Volume 85, Number 12,
December 2024

ISSN 0005-1179
CODEN: AURCAT

A JOURNAL OF RUSSIAN ACADEMY OF SCIENCES

AUTOMATION AND REMOTE CONTROL

Automation and Remote Control

Vol. 85, No. 12, December 2024

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

0005-1179/24. *Automation and Remote Control* (ISSN: 0005-1179 print version, ISSN: 1608-3032 electronic version) is published monthly by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 65 Profsoyuznaya street, Moscow 117997, Russia.
Volume 85 (12 issues) is published in 2024.
Publisher: Trapeznikov Institute of Control Sciences, Russian Academy of Sciences.
65 Profsoyuznaya street, Moscow 117997, Russia; e-mail: redacsia@ipu.rssi.ru; http://ait.mtas.ru, http://ait-arc.ru

Contents

Automation and Remote Control
Vol. 85, No. 12, 2024

Linear Systems

Stochastic Systems

Control in Social Economic Systems

Notes, Chronicles, Information

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 12, pp. 1133–1148.

LINEAR SYSTEMS

Output Feedback Design in Discrete-Time Control Systems 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 6, 2024
Revised September 26, 2024
Accepted October 22, 2024

Abstract—This paper proposes a novel approach to rejecting nonrandom bounded exogenous disturbances in linear discrete-time control systems by means of dynamic output feedback. The approach is based on reducing the original problem to a matrix optimization problem with the gain and observer matrices as the variables. A gradient descent method for finding dynamic output feedback is derived and justified.

Keywords: linear systems, discrete time, exogenous disturbances, output feedback, observer, optimization, Lyapunov equation, gradient descent method, Newton's method

DOI: 10.31857/S0005117924120013

1. INTRODUCTION

A new approach to rejecting nonrandom bounded exogenous disturbances in linear continuous-time control systems by means of dynamic output feedback using the Luenberger observer [2, 3] was proposed and discussed in [1]. Dating back to the work [4], this approach is based on reducing the original problem to a matrix optimization problem, where the variables are the gain and observer matrices. Below, this problem is solved using a gradient descent method.

The literature on the corresponding range of problems was briefly reviewed in [1]; for example, see [5] and the extensive bibliography therein. The same problem was solved in [6] via reduction to a semidefinite programming problem in terms of linear matrix inequalities (LMIs) [7, 8]; in doing so, several rough transformations were undertaken to linearize the matrix inequalities and establish the final result in terms of LMIs. Consequently, the final conditions suffered from excessive conservatism.

Among ideologically similar publications, note the recent works [9, 10], where a promising approach was proposed based on solving parameterized matrix Riccati equations.

This paper is a direct continuation of the research [1], extending the considerations presented therein to discrete-time control systems. An important applications-oriented feature of the approach is the possibility of limiting the magnitude of the observer and dynamic controller matrices.

From now on, $\mathbb{S}^n$ denotes the space of symmetric real matrices of dimensions $n \times n$; $\|\cdot\|$ is the Euclidean norm of a vector and the spectral norm of a matrix; $\|\cdot\|_F$ indicates the Frobenius norm of a matrix; $^{\mathrm{T}}$ is the transpose symbol; tr stands for the trace of a matrix; $\langle\cdot,\cdot\rangle$ means the Frobenius inner product of matrices; I denotes an identity matrix of appropriate dimension; $\lambda_i(A)$ are the eigenvalues of a matrix A; finally, $\rho(A) = \max\limits_i |\lambda_i(A)| < 1$ specifies the stability radius of a Schur matrix A.

2. PROBLEM STATEMENT

Consider a control system described by

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Dw_k, \\ y_k &= C_1x_k + D_1w_k, \\ z_k &= C_2x_k, \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$, $D \in \mathbb{R}^{n\times m}$, $D_1 \in \mathbb{R}^{\ell\times m}$, $C_1 \in \mathbb{R}^{\ell\times n}$, $C_2 \in \mathbb{R}^{r\times n}$, with the state $x_k \in \mathbb{R}^n$, an initial condition x_0, the observed output $y_k \in \mathbb{R}^\ell$, the optimized (target) output $z_k \in \mathbb{R}^r$, the control input $u_k \in \mathbb{R}^p$, and an exogenous disturbance[1] $w_k \in \mathbb{R}^m$ bounded at each time instant:

$$\|w_k\| \leqslant \delta \quad \text{for all } k = 0, 1, 2, \ldots .$$

By assumption, the pairs (A, B) and (A, C_1) are controllable and observable, respectively.

Suppose that the state vector x_k of the system is unmeasurable and information about the system is provided by its output y_k. The problem is to find a minimal (in a sense) ellipsoid containing the target output z_k.

Let us construct an observer described by a linear difference equation with the mismatch between the output y_k and its prediction $C_1\widehat{x}_k$:

$$\widehat{x}_{k+1} = A\widehat{x}_k + Bu_k + L(y_k - C_1\widehat{x}_k), \quad \widehat{x}_0 = 0, \tag{2}$$

where $L \in \mathbb{R}^{n\times\ell}$ is the observer matrix.

According to (1) and (2), the error $e_k = x_k - \widehat{x}_k$ satisfies the difference equation

$$e_{k+1} = (A - LC_1)e_k + (D - LD_1)w_k, \quad e_0 = x_0.$$

With the dynamic feedback controller

$$u_k = K\widehat{x}_k, \quad K \in \mathbb{R}^{p\times n}, \tag{3}$$

applied to system (1), we arrive at the closed loop system

$$\begin{aligned} x_{k+1} &= (A + BK)x_k - BKe_k + Dw_k, \\ e_{k+1} &= (A - LC_1)e_k + (D - LD_1)w_k, \\ z_k &= C_2x_k \end{aligned} \tag{4}$$

with the controlled output z_k.

Like in the continuous-time case, it is often impossible to design a static output-feedback controller $u_k = Ky_k$ for the system: the matrix $A + BKC_1$ may turn out nonstabilizable by the choice of K, while the dynamic controller (3) can be constructed (under the relatively mild requirements of system controllability and observability; see Section 3 for details).

3. THE SOLUTION APPROACH

We employ the invariant ellipsoid method [8, 13]. Recall that an ellipsoid

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

[1] Although the nature of disturbances in the state vector and output of the system generally differs, it is convenient to consider them the same, assuming that the matrices D and D_1 “cut out” different “pieces” from the vector w_k; the general case can also be considered at the expense of some complication.

is said to be *invariant* for a linear discrete-time dynamic system

$$\begin{aligned} x_{k+1} &= Ax_k + Dw_k, \quad \|w_k\| \leqslant 1, \\ z_k &= Cx_k \end{aligned} \tag{5}$$

with a Schur matrix A if any system trajectory starting from an initial point x_0 inside the ellipsoid $\mathcal{E}_x$ remains there at any time instant under all exogenous disturbances $w_k\colon \|w_k\| \leqslant 1$.

An invariant ellipsoid has the attraction property: a system trajectory starting from an initial point outside the ellipsoid tends to it over time.

Obviously, if $\mathcal{E}_x$ is an invariant ellipsoid with a matrix P, the linear output z_k of system (5) with $x_0 \in \mathcal{E}_x$ will belong to the ellipsoid

$$\mathcal{E}_z = \{z \in \mathbb{R}^r\colon \quad z^{\mathrm{T}}(CPC^{\mathrm{T}})^{-1}z \leqslant 1\},$$

called *bounding.* (In the case $x_0 \notin \mathcal{E}_x$, this output will tend to the bounding ellipsoid.)

When estimating the effect of exogenous disturbances on the system output, minimal bounding ellipsoids are of natural interest; a common minimality criterion for a bounding ellipsoid is the value $\operatorname{tr} CPC^{\mathrm{T}}$, equal to the sum of the squares of its semiaxes. The following result is well known; for example, see the book [8].

Theorem 1. *Assume that A is a Schur matrix, $\rho = \max_i |\lambda_i(A)| < 1$, the pair (A, D) is controllable, and a matrix $P(\alpha) \succ 0$, $\rho^2 < \alpha < 1$, satisfies the discrete Lyapunov equation*

$$\frac{1}{\alpha}APA^{\mathrm{T}} - P + \frac{1}{1-\alpha}DD^{\mathrm{T}} = 0. \tag{6}$$

Then the problem on the optimal bounding ellipsoid for system (5) *is reduced to minimizing the one-dimensional objective function*

$$f(\alpha) = \operatorname{tr} CP(\alpha)C^{\mathrm{T}}$$

on the interval $\rho^2 < \alpha < 1$.

Let us make several remarks. First, if α^* is a minimum point in the problem of Theorem 1 and x_0 satisfies the condition $x_0^{\mathrm{T}}P^{-1}(\alpha^*)x_0 \leqslant 1$, then the following bound uniformly holds:

$$\|z_k\|^2 \leqslant \|CP(\alpha^*)C^{\mathrm{T}}\| \leqslant \operatorname{tr} CP(\alpha^*)C^{\mathrm{T}} = f(\alpha^*), \quad k = 0, 1, 2, \ldots.$$

Second, equation (6) can be represented as

$$\left(\frac{1}{\sqrt{\alpha}}A\right) P \left(\frac{1}{\sqrt{\alpha}}A\right)^{\mathrm{T}} - P + \frac{1}{1-\alpha}DD^{\mathrm{T}} = 0;$$

by [8, Lemma 1.2.6], it has a unique positive definite solution if and only if $\frac{1}{\sqrt{\alpha}}A$ is a Schur matrix, i.e.,

$$\rho\left(\frac{1}{\sqrt{\alpha}}A\right) < 1.$$

Introducing the combined vector

$$g_k = \begin{pmatrix} x_k \\ e_k \end{pmatrix} \in \mathbb{R}^{2n},$$

we write system (4) in matrix form and consider the matrices $A_{K,L}$, D_L, and $\mathcal{C}$:

$$\begin{aligned}
\dot g_k &= \underbrace{\begin{pmatrix} A+BK & -BK \\ 0 & A-LC_1 \end{pmatrix}}_{A_{K,L}} g_k + \underbrace{\begin{pmatrix} D \\ D-LD_1 \end{pmatrix}}_{D_L} w_k, \quad g_0 = \begin{pmatrix} x_0 \\ x_0 \end{pmatrix}, \\
z_k &= \underbrace{\begin{pmatrix} C_2 & 0 \end{pmatrix}}_{\mathcal{C}} g_k.
\end{aligned} \tag{7}$$

Enclosing the state g_k of system (7) in the invariant ellipsoid

$$\mathcal{E}_g = \{g \in \mathbb{R}^{2n}: \quad g^{\mathrm{T}} P^{-1} g \leqslant 1\}$$

generated by a matrix $P \in \mathbb{S}^{2n}$, we minimize the size of the corresponding bounding ellipsoid

$$\mathcal{E}_z = \{z \in \mathbb{R}^{r}: \quad z^{\mathrm{T}} (\mathcal{C} P \mathcal{C}^{\mathrm{T}})^{-1} z \leqslant 1\}$$

for the output z_k with the matrix $\mathcal{C}P\mathcal{C}^{\mathrm{T}}$. As the minimality criterion we choose the trace function $\operatorname{tr} \mathcal{C}P\mathcal{C}^{\mathrm{T}}$.

Note that the matrix $A_{K,L}$ can be represented as

$$A_{K,L} = \begin{pmatrix} A+BK & -BK \\ 0 & A-LC_1 \end{pmatrix} = \underbrace{\begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix}}_{\mathcal{A}} + \underbrace{\begin{pmatrix} B \\ 0 \end{pmatrix}}_{M_1} K \underbrace{\begin{pmatrix} I & -I \end{pmatrix}}_{N_1} + \underbrace{\begin{pmatrix} 0 \\ I \end{pmatrix}}_{M_2} L \underbrace{\begin{pmatrix} 0 & -C_1 \end{pmatrix}}_{N_2}.$$

According to Theorem 1, we therefore arrive at the problem of minimizing the objective function $\operatorname{tr} \mathcal{C}P\mathcal{C}^{\mathrm{T}}$ subject to the constraint

$$\frac{1}{\alpha} (\mathcal{A} + M_1 K N_1 + M_2 L N_2) P (\mathcal{A} + M_1 K N_1 + M_2 L N_2)^{\mathrm{T}} - P + \frac{\delta^2}{1-\alpha} D_L D_L^{\mathrm{T}} = 0 \tag{8}$$

with respect to the matrix variables $0 \prec P \in \mathbb{S}^{2n}$, $K \in \mathbb{R}^{p\times n}$, and $L \in \mathbb{R}^{n\times \ell}$ and the scalar parameter $\alpha > 0$.

As the performance criterion we take the function

$$f(K, L, \alpha) = \operatorname{tr} \mathcal{C}P\mathcal{C}^{\mathrm{T}} + \chi_K \|K\|_F^2 + \chi_L \|L\|_F^2, \tag{9}$$

which incorporates, in addition to the component determining the size of the bounding ellipsoid by the trace criterion, penalties for the magnitude of the gain and observer matrices (the coefficients χ_K and $\chi_L > 0$ tune their significance). Their presence ensures the coercivity of the objective function in K and L; for details, see Section 4.

Remark 1. Note that the block matrix $A_{K,L}$ has the same eigenvalues as the matrices $A + BK$ and $A - LC_1$ standing on its diagonal. In turn, the existence of matrices K and L such that the matrices $A + BK$ and $A - LC_1$ are stable follows from the controllability and observability of the original system.

By Remark 1, there surely exist matrices K_0 and L_0 such that A_{K_0,L_0} is a Schur matrix. Matrices (K, L) with this property will be called a *stabilizing matrix pair.*

4. OPTIMIZATION OF THE FUNCTION $f(K, L, \alpha)$

Thus, the original problem of designing an observer-aided dynamic feedback controller rejecting the effect of exogenous disturbances on the output z_k of system (1) has been reduced to that of minimizing the objective function $f(K, L, \alpha)$ (9) subject to the constraint

$$\frac{1}{\alpha} A_{K,L} P A_{K,L}^{\mathrm{T}} - P + \frac{\delta^2}{1-\alpha} D_L D_L^{\mathrm{T}} = 0 \tag{10}$$

with respect to the matrix variables $P \in \mathbb{S}^{2n}$, $K \in \mathbb{R}^{p\times n}$, and $L \in \mathbb{R}^{n\times \ell}$ and the scalar parameter $\alpha > 0$. The notation $f(K, L, \alpha)$ emphasizes that given K, L, and α, the matrix P is found from the Lyapunov equation (10); therefore, K, L, and α are the independent variables.

The properties of the function $f(\alpha) = \operatorname{tr} \mathcal{C} P \mathcal{C}^{\mathrm{T}}$ established in [11] (under some fixed stabilizing pair (K, L)) can be fully transferred to the case under consideration. In particular, the function $f(\alpha)$ is well-defined, positive, and strictly convex on the interval $\rho^2(A_{K,L}) < \alpha < 1$, and its values tend to infinity at the ends of this interval.

The function $f(\alpha)$ can be effectively minimized using Newton's method. Let us specify some initial approximation $\rho^2(A_{K,L}) < \alpha_0 < 1$, e.g., $\alpha_0 = \left(1 + \rho^2(A_{K,L})\right)/2$, and apply the iterative process

$$\alpha_{j+1} = \alpha_j - \frac{f'(\alpha_j)}{f''(\alpha_j)}. \tag{11}$$

Here, according to [11],

$$f'(\alpha) = \operatorname{tr} Y \left(\frac{\delta^2}{(1-\alpha)^2} D_L D_L^{\mathrm{T}} - \frac{1}{\alpha^2} A_{K,L} P A_{K,L}^{\mathrm{T}} \right),$$

$$f''(\alpha) = 2\operatorname{tr} Y \left(\frac{\delta^2}{(1-\alpha)^3} D_L D_L^{\mathrm{T}} + \frac{1}{\alpha^3} A_{K,L} (P - X) A_{K,L}^{\mathrm{T}} \right),$$

with P, Y, and X representing the solutions of the discrete Lyapunov equations (10),

$$\frac{1}{\alpha} A_{K,L}^{\mathrm{T}} Y A_{K,L} - Y + \mathcal{C}^{\mathrm{T}} \mathcal{C} = 0,$$

and

$$\frac{1}{\alpha} A_{K,L} X A_{K,L}^{\mathrm{T}} - X + \frac{\delta^2}{(1-\alpha)^2} D_L D_L^{\mathrm{T}} - \frac{1}{\alpha^2} A_{K,L} P A_{K,L}^{\mathrm{T}} = 0,$$

respectively.

The next result ensures the global convergence of the algorithm.

Theorem 2 [11]. *In the method* (11), *we have the upper bounds*

$$|\alpha_j - \alpha^*| \leqslant \frac{f''(\alpha_0)}{2^j f''(\alpha^*)} |\alpha_0 - \alpha^*|, \qquad |\alpha_{j+1} - \alpha^*| \leqslant c|\alpha_j - \alpha^*|^2,$$

where $c > 0$ is some constant (which can be written explicitly).

Due to the first bound, the method converges globally (faster than a geometric progression with a ratio of 1/2); due to the second bound, it converges quadratically in the vicinity of the solution.

We pass to the minimization of the function

$$f(K, L) \doteq \min_{\alpha} f(K, L, \alpha),$$

after examining its properties.

Lemma 1 [1]. *The function $f(K, L)$ is well-defined and positive on the set $\mathcal{S}$ of all stabilizing matrix pairs.*

The definitional domain $\mathcal{S}$ of the function $f(K, L)$ may be nonconvex and disconnected and, as in the continuous-time case, its boundaries may be nonsmooth (see [12]).

Lemma 2. *The function $f(K, L, \alpha)$ is well-defined for $(K, L) \in \mathcal{S}$ and $\rho^2(A_{K,L}) < \alpha < 1$. On this admissible set, it is differentiable and the gradient is given by*

$$\nabla_\alpha f(K, L, \alpha) = \operatorname{tr} Y \left(\frac{\delta^2}{(1-\alpha)^2} D_L D_L^{\mathrm{T}} - \frac{1}{\alpha^2} A_{K,L} P A_{K,L}^{\mathrm{T}} \right),$$

$$\frac{1}{2} \nabla_K f(K, L, \alpha) = \frac{1}{\alpha} M_1^{\mathrm{T}} Y A_{K,L} P N_1^{\mathrm{T}} + \chi_K K, \tag{12}$$

$$\frac{1}{2} \nabla_L f(K, L, \alpha) = \frac{1}{\alpha} M_2^{\mathrm{T}} Y A_{K,L} P N_2^{\mathrm{T}} - \frac{\delta^2}{1-\alpha} \begin{pmatrix} 0 & I \end{pmatrix} Y D_L D_1^{\mathrm{T}} + \chi_L L, \tag{13}$$

where the matrix Y satisfies the discrete Lyapunov equation

$$\frac{1}{\alpha} A_{K,L}^{\mathrm{T}} Y A_{K,L} - Y + \mathcal{C}^{\mathrm{T}} \mathcal{C} = 0. \tag{14}$$

The function $f(K, L, \alpha)$ achieves minimum at an interior point of the admissible set under the necessary conditions

$$\nabla_K f(K, L, \alpha) = 0, \quad \nabla_L f(K, L, \alpha) = 0, \quad \nabla_\alpha f(K, L, \alpha) = 0.$$

In addition, $f(K, L, \alpha)$ as a function of α is strictly convex on $\rho^2(A_{K,L}) < \alpha < 1$ and achieves minimum at an interior point of this interval.

The proofs of this and subsequent statements are provided in the Appendix.

Next, to obtain simple quantitative estimates in Lemma 3, we embed a regularizing additive term ε into the objective function (9) as follows:

$$f(K, L, \alpha) = \operatorname{tr} P(\mathcal{C}^{\mathrm{T}} \mathcal{C} + \varepsilon I) + \chi_K \|K\|_F^2 + \chi_L \|L\|_F^2 \to \min_{K,L,\alpha}, \quad 0 < \varepsilon \ll 1.$$

The requirement for adding this term can be weakened considerably, but the current aim is to obtain the simplest and most straightforward results.

Lemma 3. *The function $f(K, L)$ is coercive on the set $\mathcal{S}$(i.e., tends to infinity on its boundary), and the following lower bounds hold:*

$$f(K, L) \geqslant \frac{\delta^2}{1 - \rho^2(A_{K,L})} \frac{\varepsilon}{1 - \sigma_{\min}^2(A_{K,L})} \|D_L\|_F^2, \tag{15}$$

$$f(K, L) \geqslant \chi_K \|K\|^2,$$

$$f(K, L) \geqslant \chi_L \|L\|^2.$$

We introduce the level set

$$\mathcal{S}_0 = \{(K, L) \in \mathcal{S} \colon \quad f(K, L) \leqslant f(K_0, L_0)\}.$$

Lemma 3 implies the following obvious result.

Corollary 1. *For any $(K_0, L_0) \in \mathcal{S}$ the set $\mathcal{S}_0$ is bounded.*

On the other hand, the function $f(K, L)$ achieves minimum on the set $\mathcal{S}_0$ (as a continuous function on a compact set, due to the properties of the solution of the Lyapunov equation), but the set $\mathcal{S}_0$ has no common points with the boundary of $\mathcal{S}$ by (15). According to the considerations above, $f(K, L)$ is differentiable on $\mathcal{S}_0$. Hence, we arrive at another fact.

Corollary 2. *There exists a minimum point (K_*, L_*) on the set S and the gradient of the function $f(K, L)$ vanishes at this point.*

The gradients of the function $f(K, L)$ with respect to K and L are not Lipschitz functions on the set S of all stabilizing controllers; however, it can be shown that they have this property on its subset S_0.

The obtained properties of the objective function and its derivatives allow us to build a minimization method and justify its convergence.

5. THE SOLUTION ALGORITHM

We propose the following iterative approach to solving problem (8)–(9). It is based on the alternate application of the gradient descent method with respect to the variables K and L and minimization with respect to the parameter α by Newton's method.

Algorithm for minimizing $f(K, L, \alpha)$:

Step 1. Specify values of the parameters $\varepsilon > 0$ and $0 < \tau_K, \tau_L < 1$ and an initial stabilizing matrix pair (K_0, L_0).
Calculate

$$\alpha_0 = \frac{1 + \rho^2(\mathcal{A} + M_1 K_0 N_1 + M_2 L_0 N_2)}{2}.$$

Step 2. At the jth iteration, we have K_j, L_j, and α_j known.
Calculate the gradient $H_j^K = \nabla_K f(K_j, L_j, \alpha_j)$.

Step 3. Make a step of the gradient descent method in K:

$$K_{j+1} = K_j - \gamma_j^K H_j^K,$$

with a step length $\gamma_j^K > 0$ selected by splitting γ_K until the following conditions hold:
a. K_{j+1} stabilizes the matrix $(\mathcal{A} + M_1 K N_1 + M_2 L_j N_2)/\sqrt{\alpha_j}$ (i.e., turns it into a Schur matrix);
b. $f(K_{j+1}) \leqslant f(K_j) - \tau_K \gamma_j^K \|H_j^K\|^2$.

Step 4. Using K_{j+1}, calculate the gradient $H_j^L = \nabla_L f(K_{j+1}, L_j, \alpha_j)$.

Step 5. Make a step of the gradient descent method in L:

$$L_{j+1} = L_j - \gamma_j^L H_j^L,$$

with a step length $\gamma_j^L > 0$ selected by splitting γ_L until the following conditions hold:
a. L_{j+1} stabilizes the matrix $(\mathcal{A} + M_1 K_{j+1} N_1 + M_2 L N_2)/\sqrt{\alpha_j}$;
b. $f(L_{j+1}) \leqslant f(L_j) - \tau_L \gamma_j^L \|H_j^L\|^2$.

Step 6. For the resulting matrices K_{j+1} and L_{j+1}, minimize $f(K_{j+1}, L_{j+1}, \alpha)$ with respect to α to obtain α_{j+1}. Go back to Step 2.
The *normalization condition* is given by

$$\|H_j^K\| \leqslant \varepsilon, \quad \|H_j^L\| \leqslant \varepsilon.$$

In this case, take the current pair (K_j, L_j) as an approximate solution and stop the algorithm.

An important point is the choice of the trial step of the gradient descent method. Here, it seems promising to use the following considerations. For some K, L, α, and $P \succ 0$, let

$$\frac{1}{\alpha}(\mathcal{A}+M_1KN_1+M_2LN_2)P(\mathcal{A}+M_1KN_1+M_2LN_2)^{\mathrm{T}}-P+\frac{\delta^2}{1-\alpha}D_LD_L^{\mathrm{T}}=0.$$

Consider the increment with respect to K:

$$K \to K-\gamma H^K, \quad H^K=\nabla_K f(K,L,\alpha),$$

and find for which γ the matrix $\mathcal{A}+M_1(K-\gamma H^K)N_1+M_2LN_2$ will remain stable (Schur).

To this end, it suffices to require that P remain the matrix of a quadratic Lyapunov function for $\mathcal{A}+M_1(K-\gamma H^K)N_1+M_2LN_2$, i.e.,

$$(\mathcal{A}+M_1(K-\gamma H^K)N_1+M_2LN_2)P(\mathcal{A}+M_1(K-\gamma H^K)N_1+M_2LN_2)^{\mathrm{T}}-P \prec 0,$$

or

$$\begin{pmatrix} P & \mathcal{A}+M_1(K-\gamma H^K)N_1+M_2LN_2 \\ (\mathcal{A}+M_1(K-\gamma H^K)N_1+M_2LN_2)^{\mathrm{T}} & P^{-1} \end{pmatrix} \succ 0.$$

The last relation can be written as

$$\begin{pmatrix} P & A_{K,L} \\ A_{K,L}^{\mathrm{T}} & P^{-1} \end{pmatrix} - \gamma \begin{pmatrix} 0 & M_1H^KN_1 \\ (M_1H^KN_1)^{\mathrm{T}} & 0 \end{pmatrix} \succ 0,$$

which yields, according to [14],

$$0<\gamma^K<\min_{\lambda_i>0}\lambda_i\left(\begin{pmatrix} P & A_{K,L} \\ A_{K,L}^{\mathrm{T}} & P^{-1} \end{pmatrix}, \begin{pmatrix} 0 & M_1H^KN_1 \\ (M_1H^KN_1)^{\mathrm{T}} & 0 \end{pmatrix}\right).$$

When optimizing the objective function with respect to the variable L, the trial step is selected by analogy:

$$0<\gamma^L<\min_{\lambda_i>0}\lambda_i\left(\begin{pmatrix} P & A_{K,L} \\ A_{K,L}^{\mathrm{T}} & P^{-1} \end{pmatrix}, \begin{pmatrix} 0 & M_2H^LN_2 \\ (M_2H^LN_2)^{\mathrm{T}} & 0 \end{pmatrix}\right),$$

where $H^L=\nabla_L f(K,L,\alpha)$.

6. AN ILLUSTRATIVE EXAMPLE: CONTROL OF A TWO-MASS SYSTEM

Consider a *two-mass system* consisting of two rigid bodies with masses m_1 and m_2 connected by a spring with an elasticity coefficient k. Assume that the system slides without friction along a fixed horizontal rod (see Fig. 1). A control action u is applied to the left body, and each of the bodies is subjected to a corresponding exogenous disturbance (w_1 or w_2).

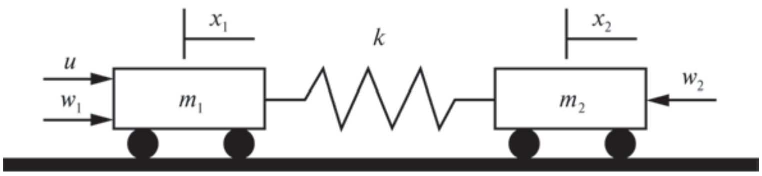


Fig. 1. A two-mass system.

Let x_1, v_1 and x_2, v_2 denote the coordinate and velocity of the left and right bodies, respectively. The observed output is the noisy two-dimensional vector

$$y = \begin{pmatrix} x_1 \\ x_2 + w_3 \end{pmatrix},$$

and the controlled (target) output is

$$z = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}.$$

Assume that the combined vector of exogenous disturbances is bounded at each time instant:

$$w = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}, \quad \|w\| \leqslant 0.2.$$

For the unit parameters of the two-mass system and the sampling interval $\Delta = 0.1$, we arrive at system (1) with the matrices

$$A = \begin{pmatrix} 0.9950 & 0.0050 & 0.0998 & 0.0002 \\ 0.0050 & 0.9950 & 0.0002 & 0.0998 \\ -0.0997 & 0.0997 & 0.9950 & 0.0050 \\ 0.0997 & -0.0997 & 0.0050 & 0.9950 \end{pmatrix}, \quad B = \begin{pmatrix} 0.0050 \\ 0.0000 \\ 0.0998 \\ 0.0002 \end{pmatrix}, \quad D = \begin{pmatrix} 0.0050 & 0 & 0 \\ 0 & 0.0050 & 0 \\ 0.0998 & 0.0002 & 0 \\ 0.0002 & 0.0998 & 0 \end{pmatrix},$$

$$C_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad C_2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

We specify an initial stabilizing pair of the form

$$K_0 = \begin{pmatrix} -6.9566 & 2.6036 & -4.1281 & -5.8682 \end{pmatrix}, \quad L_0 = \begin{pmatrix} 3.2037 & -0.6077 \\ 3.5790 & 0.3961 \\ 3.3299 & 0.1006 \\ -7.0446 & 5.3352 \end{pmatrix}$$

and let

$$\chi_K = \chi_L = 0.1.$$

In accordance with the algorithm above, the iterative procedure gives the gain matrix

$$K^* = \begin{pmatrix} -1.0311 & 0.6922 & -1.8064 & 0.0088 \end{pmatrix}, \quad \|K^*\| = 2.1921,$$

of the dynamic controller, the observer matrix

$$L^* = \begin{pmatrix} 1.1630 & -0.1759 \\ 1.8237 & 0.3432 \\ 2.0765 & 0.0714 \\ -3.7874 & 2.6783 \end{pmatrix}, \quad \|L^*\| = 5.2735,$$

and the corresponding bounding ellipse with the matrix

$$\mathcal{C}P^*\mathcal{C}^{\mathrm{T}} = \begin{pmatrix} 0.3974 & 0.1478 \\ 0.1478 & 0.4115 \end{pmatrix}, \quad \operatorname{tr} \mathcal{C}P^*\mathcal{C}^{\mathrm{T}} = 0.8089;$$

in this case, $\alpha^* = 0.9714$. The dynamics of the values of the objective function $f(K, L)$ are shown in Fig. 2.

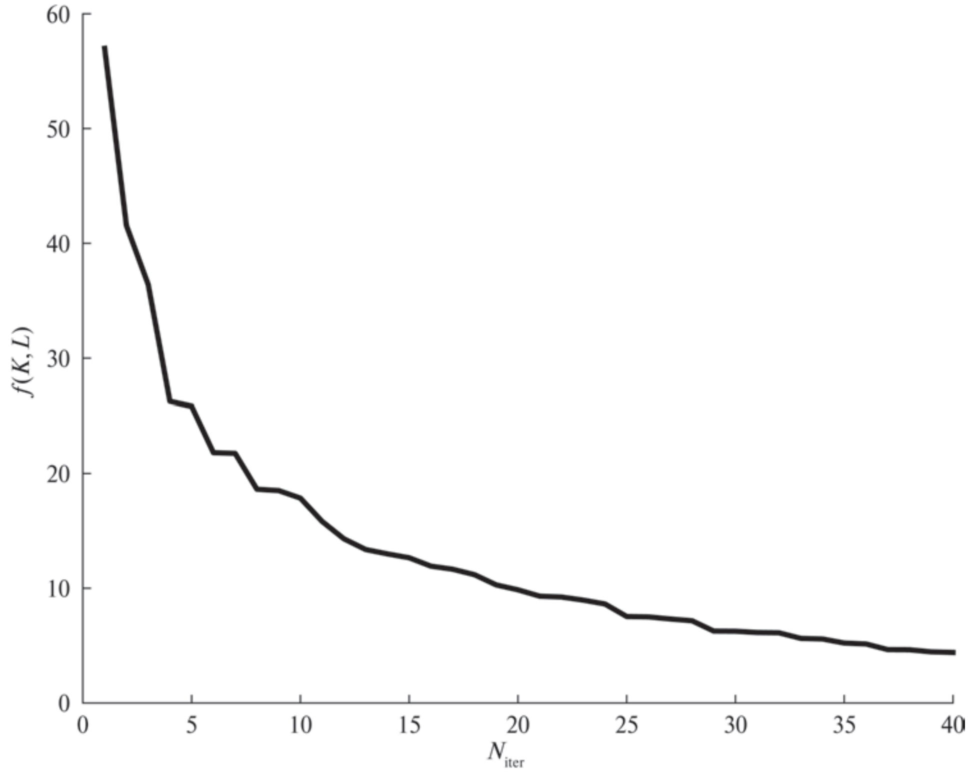


Fig. 2. The optimization procedure under $\chi_K = \chi_L = 0.1$.

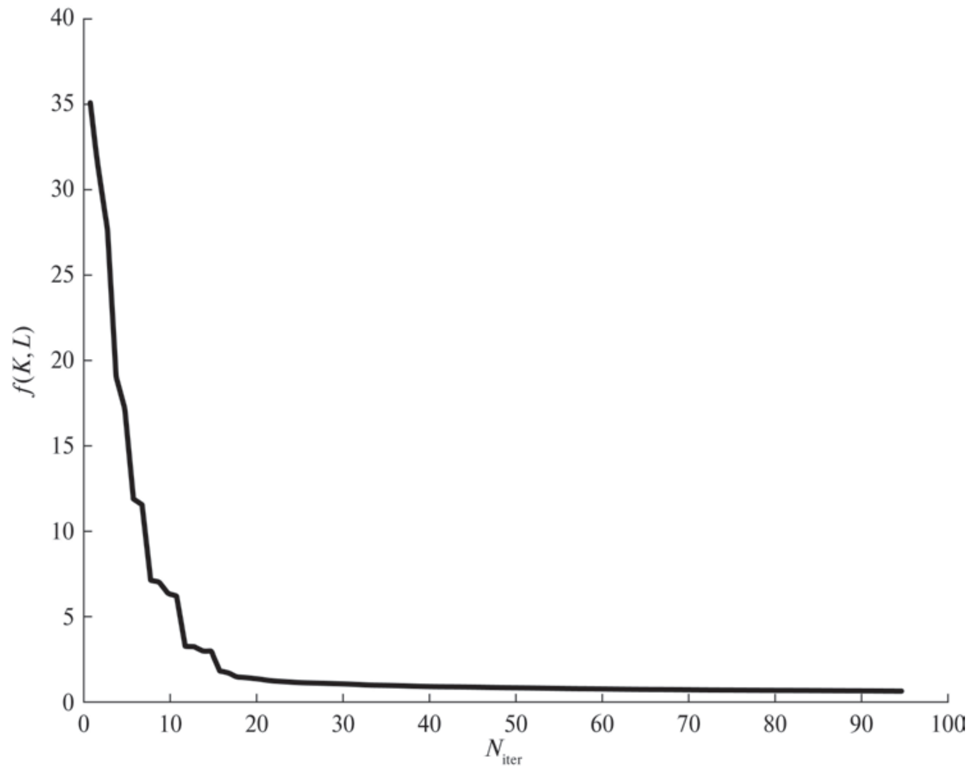


Fig. 3. The optimization procedure under $\chi_K = \chi_L = 0.01$.

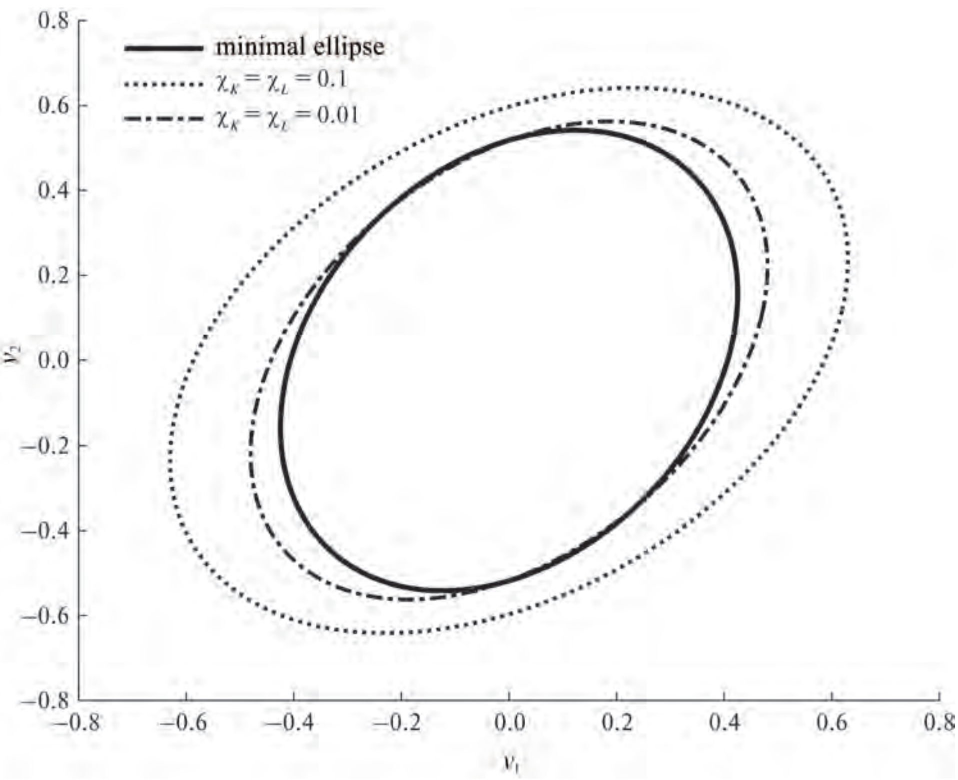


Fig. 4. Bounding ellipses.

