Volume 84, Number 10
October 2023

ISSN 0005-1179
CODEN: AURCAT

A JOURNAL OF RUSSIAN ACADEMY OF SCIENCES

AUTOMATION AND REMOTE CONTROL

Automation and Remote Control

Editor-in-Chief
Andrey A. Galyaev

Vol. 84, No. 10, October 2023

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. 10, 2023

Topical Issue

Robust, Adaptive, and Network Control

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 10, pp. 1141–1142.

To the 110th Anniversary of the Birth of Boris N. Petrov, Vice President of the USSR Academy of Sciences

In 2023, the scientific community celebrates the 110th anniversary of the birth of Boris Petrov, a great scholar in automatic control, theoretician of rocket and space technology, science organizer, and Vice President of the USSR Academy of Sciences.

Boris Petrov, in full Boris Nikolaevich Petrov, was born on March 11, 1913, in Smolensk. Having graduated from a high school in 1930, he left for Moscow. After studying at a factory apprentice school, from October 1932 to September 1933, he worked as a turner. In 1933, Boris entered Moscow Power Engineering Institute (MPEI) at the Electromechanical Department. In 1939, he graduated with honors from MPEI. Petrov's graduation work entitled *Automatic Regulation of Boilers with Pulverized Coal Furnace* was written under the guidance of his teacher, Academician Victor S. Kulebakin. The work was recognized as outstanding. In 1939, by Kulebakin's suggestion, Boris entered the Commission for Automation and Remote Control, the USSR Academy of Sciences. The same year, based on the Commission, the Institute of Automation and Remote Control (IARC) was established. (Nowadays, it is known as the Trapeznikov Institute of Control Sciences, the Russian Academy of Sciences (ICS RAS).) Petrov worked at the Institute all his life and made a career from Junior Researcher to Director. In October 1940, Boris entered the postgraduate program of IARC; his scientific supervisor was Vadim A. Trapeznikov. During the Great Patriotic War, IARC was evacuated to Ulyanovsk, where Petrov actively studied the problem of automatic product rejection. In 1945, Boris submitted for defense his candidate's dissertation in engineering entitled *Analysis of Automatic Copying Systems* and was immediately awarded the higher degree of Dr. Sci. (Eng.). After the dissertation defense, he led active R&D and educational activities. Petrov early earned a great reputation among leading scientists and had outstanding organizational skills. It was noticed by the USSR Academy of Sciences: in 1947, on the recommendation of the Bureau of the Section of Engineering, he was appointed Acting Director of IARC. In 1949, Boris became Head of the Department of Aircraft Automatic Control Systems at Ordzhonikidze Moscow Aviation Institute (MAI). He headed the Department until the end of his life and trained many famous scientists and experts in aerospace technology.

Petrov's main scientific works were devoted to the theory of dynamic objects control, particularly the theory of invariance of automatic control systems, the theory of adaptive and terminal systems, nonlinear servomechanisms and variable structure systems, automatic control systems for aircraft and spacecraft, and the design of high-precision measuring devices.

Petrov's fruitful activities were highly appreciated in the USSR and abroad. He was entitled the Hero of Socialist Labor and was awarded the Lenin Prize and two State Prizes as well as many other domestic and foreign orders. Boris was Full member of the International Academy of Astronautics and Foreign Member of the Czechoslovak, Hungarian, Bulgarian, and Polish Academies of Sciences. The Lenin Prize (1966) was awarded for his participation in the design and manufacture of Voskhod-1 and Voskhod-2 multi-manned spacecraft, their launches, and the implementation of the world's first human walk in open space; for his participation in the design and manufacture of Luna-9 and Luna-10 automatic interplanetary stations, their launch, and soft landing on the surface of the Moon, the transmission of photographic data of the lunar panorama to Earth, and the injection of the world's first artificial satellite of the Moon into lunar orbit.

Petrov was an active organizer of IFAC International Symposiums on Automatic Control in the Peaceful Uses of Spaces, held in Norway (1965), Austria (1967), France (1970), Italy (1973), the USSR (1974), German Federal Republic (1975), and England (1979). From 1966 to 1980, he was Chairman of the Interkosmos Council for International Cooperation and Use of Outer Space. As Chairman of Interkosmos at the USSR Academy of Sciences, Academician Petrov took an active part in the organization and implementation of the Apollo–Soyuz Test Project, the joint experimental manned flight of Soyuz-19 spacecraft (the USSR) and Apollo spacecraft (USA). He was a leading scholar and an outstanding science organizer. Since 1963, Petrov was permanent Academician-Secretary of the Section of Mechanical Engineering and Control, the USSR Academy of Sciences; in 1979, he was elected Vice President of the USSR Academy of Sciences.

Petrov's entire scientific life was connected with IARC (ICS). Nowadays, the Institute develops the main modern theoretical branches in the control of space objects, aircraft, and dynamic objects that were initiated by him. They include the theory of terminal and adaptive control of space objects under normal and abnormal operating conditions with different levels of a priori and current information. In the 1970s, Petrov posed the problem of developing formal models and methods for designing information and control systems of spacecraft and their software. Based on a unified methodology, formalization methods and means as well as algorithms and programs were developed to design optimal modular real-time data processing systems. The theory of optimal control with a vector criterion was further developed to design algorithms for implementing the desired motion trajectories of dynamic objects. In addition to classical methods, the theory of anisotropic control and filtering for linear discrete-time stochastic systems is used to suppress the effect of exogenous disturbances on control systems. The method of spatial and angular relative positioning was proposed for the information support of aircraft control systems. It involves the parameters of the magnetic induction gradient as measuring information. This method is currently being developed further.

The thematic issue contains several papers presenting recent results in the theoretical branches mentioned above.

Glumov, V.M., Dr. Sci. (Eng.)

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 10, pp. 1143–1152.
© The Author(s), 2023 published by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 2023.
Russian Text © The Author(s), 2023, published in Avtomatika i Telemekhanika, 2023, No. 10, pp. 6–17.

Determination of the Relative Positionong Based on Magnetic Gradiometry Measurements

A. K. Volkovitsky*[,a], E. V. Karshakov*[,b], B. V. Pavlov*[,c], and E. A. Tretyakova*[,d]

[*] Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
e-mail: [a] avolkovitsky@yandex.ru, [b] karshakov@ipu.ru, [c] pavlov@ipu.ru, [d] ekaterina_tretikova@mail.ru

Received June 19, 2023

Revised July 17, 2023

Accepted August 2, 2023

Abstract—The paper is devoted to solving the problem of determining the relative spatial arrangement and orientation of objects. The task was set: to show the fundamental possibility of spatial and angular relative positioning when using the parameters of the magnetic field gradient in tensor form and in the form gradient of an absolute value vector as measurement information for the magnetic field of a local dipole source. The solution of the problem is presented along with the features and limitations for both forms of representation are considered. The principles of construction of magnetic gradiometry measurement systems are briefly described, the limitations of technical implementation are considered, and the benefits of using an alternating magnetic field source is outlined. The results of modelling are presented, proving the possibility of using the proposed positioning method for various engineering problems.

Keywords: relative positioning, magnetic field vector, magnetic field gradient, gradient tensor

DOI: 10.25728/arcRAS.2023.97.59.001

1. INTRODUCTION

The solution to many engineering problems is in one way or another connected with the need to determine the relative position of objects during their interaction. For example, high-precision control is necessary to control movement during mid-air refueling; mooring a vessel to a pier, loading terminal or drilling platform; landing an aircraft on a limited area, docking spacecraft and underwater vehicles [1–3]. Solving the relative positioning problem assumes that in the coordinate system associated with one of the objects, it is necessary to determine the radius vector of the location point of another object, as well as their mutual angular orientation. Usually, to solve such problems, gyroinertial systems, multi-antenna GNSS receivers, optical systems, etc. are used, but in many cases the technical solution is significantly complicated by the special aspects of the application conditions, thus becoming excessively large. In many cases, positioning accuracy is insufficient. The possibility of using magnetic gradient measurements provides possibility to develop relative positioning methods. The basis of the idea is the fact that the direction and absolute magnitude of the magnetic field vector of a point dipole transmitter at a certain point in space is fully determined by the magnitude and direction of the source dipole magnetic moment vector and the position of the measurement point. A type of the dependence of the field strength allows, from the data obtained by receiver of the local transmitter, to simultaneously determine their relative spatial and angular location.

2. PROBLEM STATEMENT

Let a local dipole magnetic field transmitter with an arbitrary direction of the dipole magnetic moment vector $\mathbf{M}$ be located at the origin and let the field be measured at an arbitrary point in space determined by the radius vector $\mathbf{r}$ in this coordinate system (Fig. 1).

For the magnetic field potential U^B of a local dipole transmitter in the associated coordinate system, the following relation is valid:

$$U^B = \frac{\mu\mu_0 \mathbf{r}^{\mathrm{T}}\mathbf{M}}{4\pi\left(\mathbf{r}^{\mathrm{T}}\mathbf{r}\right)^{3/2}}. \tag{1}$$

Here μ and μ_0 are the magnetic permeability of the medium and the magnetic constant in the international system of units (SI), respectively.

Differentiating (1), we obtain the values for the field vector:

$$\nabla U^B = \frac{\mu\mu_0\left|\mathbf{M}\right|}{4\pi|\mathbf{r}|^5}\begin{pmatrix} 3y_1^2 - |\mathbf{r}|^2 \\ 3y_1y_2 \\ 3y_1y_3 \end{pmatrix}, \tag{2}$$

and gradient tensor:

$$\mathbf{U} = \nabla\nabla^{\mathrm{T}}U^B = \frac{3\mu\mu_0\left|\mathbf{M}\right|}{4\pi|\mathbf{r}|^7}\begin{pmatrix} -2y_1^3 + 3y_1y_2^2 + 3y_1y_3^2 & -4y_1^2y_2 + y_2^3 + y_2y_3^2 & -4y_1^2y_3 + y_2^2y_3 + y_3^3 \\ -4y_1^2y_2 + y_2^3 + y_2y_3^2 & y_1^3 - 4y_1y_2^2 + y_1y_3^2 & -5y_1y_2y_3 \\ -4y_1^2y_3 + y_2^2y_3 + y_3^3 & -5y_1y_2y_3 & y_1^3 + y_1y_2^2 - 4y_1y_3^2 \end{pmatrix}. \tag{3}$$

The parameters of the magnetic field at the observation point are determined by a tensor gradiometric receiver, the coordinate system of which is oriented arbitrarily relative to the field source. Determining the parameters of the gradient tensor consists of measurement the field values at several points in space near the point with the radius vector $\mathbf{r}$ [5].

The problem is by knowing the magnitude and direction of the vector of the dipole magnetic moment of the field source in the coordinate system associated with it, and also having the results of parameter measurement of the field gradient tensor in the area where the observation point is located in the coordinate system of the receiver to determine the parameters of the radius vector between the source and the field receiver, as well as the direction of the vector of the dipole magnetic moment of the transmitter in the coordinate system of the receiver.

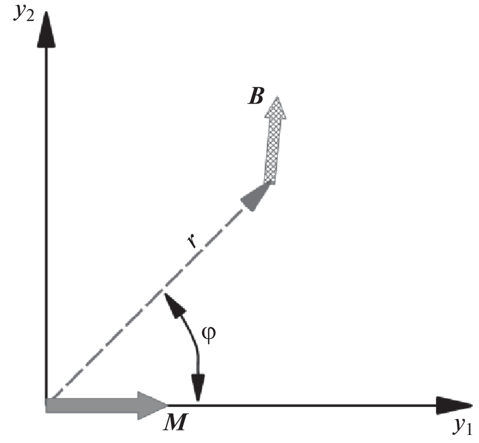


Fig. 1. The vector of the dipole magnetic moment and the vector of field in the coordinate system associated with the transmitter.

3. POSITIONING BASED ON MAGNETIC FIELD GRADIENT TENSOR MEASUREMENTS

The tensor parameters (3), which are important for solving the positioning problem, can be obtained by measurement the field at spatially separated points, but close enough as compared to the distance to the transmitter (so that one can limit ourselves to a linear approximation of the dependence of the field change on the distance).

Magnetic field potential is a harmonic function. Therefore, the tensor (3) is symmetric, and its trace is equal to zero. Thus, it contains not nine, but only five independent components. Moreover, by orthogonal transformations the coordinate system of the transemitter can be brought to the principal axes of the tensor. In this system, only its diagonal elements are nonzero.

The angular divergence α of the coordinate systems of the principal axes of the tensor y' and the y system is determined by the angle φ between the vector of the dipole magnetic moment and the radius vector $\mathbf{r}$ (Fig. 2). Knowing the angle φ between the radius vector $\mathbf{r}$ and the direction of the vector $\mathbf{M}$ from (3) it follows that the values of the angles α and φ are related to the relations of the main components of the tensor (Fig. 3). It also follows from (3) that when the coordinate

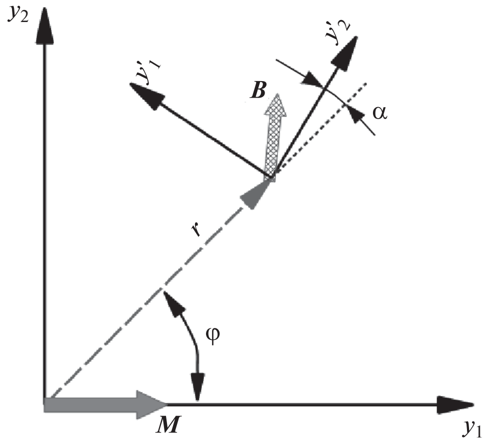


Fig. 2. To the parameters of the field gradient tensor of a point dipole: the principal axes of the tensor at the field measurement point.

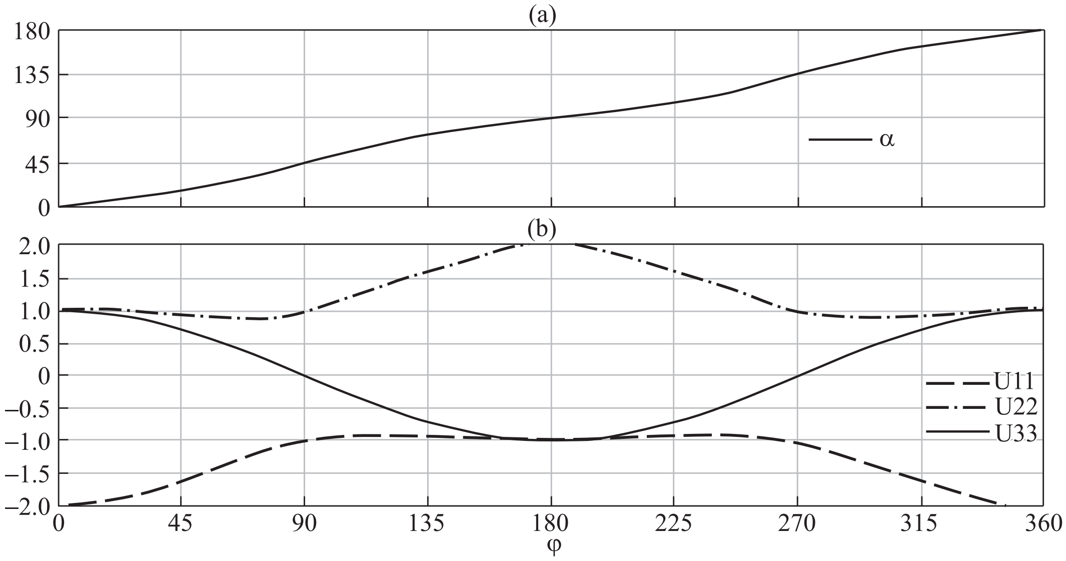


Fig. 3. Dependence of the tensor parameters on the value of the angle φ: (a) angular divergence of coordinate systems (angle α), (b) values of the main components of the tensor $\mathbf{U}'$.

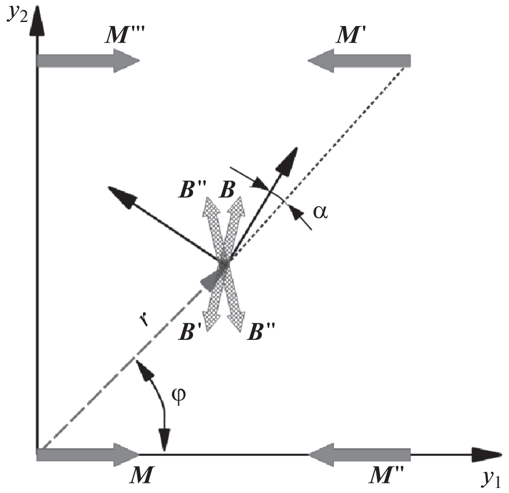


Fig. 4. Uncertainty in determining the DMM from measurements of the magnetic field gradient tensor.

system of the receiver is rotated around the y_3 axis by 180°, the tensor value remains unchanged, only its main components change places and change sign. Figure 3 clearly shows this.

The fact that the values of the angles α and φ are determined by the same ratios of the values of the main components of the tensor (the angle α is determined up to rotation for 180°) gives grounds for determining the directions of the radius from the data of magnetic gradiometry measurements vector **r** and dipole moment vector **M**. With a known absolute value of the dipole magnetic moment, the distance between the transmitter and the receiver can be determined, which constitutes the solution to the relative positioning problem.

Unfortunately, the solution to the positioning problem is ambiguous. Having the results of measurements of the components of the tensor $\mathbf{U}'$ at a certain point in space, the problem of positioning the dipole-transmitter in the system of the principal axes of the tensor can be considered as follows:

On the interval from 0° to 90° in φ, the $\mathbf{U'_{11}}$ component, corresponding to the value of the second derivative with respect to the first component, is maximum in amplitude and negative. Having determined direction of the first axis, it is necessary to choose the direction of the third so that the minimum amplitude of the gradient corresponds to it. The second axis complements the vector tripod to the right.

On the interval from 90° to 180° in φ, the component $\mathbf{U'_{22}}$, corresponding to the second derivative with respect to the second component, is maximum in amplitude and positive. Having determined direction of the second axis, it is necessary to choose the direction of the third axis so that the minimum amplitude of the gradient corresponds to it. The direction of the first axis should set the right tripod.

In the interval from 360° to 180° in φ the tensor components behave in the same way as in the interval from 0° to 180°. Thus, the angle φ can only be determined up to sign. Moreover, if for φ from 0° to 180° the angle α is determined, then for φ from 360° to 180° this angle is α.

Due to their insensitivity to 180° rotation, the components of the gradient tensor determine two possible directions of the location of the transmitter dipoles that could create the measured gradient—$\mathbf{M}$ and $\mathbf{M}'$. These possible transmitters are located opposite to the observation point, identical in size and opposite in direction. In addition, two more dipoles $\mathbf{M}''$ and $\mathbf{M}'''$ also correspond to the measurement results due to symmetry with respect to the dipole axis (Fig. 4).

Thus, the problem of determining the position of the transmitter from measurements of the gradient tensor is uniquely solved only if the quadrant of its location is known. It is also clear from Fig. 4 that additional information about the direction cosines of the field vector $\mathbf{B}$ will allow us to immediately reject incorrect hypotheses, and if we assume that the absolute value of the dipole magnetic moment of the transmitter is known, then according to (2) and (3) we can also determine the distance to the dipole, i.e. obtain the necessary information to solve the relative positioning problem.

Note, however, that the result of measurement the parameters of the field gradient tensor of a point transmitter is invariant to the rotation of the coordinate system associated with the field source around an axis whose direction coincides with the direction of the dipole magnetic moment vector. This means that the measurements taken are not enough to determine the relative angular orientation of objects.

To solve this problem, additional information can be used, which for some conditions is quite natural. Thus, when the ship approaches the berth, the directions of the vertical axes in the systems associated with the field source and the transemitter can be considered coincident. If the field source is located on the cone of the refueling hose, and the dipole moment vector is directed along it, then the effect of rotating the coordinate system around the moment vector does not change anything from the point of view of the docking process during air refueling.

A complete solution to the positioning problem can be obtained by placing not one, but several dipole transmitters on one of the interacting objects. The technical capacity of performing correct measurements in this option is discussed below.

4. POSITIONING USING A VECTOR MAGNETOGRADIOMETER

It is important to note that at the hardware level, measurement tensor components (3) involves the use of three spatially separated vector sensors—field induction meters. Today, such devices are characterized by low accuracy rates.

Scalar magnetically sensitive sensors that directly measure the absolute value of the field induction are somewhat more accurate. Their operation is based on the quantum effects of precession of atoms in polarized light (optically pumped quantum magnetometer) or protons (proton and Overhauser magnetometers) [6]. In this regard, it is interesting to consider the possibility of determining the spatial location and orientation of the field source based on the results of determining the gradient vector of the absolute value of the magnetic field induction vector. The components of this vector can be measured by a system composed of four spatially separated scalar sensors. The value of the gradient vector of the absolute value of the field and the gradient tensor are related by the equation

$$\nabla\left|\mathbf{B}\right| = \left(\nabla\mathbf{B}^{\mathrm{T}}\right)\left(\mathbf{B}/\left|\mathbf{B}\right|\right). \tag{4}$$

This equation is obtained by differentiating $|\mathbf{B}| = \sqrt{\mathbf{B}^{\mathrm{T}}\mathbf{B}}$. It turns out that to solve the positioning problem using vector gradiometry data during measurements, it is necessary to determine not only the scalar field values at four points, but also the direction of the field vector (the ratio $\mathbf{B}/\left|\mathbf{B}\right|$ in (4)).

Calculations show that the gradient vector is directed predominantly towards the source (Fig. 5). The magnitude of the angular discrepancy β between the gradient vector and the direction to the transmitter depends on the angle φ between the directions of the radius vector $\mathbf{r}$ and the dipole moment vector $\mathbf{M}$. The maximum discrepancy is about 15°.

The dependence of the angular divergence of the β radius vector and the gradient vector on the direction to the dipole is shown in Fig. 6. It is clearly seen that even with direct measurements of

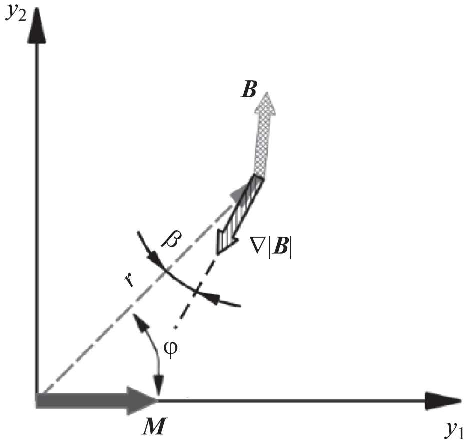


Fig. 5. Measurement of the field and gradient vector of the absolute value of the magnetic vector.

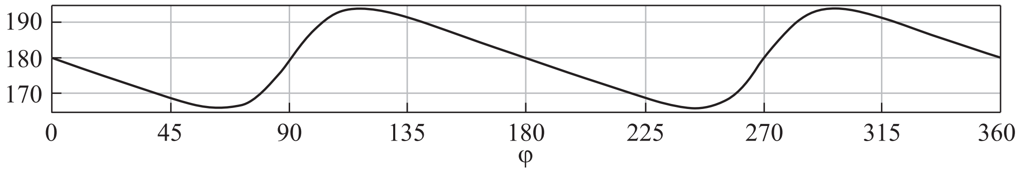


Fig. 6. Divergence of the directions of the radius vector and the gradient vector.

the gradient vector, the problem of determining the direction to the dipole-transmitter is solved, although roughly, but without the ambiguity inherent in tensor measurements.

From the measurement data of the vector with a known value of the dipole magnetic moment, the distance to the transmitter can be calculated, but additional information is needed to determine the radius vector. This additional information can be obtained from a series of measurements as objects move relative to each other. It is also possible to use readings from several spaced gradiometers. Since the symmetry axes of equivalent solutions are lines drawn through the measurement point parallel and perpendicular to the dipole axis, then for three gradiometers that do not lie on the same line, the result of determining the source position will be a single point. Note that such a scheme, although technically complex, does not require specifying the value of the dipole moment of the source, i.e. makes it possible not only to get rid of ambiguity, but also to localize the source, while determining the value of its dipole moment.

5. LIMITATIONS OF TECHNICAL IMPLEMENTATION

The choice of the form of presentation of magnetic gradient information, and therefore the method of measurement, and the structure of the magnetic measurement installation for solving the problem of relative positioning is largely determined by the working conditions. A significant role is played by the functioning of sensors, the dynamics of object movement, the presence of interference in the application area, and much more. However, it is important that in addition to the field caused by the operation of an artificial dipole-transmitter, the receiver inevitably registers the natural magnetic field of the earth, which is very large in magnitude, usually has a significant gradient, and also unpredictably changes in time under the influence of natural geomagnetic disturbances.

This fact, however, should not be considered a significant hindrance to the implementation of the methods and algorithms discussed above, since an inductor (loop dipole) powered by an alternating current of a certain shape can be used as a field source. This approach allows the use of a two-

and three-dipole transmitter, thereby overcoming the ambiguity in determining the direction to the source in the case of using a tensor receiver. The task of isolating the field vector of each transmitter individually is not significantly difficult.

Another kind of difficulty in the application of the considered algorithms turns out to be related to the perfomance features of magnetically sensitive sensors and, first of all, the influence of magnetic interference during the measurement process. The use of an alternating magnetic field allows the use of narrowband filtering algorithms, which significantly reduces this negative impact. Moreover, this approach allows the use of induction magnetometers as recievers, which are not capable of measurement the constant component of the field, but have a significantly higher sensitivity in relation to other types of sensors.

It is also important to note that the considered algorithms are basic and do not take into account fundamentally important aspects of a possible technical implementation. Thus, the field source is assumed to be local, or more precisely, a point dipole transmitter. However, a technically feasible transmitter inevitably has a non-zero size, and therefore its field differs from the field of an ideal dipole. The degree of difference decreases with distance, however, with a significant distance, the amplitude of the measured field decreases significantly, the limitations of the sensitivity and accuracy of the sensors, and the negative influence of various external interferences are fully manifest themselves.

Similar difficulties in technical implementation are typical for gradient field receivers. The definition of the gradient as the second derivative of the potential assumes that the increments of the field induction vector along the selected directions are measured at a point at infinitesimal distance increments. In the technical implementation, even at small distances between the field measurement points, the discrepancy between the values of the derivative $\frac{\partial \mathbf{B}}{\partial x}$ and the ratio $\frac{\triangle \mathbf{B}}{\triangle x}$ is also present due to the essentially nonlinear dependence of the field magnitude on distance ($|\mathbf{B}| \sim 1/\,|\mathbf{r}|^3$) inevitably increases as approach the field source. In the same context, consideration of the possibility of using scalar sensors to construct a vector gradiometer deserves special attention. High-precision and highly sensitive scalar quantum magnetometers with optical pumping could be used with a small distance between them in the structure of the installation, but their design is such that bringing the sensors closer to each other than 1.5 m radically distorts the readings. No less important factors that can destroy the harmonious scheme of basic algorithms are other imperfections of various magnetically sensitive sensors and the measurement system as a whole: orientation errors, various types of nonlinearities, temperature drift of zeros and scale factors, etc.

6. EXPERIMENTS TO EVALUATE THE ACCURACY OF DETERMINING THE RELATIVE POSITION

The above features of the technical implementation make the possibility of putting the basic algorithms into action not entirely obvious and explain the desire to conduct experiments on the actually achievable capabilities of the system in terms of: the required characteristics of the sensors and the measurement system as a whole, the available range of distances between the source and the field reciever, the degree of influence of various types of interference, the potentially achievable accuracy of determining geometric parameters, and speed of operation. To assess the technical feasibility and confirm the effectiveness of the considered algorithms, a series of experiments was carried out, the task of which was to assess the accuracy of determining distances and directions in real conditions, taking into account natural magnetic interference and the limited accuracy of magnetically sensitive sensors, as well as the limited accuracy of monitoring the dipole magnetic moment of the emitter.

A loop transmitter was used as a field source—a flat inductor with a diameter (500 mm, 100 turns), fed by a meander-shaped current with a frequency of 4 Hz. The amplitude of the

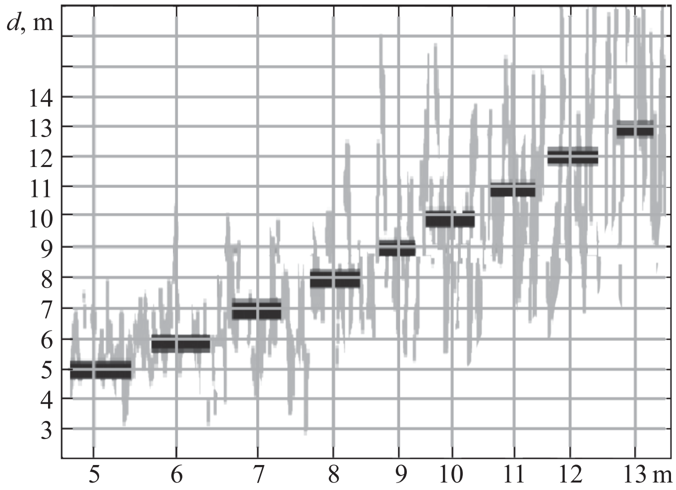


Fig. 7. Calculating the distance to the field source.

dipole magnetic moment was about 35 Am^2; to simplify control, the direction of the vector was set horizontal. The tensor-type magnetic gradiometry reciever was composed of three vector fluxgate magnetometers HB0302 [7], having a sensitivity of 1.0–5.0 nT. The sensors were installed on a rotating platform in a horizontal plane along the vertices of an equilateral triangle with an edge length of 1.0 m. The experiments were preceded by a series of calibration procedures, the coverage of the theoretical foundations and technology of which is beyond the scope of this article. The sequence of measurement procedures in the final experimental design was presented in the following series.

At a known distance from the center of the triangle of the magnetic gradiometry system to the field source (this distance ranged from 5 m), a series of measurements were performed in which the magnetic gradiometry measurement installation remained stationary, and the loop transemitter, maintaining its location in space, sequentially changed the direction of the dipole moment in azimuth. Then the measurement installation, remaining in place, changed its position in azimuth. This series made it possible to evaluate the accuracy of determining the direction to the source and the direction of the vector of its dipole moment. Measurements in this sequence were performed twice. The first part were as the basis for calibration procedures, and according to the data of the other, accuracy control was carried out.

The second series of experiments consisted of monitoring the accuracy of determining the distance to the field source for different directions of the dipole magnetic moment vector. The magnetic measurement installation remained stationary, and the loop transmitter with a step of 2.0 m moved away from the reciever at a distance of 5 to 13 m. At each position, four measurements were performed at different directions of the dipole magnetic moment vector. In this series, the accuracy of determining the distance to the source was assessed at various distances and for various directions of the dipole moment vector.

During the experiments, the following results were obtained.

Figure 7 shows the results of an experiment to determine the distance to the source from gradient measurements. The specified values of the distance between the dipole-transmitter and the measurement installation are plotted horizontally, and calculated values are displayed vertically. The curve shows the calculated value, the horizontal segments—averaged for each of the time intervals corresponding to the distance of the dipole from point to point with a step of 1.0 m.

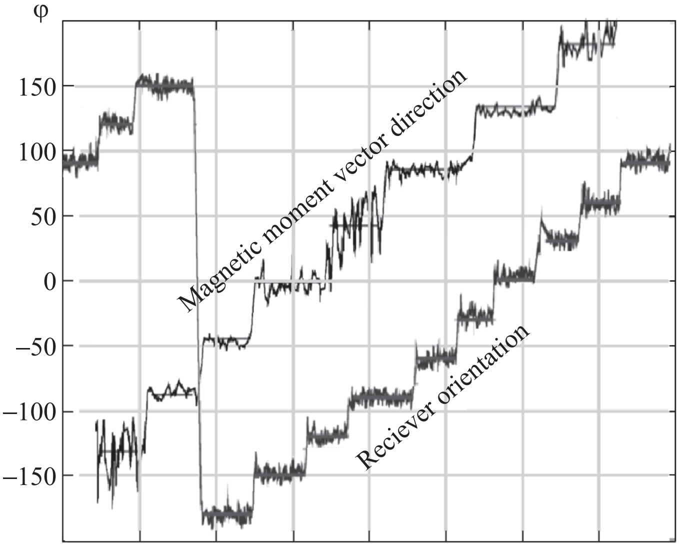


Fig. 8. Calculation of angular orientation parameters.

It can be seen from the figure that the distance to the source in the presented experimental design is generally calculated reliably. The small discrepancy is explained by the imperfection of the experimental conditions: the significant influence of magnetic interference in the measurement area, as well as the error in the placement of the transmitter dipole relative to the measurement system. The resulting accuracy in this experiment was 4–9% depending on the value of the determined distance.

Figure 8 shows two graphs showing the possibility of determining the parameters of the mutual angular orientation of the receiving system and transmitter from magnetic gradient measurements. The graphs show the results of changes over time in determining the values of the angles of the azimuthal orientation of the dipole magnetic moment vector (direction of the moment vector) and the selected axis of the measurement installation (orientation of the meter). The calculated values are plotted along the vertical axis. Line segments in the graphs show preset values. It is clearly seen from the figure that in this experiment the direction to the dipole-transmitter was determined based on the results of magnetic gradiometry measurements in general more accurately than the direction of the dipole moment vector, however, taking into account the simplicity of the measurement scheme, in general, sufficient reliability of the operation of the algorithms for determining both directions is shown.

The resulting accuracy in determining the orientation of the reciever was 3–10° depending on the distance. The resulting accuracy in determining the direction of the dipole moment vector depends not only on the distance, but also on the orientation of the reciever. It was 10–30° depending on the distance.

7. CONCLUSION

The research presented in the paper made it possible to formulate the basic principles of a promising method of relative angular and spatial positioning of objects. The above calculations show the fundamental possibility of constructing structurally and functionally simple high-precision systems useful for solving problems of controlling the movement of objects during their interaction: mooring, docking, refueling in the air, monitoring the position of the ship relative to the anchor, etc. The experiments presented in this work confirmed the perfomance capabilities of constructing systems operating on the principles of the algorithms discussed in the work.

REFERENCES

1. Obolensky, Yu.G., Pokhvalensky, V.L., and Cheglakov, D.I., Algorithm for automatic control of an aircraft during refueling in the air, *Proceedings of MAI*, 2013, no. 65, pp. 1–17.

2. Nebylov, A.V., Perlyuk, V.V., and Leontyeva, T.S., Research on the technology of mutual navigation and orientation of small spacecraft in a group, *Bulletin of Samara University. Aerospace engineering, technology and mechanical engineering*, 2019, vol. 18, no. 1, pp. 88–93.

3. Kolesnikov, M.P., Martynova, L.A., Pashkevich, I.V., and Shelest, P.S., Positioning method for an autonomous uninhabited underwater vehicle in the process of bringing it to a mooring device, *Izv. Tul. state un-ta. Technical science*, 2015, no. 11, part 2, pp. 38–48.

4. Landau, L.D. and Lifshits, E.M., *Course of Theoretical Physics. Vol. 2 (The Classsical Theory of Fields)*, London: Butterworth Heinemann, 1996.

5. Volkovitsky, A.K., Karshakov, E.V., and Pavlov, B.V., *Magnetic gradiometry measurement systems and complexes: Monograph in two volumes. Principles of measurements and structure of magnetic gradient complexes. Volume I*, Moscow: IPU RAN, 2018.

6. Pomerantsev, N.M., Ryzhkov, V.M., and Skrotsky, G.V., *Physical foundations of quantum magnetometry*, Moscow: Nauka, 1972.

7. *Magnetic devices. Three-component magnetic field induction converter NV0302* [Electronic resource]:— Manufacturer's website — Electronic data. Access mode: URL: https://www.magnetic.spb.ru/products/31125352, free—(access date 07.15.2023).

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 10, pp. 1153–1167.

Optimization of Interception Plan for Rectilinearly Moving Targets

A. A. Galyaev[*,a], **V. P. Yakhno**[*,b], **P. V. Lysenko**[*,c],
L. M. Berlin[*,d], **and M. E. Buzikov**[*,e]

[*] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
e-mail: [a]*galaev@ipu.ru,* [b]*vic_iakhno@mail.ru,* [c]*pavellysen@ipu.ru,*
[d]*berlin.lm@phystech.edu,* [e]*me.buzikov@physics.msu.ru*

Received June 19, 2023
Revised July 18, 2023
Accepted August 2, 2023

Abstract—The article considers the problem of combinatorial optimization of interception of rectilinearly moving targets as a modification of the traveling salesman problem. New macro characteristics and definitions for this problem are introduced and used to classify the obtained solutions. Vector criteria composed of several important for applications functionals are described. The principles of no-waiting and maximum velocity are proved for two types of criteria. An intelligent brute-force algorithm with dynamic programming elements for finding optimal plans according to the introduced intercept criteria is proposed and implemented. Statistics of solutions of the developed algorithm is collected for a set of different initial parameters and the proposed macro characteristics are investigated. The conclusions about their applicability as local rules for the greedy algorithm for finding a suboptimal intercept plan are drawn.

Keywords: moving targets traveling salesman problem, combinatorial optimization, simple motion model

DOI: 10.25728/arcRAS.2023.62.75.001

1. INTRODUCTION

A recent development of intelligent technologies in the field of unmanned autonomous vehicles makes it possible to use them cooperatively in a vast variety of applications and scenarios that were considered impossible before. One such application is the problem of preventing moving targets from infiltration of a given point in space by intercepting each one. The problem of optimal choice of the traverse order of targets is found to be crucial in the formalization and further solution of the problem of intercepting a set of targets. An optimal choice in terms of one criterion may turn out to be poor when another criterion is used. For example, if the optimal choice of the traverse order of moving targets is related to the requirement that targets should be intercepted in the shortest possible time at the greatest possible distance from the defended point, then an internal contradiction of such requirement can be shown. Let us consider a situation in which one defender faces two enemy targets, one of them is fast and the other is slow, and the defender's starting point is attacked from diametrically opposite directions. For the fastest execution time it is needed to intercept the slow target first, letting the fast target get closer, but if the fast target is intercepted first, then the shortest distance to the defended point will be greater than in the first case.

The Moving Targets Traveling Salesman Problem (MTTSP) [1–4] is the closest problem statement to the problem of constructing an optimal plan for intercepting moving targets studied in this paper. The MTTSP is a generalization of the traveling salesman problem (TSP). In 1972, the NP-completeness of the Hamiltonian cycle problem was shown to imply NP-completeness of the

MTTSP [5]. One of the first MTTSP statements for rectilinearly moving targets was in paper [1] where it was found that the dynamic programming apparatus could be used to construct an efficient algorithm for finding the optimal intercept plan.

In MTTSP, it is generally assumed that the control object movement is subject to simple motion model (the controlled input is a velocity vector). For some conditions [1, 4] such an assumption makes it possible to switch from a discrete-continuous optimization problem to a discrete optimization problem. The model where the controlled input is a velocity vector can be a sufficiently rough approximation for constructing reference trajectories of a real control object, but using more accurate models that take into account, for example, the maneuverability of the control object, do not allow to switch from a discrete-continuous problem to a discrete one to obtain an accurate solution even in the case of stationary targets [6, 7]. In this case, if the control object is sufficiently maneuverable and the distances between targets are large (in comparison to the minimum turning radius of the object), then taking into account maneuverability in the problem of target traverse planning does not affect the structure of the optimal plan.

Methods for solving MTTSP can be categorized as follows:

- with time sampling [3, 8, 9] or without [1, 4, 8, 10, 11];
- giving an optimal solution [3, 4, 8] or suboptimal [9–11];
- deterministic [1, 3, 4, 8, 9, 12] or random [10, 11, 13].

In [4] an algorithm for constructing a guaranteed intercept plan based on the notion of target danger is proposed for the problem of preventing intrusion of targets into a given point with the traveling salesman (TS) returning to it after each encounter with the target.

This paper considers the problem of optimizing an intercept plan for a vector criteria. The concepts of danger, convenience and complexity of intercept are formalized. The preference is given to deterministic methods without time discretization that give an optimal solution (without guarantees of fast completion).

The paper consists of an introduction, four sections and a conclusion and has the following structure. In Section 2, a new formulation of the problem of finding an optimal plan for intercepting targets moving rectilinearly to one protected point is formalized, the set of acceptable plans, vector criteria of the problem are introduced, and the definition of a guaranteed intercept plan is given. In Section 3, the theorems of guaranteed intercept and the principle of non-optimality of no-waiting are proved, and new notions of danger, convenience and complexity of intercept plan are introduced. Section 4 is devoted to an intelligent algorithm for brute-force search that significantly reduces the number of computations. Later in 5 the results of simulation based on the proposed algorithm are presented and the properties of the optimal plans are statistically investigated. In the 6 plans for further work are presented.

2. PROBLEM STATEMENT

2.1. Mathematical Model

Let us assume that the protected point is located on the plane at the coordinate origin. The targets appear on the outer boundary of a circle of radius R in a layer of width $2\Delta R$ in a sector with a central angle α. Targets move rectilinearly with known velocities of given range $[v_{\min}, v_{\max}]$. At the initial moment TS is situated at the origin and it controlled with velocity $v(t) \in [0, V]$, $V > v_{\max}$. It is assumed that moving targets must be serviced by the salesman as far from the protected object as possible, minimizing the danger as much as possible.

Definition 1. The set of initial conditions for targets and TS with all parameters of the problem being fixed is called the initial state.

The state may change as the current data about the objects is changed and refined. Since any moment when all parameters of the problem are known can be chosen as the initial moment, the situation at this moment will be called the current state.

Let us assume that there are only m targets and each of them is located at the initial moment at the point $\boldsymbol{r}_j^0 = (x_j^0, y_j^0)$, where $j = 1, \ldots, m$. Each target moves with constant velocity $\boldsymbol{v}_j = (v_{x,j}, v_{y,j})$. Thus, the trajectory of each target is a straight line

$$\mathbf{r}_j(t) = \boldsymbol{r}_j^0 + \boldsymbol{v}_j t, \quad j = 1, \ldots, m \tag{1}$$

with constrained parameters

$$\begin{gathered} ||\boldsymbol{v}_j|| \in [v_{\min}, v_{\max}], \quad v_{\max} < V, \\ ||\boldsymbol{r}_j^0|| \in [R - \Delta R, R + \Delta R], \\ \arctan \frac{y_j^0}{x_j^0} \in \left[\frac{\pi}{2} - \frac{\alpha}{2}, \frac{\pi}{2} + \frac{\alpha}{2}\right]. \end{gathered} \tag{2}$$

When the target reaches the origin, it is meaningless to service it. This moment in time for target number j can be calculated as follows:

$$t_j^0 = \frac{\|\boldsymbol{r}_j^0\|}{\|\boldsymbol{v}_j\|}. \tag{3}$$

The dynamics of TS is described with a system of differential equations of the following form:

$$\begin{aligned} \dot{\mathrm{x}}^I(t) &= v(t)\cos\psi(t), \quad v(t) \in [0, V]; \\ \dot{\mathrm{y}}^I(t) &= v(t)\sin\psi(t), \quad \psi(t) \in [0, 2\pi); \end{aligned} \tag{4}$$

where $\mathbf{r}^I(t) = (\mathrm{x}^I(t), \mathrm{y}^I(t))$ is the position of TS at the moment t; $\psi(t)$ is the velocity direction control on the plane. The salesman is at the origin at the initial moment $\mathbf{r}^I(0) = (\mathrm{x}^I(0), \mathrm{y}^I(0)) = (0, 0)$.

In order to formulate the plan construction problem as an optimization problem, it is necessary to give a formal description of the problem model, which include definitions of problem solution, its acceptability, and a quality criterion of the problem.

The principles of non-optimality of no-waiting and maximum velocity motion is proved later in the paper. The intercept function is constructed for the problem of the fastest intercept of a target moving uniformly along a straight line by TS whose movement is subject to simple motion model, and is reduced to finding the smallest positive root of the following quadratic equation with respect to the intercept time τ:

$$(\boldsymbol{r}_j + \boldsymbol{v}_j \tau)^2 = V^2 \tau^2.$$

Here $\boldsymbol{r}_j = \mathbf{r}_j(t) - \mathbf{r}^I(t)$ is a vector of relative positions of TS and the target with number j, $\mathbf{r}^I(t)$ is a current position of TS, V is a maximum velocity of TS, $\boldsymbol{v}_j$ is a velocity vector of the target. Let us denote the smallest non-negative root of this equation by $\tau(\boldsymbol{r}_j, \boldsymbol{v}_j)$. It can be shown that when $\boldsymbol{v}_j^2 < V^2$ the following expression is valid

$$\tau(\boldsymbol{r}_j, \boldsymbol{v}_j) = \frac{(\boldsymbol{v}_j, \boldsymbol{r}_j) + \sqrt{(\boldsymbol{v}_j, \boldsymbol{r}_j)^2 + \boldsymbol{r}_j^2 (V^2 - \boldsymbol{v}_j^2)}}{V^2 - \boldsymbol{v}_j^2}. \tag{5}$$

2.2. Individual Plan

An individual plan π for TS intercepting $k \in \{0, \ldots, m\}$ targets is a tuple of k different numbers from $\mathcal{M} = \{1, \ldots, m\}$. The order of the elements in the tuple determines the order according to which the targets are intercepted. The space of all individual plans intercepting $k \in \{0, \ldots, m\}$ targets can be described as follows:

$$\Pi_k = \Big\{(\pi_1, \ldots, \pi_k) \in \mathcal{M}^k : \ \forall p, q \in \{1, \ldots, k\} \quad p \neq q \to \pi_p \neq \pi_q\Big\}.$$

Let us consider for example $m = 2$:

$$\begin{aligned}
&\Pi_0 = \{()\},\\
&\Pi_1 = \{(1,), (2,)\},\\
&\Pi_2 = \{(1,2), (2,1)\}.
\end{aligned}$$

() denotes an empty tuple (an individual plan prescribing inaction). Thus, if an individual plan $\pi = (\pi_1, \ldots, \pi_k) \in \Pi_k$ is prescribed for TS, then TS should firstly intercept the target with number π_1, then – the target with number π_2, and so on.

The space of all plans for a given number of targets m is the following set:

$$\Pi = \bigcup_{k=0}^{m} \Pi_k.$$

First, let us compute the minimum time $T(\pi)$ that it takes a salesman to execute an individual plan $\pi = (\pi_1, \ldots, \pi_k)$. Using the definition of $\tau(\boldsymbol{r}_j, \boldsymbol{v}_j)$ from (5), the following recursive expression can be obtained:

$$T(\pi) = \begin{cases} 0, & k = 0;\\ \tau(\boldsymbol{r}^0_{\pi_1}, \boldsymbol{v}_{\pi_1}), & k = 1;\\ t + \tau(\mathbf{r}_{\pi_k}(t) - \mathbf{r}^I(t), \boldsymbol{v}_{\pi_k}), & k > 1, \ \text{here } t = T((\pi_1, \ldots, \pi_{k-1})). \end{cases} \tag{6}$$

Let us also write out the constraint that every target entering the plan must be intercepted on time, i.e. before reaching the coordinate origin:

$$\mathrm{OnTime}(\pi) = \Big(\forall j \in \{1, \ldots, k\} : \ T((\pi_1, \ldots, \pi_j)) \leqslant t_{\pi_j}\Big).$$

It should be noted that last restriction can be checked recurrently. If for an individual plan $\pi = (\pi_1, \ldots, \pi_k)$ the corresponding constraint is satisfied and the target j is not considered in the individual plan π, then the following expression can be used to check the constraint for the individual plan $\pi + j = (\pi_1, \ldots, \pi_k, j)$:

$$\mathrm{OnTime}(\pi + j) = \mathrm{OnTime}(\pi) \ \& \ T(\pi + j) \leqslant t_j.$$

Definition 2. The interception plan is acceptable if it allows to intercept moving targets on time. The set of acceptable plans is the following

$$\Pi_A = \{\pi \in \Pi : \ \mathrm{OnTime}(\pi)\}.$$

2.3. Criteria and Definitions

The criterion of the problem is associated with the loss functional J. Obviously, a lesser loss function corresponds to a better solution. The loss functional must be defined on the set of acceptable plans Π_A, i.e. for each plan an estimate of the losses can be made. The optimal solution to the traverse plan construction problem is the plan $\pi^* \in \Pi_A$ that minimizes the value of the loss functional:

$$\pi^* \in \arg \min_{\pi \in \Pi_A} J[\pi]. \tag{7}$$

The set of acceptable plans Π_A is finite and contains at least one element (which is the empty plan), so the minimization problem always has a solution, maybe not the only one. The equal sign in the expression $\pi^* = \arg \min\limits_{\pi \in \Pi_A} J[\pi]$ means that solution is unique.

Basic functionals that can be used to build the criterion of the problem are

- **Missed targets** (the number of missed targets that reached the origin). All targets that were not included in the individual plan $\pi \in \Pi_A$ will reach the origin, i.e. the number of missed targets is calculated as follows

  $$n_0[\pi] = m - \mathrm{card}(\pi),$$

  where $\mathrm{card}(\pi)$ is the length of the plan π.
- **Execution time.** It is calculated as the execution time of the individual plan

  $$T_{\mathrm{sum}}[\pi] = T(\pi).$$

- **Minimum interception distance.** The minimum distance from the origin to target interception point for the plan $\pi \in \Pi_A$ is calculated as follows

  $$D_{\min}[\pi] = \min_{j \in \{1,\ldots,m\}} \|\mathbf{r}_{\pi_j}(T((\pi_1, \ldots, \pi_j)))\|.$$

  If the plan π is empty, it will be formally assumed that $D_{\min}[\pi] = 0$.

Not all of the mentioned functionals are suitable for the role of the problem criterion. Indeed, minimizing only the execution time of the plan leads to an empty plan consisting in inaction, and it will be optimal because zero time units are required for its execution. Only a loss functional, describing the number of targets that reached the origin, can be used as problem criterion.

Definition 3. The interception plan is guaranteed if $n_0[\pi] = 0$. The set of guaranteed plans is denoted by Π_G.

Some generalization of the comparison method is needed to adequately compare the listed functionals in the final problem criterion. Most of the mentioned functionals make sense in the minimization problem if they are combined together to form a criterion. For example, if two plans are compared primarily on the number of missed targets that reached the origin, and secondarily, on the total interception time, then such a combined loss functional adequately capture the essence of the point defense problem. In other words, if some plan π_1 admits skipping one target to the origin and the execution time of the plan is 7, i.e. $n_0[\pi_1] = 1$ and $T_{\mathrm{sum}}[\pi_1] = 7$, and a plan π_2 admits skipping one target to the origin and the execution time of the plan is 8, i.e. $n_0[\pi_2] = 1$ and $T_{\mathrm{sum}}[\pi_2] = 8$, then plan π_1 is better than plan π_2, i.e. the tuples $(1, 7) < (1, 8)$ can be formally compared. This comparison is similar to the positional comparison of real numbers, where each digit at the corresponding position of a real number is compared to the corresponding digit of another number until no differences in values are found from left to right. Let us formalize the above on the concept of lexicographic order.

Definition 4. A tuple of numbers $\boldsymbol{a} = (a_1, a_2, \ldots, a_p)$ is less than a tuple of numbers $\boldsymbol{b} = (b_1, b_2, \ldots, b_q)$ if there exists a number $k \in \{1, \ldots, \min(p, q)\}$ such that $a_i = b_i$ for $i < k$ and $a_k < b_k$. If for all $k \in \{1, \ldots, \min(p, q)\}$ $a_k = b_k$, then for $p < q$ it is assumed that $\boldsymbol{a} < \boldsymbol{b}$. In other cases, it is assumed that $\boldsymbol{a} \geqslant \boldsymbol{b}$.

Examples:

$$(1,2) < (1,3), \quad (0,1) < (1,2), \quad (1,2) < (1,2,1), \quad () < (1,2), \quad (1,2) < (2,).$$

The main criteria of the target interception problem for an arbitrary acceptable plan $\pi \in \Pi_A$ are formulated using the definition of tuple comparison.

- **Missed targets + Execution Time.** The quality criterion for the obtained plans is the following

$$J_T[\pi] = (n_0[\pi], T_{\text{sum}}[\pi]). \tag{8}$$

  Minimizing of this loss functional is primarily aimed at minimizing the number of missed targets and time execution of the plan secondarily.

- **Missed targets + Minimum interception distance.** The quality criterion is the following

$$J_D[\pi] = (n_0[\pi], -D_{\min}[\pi]). \tag{9}$$

  Minimization of this loss functional is primarily aimed at minimizing the number of missed targets and secondarily at maximizing the distance of the closest target approaching to the origin.

- **Missed Targets + Minimum interception distance + Execution Time.** The quality criterion is the following

$$J_{DT}[\pi] = (n_0[\pi], -D_{\min}[\pi], T_{\text{sum}}[\pi]). \tag{10}$$

  Minimization of this loss functional is primarily aimed at minimizing the number of missed targets, secondarily at maximizing the distance of the closest target approaching to the origin and thirdly at minimizing of total execution time.

Let us formulate the optimization problem.

Problem 1. For m targets moving along trajectories (1) with constraints on the motion parameters (2), it is required to find an optimal according to criterion (8) or (10) intercept plan $\pi \in \Pi_A$ for TS with dynamics (4).

3. PROPERTIES OF THE OPTIMAL INTERCEPT PLAN SEARCH PROBLEM

The following definitions and terms are needed to describe the problem properties.

Using the formula (5), the time $\tau_j(t) = \tau(\boldsymbol{r}_j, \boldsymbol{v}_j)$ of intercepting the jth target from the current state and the time $t_j(t)$ of movement of the jth target to the coordinate origin are introduced.

Definition 5. A danger K_j of the jth target is defined as the inverse value of the movement time needed to reach the coordinate origin, namely

$$K_j(t) = \frac{1}{t_j(t)}.$$

The danger is a property of the target. The less time before the target enters the protected region, the more dangerous it is considered.

Definition 6. The convenience U_j of intercepting the jth target is the inverse of the time which TS needs to intercept this target from the current state, namely

$$U_j(t) = \frac{1}{\tau_j(t)}.$$

Convenience is a property of TS's action with respect to the target. The less time it takes, the more convenient it is to intercept the target.

Definition 7. The intercept complexity $C[\pi]$ of the plan π is the maximum time between two consecutive intercepts in the plan, namely

$$C[\pi] = \max_{\substack{\{\pi_j\} \in \pi, \\ 1 < j \leqslant m.}} \tau_{\pi_j}(T((\pi_1, \ldots, \pi_{j-1}))). \tag{11}$$

Complexity is a property of TS's plan. The less time there is between two consecutive target intercepts during plan execution, the less complex it is.

Definition 8. The average complexity $\widehat{C}[\pi]$ of an intercept plan π is the average time between two consecutive intercepts in the plan, namely

$$\widehat{C}[\pi] = \frac{1}{m-1} \sum_{\substack{\{\pi_j\} \in \pi, \\ 1 < j \leqslant m.}} \tau_{\pi_j}(T((\pi_1, \ldots, \pi_{j-1}))). \tag{12}$$

Average complexity characterises the durations between consecutive interceptions in a plan. If all the targets are intercepted consecutively without long interceptions then this plan is less complex in average compared to the plan containing several long interceptions.

The danger is directly related to the criteria for execution of the intercept plan, while convenience and complexity relate to the sequential selection of the next target and the quality of plan execution according to the time-optimal criterion. If there is a target traverse plan where the consecutive intercepts occur as conveniently as possible and there is no target miss, then the execution time of the plan is often close to optimal.

The notions of danger and convenience can be generalised for the current state.

Definition 9. The danger of the current state is a decreasingly ordered tuple of m target danger values

$$(K_{j_1}(t), \ldots, K_{j_m}(t)). \tag{13}$$

The order of targets in the danger tuple does not change during the execution of the plan.

Definition 10. The convenience of the current state is a decreasingly ordered tuple of m target convenience values

$$(U_{j_1}(t), \ldots, U_{j_m}(t)).$$

The convenience of the current state depends on the position of TS and changes over time.

When targets are intercepted, tuple lengths are reduced. The complexity of the plan is directly related to the convenience of the traverse. The optimal plan combines all of the above state characteristics.

Theorem 1. *For any initial state and any number of targets in the problem 1, there is a guaranteed intercept plan* $\pi \in \Pi_G$.

