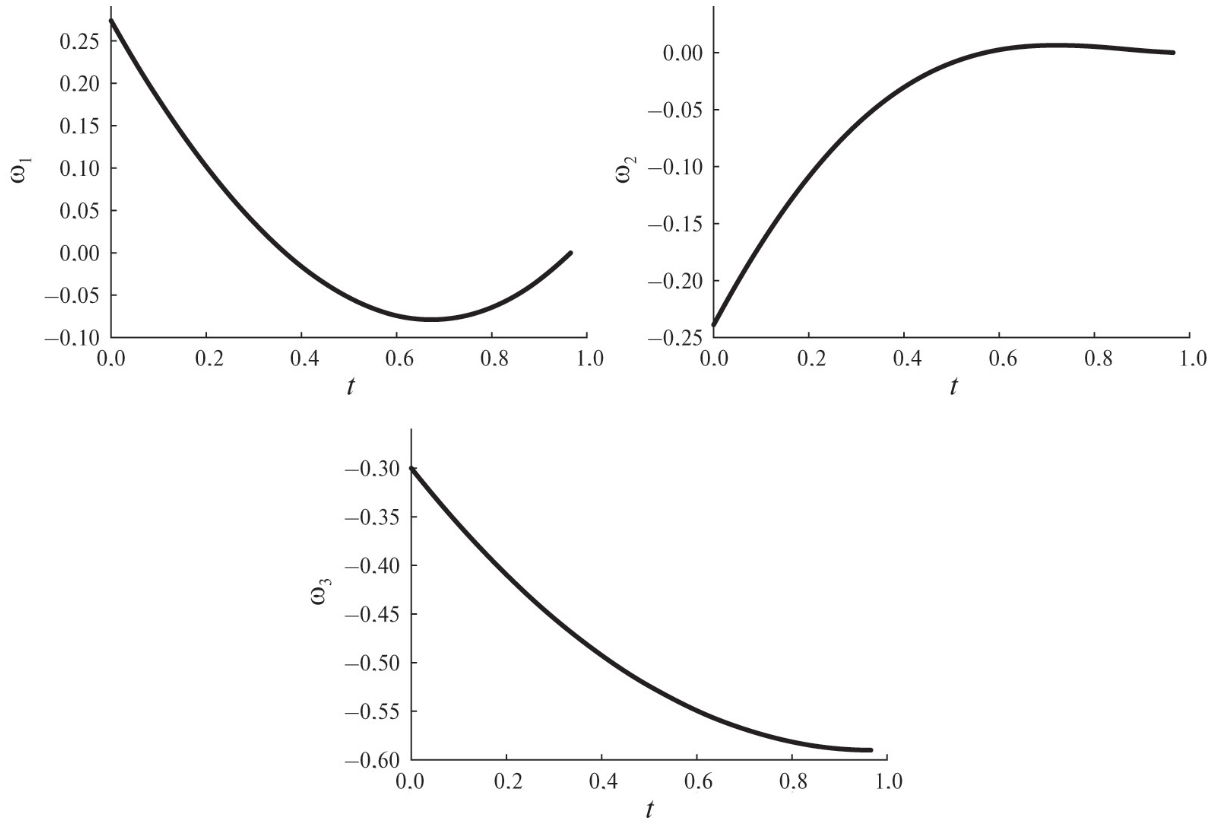

Fig. 6. Angular velocity vector components for SC 1 in the presence of translational displacement.

determining the components of the linear and angular velocity vectors was about 10^{-15} and 10^{-17} dimensionless units, respectively. The order of these values is due to the accuracy of solving the system of algebraic dual equations (4.5), (4.6), (4.16), (4.17), (4.19) for finding the undetermined dual constants $C_1, \ldots, C_5$, C_7, C_8, A_1, A_2, which was approximately 10^{-15} dimensionless units.

Figures 4 and 5 show the time-varying laws of the components of the dual quaternion of the spacecraft's finite displacement in inertial space, as well as the projections r_j $(j = \overline{1,3})$ of the radius vector and the projections v_j $(j = \overline{1,3})$ of the velocity vector of its center of mass onto the axes of the spacecraft-fixed coordinate system X for the case of a spherically symmetric spacecraft. The functions $r_j(t)$ and $\lambda_j^0(t)$ are harmonic functions of time (here and below, the laws of variation of dimensionless quantities are discussed). The laws of variation of the components of the vector $\mathbf{v}_x$ can be approximated by parabolas. One of the components is a convex downward function, while the other two are convex upward. Each of the components of the vector $\mathbf{v}_x$ changes sign twice during the spacecraft's motion (the components of this vector change sign almost simultaneously). The extremum points of the components of the vector $\mathbf{r}_x$ coincide with the time instants at which the corresponding component of the vector $\mathbf{v}_x$ is zero.

