Volume 84, Number 9
September 2023

ISSN 0005-1179
CODEN: AURCAT

AUTOMATION AND REMOTE CONTROL

Automation and Remote Control

Vol. 84, No. 9, September 2023

Editor-in-Chief
Andrey A. Galyaev

http://ait.mtas.ru

Automation and Remote Control

ISSN 0005-1179

SCOPE

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

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

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

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

Volume 84 (12 issues) is published in 2023.

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

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

Contents

Automation and Remote Control
Vol. 84, No. 9, 2023

Reviews

Linear Systems

Nonlinear Systems

Robust, Adaptive, and Network Control

Control in Technical Systems

Optimization, System Analysis, and Operations Research

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 9, pp. 1017–1040.

REVIEWS

A Historical Essay on the Scientific School of V.A. Yakubovich

A. S. Matveev[*,a], **A. L. Fradkov**[**,*,b], **and A. I. Shepeljavyi**[*,c]

[*] *St. Petersburg State University, St. Petersburg, Russia*
[**] *Institute for Problems of Mechanical Engineering, Russian Academy of Sciences, St. Petersburg, Russia*
e-mail: [a] *almat1712@yahoo.com,* [b] *fradkov@mail.ru,* [c] *aishep@mail.ru*

Received June 15, 2023
Revised July 16, 2023
Accepted July 20, 2023

Abstract—The milestones of the history of the scientific school on cybernetics (the School), established in 1959 by outstanding scientist V.A. Yakubovich at Leningrad State University (LSU), are presented. The connections of the School with other Russian and foreign scientific schools in related fields are outlined.

Keywords: history, cybernetics, control theory, St. Petersburg State University, Department of Theoretical Cybernetics

DOI: 10.25728/arcRAS.2023.71.51.002

This paper describes the main milestones in the history of the scientific school of cybernetics and control theory (the School), established in 1959 by outstanding scientist V.A. Yakubovich at Leningrad State University (LSU). The School will celebrate its 65th anniversary in 2024. The essay is partially based on the publications [1–3] on the history of the Department of Theoretical Cybernetics, St. Petersburg State University (SPbSU), as well as its scientific directions and related issues. The authors are not intended to provide a complete bibliographic survey of the School's results concerning the aspects of its activities touched upon in the paper, particularly due to its limited scope. The authors present either key works or illustrative and sometimes subjectively selected examples of research works on a certain topic. The authors apologize to colleagues whose publications are not mentioned below.

The beginning of the history of cybernetics at St. Petersburg (Leningrad) University can be considered the year 1956, when Vladimir Andreevich Yakubovich, a 30-year-old candidate of physics and mathematics, came to the Faculty of Mathematics and Mechanics. It was a time of great changes in society and in science, the beginning of the thaw. The first electronic computing machines (ECM) and publications rehabilitating cybernetics [4, 5] appeared. Cybernetics was gaining popularity, and lectures and discussions about it spread everywhere. The country's first section of cybernetics was established at the Leningrad House of Scientists; it was headed by academician and future Nobel laureate L.V. Kantorovich. The Computing Center (CC) and research laboratories were organized in Leningrad University to master and use the new (fantastic, as it seemed at that time) capabilities of computers. Following the impact of the seminal book by N. Wiener [6], cybernetics was perceived primarily as a scientific foundation for the application of computer technology and automatic devices. Not surprisingly, when the administration of the Faculty of Mathematics and Mechanics proposed to V.A. Yakubovich to gather a group of researchers in the field of advanced mathematical methods of automation and control systems, the "cybernetic flag" became

most suitable for the group. Thus, in 1959, the Laboratory of Theoretical Cybernetics (LTC, the Laboratory) appeared in the CC of LSU.

The first years of LTC research focused on pattern recognition and machine learning. The Laboratory developed and generalized Rosenblatt's concept of perceptron, which was popular at that time, and several approaches to the mathematical theory of pattern recognition [7–9].[1] A series of applied problems were successfully solved, including handwriting and aerial survey photo recognition, extraction of useful signals from noisy material, and automatic description and analysis of scenes [11–14]. The Laboratory's team owns a series of original algorithmic solutions for the entire problem and its individual aspects, such as B.N. Kozinets' algorithm for memory-saving class separation [15, § 2.6], [16, Ch. 6], A.A. Schmidt's method of algebraic invariants in image recognition problems [16, Ch. 8], and others. Comprehension of the ideas accumulated at that stage led V.A. Yakubovich to the general concept of an infinite a priori unknown recursive system of inequalities, where inequalities are added step-by-step in real time, and finitely convergent algorithms for solving such systems in real time [17]. Later on, this concept and related methods were repeatedly shown to be productive in various fields. The key approach to solving such systems developed in LTC was subsequently called *the method of recursive aim inequalities* [18].

The emergence of a new field— cybernetics—inevitably gave rise to a discussion of its relationship with the traditional theory of automatic control. A fruitful channel for that discussion was paved, among others, by the notion of adaptability, i.e., the autonomous capability of a system to adjust successfully to a priori essentially uncertain conditions of its operation (external and internal). In those years, statistical approaches to adaptive control prevailed in the Moscow school. In particular, Ya.Z. Tsypkin elaborated the theory of adaptive and learning systems based on statistical estimation and stochastic approximation methods [19, 20]. In parallel, V.A. Yakubovich developed an original alternative (deterministic) approach without involving probability theory; a key element of his approach is the method of recursive aim inequalities. V.A. Yakubovich gave historically the first general mathematical definition of an adaptive system [21, 22]. The basic material on the theory of recursive aim inequalities and adaptive control can be found in the monographs [15, 23]; a survey of subsequent works is available in [24–26].

The results of V.A. Yakubovich's team in the field of adaptive systems were naturally continued in robotics research. Initially, scientists all over the world carefully avoided the term "robot" and its derivatives, believing them to be frivolous and suitable for science fiction at most. There are grounds to state that V.A. Yakubovich pioneered "robot" as a now generally recognized scientific concept; see his paper [21] published in *Doklady Akademii Nauk USSR*. The method of recursive aim inequalities was used therein to solve the problem of self-learning of a manipulator robot (an "eye–arm" robot) and to prove several theorems about "the rationality of robots" in the sense of the definition introduced.

In almost all industrialized countries, the late 1960s and early 1970s were marked by the rapidly growing interest in production automation based on manipulator robots with elements of artificial intelligence. On the crest of that wave, in 1973, a robotics group was formed in LTC, headed by V.A. Yakubovich's students—Dr. Sci. (Eng.) A.V. Timofeev [27–29], and then Cand. Sci. (Phys.–Math.) S.V. Gusev [30, 31]. Note the main robotics achievements of LTC during that period: a mathematical theory of adaptive robots and a theory to train them to complex rational behavior [32–34]. The viability of this theory was first demonstrated by vivid examples of solving prototypical problems, such as the problem of training a robot to ride a two-wheeled bicycle, as well as training other adaptive robots. (As pets in the team they received nicknames "grasshopper," "hawk," "eye–arm," and others.) The significance of the solved problems is emphasized by the fact

[1] The paper [7] was actually the first work in Russian devoted to machine learning. It was reprinted and translated into English in 2021; see [10].

that the corresponding results in 1972 were selected by the international organizing committee for presentation at the IFAC World Congress (Paris, 1972); see [35]. (It was joked that V.A. Yakubovich went to Paris on a bicycle as a speaker.) In the future, the effectiveness of the theory developed in the team was demonstrated by experiments (one of the first in the country) with real wheeled robots [30], started in 1974; in 1980, they were continued using a more advanced experimental robot developed in LTC [36]. In the 1980s, the LTC robotics group participated in the development of a manipulator control system for Buran, the reusable space shuttle project. In the 1970s–1980s, the LTC team also elaborated the theory of adaptive control of robotic systems described by general Lagrange equations [37–40]; those studies were pioneering in many respects and underlie the subsequent large-scale development of the corresponding research area in the world.

The rapid development of cybernetics and control theory in the 1960s led to the emergence of numerous algorithms for control, adaptation, recognition, learning, estimation, and filtering. The need arose to generalize the results obtained and to unify the algorithms proposed, i.e., to identify their key ideological core. Probably, Ya.Z. Tsypkin was the first to feel that need [19, 20]: he proposed to treat various problems of recognition, estimation, control, etc. as problems of minimizing the mean of a certain loss function. As a result, the chaotic mass of then-existing disparate algorithms was represented in the form of systematized special cases of uniform probabilistic gradient iterative procedures for minimizing or estimating parameters. However, the basic adaptation (self-learning) and control algorithms related to the continuous-time case did not fit into that scheme. Following analysis attacks from different directions, it gradually became clear that the algorithms mentioned can be unified within Ya.Z. Tsypkin's scheme by passing from the gradient of the objective function to the gradient of the rate of its change along the trajectories of the controlled object. Probably, the most general and complete approach to implementing that idea was proposed and developed by A.L. Fradkov [41] and named by him *the speed gradient method*.

Initially, the method was mainly focused on adaptive control and identification problems. As a result of subsequent many-year research, it was developed and applied as a universal approach to designing various continuous dynamic systems in mathematics, physics, engineering, biology, and other fields. For example, based on this approach, control and synchronization problems were solved for a wide class of oscillatory, including chaotic, systems. The corresponding results opened new perspectives in vibration engineering, laser and chemical technologies, and information transmission systems. Due to practical simplicity and the availability of a rigorous mathematical justification of the obtained algorithms, this method became generally recognized as a research tool, both in the USSR (Russia) and abroad. The number of publications where the method is applied in one form or another has been constantly growing and now reaches several hundred. Recently, interest in the speed gradient method has increased as a tool for understanding the laws of evolution to comprehend better the dynamics of physical, biological, and other systems. In this form, the method is known as *the speed gradient principle* [42, 43].

In the early 1970s, a bionics group was established at LTC under the leadership of Dr. Sci. (Psy.) R.M. Granovskaya. The task of the group was to study and model the phenomena of perception and recognition, as well as memory mechanisms of living organisms, including humans [44, 45]. A considerable amount of experimental and theoretical studies was conducted, and the results obtained were largely motivated and actively implemented by interested organizations.

In 1970, the Department of Theoretical Cybernetics (DTC, the Department) was established on the basis of LTC. The three pioneering alumni of the Department—G.S. Aksenov, B.D. Lyubachevskii, and A.L. Fradkov—graduated soon, in 1971. LTC and DTC were in fact a single team, with common affairs and minimal influence of the formal distribution of employees between them. The LTC staff was engaged in teaching, while DTC members conducted research on topics common to LTC and very often in collaboration with LTC colleagues. Discussion of relevant components of

that research was systematically transferred from the Laboratory walls to the classrooms: while still learning the professional base, students were exposed to the cutting edge of the field. For example, the LTC-DTC staff often presented new, as yet unpublished results in lectures. Sometimes students were given proofs of theorems that had been obtained only the day before. There were a sense of lively participation in mathematical creativity and a feeling of being at the forefront of science. Sometimes, students found inaccuracies in proofs or suggested ways to improve the considerations. Such students were thanked in publications, which caused a sense of pride and a desire to move on. Note that from long ago, the motto of the department has been Docendo discimus, which means "learning by teaching."

In addition to the purely cybernetic direction (recognition, machine learning, artificial intelligence, adaptive systems, robots, etc.), the field of scientific interests of the team has been covering several classical branches of mathematics and control theory. They concern linear differential equations, dynamic systems and parametric resonance (V.A. Yakubovich, V.N. Fomin, and V.I. Derguzov), stability and oscillations in nonlinear dynamic systems, including phase synchronization and frequency auto-tuning systems, stability and oscillations in pulse-modulated systems (G.A. Leonov, A.I. Shepeljavyi, A.Kh. Gelig, and A.N. Churilov), optimal control (A.S. Matveev, A.E. Barabanov, and V.A. Yakubovich), estimation and filtering theory (V.N. Fomin and A.E. Barabanov), and others.

Even before the establishment of the Laboratory and the Department, V.A. Yakubovich obtained fundamental results on the stability of linear systems of differential equations with periodic coefficients and parametric resonance. He proved I.M. Gelfand's hypothesis that in the functional space of coefficients of two-dimensional Hamiltonian systems, the set of coefficients corresponding to stable systems decomposes into a countable number of connected domains; moreover, he showed that the Lyapunov criterion, popular in the subject in those years, applies to one of them only. V.A. Yakubovich obtained stability criteria for each domain, which, like the mentioned Lyapunov criterion, are irreducible in a certain natural sense. These results were then transferred by V.N. Fomin and V.A. Derguzov to systems with an infinite-dimensional state space. The fundamental monograph [46] summarized the intermediate outcomes of this direction and is still actively cited in the works of mathematicians, physicists, and engineers.

Among the numerous scientific results of the team, perhaps the greatest fame and influence has been gained by the achievements related to the so-called "frequency theorem," also known as the Yakubovich–Kalman lemma and the Kalman–Yakubovich–Popov (KYP) lemma. It was proved by V.A. Yakubovich and first published in 1962 [47]. This theorem gives mathematically beautiful, transparent, and constructive conditions for the solvability of a rather complex system of relations, which is found in a variety of problems in stability theory, automatic control, robotics, and other fields; in turn, the solution of this system of relations is the key to solving the main problem and its qualitative analysis. The importance and authority of the frequency theorem are the derivatives of its productivity in a whole range of diverse fields and problems, where it has given a second wind to the method of Lyapunov functions. For example, it allowed obtaining a whole series of new constructive criteria for absolute stability, instability, auto-oscillations, and the existence of globally stable periodic and almost periodic modes in a variety of nonlinear systems, as well as advancing in the study of the so-called strange attractors of such systems and developing new optimal and adaptive control methods; some of these results were presented in 1978 in the monograph [48]. This book is still relevant and interesting to scientists of different countries, as evidenced, in particular, by the publication of its English translation in 2004. Moreover, the lemma under consideration allowed establishing a kind of exhaustive results, surely covering all the conditions of a given type of system behavior, that can be obtained using Lyapunov functions from popular classes (e.g., a quadratic form, a quadratic form plus the integral of a nonlinearity, etc.).

Note that the frequency theorem is sometimes called the Great Lemma of Systems Theory: it is "officially" recognized by the international scientific community as one of the cornerstones of modern control theory. For example, this fact is reflected by the presence of V.A. Yakubovich's paper on the frequency theorem [47] in *Twenty Five Seminal Papers in Control* (Wiley—IEEE Press, 2000), a special collection containing 25 papers with the greatest impact on the development of control theory in the 20th century according to the IEEE Control Systems Society.

At first, the frequency theorem was proved for control systems described by ordinary differential equations. Subsequently, it was extended in different directions, in particular, it was transferred to many other classes of controlled systems. Among them, note discrete-time systems, stochastic systems, adaptive systems, systems with an infinite-dimensional state space (e.g., described by partial differential equations, equations with delayed argument, differential equations in an infinite-dimensional Hilbert space, integral equations, etc.), and systems over ordered fields [49–59]. These achievements were overwhelmingly not an end in themselves but a road to a scattering of new self-sufficient results pushing the boundaries of understanding of relevant fields, e.g., to the criteria of absolute stability and instability for the classes of systems under consideration. In this scientific development, the school of V.A. Yakubovich went, mutually enriching, hand in hand with other scientific schools, e.g., with the Nizhny Novgorod school (V.A. Brusin, P.V. Pakshin, V.A. Ugrinovskii, etc.) [60–64]. The history and current state of this direction were described in detail in the surveys [65, 66] and the collective monograph [3].

Note that necessary and sufficient conditions for the existence of linear output-feedback of a linear system ensuring the existence of its quadratic Lyapunov function were obtained in [55, 56]. This property of the system is equivalent to its passivity, meaning the fulfillment of some dissipation-type inequality on the trajectories of the system. Therefore, the results [55, 56] can be termed passification theorems for linear systems. These statements underlaid a general approach to system design called the passification (passivation) method. Subsequently, the method was extended to a wide class of control and estimation problems for nonlinear and adaptive systems [67–70]. The passification method is now applied by researchers from various countries [71–73]. In Russia, it is actively used particularly in the scientific school of ITMO University (V.O. Nikiforov, A.A. Bobtsov, etc.) [74, 75, 141].

The wide applicability of the frequency theorem motivated V.A. Yakubovich to construct an abstract theory of absolute stability: using the apparatus of functional analysis, such a theory generalizes the mass of known results and also creates a comfortable basis for their extension to all new types of equations. Note that research on the frequency theorem is related to the now ultra-popular method of linear matrix inequalities (LMIs). Accordingly, the authors of the book [77] called V.A. Yakubovich the "father" of the scientific direction based on this method (in honorable company with "grandfather" A.M. Lyapunov). A lot of adherents in the world have been successfully developing this direction for a long time in a surprisingly wide range of applied fields.

The frequency theorem was born in the stability analysis of equilibria of nonlinear dynamic systems as an answer to the following question: under what conditions does there exist a quadratic Lyapunov function common for a whole class of such systems described using a quadratic form? Subsequently, its fundamental character was manifested in the discovery and effective utilization of its connections with a number of other fields. The theory of optimal control was among the historically first of them. Here, the frequency theorem proved to be a powerful constructive tool for checking the solvability of linear quadratic control problems (a combination of a linear control system and a quadratic performance criterion) and designing their solutions in the engineeringly attractive form of an optimal controller.

The foundations of the linear quadratic theory of optimal control were laid by the classical works of R. Kalman [78], N.N. Krasovskii [79], and A.M. Letov [80] (in the part concerning stochastic objects, by the investigations of A.N. Kolmogorov [81], N. Wiener [82], and S. Bucy [83]); a significant contribution to its development was made by J.C. Willems, V.I. Zubov, V.M. Kuntsevich, A.B. Kurzhanski, J.-L. Lions, A.I. Lurie, V.I. Utkin, V.A. Yakubovich, and many other scientists. (For the history of the linear quadratic theory of optimal control, see the surveys [84, 85].) Methodologically, this field has important connections with complex analysis, stability, and stabilization theory of nonlinear dynamic systems ([77, 86–88], etc.). First of all, it concerns the so-called uncertain systems, where the epithet reflects a common situation for applications: complete information about the system is unavailable. Starting from the late 1960s, the flow of scientific publications on the subject has become an avalanche, with a persistent marked interest until the present time. One reason is the generally recognized practical effect of the linear quadratic theory of optimal control. For example, according to the plenary report of Prof. M. Morari at the Second European Control Conference (Groningen, the Netherlands, 1993) [89], the linear quadratic theory occupies an honorable second place in the intensity of use in civil industrial applications among all branches of modern mathematical control theory. (The first place was given to the theory of PID controllers.)

In the theory of linear-quadratic optimization, the school of V.A. Yakubovich systematically developed the approach based on the frequency theorem. In the works of V.A. Yakubovich, A.I. Shepeljavyi, A.L. Likhtarnikov, A.V. Megretsky, S.G. Semenov, D.V. Plyako, A.V. Savkin, etc., the approach was extended to a wide class of important problems and systems, including, among others, systems with continuous and discrete time, systems with infinite-dimensional state space, problems arising under conflict (differential games), and problems with the singularity effect [85], which may cause no solution in the conventional sense.

The frequency theorem as a criterion for the existence of a quadratic Lyapunov function is traditionally and often supplemented by a special technique for constructing such a function, the so-called S-procedure [90]. In [91], it was abstracted from Lyapunov functions and given the sense of replacing (in a certain interpretation) a system of several inequalities by a single inequality with a free parameter. The key question here consists in the following: is the replacement equivalent? Under the affirmative answer, the S-procedure is said to be lossless, and the corresponding statement is also called the S-lemma. This question is relevant to several fields of mathematics [77], e.g., duality in extremum problems, matrix theory, and operator theory. The case of quadratic inequalities, where the losslessness of the S-procedure adjoins the effect of hidden (non-obvious) convexity of images of quadratic mappings [66], has proved to be particularly productive for control theory. Classical results of this kind are the Dines theorem (two quadratic forms transform a real linear space into a convex set) and the Toeplitz–Hausdorff theorem (two continuous Hermite forms transform the sphere of a complex Hilbert space into a convex set).

The first studies of the School on the losslessness of the S-procedure [92, 93] started at the turn of the 1970s and dealt with no more than three inequalities. Basically, they stayed within the idea field of the Dines and Toeplitz–Hausdorff theorems, in which, according to P. Halmos [94], all known proofs are based on calculations, although it is desirable to have an idea proof, at least (or especially?) using less elementary concepts. Further research of the School on the subject can be interpreted as a movement in the above direction, where the main goal was generalization to an arbitrary number of forms (unattainable in the general case). A noticeable impetus to this research was given by the students of V.A. Yakubovich and N.K. Nikol'skii in the work [95], where the convexity of the joint image was established for an arbitrary number of forms but in a very special situation motivated by control theory. The specialization was rather quickly overcome by V.A. Yakubovich together with A.S. Matveev, who joined this subject a little later: they obtained a series of general results on the losslessness of the S-procedure and the hidden convexity of quadratic

functionals. In their works, the less elementary common reason for convexity was the invariance of forms with respect to shift operators and the weak convergence to zero of shifted space elements (as, e.g., in $L_2(0,\infty)$ when shifted by $T \to \infty$) [96]. Those results concerned important problems in control theory but did not cover the classical Dines and Toeplitz–Hausdorff theorems. Subsequently, A.S. Matveev obtained even more general criteria for the convexity of the joint image of an arbitrary number of forms; they "automatically" covered the classical results and presented certain properties (obviously fulfilled in the classical case) of the peripheral part of the spectrum of the operator bundle generated by the forms as a "less elementary" reason of convexity [97–99]. In that series of papers, the theory of approximate convexity of images of quadratic mappings with defect estimates was also developed, the property of hyper-convexity was discovered and investigated, and the results were extended to more general (non-quadratic) mappings. Several results in this direction were also obtained by outstanding scientist B.T. Polyak [100, 101]. Motivated by the theory of stochastic control, N.G. Dokuchaev (with the participation of V.A. Yakubovich) developed a parallel ideology related to A.A. Lyapunov's effect (the convexity of the image of an atomless vector measure).

At the turn of the 2000s, an important discovery presented the relationship and interaction of the S-procedure and the frequency theorem in a completely new light. Namely, a new proof of the frequency theorem was given in [102] based on the losslessness theorem of the S-procedure (S-lemma). As a result, figuratively speaking [66], the frequency theorem and the S-procedure lived for a long time as friendly neighbors, and now, after so many years, everyone has found out that they are also relatives.

The work [102] stimulated research on the so-called generalized frequency theorem (generalized KYP-lemma), establishing applications-relevant properties equivalent to the fulfillment of frequency domain inequalities in some restricted frequency range. The corresponding results provide new system analysis and design tools related to frequency domain inequalities satisfied in a finite frequency range [102, 103]. As it turned out, the standard frequency domain inequality in a finite frequency range is equivalent to some non-classical linear matrix inequalities for the pair of matrices P, Q; in a certain sense, these inequalities are analogous to and "replace" the inequalities for a single matrix P in the classical KYP lemma. According to [104], the frequency domain inequality in a finite frequency range is, in turn, equivalent to definite inequalities (of the dissipation type) only on part of the system trajectories defined by an additional integral matrix inequality (the so-called restricted dissipativity [104]). Thus, a complete extension of the classical KYP results to the finite-frequency case was obtained. In [105], the above results were further generalized to the case of the "conic" S-procedure to work with an infinite number of constraints. The finite-frequency version of the frequency theorem has already found application in several practical problems [106–108].

In the 1990s, the three main scientific directions of the School—the frequency theorem, the S-procedure, and linear-quadratic optimization—merged in the research on nonconvex global optimization methods. More precisely, the matter concerns a general approach based on these directions in order to develop efficient algorithms for special problems in the field of nonconvex global optimization in a standard way. Unlike the majority of methods in this field, which are mostly computational, often involve heuristic ideas, and do not always converge, the algorithms mentioned above rest on a mathematical theory, are analytical in their most essential part, and surely yield the global optimum. The general approach was proposed by V.A. Yakubovich in 1992 [109, 110]. It was further developed in the works of V.A. Yakubovich, A.S. Matveev, and N.G. Dokuchaev. This approach justifies the basic relations of the theory of convex duality for the nonconvex optimization problems under consideration and solves them using the (Yakubovich) rule based on the relations. The rule is not necessarily correct. It was established that the rule is correct whenever it is effective (produces a non-empty set of answers). Despite this fact, of greatest interest are the criteria to verify the applicability of the rule a priori (before its application) based on a (usually

simple) check of certain properties of the initial problem data. Several such criteria were derived. Note that in many respects, these criteria served as the main purpose of the studies of images of quadratic mappings discussed above.

In the late 1970s, V.A. Yakubovich initiated an extensive cycle of research for his team to elaborate the theory of the maximum principle in optimal control problems within an abstract approach. This approach implies the choice of some abstract model described by the language of functional analysis as the main object of study. The results obtained for the model are then supposed to be interpreted with respect to the specific models encountered by the researcher. Thus, when working with a variety of applications, this approach allows reducing the amount of considerations: much of them have already been done once and for all within the abstract theory. Another advantage of the abstract approach is a uniform procedure for deriving optimality conditions. Methodologically, it provides a more accessible and simple presentation of the main ideas: they are not obscured by the entourage of a particular model.

Such abstract theories were elaborated by many authors. V.A. Yakubovich proposed his own (original) approach to constructing an abstract theory of optimal control. Its characteristic feature consists in the apparatus of the calculus of differentials on bundles of (generally nondifferentiable) curves. On this basis, an abstract maximum principle is established for an abstract model of an optimal control problem. In particular, it explains why maximum principles analogous to Pontryagin's maximum principle naturally arise as necessary conditions of optimality in very seemingly different problems, highlighting the general properties of the problem that predetermine the specified form of the answer. This approach also yielded a uniform theory of necessary conditions of the first and higher orders in problems with constraints: all of them turn out to be parts of some single condition [111]. The approach under discussion was developed in different directions in an extensive series of works by V.A. Yakubovich and his students; a number of self-sufficient new results were obtained on its basis. Some of them (e.g., concerning the optimal control of systems described by partial differential equations) were significantly ahead of similar research results in the world. Some outcomes of those studies were systematized in the books by A.S. Matveev and V.A. Yakubovich [111, 112]. The textbook [112] was intended to teach the reader to independently apply the abstract theory to new problems. The book contains 75 problems on the application of this theory. Some of them correspond to the level of scientific publications of the recent past; at the same time, they are successfully handled by fourth-year students of the Faculty of Mathematics and Mechanics (SPbSU).

Since the inception of LTC and DTC, there were two assistant captains on their command bridge: A.Kh. Gelig and V.N. Fomin. The main interests of A.Kh. Gelig were focused on analyzing the dynamics of different types of pulse-modulated systems. In this area, he developed a new approach based on the time-averaging of the pulse signal and the absolute stability theory of continuous nonlinear systems. Unlike the classical averaging method, Gelig's averaging is not asymptotic in nature and allows estimating the required sampling frequency explicitly. Classical theorems of the absolute stability theory of nonlinear systems (such as the famous circle criterion and V.M. Popov's criterion, as well as the stability criteria of periodic modes) are obtained as limiting cases when the value of the discretization period tends to zero. Therefore, the constructed theory has a high degree of unification. The corresponding cycle of works was summarized in the joint monograph by A.Kh. Gelig and A.N. Churilov, first in Russian and then in English [113] (the extended version published by Birkhauser). A.Kh. Gelig's long-term interests also included the analytical design of controllers for nonlinear systems. In contact with his many-year collaborators I.E. Zuber and A.N. Churilov, he solved various stability and stabilization problems for continuous, pulse-modulated, and discrete systems in the cases of state- and output-feedback control [114]. A.Kh. Gelig was among the pioneers investigating nonlinear dynamics of neural networks in the

USSR [115]; together with V.A. Yakubovich and G.A. Leonov, he studied the stability of systems with a nonunique equilibrium (stationary sets) [48].

V.N. Fomin began his scientific activity with the study of parametric resonance in Hamiltonian systems described by partial differential equations. Here, he managed to construct a rather complete analog of the finite-dimensional theory based on Galerkin's method and a variant of the latter's perturbation method. After defending his doctoral dissertation on this subject in 1971, V.N. Fomin's research interests shifted to the field of mathematical theory of cybernetic systems. He paid special attention to topics related to machine learning and adaptive systems, demonstrating an encyclopedic coverage of the subject. His monograph [116] and the coherent course of lectures were among the first in the country on these very important topics and painted a broad picture of the field, not limited to a single group or approach. In 1976, the book [116] was awarded the first prize of LSU in the field of scientific works. The third main direction of V.N. Fomin's work gradually gained strength: the mathematical theory of filtering and control theory, first of all, in its probabilistic variant [117–119]. Here, he obtained numerous results concerning, among others, the stochastic linear-quadratic optimal control problem, spectral factorization, and optimal estimation of random processes and fields; he developed methods for designing optimal filters when processing a packet of random plane waves against the background of distributed noise. The results of this cycle have important applications in the theory of radar and short-wave communications, underwater acoustics, radio astronomy, seismology, geophysics, and television tracking systems. The tendency to use the power of functional analysis in control theory, general for the school of V.A. Yakubovich, did not pass over V.N. Fomin. In recent years, before his untimely death, he actively and passionately developed the operator approach to filtering problems and related control problems. In particular, he succeeded in constructing a unified theory of optimal filtering, which effortlessly encompasses the Wiener–Kolmogorov theory of optimal filtering of stationary processes and the Kalman–Bucy theory of recursive filtering and, moreover, has a wide scope of applicability. Vladimir Nikolaevich's energy, charisma, and sparkling humor made him the driver of almost any event (seminar, lecture, etc.) with his participation, and the main claim of students who were lucky enough to attend his lectures was that they could never fall asleep.

In 1969, a new postgraduate—G.A. Leonov—appeared in LTC. In 1971, he defended his candidate's dissertation and continued his work in the Laboratory and at the Department. Gradually, an individual scientific direction was formed under his leadership within the traditional LTC–DTC approaches. The fundamental results on the theory of stability and synchronization of nonlinear oscillations in phase systems [48, 51, 120, 121] were followed by pioneering works and books on the theory of control and stabilization of linear controlled systems [122, 123] and the qualitative study of global attractors in dynamic systems: instability, bifurcations, synchronization, and dimension estimation [124, 125]. In 2007, G.A. Leonov became Head of the Department of Applied Cybernetics, newly established at the Faculty of Mathematics and Mechanics (SPbSU), and subsequently part of its history.

In V.A. Yakubovich and the older generation of his students, a keen interest in practical problems was naturally combined with the I. Kant's thesis that there is as much truth in each science as there is substantial mathematics in it. Among the next generation, a bright adherent of this philosophy was A.E. Barabanov, a student of V.N. Fomin. Colleagues repeatedly admired Andrei Evgen'evich's ability to apply deeply non-trivial mathematical moves in seemingly routine but important applied problems. And, more significantly, it brought success, confirming the above thesis. The range of A.E. Barabanov's interests was vast. As an illustration, let us mention important R&D works for the defense industry, the development of interference-proof dial-up modems for highly noisy switched lines (together with employees of the Department of System Programming, SPbSU), and radar signal processing systems, first for NPO Ravenstvo and then for Transas, one of the world's

largest suppliers of maritime navigation software. (According to experts, e.g., A.N. Terekhov[2], Transas was a monopolist of the onboard software market in the 1990s–2010s.) Note also systems for analyzing dolphin sound signals and systems for speech analysis and synthesis, which were developed in creative contact with the Department of Phonetics, SPbSU. On the latter topics, A.E. Barabanov prepared and delivered advanced courses of lectures. The focus of his theoretical research was on the design of optimal and suboptimal controllers [126, 127], where he obtained a series of important and sometimes unexpected results. As an example, in the 1980s he designed an optimal controller under uniformly bounded perturbations and, on this basis, constructed a new theory of L_1-optimal control. The pioneering work of A.E. Barabanov and O.N. Granichin on this topic was far ahead of similar foreign publications. Later on, O.N. Granichin (another student of V.N. Fomin) systematically developed approaches based on randomization in control systems and obtained the conditions of system operability under "almost arbitrary" (unknown but bounded) disturbances [128, 129].

At the turn of the millennium, the scientific community realized that, on the one hand, a continuous physical process interacting with a discrete (digital) control computing device is a steadily spreading combination of the future; on the other hand, the available tools of the mathematical theory are not ready, to the extent required, to deal with this combination. Its mathematical model is the hybrid dynamic system (HDS), i.e., a system described by both continuous and discrete state variables that mutually affect the evolution of each other. In the late 1990s, the interest in the mathematical modeling and theory of such systems could be characterized as a kind of boom.

Since 1997 the students of V.A. Yakubovich—A.S. Matveev and A.V. Savkin—conducted joint studies on the qualitative theory of HDSs. They laid the foundations of such a theory for a rather general class of HDSs and obtained some of the first general proof results in this field. The results were published in leading international journals, as well as in the monograph [130], probably the first in the world on this subject. The corresponding series of works focused on a general class of switched HDSs, i.e., systems for which the continuous state variables have no jumps. Among other things, the outcomes include necessary and sufficient conditions of strong determinacy of the system and invariance of a given domain, criteria for the existence and global stability of limit cycles, analogs of the classical Poincaré–Bendixson theorem, a method for designing distributed switching algorithms for processors ensuring excitation and global stability of given (optimal) oscillatory processes in large-scale flow networks, etc. The effectiveness of the general theory was demonstrated by the productive study of a number of models of information, computer, transportation, and other networks, flexible manufacturing systems, biotechnological, and other processes of independent interest.

Various aspects of the discretization problem of continuous systems in the general context of constructing the theory of HDSs were actively studied by V.A. Bondarko [131, 132]. For example, for linear time-invariant objects, he compared different discretization methods in terms of their adequacy, interconnections, and the asymptotics of properties of discrete models when increasing the frequency of time quantization. In parallel with refining the theory of finitely convergent algorithms for solving countable systems of inequalities, V.A. Bondarko established important results in the field of adaptive control, including the control of nonlinear systems and systems with an infinite-dimensional state space.

Since the mid-1990s V.A. Yakubovich and his students (with a special role of A.V. Proskurnikov among them) breathed new life into the traditional topics of the School related to linear-quadratic optimization and the frequency theorem: a cycle of about 20 papers was published on optimal damping of oscillations, optimal signal tracking, and invariance theory [133–138]. Within this cycle, a number of pioneering aspects were introduced into quite classical control theory topics, including

[2] https://www.rbc.ru/spb{_}sz/22/03/2018/5ab26f809a7947027cb81160.

the conceptualization of a "universal controller" ensuring optimality under all a priori unknown noises and tracked signals and the invariance of the system output with respect to the exogenous disturbance. The seminal paper [133] of the cycle won the Nauka/Interperiodica's award for the best publication of the year 1995; at the 1995 European Control Conference, V.A. Yakubovich was a plenary speaker on this subject [134]. In 2008, A.V. Proskurnikov was awarded the Young Scientists Medal and Prize of the Russian Academy of Sciences for for the cycle under discussion.

Starting from about the early 2000s, the phenomenon of swarm intelligence in complex network systems, as it is now called, has attracted considerable interest from physicists, mathematicians, and computer scientists in the world. Here, the main intrigue lies in how the local interactions of uninformed and low-influence elements give birth to rational and meaningful behavior of the network as a whole. The motivation for this topic is diverse and includes investigating the dynamics of ensembles of physical particles, biological populations, and opinions in social groups, artificial intelligence systems, networked control systems, etc. The subject of one of the most mathematically substantial (to date) sections of this field is distributed consensus algorithms. Significant results of the school of V.A. Yakubovich in this field, unfortunately, still poorly represented in the Russian Federation, belong to A.V. Proskurnikov (with the initial participation of A.S. Matveev); in 2022, he defended his doctoral dissertation on this topic at SPbSU. It crowns the cycle of studies, in particular, with a remarkable advance toward a complete theory of distributed averaging consensus algorithms and a productive original method of differential and recursive averaging inequalities. A.V. Proskurnikov's contribution to the related development of mathematical sociology was recognized by a joint publication in *Science* [139].

Among the examples of initiative work by V.A. Yakubovich's students, it is necessary to mention the development and promotion of a new and, to a large extent, pioneering direction lying at the junction of physics and cybernetics. This direction emerged not by chance; on the contrary, at the turn of the 1990s, there was an explosive interest in the application of cybernetics and information and control theory methods in physics. One of its triggers was the intriguing possibility, discovered in those years, to significantly change the properties of a system, e.g., to suppress or create chaos in its behavior, to change its resonance characteristics, etc., through a (theoretically arbitrarily) small impact. The books [140, 141] published in 1998 and 1999 were the first monographs in the world in this direction, and the field of sciences on the border of cybernetics and physics was named cybernetical physics (cyber-physics) by their coauthor A.L. Fradkov. In particular, it includes the control of molecular and quantum systems (A.L. Fradkov, M.S. Anan'evskii), which play an important role in the creation of promising nanotechnologies. The paper [142] reviewing research works on chaos control[3] won the Nauka/Interperiodica's award for the best publication of the year 2003. The basic principles of cyber-physics were described in the books [42, 143]. Signs of its international recognition are the world's first international conferences on physics and control held in St. Petersburg (2003–2005), as well as the International Physics and Control Society (IPACS) established with headquarters in St. Petersburg. *Cybernetics and Physics*, an international journal indexed in Scopus, is published in St. Petersburg under the auspices of this society.

Another important direction, perhaps decisive for cybernetics itself, reflects the convergence trend of the theories of control, computation, and communication toward their unity, which took shape at the turn of the millennium. More and more problems require close interaction of the methods of these three theories; even the aphoristic formula Control×Computation×Communication= C^3 has appeared, expressing the aspiration to return the holistic perception of information, computational, and control processes, which meant so much for the successes of the "romantic" cybernetics of the 1960s. Some pioneering results in this direction were obtained in the early 2000s by

[3] The survey is the most cited article of *Avtomatika i Telemekhanika*, and its second co-author is the most cited author of the journal.

A.S. Matveev together with A.V. Savkin, a DTC alumnus and Professor of the University of New South Wales (Australia). The corresponding cycle of works was devoted to control and estimation under the capacity constraints of communication channels [144–147] and was partially summarized in the monograph [148]. In particular, it was demonstrated that an unstable linear controlled system can be stabilized if and only if the bit rate of information arrival through the communication channel exceeds the rate of information production by the system; also, a fundamental advance was made to determine the place of the basic concepts of C. Shannon's information theory (in particular, the capacity of a noisy communication channel) in the discussed topic. Subsequently, the research switched to nonlinear systems and was carried out by A.S. Matveev in coauthorship with Professor A.Yu. Pogromsky (the Eindhoven University of Technology) using the nonlinear dynamics analysis methods of A.M. Lyapunov and G.A. Leonov. In particular, a new concept of the restoration entropy of a nonlinear system was developed; in co-authorship with C. Kawan (Ludwig-Maximilians-Universitä München) it was shown that, in a certain sense and in the questions under consideration, this entropy adequately characterizes the rate of information production by the system [149, 150]. Sufficient conditions for the operability of nonlinear and adaptive systems under communication constraints were also obtained by A.L. Fradkov et al. [151–156]. Some results obtained by the team were overviewed in [157].

With all the conventionality of any rubric, it has become a kind of tradition to divide modern robotics into two sections, industrial and mobile robotics. The first (and more developed) section focuses on the orchestra of industrial systems, in which manipulation systems (mechanical arms) play first fiddle. At present, the vast majority of such systems follow the hold-down arm paradigm with a fixed operational location. At the same time, practical tasks are systematically introduced into the agenda where the soft (non-hold-down arm) approach is needed and/or the manipulation object is malleable, and mobile and manipulation functions interact operatively. Such tasks now fall into an almost unexplored field; its development requires solving a number of fundamental problems, including theoretical ones. A group of DTC graduates (A.S. Shiryaev and S.V. Gusev) has been systematically working in this direction since the 2010s. Its asset is the development of largely pioneering mathematical methods of dynamics analysis and controller design for solving the corresponding problems, in particular, the method of moving Poincaré sections and transverse linearization, high-speed methods for solving special matrix Riccati differential equations, general methods for finding periodic motions implemented in complex under-actuated mechanical systems, and other results [158–160]. The effectiveness of these R&D results was demonstrated in 2015 by the world's first experimentally validated solution of a complex prototypical problem posed in 1998 by C. Lynch: stabilize the circular motion of a ball on a rotating butterfly-shaped guide [160].[4]

The mobile robotics section concentrates on the autonomous navigation of mobile robots and their motion control in a priori unknown environments with obstacles. This direction has been systematically developed since the 2010s by the mobile robotics group of DTC (A.S. Matveev, A.A. Semakova, and P.A. Konovalov) with the participation of A.V. Savkin (until 2017). A number of fundamental results on robot navigation algorithms in complex (particularly moving and unpredictable) environments, including distributed control of their multi-agent ensembles, were obtained here. They were partially systematized in the two monographs [161, 162], released in 2015 and 2016 by the world's leading academic publishers. The specifics of the group's R&D works are resource-saving algorithms (in terms of computations, energy, sensory data about the environment, etc.) that convert current observation into current control in a reflex-like manner (as a consequence, with minimal requirements for onboard processors) and are nevertheless provided with mathematically rigorous guarantees of achieving the result. According to the WoS data for the year 2022,

[4] https://www.youtube.com/watch?v=kyvW5sOcZHU.

of the five most cited publications on robotics affiliated with Russia, four are related to DTC, including the most cited paper [163] (279 citations).

Mathematical methods have long been used for the quantitative and qualitative study of processes and systems, to a greater or lesser extent related to the field of biology and medicine. In this direction, in the early 2000s A.S. Matveev and A.V. Savkin investigated optimal protocols for chemotherapeutic treatment of cancer [164]. Starting from the mid-2000s, DTC (A.N. Churilov and A.I. Shepeljavyi) together with Uppsala University (Sweden) conducted systematic studies on modeling and analysis of biological rhythms and chaotic dynamics in neurohormonal systems [165, 166]. Since the 2010s the scientific directions of the School include neural control and neurofeedback based on the mathematical study of networks of biological neurons. These R&D works lie at the junction of cybernetics and neuroscience; here, the world expects breakthroughs in medical diagnosis, as well as in the control of robots and other devices with the power of thought (without human muscles). At present, under the guidance of A.L. Fradkov, a grant-supported project is being implemented on this topic at SPbSU. The corresponding works are being carried out jointly with the Higher Nervous Activity and Psychophysiology Department of SPbSU, the Institute of Human Brain (the Russian Academy of Sciences), Institute for Problems of Mechanical Engineering (the Russian Academy of Sciences), and Immanuel Kant Baltic Federal University. M. Lipkovich and S.A. Plotnikov, young representatives of the School, actively participate in the project.

Representatives of the School have been teaching at various universities of the country. In St. Petersburg, let us note the following persons (currently active or passed away): G.A. Leonov, Dean of the Faculty of Mathematics and Mechanics (SPbSU), USSR State Prize Laureate, Corresponding Member of the Russian Academy of Sciences; N.V. Kuznetsov, Head of the Department of Applied Cybernetics (SPbSU), Corresponding Member of the Russian Academy of Sciences; O.N. Granichin, Professor of the Department of System Programming (SPbSU); Professors A.V. Timofeev (St. Petersburg State University of Aerospace Instrumentation), A.N. Churilov (St. Petersburg State Marine Technical University), V.B. Smirnova (St. Petersburg State University of Architecture and Civil Engineering), N.E. Barabanov (St. Petersburg Electrotechnical University "LETI"); Heads of laboratories of academic institutes A.V. Timofeev (St. Petersburg Institute for Informatics and Automation, the Russian Academy of Sciences) and A.L. Fradkov (Institute for Problems of Mechanical Engineering, the Russian Academy of Sciences). In the 1970s and 1990s, several talented graduates of the Department left the country, B.G. Pittel, M.V. Levit, and B.D. Lyubachevskii were among them. Some of them became professors at foreign universities: A. Megretski (Massachusetts Institute of Technology, USA), N. Barabanov (North Dakota State University, USA), A. Savkin (University of New South Wales, Australia), A. Shiriaev (Umeå University, Sweden, and Norwegian University of Science and Technology, Trondheim, Norway)

A significant place in the School's activities is occupied by scientific and organizational work. For example, since 1967 V.A. Yakubovich was Deputy Chairman (Deputy of A.A. Vavilov, Rector of Leningrad Electrotechnical University) and part-time Chairman of the Section for the Theory of Adaptive Control Systems, the Leningrad Territorial Group of the National Committee on Automatic Control. A series of six Leningrad (St. Petersburg) symposia and one All-Union Conference on the Theory of Adaptive Systems, held on the initiative of V.A. Yakubovich and under his guidance from 1972 to 1999, occupied a notable place in the scientific and organizational landscape of the country. This series was another sign recognizing the School's merits in the field of adaptive systems, and its events were important milestones in the development of the field. In those years, it was one of the main growth points of mathematical control theory and cybernetics and attracted the interest of talented young people and venerable researchers: the number of papers and participants usually numbered in the hundreds. The symposia were attended by leaders of domestic and, since the 1990s, foreign science. Note the following persons among

them: Academicians Ya.Z Tsypkin, A.A. Krasovskii, E.P. Popov, and N.N. Moiseev; Doctors of Science D.A. Pospelov, V.Yu. Rutkovskii, Yu.I. Neimark, A.A. Pervozvanskii, and R.M. Yusupov; G. Bartolini (Italy), S. Bittanti (Italy), V. Răsvan (Romania), A. Halanay (Romania), L. Ljung (Sweden), J. Lando (France), A. Lindqvist (Sweden), D. Šiljak (USA), K. Furuta (Japan), and others. In 1972, a plenary report was delivered by M.M. Botvinnik, Doctor of Engineering, former world chess champion; he spoke about the development of a computer algorithm for chess play. Initially, the scientific secretary of the series of events was D.P. Derevitskii, Associate Professor of the Department of Automatic Control Systems (Leningrad Mechanical Institute); later, he was replaced by A.L. Fradkov. In scientific and organizational activities, DTC traditionally and closely cooperates with the Laboratory of Complex Systems Control (Institute for Problems of Mechanical Engineering, the Russian Academy of Sciences). It was established in 1990 by A.L. Fradkov, the first and present-day head. This laboratory is closely connected with the Department both in research interests and in education.

The team conducts career-oriented work with young people in the field of cybernetics. In 1999, a group of experts in automation and control systems from several universities of the city proposed to organize school olympiads in cybernetics. The idea was supported by V.P. Tarasov, Head of the department of science and technology at St. Petersburg City Palace of Youth Creativity (the Anichkov Palace), and things got rolling: 14 Olympiads were held in 1999–2013. M.S. Anan'evskii, A.L. Fradkov, and A.S. Matveev, representatives of the School, took an active part in their organization and holding from the very beginning. The materials of the Olympiads and some methodological conclusions were summarized in a series of proceedings published largely owing to the work and energy of M.S. Anan'evskii.

In 2008, a cybernetics club was organized for junior students of DTC based on LEGO Mindstorms NXT. While learning control theory, students had an opportunity to implement control algorithms on physical objects and to connect their theoretical knowledge with practice. In the class, students independently developed original designs such as a bicycle robot, a segway, a crawling robot, a predator robot, and others. The best works were presented at the Robot Show during the Week of the Faculty of Mathematics and Mechanics (SPbSU). At the same time, creative cooperation began with the Robotics Center of Presidential Physics and Mathematics Lyceum (PPML) No. 239, headed by S.A. Filippov. The results of cooperation were presented at several international conferences [167, 168]. The enthusiasm of R.M. Luchin, a DTC member and teacher, played an important role in organizing and leading the club. Robot soccer became one direction of his work. The first city competitions of radio-controlled robots were held in 2012; a year later autonomous robots were already on the field.

Experience with LEGO led a group of enthusiasts (R.M. Luchin, S.A. Filippov, and A.N. Terekhov) to the idea of developing their own constructor set, more advanced than LEGO. Cybernetic Technologies LLC was founded, where the Universal Cybernetic Constructor TRIK and the necessary software were developed, allowing to implement various projects, from basic educational to modern research projects. They are used in Russian schools and universities. According to the 2018 annual analytical review of the global robotics market by Sberbank's robotics laboratory, the company was mentioned as one of Russia's few unconditional successes in this market so far. Unfortunately, R.M. Luchin passed away prematurely at a young age due to the COVID-19 pandemic. His work is being continued by his student, I.Yu Shirokolobov, an employee of DTC. In 2019, URoboRus, the jointly created team of robotic soccer players, was the first Russian team to qualify for the RoboCup SSL, a kind of world championship. In 2020, URoboRus qualified again, but the competition was canceled due to the pandemic. In 2021, the competition was carried out online, and URoboRus managed to participate in the playoff for the first time, ranking first in the group. The year 2022 was remarkable for another successful qualification for the RoboCup

SSL World Championship and the first full-time participation. The event took place at the FEI University in Sao Paolo (Brazil) during RoboCup Brazil Open, the Brazilian Open Championship.

Thanks to the efforts of the DTC team, as well as the support of PPML No. 239, the Scientific and Educational Center (SEC) "Mathematical Robotics and Artificial Intelligence" was established at SPbSU in 2019. Since its foundation, K.S. Amelin (a student of O.N. Granichin) and A.L. Fradkov are Director and Scientific Supervisor of SEC, respectively. The Center is intended to integrate the efforts of SPbSU in fundamental research on mathematical and educational robotics and intelligent control. The directions of the Center's work include the issues of navigation of mobile robots and their multi-agent network ensembles, control of underactuated manipulators, computer vision, artificial intelligence, machine learning and big data processing, neural network control, methods and tools for programming and debugging robots, and educational and practical robotics. In 2022, the experimental park of SEC contained Geoscan Pioneer quadcopters and TRIK universal cybernetic constructor kits. With the active participation of SPbSU students and the use of this park, SEC has already implemented several applied projects, in particular, forrest inventory by robotic quadcopters, search for a person lost in the forest, semi-automatic dropping of GPS beacons on glaciers to monitor their movement, automatic overflight of protected areas, control of bridge piers, increase of the data transmission rate in large wireless networks, etc.

Additional information about the department is available in thematic issues of Russian and international journals [169–171] dedicated to the anniversaries of DTC employees and in the collection of articles [3]. The scientific product of the School counts many hundreds of publications, including over 60 books. V.A. Yakubovich's nestlings work fruitfully in many Russian and foreign research centers and universities; they have defended over 100 dissertations on physics, mathematics, and engineering, including 19 doctoral dissertations.

The influence of the Department and School's achievements is noticeable in the distribution of university places in world rankings. For example, according to the Shanghai Academic Ranking of World Universities (ARWU), SPbSU ranked 32nd in the direction "Automation and Control" in 2018. The number of publications by DTC employees in top journals on automation and control, taken into account in the ARWU ranking, approximately equals 28% (24 out of 85) of all Russian publications in such journals for 2012–2016.

Representatives of the School have been repeatedly given prestigious Russian and international awards and titles. In 1998, V.A. Yakubovich became Honored Scientist of the Russian Federation; in 2005, he was awarded the Order of Honor. V.A. Yakubovich was Member of the Russian Academy of Sciences and Academician of the Russian Academy of Natural Sciences. In 2006 he was elected Honorary Professor of SPbSU. A.L. Fradkov was awarded the international honorary titles of IFAC Fellow, IEEE Life Fellow, and AAIA Fellow. In 2015, DTC alumnus (1998) Alexey Pavlov and colleagues from the Eindhoven University of Technology received the prestigious IEEE Control Systems Technology award. In 2020, DTC alumnus A.V. Proskurnikov and coauthors were awarded the IFAC and Elsevier paper prize award for the best paper published in *Annual Reviews in Control* in 2017–2020. (Proskurnikov, A.V. and Tempo, R., A Tutorial on Modeling and Analysis of Dynamic Social Networks. Part I, *Annual Reviews in Control*, 2017, vol. 43, pp. 65–79.) The same award for the best paper of 2020–2022 was given for the survey [76]. In 2018, A.L. Fradkov was awarded the Andronov Prize of the Russian Academy of Sciences for the series of works on synchronization and control of nonlinear oscillations (together with I.I. Blekhman).

After the demise of Vladimir Andreyevich Yakubovich in 2012, the founder and long-term head of DTC, the founder of the scientific school of cybernetics and artificial intelligence in St. Petersburg, the Department was successively headed by his closest colleagues and students, A.Kh. Gelig, A.L. Fradkov, and A.S. Matveev (since 2021 until present). Several thematic collections, publications, and speeches have been devoted to the creative biography and scientific achievements of

V.A. Yakubovich, as well as the 1st International IFAC Conference on Modelling, Identification and Control of Nonlinear Systems (MICNON 2015) [172–174]. The DTC staff prepared and published a CD-ROM containing over 300 main works of V.A. Yakubovich.

ACKNOWLEDGMENTS

The authors are grateful to K.S. Amelin, V.A. Bondarko, S.V. Gusev, P.A. Konovalov, A.N. Churilov, and I.Yu. Shirokolobov for their materials and memories.

FUNDING

This work was supported in part by the Ministry of Science and Higher Education of the Russian Federation, agreement no. 075-15-2021-573.

REFERENCES

1. Shepelyavyi, A.I., The Department of Theoretical Cybernetics at the Faculty of Mathematics and Mechanics, St. Peterburg State University, *Vest. St. Peterburg Univ.*, 2000, vol. 1, no. 1, pp. 3–15.
2. Fradkov, A.L., V.A. Yakubovich's Scientific School on Theoretical Cybernetics at St. Petersburg (Leningrad) University, in *Istoriya informatiki i kibernetiki v Sankt-Peterburge (Leningrade)* (The History of Information Science and Cybernetics in St. Petersburg (Leningrad)), Yusupov, R.M., Ed., St. Petersburg, 2008, pp. 79–83.
3. *Nelineinye sistemy. Chastotnye i matrichnye neravenstva. K 80-letiyu so dnya rozhdeniya V.A. Yakubovicha* (Nonlinear Systems. Frequency Domain and Matrix Inequalities. Dedicated to the 80th Anniversary of V.A. Yakubovich's Birth), Gelig, A.Kh., Leonov, G.A., and Fradkov, A.L., Eds., Moscow: Fizmatlit, 2008.
4. Sobolev, S.L., Kitov, A.I., and Lyapunov, A.A., Basic Features of Cybernetics, *Voprosy Filosofii*, 1955, no. 4, pp. 136–148.
5. Kolmogorov, A.N., Cybernetics, in *Bol'shaya Sovetskaya Entsiklopediya. Tom 51* (The Great Soviet Encyclopedia, Vol. 51), 2nd. ed., Moscow: Bol'shaya Sovetskaya Entsiklopediya, 1958, pp. 149–151.
6. Wiener, N., *Cybernetics: or Control and Communication in the Animal and the Machine*, Cambridge: MIT Press, 1948.
7. Yakubovich, V.A., Machines Learning Pattern Recognition, *Sb. Vych. Tsentr. Leningrad. Gos. Univ.*, 1963, no. 2, pp. 95–131.
8. Yakubovich, V.A., Some General Principles to Construct Learning Recognizing Systems. I, in *Vychislitel'naya tekhnika i voprosy programmirovaniya* (Computer Engineering and Programming Issues), Leningrad: Leningrad State University, 1965, pp. 3–71.
9. Yakubovich, V.A., Three Theoretical Schemes of Learning Systems, in *Samoobuchayushchiesya avtomaticheskie sistemy* (Self-learning Automatic Systems), Moscow: Nauka, 1956, pp. 21–28.
10. Yakubovich, V.A., Machines Learning Pattern Recognition. I, II, *Vestnik of Saint Petersburg University. Mathematics. Mechanics. Astronomy*, 2021, vol. 8, no. 4, pp. 625–638; 2022, vol. 9, no. 1, pp. 94–112.
11. Kozinets, B.N., Lantsman, R.M., and Yakubovich, V.A., The Use of Electron Computers in Criminalistics for Differentiation between Very Similar Handwritings, *Dokl. Akad. Nauk SSSR*, 1966, vol. 167, no. 5, pp. 1008–1011.
12. Gelig, A.Kh. and Yakubovich, V.A., Application of Learning Recognizing Systems to Signal Extraction from Noise, in *Vychislitel'naya tekhnika i voprosy programmirovaniya* (Computer Engineering and Programming Issues), Leningrad: Leningrad State University, 1968, pp. 95–100.
13. Kharichev, V.V., Shmidt, A.A., and Yakubovich, V.Ya., A New Problem in Pattern Recognition, *Autom. Remote Control*, 1973, vol. 34, no. 1, pp. 98–109.

14. Kozinets, B.N., Lantsman, R.M., and Yakubovich, V.A., To the Problem of Recognition and Description of Complex Images, *Proc. of Intern. IFAC. Symposium*, Tbilisi, 1975, pp. 207–250.

15. Fomin, V.N., *Matematicheskaya teoriya obuchayushchikhsya opoznayushchikh sistem* (Mathematical Theory of Learning Recognizing Systems), Leningrad: Leningrad State University, 1976.

16. Gelig, A.Kh. and Matveev, A.S., *Vvedenie v matematicheskuyu teoriyu obuchaemykh raspoznayushchikh sistem i neironnykh setei* (Introduction to Mathematical Theory of Learning Recognizing Systems and Neural Networks), St. Petersburg: St. Peterburg State University, 2014.

17. Yakubovich, V.A., Recurrent Finitely Convergent Algorithms for Solving Systems of Inequalities, *Soviet Mathematics*, 1966, vol. 7, pp. 300–304.

18. Yakubovich, V.A., The Method of Recursive Aim Inequalities in the Theory of Adaptive Systems, in *Voprosy kibernetiki: Adaptivnye sistemy* (Cybernetics Issues: Adaptive Systems), Moscow–Leningrad: the USSR Academy of Sciences, Scientific Council on the Complex Problem of Cybernetics, 1976, pp. 32–64.