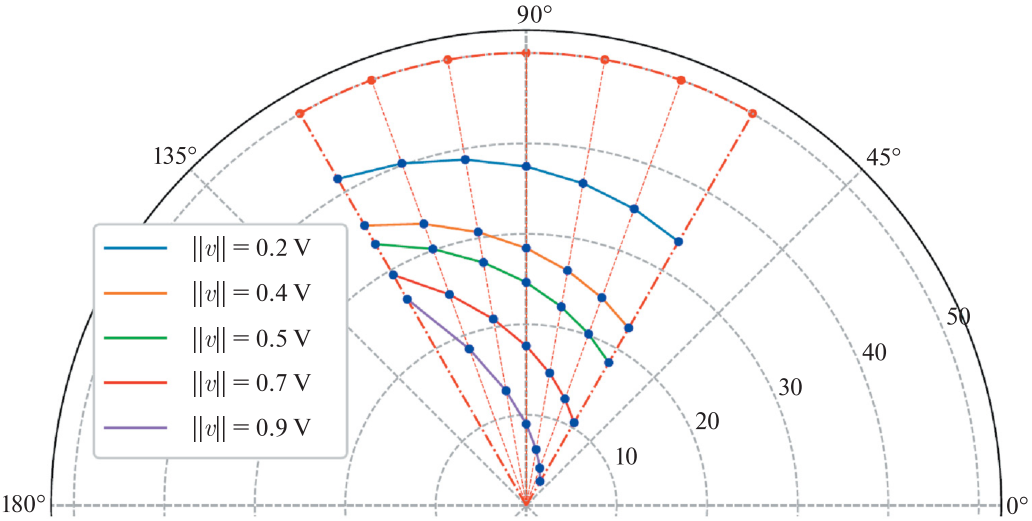


Fig. 1. Intercept targets with velocities $||\boldsymbol{v}|| = \{0.2V, 0.4V, 0.5V, 0.7V, 0.9V\}$, located on the boundary of a circular sector with a centre angle $\alpha = 60°$.

Proof. It is possible to carry out the proof using Theorem 10 of [4], but then the features of the problem will be left out.

The intercept plan is created according to the initial state danger (13) calculated similarly to [4]. Let us prove that such a plan is guaranteed.

If in the initial state the distance $||\boldsymbol{r}_j^0||$, where the index j corresponds to the most dangerous target, is not the minimum among all $||\boldsymbol{r}_k^0||$, $k = 1, \dots, j-1, j+1, \dots, m$, then the moment of the start of TS's movement is postponed. Then a radius R_0 is found such that the targets cross it in decreasing order of danger $(K_{j_1}, \dots, K_{j_m})$ by permutation of the targets $(j_1, \dots, j_m)$ with respect to the initial numbering $(1, \dots, m)$. When the most dangerous target is intercepted, all others are outside the radius of the current interception. Due to the superiority of the velocity of TS, no target will reach the origin, which is shown in the example of a situation where the next most dangerous target j_{d+1} is located diametrically opposite to the current one $||\mathbf{r}_{j_d}(t)|| < ||\mathbf{r}_{j_{d+1}}(t)||$, $t = T((j_1, \dots, j_d))$. In this case, the difference between the times it takes the target and TS to reach the origin is equal to

$$\frac{||\boldsymbol{r}_{j_{d+1}}(t)||}{||\boldsymbol{v}_{j_{d+1}}||} - \frac{||\boldsymbol{r}_{j_d}(t)||}{V} = \frac{V \cdot ||\boldsymbol{r}_{j_{d+1}}(t)|| - ||\boldsymbol{v}_{j_{d+1}}|| \cdot ||\boldsymbol{r}_{j_d}(t)||}{V \cdot ||\boldsymbol{v}_{j_{d+1}}||} > 0.$$

This means that in an extreme case, when the interception occurs along the beams of one straight line, TS will have time to get to the origin, after which TS will intercept the next target. In cases where the interception is carried out on the remaining beams, it is obvious that the targets also will not reach the origin. This proves that the danger interception plan is guaranteed. □

Example 1. Let the targets be uniformly distributed on the boundary of a circular sector with a radius R and a central angle α and move with equal velocities $||\boldsymbol{v}||$. Then the optimal intercept according to the criteria $J_T[\pi]$ and $J_{DT}[\pi]$ is carried out along a trajectory close to the logarithmic spiral [14] following the plan π on which $n_0[\pi] = 0$, as shown in Fig. 1. In this example, the danger and convenience of the initial setting are $(K_1, \dots, K_m) = (U_1, \dots, U_m) = (||\boldsymbol{v}||/R, \dots, ||\boldsymbol{v}||/R)$ and do not allow to make an initial choice of target. Once the rightmost or leftmost target in a sector has been intercepted, the remaining targets will also be equally distributed in danger. However, the intercept convenience tuple will not only have an order of the remaining targets, but also this order will not change after each intercept.

Theorem 2. *For criteria $J_T[\pi]$ and $J_{DT}[\pi]$ in problem 1 the following principles are valid:*

1) the no-waiting principle (on the optimal plan TS cannot be motionless),

2) the principle of maximum velocity (on the optimal plan TS moves at the maximum possible velocity).

Proof. The guaranteed intercept ($n_0[\pi] = 0$) minimising the number of missed targets for criteria $J_T[\pi] = (n_0[\pi], T_{\text{sum}}[\pi])$ and $J_{DT}[\pi] = (n_0[\pi], -D_{\min}[\pi], T_{\text{sum}}[\pi])$, can be obtained by the Theorem 1 by choosing the plan $\pi = (i_1, \ldots, i_m)$ according to the initial state danger $(K_{i_1}, \ldots, K_{i_m})$.

Further optimisation of vector criteria $J_T[\pi]$ and $J_{DT}[\pi]$ is performed according to guaranteed plans ($n_0[\pi] = 0$), consisting of at least one plan $\pi = (i_1, \ldots, i_m)$.

The validity of principles of no waiting and maximum velocity when minimising $T_{\text{sum}}[\pi]$ is shown in Lemma 1 of [4], which finishes the proof of the theorem for the functional $J_T[\pi]$.

Maximizing the functional $D_{\min}[\pi]$ in the criterion J_{DT} for a finite number of guaranteed plans leads to finding the plan π^*. Let us fix this plan and find the first target number j in the plan, on which the minimum distance to the origin is reached. The plan $\pi^* = (\pi_{1,j}, \pi_{j,m})$ is divided into two parts: $\pi_{1,j}$ before the goal j inclusive and $\pi_{j,m}$ after the goal j, then $D_{\min}[\pi^*] = D_{\min}[\pi_{1,j}]$. Increasing the execution time of part of the plan $\pi_{1,j}$ by waiting time or moving at less than maximum velocity leads to a decrease $D_{\min}[\pi_{1,j}]$ similar to Lemma 1 of [4]. Further, the part of the plan $\pi_{j,m}$ is optimal in execution time by Bellman's principle of optimality. The part $T_{\text{sum}}[\pi]$ of the criterion after reaching a minimum on the functional $D_{\min}[\pi]$ is responsible for this. Therefore, waiting and slowing down is impossible for TS on $\pi_{j,m}$ and hence on the whole π^*. □

4. OPTIMAL INTERCEPT PLAN FINDING ALGORITHM

The algorithm for constructing a traverse plan for a single TS and many moving targets is based on brute-force sorting of plans with an initial sorting of targets by danger and an intelligent rule for discarding obviously non-optimal branches of the search in the process of its operation. It is guaranteed that the algorithm finds the optimal intercept plan.

To introduce some important notions that are necessary to understand the algorithm, let us first consider the simplest case of a brute-force search. The work of the algorithm in this case can be illustrated by the transition matrices in Table 1, which show the complete sequence of plans considered during the operation of the algorithm. Table 1 is called the plan search table. The criterion of the problem in the algorithm is from (10):

$$J_{DT}[\pi] = (n_0[\pi], -D_{\min}[\pi], T_{\text{sum}}[\pi]).$$

In Table 1 the transition matrices are numbered from 1 to $m! = 24$. For each matrix, a vector of indices is written in the column on the left that defines the intercept plan. The resulting plan is a sequence of marked circles in the matrix according to the index vector in ascending order of row number.

The uppermost index in the column is marked with an asterisk, from which a new calculation of the next plan begins. An intermediate state, characterised by a part of the already computed plan, is stored to save computational resources.

Statement 1. *The total number of calls of the single intercept function* (5) *in a brute force search of all variants in the algorithm with intermediate saving of calculations is described by the recurrence formula*

$$f(m) = m(f(m-1) + 1). \tag{14}$$

Table 1. Plan search table consisting of transition matrices for the case of brute-force search with $m = 4$

1

1	①	2	3	4
1	②	3	4	
1	③	4		
1	④			

2

1	①	2	3	4
1	②	3	4	
*2	3	④		
1	③			

3

1	①	2	3	4
*2	2	③	4	
1	②	4		
1	④			

4

1	①	2	3	4
2	2	③	4	
*2	2	④		
1	②			

5

1	①	2	3	4
*3	2	3	④	
1	②	3		
1	③			

6

1	①	2	3	4
3	2	3	④	
*2	2	③		
1	②			

7

*2	1	②	3	4
1	①	3	4	
1	③	4		
1	④			

8

2	1	②	3	4
1	①	3	4	
*2	3	④		
1	③			

9

2	1	②	3	4
*2	1	③	4	
1	①	4		
1	④			

10

2	1	②	3	4
2	1	③	4	
*2	1	④		
1	①			

11

2	1	②	3	4
*3	1	3	④	
1	①	3		
1	③			

12

2	1	②	3	4
3	1	3	④	
*2	1	③		
1	①			

13

*3	1	2	③	4
1	①	2	4	
1	②	4		
1	④			

14

3	1	2	③	4
1	①	2	4	
*2	2	④		
1	②			

15

3	1	2	③	4
*2	1	②	4	
1	①	4		
1	④			

16

3	1	2	③	4
2	1	②	4	
*2	1	④		
	①			

17

3	1	2	③	4
*3	1	2	④	
1	①	2		
1	②			

18

3	1	2	③	4
3	1	2	④	
*2	1	②		
1	①			

19

*4	1	2	3	④
1	①	2	3	
1	②	3		
1	③			

20

4	1	2	3	④
1	①	2	3	
*2	2	③		
1	②			

21

4	1	2	3	④
*2	1	②	3	
1	①	3		
1	②			

22

4	1	2	3	④
2	1	②	3	
*2	1	③		
1	①			

23

4	1	2	3	④
*3	1	2	③	
1	①	2		
1	②			

24

4	1	2	3	④
3	1	2	③	
*2	1	②		
1	①			

Thus, the number of calls of the single intercept function is significantly reduced. Example with $m = 4$:

$$f(m) = 64, \tag{15}$$

whereas for the case of brute force search the number of calls of this function is $F(m) = m! \cdot m = 96$. For a larger number of targets $m = 10$ respectively there are

$$\begin{aligned} f(m) &= 9\,864\,100, \\ F(m) &= 36\,288\,000. \end{aligned} \tag{16}$$

Saving of the current state of the plan already significantly reduces the number of computations. However, the main gain in the efficiency of the proposed algorithm is due to its problem orientation specifics and the possibility of discarding non-optimal chains of plans, whose criterion values is worse than ones of the currently saved plan. The algorithm consists of the following sequence of actions.

Algorithm 1. Finding a traverse plan.

(1) Targets are sorted by danger K_i.

(2) An auxiliary search matrix (matrix with index 1 in Table 1) is filled in and used to form a sequence of the plans.

(3) At each new step of the algorithm, the transition through the states of the full plan search table $1 \dots m!$ (Table 1) is performed according to the criterion (10).

(4) The first case: an intercept plan for all targets has not yet been found. In this case:

  (a) The transition in the plan search table is performed according to the parameter of the number of intercepted targets.

  (b) If the next plan is impossible to complete (the target reaches the origin) and the new considered plan has the same number of intercepted targets, the distance and time criteria are checked and the best plan is stored in memory.

  (c) The branch may be discarded if the number of targets missed at the origin has become worse with respect to the saved plan.

(5) The second case: if a plan that intercepts all targets is found. Then:

  (a) Any missed target in the new plan leads to the end of consideration of the current chain of plans.

  (b) If the new plan intercepts all targets, the distance and time criteria are checked and the better of the two plans is stored in memory.

(6) The last saved plan is the optimal plan.

The proposed initial sorting of targets by danger is used to discard non-optimal plan chains in the early stages of the Algorithm 1.

5. MODELLING AND RESULTS DISCUSSION

The interception Algorithm 1 was implemented in Matlab using the functions (5) and (6). Modelling has shown that the running time of the algorithm is acceptable for real-time applications and is strongly reduced relative to the brute-force algorithm. For 1000 experiments, the running time was 200 s, which means that the average running time for one initial state is 0.2 s.

1000 different initial states are considered, for which the following basic parameters are chosen:

- The number of targets is $m = 15$.
- The central angle of the sector where the targets are located is $\alpha = 60^\circ$.
- The values $||\mathbf{r}_j||, j = 1, \dots, m$ are uniformly distributed in $[800,\ 1000]$.
- Target velocities are uniformly distributed in $[0.5V,\ 0.7V]$.

For each state, the danger, convenience, and optimal traverse plans are found according to the criteria $J_T[\pi]$ and $J_{DT}[\pi]$ using the Algorithm 1. Tables 2 and 3 give statistics on how often the first few objectives of the optimal plan turn out to be the most dangerous/convenient.

Tables 2 and 3 show that the statistics of selecting the first target in the plan differs from the statistics in the following steps, since the initial state is significantly different from the state that arise after each interception. In more than 70% of cases, according to the obtained statistics, the first target of the optimal plan matches with the most dangerous or the most convenient target, which can be used to construct greedy algorithms based on local rules according to danger or convenience instead of brute-force algorithms.

It was found for 1000 initial settings that in 24.6% of cases, the optimal plan π^* by criterion $J_{DT}[\pi^*]$ coincides with the optimal plan by criterion $J_T[\pi^*]$.

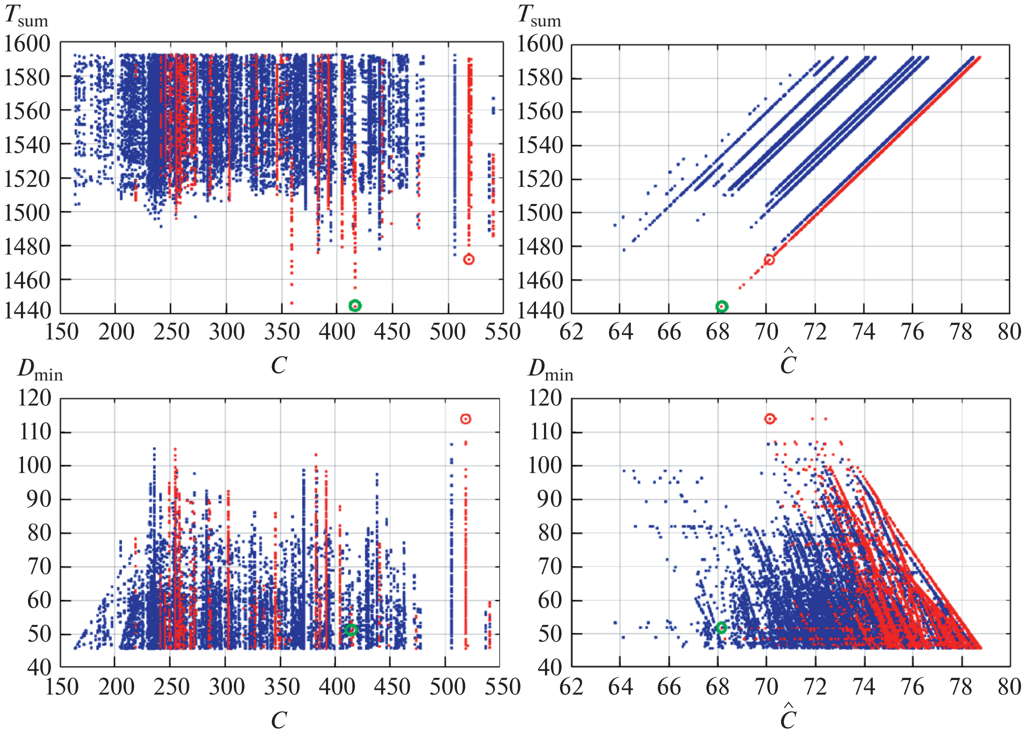


Fig. 2. Execution time of the plan and minimum interception distance for acceptable plans of the same initial situation depending on the complexity and average complexity of the plan.

Further modelling is devoted to investigating plans for a single initial state. Figure 2 presents the dependences $T_{\text{sum}}[\pi], D_{\text{min}}[\pi]$ from $C[\pi], \widehat{C}[\pi]$ for acceptable π plans, whose minimum interception distance $D_{\text{min}}[\pi] > 0.6D_{\text{min}}[\pi^*]$, where π^* is the $J_{DT}[\pi]$ optimal plan. The values $T_{\text{sum}}[\pi^*], D_{\text{min}}[\pi^*]$ are additionally circled in red, and all points $\{\pi: \ \pi_1 = \pi_1^*\}$ corresponding to acceptable plans are also highlighted in red.

Table 2. Percentage of matches of the first four objectives $\pi_1^*, \pi_2^*, \pi_3^*, \pi_4^*$ of the optimal $J_T[\pi]$ plan π^* with the corresponding dangerous and convenient targets for 1000 different initial state

Target numbers in the plan π^*	Number of matches of i target of the plan π^* with i according to			
	danger ($K_{\pi_i^*} = K_{j_i}$), %	convenience ($U_{\pi_i^*} = U_{j_i}$), %	danger and convenience, %	danger or convenience, %
First target π_1^* $(i = 1)$	65.0	65.9	56.8	74.1
Second target π_2^* $(i = 2)$	32.1	57.6	13.9	75.8
Third target π_3^* $(i = 3)$	19.4	58.5	7.1	70.8
Fourth target π_4^* $(i = 4)$	16.3	57.0	3.8	69.5

Table 3. Percentage of matches of the first four objectives $\pi_1^*, \pi_2^*, \pi_3^*, \pi_4^*$ of the optimal $J_{DT}[\pi]$ plan π^* with the corresponding dangerous and convenient targets for 1000 different initial state

Target numbers in the plan π^*	Number of matches of i target of the plan π^* with i according to			
	danger ($K_{\pi_i^*} = K_{j_i}$), %	convenience ($U_{\pi_i^*} = U_{j_i}$), %	danger and convenience, %	danger or convenience, %
First targe π_1^* $(i = 1)$	61.2	63.0	52.8	71.4
Second targe π_2^* $(i = 2)$	36.6	52.9	13.2	76.3
Third targe π_3^* $(i = 3)$	27.5	49.2	7.7	69.0
Fourth targe π_4^* $(i = 4)$	24.6	45.7	6.2	64.1

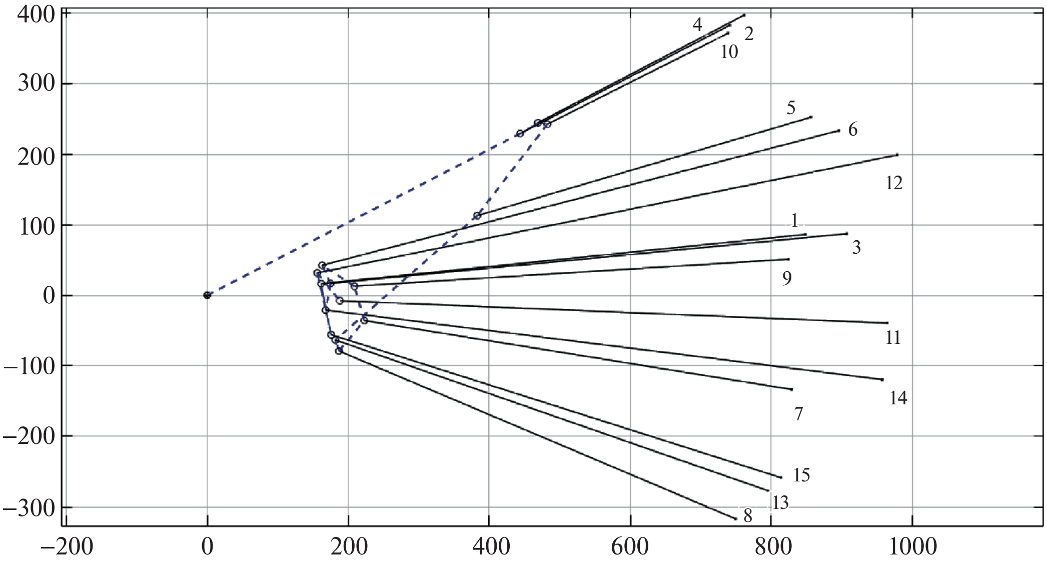


Fig. 3. Optimal interception plan of 15 targets by TS according to the criterion $J_{DT}[\pi] : \pi^* = (2, 4, 10, 5, 13, 8, 7, 9, 6, 3, 14, 1, 15, 12, 11)$, $T_{\text{sum}}[\pi^*] = 1471.049$, $D_{\min}[\pi^*] = 159.168$.

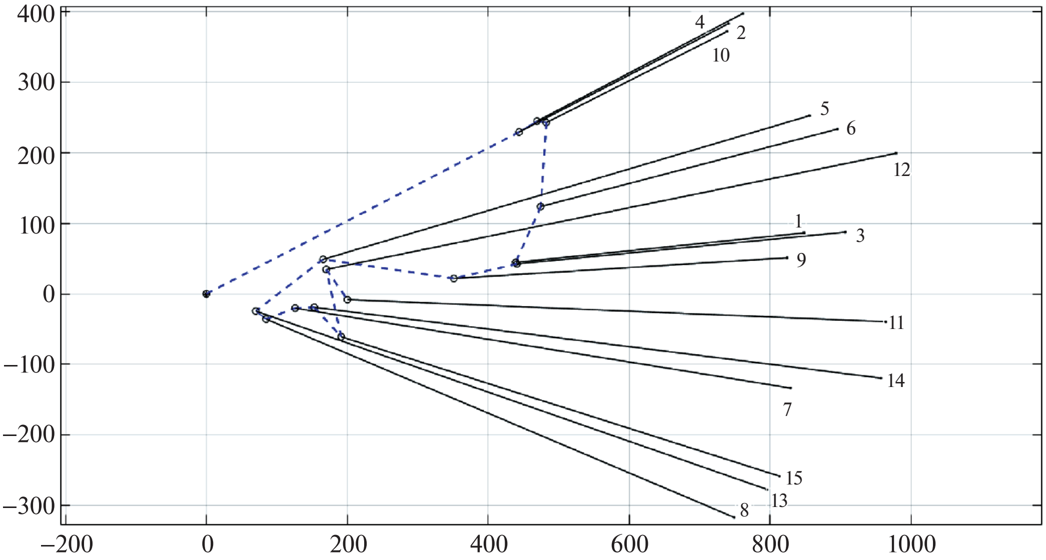


Fig. 4. Optimal interception plan of 15 targets by TS according to the criterion $J_T[\pi] : \pi^* = (2, 4, 10, 6, 1, 3, 9, 5, 13, 8, 7, 14, 15, 12, 11)$, $T_{\text{sum}}[\pi^*] = 1448.051$, $D_{\min}[\pi^*] = 74.183$.

The graph shows that the execution times of all acceptable plans are linearly dependent on $\widehat{C}[\pi]$, which makes it possible to create an optimal polynomial algorithm for constructing an intercept plan in the considered problem. The green circles on all the graphs indicate the optimal J_T plan, the execution time of which is less than $T_{\text{sum}}[\pi^*]$ by 2%, and this plan is worse than π^* by 54% according to the minimum distance of approaching the targets to the origin.

The four acceptable plans $\pi^*, \pi^1, \pi^2, \pi^3$ with the same value $D_{\min} = 113.8$, are analyzed in Fig. 2 on the left, are labeled with a single dot, and on the right they are separated by mean complexity values. Table 4 shows the considered target traverse plans in clear form, where their common part is highlighted.

Table 4. Acceptable plans with $D_{\min}[\pi] = 113.8$

i	π^i		$T_{\text{sum}}[\pi]$	$D_{\min}[\pi]$	$C[\pi]$	$\widehat{C}[\pi]$
π^*	$\{10, 1, 6, 12, 2, 13, 3, 5, 15,$	$8,\ 11, 9, 4, 14, 7\}$	1472	113.8	518.8	70.1
1		$8,\ 11, 4, 9, 14, 7\}$	1496	113.8	518.8	71.9
2		$11, 8,\ 4, 9, 14, 7\}$	1504	113.8	518.8	72.4
3		$11, 8,\ 9, 4, 14, 7\}$	1475	113.8	518.8	70.4

Table 4 shows that the maximum time between intercepts C and $D_{\min}$ in these plans were achieved in the general plan section. The differing sequences of targets finalizing the plans, however, resulted in a change in the average complexity of each plan.

The modeling section is completed with an example of constructing two optimal intercept plans according to the criteria $J_{DT}[\pi]$ and $J_T[\pi]$ in Figs. 3 and 4 for the same initial state, where the trajectory of TS is highlighted by the blue dashed line.

Optimization according to the criterion $J_T[\pi]$ leads to a slight improvement in the execution time of the plan compared to the optimal $J_{DT}[\pi]$ plan, but at the same time the value of $D_{\min}[\pi]$ is more than halved.

6. CONCLUSION

In this paper the problem of intercepting of a set of rectilinearly moving targets by a single interceptor was considered. New macro characteristics of the problem were proposed and their influence on the construction of the optimal intercept plan for different initial states was statistically investigated. An optimal plan finding algorithm based on intelligent brute-force search and dynamic programming concepts was proposed to collect statistics in adequate time. The impact of the new quantities on mission success is shown on the collected statistics and conclusions are made about its applicability to the creation of fast greedy intercept algorithms.

There are plans to investigate various local rules that take into account the state information and geometric characteristics of the target distribution, build suboptimal intercept algorithms based on them, and compare them with the brute-force optimal algorithms.

FUNDING

The work was supported by the Russian Science Foundation under grant no 23-19-00134.

REFERENCES

1. Siharulidze, G.G., On one generalization of the traveling salesman problem. I, *Avtom. Telemekh.*, 1971, no. 8, pp. 116–123.
2. Siharulidze, G.G., On one generalization of the traveling salesman problem. II, *Avtom. Telemekh.*, 1971, no. 10, pp. 142–147.
3. Picard, J.C. and Queyranne, M., The time-dependent traveling salesman problem and its application to the tardiness problem in one-machine scheduling, *Oper. Res.*, 1978, vol. 26, no. 1, pp. 86–110.
4. Helvig, C.S., Robins, G., and Zelikovsky, A., The moving-target traveling salesman problem, *J. Algorithm. Comput. Technol.*, 2003, vol. 49, no. 1, pp. 153–174.
5. Garey, M.R. and Johnson, D.S., Computers and intractability: A guide to the theory of NP-completeness, 1979, San Francisco, California: W. H. Freeman & Co.
6. Ny, J., Feron, E., and Frazzoli, E., On the Dubins traveling salesman problem, *IEEE Trans. Automat. Control*, 2012, vol. 57, no. 1, pp. 265–270.

7. Isaiah, P. and Shima, T., Motion planning algorithms for the Dubins travelling salesperson problem, *Automatica*, 2015, vol. 53, pp. 247–255.

8. Stieber, A., The multiple traveling salesperson problem with moving targets, *Brandenburg University of Technology*, Cottbus-Senftenberg, 2022.

9. Ahrens, B., The tour construction framework for the dynamic Travelling Salesman Problem, *SoutheastCon, IEEE*, 2015, pp. 1–8.

10. Choubey, N.S., Moving target travelling salesman problem using genetic algorithm, *Int. J. Comput. Appl.*, 2013, vol. 70, no. 1, pp. 30–34.

11. Smith, C.D., Assessment of genetic algorithm based assignment strategies for unmanned systems using the multiple traveling salesman problem with moving targets, *Thesis (M.S.), Department of Civil and Mechanical Engineering, University of Missouri*, Kansas City, 2021.

12. Buzikov, M.E. and Galyaev, A.A., Minimum-time lateral interception of a moving target by a Dubins car, *Automatica*, 2022, vol. 135.

13. Galyaev, A.A., Lysenko, P.V., and Rubinovich, E.Y., Optimal Stochastic Control in the Interception Problem of a Randomly Tacking Vehicle, *Mathematics*, 2021, vol. 9, no. 19.

14. Galyaev, A.A., Dobrovidov, A.V., Lysenko, P.V., Shaikin, M.E., and Yakhno, V.P., Path Planning in Threat Environment for UUV with Non-Uniform Radiation Pattern, *Sensors*, 2020, vol. 20, no. 7.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 10, pp. 1168–1176.

TOPICAL ISSUE

Terminal Control of Center of Mass Motion and Propellant Consumption in Liquid-Propellant Rocket Carriers

V. P. Ivanov[*,a], **V. K. Zavadskiy**[*,a], **A. A. Muranov**[*,a], **A. I. Chadaev**[*,a], **E. B. Kablova**[*,a], **L. G. Klenovaya**[*,a], **and I. E. Tropova**[*,a]

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

Received June 19, 2023
Revised July 14, 2023
Accepted August 2, 2023

Abstract—We dedicate this work to the memory of academician B.N. Petrov. It develops the principles of terminal control of rocket carriers formulated by him. Next-generation rocket carriers implement the principle of interconnected, coordinated terminal control of the center of mass motion and propellant consumption. In this article we consider the problem of synthesizing such control and the main principles of its implementation.

Keywords: terminal control, predictive model

DOI: 10.25728/arcRAS.2023.31.64.001

1. INTRODUCTION

The beginning of B.N. Petrov's creative activity coincided with the time when our war-exhausted country made a gigantic breakthrough, opening the way to space for humanity. Soviet science played an important role in this breakthrough. Many of the problems related to the creation of rocket carriers belong to automatic control of mobile objects. B.N. Petrov's profound knowledge in this field and his erudition allowed him to actively participate in the development of new unique automatic control problems and in the development and discussion of our country's space programs alongside leading figures in rocket and space science and technology.

He rightfully became one of the founders of domestic cosmonautics, working for many years in close contact with S.P. Korolev, V.P. Glushko, M.K. Yangel, V.N. Chelomey, V.F. Utkin, and N.A. Pilyugin.

The results of B.N. Petrov and Institute of Dynamics Research, which he headed in the development of methods of modeling and regulating liquid rocket engine thrust and propellant component ratio, are used in many onboard terminal systems. These systems significantly increase the energy of rockets by dramatically reducing the guaranteed propellant reserves. The book by Chertok "Rockets and People" [1] notes the significance of this work.

Understanding the specifics of onboard terminal systems and the peculiarities of organizing control processes allowed B.N. Petrov and his students to classify these systems as a separate class among other automatic control systems. The monograph "Onboard Terminal Control Systems" [2] develops the principles and elements of the theory of this class of systems.

The ideas of B.N. Petrov have further evolved and been applied in modern developments of the Institute in the field of rocket and space technology, resulting in the creation of terminal control systems for new-generation rocket carriers and booster blocks for space and defense purposes (upgraded carrier rockets “Soyuz-2”, “Angara” rocket family, “Sarmat” rocket, rocket boosters under development “Soyuz-5”, “Amur”, and the KVTK booster block).

Next-generation rocket carriers implement the principle of interconnected, coordinated terminal control of the center of mass motion and propellant consumption. In this article we consider the problem of synthesizing such control and the main principles of its implementation.

2. PROBLEM STATEMENT

Consider the control of the center of mass motion of the rocket carrier in the exoatmospheric flight phase.

To simplify, we assume the following:

—Aerodynamic forces are absent,

—The Earth’s gravitational field is parallel to the surface and the acceleration of the gravitational force is constant at all altitudes $\vec{g} = \text{const}$.

—Rotation of Earth is neglected.

The motion of the center of mass of the rocket stage in the longitudinal plane (the plane of the trajectory) in the exoatmospheric flight phase is described by the following equations:

$$\begin{cases} \dot{V}_x = \dfrac{P}{m_\kappa + m}\cos(\vartheta), \quad \dot{V}_y = \dfrac{P}{m_\kappa + m}\sin(\vartheta) - g, \quad P = wr, \\ \dot{x} = V_x, \quad \dot{y} = V_y, \quad \dot{m} = -r, \\ \dot{\vartheta} = \omega, \\ \dot{\omega} = \varphi(\vartheta, \omega, \vartheta_{\text{des}}), \end{cases} \tag{1}$$

where x, y are horizontal and vertical coordinates, m is propellant mass, m_κ is dry mass of the stage, r is propellant consumption rate, w is specific exhaust velocity, P is engine thrust, g is acceleration of gravity, ϑ is pitch angle, ϑ_{des} is control input (desired value of ϑ) for changing the pitch angle, V_x, V_y are horizontal and vertical velocity components.

The equation for the pitch angle ϑ and the angular velocity ω in (1) simplistically describes the operation of the stabilization system.

Coordinates x, y, m, ϑ, and their derivatives are functions of time t, $t \in [t_0, t_k]$, t_k is the terminal time.

Note that the pitch angle ϑ converges to the value $\vartheta_{\text{des}}(t)$ in a significantly shorter time than t_k.

For the final stage, reaching the specified altitude with zero vertical velocity is required:

$$\begin{cases} y(t_k) = y_k, \\ V_y(t_k) = 0. \end{cases} \tag{2}$$

No conditions are set for the horizontal velocity component. Solving the problem assumes maximizing the horizontal component.

For the lower stages of the rocket, we state the problem of hitting the designated burned-out stage impact areas. In this case, we can determine boundary conditions for deviation of the flight range L of the burned-out stage due to deviations of the motion coordinates at the end of the flight from the target values:

$$\delta L = \zeta_x(x(t_k) - x_k) + \zeta_y(y(t_k) - y_k) + \zeta_{Vx}(V_x(t_k) - V_{xk}) + \zeta_{Vy}(V_y(t_k) - V_{yk}) = 0, \tag{3}$$

where ζ_x, ζ_y, ζ_{Vx}, ζ_{Vy} are partial derivatives of δL with respect to the motion coordinates, δ is deviation of the flight range from the target value.

We can write the equations determining the apparent velocity change and the engine propellant consumption processes in the following form:

$$\begin{aligned}
&\dot{W} = \frac{rg}{m_\kappa + m} P_{sp}, \quad P_{sp} = \frac{w}{g}, \quad m = m_o + m_f, \quad r = r_o + r_f, \\
&\dot{m}_o = -r_o, \quad \dot{m}_f = -r_f, \quad K_m = \frac{\dot{m}_o}{\dot{m}_f}, \quad P_{sp} = \varphi(K_m), \\
&\dot{r}_o = f_o(r_o, \alpha_{K_m}, \alpha_R), \quad \dot{r}_f = f_f(r_f, \alpha_{K_m}, \alpha_R),
\end{aligned} \tag{4}$$

with initial conditions accounting for fueling errors and pre-launch propellant component consumption scatter at the moment of the fuel consumption control system activation $m_o(t_0)$, $m_f(t_0)$.

Here, m_o, m_f are the oxidizer and fuel masses, P_{sp} is the specific thrust of the propulsion system, r_o, r_f are the propellant consumption rates determined by the engine equations, α_{K_m}, α_R are the positions of the engine control devices determined by the specified values of the propellant component consumption ratio coefficient K_m and thrust engine condition R.

All coordinates W, m_o, m_f, r_o, r_f and their derivatives are functions of time. We consider them on a bounded time interval t, $t \in [t_0, t_k]$, t_k is terminal time.

The positions of the engine control devices that result in the desired values of the fuel component consumption ratio coefficient $K_m(t)$ and the thrust mode $R(t)$ for engine operation are determined by static nonlinear engine equations:

$$\alpha_{K_m}(t) = f_{K_m}(K_m(t), R(t)), \quad \alpha_R(t) = f_R(K_m(t), R(t)), \quad R(t) = \frac{P(t)}{P_{\text{nom}}}(t).$$

We assume here that $K_m(t)$ is calculated in the algorithm of the terminal control system for object (3), and $R(t)$ is determined by the specified thrust program.

Note that the transients of the propellant consumption rate r_o, r_f in response to the position change of the engine control devices $\alpha_{K_m}(t)$, $\alpha_R(t)$ conclude in a time significantly shorter than t_k.

We impose constraints on the value of the fuel component consumption ratio coefficient that can change during control. We determine the boundary values based on the conditions for stable engine operation and significantly depend on the thrust mode: $K_{m\,\min}(R,t) \leqslant K_m(t) \leqslant K_{m\,\max}(R,t)$.

In this case, we impose final terminal conditions on the remaining propellant components at the moment of engine shutdown and determine them based on the requirements for safe engine shutdown. We specify the conditions as inequalities meaning the necessity of positive values of the remaining propellant components at the moment of engine shutdown, generated by the control system, relative to the propellant level that ensures a safe engine shutdown:

$$m_o(t_k) - m_{o\min} > 0, \quad m_f(t_k) - m_{f\min} > 0. \tag{5}$$

Here $m_{o\min}$, $m_{f\min}$ are the remaining propellant components that are not spent due to the intake design features and accounting for the control system errors.

We include the values $m_{o\min}$, $m_{f\min}$ in m_κ. We understand $m(t)$, $m_o(t)$, $m_f(t)$ as the values of the mass of the current propellant component excluding $m_{o\min}$, $m_{f\min}$.

Let's define the vector of residuals of the specified boundary conditions (2), (3), (5) for the terminal problem solution, and the vector of control inputs:

$$\begin{aligned}
&z_0 = (y(t_k) - y_k, V_y(t_k), m_o(t_k), m_f(t_k)) \text{ — for the terminal stage,} \\
&z_0 = (\delta L, m_o(t_k), m_f(t_k)) \text{ — for the bottom stages,} \\
&u = (\vartheta_{\text{des}}, K_m, t_k).
\end{aligned} \tag{6}$$

Note that the value of R, which determines the engine thrust program, is a specified function of time and is not included in the vector of control inputs u. The terminal time moment t_k can vary and can be used as a control parameter to solve the terminal problem.

The main objective of terminal control is to minimize the residuals of the boundary conditions. In addition to satisfying the boundary conditions, terminal systems also have other requirements, the physical content of which can include energy resource costs, time costs, and control losses. In this work, we limit the problem of criterion synthesis to boundary conditions, the fulfillment of which is a priority.

The control object of the considered terminal system, in terms of transition to the specified final state, is quite inertial (it represents integrating elements).

We achieve control of these processes by influencing other coordinates of the object ϑ, r_o, r_f with rapidly decaying dynamics of their transients. The essence of such control lies in setting the desired steady-state values of these coordinates.

Control of the coordinates ϑ, r_o, r_f (by changing the positions of actuators, drives, fins, etc.) consists of stabilizing these coordinates of the object relative to the specified values determined by the vector $u(t)$. The operation of the closed stabilization loop is simplified by a system of equations for ϑ, r_o, r_f.

In this case, we consider the operation of the stabilization loop in terms of transient responses to changes in the control input. We assume that the transient process is completed in an interval significantly shorter than the terminal control interval.

3. CONTROL ALGORITHM SYNTHESIS IN THE CLASS OF PIECEWISE-CONSTANT FUNCTIONS OF PREDICTED RESIDUALS OF THE TERMINAL CONDITIONS

Let us consider the object terminal control problem (1), (4) within the class of predictive model systems.

Let us integrate (1) on prediction interval $\tau \in [t, t_{\text{command}}]$, where t_{command} is the predicted value of the terminal time moment. We define the current initial rocket center of mass coordinates x, y, V_x, V_y at time t in the inertial navigation system. We substitute propellant mass $m(t)$ equation in (1) with $m_{\text{mod}}(t)$ formed in the propellant management algorithm:

$$\begin{aligned} \dot{m}_{\text{mod}}(t) &= r_{\text{mod}}(t), \\ r_{\text{mod}}(t) &= r_{\text{cycl}}(t)(1 + \lambda(t)), \end{aligned} \tag{7}$$

where r_{cycl} is cumulative propellant consumption corresponding to a given cyclogram of the engine's thrust operation mode, $\lambda(t)$ is controlled parameter of the model that corrects $r_{\text{cycl}}(t)$ in the propellant consumption model. The physical analog of $\lambda(t)$ is the relative deviation of the cumulative consumption from its nominal value.

Note that the cumulative propellant consumption value corresponding to a given cyclogram of the engine's thrust operation mode ($r_{\text{cycl}}(t)$) can be determined based on measurements of apparent acceleration and equation for $\dot{W}$ in (4).

We integrate (1), (7) in the interval $\tau \in [t, t_{\text{command}}]$ with the assumption that $\vartheta(\tau) = \vartheta(t)$, $r(\tau) = r_{\text{cycl}}(\tau)(1 + \lambda(t))$, $m(t) = m_{\text{mod}}(t)$.

Let us define t_{command} from condition $m_{\text{mod}}(t) - \int_t^{t_{\text{command}}} r_{\text{mod}}(\tau) d\tau = 0$.

Let us define the values of the predicted residuals $y(t_{\text{command}}) - y_k$, $V_y(t_{\text{command}})$, $\delta L(t_{\text{command}})$.

When integrating (1) we can use the integral expressions presented in [3].

We take the time moment t such that $m_{\text{mod}}(t) = 0$ is the value of the terminal time moment t_k (engine shutdown).

Regarding the management of propellant components, the predictive model includes equation (7) and equations of the processes of change of the mass of propellant components (4). Taking into account the interdependence of equation (7) with (4), let us define the equations for the deviations of the current values of oxidizer and fuel masses from the model analogues, formed from the model value of the total propellant mass according to the nominal value of ratio coefficient K_m:

$$\begin{aligned}\Delta m_{\text{o}}(t) &= m_{\text{o}}(t) - m_{\text{mod}}(t)\frac{K_{m\,\text{nom}}}{K_{m\,\text{nom}}+1},\\ \Delta m_{\text{f}}(t) &= m_{\text{f}}(t) - m_{\text{mod}}(t)\frac{1}{K_{m\,\text{nom}}+1},\end{aligned} \tag{8}$$

where $m_{\text{o}}(t)$, $m_{\text{f}}(t)$ are determined based on measurements of discrete level sensors in tanks.

For the deviations (8), we can obtain equations of the following form:

$$\begin{aligned}\Delta \dot m_{\text{o}}(t) &= r_{\text{o}}(t) - r_{\text{mod}}(t)\frac{K_{m\,\text{nom}}}{K_{m\,\text{nom}}+1},\\ \Delta \dot m_{\text{f}}(t) &= r_{\text{f}}(t) - r_{\text{mod}}(t)\frac{1}{K_{m\,\text{nom}}+1}.\end{aligned} \tag{9}$$

Let us integrate equations (9) in interval $\tau \in [t, t_{\text{command}}]$ assuming $r_{\text{o}}(\tau) = r_{\text{o}}(t)$, $r_{\text{f}}(\tau) = r_{\text{f}}(t)$, $r_{\text{mod}}(\tau) = r_{\text{cycl}}(\tau)(1 + \lambda(t))$, and initial conditions $\Delta m_{\text{o}}(t)$, $\Delta m_{\text{f}}(t)$.

Determine the values of the predicted residuals $\Delta m_{\text{o}}(t_{\text{command}})$, $\Delta m_{\text{f}}(t_{\text{command}})$.

Due to predictive model of the object (1), (4), the vector of predicted boundary condition residuals (6) is defined as

$$\begin{aligned}z(t) &= (y_{pr}(t_{\text{command}}) - y_k, V_{ypr}(t_{\text{command}}), \Delta m_o(t_{\text{command}}), \Delta m_f(t_{\text{command}}))\\ &\qquad\text{— for the terminal rocket stage,}\\ z(t) &= (\delta L, \Delta m_o(t_{\text{command}}), m_f(t_{\text{command}})) \text{ — for the bottom rocket stage,}\end{aligned} \tag{10}$$

and the vector of control inputs in form $u = (\vartheta_{\text{des}}, K_m, \lambda)$.

If $t \to t_k$, $t_{\text{command}} \to t_k$, $z(t) \to z_0$.

We solve the problem of terminal control of object (1), (4) by forming feedback control based on predicted boundary condition residuals (10).

Let $x_T(t) = (x(t), y(t), V_x(t), V_y(t), \Delta m_o(t), \Delta m_f(t), m_{\text{mod}}(t))$ be the vector of coordinates of the predicted model of the object (1), (7), (9) supplemented with equations for $\dot m_o, \dot m_f, \dot r_o, \dot r_f$, which determine the boundary condition residuals, and let $x_u(t) = (\vartheta(t), r_o(t), r_f(t), \lambda(t))$ be the vector of coordinates directly influenced by the control inputs.

As shown in [4, 5], we determine the derivative with respect to time and the differential equation for the vector of predicted boundary condition residuals $z(t)$ by differentiating $z(t)$ as a composite function:

$$\frac{dz(t)}{dt} = \frac{\partial z(t)}{\partial x_T(t_{\text{command}})}\left[\frac{\partial x_T(t_{\text{command}})}{\partial x_u(t)}\frac{dx_u(t)}{d(t)} + \frac{dt_{\text{command}}}{dt}\frac{dx_T(t_{\text{command}})}{d(t)}\right].$$

We choose control inputs $\vartheta_{\text{des}}(t)$, $K_m(t)$, $\lambda(t)$ from the class of piecewise-constant functions of time. The control input for the pitch angle ϑ_{des} changes discretely at moments in time when the information is updated from the inertial navigation system. The control inputs $K_m(t)$ and $\lambda(t)$

for the fuel consumption processes change at discrete moments in time when the levels of the components in the tanks are measured. At these same moments transient processes for $r_o(t)$, $r_f(t)$ appear and the quantities $r_{\text{mod}}(t)$ and $t_{\text{command}}(t)$ change abruptly.

For piecewise-constant control, we can obtain the difference equations for $z(t)$ from the differential equations. We introduce notation for the components of the residuals vector:

$$z_y(t) = y_{pr}(t_{\text{command}}) - y_k, \quad z_V(t) = V_{ypr}(t_{\text{command}}),$$
$$z_{m_o}(t) = \Delta m_o(t_{\text{command}}), \quad z_{m_f}(t) = \Delta m_f(t_{\text{command}}), \quad z_\delta(t) = \delta(t_{\text{command}}).$$

We can express the difference equations for the components of the vector $z(t)$ as follows. In terms of controlling the motion of the center of mass, the difference equations are determined for discrete moments in time t_i when the navigation information is updated $i = 0, 1, 2, \ldots, I$, $t_{I+1} = t_k$ (when $\lambda(t) = \text{const}$, $t_{\text{command}}(t) = \text{const}$):

$$\begin{aligned} z_y(t_{i+1}) &= z_y(t_i) + \frac{\partial z_y}{\partial \vartheta}(t_i)\Delta\vartheta_i, \\ z_{Vy}(t_{i+1}) &= z_{Vy}(t_i) + \frac{\partial z_{Vy}}{\partial \vartheta}(t_i)\Delta\vartheta_i, \\ z_\delta(t_{i+1}) &= z_\delta(t_i) + \frac{\partial z_\delta}{\partial \vartheta}(t_i)\Delta\vartheta. \end{aligned} \tag{11}$$

Here

$$\Delta\vartheta_i = \int\limits_{t_i}^{t_i+\delta t} \dot{\vartheta}(\tau)d\tau,$$

where δt is time interval of the transient in object (1) with respect to coordinate ϑ during an abrupt control input ϑ_{des} change at time momentt_i.

Furthermore, at time moments t_j of discrete measurement of the propellant level in tanks, the aforementioned residuals change due to changes in $\lambda(t)$, $t_{\text{command}}(t)$.

Let us assume that the level sensors conduct discrete measurements at one of the discrete time moments of navigational information update $t_j = t_i$. Let us add terms accounting for abrupt changes of $\lambda(t)$ and $t_{\text{command}}(t)$ to (11):

$$\begin{aligned} z_y(t_{i+1}) &= z_y(t_i) + \frac{\partial z_y}{\partial \vartheta}(t_i)\Delta\vartheta_i + \frac{\partial z_y}{\partial r_{\text{mod}}}(t_i) r_{\text{cycl}}(t_j)\Delta\lambda_j + \Delta t_{\text{command}j}\dot{y}(t_{\text{command}}), \\ z_{V_y}(t_{i+1}) &= z_{V_y}(t_i) + \frac{\partial z_{V_y}}{\partial \vartheta}(t_i)\Delta\vartheta_i + \frac{\partial z_{V_y}}{\partial r_{\text{mod}}}(t_i) r_{\text{cycl}}(t_j)\Delta\lambda_j + \Delta t_{\text{command}j}\dot{V}_y(t_{\text{command}}), \\ z_\delta(t_{i+1}) &= z_\delta(t_i) + \frac{\partial z_\delta}{\partial \vartheta}(t_i)\Delta\vartheta_i + \frac{\partial z_\delta}{\partial r_{\text{mod}}}(t_i) r_{\text{cycl}}(t_j)\Delta\lambda_j \\ &\quad + \Delta t_{\text{command}j}(\zeta_x \dot{x}(t_{\text{command}}) + \zeta_y \dot{y}(t_{\text{command}}) + \zeta_{Vx}\dot{V}_x(t_{\text{command}}) + \zeta_{Vy}\dot{V}_y(t_{\text{command}})). \end{aligned} \tag{12}$$

Here $\Delta t_{\text{command}j}$ is difference of $t_{\text{command}j}$ values determined from equation (7) at t_j while $\lambda = \lambda(t_j)$ and $\lambda = \lambda(t_j) + \Delta\lambda_j$. We can determine the value of this difference with the following approximate expression: $\Delta t_{\text{command}j} = \zeta_{tk}(t_j)\Delta\lambda_j$.

In regard to propellant management, we define difference equations for discrete time moments t_j of information update of the level sensors:

$$\begin{gathered}
z_{m_o}(t_{j+1}) = z_{m_o}(t_j) + \frac{\partial z_{m_o}}{\partial r_o}(t_j)\Delta r_{oj} + \frac{\partial z_{m_o}}{\partial r_{\text{mod}}}(t_j) r_{\text{cycl}}(t_j)\Delta\lambda_j \\
+ (r_o(t_j) - r_{\text{mod}}(t_j)\frac{K_{m\,\text{nom}}}{K_{m\,\text{nom}}+1}\zeta_{tk}(t_j)\Delta\lambda_j), \\
z_{m_f}(t_{j+1}) = z_{m_f}(t_j) + \frac{\partial z_{m_f}}{\partial r_f}(t_j)\Delta r_{fj} + \frac{\partial z_{m_f}}{\partial r_{\text{mod}}}(t_j) r_{\text{cycl}}(t_j)\Delta\lambda_j \\
+ (r_f(t_j) - r_{\text{mod}}(t_j)\frac{1}{K_{m\,\text{nom}}+1}\zeta_{tk}(t_j)\Delta\lambda_j),
\end{gathered} \tag{13}$$

Here

$$\Delta r_{oj} = \int\limits_{t_j}^{t_j+\delta t} f_o(r_o, \alpha_{K_m}, \alpha_R)d\tau, \quad \Delta r_{fj} = \int\limits_{t_j}^{t_j+\delta t} f_f(r_f, \alpha_{K_m}, \alpha_R)d\tau,$$

where δt is transient time interval in object (4) with respect to coordinates r_o, r_f during abrupt change of α_{K_m} during implementation of control input $K_m(t)$ at time t_i.

For linearized engine equations under constant thrust mode, the values of propellant component flow increments due to changes in ratio coefficient K_m can be determined with the following expression [5]:

$$\Delta r_{oj} = \frac{\delta r_o(t_j)}{\delta K_m}\Delta K_{mj}, \quad \Delta r_{fj} = \frac{\delta r_f(t_j)}{\delta K_m}\Delta K_{mj}.$$

Let us rephrase the original terminal control problem. Instead of finding control $u(t)$ in the class of piecewise-constant functions, we search for the discrete sequence of coordinate increments $\vartheta(t), K_m(t), \lambda(t)$ at time points t_i, t_j.

Based on difference equations (11)–(13), we define algorithms for forming control input vector $\Delta u = (\Delta\vartheta_i, \Delta K_{mj}, \Delta\lambda_j)$ functions of the predicted boundary condition residuals.

The main disturbance in the terminal problem considered is the unknown initial conditions for the equations of the coordinates of the object (1), (4). The ability to counteract these disturbances when controlling the regions of the lower stage drop depends on the fact that the dimensions of the control vector are equal to the dimensions of the residual vector. When controlling the final stage, the dimensions of the boundary condition vector increase. In this case, to solve the terminal problem, it is necessary to choose the values of the control inputs for two discrete time points. In this case, the number of independent control inputs is larger than the dimensions of the residual vector. As a result of the analysis of possible options to form control inputs for two discrete time points, we adopted the following most obvious control algorithm. Consider a discrete time point t_j.

From the discrete equations (13) for the boundary condition residuals in terms of propellant consumption management, we determine the values of the control inputs $K_m(t_j)$, $\Delta\lambda(t_j)$. The control algorithm for the pitch angle with feedback based on predicted residual values $y_{pr}(t_{\text{command}}) - y_k$, $V_{ypr}(t_{\text{command}})$, which ensures the solution of the terminal problem under the given boundary conditions for the coordinates $y(t_k)$, $V_y(t_k) = 0$, is determined from equations (11), (12) for two discrete time points t_{i+1}, t_{i-p+1}. It should be noted that in the interval $[t_i, t_{i-p+1}]$, the residuals $y_{pr}(t_{\text{command}}) - y_k, V_{ypr}(t_{\text{command}})$ maintain their values unchanged.

The algorithm to control the pitch angle with the feedback predicted from discrepancies $y_{pr}(t_{\text{command}}) - y_k$, $V_{ypr}(t_{\text{command}})$ at discrete time points t_i, t_{i-p} is determined based on equation (12). It takes into account the value $\Delta\lambda(t_i)$, calculated in the propellant consumption control algorithm. The procedure to form this algorithm is described in [4]. In this case, the pitch angle at time t_{i-p} receives an increment $\Delta\vartheta_1$, while at time t_i it changes by an amount $\Delta\vartheta_2$.

The presence of parametric disturbances determines the errors in terminal control. We counter these disturbances by applying an iterative procedure to form the control vector $\Delta u = (\Delta\vartheta_i, \Delta K_{mj}, \Delta\lambda_j)$ with feedback on the vector of the residuals of the predicted boundary condition $z(t)$.

The main result of solving the problem considered of coordinated control of the center of mass motion and propellant consumption is the most complete utilization of available propellant reserves [6]. The essence of such coordinated control is as follows. Information about the current propellant mass is generated in accordance with (7), where $\lambda(t)$ is determined taking into account the measurements of the level sensors. We take this into account when predicting the discrepancies in the center of mass trajectory coordinates corresponding to the target of escape. In this case, equations (12) for $z_y(t_{j+1})$, $z_V(t_{j+1})$, $z_\delta(t_{j+1})$ include disturbance $\Delta\lambda_j$. By burning additional propellant, the final value of apparent velocity $W(t_{\text{command}})$ increases. The resulting error in the impact area is eliminated by varying the velocity in the neutral direction through additional pitch angle control. Note that the effectiveness of such control is maintained until the pitch angle approaches the value at which the maximum range of the spent stage is ensured.

Without taking into account the actual current value of the fuel mass in controlling the motion of the center of mass, the terminal time t_{command} is determined by the zero discrepancy in the coordinates of the trajectory. In this case, the effects of disturbing factors such as deviations in initial mass, propellant consumption, etc., on the trajectory that are countered by controlling the thrust vector up to the moment t_{command}, lead to significant unused propellant residues. The magnitude of these residues can reach 1% of the initial propellant mass.

In the propellant consumption control loop, significant random measurement errors occur when measuring the levels of propellant components in the tanks. As a result, even with error filtering, random control errors occur in the form of component residues at the moment t_k. To counteract these errors, we introduce safety reserve propellant components, which reduces the effectiveness of control. However, the implementation of coordinated terminal control for modern rocket boosters such as Angara and Soyuz-5 reduces unused propellant reserves by a factor of 3.

The principle of coordinated control of the center of mass movement and fuel consumption is implemented in the control algorithms of the Proton-M and Angara rocket boosters.

In foreign counterparts, terminal control of the center of mass movement by influencing the thrust vector and iterative procedures for feedback control based on predicted residuals was developed almost at the same time (at the end of the last century) as in the USSR and later in the Russian Federation. However, coordinated control of the center of mass movement and propellant consumption was not required. Presumably, because there were no strict constraints on the spent stages impact areas.

4. CONCLUSION

1. We consider the problem of synthesizing terminal control of the center of mass movement and propellant consumption for liquid rocket boosters. The control synthesis problem is limited by given boundary conditions, the fulfillment of which is a priority task.

When solving the problem, we assume that the system can be decomposed into interrelated processes of final-state control and object stabilization. Decomposition allows us to reveal the content of control processes in the terminal system. Terminal control is performed by specifying the object coordinate values maintained by the stabilization loop. Stabilization of the object relative to the given values is characterized by fast damping of the dynamics of transient processes. The derivative of the residuals in the decomposed system explicitly depends on the terminal control.

2. We solve the synthesis problem in the class of systems with the prediction of boundary condition residuals, which are vector functions of the current values of the object coordinates and time. We discretized the synthesis problem for control variations in the class of piecewise-constant functions. We obtain difference equations for the vector of predicted residuals. We determined algorithms to form the vector of control actions to change the pitch angle, the proportion of component consumption rates, and the controlled parameter of the object model as functions of the predicted residuals of the boundary condition based on the difference equations obtained.

3. The solution to the considered terminal problem is a jointly coordinated control of the center of mass movement and propellant consumption, ensuring the most complete use of available propellant reserves. The principle of coordinated control of the center of mass movement and fuel component consumption is implemented in the control algorithms of the Proton-M rocket booster and the Angara rocket booster family.