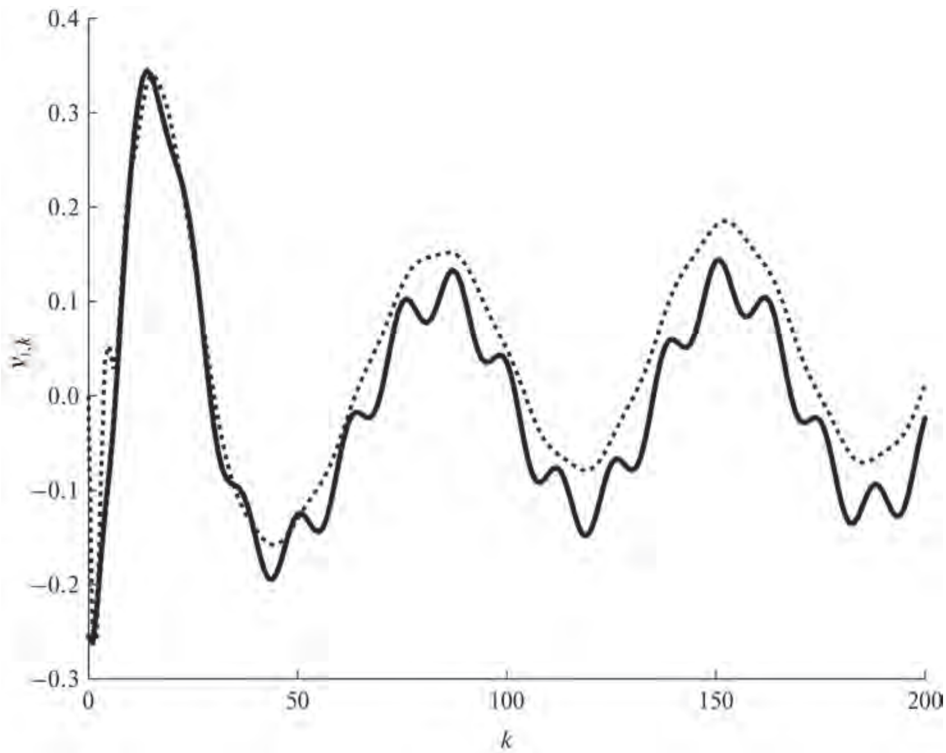


Fig. 5. The dynamics of the coordinate v_1 (solid line) and its estimate $\widehat{v}_1$ (dotted line).

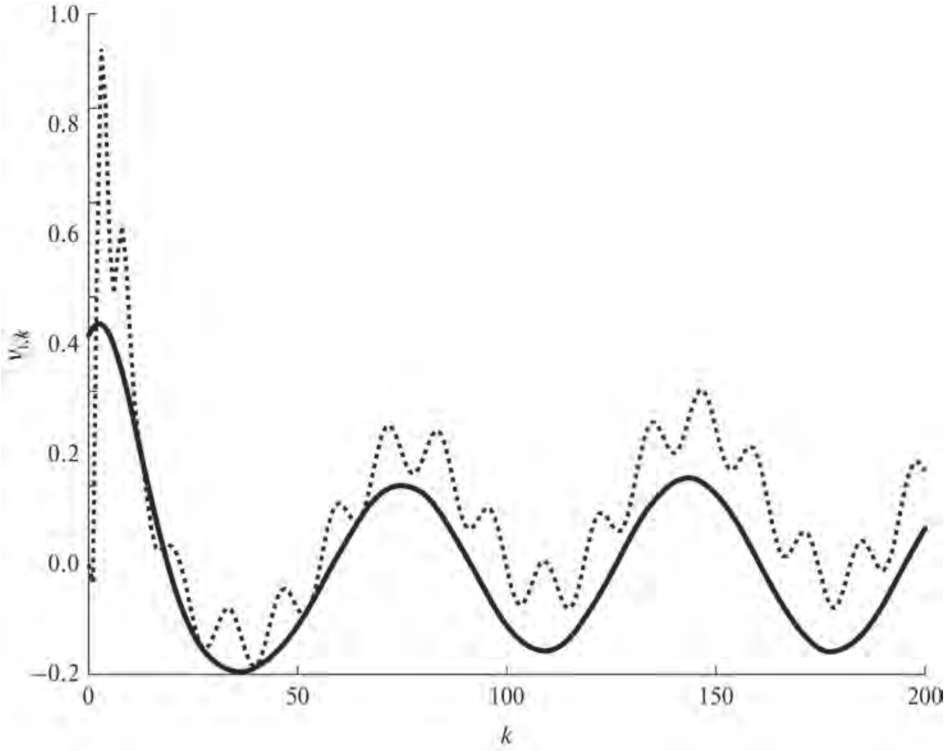


Fig. 6. The dynamics of the coordinate v_2 (solid line) and its estimate $\widehat{v}_2$ (dotted line).

Now, assigning

$$\chi_K = \chi_L = 0.01,$$

we obtain the gain matrix

$$K^{'*} = \begin{pmatrix} -1.3591 & 0.8130 & -2.5922 & 0.0186 \end{pmatrix}, \quad \|K^{'*}\| = 3.0378,$$

the observer matrix

$$L^{'*} = \begin{pmatrix} 0.6899 & -0.0066 \\ 2.7589 & 0.5861 \\ 3.0458 & 0.1396 \\ -5.3792 & 3.8014 \end{pmatrix}, \quad \|L^{'*}\| = 7.4014$$

and the corresponding bounding ellipse with the matrix

$$\mathcal{C}P^{'*}\mathcal{C}^{\mathrm{T}} = \begin{pmatrix} 0.2310 & 0.1056 \\ 0.1056 & 0.3163 \end{pmatrix}, \quad \operatorname{tr}\mathcal{C}P^{'*}\mathcal{C}^{\mathrm{T}} = 0.5473;$$

in this case, $\alpha^* = 0.9677$. The dynamics of the values of the objective function $f(K, L)$ are shown in Fig. 3.

Obviously, the reduction of the penalty coefficients produced an expected outcome: the size of the bounding ellipse decreased (by about one-third), at the cost of increasing the norm of the gain and observer matrices (by approximately 40%).

For comparison, we solved the same problem using the approach proposed in [10]; the calculations yielded the gain matrix

$$\widehat{K} = \begin{pmatrix} -15.2436 & 10.0353 & -10.7619 & -8.9984 \end{pmatrix}, \quad \|\widehat{K}\| = 23.0188,$$

the observer matrix

$$\widehat{L} = \begin{pmatrix} 3.0329 & 0.0004 \\ 14.9866 & 0.0818 \\ 20.9524 & 0.0080 \\ 6.6502 & 0.0405 \end{pmatrix}, \quad \|\widehat{L}\| = 26.7774,$$

and the corresponding minimal bounding ellipse with the matrix

$$\widehat{R} = \begin{pmatrix} 0.1807 & 0.0672 \\ 0.0672 & 0.2935 \end{pmatrix}, \quad \operatorname{tr} \widehat{R} = 0.4742.$$

Clearly, the bounding ellipse with the matrix $\mathcal{C}P'^{*}\mathcal{C}^{\mathrm{T}}$ exceeds the optimal one (with the matrix $\widehat{R}$) by merely 15%, while the corresponding matrices of the dynamic controller and observer differ in norm by several times (almost by an order of magnitude for the gain matrix).

The solid line in Fig. 4 presents the optimal bounding ellipse obtained by the method [10]; the dotted and dash-and-dot lines show the bounding ellipses found using the optimization procedure under $\chi_K = \chi_L = 0.1$ and $\chi_K = \chi_L = 0.01$, respectively.

Finally, Figs. 5 and 6 demonstrate the true trajectories of the system coordinates under some initial condition and an admissible exogenous disturbance (the solid lines) and their estimates provided by the constructed pair (K^*, L^*) (the dotted lines).

7. CONCLUSIONS

In this paper, we have proposed a novel approach to rejecting nonrandom bounded exogenous disturbances in linear discrete-time control systems by means of dynamic output feedback. The approach is based on reducing the original problem to a matrix optimization problem, further solved by the gradient descent method. Of course, it is possible to use appreciably faster first-order minimization methods (in particular, the conjugate gradient method). Detailed verification of more efficient methods will be the subject of future research; for now, it is important to check the fundamental feasibility and effectiveness of the novel approach. Its applications-oriented feature is the possibility of limiting the magnitude of the observer and gain matrices.

Further studies may aim at generalizing the proposed approach to various robust problem statements.

APPENDIX

Proof of Lemma 2. Differentiation with respect to α is performed according to the results of Section 4.

To differentiate the objective function (9) with respect to K under the discrete Lyapunov equation constraint

$$\frac{1}{\alpha} A_{K,L} P A_{K,L}^{\mathrm{T}} - P + \frac{\delta^2}{1-\alpha} D_L D_L^{\mathrm{T}} = 0 \tag{A.1}$$

for the matrix P of the invariant ellipsoid, we assign an increment ΔK to K and denote the corresponding increment of P by ΔP:

$$\frac{1}{\alpha}(\mathcal{A} + M_1(K + \Delta K)N_1 + M_2 L N_2)(P + \Delta P)(\mathcal{A} + M_1(K + \Delta K)N_1 + M_2 L N_2)^{\mathrm{T}}$$
$$-(P + \Delta P) + \frac{\delta^2}{1-\alpha} D_L D_L^{\mathrm{T}} = 0.$$

With the notation ΔP kept for the principal part of the increment, we have

$$\frac{1}{\alpha}(A_{K,L}PA_{K,L}^{\mathrm{T}} + A_{K,L}P(M_1\Delta KN_1)^{\mathrm{T}} + M_1\Delta KN_1PA_{K,L}^{\mathrm{T}} + A_{K,L}\Delta PA_{K,L}^{\mathrm{T}})$$
$$- (P + \Delta P) + \frac{\delta^2}{1-\alpha}D_L D_L^{\mathrm{T}} = 0.$$

Subtracting equation (A.1) from this expression yields

$$\frac{1}{\alpha}A_{K,L}\Delta PA_{K,L}^{\mathrm{T}} - \Delta P + \frac{1}{\alpha}(A_{K,L}P(M_1\Delta KN_1)^{\mathrm{T}} + M_1\Delta KN_1PA_{K,L}^{\mathrm{T}}) = 0. \tag{A.2}$$

Let us calculate the increment of the function $f(K, L, \alpha)$ with respect to K by linearizing the corresponding values:

$$\begin{aligned}\Delta_K f(K, L, \alpha) &= f(K + \Delta K, L, \alpha) - f(K, L, \alpha)\\ &= \operatorname{tr}\mathcal{C}(P + \Delta P)\mathcal{C}^{\mathrm{T}} + \chi_K\|K + \Delta K\|_F^2 - (\operatorname{tr}\mathcal{C}P\mathcal{C}^{\mathrm{T}} + \chi_K\|K\|_F^2)\\ &= \operatorname{tr}\mathcal{C}\Delta P\mathcal{C}^{\mathrm{T}} + \chi_K(\langle K + \Delta K, K + \Delta K\rangle - \langle K, K\rangle)\\ &= \operatorname{tr}\mathcal{C}\Delta P\mathcal{C}^{\mathrm{T}} + \chi_K(\operatorname{tr}K^{\mathrm{T}}\Delta K + \operatorname{tr}(\Delta K)^{\mathrm{T}}K) = \operatorname{tr}\Delta P\mathcal{C}^{\mathrm{T}}\mathcal{C} + 2\chi_K\operatorname{tr}K^{\mathrm{T}}\Delta K.\end{aligned}$$

From the dual Lyapunov equations (A.2) and (14) it follows that (see [11, Lemma A.1.1])

$$\operatorname{tr}\Delta P\mathcal{C}^{\mathrm{T}}\mathcal{C} = \frac{2}{\alpha}\operatorname{tr}YM_1\Delta KN_1PA_{K,L}^{\mathrm{T}}$$

and consequently,

$$\begin{aligned}\Delta_K f(K, L, \alpha) &= \frac{2}{\alpha}\operatorname{tr}YM_1\Delta KN_1PA_{K,L}^{\mathrm{T}} + 2\chi_K\operatorname{tr}K^{\mathrm{T}}\Delta K\\ &= 2\left\langle\frac{1}{\alpha}M_1^{\mathrm{T}}YA_{K,L}PN_1^{\mathrm{T}} + \chi_K K, \Delta K\right\rangle.\end{aligned}$$

Thus, we have established the relation (12).

Let us proceed to the differentiation of the function $f(K, L, \alpha)$ with respect to L: we assign an increment ΔL to L and denote the corresponding increment of P by ΔP:

$$\frac{1}{\alpha}(\mathcal{A} + M_1KN_1 + M_2(L + \Delta L)N_2)(P + \Delta P)(\mathcal{A} + M_1KN_1 + M_2(L + \Delta L)N_2)^{\mathrm{T}}$$
$$-(P + \Delta P) + \frac{\delta^2}{1-\alpha}\begin{pmatrix} D \\ D - (L + \Delta L)D_1\end{pmatrix}\begin{pmatrix} D \\ D - (L + \Delta L)D_1\end{pmatrix}^{\mathrm{T}} = 0.$$

With the notation ΔP kept for the principal part of the increment, we have

$$\frac{1}{\alpha}(A_{K,L}PA_{K,L}^{\mathrm{T}} + A_{K,L}P(M_2\Delta LN_2)^{\mathrm{T}} + M_2\Delta LN_2PA_{K,L}^{\mathrm{T}})$$
$$-(P + \Delta P) + \frac{\delta^2}{1-\alpha}\left[D_L D_L^{\mathrm{T}} - \begin{pmatrix} 0 \\ \Delta LD_1\end{pmatrix}D_L^{\mathrm{T}} - D_L\begin{pmatrix} 0 \\ \Delta LD_1\end{pmatrix}^{\mathrm{T}}\right] = 0.$$

Subtracting equation (A.1) from this expression yields

$$\frac{1}{\alpha}A_{K,L}\Delta PA_{K,L}^{\mathrm{T}} - \Delta P + \frac{1}{\alpha}(A_{K,L}P(M_2\Delta LN_2)^{\mathrm{T}} + M_2\Delta LN_2PA_{K,L}^{\mathrm{T}})$$
$$-\frac{\delta^2}{1-\alpha}\left[\begin{pmatrix} 0 \\ \Delta LD_1\end{pmatrix}D_L^{\mathrm{T}} + D_L\begin{pmatrix} 0 \\ \Delta LD_1\end{pmatrix}^{\mathrm{T}}\right] = 0. \tag{A.3}$$

We calculate the increment of the function $f(K, L, \alpha)$ with respect to L by linearizing the corresponding values:

$$\begin{gathered}\Delta_L f(K,L,\alpha) = f(K, L+\Delta L, \alpha) - f(K,L,\alpha) \\ = \operatorname{tr}\mathcal{C}(P+\Delta P)\mathcal{C}^{\mathrm{T}} + \chi_L\|L+\Delta L\|_F^2 - (\operatorname{tr}\mathcal{C}P\mathcal{C}^{\mathrm{T}} + \chi_L\|L\|_F^2) \\ = \operatorname{tr}\mathcal{C}\Delta P\mathcal{C}^{\mathrm{T}} + \chi_L(\langle L+\Delta L, L+\Delta L\rangle - \langle L, L\rangle) \\ = \operatorname{tr}\mathcal{C}\Delta P\mathcal{C}^{\mathrm{T}} + \chi_L(\operatorname{tr}L^{\mathrm{T}}\Delta L + \operatorname{tr}(\Delta L)^{\mathrm{T}}L) = \operatorname{tr}\Delta P\mathcal{C}^{\mathrm{T}}\mathcal{C} + 2\chi_L\operatorname{tr}L^{\mathrm{T}}\Delta L.\end{gathered}$$

From the dual Lyapunov equations (A.3) and (14) it follows that

$$\operatorname{tr}\Delta P\mathcal{C}^{\mathrm{T}}\mathcal{C} = 2\operatorname{tr}Y\left[\frac{1}{\alpha}M_2\Delta L N_2 P A_{K,L}^{\mathrm{T}} - \frac{\delta^2}{1-\alpha}\begin{pmatrix}0\\ \Delta L D_1\end{pmatrix}D_L^{\mathrm{T}}\right]$$

and consequently,

$$\begin{gathered}\Delta_L f(K,L,\alpha) \\ = 2\operatorname{tr}Y\left[\frac{1}{\alpha}M_2\Delta L N_2 P A_{K,L}^{\mathrm{T}} - \frac{\delta^2}{1-\alpha}\begin{pmatrix}0\\ \Delta L D_1\end{pmatrix}D_L^{\mathrm{T}}\right] + 2\chi_L\operatorname{tr}L^{\mathrm{T}}\Delta L \\ = 2\operatorname{tr}\left[\frac{1}{\alpha}N_2 P A_{K,L}^{\mathrm{T}} Y M_2 \Delta L - \frac{\delta^2}{1-\alpha}D_1 D_L^{\mathrm{T}} Y\begin{pmatrix}0\\ I\end{pmatrix}\Delta L\right] + 2\chi_L\operatorname{tr}L^{\mathrm{T}}\Delta L \\ = 2\left\langle \frac{1}{\alpha}M_2^{\mathrm{T}} Y A_{K,L} P N_2^{\mathrm{T}} - \frac{\delta^2}{1-\alpha}\begin{pmatrix}0 & I\end{pmatrix} Y D_L D_1^{\mathrm{T}} + \chi_L L, \Delta L\right\rangle,\end{gathered}$$

which finally gives formula (13).

The proof of Lemma 2 is complete.

Proof of Lemma 3. Consider a sequence of stabilizing matrix pairs $\{K_j, L_j\} \in \mathcal{S}$ such that

$$(K_j, L_j) \to (K, L) \in \partial\mathcal{S},$$

i.e., $\rho(A_{K,L}) = 1$. In other words, for any $\epsilon > 0$ there exists a number $N = N(\epsilon)$ such that

$$|\rho(A_{K_j,L_j}) - \rho(A_{K,L})| = 1 - \rho(A_{K_j,L_j}) < \epsilon \quad \text{for all } j \geqslant N(\epsilon).$$

Let P_j be the solution of the discrete Lyapunov equation (10) associated with the pair (K_j, L_j), i.e.,

$$\frac{1}{\alpha_j}A_{K_j,L_j}P_j A_{K_j,L_j}^{\mathrm{T}} - P_j + \frac{\delta^2}{1-\alpha_j}D_{L_j}D_{L_j}^{\mathrm{T}} = 0,$$

and Y_j be the solution of its dual Lyapunov equation

$$\frac{1}{\alpha_j}A_{K_j,L_j}^{\mathrm{T}} Y_j A_{K_j,L_j} - Y_j + \mathcal{C}^{\mathrm{T}}\mathcal{C} + \varepsilon I = 0.$$

By [12, Lemma A.1], we have

$$\begin{gathered}f(K_j, L_j) = \operatorname{tr}P_j(\mathcal{C}^{\mathrm{T}}\mathcal{C} + \varepsilon I) + \chi_K\|K_j\|_F^2 + \chi_L\|L_j\|_F^2 \geqslant \operatorname{tr}P_j(\mathcal{C}^{\mathrm{T}}\mathcal{C}+\varepsilon I) \\ = \operatorname{tr}Y_j\frac{\delta^2}{1-\alpha_j}D_{L_j}D_{L_j}^{\mathrm{T}} \geqslant \frac{\delta^2}{1-\alpha_j}\lambda_{\min}(Y_j)\operatorname{tr}(D_{L_j}D_{L_j}^{\mathrm{T}}) \geqslant \frac{\delta^2}{1-\alpha_j}\frac{\lambda_{\min}(\mathcal{C}^{\mathrm{T}}\mathcal{C}+\varepsilon I)}{1-\sigma_{\min}^2(A_{K_j,L_j})}\operatorname{tr}(D_{L_j}D_{L_j}^{\mathrm{T}}) \\ \geqslant \frac{\delta^2}{1-\rho^2(A_{K_j,L_j})}\frac{\varepsilon}{1-\sigma_{\min}^2(A_{K_j,L_j})}\|D_{L_j}\|_F^2 \geqslant \frac{\delta^2}{\epsilon}\frac{\varepsilon}{1-\sigma_{\min}^2(A_{K_j,L_j})}\|D_{L_j}\|_F^2 \xrightarrow[\epsilon\to 0]{} +\infty\end{gathered}$$

since

$$\rho^2(A_{K_j,L_j}) < \alpha_j < 1.$$

On the other hand,

$$\begin{aligned} f(K_j, L_j) &= \operatorname{tr} P_j(\mathcal{C}^{\mathrm{T}}\mathcal{C} + \varepsilon I) + \chi_K\|K_j\|_F^2 + \chi_L\|L_j\|_F^2 \\ &\geqslant \chi_K\|K_j\|_F^2 \geqslant \chi_K\|K_j\|^2 \xrightarrow[\|K_j\|\to+\infty]{} +\infty \end{aligned}$$

and

$$\begin{aligned} f(K_j, L_j) &= \operatorname{tr} P_j(\mathcal{C}^{\mathrm{T}}\mathcal{C} + \varepsilon I) + \chi_K\|K_j\|_F^2 + \chi_L\|L_j\|_F^2 \\ &\geqslant \chi_L\|L_j\|_F^2 \geqslant \chi_L\|L_j\|^2 \xrightarrow[\|L_j\|\to+\infty]{} +\infty. \end{aligned}$$

The proof of Lemma 3 is complete.

REFERENCES

1. Polyak, B.T. and Khlebnikov, M.V., Observer-Aided Output Feedback Synthesis as an Optimization Problem, *Autom. Remote Control*, 2022, vol. 83, no. 3, pp. 303–324.
2. Korovin, S.K. and Fomichev, V.V., *Nablyudateli sostoyaniya dlya lineinykh sistem s neopredelennost'yu* (State Observers for Linear Systems with Uncertainty), Moscow: Fizmatlit, 2007.
3. Luenberger, D.G., An Introduction to Observers, *IEEE Trans. Autom. Control*, 1971, vol. AC-16, no. 6, pp. 596–602.
4. 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.
5. Fatkhullin, I. and Polyak, B., Optimizing Static Linear Feedback: Gradient Method, *SIAM J. Control Optim.*, 2021, vol. 59, no. 5, pp. 3887–3911.
6. Polyak, B.T. and Topunov, M.V., Suppression of Bounded Exogenous Disturbances: Output Feedback, *Autom. Remote Control*, 2008, vol. 69, no. 5, pp. 801–818.
7. Boyd, S., El Ghaoui, L., Feron, E., et al., *Linear Matrix Inequalities in System and Control Theory*, Philadelphia: SIAM, 1994.
8. 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.
9. Peregudin, A. and Furtat, I., New Duality Relations in Linear Systems and Optimal Control under Bounded Disturbances, *IEEE Trans. Autom. Control*, 2024, vol. 69, no. 8, pp. 5569–5576.
10. Dogadin, E., Peregudin, A., and Shirokih, D., Optimal Control for Discrete-Time Systems under Bounded Disturbances, *Proc. 2024 American Control Conf. (ACC)*, Toronto, Canada, 2024, pp. 1940–1945.
11. Khlebnikov, M.V., A Comparison of Guaranteeing and Kalman Filters, *Autom. Remote Control*, 2023, vol. 84, no. 4, pp. 434–459.
12. 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.
13. Polyak, B.T., Khlebnikov, M.V., and Shcherbakov, P.S., Linear Matrix Inequalities in Control Systems with Uncertainty, *Autom. Remote Control*, 2021, vol. 82, no. 1, pp. 1–40.
14. Polyak, B.T. and Shcherbakov, P.S., The D-Decomposition Technique for Linear Matrix Inequalities, *Autom. Remote Control*, 2006, vol. 67, no. 11, pp. 1847–1861.

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, 2024, Vol. 85, No. 12, pp. 1149–1167.

STOCHASTIC SYSTEMS

AUV Positioning and Motion Parameter Identification Based on Observations with Random Delays

A. V. Bosov

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

Received June 20, 2024
Revised September 10, 2024
Accepted September 20, 2024

Abstract—A stochastic observation system model with random time delays between an arriving observation and the factual state of a moving object is adapted to identify its motion parameters. Equations for optimal Bayesian identification are given. A conditionally minimax nonlinear filter (CMNF) is applied to solve the problem in practice. The design procedure of the CMNF, including the choice of the filter structure, is discussed in detail on an example of autonomous underwater vehicle (AUV) positioning based on observations of stationary acoustic beacons. A computational experiment is carried out on a model close to practical needs using three variants of the filter, namely, the typical approximation of the updating process, the method of linear pseudomeasurements, and the geometric interpretation of angular measurements.

Keywords: nonlinear stochastic observation system, parameter identification, observations with random delays, conditionally minimax nonlinear filter, positioning, acoustic sensors, linear pseudomeasurements

DOI: 10.31857/S0005117924120029

1. INTRODUCTION

The theory of stochastic filtering provides methods and algorithms for solving various applied problems of estimation and control of moving objects [1]. In turn, applications may be sources of new formulations, models, and criteria, stimulating the improvement and development of the theory. One example is the applied field of autonomous underwater vehicles (AUVs) [2]. The peculiarities of the aquatic environment affect not only the nature and goals of the motion but also measuring means. For instance, different measuring means can be used for AUV positioning [3]; however, they all involve common physical laws. They are acoustic sensors, i.e., their operation depends on many factors, from water temperature to salinity and pressure [4]. As a result, data on the state of the observed AUV arrive with a random delay. This factor is not considered in the analogous problem of aircraft navigation because the high velocity of radio wave propagation allows one to neglect it. Indeed, the models of radar surveillance systems assume that data on the current position of an aircraft are acquired at the current time instant. However, the velocity of acoustic wave propagation is not so high, so the delay may be quite large and cannot be neglected. The stochastic dynamic observation system model with the delay factor of the acoustic signal was proposed in [5] and supplemented in [6]. Formally, this model can be reduced to the unified representation of a nonlinear stochastic observation system; as a result, it becomes possible to write optimal Bayesian filtering relations for it [7] and apply well-known heuristic filtering methods, including the extended Kalman filter [8], particle filters [9], and different modifications of sigma-point filters [10]. Meanwhile, both the optimal filter and any heuristic estimates cannot be used in practice even on model examples since the factor of a random time delay between an arriving observation and the factual state of a moving object is compensated via augmenting the state vector by a value proportional to the delay time and the measurement frequency. Under near-realistic conditions, one obtains a huge dimension, with no prospect of handling it. The only effective tool to cope with delays is given by Pugachev's method of conditionally optimal filtering [11, 12] and its development, Pankov's conditionally minimax nonlinear filter (CMNF) [13]. This conclusion was confirmed by calculations [5, 6]. Continuing to investigate the CMNF in the time-delay model, this paper focuses on two issues. First, the quality of conditionally optimal filtering is largely determined by a reasonable choice of the filter structure. Universal recommendations [14] can strongly lose to solutions considering the problem's specifics. This fact was demonstrated, in particular, in [5, 6], where the filter was successfully designed based on a simple geometric interpretation of measurements. The same idea is employed below not for a simplified model example but for that with real application content, and the results seem also good. However, such a variant of measurements (with direct position determination by solving a geometric problem) still has limited capabilities, and a more universal approach is needed here. Such an approach utilizes the linear pseudomeasurement method proposed in [15] and successfully developed in [16]. The corresponding results are presented in Section 4, including a stochastic observation system model describing the AUV positioning process using the angular measurements of two acoustic beacons and variants for choosing the CMNF structure that agree with the physical meaning of the problem. This example shows how angular measurements can improve the positioning accuracy considering the time delay of observations caused by the aquatic environment. The second issue concerns the traditional assumption for stochastic observation systems that the state model of a moving object is known. In practice, this assumption is incompletely valid, and the models are used with some errors. It is possible to compensate for the influence of such errors by including their model in the state equations. Many problems and methods of this kind have been created in the theory of robust estimation [17]. Another alternative is to describe the model inaccuracies by parameters, estimate them, and use the estimates together with the solution of the basic filtering problem. The corresponding methods belong to the theory of identification [18], regardless of whether this is possible and whether these parameters are estimated in advance or simultaneously with the estimation of the current position of the object. It seems rather natural to combine filtering and identification problems by forming an observation system model within the Bayesian approach. This variant of the CMNF with application to the AUV positioning problem is implemented below. Section 2 describes the formal model of a stochastic system with random time delays of measurements and the unknown parameters of the state model. In Section 3, we adapt the optimal filtering equations [5, 6] and the equations for the CMNF parameters used for positioning in the numerical experiment to the case of Bayesian parameter identification. In parallel with positioning, we identify the unknown motion parameters determining the average constant velocity of the AUV. Hence, it becomes possible to estimate the influence of the identification results of the motion model on the quality of solving the (main) positioning problem and, when using only angular position measurements, assess the fundamental possibility of estimating the motion velocity without its direct measurement, e.g., by Doppler sensors.

2. MOTION PARAMETER IDENTIFICATION MODEL BASED ON OBSERVATIONS WITH RANDOM DELAYS

The AUV motion and the set of measurements are described by a discrete stochastic dynamic system with unknown parameters. By assumption, the maximum possible time delay of arriving observations, T, is a priori given. Since the sound velocity in water can be supposed to be known, this assumption is reduced to estimating the maximum distance between the meters and the AUV, which should not be difficult. Denoting by $t = 0$ the initial positioning time instant, for the discrete time variable t we have $t = -T, -T+1, \ldots, 0,\ 1, \ldots.$

The AUV position, defined by a vector $x_t = (x_{1t}, \dots, x_{pt})' \in \mathbb{R}^p$ (e.g., the coordinates $(x(t), y(t), z(t))'$ and velocities $(v_x(t), v_y(t), v_z(t))'$ in an inertial reference frame linked to the Earth), and a random vector $\mu = (\mu_1, \dots, \mu_r)' \in \mathbb{R}^r$ of unknown motion model parameters form the state vector $(x_t', \mu')' \in \mathbb{R}^{p+r}$ of the dynamic system. Throughout this paper, "$'$" indicates the transpose of an appropriate vector or matrix. Let the distribution of μ be given.

The observation vector $y_t = (y_{1t}, \dots, y_{qt})' \in \mathbb{R}^q$ consists of measurements containing additive errors and time delays. Each observation element y_{it} has a particular shift described by a discrete random variable τ_{it} with values from the set $\{0, 1, \dots, T\}$. The values τ_{it} are combined into the vector $\tau_t = (\tau_{1t}, \dots, \tau_{qt})' \in \mathbb{R}^q$, which is a function of x_t. Exactly, each observation element y_{it} is a measurement performed for the position $x_{t-\tau_{it}}$. Thus, the observation system has the form

$$\begin{gathered} x_t = \varphi_t\left(x_{t-1}, \mu\right) + w_t, \quad x_{-T-1} = \eta, \\ y_{i_t} = \psi_{i_t}\left(x_{t-\tau_{i_t}}\right) + v_{i_t}, \quad i = 1, \dots, q, \\ \tau_t = \theta_t\left(x_t\right), \end{gathered} \tag{1}$$

where $w_t = (w_{1t}, \dots, w_{pt})' \in \mathbb{R}^p$ is a vector discrete white noise that models disturbances, $\eta \in \mathbb{R}^p$ are initial conditions, and $v_t = (v_{1t}, \dots, v_{qt})' \in \mathbb{R}^q$ is a vector discrete white noise that models measurement errors. The vectors η, μ, w_t, and v_t are independent in the aggregate, and the known functions φ_t, ψ_t, and θ_t satisfy sufficient conditions for the existence of filtering, Bayesian, and conditionally minimax estimates. These are typical conditions ensuring the existence of the second moment for the state and observation vectors, e.g., the linear growth of the system functions at infinity and the presence of the second moments for the disturbances and observation errors; see Theorems 1 and 2 in [5].

In model (1), we can introduce a compound state vector including both the AUV coordinates and the unknown random parameters of the motion model: $(x_t', \mu_t')'$. Then the observation system model takes the form

$$\begin{gathered} x_t = \varphi_t\left(x_{t-1}, \mu_{t-1}\right) + w_t, \quad x_{-T-1} = \eta, \\ \mu_t = \mu_{t-1}, \quad \mu_{-T-1} = \mu, \\ y_t = \psi_t\left(x_{t-\tau_t}\right) + v_t, \quad \tau_t = \theta_t\left(x_t\right). \end{gathered} \tag{2}$$

To make model (2) correct and usable as a formal representation of (1), we clarify that the notation $x_{t-\tau_t}$ is the vector $\left(x'_{1_{t-\tau_{1t}}}, \dots, x'_{p_{t-\tau_{qt}}}\right)'$ of the AUV positions corresponding to the time delay of each measurement included in the observation vector y_t.

In accordance with the Bayesian interpretation of the identification problem, it is required to specify a probability density for the initial condition vector $(\eta', \mu')'$. This density can be written as

$$f_0\left(X_{-T-1}, M\right) = f_\eta\left(X_{-T-1}\right) f_\mu(M), \tag{3}$$

where $X_{-T-1} \in \mathbb{R}^p$ and $M \in \mathbb{R}^r$ are the arguments of the probability densities corresponding to the random vectors η and μ, which have been supposed to be independent above.

Now, for the observation system (2), we can pose a filtering problem: it is required to estimate the state $(x_t', \mu_t')'$ based on the observations $y^t = (y_0', \dots, y_t')'$. Solving the filtering problem for the compound state vector yields a solution for both the positioning problem (estimation of the current position x_t) and the parameter identification problem (estimation of the vector μ).

3. FILTERING ALGORITHMS

Note that in the problem under consideration, the optimal Bayesian filter can be described by recurrence relations for the posterior probability density. Similar relations are well known for conventional discrete observation systems (without delays) [19]; they were generalized to the case of time-delayed systems in [7]. Clearly, the same result holds in the problem to be solved. The filter is written for the augmented position vector $\mathbf{x}_t = \left(x'_{t-T}, \ldots, x'_{t-1}, x'_t\right)'$ because all possible time delays need to be considered. For a less cumbersome notation, we simplify the time-delay model by assuming the scalar value $\tau_t \in \mathbb{R}^1$ instead of the vector $\tau_t \in \mathbb{R}^q$. This happens in the special case $\tau_{1t} = \ldots = \tau_{q_t}$. In other words, all observers are located at approximately the same distance from the AUV, e.g., when there is only one measuring complex.

It is possible to write the posterior probability density $\rho_t = \rho_t\left(\mathbf{X}_t, M \middle| Y^t\right)$ of the vector $(\mathbf{x}'_t, \mu')'$ with respect to the vector of all observations, y^t, and the filtering estimates of the position and parameters, x^*_t and μ^*_t. For this purpose, for the sake of convenience, we denote by $\mathbf{X}_t = \left(X'_{t-T}, \ldots, X'_{t-1}, X'_t\right)'$, $M = (M_1, \ldots, M_r)'$, and $Y^t = (Y'_0, \ldots, Y'_t)'$ the arguments of the posterior density; they correspond to the vectors $\mathbf{x}_t$, μ, and y^t, respectively. In addition, we employ the unnormalized posterior density $\widehat{\rho}_t = \widehat{\rho}_t\left(\mathbf{X}_t, M \middle| Y^t\right)$, so

$$\rho_t\left(\mathbf{X}_t, M \middle| Y^t\right) = \frac{\widehat{\rho}_t\left(\mathbf{X}_t, M \middle| Y^t\right)}{\int \widehat{\rho}_t\left(\mathbf{X}_t, M \middle| Y^t\right) d\mathbf{X}_t dM}.$$

Then we have [7]

$$\begin{gathered}\widehat{\rho}_t\left(\mathbf{X}_t, M \middle| Y^t\right) = f_{w_t}\left(X_t - \varphi_t\left(X_{t-1}, M\right)\right) \\ \times \sum_{i=0}^{T} I\left(\theta_t\left(X_t\right) = i\right) f_{v_t}\left(Y_t - \psi_t\left(X_{t-i}\right)\right) \int \rho_{t-1} dX_{t-T-1},\end{gathered} \tag{4}$$

where $f_{w_t}(\cdot)$ is the probability density of the disturbances w_t and $f_{v_t}(\cdot)$ is the probability density of the observation errors. Equation (4) is solved with the initial condition

$$\begin{gathered}\rho_{-1}\left(\mathbf{X}_{-1}, M \middle| Y^{-1}\right) = \rho_{-1}\left(\mathbf{X}_{-1}, M\right) = \rho_{-1}\left(X_{-T-1}, \ldots, X_{-1}, M\right) \\ = f_0\left(X_{-T-1}, M\right) f_{w_{-T}}\left(X_{-T} - \varphi_{-T}\left(X_{-T-1}, M\right)\right) \ldots f_{w_{-1}}\left(X_{-1} - \varphi_{-1}\left(X_{-2}, M\right)\right).\end{gathered}$$

The Bayesian estimates x^*_t and μ^*_t are obtained by integration:

$$\left((\mathbf{x}^*_t)', (\mu^*_t)'\right)' = \int \left(\mathbf{X}'_t, M'\right)' \rho_t\left(\mathbf{X}_t, M \middle| Y^t\right) d\mathbf{X}_t dM, \quad \mathbf{x}^*_t = \left(\ldots, (x^*_t)'\right)'.$$

According to the last equality, the current position estimate x^*_t enters as the last subvector in $\mathbf{x}^*_t$; also, the estimates of the past AUV positions have to be computed, although they are unnecessary.