Figure 6 shows the time-varying laws of the angular velocity vector components for the case of a spherically symmetric spacecraft.

Figures 7 and 8 show the time-varying laws of the optimal angular and linear accelerations, control torque, and control force for the case of a spherically symmetric spacecraft. In the case of a spherically symmetric spacecraft, the optimal laws of variation of the angular acceleration vector components completely coincide with the laws of variation of the optimal control torque vector components. The second and third components of the optimal angular acceleration vector are close to zero at the end of the motion. The second and third components of the optimal linear

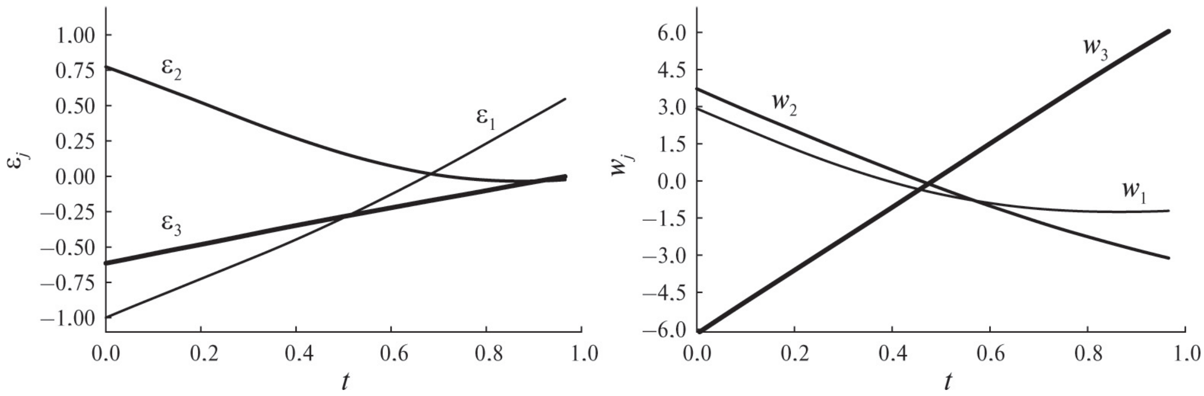

Fig. 7. Optimal control for SC 1 in the presence of translational displacement.