REFERENCES

1. Chertok, B.E., *Rakety i lyudi* (Rockets and People, I), Moscow: Mashinostroenie, 1994.

2. Petrov, B.N., Portnov-Sokolov, Yu.P., Andrienko, A.Ya., and Ivanov, V.P., *Bortovye terminal'nye sistemy upravleniya* (Board Terminal Control Systems), Moscow: Mashinostroenie, 1983.

3. Sikharulidze, Yu.G., *Ballistika i navedenie letatel'nykh apparatov* (Ballistics and Guidance of Aircraft), Moscow: Binom. Laboratoriya Znanii, 2011.

4. Ivanov, V.P. and Tabalin, D.D., On a Deterministic Terminal Control Method with Predictive Forecasting of Mismatches in the Boundary Conditions, —it Autom. Remote Control, 2022, vol. 83, no. 1, pp. 62–77. https://doi.org/10.31857/S0005231022010056

5. Ivanov, V.P., Stamenkovich, N.N., Kablova, E.B., and Klenovaya, L.G., Deterministic Synthesis of Algorithms of Propellant Consumption Control from Tanks of a Liquid Rocket Booster Considering Conditions of Stable Engine Operation, *Trudy FGUP "NPTzAP". Sistemy i pribory upravleniya*, 2020, no.3 (53), pp. 31–41.

6. Ivanov, V.P., Zavadsky, V.K., Gus'kov, A.D., Dishel', V.D., Vasyagina, I.V., and Kislik, V.D., Terminal Control of Rocket Booster Guidance and Propellant Consumption Aiming Dry Condition, *Mezhd. nauch.-tekh. konf. "Sistemy i kompleksy avtom. upr. let. apparat."* (Internat. sci. conf. "Systems and Complexes of Automatic Aircraft Control"), Moscow: OOO "Nauch.-izd. centr "Inzhener", 2008, pp. 56–65.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 10, pp. 1177–1185.

A Comprehensive Software Verification Technology for Onboard Control Systems of Spacecraft

V. V. Kul'ba^{*,a}**, E. A. Mikrin***†**, B. V. Pavlov**^{*,b}**, and S. K. Somov**^{*,c}

Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
e-mail: [a]*kulba@ipu.ru,* [b]*pavlov@ipu.ru,* [c]*ssomov2016@ipu.ru*

Received June 19, 2023
Revised July 18, 2023
Accepted August 2, 2023

Abstract—This paper conceptualizes the main principles of comprehensive software verification for an onboard spacecraft control system. An optimal comprehensive verification strategy for onboard software is selected by rigorously stating and solving the corresponding optimization problem. Software verification methods with functional correctness indicators are proposed.

Keywords: spacecraft, software, onboard control system, comprehensive software verification

DOI: 10.25728/arcRAS.2023.99.27.001

1. INTRODUCTION

A wide variety of tasks performed by spacecraft, extreme conditions of spacecraft operation, and different technologies and protocols of information interaction between onboard hardware elements and software, including various sensors and indicators, determine the need to create new approaches, methods, and technologies to support R&D works in the field of advanced space technology. The papers [1, 5], the preprint [2], and the books [3, 4] were devoted to the development and implementation of modeling methods in the aerospace industry. An important place therein was occupied by the issues of digital modeling, a relevant method for studying different aspects in the operation of onboard spacecraft control systems (OSCSs), including the design, development, and verification of their software [6, 7]. According to the experience of application of modern organizational, methodological, and technical solutions used to verify OSCS software, it is necessary to develop basic principles of a comprehensive verification methodology for OSCS software [6]. The solution of this problem is urgent for the effective development and verification of OSCS software using software prototypes, early functional integration, and the iterative checking of software requirements. This methodology is used to optimize the comprehensive software verification process in terms of time and cost criteria considering various technological constraints.

2. SELECTING AN OPTIMAL COMPREHENSIVE VERIFICATION STRATEGY

Formal problem statements on selecting an optimal comprehensive verification strategy for OSCS software often involve two optimality criteria: the minimum time of verification and the minimum cost of verification. The general problem of selecting an optimal comprehensive verification strategy is to determine the following elements: 1) an optimal partition of the software complex into separate parts, 2) the set of necessary subprograms (mocks and drivers), and 3) a scenario to verify the separate parts of the software complex. When stating the problem, constraints are used to determine the admissible partitions and unions of a special graph $\boldsymbol{\Gamma}$, whose vertices correspond to

[†] Deceased.

the program modules and whose arcs are control links between them. When developing tests and localizing errors, graph models are used to detail the graph $\boldsymbol{\Gamma}$. These models formalize the detailed flowcharts of the software complex and, in addition, the detailed flowcharts of separate program modules (PMs).

The set of different comprehensive verification strategies for OSCS software is defined as follows. At the initial stage, it is necessary to determine the set of all admissible partitions of the graph $\boldsymbol{\Gamma}$ into subgraphs. Autonomous testing is conducted for each resulting subgraph. Next, the set of all admissible unions of the resulting subgraphs is determined. These unions are used for joint software testing. Each comprehensive verification strategy is defined as follows. First, it is the set of subgraphs $p^m = \{p_1, \ldots, p_l, \ldots, p_M\}$ obtained by partitioning the graph $\boldsymbol{\Gamma}$. Second, it depends on the order in which these subgraphs are united. Uniting the subgraphs in a given order yields the original graph structure $\tilde{p}^{mn} = \{\tilde{p}_1^{mn}, \ldots, \tilde{p}_k^{mn}, \ldots, \tilde{p}_N^{mn}\}$, where $\tilde{p}_N^{mn}$ coincides with the graph $\boldsymbol{\Gamma}$.

The objective of selecting an optimal comprehensive verification strategy is to find a partition p^{m^*} of the graph $\boldsymbol{\Gamma}$ and a sequence of uniting the subgraphs $\tilde{p}^{mn^*}$ that, when used together, yield a comprehensive verification scenario with the optimal values of time and cost characteristics of the verification process.

If the verification process involves the mn-strategy, the time and cost of comprehensive verification consist of two components, $\overline{T}^p_{mn}\left(\overline{C}^p_{mn}\right)$ and $\overline{T}^o_{mn}\left(\overline{C}^o_{mn}\right)$. The first component is the time $\overline{T}^p_{mn}$ (and cost $\overline{C}^p_{mn}$) of the autonomous verification of the subgraphs obtained by partitioning the graph $\boldsymbol{\Gamma}$ for the mn-strategy. The second component is the time $\overline{T}^o_{mn}$ (and cost $\overline{C}^o_{mn}$) of implementing the uniting stages for these subgraphs and performing the subsequent joint verification of the subgraphs for the mn-strategy. The time and cost of verifying autonomously the subgraphs obtained by partitioning the graph $\boldsymbol{\Gamma}$ are given by

$$\bar{T}^p_{mn} = \sum_{\nu} t_{\nu mn}, \quad \bar{C}^p_{mn} = \sum_{\nu} C_{\nu mn},$$

where t_{vmn} and C_{vmn} denote the time and cost of the autonomous verification of the νth subgraph of the graph $\boldsymbol{\Gamma}$.

When uniting the subgraphs, the time and cost characteristics of joint verification are given by

$$\overline{T}^o_{mn} = \sum_{k} b_{kmn}, \quad \overline{C}^o_{mn} = \sum_{k} S_{kmn},$$

where b_{kmn} and S_{kmn} denote the time and cost of joint verification at the kth subgraph uniting stage.

The problem of determining an optimal verification strategy with the time criterion has the following general statement: it is required to minimize the expression

$$\sum_{mn}\left(\overline{T}^p_{mn} + \overline{T}^o_{mn}\right) x_{mn}$$

subject to the verification cost constraint

$$\sum_{mn}\left(\overline{C}^p_{mn} + \overline{C}^o_{mn}\right) x_{mn} \leqslant C.$$

In this constraint,

$$x_{mn} = \begin{cases} 1 & \text{if the } mn\text{-strategy is chosen for comprehensive verification} \\ 0 & \text{otherwise.} \end{cases}$$

The constraint also includes a constant $\mathbf{C}$, which specifies the maximum allowable cost of comprehensive verification.

In the process of solving this problem, possible partitions of the graph $\boldsymbol{\Gamma}$ into subgraphs are found by selecting the composition of V groups of the program modules of the software complex, where V is the number of program modules in the software complex of the onboard control system. When solving the problem, it is required to observe the constraints on the admissible combinations of program modules for each of $\mathbf{V}$ groups.

In the course of selecting a union of subgraphs from the subgraph set $p^m = \{p_1, \ldots, p_m, \ldots, p_M\}$ into the original graph $\boldsymbol{\Gamma}$, it is required to determine the list of stages to unite $\mathbf{V}^*$ non-empty subgraphs into $\boldsymbol{\Gamma}$. The maximum number of such stages must be equal to the number of non-empty subgraphs $\mathbf{V}^*$.

However, if the number of program modules in the software complex (the number of software components) is high, then the set of possible system verification strategies becomes very large as well. Due to this fact, estimating the time and cost characteristics of system verification strategies becomes an extremely resource-intensive and time-consuming task. To eliminate this difficulty, we propose to find particular optimal software verification strategies: such problems are most commonly encountered in practice.

We define the set $\overline{P}^p$ of admissible partitions of the graph $\boldsymbol{\Gamma}$ into subgraphs as follows:

$$\overline{P}^p = \{P^m\}, \quad m = \overline{1, M}.$$

Here, $P^m = \{p_1^m, \ldots, p_\nu^m, \ldots, p_{D_m}^m\}$ and p_ν^m denote the mth partition and the νth subgraph, respectively, and D_m is the number of subgraphs in the mth partition.

The set $\overline{P}^o = \{\tilde{P}^{mn}\}$, $(n = \overline{1, N_m},\ m = \overline{1, M})$ defines the admissible unions of the graph $\boldsymbol{\Gamma}$. The element $\tilde{P}^{mn} = \left\{\tilde{p}_1^{mn}, \ldots, \tilde{p}_k^{mn}, \ldots, \tilde{p}_{F_{mn}}^{mn}\right\}$ of the set $\overline{P}^o$ is the nth union under the mth partition of the graph $\boldsymbol{\Gamma}$. The values N_m and F_{mn} are the number of the resulting unions of subgraphs and the number of uniting stages under the mth partition of the graph $\boldsymbol{\Gamma}$.

For the nth union, the element $\tilde{p}_k^{mn}$ is defined as follows:

$$\tilde{p}_k^{mn} = \bigcup_{v \in R1_k^{mn}} p_v^m \bigcup_{i \in R2_k^{mn}} p_i^m.$$

Here, $R1_k^{mn}$ is the index set of the subgraphs from P^m and $R2_k^{mn}$ is the index set of the subgraphs from $\tilde{P}^{mn}$ included in the kth joint verification stage under the mth partition and the nth union of the graph $\boldsymbol{\Gamma}$.

On the one hand, the comprehensive verification strategy is determined by the partition of the graph $\boldsymbol{\Gamma}$ into subgraphs $P^m \in \overline{P}^p$; on the other, by the union of the resulting subgraphs $\tilde{P}^{mn} \in \overline{P}^p$ into the original graph.

The time t_ν and cost C_ν of the autonomous verification of each vth subgraph in a partition consist of three components as follows: the time and cost (t_ν^n, C_ν^n) of preparing test data, the time and cost (t_ν^p, C_ν^p) of executing the testing process, and the time and cost $\left(t_\nu^{\text{loc}}, C_\nu^{\text{loc}}\right)$ of localizing the errors detected during subgraph testing, i.e.,

$$t_\nu = t_\nu^n + t_\nu^p + t_\nu^{\text{loc}}, \quad C_\nu = C_\nu^n + C_\nu^p + C_\nu^{\text{loc}},$$

where

$$t_\nu^n = t_\nu^{\text{gen}} + t_\nu^{\text{mock}} + t_\nu^{\text{dri}}, \quad C_\nu^n = C_\nu^{\text{gen}} + C_\nu^{\text{mock}} + C_\nu^{\text{dri}}.$$

These formulas have the following notations: t_ν^{gen} and C_ν^{gen} are the time and cost of generating test data for the vth subgraph, respectively; t_ν^{mock} and C_ν^{mock} are the time and cost of developing mock subprograms to verify the vth subgraph, respectively; t_{ν}^{dri} and C_{ν}^{dri} are the time and cost of developing driver subprograms to verify the vth subgraph, respectively; t_{ν}^{p} and C_{ν}^{p} is the time and cost of carrying out tests for the vth subgraph, respectively; finally, t_{ν}^{loc} and C_{ν}^{loc} are the time and cost of localizing errors detected when testing the vth subgraph, respectively.

An optimal system verification strategy can be found in two steps as follows. The first step is to select an admissible partition of the graph $\boldsymbol{\Gamma}$ into subgraphs $P^m \in \overline{P}^p$ for their autonomous verification. The second step is to select an admissible union of these subgraphs from the set $\overline{P}^o$ for joint verification. The two steps ensure the comprehensive verification process with minimum time and cost under the existing time and cost constraints.

For the problem statement under consideration, we define the variable

$$y_{mn} = \begin{cases} 1 & \text{if for the } m\text{th partition of the graph } \Gamma \text{ the } n\text{th union is selected} \\ 0 & \text{otherwise.} \end{cases}$$

This problem is solved using the following initial data:

1) the sets $\overline{P}^p = \{P^m\}\,,\ m = \overline{1, M}$ and $\overline{P}^o = \{\tilde{P}^{mn}\},\ n = \overline{1, N_m},\ m = \overline{1, M}$,

2) the time and cost characteristics of the autonomous and joint verification processes.

The time and cost of comprehensive verification are given by

$$\overline{T}^k = \overline{T}_m^p + \overline{T}_{mn}^o, \quad \overline{C}^k = \overline{C}_m^p + \overline{C}_{mn}^o,$$

where $\overline{T}_m^p$ and $\overline{C}_m^p$ denote the time and cost of autonomous verification under the mth partition of the graph $\boldsymbol{\Gamma}$, respectively; $\overline{T}_{mn}^o$ and $\overline{C}_{mn}^o$ denote the time and cost of joint verification under the mth partition of the graph $\boldsymbol{\Gamma}$ and the nth union of the graph Γ.

The time $\overline{T}_m^p$ and cost $\overline{C}_m^p$ of autonomous verification are given by

$$\bar{T}_m^p = \sum_{\nu} \left(t_{\nu m}^{\mathrm{gen}} + t_{\nu m}^{\mathrm{mock}} + t_{\nu m}^{\mathrm{dri}} + t_{\nu m}^{\mathrm{loc}} \right),$$

$$\bar{C}_m^p = \sum_{\nu} \left(c_{\nu m}^{\mathrm{gen}} + c_{\nu m}^{\mathrm{mock}} + c_{\nu m}^{\mathrm{dri}} + c_{\nu m}^{\mathrm{loc}} \right).$$

Let the test sets to debug the subgraphs $p_{\nu}^m \in P^m$ be determined, and let the corresponding time and cost characteristics be known for them. Then the time and cost characteristics of autonomous debugging are calculated as

$$t_{\nu m}^{\mathrm{gen}} = \sum_{j=1}^{J_{\nu m}} t_{j\nu} m^{\mathrm{gen}}, \qquad c_{\nu m}^{\mathrm{gen}} = \sum_{j=1}^{J_{\nu m}} \hat{c}_{j\nu} m^{\mathrm{gen}},$$

$$t_m^{\mathrm{progr}} = \sum_{j=1}^{J_{\nu m}} t_{j\nu} m^{\mathrm{progr}}, \quad c_{\nu m}^{\mathrm{progr}} = \sum_{j=1}^{J_{\nu m}} \hat{c}_{j\nu} m^{\mathrm{progr}},$$

$$t_m^{\mathrm{loc}} = \sum_{j=1}^{J_{\nu m}} t_{j\nu} m^{\mathrm{loc}} \rho, \qquad c_{\nu m}^{\mathrm{loc}} = \sum_{j=1}^{J_{\nu m}} \hat{c}_{j\nu} m^{\mathrm{loc}} \rho.$$

In these formulas, $J_{\nu m}$ is the set of tests to verify the subgraph p_{ν}^m.

The variables $t_{\nu m}^{\mathrm{mock}}$ and $c_{\nu m}^{\mathrm{mock}}$ specify the time and cost of developing all mock subprograms to verify the subgraph p_{ν}^m, i.e.,

$$t_{\nu m}^{\mathrm{mock}} = \sum_{i=1}^{I_{\nu m}} \hat{t}_{i\nu}^{\mathrm{mock}}, \quad c_{\nu m}^{\mathrm{mock}} = \sum_{i=1}^{I_{\nu m}} \hat{c}_{i\nu}^{\mathrm{mock}}.$$

Here, I_{vm} is the number of mock subprograms to verify the subgraph p_v^m.

The time $\overline{T}^o$ and cost $\overline{C}^o$ of executing joint verification stages under the mth partition and the nth union of the graph $\boldsymbol{\Gamma}$ are given by

$$\overline{T}^o = \sum_{k=1}^{F_{mn}} \left(b_{kmn}^{n} + b_{kmn}^{\text{progr}} + b_{kmn}^{\text{loc}} \right),$$

$$\overline{C}^o = \sum_{k=1}^{F_{mn}} \left(S_{kmn}^{n} + S_{kmn}^{\text{progr}} + S_{kmn}^{\text{loc}} \right).$$

Suppose that the subgraphs $\tilde{p}_k^{mn} \in \tilde{P}^{mn}$ are verified using test sets with known time and cost characteristics. Then the time and cost of executing the kth joint verification stage under the mth partition and the nth union of the graph $\boldsymbol{\Gamma}$ are given by

$$b_{kmn}^{\text{n}} = \sum_{j=1}^{J_{kmn}} \hat{b}_{jkmn}^{\text{gen}}; \quad b_{kmn}^{\text{progr}} = \sum_{j=1}^{J_{kmn}} \hat{b}_{jkmn}^{\text{progr}}; \quad b_{kmn}^{\text{loc}} = \sum_{j=1}^{J_{kmn}} b_{jkmn}^{\text{loc}} \rho,$$

$$S_{kmn}^{\text{n}} = \sum_{j=1}^{J_{kmn}} \hat{S}_{jkmn}^{\text{gen}}; \quad S_{kmn}^{\text{progr}} = \sum_{j=1}^{J_{kmn}} \hat{S}_{jkmn}^{\text{progr}}; \quad S_{kmn}^{\text{loc}} = \sum_{j=1}^{J_{kmn}} S_{jkmn}^{\text{loc}} \rho.$$

With all these expressions for the time and cost characteristics of the software verification process, we formally state an optimization problem to find an optimal strategy for implementing a comprehensive verification scenario in terms of the minimum total time:

$$\sum_m \left(\sum_m \overline{T}_m^p \sum_{n=1}^{N_m} y_{mn} + \sum_{n=1}^{N_m} \overline{T}_m^o y_{mn} \right) \to \min.$$

This problem is solved subject to the following constraints:

—the maximum allowable cost of implementing the verification process,

$$\sum_m \left(\overline{C}_m^p \sum_m y_{mn} + \sum_{n=1}^{N_m} \overline{C}_m^o y_{mn} \right) \leqslant C;$$

—the set of M constraints on the variables y_{mn},

$$\sum_{n=1}^{N_m} y_{mn} = 1, \quad m = \overline{1, \text{M}}.$$

The problem of finding an optimal comprehensive verification strategy with the minimum cost criterion is formulated by analogy:

$$\sum_m \left(\overline{C}_m^p \sum_m y_{mn} + \sum_{n=1}^{N_m} \overline{C}_m^o y_{mn} \right) \to \min$$

subject to the time constraint imposed on the verification process,

$$\sum_m \left(\sum_m \overline{T}_m^p \sum_{n=1}^{N_m} y_{mn} + \sum_{n=1}^{N_m} \overline{T}_m^o y_{mn} \right) \leqslant T,$$

and the set of M constraints imposed on the variables y_{mn}.

These problems belong to the class of linear mathematical programming problems widely used in practice.

3. VERIFICATION METHODS FOR OSCS SOFTWARE WITH FUNCTIONAL CORRECTNESS INDICATORS

At the early functional integration stage of OSCS components, functional correctness indicators are used to assess the proper implementation of the functions of OSCS software. Each functionality of the software complex is implemented on some set of data processing routes. Along these routes, the input parameters of a function are transformed into one output result of this function or into a set of its output results. In order to check the correct operation of a function fully, it is necessary to check the entire set of data processing routes used by this function for a given set of its input parameters. Correct operation is validated if the output results of functions completely coincide with the reference results provided in the specifications of the program complex. Checking correct operation on the entire set of input data and on all data processing routes is a task of very high complexity. Therefore, one should select a bounded subset of data processing routes for their checking. This subset must be sufficient to check the implementation of the main functions of the software complex.

Nowadays, there are two approaches to check the correct operation of software: functional and structural. The functional approach involves the "black box" representation of software. The structural approach is based on checking the correct implementation of data processing routes; when preparing tests, it considers the structural peculiarities of separate modules of the software complex as well as the peculiarities of inter-module interaction within the complex. Both functional and structural approaches have significant disadvantages from the standpoint of efficient software verification implementation [2].

Due to this fact, we propose a method with the positive properties of both approaches. The method implies selecting a set of tests with the functional correctness indicators of the program complex that are necessary to check its correct operation. The quality of software operation is assessed based on the results of carrying out a set of selected tests. Consider this method in detail.

Let **F** be the set of all functions of the software complex implementing all primary and auxiliary functions. It is required to select a subset $\overline{F} < F$ of functions to be checked so that their correct operation will yield the desired values of the functional correctness indicators of the software complex.

For the software complex, an input data domain $\overline{E}$ is defined. For each function $F_j \in \overline{F}$, the corresponding subset $E_j \in \overline{E}$ of this domain is defined as well. Each such function transforms data of the input domain $E_j \in \overline{E}$ into the corresponding data of the output domain $y_j \in \overline{Y}$. Here, the set y_j contains all possible values of the output data for the function F_j $\left(j = \overline{1, J}\right)$.

The output results $y_{kj} \in Y$ of the software complex are obtained when implementing the sets of routes M_{jk} $\left(j = \overline{1, J}, k = \overline{1, K}\right)$ to process the data. Hence, to check the set of functions $\overline{F}$ of the software complex, it is necessary to check the correct operation of the set of data processing routes. Implementing these routes gives the necessary output results y_{kj} for each function F_j from the set $\overline{F}$ using the input data subsets $E_j \in \overline{E}$.

A function F_j of the software complex will be considered checked if, for all output results $y_{kj} \in Y_j$ of this function, the correctness of passing the set $M_{jk}\left(j = \overline{1, J}, k = \overline{1, K}\right)$ is successfully checked for all data processing routes yielding the output results for the function F_j. The sets $M_{jk} \in \overline{M_j}$, $k = \overline{1, K}$, of such routes will be considered the sets $\overline{M_j}$ of backbone routes for the function F_j. The correctness of obtaining the result of the jth function will be assessed using the indicator

$$N_{kj} = \frac{n_{kj}^{\text{chec}}}{n_{kj}^{\text{tot}}}.$$

In this formula, n_{kj}^{chec} is the number of checked backbone routes and n_{kj}^{gen} is the total number of backbone routes forming the results $y_{kj} \in Y_j$. The total number of backbone routes equals the cardinality of the set M_{kj}.

We will use the backbone route as the main element to be checked when assessing the functional correctness indicator of software and the graph model $\boldsymbol{\Gamma}\,(V, C)$ of the enlarged flowchart of the software complex when executing the verification scenario and determining the backbone paths for the functions of the set $\overline{F}$.

In the graph model, V is the vertex set of the graph $\boldsymbol{\Gamma}$, which corresponds to the set of blocks in the enlarged flowchart of the software complex, and C is the arc set of the graph. The arcs C represent the transfer of control between the flowchart blocks. These blocks are separate procedures and their aggregates or the program modules of the software complex. An arc between blocks i and j means the transfer of control from the former to the latter. In the model under consideration, vertex $\nu_i \in V$ of the graph $\boldsymbol{\Gamma}(V, C)$ is associated with the sets of its arguments $A_i = \{a_{\text{in}}\}$ and the sets of its results $R_i = \{r_{ij}\}$.

An information processing route m in the graph $\boldsymbol{\Gamma}(V, C)$ is a sequence $(v_0, c_0, v_1, c_1, \ldots, c_{I-1}, v_I)$ containing vertices and arcs. In this sequence, v_i $(0 \leqslant i \leqslant I)$ is a vertex of the graph $\boldsymbol{\Gamma}(V, C)$ and c_i $(1 \leqslant i \leqslant I-1)$ is an arc connecting vertices v_i and v_{i+1}. In turn, a sequence $(v_0, \ldots, v_I)$ of vertices corresponds to the transformations implemented on a data processing route m. Such a sequence is called a transformer of route m, and a sequence $(c_0, \ldots, c_{I-1})$ of arcs corresponds to the conditions to be satisfied on route m and is called the condition of route m.

For the result $y_{jk} \in Y_j$ of a function $F_j \in \overline{F}$, the backbone route m_{jk} is a route whose transformer $(v_0, \ldots, v_i)$ includes at least one of the possible sequences of external and internal information links. These external and internal links must start at vertex v_0 and end at vertex v_i to obtain the result y_{jk}.

4. SOFTWARE VERIFICATION FOR THE ONBOARD CONTROL SYSTEM OF THE RUSSIAN SEGMENT OF THE ISS

In this section, as one example, the concept described above is used to verify software configuration elements (SCEs) of the Russian Segment of the International Space Station (ISS) [3].

The following operations are carried out stage-by-stage to verify the SCEs:

(1) the autonomous testing of the software complex;

(2) the comprehensive verification of the SCEs on a ground verification bench;

(3) software verification jointly with C&C MDM (Command and Control Multiplexor DeMultiplexor, the onboard central computer of the US Segment and the entire ISS);

(4) the formal qualification tests of the SCEs.

The listed verification stages of the SCEs allow detecting, localizing, and eliminating the errors arising in the software verification process as well as confirming software operability and assessing software compliance with the technical specifications.

The **autonomous testing** of software is conducted based on an autonomous PC workstation and on the SDDF complex (software project development tools). Testing is conducted using a methodology that includes the following elements: the description of the testing procedure, initial testing conditions, and test cases. After the software testing process is finished, it is handed over to the configuration control group, which integrates the tested software into the SCEs of the onboard central computer.

The **comprehensive verification of the SCEs** is performed according to a special scenario to solve the following tasks:

(1) quality checking for the operating system;

(2) onboard control system software assembly and comprehensive verification in accordance with the flight plan and the operating modes of the Russian Segment and the service module simultaneously with flight safety control (i.e., checking the correct implementation of all subgraphs and the entire graph $\boldsymbol{\Gamma}$);

(3) spot checks of the backbone routes corresponding to the most probable abnormal situations, the localization of abnormal situations, and their elimination;

(4) checking the compliance of onboard control system software with the documents (ICD SSP 50 097);

(5) resource allocation control (memory, CPU time, and I/O channels).

Joint tests with C&C MDM were conducted on SITE-C, EGSE, and SVF, dedicated benches with special test implementation scenarios. During the tests, the onboard software of both onboard control systems (the US Segment and the Russian Segment) as well as the model software of both onboard systems (the US Segment and the Russian Segment) were used.

Formal qualification tests or acceptance tests and docking tests is a process that verifies the compliance of the SCEs of the onboard central computer with the requirement specification and ICD.

A certain subset is selected from the set of tests conducted using the NKO ground verification complex. This subset serves to verify the correctness of implementing a given set of backbone routes. Upon completion of the formal qualification testing, the Customer signs the report that the SCEs of the onboard central computer are ready for docking tests.

Docking tests were conducted in accordance with a dedicated methodology. The hardware and software means of the onboard central computer undergo docking tests with real hardware or its analogs using the NKO-2 ground verification complex. Docking tests of the hardware and software means of the onboard central computers (the Russian Segment with the US Segment) were conducted using the NKO-1 ground verification complex. They were carried out in accordance with the NASA–RSA Phase 2-3 Bilateral Integration and Verification Plan (SSP50101). The hardware and software means of the onboard central computer as part of the Zvezda service module (product index 17KSM) were tested on complex bench No. 24008 and on the control and test station in a required volume.

5. CONCLUSIONS

This paper has presented the existing experience as well as organizational, methodological, and technical solutions concerning software verification for onboard spacecraft control systems. The main features of a comprehensive software verification technology for onboard spacecraft control systems have been described. This technology ensures effective software development and debugging based on software prototypes, the iterative refinement of requirements, and early functional integration. The proposed technology has been implemented within the computer-aided software development and verification system for onboard spacecraft control systems. As a result, the total number of errors in the process of software development and verification has been significantly reduced for the Russian Segment of the ISS.

REFERENCES

1. Mikrin, E.A., Kul'ba, V.V., and Pavlov, B.V., Developing Models and Design Methods for Information Management Systems in Space Vehicles, *Autom. Remote Control*, 2013, vol. 74, no. 3, pp. 348–357.

2. Mikrin, E.A., Kul'ba, V.V., Kosyachenko, S.A., Somov, D.S., and Gladkov, Yu.M., *Kompleksnaya otrabotka programmnogo obespecheniya bortovogo kompleksa upravleniya kosmicheskimi apparatami i imitatsionnye modeli funktsionirovaniya bortovykh sistem i vneshnei sredy* (Comprehensive Software Verification for an Onboard Spacecraft Control System and Simulation Models of Onboard Systems and

Environment), Preprint of Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, 2011.

3. Kul'ba, V.V., Mikrin, E.A., Pavlov, B.V., and Platonov, V.N., *Teoreticheskie osnovy proektirovaniya informatsionno-upravlyayushchikh sistem kosmicheskikh apparatov* (Theoretical Foundations of Designing Spacecraft Information and Control Systems), Moscow: Nauka, 2006.
4. Kurenkov, V.I. and Kucherov, A.S., *Metody issledovaniya effektivnosti raketno-kosmicheskikh sistem. Problemno-orientirovannye sistemy avtomatizirovannogo proektirovaniya* (Methods for Studying the Efficiency of Rocket and Space Systems. Problem-Oriented Computer-Aided Design Systems), Samara: Samara State Aerospace University, 2012.
5. Zelentsov, V., Kovalev, A., Okhtilev, M., Sokolov, B., and Yusupov, R., Creation and Application Methodology of the Intelligent Information Technology of Complexity Objects Space and Ground Based Monitoring, *SPIIRAS Proceedings*, 2013, vol. 5, no. 28, pp. 7–81.
6. Mikrin, E.A., *Bortovye kompleksy upravleniya kosmicheskimi apparatami i proektirovanie ikh programmnogo obespecheniya* (Onboard Spacecraft Control Systems and Their Software Development), Moscow: Bauman Moscow State Technical University, 2003.
7. Mikrin, E.A., Sukhanov, N.A., Platonov, V.N., et al., Design Concepts of Onboard Control Complexes for Automatic Spacecrafts, *Control Sciences*, 2004, no. 3, pp. 62–66.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 10, pp. 1186–1195.

TOPICAL ISSUE

Parametrization of Optimal Anisotropic Controllers

A. Yu. Kustov

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

Received June 19, 2023
Revised July 18, 2023
Accepted August 2, 2023

Abstract—This paper provides a parametrization of optimal anisotropic controllers for linear discrete time invariant systems. The controllers to be designed are limited by causal dynamic output-feedback control laws. The obtained solution depends on several adjustable parameters that determine the specific type of controller, and is of the form of a system of the Riccati equations relating to a $\mathcal{H}_2$-optimal controller for a system formed by a series connection of the original system and the worst-case generating filter corresponding to the maximum value of the mean anisotropy of the external disturbance.

Keywords: linear discrete systems, anisotropy-based theory, optimal control, parametrization

DOI: 10.25728/arcRAS.2023.11.68.001

1. INTRODUCTION

The anisotropy-based control and estimation theory has been developed in the mid-90s as a response to attempts to provide the generalization of the results of the well-known $\mathcal{H}_2$- and $\mathcal{H}_\infty$-controller design theories [4, 12, 13].

It clearly shows the features of the control problems, the information theory, and various classical methods for suppressing (or mitigating) the impact of external disturbances [5]. However, unlike some approaches where it was proposed to use artificially defined in a certain sense mixed-type functionals, the anisotropy-based theory was focused on the method of describing the external disturbance driven the system. It was shown that the use of theoretical functionals makes it possible not only to describe a wide class of statistically uncertain random noises, but also generalize in a natural way the concepts of $\mathcal{H}_2$- and $\mathcal{H}_\infty$-norms making them the limiting cases of the anisotropic norm.

In this paper, the problem of parametrization of optimal anisotropic controllers for linear discrete time invariant systems is solved. The solution to the problem is based on the result associated with the parametrization of $\mathcal{H}_2$-optimal controllers as well as with the equations for the worst-case generating filter used in the anisotropy-based theory to form a signal with a given threshold level of mean anisotropy.

The paper is organized as follows. In Section 2, some preliminary mathematics from anisotropy-based theory are given. It also contains a parametrization of the $\mathcal{H}_2$-optimal controllers. In Section 3, the problem of parametrization of optimal anisotropic controllers is solved. The results are demonstrated with a numerical example. The last section contains the conclusions.

2. PRELIMINARIES

2.1. Notations

$\mathcal{H}_2^{m\times n}$ is the Hardy space of analytic rational transfer functions $P(z) = \sum\limits_{k=0}^{+\infty} P_k z^k \in \mathbb{C}^{m\times n}$ in the open unit disk $\{z \in \mathbb{C} : \ |z| < 1\}$ having the finite $\mathcal{H}_2$-norm

$$\|P\|_2 = \left(\frac{1}{2\pi} \int\limits_{-\pi}^{\pi} \operatorname{tr}\left(\widehat{P}(\omega)\widehat{P}^{\mathrm{T}}(-\omega)\right) d\omega \right)^{1/2},$$

where $\widehat{P}(\omega) = \lim\limits_{r\to 1-0} P(re^{i\omega})$; $\mathcal{RH}_2^{m\times n}$ is set of strictly proper stable rational $m \times n$ transfer functions; $\|P\|_\infty = \sup_{\omega\in[-2\pi;\pi)} \sigma_{\max}(\widehat{P}(\omega))$ is $\mathcal{H}_\infty$-norm of transfer matrix function $P(z)$ where $\sigma_{\max}(X) = \max_k \sigma_k(X)$ denotes maximum singular value of a matrix X, and $\sigma_k(X) = \lambda_k(X^{\mathrm{T}}X)$.

2.2. Basic Concepts of Anisotropy-Based Theory

Usually, the object of study in anisotropy-based theory is a stable linear discrete time invariant system

$$P_{zw} \sim \begin{cases} x_{k+1} = Ax_k + Bw_k, \\ z_k \quad\; = Cx_k + Dw_k, \end{cases} \tag{1}$$

with known matrices $A \in \mathbb{R}^{n_x\times n_x}$, $B \in \mathbb{R}^{n_x\times n_w}$, $C \in \mathbb{R}^{n_z\times n_x}$, $D \in \mathbb{R}^{n_z\times n_w}$, and, in general, zero initial conditions ($x_0 = 0$). This system describes the relation between the dynamical processes $\{x_k\}_{k\geqslant 0}$ and $\{z_k\}_{k\geqslant 0}$ driven by random input disturbance $\{w_k\}_{k\geqslant 0}$. The system (1) corresponds to its transfer function $P_{zw}(z) = D + C(zI_{n_x} - A)^{-1}B$ given by the quadruple

$$P_{zw} \sim \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] : \quad w \overset{x}{\to} z. \tag{2}$$

If necessary, we will specify the spaces of the states, the inputs, and the outputs. Within the discussion of the controller design problem, the plant (1) should be considered as the closed-loop system.

The following definitions give a basic idea of the concepts of the anisotropy-based theory. See [4, 12, 13] for more details.

Definition 1. The anisotropy of the square-integrable random vector $w \in \mathbb{L}_2^{n_w}$ is a nonzero number defined by

$$\mathbf{A}(w) = \min_{\lambda>0} \mathbf{D}(f \| p_{n_w,\lambda}), \tag{3}$$

where $\mathbf{D}(f\|g)$ is the Kullback–Leibler information divergence of f with respect to g; $f(x)$ is the probability density function (p.d.f.) of vector w; $p_{n_w,\lambda}(x) = (2\pi\lambda)^{-n_w/2} \exp\left(-\frac{|x|^2}{2\lambda}\right)$ is the p.d.f. of zero-mean Gaussian vector having scalar covariance matrix λI_{n_w}.

Definition 2. The mean anisotropy of stationary ergodic random sequence $W = \{w_k\}_{k\geqslant 0}$ is defined by the following formula:

$$\overline{\mathbf{A}}(W) = \lim_{N\to+\infty} \frac{\mathbf{A}(W_{0:N-1})}{N}. \tag{4}$$

Here, $W_{s:t} = (w_s^{\mathrm{T}}, \ldots, w_t^{\mathrm{T}})^{\mathrm{T}}$ denotes the fragment of the sequence $W = \{w_k\}_{k\geqslant 0}$ for $k = s, s+1, \ldots, t-1, t$.

It is assumed that the system (1) is driven by a disturbance with mean anisotropy constrained by nonnegative number $a \geqslant 0$, i.e. $\overline{\mathbf{A}}(W) \leqslant a$. This limitation determines the ability of the nature to generate the worst-case (in the sense of the value of the root-mean-square (RMS) gain) external disturbance, which the $\mathcal{H}_\infty$-theory works with, but at the same time allows it to have both spatial and temporal correlations, which is not covered by the classic $\mathcal{H}_2$-theory.

Definition 3. Anisotropic norm of the system (1) driven by the input disturbance whose mean anisotropy satisfy $\overline{\mathbf{A}}(W) \leqslant a$ is defined as

$$|\!|\!| P_{zw} |\!|\!|_a = \sup\left\{\frac{\|P_{zw}G\|_2}{\|G\|_2} :\ G \in \mathcal{H}_2^{n_w \times n_w} \wedge W = GV \wedge \overline{\mathbf{A}}(W) \leqslant a\right\} \tag{5}$$

where $V = \{v_k\}_{k\geqslant 0}$ denotes the standard Gaussian white noise passed through the linear system with $(n_w \times n_w)$-dimensional transfer function $G(z)$ having bounded $\mathcal{H}_2$-norm.

The anisotropic norm quantitatively reflects the ability of the system to amplify in the RMS sense the input signal with the information-theoretic constraint $\overline{\mathbf{A}}(W) \leqslant a$ imposed on it. In the case $\overline{\mathbf{A}}(W) = 0$, we have that $W = V$, and $|\!|\!| P_{zw} |\!|\!|_0 = \|P_{zw}\|_2/\sqrt{n_w}$. In the case when the restriction on mean anisotropy is removed, i.e. $\overline{\mathbf{A}}(W) < +\infty$, it can be shown that $\lim\limits_{a\to+\infty} |\!|\!| P_{zw} |\!|\!|_a = \|P_{zw}\|_\infty$. Thus, the anisotropy-based theory not only describes a wide class of external disturbances in information-theoretic terms, but also generalizes the approaches to controller design developed within the framework of $\mathcal{H}_2$- and $\mathcal{H}_\infty$-theories.

2.3. Parametrization of $\mathcal{H}_2$-Optimal Controllers

A lot of works are devoted to the study of all possible aspects of the behavior of linear systems with $\mathcal{H}_2$-optimal estimating controllers. In particular, a number of them describe methods for parameterizing the entire set of $\mathcal{H}_2$-optimal controllers. The procedure for solving this problem, as well as the accompanying difficulties, are described in detail in [2, 6, 9, 11] and many others. The main idea on which the solution is based is that $\mathcal{H}_2$-optimal controllers are directly related to the controllers that ensure the invariance of the output of some auxiliary system with respect to disturbances. Hence, by parameterizing the set of these controllers, parametrization of the $\mathcal{H}_2$-controllers can be obtained. The formulation of the problem of parametrization of $\mathcal{H}_2$-controllers, and the solution of this problem can be given in the following form.

Consider the system

$$F \sim \left[\begin{array}{c|cc} A & B_u & B_w \\ \hline C_y & 0 & D_{yw} \\ C_z & D_{zu} & 0 \end{array}\right] :\quad \begin{pmatrix} u \\ w \end{pmatrix} \xrightarrow{x} \begin{pmatrix} y \\ z \end{pmatrix}, \tag{6}$$

with matrices $A \in \mathbb{R}^{n_x \times n_x}$, $B_u \in \mathbb{R}^{n_x \times n_u}$, $B_w \in \mathbb{R}^{n_x \times n_w}$, $C_y \in \mathbb{R}^{n_y \times n_x}$, $D_{yw} \in \mathbb{R}^{n_y \times n_w}$, $C_z \in \mathbb{R}^{n_z \times n_x}$, $D_{zu} \in \mathbb{R}^{n_z \times n_w}$, where u, w, y, z are state, disturbance, measurement, and controlled vectors. Consider also non-strictly causal dynamical stabilizing output-feedback controller

$$K \sim \left[\begin{array}{c|c} A_c & B_c \\ \hline C_c & D_c \end{array}\right] :\quad y \xrightarrow{h} u, \tag{7}$$

where $h_k \in \mathbb{R}^{n_h}$, and A_c, B_c, C_c, D_c are unknown matrices. The closed-loop system formed by the system (6) and the controller (7) can be represented by

$$F_{cl}(K) \sim \left[\begin{array}{cc|c} A + B_u D_c C_y & B_u C_c & B_w + B_u D_c D_{yw} \\ B_c C_y & A_c & B_c D_{yw} \\ \hline C_z + D_{zu} D_c C_y & D_{zu} C_c & D_{zu} D_c D_{yw} \end{array}\right] :\quad w \xrightarrow{\binom{x}{h}} z. \tag{8}$$

To solve the problem of $\mathcal{H}_2$-optimal control means to find the matrices of the controller (7), such that $\mathcal{H}_2$-norm of the closed-loop system (8) is minimal, i.e. $\|F_{cl}(K)\|_2 \to \min\limits_K$.

It has already been noted that the problem of designing $\mathcal{H}_2$-optimal controller is associated with the problem of input-output invariance of some auxiliary system [2, 3, 6, 9, 10]. To solve the last problem, several additional definitions are introduced, closely related to the concepts of controllable and observable invariants [1, 7]. Namely, for the system $F:\ w \overset{x}{\to} z$ defined by the quadruple (A, B, C, D) where $x \in \mathbb{R}^{n_x}$, $w \in \mathbb{R}^{n_w}$, $z \in \mathbb{R}^{n_z}$, we define two sets: $\mathcal{W}(F)$ and $\mathcal{S}(F)$ (see, for example, [8]).

Definition 4. The stabilizable weakly unobservable subspace $\mathcal{W}(F)$ is the largest subspace $\mathcal{W} \subseteq \mathbb{R}^{n_x}$ for which there exists a matrix Π of suitable dimensions, such that $\mathcal{W} \subseteq \ker(C + D\Pi)$, $(A + B\Pi)\mathcal{W} \subseteq \mathcal{W}$ and $\rho(A + B\Pi) < 1$.

Definition 5. The detectable strongly controllable subspace $\mathcal{S}(F)$ is the smallest subspace $\mathcal{S} \subseteq \mathbb{R}^{n_x}$ for which there exists a matrix Λ of suitable dimensions, such that $\operatorname{im}(B + \Lambda D) \subseteq \mathcal{S}$, $(A + \Lambda C)\mathcal{S} \subseteq \mathcal{S}$ and $\rho(A + \Lambda C) < 1$.

Let us also introduce two auxiliary matrices P and Q associated with the system (6) as the largest in the sense of the matrix order ($X \succ Y \Leftrightarrow X - Y \succ 0$) matrices-solutions to the inequalities

$$M_1(P) = \begin{bmatrix} A^{\mathrm{T}}PA - P + C_z^{\mathrm{T}}C_z & C_z^{\mathrm{T}}D_{zu} + A^{\mathrm{T}}PB_u \\ D_{zu}^{\mathrm{T}}C_z + B_u^{\mathrm{T}}PA & D_{zu}^{\mathrm{T}}D_{zu} + B_u^{\mathrm{T}}PB_u \end{bmatrix} \succeq 0, \tag{9a}$$

$$M_2(Q) = \begin{bmatrix} AQA^{\mathrm{T}} - Q + B_wB_w^{\mathrm{T}} & B_wD_{yw}^{\mathrm{T}} + AQC_y^{\mathrm{T}} \\ D_{yw}B_w^{\mathrm{T}} + C_yQA^{\mathrm{T}} & D_{yw}D_{yw}^{\mathrm{T}} + C_yQC_y^{\mathrm{T}} \end{bmatrix} \succeq 0. \tag{9b}$$

For a pair (P, Q), we additionally define the matrices C_P, D_P, B_Q and D_Q in accordance with the formulas

$$\begin{bmatrix} C_P^{\mathrm{T}} \\ D_P^{\mathrm{T}} \end{bmatrix} [C_P \quad D_P] = M_1(P), \qquad \begin{bmatrix} B_Q \\ D_Q \end{bmatrix} [B_Q^{\mathrm{T}} \quad D_Q^{\mathrm{T}}] = M_2(Q), \tag{10}$$

provided that both $[C_P \quad D_P]$ and $[B_Q^{\mathrm{T}} \quad D_Q^{\mathrm{T}}]$ are of full rank.

The solution to the problem of parametrization of $\mathcal{H}_2$-optimal controllers is given in the form of the following theorem.

Theorem 1 [2, 10]. *For a system* (6)*, there exists an $\mathcal{H}_2$-optimal controller of the form* (7) *if and only if the following conditions are satisfied:*

(i) (A, B_u) *is stabilizable,*

(ii) (C_y, A) *is detectable,*

(iii) $\operatorname{im}(B_Q - B_uD_P^{+}R) \subseteq \mathcal{W}(F_{P_u})$,

(iv) $\mathcal{S}(F_{Q_y}) \subseteq \ker(C_P - RD_Q^{+}C_y)$,

(v) $\mathcal{S}(F_{Q_y}) \subseteq \mathcal{W}(F_{P_u})$,

(vi) $(A - B_uD_P^{+}RD_Q^{+}C_y)\mathcal{S}(F_{Q_y}) \subseteq \mathcal{W}(F_{P_u})$,

where

$$R = (D_P^{\mathrm{T}})^{+}(D_{zu}^{\mathrm{T}}C_zQC_y^{\mathrm{T}} + B_u^{\mathrm{T}}PAQC_y^{\mathrm{T}} + B_u^{\mathrm{T}}PB_wD_{yw}^{\mathrm{T}})(D_Q^{\mathrm{T}})^{+}, \tag{11}$$

and the systems F_{P_u} and F_{Q_y} are defined by

$$F_{P_u} \sim \left[\begin{array}{c|c} A & B_u \\ \hline C_P & D_P \end{array}\right], \quad F_{Q_y} \sim \left[\begin{array}{c|c} A & B_Q \\ \hline C_y & D_Q \end{array}\right]. \tag{12}$$

If the conditions of the theorem are met, the set of all dynamic $\mathcal{H}_2$-optimal controllers of the form (7) *is given by*

$$K \sim \left[\begin{array}{cc|c} A + B_u\Pi + \Lambda C_y - B_u\widetilde{D}C_y & B_u\widetilde{C} & B_u\widetilde{D} - \Lambda \\ -\widetilde{B}C_y & \widetilde{A} & \widetilde{B} \\ \hline \Pi - \widetilde{D}C_y & \widetilde{C} & \widetilde{D} \end{array}\right], \tag{13}$$

where the choice of matrices $\widetilde{A}$, $\widetilde{B}$, $\widetilde{C}$, $\widetilde{D}$ is limited by the fact that the transfer function

$$\widetilde{F}(z) = \widetilde{D} + \widetilde{C}(zI_{n_x} - \widetilde{A})^{-1}\widetilde{B} \tag{14}$$

belongs to the following algebraic sum of spaces: $\widetilde{F}(z) \in N_F + M_F$, where

$$N_F = \left\{N \in \mathbb{R}^{n_u \times n_y} : D_P N D_Q = -R\right\}, \tag{15a}$$

$$M_F = \left\{M(z) \in \mathcal{RH}_2^{n_u \times n_y} : F_1(z)M(z)F_2(z) = 0\right\}, \tag{15b}$$

and

$$F_1(z) = D_P + (C_P + D_P\Pi)(zI_{n_x} - A - B_u\Pi)^{-1}B_u, \tag{16a}$$

$$F_2(z) = D_Q + C_y(zI_{n_x} - A - \Lambda C_y)^{-1}(B_Q + \Lambda D_Q). \tag{16b}$$

A solid analysis of the statement of the theorem can be found in [11]. For the case when left/right-invertible system has no invariant zeros, the similar theorem can be formulated with a certain changes [2]. In this case, the statement will additionally include the condition of *uniqueness* of the $\mathcal{H}_2$-optimal controller if one exists.

3. PARAMETRIZATION OF ANISOTROPIC CONTROLLERS

3.1. Problem Statement and Solution

The problem of optimal anisotropic controller design for linear discrete time invariant systems was solved in [13]. The conditions under which the controller was designed ensure the existence and uniqueness of the solution, and the controller itself was specified in a strictly causal form. This section provides a solution to a similar problem, which consists of parameterizing all optimal non-strictly causal anisotropic controllers.

Problem 1. For a system (6) driven by external disturbance with the constraint $\overline{\mathbf{A}}(W) \leqslant a$, describe the parametric set of optimal anisotropic controllers of the form (7), i.e. parameterize non-strictly causal stabilizing dynamical controllers minimizing the anisotropic norm of the corresponding closed-loop system.

It is known that when solving the anisotropic analysis and synthesis problems in the optimal setting, it is necessary to consider an additional mathematical construction called the worst-case generating filter. The goal of this filter is to generate the most undesirable (in terms of RMS gain value) external disturbance for a closed-loop system. In accordance with the results obtained in [12, 13], for the systems of the form (2), the worst-case filter is of the form

$$G \sim \left[\begin{array}{c|c} A + BL & B\Sigma^{1/2} \\ \hline L & \Sigma^{1/2} \end{array}\right] : \quad v \overset{x}{\to} w, \tag{17}$$

where $L \in \mathbb{R}^{n_w \times n_x}$ and $\Sigma \succ 0$ are the matrices chosen to maximize the RMS gain $\|P_{zw}G\|_2/\|G\|_2$ under the constraint $\overline{\mathbf{A}}(W) \leqslant a$. Here and below $V = \{v_k\}_{k\geqslant 0}$ — standard Gaussian white noise.

The main idea of solving the problem 1 is to consider a system formed by a successive connection of the worst-case shaping filter and the original system F, and then to parameterize the $\mathcal{H}_2$-optimal controllers for the obtained system. First of all we note that taking into account the shaping filter (17), the system (6) is equivalent from the point of view of the corresponding dynamic processes to the system

$$\overline{F} \sim \left[\begin{array}{c|cc} \overline{A} & \overline{B}_u & \overline{B}_w \\ \hline \overline{C}_y & 0 & \overline{D}_{yw} \\ \overline{C}_z & \overline{D}_{zu} & 0 \end{array}\right] = \left[\begin{array}{c|cc} A' + B'_w L & B'_u & B'_w \Sigma^{1/2} \\ \hline C'_y + D'_{yw} L & 0 & D'_{yw}\Sigma^{1/2} \\ C'_z & D'_{zu} & 0 \end{array}\right] : \quad \begin{pmatrix} \overline{u} \\ v \end{pmatrix} \overset{\binom{x}{h}}{\rightarrow} \begin{pmatrix} \overline{y} \\ z \end{pmatrix}, \tag{18}$$

where the new variables are defined as $\overline{u}_k = \left(u_k^{\mathrm{T}} \;\; h_{k+1}^{\mathrm{T}}\right)^{\mathrm{T}}$ and $\overline{y}_k = \left(y_k^{\mathrm{T}} \;\; h_k^{\mathrm{T}}\right)^{\mathrm{T}}$; the matrices used in the expression (18) have the following structure:

$$A' = \begin{bmatrix} A & 0 \\ 0 & 0_{n_h \times n_h} \end{bmatrix}, \; B'_u = \begin{bmatrix} B_u & 0 \\ 0 & I_{n_h} \end{bmatrix}, \; B'_w = \begin{bmatrix} B_w \\ 0_{n_h \times n_w} \end{bmatrix}, \tag{19a}$$

$$C'_y = \begin{bmatrix} C_y & 0 \\ 0 & I_{n_h} \end{bmatrix}, \; D'_{yw} = \begin{bmatrix} D_{yw} \\ 0_{n_h \times n_w} \end{bmatrix}, \tag{19b}$$

$$C'_z = [C_z \quad 0_{n_z \times n_h}], \; D'_{zu} = [D_{zu} \quad 0_{n_z \times n_h}]; \tag{19c}$$

matrices L and Σ correspond to the shaping filter G (which is the worst-case one for the system (18)) generating a colored signal with the mean anisotropy less or equal to a given threshold $a \geqslant 0$ from standard Gaussian white noise $V = \{v_k\}_{k \geqslant 0}$.

Theorem 2. *For a system* (6) *with an external disturbance satisfying the constraint* $\overline{\mathbf{A}}(W) \leqslant a$, *there is an optimal anisotropic controller of the form* (7) *iff the conditions (i)–(vi) of the Theorem 1 hold true. If these conditions are met, the set of all optimal anisotropic controllers of the form* (7) *for the system* (6) *is determined by the formula* $\overline{u}_k = \left(u_k^{\mathrm{T}} \;\; h_{k+1}^{\mathrm{T}}\right)^{\mathrm{T}}$ *where control* $\overline{u}_k$ *is given by the following set of optimal anisotropic controllers for the system* (18):

$$\overline{K} \sim \left[\begin{array}{cc|c} \overline{A} + \overline{B}_u\overline{\Pi} + \overline{\Lambda}\overline{C}_y - \overline{B}_u\widetilde{D}\overline{C}_y & \overline{B}_u\widetilde{C} & \overline{B}_u\widetilde{D} - \overline{\Lambda} \\ -\widetilde{B}\overline{C}_y & \widetilde{A} & \widetilde{B} \\ \hline \overline{\Pi} - \widetilde{D}\overline{C}_y & \widetilde{C} & \widetilde{D} \end{array}\right]. \tag{20}$$

The matrices $\widetilde{A}$, $\widetilde{B}$, $\widetilde{C}$, $\widetilde{D}$ *correspond to the transfer function*

$$\widetilde{F}(z) = \widetilde{D} + \widetilde{C}(zI_{n_x+n_h} - \widetilde{A})^{-1}\widetilde{B} \tag{21}$$

belonging to the sum of subspaces $\widetilde{F}(z) \in N_{\overline{F}} + M_{\overline{F}}$, *where*

$$N_{\overline{F}} = \{\overline{N} \in \mathbb{R}^{(n_u+n_h)\times(n_y+n_h)} : \overline{D}_P\overline{N}\,\overline{D}_Q = -\overline{R}\}, \tag{22a}$$

$$M_F = \{\overline{M}(z) \in \mathcal{RH}_2^{(n_u+n_h)\times(n_y+n_h)} : \overline{F}_1(z)\overline{M}(z)\overline{F}_2(z) = 0\}, \tag{22b}$$

where

$$\overline{F}_1(z) = \overline{D}_P + (\overline{C}_P + \overline{D}_P\overline{\Pi})(zI_{n_x+n_h} - \overline{A} - \overline{B}_u\overline{\Pi})^{-1}\overline{B}_u, \tag{23a}$$

$$\overline{F}_2(z) = \overline{D}_Q + \overline{C}_y(zI_{n_x+n_h} - \overline{A} - \overline{\Lambda}\overline{C}_y)^{-1}(\overline{B}_Q + \overline{\Lambda}\overline{D}_Q) \tag{23b}$$

and

$$\overline{R} = (\overline{D}_P^{\mathrm{T}})^{+}(\overline{D}_{zu}^{\mathrm{T}}\overline{C}_z\overline{Q}\,\overline{C}_y^{\mathrm{T}} + \overline{B}_u^{\mathrm{T}}\overline{P}\,\overline{A}\,\overline{Q}\,\overline{C}_y^{\mathrm{T}} + \overline{B}_u^{\mathrm{T}}\overline{P}\,\overline{B}_w\overline{D}_{yw}^{\mathrm{T}})(\overline{D}_Q^{\mathrm{T}})^{+}. \tag{24}$$

The matrices $\overline{C}_P$, $\overline{D}_P$, $\overline{B}_Q$ and $\overline{D}_Q$ are introduced according to (10) *for the matrices $M_1(\overline{P})$ and $M_2(\overline{Q})$ associated with the system* (18)*, and the matrices $\overline{\Pi}$ and $\overline{\Lambda}$ relate to the sets $\mathcal{W}(\overline{F}_{P_u})$ and $\mathcal{S}(\overline{F}_{Q_y})$ introduced according to 4 i 5.*

The proof of the theorem is given in the Appendix.