19. Tsypkin, Ya.Z., Adaptation, Learning and Self-learning in Automatic Systems, *Avtomat. i Telemekh.*, 1966, no. 1, pp. 23–61.

20. Tsypkin, Ya.Z., *Adaptation and Learning in Automatic Systems*, Academic Press, 1971.

21. Yakubovich, V.A., Theory of Adaptive Systems, *Dokl. Akad. Nauk SSSR*, 1968, vol. 182, no. 3, pp. 518–521.

22. Yakubovich, V.A., Adaptive Systems with Multistep Goal Conditions, *Dokl. Akad. Nauk SSSR*, 1968, vol. 183, no. 2, pp. 303–306.

23. Fomin, V.N., Fradkov, A.L., and Yakubovich, V.A., *Adaptivnoe upravlenie dinamicheskimi ob"ektami* (Adaptive Control of Dynamic Objects), Moscow: Nauka, 1981.

24. Yakubovich, V.A., The Method of Recursive Aim Inequalities in Adaptive Control, in *Spravochnik po teorii avtomaticheskogo upravleniya* (Handbook of Automatic Control Theory), Moscow: Nauka, 1987, ch. 5, pp. 501–526.

25. Bondarko, V.A. and Yakubovich, V.A., The Method of Recursive Aim Inequalities in Adaptive Control Theory, *Int. J. Adaptive Control and Signal Proc.*, 1992, vol. 6, pp. 141–160.

26. Bondarko, V.A., Adaptive Suboptimal Systems with a Variable Dimension of the Vector of Adjustable Parameters, *Autom. Remote Control*, 2006, vol. 67, no. 11, pp. 1732–1751.

27. Timofeev, A.V., *Roboty i iskusstvennyi intellekt* (Robots and Artificial Intelligence), Moscow: Nauka, 1978.

28. Timofeev, A.V., *Adaptivnye robototekhnicheskie kompleksy* (Adaptive Robotic Complexes), Moscow: Mashinostroenie, 1988.

29. Timofeev, A.V., *Upravlenie robotami* (Robot Control), St. Petersburg: St. Petersburg State University, 1986.

30. Gusev, S.V., Timofeev, A.V., and Yakubovich, V.A., On a Hierarchical System of Integral Robot Control, in *Proc. of the 4th International Joint Conference on Artificial Intelligence*, Moscow, 1975, vol. 9, pp. 53–61.

31. Gusev, S.V. and Yakubovich, V.A., An Algorithm for Adaptive Control of a Manipulator Robot, *Autom. Remote Control*, 1981, vol. 41, no. 9, pp. 1268–1277.

32. Yakubovich, V.A., On Certain Problem of Self-learning Expedient Behaviour, *Autom. Remote Control*, 1969, vol. 30, no. 8, pp. 1292–1310.

33. Lyubachevskii, B.D. and Yakubovich, V.A., Adaptive Control of Stable Dynamic Plants, *Autom. Remote Control*, 1974, vol. 35, no. 4, pp. 621–631.

34. Yakubovich, V.A., On the "Brain" Organization of Adaptive Systems with a One-Step Target Condition, in *Problemy bioniki* (Problems of Bionics), Moscow: Nauka, 1973, pp. 355–360.

35. Yakubovich, V.A., On a Method of Adaptive Control under Conditions of Great Uncertainty, *Preprints of the 5th IFAC World Congress*, Paris, 1972, vol. 37, no. 3, pp. 1–6.

36. Grigor'ev, G.G., Gusev, S.V., Nesterov, V.V., and Yakubovich, V.A., Mobile Robot-Manipulator Adaptive Control, *Proc. of the Soviet Conference on Adaptive Robots*, Moscow, 1982, pp. 89–91.

37. Belenkov, B.A., Gusev, S.V., Zotov, Yu.K., Ruzhanskii, V.I., Timofeev, A.V., Frolov, R.B., and Yakubovich, V.A., An Adaptive Control System for an Autonomous Mobile Robot, *Izv. Akad. Nauk SSSR. Tekhn. Kibern.*, 1978, no. 6, pp. 52–63.

38. Timofeev, A.V. and Yakubovich, V.A., Adaptive Control of Programmed Motion of a Manipulator Robot, in *Voprosy kibernetiki: Adaptivnye sistemy* (Cybernetics Issues: Adaptive Systems), Moscow–Leningrad: the USSR Academy of Sciences, Scientific Council on the Complex Problem of Cybernetics, 1976, pp. 170–174.

39. Gelig, A.Kh., An Adaptive Control System for an Eye–Arm Robot, in *Voprosy kibernetiki: Adaptivnye sistemy* (Cybernetics Issues: Adaptive Systems), Moscow–Leningrad: the USSR Academy of Sciences, Scientific Council on the Complex Problem of Cybernetics, 1976, pp. 162–163.

40. Aksenov, G.S. and Fomin, V.N., To the Problem of Adaptive Control of a Manipulator, in *Voprosy kibernetiki: Adaptivnye sistemy* (Cybernetics Issues: Adaptive Systems), Moscow–Leningrad: the USSR Academy of Sciences, Scientific Council on the Complex Problem of Cybernetics, 1976, pp. 165–168.

41. Fradkov, A.L., A Scheme of Speed Gradient and Its Application in Problems of Adaptive Control, *Autom. Remote Control*, 1980, vol. 40, no. 9, pp. 1333–1342.

42. Fradkov, A.L., *Kiberneticheskaya fizika: printsipy i primery* (Cybernetic Physics: Principles and Examples), St. Petersburg: Nauka, 2003.

43. Fradkov, A.L. and Shalymov, D.S., Speed Gradient and MaxEnt Principles for Shannon and Tsallis Entropies, *Entropy*, 2015, vol. 17, no. 3, pp. 1090–1102.

44. Granovskaya, R.M. and Bereznaya, I.Y., Experiments on Human Pattern Recognition: a Hierarchical Sign-System Approach, *Pattern Recognition*, 1980, vol. 12, no. 1, pp. 17–26.

45. Granovskaya, R.M. and Bereznaya, I.Y., Consciousness as the Unity of Higher Psychic Processes, *Kybernetes*, 1988, vol. 17, no. 2, pp. 35–43.

46. Yakubovich, V.A. and Starzhinskii, V.M., *Lineinye differentsial'nye uravneniya s periodicheskimi koeffitsientami i ikh prilozheniya* (Linear Differential Equations with Periodic Coefficients and Their Applications), Moscow: Nauka, 1972.

47. Yakubovich, V.A., The Solution of Some Matrix Inequalities Encountered in Automatic Control Theory, *Dokl. Akad. Nauk SSSR*, 1962, vol. 143, no. 6, pp. 1304–1307.

48. Yakubovich, V.A., Leonov, G.A., and Gelig, A.Kh., *Stability of Stationary Sets in Control Systems with Discontinuous Nonlinearities*, Singapore: World Scientific, 2004.

49. Yakubovich, V.A., A Frequency Theorem in Control Theory, *Siberian Math. J.*, 1973, vol. 14, no. 2, pp. 265–289.

50. Likhtarnikov, A.L. and Yakubovich, V.A., A Frequency Theorem for Equations of Evolution Type, *Siberian Math. J.*, 1976, vol. 17, no. 5, pp. 790–803.

51. Leonov, G.A., Burkin, I.M., and Shepeljavyi, A.I., *Frequency Methods in Oscillations Theory*, Dortrecht: Kluwer, 1996.

52. Levit, M.V., A Frequency Criterion for the Absolute Stochastic Stability of Nonlinear Systems of Ito Differential Equations, *Uspekhi Mat. Nauk*, 1972, vol. 27, no. 4(166), pp. 215–216.

53. Levit, M.V. and Iakubovich, V.A., Algebraic Criterion for Stochastic Stability of Linear Systems with Parametric Action of the White Noise Type, *Journal of Applied Mathematics and Mechanics*, 1972, vol. 36, no. 1, pp. 130–136.

54. Antonov, V.G., Likhtarnikov, A.L., and Yakubovich, V.A., A Discrete Frequency Theorem for the Case of Hilbert Spaces of States and Controls. I, *Vest. Leningr. Univ. Mat.*, 1980, vol. 8, pp. 1–11.

55. Fradkov, A.L., Synthesis of Adaptive System of Stabilization for Linear Dynamic Plants, *Autom. Remote Control*, 1974, vol. 35, pp. 1960–1966.

56. Fradkov, A.L., Quadratic Lyapunov Functions in Adaptive Stabilization Problem of a Linear Dynamic Plant, *Sib. Math. J.*, 1976, vol. 17, no. 2, pp. 341–348.

57. Bondarko, V.A., Likhtarnikov, A.L., and Fradkov, A.L., Design of an Adaptive System for Stabilizing a Linear Object with Distributed Parameters, *Autom. Remote Control*, 1979, vol. 40, no. 12, pp. 1785–1792.

58. Lihtarnikov, A.L. and Jakubovic, V.A., The Frequency Theorem for Continuous One-Parameter Semigroups, *Izvestiya: Mathematics*, 1977, vol. 11, no. 4, pp. 849–864.

59. Gusev, S.V., Kalman–Popov–Yakubovich Lemma for Ordered Fields, *Autom. Remote Control*, 2014, vol. 75, no. 1, pp. 18–33.

60. Pakshin, P.V., Stability of One Class of Nonlinear Stochastic Systems, *Autom. Remote Control*, 1977, vol. 38, no. 4, pp. 474–481.

61. Brusin, V.A., Global Stability and Dichotomy of a Class of Nonlinear Systems with Random Parameters, *Siberian Math. J.*, 1981, vol. 22, no. 2, pp. 210–222.

62. Brusin, V.A. and Ugrinovskii, V.A., Investigation of Stochastic Stability of a Class of Nonlinear Differential Equations of Ito Type, *Siberian Math. J.*, 1987, vol. 28, no. 3, pp. 381–393.

63. Ugrinovskii, V.A., A Stochastic Analogue of the Frequency Theorem, *Soviet Math. (Iz. VUZ)*, 1987, vol. 31, no. 10, pp. 47–55.

64. Brusin, V.A. and Ugrinovskii, V.A., Absolute Stability Approach to Stochastic Stability of Infinite-Dimensional Nonlinear Systems, *Automatica*, 1995, vol. 31, no. 10, pp. 1453–1458.

65. Barabanov, N.E., Gelig, A.Kh., Leonov, G.A., Likhtarnikov, A.L., Matveev, A.S., Smirnova, V.B., and Fradkov, A.L., The Frequency Theorem (Kalman–Yakubovich Lemma) in Control Theory, *Autom. Remote Control*, 1996, vol. 57, no. 10, pp. 1377–1407.

66. Gusev, S.V. and Likhtarnikov, A.L., Kalman–Popov–Yakubovich Lemma and the S-procedure: A Historical Essay, *Autom. Remote Control*, 2006, vol. 67, no. 11, pp. 1768–1810.

67. Seron, M.M., Hill, D.J., and Fradkov, A.L., Adaptive Passification of Nonlinear Systems, *Proc. 33rd IEEE Conf. Dec. Contr.*, 1994, pp. 190–195.

68. Jiang, Z.P., Hill, D.J., and Fradkov, A.L., A Passification Approach to Adaptive Nonlinear Stabilization, *Syst. Control. Lett.*, 1996, vol. 28, pp. 73–84.

69. Fradkov, A.L., Passification of Nonsquare Linear Systems and Feedback Yakubovich–Kalman–Popov Lemma, *Europ. J. Contr.*, 2003, no. 6, pp. 573–582.

70. Andrievskii, B.R. and Fradkov, A.L., Method of Passification in Adaptive Control, Estimation, and Synchronization, *Autom. Remote Control*, 2006, vol. 67, no. 11, pp. 1699–1731.

71. Xie, L.H., Fu, M.Y., and Li, H.Z., Passivity Analysis and Passification for Uncertain Signal Processing Systems, *IEEE Transactions on Signal Processing*, 1998, vol. 46, no. 9, pp. 2394–2403.

72. Mahmoud, M.S. and Ismail, A., Passivity and Passification of Time-Delay Systems, *Journal of Mathematical Analysis and Applications*, 2004, vol. 292, no. 1, pp. 247–258.

73. Xia, M., Rahnama, A., Wang, A., and Antsaklis, P.J., Control Design Using Passivation for Stability and Performance, *IEEE Transactions on Automatic Control*, 2018, vol. 63, no. 9, pp. 2987–2993.

74. Pyrkin, A.A., Aranovskiy, S.V., Bobtsov, A.A., Kolyubin, S.A., and Nikolaev, N.A., Fradkov Theorem-Based Control of MIMO Nonlinear Lurie Systems, *Autom. Remote Control*, 2018, vol. 79, no. 6, pp. 1074–1085.

75. Tomashevich, S. and Belyavskyi, A., Passification Based Simple Adaptive Control of Quadrotor, *IFAC-PapersOnLine*, 2016, vol. 49, no. 13, pp. 281–286.

76. Annaswamy, A.M. and Fradkov, A.L., A Historical Perspective of Adaptive Control and Learning, *Annual Reviews in Control*, 2021, no. 52, pp. 18–41.

77. Boyd, S., Ghaoui, L.E., Feron, E., and Balakrishnan, A.V., *Linear Matrix Inequalities in Systems and Control Theory*, Philadelphia: SIAM, 1994.

78. Kalman, R.E., Contributions to the Theory of Optimal Control, *Boletín de la Sociedad Matemática Mexicana*, 1960, vol. 2, no. 2, pp. 102–119.

79. Krasovskii, N.N., The Problem of Stabilization of Controlled Motion, in *Teoriya ustoichivosti dvizheniya* (The Theory of Stability of Motion), Malkin, I., Ed., app. 4, Moscow: Nauka, 1976.

80. Letov, A.M., Analytical Design of Controllers. I, II, *Avtomat. i Telemekh.*, 1960, vol. 21, no. 4, pp. 436–441; no. 5, pp. 561–568.

81. Kolmogorov, A.N., Interpolation and Extrapolation of Stationary Random Sequences, *Izv. Akad. Nauk SSSR Ser. Mat.*, 1941, vol. 5, no. 1, pp. 3–14.

82. Wiener, N., *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, Cambridge, 1949.

83. Busy, R.S. and Joseph, P.D., *Filtering of Stochastic Processes with Application to Guidance*, New York–London, 1968.

84. Trentelman, H., Linear Quadratic Optimal Control, in *Encyclopedia of Systems and Control*, Baillieul, J. and Samad, T., Eds., London: Springer, 2013.

85. Megretski, A.V. and Yakubovich, V.A., Singular Stationary Nonhomogeneous Linear-Quadratic Optimal Control, *Transactions of the American Mathematical Society*, 1993, vol. 155, pp. 129–167.

86. Yakubovich, V.A., Minimization of Quadratic Functionals under Quadratic Constraints and the Necessity of a Frequency Condition in the Quadratic Criterion for Absolute Stability of Nonlinear Control Systems, *Dokl. Akad. Nauk SSSR*, 1973, vol. 209, no. 5, pp. 1039–1042.

87. Megretsky, A., Necessary and Suffient Conditions of Stability: A Multiloop Generalization of the Circle Criterion, *IEEE Transactions on Automatic Control*, 1993, vol. AC-38, no. 5, pp. 753–756.

88. Savkin, A.V. and Petersen, I.R., Minimax Optimal Control of Uncertain Systems with Structured Uncertainty, *Int. J. of Robust and Nonlinear Control*, 1995, vol. 5, pp. 119–138.

89. Morari, M., Some Control Problems in the Process Industries, in *Essays on Control: Perspectives in Theory and Applications*, Trentelman, H.L. and Willems, J.C., Eds., *Progress in System and Control Theory*, 1993, vol. 14, pp. 55–77.

90. Aizerman, M.A. and Gantmakher, F.R., *Absolyutnaya ustoichivost' reguliruemykh sistem* (Absolute Stability of Regulated Systems), Moscow: the USSR Academy of Sciences, 1963.

91. Gantmakher, F.R. and Yakubovich, V.A., Absolute Stability of Nonlinear Controlled Systems, *Trudy 2-go Vsesoyuznogo s"ezda po teoreticheskoi i prikladnoi mekhanike* (Proc. of the 2nd All-Union Congress on Theoretical and Applied Mechanics), Moscow: Nauka, 1965.

92. Yakubovich, V.A., The *S*-procedure in Nonlinear Control Theory, *Vest. Leningrad. Gos. Univ. Ser. Mat. Mekh. Astron.*, 1971, no. 1, pp. 62–77.

93. Fradkov, A.L. and Yakubovich, V.A., The *S*-procedure and Duality Relation in Nonconvex Quadratic Programming Problems, *Vest. Leningrad. Gos. Univ.*, 1973, no. 1, pp. 71–76.

94. Halmos, P.R., *A Hilbert Space Problem Book*, Princeton–New Jersey–Toronto–London: D. Van Nostrand Company, 1982.

95. Megretsky, A., Treil, S., and Fradkov, A.L., Power Distribution Inequalities in Optimization and Robustness of Uncertain Systems, *J. Math. Systems, Estimation, Control*, 1993, vol. 3, no. 3, pp. 301–319.

96. Matveev, A.S. and Yakubovich, V.A., Nonconvex Problems of Global Optimization, *St. Petersburg Math. J.*, 1993, vol. 4, no. 6, pp. 1217–1243.

97. Matveev, A.S., Lagrange Duality in Nonconvex Optimization Theory and Modifications of the Toeplitz–Hausdorff Theorem, *St. Petersburg Math. J.*, 1996, vol. 7, no. 5, pp. 787–815.

98. Matveev, A.S., On the Convexity of the Images of Quadratic Mappings, *St. Petersburg Math. J.*, 1999, vol. 10, no. 2, pp. 343–372.

99. Matveev, A.S., Spectral Approach to Duality in Nonconvex Global Optimization, *SIAM J. Control and Optim.*, 1998, vol. 36, no. 1, pp. 336–378.

100. Polyak, B.T., Convexity of Quadratic Transformations and Its Use in Control and Optimization, *J. Optimizat. Theor. and Appl.*, 1998, vol. 99, no. 3, pp. 553–583.

101. Polyak, B.T., Local Programming, *Comput. Math. Math. Phys.*, 2001, vol. 41, no. 9, pp. 1259–1266.

102. Iwasaki, T., Meinsma, G. and Fu, M., Generalized S-procedure and Finite Frequency KYP Lemma, *Mathematical Problems in Engineering*, 2000, no. 6, pp. 305–320.

103. Iwasaki, T. and Hara, S., Generalized KYP Lemma: Unified Frequency Domain Inequalities with Design Applications, *IEEE Transactions on Automatic Control*, 2005, vol. 50, no. 1, pp. 41–59.

104. Iwasaki, T., Hara, S., and Fradkov, A., Time Domain Interpretations of Frequency Domain Inequalities on (Semi)finite Ranges, *Systems & Control Letters*, 2005, vol. 54, no. 7, pp. 681–691.

105. Fradkov, A.L., Conic S-procedure and Constrained Dissipativity for Linear Systems, *Intern. J. of Robust and Nonlinear Control*, 2007, vol. 17, no. 5–6, pp. 405–413.

106. Sun, W., Gao, H., and Kaynak, O., Finite Frequency H_∞ Control for Vehicle Active Suspension Systems, *IEEE Transactions on Control Systems Technology*, 2011, vol. 19, no. 2, pp. 416–422.

107. Tan, Y.Z., Pang, C.K., Hong, F., et al., Integrated Servo-mechanical Design of High-Performance Mechatronics Using Generalized KYP Lemma, *Microsyst. Technol.*, 2013, vol. 19, pp. 1549–1557.

108. Paszke, W., Rogers, E., and Galkowski, K., Experimentally Verified Generalized KYP Lemma Based Iterative Learning Control Design, *Control Engineering Practice*, 2016, no. 53, pp. 57–67.

109. Yakubovich, V.A., Nonconvex Optimization Problem, *Systems & Control Letters*, 1992, vol. 19, pp. 13–22.

110. Yakubovich, V.A., On One Method for Solving Special Global Optimization Problems, *Vest. Sankt-Peterburg. Gos. Univ.*, 1992, pp. 58–68.

111. Matveev, A.S. and Yakubovich, V.A., *Abstraktnaya teoriya optimal'nogo upravleniya* (Abstract Theory of Optimal Control), St. Petersburg: St. Petersburg State University, 1994.

112. Matveev, A.S. and Yakubovich, V.A., *Optimal'nye sistemy upravleniya: Obyknovennye differentsial'nye uravneniya. Spetsial'nye zadachi* (Optimal Control Systems: Ordinary Differential Equations. Special Problems), St. Petersburg: St. Petersburg State University, 2003.

113. Gelig, A.Kh. and Churilov, A.N., *Stability and Oscillations of Nonlinear Pulse-Modulated Systems*, Boston–Basel–Berlin: Birkhauser, 1998.

114. Gelig, A.Kh., Zuber, I.E., and Churilov, A.N., *Ustoichivost' i stabilizatsiya nelineinykh sistem* (Stability and Stabilization of Nonlinear Systems), St. Petersburg: St. Petersburg State University, 2006.

115. Gelig, A.Kh., *Dinamika impul'snykh sistem i neironnykh setei* (Dynamics of Pulse-Modulated Systems and Neural Networks), Leningrad: Leningrad State University, 1982.

116. Fomin, V.N., *Matematicheskaya teoriya obuchaemykh opoznayushchikh sistem* (Mathematical Theory of Learning Recognizing Systems), Leningrad: Leningrad State University, 1976.

117. Fomin, V.N., *Discrete Linear Control Systems*, Dordrecht–Boston–London: Kluwer Academic Publishers, 1991.

118. Fomin, V., *Optimal Filtering. Vol. 1: Filtering of Stochastic Processes*, Kluwer Academic Publishers, 1998; *Optimal Filtering. Vol. 2: Spatio-Temporal Fields*, Kluwer Academic Publishers, 1999.

119. Fomin, V.N., *Optimal'naya i adaptivnaya fil'tratsiya* (Optimal and Adaptive Filtering), St. Petersburg: St. Petersburg State University, 2001.

120. Leonov, G.A. and Smirnova, V.B., *Matematicheskie problemy teorii fazovoi sinkhronizatsii* (Mathematical Problems of Phase Synchronization Theory), Petersburg: Nauka, 2000.

121. Leonov, G.A., Ponomarenko, D.V., and Smirnova, V.B., *Frequency-Domain Methods for Nonlinear Analysis. Theory and Application*, World Scientific Series on Nonlinear Science, series A, vol. 9, 1996.

122. Leonov, G.A. and Shumafov, M.M., *Metody stabilizatsii lineinykh upravlyaemykh sistem* (Methods of Stabilization of Linear Controlled Systems), St. Petersburg: St. Petersburg State University, 2005.

123. Leonov, G.A., *Mathematical Problems of Control Theory*, World Scientific, 2002.

124. Leonov, G.A., *Teoriya upravleniya* (Control Theory), St. Petersburg: St. Petersburg State University, 2006.

125. Leonov, G.A., *Khaoticheskaya dinamika i klassicheskaya teoriya ustoichivosti dvizheniya* (Chaotic Dynamics and Classical Theory of Stability of Motion), Moscow–Izhevsk, 2006.

126. Barabanov, A.E. and Granichin, O.N., An Optimal Controller of a Linear Plant Subjected to Constrained Noise, *Autom. Remote Control*, 1984, vol. 45, no. 5, pp. 578–584.

127. Barabanov, A.E., *Sintez optimal'nykh regulyatorov* (Design of Optimal Controllers), St. Petersburg: St. Petersburg State University, 1996.

128. Granichin, O.N. and Polyak, B.T., *Randomizirovannye algoritmy otsenivaniya i optimizatsii pri pochti proizvol'nykh pomekhakh* (Randomized Estimation and Optimization Algorithms under Almost Arbitrary Disturbances), Moscow: Nauka, 2003.

129. Granichin, O. and Amelina, N., Simultaneous Perturbation Stochastic Approximation for Tracking under Unknown but Bounded Disturbances, *IEEE Transactions on Automatic Control*, 2015, vol. 60, no. 6, pp. 1653–1658.

130. Matveev, A.S. and Savkin, A.V., *Qualitative Theory of Hybrid Dynamical Systems*, Boston: Birkhauser, 2000.

131. Bondarko, V.A., Discretization of Continuous Linear Dynamic Systems. Analysis of the Methods, *Systems & Control Letters*, 1984, vol. 5, no. 2, pp. 97–101.

132. Bondarko, V.A., Asymptotic Behavior of the Zeros of a Discrete Model of a Linear Continuous System with Delay, *Autom. Remote Control*, 2015, vol. 76, no. 8, pp. 1327–1346.

133. Yakubovich, V.A., Universal Controllers in Problems of Invariance and Tracking, *Dokl. Akad. Nauk*, 1995, vol. 343, no. 2, pp. 172–175.

134. Yakubovich, V.A., Universal Regulators in Linear-Quadratic Optimization Problem, in *Trends in Control: European Perspective*, Isidori, A., Ed., 1995, pp. 53–67.

135. Proskurnikov, A.V. and Yakubovich, V.A., A Problem on the Invariance of a Control System, *Dokl. Akad. Nauk*, 2003, vol. 389, no. 6, pp. 742–746.

136. Proskurnikov, A.V. and Yakubovich, V.A., The Problem of Absolute Invariance for Control System with Delays, *Dokl. Akad. Nauk*, 2004, vol. 397, no. 5, pp. 610–614.

137. Proskurnikov, A.V. and Yakubovich, V.A., Universal Regulators for Optimal Tracking of Polyharmonic Signals in Systems with Delays, *Dokl. Math.*, 2006, vol. 73, pp. 147–151.

138. Proskurnikov, A.V. and Yakubovich, V.A., Universal Regulators for Optimal Tracking of Stochastic Signals with an Unknown Spectral Density, *Dokl. Math.*, 2006, vol. 74, pp. 614–618.

139. Proskurnikov, A., Tempo, R., and Parsegov, S., Network Science on Belief System Dynamics under Logic Constraints, *Science*, 2016, vol. 354, no. 6310, pp. 321–326.

140. Fradkov, A.L. and Pogromsky, A.Yu., *Introduction to Control of Oscillations and Chaos*, Singapore: World Scientific Publishers, 1998.

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

142. Andrievskii, B.R. and Fradkov, A.L., Control of Chaos: Methods and Applications. I. Methods, *Autom. Remote Control*, 2003, vol. 64, no. 5, pp. 673–713.

143. Fradkov, A.L., *Cybernetical Physics: from Control of Chaos to Quantum Control*, Springer-Verlag, 2007.

144. Matveev, A.S. and Savkin, A.V., The Problem of State Estimation via Asynchronous Communication Channels with Irregular Transmission Times, *IEEE Transactions on Automatic Control*, 2003, vol. 48, no. 4, pp. 670–676.

145. Matveev, A.S. and Savkin, A.V., An Analogue of Shannon Information Theory for Networked Control Systems. Stabilization via a Noisy Discrete Channel, *Proc. 43th IEEE Conference on Decision and Control*, Atlantis, Paradise Island, Bahamas, 2004, pp. 4491–4496.

146. Matveev, A.S. and Savkin, A.V., Optimal Control via Asynchronous Communication Channels, *Journal of Optimization Theory and Applications*, 2004, vol. 122, no. 3, pp. 539–572.

147. Matveev, A.S. and Savkin, A.V., An Analogue of Shannon Information Theory for Detection and Stabilization via Noisy Discrete Communication Channels, *SIAM Journal on Control and Optimization*, 2007, vol. 46, no. 4, pp. 1323–1361.

148. Matveev, A.S. and Savkin, A.V., *Estimation and Control over Communication Networks*, Springer-Verlag, 2008.

149. Matveev, A.S. and Pogromskii, A.Y., Observation of Nonlinear Systems via Finite Capacity Channels, Part II: Restoration Entropy and Its Estimates, *Automatica*, 2019, vol. 103, pp. 189–199. https://doi.org/ 10.1016/j.automatica.2019.01.019.

150. Kawan, C., Matveev, A.S., and Pogromsky, A.Y., Remote State Estimation Problem: Towards the Data-Rate Limit along the Avenue of the Second Lyapunov Method, *Automatica*, 2021, vol. 125, art. no. 109467. https://doi.org/10.1016/j.automatica.2020.109467.

151. Fradkov, A.L., Andrievsky, B., and Evans, R.J., Chaotic Observer-Based Synchronization under Information Constraints, *Phys. Rev. E.*, 2006, vol. 73, art. no. 066209.

152. Fradkov, A.L., Andrievsky, B., and Evans, R.J., Synchronization of Nonlinear Systems under Information Constraints, *Chaos*, 2008, vol. 18, no. 3, art. no. 037109, pp. 1–6. https://doi.org/10.1063/1.2977459

153. Fradkov, A.L., Andrievsky, B., and Evans, R.J., Adaptive Observer-Based Synchronization of Chaotic Systems with First-Order Coder in the Presence of Information Constraints, *IEEE Trans. Circuits and Systems I*, 2008, vol. 55, no. 6, pp. 1685–1694.

154. Fradkov, A.L., Andrievsky, B., and Ananyevskiy, M.S., Passification Based Synchronization of Nonlinear Systems under Communication Constraints and Bounded Disturbances, *Automatica*, 2015, vol. 55, no. 5, pp. 287–293.

155. Andrievsky, B., Fradkov, A.L., and Liberzon, D., Robustness of Pecora-Carroll Synchronization under Communication Constraints, *Systems & Control Letters*, 2018, vol. 111, pp. 27–33.

156. Andrievsky, B., Orlov, Y., and Fradkov, A.L., Output Feedback Control of Sine-Gordon Chain over the Limited Capacity Digital Communication Channel, *Electronics*, 2023, vol. 12, p. 2269.

157. Andriesky, B.R., Matveev, A.S., and Fradkov, A.L., Control and Estimation under Information Constraints: toward a Unified Theory of Control, Computation and Communications, *Autom. Remote Control*, 2010, vol. 71, no. 4, pp. 572–633.

158. Shiriaev, A.S., Freidovich, L.B., and Spong, M.W., Controlled Invariants and Trajectory Planning for Underactuated Mechanical Systems, *IEEE Transactions on Automatic Control*, 2014, vol. 59, no. 9, pp. 2555–2561.

159. Shiriaev, A.S., Perram, J.W., and Canudas de Wit, C., Constructive Tool for Orbital Stabilization of Underactuated Nonlinear Systems: Virtual Constraints Approach, *IEEE Transactions on Automatic Control*, 2005, vol. 50, no. 8, pp. 1164–1176.

160. Surov, M.O., Shiriaev, A.S., Freidovich, L.B., Gusev, S.V., and Paramonov, L., Case Study in Non-Prehensile Manipulation: Planning and Orbital Stabilization of One-Directional Rollings for the "Butterfly" Robot, *Proceedings of the International Conference on Robotics and Automation*, May 2015, Washington, DC, pp. 1484–1489.

161. Savkin, A.V., Cheng, T.M., Xi, Z., Javed, F., Matveev, A.S., and Hguyen, H., *Decentralized Coverage Control Problems for Mobile Robotic Sensor and Actuator Networks*, Hoboken, NJ: IEEE Press and John Wiley & Sons, 2015.

162. Matveev, A.S., Savkin, A.V., Hoy, M.C., and Wang, C., *Safe Robot Navigation among Moving and Steady Obstacles*, Oxford, UK: Elsevier and Butterworth Heinemann, 2016.

163. Hoy, M., Matveev, A.S., and Savkin, A.V., Algorithms for Collision-Free Navigation of Mobile Robots in Complex Cluttered Environments: a Survey, *Robotica*, 2015, vol. 33, no. 03, pp. 463–497.

164. Matveev, A.S. and Savkin, A.V., Optimal Chemotherapy Regimens: Influence of Tumors on Normal Cells and Several Toxicity Constraints, *IMA Journal of Mathematics Applied in Medicine and Biology*, 2001, vol. 18, pp. 25–40.

165. Churilov, A., Medvedev, A., and Shepeljavyi, A., Mathematical Model of Non-basal Testosterone Regulation in the Male by Pulse Modulated Feedback, *Automatica*, 2009, vol. 45, no. 1, pp. 78–85.

166. Churilov, A., Medvedev, A., and Shepeljavyi, A., A State Observer for Continuous Oscillating Systems under Intrinsic Pulse-Modulated Feedback, *Automatica*, 2012, vol. 48, pp. 1117–1122.

167. Filippov, S.A. and Fradkov, A.L., Control Engineering at School: Learning by Examples, *IFAC Proceedings Volumes*, 2012, vol. 45, no. 11, pp. 118–123.

168. Filippov, S., Ten, N., Shirokolobov, I., and Fradkov, A., Teaching Robotics in Secondary School, *IFAC-PapersOnLine*, 2017, vol. 50, no. 1, pp. 12155–12160.

169. *Autom. Remote Control*, 2006, vol. 67, nos. 10 and 11. Special Issues Dedicated to the 80th Anniversary of V.A. Yakubovich's Birth.

170. *Vest. Sankt-Peterburg. Univ. Mat. Mekh. Astron.*, 2006, no. 4. Special Issue Dedicated to the 80th Anniversary of V.A. Yakubovich's Birth.

171. *International Journal of Robust and Nonlinear Control*, Special Issue: Frequency-domain and Matrix Inequalities in Systems and Control Theory, Dedicated to the 80th Birthday of V.A. Yakubovich, 2007, vol. 17, no. 5–6.

172. Fradkov, A.L., Scientific Biography of V.A. Yakubovich and His School at St. Petersburg (Leningrad) University, Special Session on Vladimir Andreevich Yakubovich and His Scientific School, *Trudy 12-go Vserossiiskogo soveshchaniya po problemam upravleniya (VSPU-2014)* (Proceedings of the 12th All-Russian Meeting on Control Problems (AMCP-2014)), Moscow, June 16-19, 2014.

173. Fradkov, A., Gelig, A., and Leonov, G., Vladimir Andreevich Yakubovich. Obituary, *IEEE Control Systems Magazine*, 2013, vol. 33, no. 2, pp. 89–91.

174. Fradkov, A.L., Scientific School of Vladimir Yakubovich in the 20th Century, *IFAC-PapersOnLine*, 2017, vol. 50, no. 1, pp. 5231–5237.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 9, pp. 1041–1064.

LINEAR SYSTEMS

Construction of the Time-Optimal Bounded Control for Linear Discrete-Time Systems Based on the Method of Superellipsoidal Approximation

D. N. Ibragimov[*,a] **and V. M. Podgornaya**[*,b]

[*] *Moscow Aviation Institute (National Research University), Moscow, Russia*
e-mail: [a] *rikk.dan@gmail.ru,* [b] *vita1401@outlook.com*

Received April, 2023
Revised June 25, 2023
Accepted July 20, 2023

Abstract—The speed-in-action problem for a linear discrete-time system with bounded control is considered. In the case of superellipsoidal constraints on the control, the optimal control process is constructed explicitly on the basis of the discrete maximum principle. The problem of calculating the initial conditions for an adjoint system is reduced to solving a system of algebraic equations. The algorithm for generating a guaranteeing solution based on the superellipsoidal approximation method is proposed for systems with general convex control constraints. The procedure of superellipsoidal approximation is reduced to solving a number of convex programming problems. Examples are given.

Keywords: linear discrete-time systems, speed-in-action problem, maximum principle, superellipse, ellipsoidal approximations

DOI: 10.25728/arcRAS.2023.98.83.001

1. INTRODUCTION

One of the natural control quality functions is the time spent by the system to achieve a given terminal state. In practice, the resulting optimal control problem is called the speed-in-action problem. It is essential that the speed-in-action problem for linear discrete-time systems has a number of serious differences from a similar problem for continuous systems. While in the case of continuous time, the solution obtained on the basis of the Pontryagin's maximum principle [1] for a linear system guarantees the relay nature of the optimal control in terms of speed, a similar result for a system with discrete time [2, 3] is incorrect.

The direct approach based on minimizing the norm of the terminal state for all control actions turns out to be difficult to apply for high-dimensional systems with a large time horizon and vector control. This is due to the fact that the resulting mathematical programming problem is characterized by a rapid increase in the number of constraints and optimization variables with an increase in the number of steps required for the system to reach origin. At the same time, for almost all initial states, the extremum in the speed-in-action problem is irregular [4], which also complicates the use of known numerical methods.

Consideration of the optimality conditions of the process using various classical approaches leads to two fundamentally different methods for solving the speed problem. Bellman's dynamic programming method [5] makes it possible to construct an optimal control in a positional form. In the case when the set of admissible control values is a polyhedron, the calculation of each control action is reduced to solving a linear programming problem [6]. Also, in [6], a method for forming

optimal control in the case of arbitrary convex control constraints based on polyhedral approximation is demonstrated [7]. This approach has a number of disadvantages related to computational difficulties. The accuracy of the guaranteeing solution in the speed-in-action problem is achieved by increasing the number of vertices of the polyhedral approximation, which leads to an exponential increase in the complexity of the corresponding linear programming problems. Due to this fact, such approach, when implemented on standard computing devices, is characterized by either low accuracy of the solution, or a relatively small time horizon, especially for large-dimensional systems.

On the contrary, the combination of optimality conditions in the speed-in-action problem with the discrete maximum principle [1–3] allows optimal program control to be formed [4]. An essential condition for the applicability of these methods is the strict convexity of the set of admissible control values. But the relation for calculating the initial state of a conjugate system in the case of an arbitrary structure of control constraints is difficult to solve. In [8], a special case of an ellipsoidal structure of a set of admissible control values is presented. An analytical solution to the speed-in-action problem based on the necessary and sufficient optimality conditions presented in [4].

A natural approach is to combine the ideas of constructing a guaranteeing solution from [6] on the basis of ellipsoidal approximation of the set of admissible control values in combination with methods of forming program control according to the discrete maximum principle [4, 8]. The technique of ellipsoidal approximation is widely used in the theory of optimal control [9, 10]. However, the class of ellipsoids does not allow achieving arbitrary accuracy of the approximation of the initial set, and consequently, the accuracy of solving the optimal control problem. The article considers a class of superellipsoidal sets (the exact definition is given in Section 2), which allow a higher order of the accuracy while maintaining strict convexity conditions, which guarantees the simplicity of solving the speed-in-action problem in the same way [8].

Superellipses on the plane have been known for a long time as Lame curves [11] and have a large number of different applications in natural science and technical disciplines. They are actively used, for example, in geodesy and mapping tasks [12], in botany for modeling plant growth [13] and describing natural shapes [14], designing waveguides for antenna arrays [15, 16] or for modeling bends of various structures [17]. However, the general study of the properties of these figures is usually limited to the two-dimensional case [12, 18]. This fact makes it relevant to study the properties of this class of geometric bodies in a space of arbitrary dimension in terms of convex analysis: the description of their support function, support point and normal cone, the solution of various approximation problems.

The purpose of this work is to develop a method for generating optimal control explicitly in the case of a superellipsoidal structure of a set of acceptable control values, as well as to describe an approach for constructing a superellipsoidal approximation of an arbitrary convex body with the highest possible accuracy. The fundamental difference from this paper and both classical [19–21] and modern [22, 23] results is the consideration of arbitrary vector control, which is convex constrained, and the lack of restrictions on the dimension of the phase space. It is a more general statement of the problem, expanding the range of possible applications.

The article has the following structure. Section 2 presents non-standard designations and assumptions that are used in the article. In Section 3, the speed-in-action problem is considered, the maximum principle is described as the main tool for its solution, and the formulation of the problem of superellipsoidal approximation of the set of admissible control values is formulated in order to form a guaranteeing process. Section 4 presents an exact solution to the speed-in-action problem in the case of a superellipsoidal structure of a set of admissible control values. Section 5 describes a method for reducing the problem of optimal in the sense of the Lebesgue

measure superellipsoidal approximation of a convex body to a number of convex programming problems. Section 6 demonstrates examples of constructing a guaranteeing process in a speed-in-action problem for systems of different dimensions based on the obtained theoretical results. Estimates of the accuracy of the constructed processes in comparison with the optimal solution are given.

2. DESIGNATIONS

We will assume that the phase space is a Euclidean space $\mathbb{R}^n$ with a scalar product defined by the relation

$$(x,y)=\sum_{i=1}^{n}x_iy_i.$$

For any $r\in[1;+\infty)$ define on $\mathbb{R}^n$ norm

$$\|x\|_r=\left(\sum_{i=1}^{n}|x_i|^r\right)^{\frac{1}{r}}.$$

For $r=2$ the norm $\|\cdot\|_2$ is consistent with the scalar product. From the point of view of theory, the value $r=1$ is acceptable, but it will not be considered within the paper, which allows us to define the number $q>1$ as a Helder dual of the number r:

$$\frac{1}{r}+\frac{1}{q}=1.$$

For arbitrary sets $\mathcal{X},\mathcal{U}\subset\mathbb{R}^n$ and the matrix $D\in\mathbb{R}^{n\times n}$ we denote the Minkowski sum by $\mathcal{X}+\mathcal{U}$

$$\mathcal{X}+\mathcal{U}=\{x+u\colon x\in\mathcal{X},\ \ u\in\mathcal{U}\},$$

and we denote by $D\mathcal{U}$ the image of the set $\mathcal{U}$ under the mapping D

$$D\mathcal{U}=\{Du\colon u\in\mathcal{U}\}.$$

By $\partial\mathcal{U}$ and $\operatorname{int}\mathcal{U}$ we denote the sets of boundary and interior points of $\mathcal{U}$ respectively. $\operatorname{cone}\{\mathcal{U}\}$ is the conic hull of the set $\mathcal{U}$ [24, § 2 ch. I].

If the set $\mathcal{U}\subset\mathbb{R}^n$ is a convex compact, then for an arbitrary point $u\in\mathcal{U}$ by $\mathcal{N}(u,\mathcal{U})$ we denote the normal cone of the set $\mathcal{U}$ at the point u [24, § 2 ch. I]:

$$\mathcal{N}(u,\mathcal{U})=\left\{p\in\mathbb{R}^n\setminus\{0\}\colon(p,u)=\max_{\tilde{u}\in\mathcal{U}}(p,\tilde{u})\right\}.$$

The elements of the normal cone $\mathcal{N}(u,\mathcal{U})$ are called support vectors to $\mathcal{U}$ at the point u. Note that by construction equality $\mathcal{N}(u,\mathcal{U})=\varnothing$ is valid if and only if the inclusion $u\in\operatorname{int}\mathcal{U}$ is correct. If the inclusion of $0\in\operatorname{int}\mathcal{U}$ is also true, then $\mathcal{U}$ will be called a convex body [25, Section 3 § 1 ch. IV] and for an arbitrary $x\in\mathbb{R}^n$ we introduce the Minkowski functional [25, Section 3 § 2 ch. III] or the calibration function [24, § 4 ch. I]:

$$M(x,\mathcal{U})=\inf\{t>0\colon x\in t\mathcal{U}\}=\inf\left\{t>0\colon\frac{x}{t}\in\mathcal{U}\right\}.$$

The strickly convex set $\mathcal{U} \subset \mathbb{R}^n$ is such set that for any $u^1, u^2 \in \mathcal{U}$, $\lambda \in (0;1)$ the inclusion $\lambda u^1 + (1-\lambda)u^2 \in \operatorname{int} \mathcal{U}$ is correct.

We will call a superellipse or superellipsoidal set for some $a_1 > 0, \ldots, a_n > 0$, $r > 1$ a set of the form

$$\mathcal{E}_r(a_1, \ldots, a_n) = \left\{ x \in \mathbb{R}^n \colon \sum_{i=1}^{n} \left| \frac{x_i}{a_i} \right|^r \leqslant 1 \right\}. \tag{1}$$

We will assume shortering $a = (a_1, \ldots, a_n)^{\mathrm{T}}$ and denote the corresponding superellipse by $\mathcal{E}_r(a)$. By $\operatorname{diag}(a) \in \mathbb{R}^{n \times n}$ we denote a diagonal matrix constructed by the vector $a \in \mathbb{R}^n$:

$$\operatorname{diag}(a) = \begin{pmatrix} a_1 & 0 & \ldots & 0 \\ 0 & a_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_n \end{pmatrix}.$$

3. PROBLEM STATEMENT

The linear discrete-time system with limited control $(A,\mathcal{U})$ is considered:

$$\begin{aligned} &x(k+1) = Ax(k) + u(k), \\ &x(0) = x_0, \quad u(k) \in \mathcal{U}, \quad k \in \mathbb{N} \cup \{0\}, \end{aligned} \tag{2}$$

where $x(k) \in \mathbb{R}^n$ is the state vector of the system, $u(k) \in \mathbb{R}^n$ is the control action, $A \in \mathbb{R}^{n \times n}$ is the matrix of the system, $\mathcal{U} \subset \mathbb{R}^n$ is the set of valid control values. It is assumed that $\det A \neq 0$, $\mathcal{U}$ is a convex compact, $0 \in \operatorname{int} \mathcal{U}$.

For the (2) system, the speed-in-action problem is solved, i.e. it is required to transfer the system $(A,\mathcal{U})$ from a given initial state $x_0 \in \mathbb{R}^n$ to the origin in the minimum number of steps $N_{\min}$:

$$N_{\min} = \min \left\{ N \in \mathbb{N} \cup \{0\} \colon \exists u(0), \ldots, u(N-1) \in \mathcal{U} \colon x(N) = 0 \right\}.$$

The control process $\{x^*(k), u^*(k-1), x^*(0)\}_{k=1}^{N_{\min}}$, satisfying the condition $x^*(N_{\min}) = 0$, we will call optimal. It is assumed that the speed-in-action problem for the system $(A,\mathcal{U})$ is solvable, i.e. $N_{\min} < \infty$. The issues of solvability of the speed-in-action problem for the (2) system are discussed in detail in [26].

The construction of optimal processes is closely related to the apparatus of 0-controllable sets [4, 6]. For an arbitrary $N \in \mathbb{N} \cup \{0\}$ we denote by $\mathcal{X}(N) \subset \mathbb{R}^n$ the 0-controllable set of the system (2) in N steps, i.e. the set of those initial states from which the system (2) can be transferred to 0 in N steps by acceptable control actions:

$$\mathcal{X}(N) = \begin{cases} \{x_0 \in \mathbb{R}^n \colon \exists u(0), \ldots, u(N-1) \in \mathcal{U} \colon x(N) = 0\}, & N \in \mathbb{N}, \\ \{0\}, & N = 0. \end{cases} \tag{3}$$

Then, according to the definition of $N_{\min}$ the following representation is also valid:

$$N_{\min} = \min\{N \in \mathbb{N} \cup \{0\} \colon x_0 \in \mathcal{X}(N)\}. \tag{4}$$

At the same time, the control, as demonstrated in [4, 6], is optimal if and only if for all $k = \overline{0, N_{\min} - 1}$ the inclusion is true

$$x^*(k+1) = Ax^*(k) + u^*(k) \in \mathcal{X}(N_{\min} - k - 1).$$

In [4], a number of results were obtained for the speed-in-action problem, which can be presented in the form of the maximum principle for a strictly convex $\mathcal{U}$.

Theorem 1. *Let $\mathcal{U} \subset \mathbb{R}^n$ be a strictly convex and compact set, $0 \in \operatorname{int} \mathcal{U}$, $\det A \neq 0$, a class of sets $\{\mathcal{X}(N)\}_{N=0}^{\infty}$ is defined according to (3), the control process $\{x^*(k), u^*(k-1), x^*(0)\}_{k=1}^{N_{\min}}$ and the trajectory of the conjugate system $\{\psi(k)\}_{k=0}^{N_{\min}-1}$ satisfy the relations*

$$
\begin{aligned}
&x^*(k+1) = Ax^*(k) + u^*(k),\\
&u^*(k) = \alpha \arg \max_{u \in \mathcal{U}} \left((A^{-1})^{\mathrm{T}} \psi(k), u \right),\\
&\psi(k+1) = (A^{-1})^{\mathrm{T}} \psi(k),\\
&x^*(0) = x_0,\\
&-\psi(0) \in \mathcal{N}\left(x_0, \alpha \mathcal{X}(N_{\min})\right),\\
&\alpha = M(x_0, \mathcal{X}(N_{\min})).
\end{aligned}
$$

Then

1) *$\{x^*(k), u^*(k-1), x^*(0)\}_{k=1}^{N_{\min}}$ is the optimal process for the system $(A, \mathcal{U})$;*
2) *if $\alpha = 1$, then the optimal process is the only one;*
3) *$-\psi(k) \in \mathcal{N}\left(x^*(k), \alpha \mathcal{X}(N_{\min} - k)\right)$, $k = \overline{0, N_{\min} - 1}$.*

From a computational point of view, the question of applying the Theorem 1 comes down to determining α and $\psi(0)$ from the conditions

$$
\begin{aligned}
&-\psi(0) \in \mathcal{N}\left(x_0, \alpha \mathcal{X}(N_{\min})\right),\\
&\alpha = M(x_0, \mathcal{X}(N_{\min})).
\end{aligned}
\tag{5}
$$

This problem in the case of an arbitrary strictly convex body $\mathcal{U}$ can be a nontrivial problem.

The main purpose of this paper is to construct effective methods for solving the conditions (5) with respect to $\psi(0) \in \mathbb{R}^n \setminus \{0\}$ and $\alpha > 0$ for the special case when $\mathcal{U}$ allows the representation

$$
\mathcal{U} = B\mathcal{E}_r(a), \quad B \in \mathbb{R}^{n \times n}, \quad \det B \neq 0, \quad a_1, \ldots, a_n > 0, \quad r > 1. \tag{6}
$$

Another goal of the paper is to develop a method for approximating an arbitrary convex body $\mathcal{U}$ by a nested set $\hat{\mathcal{U}}$ of the form (6), that minimizes the Lebesgue measure of the difference between two sets $\mu(\mathcal{U} \setminus \hat{\mathcal{U}})$, in order to construct a guaranteed solution in the speed-in-action problem for the system $(A, \mathcal{U})$.

4. THE OPTIMAL PROCESS IN THE CASE OF A SUPERELLIPSOIDAL STRUCTURE OF CONTROL CONSTRAINTS

The conditions (5) can be reduced to an equivalent system of algebraic equations in the case of (6). To do this, we will carry out an analytical description of 0-controllable sets and some properties of strictly convex and superellipsoidal sets.

Lemma 1 [4, Lemma 1]. *Let $\det A \neq 0$, the class of sets $\{\mathcal{X}(N)\}_{N=0}^{\infty}$ be determined by the relations (3). Then for any $N \in \mathbb{N}$ the representation is true*

$$
\mathcal{X}(N) = -\sum_{k=1}^{N} A^{-k} \mathcal{U}.
$$

Lemma 2 [27, Lemma 3]. *Let $\mathcal{U} \subset \mathbb{R}^n$ be a strictly convex compact, $0 \in \operatorname{int} \mathcal{U}$. Then for any different $u^1, u^2 \in \mathcal{U}$ it is true that*

$$\mathcal{N}(u^1,\mathcal{U}) \cap \mathcal{N}(u^2,\mathcal{U}) = \varnothing.$$

The following statement follows from [27, Lemmas 5, 6].

Lemma 3. *Let $\mathcal{U}, \mathcal{X} \subset \mathbb{R}^n$ be convex compacts, $u \in \mathcal{U}$, $x \in \mathcal{X}$, $A \in \mathbb{R}^{n\times n}$, $\det A \neq 0$.*
Then
1) $\mathcal{N}(u + x,\mathcal{U} + \mathcal{X}) = \mathcal{N}(u,\mathcal{U}) \cap \mathcal{N}(x,\mathcal{X})$;
2) $\mathcal{N}(Ax, A\mathcal{X}) = (A^{-1})^{\mathrm{T}}\mathcal{N}(x,\mathcal{X})$.

The Lemma 3 defines the transformation of the normal cone of convex sets with non-degenerate linear mapping and Minkowski addition. Taking into account the Lemma 1 this makes it possible to describe an arbitrary normal cone of any 0-controllable set in terms of the normal cones of the set $\mathcal{U}$ or $\mathcal{E}_r(a_1,\dots,a_n)$ in the case (6). On the other hand, the Lemma 2 establishes a one-to-one correspondence between a boundary point and its normal cone for a strictly convex set. If this dependence is described explicitly, then it is possible to obtain algebraic equations equivalent to the conditions (5).

We introduce for an arbitrary $r > 1$ the bijective operator $I_r\colon \mathbb{R}^n \to \mathbb{R}^n$:

$$I_r(x) = \left(\operatorname{sgn}(x_1)|x_1|^{r-1},\dots,\operatorname{sgn}(x_n)|x_n|^{r-1}\right).$$

Lemma 4. *Let the set $\mathcal{E}_r(a)$ be defined by the relations* (1). *Then*
1) *for any $x \in \partial\mathcal{E}_r(a)$ it is true that*

$$\mathcal{N}\left(x,\mathcal{E}_r(a)\right) = \left\{\gamma \operatorname{diag}(a)^{-1} I_r\left(\operatorname{diag}(a)^{-1}x\right) \in \mathbb{R}^n\colon \gamma > 0\right\};$$

2) *for any $p \in \mathbb{R}^n \setminus \{0\}$ there is a unique*

$$x^*(p) = \arg \max_{x\in\mathcal{E}_r(a)} (p,x) = \frac{\operatorname{diag}(a) I_q\left(\operatorname{diag}(a)p\right)}{\|\operatorname{diag}(a)p\|_q^{q-1}}.$$

The proof of the Lemma 4 and all other statements is given in the Appendix.

Lemma 5. *Let $\mathcal{U} = D\mathcal{E}_r(a)$, where $\mathcal{E}_r(a)$ is determined by the relations* (1), *$D \in \mathbb{R}^{n\times n}$, $\det D \neq 0$. Then*
1) *for any $u \in \partial\mathcal{U}$*

$$\mathcal{N}\left(u,\mathcal{U}\right) = \left\{\gamma (D^{-1})^{\mathrm{T}} \operatorname{diag}(a)^{-1} I_r\left(\operatorname{diag}(a)^{-1} D^{-1} u\right) \in \mathbb{R}^n\colon \gamma > 0\right\};$$

2) *for any $p \in \mathbb{R}^n \setminus \{0\}$ there is only one*

$$u^*(p) = \arg \max_{u\in\mathcal{U}} (p,u) = \frac{D\operatorname{diag}(a) I_q\left(\operatorname{diag}(a) D^{\mathrm{T}} p\right)}{\|\operatorname{diag}(a) D^{\mathrm{T}} p\|_q^{q-1}}.$$

The Lemma 5, on the one hand, allows us to calculate the optimal control according to the Theorem 1 in the case (6), when we choose $D = B$. On the other hand, the Lemma 5 in combination with Lemmas 1 and 2 connects a point on the boundary of the 0-controllable set with an element of its normal cone, when we choose $D = -A^{-k}B$, which makes it possible to reduce the conditions (5) to equivalent algebraic equations. We formulate this fact in the form of a theorem.

Theorem 2. *Let $\mathcal{U}$ be determined according to* (6)*, $x_0 \neq 0$, $\psi(0) \in \mathbb{R}^n \setminus \{0\}$, $\alpha > 0$. Then $\psi(0)$ and α satisfy the conditions* (5) *if and only if the following equality is true:*

$$-x_0 = \alpha \sum_{k=1}^{N_{\min}} \frac{A^{-k}B\operatorname{diag}(a)I_q\left(\operatorname{diag}(a)(A^{-k}B)^{\mathrm{T}}\psi(0)\right)}{\|\operatorname{diag}(a)(A^{-k}B)^{\mathrm{T}}\psi(0)\|_q^{q-1}}.$$

The system of equations presented in the Theorem 2 has not the only solution, since the right part is invariant to the multiplication of the vector $\psi(0)$ by any positive number. To use numerical methods, we can propose a modification of this system, which has a single solution.

Corollary 1. *Let $\mathcal{U}$ be determined according to* (6)*, $\psi(0) \in \mathbb{R}^n \setminus \{0\}$, $\alpha > 0$. Then for any $x_0 \neq 0$ there is a unique solution of the system of equations*

$$\begin{cases} -x_0 = \alpha \displaystyle\sum_{k=1}^{N_{\min}} \frac{A^{-k}B\operatorname{diag}(a)I_q\left(\operatorname{diag}(a)(A^{-k}B)^{\mathrm{T}}\psi(0)\right)}{\|\operatorname{diag}(a)(A^{-k}B)^{\mathrm{T}}\psi(0)\|_q^{q-1}}, \\ (\psi(0), \psi(0)) = 1, \end{cases}$$

which also satisfies the conditions (5)*.*

Example 1. Consider the procedure for calculating $\psi(0)$, α, $N_{\min}$ based on the Corollary 1. The parameters of the system (2) have the following values:

$$A = \begin{pmatrix} 3 & 1 \\ 1 & -2 \end{pmatrix}, \quad B = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}, \quad a_1 = 2, \quad a_2 = 3,$$
$$r = \frac{4}{3}, \quad q = 4, \quad x_0 = \left(\frac{1}{3}, \frac{4}{3}\right)^{\mathrm{T}}.$$

Suppose that $N_{\min} = 2$, and we make up the system of equations presented in the Theorem 1:

$$\frac{0.20\big(0.20\psi_1(0) + 0.81\psi_2(0)\big)^3 + 0.91\big(0.91\psi_1(0) - 0.61\psi_2(0)\big)^3}{\left(\big(0.20\psi_1(0) + 0.81\psi_2(0)\big)^4 + \big(0.91\psi_1(0) - 0.61\psi_2(0)\big)^4\right)^{\frac{3}{4}}}$$
$$+ \frac{0.17\big(0.17\psi_1(0) - 0.32\psi_2(0)\big)^3 + 0.17\big(0.17\psi_1(0) + 0.39\psi_2(0)\big)^3}{\left(\big(0.17\psi_1(0) - 0.32\psi_2(0)\big)^4 + \big(0.17\psi_1(0) + 0.39\psi_2(0)\big)^4\right)^{\frac{3}{4}}} = -\frac{1}{3\alpha},$$

$$\frac{0.81\big(0.20\psi_1(0) + 0.81\psi_2(0)\big)^3 - 0.61\big(0.91\psi_1(0) - 0.61\psi_2(0)\big)^3}{\left(\big(0.20\psi_1(0) + 0.81\psi_2(0)\big)^4 + \big(0.91\psi_1(0) - 0.61\psi_2(0)\big)^4\right)^{\frac{3}{4}}}$$
$$+ \frac{-0.32\big(0.17\psi_1(0) - 0.32\psi_2(0)\big)^3 + 0.39\big(0.17\psi_1(0) + 0.39\psi_2(0)\big)^3}{\left(\big(0.17\psi_1(0) - 0.32\psi_2(0)\big)^4 + \big(0.17\psi_1(0) + 0.39\psi_2(0)\big)^4\right)^{\frac{3}{4}}} = -\frac{4}{3\alpha}.$$

By supplementing this system with the equivalence $\psi_1(0)^2 + \psi_2(0)^2 = 1$ according to the Corollary of 1, we get the following solution:

$$\psi_1(0) = -0.35, \quad \psi_2(0) = -0.94, \quad \alpha = 1.08.$$

Due to (5) it is true that $\alpha = M(x_0, \mathcal{X}(2))$. It is correct that $x_0 \notin \mathcal{X}(2)$ since $\alpha > 1$. We got a contradiction, from which it follows that $N_{\min} > 2$.