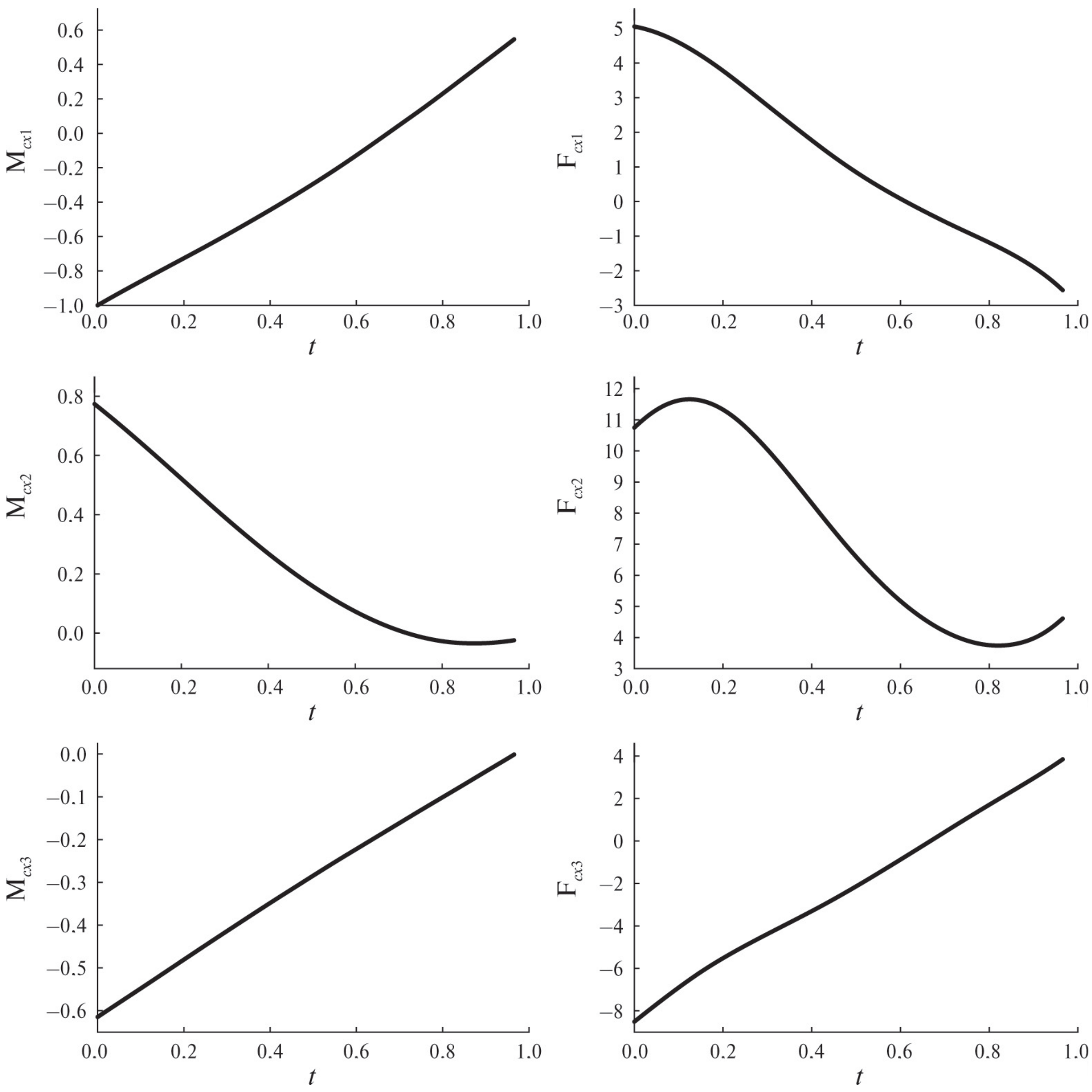

Fig. 8. Components of the control torque and control force vectors for SC 1 in the presence of translational displacement.