Corollary 1. *If in the Theorem 2 it is also true that the transfer function*

$$F^{ol}_{yw}(z) = D_{yw} + C_y(zI_{n_x} - A)^{-1}B_w \tag{25}$$

is right invertible, and the transfer function

$$F^{ol}_{zu}(z) = D_{zu} + C_z(zI_{n_x} - A)^{-1}B_u \tag{26}$$

is left reversible then the optimal anisotropic controller exists and is unique.

3.2. Numerical Example

As an example, consider a system of the form (6) with matrices

$$A = \begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix}, \quad B_u = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad B_w = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \tag{27a}$$

$$C_y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad D_{yw} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad C_z = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad D_{zu} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{27b}$$

We assume that the external disturbance has the mean anisotropy bounded by a certain number $a \geqslant 0$. Let us set the goal of the example as to solve the problem of parametrization of optimal anisotropic controllers of order not higher than the order of the system itself. Moreover, for the sake of simplicity we will require that the number of additional variables is minimal, i.e., according to (21), $\widetilde{F}(z) = \widetilde{D}$.

Following the required calculations, we can verify that the system (18) has the form

$$\overline{F} \sim \left[\begin{array}{cc|cc|c} A + B_wL_1 & B_wL_2 & B_u & 0_{2\times2} & B_w\sqrt{\sigma} \\ 0_{2\times2} & 0_{2\times2} & 0_{2\times1} & I_2 & 0_{2\times1} \\ \hline I_2 & 0_{2\times2} & 0_{2\times1} & 0_{2\times2} & 0_{2\times1} \\ 0_{2\times2} & I_2 & 0_{2\times1} & 0_{2\times2} & 0_{2\times1} \\ \hline C_z & 0_{2\times2} & D_{zu} & 0_{2\times2} & 0_{2\times1} \end{array}\right], \tag{28}$$

and the corresponding matrices $\overline{P}$ and $\overline{Q}$ defined by the formulas (9) are as follows:

$$\overline{P} = \begin{bmatrix} \overline{P}_{11} & 0_{2\times2} \\ 0_{2\times2} & 0_{2\times2} \end{bmatrix}, \quad \overline{Q} = \begin{bmatrix} \overline{Q}_{11} & 0_{2\times2} \\ 0_{2\times2} & 0_{2\times2} \end{bmatrix}, \quad \overline{Q}_{11} = B_wB_w^{\mathrm{T}}\sqrt{\sigma}, \tag{29}$$

where $\overline{P}_{11}$ is the solution to the Riccati equation

$$\overline{P}_{11} = (A + B_wL_1)^{\mathrm{T}}\overline{P}_{11}(A + B_wL_1) + C_z^{\mathrm{T}}C_z \tag{30a}$$

$$- (A + B_wL_1)^{\mathrm{T}}\overline{P}_{11}B_u(D_{zu}^{\mathrm{T}}D_{zu} + B_u^{\mathrm{T}}\overline{P}_{11}B_u)^{-1}B_u^{\mathrm{T}}\overline{P}_{11}(A + B_wL_1), \tag{30b}$$

where, in order to simplify the further calculations, *we will immediately assume* that $L_2 = 0$ (one can show this statement is true). After this, the matrices $\overline{D}_P$, $\overline{C}_P$, $\overline{D}_Q$ and $\overline{B}_Q$:

$$\overline{D}_P = \begin{bmatrix} (\overline{D}_P)_{11} & 0_{1\times 2} \\ 0_{2\times 1} & 0_{2\times 2} \end{bmatrix} = \begin{bmatrix} (D_{zu}^{\mathrm{T}} D_{zu} + B_u^{\mathrm{T}} \overline{P}_{11} B_u)^{1/2} & 0_{1\times 2} \\ 0_{2\times 1} & 0_{2\times 2} \end{bmatrix}, \tag{31a}$$

$$\overline{C}_P = \begin{bmatrix} (\overline{C}_P)_{11} & 0_{1\times 2} \\ 0_{2\times 2} & 0_{2\times 2} \end{bmatrix} = \begin{bmatrix} (\overline{D}_P)_{11}^{-1} B_u^{\mathrm{T}} \overline{P}_{11} (A + B_w L_1) & 0_{1\times 2} \\ 0_{2\times 2} & 0_{2\times 2} \end{bmatrix}, \tag{31b}$$

$$\overline{D}_Q = \begin{bmatrix} (\overline{D}_Q)_{11} & 0_{2\times 2} \\ 0_{2\times 2} & 0_{2\times 2} \end{bmatrix} = \begin{bmatrix} (B_w B_w^{\mathrm{T}})^{1/2} \sqrt{\sigma} & 0_{2\times 2} \\ 0_{2\times 2} & 0_{2\times 2} \end{bmatrix}, \tag{31c}$$

$$\overline{B}_Q = \begin{bmatrix} (\overline{B}_Q)_{11} & 0_{2\times 2} \\ 0_{2\times 2} & 0_{2\times 2} \end{bmatrix} = \begin{bmatrix} (A + B_w L_1)(\overline{D}_Q)_{11} & 0_{2\times 2} \\ 0_{2\times 2} & 0_{2\times 2} \end{bmatrix}. \tag{31d}$$

We also calculate the matrix R using (11):

$$\overline{R} = \begin{bmatrix} (\overline{C}_P)_{11} (\overline{D}_Q)_{11} & 0_{1\times 2} \\ 0_{2\times 2} & 0_{2\times 2} \end{bmatrix}. \tag{32}$$

Now one can check the conditions of the Theorem 2. Obviously, the pair $(\overline{A}, \overline{B}_u)$ is stabilizable, and the pair $(\overline{C}_y, \overline{A})$ is detectable. Now one needs to introduce the sets $\mathcal{W}(\overline{F}_P)$ and $\mathcal{S}(\overline{F}_Q)$. According to the definitions (4) and (5), the matrices $\overline{\Pi}$ and $\overline{\Lambda}$ satisfy the following conditions:

$$\overline{\Pi} = \begin{bmatrix} \overline{\Pi}_{11} & \overline{\Pi}_{12} \\ \overline{\Pi}_{21} & \overline{\Pi}_{22} \end{bmatrix} = \begin{bmatrix} -(\overline{D}_P)_{11}^{-1} (\overline{C}_P)_{11} & 0_{1\times 2} \\ \overline{\Pi}_{21} & \overline{\Pi}_{22} \end{bmatrix}, \quad \rho(\overline{\Pi}_{22}) < 1, \tag{33a}$$

$$\overline{\Lambda} = \begin{bmatrix} \overline{\Lambda}_{11} & \overline{\Lambda}_{12} \\ \overline{\Lambda}_{21} & \overline{\Lambda}_{22} \end{bmatrix} = \begin{bmatrix} -(\overline{B}_Q)_{11} (\overline{D}_Q)_{11}^{-1} & \overline{\Lambda}_{12} \\ 0_{2\times 2} & \overline{\Lambda}_{22} \end{bmatrix}, \quad \rho(\overline{\Lambda}_{22}) < 1. \tag{33b}$$

To simplify the calculations, we choose $\overline{\Pi}_{21} = \overline{\Pi}_{22} = 0_{2\times 2}$ and $\overline{\Lambda}_{12} = \overline{\Lambda}_{22} = 0_{2\times 2}$. One should keep in mind that the particular choice of these matrices leads to a narrowing of the set of the optimal anisotropic controllers. The choice made leads to the fact that $\mathcal{W}(\overline{F}_P) = \mathbb{R}^4$ and $\mathcal{S}(\overline{F}_Q) = \{0\}^4$, after which the conditions *(iii)–(vi)* of the Theorem 2 can be trivially verified.

In this example, a controller with representation (20) under the condition $\widetilde{F}(z) = \widetilde{D}$ is completely determined by the matrix $\widetilde{D}$, satisfying the requirement $\overline{D}_P \widetilde{D} \overline{D}_Q = -\overline{R}$. Substituting the previously found matrices into the last equality, we obtain that $\widetilde{D}$ have the form

$$\widetilde{D} = \begin{bmatrix} \widetilde{D}_{11} & \widetilde{D}_{12} \\ \widetilde{D}_{21} & \widetilde{D}_{22} \end{bmatrix}, \tag{34}$$

where $\widetilde{D}_{11} = \overline{\Pi}_{11}$, which completes the procedure of describing all optimal anisotropic controllers (20) associated with the relation $\overline{u}_k = \begin{pmatrix} u_k^{\mathrm{T}} & h_{k+1}^{\mathrm{T}} \end{pmatrix}^{\mathrm{T}}$.

Let us make a few important comments.

In order to obtain the final solution to the problem, the resulting system of equations must be supplemented with a system of equations determining the worst-case generating filter, thus finding the variables $L_1 \in \mathbb{R}^{1\times 2}$ and $\sigma > 0$ (see, for example, [12]).

Also, since in the framework of the considered example, the state was observed (measured) precisely, it seems natural to choose a controller in the form of the static state-feedback $u_k = K x_k$. In this case, the optimal choice of matrix K is $K = \widetilde{D}_{11}$.

Let us also present a solution to this problem for $a = 0$ (this case was chosen for simplicity, since there is no need to solve the auxiliary problem associated with the generating filter). It can be shown that all controllers with the representation

$$K \sim \left[\begin{array}{c|c} A_c & B_c \\ \hline C_c & D_c \end{array}\right] \approx \left[\begin{array}{cc|cc} 0 & 0 & -1 & 1 \\ -\varkappa - 1.4773 & -\varkappa - 1.4773 & \varkappa + 1 & \varkappa + 2.1823 \\ \hline -\varkappa - 1.4773 & -\varkappa - 1.4773 & \varkappa & \varkappa + 2.1823 \end{array}\right] \tag{35}$$

are optimal, and have the same closed-loop system that does not depend on the specific choice of $\varkappa$ (the choice of which is constrained by the inclusion $\varkappa \in (-2.4773;\ -0.4773)$ providing that the spectral radius of the matrix A_c is less than 1):

$$x_{k+1} \approx \begin{bmatrix} -1 & 1 \\ -0.4773 & 0.7051 \end{bmatrix} x_k + \begin{bmatrix} 1 \\ -1 \end{bmatrix} w_k, \tag{36a}$$

$$z_k \approx \begin{bmatrix} -1.4773 & 0.7051 \\ 0 & 1 \end{bmatrix} x_k. \tag{36b}$$

Note that $\varkappa \approx -1.4773$ in (35) corresponds to a static state-feedback controller $u_k = D_c x_k$.

4. CONCLUSION

The paper provides a parametrization of a set of optimal anisotropic controllers for linear discrete time invariant systems. The results obtained can find application in solving practical problems of navigation and control, in particular, in cases when additional constraints are imposed on the control actions. The results can also be useful to solve the problem of parameterizing a set of suboptimal anisotropic controllers and estimators.

APPENDIX

Proof of Theorem 2. First let us show that the conditions *(i)–(vi)* from the formulation of the Theorem 2 are equivalent to the following:

(a) $(\overline{A}, \overline{B}_u)$ is stabilizable,

(b) $(\overline{C}_y, \overline{A})$ is detectable,

(c) $\operatorname{im}(\overline{B}_Q - \overline{B}_u \overline{D}_P^{+} \overline{R}) \subseteq \mathcal{W}(\overline{F}_{P_u})$,

(d) $S(\overline{F}_{Q_y}) \subseteq \ker(\overline{C}_P - \overline{R}\overline{D}_Q^{+} \overline{C}_y)$,

(e) $\mathcal{S}(\overline{F}_{Q_y}) \subseteq \mathcal{W}(\overline{F}_{P_u})$,

(f) $(\overline{A} - \overline{B}_u \overline{D}_P^{+} \overline{R} \overline{D}_Q^{+} \overline{C}_y) \mathcal{S}(\overline{F}_{Q_y}) \subseteq \mathcal{W}(\overline{F}_{P_u})$

where matrices $\overline{C}_P$, $\overline{D}_P$, $\overline{B}_Q$, $\overline{D}_Q$ and $\overline{R}$, as well as systems $\overline{F}_{P_u}$ and $\overline{F}_{Q_y}$ are set in accordance to the material presented in Section 2.3 in relation to the system (18). Note that the conditions *(a)–(f)* are a direct analogue of the conditions *(i)–(vi)* of the Theorem 1 for system (18).

The equivalence of *(i)* ⇔ *(a)* and *(ii)* ⇔ *(b)* is obvious due to the notation (19). For further proof, let us determine the relation of the sets $\mathcal{W}(F_{P_u})$ and $\mathcal{S}(F_{Q_y})$ from Theorem 1 to the sets $\mathcal{W}(\overline{F}_{P_u})$ and $\mathcal{S}(\overline{F}_{Q_y})$, respectively. Given the system $\overline{F}$, using the Definitions 4 and 5, it can be verified that there exist matrices $\overline{\Pi}$ and $\overline{\Lambda}$ such that

$$\mathcal{W}(\overline{F}_{P_u}) = \mathcal{W}(F_{P_u}) \times \mathbb{R}^{n_h}, \qquad \mathcal{S}(\overline{F}_{Q_y}) = \mathcal{S}(F_{Q_y}) \times \{0\}^{n_h}, \tag{A.1a}$$

$$(\overline{A} + \overline{B}_u \overline{\Pi}) \mathcal{W}(\overline{F}_{P_u}) \subseteq \mathcal{W}(\overline{F}_{P_u}), \qquad (\overline{A} + \overline{\Lambda}\overline{C}_y) \mathcal{S}(\overline{F}_{Q_y}) \subseteq \mathcal{S}(\overline{F}_{Q_y}), \tag{A.1b}$$

$$\rho(\overline{A} + \overline{B}_u \overline{\Pi}) < 1, \qquad \rho(\overline{A} + \overline{\Lambda}\overline{C}_y) < 1. \tag{A.1c}$$

After this, we can conclude that the equivalence of the conditions *(iii)⇔(c)* and *(iv)⇔(d)* holds due to the fact that

$$\ker(\overline{C}_P - \overline{R}\overline{D}_Q^+\overline{C}_y) = \ker(C_P - RD_Q^+C_y) \times \mathbb{R}^{n_h}, \tag{A.2a}$$

$$\operatorname{im}(\overline{B}_Q - \overline{B}_u\overline{D}_P^+\overline{R}) = \operatorname{im}(B_Q - B_uD_P^+R) \times \{0\}^{n_h}. \tag{A.2b}$$

Finally, by (A.1), the equivalence of the conditions *(v)⇔(e)* and *(vi)⇔(e)* is proved.

The structure of the controller (20) is determined by the content of the Theorem 1.

The Theorem 2 is proven.

ACKNOWLEDGMENTS

The author of the paper thanks Igor Gennadievich Vladimirov and Alexander Viktorovich Yurchenkov who took part in the discussion of the part of the material.

REFERENCES

1. Basile, G. and Marro, G., Controlled and conditioned invariant subspaces in linear system theory, *J. Optim. Theory Appl.*, 1969, vol. 3, pp. 306–315. https://doi.org/10.1007/BF0093137010.1007/BF00931370

2. Chen, B.M., Saberi, A., and Shamash, Y., Necessary and sufficient conditions under which a discrete time H$_2$-optimal control problem has a unique solution, *Proc. 32nd IEEE Conf. Decision and Control*, 1993, vol. 1, pp. 805–810. https://doi.org/10.1109/CDC.1993.325038

3. Chen, B.M., Saberi, A., Shamash, Y., and Sannuti, P., Construction and parameterisation of all static and dynamic H$_2$-optimal state feedback solutions for discrete time systems, *Proc. 32nd IEEE Conf. Decision and Control*, 1993, vol. 1, pp. 126–131. https://doi.org/10.1109/CDC.1993.325177

4. Diamond, P., Kloeden, P., and Vladimirov, I., Mean anisotropy of homogeneous Gaussian random fields and anisotropic norms of linear translation-invariant operators on multidimensional integer lattices, *J. Appl. Math. Stochast. Anal.*, 2003, vol. 16:3, pp. 209–231. https://doi.org/10.1155/S1048953303000169

5. Doyle, J.C., Glover, K., Khargonekar, P.P., and Francis, B.A., State-space solutions to standard H_2 and H_∞ control problems, *IEEE Transactions on Automatic Control*, 1989, vol. 34, pp. 831–847. https://doi.org/10.1109/9.29425

6. Saberi, A., Sannuti, P., and Stoorvogel, A.A., H$_2$ optimal controllers with measurement feedback for continuous-time systems: flexibility in closed-loop pole placement, *Automatica*, 1997, vol. 33, no. 3, pp. 289–304. https://doi.org/10.1016/S0005-1098(96)00195-1

7. Schumacher, J.M., *Dynamic feedback in finite- and infinite-dimensional linear systems*, Mathematisch Centrum, 1981. ISBN: 9061962293.

8. Stoorvogel, A.A., The H_∞ control problem: a state space approach, *Phd Thesis, Mathematics and Computer Science*, Technische Universiteit Eindhoven, 1981. 229 P. https://doi.org/10.6100/IR338287

9. Stoorvogel, A.A., The singular H$_2$ control problem, *Automatica*, 1992, vol. 28, no. 3, pp. 627–631. https://doi.org/10.1016/0005-1098(92)90189-M

10. Trentelman, H.L. and Stoorvogel, A.A., Sampled-data and discrete-time H$_2$ optimal control, *Proc. 32nd IEEE Conf. Decision and Control.*, 1993, vol. 1, pp. 331–336. https://doi.org/10.1109/CDC.1993.325136

11. Trentelman, H.L. and Stoorvogel, A.A., Sampled-data and discrete-time H$_2$ optimal control, *SIAM J. Control and Optimization*, 1995, vol. 33, no. 3, pp. 834–862. https://doi.org/10.1137/S0363012992241995

12. Vladimirov, I.G., Kurdjukov, A.P., and Semyonov, A.V., On computing the anisotropic norm of linear discrete-time-invariant systems, *IFAC Proceedings Volumes*, 1996, vol. 29, no. 1, pp. 3057–3062. https://doi.org/10.1016/S1474-6670(17)58144-6

13. Vladimirov, I.G., Kurdyukov, A.P., and Semyonov, A.V., State-space solution to anisotropy-based stochastic H∞-optimization problem, *IFAC Proceedings Volumes*, 1996, vol. 29, no. 1, pp. 3816–3821.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 10, pp. 1196–1210.

Increasing the Angular Resolution and Range of Measuring Systems Using Ultra-Wideband Signals

B. A. Lagovsky[*,a] **and E. Ya. Rubinovich**[**,b]

[*] *Russian Technological University (MIREA), Moscow, Russia*
[**] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
e-mail: [a] *robertlag@yandex.ru*, [b] *rubinvch@gmail.com*

Received September 29, 2022
Revised July 23, 2023
Accepted August 2, 2023

Abstract—The problem of obtaining three-dimensional radio images of objects with increased resolution based on the use of ultra-wide-band pulse signals and new methods of their digital processing is considered. The inverse problem of reconstructing the image of a signal source with a resolution exceeding the Rayleigh criterion has been solved numerically. Mathematically, the problem is reduced to solving the Fredholm integral equation of the first kind by numerical methods based on the representation of the solution in the form of decomposition into systems of orthogonal functions. The method of selecting the systems of functions used, which increases the stability of solutions, is substantiated. Variational problems of optimizing the shape and duration of ultra-wide-band pulses have been solved, ensuring the maximum possible signal-to-noise ratio during location studies of objects with fully or partially known signal reflection characteristics. The proposed procedures make it possible to increase the range of measuring systems, and also make it possible to increase the stability of solutions to inverse problems. It is shown that the use of the developed methods for achieving super-resolution to process ultra-wideband signals dramatically improves the quality of 3D images of objects in the radio range.

Keywords: Rayleigh criterion, angular superresolution, stability of solutions to inverse problems

DOI: 10.25728/arcRAS.2023.90.53.001

1. INTRODUCTION

Increasing the effective angular resolution of radio and sonar, radio navigation, and remote sensing systems and bringing it to super-resolution makes it possible to detail images of the objects under study, solve problems of their recognition and identification, and separately observe individual targets as part of group targets. Solving these tasks makes it possible to improve the quality of existing and promising control systems for land, surface, underwater and aerospace objects. Currently, many methods of digital signal processing and analysis are known to increase the effective resolution. These are, in particular, methods of reverse convolution of signals, phase weighing coefficients, angular weighing, etc. Currently popular methods are: MUSIC (MUltiple SIgnal Classification) [1], ESPRIT (Estimation of Signal Parameters via Rotational Invariant Techniques) [2], deconvolution method [3, 4], maximum entropy method [5], maximum likelihood method [6], methods using neural networks [7], as well as nonlinear methods [8].

The listed methods are not effective in all cases. Most of them, including MUSIC and ESPRIT, turn out to be ineffective when active measuring systems use complex signals, in particular, UWB (Ultra Wide Band) signals with a duration of nanoseconds. The use of such ultra-wideband signals potentially allows for very high resolution over a range of about 1 m. As a result, the combination of the use of UWB signals and the achievement of angular super-resolution due to digital signal processing will allow obtaining high-quality three-dimensional radio images of objects. Such systems are all-weather and can operate at any time of the day.

2. SETTING THE TASK OF ACHIEVING SUPER-RESOLUTION

The signal $U(\varphi,\theta)$ received by the goniometer system when scanning a two-dimensional sector of the survey can be expressed as a linear integral transformation [9]

$$U(\varphi,\theta)=\int\limits_{\Omega} f(\varphi-\phi,\theta-\vartheta)I(\phi,\vartheta)\,d\phi\,d\vartheta, \tag{1}$$

where $\Omega=\Omega(\varphi,\theta)$—the angular area of the signal source location; $I(\varphi,\theta)$—the angular amplitude distribution of the signal reflected (or emitted) by the object of observation, equal to zero outside Ω; $f(\varphi,\theta)$—directional pattern (DP) of the measuring system. For convenience, the Cartesian coordinate system is used here and further, where the angles are calculated from the normal to the antenna plane.

It is known that the angular resolution achieved during direct measurements in accordance with (1), i.e. the ability to distinguish two closely spaced objects, is measured by the minimum angles $\delta\varphi$ and $\delta\theta$, at which the two point signal sources still differ separately. These angles are determined based on the Rayleigh criterion

$$\delta\varphi\cong\lambda/D_x,\quad \delta\theta\cong\lambda/D_y, \tag{2}$$

where D_x and D_y are the linear dimensions of the antenna at the corresponding angles φ and θ directions, λ is a wavelength. The angles $\delta\varphi$ and $\delta\theta$ they turn out to be equal to the width of the DP, determined by reducing the radiated power by half and are denoted as $\varphi_{0.5}$ and $\theta_{0.5}$.

The task is to obtain an image of the signal source $I(\varphi,\theta)$ with an angular resolution exceeding the Rayleigh criterion to the greatest extent possible, based on the intelligent analysis of the received signal $U(\varphi,\theta)$ and the known DP $f(\varphi,\theta)$ of the system. Mathematically, the problem is reduced to an approximate solution of the Fredholm integral equation (IE) of the first kind of convolution type (1) with respect to an unknown function $I(\varphi,\theta)$ with the maximum achievable accuracy.

In general, attempts to increase the resolution exceeding (2) by solving the IE lead to unstable solutions, since the task belongs to the class of inverses and does not satisfy the second and third requirements of Hadamard correctness (2nd is an unambiguity of solutions and 3rd is their stability).

The methods of digital signal processing developed by the authors, called algebraic [9–15], seem promising, allowing to obtain a stable approximate solution of the IE (1).

3. ALGEBRAIC METHODS OF SOLUTION

Algebraic methods consist in parameterizing the problem by presenting approximate solutions in the form of expansions over selected sequences of functions. The choice of systems of functions is based on a priori information about the solution.

Let's consider practically important tasks when scanning is performed using one of the angular coordinates.

The desired distribution of $I(\varphi)$ can always be represented as a decomposition over some complete system of orthogonal functions in the domain of Ω $g_m(\varphi)$ with unknown coefficients b_m

$$I(\varphi)=\sum_{m=1}^{\infty}b_m g_m(\varphi)\cong\sum_{m=1}^{M}b_m g_m(\varphi). \tag{3}$$

Then the received signal is $U(\varphi)$ is expressed as a superposition of functions $G_m(\varphi)$, which are images of $g_m(\varphi)$ when converting

$$G_m(\varphi) = \int\limits_{\Omega} f(\varphi - \phi) g_m(\phi)\, d\phi, \tag{4}$$

$$U(\varphi) = \sum_{m=1}^{\infty} b_m G_m(\varphi) \cong \sum_{m=1}^{M} b_m g_m(\varphi), \tag{5}$$

where M is the selected number of expansion terms.

Thus, the inverse problem turns out to be parameterized, and its solution is reduced to finding the coefficients b_m [10–12], which are usually found when minimizing the standard deviation of the function $U(\varphi)$ from (5) from the signal under study (1) in the corner sector $\Phi > \Omega$, where Φ is the sector in which the useful signal exceeds noise and can be measured with sufficiently high accuracy. In practice, the boundaries of the Φ sector are often determined by reducing the amplitude of the useful signal by half in relation to its maximum value.

Function system $G_m(\varphi)$ of (4), generally speaking, is not orthogonal and the minimization mentioned above reduces to solving the following system of linear algebraic equations (SLAE)

$$\mathbf{V} = \mathbf{SB},$$

where $\mathbf{B}$ is a vector column of coefficients b_m, and the components of the vector $\mathbf{V}$ and the matrix $\mathbf{S}$ are equal respectively:

$$V_j = \int\limits_{\Phi} U(\varphi) G_j(\varphi)\, d\varphi, \qquad S_{jm} = \int\limits_{\Phi} G_j(\varphi) G_m(\varphi)\, d\varphi,$$

here

$$\int\limits_{\Phi} U(\varphi) G_j(\varphi)\, d\varphi = \sum_{m=1}^{M} b_m \int\limits_{\Phi} G_j(\varphi) G_m(\varphi)\, d\varphi, \quad j = 1, 2, \ldots, M. \tag{6}$$

The principal feature of SLAE (6) is their poor conditionality, which is a consequence of an attempt to solve an incorrect inverse problem. An increase in the stability of solutions can be achieved if the functions $G_m(\varphi)$ turns out to be orthogonal in the domain of Φ. In this case, in the matrix $\mathbf{S}$, only the elements on the main diagonal are different from zero and the coefficients b_m are easily found

$$\int\limits_{\Phi} U(\varphi) G_m(\varphi)\, d\varphi = b_m \sum_{j=1}^{M} G_j^2(\varphi)\, d\varphi, \quad m = 1, 2, \ldots, M.$$

Thus, the problem arises of choosing such an orthonormal in the domain of Φ systems of functions $\widetilde{g}_m(\varphi)$, images $\widetilde{G}_m(\varphi)$ of which are orthogonal in Φ.

4. SIMULTANEOUS ORTHOGONALIZATION OF THE SYSTEMS OF FUNCTIONS USED

The orthogonal functions $g_m(\varphi)$ and $G_m(\varphi)$ can be used as their own functions and IE (1). However, the numerical search for each of them boils down to solving unstable problems and, consequently, to the appearance of significant errors in solving the entire problem. Even in the simplest case of searching for eigenfunctions, when the core of the IE is degenerate, i.e. DP of the $f(\varphi)$ measuring system is the DP of a one-dimensional antenna array (AA) [29]

$$f(\varphi) = \sum_{n=-K}^{n=K} j_n \exp(-ikdn \sin\varphi). \tag{7}$$

The SLAE obtained for searching for eigenfunctions turn out to be poorly conditioned. In (7) it is indicated: j_n is the magnitude of the current at the nth emitter, d is a distance between neighboring emitters, $2K+1$—the number of DP elements and the constant $k = 2\pi/\lambda$, where the wavelength $\lambda = 2\pi c/\omega$, c is the speed of light, ω is the frequency of radiation. It is significant that the conditionality numbers of the corresponding matrices increase exponentially with an increase in the number of eigenfunctions to be determined, i.e. with attempts to increase the effective angular resolution.

Note that the construction of an orthogonal system of functions $\widetilde{G}_m(\varphi)$ in the domain Φ can also be carried out on the basis of the Gram–Schmidt orthogonalization process. In this case, however, the resulting functions turn out to be images of functions that are not orthogonal in the domain of Φ. In this case, the source is also represented as a superposition of non-orthogonal functions, which significantly reduces the quality of the approximate solution.

The actual problem of simultaneous orthogonalization of systems of functions g_m and G_m is proposed to be solved on the basis of the following theorem, the proof of which is given in Appendix.

Theorem 1. *Let's define a system of N orthonormal functions $g_m(x)$ (hereafter $m = 1, 2, \ldots, N$) on the segment L_g and an arbitrary linear operator $\mathbf{A}$ generating a system of N functions $G = \mathbf{A}g$, on the segment L_G. Here G and g are N-dimensional vector columns with components G_m and g_m. Then there is a linear transformation, represented as a matrix $\mathbf{T}$, such that the systems of functions*

$$\widetilde{G}_m(\varphi) = \sum_{j=1}^{N} T_{jm} G_j(\varphi), \quad \widetilde{g}_m(\varphi) = \sum_{j=1}^{N} T_{mj} g_j(\varphi) \tag{8}$$

on the segments L_G and L_g, respectively, turn out to be orthogonal, while maintaining the condition $\widetilde{G} = \mathbf{A}\widetilde{g}$.

The results of the theorem allow us to simultaneously present the desired solution $I(\varphi)$ of the inverse problem under consideration and the signal under study $U(\varphi)$ in the form of decompositions over systems of orthogonal functions, which simplifies the analysis of the problem, increases the stability of numerical solutions and, ultimately, allows to increase the achieved degree of superresolution.

Using (A.1)–(A.5) (see Appendix), we obtain

$$\widetilde{G}_m(\varphi) = \int\limits_{\Omega} f(\varphi - \phi)\widetilde{g}_m(\phi)\, d\phi, \qquad \widetilde{g}_m(\varphi) = \sum_{j=1}^{N} T_{mj} g_j(\varphi). \tag{9}$$

Further, expressing the received signal in the form of decomposition

$$U(\varphi) \cong \sum_{m=1}^{N} C_m \widetilde{G}_m(\varphi),$$

we find, due to the orthogonality of the functions, the coefficients C_m

$$C_m = \frac{1}{P_m} \int\limits_{\Phi} U(\varphi)\widetilde{G}_m(\varphi)\, d\varphi, \quad \text{where} \quad P_m = \int\limits_{\Phi} \widetilde{G}_m^2(\varphi)\, d\varphi. \tag{10}$$

Taking into account the entered designations, the received signal (1) can be represented as follows

$$U(\varphi) = \int\limits_{\Omega} f(\varphi - \phi) I(\phi)\, d\phi \cong \sum_{m=1}^{N} C_m \widetilde{G}_m(\varphi) = \int\limits_{\Omega} f(\varphi - \phi) \left(\sum_{m=1}^{N} C_m \widetilde{g}_m(\phi) \right) d\phi. \qquad (11)$$

Equating the integral expressions in (11), we obtain a solution to the inverse problem under consideration in the form of decompositions both according to the introduced system of functions (A.5) and according to the original system of N functions (3)

$$I(\varphi) \cong \sum_{m=1}^{N} C_m \widetilde{g}_m(\varphi), \qquad I(\varphi) = \sum_{j=1}^{N} b_j g_j(\varphi), \qquad b_j = \sum_{m=1}^{N} C_m T_{mj}. \qquad (12)$$

Next, the algorithm uses an iterative process of increasing N to increase the degree of superresolution achieved until stable solutions can be obtained.

Since the inverse problem is considered, the solution of which, after parameterization, is reduced to a SLAE solution, all the negative properties of inverse problems are preserved and transferred eventually to SLAE solutions. In the problems under consideration, the second and third signs of the correctness of the Hadamard problem are violated, namely: unambiguity of solutions and their stability. The matrices $\mathbf{S}$ in (6) turn out to be poorly conditioned. When trying to increase the resolution, the dimension of the $\mathbf{S}$ matrices increases, while the conditioning numbers increase exponentially and reach huge values: 10^{10}–10^{13}, so even insignificant rounding errors lead to inadequate solutions. The presence of noise and measurement errors further worsens the situation. The direct solution of SLAE by known numerical methods of linear algebra does not lead to a satisfactory result.

At the same time, the values of the conditioning numbers of matrices of type $\mathbf{T}$ from (8), (9) are many times—and the orders are less than those of matrices $\mathbf{S}$. This circumstance is an indicator of the higher stability of solutions obtained on the basis of Theorem 1, in comparison with the direct solution of SLAE (5), (6). Thus, the proposed approach to solving the problem (1) provides the opportunity to use a larger number of functions in the solution representation (12) compared to (3)–(6), which increases the angular resolution. In an alternative formulation, the developed approach makes it possible to achieve the same level of exceeding the Rayleigh criterion as other methods, but at a significantly higher level of noise and interference.

5. EXAMPLES OF PROBLEM SOLUTIONS

Initially, single-stage functions were selected to represent the solution in the domain $\Omega = [-\theta_0, \theta_0]$, where $2\theta_0 = \theta_{0.5}$, and the solution was searched based on the algebraic method briefly described above (3)–(5). Then the search for solutions was carried out on the basis of Theorem 1 and the relations (8)–(12), and the solutions obtained were compared.

Figure 1a of the five original functions $g_m(\varphi)$, $m = 1, \ldots, 5$, three are shown—$g_1(\varphi), g_2(\varphi)$ and $g_4(\varphi)$. Figure 1b shows for illustration the transformed modifications of the original functions $g_1(\varphi)$ and $g_4(\varphi)$, i.e. $\widetilde{g}_1(\varphi)$ and $\widetilde{g}_4(\varphi)$. Figure 2a shows the images of $G_m(\varphi)$ source functions $g_m(\varphi)$ at $m = 1, 3, 5$, and in Fig. 2b—images of $\widetilde{G}_m(\varphi)$ functions $\widetilde{g}_m(\varphi)$ in the domain Φ.

Two point targets with the same amplitude of the emitted signal were selected as classical objects for the study of resolution. The distance between the objects was consistently reduced until it was possible to obtain sufficiently stable solutions adequate to the original objects. When objects approach each other, false sources begin to appear in the solution. Their intensity increases dramatically with further convergence. Figure 3b shows the extreme case when they can still be neglected.

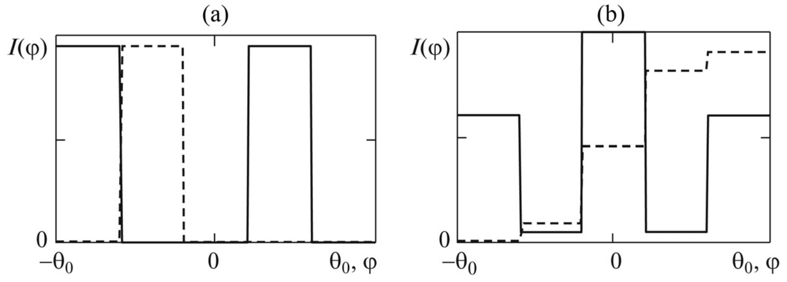


Fig. 1. Initial step functions (a), modified functions (b).

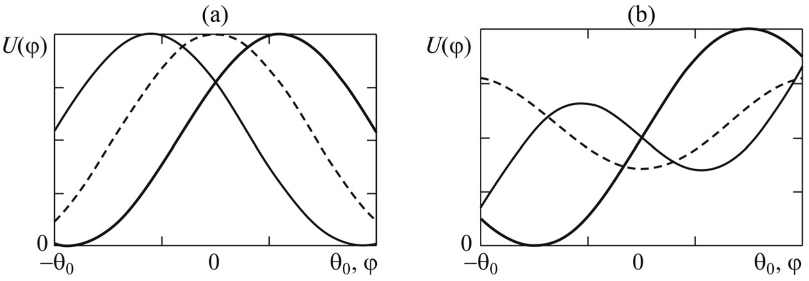


Fig. 2. Images of G_m original functions (a), images of $\widetilde{G}_m$ modified functions (b).

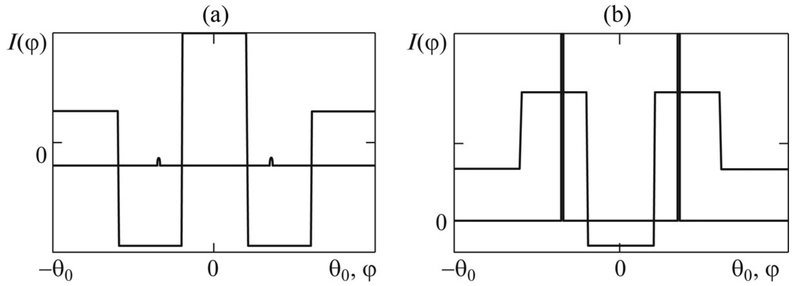


Fig. 3. Solution based on the original functions (a), solution based on modified functions (b).

Figures 3a and 3b present the solutions obtained in accordance with (3)–(6), i.e. without orthogonalization of functions and their images, as well as solutions after the procedure of simultaneous orthogonalization of $g_m(\varphi)$ and $G_m(\varphi)$. The angular position of point objects is shown as a bold vertical line, and the solution is shown as a polyline.

The results of numerical experiments have shown that by simultaneous orthogonalization it is possible to exceed the Rayleigh criterion four times (Fig. 3b). An attempt to obtain a stable solution to the same problem in accordance with (3)–(6) does not lead to a satisfactory result. The resulting inadequate solution, shown in Fig. 3a, is characterized by an oscillating character with a very large oscillation amplitude. Against the background of this solution, the true objects depicted on the same scale as in Fig. 3b are almost invisible. The type of solution is typical for cases when it is not possible to find an adequate solution. The conditioning numbers of the matrices used in solving and characterizing the stability of problems differ in the presented examples by two orders of magnitude.

It should be noted that when the number of M functions of the original system used in the solution representation (3) changes, the systems of functions $g_m(\varphi)$ and $G_m(\varphi)$ (9) themselves change. This feature has little effect on the running time of the program, since the basic calculations are performed using standard high-speed and well-developed algorithms.

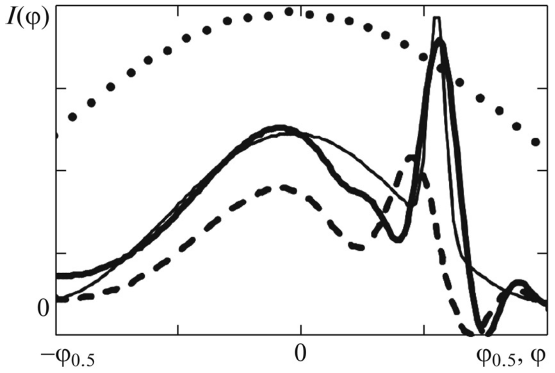


Fig. 4. A solution based on DOG wavelets.

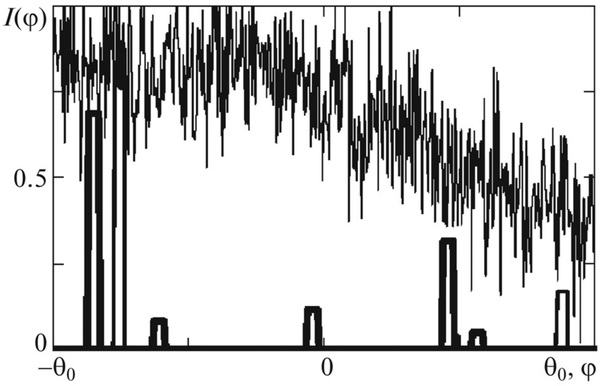


Fig. 5. A solution based on delta functions.

The choice of the initial system of functions $g_m(\varphi)$ is based on a priori information about the solution [16] and the shape of the signal received during scanning (1). Such information may include, in particular, the size and location of the signal source localization area, monotony, smoothness of the continuity area of the angular distribution of the amplitude of the emitted signal, the presence of areas with a discrete distribution, the dynamic range of intensity variation, restrictions on the gradient and other characteristics [11, 13, 16].

Figure 4 shows the solution of the inverse problem under consideration using this kind of a priori information. It was known that in the remote sensing problem, the reflecting surface is described by a smooth function with a smooth change in the amplitude of the reflected signal, with the possible presence of a small-sized area with high reflection. Based on this information, a system of functions based on DOG-wavelets was chosen to represent the solution.

Under direct observation without the proposed digital processing, the presented section has some averaged amplitude – the upper point curve. Signal processing by the algebraic method made it possible to identify the details of the amplitude distribution of $I(\varphi)$. In the form of a solid thin curve, Fig. 4 shows the true distribution of the reflected signal, the strokes represent the solution found by the algebraic method (3)–(6), the solid bold curve is the solution obtained using the considered double orthogonalization method.

Digital processing based on double orthogonalization has improved the quality of the solution, especially in the area of the area with a high gradient of the reflection coefficient.

Orthogonalization of function systems allows not only to increase the angular resolution, but, due to good stability, to obtain adequate solutions at high levels of random components.

Figure 5 shows the solution of the problem at a high noise level in the form of a bold polyline. The signal source consisted of two small-sized objects shown in the drawing with a thin polyline. The amplitudes of the signals reflected from the objects differed by five times. The objects were not resolved under direct observation. To illustrate, the figure shows the signal received during scanning in the Ω sector—the upper curve.

To represent solutions as a $g_m(\varphi)$ system, there was selected a system of delta functions located at the same distance from each other. In the process of finding a solution, these distances could be changed.

In the course of numerical experiments, the minimum value of the signal-to-noise ratio (SNR) was sought, at which it was still possible to obtain a satisfactory solution. The large difference in the amplitudes of the reflected signals significantly complicated the solution of the problem. False objects appeared, albeit with a small amplitude, the magnitude of which allowed them to be neglected when presenting the final solution. As a result, a completely satisfactory stable solution was obtained with a SNR equal to 1/3, or 10.5 dB. Many well-known methods, including [1–8], allow us to successfully solve such problems only with an SNR of at least 20 dB. Thus, the application of the method of simultaneous orthogonalization of systems of functions for solving inverse problems makes it possible to detail images of objects with an angular resolution exceeding the Rayleigh criterion at a significant level of random components of the signal. A further increase in the achieved degree of superresolution is possible with a decrease in the level of random components—noise in the studied signals.

6. IMPROVING THE SNR FOR UWB SYSTEMS

Currently existing UWB signal generation systems do not have sufficient energy to carry out measurements at significant distances [17–19]. In these conditions, an important task is to increase the range of systems by optimizing the digital processing of received UWB pulses. Optimization consists in the development of algorithms to increase the SNR in the received signals, which ultimately increases the range of the systems, and also improves the quality of images of objects with angular superresolution. Increasing the SNR increases the stability of solutions to the inverse problems discussed above, which are significantly more sensitive than direct ones to the presence and level of random components in the studied signals. Any linear algorithms for processing UWB signals that increase the SNR simultaneously provide an increase in the effective angular resolution.

Known methods of calculating characteristics and optimizing them are of little use for solving problems of optimizing the characteristics of UWB radars [20–28]. When emitting, receiving and reflecting ultra-wideband pulses from objects, it is necessary to take into account the dispersion dependences of the reflection characteristics of the studied objects, as well as antenna systems. As a result of the dispersion, the shape and spectrum of the received pulse differ significantly from the emitted one, which makes it almost impossible to use traditional methods of coherent signal processing.

Another feature of solving problems of analysis and optimization of UWB pulses is the difficulty of using well-developed spectral analysis methods in calculations, since for their successful application it is necessary to set the amplitude and phase spectra of pulses with high accuracy. The UWB signal, however, has an ultra-wide frequency band and, consequently, the spectral density of the pulse turns out to be small (often close to the magnitude of errors in calculations and measurements). In particular, when receiving a signal, its spectral density is often lower than the spectral density of noise. Under these conditions, the measurement accuracy of the amplitude-phase spectrum required for optimization cannot be achieved.

To overcome such difficulties, it is proposed to apply to calculations related to the description of the processes of radiation, reception, reflection and processing of UWB pulses a time domain analysis method based on the representation of antenna systems, reception, generation systems, etc., as linear systems described by pulse characteristics.

The proposed unconventional approach turns out to be more convenient and accurate, since when using it, it is necessary to know not the spectrum, but only the time dependence of the generated signal $U(t)$, which can be determined experimentally with sufficiently high accuracy.

7. OPTIMIZATION OF THE IMPULSE RESPONSE OF THE RECEIVING SYSTEM

Let's set the task of searching for the impulse response $h_r(t)$ of the receiving system, which ensures the maximum possible power gain—q^2. The shape of the generated UWB pulse $U(t)$, the DP of the transmitting and receiving antenna systems at each of the frequencies used are considered to be set—$f_e(\varphi, \omega)$ and $f_p(\varphi, \omega)$, as well as the complex frequency response of the reflection of the object—$R(\omega)$. The specified variance dependencies allow using the Fourier transform F[..] determine the pulse characteristics of radiation, reception and reflection of the signal:

$$h_e(\varphi, t) = \mathbf{F}[f_e(\varphi, \omega)], \quad h_p(\varphi, t) = \mathbf{F}[f_p(\varphi, \omega)], \quad h_R(t) = \mathbf{F}[R(\omega)]. \tag{13}$$

In addition to the above characteristics, for modern systems based on antenna arrays (AA), it is necessary to additionally take into account the mutual influence of emitters on each other. Mutual influence is usually described using mutual complex resistances, i.e. the intrinsic resistance of the emitter changes by the amount of a certain introduced resistance. This resistance, called mutual resistance, depends on the distance between the emitters, measured by the ratio of the physical distance to the wavelength—the electrical distance. Without taking into account the mutual influence for narrow-band AA, the error in calculating their characteristics is 3–6% and it can often be neglected. When using UWB pulses for low-frequency components, the electrical distances between the emitters decrease several times and the resistance value increases noticeably. In order to avoid significant errors—up to 40–50%—mutual influence must be taken into account when constructing the pulse characteristics of UWB radars.

For two separate AA emitters numbered m and n located at a distance of $d_{m,n}$ from each other with co-directional DP, their mutual complex resistance

$$z(kd_{m,n}) = r(kd_{m,n}) + ix(kd_{m,n}),$$

expressed as [29]

$$r(kd_{m,n}) = \frac{1}{B} \int\limits_0^{\pi} \int\limits_0^{2\pi} \phi_m(\varphi, \theta) \phi_n^*(\varphi, \theta) \cos\left(kd_{m,n} \sin\theta\right) \sin\theta \, d\varphi \, d\theta, \tag{14}$$

$$\begin{gathered} x(kd_{m,n}) = \frac{4}{kd_{m,n}} \int\limits_0^{\pi} \phi_m(\theta) \phi_n^*(\theta) \, d\theta \\ - \int\limits_0^{\pi} \int\limits_0^{2\pi} \phi_m(\varphi, \theta) \phi_n^*(\varphi, \theta) \sin\left(kd_{m,n} \sin\theta |\sin\varphi|\right) \sin\theta \, d\varphi \, d\theta, \end{gathered} \tag{15}$$

where B is the normalizing factor, $\phi_m(\varphi, \theta)$—DP of a separate emitter.

Usually, the weakly directional DP of individual AA emitters are the same and often, especially for flat and linear AA, do not depend on the azimuth angle. Then they can be described with high accuracy in the form of functions $\phi(\varphi, \theta) = \cos^{\nu}\theta$ or a superposition of similar functions, where the parameter ν describes the directivity of the emitter. In this case, the integrals in (14), (15) are taken explicitly [29], and the mutual effective resistance of two adjacent radiators (14) turns out to be equal

$$r(kd_{m,n}) = \Gamma(\nu + 3/2)\frac{J_{\nu+1/2}(kd_{m,n})}{(kd_{m,n}/2)^{\nu+1/2}}, \tag{16}$$

where $\Gamma(\nu)$ is a Gamma function, J_ν is a Bessel function of the order of ν. The mutual reactive part of the resistance, normalized to its own resistance, is reduced to the form

$$x(kd_{m,n}) = \frac{2\Gamma(\nu + 3/2)}{\sqrt{\pi}\Gamma(\nu + 1)kd_{m,n}} - \frac{\Gamma(\nu + 3/2)H_{\nu+1/2}(kd_{m,n})2^{\nu+1/2}}{(kd_{m,n})^{\nu+1/2}}, \tag{17}$$

where H_ν is a Struve function of the order ν [30].

For large AA, one can ignore the edge effects and assume that all emitters are in the same conditions. Then the resistances of all emitters turn out to be the same, and taking into account the mutual influence of the emitters leads to the need to use instead of $h_{e,p}(\varphi, t)$ from (13)

$$h_{e,p}(\varphi, t) = \mathbf{F}\left[\frac{f_{e,p}(\varphi, \omega)}{z(\varphi, \omega)}\right], \tag{18}$$

where $z(\varphi, \omega)$—the resistance of the emitter at the frequency ω, taking into account the influence of all other emitters of the AA.

Let's find the frequency dependence of the emitter resistance $z(\varphi, \omega)$. To this end, let's first consider the linear AA. For large AA with a number of $2N + 1$ emitters focused in the direction of φ to the AA axis (7), active resistance $r(\varphi, \omega)$ of each element represents the following amount, which can be extended indefinitely with almost no error

$$r(\varphi, \omega) = \sum_{n=-N}^{N} r(kdn)\cos(kdn\sin\varphi) \approx \sum_{n=-\infty}^{\infty} r(kdn)\cos(kdn\sin\varphi). \tag{19}$$

The sum of the series (19) can be represented as a closed expression. To do this, you first need to find the sum of the next row

$$W = \sum_{n=1}^{\infty} \frac{J_\nu(x)}{(nx/2)^\nu}\cos(nx\sin\varphi), \tag{20}$$

which is called the generalized Schlemilch series [30]. Note that the values of the sum (20) are given in the reference books only for a few special cases. In general, the sum of the series is found in [22]. It is shown that under the condition $kd < 2\pi/(1 + \sin\varphi)$, which is true for AA, the sum of the generalized Schlemilch series is equal to

$$W = -\frac{1}{2\Gamma(\nu + 1)} + \frac{\sqrt{\pi}}{\Gamma(\nu + 1/2)x}\cos^{2\nu-1}\varphi. \tag{21}$$

Finally,

$$r(\varphi, \omega) = \frac{2\sqrt{\pi}\Gamma(\nu + 3/2)}{\Gamma(\nu + 1)kd}\cos^{2\nu}\varphi. \tag{22}$$

The imaginary part of the resistance of each element of a large linear AA has a representation

$$X = \sum_{n=-\infty}^{\infty} x(kdn)\cos(kdn\sin\varphi) \cong \sum_{n=-N}^{N} x(kdn)\cos(kdn\sin\varphi) \tag{23}$$

Numerical estimates (23) for various ν and kd from (17) show that the value of X for large linear AA turns out to be close to zero. Thus, the resistance of each element in the composition of a large linear AA $z(kd)$ with good accuracy takes the value (22), which allows us to find the impulse response (18).

The resistance of the radiator in a large flat AA is obtained twice using the sum (22)

$$r(\varphi,\omega)=\frac{4\pi(\nu+1/2)}{(kd)^2}\cos^{2\nu-1}\varphi. \tag{24}$$

Now, using the found resistance values, we find the pulse characteristics of the radiation and AA reception in the form (18).

Typically, the SNR for pulse signals is defined as the ratio of the square of the maximum value of the useful signal to the RMS value of the noise $\overline{U_n^2}$

$$q^2=\frac{U_M^2}{\overline{U_n^2}}. \tag{25}$$

Then for narrowband signals, when the frequency band is much less than the fundamental frequency $\Delta\omega \ll \omega_0$, for the viewing angle $\varphi=0$ we get

$$q^2=\frac{f^4(0,\omega_0)|R(\omega_0)|^2}{N_0(\omega_0)}, \tag{26}$$

where N_0 is the spectral density of noise at the frequency ω_0. For UWB signals, considering noise to be a stationary random process, (25) takes the form:

$$q^2=\frac{\left(\int_{-\infty}^{\infty} h_r(t)U_r(t_0-t)\,dt\right)^2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_r(t)h_r(T)K(t-T)\,dt\,dT}, \tag{27}$$

where $U_r(t)$ is the received signal, t_0 is the time when the useful signal reaches the maximum value of U_M, $K(t)$ is the autocorrelation function of noise at the receiver input. Often in practically significant problems, when noise can be described as white, the $K(t)$ function is a delta function. Then, solving the variational problem of finding $h_r(t)$ from (27), which provides the maximum possible value of q^2 up to a constant, we obtain

$$h_r(t)=U_0(t_0-t), \qquad U_r(t)=h_e(0)\star h_R(t)\star h_e(0)\star U_g(t), \tag{28}$$

where $U_g(t)$ is the generated signal, and the symbol $\star$ denotes the convolution of functions. Note that the first two convolutions are formed by the given functions and can be replaced by a single dependency

$$H_r(t)=h_e(0)\star h_R(t)\star h_e(0), \qquad U_r(t)=H_r(t)\star U_g(t), \tag{29}$$

which determines $h_r(t)$ from (28), which can be called the impulse response of the optimal filter (OF).

Most often, in communication, radar, and remote sensing tasks, noise is considered white. However, in an ultra-wide frequency band, the spectral noise density may differ markedly from the constant, and then its shape must be taken into account when synthesizing OF. In this case, instead of (29) from (27) up to a constant, it follows

$$U_r(t_0-t)=\int\limits_{-\infty}^{\infty} h_r(\tau)K(t-\tau)\,d\tau \tag{30}$$

and now, to determine $h_r(t)$ OF, it is necessary to numerically solve the resulting integral equation (30).

8. RESULTS OF NUMERICAL EXPERIMENTS

The problem of optimal reception of UWB pulses when reflected from an object with a smooth increase in the value of the modulus of the reflection coefficient from the frequency $R(\omega)$ was considered and a fast-variable phase characteristic. As the spectral noise density $N_0(\omega)$ the distribution of atmospheric noise in the wavelength range of 1 m–3 cm was chosen. The DP of the antenna system at each of the frequencies used corresponded to the DP of the antenna array with a beam width of $2\theta_0 = 3^\circ$ at the average frequency of the range used.

The results of solving the problem are shown in Fig. 6. Shows: the initial UWB pulse is a dashed curve; the received UWB pulse without using OF is a thin solid curve; the received UWB pulse after optimal processing in the receiver (28), (29) is a bold curve. Noticeable changes occur in the received signal due to dispersion: —the pulse duration increases significantly; —the maximum modulo values of the reflected signal, pronounced in the initial pulse, disappear; —the shape of the received pulse, and therefore the shape of its spectrum, become little similar to the generated signal.

The use of an optimal receiver for UWB signals turns out to be highly effective, since it is used in an ultra-wide frequency band. In the given example of optimization of reception from the direction $\varphi = 0$, the SNR increased by more than 150 times, which corresponds to an increase in the range of the system by 3.7 times.

When receiving a signal from a direction other than $\varphi = 0$, the filter characteristic $h_r(t)$ in accordance with (28)–(30) is no longer optimal and increases the peak value of the signal to a lesser extent than from the direction $\varphi = 0$. In the given example, the optimization gain for $\varphi = 0$ decreased by 5 times at the boundary of the transmitting beam $\varphi = \theta_0$. The revealed pattern with optimal filtering of UWB pulses shows that the effective width of the receiving DP for the signal in question becomes significantly less than θ_0. This effect can be used to improve the accuracy and angular resolution of UWB systems when searching and tracking objects with known reflection characteristics.

In practice, it is difficult to expect that the complex reflection coefficient of the object under study, especially its phase characteristic, is precisely known. However, as numerical experiments have shown, taking into account even partial information about the reflective properties of an object can significantly increase the SNR—up to 0.2–0.5 from the value of the optimal q^2. In the example given, the phase characteristic of the reflection was given as a very approximate estimate. Nevertheless, it was possible to significantly increase the SNR by about 50 times.

The obtained theoretical results and the results of numerical experiments on a mathematical model show that optimizing the reception of UWB pulses makes it possible to increase the probability of correct detection and identification of the objects under study.

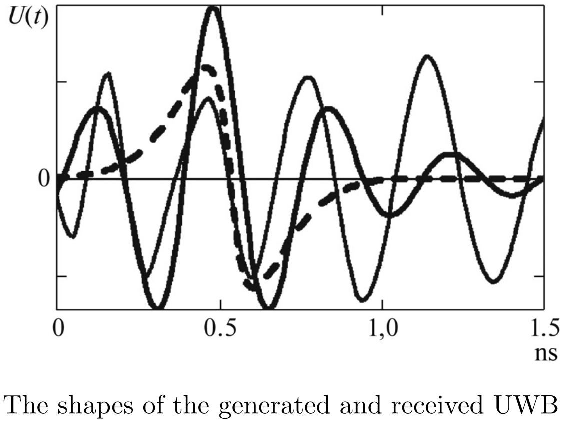


Fig. 6. The shapes of the generated and received UWB pulses.

9. CONCLUSION

1. The proposed methods of processing received signals based on the simultaneous orthogonalization of two interconnected systems of functions make it possible to increase the stability of the inverse problems being solved and restore the image of signal sources with an angular resolution several times higher than the Rayleigh criterion.

