Volume 85, Number 8,
August 2024

ISSN 0005-1179
CODEN: AURCAT

A JOURNAL OF RUSSIAN ACADEMY OF SCIENCES

AUTOMATION AND REMOTE CONTROL

Automation and Remote Control

Vol. 85, No. 8, August 2024

Editor-in-Chief
Andrey A. Galyaev

http://ait.mtas.ru

Automation and Remote Control

ISSN 0005-1179

SCOPE

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

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

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

0005-1179/24. *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. 8, 2024

Nonlinear Systems

Stochastic Systems

Robust, Adaptive, and Network Control

Control in Social Economic Systems

Optimization, System Analysis, and Operations Research

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 8, pp. 747–760.

NONLINEAR SYSTEMS

Interval Observers for Estimating Unknown Inputs in Discrete Time-Invariant Dynamic Systems

A. N. Zhirabok[*,**,a] **and A. V. Zuev**[*,**,b]

[*] *Far Eastern Federal University, Vladivostok, Russia*
[**] *Institute of Marine Technology Problems, Far Eastern Branch, Russian Academy of Sciences, Vladivostok, Russia*
e-mail: [a] *zhirabok@mail.ru*, [b] *alvzuev@yandex.ru*

Received December 16, 2024
Revised May 4, 2024
Accepted May 30, 2024

Abstract—This paper considers discrete time-invariant systems (with constant parameters) described by dynamic models in the presence of exogenous disturbances. For such systems, the problem of estimating unknown inputs using interval observers is studied. The solution is based on a minimal-dimension disturbance-insensitive model of the original system. For this model, an interval observer is designed and then used to estimate the unknown inputs. The theoretical results are illustrated with a practical example.

Keywords: linear systems, unknown inputs, interval observers, estimation

DOI: 10.31857/S0005117924080016

1. INTRODUCTION AND PROBLEM STATEMENT

The problem of estimating unknown inputs of dynamic systems has theoretical and applied significance, particularly for determining the changes in system parameters due to arising faults to parry these changes or the level of exogenous disturbances by a control system. For systems described by continuous-time models, in the past twenty-five years, this problem has been solved using sliding mode observers [1–3].

On the other hand, interval observers have been applied in recent years to estimate the components of the state vector of dynamic systems. At each time instant, such observers produce an estimate of the set of admissible values of the state vector (or a given linear function of this vector) for various classes of dynamic systems with uncertainty. Detailed reviews of the contemporary results were provided in [4–6]; solutions for various classes of systems (continuous, discrete, hybrid, and delayed), as well as practical applications, can be found in [7–14]. A peculiarity is that interval observers involve simple means to consider various types of uncertainty in systems (exogenous disturbances, measurement noises, and parametric uncertainties) and generate an interval surely containing the values of the state vector (or a given linear function of this vector).

Interval observers were proposed in [15] to estimate unknown inputs in discrete-time linear dynamic systems. This method is an alternative to sliding mode observers, which become ineffective in the discrete case.

In this paper, the approach introduced in [15] is further developed to estimate unknown inputs in discrete-time linear and nonlinear dynamic systems subjected to exogenous disturbances. In contrast to [15], the system is not required to be minimum-phase; moreover, the approach proposed therein involves several complex transformations of the system, in particular, reduction to a descriptor form, which is not required below. By analogy with [5–13], the interval observer in [15] was designed for the original transformed system and therefore has full dimension. In contrast, the solution presented below uses a reduced model of the original system, simplifying the observer design procedure and making the resulting interval observer insensitive to exogenous disturbances. In some cases, this property ensures an exact estimate even in the presence of exogenous disturbances. In addition, we overcome the matching condition requirement imposed in [15], which restricts estimation capabilities.

2. MAIN MODELS AND RELATIONS

Consider a system described by the discrete-time nonlinear model

$$\begin{aligned} x(t+1) &= Fx(t) + Gu(t) + C\Psi(x(t), u(t)) + Dd(t) + L\rho(t), \\ y(t) &= Hx(t), \end{aligned} \tag{2.1}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and $y(t) \in \mathbb{R}^l$ are the state vector, control input, and output, respectively; F, G, H, C, and L are known constant matrices; $\rho(t) \in \mathbb{R}^p$ is an unknown bounded time-varying function that describes disturbances affecting the system; the term $Dd(t)$ is an unknown input to be estimated. In particular, such an input may manifest a system fault. By assumption, $d(t)$ is a scalar bounded function, i.e., $\underline{d} \leqslant d(t) \leqslant \overline{d}$ for all $t = 0, 1, 2, \dots$ with given values $\underline{d}$ and $\overline{d}$. The nonlinear component $\Psi(x, u)$ is represented as

$$\Psi(x, u) = \begin{pmatrix} \varphi_1(A_1 x, u) \\ \cdots \\ \varphi_q(A_q x, u) \end{pmatrix},$$

where $A_1, \dots, A_q$ are known constant row matrices, and $\varphi_1, \dots, \varphi_q$ are arbitrary nonlinear functions.

Remark 1. By analogy with [15], the matching condition $\operatorname{rank}(HD) = \operatorname{rank}(D)$ is supposed to hold in the main part of this work, except for Section 7. This means that the unknown input $d(t)$ is included in the equations for the state vector components measured by system sensors. In contrast to [15], where $d(t)$ was an arbitrary-dimension function, we first assume that $d(t)$ is a scalar and then eliminate this constraint in Section 8.

The problem under consideration is first solved in the linear case, when $C = 0$. As stated, a solution will be found based on a reduced model of the original system. For this purpose, let us briefly recall the results of [14, 16]. In the cited papers, a linear minimal-dimension disturbance-insensitive model of system (2.1) was built to estimate some variable $y_*(t) = R_* y(t)$ for the matrix R_* to be determined during the solution procedure. Such a model is described by the equations

$$\begin{aligned} x_*(t+1) &= F_* x_*(t) + J_* y(t) + G_* u(t) + D_* d(t), \\ y_*(t) &= H_* x_*(t) = R_* y(t), \end{aligned} \tag{2.2}$$

where $x_*(t) \in \mathbb{R}^k$, $k < n$ denotes the model dimension and the matrices F_*, G_*, J_*, and H_* have to be determined. By assumption, there exists a matrix Φ such that $x_*(t) = \Phi x(t)$. According to [14, 16], the model matrices satisfy the equations

$$\Phi F = F_* \Phi + J_* H, \quad H_* \Phi = R_* H, \quad \Phi G = G_*, \quad \Phi D = D_*, \quad \Phi L = 0. \tag{2.3}$$

A necessary condition for building such a system is the inequality

$$\operatorname{rank} \begin{pmatrix} H \\ L_0 \end{pmatrix} < \operatorname{rank}(H) + \operatorname{rank}(L_0), \tag{2.4}$$

where L_0 is a matrix of maximal rank such that $L_0 L = 0$. Indeed, the condition $\Phi L = 0$ can be written as $\Phi = N L_0$ for some matrix N. Since $H_* \Phi = R_* H$, we have $H_* N L_0 = R_* H$, holding under condition (2.4). If the latter fails, then the disturbance-insensitive system estimating the variable $y_*(t)$ cannot be built. Below we will suppose the validity of condition (2.4).

3. BUILDING THE DISTURBANCE-INSENSITIVE MODEL

In the general case, the matrices F_* and H_* are given in the identification canonical form:

$$F_* = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}, \quad H_* = (1 \;\; 0 \;\; \dots \;\; 0). \tag{3.1}$$

If system (2.2) is observable, it can always be reduced to this form [17]. Otherwise, it can be reduced to a form with an observable subsystem; then the matrices of this subsystem can be found in the form (3.1) [17]. The problem is solved based on the equation

$$(\; R_* \;\; -J_{*1} \;\; \dots \;\; -J_{*k} \;)(W^{(k)} \;\; L^{(k)}) = 0, \tag{3.2}$$

where J_{*i} denotes the ith row of the matrix J_*,

$$W^{(k)} = \begin{pmatrix} HF^k \\ HF^{k-1} \\ \dots \\ H \end{pmatrix}, \quad L^{(k)} = \begin{pmatrix} HL & HFL & \dots & HF^{k-1}L \\ 0 & HL & \dots & HF^{k-2}L \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{pmatrix}, \quad k = 1, 2, \dots.$$

The matrix $W^{(k)}$ serves to build model (2.2) whereas the matrix $L^{(k)}$ ensures its insensitivity to the disturbances, i.e., the condition $\Phi L = 0$. Equation (3.2) has a nontrivial solution if

$$\operatorname{rank}(W^{(k)} \;\; L^{(k)}) < l(k+1) \tag{3.3}$$

because, in this case, the rows of the compound matrix $(W^{(k)} \;\; L^{(k)})$ are linearly dependent, which ensures the existence of a solution.

To build the model, it is necessary to determine a minimum number k from (3.3), starting from $k = 1$, and the row $(\; R_* \;\; -J_{*1} \;\; \dots \;\; -J_{*k} \;)$ from (3.2); next, using the relations

$$R_* H = \Phi_1, \quad \Phi_i F = \Phi_{i+1} + J_{*i} H, \quad i = 1, \dots, k-1, \quad \Phi_k F = J_{*k} H,$$

obtained from (2.3) and (3.1), the matrix Φ is constructed; finally, G_* is found from (2.3), where Φ_i denotes the ith row of the matrix Φ. Thus, the linear model (2.2) is built; by assumption, $D_* \neq 0$.

If equation (3.2) has no solution for all $k < n$, then the disturbance-insensitive model does not exist. In this case, robust methods can be used; for example, see [18]. However, the resulting estimate will be inexact.

4. PROBLEM SOLUTION

For the sake of simplicity, we begin with the assumption $k = 1$. In other words, it is possible to build a one-dimensional model described by the equations

$$\begin{aligned} x_*(t+1) &= J_* y(t) + G_* u(t) + D_* d(t), \\ y_*(t) &= x_*(t). \end{aligned} \tag{4.1}$$

(Under this assumption, $F_* = 0$ and $H_* = 1$.)

Proposition 1. *Model* (4.1) *exists under two conditions:*

$$\operatorname{rank}\begin{pmatrix} HF \\ H \end{pmatrix} < \operatorname{rank}(HF) + \operatorname{rank}(H),$$
$$R_* HL = 0.$$

Proof. For $k = 1$, from (3.2) it follows that

$$\begin{pmatrix} R_* & -J_* \end{pmatrix}\begin{pmatrix} HF & HL \\ H & 0 \end{pmatrix} = 0,$$

which implies the equality $\begin{pmatrix} R_* & -J_* \end{pmatrix}\begin{pmatrix} HF \\ H \end{pmatrix} = 0$. It holds under the first condition. The second condition is immediate from the previous equality.