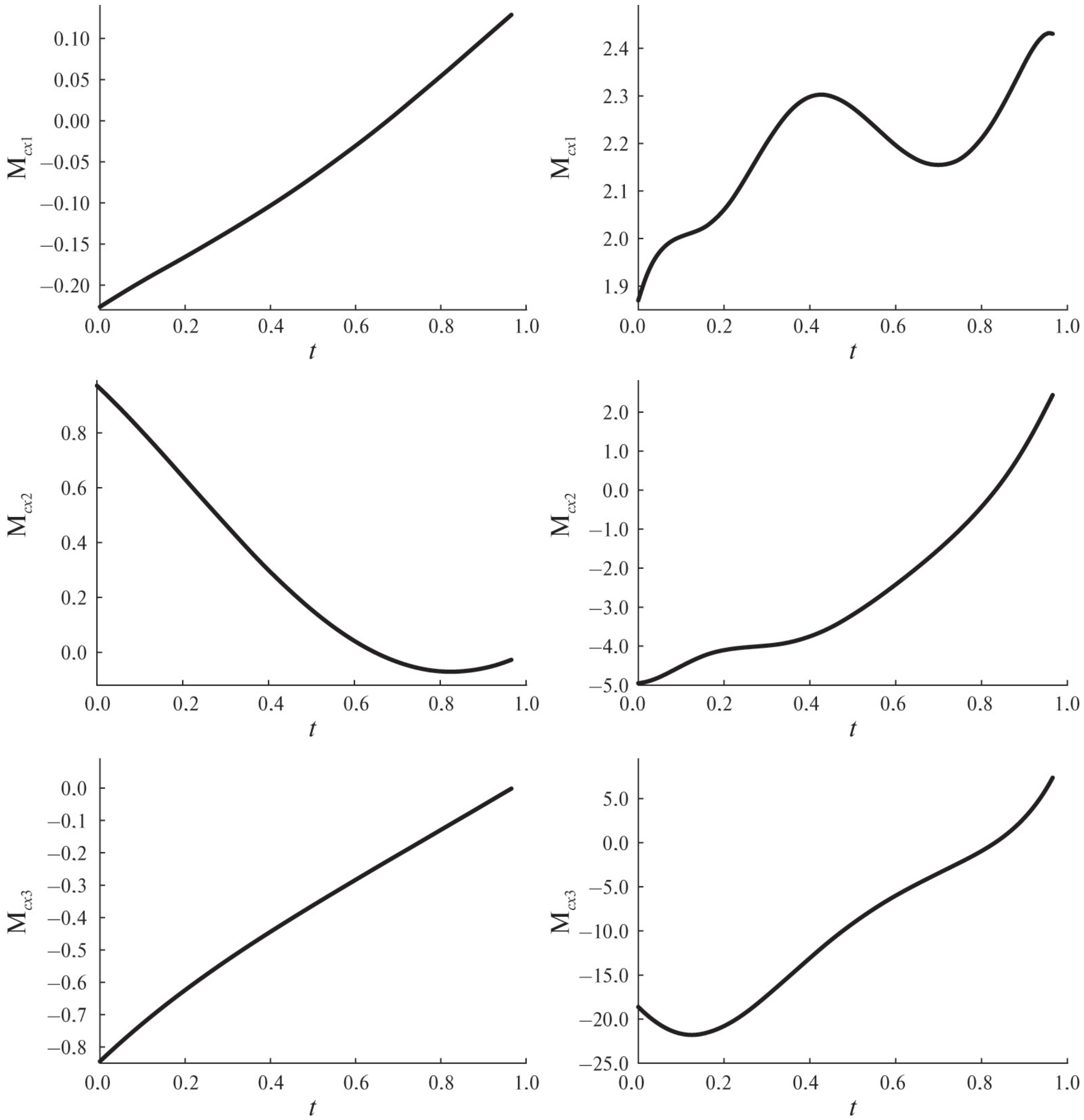

Fig. 9. Control torque vector components for SC 2 in the presence of translational displacement.

acceleration vector are close to linear functions of time. The sign change point of each component of this vector corresponds to the extremum point of the corresponding component of the spacecraft's linear velocity vector. The first component of the control force vector is a decreasing function, and the third is an increasing function. The second component of this vector varies according to a harmonic law.

Next, Figs. 9 and 10 show the time-varying laws of the control torque for the case where the mass distribution of the spacecraft corresponds to the ISS and the Space Shuttle spacecraft (left column—in the absence of gravitational torque, right column—in its presence). Note that since the inertia tensor **J** does not appear in expressions (5.1) and (5.2) for the control force, the laws of variation of the control force for the ISS and the Space Shuttle spacecraft coincide with the law shown in the second column of Fig. 8 for the case of a spherically symmetric spacecraft.