2. Algorithms based on algebraic methods make it possible to obtain satisfactory solutions with a signal-to-noise ratio of 15–20 dB, and sometimes at 11–12 dB, i.e. with significantly higher values of random components than the well-known methods described in domestic and foreign literature.

3. A priori information about signal sources allows for a targeted selection of systems of functions to represent solutions and, thereby, increase the adequacy and stability of the solutions obtained.

4. The relative simplicity of object image recovery algorithms makes it possible to use them in real time.

5. The variational problem of optimizing the pulse characteristics of the receiver of probing UWB pulses has been solved. The efficiency of using the proposed algorithms for processing UWB signals is shown, which allows 2–4 times to increase the range of UWB systems and improve the quality of radio images.

6. It is shown that optimizing the shape of the received UWB pulses allows for known object types to simultaneously increase the range of the systems, improve their angular characteristics and detection and identification characteristics.

FUNDING

The work was partially supported by the Russian Scientific Foundation, project no. 23-29-00448.

APPENDIX

Proof of Theorem 1. The system of functions $G_m(x)$ is generally non-orthogonal to L_G. Let's make a Gram matrix based on it, i.e. a matrix $\mathbf{P}$ of scalar products with elements P_{mn}:

$$P_{mn} = (G_m, G_n) = \int\limits_{\Phi} G_m(\phi)G_n(\phi)\,d\phi. \tag{A.1}$$

Since the matrix $\mathbf{P}$ is symmetric and positively defined, there is a transformation $\mathbf{T}$ that leads it to a diagonal form

$$\widetilde{\mathbf{P}} = \mathbf{T}^{\star}\mathbf{P}\mathbf{T}. \tag{A.2}$$

Using the found matrix $\mathbf{T}$, we introduce a new system of functions $\widetilde{G}_m(x)$ in the form (9). The resulting system turns out to be orthogonal on the segment L_G, which is easily verified by directly calculating scalar products:

$$(\widetilde{G}_m, \widetilde{G}_n) = \sum_{j,i=1}^{N} T_{jm}T_{in} \int\limits_{\Phi} G_j(\phi)G_i(\phi)\,d\phi = \sum_{j,i=1}^{N} T_{jm}T_{in}P_{ji} = \widetilde{P}_{mn},$$

where $\widetilde{P}_{mn}$ are the elements of the diagonal matrix (A.2).

Now let's find the system of functions $\widetilde{g}_m(x)$, which generates the resulting orthogonal in the domain L_G the system $\widetilde{G}_m(x)$, i.e.

$$\widetilde{G}_m = \mathbf{A}\widetilde{g}_m. \tag{A.3}$$

The required representation (9) follows

$$\widetilde{G}_m = \sum_{j=1}^{N} T_{mj}\mathbf{A}g_j = \mathbf{A}\left(\sum_{j=1}^{N} T_{mj}g_j\right). \tag{A.4}$$

Comparing (A.3) and (A.4), we obtain

$$\widetilde{g}_m(x) = \sum_{j=1}^{N} T_{mj}g_j(x). \tag{A.5}$$

The found system (A.5) turns out to be orthogonal on the segment L_g. Indeed, due to the orthogonality of the functions $g_m(x)$ and the orthogonality of the eigenvectors of the matrix $\mathbf{P}$ forming the matrix $\mathbf{T}$, we have

$$(\widetilde{g}_m(x), \widetilde{g}_n(x)) = \sum_{j=1}^{N} T_{mj}T_{nj}(g_j, g_i) = \begin{cases} 0, & m \neq n \\ \lambda_m, & m = n \end{cases}, \quad \lambda_m = \sum_{j=1}^{N} T_{mj}^2.$$

Note that the found system of orthogonal functions $\widetilde{g}_m(x)$ is determined by the same linear transformation $\mathbf{T}$ as the system of functions $\widetilde{G}_m(x)$.

As a result, based on a given system of N orthogonal functions $g_m(x)$ on the segment L_g, a new orthogonal system of functions on the same segment is constructed, generating an orthogonal system of functions $\widetilde{g}_m(x)$ in the domain L_g. The theorem is proved

REFERENCES

1. Odendaal, W., Barnard, E., and Pistorius, C.W.I., Two Dimensional Superresolution Radar Imaging Using the MUSIC Algorithm, *IEEE Trans.*, 1994, vol. AP-42, no. 10, pp. 1386–1391. https://doi.org/10.1109/8.320744

2. Waweru, N.P., Konditi, D.B.O., and Langat, P.K., Performance Analysis of MUSIC Root-MUSIC and ESPRIT DOA Estimation Algorithm, *Int. J. Electrical Computer Energetic Electronic and Communication Engineering*, 2014, vol. 08, no. 01, pp. 209–216.

3. Yuebo Zha, Yulin Huang, and Jianyu Yang, An Iterative Shrinkage Deconvolution for Angular Super-Resolution Imaging in Forward-Looking Scanning Radar, *Progress In Electromagnetics Research B*, 2016, vol. 65, pp. 35–48. https://doi.org/10.2528/PIERB15100501

4. Almeida, M.S. and Figueiredo, M.A., Deconvolving images with unknown boundaries using the alternating direction method of multipliers, *IEEE Trans. Image Process.*, 2013, vol. 22, no. 8, pp. 3074–3086.

5. Dudik, M., Phillips, S.J., and Schapire, R.E., Maximum entropy density estimation with generalized regularization and an application to species distribution modeling, *J. Machine Learning Research*, 2007, vol. 8, pp. 1217–1260.

6. Stoica, P. and Sharman, K.C., Maximum likelihood methods for direction-of-arrival estimation, *IEEE Trans. on Acoustics, Speech and Signal Processing*, 1990, no. 38(7), pp. 1132–1143.

7. Geiss, A. and Hardin, J.C., Radar super resolution using a deep convolutional neural network, *Journal of Atmospheric and Oceanic Technology*, 2020, vol. 37, no. 12, pp. 2197–2207.

8. Ramani, S., Liu, Z., Rosen, J., Nielsen, J., and Fessler, J.A., Regularization parameter for nonlinear iterative image restoration and MRI selection reconstruction using GCV and SURE- based methods, *IEEE Trans. on Image Processing*, 2012, vol. 21, no. 8, pp. 3659–3672.

9. Morse, P. and Feshbach, H., *Methods of Theoretical Physics*, New York: McGraw-Hill Book Company, Inc., 1953.

10. Lagovsky, B.A. and Rubinovich, E.Y., Algebraic methods for achieving super-resolution by digital antenna arrays, *Mathematics*, 2023, vol. 11, no. 4, pp. 1–9. https://doi.org/10.3390/ math11041056

11. Lagovsky, B.A., Samokhin, A.B., and Shestopalov, Y.V., Angular Superresolution Based on A Priori Information, *Radio Science*, 2021, vol. 56, no. 1, pp. 1–11. https://doi.org/10.1029/2020RS007100

12. Lagovsky, B.A., Angular superresolution in two-dimensional radar problems, *Radio Engineering and Electronics*, 2021, vol. 66, no. 9, pp. 853–858. https://doi.org/10.31857/S0033849421090102

13. Lagovsky, B.A. and Rubinovich, E.Y., Algorithms for digital processing of measurement data providing angular superresolution, *Mechatronics, automation, control*, 2021,vol. 22, no. 7, pp. 349–356. https://doi.org/10.17587/mau.22.349-356

14. Kalinin, V.I., Chapursky, V.V., and Cherepenin, V.A., Superresolution in radar and radioholographyy systems based on MIMO antenna arrays with signal recirculation, *Radio engineering and electronics*, 2021, vol. 66, no. 6, pp. 614–624. https://doi.org/10.31857/s0033849421060139

15. Shchukin, A.A. and Pavlov, A.E., Parameterization of user functions in digital signal processing to obtain angular superresolution, *Russian Technological Journal*, 2022, no. 10(4), pp. 38–43. https://doi.org/10.32362/2500-316X-2022-10-4-38-43

16. Lagovsky, B.A. and Samokhin, A.B., Superresolution in signal processing using a priori information, *IEEE Conf. Publications International Conference Electromagnetics in Advanced Applications (ICEAA)*, Italy, 2017, pp. 779–783. https://doi.org/10.1109/ICEAA.2017. 8065365.

17. Dong, J., Li, Y., Guo, Q., and Liang, X., Through-wall moving target tracking algorithm in multipath using UWB radar, *IEEE Geosci. Remote Sens. Lett.*, 2021, pp. 1–5. https://doi.org/10.1109/lgrs.2021.3050501

18. Khan, H.A., Edwards, D.J., and Malik, W.Q., Ultra wideband MIMO radar, *Proc. IEEE Intl. Radar Conf. Arlington*, VA, USA, 9 May 2005.

19. Zhou Yuan, Law Choi Look, and Xia Jingjing, Ultra low-power UWB-RFID system for precise location-aware applications, *2012 IEEE Wireless Communications and Networking Conference. Workshops (WCNCW)*, 2012, pp. 154–158.

20. Taylor, J.D., *Ultra-wideband Radar Technology*, CRC Press Boca Raton, London, New Work, Washington. 2000.

21. Holami, G., Mehrpourbernety, H., and Zakeri, B., UWB Phased Array Antennas for High Resolution Radars, *Proc. of the 2013 International Symp. on Electromagnetic Theory*, 2013, pp. 532–535.

22. Lagovsky, B.A., Samokhin, A.B., and Shestopalov, Y.V., Pulse Characteristics of Antenna Array Radiating UWB Signals, *Proceedings of the 10th European Conference on Antennas and Propagation (EuCAP 2016)*, Davos, Switzerland, 2016, pp. 2479–2482. https://doi.org/10.1109/ EuCAP.2016.7481624

23. Lagovsky, B.A., Samokhin, A.B., and Shestopalov, Y.V., Increasing accuracy of angular measurements using UWB signals. 2017 11th European Conference on Antennas and Propagation (EUCAP), *IEEE Conf. Publications. Paris*, 2017, pp. 1083–1086. https://doi.org/10.23919/ EuCAP.2017.7928204

24. Anis, R. and Tielert, M., Design of UWB pulse radio transceiver using statistical correlation technique in frequency domain, *Advances in Radio Science*, 2007, vol. 5, pp. 297–304. https://doi.org/10.5194/ars-5-297-2007

25. Niemela, V., Haapola, J., Hamalainen, M., and Iinatti, J., An ultra wideband survey: Global regulations and impulse radio research based on standards, *IEEE Communications Surveys and Tutorials*, 2016, vol. 19, no. 2, pp. 874–890. https://doi.org/10.1109/COMST.2016.2634593

26. Barrett, T., History of UWB Radar and Communications: Pioneers and Innovators, *Progress in Electromagnetics Symposium (PIERS) 2000. Microwave Journ*, January 2001.

27. Dmitriev, A.S., Efremova, E.V., and Kuzmin, L.V., Generation of a sequence of chaotic pulses under the influence of a periodic signal on a dynamic system, *Letters to the Journal of Theoretical Physics*, 2005, vol. 31, no. 22, p. 29. https://doi.org/10.1134/ S1064226906050093

28. Yang, D., Zhu, Z., and Liang, B., Vital sign signal extraction method based on permutation entropy and EEMD algorithm for ultra-wideband radar, *IEEE Access*, 2019, vol. 7. https://doi.org/10.1109/ACCESS.2019.2958600.

29. Vendik, O.G., *Antenny s nemekhanicheskimi dvizheniyami lucha* (Antennas with non-mechanical beam motion), Moskow: Sov. Radio, 1965.

30. Watson, G.N., *Theory of Bessel functions*, trans. from the 2nd English edition, Moscow: IL, 1947.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 10, pp. 1211–1220.

TOPICAL ISSUE

Control of a Free-Flying Space Manipulation Robot with a Payload

V. Yu. Rutkovskii[*†] **and M. V. Glumov**[*,a]

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

Received June 19, 2023
Revised July 14, 2023
Accepted August 2, 2023

Abstract—The control modes of a free-flying space manipulation robot during the transportation and installation of a building element on a large space structure are considered. It is proposed to save the working fluid of the gas-jet engines of the robot body when moving along the trajectory by using the mobility of a manipulator with electromechanical drives for the angular stabilization of the mechanical "robot–transported element" system. Conditions ensuring the stable motion of the manipulator in the working area when installing the element on the assembled structure are obtained. A stability domain is determined to select the initial configuration of the manipulator before installing the element and its admissible change during installation. The control algorithms are designed based on the principle of dynamic feedback systems.

Keywords: free-flying space manipulation robot, working area, technical controllability, control algorithm, motion stability

DOI: 10.25728/arcRAS.2023.67.85.001

1. INTRODUCTION

In space technology, space manipulation robots (SMRs) are used for servicing and assembling various-purpose spacecraft in orbit. Such robots fly freely in space due to their movement system independently of the spacecraft that delivered the robot to the destination point [1]. The feasibility of developing this type of space robotic devices was declared at the 6th IFAC Symposium on Space Control (1974), which was held under the leadership of Academician B.N. Petrov [2]. Currently, there are two ways of connecting spacecraft and modules in space: direct docking and berthing (docking by means of a manipulator) [3]. The latter term defines several operations, such as soft docking, payload stowage in the receiving compartment of a cargo spacecraft, etc. This paper considers the problem of attaching a building element to a large space structure (LSS) being assembled in orbit by means of an SMR, another operation of the same type. As in [3], the mass of the LSS element may significantly exceed that of the manipulator, whose gripper with the held payload may be located at a significant distance from the center of gravity of the SMR body and the entire mechanical system. The kinematic algorithm used to control the manipulator converts control signals into the required rotational velocities of the actuators; this algorithm considers the geometric and kinematic constraints determined by the current configuration of the manipulator.

Structurally, a freely-flying space manipulation robot (FSMR) is designed as a platform with one or several manipulators attached. The platform is equipped with control devices and a set of actuators that provide the required orientation and desired trajectory of the platform in outer

[†] Deceased.

space. Such SMRs are called free-flying robots in the literature [4, 5]. One of the first domestic publications [6] presented a methodology for analyzing the dynamics of a manipulator on a moving base and one solution to capture a payload in inertial space by means of an FSMR in the cases of its stabilizable and non-stabilizable body.

An assembly operation performed in space includes two stages as follows. In the first stage, an FSMR approaches the installation zone of a building element; at the end of this stage, the robot hovers in the vicinity of the docking point of the element with an LSS. The boundary of the working area is determined, on the one hand, by safety conditions (no possible contact between the hovering robot and the LSS when installing the element) and, on the other, by goal attainability conditions (the successful installation of the element with a given orientation in the required point of the LSS) [7, 8]. The latter is the content of the second stage of the assembly operation. The first stage of the assembly operation is implemented by means of a control system of the translational and angular movements of the FSMR using reaction forces and torque applied by the actuators to the robot body. The manipulator with the transported element is stationary in this stage, and its configuration should be as close to optimal as possible [7].

The list of problems arising in the design of control systems for FSMRs was described in [1]. This paper considers those of manipulator control during the FSMR movement to the working area and in it. When controlling an FSMR in its working area in the free-floating mode (i.e., the angular position control system of the robot body is disabled), the challenges include the narrowing of the working area [4, 9] and the presence of dynamic singularities [10, 11]. The dynamics and kinematics of the mechanical structure of an FSMR in this mode are significantly complicated due to the disturbing effect of the manipulator motions on the body position [12, 13]. Therefore, we consider manipulator control based on the feedback principle using information about the angular position of the robot body and estimates of the deviation of the manipulator's endpoint from the target point [14, 15].

We present solutions of two problems as follows. The first problem arises if it is necessary to save the working fluid of the gas-jet engines of the robot body when moving along the trajectory. This problem is solved using the manipulator mobility. The second problem is related to the stabilization of the manipulator movement in the working area when installing the transported element on the LSS.

2. THE MECHANICAL STRUCTURE OF A FREE-FLYING SPACE MANIPULATION ROBOT: KEY FEATURES

The mechanical structure of an FSMR is a set of elements connected through joints. The main element is the body equipped with a control system and jet engines. Multilink manipulators are attached to the body. A gripper, a device for capturing and holding a payload during manipulation operations of an FSMR, is rigidly fixed on the end link of each manipulator. This mechanical structure is characterized by many degrees of freedom and the mutual influence of the movements of its elements. The FSMR body responds to dynamic reaction forces arising from the movements of the manipulator's links. When controlling the configuration and angular movements in such a mechanical system, it is necessary to consider the dynamic coupling between the body and manipulators [15].

To illustrate the key features of its mechanical structure, we consider the plane motion of an FSMR with one three-link manipulator [16] as one possible setup. The coordinates X_0 and Y_0 of the FSMR body's center of gravity and its angle of rotation ϑ are the generalized coordinates describing the position of the FSMR body in the inertial frame CXY whose axes are associated with an LSS. They form the vector $q_K = (X_0, Y_0, \vartheta)^{\mathrm{T}}$. The vector $q_\alpha = (\alpha_1, \alpha_2, \alpha_3)^{\mathrm{T}}$ consists of the generalized coordinates of the inter-link angles that specify the manipulator's configuration. The vector $\rho_{BA} = (X_\varepsilon, Y_\varepsilon)^{\mathrm{T}}$ represents the controlled coordinates, i.e., the deviations of the endpoint of the transported payload $B = (X_B, Y_B)$ from the target point $A = (X_A, Y_A)$ in the inertial frame: $X_\varepsilon = X_A - X_B$ and $Y_\varepsilon = Y_A - Y_B$.

The plane motion of this setup is described by the equation

$$A(q)\ddot{q} = M(q,u) + F(q,\dot{q}) \tag{1}$$

for the vector $q = (q_K, q_\alpha)^{\mathrm{T}}$ with the following notations [15, 16]: the matrix $A(q) \in R^{6\times 6}$ contains blocks of symmetric matrices specifying the mass-inertia parameters of the body and manipulator ($A_{11}(q) \in R^{3\times 3}$ and $A_{22}(q) \in R^{3\times 3}$, respectively) as well as the dynamic interaction coefficients of the body and manipulator's links ($A_{12}(q) \in R^{3\times 3}$ and $A_{21}(q) \in R^{3\times 3}$, respectively, where $A_{12}(q) = A_{21}(q)$); $M(q,u) = (M_K, M_\alpha)^{\mathrm{T}}$, where $M_K \in R^3$ is the vector of control actions applied to the robot body and $M_\alpha \in R^3$ is the vector of control actions applied by actuators to the manipulator's links when feeding control voltages $u(t)$ to the former's inputs; finally, $F(q,\dot{q}) = (f_K(q,\dot{q}), f_\alpha(q,\dot{q}))^{\mathrm{T}}$ is the vector of nonlinear disturbance functions from Coriolis and centrifugal forces. The expressions for calculating the elements of these matrices and vectors were given in [16].

In this paper, the actuators of manipulator's links are assumed to have DC motors with independent excitation [16]. In the first approximation with the time constant of the motor and mechanical nonlinearities being neglected [16, 17], the dynamics of each jth actuator ($j = \overline{1,3}$) are described by the equations

$$J_j i_{pj} \ddot{\alpha}_j = (k_{bj} k_{aj})^{-1} u_j(t) - k_{aj}^{-1} i_{pj} \dot{\alpha}_j - M_{Rj}(t), \quad j = \overline{1,3}, \tag{2}$$

where $\alpha_j \in q_\alpha$, J_j is the moment of inertia of the jth actuator reduced to the motor shaft, i_{pj} is the gearing ratio, M_{Rj} is the moment of dynamic load on the motor shaft from the manipulator, and k_{bj} and k_{aj} are constants.

Self-braking mechanical gears [15, 16] are often used in the link actuators to reduce the energy cost of controlling the FSMR manipulator. The self-braking property is provided by imposing an impulse coupling on the moving link of the manipulator; as a result, $\dot{\alpha}_j = 0$ and $u_j = 0$. Due to self-braking, the equation for α_j disappears from (1), which is mathematically expressed as a decrease (or increase) in the order of system (1) by $2 \times r$, where r denotes the number of simultaneously braked (or unbraked) links of the manipulator. The FSMR model (1), (2) serves to design robot motion control algorithms in different operation modes of the manipulator [15].

When designing angular motion control algorithms for an FSMR, it is necessary to consider the property of technical controllability, a necessary condition for the performance of the robot [18]. For an FSMR, this property means that the angular motion of the robot body and the movements of the manipulator's links must be controllable. In other words, when control signals are supplied to change their positions, these changes must be implemented in a required direction and with a given speed. It is reasonable to analyze the controllability of FSMRs based on a simplified angular motion model of the robot mechanical system under the following assumptions [18]: for each q_i, there exists M_i with the constraint $|M_i| \leqslant M_i^{\max} > 0$, $i = \overline{1,6}$; given $M_j = 0$, $i,j = \overline{1,6}$, $j \neq i$, at a time instant $t = t_0$ and $q_i(t) = \dot{q}_i(t) = \ddot{q}_i(t) = 0$ ($t < t^*$), the desired response to $M_i^{\max} > 0$ is $q_i(t) \geqslant 0$ at time instants $t > t^*$; the velocities $\dot{q}$ are small enough to nullify the terms of the full motion model that depend on the products of $\dot{q}$; the motion equations of the model can be linearized with respect to the position $q = q^*$, where $q_i^* = \text{const}$, $i = \overline{1,6}$.

The angular motion model linearized in the position q^* has the form

$$A(q^*)\Delta\ddot{q} = P(q^*)M(q), \tag{3}$$

where $\Delta q = q - q^*$, $A(q^*)$ is a positive definite matrix, and the matrix $P(q^*)$ relates the generalized forces to the vector of control forces and moments [18].

The FSMR with model (3) is controllable in Δq_i $i = \overline{1,6}$ in the position $q = q^*$ if under the zero initial conditions $\Delta q_i(t) = \Delta\dot{q}_i(t) = \Delta\ddot{q}_i(t) = 0$ $\forall t < t_0$, supplying the maximum control $|M_i(t)| = M_i^{\max}$ $\forall t \geqslant t_0$ at the time instant t_0 generates an acceleration $\Delta\ddot{q}_i(t) \geqslant \eta_i \neq 0$ of the same sign as $M_i(t)$ irrespective of the other control actions $M_i(t)$ $(j = \overline{1,6}; j \neq i)$, where η_i are known characteristic values of the mechanical system. According to the theorem proved in [18], the controllability of the FSMR in the neighborhood of the point $q = q^*$ is determined only by the design parameters of the robot's mechanical system and not by the vector of control constraints $M^{\max}$.

3. TRAJECTORY MOTION CONTROL OF A FREE-FLYING SPACE MANIPULATION ROBOT

Consider a section of the FSMR trajectory that starts when turning the cruise engine off and ends when reaching the boundary of the manipulator's working area. On this section, the FSMR motion control system must eliminate the residual lateral velocity and the lateral deviation of the robot from the line of sight as well as stabilize the angular position of its body. If gas-jet engines are used as actuators, the problem is to reduce the consumption of the onboard working fluid of the engines. This problem will be solved for control design by the joint use of gas-jet nozzles and torque actuators of the manipulator. For brevity, such control will be referred to as cost-efficient control.

When the FSMR with the transported element of the LSS moves along the trajectory, its manipulators must be fixed in a position that aligns the center of gravity of the robot's mechanical system with the center of application of the control forces [7]. The manipulator with the transported element is stationary, and the trajectory and angular motion of the FSMR are controlled using basic algorithms that form the control actions M_ϑ applied to the robot body from gas-jet nozzles. Under cost-efficient control on the considered trajectory section, we propose to provide the limited mobility of the manipulator. In this case, the required angular stabilization of the body is implemented through motion exchange between the robot body and the manipulator's links by applying control torques from the electromechanical actuators of the manipulator, the electrical energy of which can be recovered. Due to restrictions on the admissible movements of the manipulator's links to control the angular position of the FSMR body, the angles of rotation of the links may reach the limit values, making further control by the electromechanical method impossible. When restoring the initial configuration of the manipulator, the required angular orientation of the body is provided by means of gas-jet nozzles. For brevity, this restoration process will be called the manipulator's unloading mode.

The following features must be considered when forming cost-efficient control algorithms: there are bounded domains of varying the coordinates of the manipulator's links ($|\alpha_i(t)| \leqslant \alpha_{i\,\max}$, $|\dot{\alpha}_i(t)| \leqslant \dot{\alpha}_{i\,\max}$); the deviation of the manipulator's links from the initial position displaces the FSMR's center of gravity relative to the center of application of the forces and, therefore, is a parametric disturbance in the robot orientation system; gas-jet actuators are relay elements, and the torques of electromechanical actuators are bounded; the conditions of technical controllability by the vector q_α hold in the entire domain of varying the coordinates of system (1).

Let $u_\vartheta(\vartheta, \dot{\vartheta}, t)$ be the basic orientation control algorithms for the FSMR and $u_\alpha(\alpha, t)$ be the manipulator's configuration control algorithms. In the case of cost-efficient control, initially implemented by the control action $M_{\alpha 1}$ from the arm link actuator only, the FSMR motion equations of motion of the SCMR have the form

$$A_1(q)\ddot{q}_1 = F_q + F_q^d, \tag{4}$$

where $q_1 = (\vartheta, \alpha_1, X_0, Y_0)^{\mathrm{T}}$, $F_q = (0, M_{\alpha 1}, 0, F_y)^{\mathrm{T}}$ is the vector of control actions used, $F_q^d = (M_\vartheta^d, 0, 0, 0)^{\mathrm{T}}$ is the vector of disturbances considered, $A_1(q) = [a_{ij}(\alpha_1, \lambda)]$ is a symmetric matrix, and λ is the vector of the parameters of the FSMR and LSS element.

The coordinate ϑ varies according to the solution of equation (4) of the form

$$\ddot{\vartheta} = k_0(k_\alpha M_{\alpha 1} + k_d M_\vartheta^d + k_y F_y), \tag{5}$$

where $k_0 = (\det[A_1(q)])^{-1}$; $k_\alpha(\alpha_1, \lambda) = -D_{21}q$ is the efficiency coefficient of $M_{\alpha 1}$ when applied to ϑ, representing the algebraic complement of the element $a_{21}(q)$ for $\det[A_1(q)]$; $k_d(\alpha_1, \lambda) = D_{11}(q)$ is the efficiency coefficient of the exogenous disturbance M_ϑ^d on ϑ, representing the algebraic complement of the element $a_{11}(q)$ for $\det[A_1(q)]$; finally, $k_y(\alpha_1, \lambda) = -D_{41}(q)$ is the efficiency coefficient of the control channel F_y, representing the algebraic complement of the element $a_{41}(q)$ for $\det[A_1(q)]$.

We construct the stabilizing control action $M_{\alpha 1}$ for the coordinate ϑ in the form

$$M_{\alpha 1}[u_{\alpha 1}(t)] = -\tilde{k}_0 k_A(\vartheta + k_{\dot{\vartheta}}\dot{\vartheta}), \tag{6}$$

where $k_A = (k_m k_b)^{-1}$ is the static gain of the actuator; $\tilde{k}_0$ is a tunable parameter of the control algorithm $u_{\alpha 1}(t)$ (if necessary); $k_{\dot{\vartheta}}$ is a constant.

If the control action (6) is implementable, then the linear part of the basic algorithm is designed to ensure stability and the desired quality of the motion (5). Considering (6), let us write (5) as

$$\ddot{\vartheta} + k_A k_{\dot{\vartheta}} \tilde{k}_0 \bar{k}_\alpha(\alpha_1, \lambda)\dot{\vartheta} + k_A \tilde{k}_0 \bar{k}_\alpha(\alpha_1, \lambda)\vartheta = \bar{M}_\Sigma^d(\alpha_1, \lambda, t), \tag{7}$$

where $\bar{k}_\alpha(\alpha_1, \lambda) = k_0 k_\alpha(\alpha_1, \lambda)$ is the reduced efficiency coefficient of the control action $M_{\alpha 1}$ (according to the technical implementability theorem and [18], this coefficient satisfies the condition $\bar{k}_\alpha(\alpha_1, \lambda) > 0 \forall (\alpha_1, \alpha_2) \in (0, \pm\pi)$); $\bar{M}_\Sigma^d(\alpha_1, \lambda, t) = \bar{k}_d(\alpha_1, \lambda) M_\vartheta^d + \bar{k}_y(\alpha_1, \lambda) F_y$ is the resulting reduced disturbing torque; finally, $\bar{k}_y(\alpha_1, \lambda) = k_0 k_y(\alpha_1, \lambda)$ and $\bar{k}_d(\alpha 1, \lambda) = k_0 k_d(\alpha_1, \lambda)$.

If the parameters λ are known and $\alpha_1(t)$ is measured, we propose an algorithm to change $k_0(t)$ in (6) based on the stationarity condition

$$\tilde{k}_0(t)\bar{k}_\alpha(\alpha_1, \lambda) = K, \tag{8}$$

where K is a constant satisfying the desired quality of motion for ϑ.

Under (8), the coefficients in (7) are constant, which ensures the stability of motion for ϑ. For the motion (7) with (8), the required static accuracy $|\vartheta(t)| \leqslant \vartheta_{\min}$ of FSMR orientation, where $\vartheta_{\min}$ is a given value, is achieved by fulfilling the condition $k_A \geqslant (K\vartheta_{\min})^{-1}\left[M_\Sigma^d(\alpha_1, \lambda, t)\right]_{\max}$.

The control action (5) is implemented by supplying the voltage $u_{\alpha 1}(t)$ to the input of the electrical actuator (2) of the manipulator's shoulder link according to the algorithm [16]

$$u_{\alpha 1}(t) = -\frac{\tilde{k}_0}{i_g}\left[\left(1 + \frac{k_{\dot{\vartheta}}}{k_m J_m}\right)\vartheta + (k_m J_m)^{-1}\int \vartheta dt + k_{\dot{\vartheta}}\dot{\vartheta} + \frac{k_b J_L}{\tilde{k}_0 J_m}\dot{\alpha}_1\right], \tag{9}$$

where $M_L \approx -J_L \ddot{\alpha}_1$ and J_L is the moment of inertia of the load reduced to the shoulder joint.

The algorithm (9) is used until reaching the domain $|\vartheta(t)| < \vartheta_{\min}$ by ϑ. The system then switches to a nonlinear algorithm containing nonlinearities (dead zone and hysteresis) to organize highly cost-efficient unilateral auto oscillations in this coordinate domain.

Generally, the residual nonzero initial conditions $\vartheta_0, \dot{\vartheta}_0$ and the forced motions generated by exogenous disturbances are damped using the algorithm (9) by changing the coordinates $\alpha_i(t)$. The damping process ends either with steady-state small oscillations in the domain $|\vartheta(t)| \leqslant \vartheta_{\min}$ or with $\alpha_i(t)$ reaching the constraints. In the latter case, it becomes necessary to return the manipulator to the initial position (the unloading mode) in order to implement the orientation control method for the FSMR using its mobility again. In the unloading mode, the manipulator's links are transferred to the initial state, $\alpha_i(t) \to \alpha_i^*$, under the action of its control $M_\alpha(u_\alpha)$ while

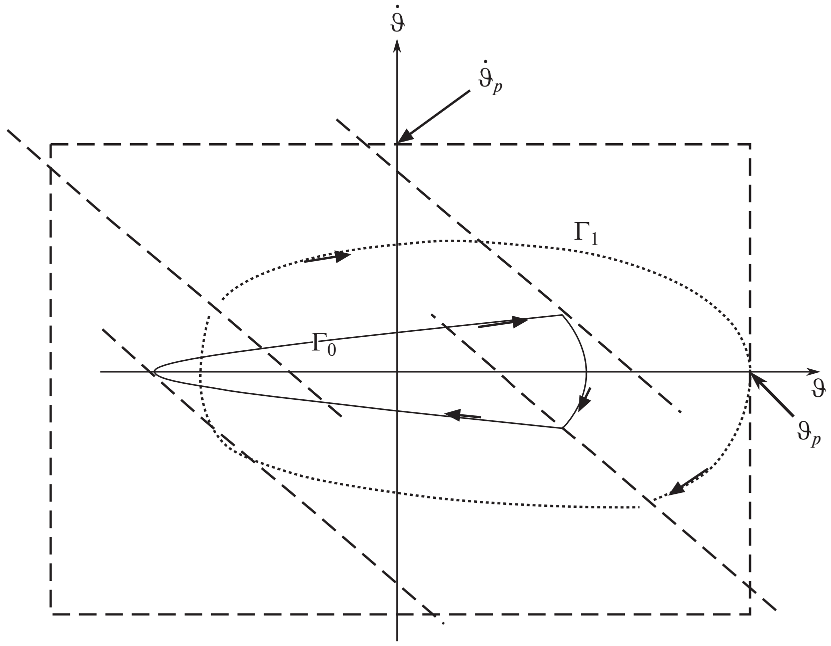


Fig. 1. Limit cycles in the manipulator's unloading mode.

keeping ϑ in the domain $|\vartheta(t)| \leqslant \vartheta_{\min}$. The angular stabilization of the body is implemented by the torque $M_\vartheta(u_\vartheta) \leqslant M_\vartheta^{\max}$, where $M_\vartheta^{\max}$ is the existing constraint. In this mode, robot control is a multilink control problem in an essentially nonlinear system with control constraints.

When describing the unloading mode in (4), it is necessary to assume

$$F_q = (M_\vartheta, M_{\alpha 1}, 0, F_y)^{\mathrm{T}}.$$

Then the behavior of the coordinate ϑ is described by the equation

$$\ddot{\vartheta} = \bar{k}_{M\vartheta}(\alpha_1, \lambda) M_\vartheta(u_\vartheta) + f_p(\alpha_1, \lambda), \tag{10}$$

where $\bar{k}_{M\vartheta}(\alpha_1, \lambda)$ is the efficiency coefficient of $M_\vartheta(u_\vartheta)$, calculated by analogy with (5); $f_p(\alpha_1, \lambda) = \bar{M}^d_\Sigma(\alpha_1, \lambda, t) + \bar{k}_\alpha(\alpha_1, \lambda) M_{\alpha 1}$ are disturbances for $|\vartheta(t)| \leqslant \vartheta_{\min}$.

When designing control algorithms for the coordinates ϑ and α, it is necessary to consider the contradictory requirements for the operation of each subsystem. Minimizing the roll time of the manipulator's link that has reached the constraint implies using the maximum achievable speeds $\dot{\alpha}_{1\max}$ of the output shaft of the actuator (under the existing constraints). However, the disturbances $f_p(\alpha_1, \lambda)$ arising in the control subsystem ϑ under the constraint $M_\vartheta^{\max}$ may violate the orientation accuracy requirements. In this case, the roll rate of the manipulator's link should be bounded by a value smaller than $\dot{\alpha}_{1\max}$. It is reasonable to use the phase plane method based on (10) to determine the optimal controller parameters ensuring the desired dynamics in the unloading mode.

Let the basic nonlinear algorithm for stabilizing the angular position of the FSMR $u_\vartheta(\vartheta, \dot{\vartheta}, t)$ generate unilateral auto oscillations represented on the phase plane $(\vartheta, \dot{\vartheta})$ as a limit cycle Γ_0 (Fig. 1). In the unloading mode of the manipulator, under the action of $f_p(\alpha_1, \lambda)$, the undisturbed cycle Γ_0 is transformed into another stable cycle Γ_1. The cycle Γ_1 is formed so that for the maximum possible value $f_{p,\max}(\alpha_1, \lambda)$, its phase trajectory would not cross the limits of the admissible deviations of the controlled coordinates ($|\vartheta| = \vartheta_p, |\dot{\vartheta}| = \dot{\vartheta}_p$); see the dashed box in Fig. 1. When the unloading process is complete, the original cycle is restored ($\Gamma_1 \to \Gamma_0$), and a return to the control action $M_\alpha(u_\alpha)$ follows.

Simultaneously with FSMR orientation control, the correction system continues to work for the transverse displacements of the body: when the deviation exceeds $Y_0 = Y_{0\,\min}$, it generates the control action F_y in (4). Since the mechanical structure of the FSMR has an unbalanced configuration, the disturbing torque $M^d_{Fy} = F_y x_c$ arises in the orientation control channel; its compensation by the action of $M_\alpha(u_\alpha)$ may be insufficient. Therefore, when F_y acts in the orientation control system, it is necessary to provide an automatic transition to the efficient nonlinear control $M_\vartheta(u_\vartheta)$.

4. MANIPULATOR CONTROL WHEN INSTALLING AN ELEMENT ON AN OBJECT

Consider FSMR manipulator control in the soft installation mode of a building element on an LSS in the working area. Here, the motion of the robot body when using manipulator's self-braking actuators does not change q_α. The controlled motion of each link must not change the values of other interlink angles. This property, characteristic of the class of mechanical systems under consideration, holds under the conditions of technical controllability (if satisfied during the design process). These conditions allow neglecting the mutual influence of joints and, consequently, treating the matrix A_{22} as a diagonal one. For $t \geqslant t_0$, where t_0 is the time of entering the working area, the coordinates of the FSMR mechanical system with the transported element of the LSS change with sufficiently small rates. Hence, linear mathematical models can be used to design algorithms [15]. The terms of the functions $f_K(q, \dot q)$ and $f_\alpha(q, \dot q)$ contain products of small values $(\dot q_i \dot q_j)$, $i, j = \overline{1, 6}$; hence, their contribution to FSMR dynamics can be neglected in the first approximation. In the presence of all these features, the motion of (1) can be approximately described by

$$A_r(q)\ddot q = M(q, u), \tag{11}$$

where the matrix $A_r(q)$ consists of the blocks $A_{r,11} = A_{11}$, $A_{r,12} = A_{12}$, $A_{r21} = 0$ and $\dot q(t_0) = 0$.

When the coordinates X_ε and Y_ε are selected as the controlled ones, manipulator control by the coordinates q_α becomes open-loop and it is possible to reach the domain $|X_\varepsilon| \leqslant X_{\varepsilon,\min}$, $|Y_\varepsilon| \leqslant Y_{\varepsilon,\min}$ by purposefully varying $q_\alpha(t)$, where $X_{\varepsilon,\min}$ and $Y_{\varepsilon,\min}$ are given values. Using only the rotating degrees of freedom of the mechanical FSMR system allows neglecting the displacement of the body's center of gravity and treating q_1 and q_2 as constants. If the control actions by α_1 and α_2 are formed in the soft installation mode of the element and the manipulator's end link is fixed (α_3 is a constant), then the motion for X_ε and Y_ε based on (11) is described by

$$\begin{aligned} \ddot X_\varepsilon &= d_{11}(q)M_{\alpha 1} + d_{12}(q)M_{\alpha 2},\\ \ddot Y_\varepsilon &= d_{21}(q)M_{\alpha 1} + d_{22}(q)M_{\alpha 2}, \end{aligned} \tag{12}$$

where

$$\begin{aligned} d_{11}(q) &= b_\Delta a_{44}^{-1}\left[(b_m - a_{23}^2)(a_{14}a_{33} - a_{13}a_{34}) + b_3(a_{24}a_{33} - a_{23}a_{34})\right],\\ d_{12}(q) &= b_\Delta a_{55}^{-1}\left[(b_m - a_{23}^2)(a_{15}a_{33} - a_{13}a_{35}) + b_3(a_{25}a_{33} - a_{23}a_{35})\right],\\ d_{21}(q) &= b_\Delta a_{44}^{-1}\left[(b_m - a_{13}^2)(a_{24}a_{33} - a_{23}a_{34}) + b_3(a_{14}a_{33} - a_{13}a_{34})\right],\\ d_{22}(q) &= b_\Delta a_{55}^{-1}\left[(b_m - a_{13}^2)(a_{25}a_{33} - a_{23}a_{35}) + b_3(a_{15}a_{33} - a_{13}a_{35})\right], \end{aligned}$$

$b_\Delta = \left[a_{33}m_S(b_m - a_{13}^2 - a_{23}^2)\right]^{-1}$, $b_m = a_{33}m_S^2$, $b_3 = a_{13}a_{23}$, and m_S denotes the FSMR mass.

The coefficients $d_{ij}(q)$, $i, j = \overline{1, 2}$, vary due to their dependence on the angular position of the FSMR body through ϑ and on the joint angles α_1 and α_2. During manipulator control in the working area, its links may take positions in which $d_{ij}(q) < 0$, causing instability for X_ε and Y_ε.

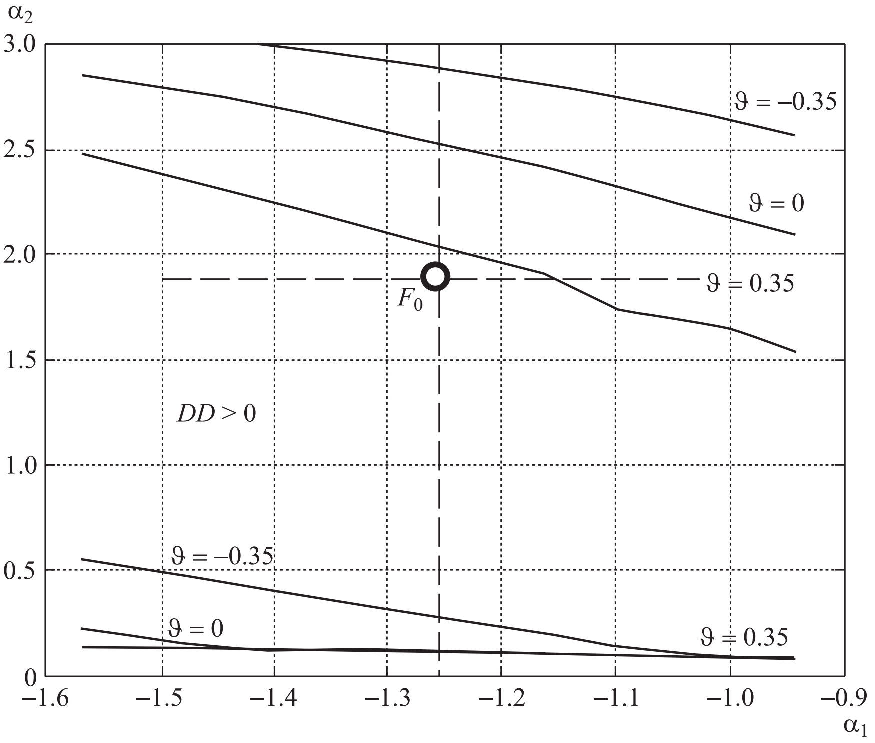


Fig. 2. The effect of the angle ϑ on the boundaries of the stability domain.

If the information about $X_\varepsilon, Y_\varepsilon$, $\dot{X}_\varepsilon$, and $\dot{Y}_\varepsilon$ is available, then stable control by X_ε and Y_ε is ensured by the PD algorithms

$$\begin{aligned} M_{\alpha 1} &= k_{0x}(k_{1x}X_\varepsilon + k_{2x}\dot{X}_\varepsilon), \\ M_{\alpha 2} &= k_{0y}(k_{1y}Y_\varepsilon + k_{2y}\dot{Y}_\varepsilon), \end{aligned} \tag{13}$$

where the gains k_{jx}, k_{jy} $(j = \overline{0,2})$ must be appropriately chosen to stabilize the trivial solution of system (12), (13). These stability requirements are defined when analyzing the characteristic equation

$$\sum_{j=0}^{4} c_j \lambda^j = 0,$$

where

$$\begin{gathered} c_0 = \Delta d k_{1x} k_{1y}; \quad c_1 = \Delta d (k_{1y}k_{2x} + k_{1x}k_{2y}); \\ c_2 = \Delta d k_{2x} k_{2y} - (k_{1x}k_{0x}d_{11} + k_{2x}k_{0y}d_{22}); \\ c_3 = -(k_{1y}k_{0x}d_{11} + k_{2y}k_{0y}d_{22}); \quad c_4 = 1; \\ \Delta d = k_{0x}k_{0y}(d_{11}d_{22} - d_{12}d_{21}). \end{gathered}$$

The necessary stability condition $c_j > 0\, \forall j = \overline{0,4}$ holds for $\Delta d > 0$ and $\text{sgn} d_{11} \neq \text{sgn} k_{0x}$, $\text{sgn} d_{22} \neq \text{sgn} k_{0y}$. The condition $\Delta d > 0$ does not depend on the gains k_{jx}, k_{jy}, $j = \overline{0,2}$, and is satisfied under the relations

$$\begin{gathered} (\text{sgn} d_{11} \neq \text{sgn} d_{22} \wedge \text{sgn} d_{12} = \text{sgn} d_{21}) \vee (\text{sgn} d_{11} = \text{sgn} d_{22} \wedge \text{sgn} d_{12} \neq \text{sgn} d_{21}); \\ (\text{sgn} d_{11} = \text{sgn} d_{22} \wedge \text{sgn} d_{12} = \text{sgn} d_{21} \wedge |d_{11}d_{22}| > |d_{12}d_{21}|) \\ \vee (\text{sgn} d_{11} \neq \text{sgn} d_{22} \wedge \text{sgn} d_{12} \neq \text{sgn} d_{21} \wedge |d_{11}d_{22}| < |d_{12}d_{21}|). \end{gathered} \tag{14}$$

If the variations of $d_{ij}(q)$ do not violate the condition $\Delta d > 0$, then the stability conditions are satisfied by varying the gains in (13). If α_1 and α_2 are measurable during FSMR manipulator control, then $d_{ij}(q)$ can be calculated and the stability conditions can be maintained by tuning the gains in (13) at appropriate time instants.

Based on (14), it is reasonable to form the stability domain in the coordinates α_1 and α_2. Information about this domain serves to choose the initial configuration of the manipulator before the element installation and the admissible variation of α_1 and α_2 during the installation process. The topology of the stability domain depends on the values ϑ and α_3, which determine the relative position of the body and the element to be installed. As one example with the data from [15], Fig. 2 shows a segment of the stability domain for $\alpha_3 = -0.2\pi$ and three initial positions of the FSMR body ($\vartheta = [-0.35; 0; 0.35]$). Here, the point F_0 indicates the initial position of the links: $\alpha_1(t_0) = -1.26$ and $\alpha_2(t_0) = 1.58$. According to Fig. 2, increasing the positive value of the angle ϑ reduces the stability domain where $\Delta d > 0$. (In this figure, the stability domain is indicated by $DD > 0$.) This fact decreases the range of varying the angles α_1 and α_2 when the element is installed by the manipulator.

Note that the variation of the angle α_3 (the gripper's position) has a smaller effect on the boundaries of the stability domain compared to the variation of the angular position of the FSMR body.

5. CONCLUSIONS

The features of the mechanical structure of the FSMR have been analyzed, and a solution has been proposed to reduce the consumption of the onboard working fluid of gas-jet engines during transportation of the LSS element and during its assembly in orbit. This solution involves the mobility of the manipulator to stabilize the angular position of the FSMR body. On separate sections of the FSMR motion trajectory, the control is jointly implemented by two types of actuators: gas-jet nozzles and torque electromechanical actuators of the manipulator. The mathematical models of FSMR motion used in this paper are convenient for designing control algorithms based on the feedback principle and studying the manipulation processes of the FSMR. The control algorithms of the FSMR satisfy the conditions of technical controllability and maintain the required configuration of the mechanical structure of the robot during the transportation and installation of the LSS element. Under sufficiently small velocities of the manipulator's joints, the algorithms presented above provide in the working area a soft installation of the element at a given point of the LSS. The stability domain in the space of the angles of the manipulator's joints has to be determined in advance in order to choose the initial configuration of the robot's mechanical system before the manipulation operation and the range of these angles during the operation that ensures stable motion.

REFERENCES

1. Papadopoulos, E., Aghili, F., Ma, O., and Lampariello, R., Manipulation and Capture in Space: A Survey, *Front. Robot. AI.*, 2021, no. 8. P. 1–36.

2. Hung, J., Irwin, J., and Moore, F., Free-flying Teleoperator for Space Missions, *Pros. of 6th IFAC Symposium on Control in Space*, 1976, vol. 2, Moscow: Nauka, pp. 173–180.

3. Yaskevich, A.V., A Mathematical Model of a Space Manipulator for the Scaled-Down Customizing of the Operations of Berthing an Effective Load, *J. Comput. Syst. Sci. Int.*, 2004, vol. 43, no. 4, pp. 644–662.

4. Dubowsky, S. and Papadopoulos, E., The Kinematics, Dynamics and Control of Free-Flying and Free-floating Space Robotic Systems, *IEEE Transact. Robot. Automat.*, 1993, vol. 9, no. 5, pp. 531–543.

5. Moosavian, S., Ali, A., and Papadopoulos, E., Free-flying Robots in Space: An Overview of Dynamics Modeling, Planning and Control, *Robotica*, 2007, vol. 25, no. 5, pp. 537–547.

6. Popov, T.P., Medvedev, V.S., and Yuschenko, A.S., Free-flying Manipulation Robot Computer Control, *Proc. of the 8th IFAC Symposium on Automatic Control in Space*, 1979, Oxford: Pergamon Press, pp. 295–301.

7. Bogomolov, V.P., Rutkovskii, V.Yu., and Sukhanov, V.M., Design of an Optimal Mechanical Structure of a Free-flying Space Robotic Module as a Control Object. I, *Autom. Remote Control*, 1998, vol. 59, no. 5, part 1, pp. 632–642.

8. Sukhanov, V.M., Rutkovskii, V.Yu., and Glumov, V.M., Determination of Workspace and Required Initial Position of Free-Flying Space Manipulator at Target Capture, *Autom. Remote Control*, 2014, vol. 75, no. 11, pp. 953–963.

9. Vafa, Z. and Dubowsky, S., On the Dynamics of Manipulators in Space Using the Virtual Manipulator Approach, *Proc. IEEE Int. Conf. Robot. Automat.*, 1985, pp. 579–585.

10. Yoshida, K. and Umetani, Y., Control of Space Free-Flying Robot, *Proc. 29 IEEE Conf. Decision Control*, 1990, pp. 97–102.

11. Papadopoulos, E. and Dubowsky, S., Dynamic Singularities in the Control of Free-floating Space Manipulators, *ASME J. Dyn., Syst. Meas., Contr.*, 1993, vol. 115, no. 1, pp. 44–52.

12. Rubus, T., Seweryn, K., and Sasiadek, J.Z., Control System for Free-floating Space Manipulator on Nonlinear Model Predictive Control (NMPC), *Intell. Robot. Syst.*, 2017, no. 85, pp. 491–509.

13. Somov, Y., Butyrin, S., Somova, T., and Somov, S., Control of a Free-flying Robot at Preparation for Capturing a Passive Space Vehicle, *IFAC-PapersOnLine*, 2018, vol. 51, no. 30, pp. 72–76.

14. Rutkovskii, V.Yu., Sukhanov, V.M., and Glumov, V.M., Some Issues of Controlling the Free-flying Manipulative Space Robot, *Autom. Remote Control*, 2013, vol. 74, no. 11, pp. 1820–1837.

15. Sukhanov, V.M., Silaev, A.V., and Glumov, V.M., Dynamic Equations of Free-flying Space Robot for Feedback Control Tasks, *Autom. Remote Control*, 2015, vol. 76, no. 8, pp. 1446–1454.

16. Rutkovskii, V.Yu., Sukhanov, V.M., and Glumov, V.M., Motion Equations and Control of the Free-flying Space Manipulator in the Reconfiguration Mode, *Autom. Remote Control*, 2010, vol. 71, no. 1, pp. 70–86.

17. Krut'ko, P.D., *Upravlenie ispolnitel'nymi sistemami robotov* (Control of Robot Actuator Systems), Moscow: Nauka, 1991.

18. Glumov, V.M., Zemlyakov, S.D., Rutkovskii, V.Yu., and Sukhanov, V.M., Technical Controllability of the Free-flying Automated Space Module, *Autom. Remote Control*, 2001, vol. 62, no. 3, pp. 370–382.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 10, pp. 1221–1231.

Suppressing Exogenous Disturbances in a Discrete-Time Control System as an Optimization Problem

M. V. Khlebnikov

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

Received June 19, 2023
Revised July 24, 2023
Accepted August 2, 2023

Abstract—This paper proposes a novel approach to suppressing bounded exogenous disturbances in a linear discrete-time control system by a static state- or output-feedback control law. The approach is based on reducing the original problem to a nonconvex matrix optimization problem with the gain matrix as one variable. The latter problem is solved by the gradient method; its convergence is theoretically justified for several important special cases. An example is provided to demonstrate the effectiveness of the iterative procedure proposed.

Keywords: linear discrete-time system, exogenous disturbances, output feedback, state feedback, optimization, gradient method, Newton's method, convergence

DOI: 10.25728/arcRAS.2023.28.37.001

1. INTRODUCTION

Consider a linear discrete-time control system described by

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Dw_k, \\ y_k &= Cx_k, \\ z_k &= C_1 x_k, \end{aligned} \tag{1}$$

with the following notations: $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$, $D \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{l\times n}$, and $C_1 \in \mathbb{R}^{r\times n}$ are given matrices of compatible dimensions; x_0 is an initial state; $x_k \in \mathbb{R}^n$ is the state vector; $y_k \in \mathbb{R}^l$ is the observed output; $z_k \in \mathbb{R}^r$ is the controlled output; $u_k \in \mathbb{R}^p$ is the control vector; $w_k \in \mathbb{R}^m$ is an exogenous disturbance bounded at each time instant:

$$|w_k| \leqslant 1 \quad \text{for all } k = 0, 1, 2, \ldots. \tag{2}$$

The problem of suppressing bounded exogenous disturbances is to find a stabilizing feedback control law that minimizes the value $\max_k |z_k|$. In this paper, we will design a linear static state- $u_k = Kx_k$ or output-feedback $u_k = Ky_k$ control law (if it exists).

The exact solution of this problem seems difficult; following the approach proposed in [1–3], we will find a suboptimal solution in terms of invariant ellipsoids. In this case, the original problem is treated as an optimization problem, where one variable is the gain matrix and the objective function to be minimized determines the performance criterion (the size of the ellipsoid containing the controlled output of the system). The corresponding approach goes back to the works [4, 5], devoted to linear quadratic control design.

This paper is a natural continuation of the publication [6], where the problem of suppressing bounded exogenous disturbances in a linear continuous-time control system was considered and solved from the same perspective.

The remainder of this paper is organized as follows. Section 2 discusses an algorithm for solving the analysis problem (finding the minimal bounding ellipsoid for the closed loop system). In Section 3, the control design problem is written as a nonconvex matrix optimization problem, and an iterative algorithm for solving it is formulated and justified. Section 4 provides an illustrative example.

2. ANALYSIS PROBLEM

Consider a discrete-time dynamic system described by

$$\begin{aligned} x_{k+1} &= Ax_k + Dw_k, \\ z_k &= Cx_k \end{aligned} \tag{3}$$

with a stable (Schur) matrix $A \in \mathbb{R}^{n\times n}$, an initial state x_0, the state vector $x_k \in \mathbb{R}^n$, the output $z_k \in \mathbb{R}^l$, and an exogenous disturbance $w_k \in \mathbb{R}^m$ that satisfies the constraint (2).

Recall that an ellipsoid of the form

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

is said to be invariant for system (3) if the condition $x_0 \in \mathcal{E}_x$ implies $x_k \in \mathcal{E}_x$ for all time instants $k = 1, 2, \dots$. If $\mathcal{E}_x$ is an invariant ellipsoid with a matrix P, then the output z_k of system (3) with $x_0 \in \mathcal{E}_x$ belongs to the so-called bounding ellipsoid

$$\mathcal{E}_z = \left\{ z \in \mathbb{R}^r\colon \quad z^{\mathrm{T}} (CPC^{\mathrm{T}})^{-1} z \leqslant 1 \right\};$$

in the case $x_0 \notin \mathcal{E}_x$, the output will tend to this ellipsoid.

The analysis problem is to assess the effect of exogenous disturbances on the system output. Within the proposed approach, we are concerned with minimal ellipsoids containing the system output. A conventional minimality criterion for ellipsoids is the value $\operatorname{tr} CPC^{\mathrm{T}}$, equal to the sum of the squares of its semi-axes. The following result holds.

Theorem 1 [1, 3]. *Assume that the matrix A is Schur, $\rho = \max_i |\lambda_i(A)| < 1$, and the matrix $P(\alpha) \succ 0$, $\rho^2 < \alpha < 1$, satisfies the discrete Lyapunov equation*

$$\frac{1}{\alpha} APA^{\mathrm{T}} - P + \frac{1}{1-\alpha} DD^{\mathrm{T}} = 0.$$

Then the optimal bounding ellipsoid for system (3) is obtained by minimizing the univariate function

$$f(\alpha) = \operatorname{tr} CP(\alpha)C^{\mathrm{T}}$$

on the interval $\rho^2 < \alpha < 1$; and if α^* is the minimum point and x_0 satisfies the condition $x_0^{\mathrm{T}} P^{-1}(\alpha^*) x_0 \leqslant 1$, then the estimate

$$|z_k| \leqslant \sqrt{f(\alpha^*)}, \quad k = 1, 2, \dots,$$

holds.