If at least one of these conditions fails, the model will be multidimensional. An additional condition is the fault sensitivity requirement $\Phi D = D_* \neq 0$, represented as $R_* HD \neq 0$.

Remark 2. Since the values of all variables in model (4.1), except for $d(t)$, are measurable, the latter can be determined from the simple relation $D_* d(t) = y_*(t+1) - J_* y(t) - G_* u(t)$. Considering the canonical form of the matrix F_* and the resulting model, similar relations for the variable $d(t)$ can be derived by introducing time shifts for the variable $y_*(t)$ and substituting some equations into others (as is done in Section 7). For $k > 2$, these relations will be quite cumbersome, and using interval observers allows significantly simplifying them. Therefore, despite the obvious solution for $k = 1$, we will illustrate the application of such observers in order to extend the result to the general case.

An interval observer is given by

$$\begin{aligned} \underline{x}_*(t+1) &= G_* u(t) + J_* y(t) + D_*^+ \underline{d} - D_*^- \overline{d} + g\underline{e}(t), \\ \overline{x}_*(t+1) &= G_* u(t) + J_* y(t) + D_*^+ \overline{d} - D_*^- \underline{d} + g\overline{e}(t), \\ \underline{y}_*(t) &= \underline{x}_*(t), \quad \overline{y}_*(t) = \overline{x}_*(t), \end{aligned} \tag{4.2}$$

where $A^+ = \max(0, A)$ and $A^- = A^+ - A$ for an arbitrary matrix A (clearly, $A^+ \geqslant 0$ and $A^- \geqslant 0$); the coefficient g is assigned to ensure the observer's stability; finally, $\underline{e}(t) = y_*(t) - \underline{y}_*(t)$ and $\overline{e}(t) = y_*(t) - \overline{y}_*(t)$.

For the interval observer, the condition $\underline{x}_*(0) \leqslant x_*(0) \leqslant \overline{x}_*(0)$ implies $\underline{x}_*(t) \leqslant x_*(t) \leqslant \overline{x}_*(t)$ for all $t \geqslant 0$ [4]; hence, there exists a number $\alpha(t) \geqslant 0$ such that, for all $t \geqslant 0$,

$$x_*(t) = \underline{x}_*(t) + \alpha(t)(\overline{x}_*(t) - \underline{x}_*(t)), \tag{4.3}$$

or

$$x_*(t) = \alpha(t)\overline{x}_*(t) + (1 - \alpha(t))\underline{x}_*(t). \tag{4.4}$$

Let us determine the value $\alpha(t+1)$. For $t := t+1$, from (4.3) it follows that

$$R_* y(t+1) = x_*(t+1) = \underline{x}_*(t+1) + \alpha(t+1)(\overline{x}_*(t+1) - \underline{x}_*(t+1)),$$

yielding

$$\alpha(t+1) = \frac{R_* y(t+1) - \underline{x}_*(t+1)}{\overline{x}_*(t+1) - \underline{x}_*(t+1)}. \tag{4.5}$$

Note that all values in this formula are measurable.

Theorem 1. *The function $d(t)$ is estimated by*

$$\begin{aligned}\hat{d}(t) = D_*^{-1}(g(\alpha(t+1)\underline{e}(t) + (1-\alpha(t+1))\overline{e}(t))\\ + \alpha(t+1)(D_*^{+}\underline{d} - D_*^{-}\overline{d}) + (1-\alpha(t+1))(D_*^{+}\overline{d} - D_*^{-}\underline{d})),\end{aligned} \tag{4.6}$$

where the coefficient $\alpha(t+1)$ is given by (4.5).

Proof. In the expression (4.4) written for $t := t+1$, we replace the variables $x_*(t+1)$, $\underline{x}_*(t+1)$ and $\overline{x}_*(t+1)$ with the right-hand sides of equations (4.1) and (4.2), respectively:

$$\begin{aligned}J_*y(t) + G_*u(t) + D_*d(t) = \alpha(t+1)(G_*u(t) + J_*y(t) + D_*^{+}\underline{d} - D_*^{-}\overline{d} + g\underline{e}(t))\\ + (1-\alpha(t+1))(G_*u(t) + J_*y(t) + D_*^{+}\overline{d} - D_*^{-}\underline{d} + g\overline{e}(t)).\end{aligned} \tag{4.7}$$

Standard transformations on the right-hand side lead to

$$\begin{aligned}J_*y(t) + G_*u(t) + D_*d(t) = g(\alpha(t+1)\underline{e}(t) + (1-\alpha(t+1))\overline{e}(t)) + G_*u(t) + J_*y(t)\\ + \alpha(t+1)(D_*^{+}\underline{d} - D_*^{-}\overline{d}) + (1-\alpha(t+1))(D_*^{+}\overline{d} - D_*^{-}\underline{d}),\end{aligned}$$

which finally gives (4.6).

Remark 3. Due to the canonical form of the matrix F_*, the observer (4.2) is stable even without introducing the feedbacks $g\underline{e}(t)$ and $g\overline{e}(t)$; the latter are necessary only to obtain the estimate $\hat{d}(t)$, and (in this case) it will be exact for all $t > 0$.

5. THE MULTIDIMENSIONAL MODEL

In the case $k > 1$, model (2.2) takes the form

$$\begin{aligned}x_{*1}(t+1) &= x_{*2}(t) + J_{*1}y(t) + G_{*1}u(t) + D_{*1}d(t),\\ x_{**}(t+1) &= F_{**}x_{**}(t) + J_{**}y(t) + G_{**}u(t) + D_{**}d(t),\\ y_*(t) &= x_{*1}(t),\end{aligned}$$

where x_{*i} denotes the ith component of the vector x_*, $x_{**} = (x_{*2}, \ldots, x_{*k})$; J_{**}, G_{**}, and D_{**} are the matrices J_*, G_*, and D_*, respectively, with the removed first rows; F_{**} is the matrix (3.1) of dimensions $(k-1) \times (k-1)$. By assumption, $D_{*1} \neq 0$ and $D_{**} = 0$, i.e., the fault figures only in the first equation. The interval observer has the form

$$\begin{aligned}\underline{x}_{*1}(t+1) &= \hat{x}_{*2}(t) + J_{*1}y(t) + G_{*1}u(t) + D_{*1}^{+}\underline{d} - D_{*1}^{-}\overline{d} + g_1\underline{e}(t),\\ \hat{x}_{**}(t+1) &= F_{**}\hat{x}_{**}(t) + J_{**}y(t) + G_{**}u(t) + g_{**}\underline{e}(t),\\ \overline{x}_{*1}(t+1) &= \tilde{x}_{*2}(t) + J_{*1}y(t) + G_{*1}u(t) + D_{*1}^{+}\overline{d} - D_{*1}^{-}\underline{d} + g_1\overline{e}(t),\\ \tilde{x}_{**}(t+1) &= F_{**}\tilde{x}_{**}(t) + J_{**}y(t) + G_{**}u(t) + g_{**}\overline{e}(t),\\ \underline{y}_*(t) &= \underline{x}_{*1}(t), \quad \overline{y}_*(t) = \overline{x}_{*1}(t),\end{aligned} \tag{5.1}$$

i.e., only the first component of the vector is described by an interval. Here $\hat{x}_{**}(t)$ and $\tilde{x}_{**}(t)$ are vectors of dimension $(k-1)$, the matrix $(g_1\ g_{**}^T)^T$ is assigned for ensuring the observer's stability. (This can always be done for the canonical form (3.1).)

In view of (5.1), the expression (4.7) is modified as follows:

$$\begin{aligned}x_{*2}(t) + J_{*1}y(t) + G_{*1}u(t) + D_{*1}d(t)\\ = \alpha(t+1)(\tilde{x}_{*2}(t) + G_{*1}u(t) + J_{*1}y(t) + D_{*1}^{+}\underline{d} - D_{*1}^{-}\overline{d} + g_1\underline{e}(t))\\ + (1-\alpha(t+1))(\hat{x}_{*2}(t) + G_{*1}u(t) + J_{*1}y(t) + D_{*1}^{+}\overline{d} - D_{*1}^{-}\underline{d} + g_1\overline{e}(t)).\end{aligned}$$

Here $\alpha(t+1)$ is determined for the first components of the state vectors by analogy with (4.5):

$$\alpha(t+1) = \frac{R_* y(t+1) - \underline{x}_{*1}(t+1)}{\overline{x}_{*1}(t+1) - \underline{x}_{*1}(t+1)}.$$

Standard transformations yield

$$\begin{aligned} D_{*1}d(t) &= g_1(\alpha(t+1)\underline{e}(t) + (1-\alpha(t+1))\overline{e}(t)) \\ &+ \alpha(t+1)(D_{*1}^{+}\underline{d} - D_{*1}^{-}\overline{d}) + (1-\alpha(t+1))(D_{*1}^{+}\overline{d} - D_{*1}^{-}\underline{d}) \\ &+ \alpha(t+1)\tilde{x}_{*2}(t) + (1-\alpha(t+1))\hat{x}_{*2}(t) - x_{*2}(t). \end{aligned} \tag{5.2}$$

Since the observer is stable, we have $\tilde{x}_{*2}(t) \to x_{*2}(t)$ and $\hat{x}_{*2}(t) \to x_{*2}(t)$ and, consequently, $\alpha(t+1)\tilde{x}_{*2}(t) + (1-\alpha(t+1))\hat{x}_{*2}(t) - x_{*2}(t) \to 0$. As a result, the estimate $\hat{d}(t)$ given by (4.6) with $D_* = D_{*1}$ satisfies $\hat{d}(t) \to d(t)$.

Remark 4. According to these relations, the use of an interval observer eliminates the effect of all components of the vector $x_{**} = (x_{*2}, \ldots, x_{*k})$ on the final result.

6. CONSIDERATION OF THE NONLINEARITIES

The model design method in the nonlinear case is based on the linear model built previously. The idea is to use the matrix Φ found for the linear model and analyze the possibility of transforming the argument of the nonlinear component $\Psi(x,u)$. This is done as follows; for details, see [18].