The main conclusion from these considerations is that, despite the fundamental possibility of obtaining the optimal filtering and Bayesian identification estimates, we cannot reckon on their practical use. In the AUV positioning model described below, implementing calculations by formula (4) would have to deal with the dimensions $p = 3$, $r = 3$, $q = 4$, and $T = 15$, which gives the augmented vector $\mathbf{X}_t$ of dimension 45. For such dimensions, it seems at least rash to expect success in computing integrals with good accuracy under the real-time arrival of observations y_t.

Thus, a different approach to filtering is needed here to obtain a practically implementable estimate, albeit not optimal, but with good accuracy. Among many known approaches to suboptimal filtering, the concept of conditionally optimal filtering [11, 12] allows considering the specifics of the model of random observation delays without making cardinal changes. The possibility to flexibly change the filter structure based on the peculiarities of a particular observation system, originally envisaged by this concept, will make it possible to consider both the time delays and the physical laws determining the AUV motion and the operation of the sensors. The minimax explanation of the properties of this filter and the idea to calculate its parameters approximately by simulation [13] give grounds for obtaining good-quality estimates both in the positioning problem and in the motion parameter identification problem as the implementability of the corresponding algorithms with reasonable computational cost is its inherent property.

Following [13] and using the notations of this paper, we write the CMNF estimate $\left((\widehat{x}_t)', (\widehat{\mu}_t)'\right)'$ of the position x_t and the random motion model parameters μ based on the observations y^t in the prediction-correction form

$$\widehat{x}_t = \widetilde{x}_t + \Delta\widehat{x}_t, \qquad \widehat{\mu}_t = \widehat{\mu}_{t-1} + \Delta\widehat{\mu}_t.$$

The position prediction $\widetilde{x}_t$ is calculated using the basic prediction function $\xi_t(X, M)$; the correction, using the basic correction function $\zeta_t(X, Y)$, $X \in \mathbb{R}^p$, $M \in \mathbb{R}^r$, $Y \in \mathbb{R}^q$. The prediction $\widetilde{x}_t$ is calculated as a function of $\xi_t = \xi_t\left(\widehat{x}_{t-1}, \widehat{\mu}_{t-1}\right)$ whereas the correction $(\Delta\widehat{x}_t', \Delta\widehat{\mu}_t')'$ as a function of $\zeta_t = \zeta_t\left(\widetilde{x}_t, y_t\right)$. Note that the correction does not involve the parameter estimate $\widehat{\mu}_{t-1}$ (no argument M), which is explained by the absence of the parameters in the observation function.

The functions implementing $\widetilde{x}_t$ and $\Delta\widehat{x}_t$ are chosen linear so that

$$\begin{gathered}\widetilde{x}_t = F_t\xi_t\left(\widehat{x}_{t-1}, \widehat{\mu}_{t-1}\right) + f_t, \\ \begin{pmatrix}\widehat{x}_t \\ \widehat{\mu}_t\end{pmatrix} = \begin{pmatrix}\widetilde{x}_t \\ \widehat{\mu}_{t-1}\end{pmatrix} + H_t\zeta_t\left(\widetilde{x}_t, y_t\right) + h_t,\end{gathered} \tag{5}$$

where

$$\begin{gathered}F_t = \operatorname{cov}\left(x_t, \xi_t\right)\operatorname{cov}^+\left(\xi_t, \xi_t\right), \ f_t = \mathrm{E}\left\{x_t\right\} - F_t\mathrm{E}\left\{\xi_t\right\}, \\ H_t = \operatorname{cov}\left(\begin{pmatrix}x_t \\ \mu\end{pmatrix} - \begin{pmatrix}\widetilde{x}_t \\ \widehat{\mu}_{t-1}\end{pmatrix}, \zeta_t\right)\operatorname{cov}^+\left(\zeta_t, \zeta_t\right), \ h_t = -H_t\mathrm{E}\left\{\zeta_t\right\}.\end{gathered} \tag{6}$$

Formulas (6) have the following notations: $\mathrm{E}\{x\}$ is the expectation of the random vector x, $\operatorname{cov}(x, y)$ is the covariance of x and y, and "$^+$" is the Moore–Penrose pseudoinverse [20]. In addition, the position prediction $\widetilde{x}_t$ and the estimates $\widehat{x}_t$ and $\widehat{\mu}_t$ are unbiased with the estimation error covariances

$$\begin{gathered}\widetilde{K}_t = \operatorname{cov}\left(x_t - \widetilde{x}_t, x_t - \widetilde{x}_t\right) = \operatorname{cov}\left(x_t, x_t\right) - F_t\operatorname{cov}\left(\xi_t, x_t\right), \\ \widehat{K}_t^x = \operatorname{cov}\left(x_t - \widehat{x}_t, x_t - \widehat{x}_t\right) = \widetilde{K}_t - \Delta\widehat{K}_t^x, \qquad \widetilde{K}_0 = \operatorname{cov}(\eta, \eta), \\ \widehat{K}_t^\mu = \operatorname{cov}\left(\mu - \widehat{\mu}_t, \mu - \widehat{\mu}_t\right) = \widehat{K}_{t-1}^\mu - \Delta\widehat{K}_t^\mu, \qquad \widetilde{K}_0 = \operatorname{cov}(\mu, \mu),\end{gathered} \tag{7}$$

where the matrices $\Delta\widehat{K}_t^x$ and $\Delta\widehat{K}_t^\mu$ are the upper and lower diagonal blocks of the matrix $H_t\operatorname{cov}\left(\zeta_t, \begin{pmatrix}x_t \\ \mu\end{pmatrix} - \begin{pmatrix}\widetilde{x}_t \\ \widehat{\mu}_{t-1}\end{pmatrix}\right) = \begin{pmatrix}\Delta\widehat{K}_t^x & \dots \\ \dots & \Delta\widehat{K}_t^\mu\end{pmatrix}$.

The linear transformations (5) of the basic prediction ξ_t and correction ζ_t have minimax justification as the best estimates on the classes of all probability distributions with known mean and covariance [13]. In practical calculations, Monte Carlo estimates of the filter coefficients F_t, f_t, H_t, and h_t are used instead of their analytical values. In other words, $\mathrm{E}\{x\}$ in (6) is replaced by $\overline{\mathrm{E}}\{x\} = \frac{1}{N}\sum_{i=1}^N x_i$, where $\{x_i\}_{i=1}^N$ are the sample values of x simulated on a computer. Finally, sufficient conditions for the existence of the estimate (5) with respect to model (2) were formulated in [5, 6].

Thus, to solve the AUV positioning problem, it is required to specify motion and observation conditions and design the CMNF by selecting the basic prediction ξ_t and correction ζ_t.

4. AUV POSITIONING USING STATIONARY ACOUSTIC BEACONS

4.1. Motion Model

As has already been mentioned, the movement of an object in a water environment is the target (and currently single) application of the considered model of the system with random time delays of observations. Water acts as a natural source of delays in measurements made by various acoustic sensors. Different problems can be formulated using such measurements, including target goal-setting and intelligent target tracking, vehicle position estimation and prediction, motion model identification, individual and group movement planning, inertial navigation and visual positioning, and others. To demonstrate the flexibility and efficiency of the CMNF, we select the positioning problem of an AUV interacting with two stationary acoustic beacons. The observation model and possible approaches to solving the estimation problem in this formulation were discussed in detail in [21–23]. Here, this model is supplemented by introducing random time delays for the measurements of each beacon and inaccurate a priori information on the motion parameters, namely, the average velocity vector of the AUV. Thus, it is possible to analyze the standard application of the CMNF for solving the state filtering problem (determination of the AUV position) as well as positioning together with parameter identification.

The AUV motion model assumes that on average the vehicle moves at a constant velocity, while the real velocity is affected by uncontrolled random factors. Their influence leads to independent deviations of the velocity, which remains constant on sampling intervals (between successive measurements) and changes at each next time instant. The same in-plane motion was used in [5] but under the assumption of known motion parameters. Here, we will identify them.

The AUV motion is described in the Cartesian reference frame $Oxyz$; its choice will be discussed below. The typical notations $x(t)$, $y(t)$, and $z(t)$ are employed for the motion trajectory coordinates; they are measured in kilometers (km). Note the distinction between these notations and $x_t,\ y_t$ used above for the observation system in the general model (1). AUV positioning starts at the time instant $t = 0$ and is performed at discrete time instants with steps δ: $\delta, 2\delta, \ldots, t\delta, \ldots$. The AUV motion starts at the time instant $-T\delta$, i.e., $t = -T$, so a measurement with any permissible delay $\tau_{0i} \leqslant T$ can be realized at the time instant $t = 0$; therefore, the AUV initial position is given by the vector $(x(-T-1), y(-T-1),\ x(-T-1))'$. In the calculations carried out, this vector is supposed to obey the Gaussian distribution with the mean $(-1, -1, 1)'$ and the covariance $\operatorname{diag}\{0.1^2; 0.1^2; 0.1^2\}$. By assumption, the AUV moves at an unknown constant velocity $(v_x, v_y, v_z)'$, which is affected by uncontrolled factors. The absence of accurate velocity information is modeled by making the vector $(v_x, v_y, v_z)'$ Gaussian with the mean $\mathrm{E}\{v_x\} = -25$ km/h, $\mathrm{E}\{v_y\} = -12.5$ km/h, $\mathrm{E}\{v_z\} = -1$ km/h and the covariance $\operatorname{diag}\{5^2, 5^2, 1\}$. Thus, each trajectory has a particular "target" motion (direction and velocity) and, along with positioning (estimation of the position $(x(t), y(t), z(t))'$), it is required to identify the parameters of this motion given by the realization of the vector $(v_x, v_y, v_z)'$.

The uncontrolled random factors affecting the velocity of motion are modeled by additive disturbances $w_x(t)$, $w_y(t)$, and $w_z(t)$; by assumption, the vector $(w_x(t), w_y(t), w_z(t))'$ obeys the Gaussian distribution with the mean $(0,\ 0,\ 0)'$ and the covariance $\operatorname{diag}\{25^2, 25^2, 25^2, 25^2\}$. Thus, we have the state vector $(x_t', \mu')' = (x(t), y(t), z(t), v_x, v_y, v_z) \in \mathbb{R}^{3+3}$ and the following dynamics:

$$\begin{cases} x(t) = x(t-1) + \delta(v_x + w_x(t)), \\ y(t) = y(t-1) + \delta(v_y + w_y(t)), \\ z(t) = z(t-1) + \delta(v_z + w_z(t)), \\ t = -T, -T+1, \ldots, 0, 1, \ldots. \end{cases} \tag{8}$$

For the other calculation parameters, as in [5], we set a sampling step of $\delta = 0.0001$ hours (h) for observations (i.e., about three measurements per second (s); positioning is performed for 1000 sampling steps, $t = -T, \ldots, 1000$, (i.e., 6 minutes (min) of motion). During this time, the AUV traverses an average distance of about 2.5–3 km, with the maximum range to the farther of the two beacons being 8 km. Accordingly, the maximum time delay is supposed to be $T = 15$, constituting 0.0015 h or 5.4 s. This assumption is based on the sound velocity in water $v_s = 5400$ km/h (1500 m/s).

4.2. Measuring Complex Model

Observers are two stationary acoustic beacons (passive acoustic devices for estimating the direction of arrival (DOA) [24]) installed in advance. Let the first ($\mathcal{F}$, $first$) and second ($\mathcal{S}$, *second*) beacons have the coordinates $(X_{\mathcal{F}},\ Y_{\mathcal{F}}, Z_{\mathcal{F}})$ and $(X_{\mathcal{S}},\ Y_{\mathcal{S}}, Z_{\mathcal{S}})$, respectively. Following the model proposed in [22], observations of the positioned AUV with the unknown coordinates $(X,\ Y, Z)$ are the directions to each of the beacons, which yield two angles

$$\begin{cases} \tan\varphi_{\mathcal{F}} = \dfrac{Y_{\mathcal{F}} - Y}{X_{\mathcal{F}} - X}, \quad \tan\varphi_{\mathcal{S}} = \dfrac{Y_{\mathcal{S}} - Y}{X_{\mathcal{S}} - X}, \\ \tan\lambda_{\mathcal{F}} = \dfrac{Z_{\mathcal{F}} - Z}{|X_{\mathcal{F}} - X|}\cos\varphi_{\mathcal{F}}, \quad \tan\lambda_{\mathcal{S}} = \dfrac{Z_{\mathcal{S}} - Z}{|X_{\mathcal{S}} - X|}\cos\varphi_{\mathcal{S}}. \end{cases} \tag{9}$$

Figure 1 shows the geometric interpretation of the measured angles, particularly how the angles $\varphi_{\mathcal{F}}$, $\varphi_{\mathcal{S}}$, $\lambda_{\mathcal{F}}$, and $\lambda_{\mathcal{S}}$ can be counted to determine the mutual position of the beacon and the positioned AUV in order to correctly find the vehicle's position based on the available tangent measurements and consider the possible crossing of the lines $X_{\mathcal{F}} - X = 0$ and $X_{\mathcal{S}} - X = 0$ when some measurements cannot be used. In this figure, positions 1-2-3-4 of the AUV (X_1, Y_1, Z_1) and the beacon $(X_{\mathcal{M}}, Y_{\mathcal{M}}, Z_{\mathcal{M}})$ correspond to possible combinations of the coordinates $X_1 < X_{\mathcal{M}}$, $X_1 > X_{\mathcal{M}}$, $Y_1 < Y_{\mathcal{M}}$, $Y_1 > Y_{\mathcal{M}}$, and the angles φ and λ are counted so that relation (9) holds in all variants.

In the experiment described below, assuming the cooperative scenario (the beacons and the AUV act jointly to fulfill their common task), it is possible to simplify the relations slightly by

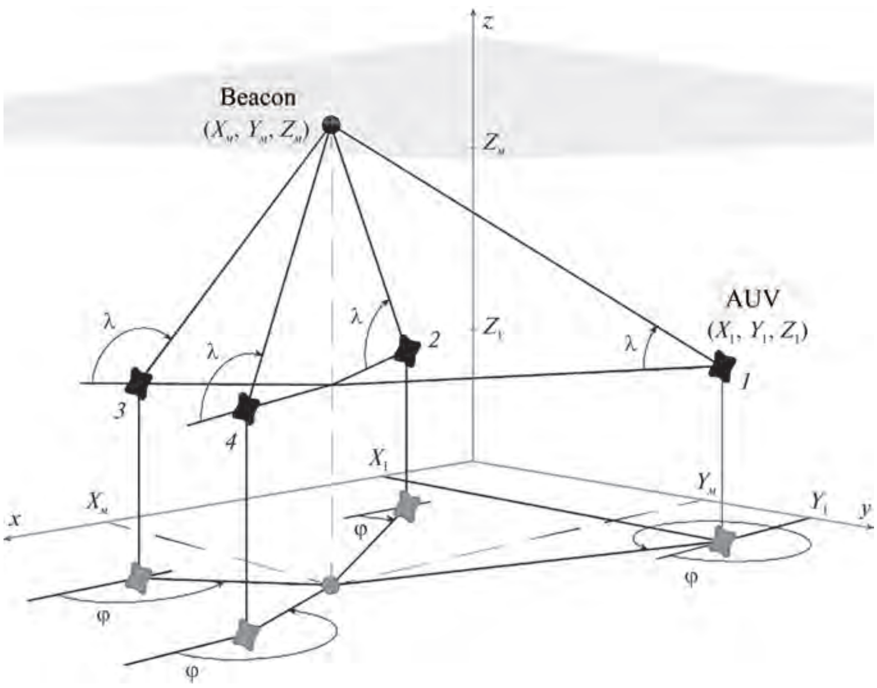


Fig. 1. Possible mutual arrangement of the AUV and beacons.

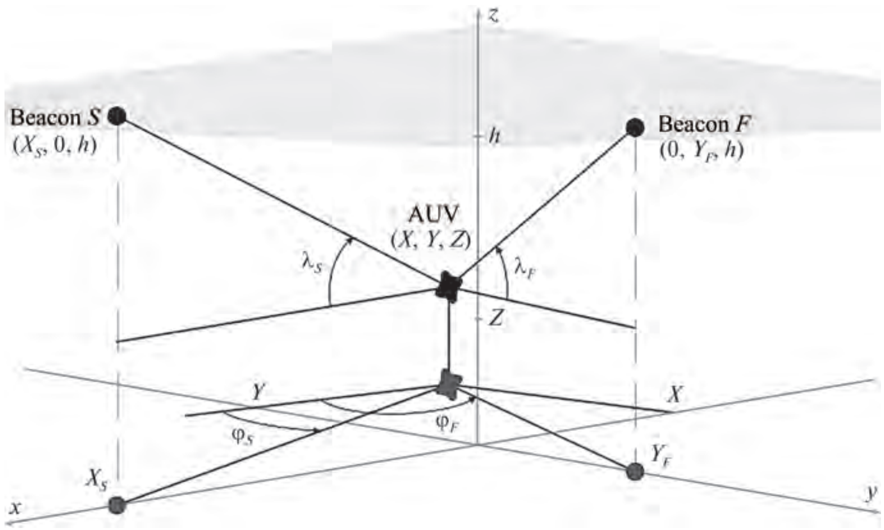


Fig. 2. Mutual arrangement of the AUV and two beacons in the experiment.

choosing a reference frame so that $(X_{\mathcal{F}}, Y_{\mathcal{F}}, Z_{\mathcal{F}}) = (0, Y_{\mathcal{F}}, h)$ and $(X_{\mathcal{S}}, Y_{\mathcal{S}}, Z_{\mathcal{S}}) = (X_{\mathcal{S}}, 0, h)$, where h is the depth at the location of the beacons (Fig. 2). Also, we may assume that $X_{\mathcal{F}} > 0$, $Y_{\mathcal{F}} > 0$ and the origin is located so that $X < 0$, $Y < 0$ throughout the entire motion time. Then all the measured angles $\varphi_{\mathcal{F}}$, $\varphi_{\mathcal{S}}$, $\lambda_{\mathcal{F}}$, and $\lambda_{\mathcal{S}}$ will vary in the range $\left(0, \frac{\pi}{2}\right)$, so $\cos\varphi_{\mathcal{F}} = \frac{1}{(1+\tan^2\varphi_{\mathcal{F}})^{1/2}}$ and $\cos\varphi_{\mathcal{S}} = \frac{1}{(1+\tan^2\varphi_{\mathcal{S}})^{1/2}}$. Thus, we arrive at the following relations of the measurements instead of (9):

$$\begin{cases} \tan\varphi_{\mathcal{F}} = \dfrac{Y - Y_{\mathcal{F}}}{X}, \quad \tan\varphi_{\mathcal{S}} = \dfrac{Y}{X - X_{\mathcal{S}}}, \\ \tan\lambda_{\mathcal{F}} = \dfrac{Z - h}{X}\cos\varphi_{\mathcal{F}}, \quad \tan\lambda_{\mathcal{S}} = \dfrac{Z - h}{X - X_{\mathcal{S}}}\cos\varphi_{\mathcal{S}}. \end{cases} \tag{10}$$

In the calculations carried out, the coordinates of the beacons are given by $Y_{\mathcal{F}} = 1$ km and $X_{\mathcal{S}} = 2$ km, the depth is $h = 2$ km. Accordingly, with the expected initial position $(-1, -1, 1)'$ and average velocity $(-25; -12.5; -1)'$ of the AUV chosen above, the vehicle will move away from both beacons and maintain negative coordinates $x(t), y(t)$ and positive coordinates $z(t)$ on average during 6 min of its positioning. Increasing the distance to the observers means increasing the values taken by the time delay τ_t.

Therefore, it remains to consider the measurement errors in (10). According to model (1), they are additive, so the factual model of the measuring complex without time delays in arriving observations has the form

$$\begin{cases} y_{1t} = \dfrac{y(t) - Y_{\mathcal{F}}}{x(t)} + v_1(t), \quad y_{2t} = \dfrac{z(t) - h}{x(t)} \dfrac{1}{\left(1 + \tan^2\varphi_{\mathcal{F}}(t)\right)^{1/2}} + v_2(t), \\ y_{3t} = \dfrac{y(t)}{x(t) - X_{\mathcal{S}}} + v_3(t), \quad y_{4t} = \dfrac{z(t) - h}{x(t) - X_{\mathcal{S}}} \dfrac{1}{\left(1 + \tan^2\varphi_{\mathcal{S}}(t)\right)^{1/2}} + v_4(t), \\ \tan\varphi_{\mathcal{F}}(t) = \dfrac{y(t) - Y_{\mathcal{F}}}{x(t)}, \quad \tan\varphi_{\mathcal{S}}(t) = \dfrac{y(t)}{x(t) - X_{\mathcal{S}}}, \end{cases} \tag{11}$$

where $y_t = (y_{1t}, y_{2t}, y_{3t}, y_{4t})'$ is the vector of the measured angular directions $\tan\varphi_{\mathcal{F}}$, $\tan\lambda_{\mathcal{F}}$, $\tan\varphi_{\mathcal{S}}$, and $\tan\lambda_{\mathcal{S}}$; the AUV coordinates $x(t)$, $y(t)$, and $z(t)$ are given by model (8), and the vector

$(v_1(t), v_2(t), v_3(t), v_4(t))'$ of the additive measurement errors obeys the Gaussian distribution with the mean $(0,0,0,0)'$ and the covariance $\operatorname{diag}\{(0,01)^2, (0,01)^2, (0,01)^2, (0,01)^2\}$. In the described experiment, the accuracy characteristics of the simulated acoustic sensors are chosen so that differences in the estimates can be visualized while maintaining an acceptable quality of all algorithms. The relative accuracy of real devices was discussed in [25].

To consider the observation delays τ_t in (11), we define two random functions $\tau_{\mathcal{F}}(t)$ and $\tau_{\mathcal{S}}(t)$, which will model the acoustic signal delay in time cycles, i.e., in the sampling steps δ, under the constant sound velocity v_s (see the assumption above). Note that for the problem under study, this simplified assumption is sufficient. In practice, it can be replaced without much difficulty by a more precise relation, e.g., the value obtained using the algorithm [26] or other simpler approximations [27]. Knowing the AUV current position $(x(t), y(t), z(t))'$, we calculate the ranges to the beacons located at the points $(0, Y_{\mathcal{F}}, h)$ and $(X_{\mathcal{S}}, 0, h)$, and the delays $\tau_{\mathcal{F}}(t)$ and $\tau_{\mathcal{S}}(t) \in \{0, 1, \ldots, T\}$ in time units of the motion model (8):

$$\begin{cases} \tau_{\mathcal{F}}(t) = \min\left\{T, \left[\dfrac{\sqrt{(x(t))^2 + (y(t) - Y_{\mathcal{F}})^2 + (z(t) - h)^2}}{(\delta v_s)}\right]\right\}, \\ \tau_{\mathcal{S}}(t) = \min\left\{T, \left[\dfrac{\sqrt{(x(t) - X_{\mathcal{S}})^2 + (y(t))^2 + (z(t) - h)^2}}{(\delta v_s)}\right]\right\}. \end{cases} \tag{12}$$

In (12) we use the notation $[x]$ for the floor function of x and the minimum to make the delays formally correspond to the model (1) and not exceed the given threshold T. In the calculations, $T = 15$ (large enough) and this minimum constraint was never activated on the simulated trajectories.

Thus, the observation model takes the final form

$$\begin{cases} y_{1t} = \dfrac{x_{2t-\tau_{\mathcal{F}}(t)} - Y_{\mathcal{F}}}{x_{1t-\tau_{\mathcal{F}}(t)}} + v_1(t),\ y_{2t} = \dfrac{x_{3t-\tau_{\mathcal{F}}(t)} - h}{x_{1t-\tau_{\mathcal{F}}(t)}} \dfrac{1}{(1 + \tan^2 \varphi_{\mathcal{F}}(t - \tau_{\mathcal{F}}(t)))^{1/2}} + v_2(t), \\ y_{3t} = \dfrac{x_{2t-\tau_{\mathcal{S}}(t)}}{x_{1t-\tau_{\mathcal{S}}(t)} - X_{\mathcal{S}}} + v_3(t),\ y_{4t} = \dfrac{x_{3t-\tau_{\mathcal{S}}(t)} - h}{x_{1t-\tau_{\mathcal{S}}(t)}} \dfrac{1}{(1 + \tan^2 \varphi_{\mathcal{S}}(t - \tau_{\mathcal{S}}(t)))^{1/2}} + v_4(t), \\ \tan \varphi_{\mathcal{F}}(t) = \dfrac{x_{2t} - Y_{\mathcal{F}}}{x_{1t}}, \quad \tan \varphi_{\mathcal{S}}(t) = \dfrac{x_{2t}}{x_{1t} - X_{\mathcal{S}}}. \end{cases} \tag{13}$$

To reduce the delay time in (13) to the form of the original model (1), in which $\tau_t \in \mathbb{R}^4$, we have to use (12) and let $\tau_{1t} = \tau_{2t} = \tau_{\mathcal{F}}(t)$ and $\tau_{3t} = \tau_{4t} = \tau_{\mathcal{S}}(t)$.

4.3. CMNF Design

The filter design procedure consists in selecting its structural functions ξ_t, ζ_t and carrying out computer simulation to calculate the coefficients by formulas (6) and determine the accuracy by formulas (7). For simplicity, we will not introduce separate notations for the CMNF estimates corresponding to different variants of the structural functions, all estimates are indicated in the same way: $\widehat{x}_t$, $\widehat{\mu}_t$, like the predictions $\widetilde{x}_t$. The particular structure of the filter is specified by its name (typical, geometric, or pseudomeasurements).

The three filter variants proposed here involve the basic prediction due to system (8), i.e., $\xi_t \in \mathbb{R}^3$, $t = 0, 1, \ldots$, and have the form

$$\xi_{1t} = X_1 + \delta M_1, \qquad \xi_{2t} = X_2 + \delta M_2, \qquad \xi_{3t} = X_3 + \delta M_3. \tag{14}$$

The motion equations (8) are linear, which does not affect the presence of parameters to be identified, so the basic prediction (14) is used in all CMNF variants and it makes sense to discuss only the correction process.

In the first variant of the filter, called *typical*, the basic correction is given by the observation residual adjusted by the estimate of the observation delay, i.e., $\zeta_t \in \mathbb{R}^4$, $t = 0, 1, \ldots$. It has the form

$$\begin{cases} \zeta_{1t} = Y_1 - \dfrac{X_2 - Y_{\mathcal{F}}}{X_1}, \quad \zeta_{2t} = Y_2 - \dfrac{X_3 - h}{X_1} \dfrac{1}{(1 + \tan^2 \widetilde{\varphi}_{\mathcal{F}})^{1/2}}, \\ \tan \widetilde{\varphi}_{\mathcal{F}} = \dfrac{X_2 - Y_{\mathcal{F}}}{X_1}, \quad (X_1, X_2, X_3)' = \widetilde{x}_{t - \widehat{\tau}_{\mathcal{F}}(t)}, \\ \zeta_{3t} = Y_3 - \dfrac{X_2}{X_1 - X_{\mathcal{S}}}, \quad \zeta_{4t} = Y_4 - \dfrac{X_3 - h}{X_1} \dfrac{1}{(1 + \tan^2 \widetilde{\varphi}_{\mathcal{S}})^{1/2}}, \\ \tan \widetilde{\varphi}_{\mathcal{S}} = \dfrac{X_2}{X_1 - X_{\mathcal{S}}}, \quad (X_1, X_2, X_3)' = \widetilde{x}_{t - \widehat{\tau}_{\mathcal{S}}(t)}, \\ \widehat{\tau}_{\mathcal{F}}(t) = \min \left\{ T, \left[\dfrac{\sqrt{(X_1)^2 + (X_2 - Y_{\mathcal{F}})^2 + (X_3 - h)^2}}{(\delta v_s)} \right] \right\} \quad (X_1, X_2, X_3)' = \widetilde{x}_t, \\ \widehat{\tau}_{\mathcal{S}}(t) = \min \left\{ T, \left[\dfrac{\sqrt{(X_1 - X_{\mathcal{S}})^2 + (X_2)^2 + (X_3 - h)^2}}{(\delta v_s)} \right] \right\} \quad (X_1, X_2, X_3)' = \widetilde{x}_t. \end{cases} \tag{15}$$

Note that to form the residual, the prediction-related argument of the basic correction ζ_t must contain not only the last position prediction $\widetilde{x}_t$ calculated using (5) but also all previous ones $\widetilde{x}_{t-1}, \ldots, \widetilde{x}_{t-T}$ since the time delay estimates $\widehat{\tau}_{\mathcal{F}}(t)$ and $\widehat{\tau}_{\mathcal{S}}(t)$ in (15) may take any value from 0 to T. The prediction shift is needed to determine the residual from the estimate of the state associated with the current observation y_t, i.e., from the delayed state. This state can be estimated by the values $\widehat{\tau}_{\mathcal{F}}(t)$ and $\widehat{\tau}_{\mathcal{S}}(t)$ for the first and second pair of the measured angles, respectively. In this case, the delay estimates can be calculated using the current position prediction $\widetilde{x}_t$ since the AUV velocity is much less than the sound velocity, so the range to the AUV during the "delivery" time of the measurement (and hence this time) will not change significantly.

The second variant of the CMNF correction employs the method of *pseudomeasurements* [15] as follows. Clearly, the relations for the measured angles can be easily transformed from the original geometric relations (10) into linear combinations of the coordinates to be determined:

$$\begin{cases} \dfrac{Y_{\mathcal{F}}}{\tan \varphi_{\mathcal{F}}} = \dfrac{Y}{\tan \varphi_{\mathcal{F}}} - X, \quad h \dfrac{\cos \varphi_{\mathcal{F}}}{\tan \lambda_{\mathcal{F}}} = Z \dfrac{\cos \varphi_{\mathcal{F}}}{\tan \lambda_{\mathcal{F}}} - X, \\ X_{\mathcal{S}} \tan \varphi_{\mathcal{S}} = X \tan \varphi_{\mathcal{S}} - Y, \quad h \cos \varphi_{\mathcal{S}} - X_{\mathcal{S}} \tan \lambda_{\mathcal{S}} = Z \cos \varphi_{\mathcal{S}} - X \tan \lambda_{\mathcal{S}}. \end{cases}$$

Applying similar transformations to (11) yields

$$\begin{cases} \dfrac{Y_{\mathcal{F}}}{y_{1t}} = \dfrac{y(t)}{y_{1t}} - x(t) + \dfrac{x(t)}{y_{1t}} v_1(t), \\ \dfrac{h}{(1 + \tan^2 \varphi_{\mathcal{F}}(t))^{1/2} \ y_{2t}} = \dfrac{z(t)}{(1 + \tan^2 \varphi_{\mathcal{F}}(t))^{1/2} \ y_{2t}} - x(t) + \dfrac{x(t)}{y_{2t}} v_2(t), \\ X_{\mathcal{S}} y_{3t} = x(t) y_{3t} - y(t) - (x(t) - X_{\mathcal{S}})\, v_3(t), \\ \dfrac{h}{(1 + \tan^2 \varphi_{\mathcal{S}}(t))^{1/2}} - X_{\mathcal{S}} y_{4t} = \dfrac{z(t)}{(1 + \tan^2 \varphi_{\mathcal{S}}(t))^{1/2}} - x(t) y_{4t} + (x(t) - X_{\mathcal{S}})\, v_4(t). \end{cases} \tag{16}$$

Finally, the precise values of $\tan\varphi_{\mathcal{F}}(t)$ and $\tan\varphi_{\mathcal{S}}(t)$ in (16) are unknown and have to be replaced by the corresponding observations y_{1t} and y_{3t}. As a result, from (16) we arrive at the approximate relations

$$\begin{cases} \dfrac{Y_{\mathcal{F}}}{y_{1t}} = \dfrac{y(t)}{y_{1t}} - x(t) + \dfrac{x(t)}{y_{1t}} v_1(t), \\ \dfrac{h}{\left(1+(y_{1t})^2\right)^{1/2} y_{2t}} = \dfrac{z(t)}{\left(1+(y_{1t})^2\right)^{1/2} y_{2t}} - x(t) + \dfrac{x(t)}{y_{2t}} v_2(t), \\ X_{\mathcal{S}} y_{3t} = x(t) y_{3t} - y(t) - (x(t) - X_{\mathcal{S}})\, v_3(t), \\ \dfrac{h}{\left(1+(y_{3t})^2\right)^{1/2}} - X_{\mathcal{S}} y_{4t} = \dfrac{z(t)}{\left(1+(y_{3t})^2\right)^{1/2}} - x(t) y_{4t} + (x(t) - X_{\mathcal{S}})\, v_4(t). \end{cases} \tag{17}$$

Following the method of pseudomeasurements, system (17) is used to write the filter; in particular, the left-hand sides are used as new observations. The reason is that the right-hand sides contain linear combinations of the estimated position $x(t)$, $y(t)$, and $z(t)$, which gives hope for a good performance of suboptimal Kalman filters, first of all, the extended Kalman filter (EKF) [8]. Indeed, this property is very attractive: when writing the EKF, there is no need to calculate the derivatives of the measuring functions (ψ_t in model (1)).

For the designed CMNF, this approach provides the basic correction in the form of the observation residual (17), i.e., $\zeta_t \in \mathbb{R}^4$ given by

$$\begin{cases} \zeta_{1t} = \dfrac{Y_{\mathcal{F}}}{Y_1} - \dfrac{X_2}{Y_1} + X_1, \quad \zeta_{2t} = \dfrac{h - X_3}{\left(1+(Y_1)^2\right)^{1/2} Y_2} + X_1, \\ (X_1, X_2, X_3)' = \widetilde{x}_{t-\widehat{\tau}_{\mathcal{F}}(t)}, \\ \zeta_{3t} = X_{\mathcal{S}} Y_3 - X_1 Y_3 + X_2, \quad \zeta_{4t} = \dfrac{h - X_3}{\left(1+(Y_3)^2\right)^{1/2}} - X_{\mathcal{S}} Y_4 + X_1 Y_4, \\ (X_1, X_2, X_3)' = \widetilde{x}_{t-\widehat{\tau}_{\mathcal{S}}(t)}. \end{cases} \tag{18}$$

In contrast to the correction term of the conventional EKF, the ultimate formula for ζ_t (18) incorporates, as the correction (15) of the typical CMNF, the observation delays with the same estimates $\widehat{\tau}_{\mathcal{F}}(t)$ and $\widehat{\tau}_{\mathcal{S}}(t)$.