Assume that $N_{\min} = 3$, and make up the system of equations presented in the Theorem 1:

$$\frac{0.20\big(0.20\psi_1(0) + 0.81\psi_2(0)\big)^3 + 0.91\big(0.91\psi_1(0) - 0.61\psi_2(0)\big)^3}{\big(\big(0.20\psi_1(0) + 0.81\psi_2(0)\big)^4 + \big(0.91\psi_1(0) - 0.61\psi_2(0)\big)^4\big)^{\frac{3}{4}}}$$
$$+ \frac{0.17\big(0.17\psi_1(0) - 0.32\psi_2(0)\big)^3 + 0.17\big(0.17\psi_1(0) + 0.39\psi_2(0)\big)^3}{\big(\big(0.17\psi_1(0) - 0.32\psi_2(0)\big)^4 + \big(0.17\psi_1(0) + 0.39\psi_2(0)\big)^4\big)^{\frac{3}{4}}}$$
$$+ \frac{0.004\big(0.004\psi_1(0) + 0.16\psi_2(0)\big)^3 + 0.11\big(0.11\psi_1(0) - 0.14\psi_2(0)\big)^3}{\big(\big(0.004\psi_1(0) - 0.16\psi_2(0)\big)^4 + \big(0.11\psi_1(0) - 0.14\psi_2(0)\big)^4\big)^{\frac{3}{4}}} = -\frac{1}{3\alpha},$$

$$\frac{0.81\big(0.20\psi_1(0) + 0.81\psi_2(0)\big)^3 - 0.61\big(0.91\psi_1(0) - 0.61\psi_2(0)\big)^3}{\big(\big(0.20\psi_1(0) + 0.81\psi_2(0)\big)^4 + \big(0.91\psi_1(0) - 0.61\psi_2(0)\big)^4\big)^{\frac{3}{4}}}$$
$$+ \frac{-0.32\big(0.17\psi_1(0) - 0.32\psi_2(0)\big)^3 + 0.39\big(0.17\psi_1(0) + 0.39\psi_2(0)\big)^3}{\big(\big(0.17\psi_1(0) - 0.32\psi_2(0)\big)^4 + \big(0.17\psi_1(0) + 0.39\psi_2(0)\big)^4\big)^{\frac{3}{4}}}$$
$$+ \frac{0.17\big(0.004\psi_1(0) + 0.16\psi_2(0)\big)^3 - 0.14\big(0.11\psi_1(0) - 0.14\psi_2(0)\big)^3}{\big(\big(0.004\psi_1(0) + 0.16\psi_2(0)\big)^4 + \big(0.11\psi_1(0) - 0.14\psi_2(0)\big)^4\big)^{\frac{3}{4}}} = -\frac{4}{3\alpha}.$$

By supplementing this system with the equivalence $\psi_1(0)^2 + \psi_2(0)^2 = 1$ according to the corollary of 1, we get the following solution:

$$\psi_1(0) = -0.50, \quad \psi_2(0) = -0.87, \quad \alpha = 0.96.$$

Then $\alpha = M(x_0, \mathcal{X}(3)) < 1$, i.e. by definition of the Minkowski functional $x_0 \in \mathcal{X}(3)$. Due to (4) it is true that $N_{\min} = 3$.

Remark 1. In the Example 1 and everywhere else, the numerical solution of systems of algebraic equations constructed according to the Corollary 1, is carried out in the Maple software environment by means of built-in procedures based on the Newton method and its modifications [28].

The Theorem 2 and the Corollary 1 in conjunction with the Theorem 1 allow us to completely solve the speed-in-action problem for a linear discrete-time system in the case of a superellipsoidal structure of the set of admissible values of control (6). The solution of the conditions (5) according to the Corollary 1 is equivalent to the numerical solution of a system of algebraic equations. At the same time, the optimal process and the trajectory of the conjugate system can be calculated from the recurrence relations presented in the Theorem 1. Optimal control is explicitly defined by point 2 of Lemma 5.

5. INTERNAL SUPERELLIPSOIDAL APPROXIMATION OF A CONVEX BODY

The case of (6) is quite special. It is often impossible to guarantee even the strict convexity of the set $\mathcal{U}$. In this connection, it turns out to be relevant to carry out an internal approximation of $\mathcal{U}$ by a set $\hat{\mathcal{U}}$ of the form (6). Transition in the speed-in-action problem from the system $(A, \mathcal{U})$ to the auxiliary system $(A, \hat{\mathcal{U}})$ allows us to construct a guaranteeing control in the original system based on the methods presented in the Section 4 in relation to the auxiliary system.

In this case, the error of the guaranteeing solution in comparison with the optimal one will be the smaller, the larger the approximating set $\hat{\mathcal{U}}$ is by inclusion. This fact leads to the need to solve the problem of optimal superellipsoidal approximation of a convex compact body $\mathcal{U} \subset \mathbb{R}^n$ by a set of the form (6). As an approximation quality criterion, we consider the Lebesgue measure of the n-dimensional set $\mu(\cdot)$ [25, Section 1 § 3 ch. V]. The resulting optimization problem will take the form

$$\begin{aligned}
&\mu(\mathcal{U} \setminus B\mathcal{E}_r(a_1, \ldots, a_n)) \to \min_{a_1,\ldots,a_n,r,B},\\
&a_i > 0,\ i = \overline{1,n},\\
&r > 1,\\
&B \in \mathbb{R}^{n\times n},\ \ \det B \neq 0,\\
&\mathcal{E}_r(a_1, \ldots, a_n) \subset \mathcal{U}.
\end{aligned}$$

This problem can be divided into two separate stages: the first stage is the selection of the orientation matrix of the superellipse $B \in R^{n\times n}$and the second stage is the selection of the numbers $a_1, \ldots, a_n > 0$, $r > 1$, parametrizing the set (1).

5.1. Selection of the Orientation Matrix of a Superellipsoidal Set

In general, the search for the optimal value of the matrix B can be a complex optimization problem, the solvability conditions of which are unknown due to its non-convexity. We propose a heuristic method for choosing B in the form of an orthogonal matrix. Since the rotation transformation preserves the Lebesgue measure, then the following equalities are valid:

$$\mu(\mathcal{U} \setminus B\mathcal{E}_r(a)) = \mu(B^{-1}(\mathcal{U} \setminus B\mathcal{E}_r(a))) = \mu(B^{-1}\mathcal{U} \setminus \mathcal{E}_r(a)).$$

They make it possible to reduce the original approximation problem to the problem of optimal internal approximation of an arbitrary convex compact body $B^{-1}\mathcal{U} \subset \mathbb{R}^n$ by the superellipse $\mathcal{E}_r(a)$. Due to the symmetry of the set $\mathcal{E}_r(a)$, it is acceptable to assume that B^{-1} should provide such a rotation of the set $\mathcal{U}$, so that the coordinate axes coincide with any axes of "symmetry" of $\mathcal{U}$, for example, with the main axes of inertia of a convex bodie $\mathcal{U}$ [29, § 32 ch. VI].

In this case, B must satisfy the condition

$$I_{\mathcal{U}} = B\mathrm{diag}(\lambda_1, \ldots, \lambda_n)B^{-1},$$

where $I_{\mathcal{U}} \in \mathbb{R}^{n\times n}$ is the inertia tensor of a convex body $\mathcal{U} \subset \mathbb{R}^n$:

$$I_{\mathcal{U}} = \begin{pmatrix} I_{11} & \ldots & I_{1n} \\ \vdots & \ddots & \vdots \\ I_{n1} & \ldots & I_{nn} \end{pmatrix}, \quad I_{ij} = \begin{cases} \int\limits_{\mathcal{U}} \sum\limits_{\substack{k=1\\k\neq i}}^{n} x_k^2 dx_1 \ldots dx_n, & i = j,\\ -\int\limits_{\mathcal{U}} x_i x_j dx_1 \ldots dx_n, & i \neq j. \end{cases}$$

Then according to [30, Theorem 3.1.11] B is determined in a unique way up to the permutation of its columns and its construction is reduced to calculating the eigenvectors of the matrix $I_{\mathcal{U}}$.

Example 2. Let us calculate the matrix B for the polyhedron $\mathcal{U} \subset \mathbb{R}^2$:

$$\mathcal{U} = \mathrm{conv}\left\{\begin{pmatrix}4\\4\end{pmatrix}, \begin{pmatrix}2\\4\end{pmatrix}, \begin{pmatrix}-2\\2\end{pmatrix}, \begin{pmatrix}-4\\-4\end{pmatrix}, \begin{pmatrix}-2\\4\end{pmatrix}, \begin{pmatrix}2\\-2\end{pmatrix}\right\}.$$

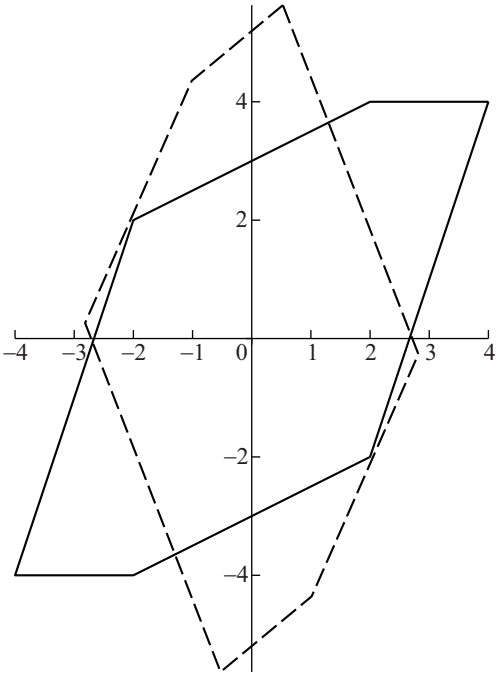


Fig. 1. The original set $\mathcal{U}$ (solid line) and the set $B^{-1}\mathcal{U}$ oriented along the axes of inertia (dotted line).

The inertia tensor $I_{\mathcal{U}}$ and the matrix B have the following numerical values:

$$I_{\mathcal{U}} = \begin{pmatrix} 153.28 & -85.03 \\ -85.03 & 121.20 \end{pmatrix}, \quad B = \begin{pmatrix} 0.64 & -0.77 \\ 0.77 & 0.64 \end{pmatrix}.$$

Then the oriented set $B^{-1}\mathcal{U}$, for which it is necessary to carry out a further superellipsoidal approximation, has the following form:

$$B^{-1}\mathcal{U} = \operatorname{conv}\left\{ \begin{pmatrix} 0.53 \\ 5.63 \end{pmatrix}, \begin{pmatrix} -1.01 \\ 4.36 \end{pmatrix}, \begin{pmatrix} -2.82 \\ 0.26 \end{pmatrix}, \begin{pmatrix} -0.53 \\ -5.63 \end{pmatrix}, \begin{pmatrix} 1.01 \\ -4.36 \end{pmatrix}, \begin{pmatrix} 2.82 \\ -0.26 \end{pmatrix} \right\}.$$

The initial set $\mathcal{U}$ and the oriented set $B^{-1}\mathcal{U}$ are shown in Fig. 1.

5.2. Selection of Parameters of a Superellipsoidal Set

Next, we will assume that the matrix B of the orientation of the superellipse is selected in the form of a rotation matrix. Then the initial approximation problem is reduced to the following optimization problem:

$$\begin{aligned} &\mu(\mathcal{U} \setminus \mathcal{E}_r(a_1, \ldots, a_n)) \to \min_{a_1, \ldots, a_n, r}, \\ &a_i > 0, \quad i = \overline{1, n}, \\ &r > 1, \\ &\mathcal{E}_r(a_1, \ldots, a_n) \subset \mathcal{U}. \end{aligned} \tag{7}$$

We formulate a number of statements that allow us to reduce the problem (7) to an equivalent convex programming problem that can be solved numerically.

Lemma 6. *Let $\mathcal{E}_r(a)$ be defined by the relations* (1). *Then equality*

$$\mu(\mathcal{E}_r(a)) = a_1 \cdot \ldots \cdot a_n \frac{\left(2\Gamma\left(\frac{1}{r}+1\right)\right)^n}{\Gamma\left(\frac{n}{r}+1\right)}$$

is correct.

Lemma 7. *Let $\mathcal{E}_r(a)$ be defined by the relations* (1), *$\mathcal{U}$ is the convex body.*
Then the inclusion of $\mathcal{E}_r(a) \subset \mathcal{U}$ is valid if and only if the inequality

$$\left(\sum_{i=1}^{n} \left|\frac{x_i}{a_i}\right|^r\right)^{\frac{1}{r}} \geqslant M(x,\mathcal{U})$$

is true for any $x \in \mathbb{R}^n$.

Based on the Lemmas 6 and 7 we present the problem (7) in an equivalent form.

Theorem 3. *Let $\mathcal{E}_r(a)$ be defined by the relations* (1), *$\mathcal{U}$ is a convex body. Then the optimization problem* (7) *is equivalent to the following problem:*

$$\begin{aligned}
&a_1 \cdot \ldots \cdot a_n \frac{\left(2\Gamma\left(\frac{1}{r}+1\right)\right)^n}{\Gamma\left(\frac{n}{r}+1\right)} \to \max_{a_1,\ldots,a_n,r},\\
&\left(\sum_{i=1}^{n} \left|\frac{x_i}{a_i}\right|^r\right)^{\frac{1}{r}} \geqslant M(x,\mathcal{U}), \ \textit{for any } x \in \mathbb{R}^n,\\
&a_i > 0, \quad i = \overline{1,n},\\
&r > 1.
\end{aligned} \tag{8}$$

Generally speaking, (8) is not a convex programming problem, which means that in general it cannot be solved by standard optimization methods [31]. We will carry out a number of transformations that will allow us to solve (8) numerically. We will also separately consider the case when $\mathcal{U}$ is a polyhedron, which will allow us to explicitly construct the Minkowski functional $M(x,\mathcal{U})$.

Lemma 8. *Let $\mathcal{E}_r(a)$ be defined by the relations* (1), *$\mathcal{U}$ is a bounded polyhedron, i.e. there are such $K \in \mathbb{N}$, $p^1, \ldots, p^K \in \mathbb{R}^n \setminus \{0\}$, $\alpha_1, \ldots, \alpha_n > 0$, which provide representation*

$$\mathcal{U} = \bigcap_{k=1}^{K} \{x \in \mathbb{R}^n \colon (p^k, x) \leqslant \alpha_k\}.$$

Then the inclusion of $\mathcal{E}_r(a) \subset \mathcal{U}$ is equivalent to the condition

$$\left\|\operatorname{diag}(a)p^k\right\|_q \leqslant \alpha_k, \quad k = \overline{1,K}.$$

The complexity of solving the problem (8) lies in the fact that the set of acceptable values of the vector of optimization variables $(r, a_1, \ldots, a_n)^{\mathrm{T}}$ is not convex in $\mathbb{R}^{n+1}$. Nevertheless, for a fixed value of $r > 1$ the corresponding set of valid values of the vector $(a_1, \ldots, a_n)^{\mathrm{T}}$ is already convex. We formulate this fact in the form of a lemma.

Lemma 9. *Let $\mathcal{E}_r(a)$ be defined by the relations* (1), *$\mathcal{U}$ is a convex and compact body, for arbitrary $r > 1$ by $\mathcal{P}_r(\mathcal{U}) = \{a \in \mathbb{R}^n \colon \mathcal{E}_r(a) \subset \mathcal{U}, a_i > 0, \ i = \overline{1,n}\}$ we denote the set of all valid values of $a_1, \ldots, a_n$ in the problems* (7) *and* (8).

Then $\mathcal{P}_r(\mathcal{U})$ is a convex and compact set.

The Lemma 9 allows us to approximate the equivalent problems (7) and (8) with a similar optimization problem in which the domain of the parameter r is narrowed to a finite set:

$$r \in \{r_1, \ldots, r_M\} \subset (1; +\infty).$$

Then the approximation problem reduces to solving N convex programming problems of the following form:

$$\begin{aligned} &a_1 \cdot \ldots \cdot a_n \to \max_{a_1,\ldots,a_n}, \\ &(a_1, \ldots, a_n)^{\mathrm{T}} \in \mathcal{P}_r(\mathcal{U}). \end{aligned} \tag{9}$$

The choice of the resulting superellipsoidal approximation corresponding to a specific value of $r^* \in \{r_1, \ldots, r_M\}$ may be made in accordance with the Lemma 6 and the idea of maximizing the measure of the nested superellipse:

$$r^* = \arg \max_{r \in \{r_1,\ldots,r_M\}} \mu(\mathcal{E}_r(a^*(r))), \tag{10}$$

where $a^*(r) \in \mathbb{R}^n$ is the maximum point in the problem (9).

Example 3. Let's construct a superellipsoidal approximation for the set $B^{-1}\mathcal{U}$, calculated in the Example 2. To use the Lemma 8 we represent $B^{-1}\mathcal{U}$ as a bounded polyhedron:

$$B^{-1}\mathcal{U} = \bigcap_{k=1}^{6} \left\{ x \in \mathbb{R}^2 \colon (p^k, x) \leqslant \alpha_k \right\},$$

$$\left(p^1, \ldots, p^6\right) = \begin{pmatrix} -1.28 & -4.09 & -5.90 & -1.28 & -4.09 & -5.90 \\ 1.54 & 1.80 & -2.29 & -1.54 & -1.80 & 2.29 \end{pmatrix},$$

$$(\alpha_1, \ldots, \alpha_6) = (8, 12, 16, 8, 12, 16).$$

We describe the set $\mathcal{P}_r(\mathcal{U})$ for $r \in \left\{\frac{4}{3}, 2, 4\right\}$ and solve the corresponding optimization problems (9).

$$\mathcal{P}_{\frac{4}{3}}(\mathcal{U})\colon \begin{aligned} &\left(2.65a_1^4 + 5.62a_2^4\right)^{\frac{1}{4}} \leqslant 8, \\ &\left(280.53a_1^4 + 10.57a_2^4\right)^{\frac{1}{4}} \leqslant 12, \\ &\left(1208.13a_1^4 + 27.48a_2^4\right)^{\frac{1}{4}} \leqslant 16, \end{aligned} \qquad \begin{cases} a_1^*\left(\dfrac{3}{4}\right) = 2.48, \\ a_2^*\left(\dfrac{3}{4}\right) = 5.16. \end{cases}$$

$$\mathcal{P}_2(\mathcal{U})\colon \begin{aligned} &\left(1.63a_1^2 + 2.37a_2^2\right)^{\frac{1}{2}} \leqslant 8, \\ &\left(16.75a_1^2 + 3.25a_2^2\right)^{\frac{1}{2}} \leqslant 12, \\ &\left(34.76a_1^2 + 5.24a_2^2\right)^{\frac{1}{2}} \leqslant 16, \end{aligned} \qquad \begin{cases} a_1^*(2) = 1.92, \\ a_2^*(2) = 4.94. \end{cases}$$

$$\mathcal{P}_4(\mathcal{U})\colon \begin{aligned} &\left(\sqrt[3]{2.65a_1^4} + \sqrt[3]{5.62a_2^4}\right)^{\frac{3}{4}} \leqslant 8, \\ &\left(\sqrt[3]{280.53a_1^4} + \sqrt[3]{10.57a_2^4}\right)^{\frac{3}{4}} \leqslant 12, \\ &\left(\sqrt[3]{1208.13a_1^4} + \sqrt[3]{27.48a_2^4}\right)^{\frac{3}{4}} \leqslant 16, \end{aligned} \qquad \begin{cases} a_1^*(4) = 1.61, \\ a_2^*(4) = 4.16. \end{cases}$$

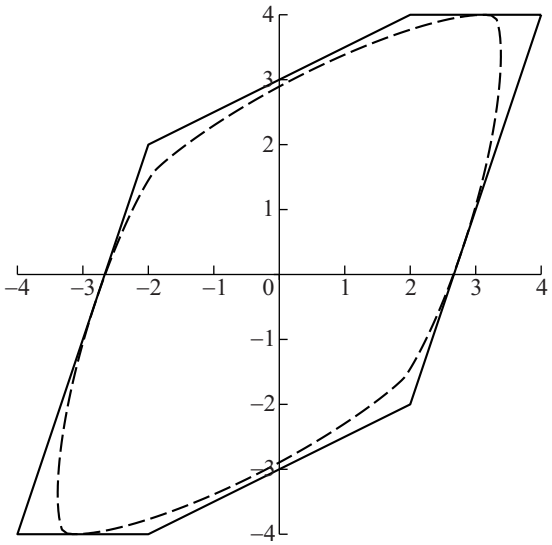


Fig. 2. The original set $\mathcal{U}$ (solid line) and its superellipsoidal approximation $B\mathcal{E}_{\frac{4}{3}}\left(a^*\left(\frac{4}{3}\right)\right)$ (dotted line).

Let us compare the obtained solutions in the sense of the Lebesgue measure of the approximating superellipse in accordance with the Lemma 6:

$$\mu\left(\mathcal{E}_{\frac{4}{3}}\left(a^*\left(\frac{4}{3}\right)\right)\right) = 32.60, \quad \mu\left(\mathcal{E}_2\left(a^*\left(2\right)\right)\right) = 29.79, \quad \mu\left(\mathcal{E}_4\left(a^*\left(4\right)\right)\right) = 24.86.$$

It follows that the best approximation of $B^{-1}\mathcal{U}$ is $\mathcal{E}_{\frac{4}{3}}\left(a^*\left(\frac{4}{3}\right)\right)$. Therefore, for the initial set $\mathcal{U}$ the most qualitative approximation of the considered ones is the set $B\mathcal{E}_{\frac{4}{3}}\left(a^*\left(\frac{4}{3}\right)\right)$. The results of the approximation are shown in Fig. 2.

6. EXAMPLES OF OPTIMAL CONTROL FORMATION

We will construct a solution to the speed problem for (2) systems of various dimensions based on the developed methods. To do this, we will use the following

Algorithm 1.

1. For a given set $\mathcal{U} \subset \mathbb{R}^n$ construct the inertia tensor $I_{\mathcal{U}}$ and calculate the orientation matrix of the superellipsoidal set $B \in \mathbb{R}^{n\times n}$ according to the Subsection 5.1.

2. Select the set of values of the superellipsoidal approximation parameter $\{r_1, \ldots, r_M\} \subset (1; +\infty)$.

3. For all $r \in \{r_1, \ldots, r_M\}$ construct optimization problems (9) for the set $B^{-1}\mathcal{U}$ and calculate the corresponding maximum points $a^*(r)$.

4. Using the Lemma 10 to determine the optimal parameter of the superellipsoidal approximation r^* according to (10).

5. For a given initial state $x_0 \in \mathbb{R}^n$ and various $N \in \mathbb{N}$ construct the systems of equations presented in the Corollary 1.

6. Determine the value of $N_{\min}$ as the smallest value of $N \in \mathbb{N}$, at which the solution of the system of equations constructed at step 5 satisfies the condition $\alpha \leqslant 1$.

7. For the value $N_{\min}$ calculated at step 6 and the corresponding $\alpha > 0$ and $\psi(0) \in \mathbb{R}^n \setminus \{0\}$ construct the optimal control $\{u^*(k)\}_{k=0}^{N_{\min}-1}$ for the system $(A, B\mathcal{E}_{r^*}(a^*(r^*)))$ according to the Theorem 1 and the Lemma 5.

Table 1. Optimal control process for a two-dimensional system

k	0	1	2	3	4	5	6	7	8	9	10
$x_1(k)$	−4.5	2.19	−3.86	3.01	−3.36	2.95	−2.82	2.54	−1.94	1.97	0
$x_2(k)$	8	−8.51	7.96	−7.86	7.28	−6.76	5.96	−4.95	3.70	−1.83	0
$u_1(k)$	3.19	0.27	2.77	−1.51	2.38	−1.95	2.21	−2.07	2.15	−2.11	–
$u_2(k)$	3.99	−2.74	3.96	−3.59	3.88	−3.75	3.83	−3.79	3.81	−3.80	–

Table 2. Results of superellipsoidal approximation for a three-dimensional system

r	$\frac{6}{5}$	$\frac{4}{3}$	2	4	6
$\mu\left(\mathcal{E}_r\left(a^*(r)\right)\right)$	57,58	61,11	57,64	41,91	35,90
$a_1^*(r)$	5,06	5,04	4,53	3,71	3,45
$a_2^*(r)$	2,48	2,24	1,58	1,20	1,10
$a_3^*(r)$	2,29	2,22	1,98	1,45	1,32

Example 4. Let $n = 2$. As $\mathcal{U}$ we choose the polyhedron considered in the Examples 2 and 3, we define the matrix of the system and the initial state as follows:

$$A = \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}, \quad x_0 = \begin{pmatrix} -4.5 \\ 8 \end{pmatrix}.$$

We can assume that the set $\mathcal{U}$ is approximated by $B\mathcal{E}_{\frac{4}{3}}\left(a^*\left(\frac{4}{3}\right)\right)$ according to the Example 3. Then the solution of the system of equations presented in the Corollary 1, for $N = 9$ has the form

$$\alpha = 1.019, \quad \psi_1(0) = 0.775, \quad \psi_2(0) = -0.632.$$

The solution obtained for $N = 10$, has the form

$$\alpha = 0.998, \quad \psi_1(0) = 0.792, \quad \psi_2(0) = -0.610.$$

From where it follows that for the auxiliary system $\left(A, B\mathcal{E}_{\frac{4}{3}}\left(a^*\left(\frac{4}{3}\right)\right)\right)$ due to (4) the equation $N_{\min} = 10$ is correct.

The optimal trajectory of the system and optimal control, calculated on the basis of the Theorem 1, are presented in Table 1.

Based on the exact methods described in [6], value $N_{\min} = 9$ was calculated for the original system $(A,\mathcal{U})$. Thus, from the point of view of control quality, the error of the guaranteeing solution is one step.

Example 5. Let $n = 3$. The set of acceptable control values, the matrix of the system and the initial state are defined as follows:

$$\mathcal{U} = \operatorname{conv}\left\{\begin{pmatrix} 4 \\ 4 \\ -3 \end{pmatrix}\begin{pmatrix} 2 \\ 4 \\ -3 \end{pmatrix}\begin{pmatrix} -2 \\ 2 \\ 0 \end{pmatrix}\begin{pmatrix} -4 \\ -4 \\ 3 \end{pmatrix}\begin{pmatrix} -2 \\ -4 \\ 3 \end{pmatrix}\begin{pmatrix} 2 \\ -2 \\ 0 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ 4 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ -4 \end{pmatrix}\right\},$$

$$A = \begin{pmatrix} 0.486 & -0.315 & 0.689 \\ -0.757 & -0.202 & 0.442 \\ 0 & -0.818 & -0.375 \end{pmatrix}, \quad x_0 = \begin{pmatrix} 26 \\ 24 \\ 30 \end{pmatrix}.$$

The inertia tensor $I_{\mathcal{U}}$ and the orientation matrix B have the form

$$I_{\mathcal{U}} = \begin{pmatrix} 526.73 & -135.75 & 132.41 \\ -135.75 & 474.79 & 164.87 \\ 132.41 & 164.87 & 439.35 \end{pmatrix}, \quad B = \begin{pmatrix} 0.49 & 0.73 & 0.48 \\ 0.59 & -0.68 & 0.43 \\ -0.64 & -0.08 & 0.76 \end{pmatrix}.$$

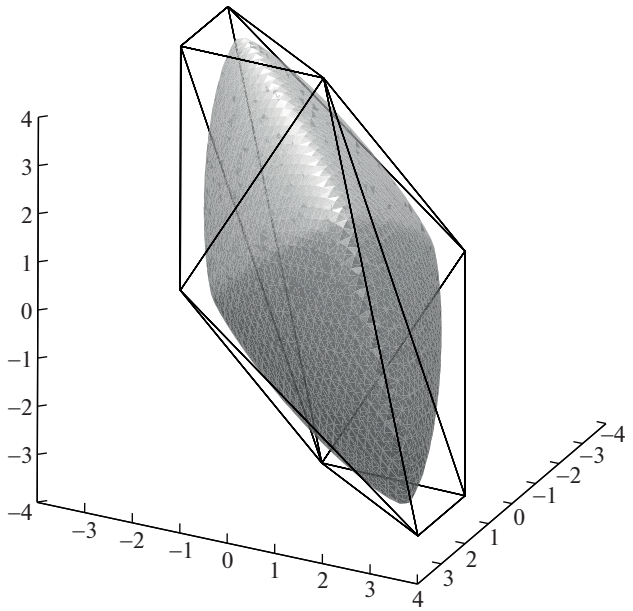


Fig. 3. The original set $\mathcal{U}$ and its super ellipsoidal approximation $B\mathcal{E}_{\frac{4}{3}}\left(a^*\left(\frac{4}{3}\right)\right)$.

The superellipsoidal approximation of the set $B^{-1}\mathcal{U}$ is carried out for $r \in \left\{\frac{6}{5}, \frac{4}{3}, 2, 4, 6\right\}$. Solutions to problems of the form (9) are presented in Table 2. It follows that the best approximation is $\mathcal{E}_{\frac{4}{3}}\left(a^*\left(\frac{4}{3}\right)\right)$. Graphically, the result of the superellipsoidal approximation of $\mathcal{U}$ by the set $B\mathcal{E}_{\frac{4}{3}}\left(a^*\left(\frac{4}{3}\right)\right)$ is shown in Fig. 3. The solution of the system of equations presented in the Corollary 1 for $N = 9$ has the form

$$\alpha = 1.038, \quad \psi_1(0) = -0.827, \quad \psi_2(0) = -0.012, \quad \psi_3(0) = -0.563.$$

The solution obtained for $N = 10$ has the form

$$\alpha = 0.890, \quad \psi_1(0) = -0.805, \quad \psi_2(0) = -0.075, \quad \psi_3(0) = -0.589.$$

It follows that for the auxiliary system $\left(A, B\mathcal{E}_{\frac{4}{3}}\left(a^*\left(\frac{4}{3}\right)\right)\right)$ due to (4) the equality $N_{\min} = 10$ is correct.

The optimal trajectory of the system and optimal control, calculated on the basis of the Theorem 1, are presented in Table 3.

Based on the exact methods described in [6], value $N_{\min} = 8$ was calculated for the original system $(A,\mathcal{U})$. Thus, from the point of view of control quality, the error of the guaranteeing solution is 2 steps.

Table 3. Optimal control process for a three-dimensional system

k	0	1	2	3	4	5	6	7	8	9	10
$x_1(k)$	26	23.64	−3.16	17.65	11.14	−0.59	9.51	3.71	1.01	2.11	0
$x_2(k)$	24	−11.76	−25.76	14.85	−10.19	−10.80	6.71	−5.75	−1.86	−0.03	0
$x_3(k)$	30	−29.28	17.93	13.24	−16.11	11.64	4.12	−6.56	4.39	0.65	0
$u_1(k)$	−2.10	1.80	−1.28	−1.88	1.88	−1.62	−1.64	1.91	−1.99	−1.48	–
$u_2(k)$	−0.48	2.70	−0.68	0.33	2.70	−1.06	0.99	2.68	−1.59	1.30	–
$u_3(k)$	1.60	−2.67	−1.13	1.01	−2.73	−0.36	0.48	−2.77	0.77	0.21	–

7. CONCLUSION

The article considers the solution of the speed-in-action problem for linear discrete-time systems with limited control. It is assumed that the set of acceptable control values is a convex compact body containing the origin, the matrix of the system is nondegenerate. For the case of strictly convex control constraints, sufficient conditions for the optimality of the control process are formulated in the form of a discrete maximum principle. At the same time, from a practical point of view, the procedure for constructing optimal control is reduced to calculating the initial conditions of the conjugate system.

A class of superellipsoidal sets, which are a generalization of the ellipsoids for normalized space, is studied in detail. In particular, the dependence of the normal cone on the support point is explicitly described, the Lebesgue measure of the superellipse in n-dimensional space is calculated. In the case when the set of admissible values of the controls of the system is a superellipsoidal set, the definition of the initial conditions of the conjugate system in the maximum principle is reduced to a system of algebraic equations with a single solution. It is essential that the dimensionality and, consequently, the complexity of the solution of this system does not depend on the optimal value of the objective function in the speed-in-action problem, but is determined only by the number of phase variables, which ensures the effectiveness of such a method in comparison with other approaches to the solution.

For systems with a general set of admissible values of controls, a superellipsoidal approximation method has been developed, which consists in constructing a superellipse of maximum measure inscribed in original convex body. The approximation procedure is divided into two stages: the selection of the orientation matrix of the superellipse and the calculation of the parameters of the superellipsoidal set. The first stage consists in calculating the inertia tensor of the approximated body, the second stage can be reduced to solving a number of convex programming problems.

The developed technique makes it possible to build optimal control processes for various discrete-time systems. Due to the general formulation of the superellipsoidal approximation problem, it is possible to generalize the discrete maximum principle, including systems with control constraints that are not strictly convex initially, for example, systems with linear constraints.

The obtained theoretical results are tested on numerical examples.

FUNDING

The work was supported by the Russian Scientific Foundation, project no. 23-21-00293.

APPENDIX

Proof of Lemma 4. Since the Minkowski functional of set (1) is a smooth function on all $\mathbb{R}^n$:

$$M(x,\mathcal{E}_r(a)=\left(\sum_{i=1}^{n}\left|\frac{x_i}{a_i}\right|^r\right)^{\frac{1}{r}},$$

then according to [24, Theorem 26.1] for an arbitrary $x\in\partial\mathcal{E}_r(a)$ the representation is correct

$$\begin{aligned}\mathcal{N}(x,\mathcal{E}_r(a))&=\operatorname{cone}\{\nabla_x M(x,\mathcal{E}_r(a))\}\setminus\{0\}\\&=\operatorname{cone}\left\{\frac{1}{r}\left(\sum_{i=1}^{n}\left|\frac{x_i}{a_i}\right|^r\right)^{\frac{1}{r}-1}\left(\frac{r|x_1|^{r-1}\operatorname{sgn}(x_1)}{|a_1|^r},\dots,\frac{r|x_n|^{r-1}\operatorname{sgn}(x_n)}{|a_n|^r}\right)^{\mathrm{T}}\right\}\setminus\{0\}\\&=\operatorname{cone}\left\{\left(\frac{|x_1|^{r-1}\operatorname{sgn}(x_1)}{|a_1|^r},\dots,\frac{|x_n|^{r-1}\operatorname{sgn}(x_n)}{|a_n|^r}\right)^{\mathrm{T}}\right\}\setminus\{0\}.\end{aligned}$$

Hence follows point 1 of the Lemma 4.

According to the definition of a normal cone, the following inclusion is true:

$$p \in \mathcal{N}(x^*(p), \mathcal{E}_r(a)).$$

Then, taking into account point 1 of Lemma 4 there will be $\alpha > 0$ such that

$$p = \alpha \left(\frac{|x_1^*(p)|^{r-1}\mathrm{sgn}(x_1^*(p))}{|a_1|^r}, \ldots, \frac{|x_n^*(p)|^{r-1}\mathrm{sgn}(x_n^*(p))}{|a_n|^r} \right)^{\mathrm{T}},$$

$$x^*(p) = \frac{1}{\alpha^{\frac{1}{r-1}}} \left(|p_1 a_1^r|^{\frac{1}{r-1}}\mathrm{sgn}(p_1), \ldots, |p_n a_n^r|^{\frac{1}{r-1}}\mathrm{sgn}(p_n) \right)^{\mathrm{T}}$$

$$= \frac{1}{\alpha^{\frac{1}{r-1}}} \left(|p_1|^{q-1} a_1^q \mathrm{sgn}(p_1), \ldots, |p_n|^{q-1} a_n^q \mathrm{sgn}(p_n) \right)^{\mathrm{T}}$$

$$= \frac{1}{\alpha^{\frac{1}{r-1}}} \mathrm{diag}(a) I_q (\mathrm{diag}(a) p).$$

The value of α can be calculated from the condition $x^*(p) \in \partial \mathcal{E}_r(a)$, which is equivalent to the equality

$$\left(\sum_{i=1}^n \left| \frac{x_i^*(p)}{a_i} \right|^r \right)^{\frac{1}{r}} = 1,$$

$$1 = \frac{1}{\alpha^{\frac{1}{r-1}}} \left(\sum_{i=1}^n \left| \frac{|p_i|^{q-1} a_i^q}{a_i} \right|^r \right)^{\frac{1}{r}} = \frac{1}{\alpha^{\frac{1}{r-1}}} \left(\sum_{i=1}^n |p_i a_i|^q \right)^{\frac{1}{r}},$$

$$\alpha^{\frac{1}{r-1}} = \left(\sum_{i=1}^n |p_i a_i|^q \right)^{\frac{1}{r}} = \|\mathrm{diag}(a) p\|_q^{q-1}.$$

The second point of the Lemma 4 is proved.

Proof of Lemma 5. Point 1 follows from point 1 of Lemma 4, point 2 of Lemma 3 and the representation

$$\mathcal{N}(u, \mathcal{U}) = \mathcal{N}(DD^{-1}u, D\mathcal{E}_r(a)).$$

Point 2 follows from point 2 of the Lemma 4 and the chain of equalities

$$\arg \max_{u \in \mathcal{U}} (p, u) = D \arg \max_{x \in \mathcal{E}_r(a)} (p, Dx) = D \arg \max_{x \in \mathcal{E}_r(a)} (D^{\mathrm{T}} p, x).$$

Lemma 5 is proved.

Proof of Theorem 2. Since $x_0 \neq 0$, then according to the definitions of the Minkowski functional and the normal cone, the conditions (5) are equivalent to the conditions

$$-\psi(0) \in \mathcal{N}\left(\frac{x_0}{\alpha}, \mathcal{X}(N_{\min}) \right), \tag{A.1}$$

$$\frac{x_0}{\alpha} \in \partial \mathcal{X}(N_{\min}). \tag{A.2}$$

The inclusion of (A.2) due to the Lemma 1 and the representation (6) is equivalent to the condition

$$\frac{x_0}{\alpha} \in \partial \left(-\sum_{k=1}^{N_{\min}} A^{-k} \mathcal{U} \right) = \partial \sum_{k=1}^{N_{\min}} A^{-k} B \mathcal{E}_r(a).$$

Then, taking into account point 1 of the Lemma 3 and the definition of the algebraic sum of sets inclusion (A.1) is equivalent to the fact that there are $x^1 \in A^{-1}B\mathcal{E}_r(a), \ldots, x^{N_{\min}} \in A^{-N_{\min}}B\mathcal{E}_r(a)$, for which the following relations are true:

$$\frac{x_0}{\alpha} = \sum_{k=1}^{N_{\min}} x^k,$$

$$-\psi(0) \in \mathcal{N}\left(\frac{x_0}{\alpha}, \mathcal{X}(N_{\min})\right) = \mathcal{N}\left(\sum_{k=1}^{N_{\min}} x^k, \sum_{k=1}^{N_{\min}} A^{-k}B\mathcal{E}_r(a)\right) = \bigcap_{k=1}^{N_{\min}} \mathcal{N}\left(x^k, A^{-k}B\mathcal{E}_r(a)\right).$$

Due to point 2 of Lemma 5 it is possible if and only if the condition

$$x^k = \frac{A^{-k}B\mathrm{diag}(a)I_q\left(-\mathrm{diag}(a)(A^{-k}B)^{\mathrm{T}}\psi(0)\right)}{\|\mathrm{diag}(a)(A^{-k}B)^{\mathrm{T}}\psi(0)\|_q^{q-1}}$$

is correct. Since $I_q(-x) = -I_q(x)$ for any $x \in \mathbb{R}^n$, we obtain equivalent relations

$$\frac{x_0}{\alpha} = \sum_{k=1}^{N_{\min}} x^k = -\sum_{k=1}^{N_{\min}} \frac{A^{-k}B\mathrm{diag}(a)I_q\left(\mathrm{diag}(a)(A^{-k}B)^{\mathrm{T}}\psi(0)\right)}{\|\mathrm{diag}(a)(A^{-k}B)^{\mathrm{T}}\psi(0)\|_q^{q-1}}.$$

That is, the conditions (5) are equivalent to the equality specified in the condition of the Theorem 2.

Proof of Corollary 1. Due to the Theorem 2 the solution of the system exists and satisfies the conditions (5). Then, due to the Lemma 1 and the symmetry of sets of the form (1) there will be such $x^1 \in \alpha A^{-1}B\mathcal{E}_r(a), \ldots, x^{N_{\min}} \in \alpha A^{-N_{\min}}B\mathcal{E}_r(a)$, which make true equality $x_0 = x^1 + \ldots x^{N_{\min}}$. From where, by point 1 of Lemma 3 it follows that any solution $(\psi(0), \alpha)$ satisfies inclusion

$$-\psi(0) \in \mathcal{N}\left(x_0, \alpha\mathcal{X}(N_{\min})\right) = \bigcap_{k=1}^{N_{\min}} \mathcal{N}\left(x^k, A^{-k}B\mathcal{E}_r(a)\right).$$

But according to point 1 of Lemma 5 for all $k = \overline{1, N_{\min}}$ sets $\mathcal{N}\left(x^k, A^{-k}B\mathcal{E}_r(a)\right)$ are one-dimensional rays with starting at 0, i.e. they contain a single vector $-\psi(0)$, satisfying the equality $(\psi(0), \psi(0)) = 1$. The uniqueness of the value $\alpha > 0$ follows from the definition of the Minkowski functional and the conditions (5).

The consequence 1 is proved.

Lemma 10. *Let $\mathcal{E}_r(a)$ be defined by the relations* (1)*. Then*

$$\mu(\mathcal{E}_r(a)) = a_1 \cdot \ldots \cdot a_n \mu(\mathcal{E}_r(1, \ldots, 1)).$$

Proof of Lemma 10. Consider replacement of variables

$$\begin{cases} x_1 = a_1 y_1, \\ \vdots \\ x_n = a_n y_n, \end{cases}$$

the Jacobian of which has the form $J = a_1 \cdot \ldots \cdot a_n$. Then

$$\mu(\mathcal{E}_r(a)) = \int\limits_{\sum\limits_{i=1}^{n}\left|\frac{x_i}{a_i}\right|^r \leqslant 1} 1dx = \int\limits_{\sum\limits_{i=1}^{n}|y_i|^r \leqslant 1} |J|dy = a_1 \cdot \ldots \cdot a_n \mu(\mathcal{E}_r(1, \ldots, 1)).$$

The Lemma 10 is proved.

Proof of Lemma 6. In the part of the space $x_i \geqslant 0$, $i = \overline{1,n}$ consider the replacement of variables

$$\begin{cases} x_1 = R(\cos\phi_2 \cdot \cos\phi_3 \ldots \cdot \cos\phi_n)^{\frac{2}{r}}, \\ x_2 = R(\sin\phi_2 \cdot \cos\phi_3 \ldots \cdot \cos\phi_n)^{\frac{2}{r}}, \\ x_3 = R(\sin\phi_3 \cdot \cos\phi_4 \ldots \cdot \cos\phi_n)^{\frac{2}{r}}, \\ \vdots \\ x_n = R(\sin\phi_n)^{\frac{2}{r}}. \end{cases} \tag{A.3}$$

$$R \geqslant 0, \quad \phi_j \in \left(0; \frac{\pi}{2}\right), \quad j = \overline{2,n}.$$

Construct a replacement Jacobian (A.3).

$$\frac{\partial x_i}{\partial R} = \frac{x_i}{R}, \quad i = \overline{1,n}, \qquad \frac{\partial x_i}{\partial \phi_j} = \begin{cases} \dfrac{2}{r}\dfrac{\cos\phi_j}{\sin\phi_j}x_i, & i = \overline{2,n}, \quad j = i, \\ -\dfrac{2}{r}\dfrac{\sin\phi_j}{\cos\phi_j}x_i, & i = \overline{1,n-1}, \quad j = \overline{i+1,n}, \\ 0, & i = \overline{3,n}, \quad j = \overline{2,i-1}, \end{cases}$$

$$J = \begin{vmatrix} \dfrac{x_1}{R} & \dfrac{x_2}{R} & \dfrac{x_3}{R} & \ldots & \dfrac{x_n}{R} \\ -\dfrac{2x_1}{r}\tan\phi_2 & \dfrac{2x_2}{r}\cot\phi_2 & 0 & \ldots & 0 \\ -\dfrac{2x_1}{r}\tan\phi_3 & -\dfrac{2x_2}{r}\tan\phi_2 & \dfrac{2x_3}{r}\cot\phi_3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\dfrac{2x_1}{r}\tan\phi_n & -\dfrac{2x_2}{r}\tan\phi_n & -\dfrac{2x_3}{r}\tan\phi_n & \ldots & \dfrac{2x_n}{r}\cot\phi_n \end{vmatrix}$$

$$= \frac{1}{R}\left(\prod_{i=1}^{n} x_i\right)\left(\prod_{j=2}^{n}\tan\phi_j\right)\left(\frac{2}{r}\right)^{n-1}\begin{vmatrix} 1 & 1 & 1 & \ldots & 1 \\ -1 & \cot^2\phi_2 & 0 & \ldots & 0 \\ -1 & -1 & \cot^2\phi_3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & -1 & \ldots & \cot^2\phi_n \end{vmatrix}$$

$$= \frac{1}{R}\left(\prod_{i=1}^{n} x_i\right)\left(\prod_{j=2}^{n}\tan\phi_j\right)\left(\frac{2}{r}\right)^{n-1}\begin{vmatrix} 1 & 1 & 1 & \ldots & 1 \\ 0 & \cot^2\phi_2 + 1 & 1 & \ldots & 1 \\ 0 & 0 & \cot^2\phi_3 + 1 & \ldots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \cot^2\phi_n + 1 \end{vmatrix}$$

$$= \frac{1}{R}\left(\prod_{i=1}^{n} x_i\right)\left(\prod_{j=2}^{n}(\tan\phi_j + \cot\phi_j)\right)\left(\frac{2}{r}\right)^{n-1} = \frac{1}{R}\left(\prod_{i=1}^{n} x_i\right)\left(\prod_{j=2}^{n}\frac{1}{\sin\phi_j\cos\phi_j}\right)\left(\frac{2}{r}\right)^{n-1}$$

$$= R^{n-1}\left(\frac{2}{r}\right)^{n-1}\prod_{j=2}^{n}(\sin\phi_j)^{\frac{2}{r}-1}(\cos\phi_j)^{\frac{2}{r}(j-1)-1}.$$

Then we can calculate the Lebesgue measure of the superellipse $\mathcal{E}_r(1,\dots,1)$ via the Lebesgue integral:

$$\mu(\mathcal{E}_r(1,\dots,1)) = \int\limits_{\sum\limits_{i=1}^{n}|x_i|^r\leqslant 1} 1dx = 2^n \int\limits_0^1 R^{n-1}\left(\frac{2}{r}\right)^{n-1} dR \prod_{j=2}^{n}\int\limits_0^{\frac{\pi}{2}} (\sin\phi_j)^{\frac{2}{r}-1}(\cos\phi_j)^{\frac{2}{r}(j-1)-1}d\phi_j.$$

For each $j=\overline{2,n}$ we calculate auxiliary integrals:

$$\int\limits_0^{\frac{\pi}{2}} (\sin\phi_j)^{\frac{2}{r}-1}(\cos\phi_j)^{\frac{2}{r}(j-1)-1}\,d\phi_j = \int\limits_0^{\frac{\pi}{2}} (\sin\phi_j)^{\frac{2}{r}-1}(\cos\phi_j)^{\frac{2}{r}(j-1)-2}\,d\sin\phi_j$$

$$= \int\limits_0^{\frac{\pi}{2}} (\sin\phi_j)^{\frac{2}{r}-1}\left(1-\sin^2\phi_j\right)^{\frac{j-1}{r}-1} d\sin\phi_j = \frac{1}{2}\int\limits_0^{\frac{\pi}{2}} \left(\sin^2\phi_j\right)^{\frac{1}{r}-1}\left(1-\sin^2\phi_j\right)^{\frac{j-1}{r}-1} d\sin^2\phi_j$$

$$= \frac{1}{2}\int\limits_0^1 t^{\frac{1}{r}-1}(1-t)^{\frac{j-1}{r}-1}dt = \frac{1}{2}\mathrm{B}\left(\frac{1}{r},\frac{j-1}{r}\right),$$

where $\mathrm{B}(x,y)$ denotes the Euler beta function.

Then the original integral has the form

$$\mu(\mathcal{E}_r(1,\dots,1)) = \frac{2^n}{n}\left(\frac{2}{r}\right)^{n-1}\prod_{j=2}^{n}\left(\frac{1}{2}\mathrm{B}\left(\frac{1}{r},\frac{j-1}{r}\right)\right) = \frac{2}{n}\left(\frac{2}{r}\right)^{n-1}\prod_{j=2}^{n}\frac{\Gamma\left(\frac{1}{r}\right)\Gamma\left(\frac{j-1}{r}\right)}{\Gamma\left(\frac{1}{r}+\frac{j-1}{r}\right)}$$

$$= \frac{2}{n}\left(\frac{2}{r}\Gamma\left(\frac{1}{r}\right)\right)^{n-1}\prod_{j=1}^{n-1}\frac{\Gamma\left(\frac{j}{r}\right)}{\Gamma\left(\frac{j+1}{r}\right)} = \frac{2}{n}\left(\frac{2}{r}\Gamma\left(\frac{1}{r}\right)\right)^{n-1}\frac{\Gamma\left(\frac{1}{r}\right)}{\Gamma\left(\frac{n}{r}\right)} = \frac{\left(\frac{2}{r}\Gamma\left(\frac{1}{r}\right)\right)^{n}}{\frac{n}{r}\Gamma\left(\frac{n}{r}\right)} = \frac{\left(2\Gamma\left(\frac{1}{r}+1\right)\right)^{n}}{\Gamma\left(\frac{n}{r}+1\right)}.$$

Taking into account the Lemma 10 we finally obtain the equality

$$\mu(\mathcal{E}_r(a)) = a_1\cdot\ldots\cdot a_n\frac{\left(2\Gamma\left(\frac{1}{r}+1\right)\right)^{n}}{\Gamma\left(\frac{n}{r}+1\right)}.$$

Lemma 6 is proved.

Lemma 11. *Let $\mathcal{U}_1,\mathcal{U}_2\subset\mathbb{R}^n$ be convex and compact bodies containing 0 as an internal point. In this case, the inclusion $\mathcal{U}_1\subset\mathcal{U}_2$ is true if and only if the following inequality is correct for any $x\in\mathbb{R}^n$:*

$$M(x,\mathcal{U}_1)\geqslant M(x,\mathcal{U}_2).$$

Proof of Lemma 11. Let $\mathcal{U}_1\subset\mathcal{U}_2$, $x\in\mathbb{R}^n$. Then by definition of the Minkowski functional

$$x\in M(x,\mathcal{U}_1)\mathcal{U}_1\subset M(x,\mathcal{U}_1)\mathcal{U}_2,$$
$$M(x,\mathcal{U}_1)\geqslant\inf\{t>0\colon x\in t\mathcal{U}_2\}=M(x,\mathcal{U}_2).$$

Let for all $x\in\mathbb{R}^n$ inequality be fair

$$M(x,\mathcal{U}_1)\geqslant M(x,\mathcal{U}_2).$$

Then by definition of the Minkowski functional

$$\mathcal{U}_1 = \{x \in \mathbb{R}^n \colon M(x,\mathcal{U}_1) \leqslant 1\} \subset \{x \in \mathbb{R}^n \colon M(x,\mathcal{U}_2) \leqslant 1\} = \mathcal{U}_2.$$

The Lemma 11 is proved.

Proof of Lemma 7. Lemma 7 follows directly from Lemma 11 and the fact that

$$M(x, \mathcal{E}_r(a)) = \left(\sum_{i=1}^{n} \left| \frac{x_i}{a_i} \right|^r \right)^{\frac{1}{r}}.$$

Lemma 7 is proved.

Proof of Theorem 3. Due to the Lemma 7 the inclusion $\mathcal{E}_r(a) \subset \mathcal{U}$ is equivalent to the condition

$$\left(\sum_{i=1}^{n} \left| \frac{x_i}{a_i} \right|^r \right)^{\frac{1}{r}} \geqslant M(x,\mathcal{U}) \text{ for any } x \in \mathbb{R}^n.$$

Also, due to this limitation, it is true that

$$\mu(\mathcal{U} \setminus \mathcal{E}_r(a)) = \mu(\mathcal{U}) - \mu(\mathcal{E}_r(a)).$$

Hence, taking into account the fact that the value of $\mu(\mathcal{U})$ does not depend on optimization variables, the statement of the Theorem 3 follows.

Lemma 12. *Let there be* $p^1, \ldots, p^K \in \mathbb{R}^n \setminus \{0\}$ *and* $\alpha_1, \ldots, \alpha_K > 0$ *such that*

$$\mathcal{U} = \bigcap_{k=1}^{K} \left\{ u \in \mathbb{R}^n \colon (p^k, u) \leqslant \alpha_k \right\}, \quad 0 \in \operatorname{int} \mathcal{U}.$$

Then

$$M(x,\mathcal{U}) = \max_{k=\overline{1,K}} \frac{(p^k, x)}{\alpha_k}.$$

Proof of Lemma 12. Since for any $t > 0$

$$t\mathcal{U} = \left\{ u \in \mathbb{R}^n \colon u = tx, \ \ x \in \mathcal{U} \right\} = \left\{ u \in \mathbb{R}^n \colon u = tx, \ \ (p^k, x) \leqslant \alpha_k, \ \ k = \overline{1,K} \right\}$$
$$= \left\{ u \in \mathbb{R}^n \colon \ \left(p^k, \frac{u}{t} \right) \leqslant \alpha_k, \ \ k = \overline{1,K} \right\} = \left\{ u \in \mathbb{R}^n \colon \ (p^k, u) \leqslant t\alpha_k, \ \ k = \overline{1,K} \right\}$$
$$= \bigcap_{k=1}^{K} \left\{ u \in \mathbb{R}^n \colon (p^k, u) \leqslant t\alpha_k \right\},$$

then according to the definition of the Minkowski functional

$$M(x,\mathcal{U}) = \inf\{t > 0 \colon \ x \in t\mathcal{U}\} = \inf \left\{ t > 0 \colon \ (p^k, x) \leqslant t\alpha_k, \ \ k = \overline{1,K} \right\}$$
$$= \inf \left\{ t > 0 \colon \ t \geqslant \frac{(p^k, x)}{\alpha_k}, \ \ k = \overline{1,K} \right\} = \max_{k=\overline{1,K}} \frac{(p^k, x)}{\alpha_k}.$$

The Lemma 12 is proved.

Proof of Lemma 8. According to Lemmas 7 and 12, the inclusion of $\mathcal{E}_r(a) \subset \mathcal{U}$ is equivalent to the fact that for all $x \in \mathbb{R}^n$ the following inequality is valid:

$$\left(\sum_{i=1}^{n} \left|\frac{x_i}{a_i}\right|^r\right)^{\frac{1}{r}} \geqslant \max_{k=\overline{1,K}} \frac{(p^k, x)}{\alpha_k}.$$

For $x = 0$, this inequality holds trivially. Consider the case of $x \neq 0$ and move on to equivalent inequalities. For all $k = \overline{1, K}$

$$\left(\sum_{i=1}^{n} \left|\frac{x_i}{a_i}\right|^r\right)^{\frac{1}{r}} \geqslant \frac{(p^k, x)}{\alpha_k},$$

$$\alpha_k \geqslant \frac{(p^k, x)}{\left(\sum\limits_{i=1}^{n} \left|\frac{x_i}{a_i}\right|^r\right)^{\frac{1}{r}}}.$$

Since these inequalities must be satisfied for any $x \in \mathbb{R}^n \setminus \{0\}$, it is possible, taking into account the Lemma 4, to proceed to the equivalent relation

$$\alpha_k \geqslant \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{(p^k, x)}{\left(\sum\limits_{i=1}^{n} \left|\frac{x_i}{a_i}\right|^r\right)^{\frac{1}{r}}} = \max_{x \in \mathbb{R}^n \setminus \{0\}} \left(p^k, \frac{x}{\left(\sum\limits_{i=1}^{n} \left|\frac{x_i}{a_i}\right|^r\right)^{\frac{1}{r}}}\right)$$

$$= \max_{y \in \partial \mathcal{E}_r(a)} \left(p^k, y\right) = \left(p^k, x^*(p^k)\right) = \frac{\left(p^k, \operatorname{diag}(a) I_q \left(\operatorname{diag}(a) p^k\right)\right)}{\|\operatorname{diag}(a) p^k\|_q^{q-1}} = \left\|\operatorname{diag}(a) p^k\right\|_q.$$

The Lemma 8 is fully proved.