We calculate the matrix $C_* = \Phi C$ and determine the numbers $j_i, \ldots, j_s$ of its nonzero columns. Next, it is necessary to verify the condition

$$\operatorname{rank}\begin{pmatrix}\Phi \\ H\end{pmatrix} = \operatorname{rank}\begin{pmatrix}\Phi \\ H \\ A'\end{pmatrix}, \tag{6.1}$$

where the matrix A' consists of the matrices $A_1, \ldots, A_q$ with numbers $j_i, \ldots, j_s$ as rows. If this condition holds, the argument transformation is possible. After that, we build the nonlinear component

$$\Psi_*(x_*, y, u) = \begin{pmatrix} \varphi_{j_1}(A_{*j_1}z, u) \\ \cdots \\ \varphi_{j_s}(A_{*j_s}z, u) \end{pmatrix},$$

where $z = \begin{pmatrix} x_* \\ y \end{pmatrix}$ and the row matrices $A_{*j_1}, \ldots, A_{*j_s}$ are found from the linear equations

$$A_j = A_{*,j}\begin{pmatrix}\Phi \\ H\end{pmatrix}, \quad j = j_i, \ldots, j_s.$$

This component is added to the linear model (2.2). If condition (6.1) fails, it is necessary to find another solution of equation (3.2) given the same or increased dimension k and repeat the described procedure with the new matrix Φ. In the general case, the nonlinear model is described by the equation

$$\begin{aligned} &\dot{x}_*(t) = F_* x_*(t) + J_* y(t) + G_* u(t) + C_* \Psi_*(x_*(t), y(t), u(t)) + D_* d(t), \\ &y_*(t) = H_* x_*(t). \end{aligned}$$

If the model is one-dimensional, its nonlinear component will depend on $y_*(t) = x_*(t)$, and the interval observer will appear in the model in the same way. As is easily observed, in this case, the exact estimate will be given by (4.6).

In view of the assumption $D_{*1} \neq 0$ and $D_{**} = 0$, we obtain the following descriptions in the multidimensional case:

—for the model,

$$\begin{aligned} x_{*1}(t+1) &= x_{*2}(t) + J_{*1}y(t) + G_{*1}u(t) + C_{*1}\Psi_{*1}(y_*(t), x_{**}(t), y(t), u(t)) \\ &\quad + D_{*1}d(t), \\ x_{**}(t+1) &= F_{**}x_{**}(t) + J_{**}y(t) + G_{**}u(t) + C_{**}\Psi_{**}(y_*(t), x_{**}(t), y(t), u(t)); \end{aligned}$$

—for the interval observer,

$$\begin{aligned} \underline{x}_{*1}(t+1) &= \hat{x}_{*2}(t) + J_{*1}y(t) + G_{*1}u(t) + C_{*1}\Psi_{*1}(y_*(t), \hat{x}_{**}(t), y(t), u(t)) \\ &\quad + D_{*1}^{+}\underline{d} - D_{*1}^{-}\overline{d} + g_1\underline{e}(t), \\ \hat{x}_{**}(t+1) &= F_{**}\hat{x}_{**}(t) + J_{**}y(t) + G_{**}u(t) + C_{**}\Psi_{**}(y_*(t), \hat{x}_{**}(t), y(t), u(t)) \\ &\quad + g_{**}\underline{e}(t), \\ \overline{x}_{*1}(t+1) &= \tilde{x}_{*2}(t) + J_{*1}y(t) + G_{*1}u(t) + C_{*1}\Psi_{*1}(y_*(t), \tilde{x}_{**}(t), y(t), u(t)) \\ &\quad + D_{*1}^{+}\overline{d} - D_{*1}^{-}\underline{d} + g_1\overline{e}(t), \\ \tilde{x}_{**}(t+1) &= F_{**}\tilde{x}_{**}(t) + J_{**}y(t) + G_{**}u(t) + C_{**}\Psi_{**}(y_*(t), \tilde{x}_{**}(t), y(t), u(t)) \\ &\quad + g_{**}\overline{e}(t), \\ \underline{y}_*(t) &= \underline{x}_{*1}(t), \quad \overline{y}_*(t) = \overline{x}_{*1}(t), \end{aligned}$$

where the matrices C_{*1} and C_{**} and the functions Ψ_{*1} and Ψ_{**} are determined by analogy with J_{*1} and J_{**}.

Considering the nonlinearities, the expression (5.2) is modified as follows:

$$\begin{aligned} D_{*1}d(t) &= g_1(\alpha(t+1)\underline{e}(t) + (1-\alpha(t+1))\overline{e}(t)) \\ &\quad + \alpha(t+1)(D_{*1}^{+}\underline{d} - D_{*1}^{-}\overline{d}) + (1-\alpha(t+1))(D_{*1}^{+}\overline{d} - D_{*1}^{-}\underline{d}) \\ &\quad + \alpha(t+1)\tilde{x}_{*2}(t) + (1-\alpha(t+1))\hat{x}_{*2}(t) - x_{*2}(t) \\ &\quad + \alpha(t+1)C_{*1}\Psi_{*1}(y_*(t), \tilde{x}_{**}(t), y(t), u(t)) \\ &\quad + (1-\alpha(t+1))C_{*1}\Psi_{*1}(y_*(t), \hat{x}_{**}(t), y(t), u(t)) \\ &\quad - C_{*1}\Psi_{*1}(y_*(t), x_{**}(t), y(t), u(t)). \end{aligned}$$

Since the gain matrix $(g_1\ g_{**}^T)^T$ is appropriately assigned to ensure the observer's stability, we have $\tilde{x}_{**}(t) \to x_{**}(t)$ and $\hat{x}_{**}(t) \to x_{**}(t)$ and, by analogy with Section 5, $\hat{d}(t) \to d(t)$.

7. NONFULFILLMENT OF THE MATCHING CONDITION

The matching condition may often fail; the proposed approach can be still applied in this case, albeit in a more complex way. For the sake of simplicity, we consider a particular case where $k = 2$ and the term $D_*d(t)$ is included in the second equation of the linear model:

$$\begin{aligned} x_{*1}(t+1) &= x_{*2}(t) + J_{*1}y(t) + G_{*1}u(t), \\ x_{*2}(t+1) &= J_{*2}y(t) + G_{*2}u(t) + D_*d(t), \\ y_*(t) &= x_{*1}(t). \end{aligned}$$

The interval observer takes the form

$$
\begin{aligned}
\underline{x}_{*1}(t+1) &= \underline{x}_{*2}(t) + J_{*1}y(t) + G_{*1}u(t) + g_1\underline{e}(t),\\
\underline{x}_{*2}(t+1) &= J_{*2}y(t) + G_{*2}u(t) + g_2\underline{e}(t) + D_*^+\underline{d} - D_*^-\overline{d},\\
\bar{x}_{*1}(t+1) &= \bar{x}_{*2}(t) + J_{*1}y(t) + G_{*1}u(t) + g_1\overline{e}(t),\\
\bar{x}_{*2}(t+1) &= J_{*2}y(t) + G_{*2}u(t) + g_2\overline{e}(t) + D_*^+\overline{d} - D_*^-\underline{d},\\
\underline{y}_*(t) &= \underline{x}_{*1}(t), \quad \overline{y}_*(t) = \overline{x}_{*1}(t),
\end{aligned}
$$

i.e., both components of the state vector are described by intervals; the variables $\underline{e}(t)$ and $\bar{e}(t)$ are determined as described above.

We find the first components of the model and the observer at the time instant $(t+2)$ by introducing time shifts in the first equations and replacing the variables $x_{*2}(t+1)$, $\underline{x}_{*2}(t+1)$, and $\overline{x}_{*2}(t+1)$ with the right-hand sides of their equations:

$$
\begin{aligned}
x_{*1}(t+2) &= J_{*1}y(t+1) + G_{*1}u(t+1) + J_{*2}y(t) + G_{*2}u(t) + D_*d(t),\\
\underline{x}_{*1}(t+2) &= J_{*1}y(t+1) + G_{*1}u(t+1) + J_{*2}y(t) + G_{*2}u(t)\\
&\quad + g_1\underline{e}(t+1) + g_2\underline{e}(t) + D_*^+\underline{d} - D_*^-\overline{d},\\
\bar{x}_{*1}(t+2) &= J_{*1}y(t+1) + G_{*1}u(t+1) + J_{*2}y(t) + G_{*2}u(t)\\
&\quad + g_1\overline{e}(t+1) + g_2\overline{e}(t) + D_*^+\overline{d} - D_*^-\underline{d}.
\end{aligned}
\tag{7.1}
$$

By analogy with (4.5), let us define

$$
\alpha(t+2) = \frac{R_*y(t+2) - \underline{x}_{*1}(t+2)}{\overline{x}_{*1}(t+2) - \underline{x}_{*1}(t+2)},
\tag{7.2}
$$

form the sum

$$
x_{*1}(t+2) = \alpha(t+2)\overline{x}_{*1}(t+2) + (1-\alpha(t+2))\underline{x}_{*1}(t+2),
$$

and replace the variables $x_{*1}(t+2)$, $\overline{x}_{*1}(t+2)$, and $\underline{x}_{*1}(t+2)$ with the right-hand sides of equations (7.1):

$$
\begin{aligned}
&J_{*1}y(t+1) + G_{*1}u(t+1) + J_{*2}y(t) + G_{*2}u(t) + D_*d(t)\\
&= \alpha(t+2)(J_{*1}y(t+1) + G_{*1}u(t+1) + J_{*2}y(t) + G_{*2}u(t)\\
&\quad + g_1\overline{e}(t+1) + g_2\overline{e}(t) + D_*^+\overline{d} - D_*^-\underline{d})\\
&+ (1-\alpha(t+2))(J_{*1}y(t+1) + G_{*1}u(t+1) + J_{*2}y(t) + G_{*2}u(t)\\
&\quad + g_1\underline{e}(t+1) + g_2\underline{e}(t) + D_*^+\underline{d} - D_*^-\overline{d}).
\end{aligned}
$$

Standard transformations yield

$$
\begin{aligned}
D_*d(t) &= g_1(\alpha(t+2)\overline{e}(t+1) + (1-\alpha(t+2))\underline{e}(t+1))\\
&\quad + g_2(\alpha(t+2)\overline{e}(t) + (1-\alpha(t+2))\underline{e}(t))\\
&+ \alpha(t+2)(D_*^+\overline{d} - D_*^-\underline{d}) + (1-\alpha(t+2))(D_*^+\underline{d} - D_*^-\overline{d}),
\end{aligned}
\tag{7.3}
$$

and we finally arrive at the desired formula for $\hat{d}(t)$.