Finally, the third structure for the CMNF is determined from geometric considerations, and the corresponding correction is called *geometric*. Again we use (10) under the assumption of no measurement errors. These relations can be interpreted as four joint equations for determining the three AUV coordinates $X,\ Y, Z$. Although the geometric problem in Fig. 1 has one solution (the surely intersecting acoustic ray lines of two beacons), system (10) can be formally solved in four ways. By combining the quantities measured by both beacons into the vector $(Y_1, Y_2, Y_3, Y_4)' = (\tan\varphi_{\mathcal{F}},\ \tan\lambda_{\mathcal{F}}, \tan\varphi_{\mathcal{S}}, \tan\lambda_{\mathcal{S}})'$, from (10) we obtain the system

$$\begin{cases} Y_1 = \dfrac{Y - Y_{\mathcal{F}}}{X}, \quad Y_2 = \dfrac{Z - h}{X} \dfrac{1}{\left(1+(Y_1)^2\right)^{1/2}}, \\ Y_3 = \dfrac{Y}{X - X_{\mathcal{S}}}, \quad Y_4 = \dfrac{Z - h}{X - X_{\mathcal{S}}} \dfrac{1}{\left(1+(Y_3)^2\right)^{1/2}}. \end{cases}$$

Its solution X, Y, Z satisfies each of the following twelve equalities:

$$\begin{cases} X = \dfrac{Y_{\mathcal{F}} + Y_3 X_{\mathcal{S}}}{Y_3 - Y_1}, \\ Y = \dfrac{Y_3\left(Y_{\mathcal{F}} + Y_1 X_{\mathcal{S}}\right)}{Y_3 - Y_1}, \\ Z = h + Y_2 \dfrac{Y_{\mathcal{F}} + Y_3 X_{\mathcal{S}}}{Y_3 - y_1}\left(1 + (Y_1)^2\right)^{1/2}, \end{cases} \qquad \begin{cases} X = \dfrac{Y_{\mathcal{F}} + Y_3 X_{\mathcal{S}}}{Y_3 - Y_1}, \\ Y = \dfrac{Y_3\left(Y_{\mathcal{F}} + Y_1 X_{\mathcal{S}}\right)}{Y_3 - Y_1}, \\ Z = h + Y_4\left(\dfrac{Y_{\mathcal{F}} + Y_3 X_{\mathcal{S}}}{Y_3 - Y_1} - X_{\mathcal{S}}\right)\left(1 + (Y_3)^2\right)^{1/2}, \end{cases}$$

$$\begin{cases} X = \dfrac{Y_4 X_{\mathcal{S}}}{Y_4 - Y_2\left(\frac{1+(Y_1)^2}{1+(Y_3)^2}\right)^{1/2}}, \\ Y = \dfrac{Y_1 Y_4 X_{\mathcal{S}}}{Y_4 - Y_2\left(\frac{1+(Y_1)^2}{1+(Y_3)^2}\right)^{1/2}} + Y_{\mathcal{F}}, \\ Z = h + \dfrac{Y_2 Y_4 X_{\mathcal{S}}}{\frac{Y_4}{\left(1+(Y_1)^2\right)^{1/2}} - \frac{Y_2}{\left(1+(Y_3)^2\right)^{1/2}}}, \end{cases} \qquad \begin{cases} X = \dfrac{Y_4 X_{\mathcal{S}}}{Y_4 - Y_2\left(\frac{1+(Y_1)^2}{1+(Y_3)^2}\right)^{1/2}}, \\ Y = \dfrac{Y_2 Y_3 X_{\mathcal{S}}}{Y_4\left(\frac{1+(Y_3)^2}{1+(Y_1)^2}\right)^{1/2} - Y_2}, \\ Z = h + \dfrac{Y_2 Y_4 X_{\mathcal{S}}}{\frac{Y_4}{\left(1+(Y_1)^2\right)^{1/2}} - \frac{Y_2}{\left(1+(Y_3)^2\right)^{1/2}}}. \end{cases}$$

Excluding the identical ones, we arrive at eight equalities based on the four measurements Y_1, Y_2, Y_3, Y_4. They can be used as the basic CMNF correction $\zeta_t \in \mathbb{R}^8$ of the form

$$\begin{cases} \zeta_{1t} = X_1 - \dfrac{Y_{\mathcal{F}} + Y_3 X_{\mathcal{S}}}{Y_3 - Y_1}, \ \ \zeta_{3t} = X_3 - h - Y_2 \dfrac{Y_{\mathcal{F}} + Y_3 X_{\mathcal{S}}}{Y_3 - Y_1}\left(1 + (Y_1)^2\right)^{1/2}, \\ \zeta_{2t} = X_2 - \dfrac{Y_3\left(Y_{\mathcal{F}} + Y_1 X_{\mathcal{S}}\right)}{Y_3 - Y_1}, \ \ \zeta_{4t} = X_3 - h - Y_4\left(\dfrac{Y_{\mathcal{F}} + Y_3 X_{\mathcal{S}}}{Y_3 - Y_1} - X_{\mathcal{S}}\right)\left(1 + (Y_3)^2\right)^{1/2}, \\ \zeta_{5t} = X_1 - \dfrac{Y_4 X_{\mathcal{S}}}{Y_4 - Y_2\left(\frac{1+(Y_1)^2}{1+(Y_3)^2}\right)^{1/2}}, \ \ \zeta_{7t} = X_2 - \dfrac{Y_2 Y_3 X_{\mathcal{S}}}{Y_4\left(\frac{1+(Y_3)^2}{1+(Y_1)^2}\right)^{1/2} - Y_2}, \\ \zeta_{6t} = X_2 - \dfrac{Y_1 Y_4 X_{\mathcal{S}}}{Y_4 - Y_2\left(\frac{1+(Y_1)^2}{1+(Y_3)^2}\right)^{1/2}} - Y_{\mathcal{F}}, \ \ \zeta_{8t} = X_3 - h - \dfrac{Y_2 Y_4 X_{\mathcal{S}}}{\frac{Y_4}{\left(1+(Y_1)^2\right)^{1/2}} - \frac{Y_2}{\left(1+(Y_3)^2\right)^{1/2}}}, \\ (X_1, X_2, X_3)' = \widetilde{x}_{t-\max\left(\widehat{\tau}_{\mathcal{F}}(t), \widehat{\tau}_{\mathcal{S}}(t)\right)}, \ \ Y_{1(2)} = y_{1(2)}{}_{t-\max\left(\widehat{\tau}_{\mathcal{F}}(t),\ \widehat{\tau}_{\mathcal{S}}(t)\right)+\widehat{\tau}_{\mathcal{F}}(t)}, \\ Y_{3(4)} = y_{3(4)}{}_{t-\max\left(\widehat{\tau}_{\mathcal{F}}(t),\ \widehat{\tau}_{\mathcal{S}}(t)\right)+\widehat{\tau}_{\mathcal{S}}(t)}. \end{cases} \tag{19}$$

Here, the time delay estimates $\widehat{\tau}_{\mathcal{F}}(t)$ and $\widehat{\tau}_{\mathcal{S}}(t)$ are calculated by analogy with the two previous variants of correction but used in a slightly more complicated way. Since the geometric interpretation "mixes" measurements from different beacons, the delays are mixed as well; therefore, both the prediction and the measurements must be shifted back for referring to the same time instant. For the prediction, this is the shift by the larger of the values $\widehat{\tau}_{\mathcal{F}}(t)$ or $\widehat{\tau}_{\mathcal{S}}(t)$; for the observations, the shift by the difference $|\widehat{\tau}_{\mathcal{F}}(t) - \widehat{\tau}_{\mathcal{S}}(t)|$ toward the earlier measurement. (Recall that the measurements y_{it} in (19) have been already shifted by $\tau_{\mathcal{F}}(t)$ or $\tau_{\mathcal{S}}(t)$ according to model (13).) Thus, the current (most "fresh") observations will be used only halfway. This "damage" is in the interest of synchronizing the measurements. In fact, of course, we introduce no damage at all: the unused part of the observations will be employed very quickly after the desynchronization time $|\tau_{\mathcal{F}}(t) - \tau_{\mathcal{S}}(t)|$.

In the case of more beacons (three, four, etc. acoustic sensors), the number of equations (19) grows exponentially. So formally scaling the geometric correction to a larger number of observers may lead to an unacceptable computing cost due to the implementation complexity instead of potentially improving the positioning quality. However, we hypothesize that there is no need to use all possible combinations of measurements, being limited to only some of them. In this case, it will be reasonable to employ all available sensors, no matter which pairs are involved in the correction. The success of the pseudomeasurement filter demonstrated in the next section indirectly confirms this hypothesis. Nevertheless, despite the simplicity of the geometric problem, the geometric variant of the correction procedure yields the best CMNF structure for solving the AUV positioning problem.

5. NUMERICAL EXPERIMENTS

To design the CMNF by the Monte Carlo method and analyze its quality, we simulated two independent samples of $N = 10\,000$ trajectories. On the first sample, the filter parameters (6) and (7) were calculated for each of the three correction structures proposed. Note that although relations (7) are not directly required to compute the filtering estimate, the values of $\widehat{K}_t$ are useful to know because they express the theoretical accuracy of the filter. On the second sample, we analyzed the real quality of the filtering estimate $\widehat{x}_t = (\widehat{x}(t), \widehat{y}(t), \widehat{z}(t))'$ and the identification estimate $\widehat{\mu}_t = (\widehat{v}_x(t), \widehat{v}_y(t), \widehat{v}_z(t))'$. The estimation accuracy was determined by the mean square deviations of the estimation errors, denoted by $\sigma_{\widehat{x}}(t)$, $\sigma_{\widehat{y}}(t)$, $\sigma_{\widehat{z}}(t)$ (indicated in meters in the figures) and $\sigma_{\widehat{v}_x}(t)$, $\sigma_{\widehat{v}_y}(t)$, $\sigma_{\widehat{v}_z}(t)$ (km/h) and calculated by the Monte Carlo method on the second sample. Thus, it is necessary to examine both the absolute values of these quantities (in particular, compare them with those calculated in the model without the time delays) and the difference between them

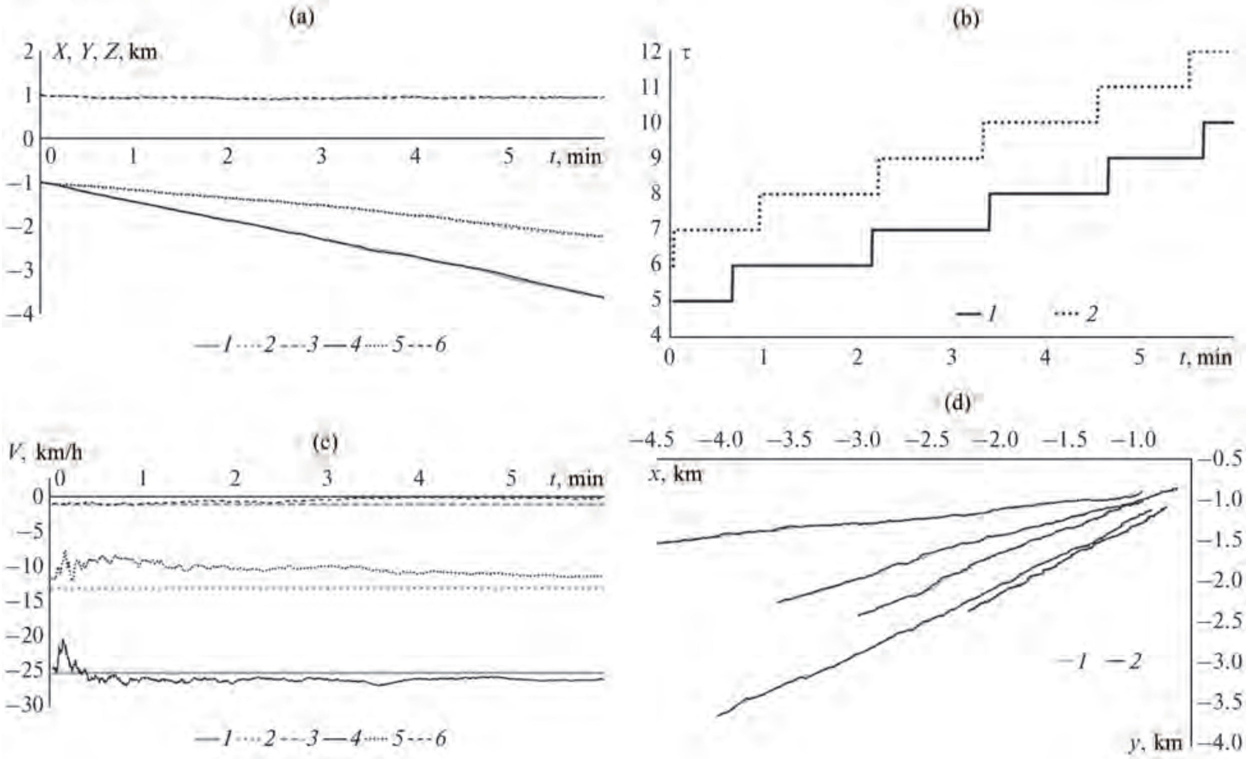


Fig. 3. AUV trajectories: (a) curves *1–3* correspond to the positions $x(t)$, $y(t)$, and $z(t)$, whereas curves *4–6* to the estimates $\widehat{x}(t)$, $\widehat{y}(t)$, and $\widehat{z}(t)$; (b) curves *1* and *2* correspond to the time delays $\tau_{\mathcal{F}}(t)$ and $\tau_{\mathcal{S}}(t)$; (c) curves *1–3* correspond to the velocities v_x, v_y, and v_z, whereas curves *4–6* to the estimates $\widehat{v}_x(t)$, $\widehat{v}_y(t)$, and $\widehat{v}_z(t)$; (d) curve *1* correspond to the AUV positions $(x(t), y(t))$, whereas curve *2* to the estimates $(\widehat{x}(t), \widehat{y}(t))$ on the trajectories.

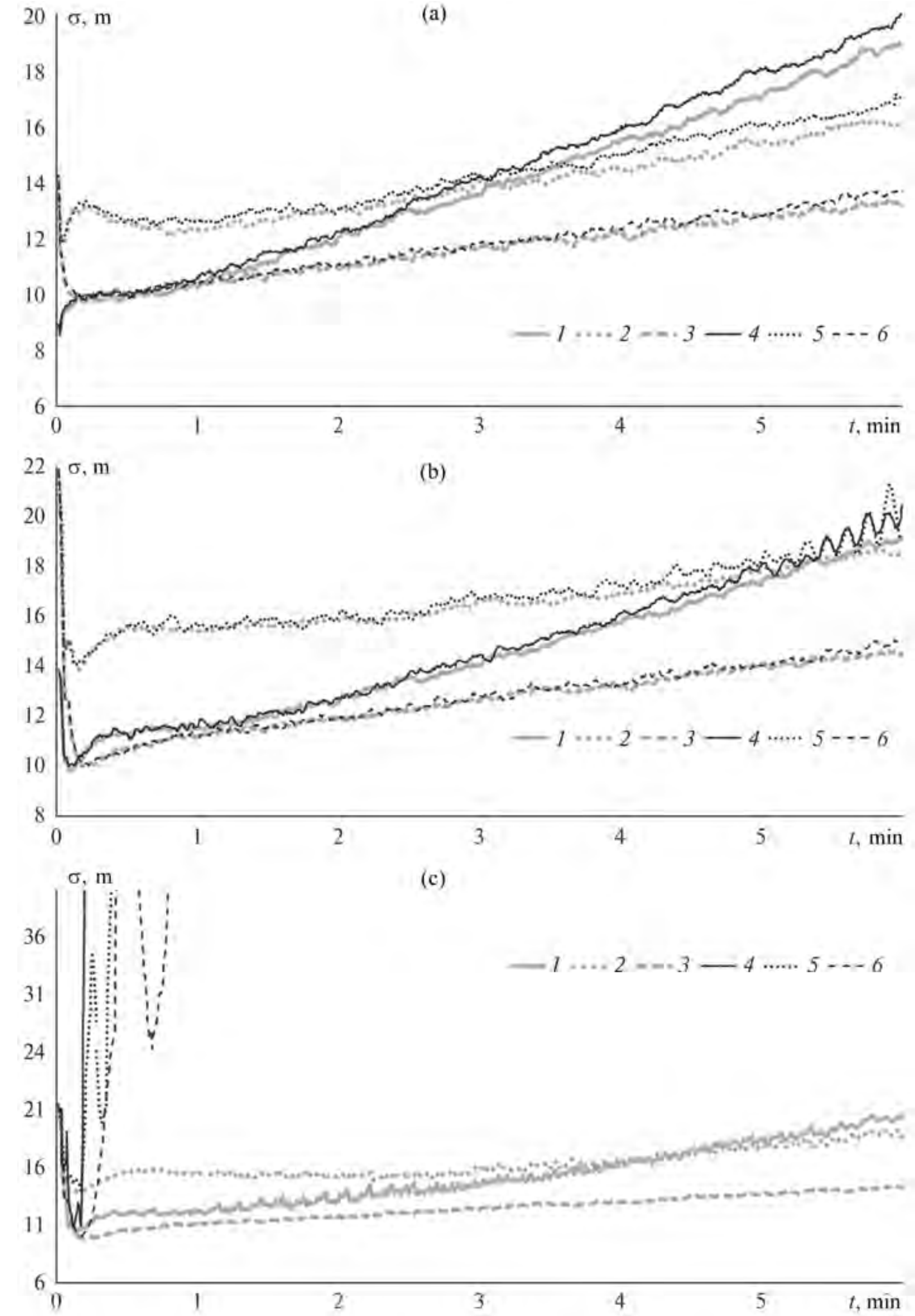


Fig. 4. Mean square deviations of the filters: (a) geometrical, (b) pseudomeasurements, and (c) typical. Curves *1–6* correspond to $\left(\widehat{K}_t\right)^{1/2}_{11}$, $\left(\widehat{K}_t\right)^{1/2}_{22}$, $\left(\widehat{K}_t\right)^{1/2}_{33}$, $\sigma_{\widehat{x}}(t)$, $\sigma_{\widehat{y}}(t)$, and $\sigma_{\widehat{z}}(t)$.

and the corresponding diagonal elements of $\widehat{K_t}$ in order to assess the practical realization of the theoretical properties of the CMNF in the case of its computer simulations-based design.

In Fig. 3, the experiment is illustrated with examples of characteristic position trajectories, time delay, filtering estimates, and parameter identification estimates. In Fig. 3a, one graph shows the coordinates of one AUV trajectory $x(t), y(t), z(t)$ and the corresponding CMNF estimate $\widehat{x}(t), \widehat{y}(t), \widehat{z}(t)$; in Fig. 3b, the time delays $\tau_{\mathcal{F}}(t)$ and $\tau_{\mathcal{S}}(t)$ for the first and second beacons on the same AUV trajectory; in Fig. 3c, the estimates $\widehat{v}_x(t), \widehat{v}_y(t), \widehat{v}_z(t)$ and the exact values of

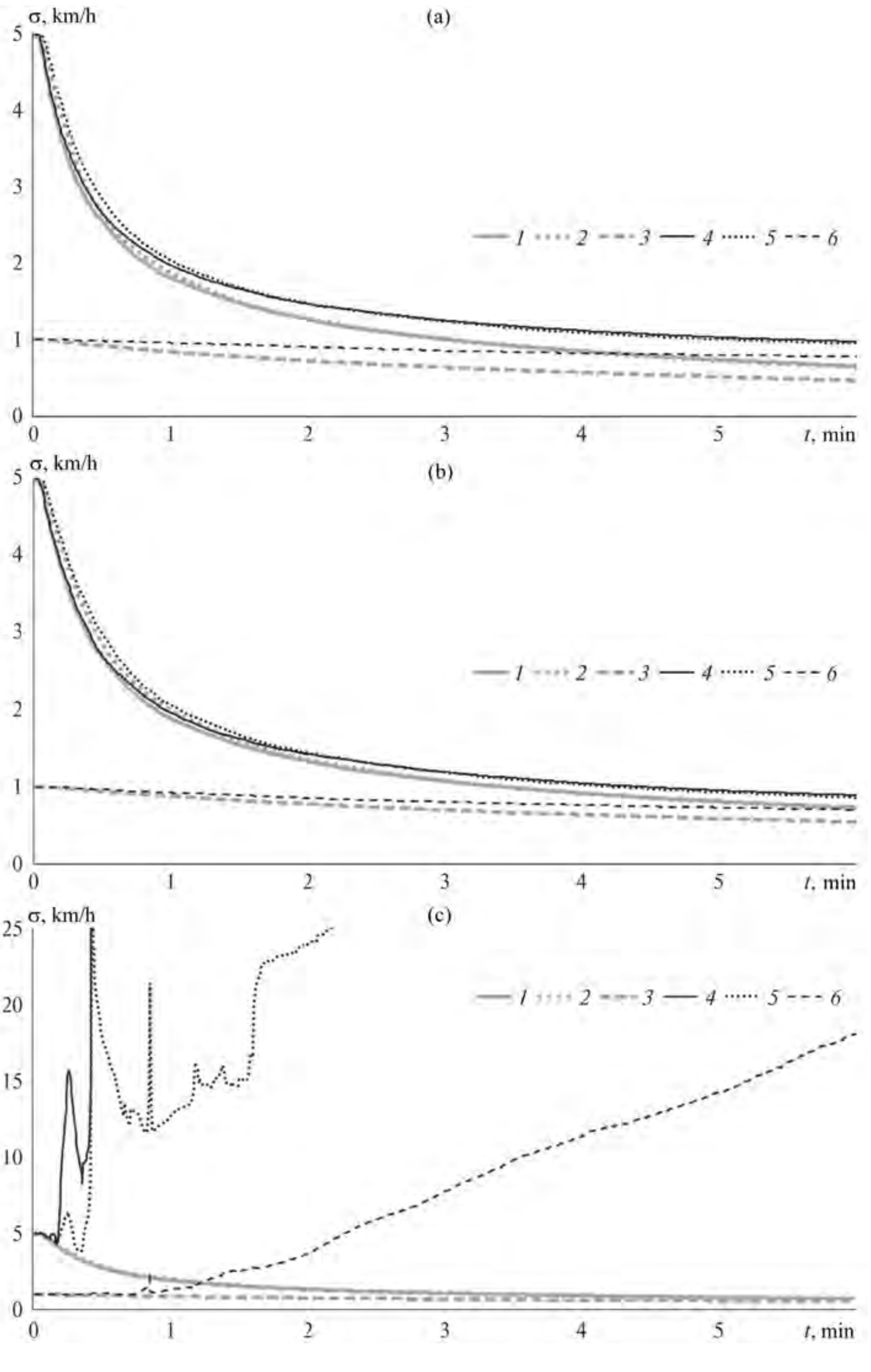


Fig. 5. Mean square deviations of the filters: (a) geometrical, (b) pseudomeasurements, and (c) typical. Curves *1–6* correspond to $\left(\widehat{K}_t\right)_{44}^{1/2}$, $\left(\widehat{K}_t\right)_{55}^{1/2}$, $\left(\widehat{K}_t\right)_{66}^{1/2}$, $\sigma_{\widehat{v}_x}(t)$, $\sigma_{\widehat{v}_y}(t)$, and $\sigma_{\widehat{v}_z}(t)$.

the velocities v_x, v_y, v_z; in Fig. 3d, several AUV trajectories in projection onto the Oxy plane. The CMNF estimates correspond to the geometric structure function. Although there exists a distinction in the filtering and identification estimates for different CMNF structures, it seems impossible to visualize this distinction on graphs. Therefore, numerical data are presented below to analyze the performance in detail.

Comparison of the estimation quality

Filter	$\sigma^{(1)}_{\widehat{x}}/\sigma^{(2)}_{\widehat{x}}$	$\sigma^{(1)}_{\widehat{y}}/\sigma^{(2)}_{\widehat{y}}$	$\sigma^{(1)}_{\widehat{z}}/\sigma^{(2)}_{\widehat{z}}$	$\sigma^{(1)}_{\widehat{v}_x}/\sigma^{(2)}_{\widehat{v}_x}$	$\sigma^{(1)}_{\widehat{v}_y}/\sigma^{(2)}_{\widehat{v}_y}$	$\sigma^{(1)}_{\widehat{v}_z}/\sigma^{(2)}_{\widehat{z}_x}$
$T = 15$, delay 54 s, unknown v_x, v_y, and v_z						
Geometric	13.80/**14.22**	13.93/**14.32**	11.61/**11.77**	**0.68**/0.99	**0.67**/0.97	**0.49**/0.79
Pseudomeasurements	14.36/14.60	16.54/16.83	12.60/12.70	0.76/**0.91**	0.75/**0.89**	0.56/**0.72**
Typical	15.22/**4300**	16.38/**3368**	12.50/**1117**	0.76/**88**	0.75/**96**	0.57/**17**
$T = 0$, delay 0 s, unknown v_x, v_y, and v_z						
Geometric	11.27/11.62	11.45/11.78	8.81/8.92	0.67/1.00	0.66/0.98	0.48/0.80
Pseudomeasurements	11.19/11.36	11.49/11.65	8.87/8.93	0.75/0.92	0.73/0.89	0.55/0.72
Typical	**11.09/11.25**	**11.45/11.61**	**8.81/8.86**	**0.74/0.92**	**0.73/0.89**	**0.55/0.72**
$T = 15$, delay 54 s, unknown $v_x = \mathrm{E}\{v_x\}$, $v_y = \mathrm{E}\{v_y\}$, and $v_z = \mathrm{E}\{v_z\}$						
Geometric	13.05/13.23	13.38/13.54	11.48/11.56	—	—	—
Pseudomeasurements	13.56/13.64	15.38/15.49	12.37/12.42	—	—	—
Typical	13.66/**45.81**	15.39/**32.22**	12.33/**22.09**	—	—	—
$T = 0$, delay 0 s, known $v_x = \mathrm{E}\{v_x\}$, $v_y = \mathrm{E}\{v_y\}$, and $v_z = \mathrm{E}\{v_z\}$						
Geometric	10.66/11.25	11.11/11.25	8.71/8.77	—	—	—
Pseudomeasurements	10.66/10.74	11.14/11.22	8.74/8.77	—	—	—
Typical	**10.65/10.72**	**11.13/11.21**	**8.71/8.75**	—	—	—

The next issue—the estimation accuracy—is illustrated in Figs. 4 and 5. Figures 4a–4c show the deviations $\sigma_{\widehat{x}}(t), \sigma_{\widehat{y}}(t), \sigma_{\widehat{y}}(t)$, and $\sigma_{\widehat{z}}(t)$ and, for comparison, the corresponding diagonal elements of the theoretical covariance $\widehat{K}_t$ of the estimation error. Figures 5a–5c present the similar characteristics $\sigma_{\widehat{v}_x}(t), \sigma_{\widehat{v}_y}(t)$, and $\sigma_{\widehat{v}_z}(t)$ for the identified velocities. Each figure contains three graphs corresponding to the three variants of the CMNF (geometric, pseudomeasurements, and typical).

Assessing these graphs visually, we state that the first two CMNF variants are successful in the positioning task. A precise comparison requires numerical indicators, which are given below. Also, the graphs confirm the fundamental possibility of velocity identification by these filters. The values characterizing the quality of identification are presented below. Obviously, the typical filter is able neither to position the AUV nor to identify the average velocity of motion. The discrepancy between the theoretical accuracy determined by the matrix $\widehat{K}_t$ and the real values of the deviations $\sigma_{\widehat{x}}(t), \sigma_{\widehat{y}}(t), \sigma_{\widehat{z}}(t)$, $\sigma_{\widehat{v}_x}(t), \sigma_{\widehat{v}_y}(t)$, and $\sigma_{\widehat{v}_z}(t)$ calculated on the second sample is too large. The reason consists in an insufficient sample size used for filter design and can be eliminated by significantly increasing N. However, in this case, the calculations become rather resource-intensive. It would make sense to consume computational resources only without the advantages of the other two CMNF structures, which are slightly more difficult to implement but much more efficient.

To conclude the experiment, it is necessary to characterize the real difference of the successful CMNF structures, the effect of the model with time delay on the quality of estimation, and the presence of identifiable parameters. For this purpose, we carried out calculations with the models in which $\tau_{\mathcal{F}}(t) = \tau_{\mathcal{S}}(t) = 0$, and the models in which $v_x = \mathrm{E}\{v_x\}$, $v_y = \mathrm{E}\{v_y\}$, $v_z = \mathrm{E}\{v_z\}$. The choice of the model without the time delay for comparison is clear; however, the assumption of a known constant velocity of the AUV on each trajectory may seem redundant. In fact, if the velocities are determined for each trajectory, as the original motion model (8) assumes, there will be no significant difference between positioning with the velocities known at the initial time instant and positioning with the velocities given by the identification estimates. This is explained in Figs. 5a and 5b, showing that the filter produces an acceptable estimate of the current velocity rather quickly. In addition, the motion model (8) yields a fairly accurate prediction using the average velocity as well; under the assumption of no time delays, the velocity does not affect the positioning quality at all, although there is the possibility to identify it.

The time-averaged mean deviations were used as an objective measure of the positioning quality (averaging over 1000 steps, 6 min). For example, for the analysis of $\widehat{x}(t)$, $\sigma^{(2)}_{\widehat{x}} = \frac{1}{1000}\sum_{t=1}^{1000}\sigma_{\widehat{x}}(t)$ were calculated to compare $\sigma_{\widehat{x}}(t)$ with the theoretical accuracy $\sigma^{(1)}_{\widehat{x}} = \frac{1}{1000}\sum_{t=1}^{1000}\left(\widehat{K}_t\right)^{1/2}_{11}$. The values of $\sigma^{(1)}_{\widehat{v}_x} = \frac{1}{100}\sum_{t=901}^{1000}\left(\widehat{K}_t\right)^{1/2}_{44}$ and $\sigma^{(2)}_{\widehat{v}_x} = \frac{1}{100}\sum_{t=901}^{1000}\sigma_{\widehat{v}_x}(t)$, etc. (the mean deviations of the estimates at the last hundred steps) were determined to analyze the identification results. Here, the superscripts (1) and (2) indicate the deviation computed on the first (the theoretical accuracy) and second (the factual accuracy) sample, respectively, and the subscript denotes the estimate. All results are combined in the table; the most interesting ones, illustrating the highest and lowest precision estimates, are set off in bold.

6. CONCLUSIONS

Of course, supplementing the model with random time delays [5, 6] with the unknown parameters to be identified has complicated the technical implementation of estimation algorithms. Even a reliable approach involving the concepts of conditionally optimal and conditionally minimax filtering suffers from certain difficulties. This is the first important conclusion from the results presented in the table above. Reproducing the Kalman filter structure, the typical filter essentially turns out to be unworkable. Interestingly, in the models without delays, both the filter based on the residual and the filters of more complex structure cope equally well with positioning (estimation of the position coordinates) and identification (estimation of the velocity). Actually, in the case of no time delays in observations, the CMNF works equally with any structure and provides very good estimation accuracy.

The second remark should be made regarding the concept of pseudomeasurements [15]. The fact that this approach would give a workable CMNF structure in the model with delays did not initially seem obvious. Its ability to compete with a purely geometric solution was all the more questionable. As the result of this study, we have obtained quite competitive estimates in terms of accuracy; in the identification problem, the CMNF with the pseudomeasurement structure even outperforms the geometric filter. Moreover, the geometric solution cannot be obtained in several cases, and increasing the number of observers will lead to an exponential growth of the dimensions. At the same time, the concept of linear pseudomeasurements has a universal character, applies to any DOA sensors, and is much easier to implement.

Finally, the third remark concerns the interpretation of the identification results. On the one hand, using the Monte Carlo method, one cannot expect the convergence of estimates in a numerical experiment. On the other hand, the Bayesian approach to parameter estimation and the CMNF properties do not imply convergence in a finite time, only ensuring that the estimation quality will improve from step to step. The real deviation of the velocity estimate turns out to be about 1 km/h; for the coordinates x and y with the initial velocity of 5 km/h and 6 min observations, this result seems to be good. The velocity along the Oz axis is poorly estimated, but it has a much smaller effect on positioning accuracy. Thus, the fundamental possibility of identifying the velocities without direct measurements, such as Doppler sensors, has been confirmed by the experiment. Velocity measurements can further improve the accuracy of both motion model identification and AUV positioning.

ACKNOWLEDGMENTS

This work was performed using the infrastructure of the Shared Research Facilities "High Performance Computing and Big Data" (CKP "Informatics") of the Federal Research Center "Computer Science and Control," Russian Academy of Sciences (Moscow).

REFERENCES

1. Bar-Shalom, Y., Li, X.-R., and Kirubarajan, T., *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software*, John Wiley & Sons, 2002.
2. *Autonomous Underwater Vehicles: Design and Practice (Radar, Sonar & Navigation)*, Ehlers, F., Ed., London: SciTech Publishing, 2020.
3. Kebkal, K.G. and Mashoshin, A.I., AUV Acoustic Positioning Methods, *Gyroscopy Navig.*, 2017, vol. 8, pp. 80–89.
4. Christ, R.D. and Wernli, R.L., *The ROV Manual: A User Guide for Remotely Operated Vehicles*, 2nd ed., Oxford: Butterworth-Heinemann, 2013.
5. Bosov, A.V., Observation-Based Filtering of State of a Nonlinear Dynamical System with Random Delays, *Autom. Remote Control*, 2023, vol. 84, no. 6, pp. 671–684.
6. Bosov, A., Tracking a Maneuvering Object by Indirect Observations with Random Delays, *Drones*, 2023, vol. 7, no. 7, art. no. 468.
7. Bosov, A.V., Nonlinear Dynamic System State Optimal Filtering by Observations with Random Delays, *Informatics and Applications*, 2023, vol. 17, no. 3, pp. 8–17.
8. Bernstein, I. and Friedland, B., Estimation of the State of a Nonlinear Process in the Presence of Nongaussian Noise and Disturbances, *J. Franklin Instit.*, 1966, vol. 281, no. 6, pp. 455–480.
9. Arulampalam, S., Maskell, S., Gordon, N.J., and Clapp, T., A Tutorial on Particle Filters for Online Non-linear/Non-Gaussian Bayesian Tracking, *IEEE Trans. Signal Processing*, 2002, vol. 50, no. 2, pp. 174–188.
10. Julier, S.J., Uhlmann, J.K., and Durrant-Whyte, H.F., A New Approach for Filtering Nonlinear Systems, *Proc. IEEE Amer. Control Conf. (ACC'95)*, 1995, pp. 1628–1632.
11. Pugachev, V.S., Recursive Estimation of Variables and Parameters in Stochastic Systems Described by Difference Equations, *Dokl. Akad. Nauk SSSR*, 1978, vol. 243, no. 5, pp. 1131–1133.
12. Pugachev, V.S., Estimation of Variables and Parameters in Discrete-Time Nonlinear Systems, *Autom. Remote Control*, 1979, vol. 40, no. 4, pp. 39–50.
13. Pankov, A.R. and Bosov, A.V., Conditionally Minimax Algorithm for Nonlinear System State Estimation, *IEEE Trans. Autom. Control*, 1994, vol. 39, no. 8, pp. 1617–1620.
14. Pugachev, V.S. and Sinitsyn, I.N., *Stochastic Differential Systems–Analysis and Filtering*, Chichester: Wiley, 1987.
15. Lin, X., Kirubarajan, T., Bar-Shalom, Y., and Maskell, S., Comparison of EKF, Pseudomeasurement, and Particle Filters for a Bearing-Only Target Tracking Problem, in *SPIE Proceedings. Signal and Data Processing of Small Targets 2002 (AEROSENSE 2002)*, Drummond, O.E., Ed., Orlando, April 1–5, 2002, SPIE: Bellingham, 2002, vol. 4728, pp. 240–250.
16. Miller, A. and Miller, B., Stochastic Control of Light UAV at Landing with the Aid of Bearing-Only Observations, in *SPIE Proceedings. Eight International Conference on Machine Vision (ICMV-2015)*, Verikas, A., Radeva, P., and Nikolaev, D., Eds., Barcelona, November 19–21, 2015, SPIE: Bellingham, 2015, vol. 9875, pp. 474–483.
17. Huber, P.J. and Ronchetti, E.M., *Robust Statistics*, 2nd. ed., Wiley, 2011.
18. Ljung, L., *System Identification—Theory for the User*, 2nd ed., NJ: PTR Prentice Hall, 1999.
19. Bertsekas, D.P. and Shreve, S.E., *Stochastic Optimal Control: The Discrete-Time Case*, NY: Academic Press, 1978.
20. Albert, A., *Regression and the Moor–Penrose Pseudoinverse*, New York: Academic Press, 1972.

21. Miller, A., Miller, B., and Miller, G., AUV Navigation with Seabed Acoustic Sensing, *Proceedings of the 2018 Australian New Zealand Control Conference (ANZCC)*, Melbourne, VIC, Australia, December 7–8, 2018, pp. 166–171.

22. Miller, A., Miller, B., and Miller, G., On AUV Control with the Aid of Position Estimation Algorithms Based on Acoustic Seabed Sensing and DOA Measurements, *Sensors*, 2019, vol. 19, art. no. 5520.

23. Miller, A., Miller, B., and Miller, G., Navigation of Underwater Drones and Integration of Acoustic Sensing with Onboard Inertial Navigation System, *Drones*, 2021, vol. 5, art. no. 83.

24. Hodges, R., *Underwater Acoustics: Analysis, Design and Performance of Sonar*, New York: Wiley, 2011.

25. Weirathmueller, M., Weber, T.C., Schmidt, V., McGillicuddy, G., Mayer, L., and Huff, L., Acoustic Positioning and Tracking in Portsmouth Harbor, New Hampshire, *Proc. OCEANS 2007*, Vancouver, BC, September 29– October 4, 2007, pp. 1–4.

26. Wong, G.S.K. and Zhu, S.-M., Speed of Sound in Seawater as a Function of Salinity, Temperature, and Pressure, *J. Acoust. Soc. Am.*, 1995, vol. 97, pp. 1732–1736.

27. Dushaw, B.D., Worcester, P.F., Cornuelle, B.D., and Howe, B.M., On Equations for the Speed of Sound in Seawater, *J. Acoust. Soc. Am.*, 1993, vol. 93, pp. 255–275.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 12, pp. 1168–1183.

STOCHASTIC SYSTEMS

Order Statistics of the Normalized Spectral Distribution for Detecting Weak Signals in White Noise

A. A. Galyaev[*,a], **L. M. Berlin**[*,b], **P. V. Lysenko**[*,c], **and V. G. Babikov**[*,d]

[*] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
e-mail: [a]*galaev@ipu.ru,* [b]*berlin.lm@phystech.edu,* [c]*pavellysen@ipu.ru,* [d]*vl.babikov@gmail.com*

Received September 8, 2024
Revised October 8, 2024
Accepted October 24, 2024

Abstract—This paper further develops the topic of the authors' previous research. In particular, order statistics of the discrete normalized spectral distribution of the additive white Gaussian noise are investigated to detect a deterministic useful signal in a noise mixture using information features. An additional relationship is established between the discrete spectral distribution of the single-window realization statistics of the white noise. Another novel result consists in exact formulas for calculating the mean and variance of normalized order statistics. Based on the analytical expressions derived, a new formula for calculating the spectral complexity is proposed and the already known one is refined. The theoretical results are verified by statistical simulation.