The optimization problem formulated in Theorem 1 can be solved using Newton's method [7]. Let us choose an initial approximation $\rho^2(A) < \alpha_0 < 1$, e.g., $\alpha_0 = (1 + \rho^2(A))/2$, and apply the iterative process

$$\alpha_{j+1} = \alpha_j - \frac{f'(\alpha_j)}{f''(\alpha_j)}, \tag{4}$$

where

$$f'(\alpha) = \operatorname{tr} Y \left(\frac{1}{(1-\alpha)^2} DD^{\mathrm{T}} - \frac{1}{\alpha^2} APA^{\mathrm{T}} \right),$$

$$f''(\alpha) = 2\operatorname{tr} Y \left(\frac{1}{(1-\alpha)^3} DD^{\mathrm{T}} + \frac{1}{\alpha^3} A(P-X)A^{\mathrm{T}} \right),$$

and P, Y, and X are the solutions of the discrete Lyapunov equations

$$\frac{1}{\alpha} APA^{\mathrm{T}} - P + \frac{1}{1-\alpha} DD^{\mathrm{T}} = 0, \qquad \frac{1}{\alpha} A^{\mathrm{T}} YA - Y + C^{\mathrm{T}} C = 0,$$

and

$$\frac{1}{\alpha} AXA^{\mathrm{T}} - X + \frac{1}{(1-\alpha)^2} DD^{\mathrm{T}} - \frac{1}{\alpha^2} APA^{\mathrm{T}} = 0,$$

respectively.

The next theorem ensures the global convergence of this algorithm. It can be established by analogy with a similar result in [6].

Theorem 2. *In the method* (4),

$$|\alpha_j - \alpha^*| \leqslant \frac{f''(\alpha_0)}{2^j f''(\alpha^*)} |\alpha_0 - \alpha^*|, \qquad |\alpha_{j+1} - \alpha^*| \leqslant c|\alpha_j - \alpha^*|^2,$$

where $c > 0$ is some constant.

3. DESIGN PROBLEM

Returning to system (1), we suppose that the matrices D and C_1 are square and nonsingular.[1] The problem is to find a linear static output-feedback control law

$$u_k = Ky_k$$

(in the case $C = I$, a linear static state-feedback control law) that stabilizes the closed loop system (1) and suppresses the exogenous disturbances (2) by minimizing the bounding ellipsoid for the controlled output z_k. As an optimality criterion we choose the value

$$\operatorname{tr} C_1 P C_1^{\mathrm{T}} + \rho \|K\|_F^2,$$

where the first component describes the size of the bounding ellipsoid and the second one is a control penalty to avoid large values of the gain matrix. (The coefficient $\rho > 0$ adjusts its significance.)

Due to Theorem 1, the original problem is reduced to the matrix optimization problem

$$\min f(K,\alpha), \quad f(K,\alpha) = \operatorname{tr} C_1 P C_1^{\mathrm{T}} + \rho \|K\|_F^2$$

subject to the constraint

$$\frac{1}{\alpha}(A+BKC)P(A+BKC)^{\mathrm{T}} - P + \frac{1}{1-\alpha} DD^{\mathrm{T}} = 0 \tag{5}$$

with respect to the matrix variables $P = P^{\mathrm{T}} \in \mathbb{R}^{n\times n}$ and $K \in \mathbb{R}^{p\times n}$ and the scalar parameter $0 < \alpha < 1$.

According to Section 2, minimization with respect to the parameter α can be performed rather effectively. (It suffices to replace the matrix A by $A + BKC$.) Therefore, we will focus on minimizing the function

$$f(K) = \min_{\alpha} f(K,\alpha).$$

[1] No doubt, this technical assumption can be relaxed; for the time being, the objective is to establish simple and visual results.

Assumption. Let K_0 be a known stabilizing controller, i.e., the matrix $A + BK_0C$ is Schur.

Note that the function $f(K)$ is well-defined and positive on the set $\mathcal{S}$ of stabilizing controllers. Its definitional domain $\mathcal{S}$ can be nonconvex and disconnected whereas its boundaries can be nonsmooth. Here, the situation completely matches the continuous-time case; see [6].

Lemma 1. *The function $f(K)$ is coercive on the set $\mathcal{S}$ of stabilizing controllers (i.e., it tends to infinity on its boundary) and, moreover,*

$$f(K) \geqslant \frac{1}{1-\rho^2(A+BKC)} \frac{\lambda_{\min}(CC^{\mathrm{T}})}{1-\sigma^2_{\min}(A+BKC)} \|D\|_F^2, \tag{6}$$
$$f(K) \geqslant \rho\|K\|^2.$$

Corollary 1. *The level set*

$$\mathcal{S}_0 = \{K \in \mathcal{S}\colon \quad f(K) \leqslant f(K_0)\}$$

is bounded for any controller $K_0 \in \mathcal{S}$.

On the other hand, the function $f(K)$ has a minimum point on the set $\mathcal{S}_0$ (as a continuous function on a compact set), but the set $\mathcal{S}_0$ shares no points with the boundary of $\mathcal{S}$ due to (6). It will be demonstrated below that $f(K)$ is differentiable on $\mathcal{S}_0$; hence, the following result is valid.

Corollary 2. *There exists a minimum point K_* on the set S, and the gradient vanishes at this point.*

The gradient and Hessian of the function $f(K, \alpha)$ have properties described by the two lemmas below.

Lemma 2. *The function $f(K, \alpha)$ is well-defined and differentiable on the set $\mathcal{S}$ of stabilizing controllers K for $\rho^2(A + BKC) < \alpha < 1$. In addition,*

$$\frac{1}{2}\nabla_K f(K,\alpha) = \rho K + \frac{1}{\alpha} B^{\mathrm{T}} Y (A+BKC) P C^{\mathrm{T}}, \tag{7}$$
$$\nabla_\alpha f(K,\alpha) = \operatorname{tr} Y \left(\frac{1}{(1-\alpha)^2} DD^{\mathrm{T}} - \frac{1}{\alpha^2}(A+BKC)P(A+BKC)^{\mathrm{T}} \right),$$

where the matrices P and Y are the solutions of the discrete Lyapunov equations

$$\frac{1}{\alpha}(A+BKC)P(A+BKC)^{\mathrm{T}} - P + \frac{1}{1-\alpha}DD^{\mathrm{T}} = 0 \tag{8}$$

and

$$\frac{1}{\alpha}(A+BKC)^{\mathrm{T}} Y (A+BKC) - Y + C_1^{\mathrm{T}} C_1 = 0, \tag{9}$$

respectively.

The function $f(K, \alpha)$ achieves minimum at an inner point of the set $\mathcal{S} \times (\rho^2(A + BKC), 1)$. This point is given by the conditions

$$\nabla_K f(K,\alpha) = \nabla_\alpha f(K,\alpha) = 0.$$

In addition, $f(K, \alpha)$ as a function of α is strictly convex on $\rho^2(A + BKC) < \alpha < 1$ and achieves minimum at an inner point of this interval.

Lemma 3. *The function $f(K,\alpha)$ is twice differentiable with respect to K, and the action of its Hessian on an arbitrary matrix*[2] $E \in \mathbb{R}^{p\times l}$ *is given by*

$$\frac{1}{2}\nabla_K^2 f(K,\alpha)[E,E] = \rho\langle E,E\rangle + \frac{1}{\alpha}\langle B^{\mathrm{T}}YBECPC^{\mathrm{T}},E\rangle + \frac{2}{\alpha}\langle B^{\mathrm{T}}Y(A+BKC)P'C^{\mathrm{T}},E\rangle,$$

where P' is the solution of the discrete Lyapunov equation

$$\frac{1}{\alpha}(A+BKC)P'(A+BKC)^{\mathrm{T}} - P' + \frac{1}{\alpha}\Big((A+BKC)P(BEC)^{\mathrm{T}} + BECP(A+BKC)^{\mathrm{T}}\Big) = 0. \quad (10)$$

The gradient of $f(K,\alpha)$ as a function of K is not Lipschitz on the set $\mathcal{S}$ of stabilizing controllers. However, like in the continuous-time case, it has the Lipschitz property on the subset $\mathcal{S}_0$. (This result can be easily obtained.)

The above properties of the objective function allow constructing a minimization method and justifying its convergence. That is, we propose an iterative approach to solve the problem that involves the gradient method with respect to the variable K and Newton's method with respect to the variable α.

The algorithm includes several steps as follows.

1. Choose some values of the parameters $\varepsilon > 0$, $\gamma > 0$, $0 < \tau < 1$, and the initial stabilizing approximation K_0. Calculate
$$\alpha_0 = \frac{1+\rho^2(A+BK_0C)}{2}.$$
2. On the jth iteration, the controller K_j and the value α_j are given. Calculate the matrix $A_j = A + BK_jC$, solve equations (8) and (9) to find the matrices P and Y. Calculate the gradient
$$H_j = \nabla_K f(K_j,\alpha_j)$$
from the relation (7).
If $\|H_j\| \leqslant \varepsilon$, then take the controller K_j as the approximate solution.
3. Perform the gradient method step:
$$K_{j+1} = K_j - \gamma_j H_j.$$
Adjust the step length $\gamma_j > 0$ by fractionating γ until the following conditions are satisfied:
a. K_{j+1} is a stabilizing controller, i.e., the matrix $(A+BK_{j+1}C)/\sqrt{\alpha_j}$ is Schur.
b. $f(K_{j+1}) \leqslant f(K_j) - \tau\gamma_j\|H_j\|^2$.
4. Minimize $f(K_{j+1},\alpha)$ with respect to α and find α_{j+1}. Revert to Step 2.

This algorithm converges in the following sense.

Theorem 3. *Only a finite number of fractions are realized for γ_j at each iteration of the algorithm, the function $f(K_j)$ is monotonically decreasing, and its gradient vanishes with an exponential rate (like a geometric progression):*

$$\lim_{j\to\infty}\|H_j\| = 0.$$

The proof is completely analogous to the continuous-time case and uses the common gradient method analysis scheme for the unconstrained minimization of functions with a Lipschitz gradient [8].

[2] In the sense of the second derivative in a direction.

4. EXAMPLE

Consider a system of the form (1) with the matrices

$$A = \begin{pmatrix} 0.9950 & 0.0050 & 0.0998 & 0.0002 \\ 0.0050 & 0.9950 & 0.0002 & 0.0998 \\ -0.0997 & 0.0997 & 0.9950 & 0.0050 \\ 0.0997 & -0.0997 & 0.0050 & 0.9950 \end{pmatrix},$$

$$B = \begin{pmatrix} 0.0050 \\ 0.0000 \\ 0.0998 \\ 0.0002 \end{pmatrix}, \qquad D = \begin{pmatrix} 0.0050 & 0.0000 \\ 0.0000 & 0.0050 \\ 0.0998 & 0.0002 \\ 0.0002 & 0.0998 \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \qquad C_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

This is a slight modification of Example 4.3.2 from the monograph [3].

Let $\rho = 0.1$ and choose

$$K_0 = \begin{pmatrix} -2.9823 \\ -3.9608 \end{pmatrix}$$

as an initial stabilizing controller.

The iterative process terminated in the 25th iteration and yielded the controller

$$K_* = \begin{pmatrix} -0.6519 \\ -1.8166 \end{pmatrix}$$

and the corresponding bounding ellipse for the controlled output of the system with the matrix

$$\begin{pmatrix} 19.2309 & -3.4643 \\ -3.4643 & 10.3506 \end{pmatrix}.$$

The dynamics of the iterative process are shown in Fig. 1.

For the initial stabilizing controller

$$K_0' = \begin{pmatrix} -0.3675 \\ -0.7106 \end{pmatrix},$$

in the 24th iteration we obtain the controller

$$K_*' = \begin{pmatrix} -0.6527 \\ -1.8166 \end{pmatrix}$$

and the corresponding bounding ellipse with the matrix

$$\begin{pmatrix} 19.2293 & -3.4638 \\ -3.4638 & 10.3543 \end{pmatrix}.$$

Note that the controllers K_* and K_*' differ in norm by fractions of a percent. The same applies to the bounding ellipses when contrasted by the trace criterion.

For comparison, we solve the same problem by constructing a dynamic feedback controller

$$u_k = K\widehat{x}_k$$

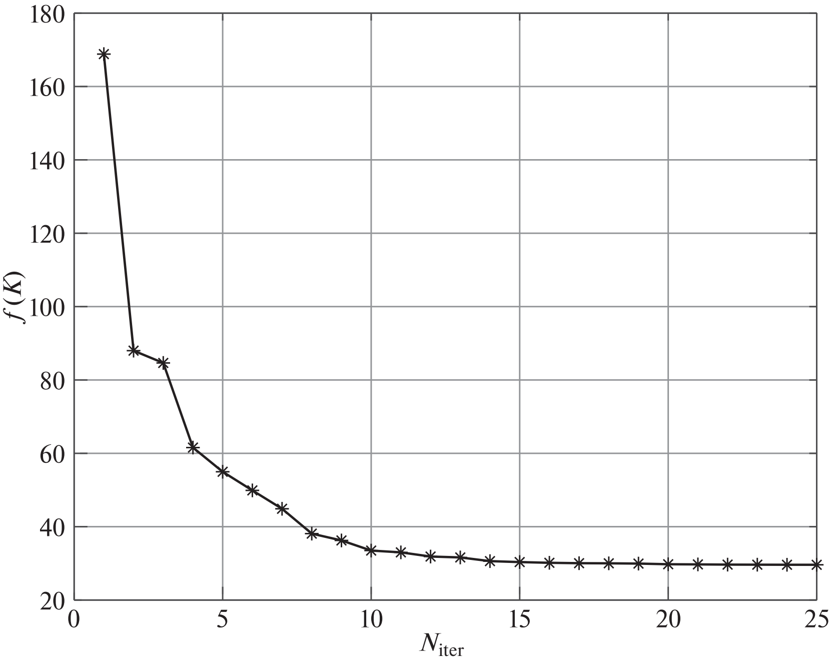


Fig. 1. Optimization procedure.

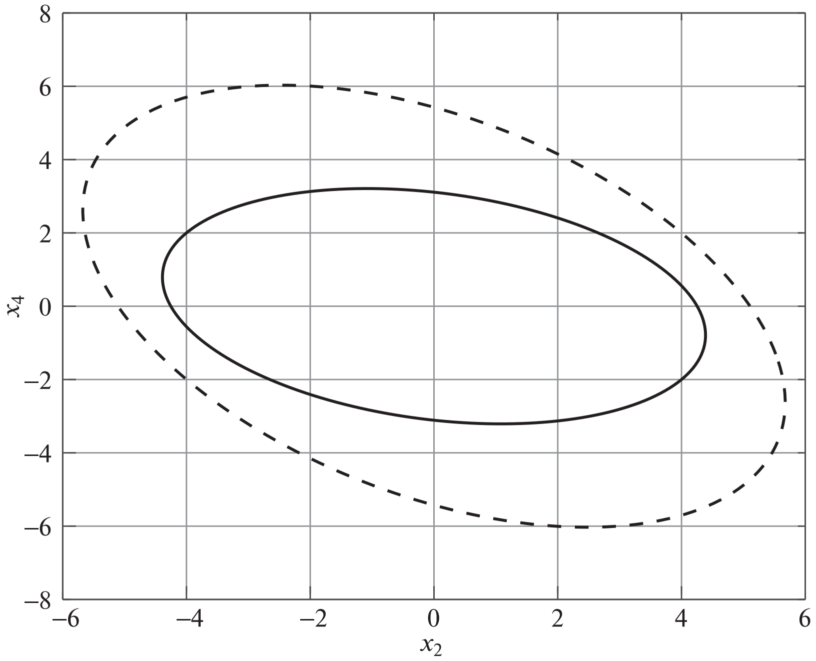


Fig. 2. Bounding ellipses.

using the observer

$$\widehat{x}_{k+1} = A\widehat{x}_k + Bu_k + L(y_k - C\widehat{x}_k), \quad \widehat{x}_0 = 0.$$

Following the approach [3] and the technique of linear matrix inequalities (LMIs), we calculate the gain matrix

$$K = \begin{pmatrix} -39.0055 & -46.7193 & -8.5074 & -98.0176 \end{pmatrix},$$

the observer matrix

$$L = \begin{pmatrix} 0.5655 & 0.0759 \\ -6.7183 & 1.8722 \\ -2.2061 & 1.0573 \\ -2.8715 & 0.7224 \end{pmatrix},$$

and the matrix

$$\begin{pmatrix} 32.2165 & -14.9238 \\ -14.9238 & 36.3654 \end{pmatrix}$$

of the ellipse containing the controlled output.

These problem statements have a small technical difference: in the latter case, the regularizing term $\rho\|K\|_F^2$ is eliminated from the objective function and an additional term with control is introduced into the regulated output of the system for the same purpose: $z_k = C_1 x_k + B_1 u_k$.

In Fig. 2, the solid line shows the bounding ellipse yielded by the iterative procedure whereas the dotted line the one provided by the dynamic controller. The rather large difference in the sizes of the ellipses can be explained as follows: when constructing a dynamic feedback controller, it is necessary to roughen several things in order to linearize the matrix inequalities, which leads to excessive conservatism.

5. CONCLUSIONS

This paper has proposed a new controller design approach for the optimal suppression of bounded exogenous disturbances in a linear discrete-time system. It is based on reducing the original problem to a matrix optimization problem with the gain matrix as one variable. Next, this problem is solved using the gradient method. Its convergence has been theoretically justified for several important special cases. A numerical example has been presented to demonstrate the effectiveness of the proposed procedure.

The problem of suppressing exogenous disturbances has been considered under fairly strict restrictions. In particular, it has been assumed that the dimension of disturbances and controlled outputs coincides with the number of states. However, the method quite effectively works in the absence of such restrictions. An important task is to justify the method in this case as well.

Since the definitional domain of the function $f(K)$ may even be disconnected, it is difficult to expect convergence to a global minimum. However, for the problem with state-feedback control, as in the continuous case, one can apparently expect that the objective function satisfies the gradient dominance condition and, hence, global convergence to a unique minimum point.

APPENDIX

Proof of Lemma 1. Consider a sequence of stabilizing controllers $\{K_j\} \in \mathcal{S}$ such that $K_j \to K \in \partial\mathcal{S}$, i.e., $\rho(A + BKC) = 1$. In other words, for any $\varepsilon > 0$ there exists a number $N = N(\varepsilon)$ such that

$$|\rho(A + BK_jC) - \rho(A + BKC)| = 1 - \rho(A - BK_jC) < \varepsilon$$

for all $j \geqslant N(\varepsilon)$.

Let P_j be the solution of equation (5) associated with the controller K_j:

$$\frac{1}{\alpha_j}(A + BK_jC)P_j(A + BK_jC)^{\mathrm{T}} - P_j + \frac{1}{1-\alpha_j}DD^{\mathrm{T}} = 0.$$

Also, let Y_j be the solution of the dual discrete Lyapunov equation

$$\frac{1}{\alpha_j}(A+BK_jC)^{\mathrm T}Y_j(A+BK_jC)-Y_j+C_1C_1^{\mathrm T}=0.$$

Using [6, Lemmas A.1 and A.2] and [7, Lemma A.1.2], we have

$$f(L_j)=\operatorname{tr}C_1P_jC_1^{\mathrm T}+\rho\|K_j\|_F^2\geqslant\operatorname{tr}P_jC_1C_1^{\mathrm T}=\operatorname{tr}\left(Y_j\frac{1}{1-\alpha_j}DD^{\mathrm T}\right)$$

$$\geqslant\frac{1}{1-\alpha_j}\lambda_{\min}(Y_j)\|D\|_F^2\geqslant\frac{1}{1-\alpha_j}\frac{\lambda_{\min}(C_1C_1^{\mathrm T})}{1-\sigma_{\min}^2(A+BK_jC)}\|D\|_F^2$$

$$\geqslant\frac{1}{1-\rho^2(A+BK_jC)}\frac{\lambda_{\min}(C_1C_1^{\mathrm T})}{1-\sigma_{\min}^2(A+BK_jC)}\|D\|_F^2$$

$$\geqslant\frac{1}{\varepsilon}\frac{1}{1+\rho(A+BK_jC)}\frac{\lambda_{\min}(C_1C_1^{\mathrm T})}{1-\sigma_{\min}^2(A+BK_jC)}\|D\|_F^2\xrightarrow[\varepsilon\to0]{}+\infty$$

since $\rho^2(A+BK_jC)<\alpha_j<1$.

On the other hand,

$$f(K_j)=\operatorname{tr}C_1P_jC_1^{\mathrm T}+\rho\|K_j\|_F^2\geqslant\rho\|K_j\|_F^2\geqslant\rho\|K_j\|^2\xrightarrow[\|K_j\|\to+\infty]{}+\infty.$$

The proof of Lemma 1 is complete.

Proof of Lemma 2. Differentiation with respect to α is performed in accordance with the results of Section 2, with A replaced by $A+BKC$.

We add the increment ΔK for K in equation (5) and denote the corresponding increment of P by ΔP:

$$\frac{1}{\alpha}(A+B(K+\Delta K)C)(P+\Delta P)(A+B(K+\Delta K)C)^{\mathrm T}-(P+\Delta P)+\frac{1}{1-\alpha}DD^{\mathrm T}=0.$$

Leaving the notation ΔP for the principal part of the increment, we have

$$\begin{aligned}\frac{1}{\alpha}\Big((A+BKC)P(A+BKC)^{\mathrm T}+B\Delta KCP(A+BKC)^{\mathrm T}\\+(A+BKC)P(B\Delta KC)^{\mathrm T}+(A+BKC)\Delta P(A+BKC)^{\mathrm T}\Big)\\-(P+\Delta P)+\frac{1}{1-\alpha}DD^{\mathrm T}=0.\end{aligned}$$

Subtracting equation (5) from this equation gives

$$\begin{aligned}\frac{1}{\alpha}(A+BKC)\Delta P(A+BKC)^{\mathrm T}-\Delta P\\+\frac{1}{\alpha}\Big((A+BKC)P(B\Delta KC)^{\mathrm T}+B\Delta KCP(A+BKC)^{\mathrm T}\Big)=0.\end{aligned}\quad\text{(A.1)}$$

The increment of $f(K)$ is calculated by linearizing the corresponding terms:

$$\begin{aligned}\Delta f(K)&=f(K)-f(K+\Delta K)\\&=\operatorname{tr}C_1(P+\Delta P)C_1^{\mathrm T}+\rho\|K+\Delta K\|_F^2-(\operatorname{tr}C_1PC_1^{\mathrm T}+\rho\|K\|_F^2)\\&=\operatorname{tr}C_1\Delta PC_1^{\mathrm T}+\rho\operatorname{tr}K^{\mathrm T}\Delta K+\rho\operatorname{tr}(\Delta K)^{\mathrm T}K=\operatorname{tr}\Delta PC_1^{\mathrm T}C_1+2\rho\operatorname{tr}K^{\mathrm T}\Delta K.\end{aligned}$$

Due to [6, Lemma A.1], from the dual equations (A.1) and (9) it follows that

$$\begin{aligned}\Delta f(K) &= 2\mathrm{tr}\, Y\frac{1}{\alpha}B\Delta KCP(A+BKC)^{\mathrm{T}} + 2\rho\mathrm{tr}\, K^{\mathrm{T}}\Delta K\\ &= 2\mathrm{tr}\,\left(\rho K^{\mathrm{T}} + \frac{1}{\alpha}CP(A+BKC)^{\mathrm{T}}YB\right)\Delta K\\ &= 2\left\langle \rho K + \frac{1}{\alpha}B^{\mathrm{T}}Y(A+BKC)PC^{\mathrm{T}}, \Delta K\right\rangle.\end{aligned}$$

Thus, we arrive at (7). The proof of Lemma 2 is complete.

Proof of Lemma 3. The value

$$\nabla^2_K f(K,\alpha)[E,E] = \langle \nabla^2_K f(K,\alpha)[E], E\rangle,$$

is calculated by differentiating $\nabla_K f(K,\alpha)[E] = \langle \nabla_K f(K,\alpha), E\rangle$ in the direction $E \in \mathbb{R}^{p\times l}$.

For this purpose, linearizing the corresponding terms, we calculate the increment of $\nabla_K f(K,\alpha)[E]$ in the direction E:

$$\begin{aligned}&\Delta\nabla_K f(K,\alpha)[E]\\ &= 2\left(\rho(K+\delta E) + \frac{1}{\alpha}B^{\mathrm{T}}(Y+\Delta Y)(A+B(K+\delta E)C)(P+\Delta P)C^{\mathrm{T}}\right)\\ &\quad -2\left(\rho K + \frac{1}{\alpha}B^{\mathrm{T}}Y(A+BKC)PC^{\mathrm{T}}\right)\\ &= 2\delta\Big(\rho E + \frac{1}{\alpha}B^{\mathrm{T}}(YBECP + Y'(K)[E](A+BKC)P\\ &\qquad + Y(A+BKC)P'(K)[E])C^{\mathrm{T}}\Big),\end{aligned}$$

where

$$\begin{aligned}\Delta P &= P(K+\delta E) - P(K) = \delta P'(K)[E],\\ \Delta Y &= Y(K+\delta E) - Y(K) = \delta Y'(K)[E].\end{aligned}$$

Thus, with $P' = P'(K)[E]$ and $Y' = Y'(K)[E]$, we have

$$\begin{aligned}&\frac{1}{2}\nabla^2_K f(K,\alpha)[E,E]\\ &= \left\langle \rho E + \frac{1}{\alpha}B^{\mathrm{T}}(YBECP + Y'(A+BKC)P + Y(A+BKC)P')C^{\mathrm{T}}, E\right\rangle.\end{aligned}$$

Furthermore, $P = P(K)$ is the solution of the discrete Lyapunov equation (5). We write it in increments in the direction E:

$$\frac{1}{\alpha}(A+B(K+\delta E)C)(P+\delta P')(A+B(K+\delta E)C)^{\mathrm{T}} - (P+\delta P') + \frac{1}{1-\alpha}DD^{\mathrm{T}} = 0$$

or

$$\begin{aligned}&\frac{1}{\alpha}\Big((A+BKC)P(A+BKC)^{\mathrm{T}} + (A+BKC)\delta P'(A+BKC)^{\mathrm{T}}\\ &\qquad + (A+BKC)P(B\delta EC)^{\mathrm{T}} + B\delta ECP(A+BKC)^{\mathrm{T}}\Big)\\ &\qquad - (P+\delta P') + \frac{1}{1-\alpha}DD^{\mathrm{T}} = 0.\end{aligned}$$

In view of (5), this expression yields equation (10).

Similarly, $Y = Y(K)$ is the solution of the discrete Lyapunov equation (9). We write it in increments in the direction E:

$$\frac{1}{\alpha}(A + B(K + \delta E)C)^{\mathrm{T}}(Y + \delta Y')(A + B(K + \delta E)C) - (Y + \delta Y') + C_1^{\mathrm{T}} C_1 = 0$$

or

$$\begin{gathered}\frac{1}{\alpha}\Big((A + BKC)^{\mathrm{T}} Y(A + BKC) + (A + BKC)^{\mathrm{T}} \delta Y'(A + BKC) \\ + (A + BKC)^{\mathrm{T}} Y B\delta EC + (B\delta EC)^{\mathrm{T}} Y(A + BKC)\Big) - (Y + \delta Y') + C_1^{\mathrm{T}} C_1 = 0.\end{gathered}$$

Due to (9), we obtain

$$\begin{gathered}\frac{1}{\alpha}(A + BKC)^{\mathrm{T}} Y'(A + BKC) - Y' \\ + \frac{1}{\alpha}\Big((A + BKC)^{\mathrm{T}} YBEC + (BEC)^{\mathrm{T}} Y(A + BKC)\Big) = 0.\end{gathered} \tag{A.2}$$

From (10) and (A.2) it follows that

$$\operatorname{tr} P'(A + BKC)^{\mathrm{T}} YBEC = \operatorname{tr} Y' BECP(A + BKC)^{\mathrm{T}},$$

so

$$\frac{1}{2}\nabla_K^2 f(K, \alpha)[E, E] = \rho\langle E, E\rangle + \frac{1}{\alpha}\langle B^{\mathrm{T}} YBECPC^{\mathrm{T}}, E\rangle + \frac{2}{\alpha}\langle B^{\mathrm{T}} Y(A + BKC)P'C^{\mathrm{T}}, E\rangle.$$

The proof of Lemma 3 is complete.

REFERENCES

1. Boyd, S., El Ghaoui, L., Feron, E., and Balakrishnan, V., *Linear Matrix Inequalities in System and Control Theory*, Philadelphia: SIAM, 1994.
2. Nazin, S.A., Polyak, B.T., and Topunov, M.V., Rejection of Bounded Exogenous Disturbances by the Method of Invariant Ellipsoids, *Autom. Remote Control*, 2007, vol. 68, no. 3, pp. 467–486.
3. Polyak, B.T., Khlebnikov, M.V., and Shcherbakov, P.S., *Upravlenie lineinymi sistemami pri vneshnikh vozmushcheniyakh: Tekhnika lineinykh matrichnykh neravenstv* (Control of Linear Systems Subjected to Exogenous Disturbances: The Technique of Linear Matrix Inequalities), Moscow: LENAND, 2014.
4. Kalman, R.E., Contributions to the Theory of Optimal Control, *Boletin de la Sociedad Matematica Mexicana*, 1960, vol. 5, no. 1, pp. 102–119.
5. Levine, W. and Athans, M., On the Determination of the Optimal Constant Output Feedback Gains for Linear Multivariable Systems, *IEEE Trans. Automat. Control*, 1970, vol. 15, no. 1, pp. 44–48.
6. Polyak, B.T. and Khlebnikov, M.V., Static Controller Synthesis for Peak-to-Peak Gain Minimization as an Optimization Problem, *Autom. Remote Control*, 2021, vol. 82, no. 9, pp. 1530–1553.
7. Khlebnikov, M.V., A Comparison of Guaranteeing and Kalman Filters, *Autom. Remote Control*, 2023, vol. 84, no. 4, pp. 434–459.
8. Polyak, B., *Introduction to Optimization*, Optimization Software, 1987.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 10, pp. 1232–1242.
© The Author(s), 2023 published by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 2023.
Russian Text © The Author(s), 2023, published in Avtomatika i Telemekhanika, 2023, No. 10, pp. 118–131.

Information Support for Aircraft Crew in Takeoff and Landing Modes

A. M. Shevchenko^{*,a}**, B. V. Pavlov**^{*,b}**, and G. N. Nachinkina**^{*,a}

* *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
e-mail: ^{a}*shev-chik@yandex.ru,* ^{b}*pavlov@ipu.ru,* ^{c}*nach_gala@ipu.ru*

Received June 19, 2023
Revised July 28, 2023
Accepted August 2, 2023

Abstract—Methods for assessing the current state and forecasting critical events are developed in order to reduce the stress load on the pilot of an aircraft. These methods are based on the energy approach to flight control. Algorithms for forecasting the possibility of safe takeoff in the presence of high-rise obstacles on the trajectory are obtained. Forecast correction algorithms are introduced. Algorithms for calculating the braking distance depending on the runway condition are found in the modes of landing or emergency braking at takeoff. Some ways to correct forecasts considering the sequence and operation time of all braking devices are proposed. Model tests are carried out for the algorithms in the entire range of operating conditions.

Keywords: takeoff, landing, forecasting methods, information support, energy approach

DOI: 10.25728/arcRAS.2023.18.69.001

1. INTRODUCTION

The issues of organizing passenger transportation have recently become more and more acute and topical. The main directions of transportation improvement are traffic intensification and the expansion of permitted weather conditions for aircraft flights. Therefore, the safety of aviation equipment comes to the forefront.

Technical and communications progress in all spheres of human activity tends to accelerate. This progress is manifested by the increase in transportation traffic and the expansion of acceptable atmospheric or climatic conditions.

According to statistical data of the Main Center for Information Technologies and Meteorological Services for Aviation (Aviamettelecom) of the Federal Service for Hydrometeorology and Environmental Monitoring (Roshydromet) [1], there were nine aviation accidents during January–March 2023, including:

(1) three fatal accidents, particularly three fatal accidents in G-class airspace, with a death toll of 5;

(2) one non-fatal accident, including one non-fatal accident in G-class airspace;

(3) two aviation incidents in total (one aircraft landing below the operational minimum and one aircraft struck by atmospheric electricity);

(4) two industrial events (emergency events);

(5) one emergency situation without investigation (one aircraft struck by atmospheric electricity).

In a statistical study of aviation accidents on passenger flights throughout the world, Boeing demonstrated that more than half of all accidents occur during takeoff and landing stages [2]. Flight control at these stages is carried out with the direct participation of the pilot, who undergoes strong psychological stress. Therefore, the human factor becomes governing. The statistics of aviation accidents based on the recent studies [3–5] shows that the share of aviation accidents caused by human participation in flight task fulfillment varies from 50 to 70% depending on the estimation methods.

One safety improvement direction is to equip aircraft with onboard systems providing instrument-based control of critical motion coordinates during two stages: ground run on the runway during landing and takeoff. Information support for the pilot and the creation of a pilot-friendly interaction environment with cockpit equipment have become necessary. For this purpose, forecasting methods and new algorithms [6–8] were developed to calculate the aircraft motion on the ground segment of the trajectory.

In particular, it was decided to supplement onboard equipment with an information measuring system (IMS) of takeoff run control [6]. This system simultaneously monitors longitudinal acceleration, speed, and distance to reach the target speed. The forecasted distance to the decision point helps the pilot to make a timely decision. But if the forecasted distance differs from the standard one by an unacceptable value, the IMS generates a signal to alert the pilot and a command signal to prohibit takeoff. In [7, 8], some variants of safe forecasted takeoff and emergency braking in unfavorable climatic conditions and geographical coordinates were developed. These solutions are conceptually based on the energy approach to controlling the spatial motion of an aircraft, first presented in [9].

This paper further refines methods for assessing the current situation and forecasting the aircraft's motion on the runway in the braking modes after landing or aborted takeoff and in the ground run stage before takeoff. In addition, we develop methods for increasing situational awareness to eliminate stress load and reduce the risks of erroneous actions of the pilot.

2. THE ENERGY BALANCE EQUATION FOR AIRCRAFT MOTION

Historically, the basic controllable coordinates of an aircraft are altitude, speed, and the direction of flight. They are natural for flight control both in visual orientation mode and in instrument flight. The theory and practice of automatic control were developed in the same line. The concept of flight control in the longitudinal channel of the aircraft using two loops—trajectory and speed—became established in aviation. In automatic flight control systems, the functions of controllers are performed by independent devices, namely, thrust automaton and autopilot. Controller design problems with classical methods neglect the nonlinear relationship between the two main variables (speed and flight altitude), which is provided by the fundamental law of conservation of energy of a body moving in a potential field of forces.

In contrast to the conventional description of the spatial motion of an aircraft by the Cauchy equations, the paper [10] proposed a control concept with the total energy of motion

$$E = mgh + \frac{mV^2}{2},$$

where m denotes the weight of the aircraft, h is the flight altitude, and V is the speed in the inertial frame.

We will consider motion in terms of the weight-normalized specific energy of motion H_E, which is also called the pseudo-energy or energy height:

$$H_E = \frac{E}{mg} = h + \frac{V^2}{2g}.$$

Being jointly solved, the dynamic equations of translational motion in the disturbed atmosphere and the total energy equation of an object yield the **energy balance equation**

$$\Delta H_E = \Delta H_E^{\mathrm{eng}} + \Delta H_E^D + \Delta H_E^{gear} + \Delta H_E^w.$$

This equation describes quantitative relations between the energy source and all its consumers in the "aircraft–engine–environment" system. The equation is written in increments and contains the following terms: ΔH_E is the increment of the energy height of the aircraft; $\Delta H_E^{\mathrm{eng}}$ is the specific work of the engine; ΔH_E^D is energy costs to overcome the aerodynamic drag; ΔH_E^{gear} is energy costs to overcome the resistance of landing gear; finally, ΔH_E^w is wind work. For each term, the following expressions were derived in [8, 9]: $\Delta H_E = \int_{t_1}^{t_2} V_B(\theta + \frac{\dot{V}_B}{g})dt$, where V_B is airspeed and θ is the angle of inclination of the trajectory in the inertial frame; $\Delta H_E^{eng} = \int_{t_1}^{t_2} V_B P_H \cos(\alpha_B + \phi_{eng})dt$, where $P_H = \frac{P}{mg}$ is the normalized thrust, α_B is the angle of attack, and ϕ_{eng} is the angle of engine installation; $\Delta H_E^D = \int_{t_1}^{t_2} V_B D_H dt$, where $D_H = \frac{D}{mg}$ is the normalized drag; $\Delta H_E^w = \int_{t_1}^{t_2} V_B f_w dt$, where the factor $f_w \approx \frac{\dot{w}_x}{g} - \frac{\dot{w}_y}{V_B}$ is called the wind factor or hazard index, and w_x and w_y are the projections of wind speed on the inertial frame axes; finally, $\Delta H_E^{gear} = \int_{t_1}^{t_2} V k_{brak} f_w dt$, where k_{brak} is the generalized normalized braking coefficient (the total resistive force of landing gear divided by aircraft weight).

3. BASIC ALGORITHMS OF ENERGY CONTROL SYSTEM

The energy height H_E has two components characterizing potential and kinetic energies, respectively. When moving in space, each component changes not independently but in according with the law of conservation of total energy. Therefore, the problem of designing flight control algorithms is naturally posed as a problem of multicriteria control. The first criterion is to minimize the deviation of the energy height: $\Delta H \to \min$. The second criterion is to minimize the mismatch between its kinetic and potential components:

$$\Delta H_E^{kin} - \Delta H_E^{pot} \to \min .$$

In the energy control system (EnCS), the thrust P is the only control variable affecting the total energy of the aircraft; the elevating rudder deviation δ_B causes a redistribution of the potential and kinetic components.

The forces in projections on the axes of the air frame satisfy the equation

$$m\dot{V}_B = P\cos(\alpha_B + \phi_{eng}) - D - mg\sin\theta_B - m(\dot{W}_{xg}\cos\theta_B + \dot{W}_{yg}\sin\theta_B),$$

where V_B is airspeed, α_B is the angle of attack in the air frame, D is drag, θ_B is the trajectory's angle of inclination in the air frame, and W_{xg} and W_{yg} are the projections of wind speed on the axes of the Earth frame. Resolving this equation for P under the assumption of small angles and passing to the normalized variables, we obtain

$$P_H = \theta + \frac{\dot{V}_B}{g} + f_w + D_H.$$

In the steady-state flight mode without wind, the simplified thrust control law in the EnCS in increments relative to the set values is given by

$$\Delta P_H^{EnCS} = \Delta\theta + \frac{\Delta\dot{V}_B}{g}.$$

Elevating rudder control is used to minimize the mismatch between the potential and kinetic components, which does not affect the first criterion:

$$\Delta\delta_H^{EnCS} = \Delta\theta - \frac{\Delta\dot{V}_B}{g}.$$

Integral terms are added to the proportional ones to ensure astatism for the controlled coordinates.

Flight control with EnCS naturally considers the mutual influence of the speed and trajectory channels; thus, correction loops for these relationships are not needed.

4. ENERGY FORECASTING METHOD FOR TAKEOFF AND OBSTACLE CLEARANCE

The pilot's goal in the takeoff stage is to overcome a high-rise obstacle at a speed at least equal to that of stable horizontal flight. In complicated conditions, the pilot needs to assess a priori the aircraft's ability to accelerate to the takeoff speed within the runway and climb sufficiently to overcome high-rise obstacles on the takeoff course. The takeoff diagram is shown in Fig. 1.

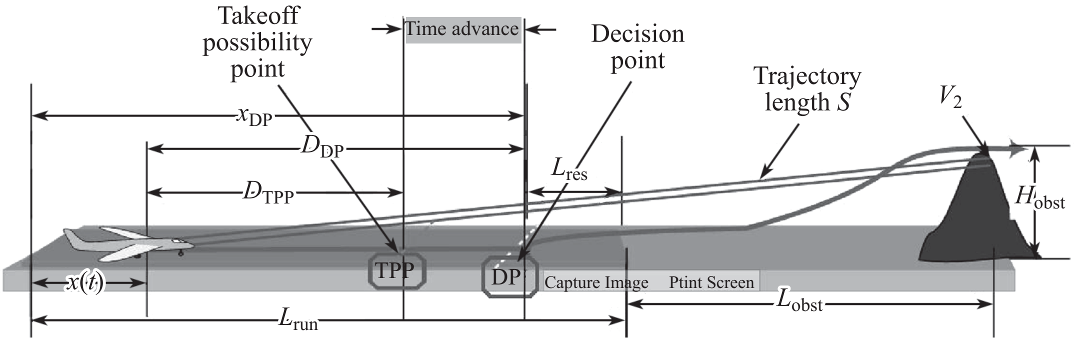


Fig. 1. Characteristic points on the takeoff trajectory.

This figure has the following notations: $x(t)$ is the current coordinate of the aircraft; H_{obst} and L_{obst} are the obstacle height and the distance to the obstacle from the runway endpoint, respectively; V_2 is the minimum speed of stable horizontal flight; S is the energy accumulation distance; L_{run} is the runway length; D_{TPP} is the distance to the takeoff possibility point (TPP); D_{DP} is the distance to the decision point (DP); x_{DP} is the coordinate of the decision point; finally, L_{res} is the takeoff run reserve from the DP to the runway endpoint.

According to the Flight Manual, takeoff is authorized when sequentially reaching the minimum horizontal flight speed V_1 and the nosewheel lift-off speed V_r regardless of the aircraft's position on the runway. However, this takeoff procedure does not ensure overcoming an obstacle safely since the speed V_r may be reached at a point in unacceptable proximity to the runway endpoint or even beyond it.

Let us inform the pilot about the possibility of a safe takeoff ahead of time by forecasting the energy state of the aircraft corresponding to the required generalized coordinates at the obstacle clearance point.

To overcome the obstacle safely, the aircraft must have a speed not less than its stable horizontal flight speed V_2. At the instant of overcoming the obstacle, the total energy $E_{H_{obst}}$ of the aircraft must contain the required minimum kinetic component and a reserve of the potential component, which gives the aircraft the necessary altitude H_{obst} for obstacle clearance:

$$E_{H_{obst}} = m\frac{V_2^2}{2} + mgH_{obst}. \tag{1}$$

The total accumulated energy of the aircraft consists of the current kinetic and potential components and the work of all external forces F_i on the trajectory of length S. Then the forecasted accumulated energy is given by

$$E(t)_{fore} = m\frac{V(t)^2}{2} + mgh(t) + S\sum_i F_i(t), \tag{2}$$

where $\sum_i F_i(t)$ is the resultant of all external forces: engine thrust, aerodynamic drag, wind force, and landing gear braking force. Equation (2) explicitly relates the energy state of the controlled object and the trajectory length to reach this state.

The resultant is naturally calculated through the longitudinal overload:

$$\sum_i F_i(t) = mgn_x(t). \tag{3}$$

Let all forces in (3) be measured during the ground run before takeoff. Equating the required (1) and forecasted (2) energies, we find the length of the forward section of the ground segment to the DP necessary to accumulate the deficient total energy:

$$D_{DP} = \frac{(g(H_{obst} - h(t)) + 0.5(V_2^2 - V(t)^2))}{gn_x(t)} - L_{obst}.$$

Note that this expression is invariant with respect to weight. The trajectory point where the forecasted length of this section becomes zero is the DP of safe takeoff: $X_{DP} = x(t)|_{D=0}$. The coordinate of this point is simply calculated as

$$X_{DP}(t) = x(t) + D_{DP}(t).$$

Total energy forecasting indicates the possibility of takeoff not at the instant of reaching the decision speed but earlier and in the distance coordinates associated with the runway.

The forecasting method based on the energy approach yields a forecast of another characteristic point on the takeoff run trajectory. Each type of aircraft is allowed to lift the front landing gear strut when reaching a known minimum takeoff run speed V_r. In abnormal situations, the pilot must assess the possibility of continuing the takeoff run and, moreover, the position of the aircraft on the runway in which it is possible to start lifting the front strut. The distance from the current position of the aircraft to reaching the rate of climb is calculated as

$$D_{V_r}(t) = \frac{V_r^2 - V^2(t)}{2gn(t)}.$$

When this forecasted distance reaches zero, it is possible to lift the front landing gear strut to turn the airplane to the takeoff angle of attack. In the course of the takeoff run, it is proposed to inform the pilot about the distance to the front strut lift point. The instrumental estimate of this distance, unlike the intuitive one, improves the pilot's situational awareness and reduces the prerequisites for erroneous actions. The distance to the front strut lift point can be shown on the instrument panel or on the display.

Situational awareness can be increased (and stress load can be reduced) when using the forecasted distance reserve to the runway endpoint at the DP:

$$L_{res}(t) = L_{run} - x(t) - D_{DP}(t).$$

A very fruitful feature of the energy method is that the current forecast considers the total energy acquired by the aircraft on the forward air segment outside the ground one. As a result, it is

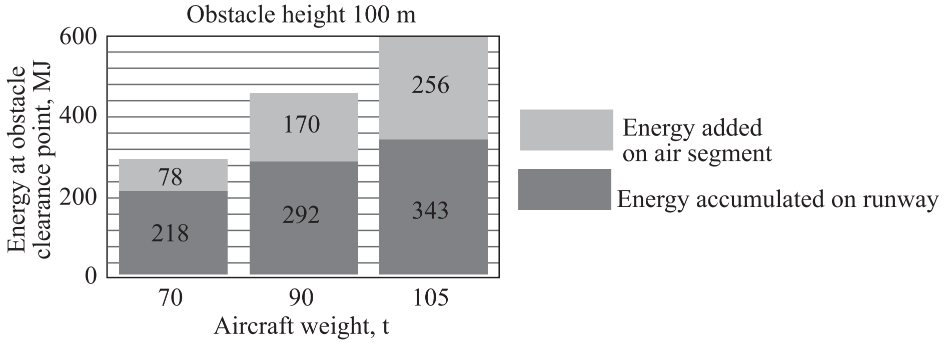


Fig. 2. Energy accumulation on the ground and air segments of the trajectory.

possible to calculate forecast values ahead of current events. Figure 2 demonstrates the energies on the ground and air segments for aircraft with three takeoff weights.

Thus, energy forecasting allows calculating the distances to all regulation events on the trajectory of a complicated takeoff ahead of time. Information about the occurrence of these events can be presented to the pilot on the cockpit indicator in text, audio, or graphic form. The pilot's awareness of the current and forecasted situation reduces the stress load and the probability of erroneous or untimely response of the pilot.

5. SIMULATION OF TAKEOFF IN THE PRESENCE OF OBSTACLES

The method for forecasting flight parameters at an obstacle clearance point was tested on a computerized bench. The bench included a complete certified model of the TU-204 aircraft, particularly the engine model and the landing gear model.

The operator's console was used to set the aircraft weight and alignment, climatic conditions, and airfield altitude and to prepare a takeoff scenario in accordance with the current flight regulations. In the bench, control during the ground run and takeoff was performed by the automatic EnCS.

The energy system saves and efficiently utilizes the resources of the controls—throttle lever and altitude channel knob—for spatial maneuvering. Therefore, the takeoff scenario contained only the required speed and altitude values.

Figure 3 shows the transients in the height YG and speed VP at takeoff in the presence of a 100 m-high obstacle at a distance of 1000 m from the runway endpoint for an aircraft with three different takeoff weights.

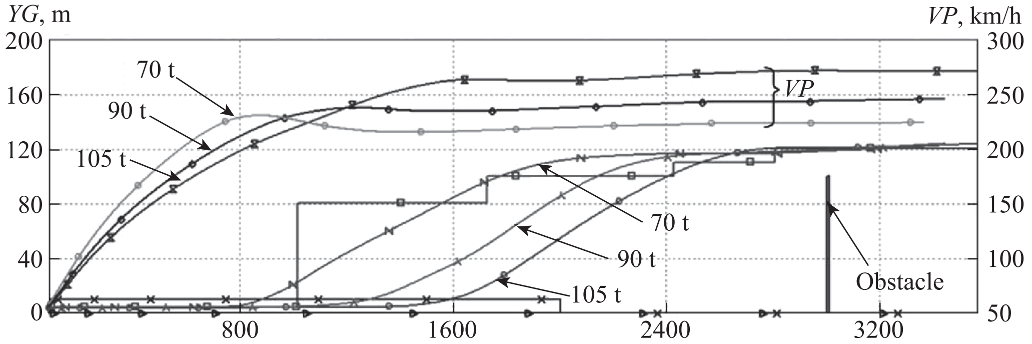
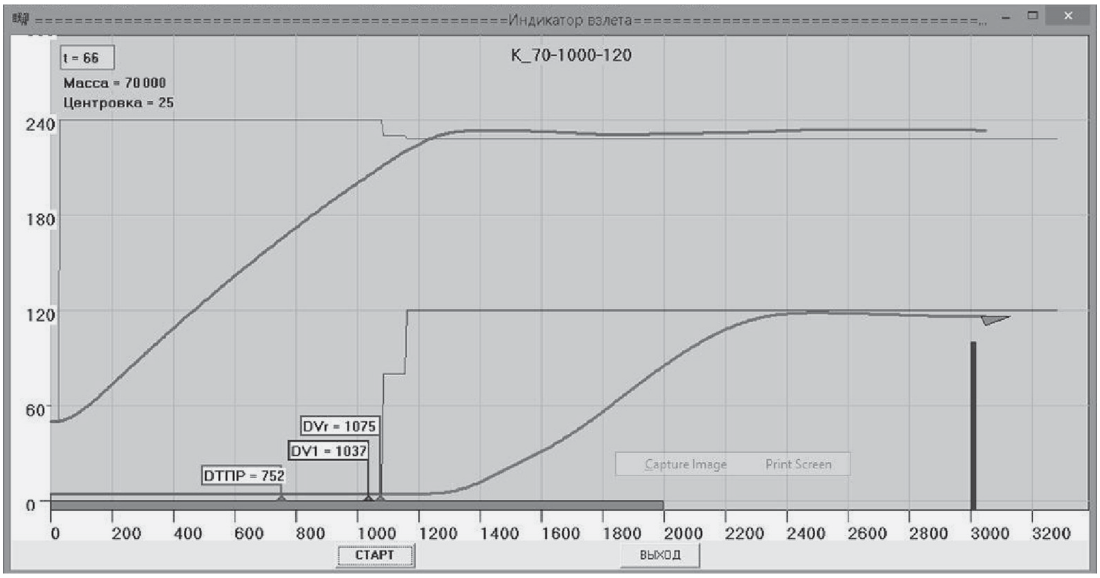


Fig. 3. Transients with the energy control system.



Индикатор взлета

K_70-1000-120

t = 66

Масса = 70 000

Центровка = 25

Fig. 4. Cockpit takeoff indicator window.

The simulation was carried out to compare the forecasted decision points for takeoff with the Flight Manual's recommendations for aircraft with different weights (from minimum to maximum) and the location of obstacles with heights of 50–150 m at a distance of 500–3000 m from the runway endpoint. During takeoff, the bench recorded the position of the aircraft on the runway (the coordinate X_{DP}) in which the current energy state was sufficient for a safe takeoff considering the forecasted motion.

Table 1 combines the coordinates of three points for an aircraft with takeoff weights of 70, 90, and 105 tons: the decision points calculated by forecasting ($X_{V_1}^{fore}$), the points of reaching the regulation takeoff speed V_1 factually ($X_{V_1}^{fact}$), and the points of reaching the nosewheel lift-off speed (X_{V_r}).

Clearly, the possibility of overcoming the obstacle, as well as the nosewheel lift-off speed, are forecasted much earlier than the aircraft gains the decision speeds V_1 and V_r prescribed by the Flight Manual.

For bench testing of takeoff modes with information support of the pilot, a prototype of a real-time indicator of aircraft movement on the ground and air segments was implemented. Figure 4 presents the indicator window at the obstacle clearance instant.

The indicator window demonstrates the histories of the set and factual values of the main flight parameters (altitude and speed). The aircraft symbol on the altitude trajectory shows its current position. The runway and obstacle are conditionally depicted as well. The prototype of the indicator successively marks in real time the forecasted distances to the decision point for takeoff (D_{DP}), to the point of reaching the regulation decision speed (DV_1), and to the point of nosewheel lift-off speed (DV_r), including their numerical values.

Table 1. Comparison of forecasted and factual coordinates

Weight, t	V_1, km/h	$X_{V_1}^{fact}$, m	$X_{V_1}^{fore}$, m	V_r, km/h	X_{V_r}, m
70	204	515	153	210	547
90	220	764	508	228	825
106	238	1095	837	245	1203

6. METHOD FOR FORECASTING SAFE BRAKING DISTANCE

Figure 5 shows the landing diagram with the following notations: $x(t)$ is the current position of the aircraft on the runway; D_{brak} is the braking path length; X_{brak} is the end point coordinate; finally, L_{res} is the ground run reserve to the runway endpoint.

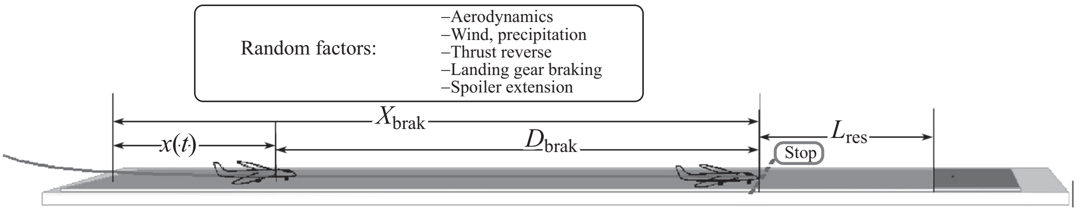


Fig. 5. Landing with braking.

Within the ground segment of the trajectory, during the run after landing or before an aborted takeoff, there may be situations with a risk of overrunning the runway. Under a time deficit, it is necessary to assess the possibility of either emergency braking and stopping within the runway or going around again. We define the braking length as the distance over which the airspeed will be canceled from the current one to some small value ϵ or the taxiing speed.

For the stopping criterion $V(t) \leqslant \epsilon$, the forecasted braking length is given by

$$D_{brak} = \frac{0.5(V^2(t) - \epsilon^2}{gn_x(t)}. \tag{4}$$

According to this estimate of the marginal stopping distance of the aircraft, the pilot may be visually informed of the safe braking distance reserve

$$L_{res} = L_{run} - x(t) - D_{brak}.$$

This information message will help the pilot to make a decision on emergency braking (and in the case of its impossibility, a decision on go-around).

In the process of braking, all forces and conditions change; therefore, the a priori estimates of the aircraft motion on the runway differ from the real ones, containing an inevitable error. Moreover, the current situation forecast is always optimistic since the main braking forces (reverse thrust and aerodynamic drag) relax with decreasing the speed.

To improve the reliability of forecasting, we propose to correct the forecast (4) by introducing a correction Q_{cor} and calculating the corrected braking distance

$$D_{brak\ cor} = Q_{cor} D_{brak}. \tag{5}$$

The highest forecasting errors occur on sections with maximum reverse and with extended spoilers, so the correction coefficients are selected separately for each configuration of the braking devices. These sections are always identifiable, and switching the type of correction is straightforward.

At the beginning of the braking path (the reverse section), the greatest impact on the forecasting errors is exerted by the friction coefficient k_{fric} (which is reported to the board for the landing calculation) and the rolling velocity V (which is bounded by the reverse speed, $V \geqslant V_{rev}$).

The correction coefficient on the reverse section, Q_{rev}, explicitly considers both factors mentioned:

$$Q_{rev} = k_{rev}(k_{fric}) k_{rev}(V).$$

The value $k_{rev}(k_{sc})$ was analytically approximated by the polynomials of the second, third, and fourth degrees. The polynomial of the third degree has the form

$$k_{rev}(k_{fric}) = 16.14(k_{fric})^3 - 22.55(k_{fric})^2 + 8.25k_{fric} + 0.716.$$

Despite the differences in the approximating polynomials, the resulting errors varied by no more than 10%.

The empirical dependence of the correction coefficient on speed was found in the form

$$k_{rev}(V) = k_1(k_0 + (1 - k_0))V/V_H,$$

where V_H is the initial braking speed, the coefficient k_1 determines the overall intensity of correction, and the coefficient k_0 changes the degree and sign of correction as the aircraft moves along the runway. The tuning coefficients k_0 and k_1 were determined by minimizing the average forecast error on the reverse section.

On the ground run section with extended spoilers, the correction was achieved by simply scaling the coefficients using the normalized average landing weight $m_{norm=m/90}$:

$$Q_{spoil} = k_i m_{norm}.$$

The values k_i were found by minimizing the error over the entire flight under all braking conditions. After retraction of the spoilers, the correction coefficient was scaled to $Q_{spoil} = 0.8K_i m_{norm}$.

The states of the braking devices and the actions of external factors change at a high rate. Therefore, to smooth out possible high-frequency bursts, all the forecasted values are passed through a damping filter, i.e., an aperiodic link with a tunable time constant Tf_{fore}.

7. STUDIES OF THE BRAKING DISTANCE FORECASTING ALGORITHM

A special simulation bench was created to study the forecasting algorithms. This bench has a set of modes to analyze the forecasting algorithms and to perform their correction and studies as well as developed service tools for setting the experimental conditions and processing and recording the results.