Remark 5. The expression (7.3) is obviously generalized to the case $k > 2$ where the term $D_*d(t)$ appears only in the last equation of the model and, considering Sections 5 and 6, to the case of its presence only in one arbitrary equation and the nonlinear case. The proposed approach is difficult to implement and simulate: it requires additional time delays for the variables $\underline{e}(t)$ and $\bar{e}(t)$ in order to obtain the expressions (7.3).

8. THE SOLUTION BASED ON THE JORDAN CANONICAL FORM

If for all solutions of equation (3.2) the variable $d(t)$ figures in two or more model components, the identification canonical form should be replaced by the Jordan diagonal form, where

$$F_* = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \\ 0 & 0 & 0 & \dots & \lambda_k \end{pmatrix},$$

$\lambda_1, \dots, \lambda_k$ denote the eigenvalues ($|\lambda_i| < 1$), which are supposed different. This form can always be obtained from the canonical form (3.1) [16]. As a result, the first equation in (2.3) turns into k independent equations:

$$\Phi_i F = \lambda_i \Phi_i + J_{*i} H, \quad i = 1, \dots, k, \tag{8.1}$$

and it becomes possible to build a one-dimensional model with the variable $d(t)$. The additional requirement $\Phi L = 0$ (insensitivity to the disturbances) is considered as follows. In view of the matrix L_0 introduced above, equation (8.1) can be written as

$$(N_i \quad -J_{*i}) \begin{pmatrix} L_0(F - \lambda_i I_n) \\ H \end{pmatrix} = 0, \quad i = 1, \dots, k, \tag{8.2}$$

where I_n means an identity matrix of dimensions $(n \times n)$.

To build the model, we select an eigenvalue $\lambda = \lambda_i$ so that the matrix $\Phi = N L_0$ given by (8.2) satisfies the requirements $D_* = \Phi D \neq 0$ and $\Phi = R_* H$ for some matrix R_*.

Remark 6. The requirements $\Phi D \neq 0$ and $\Phi = R_* H$ do not necessarily hold since the exogenous disturbances represented by the matrix L_0 may restrict the set of solutions of equation (8.2) even under the matching condition. This restriction can be removed, but the resulting solution will be inexact.

The model takes the form

$$\begin{aligned} x_*(t+1) &= \lambda x_*(t) + J_* y(t) + G_* u(t) + D_* d(t), \\ y_*(t) &= x_*(t). \end{aligned} \tag{8.3}$$

Despite the one-dimensionality of the model, the technique shown in Remark 2 is inapplicable here: the right-hand side of equation (8.3) includes the variable $x_*(t)$. Therefore, it is necessary to design an interval observer described by the equations

$$\begin{aligned} \underline{x}_*(t+1) &= \lambda \underline{x}_*(t) + G_* u(t) + J_* y(t) + D_*^+ \underline{d} - D_*^- \overline{d} + g \underline{e}(t), \\ \overline{x}_*(t+1) &= \lambda \bar{x}_*(t) + G_* u(t) + J_* y(t) + D_*^+ \overline{d} - D_*^- \underline{d} + g \overline{e}(t), \\ \underline{y}_*(t) &= \underline{x}_*(t), \quad \overline{y}_*(t) = \overline{x}_*(t). \end{aligned} \tag{8.4}$$

Note that, regardless of the value of λ, the coefficient g can always be appropriately assigned to make the observer stable.

Theorem 2. *The function $d(t)$ is estimated by*

$$\begin{aligned} \hat{d}(t) = D_*^{-1} g(\alpha(t+1)\underline{e}(t) + (1 - \alpha(t+1))\overline{e}(t)) \\ + \alpha(t+1)(D_*^+ \underline{d} - D_*^- \overline{d}) + (1 - \alpha(t+1))(D_*^+ \overline{d} - D_*^- \underline{d}) \\ + \lambda(\overline{x}_*(t) - \underline{x}_*(t))(\alpha(t+1) - \alpha(t)). \end{aligned} \tag{8.5}$$

The variable $\alpha(t)$ is calculated by analogy with $\alpha(t+1)$, and the variables $\overline{x}_*(t)$ and $\underline{x}_*(t)$ are available for measurement.

Proof. The coefficient $\alpha(t+1)$ is given by (4.5), and the relation (4.7) becomes

$$\begin{aligned}\lambda x_*(t) + J_*y(t) + G_*u(t) + D_*d(t)\\ = \alpha(t+1)(\lambda\bar{x}_*(t) + G_*u(t) + J_*y(t) + D_*^+\underline{d} - D_*^-\overline{d} + g\underline{e}(t))\\ + (1-\alpha(t+1))(\lambda\underline{x}_*(t) + G_*u(t) + J_*y(t) + D_*^+\overline{d} - D_*^-\underline{d} + g\overline{e}(t)).\end{aligned}$$

The right-hand side of this equality is transformed as follows:

$$\begin{aligned}\lambda x_*(t) + J_*y(t) + G_*u(t) + D_*d(t) = g(\alpha(t+1)\underline{e}(t) + (1-\alpha(t+1))\overline{e}(t))\\ + \alpha(t+1)(D_*^+\underline{d} - D_*^-\overline{d}) + (1-\alpha(t+1))(D_*^+\overline{d} - D_*^-\underline{d})\\ + J_*y(t) + G_*u(t) + \alpha(t+1)\lambda\bar{x}_*(t) + (1-\alpha(t+1))\lambda\underline{x}_*(t).\end{aligned} \tag{8.6}$$

The variable $x_*(t)$ on the left-hand side of (8.6) can be represented as

$$x_*(t) = \alpha(t)\overline{x}_*(t) + (1-\alpha(t))\underline{x}_*(t).$$

By combining it with the two last terms in (8.6), we obtain

$$\lambda(\overline{x}_*(t) - \underline{x}_*(t))(\alpha(t+1) - \alpha(t)),$$

which finally yields (8.5).

Remark 7. The described approach can also be applied to the unknown vector input $d(t)$ if the matching condition holds. In this case, the solution of equation (8.2) is found separately for each component of this vector, and the observer (8.4) is built for each component of the vector as well. Perhaps, it will be impossible to build the disturbance-insensitive observer for some components.

9. A PRACTICAL EXAMPLE

Consider a discretized model of an electric drive of the form

$$\begin{aligned}x_1(t+1) &= k_1x_2(t) + x_1(t),\\ x_2(t+1) &= k_2x_2(t) + k_3x_3(t) + \rho(t),\\ x_3(t+1) &= k_4x_2(t) + k_5x_3(t) + k_6u(t) + d(t),\\ y_1(t) &= x_1(t), \quad y_2(t) = x_3(t),\end{aligned} \tag{9.1}$$

where x_1 is the rotation angle of the gearbox output shaft; x_2 is the angular velocity of the electric motor shaft, and x_3 is the electric motor current. The coefficients $k_1, \ldots, k_6$ depend on the drive parameters and the sampling interval; the disturbance $\rho(t)$ is caused by an external load moment reduced to the motor shaft; the fault represented by the function $d(t)$ is caused by a change in the active resistance of the armature circuit, which is expressed through a corresponding change in the value of the coefficient k_5 : if $\Delta k_5(t)$ is the variation of the coefficient k_5, then $d(t) = x_3(t)\Delta k_5(t)$. This model is described by equations (2.1) with the matrices

$$F = \begin{pmatrix} 1 & k_1 & 0\\ 0 & k_2 & k_3\\ 0 & k_4 & k_5\end{pmatrix},\ G = \begin{pmatrix} 0\\ 0\\ k_6\end{pmatrix},\ H = \begin{pmatrix} 1 & 0 & 0\\ 0 & 0 & 1\end{pmatrix},\ L = \begin{pmatrix} 0\\ 1\\ 0\end{pmatrix},\ D = \begin{pmatrix} 0\\ 0\\ 1\end{pmatrix}.$$

As is easily verified, the matching condition holds.

The disturbance-insensitive model is obtained by solving equation (3.2) with $k = 1$ and the matrices

$$R_* = \begin{pmatrix} -k_4 & k_1\end{pmatrix},\ J_* = \begin{pmatrix} -k_4 & k_1k_5\end{pmatrix},\ \Phi = \begin{pmatrix} -k_4 & 0 & k_1\end{pmatrix},\ G_* = k_1k_6,\ D_* = k_1.$$

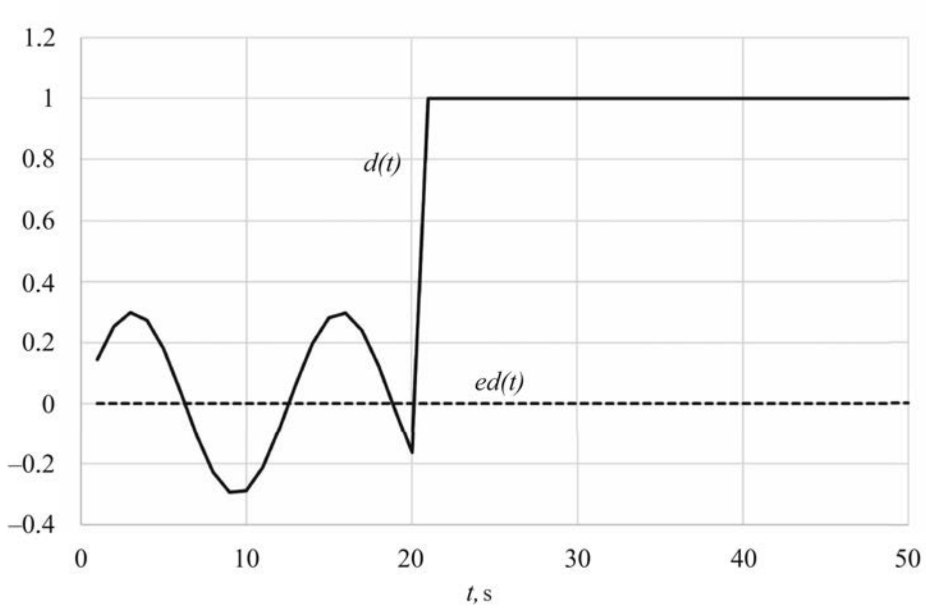

Fig. 1. The graphs of the functions $d(t)$ and $ed(t) = \hat{d}(t) - d(t)$.