Keywords: order statistics, signal processing, Fourier transform, signal detection in noise

DOI: 10.31857/S0005117924120038

1. INTRODUCTION

The problems of detecting deterministic, chaotic, and random signals have attracted the attention of researchers since the time of studying space objects [1, 2]. When solving such problems, the main difficulty lies in the unknown properties and characteristics of the observed signal. Therefore, the periodic repetition of a signal over time is often treated as a feature of the appearance of a deterministic signal. Then, based on the results of observations, the received energy is accumulated and averaged, and the presence/absence of the desired signal is concluded accordingly [3].

The statistical foundations of the theory of signal detection and classification in noise were laid in the 1950s–1960s [4]. In turn, machine learning methods allowed solving classification problems based on extracting a set of individual features [5]. Over the last 30–40 years, research results have appeared in the scientific literature in which information characteristics of temporal or spectral distributions are used as detection and classification criteria. Since entropy or information is insensitive to the permutation of discrete samples of distributions, the gaze of the scientific community is directed to studying the distributions of order statistics and describing their properties [6, 7]. In recent years, entropy and related indicators have been frequently applied in the analysis of electroencephalography (EEG) and brain activity [8]. Another example of using signal entropy and related spectral complexity is the problem of signal classification, as shown in [5, 9]. Lately, order statistics have found applications in generative neural networks [10] and the related Wasserstein distance and theory of optimal transport [11, 12].

In many detection problems, the signal being detected is often supposed known [13]. In practice, the pre-detection problem is also popular when it is important, e.g., to determine the presence of a deterministic signal in white noise [14, 15].

This paper aims to establish statistical regularities in the single-window observations of white noise and studying the evolution of these regularities when a signal appears in the mixture. An expected result is a new tool for detecting (and, subsequently, classifying) a signal under small values of the signal-to-noise ratio (SNR).

The remainder of this paper is organized as follows. In Section 2, we present the mathematical formulation of the problem and the background necessary for further considerations. Sections 3 and 4 contain the main analytical results of this research, namely, the lemmas and propositions for finding the mean and variance of the normalized order statistics of the spectral distribution of the white noise when processing the observations of a single rectangular window. In Section 5, new information characteristics of signals are introduced based on the analytical results obtained. Section 6 describes the results of numerical simulations confirming the theoretical constructs of the previous sections. In Section 7, we summarize the outcomes of the paper.

2. PROBLEM STATEMENT AND BACKGROUND

The problem of detecting a signal $s(n)$ is traditionally reduced to that of distinguishing between two hypotheses [15]:

$$\begin{cases} \Gamma_0 : x(n) = w(n) \\ \Gamma_1 : x(n) = s(n) + w(n),\ n = 1, \ldots, 2N+2. \end{cases}$$

The hypothesis Γ_0 is associated with the decision to receive only noise while the hypothesis Γ_1 with the decision to receive a mixture of a useful signal and noise, where the sequences $\{x(n)\}$, $\{s(n)\}$, and $\{w(n)\}$, $n = 1, \ldots, 2N+2$, denote the time series of the received data, useful signal, and additive random noise, respectively, and $2N+2$ is the time series length. The random variables $(x(1), \ldots, x(n), \ldots, x(2N+2))$ of the time series take values $(x_1, \ldots, x_n, \ldots, x_{2N+2}) \in \mathbb{R}^{2N+2}$.

To estimate the probability of erroneous hypotheses distinction, we can apply a modification of the Neyman–Pearson lemma in which the error function is described by the exact formula

$$\mathcal{E}r(N; \Gamma_0, \Gamma_1) = 1 - \frac{1}{2}\|P_0^{(N)} - P_1^{(N)}\| = 1 - TV(P_0, P_1), \tag{1}$$

where $P_0^{(N)}$ and $P_1^{(N)}$ are the multivariate distribution functions of observation statistics for the hypotheses Γ_0 and Γ_1, respectively, $TV(P_0, P_1)$ is the total variation of the measure with a sign, and $\|Q\| = 2\sup_A |Q(A)|$. Thus, if the carrier sets of the measures P_0 and P_1 do not overlap, then error-free hypotheses distinction is possible. If the measures $P_0^{(N)}$ and $P_1^{(N)}$ are close, then $\|P_0^{(N)} - P_1^{(N)}\| \approx 0$ and, in this case, $\mathcal{E}r(N; \Gamma_0, \Gamma_1) \approx 1$.

The peculiarity of the problem is to detect a deterministic signal under small SNR values, i.e., $\mathcal{E}r(N; \Gamma_0, \Gamma_1) \approx 1$, by analyzing the spectral properties of the received signal-noise mixture and information criteria. Due to this peculiarity, the following problem was formulated in the paper [16].

Problem 1. Consider a given realization $\{x_1, \ldots, x_{2N+2}\}$ for a sequence of independent random variables $\{\xi_1, \ldots, \xi_{2N+2}\}$ with zero mean. Let the discrete Fourier transform (DFT)

$$X_k = \sum_{n=1}^{2N+2} x_n e^{-2\boldsymbol{i}\pi k(n-1)/(2N+2)} \tag{2}$$

be applied to this sequence to obtain the random variable

$$\Xi_k = \sum_{n=1}^{2N+2} \xi_n e^{-2\boldsymbol{i}\pi k(n-1)/(2N+2)}, \tag{3}$$

where $k = 0, \dots, N$. (Due to the symmetry of the DFT of a real signal, the second half of $N+1, \dots,$ $2N+1$ complex amplitudes of the spectral samples is conjugate to the first.)

It is required to find the discrete probability function of the normalized ordered spectral distribution $n_k(N)$ as the normalized mean for each kth value of the random variable

$$\eta_k(N) = \frac{(\mathbf{T}I)_k}{E_X}, \tag{4}$$

where $I_k = \Xi_k \Xi_k^*$ (the square of the amplitude modulus or the energy of the spectral sample), E_X is half of the signal energy, and $\mathbf{T}$ is the non-descending ordering operator of the series, and to investigate the properties of the resulting distribution on different information measures.

To solve Problem 1 in the general case, one needs to calculate

$$\mathbb{E}[\eta_k(N)] = \mathbb{E}\left[\frac{(\mathbf{T}I)_k}{E_X}\right]. \tag{5}$$

In 1898 A. Schuster established that the distributions of the random variables I_k, $k = 1, \dots, N$, are exponential, and the random variable I_0 obeys the χ^2 distribution with one degree of freedom. This fact holds if the random variables ξ_n, $n = 1, \dots, 2N+2$, are independent Gaussian ones with zero mean and variance σ_0^2, as demonstrated in [17].

Remark 1. The total number of signal samples in the time domain is set equal to $2N+2$ for the convenience of analyzing the energies I_k of half of the spectral samples, only N of which obey the exponential distribution.

In this paper, the detection problem is solved with a criterion defined by an information characteristic, i.e., complexity represented as the product of entropy and the $\mathcal{L}_1$ norm of two distributions. Being only a function of the discrete spectral distribution, entropy possesses insensitivity to the permutation of its samples; being a function of two distributions, complexity is sensitive to permutation when calculating a relatively nonuniform distribution. Therefore, when postulating the order of samples in a discrete distribution, a rule is established for calculating the information characteristics. As such a rule, we will use ascending or descending ordering for spectrum samples.

3. MAIN RESULTS. CALCULATING THE MEAN OF NORMALIZED ORDER STATISTICS

The mean of an order statistic normalized by the sum of samples possesses the following property.

Lemma 1. *Let $z_1, \dots, z_N$ be the observations of a random variable Z obeying the exponential distribution $F(z) = 1 - \exp(-z)$ with the density function $f(z) = \exp(-z)$. Consider the values of the sequence $z_{(1)}, \dots, z_{(N)}$ of the same results rearranged in ascending order, where the random variable $Z_{(k)}$ is a nondecreasing kth order statistic.*

Then the mean of such a random variable normalized by the sum of all elements of the sample has the form

$$\mathbb{E}\left[\frac{Z_{(k)}}{\sum_{i=1}^N Z_{(i)}}\right] = \frac{\int\limits_0^\infty \int\limits_0^{z_N} \dots \int\limits_0^{z_3} \int\limits_0^{z_2} \frac{z_k}{\sum_{i=1}^N z_i} \exp\left(-\sum\limits_{i=1}^N z_i\right) dz_1 dz_2 \dots dz_{N-1} dz_N}{\int\limits_0^\infty \int\limits_0^{z_N} \dots \int\limits_0^{z_3} \int\limits_0^{z_2} \exp\left(-\sum\limits_{i=1}^N z_i\right) dz_1 dz_2 \dots dz_{N-1} dz_N}. \tag{6}$$

Proof. The lemma is directly verified by writing the mean for the order statistic.

Formula (6) differs from the known ones in the calculation of the mean of $\frac{z_k}{\sum_{i=1}^N z_i}$ by the joint distribution of the order statistics $Z_{(1)}, \dots, Z_{(N)}$. A difficulty arises when calculating the integral in the numerator of (6), which is overcome using the relation between the means $\mathbb{E}\left[\frac{Z_{(k)}}{\sum_{i=1}^{N} Z_{(i)}}\right]$ and $\mathbb{E}[Z_{(k)}]$.

The density function of the distribution for the kth order statistic $Z_{(k)}$ is defined as follows:

$$f_{Z_{(k)}}(z) = \frac{N!}{(k-1)!(N-k)!} f(z)[F(z)]^{k-1}[1-F(z)]^{N-k}.$$

When substituting the exponential law $F(z)$ this density takes the explicit form

$$f_{Z_{(k)}}(z) = \frac{N!}{(k-1)!(N-k)!} \exp(-z(1+N-k))(1-\exp(-z))^{k-1}. \tag{7}$$

The next formula is a well-known result for finding the mean of order statistics of the standard exponential law [7]:

$$\mathbb{E}\left[Z_{(k)}\right] = \int_0^\infty z f_{Z_{(k)}}(z)dz = \sum_{i=N-k+1}^{N} \frac{1}{i}. \tag{8}$$

On the other hand, this mean can be obtained by the formula of Lemma 1, i.e.,

$$\mathbb{E}\left[Z_{(k)}\right] = \frac{\int_0^\infty \int_0^{z_N} \ldots \int_0^{z_3}\int_0^{z_2} z_k \exp\left(-\sum_{i=1}^{N} z_i\right) dz_1 dz_2 \ldots dz_{N-1} dz_N}{\int_0^\infty \int_0^{z_N} \ldots \int_0^{z_3}\int_0^{z_2} \exp\left(-\sum_{i=1}^{N} z_i\right) dz_1 dz_2 \ldots dz_{N-1} dz_N} = \sum_{i=1}^{N}\frac{1}{i} - \sum_{i=1}^{N-k}\frac{1}{i} = \sum_{i=N-k+1}^{N}\frac{1}{i}. \tag{9}$$

Proposition 1. *The mean of the normalized order statistic* (6) *is given by*

$$\mathbb{E}\left[\frac{Z_{(k)}}{\sum_{i=1}^{N} Z_{(i)}}\right] = \frac{1}{N}\sum_{i=N-k+1}^{N}\frac{1}{i}, \tag{10}$$

which coincides, up to the factor $\frac{1}{N}$, *with the value of* $\mathbb{E}[Z_{(k)}]$.

Proof. The repeated integral in the denominator is analytically calculated for an arbitrary upper limit of the last integral:

$$\int_0^a \int_0^{z_N} \ldots \int_0^{z_3}\int_0^{z_2} \exp\left(-\sum_{i=1}^{N} z_i\right) dz_1 dz_2 \ldots dz_{N-1} dz_N = \frac{1}{N!} + O(a^{N-1}\exp(-a)).$$

Passing to the limit as $a \to \infty$ yields

$$\int_0^\infty \int_0^{z_N} \ldots \int_0^{z_3}\int_0^{z_2} \exp\left(-\sum_{i=1}^{N} z_i\right) dz_1 dz_2 \ldots dz_{N-1} dz_N = \frac{1}{N!}. \tag{11}$$

This result can be alternatively established by observing that

$$\int_0^\infty \int_0^\infty \ldots \int_0^\infty\int_0^\infty \exp\left(-\sum_{i=1}^{N} z_i\right) dz_1 dz_2 \ldots dz_{N-1} dz_N = 1.$$

At the same time, this integral consists of the sum of $N!$ integrals of the form (11) (by the number of all permutations of z_i, $i = 1, \ldots, N$).

Now we study the integral in the numerator of (10). The repeated integral in the numerator is analytically calculated for an arbitrary upper limit of the last integral:

$$\int_0^a \int_0^{z_N} \ldots \int_0^{z_3} \int_0^{z_2} \frac{z_k}{\sum_{i=1}^N z_i} \exp\left(-\sum_{i=1}^N z_i\right) dz_1 dz_2 \ldots dz_{N-1} dz_N$$
$$= \frac{1}{N!}\frac{1}{N}\left(\sum_{i=1}^N \frac{1}{i} - \sum_{i=1}^{N-k} \frac{1}{i}\right) + O(a^N \exp(-a)) + O(a^{N+1}\mathrm{Ei}(1,a)),$$

where $\mathrm{Ei}(1,a) = \int\limits_1^\infty t^{-1}\exp(-ta)dt$.

Passing to the limit as $a \to \infty$ gives the desired result.

Section 4 proposes a method for calculating the variance of a normalized order random variable, which can also be used to prove Proposition 1.

Simultaneously, several interesting results have been obtained concerning repeated improper integrals:

$$\int_0^\infty \int_0^{z_N} \ldots \int_0^{z_3} \int_0^{z_2} \exp\left(-\sum_{i=1}^N z_i\right) dz_1 dz_2 \ldots dz_N = \frac{1}{N!}, \tag{12}$$

$$\int_0^\infty \int_0^{z_N} \ldots \int_0^{z_3} \int_0^{z_2} \frac{1}{\sum\limits_{i=1}^N z_i} \exp\left(-\sum_{i=1}^N z_i\right) dz_1 dz_2 \ldots dz_N = \frac{1}{N-1}\frac{1}{N!}, \tag{13}$$

$$\int_0^\infty \int_0^{z_N} \ldots \int_0^{z_3} \int_0^{z_2} \frac{1}{\left(\sum\limits_{i=1}^N z_i\right)^2} \exp\left(-\sum_{i=1}^N z_i\right) dz_1 dz_2 \ldots dz_N = \frac{1}{N-2}\frac{1}{N-1}\frac{1}{N!}. \tag{14}$$

They can be calculated using the following property.

Proposition 2. *For an arbitrary natural degree p, $N - p \geqslant 1$, wee have*

$$M(p,N) = \int_0^\infty \int_0^\infty \ldots \int_0^\infty \int_0^\infty \frac{1}{\left(\sum\limits_{i=1}^N z_i\right)^p} \exp\left(-\sum_{i=1}^N z_i\right) dz_1 dz_2 \ldots dz_N = \frac{(N-p-1)!}{(N-1)!}. \tag{15}$$

For $p \geqslant N$, this integral becomes meaningless due to its divergence. Moreover, for an arbitrary natural degree p,

$$\int_0^\infty \int_0^\infty \ldots \int_0^\infty \int_0^\infty \left(\sum_{i=1}^N z_i\right)^p \exp\left(-\sum_{i=1}^N z_i\right) dz_1 dz_2 \ldots dz_N = \frac{(N+p-1)!}{(N-1)!}. \tag{16}$$

Proof. The change of variables $\{z_i\}_{i=1}^N \to \{x_i\}_{i=1}^N$ in the integrals (15) and (16) is performed by the following rule [18]:

$$x_1 = \frac{z_1}{x_N}, \ldots, x_{N-1} = \frac{z_{N-1}}{x_N},\ x_N = \sum_{i=1}^N z_i. \tag{17}$$

To find the Jacobian $J(x_1, \ldots, x_N)$ of the mapping (17), it is necessary to express the variables $\{z_i\}_{i=1}^N$ through $\{x_i\}_{i=1}^N$:

$$z_1 = x_1 x_N, \ldots, z_{N-1} = x_{N-1} x_N,$$
$$z_N = \sum_{i=1}^N z_i - \sum_{i=1}^{N-1} z_i = x_N - \sum_{i=1}^{N-1} x_N x_i = x_N \left(1 - \sum_{i=1}^{N-1} x_i\right).$$

After elementary transformations, the Jacobian of this mapping is reduced to the determinant of the upper-triangular matrix and equals $J(x_1, \ldots, x_N) = x_N^{N-1}$. Therefore, the integral (15) is calculated as follows:

$$M(p, N) = \int_0^\infty \int_0^\infty \ldots \int_0^\infty \int_0^\infty \left(\sum_{i=1}^N z_i\right)^{-p} \exp\left(-\sum_{i=1}^N z_i\right) dz_1 dz_2 \ldots dz_N$$
$$= \int_0^\infty x_N^{-p} x_N^{N-1} \exp\left(-x_N\right) dx_N \left(\int_0^1 \int_0^{x_{N-1}} \ldots \int_0^{x2} dx_1 \ldots dx_{N-2} dx_{N-1}\right)$$
$$= \int_0^\infty x_N^{-p} \frac{x_N^{N-1}}{\Gamma(N)} \exp\left(-x_N\right) dx_N \left(\Gamma(N) \int_0^1 \int_0^{x_{N-1}} \ldots \int_0^{x2} dx_1 \ldots dx_{N-2} dx_{N-1}\right)$$
$$= \int_0^\infty x_N^{-p} \frac{x_N^{N-1}}{\Gamma(N)} \exp\left(-x_N\right) dx_N \int_0^1 \ldots \int_0^1 dx_1 \ldots dx_{N-1} = \frac{\Gamma(N-p)}{\Gamma(N)} = \frac{(N-p-1)!}{(N-1)!}.$$

The integral (16) is found by analogy. Finally, note that the integral (15) consists of $N!$ identical integrals of the form (13) for $p = 1$.

Here is another interesting mathematical result on the calculation of the integral (13).

Lemma 2. *For any $N \in \mathbb{N}$, the integral (15) with $p = 1$ can be somehow represented by dividing into two parts containing k and $N - k$ elements:*

$$M(1, N) = \frac{1}{\Gamma(N-k)\Gamma(k)} \int_0^\infty \int_0^\infty \frac{\zeta^{N-k-1} \eta^{k-1}}{\zeta + \eta} \exp\left(-(\zeta + \eta)\right) d\zeta d\eta. \tag{18}$$

Proof. This lemma can be verified using two changes of variables in the integral, similar to the proof of Proposition 2. Note only that in this case, as before by Proposition 2, $M(1, N) = \frac{1}{N-1}$.

4. MAIN RESULTS. CALCULATING THE VARIANCE OF NORMALIZED ORDER STATISTICS

Now we have the problem of calculating the variance $\mathbb{D}\left[\frac{Z_{(k)}}{\sum_{i=1}^N Z_{(i)}}\right]$; to this end, it is necessary to find $\mathbb{E}\left[\frac{Z_{(k)}^2}{\left(\sum_{i=1}^N Z_{(i)}\right)^2}\right]$.

Proposition 3. *The desired mean can be obtained as follows:*

$$\mathbb{E}\left[\frac{Z_{(k)}^2}{\left(\sum_{i=1}^N Z_{(i)}\right)^2}\right] = \frac{1}{N(N+1)} \mathbb{E}\left[Z_{(k)}^2\right]. \tag{19}$$

Proof. The means in both parts of the equality are written by definition considering the order of integration:

$$\mathbb{E}\left[Z_{(k)}^2\right] = N! \int\limits_0^\infty \int\limits_0^{z_N} \ldots \int\limits_0^{z_3} \int\limits_0^{z_2} z_k^2 \exp\left(-\sum_{i=1}^N z_i\right) dz_1 dz_2 \ldots dz_{N-1} dz_N,$$

$$\mathbb{E}\left[\frac{Z_{(k)}^2}{\left(\sum\limits_{i=1}^N Z_{(i)}\right)^2}\right] = N! \int\limits_0^\infty \int\limits_0^{z_N} \ldots \int\limits_0^{z_3} \int\limits_0^{z_2} \frac{z_k^2}{\left(\sum\limits_{i=1}^N z_i\right)^2} \exp\left(-\sum_{i=1}^N z_i\right) dz_1 dz_2 \ldots dz_{N-1} dz_N.$$

Similar to the proof of Proposition 2, we apply the change of variables (17) in the integrals to continue the corresponding equalities:

$$\mathbb{E}\left[Z_{(k)}^2\right] = N! \int\limits_0^\infty \int \cdots \int\limits_\Omega x_N^{N-1} x_k^2 x_N^2 \exp\left(-x_N\right) dx_1 dx_2 \ldots dx_{N-1} dx_N$$

$$= N! \int\limits_0^\infty x_N^{N+1} \exp\left(-x_N\right) dx_N \int \cdots \int\limits_\Omega x_k^2 dx_1 dx_2 \ldots dx_{N-1}$$

$$= N(N+1)N! \int\limits_0^\infty x_N^{N-1} \exp\left(-x_N\right) dx_N \int \cdots \int\limits_\Omega x_k^2 dx_1 dx_2 \ldots dx_{N-1},$$

$$\mathbb{E}\left[\frac{Z_{(k)}^2}{\left(\sum\limits_{i=1}^N Z_{(i)}\right)^2}\right] = N! \int\limits_0^\infty \int \cdots \int\limits_\Omega x_N^{N-1} x_k^2 \exp\left(-x_N\right) dx_1 dx_2 \ldots dx_{N-1} dx_N$$

$$= N! \int\limits_0^\infty x_N^{N-1} \exp\left(-x_N\right) dx_N \int \cdots \int\limits_\Omega x_k^2 dx_1 dx_2 \ldots dx_{N-1},$$

where Ω denotes the domain of integration for the new variables $x_1, \ldots, x_{N-1}$ under the change. In the case under consideration, this domain is part of the simplex that arises under a similar change when deriving the Dirichlet distribution from the generating gamma distributions [19, 20].

Comparing the expressions for $\mathbb{E}\left[\frac{Z_{(k)}^2}{\left(\sum_{i=1}^N Z_{(i)}\right)^2}\right]$ and $\mathbb{E}\left[Z_{(k)}^2\right]$, we arrive at the desired result.

The method above is also suitable for proving Proposition 1.

Corollary 1. *Due to Proposition 3 and formula* (10), *the variance is given by*

$$\mathbb{D}\left[\frac{Z_{(k)}}{\sum_{i=1}^N Z_{(i)}}\right] = \frac{1}{N(N+1)} \mathbb{E}\left[Z_{(k)}^2\right] - \left(\frac{1}{N} \mathbb{E}\left[Z_{(k)}\right]\right)^2. \tag{20}$$

Now we can estimate its value.

Corollary 2. *For all N the variance* (20) *satisfies the upper bound*

$$\mathbb{D}\left[\frac{Z_{(k)}}{\sum_{i=1}^N Z_{(i)}}\right] \leqslant \frac{1}{N^2} \mathbb{D}\left[Z_{(k)}\right].$$

Proof. It follows from Corollary 1 that

$$\mathbb{D}\left[\frac{Z_{(k)}}{\sum_{i=1}^{N} Z_{(i)}}\right] = \frac{1}{N(N+1)}\mathbb{E}\left[Z_{(k)}^2\right] - \left(\frac{1}{N}\mathbb{E}\left[Z_{(k)}\right]\right)^2 = \frac{1}{N^2}\mathbb{D}\left[Z_{(k)}\right] - \frac{1}{N^2(N+1)}\mathbb{E}\left[Z_{(k)}^2\right].$$

The proof of this corollary is complete.

Proposition 4. *The second moment and variance of a non-decreasing order statistic are given by*

$$\mathbb{E}\left[Z_{(k)}^2\right] = \left(\sum_{i=N-k+1}^{N} \frac{1}{i}\right)^2 + \sum_{i=N-k+1}^{N} \frac{1}{i^2}, \tag{21}$$

$$\mathbb{D}\left[Z_{(k)}\right] = \sum_{i=N-k+1}^{N} \frac{1}{i^2}. \tag{22}$$

Proof. For $\mathbb{E}\left[Z_{(k)}^2\right]$, we have the formula [7]

$$\mathbb{E}\left[Z_{(k)}^2\right] = \int_0^\infty z^2 f_{Z_{(k)}}(z)dz = \int_0^\infty \frac{z^2 N!}{(k-1)!(N-k)!}\exp(-z(1+N-k))(1-\exp(-z))^{(k-1)}dz$$
$$= (H_N - H_{N-k})^2 + \psi^{(1)}(N-k+1) - \psi^{(1)}(N+1),$$

which contains the polygamma function $\psi^{(m)}(n) = (-1)^{(m+1)} m! \sum_{k=n}^{\infty} \frac{1}{k^{(m+1)}}$ and the harmonic series $H_N = \sum_{i=1}^{N} \frac{1}{i}$.

Substituting the expression $\psi^{(m)}(n)$ into the last equality yields (21). The variance (22) is obtained from (21) by considering formula (8) for $\mathbb{E}\left[Z_{(k)}\right]$.

Proposition 5. *The variance* (20) *is calculated using the constructive formula*

$$\mathbb{D}\left[\frac{Z_{(k)}}{\sum_{i=1}^{N} Z_{(i)}}\right] = \frac{1}{N(N+1)}\left(\sum_{i=N-k+1}^{N} \frac{1}{i^2} - \frac{1}{N}\left(\sum_{i=N-k+1}^{N} \frac{1}{i}\right)^2\right). \tag{23}$$

Proof. Let us transform the general expression (20) for the variance of a normalized order statistic using (21) and (10):

$$\mathbb{D}\left[\frac{Z_{(k)}}{\sum_{i=1}^{N} Z_{(i)}}\right] = \frac{1}{N(N+1)}\mathbb{E}\left[Z_{(k)}^2\right] - \left(\frac{1}{N}\mathbb{E}\left[Z_{(k)}\right]\right)^2$$
$$= \frac{1}{N(N+1)}\left(\left(\sum_{i=N-k+1}^{N} \frac{1}{i}\right)^2 + \sum_{i=N-k+1}^{N} \frac{1}{i^2}\right) - \left(\frac{1}{N}\sum_{i=N-k+1}^{N} \frac{1}{i}\right)^2$$
$$= \left(\sum_{i=N-k+1}^{N} \frac{1}{i}\right)^2 \left(\frac{1}{N(N+1)} - \frac{1}{N^2}\right) + \frac{1}{N(N+1)}\sum_{i=N-k+1}^{N} \frac{1}{i^2} \tag{24}$$
$$= \left(\sum_{i=N-k+1}^{N} \frac{1}{i}\right)^2 \left(\frac{-1}{N^2(N+1)}\right) + \frac{1}{N(N+1)}\sum_{i=N-k+1}^{N} \frac{1}{i^2}$$
$$= \frac{1}{N(N+1)}\left(\sum_{i=N-k+1}^{N} \frac{1}{i^2} - \frac{1}{N}\left(\sum_{i=N-k+1}^{N} \frac{1}{i}\right)^2\right).$$

Here are a few more interesting facts from the results above.

Corollary 3. *For $k = N$, formula* (22) *yields*

$$\lim_{N\to\infty} \mathbb{D}\left[Z_{(N)}\right] = \lim_{N\to\infty} \sum_{i=1}^{N} \frac{1}{i^2} = \frac{\pi^2}{6}.$$

Corollary 4. *For all N, we have*

$$\mathbb{E}\left[Z_{(1)}^2\right] = \frac{1}{N+1}\frac{2}{N^2}.$$

5. THE MEAN VALUE OF THE NORMALIZED ENERGY SPECTRUM DENSITY, ITS ESTIMATES AND SPECTRAL COMPLEXITY

Let us return to the original Problem 1. After finding the energy spectral samples of one observation window, $y_1, \ldots, y_N$, the results of observations of the exponentially distributed random variable Y generating $\eta_k(N)$, are introduced [16]. Then the values of the sequence $y_{(1)}, \ldots, y_{(N)}$ (the same results rearranged in descending order) are an inverse variational series, where the random variable $Y_{(k)}$ is a nonincreasing kth order statistic.

In this case, we have

$$\mathbb{E}\left[\frac{Z_{(k)}}{\sum_{i=1}^{N} Z_{(i)}}\right] = \mathbb{E}\left[\frac{Y_{(N-k+1)}}{\sum_{i=1}^{N} Y_{(i)}}\right],$$

or

$$\widetilde{n}_k(N) = \mathbb{E}\left[\frac{Y_{(k)}}{\sum_{i=1}^{N} Y_{(i)}}\right] = \mathbb{E}\left[\frac{Z_{(N-k+1)}}{\sum_{i=1}^{N} Z_{(i)}}\right] = \frac{1}{N}\left(\sum_{i=1}^{N}\frac{1}{i} - \sum_{i=1}^{k-1}\frac{1}{i}\right) = \frac{1}{N}\sum_{i=k}^{N}\frac{1}{i}. \tag{25}$$

Based on (25), it follows that

$$\sum_{k=1}^{N} \mathbb{E}\left[\frac{Y_{(k)}}{\sum_{i=1}^{N} Y_{(i)}}\right] = N\frac{1}{N} = 1,$$

and the normalization condition is obviously valid.

In turn, the variance is given by

$$\mathbb{D}\left[\frac{Z_{(k)}}{\sum_{i=1}^{N} Z_{(i)}}\right] = \mathbb{D}\left[\frac{Y_{(N-k+1)}}{\sum_{i=1}^{N} Y_{(i)}}\right],$$

or

$$\mathbb{D}\left[\frac{Y_{(k)}}{\sum_{i=1}^{N} Y_{(i)}}\right] = \frac{1}{N(N+1)}\left(\sum_{i=k}^{N}\frac{1}{i^2} - \frac{1}{N}\left(\sum_{i=k}^{N}\frac{1}{i}\right)^2\right). \tag{26}$$

The probability function of the normalized ordered discrete spectrum is approximately calculated using the formula [16]

$$n_k(N) = -\frac{1}{NK_N}\ln\frac{k}{N+1}, \quad \text{where} \quad K_N = -\frac{1}{N}\sum_{k=1}^{N}\ln\frac{k}{N+1}. \tag{27}$$

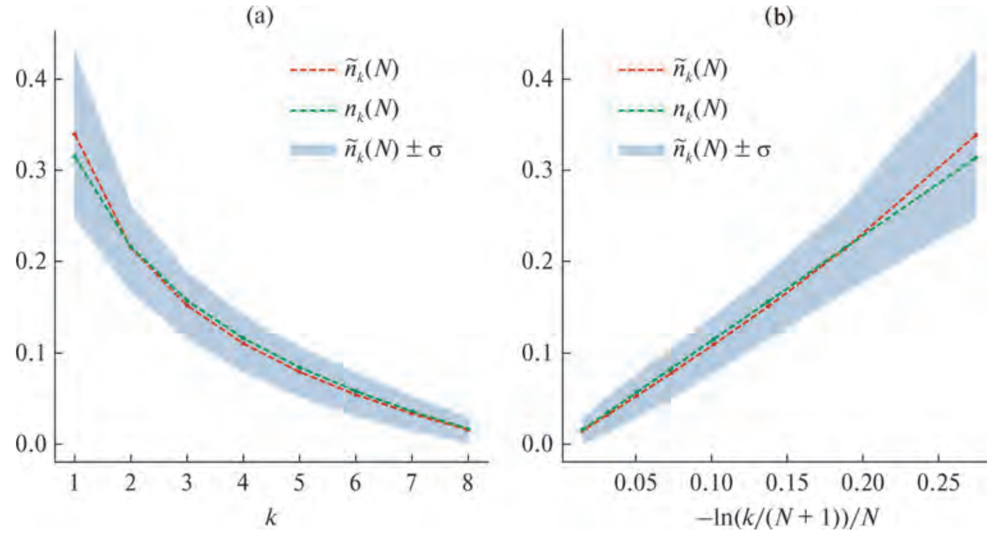


Fig. 1. Discrete distributions $\widetilde{n}_k(N)$ and $n_k(N)$ for the series of size $N = 8$: the horizontal axis in (a) linear and (b) logarithmic scale.

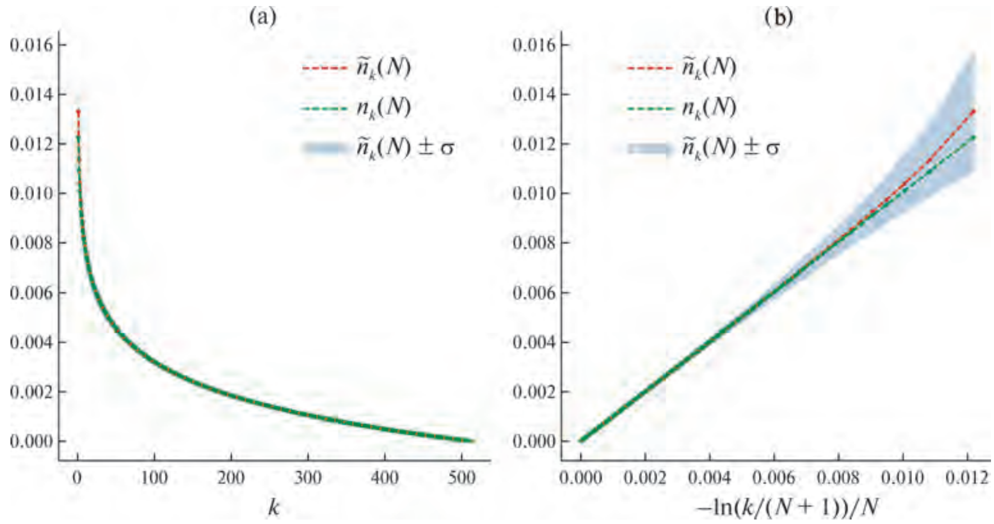


Fig. 2. Discrete distributions $\widetilde{n}_k(N)$ and $n_k(N)$ for the series of size $N = 512$: the horizontal axis in (a) linear and (b) logarithmic scale.

Now we compare the discrete distributions: the exact (25) and approximate (27) ones.

For large values of N and k, the expression (25), which is the difference of harmonic series, can be approximated in various ways. Let us choose the estimate

$$\sum_{i=1}^{N} \frac{1}{i} - \sum_{i=1}^{k-1} \frac{1}{i} \approx -\ln \frac{k}{N+1},$$

which is well-defined and makes sense for all N and $k \in [1, \ldots, N]$ and also appears in (27).

Figures 1 and 2 present the plots of the distributions (25) and (27) depending on the sample number. The symbol σ indicates the standard deviation of the distribution of the order statistic given by (26).

According to Figs. 1 and 2, the exact values of the mean of (25) slightly differ from those of the distribution (27) for a small number of points (members of the spectral series), almost coinciding for the other samples. As N grows, the share of points deviating from the estimate rapidly decreases.

Lemma 3. *In the general case, a more accurate estimate is*

$$\sum_{i=1}^{N}\frac{1}{i}-\sum_{i=1}^{k-1}\frac{1}{i}\approx -\ln\frac{k-0.5}{N+0.5}, \tag{28}$$

which is also well-defined and makes sense for all N and $k\in[2,\dots,N]$.

Proof. To show this fact, we address the theory of harmonic series.

The partial sum of the first N terms of the harmonic series is called the Nth harmonic number:

$$H_N=\sum_{i=1}^{N}\frac{1}{i}=1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\dots+\frac{1}{N}. \tag{29}$$

In 1740, Euler derived an asymptotic expression for H_N, called the Euler–Maclaren formula:

$$H_N=\ln N+\gamma+\frac{1}{2N}-\varepsilon_N, \tag{30}$$

where $\gamma=0.5772\dots$ is the Euler–Mascheroni constant and $0\leqslant\varepsilon_N\leqslant 1/8N^2$. Hence, $\varepsilon_N\to 0$ as $N\to\infty$, and for large N we have

$$H_N=\ln N+\gamma+O(N^{-1})\approx \ln N+\gamma. \tag{31}$$

This expression is called the Euler formula for the sum of the first n members of the harmonic series. Returning to (30), note that

$$\ln\left(N+\frac{1}{2}\right)=\ln N+\frac{1}{2N}+O\left(\frac{1}{N^2}\right). \tag{32}$$

Substituting the right-hand side into (30) yields

$$H_N=\ln\left(N+\frac{1}{2}\right)+\gamma+O\left(\frac{1}{N^2}\right), \tag{33}$$

since the numbers ε_N and ε_{k+1} have a close order of smallness with respect to $\frac{1}{N^2}$ and $\frac{1}{k^2}$, which rapidly vanish with increasing N and k.

Therefore, (28) takes the form

$$\sum_{i=1}^{N}\frac{1}{i}-\sum_{i=1}^{k-1}\frac{1}{i}=H_N-H_{k-1}=\ln\left(N+\frac{1}{2}\right)+O\left(\frac{1}{N^2}\right)-\ln\left(k-\frac{1}{2}\right)+O\left(\frac{1}{(k-1)^2}\right). \tag{34}$$

Thus, for relatively large N and k, we arrive at the estimate

$$\sum_{i=1}^{N}\frac{1}{i}-\sum_{i=1}^{k-1}\frac{1}{i}\approx -\ln\frac{k-0.5}{N+0.5}, \tag{35}$$

and the proof of this lemma is complete.