Proof of Lemma 9. Denote for arbitrary convex sets $\mathcal{U}$ and $p \in \mathbb{R}^n \setminus \{0\}$ via $s(p, \mathcal{U})$ support function $\mathcal{U}$:

$$s(p, \mathcal{U}) = \sup_{x \in \mathcal{U}} (p, x).$$

As demonstrated in [24, Theorem 11.5], an arbitrary convex compact set $\mathcal{U}$ is the intersection of all support half-spaces:

$$\mathcal{U} = \bigcap_{p \in \mathbb{R}^n \setminus \{0\}} \left\{x \in \mathbb{R}^n \colon (p, x) \leqslant s(p, \mathcal{U})\right\}.$$

Then the inclusion $\mathcal{E}_r(a) \subset \mathcal{U}$ is equivalent to the fact that for every $p \in \mathbb{R}^n \setminus \{0\}$ the following inequality will be satisfied

$$s(p, \mathcal{E}_r(a)) \leqslant s(p, \mathcal{U}). \tag{A.4}$$

Let $a, b \in \mathcal{P}_r(\mathcal{U})$, $\lambda \in (0; 1)$, $p \in \mathbb{R}^n \setminus \{0\}$. Then, due to point 2 of the Lemma 4 and the Minkowski inequality [25, section 1 §1 ch. II] folowing relations are correct:

$$s(p, \mathcal{E}_r(\lambda a + (1-\lambda) b)) = \max_{x \in \mathcal{E}_r(\lambda a + (1-\lambda) b)} (p, x) = \left(\sum_{i=1}^{n} |(\lambda a_i + (1-\lambda) b_i) p_i|^q\right)^{\frac{1}{q}}$$

$$\leqslant \lambda \left(\sum_{i=1}^{n} |a_i p_i|^q\right)^{\frac{1}{q}} + (1-\lambda) \left(\sum_{i=1}^{n} |b_i p_i|^q\right)^{\frac{1}{q}} = \lambda s(p, \mathcal{E}_r(a)) + (1-\lambda) s(p, \mathcal{E}_r(b)) \leqslant s(p, \mathcal{U}).$$

Then the condition $\mathcal{E}_r(\lambda a + (1-\lambda)b) \subset \mathcal{U}$ is correct, which by definition is equivalent to inclusion $\lambda a + (1-\lambda)b \in \mathcal{P}_r(\mathcal{U})$. This implies the convexity of $\mathcal{P}_r(\mathcal{U})$.

Choose as $p \in \mathbb{R}^n \setminus \{0\}$ the i-th coordinate vector:

$$p = (\underbrace{0, \ldots, 0}_{i-1}, 1, 0, \ldots, 0)^{\mathrm{T}}.$$

Then by construction it is correct that

$$s(\pm p, \mathcal{E}_r(a)) = a_i.$$

Taking into account the condition (A.4), we obtain that for any $a \in \mathcal{P}_r(\mathcal{U})$ the following inequality is correct:

$$0 \leqslant a_i \leqslant \min\{s(p,\mathcal{U}), s(-p,\mathcal{U})\}.$$

Since $\mathcal{U}$ is limited, then for any $p \in \mathbb{R}^n \setminus \{0\}$ the value of the support function $s(p,\mathcal{U})$ is finite. Then $\mathcal{P}_r(\mathcal{U})$ is limited.

The closeness of $\mathcal{P}_r(\mathcal{U})$ follows from the closeness of $\mathcal{U}$.

The Lemma 9 is proved.

REFERENCES

1. Pontryagin, L.S., Boltyansky, V.G., Gamkrelidze, R.V., and Mishchenko, B.F., *Matematicheskaya teoriya optimal'nykh protsessov* (Mathematical Theory of Optimal Processes), Moscow: Nauka, 1969.
2. Boltyanskii, V.G., *Optimal'noe upravlenie diskretnymi sistemami* (Optimal Control of Discrete Systems), Moscow: Nauka, 1973.
3. Propoi, A.I., *Elementy teorii optimal'nykh diskretnykh protsessov* (Elements of the Theory of Optimal Discrete Processes), Moscow: Nauka, 1973.
4. Ibragimov, D.N. and Sirotin, A.N., On the Problem of Operation Speed for the Class of Linear Infinite-Dimensional Discrete-Time Systems with Bounded Control, *Autom. Remote Control*, 2017, vol. 78, no. 10, pp. 1731–1756. https://doi.org/10.1134/S0005117917100010
5. Bellman, R., *Dinamicheskoe programmirovanie* (Dynamic Programming), Moscow: Inostrannaya Literatura, 1960.
6. Ibragimov, D.N., Novozhilin, N.M., and Portseva, E.Yu., On Sufficient Optimality Conditions for a Guaranteed Control in the Speed Problem for a Linear Time-Varying Discrete-Time System with Bounded Control, *Autom. Remote Control*, 2021, vol. 82, no. 12, pp. 2076–2096. https://doi.org/10.1134/S000511792112002X
7. Kamenev, G.K., *Chislennoe issledovanie effektivnosti metodov poliedral'noi approksimatsii vypuklykh tel* (Numerical Study of the Efficiency of Polyhedral Approximation Methods for Convex Bodies), Moscow: Vychisl. Tsentr Ross. Akad. Nauk, 2010.
8. Ibragimov, D.N., The minimum-time correction of the satellite's orbit, *Elektron. Zh. Tr. MAI*, 2017, no. 94. http://trudymai.ru/published.php
9. Kurzhanskiy A. and Varaiya P., Ellipsoidal Techniques for Reachability Analysis of Discrete-Time Linear Systems, *IEEE Transactions on Automatic Control*, 2007, vol. 52, no. 1, pp. 26–38. https://doi.org/10.1109/TAC.2006.887900
10. Chernous'ko, F.L., *Otsenivanie fazovogo sostoyaniya dinamicheskikh sistem. Metod ellipsoidov* (Phase State Estimation of Dynamic Systems. The Ellipsoid Method), Moscow: Nauka, 1988.
11. Gridgeman, N.T., Lame Ovals, *The Mathematical Gazette*, 1970, vol. 54, no. 387, pp. 31–37. https://doi.org/10.2307/3613154

12. Tobler, W.R., The Hyperelliptical and Other New Pseudo Cylindrical Equal Area Map Projections, *J. Geophys. Res.*, 1973, vol. 78, no. 11, pp. 1753–1759. https://doi.org/10.1029/JB078i011p01753

13. Shi, P.J., Huang, J.G., Hui, C., Grissino-Mayer, H.D., Tardif, J.C., Zhai, L.H., Wang, F.S., and Li, B.L., Capturing Spiral Radial Growth of Conifers Using the Superellipse to Model Tree-Ring Geometric Shape, *Frontiers in Plant Science*, 2015, vol. 6, no. 856, pp. 1–13. https://doi.org/10.3389/fpls.2015.00856

14. Gielis, J., A Generic Geometric Transformation That Unifies a Wide Range of Natural and Abstract Shapes, *Amer. J. Botany*, 2003, vol. 90, no. 3, pp. 333–338. https://doi.org/10.3732/ajb.90.3.333

15. Maximidis, R.T., Caratelli, D., Toso, G., and Smolders, B.A., Analysis of a Novel Class of Waveguiding Structures Suitable for Reactively Loaded Antenna Array Design, *Doklady TUSUR*, 2017, no. 1, pp. 10–13. https://doi.org/10.21293/1818-0442-2017-20-1-09-13

16. Zolotenkova, M.K. and Egorov, V.V., Development and Analysis of Ultrasound Registrating and Performing Rodent Vocalization Device, *IEEE-EDM*, 2022, pp. 506–509. https://doi.org/10.1109/EDM55285.2022.9855056

17. Sadowski, A.J., Geometric Properties for the Design of Unusual Member Cross-Sections in Bending, *Engineering Structures*, 2011, vol. 33, no. 5, pp. 1850–1854. https://doi.org/10.1016/j.engstruct.2011.01.026

18. Tobler, W.R., Superquadrics and Angle-Preserving Transformations, *IEEE-CGA*, 1981, vol. 1, no. 1, pp. 11–23. https://doi.org/10.1109/MCG.1981.1673799

19. Desoer, C.A. and Wing, J., The Minimal Time Regulator Problem for Linear Sampled-Data Systems: General Theory, *J. Franklin Inst.*, 1961, vol. 272, no. 3, pp. 208–228. https://doi.org/10.1016/0016-0032(61)90784-0

20. Lin, W.-S., Time-Optimal Control Strategy for Saturating Linear Discrete Systems, *Int. J. Control*, 1986, vol. 43, no. 5, pp. 1343–1351. https://doi.org/10.1080/00207178608933543

21. Moroz, A.I., Synthesis of Time-Optimal Control for Linear Discrete Objects of the Third Order, *Autom. Remote Control*, 1965, vol. 25, no. 9, pp. 193–206.

22. Krasnoshchechenko, V.I., Simplex Method for Solving the Brachistochroneproblem at State and Control Constraints, *Inzhenernyi zhurnal: nauka i innovatsii*, 2014, no. 6. http://engjournal.ru/catalog/it/asu/1252.html

23. Cazanova, L.A., Stability of optimal synthesis in the time-optimality problem, *Izvestiya vuzov, Matematika*, 2002, no. 2, pp. 46–57.

24. Rockafellar, R., *Vypuklyi analiz* (Convex Analysis), Moscow: Mir, 1973.

25. Kolmogorov, A.N. and Fomin, S.V., *Elementy teorii funktsii i funktsional'nogo analiza* (Elements of the Theory of Functions and Functional Analysis), Moscow: Fizmatlit, 2012.

26. Berendakova, A.V. and Ibragimov, D.N., About the Method for Constructing External Estimates of the Limit 0-Controllability Set for the Linear Discrete-Time System with Bounded Control, *Autom. Remote Control*, 2023, vol. 84, no. 2, pp. 83–104. https://doi.org/10.1134/S0005117923020030

27. Ibragimov, D.N., On the Optimal Speed Problem for the Class of Linear Autonomous Infinite-Dimensional Discrete-Time Systems with Bounded Control and Degenerate Operator, *Autom. Remote Control*, 2019, vol. 80, no. 3, pp. 393–412. https://doi.org/10.1134/S0005117919030019

28. Ostrowski, A.M., *Reshenie uravnenii i sistem uravnenii* (Solution of equations and systems of equations), Moscow: Izdatel'stvo inostrannoi literatury, 1963.

29. Landau, L.D. and Lifshitz, E.M., *Mekhanika* (Mechanics), Moscow: Nauka, 1988.

30. Horn, R. and Johnson, C., *Matrichnyi analiz* (Matrix Analysis), Moscow: Mir, 1989.

31. Ashmanov, S.A. and Timohov, S.V., *Teoriya optimizatsii v zadachakh i uprazhneniyakh* (Optimization theory in problems and exercises), Moscow: Nauka, 1991.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 9, pp. 1065–1074.

Static Feedback Design in Linear Discrete-Time Control Systems Based on Training Examples

V. A. Mozzhechkov

Tula State University, Tula, Russia
e-mail: v.a.moz@yandex.ru

Received November 14, 2022
Revised June 21, 2023
Accepted July 20, 2023

Abstract—The problem of static feedback design in linear discrete time-invariant control systems is considered. The desired behavior of the system is defined by a set of its output variation laws (training examples) and by a requirement to the degree of its stability. Controller's structural constraints are taken into account. Explicit relations are obtained and an iterative method based on these relations is proposed to find a good initial approximation of the desired gain matrix and to refine it sequentially. In the general case, simple-structure gain matrices are found: in such matrices, only those components are nonzero that are necessary and sufficient to give the system the desired properties. Some examples are provided to illustrate the method.

Keywords: linear discrete-time control systems, feedback, design

DOI: 10.25728/arcRAS.2023.79.83.001

1. INTRODUCTION

A considerable number of works are devoted to the design of static feedback in linear control systems. As a rule, the desired behavior of the system is defined by requiring that the roots of its characteristic polynomial belong to some value set or by minimizing an integral quadratic functional that assesses the quality of transients. Accordingly, the problems under consideration are placing the poles of the transfer function of a closed loop system (modal control) and designing a linear quadratic controller (LQR). There exist [1] effective methods for solving them exactly provided that all components of the state vector can be used in the controller and no explicit constraints are imposed on the choice of gain coefficients. However, these problems turn out to be intractable in the case of controller's structural constraints [2, 3], particularly under the unavailability of some state variables (e.g., when designing output feedback). In such a case, pole placement is an NP-hard problem [2, 4] that often reduces to a nonsmooth and nonconvex optimization problem in the space of controller's parameters [2, 5]. Necessary and sufficient conditions for the existence of a solution were established for this problem [6–9], but it was not possible to develop methods for obtaining an exact solution [2, 3]. At the same time, algorithms were proposed to calculate an approximate solution. A significant part of them involve Lyapunov functions for the design of stabilizing controllers and the reduction of the original problem to nonlinear matrix inequalities by repeatedly solving linear matrix inequalities (LMIs) during iterative refinement of the desired solution [9–13]. The papers [14–16] investigated the possibility of using the LMI technique to consider the sparse feedback design requirements that limit freedom in choosing the controller structure. Along with the ones mentioned above, algorithms were proposed to design stabilizing output-feedback controllers by minimizing the spectral abscissa of a closed loop system by its

direct calculation and solving the corresponding nonlinear programming problem based on methods that take into account the peculiarities of the design problem [2, 3]. The algorithms presented in [8, 17, 18] involve external algebra methods to find an initial approximation of the desired output-feedback gain matrix for the modal control problem; this approximation is then refined iteratively. For the LQR problem with output feedback, necessary conditions for the existence of a solution in the form of a system of nonlinear matrix equations were obtained [19, 20] and corresponding iterative algorithms for the approximate solution of this problem were proposed [20–24]. Numerical methods for solving the LQR problem with a sparse feedback matrix based on the LMI technique were considered in [14–16, 25]; for the first time, such a problem was solved in [26] by reducing to a nonlinear discrete programming problem. However, the algorithms mentioned do not ensure an exact solution and are heuristic: their convergence was not proved rigorously.

The problem considered below essentially differs from classical static feedback design problems as follows. The desired behavior of the system is defined by a set of its output variation laws, acting as training examples. They can be trajectories corresponding, e.g., to a feedback control law that should be simplified using a simpler controller in the designed system (in particular, a system with state feedback can be a source of training examples for output feedback design) or to a program control law or human control that should be implemented in the designed system based on a feedback control law. Together with the closeness of the system trajectories to the trajectories given as training examples, the requirement to ensure a given degree of its stability is considered. In addition, the constraints imposed on the feedback structure are taken into account. They can be expressed as the requirement to use output feedback, the requirement that some elements of the gain matrix be zero, and the requirement to eliminate its structural redundancy. The latter is equivalent to obtaining a simple-structure gain matrix [27–31]: in such matrices, only those components are nonzero that are necessary and sufficient to give the system the desired properties. The goal of design is to approximate the system behavior to the desired one by choosing the elements and structure of the gain matrix. This problem statement is novel and has not been considered in the works devoted to controller design, including those involving machine learning methods [32–35].

In this paper, we derive explicit relations and propose a corresponding iterative method to find a good initial approximation of the desired gain matrix and to refine it sequentially. The novel method allows designing all possible simple-structure gain matrices.

2. PROBLEM STATEMENT

Consider a control system described by the equations

$$x_{k+1} = Ax_k + Bu_k, \tag{1}$$

$$y_k = Cx_k, \tag{2}$$

$$u_k = Ky_k, \tag{3}$$

where k denotes discrete time from the set of natural numbers; x_k, y_k, and u_k are the state, output, and control vectors, respectively; the components of the vectors x_k, y_k, and u_k as well as the elements of constant matrices A, B, C, and K are real numbers; the controller's gain matrix K has to be determined, the other matrices are supposed given.

Consider structural constraints imposed on the controller (3). They are usually reduced [2, 14, 15, 26] to requiring zero value for some elements of the gain matrix $K = (k_{i,j})$. Therefore, we introduce the condition

$$k_{i,j} = 0, \ \forall (i,j) \notin \check{S}, \tag{4}$$

where $\check{S}$ is the set of number pairs (i,j) for the elements of the gain matrix K that are not required to be zero.

We define the desired behavior of system (1)–(4) by specifying the corresponding desired trajectories $Y_\gamma = (y_k^\gamma)$, $k \in \{\overline{1,N}\}$, of the output (2) of system (1)–(4) for some set of initial conditions x_0^γ, $\gamma \in \{\overline{1,q}\}$.

In other words, we define a set

$$Q = \{(x_0^\gamma, Y_\gamma)\}, \quad \gamma \in \{\overline{1,q}\}, \tag{5}$$

in which the pairs (x_0^γ, Y_γ) are training examples.

In system (1)–(4), perfectly matching the desired behavior, the equality $y(x_0^\gamma, K)_k = y_k^\gamma$ holds for the initial conditions $x(0) = x_0^\gamma$ at each time instant $k \in \{\overline{1,N}\}$. Let us require this condition for each pair $(x_0^\gamma, Y_\gamma) \in Q$, i.e.,

$$y(x_0^\gamma, K)_k = y_k^\gamma, \ \forall k \in \{\overline{1,N}\}, \ \forall \gamma \in \{\overline{1,q}\}. \tag{6}$$

The possibility that (6) is satisfied approximately will be described as follows:

$$\varepsilon_k^{\gamma-} \le y(x_0^\gamma, K)_k - y_k^\gamma \le \varepsilon_k^{\gamma+}, \ \forall k \in \{\overline{1,N}\}, \ \forall \gamma \in \{\overline{1,q}\}, \tag{7}$$

where $\varepsilon_k^{\gamma-}$ and $\varepsilon_k^{\gamma+}$ are given constant vectors.

Generally, conditions (7) do not ensure the stability of system (1)–(4). Therefore, together with (7), we require the necessary degree of Schur stability for the matrix $A_c = A + BKC$ of the closed loop system (1)–(4), i.e.,

$$\rho(A_c(K)) \le 1 - \sigma, \tag{8}$$

where $\rho(A_c(K))$ denotes the spectral radius of the matrix $A_c(K)$ and σ is a given degree of stability.

Let the matrix K be chosen through the best approximation of the behavior of system (1)–(4) to the desired one by minimizing the Euclidean norm of the vector $\Delta y(K)$ composed of the residuals $y(x_0^\gamma, K)_k - y_k^\gamma$ of all equations (6):

$$|\Delta y(K)| \to \min_K. \tag{9}$$

In the case of a given structure of the controller (a fixed set $\check{S}$ defining its structure), the problem under consideration is to find the matrix K in system (1)–(4) that satisfies the requirements (7)–(9).

In general, we will solve the structural design problem: determine all sets $\check{S}$ and the corresponding matrices K for which conditions (7)–(9) hold and the structure of the controller (3), (4) is simple. This means [27–31] that only those components of the matrix K are nonzero that are necessary and sufficient to give system (1)–(4) the desired properties. Formally, the problem of determining a set Ω of simple structures of the controller (3), (4) consists in the following: find admissible structures $\check{S} \in \zeta$ for which a less complex admissible structure cannot be specified. (A structure $\check{S}'$ is considered simpler than $\check{S}$ if $\check{S}' \subset \check{S}$.) In other words, it is required to find

$$\Omega = \left\{\check{S} \in \zeta \mid \{\check{S}' \in \zeta \mid \check{S}' \subset \check{S}\} = \varnothing\right\}, \tag{10}$$

where ζ denotes the set of admissible structures, i.e., those for which there exists a matrix K satisfying conditions (1)–(4) and (7)–(9). The formula $\{\check{S}' \in \zeta \mid \check{S}' \subset \check{S}\} = \varnothing$ indicates the absence of an admissible structure $\check{S}'$ simpler than a structure $\check{S} \in \Omega$.

3. ANALYSIS OF THE PROBLEM

Given x_0^γ and K, the solution of system (1)–(3) can be written [1, p. 20] as follows:

$$y(x_0^\gamma, K)_k = CA^k x_0^\gamma + C\sum_{i=0}^{k-1} A^{k-i-1}BKy(x_0^\gamma, K)_i,\ \forall k \in \{\overline{1,N}\}. \tag{11}$$

In view of (11), condition (6) is equivalent to the system of equations

$$CA^k x_0^\gamma + C\sum_{i=0}^{k-1} A^{k-i-1}BKy(x_0^\gamma, K)_i = y_k^\gamma,\ \forall k \in \{\overline{1,N}\},\ \forall \gamma \in \{\overline{1,q}\}. \tag{12}$$

Applying identity transformations yields the system

$$CA^k x_0^\gamma + C\sum_{i=0}^{k-1} \left(y(x_0^\gamma, K)_i^{\mathrm{T}} \otimes A^{k-i-1}B\right)\mathrm{vec}(K) = y_k^\gamma,\ \ \forall k \in \{\overline{1,N}\},\ \forall \gamma \in \{\overline{1,q}\}, \tag{13}$$

where $\otimes$ denotes the Kronecker product [36, p. 83] and $\mathrm{vec}(\cdot)$ is the vectorization function [36]. (It produces a column vector by the successive connection of all columns of the argument matrix.) We write system (13) as

$$Y_{0\gamma} + G_\gamma(K)\mathrm{vec}(K) = Y_\gamma,\ \forall \gamma \in \{\overline{1,q}\}, \tag{14}$$

where $Y_{0\gamma}$, Y_γ, and $G_\gamma(K)$ are the column vectors composed of the blocks $CA^k x_0^\gamma$, y_k^γ, and $G_{k\gamma}(K) = C\sum_{i=0}^{k-1}\left(y(x_0^\gamma, K)_i^{\mathrm{T}} \otimes A^{k-i-1}B\right)$, respectively, $k \in \{\overline{1,N}\}$.

From (14) and (4) it follows that

$$G_\gamma(K)_S\mathrm{vec}(K)_S = \hat{Y}_\gamma,\ \forall \gamma \in \{\overline{1,q}\}, \tag{15}$$

where the matrix $G_\gamma(K)_S$ and the vector $\mathrm{vec}(K)_S$ contain the columns of the matrix $G_\gamma(K)$ and the coordinates of the vector $\mathrm{vec}(K)$, respectively, whose numbers are specified in the set S. (In accordance with the set $\check{S}$, the former set determines the numbers of the coordinates of the vector $\mathrm{vec}(K)$ that are not required to be zero.) In addition, $\hat{Y}_\gamma = Y_\gamma - Y_{0\gamma}$.

Let all the desired trajectories $Y_\gamma = (y_k^\gamma)$, $k \in \{\overline{1,N}\}$, $\gamma \in \{\overline{1,q}\}$, belong to the set of solutions of system (1)–(4). Then $y(x_0^\gamma, K)_i$ in the expressions (12), (13) can be replaced by y_i^γ; as a result, the matrix $G_\gamma(K)$ in (15) becomes constant and independent of the desired unknown matrix K. In this case, system (15) can be represented as

$$\bar{G}_{\gamma S}\mathrm{vec}(K)_S = \hat{Y}_\gamma,\ \forall \gamma \in \{\overline{1,q}\}, \tag{16}$$

where $\bar{G}_\gamma$ is the column vector of the blocks $\bar{G}_{k\gamma} = C\sum_{i=0}^{k-1}\left(y^{\gamma\,\mathrm{T}}_i \otimes A^{k-i-1}B\right)$, $k \in \{\overline{1,N}\}$.

Proposition 1. *System* (1)–(4) *perfectly matches the desired behavior given by the set of training examples* (5)*, i.e., the requirement* (6) *holds, if and only if all the desired trajectories* Y_γ, $\gamma \in \{\overline{1,q}\}$, *belong to the set of solutions of system* (1)–(4) *and the matrix* K *given* (4) *is the solution of the system of linear equations* (16)*.*

The proof of Proposition 1 is postponed to the Appendix.

According to Proposition 1, the feasibility of system (16) is a necessary and sufficient condition for equalities (6), i.e., a condition for the exact reproduction of all training examples by the designed system.

Due to the equivalence of equations (6) and (15), conditions (9) and (7) are equivalent to the requirements

$$\sum_{\gamma=1}^{q} \mid G_\gamma(K)_S \mathrm{vec}(K)_S - \hat{Y}_\gamma \mid^2 \to \min_K, \tag{17}$$

$$\hat{Y}_\gamma + \varepsilon_\gamma^- \le G_\gamma(K)_S \mathrm{vec}(K)_S \le \hat{Y}_\gamma + \varepsilon_\gamma^+, \quad \gamma \in \{\overline{1,q}\}. \tag{18}$$

Proposition 2. *The behavior of system* (1)–(4) *best approximates the desired one specified by the set of training examples* (5)*, i.e., the requirements* (7)–(9) *hold, if and only if the matrix* K *given* (4) *is the solution of the nonlinear least-squares problem* (17) *with the constraints* (18) *and* (8)*.*

The proof of Proposition 2 is given in the Appendix.

4. THE SOLUTION METHOD

4.1. Solution of the Problem with a Given Controller Structure

Assume that the controller has a given structure, i.e., the set $\check{S}$ is specified. The desired matrix K corresponding to conditions (4) and (7)–(9) can be determined by solving problem (4), (17), (18), (8) (see Proposition 2). It will be called the statical controller training (SCT) problem. The success in solving this problem will significantly depend on the choice of the initial approximation (on how close the initial values of the desired unknowns are to the solution).

The solution of system (16) is a good initial approximation in the SCT problem. In general, we can take its approximate solution, i.e., the matrix $\underline{K}$ for which the vector $\mathrm{vec}(\underline{K})_S$ minimizes the Euclidean norm of the difference between the left- and right-hand sides of system (16) (the normal pseudosolution)

$$\mathrm{vec}(\underline{K})_S = \bar{G}_S^+ \hat{Y}, \tag{19}$$

where $\bar{G}_S^+$ is the Moore–Penrose pseudoinverse of the matrix of system (16) and $\hat{Y}$ is the right-hand side of system (16).

The closeness of the matrix $\underline{K}$ to the desired solution can be argued as follows. Let conditions (7)–(8) be feasible and $\check{K}$ be the solution of the SCT problem. If the desired trajectories belong to the set of trajectories possible in system (1)–(4), by Proposition 1 the matrix $\check{K}$ will coincide with the solution of system (16), i.e., $\check{K} = \underline{K}$. A small discrepancy between the desired and possible trajectories leads to a small discrepancy between the matrices $\check{K}$ and $\underline{K}$ since small changes in the parameters of system (1)–(4) correspond to small changes in its solutions and vice versa. The feasibility of conditions (7)–(8) means the closeness of the desired and possible trajectories in system (1)–(4); hence, if the desired solution of the SCT problem exists, it will be close to $\underline{K}$. (Hereinafter, we estimate the closeness of matrices by the Frobenius norm.)

Generally speaking, the matrix $\underline{K}$ differs from the desired solution because its definition does not fully considers conditions (7)–(9). Therefore, using it as a starting point, we will find a solution corresponding to the entire set of requirements.

The efficiency of solving the SCT problem can be improved by taking into account its peculiarities. Note that this problem turns into a linear least-squares problem with linear constraints [37, p. 225] (hereinafter, the LSL problem) when replacing, first, $G_\gamma(K)_S$ in (17), (18) with a fixed matrix $G_\gamma(K^*)_S$ corresponding to the fixed matrix K^* and, second, the function $\rho(A_c(K))$ in (8) with its linear approximation near of K^*. Such a linearization procedure is acceptable when seeking a solution in a small neighborhood of the matrix K^*. Therefore, it is possible to approach the solution of the SCT problem sequentially at each search step by solving the LSL problem with the matrix K^* found at the previous step.

The algorithm for solving the SCT problem proposed in this paper includes the following stages.

1. Choose the normal pseudosolution of system (16) as an initial approximation of the desired vector $\mathrm{vec}(K)_S$.

2. Perform an iterative search for the solution. At the 0th iteration, take $\mathrm{vec}(K^{(0)})_S = \mathrm{vec}(\underline{K})_S$. (The iteration number is specified as the superscript in brackets.)

At each jth iteration, solve the LSL problem

$$\sum_{\gamma=1}^{q} \mid G_\gamma^{(j-1)}\alpha^{(j)} - \hat{Y}_\gamma \mid^2 \to \min_K, \tag{20}$$

$$\hat{Y}_\gamma + \varepsilon_\gamma^- \le G_\gamma^{(j-1)}\alpha^{(j)} \le \hat{Y}_\gamma + \varepsilon_\gamma^+, \quad \gamma \in \{\overline{1,q}\}, \tag{21}$$

$$r_0^{(j-1)} + r_1^{(j-1)}\alpha^{(j)} \le 1 - \sigma, \tag{22}$$

where $\alpha^{(j)} \equiv \mathrm{vec}(K^{(j)})_S$ is the vector of unknowns, $G_\gamma^{(j-1)}$ is the column composed of the blocks $G_{k\gamma}(K^{(j-1)}) = C\sum_{i=0}^{k-1}\left(y\left(x_0^\gamma, K^{(j-1)}\right)_i^{\mathrm{T}} \otimes A^{k-i-1}B\right)$, $k \in \{\overline{1,N}\}$, and $r_0^{(j-1)} + r_1^{(j-1)}\alpha^{(j)}$ is the linear approximation of the function $\rho(A_c(K))$ near $K^{(j-1)}$. Conditions (21), (22) may fail when solving the LSL problem (20)–(22). In this case, the search procedure is stopped with stating that the solution of the SCT problem could not be found (because it does not exist or the algorithm is not efficient enough).

3. The search procedure is successfully completed when the vector of unknowns $\alpha^* = \alpha^{(j)}$ satisfying conditions (21) and (22) is obtained and either the difference $\mid \alpha^{(j)} - \alpha^{(j-1)} \mid$ or the objective function (20) becomes small enough, or a given number of iterations is exhausted. Take the matrix $K = \mathrm{vec}_S^{-1}(\alpha^*)$ as the solution, where $\mathrm{vec}_S^{-1}(\cdot)$ is the inverse of the vectorization function. (Given (4), it reconstructs the matrix K from the argument vector.)

The method presented above is substantially similar to the Gauss–Newton iterative algorithm for solving the unconstrained nonlinear least-squares problem. At each iteration of this algorithm, Taylor's theorem is applied to linearize the objective function and solve the resulting linear least-squares problem. In contrast, the novel method essentially exploits the peculiarities of problem (17), (18), (8) and, consequently, requires no differentiation to linearize the objective function. For this purpose, as stated above, it suffices to fix the matrix K within the next iteration. In addition, the novel method is a constrained optimization method: it considers conditions (18) and (8)) when solving the nonlinear least-squares problem. At each iteration of the novel method, the LSL problem is solved, which belongs to the class of convex programming problems [38, 39]. For such problems, the existing effective optimization procedures yield the solution or state its absence. (For example, we mention the lsqlin function in Matlab.)

4.2. Solution of the Structural Design Problem

Assume that the controller structure is not given: the set $\check{S}$ is not specified in the initial problem data and must be determined. In this case, we have the structural design problem. Within the adopted formalization (10), it consists in finding sets $\check{S}$ and corresponding matrices K for which conditions (7)–(9) hold and the structure of the controller (3), (4) is simple [27–31]. It can be solved using the algorithm for designing general-form simple structures [31]. The procedure proposed in subsection 4.1 may serve to assess the acceptability of the controller structure and calculate the corresponding parameters.

5. EXAMPLES

Example 1. Consider the model of a two-mass system [1, p. 52, p. 125]. Assume that the output is composed of all components of the state vector except the second one. Given a time discretization step of 0.01, unit masses, and a stiff spring linking them, we obtain the following matrices of system (1), (2):

$$A = \begin{pmatrix} 1 & 0 & 0.01 & 0 \\ 0 & 1 & 0 & 0.01 \\ -0.01 & 0.01 & 1 & 0 \\ 0.01 & -0.01 & 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 0 \\ 0.01 \\ 0 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

In (3), the desired matrix K has dimensions 1×3. All its components are allowed to be nonzero; therefore, $\check{S} = \{(1,1)\ (1,2)\ (1,3)\}$ in (4).

We define the desired behavior of system (1)–(4) as follows. Let the desired trajectories correspond to the optimal control by the minimum energy criterion that transfers the system from the initial states $x_0^1 = (-1; 1; 1; -1)$ and $x_0^2 = (-10; 10; -1; -1)$ to the origin in time $k = 250$. Together with the initial conditions, these trajectories $Y_1 = (y_k^1)$, $Y_2 = (y_k^2)$, $k \in \{\overline{1{,}500}\}$, form the set of training examples (5) $Q = \{(x_0^1, Y_1), (x_0^2, Y_2)\}$. They can be calculated using the known dependencies [1, p. 128].

First, we solve the design problem without the constraints (7), (8) (i.e., the unconstrained optimization problem of the objective function (9)). After three iterations, the novel method described in Section 4.1 yields the gain matrix $K = (-10.671 \quad -4.124 \quad -13.745)$. The corresponding degree of stability is $\sigma = 0.964 \times 10^{-2}$, and the objective function takes a value of 37.25.

Example 2. To improve stability, we increase the value σ to 1.2×10^{-2} and reduce the amplitude of oscillations on the final interval (for $k \in \{300, \ldots, 500\}$), restricting the admissible deviation of the output coordinates from the desired trajectories to the values ± 0.5 and ± 1.5 for x_0^1 and x_0^2, respectively. (In Example 1, these deviations are 0.68 and 2.23.) Given the above requirements, it is therefore necessary to solve the constrained optimization problem (9), (7), (8). Three iterations of the novel method result in $K = (-13.012 \quad -5.310 \quad -16.821)$; in addition, $\sigma = 1.2 \times 10^{-2}$, conditions (7) and (8) hold, and the value of the objective function is 120.32.

Example 3. Consider the lateral motion model of an aircraft presented in [26, p. 182]. For a time discretization step of 0.001, we obtain the following matrices of system (1):

$$A = \begin{pmatrix} 1000 & 0 & 1 & 0.044 & 0 \\ -1.215 & 999 & 0.131 & 0 & 0 \\ 0.430 & 0.021 & 1000 & 0 & 0 \\ 0 & 1 & 0 & 1000 & 0 \\ 0 & 0 & 1 & 0 & 1000 \end{pmatrix} \times 10^{-3}, \quad B = \begin{pmatrix} 0 & 0 \\ -0.040 & 1.587 \\ 0.381 & -0.067 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \times 10^{-3}.$$

In equation (2), C is an identity matrix of dimensions 5×5.

Solving for system (1)–(3) the LQR problem with the minimization criterion $\sum_{k=1}^{\infty} x_k^{\mathrm{T}} x_k$, we find the gain matrix of the controller (3)

$$K_{\mathrm{LQR}} = -\begin{pmatrix} 2.049 & 0.098 & 3.937 & 0.096 & 0.766 \\ -0.110 & 1.100 & -0.168 & 1.031 & -0.642 \end{pmatrix}.$$

Let the first component of the state vector be excluded from the controller's variables for its structural simplification. This can be done by writing condition (4) of the design problem as

$k_{1,1} = 0$, $k_{2,1} = 0$ (equivalently, the design problem of an output feedback containing all components of the state vector except the first one). Accordingly, the set $\check{S}$ in (4) is given and includes all number pairs of the elements of the matrix K except (1,1) and (2,1).

The set of training examples Q consists of the trajectories $Y_\gamma = (y_k^\gamma)$, $\gamma \in \{\overline{1,5}\}$, $k \in \{\overline{1,10^4}\}$, of system (1)–(3) with the LQR controller (with the gain matrix $K = K_{\mathrm{LQR}}$) corresponding to initial conditions x_0^γ where the component with number γ is 1 and the others are zero. Let the components of the vectors $\varepsilon_k^{\gamma\,-}$ and $\varepsilon_k^{\gamma\,+}$ in (7) be assigned by requiring that the admissible deviation of the trajectories y_k^γ of system (1)–(3) from the desired ones lies within $\pm 1\%$ of their maximum absolute values at each time instant k. In addition, the degree of stability of the designed system must be not smaller than that of system (1)–(3) with the LQR controller. For this purpose, $\sigma = 4 \times 10^{-5}$ is chosen in (8).

Using the novel method, we find the gain matrix

$$K = \begin{pmatrix} 0 & 6.622 & -13.519 & 8.180 & -8.181 \\ 0 & -1.420 & 0.621 & -1.414 & 1.001 \end{pmatrix}.$$

The solution is obtained after four iterations upon satisfying the assigned constraints without progress in decreasing the objective function.

Example 4. We modify the problem of Example 3 as follows. Let the first component of the state vector be excluded from the output by redefining the matrix C in equation (2) as a matrix of dimensions 4×5 obtained by eliminating the first row from the matrix C of Example 3. In this case, the desired matrix K has dimensions 2×4. We solve the structural design problem of the system in the statement presented in subsection 4.2. The novel method yields the sets $\check{S}$ and the corresponding matrices K (see the table) for which conditions (21) and (22) are satisfied and the controller (3), (4) has a simple structure [27–31].

Table

No.	Gain matrix	No.	Gain matrix
1	$\begin{pmatrix} 4.819 & -10.920 & 5.898 & -6.696 \\ 0 & -1.444 & 0.403 & -0.184 \end{pmatrix}$	3	$\begin{pmatrix} 5.230 & -11.510 & 6.413 & -7.030 \\ -0.318 & -0.984 & 0 & 0.0786 \end{pmatrix}$
2	$\begin{pmatrix} 6.084 & -12.742 & 7.500 & -7.736 \\ -0.994 & 0 & -0.867 & 0.644 \end{pmatrix}$	4	$\begin{pmatrix} 5.112 & -11.344 & 6.268 & -6.936 \\ -0.225 & -1.120 & 0.120 & 0 \end{pmatrix}$

6. CONCLUSIONS

This paper has proposed a novel approach to designing static feedback in linear discrete time-invariant control systems. Within this approach, the desired behavior of the system is defined by a set of its output variation laws (training examples). The problem statement and solution method can be generalized to the case dynamic controllers based on the known procedure [3] for reducing dynamic feedback design to an equivalent static feedback design.

The algorithm for solving the static controller learning problem (see subsection 4.1) is heuristic: its convergence has been confirmed by computational experiments without rigorous proof.

APPENDIX

Proof of Statement 1. Let the matrix K be the solution of system (16). Equations (16) and (6) are equivalent if all the desired trajectories Y_γ, $\gamma \in \{\overline{1,q}\}$, belong to the set of solutions of system (1)–(4); see the considerations above. Hence, under all other hypotheses of the proposition, choosing the matrix K based on equalities (16) ensures the requirements (6). This proves the

sufficiency part of Proposition 1. If the matrix K is not the solution of system (16), violating equations (16) will also violate conditions (6). If some of the desired trajectories Y_γ, $\gamma \in \{\overline{1,q}\}$, do not belong to the set of solutions of system (1)–(4), the equality $y_k = y_k^\gamma$ will not hold for them at each time instant $k \in \{\overline{1,N}\}$. Therefore, conditions (6) will fail as well. This proves the necessity part of Proposition 1.

Proof of Statement 2. This result follows from the equivalence of conditions (1)–(4) and (7)–(9) (on the one hand) and conditions (4), (8), (17), and (18) (on the other hand).

REFERENCES

1. Polyak, B.T., Khlebnikov, M.V., and Rapoport, L.B., *Matematicheskaya teoriya avtomaticheskogo upravleniya* (Mathematical Theory of Automatic Control), Moscow: Lenand, 2019.
2. Sadabadi, M.S. and Peaucelle, D., From Static Output Feedback to Structured Robust Static Output Feedback: A Survey, *Ann. Rev. Control*, 2016, vol. 42, pp. 11–26.
3. Syrmos, V.L., Abdallah, C.T., Dorato, P., and Grigoriadis, K., Static Output Feedback—a Survey, *Automatica*, 1997, vol. 33, no. 2, pp. 125–137.
4. Toker, O. and Ozbay, H., On the Np-Hardness of Solving Bilinear Matrix Inequalities and Simultaneous Stabilization with Static Output Feedback, *IEEE American Control Conference*, Seattle, 1995, pp. 2525–2526.
5. Toscano, R., *Structured Controllers for Uncertain Systems: A Stochastic Optimization Approach*, New York: Springer-Verlag, 2013.
6. Rosinova, D., Vesely, V., and Kucera, V., A Necessary and Sufficient Condition for Static Output Feedback Stabilizability of Linear Discrete-Time Systems, *Kybernetika*, 2003, vol. 39, pp. 447–459.
7. Cao, Y.Y., Lam, J., and Sun, Y.X., Static Output Stabilization: An ILMI Approach, *Automatica*, 1998, vol. 34, no. 12, pp. 1641–1645.
8. Wang, X., Pole Placement by Static Output Feedback, *J. Math. Syst. Estim. Control*, 1992, vol. 2, no. 2, pp. 205–218.
9. Pakshin, P.V. and Ryabov, A.V., A Static Output Feedback Control for Linear Systems, *Autom. Remote Control*, 2004, vol. 65, no. 4, pp. 559–566.
10. Agulhari, C.M., Oliveira, R.C., and Peres, P.L., LMI Relaxations for Reduced-Order Robust H^∞-control of Continuous-Time Uncertain Linear Systems, *IEEE Trans. Autom. Control*, 2012, vol. 57, no. 6, pp. 1532–1537.
11. Ebihara, Y., Tokuyama, K., and Hagiwara, T., Structured Controller Synthesis Using LMI and Alternating Projection Method, *Int. J. Control*, 2004, vol. 77, no. 12, pp. 1137–1147.
12. Grigoriadis, K.M. and Beran, E.B., Alternating Projection Algorithms for Linear Matrix Inequalities Problems with Rank Constraints, in *Advances in Linear Matrix Inequality Methods in Control*, Philadelphia: SIAM, 2000, pp. 251–267.
13. Leibfritz, F., An LMI-Based Algorithm for Designing Suboptimal Static H^2/H^∞ Output Feedback Controllers, *SIAM J. Control Optim.*, 2001, vol. 39, no. 6, pp. 1711–1735.
14. Polyak, B.T., Khlebnikov, M.V., and Shcherbakov, P.S., Sparse Feedback in Linear Control Systems, *Autom. Remote Control*, 2014, vol. 75, no. 12, pp. 2099–2111.
15. Bykov, A.V. and Shcherbakov, P.S., Sparse Feedback Design in Discrete-Time Linear Systems, *Autom. Remote Control*, 2018, vol. 79, no. 7, pp. 1175–1190.
16. Lin, F., Fardad, M., and Jovanović, M.R., Design of Optimal Sparse Feedback Gains via the Alternating Direction Method, *IEEE Trans. Autom. Control*, 2013, vol. 58, no. 9, pp. 2426–2431.
17. Belozyorov, V.Y., New Solution Method of Linear Static Output Feedback Design Problem for Linear Control Systems, *Linear Algebra Appl.*, 2016, vol. 504, pp. 204–227.

18. Blumthaler, I. and Oberst, U., Design, Parametrization, and Pole Placement of Stabilizing Output Feedback Compensators via Injective Cogenerator Quotient Signal Modules, *Linear Algebra Appl.*, 2012, vol. 436, pp. 963–1000.

19. Johnson, T. and Athans, M., On the Design of Optimal Constrained Dynamic Compensators for Linear Constant Systems, *IEEE Trans. Autom. Control*, 1970, vol. 15, pp. 658–660.

20. Moerder, D. and Calise, A., Convergence of a Numerical Algorithm for Calculating Optimal Output Feedback Gains, *IEEE Trans. Autom. Control*, 1985, vol. 30, pp. 900–903.

21. Choi, S. and Sirisena, H., Computation of Optimal Output Feedback Gains for Linear Multivariable Systems, *IEEE Trans. Autom. Control*, 1974, vol. 19, pp. 254–258.

22. Kreisselmeier, G., Stabilization of Linear Systems by Constant Output Feedback Using the Riccati Equation, *IEEE Trans. Autom. Control*, 1975, vol. 20, pp. 556–557.

23. Toivonen, H.T., A Globally Convergent Algorithm for the Optimal Constant Output Feedback Problem, *Int. J. Control*, 1985, vol. 41, no. 6, pp. 1589–1599.

24. Geromel, J., Peres, P., and Souza, S., Convex Analysis of Output Feedback Structural Constraints, *Proc. IEEE Conf. on Decision and Control*, San Antonio, 1993, pp. 1363–1364.

25. Iwasaki, T. and Skelton, R., All Controllers for the General H^∞ Control Problem: LMI Existence Conditions and State Space Formulas, *Automatica*, 1994, vol. 30, pp. 1307–1317.

26. Paraev, Yu.I. and Smagina, V.I., Problems of Simplifying the Structure of Optimal Controllers, *Avtomat. i Telemekh.*, 1975, no. 6, pp. 180–183.

27. Mozzhechkov, V.A., *Prostye struktury v teorii upravleniya* (Simple Structures in Control Theory), Tula: Tula State University, 2000.

28. Mozzhechkov, V.A., Design of Simple-Structure Linear Controllers, *Autom. Remote Control*, 2003, vol. 64, no. 1, pp. 23–36.

29. Mozzhechkov, V.A., Design of Simple Robust Controllers for Time-Invariant Dynamic Systems, *J. Comput. Syst. Sci. Int.*, 2021, vol. 60, pp. 353–363.

30. Mozzhechkov, V.A., Synthesis of Simple Relay Controllers in Self-oscillating Control Systems, *Autom. Remote Control*, 2022, vol. 83, no. 9, pp. 1393–1403.

31. Mozzhechkov, V.A., Simple Structures in Problems of Control Theory: Formalization and Synthesis, *J. Comput. Syst. Sci. Int.*, 2022, vol. 61, pp. 295–312.

32. Vapnik, V.N., An Overview of Statistical Learning Theory, *Transactions on Neural Networks*, 1999, vol. 10, no. 5, pp. 988–999.

33. Vorontsov, K.V., Combinatorial Bounds for the Quality of Learning by Precedents, *Dokl. Akad. Nauk*, 2004, vol. 394, no. 2, pp. 175–178.

34. Mohri, M., Rostamizadeh, A., and Talwalkar, A., *Foundations of Machine Learning*, Massachusetts: MIT Press, 2012.

35. Schmidhuber, J., Deep Learning in Neural Networks, *Neural Networks*, 2015, vol. 61, pp. 85–117.

36. Ikramov, Kh.D., *Chislennoe reshenie matrichnykh uravnenii* (Numerical Solution of Matrix Equations), Moscow: Nauka, 1984.

37. Gill, Ph.E., Murray, W., Wright, M.H., *Practical Optimization*, London: Academic Press, 1981.

38. Bertsekas, D.P., *Convex Optimization Algorithms*, Belmont: Athena Scientific, 2015.

39. Polyak, B., *Introduction to Optimization*, Optimization Software, 1987.

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, 2023, Vol. 84, No. 9, pp. 1075–1084.

NONLINEAR SYSTEMS

Fault Identification: An Approach Based on Optimal Control Methods

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

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

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

Abstract—This paper considers the problem of identifying (estimating) faults in systems described by linear models under exogenous disturbances. It is solved using optimal control methods; in comparison with sliding mode observers, they avoid high-frequency switching. The solution method proposed below involves a reduced model of the original system that is sensitive to faults and insensitive to disturbances. The corresponding theory is illustrated by an example.

Keywords: linear systems, faults, identification, observers, optimal systems

DOI: 10.25728/arcRAS.2023.89.57.001

1. INTRODUCTION

For the last two decades, the problem of fault identification has been solved based on sliding mode observers [1–7]. In the works cited, certain constraints were imposed on the system under consideration. The most typical ones include the matching condition and the minimum phase property of the system. This restricts the class of systems for which such observers can be constructed. In addition, the implementation of such observers implies high-frequency switching and, consequently, high-frequency data exchange in the control system, which is not always practicable. The method based on the optimal control theory proposed below is free from this disadvantage.

Consider control systems described by the linear model

$$\begin{aligned} &\dot{x}(t) = Ax(t) + Bu(t) + Dd(t) + L\rho(t), \quad x(t_0) = x_0, \\ &y(t) = Cx(t) \end{aligned} \tag{1.1}$$

with the following notations: $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^m$ is the control vector, and $y \in \mathbb{R}^l$ is the output; $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{l\times n}$, $D \in \mathbb{R}^{n\times q}$, and $L \in \mathbb{R}^{n\times s}$ are known constant matrices; the vector function $d(t) \in \mathbb{R}^q$ describes faults, i.e., $d(t) = 0$ if there are no faults, and $d(t)$ becomes an unknown time-varying function otherwise; finally, $\rho(t) \in \mathbb{R}^s$ is an unknown time-varying function of exogenous disturbances affecting the system.

In this paper, the problem is to design an observer for estimating the function $d(t)$. In contrast to the conventional approach, the solution proposed below is based on optimal control methods. By analogy with [5–7], the problem is solved not for the original system but for its reduced model insensitive to disturbances. Such a model has a smaller dimension than the original system.

2. BUILDING THE REDUCED MODEL

The reduced model has the form

$$\begin{aligned} \dot{x}_*(t) &= A_* x_*(t) + B_* u(t) + J_* y(t) + D_* d(t), \\ y_*(t) &= C_* x_*(t), \end{aligned} \tag{2.1}$$

where $x_*(t) \in \mathbb{R}^k$ is the state vector; A_*, B_*, J_*, C_*, and D_* are matrices of compatible dimensions to be determined. By analogy with [5–7], the matrices A_* and C_* are found in the canonical form

$$A_* = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}, \quad C_* = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \end{pmatrix}. \tag{2.2}$$

Also following [5–7], we suppose the existence of matrices Φ and R_* such that $x_*(t) = \Phi x(t)$, $y_*(t) = R_* y(t)$, and

$$\Phi A = A_* \Phi + J_* C, \quad R_* C = C_* \Phi, \quad \Phi B = B_*, \quad \Phi D = D_*. \tag{2.3}$$

In view of the canonical form (2.2), equations (2.3) imply [5–7] the equations

$$\begin{aligned} &\Phi_1 = R_* C, \quad \Phi_i A = \Phi_{i+1} + J_{*i} C, \quad i = 2, \dots, k-1, \\ &\Phi_k A = J_{*k} C, \end{aligned} \tag{2.4}$$

where Φ_i and J_{*i} are the ith rows of the matrices Φ and J_*, respectively, $i = 1, \dots, k$. The matrix R_* must be chosen so that $D_* \neq 0$. The corresponding procedure will be given below.

Assumption 1. $\mathrm{Im}(D) \not\subset \mathrm{Ker}(V^{(n)})$, where

$$V^{(n)} = \begin{pmatrix} C \\ CA \\ \dots \\ CA^{n-l} \end{pmatrix}$$

is the observability matrix.

Assumption 1 holds if system (1.1) is observable: in this case, $\mathrm{Ker}(V^{(n)}) = 0$ and, consequently, $V^{(n)} D \neq 0$. Let p be the smallest integer satisfying $CA^p D \neq 0$ and j be an integer for which $C_j A^p D \neq 0$. It can be demonstrated that (2.4) implies $\Phi = QV^{(n)}$ for some matrix Q, and then we obtain $D_* = \Phi D \neq 0$ from $C_j A^p D \neq 0$. According to the aforesaid, the pth derivative of the variable y_j is sensitive to faults: this derivative will change value if a fault occurs. Also, obviously, if the jth position of the matrix R_* is nonzero, model (2.1) with this matrix will be sensitive to faults.

As was shown in [5–7], insensitivity to exogenous disturbances holds if $\Phi L = 0$. Together with (2.4), this condition can be reduced to the equation

$$\begin{pmatrix} R_* & -J_{*1} & \dots & -J_{*k} \end{pmatrix} \begin{pmatrix} W^{(k)} & L^{(k)} \end{pmatrix} = 0, \tag{2.5}$$

where

$$W^{(k)} = \begin{pmatrix} CA^k \\ CA^{k-1} \\ \dots \\ C \end{pmatrix}, \quad L^{(k)} = \begin{pmatrix} CL & CAL & \dots & CA^{k-1}L \\ 0 & CL & \dots & CA^{k-2}L \\ & & \dots & \\ 0 & 0 & \dots & 0 \end{pmatrix}.$$

Equation (2.5) has a solution under

$$\mathrm{rank}\,(W^{(k)} \quad L^{(k)}) < l(k+1).$$

This inequality serves to determine the minimal dimension $k > p$; equation (2.5), to determine the row $(\ R_* \ \ -J_{*1} \ \ \dots \ \ -J_{*k}\ )$. If the jth position of the matrix R_* is nonzero, then the matrices Φ, B_*, and D_* are obtained using (2.3) and (2.4). Otherwise, it is necessary to find another solution of equation (2.5).

The stability of the model is ensured by feedback on the residual signal $r_*(t) = R_*y(t) - y_*(t)$:

$$\dot{x}_*(t) = A_*x_*(t) + B_*u(t) + J_*y(t) + D_*d(t) + Kr_*(t), \tag{2.6}$$

where the matrix K has the form $K = (\ k_1 \ \ k_2 \ \ \dots \ \ k_k\ )^{\mathrm{T}}$. The coefficients $k_1, k_2, \dots, k_k$ are determined from given eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_k$:

$$\begin{aligned} k_1 &= -(\lambda_1 + \lambda_2 + \ldots + \lambda_k),\\ k_2 &= \lambda_1\lambda_2 + \lambda_1\lambda_3 + \ldots + \lambda_{k-1}\lambda_k,\\ &\dots,\\ k_k &= (-1)^k\lambda_1\lambda_2\ldots\lambda_k. \end{aligned}$$

Consider the expression for the residual $r_*(t)$, equation (2.6) can be transformed into

$$\dot{x}_*(t) = (A_* - KC_*)x_*(t) + B_*u(t) + (J_* + KR_*)y(t) + D_*d(t).$$

3. AN AUXILIARY OPTIMAL CONTROL PROBLEM

As has been emphasized, the problem of fault identification is solved using optimal control methods. Consider the corresponding problem for the system

$$\begin{aligned} &\dot{z}(t) = (A_* - KC_*)z(t) + B_*u(t) + (J_* + KR_*)y(t) + D_*w(t),\ z(t_0) = \Phi x_0,\\ &y_z(t) = C_*z(t), \end{aligned} \tag{3.1}$$

where the auxiliary control variable $w(t)$ plays the role of the unknown function $d(t)$. It is chosen to transfer system (3.1) from the state $z(t_0)$ to a target state with the output $y_z(t_f)$ such that $y_z(t_f) \to y_*(t_f)$ as $t_f \to \infty$ and

$$J = \frac{1}{2}\int\limits_{t_0}^{\infty}\left(e_y^{\mathrm{T}}Qe_y + w^{\mathrm{T}}Rw\right)dt \to \min_v. \tag{3.2}$$

Here, $e_y(t) = y_z(t) - y_*(t)$ denotes the residual, and Q and $R \in \mathbb{R}^{q\times q}$ are a positive number and a positive definite matrix, respectively. The relation $y_z(t_f) \to y_*(t_f)$, $t_f \to \infty$, is understood as convergence in the Euclidean norm: $\|y_z(t_f) - y_*(t_f)\| \to 0$ as $t_f \to \infty$. The convergence of other time-varying functions in this paper is interpreted by analogy.

The identification problem is to construct the optimal control $w(t)$ in the sense of the performance criterion (3.2) such that $y_z(t) \to y_*(t)$ and $w(t) \to d(t)$ as $t \to \infty$. The criterion (3.2) must be minimized for a sufficiently large value of the constant Q, in particular, $Q = 10^{20}$ in the example below. This practically ensures the property $e_y(t) \to 0$ as $t \to \infty$. With this in mind, we denote by e_{y*} the asymptote for $e_y(t)$. According to the previous considerations, $e_{y*} = 0$ can be taken with a sufficient degree of accuracy.

Introducing the error vector $e(t) = z(t) - x_*(t) \in R^k$, we write the corresponding equation

$$\begin{aligned} \dot{e}(t) &= A_*e(t) + D_*(w(t) - d(t)) - Ke_y(t)\\ &= (A_* - KC_*)e(t) + D_*(w(t) - d(t)), \quad e(t_0) = 0,\\ e_y(t) &= C_*e(t). \end{aligned} \tag{3.3}$$

Assumption 2. System (3.3) is strongly observable.

Strong observability means the absence of invariant zeros. In other words, there exist no s for which

$$\operatorname{rank}(R(s)) < k + \operatorname{rank}\begin{pmatrix} -D_* \\ 0 \end{pmatrix},$$

where $R(s)$ is the Rosenbrock matrix [8, 9]:

$$R(s) = \begin{pmatrix} sI - (A_* - KC_*) & -D_* \\ C_* & 0 \end{pmatrix}.$$

Theorem 1. *If system* (3.3) *is strongly observable, then* $e_y(t) \to 0$ *implies* $w(t) \to d(t)$ *as* $t \to \infty$.

Proof. Let $H(s)$ be the transfer function of system (3.3):

$$E_y(s) = H(s)(W(s) - D(s)), \tag{3.4}$$

where $E_y(s)$, $W(s)$, and $D(s)$ are the Laplace images of the functions e_y, $w(t)$, and $d(t)$, and s denotes the complex variable. Since $e_{y*} = 0$, it follows that $E_y(s) = 0$. System (3.3) has no invariant zeroes; hence, for all s, the function $H(s)$ is nonzero and, consequently, $W(s) = D(s)$. According to [10], functions with identical images coincide for all $t > 0$ except a set of measure zero. Therefore, $w(t)$ and $d(t)$ coincide for all $t > 0$ except a set of measure zero. The asymptotic convergence of the function $e_y(t)$ will be written as $w(t) \to d(t)$.

Clearly, the converse is true: if the system is not strongly observable, its transfer function will have a zero, $H(s) = 0$ for some s. Then equality (3.4) will hold for $d(t) + e^{at}$ with $s = a$. In this case, the fault will be reconstructed within the exponent.