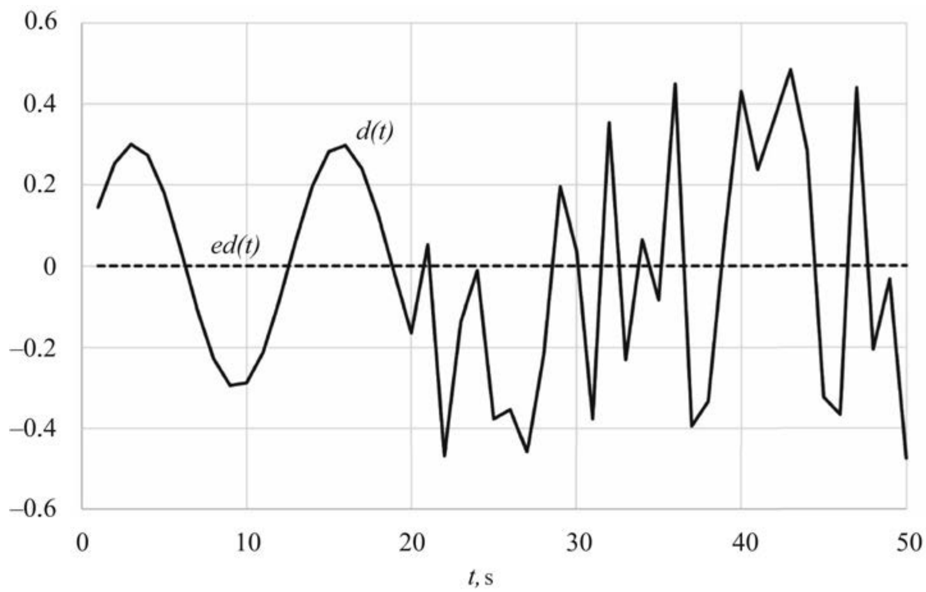

Fig. 2. The graphs of the functions $d(t)$ and $ed(t) = \hat{d}(t) - d(t)$.

The required model has the form

$$\begin{aligned} x_*(t+1) &= -k_4 y_1(t) + k_1 k_5 y_2(t) + k_1 k_6 u(t) + k_1 d(t), \\ y_*(t) &= x_*(t) = -k_4 y_1(t) + k_1 y_2(t). \end{aligned}$$

The interval observer is described by

$$\begin{aligned} \underline{x}_*(t+1) &= -k_1 y_1(t) + k_1 k_5 y_2(t) + k_1 k_6 u(t) + k_1 \underline{d} + g\underline{e}(t), \\ \overline{x}_*(t+1) &= -k_1 y_1(t) + k_1 k_5 y_2(t) + k_1 k_6 u(t) + k_1 \overline{d} + g\overline{e}(t), \\ \underline{y}_*(t) &= \underline{x}_*(t), \quad \overline{y}_*(t) = \overline{x}_*(t), \end{aligned} \tag{9.2}$$

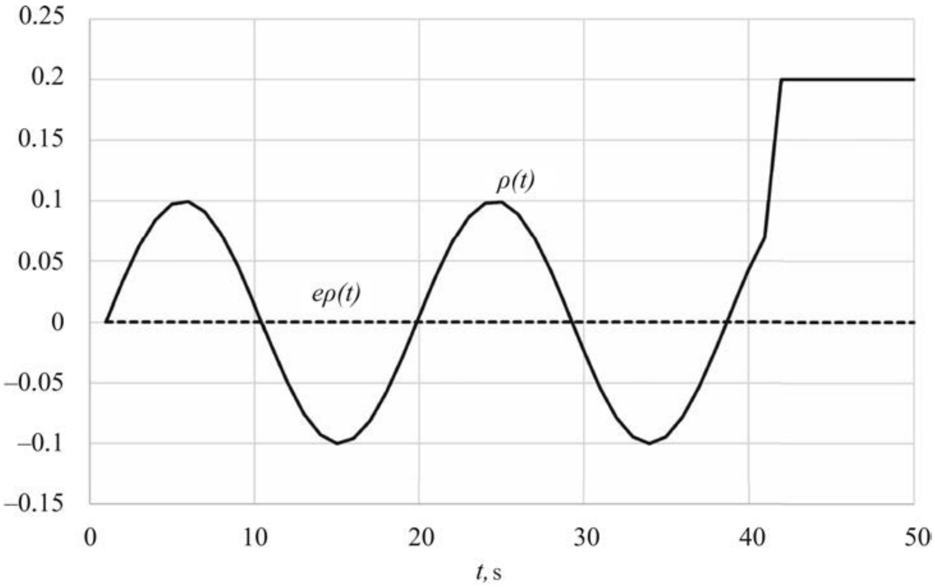

Fig. 3. The graphs of the functions $\rho(t)$ and $e\rho(t) = \hat{\rho}(t) - \rho(t)$.

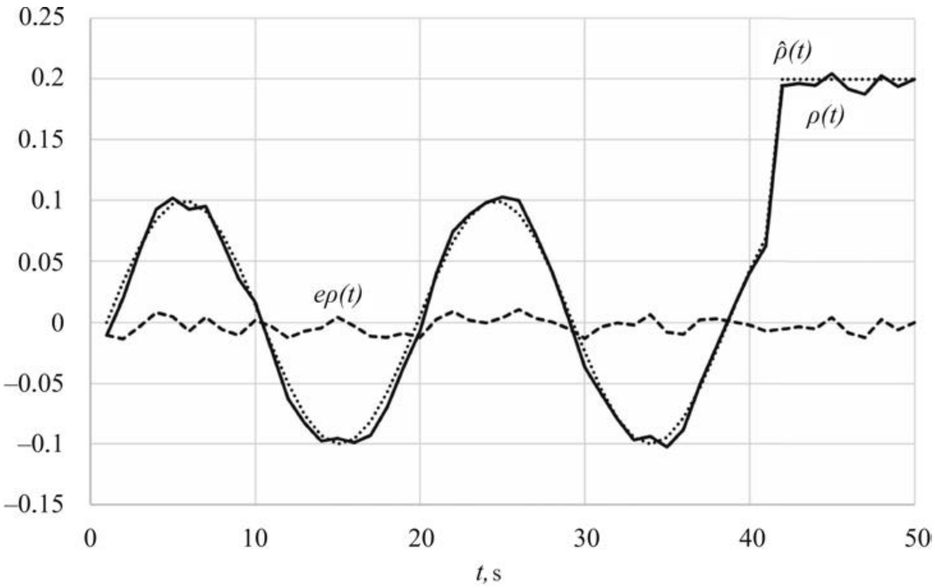

Fig. 4. The graphs of the functions $\rho(t)$, $\hat{\rho}(t)$, and $e\rho(t) = \hat{\rho}(t) - \rho(t)$.

where $\underline{e}(t) = y_*(t) - \underline{y}_*(t)$ and $\overline{e}(t) = y_*(t) - \overline{y}_*(t)$. The estimate of the variable $d(t)$ is given by (4.6).

For Matlab simulations, we choose system (9.1) and the observer (9.2) with the control input $u(t) = 0.2\sin(t/10)$, the random disturbance $\rho(t)$ with the uniform distribution on the interval $[-0.05; 0.05]$, $\underline{d} = -0.4$, $\bar{d} = 0.4$, and $g = -0.7$. The simulation results with the initial conditions $x(0) = (0\ \ 0\ \ 0)^T$, $\underline{x}_*(0) = -1$, and $\overline{x}_*(0) = 1$ are presented in Figs. 1 and 2. In particular, Fig. 1 shows the behavior of the function $d(t)$ and the estimation error $\hat{d}(t) - d(t)$ in the case $d(t) = 0.3\sin(t/2)$, $t \leqslant 20$, and $d(t) = 1$, $t > 20$; Fig. 2, the same graphs in the case $d(t) = 0.3\sin(t/2)$, $t \leqslant 20$, and the random function $d(t)$, $t > 20$, with the uniform distribution on the interval $[-0.5, 0.5]$. The estimation error is almost zero.

In addition, we estimate the unknown input $\rho(t)$ of system (9.1) under $d(t)=0$; by assumption, $|\rho(t)| \leqslant \rho_*$, i.e., $\underline{\rho} = -\rho_*$ and $\overline{\rho} = \rho_*$. As is easily checked, the matching condition fails here. The model solving the estimation problem takes dimension 2:

$$\begin{aligned} x_{*1}(t+1) &= x_{*2}(t) + Ky_2(t), \\ x_{*2}(t+1) &= y_1(t) + (k_1k_3 - Kk_5)y_2(t) + k_1\rho(t), \\ y_*(t) &= x_{*1}(t) = y_1(t), \end{aligned}$$

where $K = k_1(1+k_2)/k_4$. The interval observer has the form

$$\begin{aligned} \underline{x}_{*1}(t+1) &= \underline{x}_{*2}(t) + Ky_2(t) + g_1\underline{e}(t), \\ \underline{x}_{*2}(t+1) &= y_1(t) + (k_1k_3 - Kk_5)y_2(t) - k_1\rho_* + g_2\underline{e}(t), \\ \bar{x}_{*1}(t+1) &= \bar{x}_{*2}(t) + Ky_2(t) + g_1\overline{e}(t), \\ \bar{x}_{*2}(t+1) &= y_1(t) + (k_1k_3 - Kk_5)y_2(t) + k_1\rho_* + g_2\bar{e}(t), \end{aligned}$$

where $\underline{e}(t) = y_*(t) - \underline{x}_{*1}(t)$, $\overline{e}(t) = y_*(t) - \overline{x}_{*1}(t)$. The coefficient $\alpha(t+2)$ is given by (7.2) with $R_* = (1 \;\; 0)$. The unknown input $\rho(t)$ is estimated using the expression (7.3):

$$\begin{aligned} \hat{\rho}(t) = k_1^{-1}(&g_1(\alpha(t+2)\overline{e}(t+1) + (1-\alpha(t+2))\underline{e}(t+1)) \\ &+ g_2\overline{e}(t)(\alpha(t+2)\overline{e}(t) + (1-\alpha(t+2))\underline{e}(t)) \\ &+ \alpha(t+2)k_1\rho_* - (1-\alpha(t+2))k_1\rho_*). \end{aligned}$$

For Matlab simulations, we choose $\rho_* = 0.2$, $g_1 = -0.4$, and $g_2 = -0.1$. The simulation results are presented in Figs. 3 and 4, which show the behavior of the function $\rho(t)$ and the estimation error $\hat{\rho}(t) - \rho(t)$. In Fig. 3, the function $\rho(t)$ is defined by $\rho(t) = 0.3\sin(t/2)$, $t \leqslant 40$, and $\rho(t) = 1$, $t > 40$; as above, the estimation error is almost zero. In Fig. 4, additional measurement noises are introduced in the form of two independent random variables with the uniform distribution on the interval $[-0.01, 0.01]$. Clearly, in this case, the estimation quality gets worse.