To proceed, we approximately calculate the probability function of the normalized ordered discrete spectrum by the formula

$$n_k(N)=-\frac{1}{NK_N}\ln\frac{2k-1}{2N+1},\quad\text{where}\quad K_N=-\frac{1}{N}\sum_{k=1}^{N}\ln\frac{2k-1}{2N+1}. \tag{36}$$

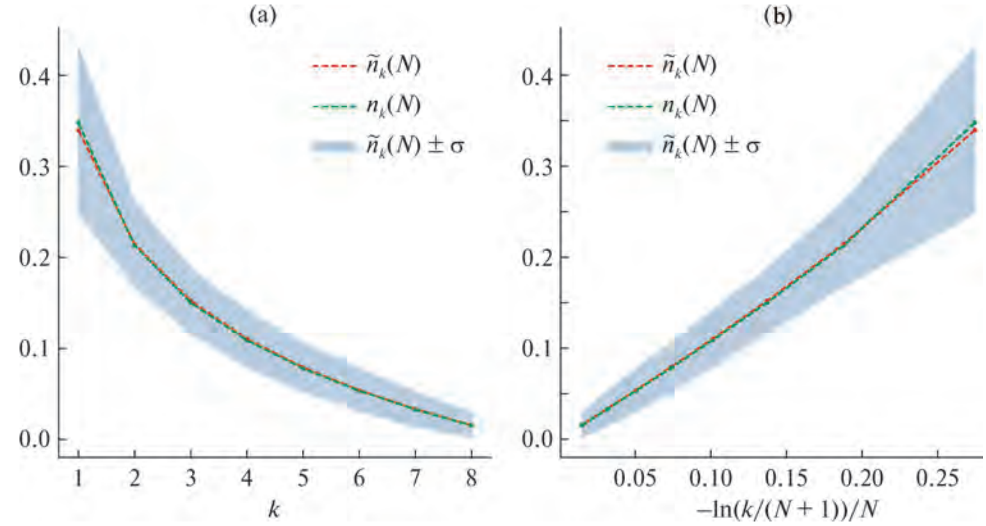


Fig. 3. Discrete distributions $\widetilde{n}_k(N)$ and $n_k(N)$ for the series of size $N = 8$: the horizontal axis in (a) linear and (b) logarithmic scale.

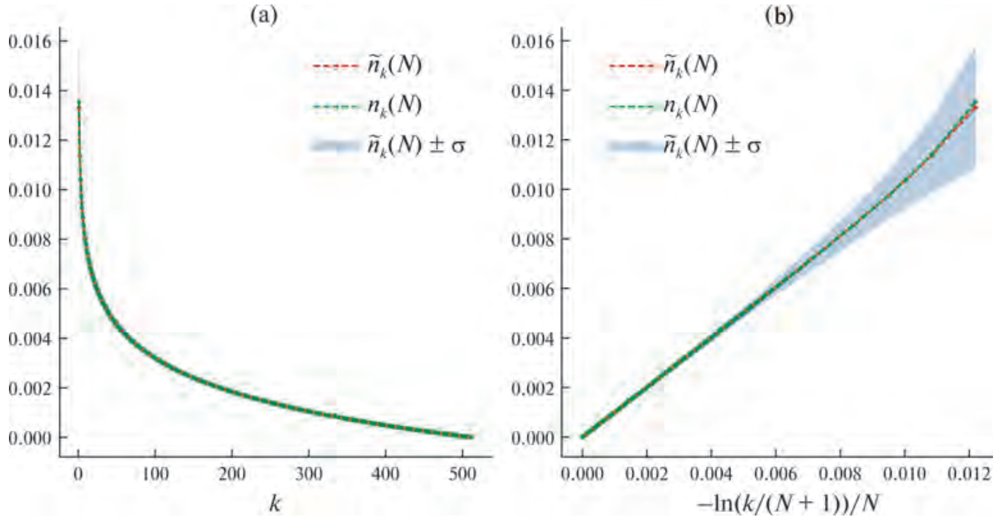


Fig. 4. Discrete distributions $\widetilde{n}_k(N)$ and $n_k(N)$ for the series of size $N = 512$: the horizontal axis in (a) linear and (b) logarithmic scale.

The result of this lemma is illustrated in Figs. 3 and 4.

Obviously, the two distributions almost coincide even for small N. Now, assuming that the normalized order statistics $(\eta_1, \ldots, \eta_N)$ take values $p = (p_1, \ldots, p_N)$, the spectral complexity can also be set based on the distribution (25) by the formula

$$C_{SS}(p) = -\frac{1}{4\log_2 N}\left(\sum_{k=1}^{N} p_k \log_2 p_k\right)\left(\sum_{k=1}^{N} |p_k - \widetilde{n}_k(N)|\right)^2 \tag{37}$$

or its approximate analog (36) by the formula

$$C_S(p) = -\frac{1}{4\log_2 N}\left(\sum_{k=1}^{N} p_k \log_2 p_k\right)\left(\sum_{k=1}^{N} |p_k - n_k(N)|\right)^2; \tag{38}$$

this will be demonstrated in Section 6.

6. STATISTICAL SIMULATION OF DETERMINISTIC SIGNAL DETECTION

To illustrate the analytical results of this paper, we use the statistical simulation procedure based on the analysis of the generated numerical data. It was described in detail in the authors' previous paper [16]. As before, all numerical results were obtained using the Python language and the Numpy and Scipy libraries.

Consider pairs of data sequences corresponding to two hypotheses of signal receipt:

$$\begin{cases} \Gamma_0 : x_n = w_n \\ \Gamma_1 : x_n = s_n + w_n, n = 1, \ldots, N. \end{cases} \tag{39}$$

The hypothesis Γ_0 is associated with the decision to receive only noise while the hypothesis Γ_1 with the decision to receive a mixture of a useful signal and noise, where the sequences $\{x(n)\}$, $\{s(n)\}$, and $\{w(n)\}$, $n = 1, \ldots, N$, denote the time series of the received data, useful signal, and additive white Gaussian random noise, respectively, and N is the time series (frame) length.

To test the quality of distinguishing between the useful deterministic signal and noise, statistics were collected on $Q = 50\,000$ numerically generated frames $\{x_n\}$ of the signal-noise mixture of length $2N = 16\,384$ with spectra size $N = 8192$, respectively. In all realizations, the signal $\{s_n\}$ remained the same, i.e., a fixed (in number and amplitude) set of $K = 30$ sinusoids evenly spaced along the spectrum with random phases. The additive white Gaussian noise $\{w_n\}$ was obtained by a Gaussian sequence generator with the mean $\mu = 0$ and variance Σ (within one set of Q frames).

Figure 5 shows the ordered discrete spectral distributions p_k for particular signal realizations and the theoretical means of $\widetilde{n}_k(N)$ described by (25). Obviously, in each realization, the real spectral distribution of white noise may appreciably differ from the theoretical one $\widetilde{n}_k(N)$. However, all such realizations lie within the confidence interval defined by the standard deviation σ of the distribution of the order statistic (26). The blue color in Fig. 5 indicates the domain bounding the confidence interval for each sample number k with the upper and lower bounds $\widetilde{n}_k(N) \pm 3\sigma$.

For each resulting sequence $\{x_n\}$, the order statistic of the discrete normalized spectral distribution p_k was calculated; for details, see [9, 15]. Next, based on p_k, the values of C_{SS} (37) and C_S (38) were obtained for the noise and noise-signal mixture satisfying the two hypotheses from the

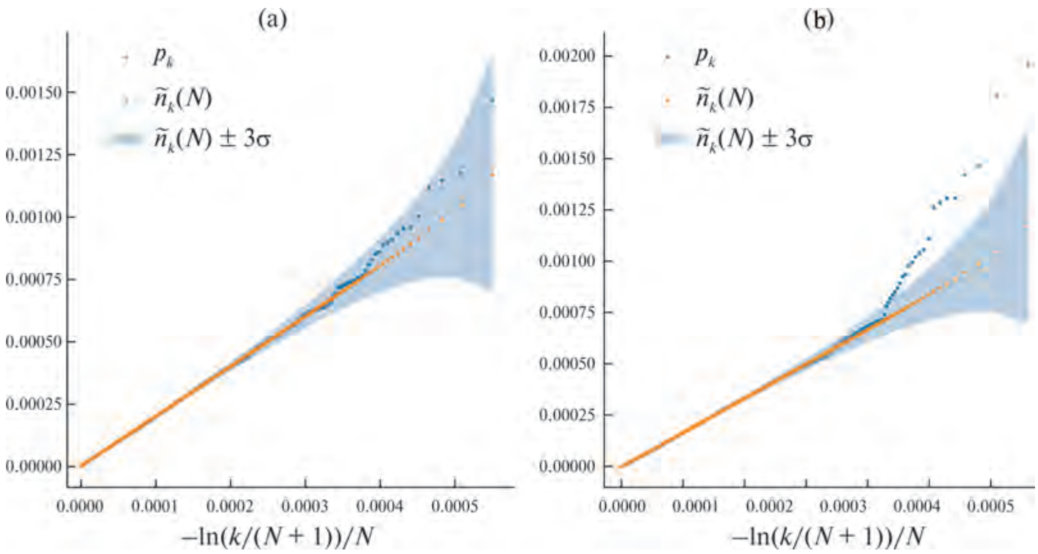


Fig. 5. Discrete distributions $\widetilde{n}_k(N)$ and p_k for the series of size $N = 8192$ and $SNR = -15.22$ dB: the time sequence of (a) noise data w_n and (b) signal and noise data $s_n + w_n$. The horizontal axis in logarithmic scale.

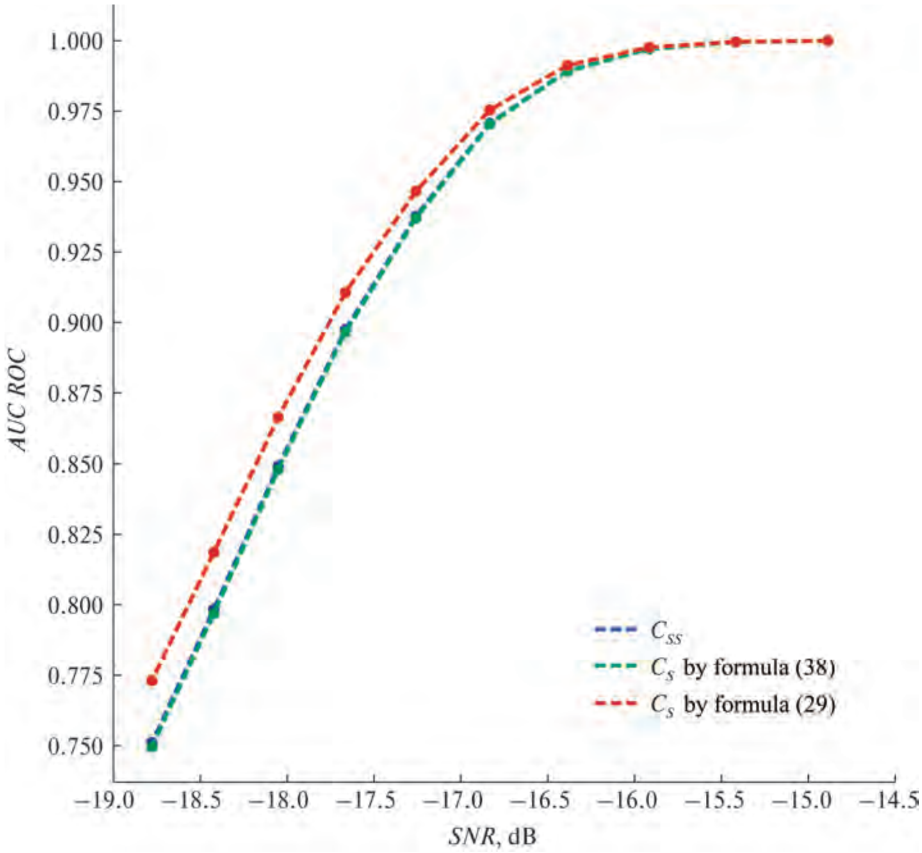


Fig. 6. The quality of binary classification in the *AUC ROC* metric depending on SNR for the spectral complexity functions under consideration.

expressions (39). The value of C_S was found for the two different approximations (27) and (36) of the order statistics of the spectral distribution of the white noise: it seems interesting to compare them as well.

The final result of simulating and comparing the calculated information metrics is the quality of binary classification in the *AUC ROC* (Area Under the Receiver Operating Characteristic Curve) metric depending on SNR, defined by

$$SNR = 10 \log_{10} \left(\frac{E_{signal}}{E_{noise}} \right), \tag{40}$$

where E_{signal} and E_{noise} denote the total signal and noise energies, respectively, calculated as the sum of the powers of the spectral decomposition of the sequences $\{s_n\}$ and $\{w_n\}$.

To obtain this dependence, the statistic Q of frames described above was collected for several values of the noise variance Σ, the histograms of the distributions of C_{SS} and C_S were built, and the values of *AUC ROC* were finally calculated [16].

Figure 6 shows the comparison of the quality of binary classification of the signal and noise for the information metrics C_{SS} and C_S. Obviously, the indicators for the functions C_{SS} and C_S, calculated using (36) with theoretical justification by Lemma 3, coincide almost completely. In addition, the information characteristic C_S calculated using formula (36) from the authors' previous paper [16] has a slight improvement in the quality of detection. All the analytical information criteria introduced in this paper demonstrate a high quality of detection of the deterministic signal in noise under small SNR values.

7. CONCLUSIONS

This paper has established that the order statistic of a discrete normalized spectral distribution is a powerful tool for detecting a deterministic signal under small SNR values in single-window observations. The spectral complexity calculated on a particular realization of the signal-noise mixture has been used as a detection criterion.

An attempt to separate the two problems (the detection/pre-detection of deterministic signals in white noise under small SNR values and classification) has demonstrated that increasing the sensitivity of the detection method causes the loss of all physical properties of the signal: further work involves only the informational ones. A simple analogy with quantum systems is suggested here, in which there exists an uncertainty relation when the researcher cannot accurately measure the speed and coordinate simultaneously. In fact, the reduced informativeness of particular signal frequencies makes it impossible to classify the signal, let alone recover the signal from a measurement together with noise. However, the ordering of the noise spectrum has a particular rigid structure allowing one to judge the presence of a signal even beyond the sensitivity of classical energy receivers.

Moreover, it seems promising to apply the proposed detection method for other types of noise as well as to construct classification grids for signals based on the information characteristics of the spectra of signal-noise mixtures. These will be the subjects of the next research by the authors.

FUNDING

This work was supported in part by the Russian Science Foundation, project no. 23-19-00134.

REFERENCES

1. Amigó, J.M. and Rosso, O.A., Ordinal Methods: Concepts, Applications, New Developments, and Challenges – in Memory of Karsten Keller (1961–2022), *Chaos*, 2023, vol. 33, no. 8, art. no. 080401. https://pubs.aip.org/cha/article/33/8/080401/2905538/Ordinal-methods-Concepts-applications-new

2. Rosso, O.A., Larrondo, H.A., Martin, M.T., Plastino, A., and Fuentes, M.A., Distinguishing Noise from Chaos, *Phys. Rev. Lett.*, 2007, vol. 99, no. 15, art. no. 154102. https://doi.org/10.1103/PhysRevLett.99.154102

3. Terebizh, V.Yu., *Analiz vremennykh ryadov v astrofizike* (Analysis of Time Series in Astrophysics), Moscow: Nauka, 1992.

4. Dobrushin, R.L., Pinsker, M.S., and Shiryaev, A.N., Application of the Notion of Entropy in the Problems of Detecting a Signal in Noise, *Lithuanian Math. J.*, 1963, vol. 3, no. 1, pp. 107–122.

5. Chen, Z., Wu, C., Wang, J., and Qiu, H., Tsallis Entropy-Based Complexity-IPE Casualty Plane: A Novel Method for Complex Time Series Analysis, *Entropy*, 2024, vol. 26, no. 6, art. no. 521. https://www.mdpi.com/1099-4300/26/6/521

6. Hoeffding, W., On the Distribution of the Expected Values of the Order Statistics, *Ann. Math. Stat.*, 1953, vol. 24, no. 1, pp. 93–100. https://doi.org/10.1214/aoms/1177729086

7. Balakrishnan, N. and Cohen, A.C., *Order Statistics & Inference: Estimation Methods*, Academic Press, 1991. https://doi.org/10.1016/C2009-0-22411-1

8. Vecchio, F., Miraglia, F., Pappalettera, C., Orticoni, A., Alù, F., Judica, E., Cotelli, M., and Rossini, P.M., Entropy as Measure of Brain Networks' Complexity in Eyes Open and Closed Conditions, *Symmetry*, 2021, vol. 13, no. 11, art. no. 2178. https://www.mdpi.com/2073-8994/13/11/2178

9. Berlin, L.M., Galyaev, A.A., and Lysenko, P.V., Comparison of Information Criteria for Detection of Useful Signals in Noisy Environments, *Sensors*, 2023, vol. 23, no. 4, art. no. 2133. https://www.mdpi.com/1424-8220/23/4/2133

10. Zhang, W., Huang, D., Zhou, M., Lin, J., and Wang, X., Open-Set Signal Recognition Based on Transformer and Wasserstein Distance, *Appl. Sci.*, 2023, vol. 13, no. 4, art. no. 2151. https://www.mdpi.com/2076-3417/13/4/2151

11. Cazelles, E., Robert, A., and Tobar, F., The Wasserstein–Fourier Distance for Stationary Time Series, *IEEE Transact. Signal Proc.*, 2021, vol. 69, pp. 709–721. https://ieeexplore.ieee.org/document/9303405

12. Ferracuti, F., Freddi, A., Monteriu, A., and Romeo, L., Fault Diagnosis of Rotating Machinery Based on Wasserstein Distance and Feature Selection, *IEEE Transact. Autom. Sci. Engin.*, 2022, vol. 19, no. 3, pp. 1997–2007. https://doi.org/10.1109/TASE.2021.3069109

13. Quan, Z., Zhang, W., Shellhammer, S.J., and Sayed, A.H., Optimal Spectral Feature Detection for Spectrum Sensing at Very Low SNR, *IEEE Transact. Commun.*, 2011, vol. 59, no. 1, pp. 201–212. https://doi.org/10.1109/TCOMM.2010.112310.090306

14. Bardenet, R., Flamant, J., and Chainais, P., On the Zeros of the Spectrogram of White Noise, *Appl. Comput. Harmon. Anal.*, 2020, vol. 48, no. 2, pp. 682–705. https://doi.org/10.1016/j.acha.2018.09.002

15. Galyaev, A.A., Lysenko, P.V., and Berlin, L.M., Statistical Complexity as a Criterion for the Useful Signal Detection Problem, *Autom. Remote Control*, 2023, vol. 84, no. 7, pp. 852–871. https://doi.org/10.25728/arcRAS.2023.27.32.001

16. Galyaev, A.A., Babikov, V.G., Lysenko, P.V., and Berlin, L.M., A New Spectral Measure of Complexity and Its Capabilities for Detecting Signals in Noise, *Dokl. Math.*, 2024, vol. 110. https://doi.org/10.1134/S1064562424702235

17. Richards, M.A., The Discrete-Time Fourier Transform and Discrete Fourier Transform of Windowed Stationary White Noise, *Technical Memorandum*, 2013, pp. 1–24.

18. DasGupta, A., *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*, New York: Springer, 2011. https://doi.org/10.1007/978-1-4419-9634-3

19. Frigyik, B., Kapila, A., and Gupta, M., Introduction to the Dirichlet Distribution and Related Processes, UWEE Technical Report no. UWEETR-2010-0006, 2010, pp. 1–28.

20. Hafeez, M., Kamal, S., and Shahbaz, M., The Multivariate Order Statistics for Exponential and Weibull Distributions, *Pak. J. Stat. Oper. Res.*, 2014, vol. 10, pp. 361–368. https://doi.org/10.18187/pjsor.v10i3.825

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 12, pp. 1184–1198.

STOCHASTIC SYSTEMS

On the Features of Service Rate Control in Fork-Join Queueing System

A. V. Gorbunova[*,a] **and A. V. Lebedev**[**,b]

[*] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
[**] *Lomonosov Moscow State University, Moscow, Russia*
e-mail: [a] *avgorbunova@list.ru,* [b] *avlebed@yandex.ru*

Received October 7, 2024
Revised October 16, 2024
Accepted October 18, 2024

Abstract—A classical fork-join queueing system is considered. The model is proposed for determining the optimal cost of such a system, taking into account the need to minimize the average response time simultaneously with reasonable costs for the resources required for this. The term "resources" within the framework of the mathematical model under study means the intensities of service on servers, the cost of which is directly proportional to the system performance, i.e., the rate of service requests. For the special case when the number of subsystems is equal to two, an exact analytical expression for determining the optimal cost is presented, for a more general case when the number of subsystems of a fork-join queueing system is greater than two, the equation is obtained, the numerical solution of which allows calculating the desired value. In addition, the asymptotic analysis of the obtained solutions behavior is carried out.

Keywords: fork-join queueing system, queueing system, optimal cost, control

DOI: 10.31857/S0005117924120043

1. INTRODUCTION

A classical fork-join queueing system (QS) is considered. Upon receipt in this system, a task is divided into K identical parts (subtasks), the number of which corresponds to the number of subsystems. Each subsystem is a system with an infinite storage and a single server. It is assumed that the service rates on all available servers are identical. A task is considered to be serviced after all its constituent parts have been serviced. Accordingly, the response time of the system (the time a task stays in the system) is determined by the maximum of K random times of subtasks staying in the subsystems.

Such systems are widely used to model various types of processes in which the task is divided or parallelized, in particular in the field of information technology we can talk about parallel or distributed computing, and they can also be used to model various work processes in manufacturing (for example, assembling a multi-component order in a warehouse or complex mechanisms consisting of many parts in production), finance (for example, processing a loan application in several divisions of a financial institution), healthcare (carrying out the necessary tests and collecting anamnesis upon admission of a patient to a medical institution), etc.

The key feature of this system, which complicates its analysis, is the existence of a dependence between the times of stay of subtasks in subsystems. Therefore, despite the relevance and demand for studying fork-join systems, the exact characteristics of the system's functioning were obtained only for the case of two subsystems ($K = 2$), in particular, the formula is known for the average response time of the system in the case of a Poisson input flow and exponential distributions of service times [1]. In other cases, only approximate expressions were found to approximate the main indicators of system performance [1–3]. The review [4] provides a more detailed overview of the known results. As for more recent studies, the works [5–12], including those by the authors of this article, in addition to formulas that refine the known estimates of the average response time or its variance present expressions for estimating characteristics such as quantiles of the response time distribution for a wider range of input flows or of service times. In addition, in [11] the exact expression for the correlation coefficient between the sojourn times of subtasks in subsystems of the $M|M|1$ type is presented, and in [9] for a fork-join system with a Pareto distribution of service time the estimate for the correlation coefficient was derived already in the case of a Pareto distribution of service time. It is also worth noting the Russian-language works [13–19]. In the series of papers, including [13–16], the analysis of a fork-join system with the infinite-server subsystems was performed in terms of generating functions for the probability distribution of the number of subtasks in each subsystem. In [17], one of the modifications of the fork-join system was analyzed, when, upon receipt, an task is divided not into a fixed, but into a variable number of subtasks, which is determined by the state of the system. In [18], an approach based on relation invariants was proposed for approximating the average response time of the fork-join QS, and in [19] the fork-join system is an integral part of the network and is used to model and study the performance characteristics of transaction service platforms.

This article examines another aspect of fork-join system performance, namely, the model of the cost of operating a system with division and parallel service is constructed, which allows determining optimal control in terms of optimizing its financial indicators. The model is based on natural assumptions about the need to minimize the average response time of the system to maintain its competitiveness at reasonable costs for the resources required for this. In particular, resources can be understood as the capacity of the necessary equipment that allows faster processing of a client request, if we are talking about information and computing or production systems, for example. It is clear that the more powerful the equipment, the more costs are required for its purchase, technical maintenance and maintenance in general. Thus, the rate of equipment operation (or in terms of QS, the rate of service) is proportional to the growth of its cost. In addition, as the rate of service increases, the response time of the system decreases. Thus, the cost of system operation consists of the optimal balance between the response time and the rate of the service servers.

In more detail, it is assumed that 1) the price (penalty) per unit of average response time and 2) the price per unit of service rate are set. In this case, the first price is assumed to be equal to one for simplicity. Then the response time and service costs are calculated and added to the total cost of expenses, which we want to minimize here. Thus, the task of cost optimization of management is set.

Similar problem statements can be found in the monograph [20], devoted to the optimal design of QS, including the optimal choice of arrival and service rates of tasks (assuming that these parameters are controllable) for various systems and queueing networks. However, for fork-join systems such problems have not been considered before.

The proposed functional dependence for the cost of operating a fork-join queueing system allows one to obtain an explicit expression for determining the optimal service rate on the system's servers in the case when the number of subsystems is two. If the number of subsystems is greater than two, then obtaining the optimal solution is possible in numerical form. The behavior of the optimal solution in extreme cases was also considered.

The paper is organized as follows: Section 2 describes a mathematical model for determining the optimal cost of operating a fork-join system in general; the next two sections derive the optimal value of the system load factor, which allows to express the optimal value of service intensity for the case of the Nelson–Tantawi formula for determining the system response time — for a particular case, i.e., when the number of subsystems corresponds to two and the formula is exact, and for a more general case when the formula is approximate; then the case of a generalization of the Nelson–Tantawi formula is analyzed, and then an asymptotic analysis of the obtained solutions is carried out, both in the general case and for specific expressions; in the conclusion, some results are summarized.

2. MATHEMATICAL MODEL FOR DETERMINING THE OPTIMAL COST OF FUNCTIONING OF A FORK-JOIN SYSTEM

We analyze a fork-join system with a Poisson input flow with intensity $\lambda > 0$ and an exponential distribution of service times on $K \geqslant 2$ homogeneous servers with intensity $\mu > 0$. The system load factor is $\rho = \lambda/\mu < 1$.

Let us present a unified mathematical approach that will be further applied in the work.

Let us denote the cost of the system operation by S and introduce the function $f(\rho)$, which defines the expression for the average response time of the system in the case of $\lambda = 1$, i.e., $f(\rho) = E[R_K]$ when $\lambda = 1$. Then, in the general case, the following expression will be valid

$$E[R_K] = \frac{1}{\lambda} f(\rho).$$

Since the cost of operating the system S depends on the average response time of the system (we take the price per unit of time as one unit) and the cost of maintenance costs, we can write for it

$$S = E[R_K] + c\mu,$$

where c is the cost of a unit of service intensity. Accordingly, taking into account the introduced function $f(\rho)$, we can rewrite this expression as follows:

$$S = \frac{1}{\lambda} f(\rho) + c\frac{\lambda}{\rho} = \frac{1}{\lambda}\left(f(\rho) + \frac{c\lambda^2}{\rho}\right).$$

Let $c_1 = c\lambda^2$, then

$$S = \frac{1}{\lambda}\left(f(\rho) + \frac{c_1}{\rho}\right). \tag{1}$$

Next, in order to find the optimal value of the system load factor we determine the extremum point of the cost function $S(\rho)$, namely the minimum point. Fo this purpose we find the derivative of the last expression and equate it to zero

$$S'_\rho = \frac{1}{\lambda}\left(f'(\rho) - \frac{c_1}{\rho^2}\right) = 0,$$

as a result we obtain the equation

$$f'(\rho)\rho^2 = c_1, \tag{2}$$

and having solved the equation (2) we find the optimal value of ρ_0 and, accordingly, the optimal value of the service intensity $\mu_0 = \lambda/\rho_0$.

Let us consider the basic case when $K = 1$, i.e., in fact, the $M|M|1$ system. The average response time in such a system is equal to

$$E[R] = \frac{1}{\mu - \lambda} = \frac{1}{\lambda}\frac{1}{\mu/\lambda - 1}.$$

Then the function $f(\rho)$, by taking into account that $\rho = \lambda/\mu$, is defined as

$$f(\rho) = \frac{1}{1/\rho - 1} = \frac{\rho}{1 - \rho} = \frac{1}{1 - \rho} - 1,$$

and the derivative of this function has the form

$$f'(\rho) = \frac{1}{(1-\rho)^2}.$$

Now we substitute the obtained expressions into the equation (2)

$$\left(\frac{\rho}{1-\rho}\right)^2 = c_1,$$

and obtain the value for the desired optimal system load at $K = 1$

$$\rho_0 = \frac{\sqrt{c_1}}{1+\sqrt{c_1}}. \tag{3}$$

In the next step we can calculate the optimal service rate, namely

$$\mu_0 = \frac{\lambda}{\rho_0} = \lambda\left(1 + \frac{1}{\sqrt{c_1}}\right) = \lambda + \frac{1}{\sqrt{c}}.$$

The similar problem was solved in [20, §1.1] for the case where the cost takes into account not the average response time, but the average waiting time.

Let us emphasize that since the equation (2) was obtained with respect to the load ρ, then in the future we will only look for the optimal load, understanding that it can, if necessary, be recalculated into the optimal service intensity.

3. ANALYSIS OF FORK-JOIN QS WITH TWO SUBSYSTEMS OF TYPE $M|M|1$. NELSON–TANTAWI FORMULA

To begin with, let us consider a special case and determine the optimal value of the system load factor when the number of subsystems is two. For the average response time at $K = 2$ the exact formula obtained in [1] is known, namely

$$E[R_2] = \frac{12-\rho}{8}\frac{1}{\mu-\lambda} = \frac{1}{\lambda}\frac{12-\rho}{8}\frac{\rho}{1-\rho} = \frac{1}{\lambda}\frac{\rho(12-\rho)}{8(1-\rho)}.$$

Thus, we have

$$f(\rho) = \frac{\rho(12-\rho)}{8(1-\rho)}.$$

Next, we find the derivative with respect to ρ and, substituting the expression for $f'(\rho)$ into (2), we can write the following equation:

$$\frac{\rho^2(\rho^2 - 2\rho + 12)}{(1-\rho)^2} = 8c_1.$$

For convenience, we will make the change

$$c_2 = 8c_1$$

and after simplification we obtain a fourth-degree equation

$$\rho^4 - 2\rho^3 + (12 - c_2)\rho^2 + 2c_2\rho - c_2 = 0, \tag{4}$$

which we will solve, since, as is known, for a fourth-degree equation there is an analytical solution in radicals. For this, we will use the Ferrari method [21–23].

Let us introduce the following notations:

$$A = -2, \quad B = 12 - c_2, \quad C = c_2, \quad D = -c_2,$$

then we get

$$\rho^4 + A\rho^3 + B\rho^2 + C\rho + D = 0.$$

By changing $\rho = y - A/4$ we reduce the fourth-degree equation (4) to the canonical form

$$y^4 + A_1 y^2 + B_1 y + C_1 = 0, \tag{5}$$

where

$$A_1 = B - \frac{3A^2}{8} = -c_2 + \frac{21}{2},$$
$$B_1 = \frac{A^3}{8} - \frac{AB}{2} + C = c_2 + 11,$$
$$C_1 = -\frac{3A^4}{256} + \frac{A^2 B}{16} - \frac{AC}{4} + D = -\frac{1}{4}c_2 + \frac{45}{16}.$$

Next, according to the Ferrari method, it is necessary to find one particular real solution of the cubic equation

$$A_2 t^3 + B_2 t^2 + C_2 t + D_2 = 0, \tag{6}$$

where

$$A_2 = 2, \quad B_2 = -A_1 = c_2 - \frac{21}{2},$$
$$C_2 = -2C_1 = \frac{1}{2}c_2 - \frac{45}{8}, \quad D_2 = A_1 C_1 - \frac{B_1^2}{4} = -\frac{175}{16}c_2 - \frac{23}{32}.$$

By changing $t = z - B_2/(3A_2)$ we reduce the equation (6) to the canonical form of a third-degree equation

$$z^3 + A_3 z + B_3 = 0, \tag{7}$$

where

$$A_3 = \frac{3A_2C_2 - B_2^2}{3A_2^2} = -\frac{1}{12}c_2^2 + 2c_2 - 12,$$
$$B_3 = \frac{2B_2^3 - 9A_2B_2C_2 + 27A_2^2 D_2}{27A_2^3} = \frac{1}{108}c_2^3 - \frac{1}{3}c_2^2 - \frac{3}{2}c_2 - 16.$$

The real root of the equation (7) according to Cardano's method [21–23] is determined as follows

$$z_1 = \left(-\frac{B_3}{2} + \sqrt{Q}\right)^{\frac{1}{3}} + \left(-\frac{B_3}{2} - \sqrt{Q}\right)^{\frac{1}{3}}, \quad Q > 0, \tag{8}$$

where

$$Q = \left(\frac{A_3}{3}\right)^3 + \left(\frac{B_2}{2}\right)^2 = -\frac{11}{432}c_2^4 + \frac{11}{12}c_2^3 - \frac{55}{16}c_2^2 + 44c_2. \tag{9}$$

At the same time, we must not forget that the following condition must be met

$$\left(-\frac{B_3}{2} + \sqrt{Q}\right)^{\frac{1}{3}} \left(-\frac{B_3}{2} - \sqrt{Q}\right)^{\frac{1}{3}} = -\frac{A_3}{3}.$$

Note that for the case $Q < 0$ for z_1 after the transformations the following formula will ultimately be valid:

$$z_1 = 2\sqrt{-\frac{A_3}{3}}\cos\frac{w}{3}, \tag{10}$$

where

$$w = \begin{cases} \arctan\left(\dfrac{-2\sqrt{-Q}}{B_3}\right), & B_3 < 0, \\ \arctan\left(\dfrac{-2\sqrt{-Q}}{B_3}\right) + \pi, & B_3 > 0, \\ \dfrac{\pi}{2}, & B_3 = 0, \end{cases}$$

and for the case $Q = 0$ we have

$$z_1 = 2\left(-\frac{B_3}{2}\right)^{\frac{1}{3}}. \tag{11}$$

The expression for Q from (9) is transformed to the form

$$-\frac{11}{432}c_2\left(c_2^3 - 36c_2^2 + 135c_2 - 1728\right)$$

and changes its sign depending on the value of c_2 (although c_2 should be positive based on its physical meaning), so we will explicitly indicate the regions of sign constancy of the expression for Q, the zeros of which can again be obtained using Cardano's method, namely

$$\begin{aligned} &Q > 0, \quad c_2 \in (0; \tilde{c}_2), \\ &Q < 0, \quad c_2 \in (-\infty; 0) \cup (\tilde{c}_2; +\infty), \end{aligned}$$

where $\tilde{c}_2 = (3 \times 11^{\frac{1}{3}} + 3 \times 11^{\frac{2}{3}} + 12) \approx 33.51$.

Accordingly, we get

$$t_1 = z_1 - \frac{B_2}{3A_2},$$

where z_1, depending on the value of Q, is determined by the expressions (8), (10) or (11). The particular solution t_1 allows us to represent the canonical equation of the fourth degree (5) as a product of two square trinomials

$$\left(y^2 - y\sqrt{2t_1 - A_1} + \frac{B_1}{2\sqrt{2t_1 - A_1}}\right)\left(y^2 + y\sqrt{2t_1 - A_1} - \frac{B_1}{2\sqrt{2t_1 - A_1}}\right) = 0.$$

The solution of one of the quadratic equations (taking into account the inverse substitution) will be the desired solution of the original equation of the fourth degree (4). In particular, the check showed that the desired one is the root of the second equation, so we can write the final solution as

$$y_0 = y_3 = \frac{-\sqrt{2t_1 - A_1} + \sqrt{Dis_2}}{2},$$

where the discriminant of the second equation is equal to

$$Dis_2 = -2t_1 - A_1 + \frac{2B_1}{\sqrt{2t_1 - A_1}}$$

and, accordingly,

$$\rho_0 = y_0 - \frac{A}{4} = y_0 + \frac{1}{2}.$$

4. ANALYSIS OF FORK-JOIN QS WITH $K > 2$ SUBSYSTEMS OF TYPE $M|M|1$. NELSON–TANTAWI FORMULA

Let us derive the equation for determining the optimal value of the system load factor in the general case when $K > 2$. For the mathematical expectation of the response time of a fork-join system, the Nelson–Tantawi approximate formula, which is considered one of the most accurate among those known, has the form [1]

$$E[R_K] \approx \left[\frac{H_K}{H_2} + \frac{4}{11}\left(1 - \frac{H_K}{H_2}\right)\rho\right]\frac{12-\rho}{8}\frac{1}{\mu-\lambda}, \tag{12}$$

where $H_K = \sum_{i=0}^{K} 1/i$ is the partial sum of the harmonic series.

We will solve the problem for the Nelson–Tantawi approximation under the assumption that the formula (12) is exact.

Let us introduce the following notations:

$$H = \frac{H_K}{H_2}, \quad M = \frac{4}{11}\left(1 - \frac{H_K}{H_2}\right) = \frac{4}{11}(1-H).$$

Then

$$E[R_K] = \frac{1}{\lambda}(H + M\rho)\frac{12-\rho}{8}\frac{\rho}{1-\rho},$$

therefore,

$$f(\rho) = (H + M\rho)\frac{12-\rho}{8}\frac{\rho}{1-\rho},$$

and for $f'(\rho)$ after transformations we obtain

$$f'(\rho) = \frac{1}{8(1-\rho)^2}(2M\rho^3 + (H - 15M)\rho^2 + (24M - 2H)\rho + 12H). \tag{13}$$

Then we substitute the resulting expression into (2)

$$\frac{\rho^2}{8(1-\rho)^2}(2M\rho^3 + (H - 15M)\rho^2 + (24M - 2H)\rho + 12H) = c_1.$$

Also, for convenience, we introduce the notation $c_2 = 8c_1$ and obtain a fifth-degree equation, which, after transformations taking into account that $M = 4(1-H)/11$, is reduced to the form

$$8(H-1)\rho^5 + (60 - 71H)\rho^4 + (118H - 96)\rho^3 + 11(c_2 - 12H)\rho^2 - 22c_2\rho + 11c_2 = 0. \tag{14}$$

The resulting equation (14) can be solved numerically and the optimal value of ρ_0 and, accordingly, the desired value of μ_0 can be determined.