First of all, the bench is used to determine the correction coefficients in terms of the selected optimality criteria (forecasting errors on any trajectory section). The program module of the forecasting algorithms contains a base of settings for the coefficients of the algorithm (5) on a discrete set of braking conditions. To make the coverage of the settings domain continuous, the software includes a module for interpolating the correction coefficients as a function of three variables: $[k_0, k_1, k_i] = INTERPOL[m, k_{fric}, V_{pos}]$.

The service software of the simulation bench includes a module for analyzing the results of statistical tests of the forecasting algorithms. The statistical testing module is configured to analyze the forecasting errors of the stopping point during aircraft braking on the runway. The random disturbances are the variations in aircraft weight and friction coefficient. The distribution law can be assigned as Gaussian or uniform. When displaying the curves on the screen, the experimental distribution function is plotted along with the analytical Gaussian function with the same moments.

Figure 6 shows the experimental distribution functions and the corresponding probability densities of the forecasting errors of the braking distance (ΔD_{brak}) for a 90-ton aircraft from an initial speed of 220 km/h. The analytical approximation of the distribution function by the Gaussian law is plotted on each graph. The mean and the width of the 5% error tolerance are also provided.

According to the graphs, random forecasting errors have distributions close to Gaussian. Small values of the mean and standard deviation indicate high forecasting accuracy, which is achieved by the effective correction of forecasting algorithms.

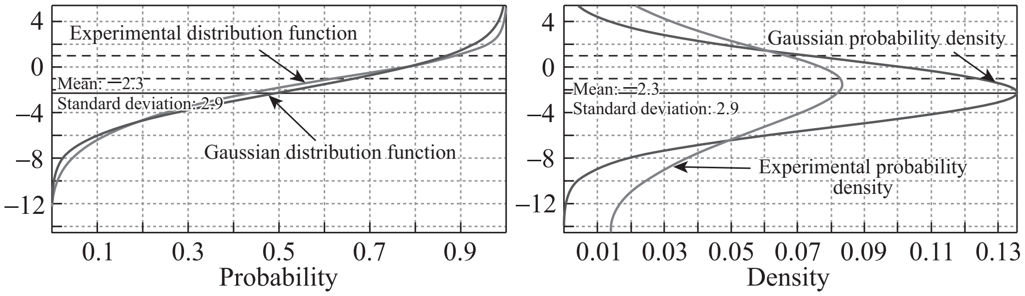


Fig. 6. Distribution functions and probability densities of the forecasting errors of braking distance.

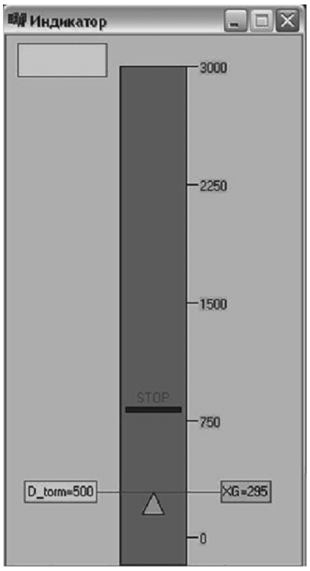


Fig. 7. Braking indicator.

Engine thrust reverse has the strongest impact on the dynamics of the braking process. Increasing the reliability of forecasting on the reverse section is very important: on this section, the speed takes the highest values, increasing the stress load on the pilot. Forecasting errors during the entire braking stage (total errors) and those in the reverse mode only (reverse errors) were investigated and compared. The correction coefficients were determined using two different optimality criteria: by minimizing the errors on the reverse section, min(reverse errors), and by minimizing the errors on the complete braking trajectory, min(total errors).

Table 2 presents the average forecasting errors on the reverse section and on the complete braking trajectory of an aircraft with a landing weight of 90 t, an initial speed of 220 km/h, and friction coefficients of 0.3, 0.5, and 0.75.

These data confirm that the reverse section contributes most to the forecasting error, and optimization in terms of the minimum error on the reverse section also significantly reduces the total error over the entire run.

Table 2. Forecasting errors on the reverse section and complete braking trajectory

Friction coefficient	0.3	0.3	0.5	0.5	0.75	0.75
Optimality criterion	Reverse errors	Total errors	Reverse errors	Total errors	Reverse errors	Total errors
min(reverse errors)	−8.97	−8.94	−0.48	10.27	−0.23	6.03
min(total errors)	−21.35	−3.81	−3.54	−2.0	1.55	0.55

Figure 7 demonstrates the prototype of the braking indicator for information support of the pilot of an aircraft moving in real time along the runway. There are marks of the current position of the aircraft and the forecasted braking endpoint. The numerical value of the aircraft coordinate on the runway and the estimated distance to the stopping point are also shown.

If the forecasted stopping point goes beyond the runway endpoint, it is a signal to go around.

8. CONCLUSIONS

To increase situational awareness of the pilot and reduce stress load, we have developed algorithms for forecasting terminal states during takeoff and landing operations. The algorithms are based on the energy approach to aircraft flight control. This approach allows assessing the current situation and, moreover, the future situation on the forward section of the trajectory, including the air segment of climbing and overcoming a high-rise obstacle. The idea is to inform the pilot of the forecasting results in the form of text, graphic, or audio alerts. In the ground run mode before takeoff, the distance to the decision point on the possibility of safe takeoff and clearance of a high-rise obstacle has been determined. In the braking mode, algorithms for forecasting the distance to the stopping point or to the taxiing speed have been developed. In each mode mentioned, the possibility of safely reaching critical points of the maneuver has been forecasted ahead of their factual occurrence on the trajectory. This gives confidence in fulfilling the flight task in nonstandard or complicated conditions on the runway.

REFERENCES

1. Aviation Accidents and Incidents in the First Quarter of Year 2023. http://old.aviamettelecom.ru.

2. Accidents Statistics. http://www.planecrashinfo.com/cause.

3. Borodkin, S., Volynchuk, A., Ganiev, Sh., et al., Modern Methods of Preventing Aircraft Overrunning the Runway, *Civil Aviation High Technologies*, 2022, vol. 25, no. 2, pp. 1–12.

4. Grebenkin, A. and Burdun, I., Landing under Extreme Conditions: Early Safety Screening by Means of the "Pilot–Automaton–Aircraft–Operating Environment" System Dynamics Model, *Proc. SAE 2019 Aviation Technology Forum*, 2019.

5. Grebenkin, A.V. and Lushnikov, A.A., Human Factor Consideration in the Integration Problems of Manual and Automatic Control of a Trunk Airliner under Complex Multifactor Landing Conditions, *Trudy 2-go Vserossiiskogo foruma s mezhdunarodnym uchastiem "Akademicheskie Zhukovskie chteniya"* (Proc. 2nd All-Russian Forum with International Participation "Academic Readings from Zhukovskii"), Voronezh: Military Training and Research Center of the Air Force, Air Force Academy, 2022, pp. 224–231.

6. Nikiforov, S.P., An Onboard Ground Run Control System as an Effective Means of Improving the Safety of Transport Aircraft Takeoffs, *Aviation Science and Technology*, 2020, nos. 3–4, pp. 47–54.

7. Shevchenko, A., Some Means for Informational Support of Airliner Pilot, *Proc. 5th Int. Conf. on Physics and Control (PhysCon 2011)*, Leon, 2011, pp. 1–5.

8. Kuznetsov, A., Shevchenko, A., and Solonnikov, Ju., The Methods of Forecasting Some Events During the Aircraft Takeoff and Landing, *Proc. 19th IFAC Symposium on Automatic Control in Aerospace (ACA2013)*, Germany, 2013, pp. 183–187.

9. Kurdjukov, A., Nachinkina, G., and Shevtchenko, A., Energy Approach to Flight Control, *Proc. AIAA Conf. on Navigation, Guidance and Control*, AAIA paper no. 98-4211, Boston, 1998.

10. Lambregts, A., Vertical Flight Path and Speed Control Autopilot Design Using Total Energy Principles, AIAA paper no. 83-2239CP, 1983, pp. 559–569.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 10, pp. 1243–1258.
© The Author(s), 2023 published by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 2023.
Russian Text © The Author(s), 2023, published in Avtomatika i Telemekhanika, 2023, No. 10, pp. 132–149.

TOPICAL ISSUE

Spectral Decompositions of Gramians and Energy Metrics of Continuous Unstable Control Systems

I. B. Yadykin[*,a] **and I. A. Galyaev**[*,b]

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

Received June 19, 2023
Revised July 15, 2023
Accepted August 2, 2023

Abstract—Deterministic continuous finite-dimensional stationary linear dynamic control systems with many inputs and many outputs are considered. Authors assume that the dynamics matrix can be both stable and unstable, but its eigenvalues are different, do not belong to the imaginary axis, and their pairwise sum is not equal to 0. The problems of constructing spectral solutions of the equations of state and matrices of gramian controllability of these systems, as well as the associated energy functionals of the degree of stability and reachability with the aim of optimal placement of sensors and actuators of multi-connected control systems and complex networks are considered. To solve the listed problems, the article uses various models of the system in state space: a general representation, as well as a representation in various canonical forms. To calculate the spectral decompositions of controllability gramians, pseudo-Hankel matrices (Xiao matrices) are used. New methods have been proposed and algorithms have been developed for calculating controllability gramians and energy metrics of linear systems. The research results can be used for the optimal placement of sensors and actuators of multi-connected control systems or for control with minimal energy in complex networks of various natures.

Keywords: spectral decompositions of gramians, energy functionals, inverse matrix of gramians, stability that takes into account the interaction between modes, Lyapunov equation, unstable control systems

DOI: 10.25728/arcRAS.2023.99.15.001

1. INTRODUCTION

Monitoring the state of control objects and controlling the damping of dangerous vibrations are important areas of research in various fields of industry (energy, mechanical engineering, aviation and astronautics, robotics). New modeling technologies require the development of tools for approximating mathematical models of complex systems of various natures. When solving these problems, an important role is played by the methods of calculating the Lyapunov and Sylvester matrix equations and the study of the structural properties of solutions to these equations [1–4]. The fundamental properties of linear dynamic systems associated with solutions to these equations are controllability, observability and stability. Important results in this area were obtained for methods for calculating the gramians of systems, the models of which are presented in the canonical forms of controllability and observability. The application of gramians for constructing simplified models of high-dimensional dynamic systems and for calculating the norms of transfer functions of linear and bilinear dynamic systems is well known [1, 2, 5–8]. Controllability gramians play an important role in calculating output deviations caused by Gaussian random disturbances. In recent years, interest has arisen in the development of methods for calculating various energy indicators to analyze the stability and degree of controllability and observability of these systems.

Such indicators for linear stable systems and unstable linear systems were proposed in [1, 8–11]. Simplified models for large networks based on output controllability gramians, that allow to calculate the energy indicators, were proposed in [12]. The balanced truncation method, based on the gramians of stable and anti-stable systems, was proposed in [13]. The important problem of optimal placement of sensors and actuators based on various energy functionals, including invariant ellipsoids, and estimation of the degree of controllability was studied in [14–18]. It is important to note that all these works used the spectrum of the system dynamics matrix.

B.N. Petrov and his students developed methods, based on Lyapunov direct method, for synthesizing adaptation algorithms that guarantee the stability of the movement of a self-adjusting system relative to the movement of its reference model [19, 20]. He developed the principle of coordinate-parametric control, which implements double invariance in non-search self-tuning systems (NSTS). In the theory of NSTS, the concept of a generalized customizable object was used, which was based on identifying the structures of a specially formed main circuit and a circuit of a customizable controller. Linearized mathematical models of circuits included coordinate, parametric and coordinate-parametric models, including parametric feedbacks in controller tuning circuits. These models are called bilinear dynamic models and are used in optimization, identification theory, and adaptive control. To calculate the gramians of these systems, generalized Lyapunov equations were developed and spectral methods for solving them were proposed [2, 10, 11]. A significant contribution was made by the school of B.N. Petrov in the formation of control theory, based on the use of the structural properties of the reference model, and in other areas of control theory, in particular in the theory of invariant systems.

2. FORMULATION OF THE PROBLEM

Consider a continuous time-invariant linear dynamic MIMO LTI system with a simple spectrum with many inputs and many outputs

$$\Sigma_1 \colon \begin{cases} \dot{x} = Ax\,(t) + Bu\,(t)\,, \quad x(0) = 0, \\ y\,(t) = Cx\,(t)\,, \end{cases} \tag{2.1}$$

where $x(t) \in R^n, u(t) \in R^m, y(t) \in R^m$.

If all eigenvalues s_r of matrix A are different, then the linear system can be reduced to diagonal form using a non-degenerate coordinate transformation

$$\begin{aligned} &x_d = Tx, \quad \dot{x}_d = A_d x_d + B_d u, \quad y_d = C_d x_d, \\ &A_d = T^{-1}AT, \quad B_d = T^{-1}B, \quad C_d = CT, \quad Q_d = T^{-1}BB^{\mathrm{T}}(T^{-1})^{\mathrm{T}}, \end{aligned}$$

or

$$\mathrm{A} = \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix} \begin{bmatrix} s_1 & 0 & 0 & 0 \\ 0 & s_2 & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & s_n \end{bmatrix} \begin{bmatrix} \nu_1^* \\ \nu_2^* \\ \vdots \\ \nu_n^* \end{bmatrix} = T\Lambda T^{-1},$$

where the matrix T^{-1} is composed of right eigenvectors u_i, and the matrix T is composed of left eigenvectors ν_i^* corresponding to the eigenvalue s_i.

Definition [21]. The square matrix $Y = [y_{j\eta}]$ is called the Xiao matrix (Zero plaid structure) and has the form:

$$Y = \begin{bmatrix} y_1 & 0 & -y_2 & 0 & y_3 \\ 0 & y_2 & 0 & -y_3 & 0 \\ -y_2 & 0 & y_3 & 0 & \dots \\ 0 & -y_3 & 0 & \dots & 0 \\ y_3 & 0 & \dots & 0 & y_n \end{bmatrix},$$

its elements are specified using the elements of the Routh table [21]:

$$y_{j\eta} = \begin{cases} 0, \quad \text{if } \ j+\eta = 2k+1, \quad k{=}1,\dots,n; \\ y_n = \dfrac{1}{2R_{n,1}}, \\ \\ y_{n-l} = \dfrac{-\sum_{i=1}^{m-1}(-1)^i R_{n-l,i+1} y_{n-l+i}}{R_{n-l,1}}, \\ \qquad \text{if } \ j+\eta{=}\ 2k, \ \ k=1,\dots,n, \ \ l=\overline{1,n-1}, \end{cases}$$

where $R_{i,j}$ is the Routh table element for the system, located at the intersection of row i and column j. In [11], a spectral decomposition of the controllability gramian of a continuous linear system with many inputs and many outputs was obtained based on the method for calculating the gramian proposed in [21, 22].

Theorem 1 [11, 21]. *We consider a continuous linear MIMO LTI system of the form* (2.1)*. Let us assume that the system is stable and all the roots of its characteristic equation are different. Then the matrices of its controllability gramian are Xiao matrices, the diagonal elements of which are defined as*

$$p_{11} = \sum_{k=1}^{n} \frac{1}{2s_k \prod_{\rho=1,\rho\neq k}^{n}\left(s_k^2 - s_\rho^2\right)},$$

$$p_{22} = \sum_{k=1}^{n} \frac{(-1)^1 (s_k)^2}{2s_k \prod_{\rho=1,\rho\neq k}^{n}\left(s_k^2 - s_\rho^2\right)},$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots,$$

$$p_{nn} = \sum_{k=1}^{n} \frac{(-1)^{n-1} (s_k)^{2(n-1)}}{2s_k \prod_{\rho=1,\rho\neq k}^{n}\left(s_k^2 - s_\rho^2\right)}.$$

The elements of the side diagonals of the gramian matrices are defined as:

$$p_{j\eta} = (-1)^{\frac{j-\eta}{2}} p_{ll}, \quad j+\eta = 2l, \quad l = \overline{1,n}.$$

The remaining elements of the gramian matrix are equal to zero.

Corollary 1. Consider a stable continuous stationary linear dynamic MIMO LTI system with a simple spectrum with many inputs and many outputs of the form (2.1). Then its controllability gramian is a matrix of the form [11]

$$P_c = \sum_{j=0}^{n-1}\sum_{\eta=0}^{n-1} P_{cj,\eta},\ P_{cj,\eta} = \omega(n, s_k, j, \eta) A_j B B^{\mathrm{T}} A_\eta^{\mathrm{T}}, \tag{2.2}$$

$$\omega\left(n, s_k, j, \eta\ \right) = \begin{cases} 0,\ \text{if } j+\eta = 2k+1,\ \ k = 1,\ldots,n, \\ \displaystyle\sum_{k=1}^{n} \frac{s_k^j (-s_k)^\eta}{2 s_k \prod_{\rho=1,\rho\neq k}^{n}\left(s_k^2 - s_\rho^2\right)},\ \text{if } j+\eta = 2k,\ \ k = 1,\ldots,n. \end{cases}$$

We will call spectral decompositions (2.2) the gramian decompositions in the form of Xiao matrices. In the expansion (2.2) a scalar multiplier function $\omega\left(n, s_k, j, \eta\right)$ appears, which determines the structure of the Hadamard matrices [21].

Let us transform the system (2.1) into the upper block-diagonal Schur form with a unitary transformation matrix U [23,24].

$$\begin{gathered} x = U x_{Sch},\ \ \dot{x}_{Sch} = A_{Sch} x_{Sch} + B_{Sch} u,\ \ y_{Sch} = C_{Sch} x_{Sch}, \\ A_{Sch} = U^{\mathrm{T}} A U,\ \ B_{Sch} = U^{\mathrm{T}} B,\ \ C_{Sch} = C U, \end{gathered} \tag{2.3}$$

$$A_{Sch} = \begin{bmatrix} A_{Sch11} & A_{Sch12} \\ 0 & A_{Sch22} \end{bmatrix},\ \ B_{Sch} = \begin{bmatrix} B_{Sch1} \\ B_{Sch2} \end{bmatrix},\ \ C_{Sch} = \begin{bmatrix} C_{Sch1} & C_{Sch2} \end{bmatrix}.$$

In order to obtain a block-diagonal representation, it is necessary to transform the equations (2.3) so that the place of the A_{Sch12} block is replaced by a zero matrix. To do this, we perform a second transformation

$$\begin{gathered} x_{Sch} = W_{bl} x_{bl},\ \ \dot{x_{bl}} = A_{bl} x_{bl} + B_{bl} u,\ \ y_{bl} = C_{bl} x_{bl}, \\ A_{bl} = W_{bl}^{-1} A_{Sch} W_{bl},\ \ B_{bl} = W_{bl}^{-1} B_{Sch},\ \ C_{bl} = C_{Sch} W_{bl}, \end{gathered} \tag{2.4}$$

$$A_{bl} = \begin{bmatrix} A_{Sch11} & 0 \\ 0 & A_{Sch22} \end{bmatrix},\ \ B_{bl} = \begin{bmatrix} B_{bl1} \\ B_{bl2} \end{bmatrix},\ \ C_{bl} = \begin{bmatrix} C_{bl1} & C_{bl2} \end{bmatrix},$$

$$W_{bl} = \begin{bmatrix} I_r & S \\ 0 & I_{n-r} \end{bmatrix}, W_{bl}^{-1} = \begin{bmatrix} I_r & -S \\ 0 & I_{n-r} \end{bmatrix}.$$

In order for the block A_{Sch12} to be replaced by a zero matrix, the matrix S must satisfy the Sylvester equation

$$-A_{Sch11} S + S A_{Sch22} + A_{Sch12} = 0. \tag{2.5}$$

A necessary condition for the existence of a solution to this equation is the following spectral condition:

$$\lambda_s + \lambda_u \neq 0,\ \ \forall s : s = \overline{1, r}, \forall u : u = \overline{r+1, n}.$$

In order to transform a system (2.4) with a block diagonal matrix into a system with a diagonal matrix, it is necessary to perform a third transformation

$$x_{bl} = W_d x_d,$$

where W_d is the transformation matrix of a system in block-diagonal form, which diagonal blocks have an upper-triangular shape

$$\begin{gathered} \dot{x_d} = A_d x_d + B_d u,\ \ y_d = C_d x_d, \\ A_d = W_d^{-1} A_{bl} W_d,\ \ B_d = W_d^{-1} B_{bl},\ \ C_d = C_{bl} W_d, \end{gathered}$$

$$A_d = \begin{bmatrix} \Lambda_- & 0 \\ 0 & \Lambda_+ \end{bmatrix},\ \ B_d = \begin{bmatrix} B_{d1} \\ B_{d2} \end{bmatrix},\ \ C_d = \begin{bmatrix} C_{d1} & C_{d2} \end{bmatrix}, \tag{2.6}$$

where Λ_- and Λ_+ are diagonal matrices consisting of negative and positive eigenvalues, respectively.

After the first transformation we have the relation

$$P = UP_{Sch}U^{\mathrm{T}}. \tag{2.7}$$

After the second transformation we get

$$P_{Sch} = TP_{bl}T^{\mathrm{T}},$$

or

$$P = T_2P_{bl}T_2^{\mathrm{T}}, \;\; T_2 = UT. \tag{2.8}$$

After the third transformation using (2.7),(2.8) we get

$$P = UT_3P_dT_3^{\mathrm{T}}, \;\; T_3 = UTW_d.$$

The structured Lyapunov equation after the second transformation has the form

$$A_{Sch11}P_1 + P_1A_{Sch11}^{\mathrm{T}} = -B_1B_1^{\mathrm{T}}, \tag{2.9}$$

$$A_{Sch22}P_2 + P_2A_{Sch22}^{\mathrm{T}} = B_2B_2^{\mathrm{T}}, \tag{2.10}$$

$$P_{cm} = T_2^{-1}\begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}T_2. \tag{2.11}$$

The matrix P_{cm} is called the mixed controllability gramian [13–17, 25]. The purpose of the article is to develop a method and algorithm for calculating spectral decompositions of the controllability gramians of unstable linear systems, based on the method described above, for calculating the specified gramians using the transformation of the original system into a block-diagonal form [13].

Many applications of spectral decompositions of gramians are associated with energy indicators of the structural properties of controllability, observability and stability. We consider the problem of selecting and optimizing the placement of sensors and actuators in complex automatic systems and complex networks [18, 26–28]. To solve this problem one could use the input and output energy of the system, traces of the controllability and observability gramian matrices and traces of their inverse matrices, minimum and maximum eigenvalues of the gramian. Another problem is to estimate the controllability measure of a dynamic system using controllability gramians [25]. This measure is defined as the minimum input energy required to move the system from an arbitrary initial state to an arbitrary final state.

Another goal of the article is to develop a method and algorithms for calculating spectral decompositions of energy metrics related to the above problems. It is required to find spectral decompositions of the following energy metrics from the simple (or paired) spectrum of the system dynamics matrix and the controllability and observability gramian matrices:

- metric of the input minimum energy of the system [2, 18]

$$J_1 = E_{\min}\left(P_c\right),$$

- output energy metric of the system [2, 3]

$$J_2 = E_{out},$$

- trace metric of the gramian matrix [26, 27]

$$J_3 = \mathrm{tr}\,(P_c),$$

- trace metric of inverse matrices of controllability gramians [2, 12, 18]

$$J_4 = \operatorname{tr}(P_c)^{-1},$$

- reachability metric

$$J_5 = \operatorname{tr}(P_{cm}),$$

where P_{cm} is the mixed controllability gramian [18, 25].

2.1. Main Results

Let us consider a finite-dimensional linear stationary continuous system with many inputs and many outputs of the form (2.1). We suppose that the spectrum of the dynamics matrix contains r stable eigenvalues $\lambda_{i-} \in \mathbb{C}^-$ and $n-r$ unstable eigenvalues $\lambda_{i+} \in \mathbb{C}^+$. We will assume that the spectrum does not contain eigenvalues belonging to the imaginary axis, and the general condition is satisfied

$$\lambda_{i-} + \lambda_{j+} \neq 0, \forall i : i = \overline{1, r}, \forall j : j = \overline{r+1, n}.$$

The last condition means that the spectrum does not contain eigenvalues that are mirror images of each other relative to zero. The simplest way to calculate the spectral decompositions of gramians in the case of a simple spectrum of the dynamics matrix is to reduce it to diagonal form [1, 11]. If unstable eigenvalues appear in the spectrum, this requires several structural transformations of the (2.1) equations. Let us introduce the notation

$$B_{d11}B_{d11}^{\mathrm{T}} = [\beta_{d-\nu\eta}]_{[r\times r]},$$

$$B_{d22}B_{d22}^{\mathrm{T}} = [\beta_{d+\nu\eta}]_{[(n-r)\times(n-r)]}.$$

Theorem 2 [8]. *Let us consider a finite-dimensional linear stationary continuous system with many inputs and many outputs of the form* (2.1)*, reduced to the diagonal form* (2.6)*. Let us assume that the system has a simple spectrum, the system is unstable, and the eigenvalues of its dynamics matrix A are not on the imaginary axis, but can be in the left and/or right half-planes $\lambda_{i-} \in \mathbb{C}^-$, $i = r$; $\lambda_{i+} \in \mathbb{C}^+$, $i = n-r$.*

In addition, assume that the condition is satisfied

$$\lambda_i \neq -\lambda_j, \ \forall i, j : i = \overline{1, n}, \ j = \overline{1, n}.$$

Let us define the mixed controllability gramian in the form

$$P_{cm} = \frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} (Ij\omega - A)^{-1} B B^{\mathrm{T}} (-Ij\omega - A^{\mathrm{T}})^{-1} d\omega. \tag{2.12}$$

The following statements are valid and equivalent.

- *The following separable spectral decompositions of the matrices of solutions to the equation* (2.9), (2.10)*, corresponding to the stable and anti-stable subsystems, are valid.*

$$p_{c-}^{(\mu\nu)} = e_\mu^{\mathrm{T}} P_{c-} e_\nu, \quad \forall \mu, \nu = \overline{1, r},$$

$$p_{c-}^{(\mu\nu)} = \frac{-\beta_{\mu\nu-}}{\lambda_{\mu-} + \lambda_{\nu-}},$$

$$p_{c+}^{(\mu\nu)} = e_\mu^{\mathrm{T}} P_{c+} e_\nu, \quad \forall \mu, \nu = \overline{r+1, n},$$

$$p_{c+}^{(\mu\nu)} = \frac{\beta_{\mu\nu+}}{\lambda_{\mu+} + \lambda_{\nu+}};$$

- *The following separable spectral expansions of the mixed gramian of controllability in the pair and simple spectra of the matrix A are valid:*

$$P_{cm} = T_3^{-1} \left[P_- \oplus P_+\right] T_3. \tag{2.13}$$

According to the pair spectrum:

$$P_- = \sum_{\nu=1}^{r} \sum_{\mu=1}^{r} p_{c-}^{(\nu\mu)} \mathbb{1}_{\nu\mu}, \tag{2.14}$$

$$P_+ = \sum_{\nu=r+1}^{n} \sum_{\mu=r+1}^{n} p_{c+}^{(\nu\mu)} \mathbb{1}_{\nu\mu}.$$

According to a simple spectrum:

$$P_- = \sum_{\nu=1}^{r} \boldsymbol{p}_{c-}^{(\nu)}, \quad \boldsymbol{p}_{c-}^{(\nu)} = \sum_{\mu=1}^{r} \boldsymbol{p}_{c-}^{(\nu\mu)} \mathbb{1}_{\nu\mu}, \tag{2.15}$$

$$P_+ = \sum_{\nu=r+1}^{n} \boldsymbol{p}_{c+}^{(\nu)}, \quad \boldsymbol{p}_{c+}^{(\nu)} = \sum_{\mu=r+1}^{n} \boldsymbol{p}_{c+}^{(\nu\mu)} \mathbb{1}_{\nu\mu}.$$

Proof of Theorem. The Lyapunov equations for the diagonalized system in this case have the form

$$\Lambda P_{cm} + P_{cm}\Lambda^* = -Q_d = \left[-B_- B_-^{\mathrm{T}} \oplus B_+ B_+^{\mathrm{T}}\right].$$

For a diagonalized system, this equation splits into two equations for stable and antistable subsystems

$$\Lambda_- P_{c-} + P_{c-}\Lambda_-^* = Q_{d-} = -B_- B_-^{\mathrm{T}},$$
$$\Lambda_+ P_{c+} + P_{c+}\Lambda_+^* = Q_{d+} = B_+ B_+^{\mathrm{T}}.$$

Integral formulas for solutions of Lyapunov equations [8]:

$$P_{cm} = [P_{c-} \oplus P_{c+}],$$

$$P_{c-} = \int_0^\infty e^{\Lambda_- \tau}\, B_- B_-^{\mathrm{T}} e^{\Lambda_-^* \tau} d\tau, \quad P_{c+} = \int_{-\infty}^{0} e^{\Lambda_+ \tau}\, B_+ B_+^{\mathrm{T}} e^{\Lambda_+^* \tau} d\tau. \tag{2.16}$$

Let’s transform the second integral in the formula (2.16) using the replacing of variables $\tau = -t$:

$$\int_{-\infty}^{0} e^{\Lambda_+ \tau}\, Q_{d+} e^{\Lambda_+ \tau} d\tau = -\int_0^\infty e^{-\Lambda_+ t}\, Q_{d+} e^{-\Lambda_+^* t} dt.$$

With such a change of variables, the unstable eigenvalues of the antistable subsystem become stable eigenvalues of the stable subsystem and the calculation of the second integrals is reduced to the scheme for calculating the first integrals (2.16). This implies

$$(-\Lambda_+)P_{c+} + P_{c+}(-\Lambda_+^*) = -B_+ B_+^{\mathrm{T}}.$$

The matrix $[\Lambda_- \oplus (-\Lambda_+)]$ is Hurwitz. Spectral expansions of the gramians of a stable subsystem were previously obtained in [9]. First, we obtain spectral decompositions of the gramians in (2.16), and then we obtain the spectral decomposition of the gramians of the original system according to the formula for transforming the gramians of controllability for a nondegenerate transformation of states with matrix T

$$P_{cm} = T[P_- \oplus P_+]\, T^{\mathrm{T}}. \tag{2.17}$$

The first step of spectral decompositions is based on transforming the equations of state of a stable subsystem into a diagonal canonical form. In this case, the Lyapunov equations take on a simple form and the elements $p_{c-}^{(\mu\nu)}$ of the solution matrix P_{c-} can be calculated using the formulas [9]

$$p_{c-}^{(\mu\nu)} = e_\mu^{\mathrm{T}} P_{c-} e_\nu, \ \forall \mu, \nu = \overline{1, r}, \tag{2.18}$$

where $e_\mu^{\mathrm{T}}, e_\nu$ are unit vectors,

$$\begin{gathered} e_\mu^{\mathrm{T}} Q_{d-} e_\nu = \beta_{\mu\nu-}, \ \forall \mu, \nu = \overline{1, r}, \\ p_{c-}^{(\mu\nu)} = \frac{-\beta_{\mu\nu-}}{\lambda_{\mu-} + \lambda_{\nu-}}. \end{gathered} \tag{2.19}$$

Since, taking into account the change of variables, the calculation of spectral decompositions of the solution matrix P_{c+} is reduced to considering the approach proposed for calculating the solution matrix P_{c-}, we present the final formulas for calculating the spectral decompositions for this case.

This approach is based on transforming the equations of state of an antistable subsystem into a diagonal canonical form. In this case, the elements $p_{c+}^{(\mu\nu)}$ of the solution matrix P_{c+} are calculated using the formulas

$$p_{c+}^{(\mu\nu)} = e_\mu^{\mathrm{T}} P_{c+} e_\nu, \quad \forall \mu, \nu = \overline{r+1, n},$$

where $e_\mu^{\mathrm{T}}, e_\nu$ are unit vectors,

$$\begin{gathered} e_\mu^{\mathrm{T}} Q_{d+} e_\nu = \beta_{\mu\nu+}, \\ p_{c+}^{(\mu\nu)} = \frac{\beta_{\mu\nu+}}{\lambda_{\mu+} + \lambda_{\nu+}}, \ \forall \mu, \nu = \overline{r+1, n}. \end{gathered} \tag{2.20}$$

The proof of the validity of spectral expansions for the antistable subsystem completely repeats the proof for the stable subsystem. The proof of the validity of the spectral decompositions (2.13)–(2.15) follows from the validity of the formula (2.19) and the transformation of the antistable subsystem to the form of a stable subsystem, the eigenvalues of which are a mirror image of the eigenvalues of the first subsystem with respect to the imaginary axis. Theorem 2 is proven.

Corollary 2. If the conditions of the theorem are satisfied, the mixed gramian is positive definite, since the matrix $[\Lambda_- \oplus (-\Lambda_+)]$ is Hurwitz. In this case, the trace of the mixed controllability gramian is equal to

$$J = \sum_{i=1}^{r} \frac{\beta_{d-ii}}{-2Re\ \lambda_i} + \sum_{i=r+1}^{n} \frac{\beta_{d+ii}}{2Re\ \lambda_i}. \tag{2.21}$$

The coefficients $\beta_{d-ii}, \beta_{d+ii}$ are always positive due to the formation of the matrices of the right sides of the Lyapunov equations. It follows that the diagonal terms of the mixed gramian matrix are positive. Then the estimates are valid

$$\begin{gathered} \max_i \beta_{d-ii}, \beta_{d+ii} = \beta_{ii} \max, \\ J \leqslant \frac{\beta_{ii}\mathrm{max}}{2\min\limits_i |Re\ \lambda_i|} n = \frac{\beta_{ii}\mathrm{max}}{\left(\frac{2\min\limits_i |Re\ \lambda_i|}{n}\right)}. \end{gathered}$$

Thus, the trace of a mixed gramian is directly proportional to the maximum value of the diagonal element of the matrix $\left[B_-B_-^{\mathrm{T}} \oplus B_+B_+^{\mathrm{T}}\right]$ and is inversely proportional to the doubled average value of the modulus of the eigenvalue of the spectrum of the matrix $[\Lambda_- \oplus (-\Lambda_+)]$, which confirms the research results of [28].

Illustrative example. Consider the problem of controlling a dynamic object with four inputs and four outputs. The model of the control object can be described by equations of state of the form

$$\Sigma_1\colon \begin{cases} \dfrac{dx}{dt} = Ax\left(t\right) + Bu\left(t\right), \quad x\left(0\right) = 0, \\ y\left(t\right) = Cx\left(t\right). \end{cases}$$

$$A = \begin{bmatrix} -0{,}33 & -2{,}67 & -4 & 1{,}33 \\ 21{,}17 & -23{,}33 & -30{,}2 & 1{,}5 \\ -14{,}67 & 14 & 17{,}83 & -1{,}17 \\ 2 & -1{,}33 & -1{,}83 & -2{,}17 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 2 \\ 5 \\ -3 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Let us transform the system into the upper block-diagonal Schur form. In this case, the unitary transformation matrix will be expressed as follows:

$$U = \begin{bmatrix} 0{,}125 & 0{,}943 & -0{,}169 & -0{,}258 \\ 0{,}814 & -0{,}26 & -0{,}056 & -0{,}516 \\ -0{,}564 & -0{,}178 & -0{,}225 & -0{,}775 \\ 0{,}063 & -0{,}109 & -0{,}958 & 0{,}258 \end{bmatrix}.$$

The system will take the form

$$A_{Sch} = \begin{bmatrix} 1 & 37{,}64 & 3{,}255 & 35{,}17 \\ 0 & -4 & -0{,}97 & -0{,}212 \\ 0 & 0 & -2 & 0{,}436 \\ 0 & 0 & 0 & -3 \end{bmatrix}, \quad B_{Sch} = \begin{bmatrix} -1{,}25 \\ -0{,}137 \\ 1{,}465 \\ -5{,}939 \end{bmatrix}.$$

The next transformation occurs in such a way that the matrix A_{Sch12} becomes zero. We select the transformation matrix W_{bl} so that the matrix A_{bl} is divided into two blocks, stable and antistable subsystems.

$$W_{bl} = \begin{bmatrix} 1 & -7{,}53 & 1{,}35 & -8{,}25 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad A_{bl} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -4 & -0{,}97 & -0{,}21 \\ 0 & 0 & -2 & 0{,}436 \\ 0 & 0 & 0 & -3 \end{bmatrix},$$

$$B_{bl} = \begin{bmatrix} -53{,}2 \\ -0{,}14 \\ 1{,}47 \\ -5{,}94 \end{bmatrix}.$$

Let's check the execution of Sylvester's equation (2.5). We transpose all components of the equation

$$\begin{bmatrix} -7{,}529 \\ 1{,}35 \\ -8{,}25 \end{bmatrix} + \begin{bmatrix} -4 & 0 & 0 \\ -0{,}97 & -2 & 0 \\ -0{,}212 & 0{,}436 & -3 \end{bmatrix} \times \begin{bmatrix} 7{,}529 \\ -1{,}35 \\ 8{,}25 \end{bmatrix} + \begin{bmatrix} 37{,}64 \\ 3{,}255 \\ 35{,}167 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

For the system in this case, the mixed gramian is given by the equation (2.11)

$$P_{cm} = T_2^{-1} \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} T_2.$$

$$P_{cm} = \begin{bmatrix} 5{,}32 & -5{,}32 & -7{,}98 & 2{,}66 \\ 0{,}94 & -0{,}26 & -0{,}18 & -0{,}11 \\ -0{,}17 & -0{,}056 & -0{,}23 & -0{,}96 \\ -0{,}26 & -0{,}52 & -0{,}78 & 0{,}26 \end{bmatrix} \times \begin{bmatrix} 1417 & 0 & 0 & 0 \\ 0 & 0{,}0067 & -0{,}057 & 0{,}18 \\ 0 & -0{,}057 & 0{,}52 & -1{,}72 \\ 0 & 0{,}18 & -1{,}72 & 5{,}88 \end{bmatrix} \times$$

$$\times \begin{bmatrix} 0{,}13 & 0 & 0 & -1{,}29 \\ 0{,}81 & -6{,}39 & 1{,}04 & -7{,}23 \\ -0{,}56 & 4{,}07 & -0{,}99 & 3{,}87 \\ 0{,}063 & -0{,}58 & -0{,}87 & -0{,}26 \end{bmatrix} = \begin{bmatrix} 32 & -203 & -132{,}5 & 11{,}3 \\ -203 & 1290 & -844 & 73 \\ -132{,}5 & -844 & -349 & -50{,}5 \\ 11{,}3 & 73 & -50{,}5 & 5{,}57 \end{bmatrix}.$$

Let's check the correctness of the gramian calculation. The matrix of the third transformation and the system itself will take the form

$$W_d = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -0{,}44 & 0{,}19 \\ 0 & 0 & 0{,}9 & -0{,}39 \\ 0 & 0 & 0 & 0{,}9 \end{bmatrix}, \quad A_d = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -4 & 0 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -3 \end{bmatrix}, \quad B_d = \begin{bmatrix} -53{,}2 \\ 0{,}57 \\ -1{,}25 \\ -6{,}6 \end{bmatrix}.$$

Then the gramian for the diagonalized system becomes equal to

$$[P_- \oplus P_+] = \begin{bmatrix} 1417 & 0 & 0 & 0 \\ 0 & 0{,}04 & -0{,}12 & -0{,}54 \\ 0 & -0{,}12 & 0{,}39 & 1{,}65 \\ 0 & -0{,}54 & 1{,}65 & 7{,}26 \end{bmatrix}.$$

The general expression of the mixed gramian after the third transformation will be written as follows:

$$P_{cm} = \begin{bmatrix} 5{,}32 & -5{,}32 & -7{,}98 & 2{,}66 \\ 0{,}86 & -0{,}29 & -0{,}29 & -0{,}57 \\ -0{,}31 & -0{,}31 & -0{,}63 & -0{,}94 \\ -0{,}29 & -0{,}57 & -0{,}86 & 0{,}29 \end{bmatrix} \times \begin{bmatrix} 1417 & 0 & 0 & 0 \\ 0 & 0{,}041 & -0{,}12 & -0{,}54 \\ 0 & -0{,}12 & 0{,}39 & 1{,}65 \\ 0 & -0{,}54 & 1{,}65 & 7{,}26 \end{bmatrix} \times$$

$$\times \begin{bmatrix} 0{,}125 & 0 & 0 & -1{,}16 \\ 0{,}81 & -6{,}39 & 3{,}73 & -8{,}13 \\ -0{,}56 & 4{,}07 & -2{,}66 & 4{,}65 \\ 0{,}063 & -0{,}58 & -0{,}53 & 0 \end{bmatrix} = \begin{bmatrix} 32 & -203 & -132{,}5 & 11{,}3 \\ -203 & 1290 & -844 & 73 \\ -132{,}5 & -844 & -349 & -50{,}5 \\ 11{,}3 & 73 & -50{,}5 & 5{,}57 \end{bmatrix}.$$

The mixed gramians coincided. Let us check whether the Sylvester criterion is satisfied for the gramian of stable and antistable systems. To do this, the matrices P_1 and P_2 must be positive definite. For compactness, we write them into one matrix.

$$[P_1 \oplus P_2] = \begin{bmatrix} 1417 & 0 & 0 & 0 \\ 0 & 0{,}0067 & -0{,}057 & 0{,}18 \\ 0 & -0{,}057 & 0{,}52 & -1{,}72 \\ 0 & 0{,}18 & -1{,}72 & 5{,}88 \end{bmatrix}, \quad \lambda_{P_1} = 1417, \quad \lambda_{P_2} = \begin{bmatrix} 0{,}0001 \\ 0{,}018 \\ 6{,}39 \end{bmatrix}.$$

All eigenvalues are greater than zero. The criterion is met. Let's calculate the trace using the formula (2.21)

$$J = \sum_{i=1}^{r} \frac{\beta_{d-ii}}{-2Re\ \lambda_i} + \sum_{i=r+1}^{n} \frac{\beta_{d+ii}}{2Re\ \lambda_i} = 0{,}0067 + 0{,}52 + 5{,}88 + 1417 \approx 1423.$$

Let us compare the value of the spectrum trace with the estimate

$$J = 1423 \leqslant \frac{2834}{\frac{2*1}{4}} = 5668.$$

The reciprocal of the average value of the modules of the eigenvalues of the dynamics matrix estimates the degree of dispersion of the real parts of the eigenvalues relative to the imaginary axis. The smaller this value is, the higher its influence on the trace of the mixed controllability gramian. The formula for the spectral decomposition of the trace allows us to perform a more refined analysis of the influence of the distribution of eigenvalues on the energy metric of the degree of reachability [25, 27].

3. SPECTRAL EXPANSIONS OF ENERGY METRICS OF CONTROLLABILITY AND OBSERVABILITY GRAMIANS

We consider the application of the obtained results to solve some problems of state estimation and control. We obtain spectral decompositions of energy metrics.

Theorem 3 [8]. *Let us consider a finite-dimensional linear stationary continuous system with many inputs and many outputs of the form* (2.1)*, reduced to the diagonal form* (2.6)*. Let us assume that the system has a simple spectrum, the system is completely controllable and unstable, and the eigenvalues of its dynamics matrix A are not on the imaginary axis, but can be in the left and/or right half-planes*

$$\lambda_{i-} \in \mathbb{C}^-, \;\; i = r; \;\; \lambda_{i+} \in \mathbb{C}^+, \;\; i = n - r.$$

In addition, we assume that the condition is satisfied

$$\lambda_i \neq -\lambda_j, \; \forall i, j : i = \overline{1, n}, \; j = \overline{1, n}.$$

The following spectral expansions of energy functionals are valid and equivalent [18]*:*

$$J_1 = E_{\min}(\infty) = \begin{bmatrix} x_{f-} & x_{f+} \end{bmatrix}^{\mathrm{T}} (P_{cm})^{-1} \begin{bmatrix} x_{f-} & x_{f+} \end{bmatrix} =$$
$$= \begin{bmatrix} x_{f-} & x_{f+} \end{bmatrix}^{\mathrm{T}} \left[\sum_{i=1}^{n} V_c^* |\sigma_i|^{-1} \mathbb{1}_{ii} U_c \right] \begin{bmatrix} x_{f-} & x_{f+} \end{bmatrix}.$$

$$J_3 \; (\textit{for SISO LTI stable systems}) = \operatorname{tr} \sum_{k=1}^{n} P_{c,k} = \sum_{k=1}^{n} \operatorname{tr} P_{c,k} =$$
$$\left(\frac{1}{\sum_{k=1}^{n} \dot{N}(s_k) N(-s_k)} - \frac{\sum_{k=1}^{n} s_k^2}{\sum_{k=1}^{n} \dot{N}(s_k) N(-s_k)} + \cdots \right.$$
$$\left. \cdots + \frac{(-1)^{n-1} \sum_{k=1}^{n} s_k^{2n}}{\sum_{k=1}^{n} \dot{N}(s_k) N(-s_k)} \right),$$

$$J_4 = \operatorname{tr} \sum_{i=1}^{n} (P_c)_i^{-1} = \sum_{i=1}^{n} \operatorname{tr}\, (P_c)_i^{-1} = \left[\sum_{i=1}^{n} \operatorname{tr} \left[V_c^* |\sigma_i|^{-1} \mathbb{1}_{ii} U_c \right] \right],$$

where $N(s)$ is characteristic polynomial of the system(2.1).

Proof of Theorem. Let us return to stable continuous MIMO LTI systems with a simple spectrum and note that the controllability and observability gramians are symmetric complex-valued matrices. In this case, there are their singular decompositions of the form [1]

$$P_c = V_c \Lambda V_c^*,$$

where the matrix V_c is formed by the right singular vectors of the matrix P_c, and the matrix Λ is a diagonal matrix of the form

$$\Lambda = \operatorname{diag}\left\{|\sigma_1|\,|\sigma_2|\ldots|\sigma_{\mathrm{n}}|\right\}.$$

We define matrices S and U in the form

$$S = \operatorname{diag}\left\{\operatorname{sgn}\sigma_1 \;\; \operatorname{sgn}\sigma_2 \;\ldots\; \operatorname{sgn}\sigma_{\mathrm{n}}\;\right\},\; U_c = V_c S,$$
$$\operatorname{sgn}\sigma = \begin{cases} +1, & \text{if } \sigma \geqslant 0 \\ -1, & \text{if } \sigma < 0. \end{cases}$$

Then

$$P_c = U_c \Lambda V_c^*, \tag{3.1}$$

where the matrix U_c is formed by the left singular vectors of the matrix P_c. Since Λ, U_c, V_c are nonsingular matrices, then

$$(P_c)^{-1} = (U_c)^{-1}\Lambda^{-1}(V_c^*)^{-1} = V_c^*\Lambda^{-1}U_c. \tag{3.2}$$

Since the matrix Λ is diagonal, its inverse matrix can be represented as

$$\Lambda^{-1} = \left[|\sigma_1|^{-1}\mathbb{1}_{11} + |\sigma_2|^{-1}\mathbb{1}_{22} + \cdots + |\sigma_n|^{-1}\mathbb{1}_{nn}\right]. \tag{3.3}$$

Substituting (3.3) into (3.1), (3.2), we obtain the following spectral expansions of the inverse controllability gramians in a simple spectrum:

$$(P_c)^{-1} = (P_c)_1^{-1} + (P_c)_2^{-1} + \cdots + (P_c)_n^{-1},$$
$$(P_c)_1^{-1} = V_c^*|\sigma_1|^{-1}\mathbb{1}_{11}U_c,\; (P_c)_2^{-1} = V_c^*|\sigma_2|^{-1}\mathbb{1}_{22}U_c,\; \ldots,\; (P_c)_n^{-1} = V_c^*|\sigma_n|^{-1}\mathbb{1}_{nn}U_c.$$

This implies the following spectral expansions of energy functionals [11]:

$$J_1 = E_{\min}(\infty) = \begin{bmatrix} x_{f-} & x_{f+} \end{bmatrix}^{\mathrm{T}} (P_c)^{-1} \begin{bmatrix} x_{f-} & x_{f+} \end{bmatrix} =$$
$$= \begin{bmatrix} x_{f-} & x_{f+} \end{bmatrix}^{\mathrm{T}} \left[\sum_{i=1}^{n} V_c^*|\sigma_i|^{-1}1_{ii}U_c\right] \begin{bmatrix} x_{f-} & x_{f+} \end{bmatrix},$$
$$J_2 = \operatorname{tr}\sum_{i=1}^{n}(P_c)_i^{-1} = \sum_{i=1}^{n}\operatorname{tr}\,(P_c)_i^{-1} = \left[\sum_{i=1}^{n}\operatorname{tr}\left[V_c^*|\sigma_i|^{-1}\mathbb{1}_{ii}U_c\right]\right],$$
$$J_3 \text{ (for SISO LTI systems)} = \operatorname{tr}\sum_{k=1}^{n}P_{c,k} = \sum_{k=1}^{n}\operatorname{tr}P_{c,k} =$$
$$\left(\frac{1}{\sum_{k=1}^{n}\dot{N}(s_k)\,N(-s_k)} - \frac{\sum_{k=1}^{n}s_k^2}{\sum_{k=1}^{n}\dot{N}(s_k)\,N(-s_k)} + \ldots \right.$$
$$\left. \ldots + (-1)^{n-1}\frac{\sum_{k=1}^{n}s_k^{2n}}{\sum_{k=1}^{n}\dot{N}(s_k)\,N(-s_k)}\right),$$
$$J_5 = \operatorname{tr}\,(P_{cm}).$$

Theorem 3 is proven.

Theorem 4 [2]. *Let us consider a finite-dimensional linear stationary continuous system with many inputs and many outputs of the general form* (2.1). *Let us assume that the system has a simple spectrum, is completely controllable and stable. Then the following spectral expansions of the energy functionals of the input and output energies $\hat{J}_1$ and $\hat{J}_2$ are valid and equivalent over the simple spectrum of the controllability gramian:*

$$\hat{J}_1 = \sum_{i=1}^{n} x_0^T \left[V_c^* |\sigma_i|^{-1} 1_{ii} U_c \right] x_0, \tag{3.4}$$

or a simple spectrum of the dynamics matrix A:

$$\hat{J}_2 = \sum_{i=1}^{n} x_0^T \left[\sum_{j=0}^{n-1} \sum_{\eta=0}^{n-1} \frac{\lambda_i^j (-\lambda_i)^\eta}{\dot{N}(\lambda_i) N(-\lambda_i)} A_j^{\mathrm{T}} C^{\mathrm{T}} C A_\eta \right] x_0. \tag{3.5}$$

Proof of Theorem. It was proven in [2] that the energy functionals of the input and output energies $\hat{J}_1$ and $\hat{J}_2$ are equal

$$\hat{J}_1 = \inf_{u,x} \int_{-\infty}^{0} \|u(t)\|^2 dt, \quad \hat{J}_2 = \int_{0}^{\infty} \|y(t), 0, \ x_0\|^2 dt.$$

Under the conditions of the theorem, they can be represented in the form of quadratic forms

$$\hat{J}_1 = E_c(x_0) = x_0^{\mathrm{T}} P_c^{\#} x_0, \tag{3.6}$$

$$\hat{J}_2 = E_o(x_0) = x_0^{\mathrm{T}} P_o x_0, \tag{3.7}$$

where $P_c^{\#}$ is the Moore-Penrose pseudo-inversion of the gramian controllability matrix, and P_o is the gramian observability matrix. Under the conditions of the theorem, the gramian controllability matrix is a non-singular matrix, therefore the equality

$$P_c^{\#} = P_c^{-1}.$$

Substituting the spectral decomposition of the inverse gramian matrix into the formula (3.6), we obtain the desired spectral decomposition of the input energy functional. In [11], a spectral decomposition of the observability gramian of system was obtained in the form of Xiao Hankel matrices [11, 22, 23]

$$P_o = \sum_{i=1}^{n} \sum_{j=0}^{n-1} \sum_{\eta=0}^{n-1} \frac{\lambda_i^j (-\lambda_i)^\eta}{\dot{N}(\lambda_i) N(-\lambda_i)} A_j^{\mathrm{T}} C^{\mathrm{T}} C A_\eta.$$

Substituting the spectral decomposition of the gramian matrix P_o (3.7), we obtain the desired spectral decomposition of the output energy functional. Theorem 4 is proven.

The functionals $\hat{J}_1$ and $\hat{J}_2$ were used in [10] to analyze the degree of stability of a linear system based on the analysis of anomalies of the square H_2 is the norm of the transfer function of the system, caused by the influence of the following weakly stable modes:

- modes close to the origin of coordinates,
- modes close to the imaginary axis,
- several aperiodic and oscillatory modes close to each other.

As the main tool for anomaly analysis, it was proposed to use asymptotic models of spectral expansions of the functionals J_1 and J_2 over the simple and/or pair spectrum of the system dynamics matrix. A similar approach can be extended to the analysis of anomalies in the spectral decompositions of the metrics of the traces of the gramians J_3 and J_4, as well as to the analysis of the degree of reachability of a linear system based on the anomalies of the spectral decompositions of the metrics of the mixed gramians J_5. Note that the spectral decompositions of the metrics depend on the eigenvalues of the dynamics matrix, which are tied to a specific node in the system graph, which makes it possible to associate the problem of optimal placement of sensors and actuators with certain nodes in the system graph.

4. CONCLUSION

The article generalizes the known results of gramian decomposition for unstable continuous linear systems to calculate their spectral decompositions of the simplest case of decompositions over the pair spectrum of the dynamics matrix. Most energy metrics associated with the use of gramians are based on calculating the spectrum of dynamics matrices and measures of the minimum energy required for the system to transition from the initial to the final point. The paper shows that spectral decompositions of controllability gramians and their inverse gramians make it possible to calculate the energy components corresponding to the characteristic eigenvalues of the gramian matrices, which determine the main contribution to the value of the reachability metric and the energy metric of stability. These spectral decompositions are presented in the form of formulas that allow one to analyze the influence of various nodes of the system graph on the formation of energy metrics of reachability and stability. The results obtained can find application in problems of localization and optimal placement of sensors and actuators on the graph of a complex multiconnected control system or in problems of placement of control nodes in the graph of a complex social, transport, energy or biological network [25].

FUNDING

This work was carried out with financial support from the Russian Science Foundation, project no. 19-19-00673.

REFERENCES

1. Antoulas, A.C., *Approximation of Large-Scale Dynamical Systems*, Philadelphia: SIAM, 2005.
2. Benner, P. and Damm, T., Lyapunov equations, Energy Functionals and Model Order Reduction of Bilinear and Stochastic Systems, *SIAM J. Control Optim.*, 2011, vol. 49, pp. 686–711.
3. Polyak, B.T., Khlebnikov, M.V., and Rapoport, L.B., *Teoriya avtomaticheskogo upravleniya* (Theory of Automatic Control), Moscow: LENAND, 2019.
4. Kailath, T., *Linear Systems Englewood Cliffs*, New Jersey: Prentice Hall, 1980.
5. Zubov, N.E., Zybin, E.Yu., Mikrin, E.A., Misrikhanov, M.Sh., and Ryabchenko, V.N., General analytical forms for solving the Sylvester and Lyapunov equations for continuous and discrete dynamic systems, *Theory and control systems*, 2017, no. 1, pp. 3–20.
6. Paraev, Yu.I. and Perepelkin, E.A., *Lineinye matrichnye uravneniya v zadachakh analiza mnogosvyaznykh dinamicheskikh sistem* (Linear matrix equations in problems of analysis of multiply connected dynamic systems), Barnaul: Altaisk. GTU, 2000.
7. Bukov, V.N., *Vlozhenie sistem. Analiticheskii podkhod k analizu i sintezu matrichnykh sistem* (Nesting systems. Analytical approach to the analysis and synthesis of matrix systems), Kaluga: Izd. Nauch. Lit. Bochkarevoi, 2006.
8. Godunov, S.K., Modern Aspects of Linear Algebra, *Trans. of Math. Monografs. V. 175*, Providence RI: Amer. Math. Soc., 1998.
9. Yadykin, I.B. Galyaev, A.A., On the methods for calculation of grammians and their use in analysis of linear dynamic systems, *Autom. Remote Control*, 2013, vol. 74, no. 2, pp. 207–224.

10. Yadykin, I.B. and Iskakov, A.B., Energy Approach to Stability Analysis of the Linear Stationary Dynamic Systems, *Autom. Remote Control*, 2016, no. 12, pp. 37–58.

11. Yadykin, I.B., Spectral Decompositions of Gramians of Continuous Stationary Systems Given by Equations of State in Canonical Forms, *Mathematics*, 2022, vol. 10, no. 13, 2339.

12. Casadei, G., Wit, C., and Zampieri, S., Model Reduction Based Approximation of the Output Controllability Gramian in Large-Scale Networks, *IEEE Transactions on Control of Network Systems*, 2020, vol. 7, no. 4, pp. 1778–1788.

13. Zhou, K., Salomon, G., and Wu, E., Balanced realization and model reduction for unstable systems, *International Journal of Robust and Nonlinear Control*, 1999, vol. 9, no. 3, pp. 183–198.