4. SOLUTION OF THE AUXILIARY PROBLEM

Here is its solution. For problem (3.1) and (3.2), the Hamiltonian has the form

$$H = \frac{1}{2}(z - x_*)^{\mathrm{T}} C_*^{\mathrm{T}} Q C_* (z - x_*) + \frac{1}{2} w^{\mathrm{T}} R w + \lambda^{\mathrm{T}}(\overline{A}_* z + \overline{J}_* y + D_* w + B_* u),$$

where $\overline{A}_* = A_* - KC_*$ and $\overline{J}_* = J_* + KR_*$. The optimal control law is given by

$$\frac{\partial H}{\partial w} = 0 \Rightarrow Rw + D_*^{\mathrm{T}}\lambda = 0 \Rightarrow w = -R^{-1}D_*^{\mathrm{T}}\lambda. \tag{4.1}$$

The state and conjugate variables satisfy the equations

$$\begin{aligned} \dot{z}(t) &= \frac{\partial H}{\partial \lambda} = \overline{A}_* z + \overline{J}_* y + D_* w + B_* u = \overline{A}_* z + \overline{J}_* y - D_* R^{-1} D_*^{\mathrm{T}}\lambda + B_* u, \\ z(t_0) &= \Phi x_0, \\ \dot{\lambda}(t) &= \frac{\partial H}{\partial z} = -\overline{A}_*^{\mathrm{T}}\lambda - C_*^{\mathrm{T}} Q C_* z + C_*^{\mathrm{T}} Q y_*. \end{aligned}$$

We write the latter relations in matrix form:

$$\begin{pmatrix} \dot{z}(t) \\ \dot{\lambda}(t) \end{pmatrix} = \begin{pmatrix} \overline{A}_* & -D_* R^{-1} D_*^{\mathrm{T}} \\ -C_*^{\mathrm{T}} Q C_* & -\overline{A}_*^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} z(t) \\ \lambda(t) \end{pmatrix} + \begin{pmatrix} B_* \\ 0 \end{pmatrix} u(t) + \begin{pmatrix} \overline{J}_* y(t) \\ C_*^{\mathrm{T}} Q y_*(t) \end{pmatrix}, \tag{4.2}$$
$$z(t_0) = \Phi x_0.$$

Equation (4.2) can be considered a diagnostic observer. By integrating (4.2) in forward time, it is possible to find and then reconstruct based on (4.1) the function describing the fault:

$$w(t) = -R^{-1}D_*^{\mathrm{T}}\lambda(t) \to d(t). \tag{4.3}$$

An open issue is the choice of initial conditions for the conjugate variable when integrating (4.2). Since the initial conditions are unknown, we introduce the following Riccati transformation [10] to find the control law:

$$z(t) = M(t)\lambda(t) + g(t), \tag{4.4}$$

where $M(t)$ and $g(t)$ are a nonsingular matrix and some vector function, respectively. Differentiating (4.4) and performing several transformations yield

$$\begin{aligned}&\left(-\dot{M}(t) + \overline{A}_*M(t) + M(t)\overline{A}_*^{\mathrm{T}} - D_*R^{-1}D_*^{\mathrm{T}} + M(t)C_*^{\mathrm{T}}QC_*M(t)\right)\lambda(t)\\ &= \dot{g}(t) - \overline{A}_*g(t) - \overline{J}_*y(t) - B_*u(t) - M(t)C_*^{\mathrm{T}}QC_*g(t) + M(t)C_*^{\mathrm{T}}Qy_*(t).\end{aligned}$$

This relation must hold for any $\lambda(t)$; hence, we obtain the equations

$$\begin{aligned}\dot{M}(t) &= \overline{A}_*M(t) + M(t)\overline{A}_*^{\mathrm{T}} - D_*R^{-1}D_*^{\mathrm{T}} + M(t)C_*^{\mathrm{T}}QC_*M(t),\\ \dot{g}(t) &= \overline{A}_*g(t) + \overline{J}_*y(t) + B_*u(t) + M(t)C_*^{\mathrm{T}}QC_*g(t) - M(t)C_*^{\mathrm{T}}Qy_*(t).\end{aligned} \tag{4.5}$$

For $t = t_0$, it follows from (4.4) that $z(t_0) = M(t_0)\lambda(t_0) + g(t_0)$. Since $\lambda(t_0)$ is unknown, the initial conditions will be satisfied by letting $M(t_0) = 0$ and $z(t_0) = g(t_0)$. Substituting (4.4) into (4.3) finally gives

$$w(t) = -R^{-1}D_*^{\mathrm{T}}M^{-1}(t)(z(t) - g(t)). \tag{4.6}$$

The ultimate expression for the desired observer has the form

$$\begin{aligned}\dot{z}(t) &= \overline{A}_*z(t) - D_*R^{-1}D_*^{\mathrm{T}}M^{-1}(t)(z(t) - g(t)) + \overline{J}_*y(t) + B_*u(t),\\ z(t_0) &= \Phi x(t_0),\\ y_z(t) &= C_*z(t).\end{aligned} \tag{4.7}$$

Here, $M(t)$ and $g(t)$ are determined from equations (4.5) with the initial conditions $M(t_0) = 0$ and $z(t_0) = g(t_0)$. On an infinite time interval, when system (3.1) is controllable and observable, the solution of equation (4.5) will tend to the steady-state value $\overline{M}$ as $t \to \infty$, which is the solution of the algebraic equation [11–13]

$$\overline{A}_*\overline{M} + \overline{M}\overline{A}_*^{\mathrm{T}} - D_*R^{-1}D_*^{\mathrm{T}} + \overline{M}C_*^{\mathrm{T}}QC_*\overline{M} = 0;$$

the function $g(t)$ from the second equation in (4.5) will tend to the bounded solution $\overline{g}(t)$ of the differential equation

$$\dot{\overline{g}}(t) = (\overline{A}_* + \overline{M}C_*^{\mathrm{T}}QC_*)\overline{g}(t) + \overline{J}_*y(t) + B_*u(t) - \overline{M}C_*^{\mathrm{T}}Qy_*(t) \tag{4.8}$$

with the initial conditions $\overline{g}(t_0) = z(t_0)$. The desired observer on an infinite time interval takes the form (4.7), where $M(t)$ and $g(t)$ are replaced by $\overline{M}$ and $\overline{g}(t)$.

The convergence of $g(t)$ to the bounded solution $\overline{g}(t)$ is immediate from the following considerations. Multiplying the equation for M (4.5) by -1 on the left and right and denoting $P = -M$, we obtain the Riccati equation, which typically arises in optimal estimation problems [14]. Under the conditions $R > 0$, $Q > 0$, and the observability of system (3.1), the solution of this equation is known to converge to the steady-state solution $\overline{P}$ representing the unique positive definite solution of the algebraic Riccati equation $\overline{A}_*\overline{P} + \overline{P}\overline{A}_*^{\mathrm{T}} + D_*R^{-1}D_*^{\mathrm{T}} - \overline{P}C_*^{\mathrm{T}}QC_*\overline{P} = 0$, and $\overline{A}_* - \overline{P}C_*^{\mathrm{T}}QC_*$ is a Hurwitz matrix. Thus, due to $\overline{P} > 0$, $P \to \overline{P}$, and $P = -M$, we obtain $M \to \overline{M}$, $\overline{M} < 0$, and the matrix $\overline{A}_* + \overline{M}C_*^{\mathrm{T}}QC_*$ will be Hurwitz as well. With the error $e_g(t) = g(t) - \overline{g}(t)$, from (4.5) and (4.8) it follows that $\dot{e}_g(t) = (\overline{A}_* + \overline{M}C_*^{\mathrm{T}}QC_*)e_g(t)$. Since $\overline{A}_* + \overline{M}C_*^{\mathrm{T}}QC_*$ is a Hurwitz matrix, we have $e_g(t) \to 0$ and $g(t) \to \overline{g}(t)$ as $t \to \infty$.

5. SIMULATION RESULTS

Let us construct an observer for a two-wheeled inverted pendulum (TWIP) robot with self-balancing [15]. The kinematic diagram of this robot is shown in Fig. 1. The mathematical model of the TWIP robot linearized at the equilibrium takes the form (1.1) with the following notations:

$$A=\begin{pmatrix}0&1&0&0&0&0\\0&0&a_2&0&0&0\\0&0&0&1&0&0\\0&0&a_4&0&0&0\\0&0&0&0&0&1\\0&0&0&0&0&0\end{pmatrix},\quad B=\begin{pmatrix}0&0\\b_2&b_2\\0&0\\b_4&b_4\\0&0\\b_6&-b_6\end{pmatrix},\quad L=\begin{pmatrix}0\\b_2\\0\\b_4\\0\\-b_6\end{pmatrix},\quad C=\begin{pmatrix}1&0&0&0&0&0\\0&0&1&0&0&0\\0&0&0&0&1&0\end{pmatrix};$$

$x=(x_l\ \dot{x}_l\ \vartheta\ \dot{\vartheta}\ \psi\ \dot{\psi})$ and $u=(T_L\ T_R)$ are the state and control vectors, respectively; T_L and T_R are the left and right wheel torques, respectively; x_l is the linear displacement; ϑ and ψ are the pitch and heading angles, respectively; $y=(x_l\ \vartheta\ \psi)$ is the system output; $d(t)=\tilde{T}_L$ is an unknown additional torque applied to the left wheel to be determined; $\rho(t)=\tilde{T}_R$ is an unknown additional torque applied to the right wheel. The coefficients a_2, a_4, b_2, b_4, and b_6 can be found using the expressions below [15]:

$$a_2=-m_B^2gl^2/\mu_1,\quad a_4=\left(m_B+2m_W+\frac{2J}{r^2}\right)m_Bgl/\mu_1,$$

$$b_2=\left((I_2+m_Bl^2)/r+m_Bl\right)/\mu_1,\quad b_4=-\left(\frac{m_Bl}{r}+m_B+2m_W+\frac{2J}{r^2}\right)/\mu_2,$$

$$b_6=-\frac{d}{r\mu_2},$$

$$\mu_1=\left(m_B+2m_W+\frac{2J}{r^2}\right)(I_2+m_Bl^2)-m_B^2l_2,$$

$$\mu_2=I_3+2K_*+2\left(m_W+\frac{J}{r^2}\right)d^2.$$

The parameter values of the robot are combined in the table.

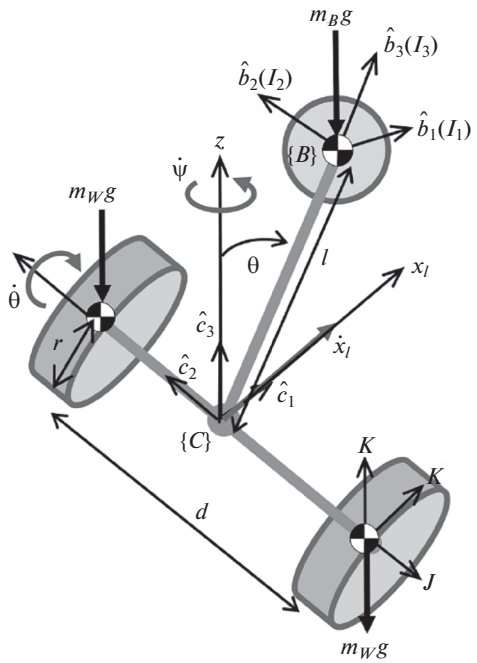


Fig. 1. The kinematic configuration of a TWIP robot.

Table

Notation	Description	Value
m_B	Weight of the pendulum body	45 kg
m_W	Wheel weight	2 kg
l	Pendulum length	0.135 m
r	Wheel radius	8 in
d	Wheel-to-wheel spacing	0.6 m
I_1, I_2, I_3	The moments of inertia of the pendulum body relative to the axes X, Y, Z	1.9; 2.1; 1.6 kg*m^2
K_*, J	The moments of inertia of the pendulum wheels relative to the vertical axis and the wheel rotation axis	0.04; 0.02 kg*m^2

To find the matrices of the reduced model (2.1), it is necessary to solve equation (2.5). For $k = 1$, it takes the form

$$(-j_1 \;\; r_1 \;\; - j_2 \;\; r_2 \;\; - j_3 \;\; r_3 \;\; 0) = 0,$$

where j_i are elements of the vector $J_* = (j_1 \;\; j_2 \;\; j_3)$ and r_i are elements of the vector $r = (r_1 \;\; r_2 \;\; r_3)$. Obviously, this equation possesses the trivial solution only. For $k = 2$, equation (2.5) takes the form

$$(-j_{21} \;\; - j_{11} \;\; a_2 r_1 - j_{22} + a_4 r_2 \;\; - j_{12} \;\; - j_{23} \;\; - j_{31} \;\; 0 \;\; b_2 r_1 + b_4 r_2 - b_6 r_3) = 0,$$

where j_{ik} are elements of the matrix J_* of dimensions 2×3. In this case, the vector R_* can be chosen as $R = (0 \;\; b_6 \;\; b_4)$; then the matrix J_* becomes

$$J_* = \begin{pmatrix} 0 & 0 & 0 \\ 0 & a_4 b_6 & 0 \end{pmatrix}.$$

The next step is to find the matrices Φ and B_* using the expressions (2.4) and (2.3):

$$\Phi = \begin{pmatrix} 0 & 0 & b_6 & 0 & b_4 & 0 \\ 0 & 0 & 0 & b_6 & 0 & b_4 \end{pmatrix}, \quad B_* = \Phi B = \begin{pmatrix} 0 & 0 \\ 2a_4 b_6 & 0 \end{pmatrix}.$$

Thus, the reduced robot model (2.6), sensitive to the function $d(t)$ and insensitive to the function $\rho(t)$, has the form

$$\begin{aligned} \dot{x}_{*1}(t) &= x_{*2}(t) + k_1 r_*(t), \\ \dot{x}_{*2}(t) &= a_4 b_6 y_2(t) + 2a_4 b_6 T_L(t) + k_2 r_*(t), \\ r_*(t) &= b_6 y_2(t) + b_4 y_3(t) - x_{*1}(t). \end{aligned}$$

The diagnostic observer for fault identification is given by (4.6)–(4.8), where $R = 10^{-2}$, $Q = 10^{20}$, and $K = (1 \;\; 1)^T$. Figure 2 shows the structural diagram of this observer.

Assume that the fault (an additional moment applied to the left wheel) is a rectangular pulse with a duration of 4 s that appears at $t = 2$.

Having adjusted the observer parameters, we obtain the identification result in Fig. 3. Clearly, the observer design approach provides an acceptable result. In addition, the graphs of the model state and observer states and observation errors can be found in Figs. 4 and 5, respectively.

Note that the quality of identification based on the optimal observer (4.6)–(4.8) depends on the choice of the penalty matrices Q and R and the matrix K. When selecting them, it is recommended to use the following considerations. The cross relations between the output and fault variables are reflected in the off-diagonal elements of these matrices. In the absence of information on such

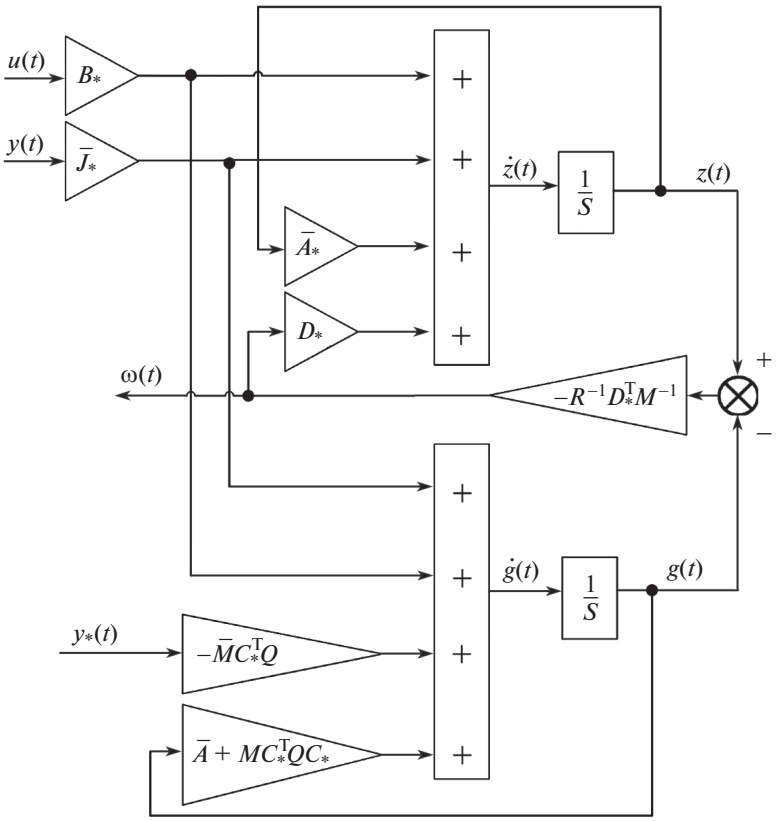


Fig. 2. Structural diagram.

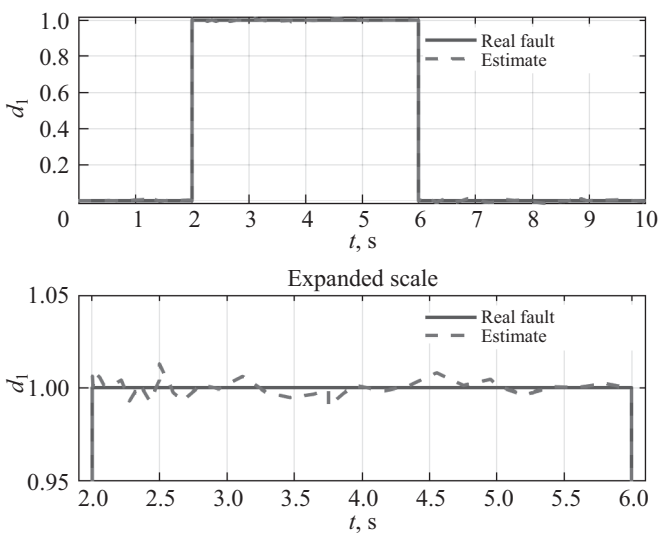


Fig. 3. Operation of the observer under a rectangular pulse fault.

relations, the diagonal form of the matrix R is recommended. The same recommendation applies to the matrix Q in the case of no cross relations between the observed outputs. If the resulting fault estimate has a large value, it is required to reduce the corresponding diagonal elements of the matrix R. Given large values of the residual $e(t) = x_*(t) - z(t)$, the elements of the matrix Q must be increased. The coefficients of the matrix K are assigned to ensure a higher performance of the observer.

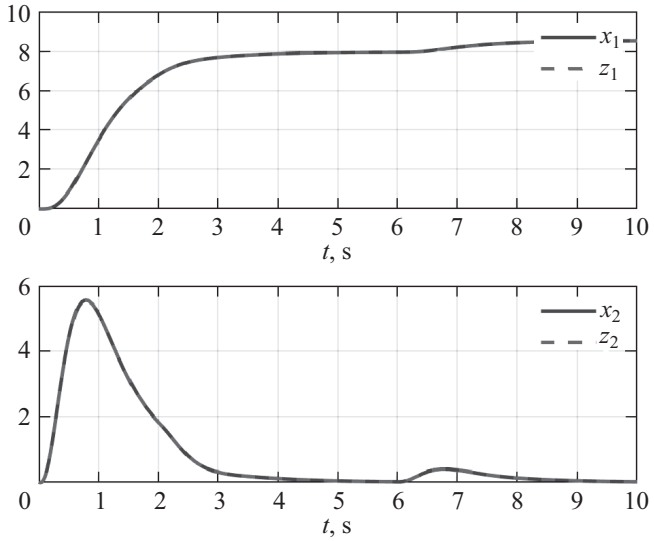


Fig. 4. The graphs of system state and diagnostic observer.

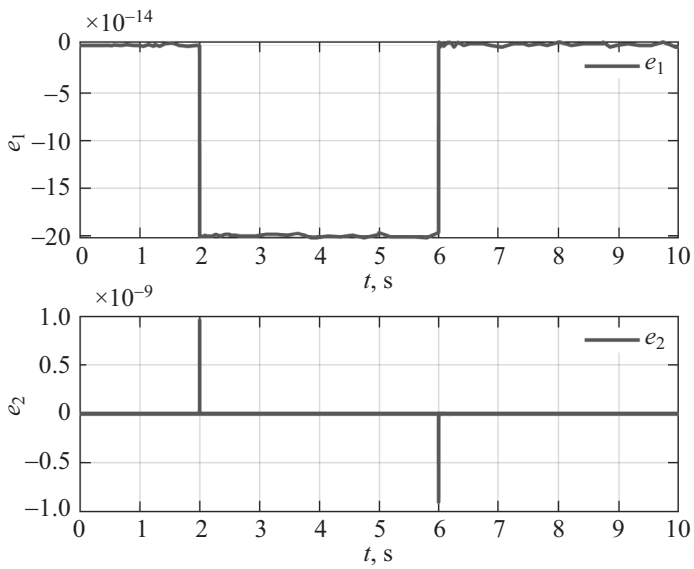


Fig. 5. The graphs of observation errors $e(t) = x_*(t) - z(t)$.

6. CONCLUSIONS

In this paper, we have estimated (identified) faults in systems described by linear models with constant coefficients under exogenous disturbances. In contrast to well-known methods based on sliding mode observers, the approach developed above expands the class of systems for which identification can be performed: the method of constructing sliding mode observers imposes restrictions on the systems for fault identification.

FUNDING

This work was supported by the Russian Science Foundation, project no. 22-19-00392; https://rscf.ru/project/22-19-00392/.

REFERENCES

1. Edwards, C., Spurgeon, S., and Patton, R., Sliding Mode Observers for Fault Detection and Isolation, *Automatica*, 2000, vol. 36, pp. 541–553.
2. Floquet, T., Barbot, J., Perruquetti, W., and Djemai, M., On the Robust Fault Detection via a Sliding Mode Disturbance Observer, *Int. J. Control*, 2004, vol. 77, pp. 622–629.
3. Yan, X. and Edwards, C., Nonlinear Robust Fault Reconstruction and Estimation Using a Sliding Modes Observer, *Automatica*, 2007, vol. 43, pp. 1605–1614.
4. Rios, H., Efimov, D., Davila, J., Raissi, T., Fridman, L., and Zolghadri, A., Non-minimum Phase Switched Systems: HOSM Based Fault Detection and Fault Identification via Volterra Integral Equation, *Int. J. Adapt. Contr. and Signal Proc.*, 2014, vol. 28, pp. 1372–1397.
5. Zhirabok, A.N., Zuev, A.V., Filaretov, V.F., et al., Fault Identification in Nonlinear Systems Based on Sliding Mode Observers with Weakened Existence Conditions, *J. Comput. Syst. Sci. Int.*, 2022, vol. 61, no. 3, pp. 313–321.
6. Zhirabok, A., Zuev, A., Sergiyenko, O., and Shumsky, A., Identification of Faults in Nonlinear Dynamical Systems and Their Sensors Based on Sliding Mode Observers, *Autom. Remote Control*, 2022, vol. 83, pp. 214–236.
7. Zhirabok, A.N., Zuev, A.V., and Shumskii, A.E., Diagnosis of Linear Systems Based on Sliding Mode Observers, *J. Comput. Syst. Sci. Int.*, 2019, vol. 58, no. 6, pp. 898–914.
8. Mironovskii, L.A., *Funktsional'noe diagnostirovanie dinamicheskikh sistem* (Functional Diagnosis of Dynamic Systems), Moscow–St. Petersburg: MGU-GRIF, 1998.
9. Hautus, M., Strong Detectability and Observers, *Linear Algebra and Its Applications*, 1983, vol. 50, pp. 353–368.
10. Korn, G.A. and Korn, Th.M., *Manual of Mathematics*, McGraw-Hill, 1967.
11. Mufti, I.H., Chow, C.K., and Stock, F.T., Solution of Ill-Conditioned Linear Two-Point Boundary Value Problems by the Riccati Transformation, *SIAM Rev.*, 1969, vol. 11, no. 4, pp. 616–619.
12. Naidu, D.S., *Optimal Control Systems*, Electrical Engineering Handbook, Florida, Boca Raton: CRC Press, 2003.
13. Bryson, A.E. and Ho, Y.-C., *Applied Optimal Control*, Routledge, 1975.
14. Kwakernaak, H. and Sivan, R., *Linear Optimal Control Systems*, Wiley-Interscience, 1972.
15. Kim, S. and Kwon, S.J., Nonlinear Optimal Control Design for Underactuated Two-Wheeled Inverted Pendulum Mobile Platform, *IEEE/ASME Transactions on Mechatronics*, 2017, vol. 22, no. 6, pp. 2803–2808.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 9, pp. 1085–1093.

NONLINEAR SYSTEMS

Global Stability of a Second-Order Affine Switching System

A. V. Pesterev

Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
e-mail: alexanderpesterev.ap@gmail.com

Received March 17, 2023
Revised June 13, 2023
Accepted June 29, 2023

Abstract—Stability of an affine switching system is studied. The system comes to existence when stabilizing a chain of two integrators by means of a feedback in the form of nested saturators. The use of such a feedback allows one to easily take into account boundedness of the control resource, to constrain the maximum velocity of approaching the equilibrium state, which is especially important in the case of large initial deviations, and to ensure desired characteristics of the transient process, such as a given exponential rate of the deviation decrease near the equilibrium state. It is proved that the closed-loop system is globally stable.

Keywords: stabilizing a chain of two integrators, affine switching system, global stability, nested saturators

DOI: 10.25728/arcRAS.2023.96.94.001

1. INTRODUCTION

Hybrid systems are dynamical systems that exhibit both continuous-time and discrete-time behavior; i.e., systems whose states vary continuously but may also jump [1]. A switching system is a hybrid dynamical system consisting of a number of subsystems and a switching law determining which subsystem is active at a current moment of time [2]. Systems of this kind are encountered in many control problems in various fields of science and technology [1, 2]. One of the most important problems in study of switching systems is that of stability [2–4]. It is stability of the switching system under consideration that is discussed in this work.

The affine switching system under study comes to existence when stabilizing a chain of two integrators by means of a feedback in the form of nested saturators. The problem of stabilizing chains of integrators was widely discussed in the literature during last several decades (see, e.g., [5–7] and references therein). The interest to this problem is motivated by the fact that many real-life systems in applications (e.g., mechanical planar ones) are modeled by chains of integrators; moreover, controls developed for chains of integrators can be easily extended to larger classes of systems.

Feedbacks in the form of nested saturators were studied and used for stabilizing integrators in many publications (see, e.g., [5, 6, 8–11] and references therein). However, the author is not aware of the works the results of which could be used for establishing stability of the system closed by the feedback considered in the paper. The general case of the nth-order integrator was discussed, for example, in [5, 6]. However, global stability of the system closed by a feedback in the form of n nested saturators was proved only for the special case where the limits of the saturation functions satisfy certain inequalities [5, Theorem 2.1], which are not fulfilled for the feedback used in this study. Global stability of the second-order integrator stabilized by a feedback in the form of nested saturators, but with the reversed order of the arguments (see the next section for more

detail), was proved in [8, 9]. However, the approach employed in these works is not applicable to the case of the feedback considered in this paper.

The feedback of the form considered in the paper was studied earlier in some works. For example, in [10, 11], optimization problems of selecting the feedback coefficients were discussed; in [12–14], it was used in the synthesis of controllers for stabilizing higher-order integrators. In these works, stability of the system closed by such a feedback was implied but was not proved. The goal of this study is to prove global stability of the system discussed filling thus this gap. The interest in the feedback in the form of nested saturators is explained by a number of remarkable features of the closed-loop system obtained, which is discussed in the next section.

2. PROBLEM STATEMENT

Consider the problem of stabilizing a chain of two integrators:

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = U(x), \tag{1}$$

where $x \equiv [x_1, x_2]^{\mathrm{T}}$, by means of a continuous feedback with a constrained control resource $U_{\max} = k_4 : U(x) = -k_3(x_2 + k_2\mathrm{sat}(k_1x_1))$ for $|U(x)| \leqslant k_4$ and $U(x) = -k_4\mathrm{sign}(k_3(x_2 + k_2\mathrm{sat}(k_1x_1)))$ for $|U(x)| > k_4$. In the compact form, control $U(x)$ is written as

$$U(x) = -k_4\mathrm{sat}\left(\frac{k_3}{k_4}(x_2 + k_2\mathrm{sat}(k_1x_1))\right), \tag{2}$$

where $\mathrm{sat}(\cdot)$ is the nonsmooth saturation function: $\mathrm{sat}(w) = w$ for $|w| \leqslant 1$ and $\mathrm{sat}(w) = \mathrm{sign}(w)$ for $|w| > 1$. The advantage of the feedback in the form of nested saturators is that not only the control constraint is automatically satisfied but also the maximum speed of approaching the equilibrium is limited: if we set $k_2 = V_{\max}$, then, for any initial deviation, $\dot{x}_1(t) \leqslant V_{\max}$ as long as $x_2(0) \leqslant V_{\max}$ [11]. Moreover, any desired type of the equilibrium (node, pole, or center) and any desired value of the exponential rate of deviation decrease near the origin can be ensured by appropriate choice of the coefficients k_1 and k_3 [11].

As noted in the Introduction, in [8, 9], a feedback of form (2) with the reverse compared to (2) order of arguments, where the argument of the internal saturator is velocity x_2 and that of the external saturator is deviation x_1, was considered (in [9], the argument of the external saturator depends additionally on the velocity squared x_2^2), and it was proved that the second-order integrator closed by such a feedback is globally stable. The proof in both works is based on the existence of a Lyapunov function in the form of the sum of a quadratic and integral terms. For system (1), (2), however, this expedient is not applicable, since no Lyapunov function is available.

2.1. Equivalent Representation in the Form of an Affine Switching System

Let us first show that system (1), (2) is an affine switching one. Consider partitioning of plane (x_1, x_2) into five sets (Fig. 1). In the set D_1, we include all points where both saturators are not saturated:

$$D_1 = \{(x_1, x_2) : |x_1| < 1/k_1,\ |x_2 + k_1x_1| < k_4/k_3\}$$

(the inclined strip bounded by the dashed lines in Fig. 1). The set D_2 consists of all points where the internal saturator reaches saturation, while the external one does not:

$$D_2 = \{(x_1, x_2) : |x_1| \geqslant 1/k_1,\ |x_2 + k_2\mathrm{sgn}(x_1)| < k_4/k_3\}.$$

As can be seen from the figure, D_2 consists of the two disjoint sets D_2^- and D_2^+ belonging to the left and right half-planes, respectively (two disjoint horizontal strips in Fig. 1). The set

$$D_3 = \{(x_1, x_2) : |x_2 + k_2\mathrm{sat}(k_1x_1))| > k_4/k_3\}$$

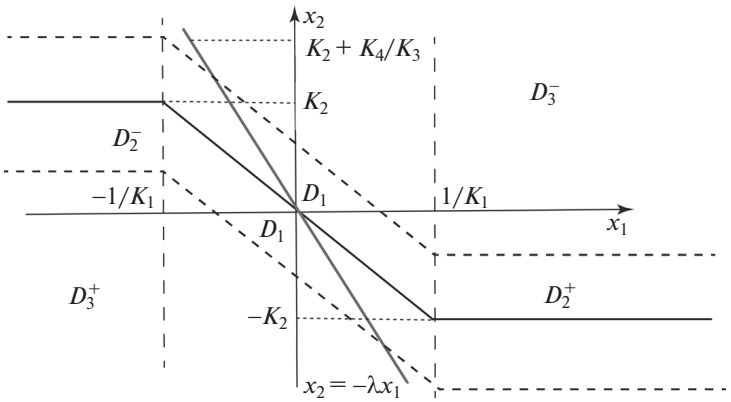


Fig. 1.

includes all points where the external saturator reaches saturation. Like D_2, D_3 consists of two nonintersecting sets D_3^- and D_3^+ lying above and below the line $x_2 = -\mathrm{sat}(k_1 x_1)$ (the solid broken line in Fig. 1), in which $U_1(x) \equiv -k_4$ and $U_1(x) \equiv +k_4$, respectively.

From formula (2), it can be seen that $U(x)$ is a piecewise continuous function:

$$U(x) = \begin{cases} -k_3 x_2 - k_1 k_2 k_3 x_1, & (x_1, x_2) \in D_1, \\ -k_3(x_2 - k_2), & (x_1, x_2) \in D_2^-, \\ -k_3(x_2 + k_2), & (x_1, x_2) \in D_2^+, \\ -k_4, & (x_1, x_2) \in D_3^-, \\ +k_4, & (x_1, x_2) \in D_3^+, \end{cases} \tag{3}$$

and the closed-loop system (1), (2) includes five linear systems, with the switching between them being state dependent according to equation (3). The goal of this study is to prove global stability of this system.

The standard method of proving stability of linear switching systems—determining a common Lyapunov function for all systems—is not applicable in this case, since the origin is an equilibrium point for only the first system with the domain D_1. The other four systems, although linear ones, have no equilibria at all; i.e., we deal with an *affine switching system*. The standard method of proving stability of general-form nonlinear systems with the help of a Lyapunov function (like, e.g., in [9]) cannot be applied either since we failed to find one for the system under consideration.

2.2. Representation in Dimensionless Form

To begin with, we simplify the task by reducing the number of the system parameters. Clearly, stability of the system does not depend on particular values of the control resource k_4 and maximum velocity k_2, so that we can set them equal to one. Indeed, turning to dimensionless variables $\tilde{x}_1 = k_4 x_1/k_2^2$, $\tilde{x}_2 = x_2/k_2$ and time $\tilde{t} = k_4 t/k_2$, we reduce system (1), (2) to the form

$$\frac{d\tilde{x}_1}{d\tilde{t}} = \tilde{x}_2, \quad \frac{d\tilde{x}_2}{d\tilde{t}} = -\mathrm{sat}(\tilde{k}_3(\tilde{x}_2 + \mathrm{sat}(\tilde{k}_1 \tilde{x}_1))), \tag{4}$$

where $\tilde{k}_1 = k_1 k_2^2/k_4$ and $\tilde{k}_3 = k_2 k_3/k_4$, with unitary dimensionless control resource $\tilde{k}_4 = 1$ and unitary maximum velocity $\tilde{k}_2 = 1$. In what follows, all variables and constants are assumed dimensionless and are denoted by the same symbols (without tilde) as dimensional ones. As before, we use the dot notation to denote the derivatives with respect to the dimensionless time. Moreover, without

loss of generality, we will select coefficients k_1 and k_3 from a one-parameter family parameterized by the exponential rate λ of the deviation decrease near the origin:

$$k_1 = \lambda/2, \quad k_3 = 2\lambda, \ \lambda > 0. \tag{5}$$

With these coefficients, system (1) closed by the feedback (2) takes the form

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = -\mathrm{sat}(2\lambda(x_2 + \mathrm{sat}(\lambda x_1/2))). \tag{6}$$

In D_1, we have linear system

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = -\lambda^2 x_1 - 2\lambda x_2, \tag{7}$$

the characteristic equation of which has two identical roots $\lambda_1 = \lambda_2 = -\lambda$; i.e., the origin is a stable degenerate node. In other domains, we have the following systems:

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = -2\lambda(x_2 - 1), \quad (x_1, x_2) \in D_2^-, \tag{8}$$

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = -2\lambda(x_2 + 1), \quad (x_1, x_2) \in D_2^+, \tag{9}$$

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = -1, \quad (x_1, x_2) \in D_3^-, \tag{10}$$

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = 1, \quad (x_1, x_2) \in D_3^+. \tag{11}$$

Equation (6) is an equivalent representation of the switching system (7)–(11).

3. PROOF OF GLOBAL STABILITY

First, we prove that, in the study of stability of the system. we can confine ourselves to the consideration of the trajectories beginning in the set D_1.

Proposition. System (7)–(11) is globally asymptotically stable if and only if any trajectory beginning in the set D_1 asymptotically tends to the origin.

The necessity of the assertion is evident. The sufficiency is proved by showing that any trajectory with arbitrary initial conditions occurs in the set D_1 in a finite time. Let us prove this. Indeed, from equations (10) and (11), it is seen that trajectories of the system in D_3^- and D_3^+ are parabolas

$$x_1 = \mp\frac{1}{2}x_2^2 + C. \tag{12}$$

Since any parabola cannot lie entirely in D_3^- or D_3^+ (see Fig. 1) and the system moves with the constant acceleration, it inevitably occurs in a finite time in either D_1 or D_2. Further, from equations (8) and (9), it is seen that, in D_2^- (D_2^+), the system moves in the positive (negative) direction of x_1 and $x_2(t) \to 1$ ($x_2(t) \to -1$). Then, it follows that the system inevitably enters D_1 in a finite time. Thus, for any initial conditions, after at most two switchings, the system occurs in the set D_1. Further, only trajectories beginning in D_1 are considered.

Theorem 1. *System* (6) *is globally asymptotically stable for any* $\lambda > 0$.

Proof. Let us find out whether the system can enter D_2 from D_1. For definiteness, consider the boundary between D_1 and D_2^+. From the first equation in (9), it is seen that the trajectory can intersect the boundary only if x_2 is positive, i.e., when the half-width $1/2\lambda$ of the strip D_2 is greater than one, like in the case shown in Fig. 2, which takes place only when $\lambda < 1/2$. Since the right-hand side of the second equation in (9) in this case is negative, $x_2(t)$ will change sign in a finite time, This, in turn, will change the direction of motion along x_1-axis, bringing thus the system to D_1 again. Note that the segment of the asymptote $x_2 = -\lambda x_1$ (the bold line in Fig. 2) of the linear system (7) for which $|x_1| \leqslant 2/\lambda$ lies completely in D_1. Since no trajectory of the system can intersect the asymptote in D_1, all trajectories asymptotically tend to the origin. The case of

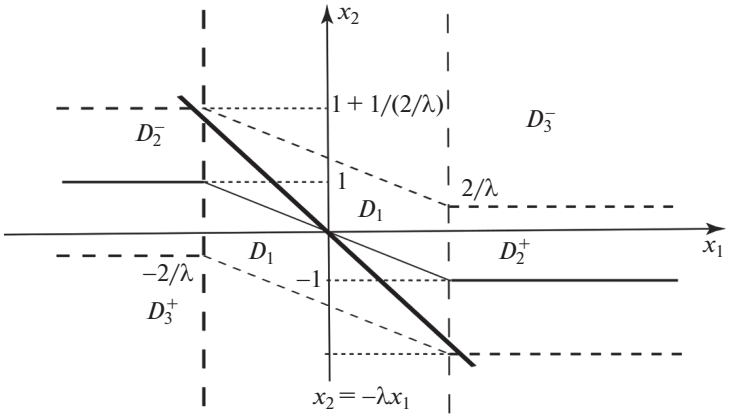


Fig. 2.

negative x_1 is considered similarly. Thus, for small $\lambda < 1/2$, the system is globally stable. Note that, in this case, any trajectory beginning in the set D_1 can intersect the boundary between the sets (D_1 and D_2) not more than twice.

Let us determine conditions under which the system can switch from D_1 to D_3. The boundary between the sets (dashed lines in Fig. 1) is given by the equations

$$x_2 = -\frac{\lambda}{2}x_1 \pm \frac{1}{2\lambda}, \quad -\frac{2}{\lambda} \leqslant x_1 \leqslant \frac{2}{\lambda}, \tag{13}$$

where the plus sign before the second addend corresponds to the upper boundary (the boundary between D_1 and D_3^-), and the minus sign, to the lower boundary. A trajectory can intersect the boundary only if its slope is less than that of the boundary, which is $\lambda/2$. From equations (10) and (11), we find that the slope of the trajectory on the boundary is $1/x_2$, from which it follows that the trajectory can intersect the upper (lower) boundary only at the points with ordinates satisfying the inequality $x_2 > 2/\lambda$ ($x_2 < -2/\lambda$), i.e., in the region

$$|x_2| > 2/\lambda. \tag{14}$$

Since the maximum value of $|x_2|$ in D_1 is achieved in two angular points with ordinates $\pm(1 + 1/(2\lambda))$, trajectories cannot intersect the boundary when $\lambda \leqslant 3/2$.

Thus, global stability of the system is proved for all $\lambda \leqslant 3/2$. Moreover, we proved the following nontrivial assertion.

Lemma 1. *Let $1/2 \leqslant \lambda \leqslant 3/2$. Then, D_1 is an invariant set of the switching system* (6).

Note that D_1 in this case is an invariant set of the linear system (7) either. Now, let us prove that the system is globally stable for any greater values of λ. From the above calculations, it follows that, for $\lambda > 3/2$, the system can pass from D_1 to only D_3. Consider, for definiteness, the upper part of the phase plane, where $U(x) < 0$. Constant C on the right-hand side of (12) depends on the coordinates of the point where the system passes from D_1 to D_3^-. Let x_{2*} denote the ordinate of the point where the trajectory intersects the boundary (the abscissa is uniquely determined from the equation of the boundary (13)). Then,

$$C \equiv C(x_{2*}) = \frac{1}{2}\left(x_{2*}^2 - \frac{4x_{2*}}{\lambda} + \frac{2}{\lambda^2}\right).$$

Substituting the right-hand side of this formula for C in (12) and solving the quadratic equation obtained, we find the ordinate (denote it as x_{2**}) of the second intersection point of the parabola

and the boundary (13), where the system switches from (10) to (7):

$$x_{2**} = \frac{4}{\lambda} - x_{2*}. \tag{15}$$

With regard to the inequalities

$$\frac{2}{\lambda} < x_{2*} \leqslant 1 + \frac{1}{2\lambda},$$

it follows from (15) that x_{2**} satisfies the inequalities

$$\frac{2}{\lambda} > x_{2**} \geqslant -1 + \frac{7}{2\lambda} > -1 + \frac{1}{2\lambda};$$

i.e., the second intersection point of the parabola and the line (13) belongs to the boundary between D_3^- and D_1, and, hence, the trajectory passing from D_1 to D_3 cannot occur in D_2. Thus, when $\lambda > 3/2$, switchings are possible only between the three systems with the domains D_1, D_3^-, and D_3^+. Similarly, two successive points of intersection of the boundry between D_1 and D_3^+ are found to be

$$x_{2**} = -\frac{4}{\lambda} - x_{2*}, \quad -\frac{2}{\lambda} < x_{2*} \leqslant -1 - \frac{1}{2\lambda}. \tag{16}$$

Since any trajectory cannot have self-intersections and does not go to infinity, it will suffice to prove that no closed trajectory (cycle) exists [15]. Let us assume the contrary: suppose that there exists a closed trajectory. From the above discussions, it follows that such a trajectory consists of four segments: two segments in D_1, one segment in D_3^+, and one segment in D_3^-, with the motion along the trajectory being clockwise.

Let us show that there exists a positive definite function that decreases on all segments of the cycle, from which it follows that the trajectory cannot be closed. Note that we do not mean a Lyapunov function, since we do not require negativeness of its derivative by virtue of system (6) at all points of the trajectory. We seek for a function the total variation of which ater passing the entire segment completely lying in one of the regions D_1, D_3^-, or D_3^+ is negative. For a candidate of the desired function, we take a quadratic Lyapunov function $F = \lambda^3 x^{\mathrm{T}} P x$ (the multiplier λ^3 is introduced for the convenience of notation) of the linear system (7). Here, P is a positive definite matrix of order two satisfying the linear matrix inequality (LMI) $A^{\mathrm{T}}P + PA < 0$ [16], and A is the matrix of the linear system (7):

$$A = \begin{pmatrix} 0 & 1 \\ -\lambda^2 & -2\lambda \end{pmatrix}.$$

In [17], it was shown that matrix P can be represented in the form

$$P = \begin{pmatrix} \lambda & q_1/2 \\ q_1/2 & q_2/\lambda \end{pmatrix}, \tag{17}$$

where $q_1, q_2 > 0$ belong to the ellipse Ω (Fig. 3) defined by the inequality

$$(q_2 - q_1 - 1)^2 + (q_1 - 2)^2 \leqslant 4. \tag{18}$$

Let us find out whether there exist $(q_1, q_2) \in \Omega$ such that function F decreases on each of the four segments.

The derivative of function F by virtue of system (7) is negative by definition, which guarantees that function F decreases on two trajectory segments lying in D_1. On the segments lying in D_3, negativeness of the derivative of F is not guaranteed; however, we prove further that the integral variation of F on each of these segments is negative; i.e., the value of the function at the boundary

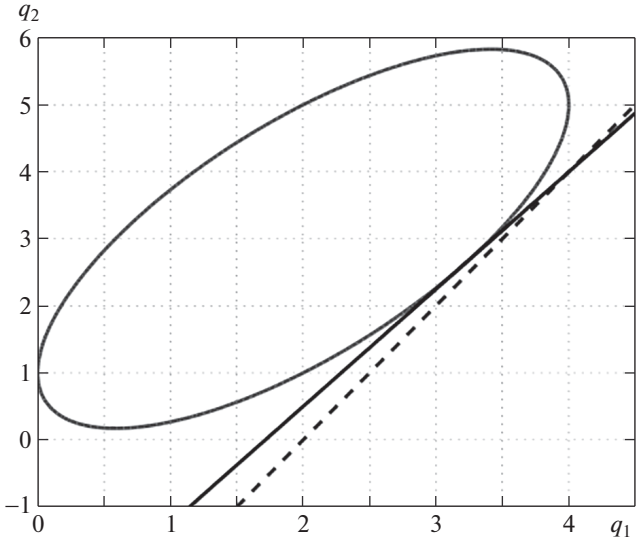


Fig. 3.

point where the system passes from D_1 to D_3 is greater than that at the point where it returns from D_3 to D_1.

Substituting the right-hand side of (17) into F, we get $F(x) = \lambda^2(\lambda^2 x_1^2 + \lambda q_1 x_1 x_2 + q_2 x_2^2)$. Expressing x_1 in terms of x_2 from the equation of boundary (13) and substituting it into the right-hand side of the formula for function F, we obtain the value of F on the upper boundary of D_1:

$$F(x_2) = c_1 x_2^2 - c_2 x_2 + 1, \tag{19}$$

where $c_1 = \lambda^2(q_2 - 2q_1 + 4)$ and $c_2 = \lambda(4 - q_1)$. It follows from inequality (18) that $q_1 < 4$ and, hence, $c_2 > 0\ \forall q_1, q_2 \in \Omega$. It is easy to show that ellipse (18) has no intersections with the straight line $q_2 - 2q_1 - 4 = 0$ (the dashed line in Fig. 3) and lies above it; hence, $c_1 > 0\ \forall q_1, q_2 \in \Omega$. Let us find the variation ΔF of function F on the trajectory segment lying in D_3^-. With regard to (19) and (15), at the beginning and end points of the segment, the function takes values

$$F(x_{2*}) = c_1 x_{2*}^2 - c_2 x_{2*} + 1, \quad F(x_{2**}) = c_1(4/\lambda - x_{2*})^2 - c_2(4/\lambda - x_{2*}) + 1.$$

Then, it follows that

$$\begin{aligned}\Delta F = F(x_{2**}) - F(x_{2*}) &= c_1(16/\lambda^2 - 8x_{2*}/\lambda) + 2c_2 x_{2*} - 4c_2/\lambda \\ &= (2c_2 - 8c_1/\lambda)x_{2*} + 16c_1/\lambda^2 - 4c_2/\lambda = -(8q_2 - 14q_1 + 24)(\lambda x_{2*} - 2).\end{aligned}$$

It is easy to verify that the straight line $8q_2 - 14q_1 + 24 = 0$ (solid line in Fig. 3) touches ellipse Ω and lies below it, so that the first multiplier is positive. Since, according to (14), $x_{2*} > 2/\lambda$, the second multiplier is positive either, so that $\Delta F < 0$ for any $(q_1, q_2) \in \Omega$.

Repeating these calculations for the lower boundary of set D_1 and taking into account (16), we find

$$F(x_2) = c_1 x_2^2 + c_2 x_2 + 1 \tag{20}$$

and

$$\begin{aligned}\Delta F &= c_1(4/\lambda + x_{2*})^2 - c_2(4/\lambda + x_{2*}) - c_1 x_{2*}^2 - c_2 x_{2*} \\ &= (8q_2 - 14q_1 + 24)(\lambda x_{2*} + 2).\end{aligned} \tag{21}$$

According to (14), on the lower boundary, $x_{2*} < -2/\lambda$, the second multiplier in (21) is negative; hence, the variation of function F on the trajectory segment lying in D_3^+ is also negative. Thus, for any $(q_1, q_2) \in \Omega$, the value of the quadratic Lyapunov function of the linear system (7) decreases after passing each trajectory segment, from which it follows that the trajectory cannot be a closed curve. The theorem is proved.

Numerical examples illustrating behavior of the trajectories of integrator (1) stabilized by means of feedback (2) can be found in [10, 11].

4. CONCLUSIONS

The problem of stabilizing a second-order affine system consisting of five subsystems, of which only one has a stable equilibrium, with a state-depending switching law has been considered. The system under study comes to existence when applying a feedback in the form of nested saturators for stabilizing a chain of two integrators. The advantages of the considered feedback are its continuity and boundedness, as well as the possibility to ensure desired characteristics of the transient process. By means of an appropriate selection of the four feedback coefficients, it is easy to ensure a desired type of the equilibrium and a desired exponential rate of the deviation decrease near the equilibrium state, as well as to constrain the maximum speed of approaching the equilibrium state, which is especially important in the case of large initial deviations. The main result of the study is the proof of global stability of the considered affine switching system.

REFERENCES

1. Goebel, R., Sanfelice, R.G., and Teel, A.R., *Hybrid Dynamical Systems: Modeling, Stability, and Robustness*, Princeton University Press, 2012.
2. Liberzon, D., *Switching in Systems and Control*, Boston: Birkhauser, 1973.
3. Lin, H. and Antsaklis, P.J., Stability and stabilizability of switched linear systems: A survey of recent results, *IEEE Trans. Automat. Control*, 2009, vol. 54, pp. 308–322.
4. Pyatnitskiy, E. and Rapoport, L., Criteria of asymptotic stability of differential inclusions and periodic motions of time-varying nonlinear control systems, *IEEE Trans. Circuits Systems I: Fundamental Theory and Applications*, 1996, vol. 43, no. 3, pp. 219–229.
5. Teel, A.R., Global stabilization and restricted tracking for multiple integrators with bounded controls, *Sys. & Cont. Lett.*, 1992, vol. 18, no. 3, pp. 165–171.
6. Teel, A.R., A nonlinear small gain theorem for the analysis of control systems with saturation, *Trans. Autom. Contr., IEEE*, 1996, vol. 41, no. 9, pp. 1256–1270.
7. Kurzhanski, A.B. and Varaiya, P., *Solution Examples on Ellipsoidal Methods: Computation in High Dimensions*, Cham, Switzerland: Springer, 2014.
8. Olfati-Saber, R., Nonlinear control of underactuated mechanical systems with application to robotics and aerospace vehicles, *Ph.D. Dissertation*, Massachusetts Institute of Technology. Dept. of Electrical Engineering and Computer Science, 2001.
9. Hua, M.-D. and Samson, C., Time sub-optimal nonlinear pi and pid controllers applied to longitudinal headway car control, *Int. J. Control*, 2011, vol. 84, pp. 1717–1728.
10. Pesterev, A.V., Morozov, Yu.V., and Matrosov, I.V., On Optimal Selection of Coefficients of a Controller in the Point Stabilization Problem for a Robot-wheel, *Communicat. Comput. Inform. Sci. (CCIS)*, 2020, vol. 1340, pp. 236–249.
11. Pesterev, A.V. and Morozov, Yu.V., Optimizing coefficients of a controller in the point stabilization problem for a robot-wheel, *Lect. Notes Comput. Sci.*, vol. 13078, Cham, Switzerland: Springer, 2021, pp. 191–202.

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

13. Pesterev, A.V. and Morozov, Yu.V., Global Stability of a Switched Affine System, *Proc. of the 16th Int. Conf. on Stability and Oscillations of Nonlinear Control Systems (Pyatnitskiy's Conference)*, 2022, pp. 1–4.

14. Pesterev, A.V. and Morozov, Yu.V., Stabilization of a cart with inverted pendulum, *Autom. Remote Control*, 2022, vol. 83, no. 1, pp. 78–91.

15. Andronov, A.A., Leontovich, E., Gordon, I.I., and Maier, A., *Qualitative Theory of Second-order Dynamic Systems*, Wiley, 1973.

16. Boyd, S., Ghaoui, L.E., Feron, E., and Balakrishnan, V., *Linear Matrix Inequalities in System and Control Theory*, Philadelphia: SIAM, 1994.

17. Pesterev, A.V., Construction of the Best Ellipsoidal Approximation of the Attraction Domain in Stabilization Problem for a Wheeled Robot, *Autom. Remote Control*, 2011, vol. 72, no. 3, pp. 512–528.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 9, pp. 1094–1103.

ROBUST, ADAPTIVE, AND NETWORK CONTROL

Velocity of Flow on Regular Non-Homogeneous Open One-Dimensional Net with Non-Symmetrical Arrangement of Nodes

A. S. Bugaev[*,a], **M. V. Yashina**[**,***,****,b], **and A. G. Tatashev**[**,***,c]

[*] *Moscow Institute of Physics and Technology, Moscow, Russia*
[**] *Moscow Automobile and Road Construction Technical University, Moscow, Russia*
[***] *Moscow Technical University of Communications and Informatics, Moscow, Russia*
[****] *Moscow Aviation Institute (National Research University), Moscow, Russia*
e-mail: [a] *bugaev@cos.ru,* [b] *mv.yashina@madi.ru,* [c] *a-tatashev@yandex.ru*

Received May 15, 2022
Revised June 30, 2023
Accepted July 20, 2023

Abstract—A system is studied such that this system belongs to the class of dynamical systems called the Buslaev nets. This class has been developed for the purpose of creating traffic models on network structures such that, for these models, analytical results can be obtained. There may be other network applications of Buslaev nets. The considered system is called an open chain of contours. Segments called clusters move along circumferences (contours) according to prescribed rules. For each contour (except the leftmost and rightmost contours) there are two adjacent contours. Each of the leftmost and rightmost contours has one adjacent contour. There is a common point (node) for any two adjacent contours. Results have been obtained on the average velocity of cluster movement, taking into account delays during the passage through nodes. These results generalize the results obtained previously for a particular case of the system under consideration.

Keywords: dynamical systems, mathematical traffic models, Buslaev nets

DOI: 10.25728/arcRAS.2023.27.84.001

1. INTRODUCTION

A class of mathematical traffic models consists of models in which particles move in a one-dimensional or two-dimensional lattice. These models can be interpreted as cellular automata [1] or exclusion processes [2]. In [3], the Nagel-Schreckenberg traffic model has been introduced. This model were studied by a number of authors. In this model, an infinite or closed lattice is a sequence of cells, and particles move along the lattice according to prescribed rules. Analytical results for simple versions of models of this class have been obtained, e.g., in [4–10]. Models with network structures, belonging to this class of models, were studied mainly by simulation.

The paper [7] (a preprint of this paper was published in 1999), the movement of particles along closed lattice is considered. It is assumed that, at each step, each particle moves onto one cell forward if the cell ahead is vacant, and the particle does not move if this cell is occupied. Suppose a cellular automaton corresponds to the system. If a cell of the automaton is in the state 1, then the cell corresponds to an occupied cell of the system, and, if a cell of the automaton is in the state 0, then the cell corresponds to a vacant cell. As it noted in [7], this automaton is the elementary cellular automaton 184 in terms of S Wolfram classification [1]. According to the results obtained in [7], if the ratio of the number of particles to the number of cells (particle density) does not

exceed 1/2, then, for any initial state, from a certain point in time, all particles move at each moment without delays (free movement, self-organization). If the density is greater than 1/2, then the average velocity of particles (the ratio of average number of moving particles per a time unit) is equal to $(1-\rho)/\rho$ where ρ is the density. Analogous results were obtained independently in [5], where moreover an upper bound for the time it takes for the system to reach limit mode. In [6], analytical results were obtained for more general model. In this model, with prescribed probability, a particle moves from the cell i to the cell $i+1$ provided that the cell $i+1$ is vacant. This probability depends on the states of cells $i-1$ and $i+2$ (cells are numbered in the direction of movement). The behavior of the system was studied in [6] for particular cases. In [8], a formula has been obtained for the average velocity of particles in the stochastic traffic model such that, in this model, at any step, with prescribed probability, each particle moves onto one cell forward if the cell ahead is vacant. Some generalisations of results found in [5, 7, 8] were obtained in [9] where a dynamical system with continuous state space was studied. In particular cases, this system is equivalent to systems considered in [5, 7, 8]. In [10], a stochastic traffic model is studied. In this model, particles move along a not closed lattice containing a finite number of cells. Particles appear at one end of the lattice and move in the direction to the opposite end. After reaching this end, the particles leave the lattice. In [10], a matrix approach has been developed to analyze the system.

In [11], a two-dimensional traffic model was proposed (the Biham–Middleton–Levine model, BML model). In this model, particles move along two-dimensional lattice in orthogonal directions. Particles of the first type move along rows, and particles of the second type move along columns. The rule of particle movement is a two-dimensional counterpart of the elementary automaton 184. This analogy is noted in [11]. In [11–18], different versions of BML model was considered, and analytical results have obtained mainly regarding conditions for self-organization and conditions for jam).

In [19], a graph with a variable configuration of particles has been introduced. The developed approach makes it possible to simulate phenomena that arise in complex networks (for example, in transport and social network).

The book [20] is a monograph on mathematical modeling of traffic flows. Based on the material presented in the book, one can see that quite few approaches are known to analytical study of traffic flows with network structure. This determines that it is relevant to develop new such approaches. One of these approaches is the development of models based on Buslaev nets.

In [21], the concept of cluster movement in transport models was introduced. In discrete version, clusters are groups of particles located in adjacent cells and moving simultaneously. In this case, the movement of particles corresponds to the rule of an elementary cellular automaton 240. In the continuous version, clusters are moving segments and are called clusters by analogy with a discrete version.