10. CONCLUSIONS

This paper has considered discrete time-invariant systems described by linear and nonlinear dynamic models with exogenous disturbances under the so-called matching condition. Unknown inputs in such systems have been estimated based on a minimal-dimension disturbance-insensitive model of the original system. For this model, an interval observer has been designed; it produces an exact estimate of the unknown inputs in several cases. The initial solution obtained initially for the linear one-dimensional model has been extended to the multidimensional and nonlinear cases and to the systems not satisfying the matching condition. The theoretical results have been illustrated with a practical example.

FUNDING

This work was supported by the Russian Science Foundation, project no. 23-29-00191 (https://rscf.ru/project/23-29-00191/).

REFERENCES

1. Edwards, C., Spurgeon, S., and Patton, R., Sliding Mode Observers for Fault Detection and Isolation, *Automatica*, 2000, vol. 36, pp. 541–553.

2. Yan, X. and Edwards, C., Nonlinear Robust Fault Reconstruction and Estimation Using Sliding Mode Observers, *Automatica*, 2007, vol. 43, pp. 1605–1614.

3. Zhirabok, A., Zuev, A., Sergiyenko, O., and Shumsky, A., Identification of Faults in Nonlinear Dynamical Systems and Their Sensors Based on Sliding Mode Observers, *Autom. Remote Control*, 2022, vol. 83, no. 2, pp. 214–236.

4. Efimov, D. and Raissi, T., Design of Interval State Observers for Uncertain Dynamical Systems, *Autom. Remote Control*, 2016, vol. 77, no. 2, pp. 191–225.

5. Khan, A., Xie, W., Zhang, L., and Liu, L., Design and Applications of Interval Observers for Uncertain Dynamical Systems, *IET Circuits Devices Syst.*, 2020, vol. 14, pp. 721–740.

6. Khan, A., Xie, W., Bo Zhang, C., and Liu, L., A Survey of Interval Observers Design Methods and Implementation for Uncertain Systems, *J. Franklin Institute*, 2021, vol. 358, pp. 3077–3126.

7. Chebotarev, S., Efimov, D., Raissi, T., and Zolghadri, A., Interval Observers for Continuous-Time LPV Systems with L1/L2 Performance, *Automatica*, 2015, vol. 51, pp. 82–89.

8. Efimov, D., Polyakov, A., and Richard, J., Interval Observer Design for Estimation and Control of Time-Delay Descriptor Systems, *Eur. J. Control*, 2015, vol. 23, pp. 26–35.

9. Efimov, D., Fridman, L., Raissi, T., Zolghadri, A., and Seydou, R., Interval Estimation for LPV Systems Applying High Order Sliding Mode Techniques, *Automatica*, 2012, vol. 48, pp. 2365–2371.

10. Kremlev, A.S. and Chebotarev, S.G., Synthesis of Interval Observer for Linear System with Variable Parameters, *Journal of Instrument Engineering*, 2013, vol. 56, no. 4, pp. 42–46.

11. Mazenc, F. and Bernard, O., Asymptotically Stable Interval Observers for Planar Systems with Complex Poles, *IEEE Trans. Automatic Control*, 2010, vol. 55, no. 2, pp. 523–527.

12. Zheng, G., Efimov, D., and Perruquetti, W., Interval State Estimation for Uncertain Nonlinear Systems, *Proc. IFAC NOLCOS 2013*, Toulouse, 2013.

13. Zhang, K., Jiang, B., Yan, X., and Edwards, C., Interval Sliding Mode Based Fault Accommodation for Non-Minimal Phase LPV Systems with Online Control Application, *Int. J. Control*, 2019. https://doi.org/10.1080/00207179.2019.1687932.

14. Zhirabok, A.N., Zuev, A.V., and Ir, K.C., Method to Design Interval Observers for Linear Time-Invariant Systems, *J. Comput. Syst. Sci. Int.*, 2022, vol. 61, no. 4, pp. 485–495.

15. Zhu, F., Zhang, W., Zhang, J., and Guo, S., Unknown Input Reconstruction via Interval Observer and State and Unknown Input Compensation Feedback Controller Designs, *Int. J. Control, Autom. Syst.*, 2020, vol. 18, pp. 1–13.

16. Zhirabok, A., Zuev, A., Filaretov, V., Shumsky, A., and Kim, C., Interval Observers for Continuous-Time Systems with Parametric Uncertainties, *Autom. Remote Control*, 2023, vol. 84, no. 11, pp. 1275–1286.

17. Kwakernaak, H. and Sivan, R., *Linear Optimal Control Systems*, Wiley, 1972.

18. Zhirabok, A., Shumsky, A., Solyanik, S., and Suvorov, A., Design of Nonlinear Robust Diagnostic Observers, *Autom. Remote Control*, 2017, vol. 78, no. 9, pp. 1572–1584.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 8, pp. 761–772.

Partial Stability in Probability of Nonlinear Stochastic Discrete-Time Systems with Delay

V. I. Vorotnikov[*,a] **and Yu. G. Martyshenko**[**,b]

[*] *Sochi Institute, Peoples' Friendship University (RUDN), Sochi, Russia*
[**] *Gubkin Russian State University of Oil and Gas, Moscow, Russia*
e-mail: [a] *vorotnikov-vi@rambler.ru,* [b] *j-mart@mail.ru*

Received September 20, 2023
Revised March 15, 2024
Accepted May 30, 2024

Abstract—A system of nonlinear stochastic functional-difference equations with finite delay is considered. By assumption, this system has a "partial" trivial equilibrium (with respect to part of the state variables). The problem under study is to analyze partial stability in probability of this equilibrium: stability is considered with respect to part of the variables determining it. The problem is solved using a discrete-stochastic modification of the method of Lyapunov–Krasovskii functionals. Conditions for partial stability in probability are established. An example is provided to illustrate the features of the proposed approach and the rationale for introducing a one-parameter family of functionals.

Keywords: system of nonlinear stochastic functional-difference equations with finite delay, partial stability in probability, method of Lyapunov–Krasovskii functionals

DOI: 10.31857/S0005117924080026

1. INTRODUCTION

In the qualitative analysis and design of nonlinear dynamic systems, a separate area of research is associated with investigating the stability of systems of stochastic discrete (finite-difference) equations [1–9]. The interest in such systems is caused by the implementation of digital control systems, modeling problems in various fields, and numerical solution problems for systems of stochastic differential equations.

Within this research area, systems of discrete equations of order $m \geqslant 1$,

$$x(k+1) = X(k, x(k), x(k-1), \ldots, x(k-m)),$$

are interpreted as systems of *discrete equations with finite delay*; for example, see the monographs [9–11]. This interpretation provides new capabilities for the qualitative analysis of such systems, although they can be transformed into standard one-step systems of discrete equations by introducing new variables and extending the state space.

On the other hand, in networked control applications, there arise systems of discrete equations with *variable* delay:

$$x(k+1) = X(k, x(k), x(k-\tau_1(k)), \ldots, x(k-\tau_l(k))),$$

where functions $\tau_i(k)$ take an integer value from an interval $0 < \tau_i(k) \leqslant m$ at each discrete time instant k. When fixing the delay value, these systems can be treated as one-step discrete switched systems in the extended state space [12].

The general class of systems of nonlinear discrete equations with finite delay is defined by a system of *functional-difference* equations [9, 11, 13–19] of the form

$$x(k+1) = X(k, x_k),$$

whose state at each discrete time instant $k \in \mathbb{Z}_+ = \{0, 1, 2, \dots\}$ is described by a discrete vector function $x_k = x(k+j)$, $j \in \mathbb{Z}_0 = \{-m, -m+1, \dots, 0\}$ with the delayed argument; the number $m \geqslant 1$ specifies the delay value. This class of systems has been studied separately since the early 1990s and includes systems of discrete equations with constant, variable, and interval-type (single or multiple) delays.

With the transition to the functional-difference treatment, the stability of systems of nonlinear stochastic discrete equations with delay can be analyzed using a stochastic modification of the method of Lyapunov–Krasovskii functionals [9, 20–25] in the space of discrete (grid) functions. For deterministic nonlinear systems of discrete equations with delay, this approach was presented in [13–19]. The results obtained therein relate to the stability problem of the trivial equilibrium *with respect to all variables*. More general problems of *partial* stability have been actively considered in recent time. Note that their analysis significantly differs, see the review [26]. However, partial stability problems have not been studied for systems of nonlinear stochastic discrete equations with delay.

This paper is devoted to a general class of systems of nonlinear stochastic discrete equations with finite delay. By assumption, the system under consideration has a "partial" trivial equilibrium (with respect to some part of the state variables). We formulate the problem of stability in probability of this equilibrium with respect to part of the variables determining it. This problem is solved using the method of functionals representing discrete analogs of Lyapunov–Krasovskii functionals. They are widespread in the analysis of systems of nonlinear functional-differential equations with aftereffect (delay).

2. PROBLEM STATEMENT

We consider the linear finite-dimensional space $\mathbb{R}^n$ of vectors x with the Euclidean norm $|x|$ and divide the vector x into two parts: $x = (y^{\mathrm{T}}, z^{\mathrm{T}})^{\mathrm{T}}$. (The symbol T indicates the transpose.) The following class of nonlinear stochastic discrete (finite-difference) equations of the first order has been sufficiently investigated by now [1, 2]:

$$x(k+1) = X(k, x(k), \xi(k)),$$

where $k \in \mathbb{Z}_+$ denotes the discrete time; $x(k)$ is a sequence of state vector values; $\xi(k)$ is a sequence of independent random vectors defined on a probability space (Ω, F, P) with identical distribution laws for each $k \in \mathbb{Z}_+$. Here Ω is a space of elementary events $\{\omega\}$ with measurable sets with respect to a filtration F_k defined by a σ-algebra F and a given probability measure $P : F \to [0, 1]$.