Figure 1 presents graphs of the dependence of the behavior of the optimal value of the load factor $\rho = \rho_0$, which is the solution of the equation (14), on the value of the parameter c_1 for a different number of subsystems K of the fork-join QS. At the initial stage, there is a fairly rapid growth of the optimal load with an increase in the cost of a unit of resource and, accordingly, the performance of the system as a whole. Moreover, even for the number of subsystems $K = 1000$, i.e., in other words, given the division of the task and the processing of its subtasks on a fairly large number of servers, already at $c_1 > 5.5$ the required load level will be $\rho > 0.5$. At the same time, with further growth of the parameter c_1, a rather slow growth of the optimal load is observed, for example, with $c_1 \approx 158$, which is almost 29 times greater than $c_1 = 5.5$, the value $\rho_0 = 0.85$ will be, i.e., still does not exceed 90%. In addition, based on the type of graphs, the effect is observed

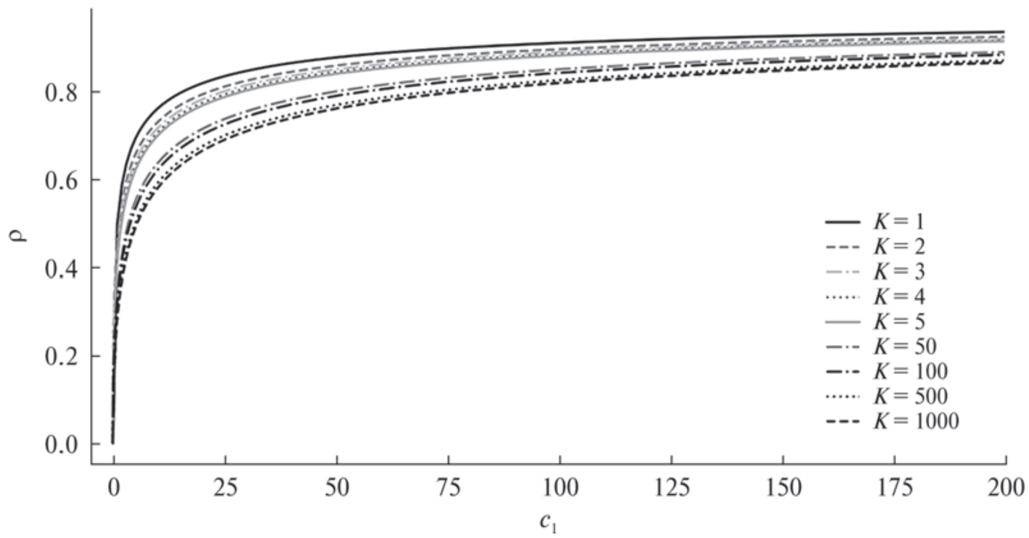


Fig. 1. Graph of the dependence of the optimal system load value ρ (solution of the equation (14)) on the parameter c_1 for a different number of subsystems K.

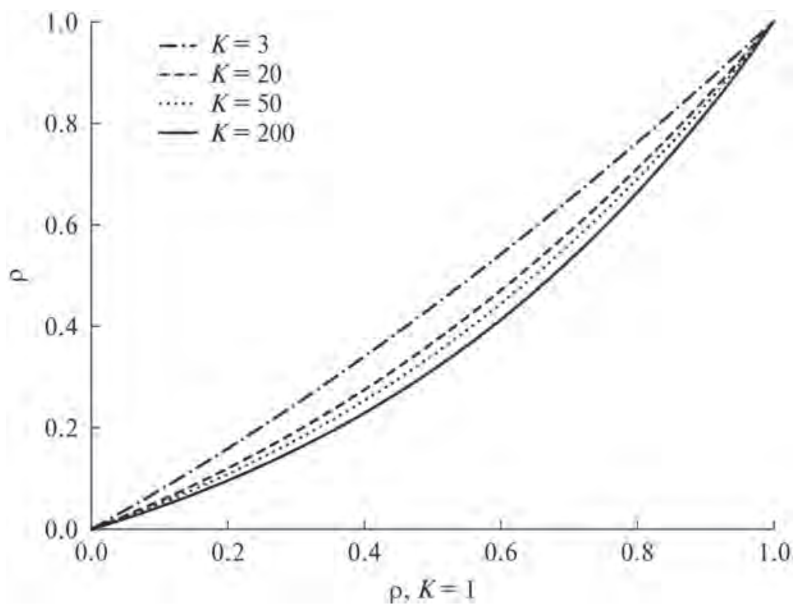


Fig. 2. Graph of the dependence of the optimal system load value ρ on the optimal value of the load ρ at $K = 1$.

that with the growth of K, the optimal level of load on the system decreases, which was expected and quite natural.

Figure 2 presents the graph of the dependence of the behavior of the optimal value of the system load factor on the optimal value of ρ for $K = 1$ according to (3). The graph allows us to compare the level of optimal load for different values of $K \geqslant 2$ with the level of optimal load in the case of $K = 1$. As can be seen, with an increase in the number of subsystems, the level of load required for optimal system operation decreases, i.e., the line bends more and more under the straight line $\rho = \rho_{K=1}$ and becomes more and more convex (below any chord connecting any two points on the graph selected on the interval under consideration).

5. ANALYSIS OF FORK-JOIN QS WITH $K > 2$ SUBSYSTEMS OF TYPE $M|M|1$. GENERALIZATION OF NELSON–TANTAWI FORMULA

Consider a generalization of the Nelson–Tantawi formula from [10], in which it was shown that the refined formula gives a better approximation for the average response time. The improvement is achieved by making some correction to the expression from (12), which we denote by $E[R_K]_{NT}$. Thus, the approximation is

$$E[R_K] = \frac{\rho}{\mu-\lambda}\left(\frac{H_K}{H_2}-1\right)\left(Q_1 - Q_2\left(\frac{H_K}{H_2}-1\right)+Q_3\rho\right)+E[R_K]_{NT}, \tag{15}$$

where

$$Q_1 \approx 0.087197, \quad Q_2 \approx 0.070236, \quad Q_3 \approx 0.09638.$$

In order to define an expression for $f(\rho)$ for this case, we take $1/\lambda$ out of the brackets

$$E[R_K] = \frac{1}{\lambda}\left[-\frac{\rho^2}{1-\rho}\left(1-\frac{H_K}{H_2}\right)\left(Q_1+Q_2\left(1-\frac{H_K}{H_2}\right)+Q_3\rho\right)+\lambda E[R_K]_{NT}\right],$$

and as the result is

$$f(\rho) = -\frac{\rho^2}{1-\rho}\left(1-\frac{H_K}{H_2}\right)\left(Q_1+Q_2\left(1-\frac{H_K}{H_2}\right)+Q_3\rho\right)+\lambda E[R_K]_{NT},$$

where $\lambda E[R_K]_{NT}$ was calculated earlier and has the form

$$\lambda E[R_K]_{NT} = f_2(\rho) = (H+M\rho)\frac{12-\rho}{8}\frac{\rho}{1-\rho}.$$

Thus, taking into account the previously proposed replacement $H = H_K/H_2$, we have

$$f(\rho) = -\frac{\rho^2}{1-\rho}(1-H)(Q_1+(1-H)Q_2+Q_3\rho)+f_2(\rho) = f_1(\rho)+f_2(\rho).$$

Next we take the derivative of the obtained expression

$$f'(\rho) = f_1'(\rho)+f_2'(\rho),$$

where

$$f_1'(\rho) = \frac{\rho(1-H)}{(1-\rho)^2}\Big(2Q_3\rho^2-\rho(3Q_3-(1-H)Q_2-Q_1)-(2Q_1+2(1-H)Q_2)\Big),$$
$$f_2'(\rho) = \frac{1}{8(1-\rho)^2}\Big(2M\rho^3+(H-15M)\rho^2+(24M-2H)\rho+12H\Big).$$

Accordingly, after substituting the obtained expressions into (2) we obtain that

$$c_1 = \rho^2 f'(\rho) = \rho^2(f_1'(\rho)+f_2'(\rho)). \tag{16}$$

The equation (16), as in the case of the Nelson–Tantawi formula from the previous section, is a fifth-degree equation that can be solved numerically, after which the optimal value of μ_0 will be found.

Figure 3 presents the graphs of the optimal load value $\rho = \rho_0$, which is the solution of (16). The behavior of the graphs is generally similar to the case of the Nelson–Tantawi formula (12). In order

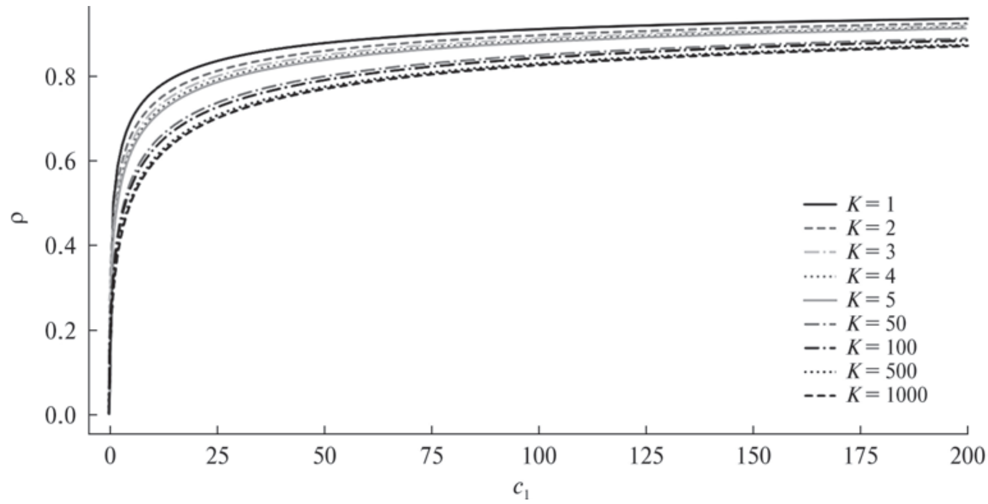


Fig. 3. Graph of the dependence of the optimal system load value ρ (solution of the equation (16)) on the parameter c_1 for a different number of subsystems K.

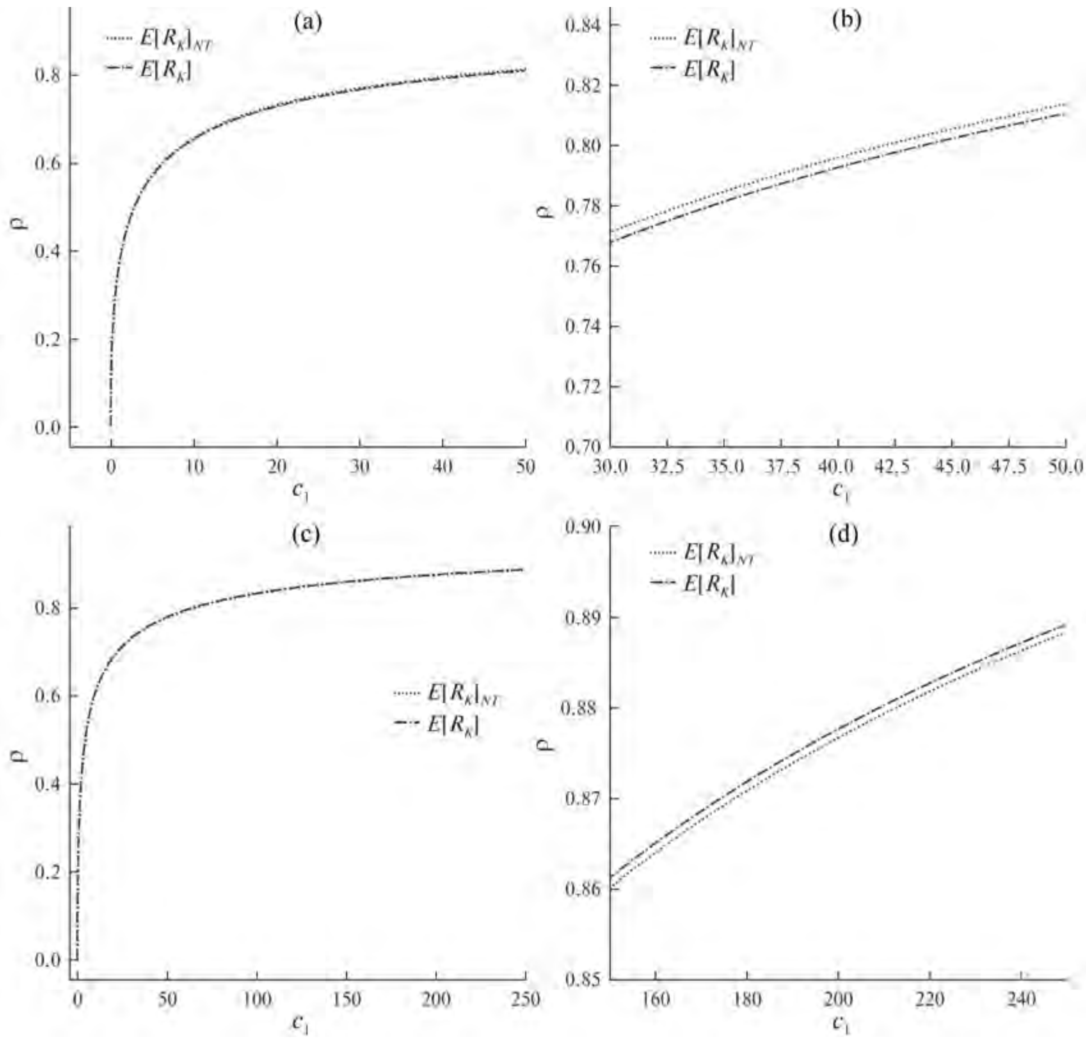


Fig. 4. Graphs of the optimal load value $\rho = \rho_0$, which is the solution to the equation (16) for $K = 20$ (Figs. 4a and 4b) and $K = 200$ (Figs. 4c and 4d).

to analyze the difference between the obtained results in more detail, we will compare the behavior of the optimal solution for the special cases of the number of subsystems $K = 20$ and $K = 200$.

Figure 4 presents the graphs of the dependence of the optimal system load level $\rho = \rho_0$ on the parameter c_1 for the cases of the Nelson–Tantawi formula for the average system response time $E[R_K]_{NT}$ from (12) and the equation (14), as well as a generalization of the Nelson–Tantawi formula for the average system response time $E[R_K]$ from (15) and the equation (16) for $K = 20$ (Figs. 4a and 4b) and $K = 200$ (Figs. 4c and 4d), including the scaled version (Figs. 4b and 4d).

In Figs. 4a and 4b for $K = 20$ it is evident that for the case of the Nelson–Tantawi formula (12) for estimating the average system response time, the optimal load value ρ exceeds the optimal value calculated by the generalized formula (15). However, for graphs 4c and 4d for $K = 200$ the opposite situation can be observed.

In order to thoroughly understand the behavior of optimal solutions in both cases, we will further analyze their asymptotics.

6. ASYMPTOTICS OF THE BEHAVIOR OF THE OPTIMAL SOLUTION

Let us consider the equation (2) and determine the behavior of its solution as $c_1 \to 0$ ($\rho \to 0$) and $c_1 \to +\infty$ ($\rho \to 1$) tend in general. First, let us analyze the case of $\rho \to 0$, respectively $c_1 \to 0$. Then we have

$$\rho^2 f'(\rho) \sim \rho^2 f'(0),$$

then we substitute the obtained result into (2), i.e., $\rho^2 f'(\rho) = c_1$, whence it follows that

$$\rho \sim \sqrt{\frac{c_1}{f'(0)}}, \quad c_1 \to 0. \tag{17}$$

Now let's analyze the case $\rho \to 1$, respectively $c_1 \to +\infty$. In the general case we have

$$\rho^2 f'(\rho) \sim f'(\rho).$$

Thus, if there exists a number $L \in (0, +\infty)$ such that

$$L = \lim_{\rho \to 1} (1-\rho)^2 f'(\rho),$$

then

$$f'(\rho) \sim \frac{L}{(1-\rho)^2},$$

therefore, when substituting the obtained expression for $f'(\rho)$ c L into (2) and taking into account the asymptotics, we have the following

$$\begin{gathered} \rho^2 f'(\rho) = c_1, \\ f'(\rho) \sim c_1, \quad \rho \to 1, c_1 \to +\infty, \\ (1-\rho)^2 \sim \frac{L}{c_1}, \quad \rho \to 1, c_1 \to +\infty, \end{gathered}$$

therefore

$$\rho = 1 - \sqrt{\frac{L}{c_1}}(1 + o(1)), \quad c_1 \to \infty. \tag{18}$$

Next, we will consider the existing models in detail, namely the cases of the Nelson–Tantawi formula and its generalization, and we will define specific expressions for the obtained equivalences from the general case.

For the Nelson–Tantawi formula for $K \geqslant 2$, taking into account the expression (13), it is true

$$f'(0) = \frac{3}{2}\frac{H_K}{H_2},$$

therefore, when substituting into (17) for $c_1 \to 0$ ($\rho \to 0$), we obtain

$$\rho \sim \sqrt{\frac{2}{3}\frac{H_2}{H_K}c_1}, \quad c_1 \to 0. \tag{19}$$

Now we will determine the asymptotics of the solution for $c_1 \to +\infty$ ($\rho \to 1$). To do this, we will find the value of L

$$L = \lim_{\rho\to 1}(1-\rho)^2 f'(\rho) = \lim_{\rho\to 1}\frac{2M\rho^3 + (H-15M)\rho^2 + (24M-2H)\rho + 12H}{8} = \frac{7}{8}\frac{H_K}{H_2} + \frac{1}{2}. \tag{20}$$

Finally we get

$$\rho = 1 - \sqrt{\frac{1}{c_1}\left(\frac{7}{8}\frac{H_K}{H_2} + \frac{1}{2}\right)}(1 + o(1)).$$

Now let us analyze the generalization of the Nelson–Tantawi formula. Similarly, consider the equation (16) and determine the behavior of its solution as $c_1 \to 0$ ($\rho \to 0$) and $c_1 \to \infty$ ($\rho \to 0$) tend. We analyze the case of $\rho \to 0$, respectively, $c_1 \to 0$.

For the formula (15) with $K \geqslant 2$ it is true

$$f'(0) = f_1'(0) + f_2'(0) = \frac{3}{2}\frac{H_K}{H_2},$$

therefore, as before,

$$\rho \sim \sqrt{\frac{2}{3}\frac{H_2}{H_K}c_1}, \quad c_1 \to 0.$$

Now let us analyze the case $\rho \to 1$ ($c_1 \to \infty$). For (15) with $K \geqslant 2$ we get

$$L = \lim_{\rho\to 1}(1-\rho)^2 f'(\rho) = \lim_{\rho\to 1}(1-\rho)^2(f_1'(\rho) + f_2'(\rho)) = L_1 + L_2.$$

In fact, the value of L_2 was calculated earlier and corresponds to the value of L from (20)

$$L_2 = \frac{7}{8}\frac{H_K}{H_2} + \frac{1}{2}.$$

Now let's calculate L_1

$$\begin{aligned} L_1 &= \lim_{\rho\to 1}(1-\rho)^2 f_1'(\rho) \\ &= \lim_{\rho\to 1}\rho(1-H)\Big(2Q_3\rho^2 - \rho(3Q_3 - (1-H)Q_2 - Q_1) - (2Q_1 + 2(1-H)Q_2)\Big) \\ &= (H-1)(Q_1 - (H-1)Q_2 + Q_3) = \left(\frac{H_K}{H_2} - 1\right)\left[Q_1 - \left(\frac{H_K}{H_2} - 1\right)Q_2 + Q_3\right] \\ &\approx \left(\frac{H_K}{H_2} - 1\right)\left[0.183577 - 0.070236\left(\frac{H_K}{H_2} - 1\right)\right]. \end{aligned}$$

Thus, we obtain that

$$\rho = 1 - \sqrt{\frac{L}{c_1}}(1 + o(1)),$$

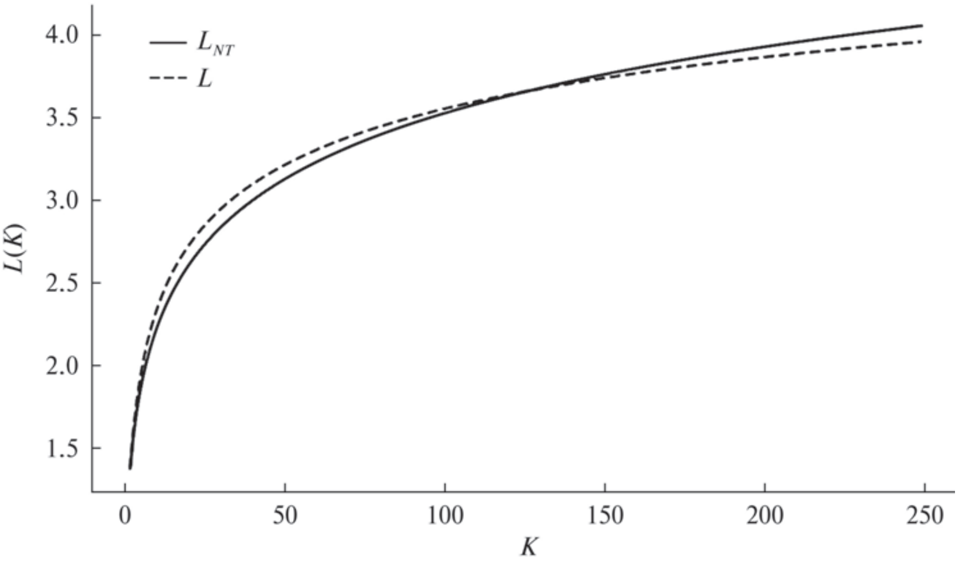


Fig. 5. Graph of the dependence of the value of L on K for the case of the Nelson–Tantawi formula (12) and the case of the generalized formula (15).

where

$$L = \left(\frac{H_K}{H_2} - 1\right)\left[Q_1 - \left(\frac{H_K}{H_2} - 1\right)Q_2 + Q_3\right] + \frac{7}{8}\frac{H_K}{H_2} + \frac{1}{2}. \tag{21}$$

Figure 5 shows the graph of the dependence of the value of L on the number of subsystems for the case of the Nelson–Tantawi formula (20) and the case of the generalized formula (21). After the value $K = 126$, a "break" in the graphs occurs, and if at first the value of L_{NT} exceeded the value of L for the generalized formula, then for values $K \geqslant 127$ the situation changed exactly the opposite, which is also confirmed by the graphs presented earlier for the optimal value of the load depending on the parameter c_1 in Fig. 4 for $K = 20$ and $K = 200$.

As for the practical application of the obtained asymptotic results, both cases of high ($\rho \to 1$) and low ($\rho \to 0$) load may be of interest. For example, in systems with intensive use of data, parallel and distributed computing have become widespread as one of the main ways to improve performance in processing big data. Therefore, the owners of such high-performance computing environments are interested in obtaining forecasts of the system behavior under peak loads, as well as in the opposite situations, when the platforms are least in demand, in order to develop strategies for the efficient operation of the systems. In this context, asymptotic formulas can suggest when (at what prices per unit of service rate) such modes (high and low load) are optimal, i.e., actually desirable for the functioning of the system, in terms of cost, and when not.

7. CONCLUSION

The paper studies a fork-join queueing system from the point of view of managing the optimal cost of its operation depending on the price of a unit of resource affecting the system's performance and, accordingly, its response time. The mathematical model is constructed that takes into account the optimal ratio of cost and efficiency of the system's operation. The analysis is carried out based on previously obtained approximations for the average response time, both one of the most well-known and accurate, and its generalization obtained by the authors of this paper. For the particular case of the system, i.e., when the number of subsystems is two, it is possible to derive in explicit form expressions for the optimal system load. For a larger number of subsystems, equations of the fifth degree are presented, the numerical solution of which allows one to determine the sought values of the system load and, accordingly, the service intensity, which actually characterizes the "power" of the necessary resources. The asymptotic behavior of the system is also analyzed.

Since the approximations of the average response time used here are quite accurate, they also describe the total cost of expenses well, so that with the calculated values of the optimal load (and, accordingly, the optimal service rate), the cost will be close to the optimal one. Therefore, the calculated values of the load can be recommended for use as estimates of their actual values for fork-join systems. In addition, if there is an estimate of the optimal load for one number of subsystems, it can be recalculated into an estimate for another number of subsystems (see Fig. 2). Since in the general case the problem is solved only numerically, then in cases of high and low prices per unit of service rate, it can be recommended to use asymptotic formulas that give simpler, but at the same time rougher estimates of the optimal load.

The mathematical model proposed in the article for control the optimal functioning of the system will also be valid in more general cases of fork-join QS, namely with a non-Poisson input flow and/or a non-exponential distribution of service time on servers, as in [9]. In this case, of course, analytical approximations for the value of the average response time of the system are necessary to derive specific relationships or equations. In addition, one can consider models with a nonlinear dependence of the cost of service costs on the service rate (for example, a power dependence, as mentioned in [20, §1.1]).

REFERENCES

1. Nelson, R. and Tantawi, A.N., Approximate Analysis of Fork/Join Synchronization in Parallel Queues, *IEEE Transact. Comput.*, 1988, vol. 37, pp. 739–743.

2. Varma, S. and Makowski, A.M., Interpolation Approximations for Symmetric Fork-Join Queues, *Performance Evaluat.*, 1994, vol. 20, pp. 245–265.

3. Varki, E., Merchant, A., and Chen, H., The M/M/1 Fork-Join Queue with Variable Subtasks, *Unpublished*, available online: https://www.cs.unh.edu/?varki/publication/2002-nov-open.pdf

4. Thomasian, A., Analysis of Fork/Join and Related Queueing Systems, *ACM Computing Surveys (CSUR)*, 2014, vol. 47, no. 2, pp. 1–71.

5. Qiu, Z., Pérez, J.F., and Harrison, P.G., Beyond the Mean in Fork-Join Queues: Efficient Approximation for Response-Time Tails, *Performance Evaluat.*, 2015, vol. 91, pp. 99–116.

6. Nguyen, M., Alesawi, S., Li, N., Che, H., and Jiang, H., ForkTail: A Black-Box Fork-Join Tail Latency Prediction Model for User-Facing Datacenter Workloads, *Proc. 27th Int. Symp. High-Perform. Parallel Distrib. Comput.*, 2018, pp. 206–217.

7. Nguyen, M., Alesawi, S., Li, N., Che, H., and Jiang, H., A Black-Box Fork-Join Latency Prediction Model for Data-Intensive Applications, *IEEE Transact. Parall. Distrib. Syst.*, 2020, vol. 31, no. 9, pp. 1983–2000.

8. Enganti, P., Rosenkrantz, T., Sun, L., Wang, Z., Che, H., and Jiang, H., ForkMV: Mean-and-Variance Estimation of Fork-Join Queuing Networks for Datacenter Applications, *Proc. IEEE International Conference on Networking, Architecture and Storage (NAS)*, 2022, pp. 1–8.

9. Gorbunova, A.V. and Lebedev, A.V., Nonlinear Approximation of Characteristics of a Fork-Join Queueing System with Pareto Service as a Model of Parallel Structure of Data Processing, *Math. Comput. Simulat.*, 2023, vol. 214, pp. 409–428.

10. Gorbunova, A.V. and Lebedev, A.V., On Estimating the Characteristics of a Fork-Join Queueing System with with Poisson Input and Exponential Service Times, *Advanc. Syst. Sci. Appl.*, 2023, vol. 23, pp. 99–114.

11. Gorbunova, A.V. and Lebedev, A.V., Correlations of the Sojourn Times of Subtasks in Fork-Join Queueing Systems with $M|M|1$-type Subsystems, *Advanc. Syst. Sci. Appl.*, 2024, vol. 24, no. 2, pp. 1–18.

12. Gorbunova, A.V. and Lebedev, A.V., Copulas and Quantiles in Fork-Join Queueing Systems, *Advanc. Syst. Sci. Appl.*, 2024, vol. 24, no. 1, pp. 1–19.

13. Ivanovskaya, I.A. and Moiseeva, S.P., Research of a mathematical model of parallel servicing of mixed-type requests, *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika*, 2010, vol. 317, no. 5, pp. 32–34.

14. Zhidkova, L.A. and Moiseeva, S.P., Study of the parallel service system of multiple requests of the simplest flow, *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika*, 2011, vol. 17, no. 4, pp. 49–54.

15. Moiseeva, S.P. and Zahorolnaya, I.A., Mathematical model of parallel servicing of multiple requests with repeated requests, *Avtometriya*, 2011, vol. 47, no. 6, pp. 51–58.

16. Moiseeva, S.P., Pankratova, E.V., and Ubonova, E.G., Study of infinite queueing system with heterogeneous service and input Markov renewal flow, *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika*, 2016, vol. 35, no. 2, pp. 46–53.

17. Osipov, O.A., A Heterogeneous Fork-Join Queueing System in Which Each Job Occupy All Free Servers, *RUDN Journal of Mathematics, Information Sciences and Physics*, 2018, vol. 26, no. 1. pp. 28–38.

18. Khabarov, R.S., Lokhvitsky, V.A., and Dudkin, A.S., Sojourn Time Approximation for Fork-Join Queue Based on Relationship Invariants, *Intellectual Technologies on Transport*, 2020, vol. 22, no. 2, pp. 46–50.

19. Redrugina, N.A., Method for Time Characteristics Calculating in the Service Platforms of Infocommunication Transactional Services with Parallel Requests Processing, *Proceedings of Telecommunication Universities*, 2023, vol. 9, no. 3, pp. 82–90.

20. Stidham, S., *Optimal design of queueing systems*, Boca Raton: CRC Press/Taylor & Francis, 2009.

21. Spiegel, M.R., Lipschutz, S., and Liu, J., *Mathematical Handbook of Formulas and Tables, McGraw Hill Professional, 3rd (Third) edition*, Schaum's Outline Series, 2008.

22. Kurosh, A.G., *Algebraicheskie uravneniya proizvol'nyh stepenei (Populyarnye lektsii po matematike; vyp. 7)* (Algebraic equations of arbitrary degrees (Popular lectures on mathematics; issue 7)), Moscow: Nauka, 1975.

23. Kurosh, A.G., *Kurs vysshei algebry: Uchebnik* (Higher Algebra Course: Textbook), Saint Petersburg: Lan', 2008.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 12, pp. 1199–1208.

A Method of Optimal Investment with Stage-by-Stage Conditional Value at Risk (CVaR) Constraints and Known Parameters of Return Vectors

A. Y. Golubin[*,**,a], **V. N. Gridin**[**,b], **D. S. Smirnov**[**,c], **and S. A. Bulgakov**[*,d]

[*] *National Research University Higher School of Economics, Moscow, Russia*
[**] *Design Information Technologies Center Russian Academy of Sciences, Odintsovo, Moscow region, Russia*
e-mail: [a] *agolubin@hse.ru,* [b] *info@ditc.ras.ru,* [c] *info@ditc.ras.ru,* [d] *s.a.bulgakov@gmail.com*

Received February 13, 2024
Revised September 30, 2024
Accepted October 9, 2024

Abstract—The paper investigates a multi-stage investment problem under Conditional Value at Risk (CVaR) constraints with: a given security level for bankruptcy, short selling permission, a normal and an elliptical total return distribution models. The purpose of the work is to find a method of determining the optimal investment in this problem at each stage. As a result of the study, an optimal investment strategy is found and it is shown that the optimal investment portfolio at each stage does not depend on the value of the investor's capital, but depends only on the number of stage. It is shown that the multi-stage problem can be reduced to a finite number of one-stage optimization problems, which are problems of conic programming. For the one-stage problem, conditions for the non-emptiness of the set of admissible portfolios are given and the Kuhn–Tucker theorem is applied. Additionally, this paper presents a numerical example of finding the optimal investment based on the open data on rates of assert prices of companies on the stock exchange.

Keywords: optimal investment portfolios, risk measure, probability constraint CVaR, bankruptcy

DOI: 10.31857/S0005117924120058

1. INTRODUCTION

The formalization of the concept of "optimal investment portfolio" and methods for its finding were first proposed by Markowitz in 1952 in [1], where a mathematical model for the determination of an optimal investment portfolio was described. Markowitz used the variance of returns as a risk measure that allows one to obtain an estimate of the financial risk for a portfolio of assets. After using variance as a risk measure, many scientists have explored alternative risk measures [2, 3]. One such risk measure is the Value at Risk (VaR), which has been widely used since the 1980s. VaR is an estimate of the amount that expected losses will not exceed with a given probability, expressed in monetary terms. In the [4] quantiles, i.e., VaR constraints, were used in the formulation and solutions of optimization problems. In [5], variance as a risk measure, which Markowitz used in his work, is compared with the risk measure VaR. The authors conclude that, in general, the risk measure VaR has advantages, but is not free from disadvantages; for example, using VaR it is impossible to estimate the size of losses outside the confidence level. Some researchers [6] have pointed out that VaR is not a coherent measure of risk (see the definition, for example, in [6]), since it does not satisfy the subadditivity property except in the case of a normal risk distribution. On the other hand, [5, 7] noted that the risk measure VaR is widely used by regulators in the financial industry.

The risk measure CVaR introduced in the literature in recent decades is a coherent risk measure; in [6], the authors compared the risk measure VaR and CVaR and came to the conclusion that in the absence of a risk-free asset in the investment portfolio, the risk measure CVaR is more effective than the risk measure VaR. In addition, the authors note that CVaR takes into account outliers, which is critical for valuing risky and volatile assets. In [8], the authors conclude that despite the advantage of the risk measure CVaR over the risk measure VaR in the general case, it makes sense to take into account both of these measures to estimate the risk of an investment portfolio.

In [9], a multi-stage investment problem with VaR constraints but without the possibility of bankruptcy was investigated. On the other hand, (see, for example, [10]) bankruptcy, i.e., a decrease in the investor's capital below a given value, when further investment transactions are prohibited for the remaining time interval, plays a significant role in the estimation of financial strategy. An optimal investment strategy with stage-by-stage quantile (i.e., VaR) constraints was found in [11], where the possibility of bankruptcy is assumed.

In the presented work, a multi-stage investment problem with the possibility of bankruptcy is studied, where, unlike [11], stage-by-stage CVaR constraints are imposed; the advantages of this approach over the introduction of VaR constraints are described above. It is shown that the considered multi-stage risk sharing problem has a solution in which each optimal portfolio at stage t depends only on the stage number t and does not depend on the value of the investor's current capital $(x > 0)$. It is proved that the original multi-stage problem is reduced to solving a finite number of one-stage cone optimization problems, where the objective functions are determined by a recurrent formula, which makes it relatively easy to find the optimal investment strategy. Unlike [12, 13], where the risk of bankruptcy at each stage was only estimated by the use of Chebyshev's inequality, here the authors investigate the problem by solving dynamic programming equations that take into account the possibility of bankruptcy.

Section 2 considers a one-stage problem with CVaR constraint. Necessary and sufficient conditions for the set of admissible portfolios to be non-empty and for the Slater regularity condition to be satisfied are given, and the Kuhn–Tucker theorem is used to determine the solution to this cone optimization problem. Section 3 is devoted to the study of optimization of investment strategy in a multi-stage problem with CVaR constraints. It is shown that each component of the optimal portfolio depends only on the investment stage number and does not depend on the investor's current capital; the necessary conditions for the optimality of the strategy are found. A model different from the previously considered normal model is investigated in Section 4, where a generalization of the normal model to elliptical distributions is studied. In Section 5, a numerical example is solved, illustrating, based on real data, the finding of an optimal strategy for investing in three large companies. Section 6 contains concluding remarks.

2. ONE-STAGE OPTIMAL PORTFOLIO CONSTRUCTION PROBLEM

Let us first study a one-stage model for choosing an optimal investment portfolio (see, for example, [6, 11]), where the random vector of asset returns $R = (R_0, \dots, R_n)$, and R_i represents the change in the value of the ith asset from the current value as a percentage. In terms of stock prices, this means that $R_i = p_1/p_0$, where p_0 is the current price of the ith asset (a deterministic variable), p_1 is the price of the ith asset at the next quotation (a random variable). Let $R_0 = m_0$ almost surely (a.s.), i.e., it is a risk-free asset. Let $\overline{a} \in R^{n+1}$ is an investment portfolio where a_i is the percentage of the initial capital $x_0 > 0$ invested in the ith asset. The budget constraint $\sum_{i=0}^{n} a_i = 1$ means the investor's self-financing (there is no inflow of funds from outside, and the investor's available funds are invested only in the assets of this market) and, at the same time, permission for "short sales," i.e., the possibility of borrowing some assets at their current value with the purpose of investing this money in other assets. The function to be maximized is the mathematical expectation of the final investor's capital

$$EX_{\overline{a}} = Ex_0 \sum_{i=0}^{n} a_i R_i = x_0 \sum_{i=0}^{n} a_i m_i,$$

where $m_i = ER_i$ and $x_0 > 0$ is the initial capital.