A.P. Buslaev has developed a class of dynamical systems, which now are called Buslaev nets [22]. A Buslaev net is a dynamic system containing a system of contours. Adjacent contours have common points, called nodes. In the discrete version, a contour is a closed sequence of particles. In the continuous version, the contour is represented as a circumference. Segments called clusters move along contours according to prescribed rules. As particles pass through nodes, delays occur due to the fact that more than one particle cannot pass through a node at the same time. The main problems in study of Buslaev nets are to find the average velocity of particles (clusters), conditions for the system to enter a state of free movement (starting from a certain moment, all clusters move without delay at the current moment and in the future) or collapse (from a certain moment no particle moves). Analytical results have been obtained for two-contour nets with one [23] or two [24] common nodes and Buslaev nets with regular periodic structures [25–30].

In [26], a Buslaev net called an open chain of contours is considered. In [26], a version of open chain is studied such that this net is an open chain with continuous time scale. There is a common point (node) for a contour and each adjacent contour except for the leftmost and rightmost contours. Each of two later contours have one adjacent contour. There is a common node for two adjacent contours. In each contour, there is a cluster. The length of each contour is the same. Each contour is divided by two nodes into parts of the same length. It has been proved that if the length of cluster is not greater than half the length of the contour, then the system from a certain moment is in a state free movement, i.e., all clusters move without delay, and if the cluster length is greater than half contour length, then all clusters move with the same average velocity, which is less than the velocity of free movement and independent of the initial state of the system. A formula for the average velocity of clusters was obtained. In [27], the average velocity of clusters has been obtained for a version of open chain with discrete state space and discrete time provided that the length of each contour is the same (the same number of cells in contour) and each contour is divided by nodes into two parts of the same length. Examples are given showing that, in the general case, clusters can move with unequal average velocities, and the average velocity can depend on the initial state of the system. In [30], the limit distribution (invariant measure) has been found for an open chain with contours of the same length and clusters of unequal length.

In [23], an asymmetrical two-contour system with one node was considered. This system is a particular case of a heterogeneous open chain of contours, for which the number of contours is equal to two. The following example of a possible application of the results on contour networks is given. Let, during the working day, raw materials or fuel are constantly delivered to two departments of the enterprise from the warehouse. The cargo is delivered by vehicles such as trucks along narrow-gauge tracks. The warehouse is located at the intersection of these paths. A vehicle that arrives to a warehouse while another vehicle is loading waits for service to complete and then begins loading. Suppose that the ith vehicle travels from the warehouse to the department and back in time $c_i - l_i$, taking into account the unloading time in the department, and loading time for the vehicle at the warehouse lasts l_i units of time, $i = 1, 2$. Then the process of vehicles movement is modeled by a system of the type under consideration with contour lengths equal to c_1 and c_2 and cluster lengths equal to l_1 and l_2, respectively, under the assumption that, in the absence of delays, each cluster moves at unit velocity.

This example can be generalized. Let us assume that there are three departments of the enterprise, two warehouses and three vehicles, each of which delivers cargo to the department to which the vehicle is related. Cargo is delivered to the department 1 from the warehouse 1. Cargo is delivered to the department 3 from the warehouse 2. Cargo is delivered to the department 2 alternately from both warehouses. The transport work process may be modeled by a heterogeneous three-contour open chain of contours. Each contour corresponds to a vehicle path. Each of the two nodes corresponds to one of the warehouses.

In this paper, we prove a theorem regarding the behavior of the system in the case when the length of the contour and the length of moving cluster depend on the contour number, and contours can be divided by the nodes into parts of unequal lengths. It is assumed that the clusters have sufficiently large lengths. The formula is obtained for the cluster velocity. This formula is a generalization of the formula obtained in the paper [27] for the particular case considered in that paper.

2. DESCRIPTION OF SYSTEM

Suppose a dynamical system, Fig. 1. The system contains N circumferences called *contours*. The length of the cluster i is equal to c_i, $i = 1, \ldots, N$. The coordinate system $[0, c_i)$ is given for the contour i, $i = 1, \ldots, N$. There is a common point of the nodes i and $i + 1$ called the *node*

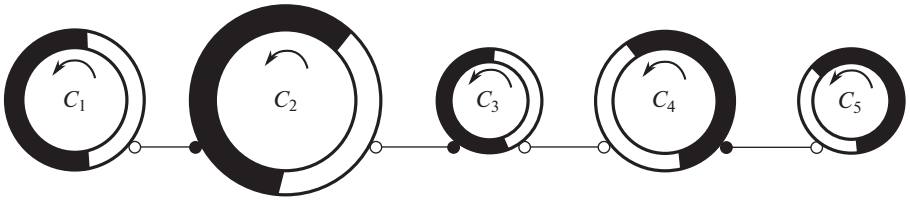


Fig. 1. An open chain of contours, $N = 5$, c_i is the length of the contour i, l_i is the length of the cluster i, $i = 1, \ldots, 5$.

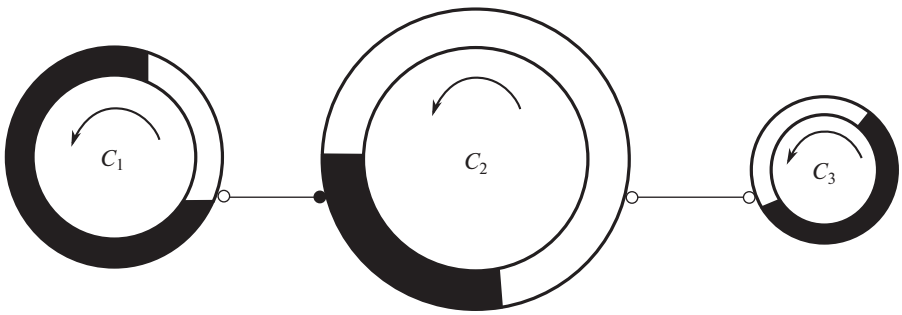


Fig. 2. The cluster 2 occupies the node $(1, 2)$, $l_2 + d_2 < c_2$. A delay of the cluster 1 at the node (1,2).

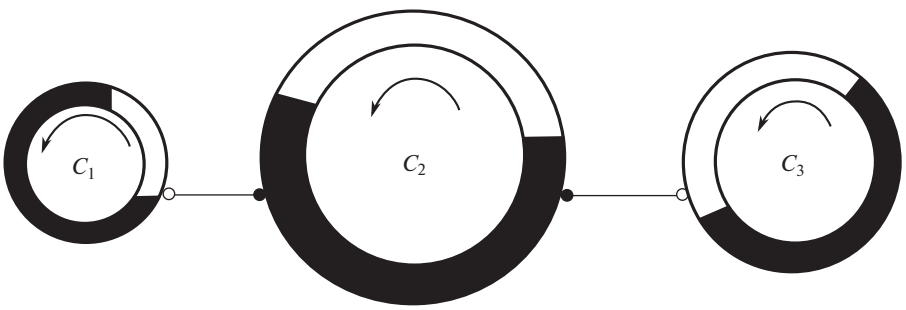


Fig. 3. The cluster 2 occupies the node $(1, 2)$, $l_2 + d_2 > c_2$. A delay of the cluster 1 at the node (1,2).

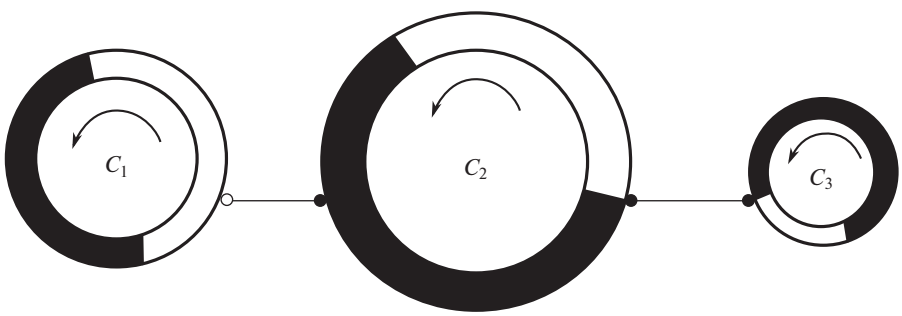


Fig. 4. A competition of the clusters 2 and 3.

$(i, i + 1)$, $i = 1, \ldots, N - 1$. For the contour i, the coordinate of the node $(i, i + 1)$ is equal to 0, and, for the contour $i+1$, the coordinate of the node $(i, i + 1)$ is equal to $d_{i+1} > 0$, $i = 1, \ldots, N - 2$. Suppose the coordinate of the node $(N - 1, N)$ is equal to 0 in both the contour $N - 1$ and the contour N. There is a moving segment of the length l_i, $i = 1, \ldots, N$. The segment is called a *cluster*. The direction of the coordinate axis i is the same as the direction of the cluster movement, $i = 1, \ldots, N$. If delays of a cluster do not occur, then the cluster moves with velocity 1, i.e., the cluster i makes a full revolution in c_i units of time, $i = 1, \ldots, N$. If, at time $t \geq 0$, the coordinate i

is equal to $x_i(t)$, then the cluster i is located on the arc $(x_i(t) - l_i, x_i(t))$ (subtraction modulo c_i), $i = 1, \ldots, N$. The *state* of the system at time t is the vector $x(t) = (x_1(t), \ldots, x_N(t))$, where $x_i(t)$ is the coordinate of the leading point of the cluster i, $i = 1, \ldots, N$. We say that, at time t, the cluster i occupies the node $(i, i+1)$ if $0 < x_i(t) < l_i$, $i = 1, \ldots, N-1$. We say that, at time t, the cluster i occupies the node $(i-1, i)$ if $d_i < x_i(t) < d_i + l_i$ (for $d_i + l_i < c_i$), Fig. 2, or $0 \le x_i(t) < d_i + l_i - c_i$ (for $d_i + l_i \ge c_i$), Fig. 3, $i = 2, \ldots, N-1$, Fig. 4, $0 < x_i(t) < l_i$, $i = N$. We say that, at time t, the cluster i is at the node $(i, i+1)$ if $x_i(t) = 0$, $i = 1, \ldots, N-1$. We say that, at time t, the cluster i is at the node $(i-1, i)$ if $x_i(t) = d_i$, $i = 2, \ldots, N-1$; $x_N(t) = 0$. The state is *admissible* if no node is occupied by two clusters. The system state at time $t = 0$ (*the initial state*) prescribed and must be admissible. A delay of cluster is at a node such that, at the moment, this node is occupied by the cluster of adjacent contour. The delay ends when this node stops being busy. If clusters i and $i+1$ are simultaneously located at node $(i, i+1)$, then *competition* between these clusters occurs (Fig. 4) and the cluster passes through the node first, chosen in accordance with deterministic or stochastic competition resolution rule. If the conditions of Theorem 1, which is proved in Section 4, hold, then, for any competition resolution rule, and the cycle is implemented such that no competitions occur, and the cycle is independent of the initial state.

3. LIMIT CYCLES. THE AVERAGE VELOCITY

The system is deterministic, and its behavior in the future is determined by its state at present. Therefore, if at some moment a state is repeated, then from the moment the states of the system will be periodically repeated, forming a cycle (limit cycle).

Let $H_i(t)$ be a total distance traveled by the cluster i in the time interval $(0, t)$. Then, if there exists the limit

$$v_i = \lim_{t \to \infty} \frac{H_i(t)}{t},$$

then this limit is called the average velocity of the cluster i, $i = 1, \ldots, N$. It is obvious that, if a cycle is implemented, then this limit exists, and the limit equals the ratio of the distance traveled by the cluster during the cycle to the period.

4. BEHAVIOR OF THE SYSTEM

In Section 4, statements about the behavior of the system are proved.

Lemma 1. *If*

$$\frac{l_i}{c_i} + \frac{l_{i+1}}{c_{i+1}} > 1 \tag{1}$$

i and $i+1$ $(1 \le i \le N-1)$, then the average velocity of at least one of the clusters i and $i+1$ is less than 1.

Proof. Suppose there exists i such that $v_i = v_{i+1} = 1$, and therefore the clusters i and $i+1$ $(1 \le i \le N-1)$ move without delays. Then the first term in the left side of the inequality (1) is the limit of the ratio of the total time in the interval $(0, t)$ during that the node $(i, i+1)$ is occupied by the cluster i to t as $t \to \infty$, and the second term is the limit of the ratio of the total time in the interval $(0, t)$ during that the node $(i, i+1)$ is occupied by the cluster $i+1$ to t as $t \to \infty$. Therefore the sum of these terms is not greater than 1 since the node cannot be occupied by two clusters simultaneously. The contradiction proves Lemma 1.

The following theorem characterizes the behavior of the system for sufficiently large cluster lengths (heavy load).

Theorem 1. *Suppose the following conditions hold*

$$l_i > \max(d_i, c_i - d_i), \quad i = 2, \dots, N-1, \tag{2}$$

$$l_1 + l_i > c_i, \quad i = 2, \dots, N, \tag{3}$$

$$l_i + l_N > c_i, \quad i = 1, \dots, N-1, \tag{4}$$

and at least one of two conditions holds

$$\frac{l_1}{c_1} + \frac{l_2}{c_2} > 1, \tag{5}$$

$$\frac{l_{N-1}}{c_{N-1}} + \frac{l_N}{c_N} > 1. \tag{6}$$

Then, for any deterministic or stochastic competition resolution rule, the same limit cycle is implemented, in which no competitions occur. The period of the cycle is equal to

$$T = l_1 + l_N + 2\sum_{j=2}^{N-1} l_j - \sum_{j=2}^{N-1} c_j. \tag{7}$$

The average velocity of the cluster i equals

$$v_i = \frac{c_i}{l_1 + l_N + 2\sum\limits_{j=2}^{N-1} l_j - \sum\limits_{i=2}^{N-1} c_i}, \quad i = 1, \dots, N. \tag{8}$$

Proof. Suppose the conditions (2)–(5) hold. According to Lemma 1, a delay of at least one of the clusters 1 and 2 occur.

Suppose, at time t_1, a delay of the cluster 2 begins at the node $(1, 2)$. Then we have $0 < x_1(t_1) < l_1$, and the movement of cluster 2 resumes at time $t_0 = t_1 + l_1 - x_1(t_1)$, and

$$x_1(t_0) = l_1, \ x_2(t_0) = d_2. \tag{9}$$

Suppose, at time t_1, a delay of the cluster 1 at the node $(1, 2)$. If, at time $t_2 > t_1$, this delay ends, then $x_1(t_2) = 0$, $x_2(t_2) = d_2 + l_2 - c_2$. For $t_3 = t_2 + c_2 - l_2$, we have $x_2(t_3) = d_2$. According to (3), we have $c_2 - l_2 < l_1$. Therefore, at time t_2, the cluster 1 occupies the node $(1, 2)$, and a delay of the cluster 2 begins at the node $(1, 2)$. Let us make sure that the delay of the cluster 1, starting at time t_1, will end. If the delay of the cluster 1 never ends, then, from time t, the cluster 1 is at the node $(1, 2)$. We can prove by induction that, for $2 \le i \le N-1$, from some moment, the cluster i is, at the node $(i, i+1)$ at present and in the future. There exists t such that, at time t, the system is in the state such that $x_{N-1}(t) = 0$, $x_N(t) = l_N - \frac{c_N}{2}$. After this moment, the cluster $N-1$ begins to move. Hence there exists a moment such that the movement of a cluster is resumed. Let t_4 be the minimum value $t_4 > t_0$ such that, at time t_4, the movement of the cluster $1 \le i_0 \le N-1$ is resumed, and the clusters $1, 2, \dots, i_0 - 1$ do not move. Then, at time

$$t_4 + \sum_{k=0}^{i_0-j-1} (d_{i_0-k} + l_{i_0-j} - c_{i_0-k}),$$

the movement of the cluster $j = 1, \dots, i_0 - 1$ is resumed. In particular, the movement of the cluster 1 is resumed.

Thus, there exists t_0 such that (9) holds.

At time $t_0 - l_1$, the cluster 2 was at the node $(1, 2)$, and, from this time, the cluster occupied this node. If (2) holds, then at least each of the clusters $i = 2, \ldots, N-1$ occupies at least one node. Hence, at time $t_0 - l_1$, the cluster 2 occupies the node $(2, 3)$. At time $t_0 - l_1$, each of the clusters $i = 2, \ldots, N-1$ occupies the node $(i, i+1)$. The proof is by induction. Therefore,

$$d_i + l_i - c_i \leqslant x_i(t_0) \leqslant d_i, \quad i = 3, \ldots, N-1. \tag{10}$$

Combining (2), (3), (9), (10), we get that, at time t_0, the system is in the state

$$x(t_0) = (l_1, d_2, \ldots, d_{N-1}, 0).$$

At time t_0, the movement of the cluster 2 begins and the movement of the cluster 1 continues. The movement of the cluster i is resumed at time

$$t(i) = t_0 + \sum_{j=2}^{i-1} (l_j - d_j), \ i = 3, \ldots, N.$$

The cluster this cluster finds the node occupied and waits for it to become free. According to (2), (4), each of the clusters $i = 1, \ldots, N$, approaching the node $(i, i+1)$, waits for its release. At time $u_0 + a$

$$a = l_N + \sum_{j=2}^{N-1} (l_j - d_j), \tag{11}$$

the system is in the state

$$x(t_0 + a) = (0, \ldots, 0, l_N).$$

The movement of the cluster i is resumed at time

$$u(i) = t_0 + a + \sum_{j=i+1}^{N-1} (d_j + l_j - c_j), \quad i = 2, \ldots, N-1.$$

According (2), (3), at time

$$u(1) = t_0 + \sum_{i=2}^{N-1} (d_i + l_i - c_i),$$

the movement of the cluster 1 is resumed at time $u_0 = t_0 + a + b$, the system enters the state

$$x(u_0) = x(t_0 + a + b) = (l_1, d_2, \ldots, d_{N-1}, 0) = u(t_0), \tag{12}$$

$$b = l_1 + \sum_{i=2}^{N-1} (d_i + l_i - c_i). \tag{13}$$

Combining (11), (13), we obtain

$$a + b = l_1 + l_N + 2 \sum_{j=2}^{N-1} l_i - \sum_{i=2}^{N-1} c_i. \tag{14}$$

Therefore a cycle with period $a + b$ is implemented. During the cycle, there is no competition between clusters simultaneously located at the same node. Each cluster makes one revolution during the cycle. Using (14), we obtain the theorem under the conditions (2)–(5).

The proof does not take into account the competition resolution rule. Therefore, due to symmetry, condition (5) should be replaced by condition (6). The theorem has been proved.

Corollary 1. *Under the conditions of the theorem, the period* (7) *of the implemented cycle and the average velocities* (8) *of the clusters do not depend on* $d_2, \dots, d_{N-1}$.

The statement follows from Theorem 1.

Suppose

$$c_1 = \cdots = c_N = 1, \quad d_2 = \cdots = d_{N-1} = \frac{1}{2}, \tag{15}$$
$$l_i > \frac{1}{2}, \quad i = 1, \dots, N,$$

i.e., the length of each contours is the same(without loss of generality, we assume that this length equals 1), the nodes divide the contour into two equal parts, and the length of each cluster is greater than half the length of the contour. If (15) holds, then (7), (8) has form [21]

$$T = l_1 + l_N + 2\sum_{i=2}^{N-1} l_i - N + 2, \tag{16}$$

$$v_i = \frac{1}{l_1 + l_N + 2\sum\limits_{i=2}^{N-1} l_i - N + 2}, \quad i = 1, \dots, N. \tag{17}$$

Note that, in this case, the average velocity of any cluster is the same. In [22], it has been shown by examples that, for a discrete open chain with the contours of the same length and the clusters of different lengths, among which there can be clusters of lengths greater than half the length of the contours, and clusters of lengths not greater than half the length of the contours the average velocity can depend on the initial state, and the average velocities of clusters can be different for the same initial state of the system. Let us give an example for the considered continuous chain.

Example 1. Suppose the conditions (15), $N = 3$, $l_1 = l_3 = 0.75$, $l_2 = 25$ hold. Let $(0.75, 0.5, 0.25)$ be the initial state. Then any cluster comes to a node when the node is not occupied. Indeed, we have $x(0) = (0.75, 0.5, 0.75)$, $x(0,5) = (0.25, 0, 0.25)$, $x(1) = (0.75, 0.5, 0.75) = x(0)$. Hence the initial state belongs to the cycle with the velocity 1. If the initial state is $(0.25, 0.5, 0.75)$, then we have $x(0) = (0.25, 0.5, 0.75)$, $x(0.5) = (0.75, 0.5, 0.25)$, $x(1) = (0.25, 0, 0.75)$, $x(1.5) = (0.75, 0, 0.25)$, $x(2) = (0.25, 0.5, 0.75) = x(0)$. Hence, during the cycle with period 2, the clusters 1 and 3 move without delays and make two revolutions, and the cluster 2 makes only one revolution. Therefore, $v_1 = 1$, $v_2 = 1/2$, $v_3 = 1$. Hence, there exists an initial state such that the velocity of any cluster is equal to 1, and there exists an initial state such that the velocity of the clusters 1 and 3 is equal to 1, and the average velocity of the cluster 2 is equal to 1/2.

If the conditions (15) and $l_1 = \cdots = l_n = l > 1/2$ hold, then (16), (17) have the form [18]

$$T = 2(N-1)l - N + 2,$$
$$v_i = \frac{1}{2(N-1)l - N + 2}, \quad i = 1, \dots, N.$$

Under the conditions (15) and $l_1 = \cdots = l_N = l \le 1/2$, it has been proved in [19] that the system results in the state of free movement from any initial state.

Thus, for an open chain with contours of the same length and clusters of the same length, the average velocity is the same for all clusters and does not depend on the initial state of the system in contrast to a closed chain with contours of equal length and clusters of different lengths [27], for which the average velocity of clusters depends on the initial state. In [23], the limit state distribution (invariant measure) for an open chain with contours of the same length and clusters with the same length under the condition $l > 1/2$ has been found.

5. CONCLUSION

A theorem has been proved about the behavior of a dynamical system called an open chain of contours. The system belongs to the class of Buslaev nets. Previously, the system was considered under the assumption that all contours have the same length and the nodes divide the contours into parts of the same length. In this paper, it is supposed that the lengths of the contours may be different. Each contour can be divided into parts of different lengths. The system is considered under the assumption that clusters located in the contours have sufficiently large lengths. It has been proved that under the considered assumptions the limit cycle of the system is unique. The average cluster velocities and the period of the limit cycle are found. The results of the work can be used in traffic modeling, and also have other applications, in particular, in infocommunication systems modeling.

FUNDING

The work was carried out with financial support as part of a government assignment Ministry of Science and Higher Education of the Russian Federation (FSFM-2023-0003).

REFERENCES

1. Wolfram, S., Statistical mechanics of cellular automata, *Rev. Mod. Phys.*, 1983, vol. 55, pp. 601–644. https://doi.org/10.1103/RevModPhys.55.601
2. Spitzer, F., Interaction of Markov processes, *Advances in Mathematics*, 1970, vol. 5, no. 2, pp. 246–290.
3. Nagel, K. and Schreckenberg, M., A cellular automaton model for freeway traffic, *J. Phys. I*, 1992, vol. 2, no. 12, pp. 2221–2229. https://doi.org/10.1051/jp1:1992277
4. Schreckenberg, M., Schadschneider, A., Nagel, K., and Ito, N., Discrete stochastic models for traffic flow, *Phys. Rev. E*, 1995, vol. 51, pp. 2939–2949. https://doi.org/10.1103/PhysRevE.51.2939
5. Blank, M.L., Exact analysis of dynamical systems arising in models of traffic flow, *Russian Mathematical Surveys*, 2000, vol. 55, no. 3, pp. 562–563. https://doi.org/10.1070/RM2000v055n03ABEH000295
6. Gray, L. and Griffeath, D., The ergodic theory of traffic jams, *J. Stat. Phys.*, 2001, vol. 105, no. 3/4, pp. 413–452.
7. Belitsky, V. and Ferrari, P.A., Invariant measures and convergence properties for cellular automation 184 and related processes, *J. Stat. Phys.*, 2005, vol. 118, no. 3/4, pp. 589–623. https://doi.org/10.1007/s10955-004-8822-4
8. Kanai, M., Nishinari, K., and Tokihiro, T., Exact solution and asymptotic behaviour of the asymmetric simple exclusion process on a ring, *J. Phys. A: Mathematical and General*, 2006, vol. 39, no. 29, 9071. https://doi.org/10.1088/0305-4470/39/29/004
9. Blank, M., Metric properties of discrete time exclusion type processes in continuum, *J. Stat. Phys.*, 2010, vol. 140, no. 1, pp. 170–197. https://doi.org/10.1007/s10955-010-9983-y
10. Evans, M.R., Rajewsky, N., and Speer, E.R., Exact solution of a cellular automaton for traffic, *J. Stat. Phys.*, 2010, vol. 95, pp. 45–56. https://doi.org/10.1023/A:1004521326456
11. Biham, O., Middleton, A.A., and Levine, D., Self-organization and a dynamic transition in traffic-flow models, *Phys. Rev. A*, 1992, vol. 46, no. 10, pp. R6124–6127. https://doi.org/10.1103/PhysRevA.46.R6124
12. Angel, O., Holroyd, A.E., and Martin, J.B., The Jammed Phase of the Biham-Middleton-Levine Traffic Model, *Electronic Communications in Probability*, 2005, vol. 10, paper 17, pp. 167–178. https://doi.org/10.48550/arXiv.math/0504001
13. D'Souza, R.M., Coexisting phases and lattice dependence of a cellular automata model for traffic flow, *Physical Review E*, 2005, vol. 71, 0066112.
14. D'Souza, R.M., BML revisited: Statistical physics, computer simulation and probability, *Complexity*, 2006, vol. 12, no. 2, pp. 30–39.

15. Austin, T. and Benjamini, I., For what number must self organization occur in the Biham-Middleton-Levine traffic model from any possible starting configuration?, *arXiv preprint math/0607759*, 2006.

16. Pan Wei, Xue Yu, Zhao Rui, and Lu Wei-Zhen, Biham–Middleton–Levine model in consideration of cooperative willingness, *Chin. Phys. B*, 2014, vol. 23, no. 5, 058902. https://doi.org/10.1088/1674-1056/23/5/058902

17. Wenbin Hu, Liping Yan, Huan Wang, Bo Du, and Dacheng Tao, Real-time traffic jams prediction inspired by Biham, Middeleton and Levine (BML), *Information Sciences*, 2017, pp. 209–228. https://doi.org/10.1016/j.ins.2016.11.023

18. Moradi, H.R., Zardadi, A., and Heydarbeygi, Z., The number of collisions in Biham–Middleton–Levine model on a square lattice with limited number of cars, *Appl. Math. E-Notes*, 2019, vol. 19, pp. 243–249.

19. Malecky, K., Graph cellular automata with relation-based neighbourhoods of cells for complex systems modelling: A case of traffic simulation, *Symmetry*, 2017, vol. 9, 322. https://doi.org/10.3390/sym9120322

20. Gasnikov, A.V. et al., *Introduction to mathematical modeling of traffic flows*, 2nd ed., Gasnikov, A.V., Ed., Moscow: MTsNMO, 2013.

21. Bugaev, A.S., Buslaev, A.P., Kozlov, V.V., and Yashina, M.V., Distributed problems of monitoring and modern approaches to traffic modeling, *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, Washington, USA, 5–7 October 2011, pp. 477–481. https://doi.org/10.1109/ITSC.2011.6082805

22. Kozlov, V.V., Buslaev, A.P., and Tatashev, A.G., On synergy of totally connected flows on chainmails, *Proc. of the 13th International Conference of Computational and Applied Methods in Science and Engineering*, Almeria, Spain, 24–27 June 2013, vol. 3, pp. 861–874.

23. Myshkis, P.A., Tatashev, A.G., and Yashina, M.V., Cluster motion in a two-contour system with priority rule for conflict resolution, *Journal of Computer and Systems Sciences International*, 2020, vol. 59, no. 3, pp. 311–321. Translated from *Izvestiya RAN. Teoriya i Sistemy Upravleniya*, 2020, vol. 20, no. 3, pp. 3–13. https://doi.org/10.1134/S1064230730030119

24. Yashina, M. and Tatashev, A., Spectral cycles and average velocity of clusters in discrete two-contours system with two nodes, *Math. Meth. Appl. Sci.*, 2020, vol. 43, no. 7, pp. 4303–4316. https://doi.org/10.1002/mma.6194

25. Buslaev, A.P., Tatashev, A.G., and Yashina, M.V., Qualitative properties of dynamical system on toroidal chainmails, *AIP Conference Proceedings*, 2013, vol. 1558, pp. 1144–1147. https://doi.org/10.1063/1.4825710

26. Buslaev, A.P. and Tatashev, A.G., Spectra of local cluster flows on open chain of contours, *Eur. J. Pure Appl. Math.*, 2018, vol. 11, no. 3, pp. 628–641. https://doi.org/10.29020/nybg.ejpam.11i3.3292

27. Yashina, M. and Tatashev, A., Discrete open Buslaev chain with heterogeneous loading, *2019 7th International Conference on Control, Mechatronics and Automation (ICCMA)*, 6–8 Nov. 2019, Delft, Netherlands, pp. 283–288. https://doi.org/10.1109/ICCMA46720.2019.8988654

28. Bugaev, A.S., Yashina, M.V., Tatashev, A.G., and Fomina, M.Yu., On velocity spectrum for saturated flows on a regular open one-dimensional network, *XI All-Russian Multiconference on Management Problems MKPU-2021, Material of the XIV Multiconference: in 4 volumes*, Rostov-on-Don, 2021, pp. 41–44.

29. Bugaev, A.S., Tatashev, A.G., and Yashina, M., Spectrum of a continuous closed symmetric chain with an arbitrary number of contours, *Mathematical Models and Computer Simulation*, 2021, vol. 13, no. 6, pp. 1014–1027. Translated from *Matematicheskoe Modelirovanie*, 2021, vol. 33, no. 14, pp. 21–44. https://doi.org/10.1134/S207004822106003X

30. Yashina, M.V. and Tatashev, A.G., Invariant measure for continuous open chain of contours with discrete time, *Computational and Mathematical Methods. e1197*, First published: 28 September 2021. https://doi.org/10.1002/cmm4.1197

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 9, pp. 1104–1114.

CONTROL IN TECHNICAL SYSTEMS

An Indirect Single-Position Coordinate Determination Method Considering Motion Invariants under Singular Measurement Errors

Yu. G. Bulychev

JSC Concern Radioelectronic Technologies, Moscow, Russia
e-mail: profbulychev@yandex.ru

Received May 18, 2022
Revised March 22, 2023
Accepted June 9, 2023

Abstract—The problem of indirect single-position coordinate determination based on the smoothed measurements of bearing and the radial velocity of an object is solved considering motion invariants and singular measurement errors. Such errors are represented as an appropriate linear combination with unknown spectral coefficients in a given finite-dimensional functional space. Possible application of the developed method to different models of motion and observation is considered. Analytical relations are derived for estimating accuracy characteristics and methodological errors. A comparative evaluation of computational cost is presented.

Keywords: single-position coordinate determination, indirect method, bearing, radial velocity, invariant, first integral, unbiasedness condition, invariance condition, optimality criterion, Lagrange's multiplier method, posterior variance of estimation error

DOI: 10.25728/arcRAS.2023.20.26.001

1. INTRODUCTION

Nowadays, the issues of single-position (active and passive) coordinate determination are still topical in a wide range of location and navigation problems. As a rule, these methods are implemented based on direct and indirect measurements of bearing, phase differences, Doppler frequencies, relative signal powers, and their derivatives. Also, additional information from various illumination sources, reflectors (of natural and artificial origin), external control systems, as well as a priori data on the structure and some parameters of the emitted signal, object speed, the initial or final point of its route, the presence of barrage and maneuvering areas, etc. are used. (For example, we refer to the publications [1–33].)

The solution of single-position coordinate determination problems under various types of interference fits well into the optimal Kalman estimation scheme in a stochastic formulation (as a rule, with state space expansion) with direct and pseudo measurements [9, 10, 17–26, 30, 36]. In practice, however, rather simple suboptimal indirect coordinate determination methods with smoothed measurements are often used for a wide class of problems (e.g., those related to express and post-processing of trajectory and telemetry data in range command and measurement complexes, real-time tracking of maneuvering objects, etc. [2, 7, 8, 11]). These methods are based on simple deterministic motion models (linear, piecewise linear, polynomial, piecewise polynomial, differential, piecewise differential, group, piecewise group, and many others), known analytical relationships between the estimated and measured parameters, and simple procedures for smoothing observations based on the least squares method (LSM) and its various modifications. Being inferior to optimal (linear and nonlinear) filtering methods in terms of potential accuracy, they are easy to implement

in practice in real time under high-quality smoothed measurements. Furthermore, their numerical implementation causes no problems related to transients, convergence, and strict requirements for the volume and quality of initial a priori information (which is often characteristic of optimal methods, e.g., when considering the effects of "smearing accuracy" or "rigidity" [9, 32]). For instance, the complexes mentioned above traditionally involve multistage information processing; indirect methods are applied therein at the stage of express processing whereas optimal methods are usually implemented at the stage of post-processing.

Some indirect single-position coordinate determination methods do not use bearing but operate with periodic pulse radio signals and are oriented to measuring the continuous frequency bias of the received signal at the observation point due to the movement of either the radiation source or the observer [3, 7]. The fundamental disadvantage of these methods is the necessity to consider a priori information about the speed of the object (or that of the source, or that of the observer), which is often unacceptable in practice. In addition, the coordinate determination problem is limited to finding the range and heading angle within the uniform rectilinear motion model; hence, it is impossible to estimate all object location parameters for an arbitrary time instant. An attempt to eliminate the speed-related limitation was undertaken in [27]; but in this case, it is required to track the evolution of the Doppler frequency considering the continuous accumulation (counting) of pulses of the received signal at the observation point. Obviously, the matter concerns only high-speed objects and severe constraints on observation conditions, and the uniform rectilinear motion model is also used. The general drawback of the indirect methods discussed in [3, 7, 27] is the technical complexity of their practical implementation.

There are goniometric Doppler methods for the single-position determination of motion parameters (e.g., see [3, 4, 7, 9, 29]) with direct (radial velocity and bearing) and indirect (the derivatives of different orders) measurements without the a priori information mentioned above. These methods are focused on the simplest motion models (e.g., orbital) and neglect the possibility of constructing several independent coordinate determination channels and the appearance of singular primary measurement errors that devalue the information contained in indirect measurements (the derivatives of radial velocity and bearing).

Note the method [29], which operates with derivatives up to the second order inclusive and forms adaptive coordinate determination algorithms based on several parallel algorithms corresponding to the invariants of object motion. However, according to the analysis, the explicit-form relations and the corresponding algorithms obtained in [29] are dependent and redundant; they are also focused on the uniform rectilinear motion model only.

This paper develops an indirect single-position coordinate determination method invariant with respect to singular errors of a given class. (Such errors are represented as an appropriate linear combination with unknown spectral coefficients in a given finite-dimensional functional space.) Based on a complete set of invariants (for a wide class of motion models), the method forms a family of independent quasi-optimal solutions and the resulting estimate of object motion parameters using these solutions. The comparative computational gain is demonstrated.

According to [8, 11, 33, 37], invariants can effectively serve to solve a whole class of applied target problems of single- and multi-position location and navigation based on indirect methods. Here, we show the possibility of decentralization, parallelization, and reduction of computational cost in processing measurements in various-type systems based on invariants of continuous groups of Lie transformations (CGLT) and first integrals used to describe the motion of various objects.

2. PROBLEM STATEMENT

Consider an object whose motion in a separate observation area is described in the Cartesian rectangular frame by some operator equation (e.g., in the vector-algebraic or vector-differential

form)

$$\mathbf{G}\,(t, \boldsymbol{\rho}, \boldsymbol{\eta}) = 0 \quad \forall\, t \in [0, T]\,, \tag{2.1}$$

where $\boldsymbol{\rho} = [x, y, z]^{\mathrm{T}}$ ($x = x(t)$, $y = y(t)$, $z = z(t)$) denotes the object's coordinate vector and $\boldsymbol{\eta}$ is the vector of unknown real-valued parameters.

Assume that the coordinates x, y, z are smooth and differential functions (a required number of times) and the vector $\boldsymbol{\rho}$ is assigned the vector of spherical coordinates $\boldsymbol{\varsigma} = [r, \lambda, \varphi]^{\mathrm{T}}$, where r, λ, and φ are inclined range, longitude, and latitude, respectively. Let $\mathbf{X} = \left[r^{(1)}, \lambda, \varphi\right]^{\mathrm{T}}$ be the vector of direct measurements, where $r^{(1)} = dr/dt$), and let $\mathbf{Y}$ be the vector of indirect measurements, whose coordinates are the derivatives of $r^{(1)}, \lambda, \varphi$ of different orders, necessary to implement a version of the method developed below. We choose a grid (sliding window, further termed the window for simplicity) $\{t_{n+i}, i = \overline{-m, m}\}$, where $n \geq m$, $m \in \{1, 2, \ldots\}$, $t_{n+i} \in [0, T]$, and $2m + 1$ is the window size. Introducing the notation $\mu \in \left\{r^{(1)}, \lambda, \varphi\right\}$, we adopt the additive observation equation

$$\mathbf{H}_\mu = \boldsymbol{\mu} + \mathbf{s}_\mu + \boldsymbol{\xi}_\mu, \tag{2.2}$$

where

$$\boldsymbol{\mu} = [\mu_{n+i}, i = \overline{-m, m}]^{\mathrm{T}}, \quad \mathbf{s}_\mu = [s_{\mu,n+i}, i = \overline{-m, m}]^{\mathrm{T}},$$
$$\boldsymbol{\xi}_\mu = [\xi_{\mu,n+i}, i = \overline{-m, m}]^{\mathrm{T}}, \quad \mu_{n+i} = \mu\,(t_{n+i}),$$
$$s_{\mu,n+i} = s_\mu\,(t_{n+i}), \quad \xi_{\mu,n+i} = \xi_\mu\,(t_{n+i}).$$

In (2.2), $s_\mu(t)$ means the singular error

$$s_\mu(t) = \mathbf{D}_\mu^{\mathrm{T}} \boldsymbol{\Theta}_\mu(t), \tag{2.3}$$

where
$\mathbf{D}_\mu = \left[d_{\mu k}, k = \overline{0, K}\right]^{\mathrm{T}}$ is the vector of unknown spectral coefficients and
$\boldsymbol{\Theta}_\mu(t) = \left[\theta_{\mu k}(t), k = \overline{0, K}\right]^{\mathrm{T}}$ is the vector of given basis functions.

The function $\mu = \mu(t)$ has the spectral representation

$$\mu(t) = \mathbf{A}_\mu^{\mathrm{T}} \boldsymbol{\Psi}_\mu(t), \tag{2.4}$$

where
$\mathbf{A}_\mu = \left[a_{\mu b}, b = \overline{0, B}\right]^{\mathrm{T}}$ is the vector of unknown coefficients and
$\boldsymbol{\Psi}_\mu(t) = \left[\psi_{\mu b}(t), b = \overline{0, B}\right]^{\mathrm{T}}$ is the vector of given basis functions.

The vector $\boldsymbol{\xi}_\mu$ consists of random errors with zero and the correlation matrix $\mathbf{K}_\mu = [k_{\mu,n+i,n+j}, i, j = \overline{-m, m}]$.

Models (2.1)–(2.4) are widely used in various localization and navigation problems. Complex trajectories (e.g., those of maneuvering objects) can be described by applying a separate model (2.1) for each observation area. In particular, a very promising approach is to describe such trajectories via the simplest groups of Lie transformations (e.g., shift, rotation, and stretching [8, 11, 31, 33–37]).

Based on the set of invariants of equation (2.1) (in particular, the first integrals of motion or the invariants of CGLT), it is required to develop an indirect coordinate determination method considering (2.2)–(2.4) and the accepted constraints with the following features: the method involves no state space expansion; the method is robust to the singular error; the method allows estimating the object's motion parameters from the extended vector of direct and indirect measurements $\mathbf{Z} = \left[\mathbf{X}^{\mathrm{T}}, \mathbf{Y}^{\mathrm{T}}\right]^{\mathrm{T}}$, whose coordinates are estimated with minimum posterior variances.

3. THE PRINCIPLE OF DETERMINING MOTION PARAMETERS BASED ON INVARIANTS

Let us associate with equation (2.1) a scalar invariant $I = I(t, \boldsymbol{\rho}, \boldsymbol{\gamma}_I)$, where the vector $\boldsymbol{\gamma}_I$ consists of some derivatives of the coordinates of the vector $\boldsymbol{\rho}$. On the solutions $\boldsymbol{\rho}(t)$ and $\boldsymbol{\gamma}_I(t)$ of equation (2.1), this invariant satisfies the condition

$$I(t, \boldsymbol{\rho}(t), \boldsymbol{\gamma}_I(t)) = C = \text{const} \quad \forall\, t \in [0, T]\,. \tag{3.1}$$

Passing to spherical coordinates in (3.1) gives

$$Q\left(t, \boldsymbol{\varsigma}(t), \boldsymbol{\gamma}_Q(t)\right) = C = \text{const} \quad \forall\, t \in [0, T]\,, \tag{3.2}$$

where the vector $\boldsymbol{\gamma}_Q$ consists of some derivatives of the coordinates of the vector $\boldsymbol{\varsigma}$.

The way to find the invariants is entirely determined by the kind of equation (2.1).

We calculate the total derivative of the left- and right-hand sides of equation (3.2):

$$\frac{\partial Q}{\partial t} + \frac{\partial Q}{\partial \boldsymbol{\varsigma}}\left(\frac{d\boldsymbol{\varsigma}}{dt}\right)^{\mathrm{T}} + \frac{\partial Q}{\partial \boldsymbol{\gamma}_Q}\left(\frac{d\boldsymbol{\gamma}_Q}{dt}\right)^{\mathrm{T}} = 0 \quad \forall\, t \in [0, T]\,. \tag{3.3}$$

Expanding all derivatives in (3.3) yields the equation

$$W(t, \boldsymbol{\varsigma}, \boldsymbol{\gamma}_W) = 0 \quad \forall\, t \in [0, T]\,, \tag{3.4}$$

where the vector $\boldsymbol{\gamma}_W$ consists of all possible derivatives of r, λ, φ.

Solving this equation for r, we determine the inclined range (distance to the object):

$$r = W^{-1}(t, \mathbf{Z})\,. \tag{3.5}$$

Associating with equation (2.1) the set of independent invariants $I_l = I_l(t, \boldsymbol{\rho}, \boldsymbol{\gamma}_I)$, $l = \overline{1, L}$, by analogy with (3.1)–(3.5), we obtain the set of formulas

$$r[l] = W_{[l]}^{-1}\left(t, \mathbf{Z}_{[l]}\right), \quad l = \overline{1, L}. \tag{3.6}$$

This set can be used in an adaptive version of the inclined range estimation procedure in order to improve the accuracy of estimation considering measurement errors. For example, if for a fixed time instant t, the vector $\mathbf{Z}_{[l]}$ is estimated with an error characterized by zero mean the correlation matrix $\mathbf{K}_{\mathbf{Z}}$, then the variance of the inclined range estimate is given by

$$\sigma_{r[l]}^2 = \mathbf{H}_{[l]}^{\mathrm{T}} \mathbf{K}_{\mathbf{Z}[l]} \mathbf{H}_{[l]}, \quad l = \overline{1, L}, \tag{3.7}$$

where the column vector $\mathbf{H}_{[l]}$ consists of the partial derivatives of (3.6) with respect to the elements of the vector $\mathbf{Z}_{[l]}$ calculated on their mathematical expectations.

As an optimal version of the inclined range estimation procedure we select the one for which

$$l^* = \arg\min_l \sigma_{r[l]}^2, \quad l^* \in \{1, 2, \ldots, L\}\,. \tag{3.8}$$

The Cartesian coordinates of the object can be determined using the dependencies

$$x[l^*] = r[l^*] \cos\varphi \cos\lambda, \quad y[l^*] = r[l^*] \cos\varphi \sin\lambda, \quad z[l^*] = r[l^*] \sin\varphi, \tag{3.9}$$

where the angular coordinates λ and φ are replaced by either direct measurements or their smoothed values.

Generally, we can use the set of probable models $\mathbf{G}_k(t,\boldsymbol{\rho},\boldsymbol{\eta})=0$, $k=\overline{0,K}$, for a given observation area instead of (2.1). In this case, the algorithm (3.7)–(3.9) takes the form

$$\begin{cases} \sigma^2_{r[k,l]}=\mathbf{H}^{\mathrm{T}}_{[k,l]}\mathbf{K}_{\mathbf{Z}[k,l]}\mathbf{H}_{[k,l]}, \quad k=\overline{1,K}, \quad l=\overline{1,L_k}, \\ [k^*,l^*]=\arg\min\limits_{[k,l]}\sigma^2_{r[k,l]}, \quad k^*\in\{1,2,\ldots,K\}, \quad l^*\in\{1,2,\ldots,L_k\}, \\ x[k^*,l^*]=r[k^*,l^*]\cos\varphi\cos\lambda, \\ y[k^*,l^*]=r[k^*,l^*]\cos\varphi\sin\lambda, \\ z[k^*,l^*]=r[k^*,l^*]\sin\varphi. \end{cases} \tag{3.10}$$

The algorithm (3.10) parallelizes the computational process considering the number of invariants used and adapts the estimation procedure of the object motion parameters to the observation conditions.

4. DESIGN AND USE OF INVARIANTS: SOME EXAMPLES

Consider a separate observation area in which the following general CGLT [32–34] corresponds to equation (2.1):

$$T_a: \quad \boldsymbol{\rho}'=\mathbf{f}(a,\boldsymbol{\rho}_0,\boldsymbol{\eta}_0) \quad \forall\, a\in\Delta_a\subset R^1, \tag{4.1}$$

where $\boldsymbol{\rho}'=[x',y',z']^{\mathrm{T}}$, $\mathbf{f}(a,\boldsymbol{\rho}_0,\boldsymbol{\eta}_0)=[f_x,f_y,f_z]^{\mathrm{T}}$, $\boldsymbol{\eta}_0$ is the vector of numerical parameters of the group and a is a real-valued group parameter such that $\mathbf{f}(a_0,\boldsymbol{\rho},\boldsymbol{\eta}_0)=\boldsymbol{\rho}$ for $a=a_0$, $a_0\in\Delta_a$.

Model (4.1) describes the object's trajectory; when treating the group parameter as a time-varying function $a=a(t,\boldsymbol{\chi}_0)$, where $\boldsymbol{\chi}_0$ is the vector of generally unknown numerical parameters, we can describe the time law of motion along this trajectory. With the change of coordinates $\boldsymbol{\rho}'=\boldsymbol{\rho}$, $\boldsymbol{\rho}=\boldsymbol{\rho}_0$, due to (2.1), it follows that $\boldsymbol{\rho}-\mathbf{f}(a,\boldsymbol{\rho}_0,\boldsymbol{\eta}_0)=\mathbf{G}(t,\boldsymbol{\rho},\boldsymbol{\eta})$, $\boldsymbol{\eta}=[\boldsymbol{\rho}_0,\boldsymbol{\eta}_0,\boldsymbol{\chi}_0]^{\mathrm{T}}$.

The invariants $I=I(\boldsymbol{\rho},\boldsymbol{\eta}_0)$ of model (4.1), independent of the parameters t, $\boldsymbol{\rho}_0$, and $\boldsymbol{\chi}_0$, are found by solving the linear partial differential equation

$$XI(\boldsymbol{\rho},\boldsymbol{\eta}_0)=\phi_x\frac{\partial I}{\partial x}+\phi_y\frac{\partial I}{\partial y}+\phi_z\frac{\partial I}{\partial z}=0, \tag{4.2}$$

where $X=\phi_x\partial/\partial x+\phi_y\partial/\partial y+\phi_z\partial/\partial z$ denotes the infinitesimal CGLT operator. Its coordinates are given by $\phi_x=\partial f_x/\partial a$, $\phi_y=\partial f_y/\partial a$, and $\phi_z=\partial f_z/\partial a$ at the point $a=a_0$.

In view of (4.2), an extended operator and the corresponding partial differential equation are constructed to find invariants considering the temporal nature of motion along the trajectory (4.1) and various derivatives of the vector $\boldsymbol{\rho}$; for details, see [8, 11, 31–34].

This method will be demonstrated on an example of a shift group. Let

$$T_{a=t}: \boldsymbol{\rho}'=\boldsymbol{\rho}+\boldsymbol{\eta}_0 t\ \forall\ a=t\in\Delta_a=[0,T]\subset R^1,$$

where $\boldsymbol{\eta}_0=\mathbf{V}_0=[V_{x0},V_{y0},V_{z0}]^{\mathrm{T}}$ is the velocity vector of an object moving straight and uniformly. In this case, we have $\phi_x=V_{x0}$, $\phi_y=V_{y0}$, $\phi_z=V_{z0}$, and two independent invariants, $I_{[1]}=xV_{y0}-yV_{x0}=xy^{(1)}-yx^{(1)}$ and $I_{[2]}=xV_{z0}zV_{x0}=xz^{(1)}-zx^{(1)}$. In addition, due to (3.1), $\gamma_{I[1]}(t)=\gamma_{I[1]}=\left[x^{(1)},y^{(1)}\right]^{\mathrm{T}}$ and $\gamma_{I[2]}(t)=\gamma_{I[2]}=\left[x^{(1)},z^{(1)}\right]^{\mathrm{T}}$. Using (3.2) and straightforward but cumbersome transformations, we obtain

$$Q_{[1]}\left(t,\boldsymbol{\varsigma}(t),\boldsymbol{\gamma}_{Q[1]}(t)\right)=r^2\lambda^{(1)}\cos^2\varphi,$$

$$Q_{[2]}\left(t,\boldsymbol{\varsigma}(t),\boldsymbol{\gamma}_{Q[2]}(t)\right)=r^2\left(\varphi^{(1)}\cos\lambda+\lambda^{(1)}\sin\lambda\sin\varphi\cos\varphi\right),$$

where

$$\boldsymbol{\gamma}_{Q[1]} = \left[\varphi^{(1)}\right] \text{ and } \boldsymbol{\gamma}_{Q[2]} = \left[\lambda^{(1)}, \varphi^{(1)}\right]^{\mathrm{T}}.$$

In view of (3.3) and (3.4), it follows that

$$W_{[1]} = 2r^{(1)}\lambda^{(1)} + r\left(\lambda^{(2)} - 2\lambda^{(1)}\varphi^{(1)}\tan\varphi\right),$$
$$W_{[2]} = 2r^{(1)}\varphi^{(1)} + r\left(\varphi^{(2)} + \left(\lambda^{(1)}\right)^2\sin\varphi\cos\varphi\right),$$

where

$$\boldsymbol{\gamma}_{W[1]} = \left[r^{(1)}, \lambda^{(1)}, \lambda^{(2)}, \varphi^{(1)}\right]^{\mathrm{T}} \text{ and } \boldsymbol{\gamma}_{W[1]} = \left[r^{(1)}, \lambda^{(1)}, \varphi^{(1)}, \varphi^{(2)}\right]^{\mathrm{T}}.$$

Finally, concretizing (3.5) and (3.6) yields two independent formulas for the inclined range:

$$r\,[1] = \frac{2r^{(1)}\lambda^{(1)}\cos\varphi}{2\lambda^{(1)}\varphi^{(1)}\sin\varphi - \lambda^{(2)}\cos\varphi}, \tag{4.3}$$

$$r\,[2] = -\frac{2r^{(1)}\varphi^{(1)}}{\varphi^{(2)} + \left(\lambda^{(1)}\right)^2\sin\varphi\cos\varphi}. \tag{4.4}$$

In the special cases $\varphi = \varphi^{(1)} = \varphi^{(2)} = 0$ and $\lambda = \lambda^{(1)} = \lambda^{(2)} = 0$, the expressions (4.3) and (4.4) directly imply the well-known ranging formulas

$$r\,[1] = -2r^{(1)}\lambda^{(1)}/\lambda^{(2)}, \quad r\,[2] = -2r^{(1)}\varphi^{(1)}/\varphi^{(2)}.$$

(For example, see the differential-geometrical method [4].)

Other ranging formulas can be derived using three new invariants, $I_{[3]} = x^{(1)} = V_{x0}$, $I_{[4]} = y^{(1)} = V_{y0}$, and $I_{[5]} = z^{(1)} = V_{z0}$. They lead to the independent ranging formulas

$$r\,[3] = \frac{r^{(2)}\cos\varphi - 2r^{(1)}\varphi^{(1)}\sin\varphi}{\varphi^{(2)}\sin\varphi + \left[\left(\lambda^{(1)}\right)^2 + \left(\varphi^{(1)}\right)^2\right]\cos\varphi}, \tag{4.5}$$

$$r\,[4] = \frac{r^{(2)}\sin\varphi + 2r^{(1)}\varphi^{(1)}\cos\varphi}{\left(\varphi^{(1)}\right)^2\sin\varphi - \varphi^{(2)}\cos\varphi}, \tag{4.6}$$

$$r\,[5] = \frac{r^{(2)}}{\left(\varphi^{(1)}\right)^2 + \left(\lambda^{(1)}\right)^2\cos^2\varphi}. \tag{4.7}$$

In contrast to [29], the set of formulas (4.3)–(4.7) is necessary and sufficient for constructing a parallel independent adaptive ranging algorithm, and the resulting relations are written in a compact (nonredundant) form.

Remark 1. There is no complete coincidence of the sets of measured parameters in all formulas (4.3)–(4.7). In view of (3.7)–(3.10), it is therefore possible to organize five independent channels for range calculation and adaptation to variable observation conditions.

Remark 2. The longitude λ is not explicitly included in any of the formulas; hence, the constant systematic errors in the measurements of the coordinate λ can be effectively dealt with. The latitude φ explicitly figures in all the formulas.

Remark 3. For more complex motion models with general CGLT, all possible invariants corresponding to the trajectory and the object's motion law along this trajectory, as well as independent expressions for determining the inclined range, can be found similar to the shift group.

Remark 4. For a maneuvering object, it is necessary to use a compound model based on an admissible set of a particular CGLT (e.g., shift, rotation, and stretching). An appropriate particular CGLT in some observation area is chosen by solving the identification problem with minimizing a decision function (e.g., the residual of the least squares method). Such an approach using the rotation group was considered in [31]; the object trajectory was approximated by pieces of circles of different radii.

If model (2.1) is some differential equation, then all possible invariants in the dynamic case can be found within the well-known theory of group analysis [34–36]. In practice, however, it often suffices to use particular invariants of motion, i.e., the so-called first integrals of the differential equation. We will demonstrate this approach on an example of circular orbital motion: $\mathbf{G}(t, \boldsymbol{\rho}, \boldsymbol{\eta}) = \boldsymbol{\rho}^{(2)} + \eta_0 R_0^{-3} \boldsymbol{\rho} = 0$, where $\boldsymbol{\eta} = [R_0, \eta_0]^{\mathrm{T}}$ and R_0 and η_0 are the radius and gravitational parameter of the Earth, respectively. As is well known, the invariants (first integrals) of this motion are $I_{[1]} = xy^{(1)} - yx^{(1)}$ and $I_{[2]} = xz^{(1)} - zx^{(1)}$, identical in form to the shift group invariants discussed above. However, the derivatives $x^{(1)}$, $y^{(1)}$, and $z^{(1)}$ here are not constants and the invariants $I_{[3]} = x^{(1)} = V_{x0}$, $I_{[4]} = y^{(1)} = V_{y0}$, and $I_{[5]} = z^{(1)} = V_{z0}$ used previously become inapplicable. With this fact in mind, we accept only the expressions (4.3) and (4.4) as ranging formulas in the dynamic case.

Remark 5. The single-position indirect method developed in this paper can be generalized to the class of stochastic models, for which the application of classical invariants is often very limited. At the same time, it is possible to use the so-called ε-invariants [37]. Within this approach, the invariance condition holds approximately (with accuracy up to ε), and the coordinate determination problem can be solved approximately as well.

5. CONSIDERATION OF FLUCTUATING MEASUREMENT ERRORS

We take an example of the shift group and the condition $\varphi = \varphi^{(1)} = \varphi^{(2)} = 0$ to demonstrate the implementation of the algorithm (3.7), (3.8). Clearly, in this particular case, the entire set of formulas (4.3)–(4.7) reduces to the two informative ones:

$$r[1] = -2r^{(1)}\lambda^{(1)}/\lambda^{(2)}, \quad r[2] = -r^{(2)}/\left(\lambda^{(1)}\right)^2.$$

Accordingly, we have two vectors of measured parameters: $\mathbf{Z}_{[1]} = \left[r^{(1)}, \lambda^{(1)}, \lambda^{(2)}\right]^{\mathrm{T}}$ and $\mathbf{Z}_{[2]} = \left[r^{(2)}, \lambda^{(1)}\right]^{\mathrm{T}}$. Let the matrices $\mathbf{K}_{\mathbf{Z}[1]}$ and $\mathbf{K}_{\mathbf{Z}[2]}$ be diagonal, i.e.,

$$\mathbf{K}_{\mathbf{Z}[1]} = \operatorname{diag}\left[\sigma^2_{r^{(1)}}, \sigma^2_{\lambda^{(1)}}, \sigma^2_{\lambda^{(2)}}\right] \text{ and } \mathbf{K}_{\mathbf{Z}[2]} = \operatorname{diag}\left[\sigma^2_{r(2)}, \sigma^2_{\lambda(1)}\right].$$

(With this supposition, the presentation below will be less cumbersome.) Due to $x = r\cos\lambda$, $y = r\sin\lambda$, and (3.7), we find

$$\sigma^2_{r[1]} = 4\left(\lambda^{(2)}\right)^{-2}\left\{\left(\lambda^{(1)}\right)^2\sigma^2_{r^{(1)}} + \left(r^{(1)}\right)^2\left[\sigma^2_{\lambda^{(1)}} + \left(\lambda^{(1)}/\lambda^{(2)}\right)^2\sigma^2_{\lambda^{(2)}}\right]\right\}, \tag{5.1}$$

$$\sigma^2_{r[2]} = \left(\lambda^{(1)}\right)^{-4}\left[\sigma^2_{r^{(2)}} + 4\left(r^{(2)}\right)^2\left(\lambda^{(1)}\right)^{-4}\sigma^2_{\lambda^{(1)}}\right]. \tag{5.2}$$

The priority is given to the ranging formula for which

$$l^* = \arg\min_l \sigma^2_{r[l]}, \quad l^* \in \{1, 2\}. \tag{5.3}$$

According to (5.1)–(5.3), the method involves derivatives up to the second order inclusive and can be effectively applied only on smoothed measurements. In addition, the class of high-speed objects is considered: the necessary increment of angular coordinates and radial velocity on a given observation interval must be provided [29].