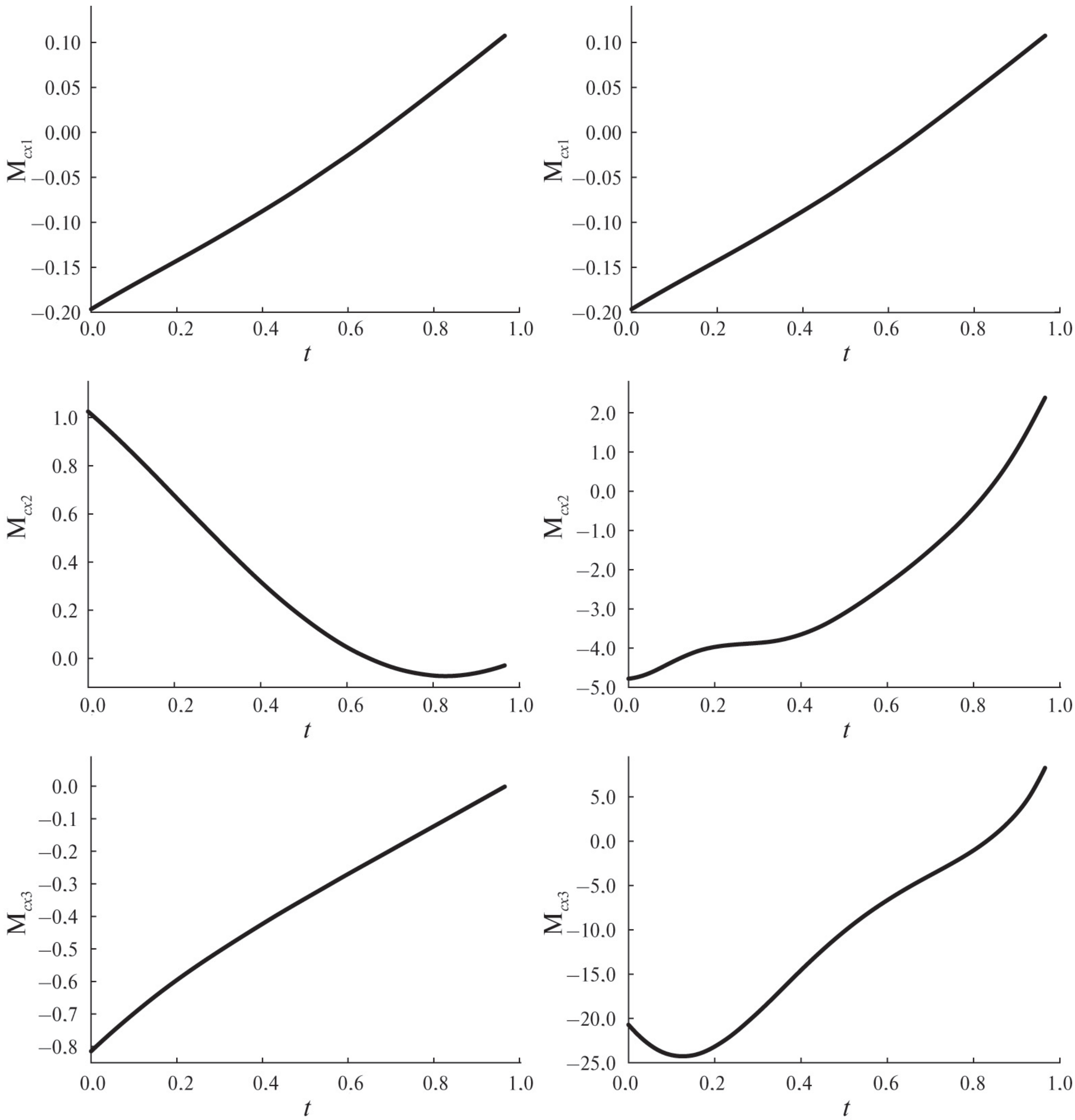

Fig. 10. Control torque vector components for SC 3 in the presence of translational displacement.

Note that the laws of variation of the other sought quantities qualitatively remained the same as in the previous case.

6. CONCLUSION

A dual quaternion (biquaternion) theory of optimal control of the spatial motion of a free rigid body (spacecraft) as an interrelated control of its spatial angular (rotational) and translational (orbital) motions has been constructed. The second derivatives of the parameters of the generalized helical conic motion of the spacecraft, equivalent to its general spatial motion, are used as optimized controls. The vectors of programmed control force and programmed control torque are found in accordance with the concept of solving inverse problems of dynamics.

The analytical algorithm for controlling the spatial motion (maneuvering) of a spacecraft as a free rigid body of arbitrary dynamic configuration with arbitrary boundary conditions, obtained on this basis, is optimal in the class of generalized helical conic motions and is applicable in spacecraft control systems. Moreover, it does not require a numerical solution of a complex high-dimensional differential boundary value optimization problem or any other complex numerical solution. This is important, in particular, for spacecraft spatial maneuvering at exceptionally high rates of displacement, when the time for calculating the optimal spatial trajectory of the spacecraft maneuver and for calculating the control laws enabling this trajectory is extremely limited.