To formulate the CVaR risk measure (see [6]) in the problem under investigation, we first need to define the risk measure Value at Risk (VaR) for random income Z:

$$\mathrm{VaR}_\alpha(Z) = -z_\alpha = -\inf\left\{t : P\left(Z < t\right) \geqslant \alpha\right\},$$

where $\alpha \in (0, 1/2)$ is a given significance level. Conditional Value at Risk (CVaR) is a risk measure that has the meaning of expected losses in the case of exceeding the conditional risk measure VaR with a given significance level α:

$$\mathrm{CVaR}_\alpha\left(Z\right) = E\left\{Z|\, Z \geqslant \mathrm{VaR}_\alpha\left(Z\right)\right\}.$$

Let $f_Z(z)$ denote the density function of the standardized income distribution $(Z - EZ)/\sqrt{DZ}$. Then [6]

$$\mathrm{CVaR}_\alpha\left(Z\right) = -EZ + \sqrt{DZ}k,$$

where

$$k = \frac{-\int\limits_{-\infty}^{-z_\alpha} z f_Z(z)dz}{\alpha}.$$

Assuming a normal approximation of the investor's final capital $X_a = x_0 \sum_{i=0}^{n} a_i R_i$, which is widely used in portfolio theory (see, for example, [6, 9]), we obtain an expression for the parameter k (see its definition above). Since $\phi'(x) = -x\phi(x)$, where $\phi(x)$ is the density of the standard normal distribution, it is easy to see that $k = \phi\left(\Phi^{-1}\left(\alpha\right)\right)/\alpha$. Then

$$\mathrm{CVaR}_\alpha\left(X_a\right) = -EX_a + \sqrt{DX_a}\frac{\varphi(\Phi^{-1}(\alpha))}{\alpha},$$

where $\Phi\left(x\right)$ is the standard normal distribution function, α is the specified significance level.

Let us introduce the constraint CVaR for the problem under consideration:

$$\mathrm{CVaR}_\alpha\left(X_a\right) \equiv -(x_0\langle a, \Delta m\rangle + x_0 m_0) + x_0\sqrt{aCa'}\frac{\phi\left(\Phi^{-1}\left(\alpha\right)\right)}{\alpha} \leqslant dx_0.$$

Here Δm denotes the vector $(m_1 - m_0, \ldots, m_n - m_0)$, $\langle a, \Delta m\rangle$ is the scalar product $\sum_{i=1}^{n} a_i \Delta m_i$, d is the share in percent of the initial capital $x_0 > 0$, C is the covariance matrix of $n \times n$ risky assets.

Summarizing all the above considerations, we formulate a one-stage optimization problem:

$$\left\{\begin{array}{l} EX_a \equiv \langle a, \Delta m\rangle + m_0 \;\rightarrow \max, \\ a \in D = \left\{a \in R^n :\, \sqrt{aCa'} \leqslant \dfrac{\langle a, \Delta m\rangle + m_0 + d}{\varphi(\Phi^{-1}(\alpha))/\alpha}\right\}. \end{array}\right. \tag{1}$$

Below we will use the following natural assumptions: $0 < m_0 < \min_{i=1,\ldots,n} m_i$, the covariance matrix of risky assets C is positive definite.

Further in this section, we simply present modifications of the statements obtained earlier (see [11, 14]) for the VaR constraints and for the case of CVaR constraints (see (1)), which will be used in solving a multi-stage problem.

By definition, a second-order cone (see, e.g., [15]) is $K = \{(a,t) \in R^{n+1} : \sqrt{aCa'} \leqslant t\}$. It is known that such a cone is regular, in other words, it is convex, closed, $IntK \neq \emptyset$ and, if $x \in K$, $-x \in K$, then $x = 0$. Problem (1) can be rewritten as a cone programming problem [15]

$$\max\langle a, \Delta m\rangle \text{ under constraints } \left(a, \frac{\langle a, \Delta m\rangle + m_0 + d}{\frac{\varphi(\Phi^{-1}(\alpha))}{\alpha}}\right) \in K. \tag{2}$$

To solve (2), we need a description of the cone dual to K. By definition, the dual cone is $K^* = \{x \in R^{n+1} : \langle x, y\rangle \geqslant 0 \text{ for all } y \in K\}$. In problem (2), it is defined [3] as

Lemma 1. *The dual cone is equal to*

$$K^* = \left\{(u,v) \in R^{n+1} : \sqrt{uC^{-1}u'} \leqslant v\right\}.$$

Below we present a condition sufficient for the fulfillment of the Slater condition (see the definition, for example, in [15]) in problem (2). Let j denote the index at which the function $\sigma_i \frac{\varphi(\Phi^{-1}(\alpha))}{\alpha} - \Delta m_i$ reaches its minimum, $i = 1, \ldots n$, where $\sigma_i = \sqrt{DR_i}$ is the standard deviation.

Statement 1. *If*

$$\sigma_j \frac{\varphi(\Phi^{-1}(\alpha))}{\alpha} - \Delta m_j < 0, \tag{3}$$

then the interior $IntD \neq \emptyset$, i.e., the Slater condition in problem (1) *is met and, therefore, in problem* (2) *also.*

Proof. Let a^j be an investment portfolio of the form $(0, \ldots, 0, a_j, 0, \ldots 0)$, where a_j is in the jth place, and all other components of the investment portfolio are zeros. Note that in the case under consideration short sales are allowed, i.e., a_j can be less than zero. CVaR the constraint in the one-stage problem (1) takes the form

$$|a_j|\,\sigma_j \frac{\varphi(\Phi^{-1}(\alpha))}{\alpha} - a_j \Delta m_j \leqslant d + m_0.$$

Let us consider the minimum of the left-hand side of this expression:

$$\rho = \min_a \left\{ |a_j|\,\sigma_j \frac{\varphi\left(\Phi^{-1}\left(\alpha\right)\right)}{\alpha} - a_j \Delta m_j \right\}.$$

Show that

$$\rho = -\infty.$$

Indeed, consider the case $a_j \geqslant 0$. The left-hand side of the CVaR constraint is rewritten as follows:

$$a_j \sigma_j \frac{\varphi\left(\Phi^{-1}\left(\alpha\right)\right)}{\alpha} - a_j \Delta m_j = a_j \left(\sigma_j \frac{\varphi\left(\Phi^{-1}\left(\alpha\right)\right)}{\alpha} - \Delta m_j\right).$$

Since by assumption $\sigma_j \frac{\varphi(\Phi^{-1}(\alpha))}{\alpha} - \Delta m_j < 0$, we have

$$\min_{a_j} \left\{ a_j \left(\sigma_j \frac{\varphi\left(\Phi^{-1}\left(\alpha\right)\right)}{\alpha} - \Delta m_j\right)\right\} = -\infty.$$

Summarizing the above reasoning, we obtain: if $\sigma_j \frac{\varphi(\Phi^{-1}(\alpha))}{\alpha} - \Delta m_j < 0$, then $IntD \neq \emptyset$. Statement 1 is proved.

Note that from the proof of Statement 1 it obviously follows (see (3)) a sufficient condition for the non-emptiness of an admissible set D, $\sigma_j \frac{\varphi(\Phi^{-1}(\alpha))}{\alpha} - \Delta m_j \leqslant 0$.

Statement 2. *If condition* (3) *is satisfied, then the admissible portfolio* a^* *is optimal in problem* (2) *if and only if there exists a vector* $(\lambda_1, \ldots, \lambda_n, \mu) \in K^*$ *such that*

$$\Delta m \left(1 + \frac{\mu}{\frac{\varphi(\Phi^{-1}(\alpha))}{\alpha}}\right) + \lambda = 0$$

and

$$a^* \lambda' + \frac{\mu(\langle a^*, \Delta m\rangle + m_0 + d)}{\frac{\varphi(\Phi^{-1}(\alpha))}{\alpha}} = 0, \text{ where } \lambda = (\lambda_1, \ldots, \lambda_n).$$

This statement is an obvious consequence of the Kuhn–Tucker theorem [15] for the concave cone programming problem (2). Note that without the concavity condition of the objective function in (2), which is now linear, the statement gives only necessary optimality conditions.

3. MULTI-STAGE OPTIMAL PORTFOLIO CONSTRUCTION PROBLEM

In a multi-stage problem, the investment horizon is divided into T parts or time moments $0, 1, \ldots, T$. The random vector of asset returns for time moment t is denoted as $R^t = (R_0^t, \ldots, R_n^t)$, where $R_0^t = m_0^t$ a.s. is the return of the risk-free asset. The vectors R^t are assumed to be independent, as in the works on investment optimization in multi-stage models with bankruptcy [7, 12]. We denote the mathematical expectation of the return on the ith asset by $m_i^t = ER_i^t$. The covariance matrices of risky asset returns C_t are positive definite at each stage t.

The term "investor bankruptcy" means the following. If the investor's current capital X_t falls below the threshold $r = 0$, then the investor cannot make any transactions with assets (buy, sell, borrow) and the value of the capital X_t is fixed from the moment t until the last time moment T. After applying the normal approximation to the percentage increase in capital over the interval $[t, t+1]$, we obtain a normally distributed random variable $Y_a^t := \sum_1^n a_i(R_i^t - m_0^t) + m_0^t$ with parameters $\mu^t(a) = \langle a, \Delta m^t\rangle + m_0^t$ and $\sigma^t(a) = \sqrt{aC_t a'}$. Then, the stage-by-stage CVaR constraints (see Section 2) have the form

$$-X_t\langle a, \Delta m^t\rangle + X_t m_0^t + X_t\sqrt{aC_t a^T}\frac{\varphi\left(\Phi^{-1}\left(\alpha^t\right)\right)}{\alpha^t} \leqslant d^t X_t \quad \text{for all} \quad X_t = x > 0.$$

Here $\Delta m^t = (m_1^t - m_0^t, \ldots, m_n^t - m_0^t)$, $d^t > 0$ is the share in percentage of the current capital, limiting the risk measure CVaR; $\alpha^t \in (0, 1)$ is the significance level, $t = 0, \ldots, T-1$. We denote the investment strategy by $A = (a^0, \ldots, a^{T-1})$, recall that short sales are allowed.

Then, the equation of capital dynamics for the chosen investment strategy is

$$X_{t+1} = \begin{cases} X_t\left[\sum\limits_{i=1}^n a_i^t\left(R_i^t - m_0^t\right) + m_0^t\right], & \text{when } X_t > 0, \\ X_t, & \text{when } X_t \leqslant 0, \\ t = 0, \ldots, T-1; X_0 = x_0, & \text{when } x_0 > 0. \end{cases} \tag{4}$$

It is assumed that the investor's goal is to maximize the average value of the final capital. Thus, the problem under consideration is the problem of optimal control of a Markov chain in the presence of a set of absorbing states $\{x : x \leqslant 0\}$:

$$\begin{cases} J[A] \equiv E(X_T) \to \max, \quad A \in \mathbf{A} \text{ under constraints (4) and} \\ -(\langle a, \Delta m^t\rangle + m_0^t) + \sqrt{aC_t a^T}\frac{\varphi\left(\Phi^{-1}\left(\alpha^t\right)\right)}{\alpha^t} \leqslant d^t, \end{cases} \tag{5}$$

where $\mathbf{A}$ is the set of all investment strategies that are predictable in the sense of natural filtering.

Define the Bellman functions (value functions) as $V_t(x) = \max_A EX_T$ for the controlled process on the interval $[t,T]$ with the initial state $X_t = x$. Then $V_T(x) = x$;

$$\begin{aligned}
V_{T-1}(x) &= \max_{a\in D_{T-1}} xEY_a^{T-1} = \max_{a\in D_{T-1}} xG_{T-1}(a) = xG_{T-1}\left(a_*^{T-1}\right), \text{ if } x>0,\\
V_{T-1}(x) &= x, \text{ if } x \leqslant 0;\\
V_{T-2}(x) &= \max_{a\in D_{T-2}} x\left\{E\left[G_{T-1}\left(a_*^{T-1}\right)Y_a^{T-2}|Y_a^{T-2}>0\right]\right.\\
&\quad \left.\times P\left(Y_a^{T-2}>0\right) + E\left[Y_a^{T-2}|Y_a^{T-2}\leqslant 0\right]P\left(Y_a^{T-2}\leqslant 0\right)\right\}\\
&= \max_{a\in D_{T-2}} xG_{T-2}(a)\\
&= xG_{T-2}\left(a_*^{T-2}\right) \text{ when } x>0 \text{ and } V_{T-2}(x) = x \text{ when } x\leqslant 0;\\
&\dots\\
V_0(x) &= \max_{a\in D_0} x\left\{E\left[G_1\left(a_*^1\right)Y_a^0|Y_a^0>0\right]P\left(Y_a^0>0\right) + E\left[Y_a^0|Y_a^0\leqslant 0\right]P\left(Y_a^0\leqslant 0\right)\right\}\\
&= \max_{a\in D_0} xG_0(a) = xG_0\left(a_*^0\right), \text{ where } x>0.
\end{aligned}$$

Here D_t, the set of admissible portfolios at stage t, is (see (5))

$$D_t = \left\{a \in R^n : \sqrt{aC_ta'} \leqslant \frac{\langle a, \Delta m^t\rangle + m_0^t + d^t}{\frac{\varphi(\Phi^{-1}(\alpha^t))}{\alpha^t}}\right\}, \tag{6}$$

where, recall, $\Delta m^t = (m_1^t - m_0^t, \dots, m_n^t - m_0^t)$, $t = 0,\dots,T-1$, random variables Y_a^t are defined above. Note that D_t does not depend on the current state $x>0$ of the process X_t.

The functions $G_t(a)$ introduced above are defined by the recurrence formula

$$\begin{aligned}
G_t(a) &= E\left[Y_a^t|Y_a^t>0\right]P\left(Y_a^t>0\right)G_{t+1}\left(a_*^{t+1}\right)\\
&\quad + E\left[Y_a^t|Y_a^t\leqslant 0\right]P\left(Y_a^t\leqslant 0\right), \; t = 0,\dots,T-1,\\
G_T(a) &\equiv 1,
\end{aligned} \tag{7}$$

where a_*^{t+1} is the portfolio that maximizes $G_{t+1}(a)$ on the set D_{t+1}.

From the given expressions for the Bellman functions it follows that each portfolio a_*^t in the optimal strategy $A_* = (a_*^0,\dots,a_*^{T-1})$ depends only on the moment t of decision making and does not depend on the value of the current state $x>0$ of the process X_t.

The following theorem gives an explicit expression for the functions $G_t(a)$ and provides the necessary optimality conditions in the optimal control problem (5).

Theorem 1. *Let the condition* (3) *be satisfied for all* α^t, σ_j^t, Δm_j^t.

If the investment strategy $(a_*^0,\dots,a_*^{T-1})$ *is optimal, then there exists a vector*

$$(\lambda_1,\dots,\lambda_n,\mu) \in K^* = \left\{(\lambda,\mu) \in R^{n+1} : \sqrt{\lambda C_t^{-1}\lambda'} \leqslant \mu\right\},$$

such that:

$$\nabla G_t\left(a_*^t\right) + \frac{\Delta m^t \mu}{\frac{\varphi(\Phi^{-1}(\alpha^t))}{\alpha^t}} + \lambda = 0 \tag{8}$$

and

$$a_*^t \lambda' + \frac{\mu\left(\langle a_*^t, \Delta m^t\rangle + m_0^t + d^t\right)}{\frac{\varphi(\Phi^{-1}(\alpha^t))}{\alpha^t}} = 0, \tag{9}$$

where $\nabla G_t\,(a)$ *denote gradient* $G_t\,(a)$,

$$\begin{aligned} G_t\,(a) = &\left\{\overline{\Phi}\left[-\frac{\mu^t\,(a)}{\sigma^t\,(a)}\right]\mu^t\,(a) + \sigma^t\,(a)\,\varphi\left[-\frac{\mu^t\,(a)}{\sigma^t\,(a)}\right]\right\} \\ &\times G_{t+1}\left(a_*^{t+1}\right) + \Phi\left[-\frac{\mu^t\,(a)}{\sigma^t\,(a)}\right]\mu^t\,(a) - \sigma^t\,(a)\,\varphi\left[-\frac{\mu^t\,(a)}{\sigma^t\,(a)}\right], \\ G_T(a) \equiv 1.& \end{aligned} \tag{10}$$

Here $\overline{\Phi}\,(\cdot) = 1 - \Phi\,(\cdot)$, $\Phi(\cdot)$ *and* $\varphi(\cdot)$ *denote the distribution function and density of the standard normal random variable, respectively,* $\mu^t\,(a) = \langle a, \Delta m^t\rangle + m_0^t$ *and* $\sigma^t(a) = \sqrt{aC_t a'}$.

Proof. The expression for $G_t\,(a)$ (see the recurrence formula (7)) is easily transformed into (10) taking into account the expressions for the mathematical expectation of the normal random variable Y_a^t, truncated to the intervals $(0, \infty)$ and $(-\infty, 0]$ (see, for example, [16]). Since the denominators in (10) contain the standard deviation $\sigma^t\,(a) = \sqrt{aC_t a'}$, it is necessary to first show that the degenerate investment portfolio $a^d = (0, \ldots, 0)$ cannot be optimal in the problem

$$\max_{a \in D_t} G_t\,(a)\,. \tag{11}$$

Indeed, consider the portfolio $a_\gamma = \gamma \Delta m^t$, which is admissible for sufficiently small $\gamma > 0$ (see (6)). Next,

$$\mu^t\,(a_\gamma) = \gamma \|\Delta m^t\|^2 + m_0^t, \quad \sigma^t\,(a_\gamma) = \gamma\sqrt{\Delta m_t C_t \Delta m_t'} \text{ and } \Phi\left(\frac{-\mu^t\,(a_\gamma)}{\mu^t\,(a_\gamma)}\right) \to 0,$$

$$\varphi\left(\frac{-\mu^t\,(a_\gamma)}{\sigma^t\,(a_\gamma)}\right) \to 0 \text{ when } \gamma \to 0 + 0.$$

Then

$$G_t\,(a_\gamma) = G_t\left(a_*^{t+1}\right)\left(\gamma\|\Delta m_t\|^2 + m_0^t\right) + o(\gamma) > G_t\left(a^d\right) = G_t\left(a_*^{t+1}\right) m_0^t$$

for sufficiently small values of $\gamma > 0$. Thus, to solve the problem (11) it is sufficient to limit ourselves to the set of admissible portfolios $D_t \setminus a^d$.

Condition (3) means that $IntD_t \neq \emptyset$, i.e., Slater's condition in problem (11) is satisfied. Applying Statement 2 to this cone programming problem, where the objective function $G_t\,(a)$ is not, in general, concave, we obtain (8), (9) as necessary conditions for optimality in (11). Theorem 1 is proved.

4. NON-NORMAL RISK ASSET RETURN MODEL

Let the returns of n risky assets have a multivariate elliptical distribution (normal distribution, Laplace distribution, Bessel distribution, etc. [17]), which allows, in particular, to take into account the "heavy tails" in the return distributions. A convenient property of this class of distributions for risk analysis is that a linear combination of elliptically distributed assets again has an elliptical distribution. Let $F(x)$ denote the distribution function of a "standard" elliptical random variable with zero mean and unit variance, and let $f(x)$ denote its density.

Then (see Section 2) the risk measure CVaR has the form

$$\mathrm{CVaR}_\alpha\left(X_a\right) = -EX_a + \sqrt{DX_a}k^*,$$

where now

$$k^* = \frac{-\int\limits_{-\infty}^{-z_\alpha^*} xf(x)dx}{\alpha} \quad \text{and} \quad z_\alpha^* : F(-z_\alpha^*) = \alpha.$$

Thus, the results obtained above remain valid in this case with $k = \phi\left(\Phi^{-1}\left(\alpha\right)\right)/\alpha$ replaced by k^*. Truncated normal distributions are replaced by corresponding truncated elliptical distributions.

5. EXAMPLE[1]

Let us illustrate the results obtained in Section 3 by solving a numerical example of finding optimal portfolios for a market of three companies: Apple, Microsoft, Facebook. Data on the stock price for the period 05/07/2022–05/07/2023 (one year) were taken in open access from the Nasdaq exchange website (https://www.nasdaq.com/market-activity/stocks/). The obtained realizations of the returns of these assets approximately follow normal distributions (but in this work, a strict test of hypotheses about the normality of the distributions of returns by methods of mathematical statistics remains out of consideration), therefore the assumption about the normal distribution of the investor's total capital at each stage seems justified. Based on these empirical data, estimates of the vector of mathematical expectations $m = (m_0^t, \ldots, m_3^t)$ and the covariance matrix $C = C_t$ of the three risky assets are constructed. It is assumed that the vectors of mathematical expectations and covariance matrices do not depend on the stage number over the entire time interval $[0, T] = [0, 4]$, and the risk-free asset has a yield of $m_0 = 1$, i.e., investments in this asset are preserved and do not bring losses/profits. The initial capital of the investor $x_0 = 1$; the significance level $\alpha = 0.05$; the constant d, which limits the risk measure CVaR defined above, is a variable parameter.

According to the results of Section 3 (see Theorem 1), the optimal portfolio at stage t is defined as the solution to the problem

$$\max_{D_t} G_t\left(a\right), \quad t = 3, 2, 1, 0,$$

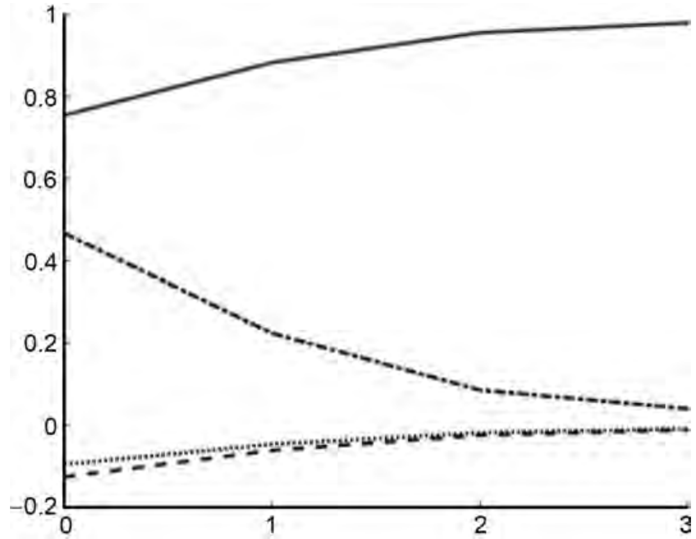
where the recurrence expressions for $G_t\left(a\right)$ are defined in (10). The table below shows how the optimal portfolio changes as the constant d, which constrains CVaR, increases.

Optimal investment portfolios in a multi-stage problem with a change in the constant d, which limits the risk measure CVaR at each stage

t	$t=0$	$t=1$	$t=2$	$t=3$
d	0.2	0.25	0.28	0.29
Apple, $a_*^t(1)$	0.7536	0.8820	0.9548	0.9786
Microsoft, $a_*^t(2)$	−0.0950	−0.0457	−0.0175	−0.0082
Facebook, $a_*^t(3)$	−0.1253	−0.0607	−0.0233	−0.0110
Risk-free asset, $a_*^t(0)$	0.4667	0.2244	0.0860	0.0406

With the increase of the constant d at each stage, the investor becomes more inclined to risk, so the share of investment in a risk-free asset is expected to decrease from stage to stage, and the share of investment in risky assets increases. For illustration, the obtained optimal investment portfolios are shown in figure.

[1] The data for this example and the solution itself were obtained with the help of Silventoinen D.P.



Optimal investment portfolios at stages $\{0, 1, 2, 3\}$. The solid line denotes the share of investments in Apple shares, the share of investments in the risk-free asset — dash-dotted line, the share of investments in Microsoft — dotted line, the share of investments in Facebook — dashed line.

The optimal value of the objective functional $J[A_*] = E\,X_4$ is equal to 1.031, which means obtaining 3% of the average profit for the entire investment period. This low value is explained by the fact that the observed annual returns on risky asset stocks are close to one and have low volatility — the differences in the empirical returns of each risky asset are only fractions of a percent. Note that the solved example is only an illustration of Theorem 1, and the methodology for calculating the optimal investment strategy given in it would give different results in the case of risky assets other than the three considered above.

6. CONCLUSION

The paper formulates and solves a multi-stage investment problem with probabilistic (CVaR) constraints and the possibility of bankruptcy. It is assumed that short selling is allowed and that the distribution of total returns at each stage is normal (or elliptical). It is shown that the optimal investment portfolio at each stage does not depend on the value of the investor's capital, but depends only on the number of the investment stage. In the context of finding optimal investment portfolios, it is shown that a multi-stage problem is reduced to a finite number of one-stage optimization problems that are recursively connected and are conical programming problems. For the auxiliary one-stage problem, sufficient conditions for the non-emptiness of the set of admissible portfolios and the fulfillment of the Slater regularity conditions are presented and proven. In solving the one stage problem, a modified Kuhn–Tucker theorem is applied for the case of generalized inequalities. The result of solving a multi-stage problem can be considered a method for constructing optimal investment portfolios, applicable in practice with known estimates of the vectors of mathematical expectations and matrices of covariances of returns. A numerical example based on one year of stock exchange price data for three companies is also presented to illustrate the theoretical results. The following directions of development of this work seem interesting: using other distributions instead of the normal or elliptical model for returns, for example, gamma distributions; using another risk measure instead of CVaR, for example, shortfall probability [3]. Another direction could be the analysis of a similar problem, but without of short sales.

FUNDING

This work is being carried out as part of the topic No. FFSM-2024-0003.

REFERENCES

1. Markowitz, H., Portfolio Selection, *J. Finance*, 1952, vol. 7, pp. 77–91.

2. Gaivoronski, A. and Pflug, G., Value at Risk in Portfolio Optimization: Properties and Computational Approach, *J. Risk*, 2004, vol. 7, no. 2, pp. 1–31.

3. Golubin, A.Y. and Gridin, V.N., Design of Efficient Investment Portfolios with a Shortfall Probability as a Measure of Risk, *Autom. Remote Control*, 2023, vol. 84, no. 4, pp. 486–495.

4. Kibzun, A.I. and Kan, Yu.S., *Zadachi stokhasticheskogo programmirovaniya s veroyatnostnymi kriteriyami* (Stochastic programming problems with probabilistic criteria), Moscow: Fizmatlit, 2009.

5. Alexander, G.J. and Baptista, A.M., Economic Implications of Using a Mean-VaR Model for Portfolio Selection: A Comparison with Mean-variance Analysis, *J. Econom. Dynam. Control*, 2002, vol. 26, pp. 1159–1193.

6. Alexander, G.J. and Baptista, A.M., A comparison of VaR and CVaR constraints on portfolio selection with the mean-variance model, *Management Sci.*, 2004, vol. 50, pp. 1261–1273.

7. Shiba, N., Xu, C., and Wang, J., Multistage Portfolio Optimization with VaR as Risk Measure, *Int. J. Innovat. Comput., Inform. Control*, 2007, vol. 3, no. 3, pp. 709–724.

8. Tyrrell, R., Rockafellar, R.T., and Uryasev, S., Conditional value-at-risk for general loss distributions, *J. Banking Finance*, 2002, vol. 26, no. 7, pp. 1443–1471.

9. Pinar, M.C., Static and Dynamic VaR Constrained Portfolios with Application to Delegated Portfolio Management, *Optimization*, 2013, vol. 62, no. 11, pp. 1419–1432.

10. Gardoni, P. and Murphy, C., Gauging the societal impacts of natural disasters using a capabilities-based approach, *Disasters: Disaster Studies, Policy, Management*, 2010, vol. 34, no. 3, pp. 619–636.

11. Golubin, A.Y., Optimal Investment Policy in a Multi-stage Problem with Bankruptcy and Stage-by-stage Probability Constraints, *Optimization*, 2022, vol. 70, no. 10, pp. 2963–2977.

12. Li, C. and Li, Z., Multi-period portfolio optimization for asset–liability management with bankrupt control, *Apppl. Math. Comput.*, 2012, vol. 218, pp. 11196–11208.

13. Wei, S. and Ye, Z., Multi-period optimization portfolio with bankruptcy control in stochastic market, *Apppl. Math. Comput.*, 2007, vol. 186, pp. 414–425.

14. Golubin, A.Y. and Gazov, A.I., Usloviya optimal'nosti v zadache vybora investitsionnogo portfelya pri veroyatnostnom ogranichenii na kapital investora, *Informatsionnye tekhnologii v proektirovanii i proizvodstve*, 2018, no. 4, pp. 53–57.

15. Boyd, S. and Vandenberghe, L., *Convex Optimization*, Cambridge: University Press, 2009.

16. Burkardt, J., The Truncated Normal Distribution, *Department of Scientific Computing*, Florida State University, 2014.

17. Landsman, Z. and Valdez, E.A., Tail Conditional Expectations for Elliptical Distributions, *North Amer. Actuarial J.*, 2003, vol. 7, no. 4, pp. 55–71.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 12, pp. 1209–1211.

The 11th International Conference on Physics and Control

A. L. Fradkov[*,**]

[*]*Institute for Problems of Mechanical Engineering, Saint Petersburg, Russia*
[**]*Saint Petersburg University, Saint Petersburg, Russia*
e-mail: fradkov@mail.ru

Abstract—The scientific and organizational issues of the 11th International Conference on Physics and Control (PhysCon 2024) held on Sept. 9–12, 2024 in Istanbul, Turkey are considered. A brief survey of keynote and plenary talks is presented.

Keywords: International Conference, Physics and Control

DOI: 10.31857/S0005117924120067

The 11th International Conference on Physics and Control (PhysCon 2024) was held from September 9 to 12 at Kadir Has University in Istanbul. The event was organized by Turkish universities Kadir Has and Bilkent with the support of the International Physics and Control Society (IPACS). Professor Fatihan Atay from Bilkent University served as the chairperson of the organizing committee. Key decisions regarding the organization were made by the supervisory board led by Eckehard Schöll, President of the IPACS and a professor at Technical University Berlin, which included the author of this report. PhysCon 2024 was dedicated to the memory of Hermann Haken (1927–2024), founder of synergetics and an inspiring pioneer in innovative work on nonequilibrium phase transitions, who was an honorary member of IPACS.

A total of 87 participants from 15 countries attended the conference, presenting 72 papers including three plenary keynotes, four invited talks, and two lectures. The distribution of presentations by country is as follows: Russian Federation – 24, Algeria – 12, Turkey – 10, Germany and Spain – six each, England – four, USA, Brazil, and Australia – two each, China, Denmark, Mexico, Uzbekistan, Iran, Italy – one each. In addition to these, five mini-symposia were also part of the program, each consisting of 3–4 thematically related presentations.

The conference covered interdisciplinary topics at the intersection of mathematics, physics, biology, electronics, and computer science. A significant portion of the presentations focused on the interface between control theory, applied mathematics, and neurobiology, particularly studying models of brain processes and their control. This emerging field can be termed cybernetical neuroscience.

The opening plenary talk by Henrik Jensen (Imperial College London) titled "Self-organized Criticality and Control" provided a brief overview of Self-Organized Criticality (SOC) followed by new insights into one of its paradigmatic models, the Forest Fire Model (FFM). The relationship between observed power-law behavior and true criticality has been debated since the model's introduction in 1992. Recent analysis shows that it is possible to establish critical scale invariance in the model if the coupling between driving force and system size is carefully managed.

Mark Timme from Dresden Technical University presented his research on fluctuation responses and tipping points in strongly perturbed nonlinear systems. He proposed an integral consistency condition and a method for predicting the tipping point using large perturbation expansions evaluated within the consistency condition framework. This novel approach could help quantitatively predict significantly nonlinear response dynamics and bifurcations arising under high-amplitude forcing in non-autonomous dynamical systems.

Michael Small from the University of Western Australia delivered a talk titled "Delay Embedding Choice and Why It Matters," building upon Takens' theorem, which guarantees accurate embedding of a deterministic nonlinear dynamic system given a time series under rather general conditions. Since the 1980s, many methods have been suggested to estimate the observation interval (delay) needed to ensure accurate embedding, but most are based on heuristic approaches. Michael introduced a new topologically grounded method for choosing delay, leveraging concepts from persistent homology and topological data analysis, ensuring the best attractor reconstruction for given data.

Eckehard Schöll's lecture "Nonequilibrium Phase Transitions and Nucleation Phenomena in Synchronizing Networks" addressed phase transitions in nonlinear dynamical systems far from thermodynamic equilibrium. The lecture was dedicated to the memory of Hermann Haken. Although concepts from thermodynamics and statistical physics had been used to describe self-organization, formation of spatio-temporal structures, coexistence of phases, critical phenomena, and first-order and second-order nonequilibrium phase transitions since the 1970s, phase transitions and critical phenomena in dynamic networks, where synchronization transitions may occur, leading to partially synchronized states and complex collective behaviors applicable to various natural, socio-economic, and technological systems, began to be studied much later. The lecture discussed these works and established connections between tipping transitions, explosive synchronization, nucleation, critical slowing down, etc., with nonequilibrium thermodynamics. In particular, the Kuramoto model with inertia, relevant to power grids, was examined, showing first-order phase transitions to synchronization through partially synchronized states, and it was demonstrated that it can be viewed as an adaptive network of phase oscillators similar to neural networks with plasticity.

Alexander Fradkov's (IPME RAS) lecture (see "Definition of Cybernetical Neuroscience" by Alexander Fradkov, https://arxiv.org/abs/2409.16314) introduced a new scientific area — cybernetical neuroscience, a branch of computational neuroscience aimed at studying neurobiological systems using cybernetic methods. Cybernetical neuroscience is based on mathematical models adopted in computational neuroscience (Hodgkin-Huxley model, FitzHugh-Nagumo model, Morris-Lecar model, Hindmarsh-Rose model, Landau-Stuart model, neural mass model, etc.) and the methods of cybernetics – the science of control and communication in living organisms, machines, and society. The lecture outlined the main problems, methods, and some results of cybernetical neuroscience obtained primarily at IPME RAS and St. Petersburg State University, including findings on neurointerface control ("brain-controlled machines"). The primary objectives of cybernetical neuroscience include:

1. Analyzing the conditions under which neuronal ensemble models exhibit specific regimes corresponding to real neuronal ensembles' behavior: synchronization, spiking, bursting, solitons, chaos, chimeras, etc.

2. Synthesizing external (control) actions that create these regimes in the models.

3. Estimating the state and parameters of models based on input and output variable measurements.

4. Classifying human brain states and future behaviors based on observations using adaptation and machine learning techniques.

5. Finding control algorithms (feedback synthesis) that ensure specified properties of closed-loop systems composed of interacting controlled systems and controlling agents.

In neurobiological studies, the controlled system is the nervous system or human brain, while the controlling agent can be implemented in a computer device. For the entire system to function, the nervous system or brain must be connected to external computer communication devices called neurointerfaces (brain-computer interfaces). The methodology of cybernetic neuroscience shares many similarities with that of cybernetical physics. The lecture provided several examples from cybernetical neuroscience.

Overall, PhysCon 2024 showcased cutting-edge research at the crossroads of multiple disciplines, highlighting the importance of understanding complex systems, self-organization, and nonlinear dynamics, all crucial for advancing modern technology and science.

A brief history of the conference is as follows. The first Physics and Control Conference took place in 2003 in Saint Petersburg. The second conference was held in 2005, again in Saint Petersburg. The Program Committee Co-Chairs of the first conference were the renowned physicist and Nobel laureate Zhores Ivanovich Alferov, as well as Vladimir Grigoryevich Peshekhonov, General Director of the Central Research Institute "Elektropribor," creator of gyroscopic equipment and control systems. Organization of those conferences was inspired by the emergence in the 1990s of a new scientific field at the intersection of physics and control theory. This area encompassed the control of oscillations, chaotic processes, quantum systems, and more. Later, it came to be known as cybernetic physics. Cybernetic physics explores issues of dynamics and control in complex dynamical systems arising in physics and other natural sciences: synchronization processes, nonlinear waves, chaos, solitons, quantum-mechanical processes, and others. All these topics remain highly relevant today. Subsequent conferences were held every two years in different cities and countries. In 2007, it was held in Potsdam, Germany; in 2009, in Catania, Sicily; in 2011, in León, Spain; in 2013, in San Luis Potosí, Mexico; in 2015, in Istanbul, Turkey; and in 2017, in Florence, Italy. In 2019, the conference took place in Russia, at the Innopolis University, and in 2021, due to the pandemic, it was conducted online by Fudan University in Shanghai, China. Selecting the venue and organizing team for the next conference took an extra year due to complicated political circumstances. However, the outcomes showed that the choice was correct: it allowed a substantial number of Russian specialists to maintain creative and personal ties with foreign colleagues. One successful organizational decision was allowing participants from sanctioned countries to pay the registration fee in cash upon arrival. The conference results also indicated that in the scientific field, which first emerged in our country in 2003, Russia remains among the leading nations in 2024.

The programme and the abstracts of PhysCon 2024 conference are posted on its website, see https://physcon2024.khas.edu.tr/. The proceedings will be published in the special issues of the journals: Cybernetics And Physics (http://cap.physcon.ru) and The European Physical Journal Special Topics (https://link.springer.com/journal/11734).