6. AN AUTO-COMPENSATION ALGORITHM FOR SMOOTHING PRIMARY MEASUREMENTS

In view of (2.1)–(2.4), we consider an auto-compensation unbiased smoothing algorithm for the parameter $\mu \in \left\{r^{(1)}, \lambda, \varphi\right\}$ and its derivatives $\mu^{(q)}$, $q \in \{0,1,2\}$, at a point t_n using the window $\{t_{n+i}, i = \overline{-m,m}\}$. Let us rest on the general approach to estimating the values of linear functionals; for details, see [38, 39].

Within this approach, the estimate $\mu^{(q)*}$ of $\mu^{(q)}$ has the form

$$\mu^{(q)*} = \mathbf{P}_{\mu q}^{\mathrm{T}} \mathbf{H}_{\mu}, \tag{6.1}$$

where $\mathbf{P}_{\mu q} = [p_{\mu q, n+i}, i = \overline{-m,m}]^{\mathrm{T}}$ is the vector of unknown weight coefficients assigned by minimizing the variance $\sigma_{\mu q}^2$ of the estimate $\mu^{(q)*}$.

This estimate belongs to the linear class; therefore,

$$\sigma_{\mu q}^2 = \mathbf{P}_{\mu q}^{\mathrm{T}} \mathbf{K}_{\mu} \mathbf{P}_{\mu q}. \tag{6.2}$$

Furthermore, we require the unbiasedness conditions of the estimate ($\mu^{(q)} - \mathbf{P}_{\mu q}^{\mathrm{T}} \boldsymbol{\mu} = 0$) and its invariance with respect to the singular error ($\mathbf{P}_{\mu q}^{\mathrm{T}} \mathbf{s}_{\mu} = 0$). The constrained optimization problem is solved using Lagrange's multiplier method with the decision function

$$J\left(\mathbf{P}_{\mu q}, \boldsymbol{\zeta}_{\mu q}, \boldsymbol{\omega}_{\mu q}\right) = \mathbf{P}_{\mu q}^{\mathrm{T}} \mathbf{K}_{\mu} \mathbf{P}_{\mu q} + \boldsymbol{\zeta}_{\mu q}^{\mathrm{T}} \Theta_{\mu}^{\mathrm{T}} \mathbf{P}_{\mu q} + \left[\left(\boldsymbol{\Psi}_{\mu}^{\mathrm{T}}\right)^{(q)} - \mathbf{P}_{\mu q}^{\mathrm{T}} \boldsymbol{\Psi}_{\mu}\right] \boldsymbol{\omega}_{\mu q}, \tag{6.3}$$

where $\boldsymbol{\zeta}_{\mu q}$ and $\boldsymbol{\omega}_{\mu q}$ are the column vectors of the Lagrange multipliers, $\boldsymbol{\Theta}_{\mu} = [\theta_{\mu k}(t_{n+i}), i = \overline{-m,m}, k = \overline{0,K}]$ is the basis matrix of the singular error, and $\boldsymbol{\Psi}_{\mu} = \left[\psi_{\mu b}\left(t_{n+i}\right),\ i = \overline{-m,m},\ b = \overline{0,B}\right]$ is the basis matrix of the parameter $\mu = \mu(t)$.

The vector $\mathbf{P}_{\mu q}$ minimizing $\sigma_{\mu q}^2$ subject to the unbiasedness and invariance conditions has the form

$$\mathbf{P}_{\mu q} = \boldsymbol{\Lambda}_{\mu} \mathbf{K}_{\mu}^{-1} \boldsymbol{\Psi}_{\mu} \left(\boldsymbol{\Psi}_{\mu}^{\mathrm{T}} \boldsymbol{\Lambda}_{\mu} \mathbf{K}_{\mu}^{-1} \boldsymbol{\Psi}_{\mu}\right)^{-1} \boldsymbol{\Psi}_{\mu n}^{(q)}, \tag{6.4}$$

where $\boldsymbol{\Lambda}_{\mu} = E_{2m+1} - \mathbf{K}_{\mu}^{-1} \Theta_{\mu} \left(\Theta_{\mu}^{\mathrm{T}} \mathbf{K}_{\mu}^{-1} \Theta_{\mu}\right)^{-1} \Theta_{\mu}^{\mathrm{T}}$, E_{2m+1} is an identity matrix of dimensions $(2m+1) \times (2m+1)$, and $\boldsymbol{\Psi}_{\mu n}^{(q)} = d^q \boldsymbol{\Psi}_{\mu}(t)/dt^q|_{t=t_n}$.

The variance of the estimate $\mu^{(q)*}$ is given by

$$\sigma_{\mu q}^2 = \left(\boldsymbol{\Psi}_{\mu n}^{(q)}\right)^{\mathrm{T}} \left[\left(\mathbf{K}_{\mu}^{-1} \boldsymbol{\Psi}_{\mu}\right)^{\mathrm{T}} \left(\boldsymbol{\Lambda}_{\mu}\right)^{\mathrm{T}} \boldsymbol{\Psi}_{\mu}\right]^{-1} \mathbf{H}_{\mu} \left(\boldsymbol{\Psi}_{\mu}^{\mathrm{T}} \boldsymbol{\Lambda}_{\mu} \mathbf{K}_{\mu}^{-1} \boldsymbol{\Psi}_{\mu}\right)^{-1} \boldsymbol{\Psi}_{\mu n}^{(q)}, \tag{6.5}$$

where

$$\mathbf{H}_{\mu} = \left(\mathbf{K}_{\mu}^{-1} \boldsymbol{\Psi}_{\mu}\right)^{\mathrm{T}} \left(\boldsymbol{\Lambda}_{\mu}\right)^{\mathrm{T}} \mathbf{K}_{\mu} \boldsymbol{\Lambda}_{\mu} \mathbf{K}_{\mu}^{-1} \boldsymbol{\Psi}_{\mu}.$$

Clearly, the methodological error due to neglecting the tail of the series (2.4) has the mathematical expectation

$$\varepsilon_{\mu q} = \Delta_{\mu n}^{(q)} - \mathbf{P}_{\mu q}^{\mathrm{T}} \boldsymbol{\Delta}_{\mu n}, \tag{6.6}$$

where $\Delta_\mu = \Delta_\mu(t)$ is the series tail and $\Delta_{\mu n}^{(q)}$ denotes its qth derivative at the point $t = t_n$, $\boldsymbol{\Delta}_{\mu n} = [\Delta_\mu(t_{n+i}),\ i = \overline{-m, m}]^{\mathrm{T}}$.

According to [38, p. 62], with increasing the number of spectral coefficients in the singular error model (2.3), the algorithm (6.1)–(6.6) reduces computational cost by 47% compared to the traditional extended least-squares method. As a result, the smoothing problem is solved faster.

Considering (6.1)–(6.6), we can construct the desired estimates of the object motion parameters invariant to singular measurement errors. For example, formulas (4.3) and (4.4) yield the following robust estimates of the inclined range for two invariants:

$$r[1] = \frac{2\left(\mathbf{P}_{r1}^{\mathrm{T}}\mathbf{H}_r\right)\left(\mathbf{P}_{\lambda 1}^{\mathrm{T}}\mathbf{H}_\lambda\right)\cos\left(\mathbf{P}_{\varphi 0}^{\mathrm{T}}\mathbf{H}_\varphi\right)}{2\left(\mathbf{P}_{\lambda 1}^{\mathrm{T}}\mathbf{H}_\lambda\right)\left(\mathbf{P}_{\varphi 1}^{\mathrm{T}}\mathbf{H}_\varphi\right)\sin\left(\mathbf{P}_{\varphi 0}^{\mathrm{T}}\mathbf{H}_\varphi\right) - \left(\mathbf{P}_{\lambda 2}^{\mathrm{T}}\mathbf{H}_\lambda\right)\cos\left(P_{\varphi 0}^{\mathrm{T}}\mathbf{H}_\varphi\right)}, \tag{6.7}$$

$$r[2] = -\frac{2\left(\mathbf{P}_{r1}^{\mathrm{T}}\mathbf{H}_r\right)\left(\mathbf{P}_{\varphi 1}^{\mathrm{T}}\mathbf{H}_\varphi\right)}{\left(P_{\varphi 2}^{\mathrm{T}}\mathbf{H}_\varphi\right) + \left(\mathbf{P}_{\lambda 1}^{\mathrm{T}}\mathbf{H}_\lambda\right)^2\sin\left(\mathbf{P}_{\varphi 0}^{\mathrm{T}}\mathbf{H}_\varphi\right)\cos\left(\mathbf{P}_{\varphi 0}^{\mathrm{T}}\mathbf{H}_\varphi\right)}. \tag{6.8}$$

The ranges for the variants (4.5)–(4.7) and the Cartesian coordinates (3.9) of the observed object are determined by analogy with (6.7) and (6.8).

According to the results of computational experiments [38, 39], the auto-compensation smoothing algorithm demonstrates high effectiveness in anomalous measurement conditions. Hence, it is possible to form stable estimates of the derivatives of the radial velocity and angular coordinates necessary for the successful application of the single-position indirect coordinate determination method. Simulation results for the adaptive algorithm (3.7), (3.8) in the case of rectilinear uniform object motion were presented in [29]. They show that the method is applicable to high-precision measurements, while the reliability of coordinate determination significantly depends on the object's dynamics and observation conditions.

7. CONCLUSIONS

The method developed above considerably expands the scope of quasi-optimal indirect fast estimation methods robust to singular measurement errors and observation conditions of high-speed objects for their single-position coordinate determination. This method can be effectively used as a tool for intelligent and analytical improvement of the existing and next-generation single-position systems of active and passive location and navigation, independently or in combination with traditional statistical methods (e.g., least squares, maximum likelihood, maximum posterior probability density, and dynamic filtering).

The method has limitations on the classes of single-position systems in terms of measurement accuracy, observation conditions, and the types of objects to be tracked.

REFERENCES

1. *Osnovy manevrirovaniya korablei* (Fundamentals of Ship Maneuvering), Skvortsov, M., Ed., Moscow: Voenizdat, 1966.

2. Brandin, V.N. and Razorenov, G.N., *Opredelenie traektorii kosmicheskikh apparatov* (Determination of Spacecraft trajectories), Moscow: Mashinostroenie, 1978.

3. Shebshaevich, V.S., *Vvedenie v teoriyu kosmicheskoi navigatsii* (Introduction to the Theory of Space Navigation), Moscow: Sovetskoe Radio, 1971.

4. Gromov, G.N., *Differentsial'no-geometricheskii metod navigatsii* (The Differential-Geometric Method of Navigation), Moscow: Radio i Svyaz', 1986.

5. Khvoshch, V.A., *Taktika podvodnykh lodok* (Tactics of Submarines), Moscow: Voenizdat, 1989.

6. Solov'ev, Yu.A., *Sputnikovaya navigatsiya i ee prilozheniya* (Satellite Navigation and Its Applications), Moscow: Ekotrends, 2003.

7. Mel'nikov, Yu.P. and Popov, S.V., *Radiotekhnicheskaya razvedka* (Radio Intelligence), Moscow: Radiotekhnika, 2008.

8. Bulychev, Yu.G. and Manin, A.P., *Matematicheskie aspekty opredeleniya dvizheniya letatel'nykh apparatov* (Mathematical Aspects of Determining the Motion of Aircraft), Moscow: Mashinostroenie, 2000.

9. Yarlykov, M.S., *Statisticheskaya teoriya radionavigatsii* (Statistical Theory of Radio Navigation), Moscow: Radio i Svyaz', 1985.

10. Sosulin, Yu.G., Kostrov, V.V., and Parshin, Yu.N., *Otsenochno-korrelyatsionnaya obrabotka signalov i kompensatsiya pomekh* (Evaluation and Correlation Signal Processing and Interference Compensation), Moscow: Radiotekhnika, 2014.

11. Bulychev, Yu.G., Vasil'ev, V.V., Dzhugan, R.V., et al., *Informatsionno-izmeritel'noe obespechenie naturnykh ispytanii slozhnykh tekhnicheskikh kompleksov* (Information and Measurement Support for Live Testing of Complex Technical Systems), Manin, A.P. and Vasil'ev, V.V., Eds., Moscow: Mashinostroenie-Polet, 2016.

12. Gel'tser, A.A, A Single-Position Method for Determining the Location of a Radio Emission Source Using Signal Reflections from a Variety of Relief Elements and Local Objects, *Extended Abstract of Cand. Sci. (Eng.) Dissertation*, Tomsk State University of Control Systems and Radioelectronics, 2012.

13. Sirenko, I.L., Donets, I.V., Reisenkind, Ya.A., et al., Single-Position Determination of Coordinates and Velocity Vector of Radio-Emitting Objects, *Radiotekhnika*, 2019, no. 10 (16), pp. 28–32.

14. Bulychev, Yu.G., Bulychev, V.Yu., Ivakina, S.S., and Nicholas, P.I, Estimation of the Inclined Range to the Target with the Polynomial Law of Motion, *Vestn. Kazan. Gos. Univ.*, 2013, no. 1, pp. 67–74.

15. Bulychev, V.Y., Bulychev, Y.G., Ivakina, S.S., et al., Estimation of Parameters of Object Motion Based on Stationary Quasi-autonomous Direction Finder, *J. Comput. Syst. Sci. Int.*, 2013, vol. 52, no. 5, pp. 811–818.

16. Bulychev, Yu.G., Bulychev, V.Yu., Ivakina, S.S., and Nasenkov, I.G., Passive Location of a Group of Moving Targets with One Stationary Bearing with Prior Information, *Autom. Remote Control*, 2017, vol. 78, no. 1, pp. 125–137.

17. Lin, X., Kirubarajan, T., Bar-Shalom, Y., and Maskell, S., Comparison of EKF, Pseudomeasurement and Particle Filters for a Bearing-only Target Tracking Problem, *Proc. SPIE-Int. Soc. Optic. Eng.*, 2002, vol. 4728, pp. 240–250.

18. Miller, B.M., Stepanyan, K.V., Miller, A.B., Andreev, K.V., and Khoroshenkikh, S.N., Optimal Filter Selection for UAV Trajectory Control Problems, *Proc. 37th Conference on Inform. Techn. Syst.*, September 1–6, 2013, Kaliningrad, pp. 327–333.

19. Miller, B.M. and Miller, A.B., Tracking of the UAV Trajectory on the Basis of Bearing-only Observations, *Sensors*, 2015, no. 15 (12), pp. 29802–29820. https://doi.org/10.3390/s151229768

20. Amelin, K.S. and Miller, A.B., An Algorithm for Refinement of the Position of a Light UAV on the Basis of Kalman Filtering of Bearing Measurements, *Commun. Technol. Electron.*, 2014, vol. 59, no. 6, pp. 622–631.

21. Karpenko, S., Konovalenko, I., Miller, A., Miller, B., and Nikolaev, D., UAV Control on the Basis of 3D Landmark Bearing-Only Observations, *Sensors*, 2015, no. 15 (12), pp. 29802–29820. https://doi.org/10.3390/s151229768

22. Karpenko, S., Konovalenko, I., Miller, A., Miller, B., and Nikolaev, D., Visual Navigation of the UAVs on the Basis of 3D Natural Landmarks, *Proc. SPIE. 8th Int. Conf. Machine Vision (ICMV 2015)*, 2015, vol. 9875, pp. 1–10. https://doi.org/10.111712.2228793

23. Bar-Shalom, Ya., Willet, P.K., and Tian, X., *Tracking and Data Fusion: A Handbook of Algorithms*, YBS Publishing, 2011. ISBN-13: 978-0964831278.

24. Ried, D., An Algorithm for Tracking Multiple Targets, *IEEE Transact. Autom. Control*, 1979, vol. 24, no. 6, pp. 843–854.

25. Nardone, S.C. and Aidala, V.J., Observability Criteria for Bearings-Only Target Motion Analysis, *IEEE Transact. Aerospac. Electron. Syst.*, 1981, vol. AES-17, pp. 162–166.

26. Murty, K.G., An Algorithm for Ranking All the Assignments in Order of Increasing Cost, *Oper. Res.*, 1968, vol. 16, no. 3, pp. 682–687.

27. Bulychev, Y.G. and Mozol', A.A., Single-Position Passive Location and Navigation with Allowance for the Evolution of the Radio Signal Period at the Reception Point, *J. Commun. Technol. Electron.*, 2021, vol. 66, no. 5, pp. 591–598.

28. Dyatlov, A.P. and Dyatlov, P.A., Doppler Detectors of Moving Objects Using an "Extraneous" Radiation Source, *Spets. Tekhnika*, 2010, no. 5, pp. 16–22.

29. Bulychev, Yu.G., Korotun, A.A., and Manin, A.P., Filtering of Trajectory Parameters in the Angular-Doppler Location Systems, *Radiotekhnika*, 1990, no. 12, pp. 22–26.

30. Zhdanyuk, B.F., *Osnovy statisticheskoi obrabotki traektornykh izmerenii* (Foundations of Statistical Processing of Trajectory Measurements), Moscow: Sovetskoe Radio, 1978.

31. Shloma, A.M., Frolov, S.M., and Preobrazhenskii, L.A., Adaptive Parametric Filtering of Curvilinear Trajectories, *Izv. VUZ. Radioelectronics*, 1986, no. 12, pp. 56–60.

32. Bulychev, Yu.G. and Eliseev, A.V., Rigidity Problems of Equations of Approximate Nonlinear Filtering, *Autom. Remote Control*, 1999, vol. 60, no. 1, pp. 35–45.

33. Bulychev, V.Y., Bulychev, Y.G., Ivakina, S.S., et al., Classification of Passive Location Invariants and Their Use, *J. Comput. Syst. Sci. Int.*, 2015, vol. 54, no. 6, pp. 905–915.

34. Miller, W., *Symmetry and Separation of Variables*, Reading, Massachusetts: Addison-Wesley, 1977.

35. Ovsyannikov, L.V., *Gruppovoi analiz differentsial'nykh uravnenii* (Group Analysis of Differential Equations), Moscow: Nauka, 1978.

36. Pavlovskii, Yu.N., Aggregation, Decomposition, Group Properties, and Decomposition Structures of Dynamic Systems, *Kibern. Vychisl. Tekhnika*, 1978, no. 39, pp. 53–62.

37. Bulychev, Y.G. and Burlai, I.V., A System Approach to Modeling Stochastic Objects with Invariants, *Autom. Remote Control*, 2001, vol. 62, no. 12, pp. 1939–1946.

38. Bulychev, Yu.G. and Eliseev, A.V., Computational Scheme for Invariantly Unbiased Estimation of Linear Operators in a Given Class, *Comput. Math. Math. Phys.*, 2008, vol. 48, no. 4, pp. 549–560.

39. Bulychev, Y.G., Application of Supporting Integral Curves and Generalized Invariant Unbiased Estimation for the Study of a Multidimensional Dynamical System, *Comput. Math. and Math. Phys.*, 2020, vol. 60, no. 7, pp. 1116–1133.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 9, pp. 1115–1127.

On the Algorithm of Cargoes Transportation Scheduling in the Transport Network

A. N. Ignatov

Moscow Aviation Institute, Moscow, Russia
e-mail: alexei.ignatov1@gmail.com

Received April 6, 2023
Revised June 19, 2023
Accepted July 20, 2023

Abstract—The problem of cargoes transportation scheduling in the transport network represented by an undirected multigraph is considered. Transportations between vertices are provided at predefined time intervals. The iterative algorithm to search for a solution approximate to the optimal one by criterion value is proposed in the problem under consideration. The algorithm is constructed on the base of solutions of mixed integer linear programming problems. The applicability of the algorithm is tested by the example with more than 90 million binary variables.

Keywords: transport network, multigraph, cargoes transportation, schedule, mixed integer linear programming

DOI: 10.25728/arcRAS.2023.74.63.001

1. INTRODUCTION

The scheduling problem (of cargoes, trains, locomotives) is a widespread problem both in theory and in practice. Publications on this topic can be divided into several groups: by the presence of movement time in the problem, by the fixedness of the movement time between vertices, by the fixedness of the movement route at optimization, by structure of the transport network (multi)graph. For example, [1] used only duration of movement time along transport network graph arcs, the graph of the special structure (one-way railway) is considered in [2, 3]. The scheduling problem for the railway network of general structure with a fixed set of routes for trains is researched in [4, 5]. The problem to construct train routes and their movement times along the railway network is solved simultaneously in [6, 7]. Time in [6, 7] is set to be discrete, that may cause to the huge dimension of the problem. The simultaneous problem of scheduling and routing for general structure railway networks is researched in [8–11]. Transportations between vertices in [8–11] are carried out at only predetermined time intervals.

Difference of problem statements with fixed movement time between vertices from problem statements with arbitrary time is very principal. In the latter there is supposed that some transport is able for the carriage at any interval of time. But it is not always physically realizable. The principal difference [11] from other researches is in possibility to not come in the arrival vertex before the end of time interval for which the timetable is scheduling (hereinafter referred as planning horizon). Such possibility is relevant when there is a cargo that needs to be departed shortly before the end of the planning horizon. But such possibility complicates not only the mathematical model of carriages but also increases the computation time [11]. That's why it is relevant to construct a faster algorithm than the algorithm from [11]. Such algorithm is being constructed in the present paper.

Within the framework of the transportation model under consideration time of readiness for departure, starting and ending times of movement of any vehicle carrying out transportation between

vertices are fixed. These characteristics are real numbers. Optimization in the future will be carried out with the goal to find a particular vehicle for a particular cargo. Other optimization variables will also be considered. For example it will be the parking time of cargo at various vertices, the expected quantity of time before delivery after the end of the final planning horizon, delivery of cargo to the destination vertex.

A system from linear equalities and inequalities is formed to construct the algorithm. This system contains binary and continuous variables and sets a mathematical model for the carriage of cargoes along a transport network of general structure. The transport network is represented by an undirected multigraph. The algorithm performs a decomposition of a set of cargoes, as well as a decomposition of the planning horizon to reduce the computation time. The algorithm contains one more possibility to accelerate the computation time. This possibility is based on the cut out of transportations that are unlikely to be used by cargoes due to the beginning time of these transportations is earlier than expected arrival time in the respective to these transportations vertices. The developed algorithm is tested on a meaningful example with millions of binary variables.

2. BASIC DESIGNATIONS AND ASSUMPTIONS

Let us consider a transport network represented by an undirected multigraph $G =< V, E >$, where V is a set of vertices (cities, railway stations, plants, airports, seaports) and E is a set of edges (highways, railway tracks, seaways, airways), connecting these vertices. Let $|V| = M \geqslant 2$. By renumbering vertices of multigraph G from 1 to M, we compose a set of indices $V' = \{1, 2, \ldots, M\}$. Each element of this set uniquely determines the vertex of multigraph G. Note that the need in multigraphs for modelling transport systems follows from applications. Namely, oncoming traffic between two railway stations in the same period of time, for safety reasons, should be separated along different railway tracks. Therefore for modelling of transportations, it is necessary to separately consider all railway tracks (edges) from one vertex (station) to another (station).

We will count the time in minutes relative to a certain moment of reference. By the planning horizon we mean the time interval $[0, T_{\max})$, for which the timetable is scheduling. If the timetable is scheduled on a day (1440 minutes), then $T_{\max} = 1440$.

We divide the planning horizon into P non-overlapping intervals (half-open intervals) $\mathcal{T}_1, \ldots, \mathcal{T}_P$, i.e., $[0, T_{\max}) = \bigcup_{p=1}^{P} \mathcal{T}_p$, where $\forall p_1, p_2 \in \{1, \ldots, P\} : p_1 \neq p_2 \;\; \mathcal{T}_{p_1} \bigcap \mathcal{T}_{p_2} = \varnothing$. These intervals we will name as *partition intervals*. Let us introduce auxiliary variables $\underline{\mathcal{T}}_p \stackrel{\text{def}}{=} \inf \mathcal{T}_p$, $\overline{\mathcal{T}}_p \stackrel{\text{def}}{=} \sup \mathcal{T}_p$, $p = \overline{1, P}$. We construct sets $\mathcal{T}_1, \ldots, \mathcal{T}_P$ in such manner that

$$\underline{\mathcal{T}}_1 = 0, \;\; \overline{\mathcal{T}}_P = T_{\max}, \;\; \underline{\mathcal{T}}_{p+1} = \overline{\mathcal{T}}_p, \;\; p = \overline{1, P-1}.$$

Let us have I cargoes (parcels, containers, trains), for each of that there are given:

- index of departure vertex $v_i^{\text{dep}} \in V'$;
- index of arrival (destination) vertex $v_i^{\text{arr}} \in V'$;
- time of readiness for departure $t_i^{\text{dep}} \in [0, T_{\max})$;
- maximal amount of time d_i during which the cargo is allowed to be at the departure vertex from the moment of readiness;
- cargo travel time T_i, i.e. maximal amount of time during which the cargo is allowed to be on the transport network (excluding time at the departure vertex) computed in minutes;
- mass of the cargo $w_i \in \mathbb{R}_+$,

$i = \overline{1, I}$. The cargo is assumed to be indivisible in sense that it can not be sent in parts.

Cargoes carriages between vertices can only be carried out at certain intervals. Let K movements/transportations (by aircrafts, sea ships, trains, trucks) between vertices are available. Pa-

rameters of transportation mathematically can be represented by 7-element row $z_k \stackrel{\text{def}}{=} (v_k^{\text{beg}}, v_k^{\text{end}}, n_k, t_k^{\text{beg}}, t_k^{\text{end}}, W_k, C_k)$, where $v_k^{\text{beg}} \in V'$ is the index of starting vertex of movement, $v_k^{\text{end}} \in V'$ is the index of ending vertex of movement, moreover v_k^{beg} and v_k^{end} are indices of adjacent vertices in multigraph G, n_k is the number of the track (edge), connecting vertices with indices v_k^{beg} and v_k^{end}, $t_k^{\text{beg}} \in [0, T_{\max})$ is starting time of movement, t_k^{end} is ending time of movement, W_k is maximum transportable mass during transportation, C_k is the transportation cost of unit mass, $k = \overline{1, K}$. Let us designate using $\mathcal{Z}$ the set of all vectors z_k, $k = \overline{1, K}$. We renumber elements of set $\mathcal{Z}$ from 1 to K. Thus number from 1 to K determines the transportation and its transportation uniquely

In the future, as *timetable* of cargo we will understand the chain of transportation numbers that are used by it. One can easily determine by transportation numbers the vertices visited by the cargo, the time of visiting these vertices, edges of the multigraph used for movement, as well as other characteristics of the movement.

According to introduced partition intervals $\mathcal{T}_1, \ldots, \mathcal{T}_P$ we split the set of transportations for several parts, namely $\{1, \ldots, K\} = \mathcal{K}_1 \bigcup \mathcal{K}_2 \bigcup \ldots \bigcup \mathcal{K}_P$, where $\mathcal{K}_p \stackrel{\text{def}}{=} \{k \in \mathbb{N} : k \leqslant K, t_k^{\text{beg}} \in \mathcal{T}_p\}$, $p = \overline{1, P}$.

When transportations are carried out, the warehouses in which goods are stored can be filled. In addition some operations may be performed with cargoes, for example, repacking. Therefore we introduce minimal and maximal possible duration of stay at the vertex with index v_k^{end} after using of transportation with number k by cargo with number i: $t_{i,k}^{\text{st min}}$ and $t_{i,k}^{\text{st max}}$, $i = \overline{1, I}$, $k = \overline{1, K}$. Obviously, $\forall i = \overline{1, I}$, $k = \overline{1, K}$ $0 \leqslant t_{i,k}^{\text{st min}} \leqslant t_{i,k}^{\text{st max}}$.

Let τ_{m_1,m_2} is expected duration (starting from the moment of readiness for departure) of a cargo carriage from vertex with index m_1 to vertex with index m_2, $m_1, m_2 = \overline{1, M}$. Obviously that $\tau_{m_1,m_1} = 0$, $m_1 = \overline{1, M}$. If historical observations on carriages from vertex with index m_1 to vertex with index m_2 are available then as τ_{m_1,m_2} one can select sample mean by existing observations, $m_1, m_2 = \overline{1, M}$. If this data is unavailable then the indicated value can be estimated by an expert. Also we introduce value η_{m_1,m_2} that designates expected duration from the moment of readiness for departure to departure from vertex with index m_1 to vertex with index m_2. This value is set by analogy with τ_{m_1,m_2}, $m_1, m_2 = \overline{1, M}$.

As the route of cargo with number i we will understand the chain from transportations numbers used in series by this cargo, $i = \overline{1, I}$. As consequence one can determine the chain of vertices traversed in series by this cargo using the route. We limit the maximal quantity of transportations in the route during the planning horizon by some predetermined value J. As jth phase of the route of ith train we will mean movement of this train when there is used jth transportation in the route, $i = \overline{1, I}$, $j = \overline{1, J+1}$. Phase $J+1$ is technical, movement in that is not provided, it is needed for accuracy in the mathematical model formulation. We will name the vertex intermediate for ith cargo if it's neither the vertex of departure nor the vertex of arrival for that, $i = \overline{1, I}$.

We also introduce value $\mathcal{D}_i$, characterizing the denial in carriage to ith cargo: 0, when cargo is denied to carriage, 1 is otherwise, $i = \overline{1, I}$. The denial in carriage may be caused by there are not enough transportations to achieve the destination vertex with restrictions on travel time and other physical limitations. In the ideal case any of values $\mathcal{D}_i$ is equal to one, $i = \overline{1, I}$, but it is not always realizable or it was not successful to find schedule that leads to this result,

3. AUXILIARY RESULTS TO CONSTRUCT THE ALGORITHM

3.1. Mathematical Model of Movements Along Transport Network

We divide set of cargoes numbers $\mathcal{I}$ into S non-overlapping subsets $\mathcal{I}_s$, i.e. $\mathcal{I} \stackrel{\text{def}}{=} \{1, \ldots, I\} = \bigcup_{s=1}^{S} \mathcal{I}_s$, and besides $\forall s_1, s_2 \in \{1, \ldots, S\} : s_1 \neq s_2$ $\mathcal{I}_{s_1} \bigcap \mathcal{I}_{s_2} = \varnothing$. In [10–12] there was a proposal to divide set $\mathcal{I}$ by principle of having cargo numbers with the same departure and destination vertices

in subsets. In addition one can only construct as many subsets as quantity of cargoes. In this case, in the subset with index 1 there will be a cargo number with the earliest/latest time of readiness for departure, with index 2—the second/penultimate time, etc.

We suppose that for every cargo with number from sets $\mathcal{I}_1, \ldots, \mathcal{I}_{\tilde{s}-1}$ there is the denial in carriage or a timetable, i.e. the chain from transportations numbers. If there is the denial in carriage for cargo with number $\hat{i} \in \bigcup_{s=1}^{\tilde{s}-1} \mathcal{I}_s$ then we assign $\hat{\delta}_{\hat{i},j,k} = 0$, $j = \overline{1, J+1}$, $k = \overline{1, K}$, and $\mathcal{D}_{\hat{i}} = 0$. If cargo with number $\hat{i} \in \bigcup_{s=1}^{\tilde{s}-1} \mathcal{I}_s$ is permitted to carriage, then value $\hat{\delta}_{\hat{i},j,k}$ is equal to one, if this cargo uses transportation with number k at the jth phase, and to zero, otherwise, $j = \overline{1, J+1}$, $k = \overline{1, K}$. At the same time we assign $\mathcal{D}_{\hat{i}} = 1$.

Initially we will construct the timetable for time interval $[0, \overline{\mathcal{T}}_1)$ to reduce the dimension of optimization problems to be solved in the future. To construct the timetable for time interval $[0, \overline{\mathcal{T}}_2) = [0, \overline{\mathcal{T}}_1) \bigcup \mathcal{T}_2$ we will take into account (freeze) the timetable for time interval $[0, \overline{\mathcal{T}}_1)$, To construct the timetable for time interval $[0, \overline{\mathcal{T}}_3) = [0, \overline{\mathcal{T}}_2) \bigcup \mathcal{T}_3$ we will take into account (freeze) the timetable for time interval $[0, \overline{\mathcal{T}}_2)$ and so on.

For this reason we consider only transportations from the beginning of the planning horizon until end of the interval $\mathcal{T}_{\tilde{p}}$, where $\tilde{p}$ is an arbitrary number from set $\{1, \ldots, P\}$. Let us formulate a set of constraints stating movements along the multigraph for cargoes with numbers from set $\mathcal{I}_{\tilde{s}}$ in this time, i.e. in the planning subhorizon $[0, \overline{\mathcal{T}}_{\tilde{p}})$. Let us suppose initially, that a timetable for cargoes with numbers from set $\mathcal{I}_{\tilde{s}}$ for subhorizon $[0, \overline{\mathcal{T}}_{\tilde{p}-1})$ $(\tilde{p} > 1)$ is not available.

By $\mathcal{K}^{\tilde{s},\tilde{p}}$ we will mean some non-empty set of transportations set $\bigcup_{p=1}^{\tilde{p}} \mathcal{K}_p$, selected for cargoes with numbers from set $\mathcal{I}_{\tilde{s}}$.

For this purpose we introduce auxiliary $\delta^{\tilde{p}}_{i,j,k}$, characterizing the usage of kth transportation by cargo with number i at jth phase when timetable is formed for the planning subhorizon $[0, \overline{\mathcal{T}}_{\tilde{p}})$, $i \in \mathcal{I}_{\tilde{s}}$, $j = \overline{1, J+1}$, $k \in \mathcal{K}^{\tilde{s},\tilde{p}}$. Variable $\delta^{\tilde{p}}_{i,j,k}$ is equal to one, if transportation with number k is used by ith cargo at jth phase, and to zero, otherwise.

We have by defnition of variables $\delta^{\tilde{p}}_{i,j,k}$

$$\delta^{\tilde{p}}_{i,j,k} \in \{0, 1\}, \quad i \in \mathcal{I}_{\tilde{s}}, \quad j = \overline{1, J+1}, \quad k \in \mathcal{K}^{\tilde{s},\tilde{p}}. \tag{1}$$

Movements of cargoes along multigraph G can be performed only along adjacent vertices

$$\sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta^{\tilde{p}}_{i,j,k} v^{\text{end}}_k \leqslant \sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta^{\tilde{p}}_{i,j+1,k} v^{\text{beg}}_k + \left(1 - \sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta^{\tilde{p}}_{i,j+1,k}\right) M^3, i \in \mathcal{I}_{\tilde{s}}, \quad j = \overline{1, J-1}, \tag{2}$$

$$\sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta^{\tilde{p}}_{i,j,k} v^{\text{end}}_k \geqslant \sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta^{\tilde{p}}_{i,j+1,k} v^{\text{beg}}_k - \left(1 - \sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta^{\tilde{p}}_{i,j+1,k}\right) M, i \in \mathcal{I}_{\tilde{s}}, \quad j = \overline{1, J-1}. \tag{3}$$

Let us remind that M is quantity of vertices in multigraph G. Constraints (2), (3) cause [10] to the fact that if for some $\tilde{i} \in \mathcal{I}_{\tilde{s}}$ and some $\tilde{j} \in \{1, \ldots, J\}$ it is true $\sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta^{\tilde{p}}_{\tilde{i},\tilde{j},k} = 0$, then $\sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta^{\tilde{p}}_{\tilde{i},j+1,k} = 0$, $j = \overline{\tilde{j}, J}$. If $\sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta^{\tilde{p}}_{\tilde{i},\tilde{j},k} = 1$, then $\sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta^{\tilde{p}}_{\tilde{i},\tilde{j}+1,k} = 0$ or $\sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta^{\tilde{p}}_{\tilde{i},\tilde{j}+1,k} = 1$. Constraints (2), (3) are identical to [10, 11] taking into account that the mathematical model is constructed for the planning subhorizon. Let us note that the third power of M in (2) ensures correctness of the mathematical model of movements along the multigraph [10].

Arrival at the destination vertex is possible in no more than J phases. Therefore we introduce constraints

$$\sum_{i \in \mathcal{I}_{\tilde{s}}} \sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta^{\tilde{p}}_{i,J+1,k} = 0. \tag{4}$$

Due to indivisibility of cargoes one can use no more than one transportation at any phase (including the first one)

$$\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}}\delta^{\tilde{p}}_{i,1,k}\leqslant 1,\quad i\in\mathcal{I}_{\tilde{s}}. \tag{5}$$

If carriage is begun then it must be performed from the respective departure vertex

$$\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}}\delta^{\tilde{p}}_{i,1,k}v^{\mathrm{beg}}_{k}=v^{\mathrm{dep}}_{i}\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}}\delta^{\tilde{p}}_{i,1,k},\quad i\in\mathcal{I}_{\tilde{s}}. \tag{6}$$

If cargo readiness to depart happens after the upper bound of interval $\mathcal{T}_{\tilde{p}}$, then for this cargo usage of transportations are prohibited until the end of $\mathcal{T}_{\tilde{p}}$, i.e.

$$\sum_{j=1}^{J}\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}}\delta^{\tilde{p}}_{i,j,k}=0,\quad \forall i\in\mathcal{I}_{\tilde{s}}:t^{\mathrm{dep}}_{i}\geqslant\overline{\mathcal{T}}_{\tilde{p}}. \tag{7}$$

Cargoes must be departed not earlier than the respective moments of readiness taking into account maximal duration of stay in departure vertices. At the same time it is possible to not depart cargo in interval $[0,\overline{\mathcal{T}}_{\tilde{p}})$, if it is admissible, taking into account maximal duration of stay in departure vertex. That's why we have constraints

$$t^{\mathrm{dep}}_{i}\leqslant\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}}\delta^{\tilde{p}}_{i,1,k}t^{\mathrm{beg}}_{k}+\left(1-\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}}\delta^{\tilde{p}}_{i,1,k}\right)\overline{\mathcal{T}}_{\tilde{p}}\leqslant t^{\mathrm{dep}}_{i}+d_i,\quad \forall i\in\mathcal{I}_{\tilde{s}}:t^{\mathrm{dep}}_{i}<\overline{\mathcal{T}}_{\tilde{p}}. \tag{8}$$

Let us comment constraints (8). For this reason we consider cargo with number $\tilde{i}\in\mathcal{I}_{\tilde{s}}:t^{\mathrm{dep}}_{\tilde{i}}<\overline{\mathcal{T}}_{\tilde{p}}$. Due to constraints (1) and (5) there are only two possible variants: $\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}}\delta^{\tilde{p}}_{\tilde{i},1,k}$ is equal to zero or one. At the same time equality of this sum to zero (i.e. cargo with number $\tilde{i}$ is not departed) causes to the fact that the following must be true: $\overline{\mathcal{T}}_{\tilde{p}}\leqslant t^{\mathrm{dep}}_{\tilde{i}}+d_{\tilde{i}}$. If this sum is equal to one, then according to (5) only one transportation can be used and its beginning time will be in the interval $[t^{\mathrm{dep}}_{\tilde{i}},t^{\mathrm{dep}}_{\tilde{i}}+d_{\tilde{i}}]$. It corresponds to the sense of constraints (8) introduced above.

From the same vertex cargo can only be departed once [1]

$$\sum_{j=1}^{J+1}\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}:v^{\mathrm{beg}}_{k}=m}\delta^{\tilde{p}}_{i,j,k}\leqslant 1,\quad i\in\mathcal{I}_{\tilde{s}},\quad m=\overline{1,M}. \tag{9}$$

Arriving in the same vertex for cargo more than once is prohibited

$$\sum_{j=1}^{J+1}\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}:v^{\mathrm{end}}_{k}=m}\delta^{\tilde{p}}_{i,j,k}\leqslant 1,\quad i\in\mathcal{I}_{\tilde{s}},\quad m=\overline{1,M}. \tag{10}$$

Departure from intermediate vertices of the route must not be earlier than arrival in these vertices. Therefore we have, taking into account minimal and maximal duration of stay, the following

$$\begin{aligned}\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}}\delta^{\tilde{p}}_{i,j,k}(t^{\mathrm{end}}_{k}+t^{\mathrm{st\,min}}_{i,k})\leqslant\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}}\delta^{\tilde{p}}_{i,j+1,k}t^{\mathrm{end}}_{k}\\+\left(1-\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}}\delta^{\tilde{p}}_{i,j+1,k}\right)\underline{T},\quad i\in\mathcal{I}_{\tilde{s}},\quad j=\overline{1,J-1},\end{aligned} \tag{11}$$

[1] Here and below it is assumed that the sum of any variables over an empty set is equal to zero.

where

$$\underline{T} = \max_{i\in\{1,\ldots,I\},k\in\{1,\ldots,K\}} t_k^{\text{end}} + t_{i,k}^{\text{st min}},$$

$$\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}} \delta_{i,j,k}^{\tilde{p}}(t_k^{\text{end}} + t_{i,k}^{\text{st max}}) \geqslant \sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}} \delta_{i,j+1,k}^{\tilde{p}} t_k^{\text{end}}, \quad i \in \mathcal{I}_{\tilde{s}}, \quad j = \overline{1, J-1}. \tag{12}$$

Constraints (11) and (12) are identical to the respective ones from [11].

To ensure allowability of parking (if it takes place) after the end of subhorizon $[0, \overline{\mathcal{T}}_{\tilde{p}})$ we impose constraints

$$\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}:v_k^{\text{end}}\neq v_i^{\text{arr}}} \delta_{i,j,k}^{\tilde{p}} \left(t_k^{\text{end}} + t_{i,k}^{\text{st max}} - \overline{\mathcal{T}}_{\tilde{p}}\right) + \overline{\mathcal{T}}_{\tilde{p}} \sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}} \delta_{i,j+1,k}^{\tilde{p}} \geqslant 0, \quad i \in \mathcal{I}_{\tilde{s}}, \quad j = \overline{1, J}. \tag{13}$$

To prohibit carriages after arrival in the destination vertex we use constraints

$$\sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}:v_k^{\text{end}}=v_i^{\text{arr}}} \delta_{i,j,k}^{\tilde{p}} \leqslant 2\left(1 - \sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}} \delta_{i,j+1,k}^{\tilde{p}}\right), \quad i \in \mathcal{I}_{\tilde{s}}, \quad j = \overline{1, J}. \tag{14}$$

Let us comment constraints (14). For this reason we consider cargo with number $\tilde{i} \in \mathcal{I}_{\tilde{s}}$. If this cargo arrived in the destination vertex after some phase then left part of (14) is equal to one. Therefore for compatibility of (14) it is needed that right side would be equal to zero. It means due to constraints (1) and (5) that the next after arrival phase will not be used as other phases. If cargo did not arrive in the destination vertex then left side of (14) is equal to zero. In this case the constraint is satisfied, because at any phase it is possible to use not more than one transportation. It means that right side will be equal to zero or one.

Let us introduce variable $\hat{T}_{i,j}^{\tilde{p}}$ that means duration spent by cargo with number i at jth (by order of traversing) intermediate vertex of its route during the planning subhorizon

$$\hat{T}_{i,j}^{\tilde{p}} = \sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}} \delta_{i,j+1,k}^{\tilde{p}}(t_k^{\text{beg}} - \overline{\mathcal{T}}_{\tilde{p}}) + \sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}:v_k^{\text{end}}\neq v_i^{\text{arr}}, t_k^{\text{end}} < \overline{\mathcal{T}}_{\tilde{p}}} \delta_{i,j,k}^{\tilde{p}}(\overline{\mathcal{T}}_{\tilde{p}} - t_k^{\text{end}}), \quad i \in \mathcal{I}_{\tilde{s}}, \quad j = \overline{1, J}. \tag{15}$$

We also assign $\hat{T}_{i,J+1}^{\tilde{p}} = 0$ for convenience of modelling.

Further we introduce new variables $\mathcal{F}_i^{\tilde{p}}$, characterizing the expected duration of time needed until arrival in the destination vertex for cargo with number i after the end of the planning subhorizon $[0, \overline{\mathcal{T}}_{\tilde{p}})$:

$$\begin{aligned}\mathcal{F}_i^{\tilde{p}} = \tau_{v_i^{\text{dep}},v_i^{\text{arr}}} &+ \sum_{j=1}^{J} \sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}} \delta_{i,j,k}\left(\tau_{v_k^{\text{end}},v_i^{\text{arr}}} - \tau_{v_k^{\text{beg}},v_i^{\text{arr}}}\right) \\ &+ \sum_{j=1}^{J} \sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}:t_k^{\text{end}}\geqslant\overline{\mathcal{T}}_{\tilde{p}}} \delta_{i,j,k}(t_k^{\text{end}} - \overline{\mathcal{T}}_p), \quad i \in \mathcal{I}_{\tilde{s}}.\end{aligned} \tag{16}$$

Next constraints are needed to not exceed cargo travel time

$$\begin{aligned}\mathcal{F}_i^{\tilde{p}} + \sum_{j=1}^{J} \sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}:t_k^{\text{end}}<\overline{\mathcal{T}}_p, v_k^{\text{end}}=v_i^{\text{arr}}} \delta_{i,j,k}^{\tilde{p}}\left(t_k^{\text{end}} - \overline{\mathcal{T}}_p\right) + \sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}} \delta_{i,1,k}^{\tilde{p}}(\overline{\mathcal{T}}_p - t_k^{\text{beg}}) \\ \leqslant T_i + \left(1 - \sum_{k\in\mathcal{K}^{\tilde{s},\tilde{p}}} \delta_{i,1,k}^{\tilde{p}}\right)\eta_{v_i^{\text{dep}},v_i^{\text{arr}}}, \quad \forall i \in \mathcal{I}_{\tilde{s}} : t_i^{\text{dep}} < \overline{\mathcal{T}}_p.\end{aligned} \tag{17}$$

Constraints (17) are identical to the respective ones from [11].

We introduce variables $\omega_i^{\tilde{p}}$, characterizing arrival of cargo with number i in the respective destination vertex on the base of used transportations during the planning subhorizon $[0, \overline{\mathcal{T}}_{\tilde{p}})$: 0—arrived, 1—did not arrive:

$$\omega_i^{\tilde{p}} = 1 - \sum_{j=1}^{J} \sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}: t_k^{\text{end}} < \overline{\mathcal{T}}_p, v_k^{\text{end}} = v_i^{\text{arr}}} \delta_{i,j,k}^{\tilde{p}}, \quad i \in \mathcal{I}_{\tilde{s}}. \tag{18}$$

Next constraints are caused by the need in not exceeding maximal allowable mass at transportation with number k

$$\sum_{i \in \mathcal{I}_{\tilde{s}}} \sum_{j=1}^{J+1} \delta_{i,j,k}^{\tilde{p}} w_i \leqslant W_k - \sum_{i \in \bigcup_{s=1}^{\tilde{s}-1} \mathcal{I}_s} \sum_{j=1}^{J+1} \hat{\delta}_{i,j,k} w_i, k \in \mathcal{K}^{\tilde{s},\tilde{p}}. \tag{19}$$

3.2. Optimality Criterion

Potentially system of equalities and inequalities (1)–(19) may not have a unique solution. Therefore a criterion is required to select among solutions. Let us compose from all $\delta_{i,j,k}^{\tilde{p}}$ vector $\delta^{\tilde{s},\tilde{p}}$, $i \in \mathcal{I}_{\tilde{s}}$, $j = \overline{1, J+1}$, $k \in \mathcal{K}^{\tilde{s},\tilde{p}}$. Also we compose from all $\mathcal{F}_i^{\tilde{p}}$ vector $\mathcal{F}^{\tilde{s},\tilde{p}}$, from $\omega_i^{\tilde{p}}$ vector $\omega^{\tilde{s},\tilde{p}}$, $i \in \mathcal{I}_{\tilde{s}}$. We unite all $\hat{T}_{i,j}^{\tilde{p}}$ in vector $\hat{T}^{\tilde{s},\tilde{p}}$, $i \in \mathcal{I}_{\tilde{s}}$, $j = \overline{1, J+1}$.

Let us choose the criterial function of the following form

$$\begin{aligned} & J_{\tilde{s}}^{\tilde{p}}\left(\delta^{\tilde{s},\tilde{p}}, \mathcal{F}^{\tilde{s},\tilde{p}}, \omega^{\tilde{s},\tilde{p}}, \hat{T}^{\tilde{s},\tilde{p}}\right) \\ & = c_1 \underbrace{\sum_{i \in \mathcal{I}_s} \sum_{j=1}^{J+1} \sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta_{i,j,k}^{\tilde{p}} (\min\{t_k^{\text{end}}, \overline{\mathcal{T}}_{\tilde{p}}\} - t_k^{\text{beg}})}_{\substack{\text{the total time in movement} \\ \text{during the planning subhorizon} [0, \overline{\mathcal{T}}_{\tilde{p}})}} + c_2 \underbrace{\sum_{i \in \mathcal{I}_s} \sum_{j=1}^{J+1} \hat{T}_{i,j}^{\tilde{p}}}_{\substack{\text{the total parking} \\ \text{time in} \\ \text{intermediate} \\ \text{vertices}}} \\ & + c_3 \underbrace{\sum_{i \in \mathcal{I}_s} \left(\sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta_{i,1,k}^{\tilde{p}} t_k^{\text{beg}} + \left(1 - \sum_{k=1}^{K} \delta_{i,1,k}^{\tilde{p}}\right) \overline{\mathcal{T}}_{\tilde{p}} - t_i^{\text{dep}} \right)}_{\substack{\text{the total parking time in departure vertices} \\ \text{from the time of readiness for departure} \\ \text{until the end of the planning subhorizon} [0, \overline{\mathcal{T}}_{\tilde{p}})}} \\ & + c_4 \underbrace{\sum_{i \in \mathcal{I}_s} \sum_{j=1}^{J+1} \sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}}} \delta_{i,j,k}^{\tilde{p}} w_i C_k}_{\substack{\text{the total cost} \\ \text{of transportations}}} + c_5 \underbrace{\sum_{i \in \mathcal{I}_s} \mathcal{F}_i^{\tilde{p}}}_{\substack{\text{the total} \\ \text{expected} \\ \text{time until} \\ \text{delivery}}} + c_6 \underbrace{\sum_{i \in \mathcal{I}_s} \omega_i^{\tilde{p}}}_{\substack{\text{the total} \\ \text{quantity} \\ \text{of undelivered} \\ \text{cargoes during} \\ \text{the planning} \\ \text{subhorizon}}}, \end{aligned} \tag{20}$$

where $c_1, \ldots, c_6$ are non-negative values chosen by a decision-maker. The choice of $c_1, \ldots, c_6$ impacts on the sense of optimization. When $c_1 = c_2 = c_3 = c_5 = c_6 = 0$, $c_4 = 1$ there is a problem to minimize the total cost of transportations. When $c_1 = c_2 = c_3 = c_5 = 1$, $c_4 = c_6 = 0$ there is a problem to minimize the sum of already spent time by cargoes in the transport network within the planning subhorizon and the expected time to delivery after the end of the planning subhorizon. We will mean as rth *criterion component* multiplier of c_r in (20), $r = \overline{1, 6}$. It should be noted

that not all criterion components are homogeneous. The first, second, third and fifth are measured in minutes, the fourth is in units of cost, and the sixth is in pieces. If optimization problem is connected only with homogeneous components, then dimension of coefficients $c_1, \ldots, c_6$ is not important. If it is needed for optimization to take into account heterogeneous components then problem to minimize the total cost should be considered, i.e. values c_1, c_2, c_3, c_5 will be measured in units of cost/minute, and c_6 in units of cost/piece.

If one takes $\tilde{p} = P$, then the planning subhorizon will coincide with $[0, \overline{\mathcal{T}}_{\tilde{p}})$. If we do not split cargoes numbers set, i.e. $\mathcal{I} = \mathcal{I}_{\tilde{s}}$ and $\mathcal{K}^{\tilde{s},P} = \bigcup_{p=1}^{P} \mathcal{K}_p$, then criterion (20) and system of constraints (1)–(19) will be precisely the same as criterion and system of constraints in [11]. But for this split (more accurately—for the absence of the split) of cargoes numbers set and value of $\tilde{p}$, that is suitable to decrease quantity elements in transportation set, used for scheduling, direct optimization of criterion (20) with the purpose to find a timetable on the entire planning horizon may be very prolonged. That's why we will form the algorithm to search although not optimal but relatively fast solution on the base of obtained in the paper constraints.

The presence of linear on optimization variables constraints (2)–(19) and linear criterion (20), binary variables vectors $\delta^{\tilde{s},\tilde{p}}$ and $\omega^{\tilde{s},\tilde{p}}$, real variable vectors $\mathcal{F}^{\tilde{s},\tilde{p}}$ and $\hat{T}^{\tilde{s},\tilde{p}}$ makes problem (20) with constraints (1)–(19) mixed integer linear programming problem.

4. THE ALGORITHM FOR SCHEDULING

At formation of the algorithm we will take into account the possibility of more fast computation time by cut out of transportations that are unlikely to be used by cargoes.

It makes no sense at scheduling for a given subhorizon to take into account transportations from vertices to which none of the cargo in this subhorizon will arrive. Generally speaking, in order to determine whether a particular cargo will vertex a specific vertex in a given time, it is necessary to solve the corresponding optimization problem. However, solving these types of problems takes time. Therefore, to establish the fact that the loads will not arrive a certain vertex, we will use the values τ_{m_1,m_2}, $m_1 = \overline{1,M}$, $m_2 = \overline{1,M}$. Of course, the conclusion on possibility to arrive at a certain vertex based on the values τ_{m_1,m_2} is not always true, $m_1 = \overline{1,M}$, $m_2 = \overline{1,M}$. This is because these values are based on past transportation history rather than the transportation currently available. Nevertheless, this significantly reduces the computation time, although with a deterioration in the value of the criterion function/inability to accept some cargoes for transportation. We will compare values τ_{m_1,m_2} with the ratio of the length of the corresponding partition interval to an acceleration parameter, $m_1 = \overline{1,M}$, $m_2 = \overline{1,M}$. The acceleration parameter, which is dimensionless, will be denoted by A. The lower A the less transportations will be crossed out, but the more cargoes will likely be accepted for carriage. And, on the contrary, the larger A the faster computation time will be, but the quality (in terms of cargoes accepted for carriage) of obtained solution will be worse. If $A = 0$ there will be no strikeouts. When solving optimization problems, it seems most rational to set A equal to one. In this case, the expected time before arrival at a certain vertex will be compared with the duration of the corresponding partition interval, i.e. a period of time in which the timetable has not yet been frozen and is being searched.

1. Values $c_1, \ldots, c_6 \in \mathbb{R}_+$ are initialized. Numbers P, $J \in \mathbb{N}$ are stated. The number $A \in \mathbb{R}_+$ is set.

2. Set of cargoes numbers is divided into $S \in \mathbb{N}$ non-overlapping subsets $\mathcal{I}_s$, i.e. $\{1, \ldots, I\} = \bigcup_{s=1}^{S} \mathcal{I}_s$, and besides $\forall s_1, s_2 \in \{1, \ldots, S\} : s_1 \neq s_2\ \mathcal{I}_{s_1} \bigcap \mathcal{I}_{s_2} = \varnothing$.

3. Partition intervals $\mathcal{T}_1, \ldots, \mathcal{T}_P$ are formed in such manner, that $[0, T_{\max}) = \bigcup_{p=1}^{P} \mathcal{T}_p$, where $\forall p_1, p_2 \in \{1, \ldots, P\} : p_1 \neq p_2\ \mathcal{T}_{p_1} \bigcap \mathcal{T}_{p_2} = \varnothing$, and besides $\underline{\mathcal{T}}_1 = 0$, $\overline{\mathcal{T}}_P = T_{\max}$, $\underline{\mathcal{T}}_{p+1} = \overline{\mathcal{T}}_p$, $p = \overline{1, P-1}$.

4. Sets $\mathcal{K}_p = \{k \in \mathbb{N} : k \leqslant K, t_k^{\text{beg}} \in \mathcal{T}_p\}$ are formed, $p = \overline{1,P}$.

5. Parameter $\tilde{s} = 1$ is initialized by 1.

6. Parameter $\tilde{p} = 1$ is initialized by 1.

7. If $\tilde{p}$ is equal to one, then set $\mathcal{V}^{\tilde{s},\tilde{p}} = \bigcup_{i \in \mathcal{I}_{\tilde{s}}} v_i^{\text{dep}}$ is formed. If $\tilde{p}$ is greater than one, then

$$\mathcal{V}^{\tilde{s},\tilde{p}} = \bigcup_{i \in \mathcal{I}_{\tilde{s}}} \begin{cases} v_i^{\text{dep}}, & \sum\limits_{j=1}^{J+1} \sum\limits_{k \in \mathcal{K}^{\tilde{s},\tilde{p}-1}} \overline{\delta}_{i,j,k}^{\tilde{p}-1} = 0, \\ \sum\limits_{k \in \mathcal{K}^{\tilde{s},\tilde{p}-1}} \overline{\delta}_{i,j_i,k}^{\tilde{p}-1} v_k^{\text{end}}, & \sum\limits_{j=1}^{J+1} \sum\limits_{k \in \mathcal{K}^{\tilde{s},\tilde{p}-1}} \overline{\delta}_{i,j,k}^{\tilde{p}-1} > 0, \end{cases}$$

where

$$j_i = \sum_{j=1}^{J+1} \sum_{k \in \mathcal{K}^{\tilde{s},\tilde{p}-1}} \overline{\delta}_{i,j,k}^{\tilde{p}-1}, \quad i \in \mathcal{I}_{\tilde{s}}.$$

Set $\mathcal{V}^{\tilde{s},\tilde{p}}$ consists of departure vertices indices for those cargoes that have not been in movement yet and from indices of last (on the current moment) vertices for those cargoes that had at least one transportation.