Applications often lead to more general systems of stochastic nonlinear discrete equations with finite delay [9, 20–24, 27, 28] described by

$$x(k+1) = X(k, x_k, \xi(k)). \tag{1}$$

At each discrete time instant $k \in \mathbb{Z}_+$, their state is defined by a discrete vector function $x_k = x(k+j)$, $j \in \mathbb{Z}$, with the delayed argument.

For each $k \in \mathbb{Z}_+$, let the operator $X(k, \psi, \xi)$ defining the right-hand side of system (1) in the space $\{\psi\}$ of discrete (grid) functions $\psi(\theta)$, $\theta \in \mathbb{Z}_0$ with the norm $\|\psi\| = \max\{|\psi(0)|, |\psi(-1)|, \dots, |\psi(-m)|\}$ be *continuous* in ψ, ξ in the domain $||\psi|| < \infty$. The initial state x_0 of system (1) is given by a set of values $x_0 = \{x(k_0), x(k_0-1), \ldots, x(k_0-m)\}$, which form a matrix of dimensions $n \times (m+1)$. Assume that this state is deterministic. Then for all $k_0 \geqslant 0, x_0$, there exist a unique random discrete process $x(k_0, x_0)$ adapted to the flow of σ-algebras F_k that is the solution of system (1) and a corresponding set of the sampled trajectories of system (1). We denote by $x(k) = x(k; k_0, x_0)$ the values of the random vector function $x(k_0, x_0)$ at step k of the process.

Due to the partition $x = (y^{\mathrm{T}}, z^{\mathrm{T}})^{\mathrm{T}}$ and the corresponding partition $x_k = (y_k^{\mathrm{T}}, z_k^{\mathrm{T}})^{\mathrm{T}}$, the system under consideration can be written as the two groups of equations

$$y(k+1) = Y(k, y_k, z_k, \xi(k)), \quad z(k+1) = Z(k, y_k, z_k, \xi(k)).$$

Under the condition $Y(k, 0, z_k, \xi(k)) \equiv 0$, the set $M\{x_k : y_k = 0\}$ is a "partial" equilibrium of system (1), i.e., an invariant set of this system. The existence of a "full" equilibrium $x_k = 0$ of system (1) is not required: this assumption may even contradict the common sense of the problems being solved.

The stability of the "partial" equilibrium $y_k = 0$ is considered with respect to a given part of the variables determining it (not all of them). For this purpose, we suppose that $y = (y_1^{\mathrm{T}}, y_2^{\mathrm{T}})^{\mathrm{T}}$, where the vector y_1 includes only those components of the vector y with respect to which stability will be analyzed. To expand the circle of concepts for the y_1-stability of the "partial" equilibrium $y_k = 0$, we introduce an arbitrary partition $z = (z_1^{\mathrm{T}}, z_2^{\mathrm{T}})^{\mathrm{T}}$ of the vector z into two groups of variables.

Let D_δ denote the domain of values x_0 such that $||y_0|| < \delta$, $||z_{10}|| < L$, and $||z_{20}|| < \infty$. Here the norms are defined by

$$||y_0|| = \max |y(k_0 + j)|, \quad ||z_{i0}|| = \max |z_i(k_0 + j)| \;\; (i = 1, 2) \;\; \text{for} \;\; j \in \mathbb{Z}_0.$$

Definition 1. The "partial" equilibrium $y_k = 0$ of system (1) is said to be:

1) y_1-stable in probability for large values of z_{10} and on the whole with respect to z_{20} if, for each $k_0 \in \mathbb{Z}_+$, any arbitrarily small numbers $\varepsilon > 0$ and $\gamma > 0$, and any given number $L > 0$, there exists a number $\delta(\varepsilon, \gamma, L, k_0) > 0$ such that

$$P\left\{\sup_{k \geqslant k_0} |y_1(k; k_0, x_0)| > \varepsilon\right\} < \gamma \quad \text{for all } k \geqslant k_0 \text{ and } x_0 \in D_\delta; \tag{2}$$

2) uniformly y_1-stable in probability for large values of z_{10} and on the whole with respect to z_{20} if $\delta = \delta(\varepsilon, \gamma, L)$.

We formulate the following problem: find conditions under which the "partial" equilibrium $y(k) = 0$ of system (1) is y_1-stable in probability for large values of z_{10} and on the whole with respect to z_{20}, using the method of discrete Lyapunov–Krasovskii functionals.

This problem can also be treated as auxiliary in the stability analysis of the "partial" equilibrium $y_k = 0$ of system (1) with respect to all variables; when adding the control inputs $u = u(k, x_k)$ to system (1), a corresponding partial stabilization problem arises naturally.

Remark 1. If x_0 is a random variable with values in $\mathbb{R}^{n \times (m+1)}$ (independent of $\xi(k)$) and the inclusion $x_0 \in D_\delta$ holds almost surely, we arrive at the definitions similar to those of partial stability (see [29]).

Remark 2. These notions are very close to the following ones of partial stability: with respect to all [30] and some [31] variables of the "partial" equilibrium of systems of stochastic Itô differential equations and systems of stochastic discrete equations [8]. Also, they are very close to stability with respect to part of the variables of systems of functional-differential equations with aftereffect (delay) [32, 33].

3. CONDITIONS OF PARTIAL STABILITY IN PROBABILITY

In the space $\{\psi\}$ of discrete (grid) functions $\psi(\theta)$, $\theta \in \mathbb{Z}_0$, we consider single-valued scalar functionals $V = V(k,\psi)$, $V(k,0) \equiv 0$, that are continuous in ψ for each $k \in \mathbb{Z}_+$ and are defined in the domain

$$k \geqslant 0, \quad ||\psi_{y1}|| < h, \quad ||\psi_{y2}|| + ||\psi_z|| < \infty. \tag{3}$$

The partition $\psi = (\psi_{y1}^{\mathrm{T}}, \psi_{y2}^{\mathrm{T}}, \psi_z^{\mathrm{T}})^{\mathrm{T}}$ corresponds to the above partition $x = (y_1^{\mathrm{T}}, y_2^{\mathrm{T}}, z^{\mathrm{T}})^{\mathrm{T}}$ of the state vector x; $||\psi_{yi}|| = \max|\psi_{yi}(\theta)|$ $(i = 1,2)$, $||\psi_z|| = \max|\psi_z(\theta)|$ for $\theta \in \mathbb{Z}_0$.

The vector function $x_k = x_k(k_0, x_0)$ defines a discrete element of the trajectory of system (1) at step k of the process. Let us substitute this function into the functional $V(t,\psi)$. An analog of its derivative due to system (1) is the averaged differences (increments) [1, 2, 9]

$$LV(k,\psi) = E_{k,\psi}[V(k+1, x_{k+1}(k_0,x_0))] - V(k,\psi),$$

where the operator $E_{k,\psi}$ defines the conditional expectation of the random variable $V(k+1, x_{k+1}(k_0,x_0))$ given $x_k(k_0,x_0) = \psi$.

In addition, to formulate partial stability conditions, we will also utilize the following auxiliary functionals and functions:

1. Scalar functionals $V^*(k,\psi_y,\psi_{z1})$ and $V^*(\psi_y,\psi_{z1})$, continuous in the domain (3), to indicate an *upper bound* for the functional V and an auxiliary vector function $\mu(k,\psi), \mu(k,0) \equiv 0$, to correct the construction domain of the functional V. The vector function ψ_{z1} is given by the partition $\psi_z = (\psi_{z1}^{\mathrm{T}}, \psi_{z2}^{\mathrm{T}})^{\mathrm{T}}$ corresponding to the partition $z = (z_1^{\mathrm{T}}, z_2^{\mathrm{T}})^{\mathrm{T}}$. Let us define $||\mu(k,\psi)|| = \sup|\mu(k,\psi(\theta))|$ for $k \in \mathbb{Z}_+,\ \theta \in \mathbb{Z}_0$;

2. A Continuous monotonically increasing in $r > 0$ scalar function $a(r)$, $a(0) = 0$, which specifies standard requirements for the main functional V in the form of a *lower bound*.

The auxiliary function $\mu(k,\psi)$ is introduced since y_1-stability analysis for the "partial" equilibrium $y_k = 0$ of system (1) in the common domain

$$||\psi_{y1}|| < h_1 < h, \quad ||\psi_{y2}|| + ||\psi_z|| < \infty \tag{4}$$

of the function space neither always reveals the desired properties of the functional V nor endows it with these properties. It is rational to study the functional V in the narrower domain

$$||\psi_{y1}|| + ||\mu(k,\psi)|| < h_1 < h, \quad ||\psi_{y2}|| + ||\psi_z|| < \infty \tag{5}$$

based on the following considerations: in fact, the y_1-stability of the "partial" equilibrium $y_k = 0$ of system (1) means satisfaction of the corresponding probability estimates (2) not only for the components of the vector y_1 but also for those of some function $\mu(k,x)$ of the state variables of system (1). In some cases, such a function $\mu(k,x)$ cannot be specified in advance. Therefore, the corresponding function $\mu(k,\psi)$ in the discrete function space should be naturally interpreted as an additional vector function, which is determined when solving the original y_1-stability problem (like a suitable functional V). This leads to the rationale for correcting the construction domain (4) of the functional V via an *additional* auxiliary function $\mu(k,\psi)$.

Theorem 1. *Assume that for system* (1), *it is possible to indicate a functional V and an additional vector function $\mu(k,\psi), \mu(k,0) \equiv 0$, so that, for each $k \in \mathbb{Z}_+$ and a sufficiently small number $h_1 > 0$, the following conditions will hold in the domain* (5) :

$$V(k,\psi) \geqslant a(|\psi_{y1}(0)| + |\mu(k,\psi(0))|), \tag{6}$$

$$V(k,\psi) \leqslant V^*(k,\psi_y,\psi_{z1}), \quad V^*(k,0,\psi_{z1}) \equiv 0, \tag{7}$$

$$LV(k,\psi) \leqslant 0. \tag{8}$$

Then the "partial" equilibrium $y_k = 0$ of system (1) *is y_1-stable in probability for large values of z_{10} and on the whole with respect to z_{20}.*

Theorem 2. *With conditions* (7) *being replaced by*

$$V(k,\psi) \leqslant V^*(\psi_y, \psi_{z1}), \quad V^*(0,\psi_{z1}) \equiv 0, \tag{9}$$

under conditions (6) *and* (8), *the "partial" equilibrium $y_k = 0$ of system* (1) *is uniformly y_1-stable in probability for large values of z_{10} and on the whole with respect to z_{20}.*

The proofs of Theorems 1 and 2 are provided in the Appendix.