14. Lee, H. and Park, Y., Degree of controllability for linear unstable systems, *Journal of Vibration and Control*, 2016, vol. 22, no. 7, pp. 1928–1934. https://doi.org/10.1177/1077546314545101

15. Shaker, H. and Tahavori, M., Optimal sensor and actuator location for unstable systems, *Journal of Vibration and Control*, 2013, vol. 19, no. 12, pp. 1915–1920. https://doi.org/10.1177/1077546312451302

16. Wal, M. and Jager, B., A review of methods for input/output selection, *Automatica*, 2001, vol. 37, no. 4, pp. 487–510. https://doi.org/10.1016/S0005-1098(00)00181-3

17. Birk, W. and Medvedev, A., A note on gramian-based interaction measures, *European Control Conference (ECC). Cambridge, UK*, 2003, pp. 2625–2630. https://doi.org/10.23919/ECC.2003.7086437.

18. Mehr, F., *A Determination of Design of Optimal Actuator Location Based on Control Energy*, London: University of London, 2018.

19. Petrov, B.N., *Izbrannye trudy, I. Teoriya avtomaticheskogo upravleniya* (Selected works T. 1. Theory of automatic control), Moscow: Nauka, 1983. P. 432 (163–178, 223–227 (double), 294–323).

20. Petrov, B.N., Rutkovsky, V.Yu., and Zemlyakov, S.D., *Adaptivnoe koordinatno-parametricheskoe upravlenie nestatsionarnymi ob"ektami* (Adaptive coordinate-parametric control of non-stationary objects), Moscow: Nauka, 1980.

21. Xiao, C.S., Feng, Z.M., and Shan, X.M., On the Solution of the Continuous-Time Lyapunov Matrix Equation in Two Canonical Forms, *IEE Proc.*, 1992, vol. 139, no. 3, pp. 286–290. https://doi.org/10.1049/ip-d.1992.0038

22. Hauksdottir, A. and Sigurdsson, S., The continuous closed form controllability Gramian and its inverse, *2009 American Control Conference Hyatt Regency Riverfront, St. Louis, MO, USA June 10–12*, 2009, pp. 5345–5351. https://doi.org/978-1-4244-4524-0/09

23. Hsu, C. and Hou, D., Reducing Unstable Linear Control Systems via Real Schur Transformation, *Electronics Letters*, 1991, vol. 27, no. 11. https://doi.org/10.1049/el:19910614

24. Safonov, M. and Chiang, G., A schur method for balanced-truncation model reduction, *IEEE Trans. Autom. Control*, 1989, vol. 34, no. 7, pp. 729–733. https://doi.org/10.1109/9.29399

25. Lindmark, G. and Altafini, C., Minimum energy control for complex networks, *Scientific Reports*, 2018, vol. 8, no. 3188, pp. 1–14. https://doi.org/10.1038/s41598-018-21398-7

26. Dilip, A., The Controllability Gramian, the Hadamard Product, and the Optimal Actuator/Leader and Sensor Selection Problem, *Nature Physics*, 2019, vol. 3, no. 4, pp. 883–888. https://doi.org/10.1109/LCSYS.2019.2919278

27. Pasqualetti, F., Zampieri, S., and Bullo, F., Controllability metrics, limitations and algorithms for complex networks, *IEEE Transactions on Control of Network Systems*, 2014, vol. 1, no. 1, pp. 40–52. https://doi.org/10.1109/ACC.2014.6858621

28. Hac, A. and Liu, L., Sensor and actuator location in motion control of flexible structures, *Journal of Sound and Vibration*, 1993, vol. 167, no. 2, pp. 239–261. https://doi.org/10.1006/jsvi.1993.1333

29. Faddeev, D.K. and Faddeeva, V.N., *Computational Methods of Linear Algebra*, San Francisco: Freeman, 2016.

30. Hanson, B. and Peeters, R., A Faddeev Sequence Method for solving Lyapunov and Sylvester Equations, *Linear Algebra and its Applications*, 1996, vol. 241–243, pp. 401–430.

31. Nagar, S. and Singh, S., An algorithmic approach for system decomposition and balanced realized model reduction, *Journal of the Franklin Institute*, 2004, vol. 341, no. 7, pp. 615–630. https://doi.org/10.1016/j.jfranklin.2004.07.005

32. Robust Control Tool Box, Mathworks, Version 2, 1997.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2023, Vol. 84, No. 10, pp. 1259–1274.

Numerical Methods for Robust Performance Analysis of Linear Discrete-Time Polytopic Systems with Respect to Random Disturbances

A. A. Belov

Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
e-mail: a.a.belov@inbox.ru

Received May 16, 2023
Revised July 19, 2023
Accepted August 31, 2023

Abstract—Discrete linear polytopic systems affected by random correlated stationary disturbances are considered. New numerical methods for estimating of the anisotropic norm of a polytopic system using linear matrix inequalities are proposed.

Keywords: polytopic uncertainties, LMI, mean anisotropy, random processes, analysis

DOI: 10.25728/arcRAS.2023.65.73.001

1. INTRODUCTION

Mathematical models of control systems are designed on the basis of known physical laws, as well as measurable parameters of the plant. Technological tolerances and measurement errors in the control system may lead to a mismatch between the mathematical model and the real plant. In some cases, this mismatch is significant and may lead to loss of system performance, and to the loss of stability of the closed-loop system. Thus, analysis and control problems subject to inexact knowledge about the parameters of mathematical models, called robust analysis and control problems, arise.

Depending on the initial assumptions regarding the type of uncertainties of the control system, there are various approaches to the analysis of its robust properties. One of the popular means of describing uncertainties in linear systems is polytopic uncertainty. This uncertainty is characterized by the fact that the unknown parameters of the system lie on the given simplex. If a system with polytopic uncertainty is lienar time invariant one, then in the literature it is called a polytopic system. For discrete-time linear systems, there are many methods for checking robust stability [1–4]. Papers [1–3] are devoted to the study of the stability of systems with polytopic time-invariant and time varying uncertainties using parametric Lyapunov functions. In [4] the results of robust analysis of polytopic systems using of linear matrix inequalities are presented. The results are given in terms of nonparametric matrix inequalities.

Along with the problems of studying robust stability of uncertain systems, one of the important aspects analysis of control systems is the ability to suppress external disturbances. Thus, in the literature it is known methods for analyzing the quality of suppression of external disturbances in terms of $\mathcal{H}_2$- and $\mathcal{H}_\infty$-norms [5]. Assuming that correlated random disturbances act at the input of the system, an anisotropy-based approach can be used to analyze the quality of its suppression by the system [6–8]. Feature of anisotropy-based approach is to study the quality of system performance under impact of correlated stationary random disturbances with a known mean anisotropy level. Methods of anisotropy-based analysis and control of polytopic systems were studied in [9–11].

In [9] parametric version of the anisotropy-based bounded real lemma, one of the results of non-parametric numerical analysis anisotropy-based performance analysis was obtained in [10], [11] is devoted to solving problem of anisotropy-based state-feedback control design with closed-loop pole placement.

This paper proposes numerical methods for solving the problem of anisotropy-based analysis for polytopic systems using linear matrix inequalities. All these methods are derived from parametric anisotropy-based bounded real lemma. The degree of conservatism of the obtained conditions is analyzed, and also estimates of their computational complexity are given.

2. PROBLEM STATEMENT

Consider linear system with state space representation as

$$x(k+1) = A(\Theta)x(k) + B_w(\Theta)w(k), \tag{1}$$
$$y(k) = C(\Theta)x(k) + D_w(\Theta)w(k), \tag{2}$$

where $x(k) \in \mathbb{R}^n$ is a state, $w(k) \in \mathbb{R}^m$ is external random disturbance with zero mean and bounded mean anisotropy level $\overline{\mathbf{A}}(W) \leqslant a$ $(a \geqslant 0)$, $y(k) \in \mathbb{R}^p$ is output.

Matrices $A(\Theta)$, $B_w(\Theta)$, $C(\Theta)$, $D_w(\Theta)$ are defined from the expressions

$$\begin{gathered} A(\Theta) = \sum_{i=1}^{r} \theta_i A_i, \quad B_w(\Theta) = \sum_{i=1}^{r} \theta_i B_{wi}, \\ C(\Theta) = \sum_{i=1}^{r} \theta_i C_i, \quad D_w(\Theta) = \sum_{i=1}^{r} \theta_i D_{wi}, \end{gathered} \tag{3}$$

with known constant matrices A_i, B_{wi}, C_i, D_{wi} of appropriate dimensions and vector Θ of unknown parameters which satisfies relations

$$\sum_{i=1}^{r} \theta_i = 1, \quad \theta_i \geqslant 0, \quad \theta_i \in \mathbb{R}, \quad \forall i = \overline{1, r}. \tag{4}$$

Mean anisotropy characterizes a measure of difference between Gaussian random sequence and white Gaussian noise with zero mean and identity covariance (we call it standard) in terms of relative entropy and is calculated using the formula

$$\overline{\mathbf{A}}(W) = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \ln\det \frac{m S_w(\omega)}{\frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{Trace} S_w(\lambda) d\lambda} d\omega, \tag{5}$$

where $S_w(\omega)$ is a spectral density of sequence $W = \{w(k)\}_{k\in\mathbb{Z}}$.

Thus, parameter $a \geqslant 0$ defines the set of all Gaussian signals whose measure of difference from standard Gaussian noise defined by expression (5), does no exceed value of a. It should be noted that mean anisotropy functional is nonnegative and goes to zero if W is standard Gaussian noise [8].

Denote the set of all parameters Θ, satisfying (3) and (4), by $\mathfrak{Q}$ and consider the mapping $Y = F_\Theta W$, defined by expressions (1)–(2).

Definition 1. Anisotropic norm of polytopic system (1)–(4) is norm of operator F_Θ, defined by expression

$$|\!|\!| F_\Theta |\!|\!|_a = \sup_{\Theta \in \mathfrak{Q}} \ \sup_{W:\, A(W) \leqslant a} \frac{\|Y\|_{\mathcal{P}}}{\|W\|_{\mathcal{P}}}, \tag{6}$$

where

$$\|W\|_{\mathcal{P}} = \sqrt{\lim_{N\to\infty} \frac{1}{2N+1} \sum_{k=-N}^{N} \mathbf{E}|w(k)|^2}$$

is power norm of signal W.

One of the most important features of anisotropic norm is that it lies between scaled $\mathcal{H}_2$-norm and $\mathcal{H}_\infty$-norm, i.e. [8]

$$\frac{\|F_\Theta\|_2^2}{m} \leqslant |\!|\!|F_\Theta|\!|\!|_a^2 \leqslant \|F_\Theta\|_\infty^2.$$

It mean that by varying value of mean anisotropy a from 0 to ∞, one can reduce or expand the set of random signals, selecting the most favorable bandwidth and stability margins of the system in the range between $\mathcal{H}_2$- and $\mathcal{H}_\infty$-norms.

In problem of robust anisotropy-based analysis of polytopic systems it's necessary to obtain conditions for checking robust stability and anisotropic norm bounds of open-loop system (1)–(2) for known mean anisotropy level $a \geqslant 0$ and given scalar $\gamma > 0$. Thus, the problem is formulated as follows.

Problem 1. For known mean anisotropy level $a \geqslant 0$ of random external disturbance $w(k)$ and given scalar $\gamma > 0$ the problem is to check:

1) if the system robustly stable;

2) if the condition holds

$$|\!|\!|F_\Theta|\!|\!|_a < \gamma.$$

Known results which are necessary for the further exposition are listed below. Let us consider the system with known parameters, for which all of the vectors and matrices dimensions coincide with ones in system (1)–(2):

$$x(k+1) = Ax(k) + B_w w(k), \tag{7}$$

$$y(k) = Cx(k) + D_w w(k). \tag{8}$$

Now provide formulation of anisotropy-based bounded real lemma in terms of LMI [13].

Lemma 1. *System* (7)–(8) *is stable and its anisotropic norm for given mean anisotropy level of external disturbance* $a \geqslant 0$ *is bounded by scalar* $\gamma > 0$, *if there exist such matrices* $X > 0$, $Y > 0$, $\Phi > 0$, *and scalar* $\mu > \gamma^2$, *for which the following relations hold true:*

$$\mu - \left(\mathrm{e}^{-2a} \det \Phi\right)^{1/q} < \gamma^2, \tag{9}$$

$$\begin{bmatrix} \Phi - \mu I_m & \star & \star \\ B_w & -Y & \star \\ D_w & 0 & -I_p \end{bmatrix} < 0, \tag{10}$$

$$\begin{bmatrix} -X & \star & \star & \star \\ 0 & -\mu I_m & \star & \star \\ A & B_w & -Y & \star \\ C & D_w & 0 & -I_p \end{bmatrix} < 0, \tag{11}$$

$$XY = I_n. \tag{12}$$

3. PROBLEM SOLUTION

3.1. Parametric Anisotropy-Based Bounded Real Lemma

Let us formulate parametric conditions for anisotropy-based analysis of a polytopic system (1)–(2), on the basis of which the main results of this paper will be obtained.

Theorem 1. *System* (1)–(2) *is robustly stable and its anisotropic norm does not exceed given scalar value $\gamma > 0$ for known mean anisotropy level $a \geqslant 0$ if there exist such matrices $P(\Theta) > 0$, $\Psi(\Theta) > 0$, nonsingular matrices $G_1(\Theta)$, $G_2(\Theta)$ and scalar $\eta > \gamma^2$, such that the following inequalities hold true*

$$\eta - \left(\mathrm{e}^{-2a} \det \Psi(\Theta)\right)^{1/m} < \gamma^2, \tag{13}$$

$$\begin{bmatrix} \Psi(\Theta) - \eta I_m & \star & \star \\ G_1(\Theta)B_w(\Theta) & L_1(\Theta) & \star \\ D_w(\Theta) & 0 & -I_p \end{bmatrix} < 0, \tag{14}$$

$$\begin{bmatrix} -P(\Theta) & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ G_2(\Theta)A(\Theta) & G_2(\Theta)B_w(\Theta) & L_2(\Theta) & \star \\ C(\Theta) & D_w(\Theta) & 0 & -I_p \end{bmatrix} < 0, \tag{15}$$

where $L_1(\Theta) = -G_1(\Theta) - G_1^{\mathrm{T}}(\Theta) + P(\Theta)$ and $L_2(\Theta) = -G_2(\Theta) - G_2^{\mathrm{T}}(\Theta) + P(\Theta)$, for each $\Theta \in \mathfrak{Q}$.

The proof of the theorem is listed in Appendix.

Conditions of Theorem 1 depend on parameters Θ explicitly. Existence of any parametric matrices $\Psi(\Theta)$, $P(\Theta)$, $G_1(\Theta)$, and $G_2(\Theta)$, which satisfy all the conditions of the Theorem 1, would allow to check robust stability of the system and establish the fact that its anisotropic norm is bounded by γ for known mean anisotropy level $a \geqslant 0$ of input disturbance W. There is currently no formal method for determining the exact type matrices $P(\Theta)$ as a function of the parameter vector Θ. In scientific literature such function $P(\Theta)$ is called parametric Lyapunov matrix [1, 2, 4]. Similar statement holds for the rest parametric matrices. Unfortunately, such parametric dependence may substantially complicate analysis of initial plant. It is possible to reduce numerical complexity of the algorithm by introducing supplementary restrictions, for example, by using different approximations of matrices $\Psi(\Theta)$, $P(\Theta)$, $G_1(\Theta)$ and $G_2(\Theta)$. On the one hand, this approach allows to get rid of explicit appearance of parameter vector Θ, on the other hand, it bring some conservatism. Below we present several methods of nonparametric anisotropy-based analysis of the polytopic system (1)–(2) depending on various approximations.

3.2. Nonparametric Variations of Anisotropy-Based Bounded Real Lemma

Let $\Psi(\Theta) = \Psi$, $G_1(\Theta) = G_1$, $G_2(\Theta) = G_2$, $P(\Theta) = P$. Then parameters θ_i can be factorized in expressions (14)–(15). The following result is obtained directly.

Theorem 2. *System* (1)–(2) *is robustly stable and its anisotropic norm does not exceed given scalar value $\gamma > 0$ for known mean anisotropy level $a \geqslant 0$ if there exist such matrices $P > 0$, $\Psi > 0$,* *nonsingular matrices G_1, G_2, and scalar $\eta > \gamma^2$, for which the following inequalities hold true:*

$$\eta - \left(e^{-2a}\det\Psi\right)^{1/m} < \gamma^2, \tag{16}$$

$$\begin{bmatrix} \Psi - \eta I_m & \star & \star \\ G_1 B_{wi} & L_1 & \star \\ D_{wi} & 0 & -I_p \end{bmatrix} < 0, \tag{17}$$

$$\begin{bmatrix} -P & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ G_2 A_i & G_2 B_{wi} & L_2 & \star \\ C_i & D_{wi} & 0 & -I_p \end{bmatrix} < 0, \tag{18}$$

where $L_1 = -G_1 - G_1^{\mathrm{T}} + P$, $L_2 = -G_2 - G_2^{\mathrm{T}} + P$ and $i = \overline{1,r}$.

The proof is trivial and is not given in the paper. Theorem 2 represents the simplest and most conservative solution to the Problem 1.

Now we will use linear approximation for parametric Lyapunov matrix and some auxiliary variables.

Theorem 3. *System (1)–(2) is robustly stable and its anisotropic norm does not exceed given scalar value $\gamma > 0$ for known mean anisotropy level $a \geqslant 0$ and all possible uncertainties which satisfy (3)–(4), if there exist matrices $P_i > 0$, $\Psi > 0$, nonsingular matrices G_{1i}, G_{2i}, and scalar value $\eta > \gamma^2$, for which the following inequalities hold true:*

$$\eta - \left(e^{-2a}\det\Psi\right)^{1/m} < \gamma^2, \tag{19}$$

$$\begin{bmatrix} \Psi - \eta I_m & \star & \star \\ G_{1i} B_{wi} & -G_{1i} - G_{1i}^{\mathrm{T}} + P_i & \star \\ D_{wi} & 0 & -I_p \end{bmatrix} < 0, \tag{20}$$

$$\begin{bmatrix} -P_i & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ G_{2i} A_i & G_{2i} B_{wi} & -G_{2i} - G_{2i}^{\mathrm{T}} + P_i & \star \\ C_i & D_{wi} & 0 & -I_p \end{bmatrix} < 0, \tag{21}$$

$$\begin{bmatrix} \Psi - \eta I_m & \star & \star \\ G_{1i} B_{wj} & -G_{1i} - G_{1i}^{\mathrm{T}} + P_i & \star \\ D_{wj} & 0 & -I_p \end{bmatrix} + \begin{bmatrix} \Psi - \eta I_m & \star & \star \\ G_{1j} B_{wi} & -G_{1j} - G_{1j}^{\mathrm{T}} + P_j & \star \\ D_{wi} & 0 & -I_p \end{bmatrix} < 0, \tag{22}$$

$$\begin{bmatrix} -P_i & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ G_{2i} A_j & G_{2i} B_{wj} & -G_{2i} - G_{2i}^{\mathrm{T}} + P_i & \star \\ C_j & D_{wj} & 0 & -I_p \end{bmatrix} + \begin{bmatrix} -P_j & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ G_{2j} A_i & G_{2j} B_{wi} & -G_{2j} - G_{2j}^{\mathrm{T}} + P_j & \star \\ C_i & D_{wi} & 0 & -I_p \end{bmatrix} < 0, \tag{23}$$

where $i, j = \overline{1,r}$, $i < j$.

The proof of the theorem is listed in Appendix.

Conditions derived in Theorem 3 do not depend on the parameter vector Θ and allow us to estimate anisotropic norm of the polytopic system by checking fulfilment of $2r + r(r-1) + 1$ inequalities. The number of inequalities as well as decision variables can be reduced by increasing the

Table 1. Estimation of numerical complexity of analysis methods

Method	Number of inequalities	Number of decision variables	Number of unknown parameters
Theorem 2	$2r+2$	6	$1+\dfrac{m^2+m}{2}+\dfrac{5n^2+n}{2}$
Theorem 3	$2r+r(r-1)+2$	$2+3r$	$1+\dfrac{m^2+m}{2}+r\dfrac{5n^2+n}{2}$
Theorem 4	$2r+\dfrac{r(r-1)}{2}+2$	$2+2r$	$1+\dfrac{m^2+m}{2}+r\dfrac{3n^2+n}{2}$

conservatism of the estimation, taking into account the fact that $\Phi(\Theta)=P^{-1}(\Theta)$. Let us formulate a theorem.

Theorem 4. *System* (1)–(2) *is robustly stable and its anisotropic norm is strictly less than scalar* $\gamma>0$ *for known mean anisotropy level* $a\geqslant 0$ *and all possible inequalities, satisfying* (3)–(4), *if there exist such matrices* $\Phi_i>0$, $\Psi>0$, *nonsingular matrices* G_i, *and scalar* $\eta>\gamma^2$, *for which the following inequalities hold true:*

$$\eta-\left(\mathrm{e}^{-2a}\det\Psi\right)^{1/m}<\gamma^2, \tag{24}$$

$$\begin{bmatrix} \Psi-\eta I_m & \star & \star \\ B_{wi} & -\Phi_i & \star \\ D_{wi} & 0 & -I_p \end{bmatrix}<0, \tag{25}$$

$$\begin{bmatrix} -G_i-G_i^{\mathrm{T}}+\Phi_i & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ A_iG_i & B_{wi} & -\Phi_i & \star \\ C_iG_i & D_{wi} & 0 & -I_p \end{bmatrix}<0, \tag{26}$$

$$\begin{bmatrix} -G_i-G_i^{\mathrm{T}}+\Phi_i & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ A_jG_i & B_{wj} & -\Phi_i & \star \\ C_jG_i & D_{wj} & 0 & -I_p \end{bmatrix}+\begin{bmatrix} -G_j-G_j^{\mathrm{T}}+\Phi_j & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ A_iG_j & B_{wi} & -\Phi_j & \star \\ C_iG_j & D_{wi} & 0 & -I_p \end{bmatrix}<0, \tag{27}$$

where $i,j=\overline{1,r}$, $i<j$.

The proof of the theorem is listed in Appendix.

Conditions derived in Theorem 4 allow to estimate anisotropic norm of polytopic system by checking of fulfilment of $2r+\frac{r(r-1)}{2}+2$ inequalities. Data on the computational complexity of using each of the theorems formulated above are given in Table 1.

To estimate the anisotropic norm of the system (1)–(2) one can solve the problem of minimizing the variable γ on the set of convex constraints specified by the theorems derived above.

Unfortunately, analytical methods do not allow to evaluate the degree of conservatism of conditions obtained in Theorems 3 and 4. The degree of conservatism of the conditions can only be assessed for specific examples using numerical tools. These tools can be developed based on the Theorem 1. Consider the grid method analysis of polytopic systems based on Theorem 1. The algorithm can be presented as follows.

Algorythm 1 (grid method)

Step 1. Set mean anisotropy level $a\geqslant 0$ and step of grid h. Define set Ω, lying inside unit cube of dimensions $\mathbb{R}^{r-1}$ and consisting of mesh points. Fix parameter Θ, by setting first $(r-1)$

components coordinates of a point from the set Ω, the last component is calculated by formula

$$\theta_r = 1 - \sum_{i=1}^{r-1} \theta_i = 1.$$

Step 2. Set $k = 1$.

Step 3. While $k \leqslant N$, choose element from the set Ω_k, fix system matrices $A_k = \sum_{i=1}^{r} \theta_i A_i$, $B_k = \sum_{i=1}^{r} \theta_i B_{wi}$, $C_k = \sum_{i=1}^{r} \theta_i C_{zi}$, $D_k = \sum_{i=1}^{r} \theta_i D_{zwi}$.

Step 4. For fixed values A_k, B_k, C_k, D_k solve optimization problem:

$$\gamma_k^2 = \min \gamma^2$$

on the set of variables $\{\eta, \gamma^2, P, \Psi, G_1, G_2\}$, satisfying (9)–(11).

Step 5. If system of matrix inequalities is not feasible at the Step 4, then initial plant is not stable for given parameter values, algorithm stops. If the solution is found, then value $\gamma_* = \max\{\gamma_k, \gamma_{k-1}\}$ is calculated. If $k < N$, then $k = k + 1$, and go to Step 4. If $k = N$, then go to Step 6.

Step 6. Upper bound of anisotropic norm is defined as γ_*.

One of the disadvantages of this method is that a sufficiently large grid step will not allow one to estimate the anisotropic norm with satisfactory accuracy and give an answer about the stability of the system. Therefore, it is recommended to first check the system for robust stability using one of the existing methods.

4. NUMERICAL EXAMPLE

In the following example, we will investigate the degree of conservatism of the methods for estimating the anisotropic norm of a polytopic system, formulated in the Theorems 2–4.

Example 1. Let the system be given by the following matrices:

$$A_1 = \begin{bmatrix} 0.9 & -0.7 \\ 0.5 & -0.3 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & 1 \\ -0.5 & -0.7 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0.7 & 0.4 \\ -0.5 & -0.5 \end{bmatrix},$$

$$B_{w1} = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}, \quad B_{w2} = \begin{bmatrix} -0.5 \\ 2 \end{bmatrix}, \quad B_{w3} = \begin{bmatrix} 0 \\ -2 \end{bmatrix},$$

$$C_1 = C_2 = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad C_3 = \begin{bmatrix} 1 & 0.3 \end{bmatrix}, \quad D_{w1} = 0, \ D_{w2} = 0.1, \ D_{w3} = -0.1.$$

Note that this system is stable for all possible values of the parameters Θ. To assess the degree of conservatism of methods, proposed in Theorems 2–4, we will use the grid method for analyzing the system with grid step $h = 0.01$. Figures 1–3 illustrate the results of minimizing the value of γ at various grid nodes. When calculating the norm, Theorem 1 was used for selected numerical values of the parameter vector Θ at various grid nodes.

As can be seen in figures, the double supremum (6) for different values of mean anisotropy a is reached at points Θ which do not coincide with each other. The variation of the norm occurs smoothly and without jumps. Checking stability conditions and an attempt to estimate the anisotropic norm using the Theorem 2 leads to an infeasible problem, therefore, numerical results are given only for Theorems 3 and 4. Results of numerical experiments for calculating anisotropic norm of the system are given in Table. 2.

The conditions of the Theorem 2 are the most conservative, which led to an infeasible problem. Theorems 3 and 4 allow us to numerically estimate the anisotropic norm of a given system using linear matrix inequalities. It can be seen from Table 2, the conditions of the Theorem 4 provide

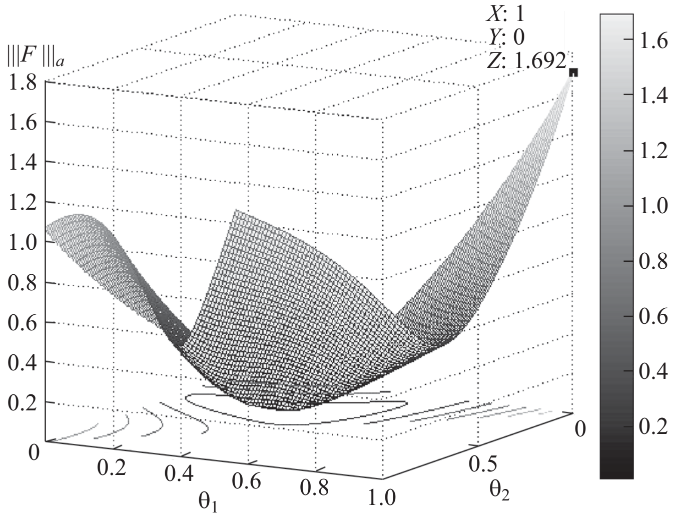


Fig. 1. Dependence of the minimum value γ on the parameters Θ at $a = 0$.

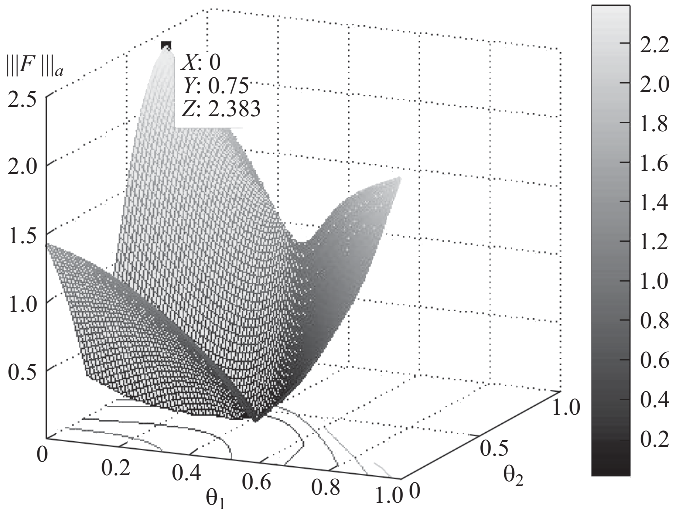


Fig. 2. Dependence of the minimum value γ on the parameters Θ at $a = 0.5$.

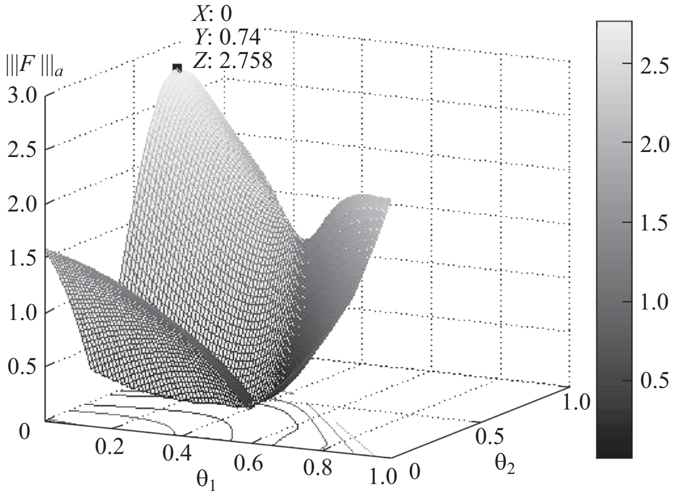


Fig. 3. Dependence of the minimum value γ on the parameters Θ at $a = 1.5$.

Table 2. Results of calculating the anisotropic norm in Example 1

Mean anisotropy a	0	0.1	0.5	1	1.5	100
$\lvert\lvert\lvert F_\Theta \rvert\rvert\rvert_a$ based on Theorem 1	1.6921	1.9258	2.3825	2.6616	2.7564	2.8100
$\lvert\lvert\lvert F_\Theta \rvert\rvert\rvert_a$ based on Theorem 3	4.6495	6.2913	7.9585	8.6163	8.8423	8.9707
$\lvert\lvert\lvert F_\Theta \rvert\rvert\rvert_a$ based on Theorem 4	6.7304	8.1552	9.0582	9.3701	9.4742	9.5327

Table 3. Results of calculating the anisotropic norm in Example 2

Mean anisotropy a	0	0.1	0.3	0.7	1.5	10
$\lvert\lvert\lvert F_\Theta \rvert\rvert\rvert_a$ based on Theorem 2	0.0771	1.1699	1.9277	2.6854	3.3354	3.7838
$\lvert\lvert\lvert F_\Theta \rvert\rvert\rvert_a$ based on Theorem 3	0.0728	0.3581	0.5827	0.8088	1.0032	1.1375
$\lvert\lvert\lvert F_\Theta \rvert\rvert\rvert_a$ based on Theorem 4	0.0727	0.3579	0.5820	0.8083	1.0028	1.1366

more conservative results. Despite this, the asymptotic behavior of the anisotropic norm for the given numerically implementable methods is preserved with a significantly lower computational complexity. Thus, these methods can be used to estimate the anisotropy-based performance of polytopic systems.

Example 2. Consider now mathematical model of damped oscillations of a spring pendulum:

$$\begin{aligned}\dot{x}(t) &= Ax(t) + B_w w(t),\\ y(t) &= x_1(t) + D_w w(t).\end{aligned}$$

Here

$$A = \begin{bmatrix} 0 & 1 \\ -\omega^2 & -2\xi\omega \end{bmatrix},$$

where ω is natural frequency of the system, ξ is attenuation coefficient, $x_1(t)$ is pendulum's center of mass position, and $x_2(t)$ is pendulum's center of mass speed.

Disturbance $w(t) \in \mathbb{R}^2$ consists of external disturbance, acting on position $x_1(t)$, and measurement noise. Then

$$B_w = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad D_w = \begin{bmatrix} 0 & 0,1 \end{bmatrix}.$$

Let the system parameters are $\xi = 0.1$, $\omega \in [4.5;\ \ 5.2]$.

Initial plant given continuous time is discretized using zero order hold as

$$A^d = \mathrm{e}^{A^f h}, \quad B_w^d = \int_0^h \mathrm{e}^{A^f(h-\tau)} B^w d\tau, \tag{28}$$

where h is discretization step.

Initial continuous plant was discretized with discretization step $h = 10^{-3}$ sec. The following parameters were obtained:

$$A_1^d = \begin{bmatrix} 1 & 0.0010 \\ -0.0202 & 0.9991 \end{bmatrix}, \quad A_2^d = \begin{bmatrix} 1 & 0.0010 \\ -0.0270 & 0.9989 \end{bmatrix},$$

$$B_{w1}^d = 10^{-3} \times \begin{bmatrix} 1 & 0 \\ -0.0101 & 0 \end{bmatrix}, \quad B_{w2}^d = 10^{-3} \times \begin{bmatrix} 1 & 0 \\ -0.0135 & 0 \end{bmatrix},$$

$$C_1 = C_2 = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad D_{w1} = D_{w2} = \begin{bmatrix} 0 & 0.1 \end{bmatrix}.$$

Note that initial system is stable. The results of calculating of the anisotropic norm for a spring pendulum are summarized in Table 3.

5. CONCLUSIONS

In this paper, conditions for the boundedness of the anisotropic norm of a linear polytopic system are obtained in terms of linear matrix inequalities. Various options for nonparametric estimation were considered anisotropic norm, and also analyzed the estimation accuracy and computational complexity of these methods. The conditions are convex and formulated in terms of matrix inequalities, the number of which depends on the number of vertices of the polytope.

FUNDING

This work was supported by the Russian Science Foundation, project no. 23-21-00306.
https://rscf.ru/project/23-21-00306/

APPENDIX

Proof of Theorem 1. The proof of the theorem consists of two parts. In the first one we will obtain conditions under which the polytopic the system (1)–(2) is robustly stable, and its $\mathcal{H}_\infty$-norm is bounded some number $\sqrt{\eta}$, i.e. $F_\Theta \in \mathcal{H}_\infty{}^{p\times m}$. In the second part of the proof we obtain conditions for the boundedness of the anisotropic norm for the robustly stable system $F_\Theta \in \mathcal{H}_\infty{}^{p\times m}$.

Consider the following parametric function as Lyapunov function candidate

$$V(k) = x^{\mathrm{T}}(k)P(\Theta)x(k), \quad P(\Theta) > 0. \tag{A.1}$$

Since we first require to prove the stability of the system and the boundedness of its $\mathcal{H}_\infty$-norm, then, to simplify calculations and without loss of generality, we assume that $W = \{w(k)\}_{k\,\in\,\mathbb{Z}} \in \mathcal{L}_2$. The difference between $V(k+1)$ and $V(k)$ is determined by the formula

$$V(k+1) - V(k) = x^{\mathrm{T}}(k+1)P(\Theta)x(k+1) - x^{\mathrm{T}}(k)P(\Theta)x(k). \tag{A.2}$$

Now we consider the expression:

$$\begin{aligned}
&V(k+1) - V(k) + z^{\mathrm{T}}(k)z(k) - \eta w^{\mathrm{T}}(k)w(k)\\
&\quad = \{\text{substitute}\, x(k+1) = A(\Theta)x(k) + B_w(\Theta)w(k) \text{ and } z(k) = C(\Theta)x(k) + D_w(\Theta)w(k)\}\\
&\qquad = \begin{bmatrix} x^{\mathrm{T}}(k) & w^{\mathrm{T}}(k) \end{bmatrix} \Bigg(\begin{bmatrix} A(\Theta) & B_w(\Theta) \end{bmatrix}^{\mathrm{T}} P(\Theta) \begin{bmatrix} A(\Theta) & B_w(\Theta) \end{bmatrix}\\
&\qquad\quad + \begin{bmatrix} C(\Theta) & D_w(\Theta) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} C(\Theta) & D_w(\Theta) \end{bmatrix} - \begin{bmatrix} P(\Theta) & 0 \\ 0 & \eta I_m \end{bmatrix} \Bigg) \begin{bmatrix} x(k) \\ w(k) \end{bmatrix}.
\end{aligned} \tag{A.3}$$

Thus, inequality

$$V(k+1) - V(k) + z^{\mathrm{T}}(k)z(k) - \eta w^{\mathrm{T}}(k)w(k) < 0 \tag{A.4}$$

holds for all $x(k)$ and $w(k)$ if

$$\begin{gathered}
\begin{bmatrix} A(\Theta) & B_w(\Theta) \end{bmatrix}^{\mathrm{T}} P(\Theta) \begin{bmatrix} A(\Theta) & B_w(\Theta) \end{bmatrix}\\
+ \begin{bmatrix} C(\Theta) & D_w(\Theta) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} C(\Theta) & D_w(\Theta) \end{bmatrix} - \begin{bmatrix} P(\Theta) & 0 \\ 0 & \eta I_m \end{bmatrix} < 0.
\end{gathered} \tag{A.5}$$

Let us transform the inequality (A.5) to the form

$$\begin{bmatrix} -P(\Theta) & 0 \\ 0 & -\eta I_m \end{bmatrix} - \begin{bmatrix} A(\Theta) & B_w(\Theta) \\ C(\Theta) & D_w(\Theta) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} -P(\Theta) & 0 \\ 0 & -I_p \end{bmatrix} \begin{bmatrix} A(\Theta) & B_w(\Theta) \\ C(\Theta) & D_w(\Theta) \end{bmatrix} < 0, \tag{A.6}$$

where matrix $\begin{bmatrix} -P^{-1}(\Theta) & 0 \\ 0 & -I_p \end{bmatrix}$ is negative definite. Applying to the inequality (A.6) Schur complement, we have

$$\begin{bmatrix} -P(\Theta) & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ A(\Theta) & B_w(\Theta) & -P^{-1}(\Theta) & \star \\ C(\Theta) & D_w(\Theta) & 0 & -I_p \end{bmatrix} < 0. \tag{A.7}$$

Fulfilling of the inequality (A.6) for zero input makes true inequalities of the form (A.4) for any $k \in \mathbb{Z}_+ \bigcup \{0\}$ and allows to sum up them from $k = 0$ to $k = \infty$. This implies the inequality

$$V(\infty) - V(0) + \sum_{k=0}^{\infty} z^{\mathrm{T}}(k) z(k) - \eta \sum_{k=0}^{\infty} w^{\mathrm{T}}(k) w(k) < 0. \tag{A.8}$$

For zero initial condition ($x(0) = 0$) $V(0) = 0$, assuming that $V(\infty) = 0$, inequality (A.8) transforms to the form

$$\sum_{k=0}^{\infty} z^{\mathrm{T}}(k) z(k) < \eta \sum_{k=0}^{\infty} w^{\mathrm{T}}(k) w(k).$$

Therefore,

$$\sup_{\Theta \in \mathfrak{Q}} \sup_{W \in \mathcal{L}_2} \frac{\sum_{k=0}^{\infty} z^{\mathrm{T}}(k) z(k)}{\sum_{k=0}^{\infty} w^{\mathrm{T}}(k) w(k)} < \eta. \tag{A.9}$$

Fulfilment of inequality (A.7) guarantees stability of the open loop system (1)–(2) and boundedness its $\mathcal{H}_\infty$-norm by scalar $\sqrt{\eta}$.

At the second step, it is necessary to find out conditions that guarantee the boundedness of the anisotropic norm for the mean anisotropy level $\overline{\mathbf{A}}(W) \leqslant a$ of input disturbances. Then conditions of anisotropic norm boundedness can be defined by anisotropy-based bounded real lemma [12] as follows:

$$-(\det(\Sigma(\Theta)))^{1/m} < -(1 - q\gamma^2)\mathrm{e}^{2a/m}, \tag{A.10}$$

$$\begin{gathered} \begin{bmatrix} A(\Theta)R(\Theta)A(\Theta) - R(\Theta) & A^{\mathrm{T}}(\Theta)R(\Theta)B_w(\Theta) \\ B_w^{\mathrm{T}}(\Theta)R(\Theta)A(\Theta) & B_w^{\mathrm{T}}(\Theta)R(\Theta)B_w(\Theta) - I_m \end{bmatrix} \\ + q \begin{bmatrix} C^{\mathrm{T}}(\Theta) \\ D_w^{\mathrm{T}}(\Theta) \end{bmatrix} \begin{bmatrix} C(\Theta) & D_w(\Theta) \end{bmatrix} < 0, \end{gathered} \tag{A.11}$$

where $q \in (0, \min(\gamma^{-2}, \|F_\Theta\|_\infty^{-2}))$, and $\Sigma(\Theta)$ defined by

$$\Sigma(\Theta) = (I_m - B_w^{\mathrm{T}}(\Theta)R(\Theta)B_w(\Theta) - qD_w^{\mathrm{T}}(\Theta)D_w(\Theta)). \tag{A.12}$$

Inequality (A.11) coincides with inequality (A.5) taking into account the change of variables $P(\Theta) = \eta R(\Theta)$ and $\eta = q^{-1}$. Thus, anisotropic norm of the system is bounded if inequalities (A.7) and (A.11) hold true.

Consider inequality (A.10) in detail. Taking into account introduced notations, it can be rewritten as

$$\eta - (\mathrm{e}^{-2a} \det(\eta I_m - B_w^{\mathrm{T}}(\Theta)P(\Theta)B_w(\Theta) - D_w^{\mathrm{T}}(\Theta)D_w(\Theta)))^{1/m} < \gamma^2. \tag{A.13}$$

Introducing new variable

$$\Psi(\Theta) < \eta I_m - B_w^{\mathrm{T}}(\Theta) P(\Theta) B_w(\Theta) - D_w^{\mathrm{T}}(\Theta) D_w(\Theta),$$

where $\Psi(\Theta) = \Psi(\Theta)^{\mathrm{T}} > 0$ [13], we ascertain inequality (A.13) fulfilled, if two following inequalities hold:

$$\eta - (\mathrm{e}^{-2a} \det(\Psi(\Theta)))^{1/m} < \gamma^2, \tag{A.14}$$

$$\Psi(\Theta) < \eta I_m - B_w^{\mathrm{T}}(\Theta) P(\Theta) B_w(\Theta) - D_w^{\mathrm{T}}(\Theta) D_w(\Theta). \tag{A.15}$$

Rewrite (A.15) as

$$\Psi(\Theta) - \eta I_m - \left[\begin{array}{cc} B_w^{\mathrm{T}}(\Theta) & D_w^{\mathrm{T}}(\Theta) \end{array} \right] \left[\begin{array}{cc} -P(\Theta) & 0 \\ 0 & -I_p \end{array} \right] \left[\begin{array}{c} B_w(\Theta) \\ D_w(\Theta) \end{array} \right] < 0. \tag{A.16}$$

Applying Schur complement to the expression (A.16), we obtain

$$\left[\begin{array}{ccc} \Psi(\Theta) - \eta I_m & B_w^{\mathrm{T}}(\Theta) & D_w^{\mathrm{T}}(\Theta) \\ B_w(\Theta) & -P^{-1}(\Theta) & 0 \\ D_w(\Theta) & 0 & -I_p \end{array} \right] < 0. \tag{A.17}$$

By right and left multiplying inequality (A.17) by matrix $\left[\begin{array}{ccc} I & 0 & 0 \\ 0 & G_1(\Theta) & 0 \\ 0 & 0 & I \end{array} \right]$ and its transposed, we get

$$\left[\begin{array}{ccc} \Psi(\Theta) - \eta I_m & \star & \star \\ G_1(\Theta) B_w(\Theta) & \Lambda_1(\Theta) & \star \\ D_w(\Theta) & 0 & -I_p \end{array} \right] < 0, \tag{A.18}$$

where $\Lambda_1(\Theta) = -G_1(\Theta) P^{-1}(\Theta) G_1^{\mathrm{T}}(\Theta)$.

Note that for $P(\Theta) > 0$ it follows from inequality

$$-(G_1(\Theta) - P(\Theta))^{\mathrm{T}} P^{-1}(\Theta)(G_1(\Theta) - P(\Theta)) \leqslant 0$$

that

$$-G_1(\Theta) P^{-1}(\Theta) G_1^{\mathrm{T}}(\Theta) \leqslant -G_1(\Theta) - G_1^{\mathrm{T}}(\Theta) + P(\Theta).$$

Introducing notation $L_1(\Theta) = -G_1(\Theta) - G_1^{\mathrm{T}}(\Theta) + P(\Theta)$ and replacing $\Lambda_1(\Theta)$ by $L_1(\Theta)$ at the inequality (A.18), we get inequality (14).

Let's get rid of the inversion of the matrix $P(\Theta)$ in the inequality (A.7).To do this, we introduce a new nonsingular matrix $G_2(\Theta)$. By right and left multiplying inequality (A.7) by nonsingular matrix

$$\left[\begin{array}{cccc} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & G_2(\Theta) & 0 \\ 0 & 0 & 0 & I \end{array} \right] \tag{A.19}$$

and its transposed respectively, we get:

$$\begin{bmatrix} -P(\Theta) & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ G_2(\Theta)A(\Theta) & G_2(\Theta)B_w(\Theta) & \Lambda_2(\Theta) & \star \\ C(\Theta) & D_w(\Theta) & 0 & -I_p \end{bmatrix} < 0, \tag{A.20}$$

where

$$\Lambda_2(\Theta) = -G_2(\Theta)P^{-1}(\Theta)G_2^{\mathrm{T}}(\Theta). \tag{A.21}$$

Similar to previous case, we replace $\Lambda_2(\Theta)$ by expression $L_2(\Theta) = -G_2(\Theta) - G_2^{\mathrm{T}}(\Theta) + P(\Theta)$. As a result, we have expression (15).

Theorem 1 is proved.

Proof of Theorem 3. Define matrices $\Psi(\Theta)$, $G_1(\Theta)$, $G_2(\Theta)$, and $P(\Theta)$ in the form $\Psi(\Theta) = \Psi$, $G_1(\Theta) = \sum_{i=1}^{r}\theta_i G_{1i}$, $G_2(\Theta) = \sum_{i=1}^{r}\theta_i G_{2i}$, $P(\Theta) = \sum_{i=1}^{r}\theta_i P_i$. Rewrite inequalities (14) and (15) taking into account introduced assumptions.

It should be noted that the inequalities (14) and (15) contain blocks of constant matrices, parametric matrices and products of two parametric matrices. Taking into account the introduced appearance of parametric variables, and also taking into account the fact that $(\sum_{i=1}^{s}\theta_i)^2 = 1$, constant matrices can be written in the form

$$I_p = \left(\sum_{i=1}^{s}\theta_i\right)^2 I_p.$$

Because of identity $\sum_{j=1}^{r}\theta_j = 1$, parametric matrices can be rewritten as

$$\sum_{i=1}^{r}\theta_i A_i = \sum_{i=1}^{r}\theta_i\left(\sum_{j=1}^{r}\theta_j\right)A_i.$$

Expressions of the form $G_1(\Theta)B_w(\Theta)$ are written as follows:

$$G_1(\Theta)B_w(\Theta) = \sum_{i=1}^{r}\theta_i^2(G_{1i}B_i) + \sum_{i=1}^{r}\sum_{i<j}^{r}\theta_i\theta_j(G_{1i}B_{wj} + G_{1j}B_{wi}).$$

Applying all above mentioned transformation to each element of inequalities (14) and (15), we get:

$$\sum_{i=1}^{r}\theta_i^2\begin{bmatrix} \Psi - \eta I_m & \star & \star \\ B_{wi} & -G_{1i} - G_{1i}^{\mathrm{T}} + P_i & \star \\ D_{wi} & 0 & -I_p \end{bmatrix}$$
$$+\sum_{i=1}^{r}\sum_{i<j}^{r}\theta_i\theta_j\left(\begin{bmatrix} \Psi - \eta I_m & \star & \star \\ G_{1i}B_{wj} & -G_{1i} - G_{1i}^{\mathrm{T}} + P_i & \star \\ D_{wj} & 0 & -I_p \end{bmatrix}\right.$$
$$\left.+\begin{bmatrix} \Psi - \eta I_m & \star & \star \\ G_{1j}B_{wi} & -G_{1j} - G_{1j}^{\mathrm{T}} + P_j & \star \\ D_{wi} & 0 & -I_p \end{bmatrix}\right) < 0,$$

$$\sum_{i=1}^{r}\theta_i^2\begin{bmatrix} -P_i & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ G_{2i}A_i & G_{2i}B_{wi} & -G_{2i}-G_{2i}^{\mathrm{T}}+P_i & \star \\ C_i & D_{wi} & 0 & -I_p \end{bmatrix}$$

$$+\sum_{i=1}^{r}\sum_{i<j}^{r}\theta_i\theta_j\left(\begin{bmatrix} -P_i & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ G_{2i}A_j & G_{2i}B_{wj} & -G_{2i}-G_{2i}^{\mathrm{T}}+P_i & \star \\ C_j & D_{wj} & 0 & -I_p \end{bmatrix}\right.$$

$$\left.+\begin{bmatrix} -P_j & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ G_{2j}A_i & G_{2j}B_{wi} & -G_{2j}-G_{2j}^{\mathrm{T}}+P_j & \star \\ C_i & D_{wi} & 0 & -I_p \end{bmatrix}\right) < 0.$$

Since $\theta_i \geqslant 0$, $i=\overline{1,r}$, it obvious that inequalities (13)–(15) hold, when inequalities (19)–(23) hold.

Proof of Theorem 4. Let us consider inequalities (A.7) and (A.17), obtained in the proof of Theorem 1. Introduce new variable $\Phi(\Theta) = P^{-1}(\Theta)$, and fix parameter η and matrix Ψ. Then inequalities (A.7) and (A.17) will be rewritten as follows:

$$\begin{bmatrix} \Psi-\eta I_m & B_w^{\mathrm{T}}(\Theta) & D_w^{\mathrm{T}}(\Theta) \\ B_w(\Theta) & -\Phi(\Theta) & 0 \\ D_w(\Theta) & 0 & -I_p \end{bmatrix} < 0 \qquad \text{(A.22)}$$

and

$$\begin{bmatrix} -\Phi^{-1}(\Theta) & \star & \star & \star \\ 0 & -\gamma^2 I_m & \star & \star \\ A(\Theta) & B_w(\Theta) & -\Phi(\Theta) & \star \\ C(\Theta) & D_w(\Theta) & 0 & -I_p \end{bmatrix} < 0. \qquad \text{(A.23)}$$

The latest inequality contains matrix $\Phi^{-1}(\Theta)$. To get rid of it, we will left and right multiply inequality (A.23) by matrix

$$\begin{bmatrix} G^{\mathrm{T}}(\Theta) & 0 & 0 & 0 \\ 0 & I_m & 0 & 0 \\ 0 & 0 & I_n & 0 \\ 0 & 0 & 0 & I_p \end{bmatrix}$$

and its transposed respectively. It results to:

$$\begin{bmatrix} \Lambda(\Theta) & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ A(\Theta)G(\Theta) & B_w(\Theta) & -\Phi(\Theta) & \star \\ C(\Theta)G(\Theta) & D_w(\Theta) & 0 & -I_p \end{bmatrix} < 0, \qquad \text{(A.24)}$$

where $\Lambda(\Theta) = -G^{\mathrm{T}}(\Theta)\Phi^{-1}(\Theta)G(\Theta)$.

Note that $\Phi(\Theta) > 0$, therefore fulfilment of inequality

$$-(G(\Theta) - \Phi(\Theta))^{\mathrm{T}}\Phi^{-1}(\Theta)(G(\Theta) - \Phi(\Theta)) \leqslant 0$$

results to $-G^{\mathrm{T}}(\Theta)\Phi^{-1}(\Theta)G(\Theta) \leqslant -G(\Theta) - G^{\mathrm{T}}(\Theta) + \Phi(\Theta)$. From the latter it follows that inequality (A.24) holds, if inequality (A.23) holds.

Consider matrix $\Phi(\Theta)$ be appeared in the form $\Phi(\Theta) = \sum_{i=1}^{r} \theta_i \Phi_i$ taking into account expressions for parametric uncertainties (3)–(4). Then inequalities (A.22) and (A.24) take form:

$$\sum_{i=1}^{r} \theta_i \begin{bmatrix} \Psi - \eta I_m & \star & \star \\ B_{wi} & -\Phi_i & \star \\ D_{wi} & 0 & -I_p \end{bmatrix} < 0, \tag{A.25}$$

$$\begin{gathered} \sum_{i=1}^{r} \theta_i^2 \begin{bmatrix} -G_i - G_i^{\mathrm{T}} + \Phi_i & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ A_i G_i & B_{wi} & -\Phi_i & \star \\ C_i G_i & D_{wi} & 0 & -I_p \end{bmatrix} \\ + \sum_{i=1}^{r} \sum_{i<j}^{r} \theta_i \theta_j \left(\begin{bmatrix} -G_i - G_i^{\mathrm{T}} + \Phi_i & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ A_j G_i & B_{wj} & -\Phi_i & \star \\ C_j G_i & D_{wj} & 0 & -I_p \end{bmatrix} \right. \\ \left. + \begin{bmatrix} -G_j - G_j^{\mathrm{T}} + \Phi_j & \star & \star & \star \\ 0 & -\eta I_m & \star & \star \\ A_i G_j & B_{wi} & -\Phi_j & \star \\ C_i G_j & D_{wi} & 0 & -I_p \end{bmatrix} \right) < 0. \end{gathered} \tag{A.26}$$

Note that inequality (A.26) can be obtained using property (4) and considering that $(\sum_{i=1}^{r} \theta_i)^2 = 1$. Obviously, fulfilment inequalities (25)–(27) automatically leads to fulfilment inequalities (A.25) and (A.26), that completes the proof.

REFERENCES

1. Gahinet, P., Apkarian, P., and Chilali, M., Affine parameter-dependent Lyapunov functions and real parametric uncertainty, *IEEE Trans. Automat. Control*, 1996, vol. 41, no. 3, pp. 436–442.

2. Daafouz, J. and Bernussou, J., Parameter dependent Lyapunov functions for discrete time systems with time varying parametric uncertainties, *Systems & Control Letters*, 2001, vol. 43, no. 5, pp. 355–359.

3. Peaucelle, D. and Arzelier, D., Robust performance analysis with LMI-based methods for real parametric uncertainty via parameter-dependent Lyapunov functions, *IEEE Trans. Automat. Control*, 2001, vol. 46, pp. 624–630.

4. de Oliveira, M.C., Bernussou, J., and Geromel, J.C., A new discrete-time robust stability condition, *Systems & Control Letters*, 1999, vol. 37, no. 4, pp. 261–265.

5. Oliveira, R.C.L.F. and Peres, P.L.D., A convex optimization procedure to compute $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norms for uncertain linear systems in polytopic domains, *Optim. Control Appl. Meth.*, 2008, vol. 29, pp. 295–312.

6. Vladimirov, I.G., Kurdyukov, A.P., and Semyonov, A.V., Anisotropy of signals and the entropy of linear stationary systems, *Doklady Math.*, 1995, vol. 51, no. 3, pp. 388–390.

7. Vladimirov, I.G., Kurdyukov, A.P., and Semyonov, A.V., On computing the anisotropic norm of linear discrete-time-invariant systems, *Proc. 13th IFAC World Congress (San-Francisco, USA)*, 1996, pp. 179–184.

8. Diamond, P., Vladimirov, I.G., Kurdyukov, A.P., and Semyonov, A.V., Anisotropy-based performance analysis of linear discrete time-invariant control systems, *Int. J. of Control*, 2001, vol. 74, no. 1, pp. 28–42.

9. Andrianova, O.G. and Belov, A.A., On Robust Performance Analysis of Linear Systems with Polytopic Uncertainties Affected by Random Disturbances, *Proceedings of the 20th International Carpathian Control Conference (ICCC 2019, Krakow-Wieliczka, Poland)*, 2019, pp. 1–6.

10. Belov, A.A., Random Disturbance Attenuation in Discrete-time Polytopic Systems: Performance Analysis and State-Feedback Control, *Proceedings of the 2020 European Control Conference (ECC 20, Saint Petersburg, Russia)*, 2020, pp. 633–637.

11. Belov, A.A., Robust pole placement and random disturbance rejection for linear polytopic systems with application to grid-connected converters, *European Journal of Control*, 2022, vol. 63, pp. 116–125.

12. Tchaikovsky, M.M. and Kurdyukov, A.P., Strict Anisotropic Norm Bounded Real Lemma in Terms of Matrix Inequalities, *Doklady Math.*, 2011, vol. 48, no. 3, pp. 895–898.

13. Tchaikovsky, M.M., Design of anisotropic stochastic robust control using convex optimization, *Dr. Sci. Thesis*, 2012 (In Russian).

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