REFERENCES

1. Strelkova, N.A., Time-optimal kinematic control of the helical displacement of a rigid body, *Izv. AN SSSR. MTT*, 1982, no. 4, pp. 73–76.
2. Malanin, V.V. and Strelkova, N.A., *Optimal'noe upravlenie orientatsiei i vintovym dvizheniem tverdogo tela (Optimal Control of Orientation and Helical Motion of a Rigid Body)*, Moscow–Izhevsk: R&C Dynamics, 2004.
3. Chelnokov, Yu.N., On integration of the kinematic equations of helical motion of a rigid body, *PMM*, 1980, vol. 44, no. 1, pp. 32–39.
4. Chelnokov, Yu.N., On a form of inertial navigation equations, *Izv. AN SSSR. MTT*, 1981, no. 5, pp. 20–28.
5. Chelnokov, Yu.N., *Kvaternionnye i bikvaternionnye modeli i metody mekhaniki tverdogo tela i ikh prilozheniya: Geometriya i kinematika dvizheniya* (Quaternion and Biquaternion Models and Methods of Rigid Body Mechanics and Their Applications: Geometry and Kinematics of Motion), Moscow: Fizmatlit, 2006.
6. Chelnokov, Yu.N., Control of spatial motion of a rigid body using biquaternions and dual matrices, *Izv. RAN. MTT*, 2021, no. 1, pp. 17–43.
7. Chelnokov, Yu.N., Synthesis of control of spatial motion of a rigid body using dual quaternions, *PMM*, 2019, vol. 83, no. 5–6, pp. 704–733.
8. Chelnokov, Yu.N., Control of spatial motion of a rigid body using dual quaternions, *Proc. XII All-Russ. Congr. Fundam. Probl. Theor. Appl. Mech.*, 2019, pp. 288–290.
9. Molodenkov, A.V. and Sapunkov, Ya.G., Analytical quasi-optimal solution of the problem of rotating an arbitrary rigid body under arbitrary boundary conditions, *Izv. RAN. MTT*, 2019, no. 2, pp. 140–154.
10. Molodenkov, A.V. and Sapunkov, Ya.G., Analytical quasi-optimal algorithm for programmed control of spacecraft angular motion, *Izv. RAN. TiSU*, 2023, no. 4, pp. 125–136.
11. Sapunkov, Ya.G. and Molodenkov, A.V., Quasioptimal spacecraft attitude control constructed according to the Poinsot concept, *Aerospace*, 2023, vol. 10, no. 5, pp. 402–417.
12. Chelnokov, Yu.N., Molodenkov, A.V., and Loginov, M.Yu., Biquaternion quasi-optimal analytical solution of the programmed control problem for spacecraft spatial motion, *Proc. XXX St. Petersburg Int. Conf. Integr. Navig. Syst.*, 2023, pp. 411–414.
13. Molodenkov, A.V. and Sapunkov, Ya.G., Solution of the optimal rotation problem for a spherically symmetric rigid body under arbitrary boundary conditions in the class of generalized conic motions, *Izv. RAN. MTT*, 2014, no. 5, pp. 22–34.
14. Branets, V.N. and Shmyglevsky, I.P., *Primenenie kvaternionov v zadachakh orientatsii tverdogo tela* (Application of Quaternions in Problems of Rigid Body Orientation), Moscow, Nauka, 1973.
15. Molodenkov, A.V. and Sapunkov, Ya.G., Analytical solution of the optimal rotation problem for a spherically symmetric spacecraft in the class of conic motions, *Izv. RAN. TiSU*, 2013, no. 3, pp. 167–176.
16. Molodenkov, A.V. and Sapunkov, Ya.G., Analytical approximate solution of the optimal spacecraft rotation problem under arbitrary boundary conditions, *Izv. RAN. TiSU*, 2015, no. 3, pp. 131–141.
17. Beletsky, V.V., *Dvizhenie iskusstvennogo sputnika otnositel'no tsentra mass* (Motion of an Artificial Satellite about Its Center of Mass), Moscow: Nauka, 1965.

18. Banit, Yu.R., Belyaev, M.Yu., Dobrinskaya, T.A., et al., Determination of the inertia tensor of the International Space Station from telemetric information, *Preprint No. 57*, Moscow: Keldysh Inst. Appl. Math. RAS, 2002.

19. Bordovitsyna, T.V., *Sovremennye chislennye metody v zadachakh nebesnoi mekhaniki* (Modern Numerical Methods in Problems of Celestial Mechanics), Moscow: Nauka, 1984.

20. Pankratov, I.A., Sapunkov, Ya.G., and Chelnokov, Yu.N., Quaternion models and algorithms for solving the general problem of optimal reorientation of a spacecraft orbit, *Izv. Saratov Univ. (N.S.), Ser. Math. Mech. Inform.*, 2020, vol. 20, no. 1, pp. 93–104.

21. Bordovitsyna, T.V. and Avdyushev, V.A., *Teoriya dvizheniya iskusstvennykh sputnikov Zemli. Analiticheskie i chislennye metody* (Theory of Motion of Artificial Earth Satellites. Analytical and Numerical Methods), Tomsk: Tomsk State Univ. Publ., 2016.

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