8. If $A = 0$, then $\mathcal{K}^{\tilde{s},\tilde{p}} = \mathcal{K}_{\tilde{p}}$. If $A > 0$, then set $\mathcal{K}^{\tilde{s},\tilde{p}} = \left\{ k \in \mathcal{K}_{\tilde{p}} : \min\limits_{m \in \mathcal{V}^{\tilde{s},\tilde{p}}} \tau_{m,v_k^{\text{beg}}} \leqslant (\overline{\mathcal{T}}_{\tilde{p}} - \underline{\mathcal{I}}_{\tilde{p}})/A, \right.$ $\left. \min\limits_{i \in \mathcal{I}_{\tilde{s}}} t_i^{\text{dep}} \leqslant t_k^{\text{beg}} \right\}$ is formed.

9. If $\tilde{p} > 1$, then set $\mathcal{K}^{\tilde{s},\tilde{p}} = \bigcup_{p=1}^{\tilde{p}-1} \mathcal{K}^{\tilde{s},p} \bigcup \mathcal{K}^{\tilde{s},\tilde{p}}$ is formed. If $\tilde{p} = 1$, then $\mathcal{K}^{\tilde{s},\tilde{p}} = \mathcal{K}^{\tilde{s},\tilde{p}}$.

10. If set $\mathcal{K}^{\tilde{s},\tilde{p}}$ is empty and $\tilde{p} < P$, then value $\tilde{p}$ is increased by 1, go to step 7.

If set $\mathcal{K}^{\tilde{s},\tilde{p}}$ is empty and $\tilde{p} = P$, then $\hat{\delta}_{i,j,k} = 0$, $\mathcal{D}_i = 0$, $i \in \mathcal{I}_{\tilde{s}}$, $j = \overline{1, J+1}$, $k = \overline{1, K}$. If $\tilde{s} = S$, then the algorithm is finished. If $\tilde{s} < S$, then value $\tilde{s}$ is increased by 1, go to step 6.

If set $\mathcal{K}^{\tilde{s},\tilde{p}}$ is not empty, go to step 11.

11. The problem

$$J_{\tilde{s}}^{\tilde{p}}(\delta^{\tilde{s},\tilde{p}}, \mathcal{F}^{\tilde{s},\tilde{p}}, \omega^{\tilde{s},\tilde{p}}, \hat{T}^{\tilde{s},\tilde{p}}) \to \min_{\delta^{\tilde{s},\tilde{p}}, \mathcal{F}^{\tilde{s},\tilde{p}}, \omega^{\tilde{s},\tilde{p}}, \hat{T}^{\tilde{s},\tilde{p}}}$$

with constraints (1)–(19), and also (when $\tilde{p} > 1$) constraint

$$\delta_{i,j,k}^{\tilde{p}} = \overline{\delta}_{i,j,k}^{\tilde{p}-1}, \quad i \in \mathcal{I}_{\tilde{s}}, \quad j = \overline{1, J+1}, \quad k \in \mathcal{K}^{\tilde{s},\tilde{p}-1} \tag{21}$$

is solved.

If a solution of this problem does not exist then $\hat{\delta}_{i,j,k} = 0$, $\mathcal{D}_i = 0$, $i \in \mathcal{I}_{\tilde{s}}$, $j = \overline{1, J+1}$, $k = \overline{1, K}$. If $\tilde{s} = S$, then the algorithm is finished. If $\tilde{s} < S$, then value $\tilde{s}$ is increased by 1, go to step 6.

If a solution was found and $\tilde{p} < P$, then values $\overline{\delta}_{i,j,k}^{\tilde{p}}$ are set: $\overline{\delta}_{i,j,k}^{\tilde{p}}$ is equal to one, if cargo with number i at jth phase uses transportation with number k, and is equal to zero, otherwise, $i \in \mathcal{I}_{\tilde{s}}$, $j = \overline{1, J+1}$, $k \in \mathcal{K}^{\tilde{s},\tilde{p}}$. Value $\tilde{p}$ is increased by 1, go to step 7.

If a solution was found and $\tilde{p} = P$, then $\mathcal{D}_i = 1$, values $\hat{\delta}_{i,j,k}$ are set: $\hat{\delta}_{i,j,k}$ is equal to one, if cargo with number i at jth phase uses transportation with number k, and is equal to zero, otherwise, $i \in \mathcal{I}_{\tilde{s}}$, $j = \overline{1, J+1}$, $k = \overline{1, K}$. If $\tilde{s} = S$, then the algorithm is finished. If $\tilde{s} < S$, then value $\tilde{s}$ is increased by 1, go to step 6.

Is should be noted that constraint (21) allows to freeze the timetable for time interval $[0, \overline{\mathcal{T}}_1)$ when the timetable for interval $[0, \overline{\mathcal{T}}_2)$ is searched, the timetable for time interval $[0, \overline{\mathcal{T}}_2)$ when the timetable for interval $[0, \overline{\mathcal{T}}_3)$ is searched and so on.

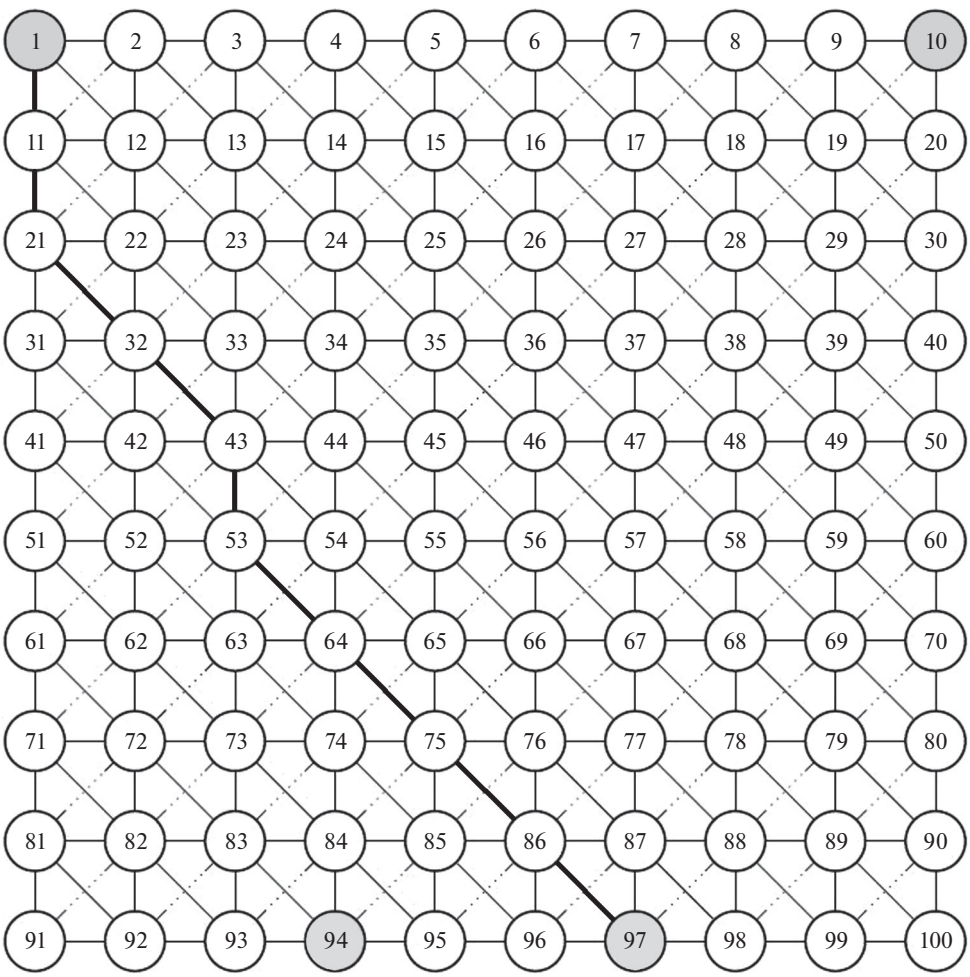
As the *minimal/maximal time* algorithm we will name such version of the proposed above algorithm when at step 2 the split is carried out in ascending and descending order of cargoes readiness moments for departure. Namely, set $\mathcal{I}_1$ consists of cargo number with the earliest/latest time of readiness for departure, set $\mathcal{I}_2$—with the second/penultimate and so on.

5. THE EXAMPLE

Let us consider a model example.

Let the multigraph of the transport network has the form shown in figure. For greater clarity the second track (the second edge) between adjacent vertices is omitted. The graph shows tracks with number 1. Some edges are indicated by a dashed line to show the multilevel intersection of edges in the transport network.

Suppose that some point of reference is chosen and $T_{\max} = 1440$ minutes. Starting from the point of reference: 5 cargoes of the same mass in 1 unit appear every 60 minutes at the vertex with index 1, these cargoes need to be transported to the vertex with index 97; 5 cargoes of the same mass in 1 unit appear every 60 minutes at the vertex with index 10, these cargoes need to be transported to the vertex with index 94.



Multigraph G of the transport network (by orange color departure and destination vertices are highlighted, by blue color the most frequent path of delivered cargoes for the one of the obtained results are highlighted).

Table 1. Properties of an approximate solution found by minimal time algorithm in the format: the total carriages time/the quantity of cargoes accepted for delivery/the quantity of delivered cargoes/the total cost of carriages/the computation time in minutes for various P and J

P \ J	9	12	15
6	147 960/192/68/21 550/75	**183 855**/240/122/28 840/91	**183 600**/240/122/28 840/99
12	141 675/182/60/19 720/45	**186 435**/240/117/28 520/52	**186 435**/240/117/28 520/58
24	170 230/220/94/25 200/54	**195 590**/240/96/28 160/59	**195 590**/240/96/28 160/64

Table 2. Properties of an approximate solution found by minimal time algorithm in the format: the total carriages time/the quantity of cargoes accepted for delivery/the quantity of delivered cargoes/the total cost of carriages/the computation time in minutes for various P and J

P \ J	9	12	15
6	148 475/192/64/21 530/73	**183 580**/240/122/29 040/88	**183 580**/240/122/29 040/107
12	124 245/156/26/15 560/50	187 870/236/106/28 060/54	**191 685**/240/110/28 770/62
24	171 470/222/96/25 760/55	171 470/222/96/25 760/57	171 470/222/96/25 760/62

Tranportations between vertices with an index difference equal to 1 or 10 by the absolute value are carried out every 30 minutes, cost of such transportations is 10 per unit of mass, maximal mass to transport is 2 units, duration of transportation is 60 minutes. Tranportations between vertices with an index difference equal to 9 or 11 by the absolute value are carried out every 30 minutes, cost of such transportations is 20 per unit of mass, maximal mass to transport is 2 units, duration of transportation is 85 minutes. Thus $I = 240$, $K = 32\,832$, $M = 100$.

Suppose also that $d_i = 180$, $T_i = 960$, $t_{i,k}^{\text{st}\,\min} = 0$, $t_{i,k}^{\text{st}\,\max} = 120$, $i = \overline{1,I}$, $k = \overline{1,K}$.

Suppose $\eta_{m_1,m_2} = 0$, $m_1, m_2 = \overline{1,100}$. Let

$$\tau_{m_1+1,m_2+1} = \begin{cases} 90, & |m_1\%10 - m_2\%10| = 1 \text{ and } ||\lfloor m_1/10\rfloor - \lfloor m_2/10\rfloor| = 1 \\ 60|m_1\%10 - m_2\%10| + 60|\lfloor m_1/10\rfloor - \lfloor m_2/10\rfloor|, & \text{otherwise,} \end{cases}$$

where $x\%y$ is remainder of x divided by y, $\lfloor x\rfloor$ is the integer part of x, $m_1, m_2 = \overline{0,99}$. Such choice of values τ_{m_1,m_2} provides that expected carriage duration from one adjacent vertex to another (if they are connected diagonally) is 90 minutes. In all other cases the expected carriage duration is proportional to the minimum number of edges when travelling from one vertex to another is carried out without using diagonal edges.

We consider the case where $c_1 = c_2 = c_3 = c_5 = 1$, $c_4 = c_6 = 0$. We set $A = 1$. Let us analyze, how results of applying proposed algorithms depend on P and J. The duration of intervals $\mathcal{T}_1, \ldots, \mathcal{T}_P$ will be the same. Let us preliminarily note that with available transportations, direct carriage from the vertex with index 1 to the vertex with index 12 costs the same as carriage through the intermediate vertex with index 2 (or index 11), while carriage directly takes less time. However, the fastest carriage from departure vertices–diagonally–due to the declared maximal mass and the frequency of transportations is not available for every cargo, so the optimization problem, generally speaking, is non-trivial.

In Tables 1 and 2 by bold font there are highlighted cases where all cargoes were accepted for delivery. As follows from Tables 1 and 2 the best result was obtained for maximal time algorithm with $P = 6$, $J = 12$. This solution we will name *basic*. For the basic solution the most frequent chain of vertices indices traversed at movement by delivered cargoes is

$$1 \to 11 \to 21 \to 32 \to 43 \to 53 \to 64 \to 75 \to 86 \to 97.$$

Table 3. Further improve of obtained solution by maximal time algorithm

Parameters of the algorithm	The total time of carriages	The quantity of cargoes accepted for delivery	The quantity of delivered cargoes	The total cost of carriages	The computation time, minutes
$A = 1, P = 4,$ $J = 12$	**182 455**	240	122	28 830	222
$A = 0{,}5, P = 6,$ $J = 12$	**183 165**	240	124	29 000	187

This chain appeared for 8 cargoes. Among delivered during the planning horizon cargoes: for 74 cargoes there were used 9 transportations, for 41 cargoes there were used 10 transportations, for 7 cargoes there were used 11 transportations. Exactly half of delivered cargoes was sent from the vertex with index 1.

Another result of the study is the fact that for cases when all cargoes are accepted for delivery, with a fixed J with decreasing P the computation time (as expected) increases, since mathematical programming problems of higher dimension are solved. Decreasing in the criterion is also observed. The growth of J at a fixed P causes to the fact that more cargoes are accepted for delivery. However, an increase in J from 12 to 15 in this problem did not allow us to decrease the criterion value always. This observation can be caused by the fact that T_i is relatively small, $i = \overline{1, I}$. Therefore, routes with a large number of transportations and travel time from the moment of readiness can not be used. In addition, the goal of optimization is to minimize the total travel time, and diagonal movement, as noted earlier, faster.

Note that even at $J = 12$ taking into account constraint (4) there are $I \cdot J \cdot K = 94\,556\,160$ binary variables in the problem under study. At the same time the solution search time is about an hour, which can be considered as an acceptable speed. To speed up the search for a solution, one can, for example, fix a certain set of vertices through which this or that cargo must travel. If a timetable is found for this set in such manner, then it is possible not to search for a timetable for this set of cargoes on the entire set of transportations. It is also possible to reduce the number of elements in the set $\mathcal{K}^{\tilde{s},\tilde{p}}$, formed at the 8th step of the proposed algorithm, $\tilde{s} = \overline{1, S}$, $\tilde{p} = \overline{1, P}$. For example, one can exclude transportations with starting or ending vertices that have already been visited by all cargoes from the set $\mathcal{I}_{\tilde{s}}$, $\tilde{s} = \overline{1, S}$. However, such (and similar) modifications, leading to the increasing of the obtaining solution speed, may degrade the solution in terms of quality.

Let us investigate the question about quality of basic solution. For this purpose we reduce P or A.

As follows from Table 3 decreasing A and P allowed to find a bit better (around 0.5 %) solution by criterion value than basic solution. But the search time for any of improved solutions has increased several times. Increasing computation time was caused by increasing dimension of solved problems at the algorithm work.

Note that the proposed algorithm can potentially be used not only for strategic but also operational planning. Operational planning is possible in situations with fewer transports/fewer multigraph vertices than those considered in this example [11]. The question of the maximum dimension of the problem being solved, at which operational planning is possible using the developed algorithm, is of separate scientific interest. It must be said that it is possible to speed up the work of the proposed algorithm with a new/different version of the mixed integer linear programming problem solver.

All results were obtained using ILOG CPLEX 12.5.1 mathematical package on the personal computer (Intel Core i5 4690, 3.5 GHz, 8 GB DDR3 RAM).

6. CONCLUSION

In this paper we have studied the problem of cargoes transportation scheduling in the transport network represented by the undirected multigraph. Transportation between vertices were carried out at predetermined time intervals. To solve this problem the mathematical model of carriages along a multigraph was proposed. This model was constructed using linear equalities and inequalities containing binary and continuous variables. The optimization criterion was formulated. The algorithm to find an approximate solution was proposed due to possible high dimension of the obtained problem. The algorithm is based on the decomposition of the cargoes set and the planning horizon. Additionally, a parameter is introduced into the algorithm for the acceleration of its work. This parameter controls the number of transportations on which the timetable is built at one or another step of the algorithm. A study of the quality of decomposition was carried out on a meaningful example with millions of binary variables.

FUNDING

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

REFERENCES

1. Archetti, C., Sperenza, G., and Vigo, D., Vehicle routing problems with profits, in *Vehicle Routing: Problems, Methods, and Applications*, Toth, P. and Vigo, D., Eds., 2nd ed., 2014, pp. 273–297.
2. Cacchiani, V., Caprara, A., and Toth, P., A column generation approach to train timetabling on a corridor, *4OR*, 2008, vol. 6, no. 2, pp. 125–142.
3. Gao, Yu., Kroon, L., et. al., Three-stage optimization method for the problem of scheduling additional trains on a high-speed rail corridor, *Omega*, 2018, vol. 80, pp. 175–191.
4. Mu, S. and Dessouky, M., Scheduling freight trains traveling on complex networks, *Transport. Res. Part B: Methodological*, 2011, vol. 45, no. 7, pp. 1103–1123.
5. Forsgren, M., Aronsson, M., and Gestrelius, S., Maintaining tracks and traffic flow at the same time, *J. Rail Transport Planning & Management*, 2013, vol. 3, no. 3, pp. 111–123.
6. Meng, L. and Zhou, X., Simultaneous train rerouting and rescheduling on an N-track network: A model reformulation with network-based cumulative flow variables, *Transportation Research Part B: Methodological*, 2014, vol. 67, pp. 208–234.
7. Cacchiani, V., Caprara, A., and Toth, P., Scheduling extra freight trains on railway networks, *Transport. Res. Part B: Methodological*, 2010, vol. 44, no. 2, pp. 215–231.
8. Lazarev, A.A. and Musatova, E.G., The problem of trains formation and scheduling: Integer statements, *Autom. Remote Control*, 2013, vol. 74, no. 12, pp. 2064–2068.
9. Gainanov, D.N., Ignatov, A.N., et al., On track procession assignment problem at the railway network sections, *Autom. Remote Control*, 2020, vol. 81, no. 6, pp. 967–977.
10. Ignatov, A.N., On the scheduling problem of cargo transportation on a railway network segment and algorithms for its solution, *Bul. of the South Ural State Univ. Ser. Mat. Model. Progr.*, 2021, vol. 14, no. 3, pp. 61–76.
11. Ignatov, A.N., On the general problem statement of cargo carriages scheduling and ways to solve it, *Autom. Remote Control*, 2023, vol. 84, no. 4, pp. 496–510.
12. Bosov, A.V., Ignatov, A.N., and Naumov, A.V., Algorithms for an approximate solution of the track possession problem on the railway network segment, *Informatika i ee primeneniya*, 2021, vol. 15, no. 4, pp. 3–11.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 9, pp. 1128–1139.

Minimizing the Total Weighted Duration of Courses in a Single Machine Problem with Precedence Constraints

E. G. Musatova[*,a] **and A. A Lazarev**[*,b]

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

Received February 8, 2023
Revised June 20, 2023
Accepted July 20, 2023

Abstract—A single machine scheduling problem with a given partial order of jobs is considered. There are subsets of jobs called courses. It is necessary to schedule jobs in such a way that the total weighted duration of all courses is minimal. We consider the case when the initial job and the final one of each course are uniquely determined. The NP-hardness of the problem under consideration is proved. We propose an algorithm for solving the problem, the complexity of which depends polynomially on the total number of jobs, but exponentially on the number of courses, which makes it possible to use it efficiently with a fixed small number of courses and an arbitrary number of jobs.

Keywords: scheduling theory, single machine problem, NP-hard problems, downtime of resources minimization

DOI: 10.25728/arcRAS.2023.90.12.001

1. INTRODUCTION

We consider a set of jobs that need to be executed on one machine and a precedence graph that sets a partial order of jobs. Some of the jobs are combined into subsets, which we will call courses. It is necessary to build a schedule in which the total weighted duration of all courses is minimal. The duration of a course is the length of the time interval between the start of processing the initial job from this course and the end of processing its final one. The article considers the case when the initial and the final jobs of each course are uniquely determined.

Single machine problems are comprehensively investigated in scheduling theory [1, 2]. At the same time, the single machine problem of minimizing the weighted total duration of courses has not been considered before.

The need to minimize the total duration of courses arises in different areas of production, education and services. In [3], a resource-constrained project scheduling problem with such objective function is considered in relation to constructing a schedule for preparing cosmonauts to work at the International Space Station. It is necessary to minimize the length of each course (or an on-board system in the terminology of the Gagarin Cosmonaut Training Centre). If too much time passes from the beginning of a course to the exam, the cosmonauts' skills are considered lost and they have to add additional hours to the preparatory process, which leads to large time and financial losses. In that publication, a heuristic algorithm for solving the problem is proposed.

We can also interpret such a problem as a problem of minimizing the total downtime of resources. Suppose that all jobs of each course requires their own specific resource (for example, processing on additional equipment). This resource is taken on a temporary lease, which begins to be paid

simultaneously with the start of the first job from the course and ends with the completion of the last job from this course. Then the duration of a course can be associated with the total payment for the resource, and the objective function characterizes the total payments for all leased resources. In addition to payments for additional resources, this goal function can be considered as a fee for storage and rental of premises.

For the first time, the question of the duration of courses was considered in [4]. Here, along with the usual concept of "activity", the concept of "hammock activity" is introduced. The duration of a hammock activity is determined by the beginning and the end of some fixed activities. This article discusses project scheduling without resource constraints and provides methods for calculating the duration of hammock activities. Note that in the case when the first and the last jobs of a course are uniquely defined, the concepts of hammock activity and course coincide.

The concept of hammock activity was developed in [5]. The authors of that article consider the problem of minimizing the total cost of several hammock activities in a project both in the presence of resource constraints (Resource-Constrained Hammock Cost Problem, RCHCP) and without them. The cost of a hammock activity means its weighted duration. In the absence of resource constraints, the problem is reduced to a linear programming problem. In case of resource constraints, the formulation of the problem in the form of a mixed integer linear programming problem is proposed. In the dissertation [6] the research of the RCHCP is continued. The author suggests metaheuristics for solving this problem, and also provides an extensive review of publications on problems of the RCHCP type.

Some studies use terms other than "hammock activity" to describe a similar objective function. So in theory of scheduling repetitive jobs (see, for example, [7]), such problems are known as project scheduling problems with work continuity constraints. For example, in [8] the following project scheduling problem with repetitive jobs is considered. There is some basic precedence graph, which is duplicated k times. Some of the repetitive jobs requires additional resources (equipment, teams of workers, etc.), and it is necessary to complete the project by the specified directive deadline with minimizing the duration of these repetitive jobs. Examples of practical applications are given for construction of multi-storey buildings, where identical jobs are performed on each storey, construction of bridges, roads, etc. In [9] the terms used are "minimizing crew idle time" and "minimizing resource idle time". The authors describe a practical use of algorithms developed for such a problem during the construction of the Westerschelde Tunnel in the Netherlands. Crews of workers and freezing machines were selected as resources whose total duration of use had to be minimized.

Thus, if we talk about minimizing the total duration of courses, the major attention in the literature is paid to project scheduling problems either with or without resource constraints. In the first case, we have to deal with an NP-hard problem [5] and the emphasis in such studies is on the development of heuristic algorithms, whereas in the second case polynomial algorithms are built.

Our article discusses a single machine problem that can be interpreted as a project scheduling problem with single resource available in the amount of one unit at any given time, provided that each job also requires one unit of the resource. It is shown that this problem is NP-hard. We propose an algorithm which allows to find an exact solution in the case of a large number of jobs, but a small fixed number of courses. Section 2 provides a formulation of the problem. Section 3 proves NP-hardness of the problem under consideration, as well as some of its properties. Section 4 is devoted to solving an auxiliary problem, and in section 5 an algorithm for solving the original problem based on solving an auxiliary problem is described and the results of a numerical experiment are presented.

2. PROBLEM STATEMENT

There is a set of jobs $I = \{1, \ldots, n\}$ that need to be executed on one machine. For each job $i \in I$, its processing time is equal to $p_i > 0$. All jobs are available at zero time. The processing of any job cannot be interrupted.

A directed acyclic precedence graph $G(I, E)$ is given, where I is a set of vertices, and E is a set of arcs. We say that for a pair of jobs $i, j \in I$, job i precedes job j, denoting by $i \to j$, if there exists a directed path from vertex i to vertex j in the graph $G(I, E)$. Denote by $A(i)$ the set of all jobs preceding job i and by $D(i)$ the set of jobs preceded by job i. Each job i must be executed after all jobs from set $A(i)$ and before all jobs from $D(i)$.

In addition, there are sets $I_k \subset I$, $|I_k| > 1$, $k \in \{1, \ldots, K\}$, called *courses*. Figure 1 gives an example of a precedence graph for a problem with three courses. Each course I_k, $k \in \{1, \ldots, K\}$, has its own weight $w_k > 0$. Depending on the interpretation, the weight is either a price of a leased resource per unit of time or an index of importance (significance) of the course. For each job $i \in I$, a schedule π determines its processing sequence number on the machine, which we will denote by $\pi(i)$, a start time of processing $S_i(\pi)$ and a completion time of processing $C_i(\pi) = S_i(\pi) + p_i$. We will call a schedule feasible if it does not contradict the precedence relations of jobs and the machine does not serve more than one job at any given time. The problem of minimizing the total weighted duration of all courses implies minimizing the following objective function:

$$H(\pi) = \sum_{k=1}^{K} w_k \left(\max_{i \in I_k} C_i(\pi) - \min_{i \in I_k} S_i(\pi) \right). \tag{1}$$

The article considers the case when the first and the last jobs of each course are uniquely determined, i.e. the following assumption is true.

Assumption 1. For each course I_k, $k \in \{1, \ldots, K\}$, there are jobs i_k^a and i_k^d, such that $i_k^a \in A(j)$ for any $j \in I_k \setminus \{i_k^a\}$ and $i_k^d \in D(j)$ for any $j \in I_k \setminus \{i_k^d\}$.

This condition is often fulfilled in practice. For example, in an educational process, the first lesson is usually introductory, and the last one implies a general knowledge test, while the sequence of

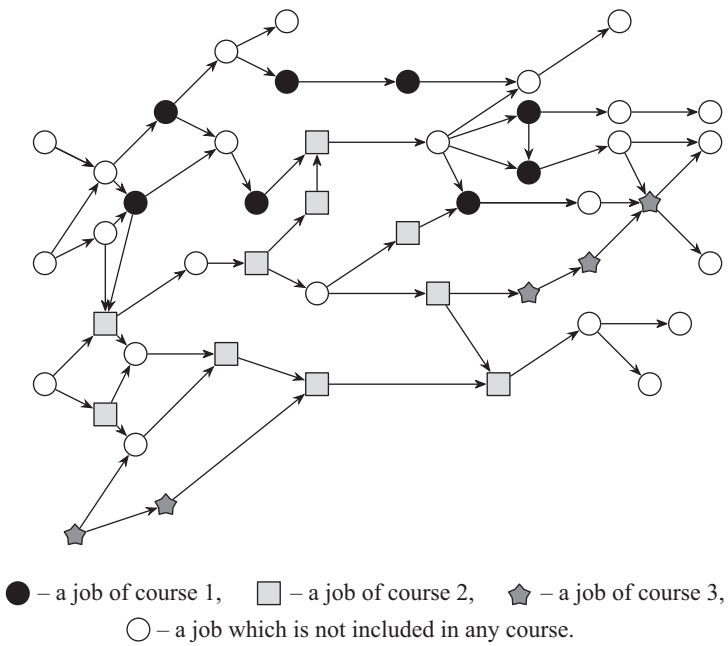


Fig. 1. A precedence graph and courses.

other classes in the course may vary. In this case, the objective function can be written as

$$H(\pi) = \sum_{k=1}^{K} w_k \Big(C_{i_k^d}(\pi) - S_{i_k^a}(\pi) \Big) = \sum_{k=1}^{K} w_k \Big(C_{i_k^d}(\pi) - C_{i_k^a}(\pi) + p_{i_k^a} \Big). \tag{2}$$

We will call i_k^a and i_k^d *extreme* jobs (vertices) of course k, $k \in \{1, \ldots, K\}$. The set of all extreme jobs of courses is denoted by I_{ad}, and their number—by e. Since some jobs from I_{ad} can be extreme in several courses, $e \leqslant 2K$.

The minimization of function (2) exactly coincides with the minimization of the hammock activities cost described in the introduction. In the standard scheduling theory notation [10] this problem can be classified as $1|prec|H$, where 1 means one machine, *prec*—the presence of precedence constraints, and H—the objective function (2).

3. PROBLEM PROPERTIES

Remark 1. Since all jobs in problem $1|prec|H$ are available at the same time and downtime does not improve the value of the objective function, we can only consider schedules without breaks between jobs, with the start of the first job at zero time. Indeed, let there exist an optimal schedule π_1, in which there are machine downtimes or the first job does not start at zero time. Then we can consider the schedule π_2, in which all jobs are performed in the same order as in π_1, but the first job starts from zero time and there are no breaks between jobs. Schedule π_2 is also optimal.

Let's show that even with the same processing time of all jobs, the problem under consideration is NP-hard.

Theorem 1. *Problem* $1|prec, p_i = 1|H$ *is strongly NP-hard.*

Proof. Let's consider the classical single machine problem of minimizing the weighted total flow time $1|prec, p_i = 1| \sum w_i C_i$. The problem is formulated as follows. One machine and a set of jobs $I' = \{1, \ldots, n'\}$ are given, each job i has weight w_i' and processing time $p_i' = 1$, $i \in \{1, \ldots, n'\}$. Let there also be given a directed precedence graph $G'(I', E')$. It is necessary to find a schedule π' that minimizes the objective function $\sum_{i=1}^{n'} w_c' C_i'(\pi')$, where $C_i'(\pi')$ is the completion time of the ith job in the schedule π'.

This problem is strongly NP-hard [11]. Let's reduce it to the following problem $1|prec, p_i = 1|H$. There is a set of jobs $I = \{1, \ldots, n\}$, $n = 2n'$. Each job i has processing time $p_i = 1$. Graph $G(I, E)$ has the following structure. There are $|E'|$ arcs defined by the following rule: if in graph $G'(I', E')$ there is an arc (j, k), then in graph $G(I, E)$ there is also arc (j, k). In addition, there are $n' - 1$ arcs of the form $(i, i+1)$ for $i \in \{n'+1, \ldots, 2n'-1\}$ and $|L|$ arcs of the form $(2n', l)$, where L is the set of root vertices (sources) in graph G', $l \in L$. The structure of graph $G(I, E)$ is shown in Fig. 2.

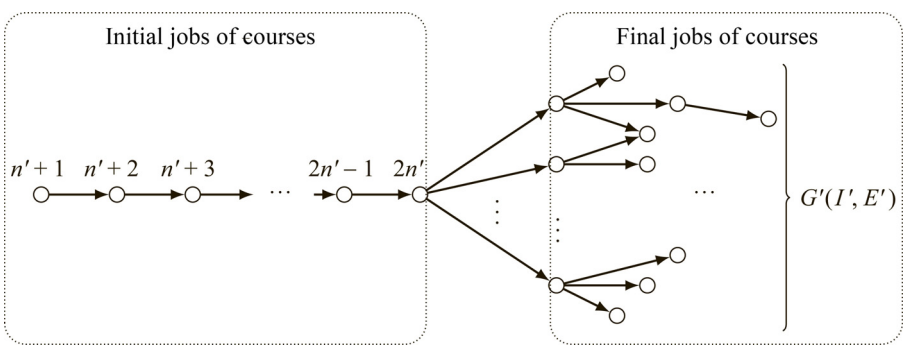


Fig. 2. The structure of graph $G(I, E)$ from the proof of theorem 1.

Let's set the courses as follows: jobs $n'+1$ and 1 belong to the first course, which has weight w'_1, jobs $n'+2$ and 2 belong to the second course, which has weight $w'_2, \ldots,$ jobs $2n'$ and n' refer to the n'th course, which has weight $w'_{n'}$, i.e. each course consists of two jobs, the first of which lies in set $\{n'+1, \ldots, 2n'\}$, and the final—in set I'. Let π be an arbitrary optimal schedule for this problem. The value of the objective function is

$$H(\pi) = \sum_{i=1}^{n'} w'_i \left(C_i(\pi) - C_{i+n'}(\pi) + 1\right) = \sum_{i=1}^{n'} w'_i C_i(\pi) - \sum_{i=1}^{n'} w'_i C_{i+n'}(\pi) + \sum_{i=1}^{n'} w'_i. \tag{3}$$

Due to the structure of the precedence graph, job $n'+1$ will be processed first. Then taking into account Remark 1 we have $C_{n'+1}(\pi) = 1$. Since there are arcs $(i, i+1)$ in the graph G for $i \in \{n'+1, \ldots, 2n'-1\}$, and all jobs from $\{1, \ldots, n'\}$ are processed after job $2n'$, the order of jobs $n'+2, \ldots, 2n'$ is known, moreover,

$$C_i(\pi) = i - n', \quad i \in \{n'+2, \ldots, 2n'\}, \tag{4}$$

that is, all these jobs are processed one after another without breaks. Thus, in schedule π, the execution of jobs $n'+1, \ldots, 2n'$ is predetermined and will end at time n'. But then, taking into account (3), schedule π is optimal if and only if (4) is executed and the minimum of the function $\sum_{i=1}^{n'} w'_i C_i(\pi)$ is reached with schedule π, i.e. when in the schedule π jobs $1, \ldots, n'$ are executed starting from the moment n' in a way minimizing their weighted total flow time. As a result, the optimal schedule π' in the problem $1|pres, p_i = 1|\sum w_i C_i$ can be obtained from the optimal schedule π of the described problem $1|prec, p_i = 1|H$. Completion times of the jobs in the problem $1|prec, p_i = 1|\sum w_i C_i$ are as follows:

$$C'_i(\pi') = C_i(\pi) - n', \quad i \in \{1, \ldots, n'\}.$$

The theorem is proven.

Remark 2. Similarly, it is possible to prove NP-hardness in the strong sense of the one-machine problem of minimizing the total (unweighted) duration of courses with different jobs processing times, using a reduction from the NP-hard problem $1|prec|\sum C_i$ to it.

As noted earlier, due to Remark 1 next, we will consider only schedules without breaks between jobs, with the start of the first job at time zero. In this case, for each job $j \in I$ its sequence number $\pi(j)$ in the schedule π uniquely sets the start time and the end of the job. Note also that only completion times of extreme jobs of courses are included in the definition of the objective function (2), moreover, only differences, and not absolute values, are decisive. Since in the absence of breaks the duration of a course is determined by the jobs started after the first job of the course and completed before the completion of the last job of the course, let's rewrite objective function (2) in a different form, without using the job completion times :

$$H(\pi) = \sum_{k=1}^{K} w_k \left(\sum_{j:\pi(i_k^a) \leqslant \pi(j) \leqslant \pi(i_k^d)} p_j \right). \tag{5}$$

Then we can write

$$H(\pi) = \sum_{j=1}^{n} W_j(\pi) p_j, \tag{6}$$

where

$$W_j(\pi) = \sum_{\substack{k \in K: \\ \pi(j) \geqslant \pi(i_k^a)}} w_k - \sum_{\substack{k \in K: \\ \pi(j) > \pi(i_k^d)}} w_k. \tag{7}$$

So the contribution $W_j(\pi)$ of each job $j \in I$ to the objective function depends on the mutual order of extreme jobs of courses and its place among the extreme jobs. In this regard, the following idea of solving the problem arises: for each feasible permutation of extreme jobs of courses, it is necessary to find an optimal order of jobs relative to the extreme jobs. The next section will present a polynomial algorithm for constructing an optimal schedule for a given order of extreme jobs of courses.

4. SOLVING AN AUXILIARY PROBLEM

As it was shown in the previous section, the value of the objective function of the initial problem depends on the mutual order of extreme jobs of courses. Denote by $\Lambda = (\lambda_1, \ldots, \lambda_e)$ an arbitrary permutation of extreme jobs that does not contradict the precedence relations given by graph $G(I, E)$, and introduce a directed acyclic graph $G'(I, E')$ such that $E \subset E'$ and for the set of extreme jobs, the following condition is fulfilled:

$$\lambda_1 \to \lambda_2 \to \ldots \lambda_{e-1} \to \lambda_e. \tag{8}$$

The graph $G'(I, E')$ is obtained from $G(I, E)$ by sequentially connecting extreme vertices $\lambda_1, \ldots, \lambda_e$ by arcs in accordance with the order given by Λ. If such order does not contradict the precedence constraints of the problem, the resulting graph $G'(I, E')$ will be acyclic. Due to the acyclicity of the original graph $G(I, E)$, there is always at least one sequence of extreme jobs Λ that does not contradict the precedence constraints. Then in any schedule π for the graph $G'(I, E')$ we will have

$$\pi(\lambda_1) < \pi(\lambda_2) < \ldots < \pi(\lambda_{e-1}) < \pi(\lambda_e).$$

The problem is to minimize function (6)–(7) relative to the new graph $G'(I, E')$. We will denote the auxiliary problem by P_Λ.

For all jobs that are not extreme, it is necessary to determine their places in the sequence $\lambda_1, \lambda_2, \ldots, \lambda_e$. For each job, there are no more than $e + 1$ options (the job is processed before λ_1, between λ_1 and λ_2, etc.). Say that job j is placed in *cell* q, $q \in \{1, \ldots, e-1\}$ if it is executed after extreme job λ_q and before extreme job λ_{q+1}. Assume that $q = 0$ if job j is executed before λ_1, and $q = e$ if j is executed after λ_e.

Consider for each extreme job $\lambda_i \in I_{ad}$ the sets $A(\lambda_i)$ and $D(\lambda_i)$ in the graph $G'(I, E')$. Note several obvious statements that will be used later and the proof of which follows directly from (8).

Lemma 1. *a) If $j \in A(\lambda_i)$, then $j \in A(\lambda_k)$ for all $k \geqslant i$.*

b) If $j \in D(\lambda_i)$, then $j \in D(\lambda_k)$ for all $k \leqslant i$.

c) If $j \notin D(\lambda_i)$, then $j \notin D(\lambda_k)$ for all $k \geqslant i$.

d) If $j \notin A(\lambda_i)$, then $j \notin A(\lambda_k)$ for all $k \leqslant i$.

To determine boundaries of the possible location of the non-extreme jobs by cells in the row of extreme jobs, we introduce the following notation:

$$q_1(j) = \begin{cases} 0, & \text{if } j \notin D(\lambda_1), \\ \max\{g \in \{1, \ldots, e\} \mid j \in D(\lambda_g)\}, & \text{otherwise;} \end{cases}$$

$$q_2(j) = \begin{cases} e, & \text{if } j \notin A(\lambda_e), \\ \min\{g \in \{1, \ldots, e\} \mid j \in A(\lambda_g)\} - 1, & \text{otherwise.} \end{cases}$$

Lemma 2. *For each job $j \in I \setminus I_{ad}$ the inequality $q_1(j) \leqslant q_2(j)$ is satisfied.*

Proof. If either $j \notin D(\lambda_1)$ or $j \notin A(\lambda_e)$, the statement is obvious. For a proof in the other cases, we assume the opposite. Let $q_1(j) > q_2(j)$. By definition of $q_1(j)$ we have $j \in D(\lambda_{q_1(j)})$. On the other hand, by definition of $q_2(j)$ we have $j \in A(\lambda_{q_2(j)+1})$. But then by lemma 1 we get $j \in A(\lambda_k)$ for all $k \geqslant q_2(j) + 1$, which means $j \in A(\lambda_{q_1(j)})$. The resulting contradiction proves the lemma.

Lemma 3. *If in a feasible solution of problem P_λ job j is placed into cell q, then $q_1(j) \leqslant q \leqslant q_2(j)$.*

Proof. Assume the opposite. Let's suppose that in some feasible schedule job j is placed into cell q, for which either $q < q_1(j)$ or $q > q_2(j)$ is holds. Let $q < q_1(j)$. Then $q_1(j) > 0$ and by definition $j \in D(\lambda_{q_1(j)})$, which means that job j cannot be executed before $\lambda_{q_1(j)}$, which contradicts the choice of cell q. Let $q > q_2(j)$. Then $q_2(j) < e$ and by definition $j \in A(\lambda_{q_2(j)+1})$, which means that the job cannot be executed after $\lambda_{q_2(j)+1}$, which contradicts the choice of cell q. The lemma is proven.

For each cell $q \in \{0, \ldots, e\}$ let's introduce its price $f(q)$ according to the following rule:

$$f(q) = \sum_{\substack{k \in K: \\ x(i_k^a) \leqslant q}} w_k - \sum_{\substack{k \in K: \\ x(i_k^d) \leqslant q}} w_k,$$

where $x(i_k^a)$ and $x(i_k^d)$ are the numbers of extreme jobs of course k in permutation $\Lambda = (\lambda_1, \ldots, \lambda_e)$. This value determines the "contribution" of a job in the original objective function (6) if this job is placed into cell q. Denote by $q^*(j)$ the first number from $q_1(j)$ to $q_2(j)$ for which the minimum of f is reached:

$$q^*(j) = \min \left\{ t \mid f(t) = \min_{q_1(j) \leqslant q \leqslant q_2(j)} f(q) \right\}. \tag{9}$$

We will call $q^*(j)$ the optimal cell for job j, $j \in I$. The following lemma shows that if one job should precede another one in a schedule, then its optimal cell is not greater than the optimal cell for another job.

Lemma 4. *If $j \to g$ in graph $G'(I, E')$ for two non-extreme jobs j and g, then*

a) $q_1(j) \leqslant q_1(g)$;

b) $q_2(j) \leqslant q_2(g)$;

c) $q^*(j) \leqslant q^*(g)$.

Proof. a) Since $j \to g$, then $g \in D(j)$. If $q_1(j) = 0$, then the statement is obvious. If $q_1(j) > 0$, then $j \in D(\lambda_{q_1(j)})$. It means that $g \in D(\lambda_{q_1(j)})$. Then by definition of $q_1(g)$ we get $q_1(g) \geqslant q_1(j)$.

b) Since $j \to g$, then $j \in A(g)$. If $q_2(g) = e$, then the statement is obvious. If $q_2(g) < e$, then $g \in A(\lambda_{q_2(g)})$. It means that $j \in A(\lambda_{q_2(g)})$. Then by definition of $q_2(j)$ we get $q_2(j) \leqslant q_2(g)$.

c) Let $q^*(j) > q^*(g)$. Taking into account a) and b), we obtain

$$q_1(j) \leqslant q_1(g) \leqslant q^*(g) < q^*(j) \leqslant q_2(j) \leqslant q_2(g).$$

This means that both cells $q^*(j)$ and $q^*(g)$ are available for jobs j and g. This contradicts cell selection rule (9). Indeed, if $f(q^*(j)) = f(q^*(g))$, then cell $q^*(g)$ should be selected for both jobs as the earlier one. If $f(q^*(j)) \neq f(q^*(g))$, then the cell with the minimum value of f should be selected for both jobs. The lemma is proven.

For each cell $q \in \{0, \ldots, e\}$ we introduce a set of jobs I_q for which this cell is optimal:

$$I_q = \{j \in I \setminus I_{ad} \ : \ q^*(j) = q\}, \ q \in \{0, \ldots, e\}.$$

Let $E_q \subset E$ be the set of arcs connecting the vertices of I_q, $q \in \{0, \ldots, e\}$. Denote by $\bar{\pi}(I_q, E_q)$ an arbitrary topological sorting of the graph $G_q(I_q, E_q)$, i.e. some permutation of jobs from I_q satisfying the partial order given by the set of arcs E_q. Due to acyclicity of the original graph $G(I, E)$, a topological sorting of any of its subgraphs $G_q(I_q, E_q)$ exists. The following theorem shows that, by ordering jobs in each cell separately, we can get an optimal schedule for problem P_Λ as follows: first we need to complete all jobs from set I_0, then complete the extreme job λ_1, then—all jobs from set I_1, the extreme job λ_2, etc.

Theorem 2. *The schedule* $\pi^\Lambda = (\bar\pi(I_0, E_0), \lambda_1, \bar\pi(I_1, E_1), \lambda_2, \ldots, \lambda_e, \bar\pi(I_e, E_e))$ *is an optimal solution of problem* P_Λ.

Proof. The schedule π^Λ is feasible in problem P_Λ. Indeed, consider any two jobs $i, j \in I \setminus I_{ad}$ such that $i \to j$. If these jobs are in the same cell, then the precedence constraint is satisfied due to the construction of the topological sorting of all jobs from this cell. If i and j are in different cells, then by virtue of lemma 4 we have $q^*(i) < q^*(j)$, which means that in schedule π^Λ job i will be executed before job j. The precedence constraints between extreme jobs and all other jobs are satisfied by virtue of the rule of constructing optimal cells.

The optimality of the solution follows from the definition of an optimal cell. Indeed,

$$\sum_{j \in I\setminus I_{ad}} f(q^*(j))p_j = \sum_{j \in I\setminus I_{ad}} \min_{q_1(j)\leqslant q\leqslant q_2(j)} \left(\sum_{\substack{k \in K:\\ x(i_k^a)\leqslant q}} w_k - \sum_{\substack{k \in K:\\ x(i_k^d)\leqslant q}} w_k \right) p_j = \min_{\pi} \sum_{j \in I\setminus I_{ad}} W_j(\pi)p_j.$$

Since for the given order Λ, the contribution to the objective function of extreme jobs is fixed and equal to

$$\sum_{j \in I_{ad}} \left(\sum_{\substack{k \in K:\\ x(j)\geqslant x(i_k^a)}} w_k - \sum_{\substack{k \in K:\\ x(j)> x(i_k^d)}} w_k \right) p_j,$$

this means that the schedule π^Λ, which corresponds to the distribution of jobs across cells, delivers the minimum of objective function (6)–(7). The theorem is proven.

Thus, solving problem P_Λ can be reduced to calculating the optimal cell for each job and ordering jobs in each cell separately. A general scheme of finding a solution to the auxiliary problem is described by Algorithm 1. Let's evaluate the complexity of this approach. It is necessary to construct sets $A(\lambda)$, $D(\lambda)$ for each extreme job λ, which in total will require $O(nK)$ operations. Next, for each job j, it is necessary to define the boundaries $q_1(j)$, $q_2(j)$ and the optimal cell $q^*(j)$ that will required $O(nK)$ operations. Building partial schedules in each cell needs no more than $O(n + |E|)$ operations [12].

Algorithm 1 Procedure $Solv(\Lambda)$

$\pi^\Lambda := ()$
for all $q \in \{0, 1, \ldots, e\}$ **do**
    $I_q := \emptyset$
end for
Generate graph $G'(I, E')$ by permutation $\Lambda = (\lambda_1, \ldots, \lambda_e)$
for all $j \in I \setminus I_{ad}$ **do**
    Calculate $q^*(j)$
    $I_{q^*(j)} := I_{q^*(j)} \cup \{j\}$
end for
Build $\bar\pi(I_0, E_0)$
$\pi^\Lambda := \bar\pi(I_0, E_0)$
for all $q \in \{1, \ldots, e\}$ **do**
    Build $\bar\pi(I_q, E_q)$
    $\pi^\Lambda := \pi^\Lambda \cup (\lambda_q, \bar\pi(I_q, E_q))$
end for
Return π^Λ

5. AN ALGORITHM FOR SOLVING PROBLEM $1|prec|H$

Denote by B the set of all possible permutations of extreme jobs Λ that do not contradict the precedence constraints of the original problem. If the number of courses in the problem is small, or due to the structure of graph $G(I, E)$, the mutual order of extreme jobs does not allow a large number of options, an efficient search for a solution to the problem is possible. It is based on iterating through permutations of extreme jobs and solving an auxiliary problem for each permutation. Thus, the scheme of solving the problem can be represented as Algorithm 2, where H^*_Λ is the optimal value of the objective function in the auxiliary problem P_Λ, and H^*, π^* are the optimal value and optimal the schedule of the original problem $1|prec|H$, respectively. Note that in Algorithm 2 there is no need to find a schedule for each permutation Λ under consideration, since to calculate the value of H^*_Λ, it is enough to know the optimal cells for each job.

Algorithm 2 Solving problem $1|prec|H$

```
H* := +∞
for all Λ ∈ B do
    Calculate H*_Λ
    if H*_Λ < H* then
        H* := H*_Λ
        Λ* := Λ
    end if
end for
π* := Solv(Λ*)
Return π*
```

The algorithm for solving the problem has the complexity $O(|B|(nK+|E|))$, where n is the total number of jobs, K is the number of courses, $|E|$ is the number of edges in the precedence graph and $|B|$ is the number of feasible permutations of the extreme jobs of the courses. The largest contribution to the complexity is given by the value $|B|$. The maximum possible value of $|B|$ is $\frac{(2K)!}{2^K}$, when all possible permutations of extreme jobs of courses are considered without taking into account their precedence relations. However, in the case of, for example, dense precedence graphs, the value $|B|$ may be acceptable for using Algorithm 2 even with a large number of courses.

During computational experiments, the proposed algorithm was compared with the optimization solver IBM ILOG CPLEX 22.1.0.0 [13]. To apply this solver, the following formulation of the problem was used in the form of an integer linear programming problem:

$$\sum_{k\in K}\sum_{j\in I, j\neq i^a_k} w_k p_j x_{i^a_k,j} + \sum_{k\in K}\sum_{j\in I, j\neq i^d_k} w_k p_j x_{j,i^d_k}$$
$$+ \sum_{k\in K}\left(p_{i^a_k} + p_{i^d_k} - \sum_{i\in I} p_i\right) \to \min,$$
$$x_{i,j} + x_{j,i} = 1 \ \forall i, j \in I;$$
$$x_{i,j} + x_{j,k} + x_{k,i} \geqslant 1 \ \forall i, j, k \in I;$$
$$x_{i,j} = 1 \ \forall (i, j) \in E;$$
$$x_{i,j} \in \{0, 1\} \ \forall i, j \in I,$$

where variable $x_{i,j}$, $i \neq j \in I$, takes the value 1 if job i is executed before job j, and the value 0 otherwise. Such variables and constraints are standard for integer formulations of single machine problems with precedence constraints (see, for example, [14]).

Table 1. Test results for sparse graphs

n	K	$\|E\|$	$\|B\|$	Algorithm 2	CPLEX
100	3	481	6	0.007	13.218
	5	467	4299	5.124	13.436
	7	525	26 244	35.191	12.774
200	3	1864	6	0.034	106.282
	5	1942	150	0.516	105.152
	7	1990	870	3.169	102.967
300	3	4550	5	0.050	426.563
	5	4423	10	0.069	404.386
	7	4410	950	6.834	396.179
400	3	7765	1	0.022	>10 min
	5	7662	4	0.051	>10 min
	7	7746	280	3.362	>10 min
500	3	12 241	2	0.037	>10 min
	5	12 118	2	0.065	>10 min
	7	12 194	96	1.854	>10 min

Table 2. Test results for dense graphs

n	K	$\|E\|$	—B—	Algorithm 2	CPLEX
100	3	2514	2	0.009	11.144
	5	2541	4	0.054	10.717
	7	2515	3	0.025	10.777
	9	2487	8	0.048	10.668
200	3	10 103	1	0.028	94.987
	5	10 194	2	0.071	95.790
	7	10 199	2	0.043	99.734
	9	10 123	4	0.165	94.775
300	3	22 846	1	0.034	386.899
	5	22 943	1	0.039	398.339
	7	22 933	6	0.225	378.841
	9	22 845	4	0.156	400.153
400	3	40 862	1	0.062	¿10 min
	5	40 692	2	0.135	¿10 min
	7	40 779	1	0.094	¿10 min
	9	40 806	2	0.147	¿10 min
500	3	63 657	1	0.090	¿10 min
	5	63 424	1	0.098	¿10 min
	7	63 676	1	0.136	¿10 min
	9	63 802	3	0.316	¿10 min

The calculations were performed on a personal computer (Intel Core i7-7700K, 4.2 GHz, 32.0 GB), the algorithm was implemented in Python using NetworkX library for working with graphs. In Tables 1 and 2 the results of solving randomly generated problems are given. Random integers from range [1; 10] were chosen for weights of courses and processing times of jobs. Random vertices of graphs were chosen as extreme jobs of courses in such a way that precedence constraints between the first and the last vertices of each course were not violated. The running time of the algorithm and the CPLEX solver was limited to 10 minutes.

Notations n, K, $|E|$, $|B|$ used in the tables coincide with the notations adopted earlier in the article, and columns “Algorithm 2” and “CPLEX” indicate the time of solving problems in seconds by the algorithm proposed in the article and by CPLEX, respectively.

We can see from Table 1 that running time of the algorithm depends on the cardinality of set B more than on the total number of jobs. So, the algorithm gives an exact solution to the problems of high dimension in a fraction of a second, if the number of feasible permutations of extreme jobs is small. In the case of large value of $|B|$ (see, for example, the problem with $n = 100$, $K = 7$), the running time of the algorithm increases significantly. CPLEX, on the contrary, is insensitive to changes of $|B|$ and K, but with an increase of n, its running time increases greatly.

In Table 2 results of solving problems with denser graphs are presented. Here, as expected, the number of feasible permutations of extreme jobs is less, so the algorithm found solutions in all problems in less than a second.

Thus, the results of the computational experiment confirm the theoretical estimation of the complexity of the developed algorithm and show that the algorithm can be efficiently applied to problems with a small set B, which corresponds to the case of either problems with dense precedence graphs, or problems with a small number of courses, or problems in which the positions of extreme jobs are fixed relative to each other. In these cases, the algorithm allows us to quickly solve high-dimensional problems.

6. CONCLUSION

The article considers the single machine problem with precedence constraints, in which it is necessary to minimize the total weighted duration of courses (some subsets of jobs). The NP-hardness of the problem under consideration is proved. An exact algorithm for its solving is proposed. This algorithm depends polynomially on the total number of jobs and allows solving problems efficiently, if there is a small number of options for the relative location of extreme jobs of courses. The direction of further research may concern a general formulation of the problem, when extreme jobs of courses are not clearly defined. Resources constrained project scheduling problem with the considered goal function can also be considered.

REFERENCES

1. Brucker, P., *Scheduling algorithms*, Springer: Heidelberg, 2007.
2. Lazarev, A.A., *Teoriya raspisanii. Metody i algoritmy* (Theory of Schedules. Methods and Algorithms), Moscow: ICS RAS, 2019.
3. Lazarev, A., Khusnullin, N., Musatova, E., Yadrentsev, D., Kharlamov, M., and Ponomarev K., Minimization of the weighted total sparsity of cosmonaut training courses, *OPTIMA 2018. Communications in Computer and Information Science*, 2019, pp. 202–215.
4. Harhalakis, G., Special features of precedence network charts, *Eur. J. Oper. Res.*, 1990, vol. 49, no. 1, pp. 50–59.
5. Csébfalvi, A.B. and Csébfalvi, G., Hammock activities in project scheduling, *Proceedings of the Sixteenth Annual Conference of POMS*, 2005.
6. Eliezer, O., A new bi-objective hybrid metaheuristic algorithm for the resource-constrained hammock cost problem (RCHCP), *Doctoral Dissertation*, Pécs, 2011.
7. El-Rayes, K. and Moselhi, O., Resource-driven scheduling of repetitive activities, *Construction Management and Economics*, 1998, vol. 16, no. 4, pp. 433–446.
8. Vanhoucke, M., Work continuity constraints in project scheduling, *Working Paper 04/265*, Belgium: Ghent University, Faculty of Economics and Business Administration, 2004.
9. Vanhoucke, M. and Van Osselaer, K., Work continuity in a real-life schedule: the Westerschelde Tunne, *Working Paper 04/271*, Belgium: Ghent University, Faculty of Economics and Business Administration, 2005.

10. Graham, R.L., Lawler, E.L., Lenstra, J.K., and Rinnooy Kan, A.H.G., Optimization and approximation in deterministic sequencing and scheduling: a survey, *Annals of Discrete Mathematics*, 1979, vol. 5, pp. 287–326.

11. Lenstra, J.K. and Rinnooy Kan, A.H.G., Complexity of scheduling under precedence constraints, *Oper. Res.*, 1978, vol. 26, no. 1, pp. 22–35.

12. Cormen, T.H., Leiserson, C.E., Rivest, R.L., and Stein, C., *Introduction to algorithms*, MIT press, 2022.

13. IBM ILOG CPLEX Optimization Studio,
URL: https://www.ibm.com/products/ilog-cplex-optimization-studio.

14. Potts, C.N., An algorithm for the single machine sequencing problem with precedence constraint, *Combinatorial Optimization II. Mathematical Programming Studies*, Springer: Berlin, Heidelberg, 1980, vol. 13, pp. 78–87.

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