Remark 3. In Theorems 1 and 2, the auxiliary functional V and its averaged difference (increment) $LV(k,\psi)$ due to system (1) are, generally speaking, *alternating* in the domain (4). Along with the main functional V, the additional auxiliary function μ is introduced for the most rational replacement of the domain (4) with the domain (5). Conditions (7) and (9) identify an *admissible structure* of the functional V, which is determined by the specifics of the partial stability problem posed above: under these conditions, it is possible to use an *arbitrary* continuous functional V^* with $V^*(k,0,\psi_{z1}) \equiv 0$ or $V^*(0,\psi_{z1}) \equiv 0$ that bounds the functional V from above.

Remark 4. We can take sign-definite (in all variables) *quadratic* functionals (or those of higher order) $V(k,\psi) \equiv V^*(k,\psi_{y1},\mu(k,\psi))$ as admissible ones. In this case, the choice of the functions μ must be consistent with conditions (7) and (9): for example, the functions of the form $\mu = \mu(\psi_{y2},\psi_{z1}), \mu(0,\psi_{z1}) \equiv 0$, are admissible.

Remark 5. Let system (1) have the "full" equilibrium $x_k = 0$, $\mu(k,\psi) \equiv 0$, $\xi(k) \equiv 0$, and the initial condition $x_0 \in D_\delta$ be replaced by $||x_0|| < \delta$. Then, under conditions (6) and (8), we arrive at a discrete variant [34, 35] of the classical Rumyantsev theorem [36] on stability with respect to a given part of the variables and its modification for $\mu(k,\psi) \neq 0$ [37]. In this case, conditions (7) and (9) are not required.

Remark 6. Azbelev's scientific school developed another approach to the interpretation and analysis of the stability of stochastic functional-difference systems; for details, see [38].

4. EXAMPLES

We distinguish two classes of nonlinear discrete systems of a given structure, for which partial stability is analyzed in the parameter space. At the same time, we will demonstrate the rationale for using a one-parameter family of functionals.

Example 1. Consider the discrete system (1) composed of the equations

$$\begin{gathered} y_1(k+1) = [a_1 + \alpha_1\xi_1(k)]y_1(k) + a_2y_1(k-1) + ly_2(k-1)z_1(k-1) + \alpha_2y_1(k-1)\xi_2(k), \\ y_2(k+1) = [b + dy_1(k-1)]y_2(k), \\ z_1(k+1) = [c + ey_1(k-1)]z_1(k), \quad z_2(k+1) = Z_2(k, x_k, \xi(k)), \end{gathered} \tag{10}$$

where $\xi_1(k)$ and $\xi_2(k)$ are mutually uncorrelated sequences of independent random variables with identical standard Gaussian distributions for each $k \in \mathbb{Z}_+$; $a_1, a_2, b, c, d, e, l, \alpha_1$, and α_2 are constant parameters. System (10) is a special case of system (1) with $m = 1$; the operator Z_2 satisfies only the general requirements of system (1) for $m = 1$.

System (10) has the "partial" equilibrium

$$y_{1k} = y_{2k} = 0. \tag{11}$$

Consider the family ($M, \beta_1, \beta_2 = \text{const} > 0$) of functionals

$$V(\psi) = \psi_{y1}^2(0) + M\psi_{y2}^2(0)\psi_{z1}^2(0) + (\beta_1 + \alpha_2^2)\psi_{y1}^2(-1) + \beta_2\psi_{y2}^2(-1)\psi_{z1}^2(-1), \tag{12}$$

representing discrete analogs of the Lyapunov–Krasovskii functionals, and the auxiliary scalar discrete function

$$\mu_1(\psi(\theta)) = \psi_{y2}(\theta)\psi_{z1}(\theta), \quad \theta \in \mathbb{Z}_0 = \{k = -1, 0\}; \tag{13}$$

let $\mu_1(0)$ and $\mu_1(-1)$ denote the values of the function $\mu_1(\psi(\theta))$ for $\theta = 0$ and $\theta = -1$, respectively.

We have the relations

$$\psi_{y1}^2(0) + M\mu_1^2(0) \leqslant V(\psi) = V^*(\psi_{y1}, \psi_{y2}, \psi_{z1}), \quad V^*(0, 0, \psi_{z1}) \equiv 0.$$

The functional V (12) satisfies conditions (6) and (7) in the domain (5), and for all $k \in \mathbb{Z}_+ \cup \mathbb{Z}_0$ its averaged difference (increment) $LV(\psi)$ due to system (10) is defined by

$$\begin{gathered}
LV(\psi) = E_{k,\psi}\{V(\psi(0), X(k, \psi(-1), \psi(0), \xi(k)))\} - V(\psi(-1), \psi(0)) \\
= E_{k,\psi}\{[a_1\psi_{y1}(0) + a_2\psi_{y1}(-1) + l\psi_{y2}(-1)\psi_{z1}(-1) + \alpha_1\psi_{y1}(0)\xi_1(k) + \alpha_2\psi_{y1}(-1)\xi_2(k)]^2 \\
+ M\psi_{y2}^2(0)\psi_{z1}^2(0)[b + d\psi_{y1}(-1)]^2[c + e\psi_{y1}(-1)]^2\} - \psi_{y1}^2(0) - M\psi_{y2}^2(0)\psi_{z1}^2(0) \\
+(\beta_1 + \alpha_2^2)[\psi_{y1}^2(0) - \psi_{y1}^2(-1)] + \beta_2[\psi_{y2}^2(0)\psi_{z1}^2(0) - \psi_{y2}^2(-1)\psi_{z1}^2(-1)] \\
= a_1^2\psi_{y1}^2(0) + 2a_1a_2\psi_{y1}(0)\psi_{y1}(-1) + 2a_1l\psi_{y1}(0)\mu_1(-1) \\
+ 2a_2l\psi_{y1}(-1)\mu_1(-1) + l^2\mu_1^2(-1) + \alpha_1^2\psi_{y1}^2(0) + \alpha_2^2\psi_{y1}^2(-1) + Mb^2c^2\mu_1^2(0) \\
+ r_1\psi_{y1}(-1)\mu_1^2(0) + r_2\psi_{y1}^2(-1)\mu_1^2(0) + r_3\psi_{y1}^3(-1)\mu_1^2(0) + Md^2e^2\psi_{y1}^4(-1)\mu_1^2(0) \\
- \psi_{y1}^2(0) - M\mu_1^2(0) + (\beta_1 + \alpha_2^2)[\psi_{y1}^2(0) - \psi_{y1}^2(-1)] + \beta_2[\mu_1^2(0) - \mu_1^2(-1)] \\
= (a_1^2 + \alpha_1^2 + \alpha_2^2 - 1 + \beta_1)\psi_{y1}^2(0) + 2a_1a_2\psi_{y1}(0)\psi_{y1}(-1) + (a_2^2 - \beta_1)\psi_{y1}^2(-1) \\
+ 2a_1l\psi_{y1}(0)\mu_1(-1) + 2a_2l\psi_{y1}(-1)\mu_1(-1) + (Mb^2c^2 - M + \beta_2)\mu_1^2(0) \\
+ (l^2 - \beta_2)\mu_1^2(-1) + r_1\psi_{y1}(-1)\mu_1^2(0) + r_2\psi_{y1}^2(-1)\mu_1^2(0) \\
+ r_3\psi_{y1}^3(-1)\mu_1^2(0) + Md^2e^2\psi_{y1}^4(-1)\mu_1^2(0), \\
r_1 = bcr_0, r_2 = M(b^2e^2 + 4bcde + c^2d^2), r_3 = der_0, r_0 = 2M(be + cd).
\end{gathered}$$

In these formulas, the conditional expectation has been calculated considering the relations $E[\xi_i(k)] = 0$, $E[\xi_i^2(k)] = 1$, corresponding to the standard Gaussian distributions of the mutually uncorrelated random variables $\xi_i(k)$ $(i = 1, 2)$.

For the sake of simpler analysis, we utilize the inequalities

$$\begin{gathered}
2a_1a_2\psi_{y1}(0)\psi_{y1}(-1) \leqslant |a_1a_2|[\psi_{y1}^2(0) + \psi_{y1}^2(-1)], \\
2a_2l\psi_{y1}(-1)\mu_1(-1) \leqslant |a_2l|[\psi_{y1}^2(-1) + \mu_1^2(-1)]
\end{gathered}$$

to obtain the following upper bound for the quadratic part $(LV)_2$ of the above expression for $LV(\psi)$:

$$\begin{gathered}
(LV)_2 \leqslant (a_1^2 + |a_1a_2| + \alpha_1^2 + \alpha_2^2 - 1 + \beta_1)\psi_{y1}^2(0) \\
+ 2a_1l\psi_{y1}(0)\mu_1(-1) + (l^2 + |a_2l| - \beta_2)\mu_1^2(-1) \\
+ (a_2^2 + |a_1a_2| + |a_2l| - \beta_1)\psi_{y1}^2(-1) + (Mb^2c^2 - M + \beta_2)\mu_1^2(0).
\end{gathered}$$

Under the conditions

$$\begin{gathered}
a_1^2 + |a_1a_2| + \alpha_1^2 + \alpha_2^2 - 1 + \beta_1 < 0, \\
(a_1^2 + |a_1a_2| + \alpha_1^2 + \alpha_2^2 - 1 + \beta_1)(l^2 + |a_2l| - \beta_2) > a_1^2l^2, \\
a_2^2 + |a_1a_2| + |a_2l| - \beta_1 < 0, \quad Mb^2c^2 - M + \beta_2 < 0,
\end{gathered}$$

$(LV)_2$ is a negative definite function of the variables $\psi_{y1}(0)$, $\psi_{y1}(-1), \mu_1(0)$, and $\mu_1(-1)$ based on Sylvester's criterion. Therefore, for a sufficiently small number $h_1 > 0$, the averaged difference (increment) $LV(\psi)$ of the functional (12) satisfies the inequality $LV(\psi) \leqslant 0$ in the domain (5).

Assuming that the parameters of system (10) satisfy the conditions

$$\begin{gathered}(|a_1| + |a_2|)^2 + |a_2 l| + a_1^2 + a_2^2 < 1, \\ [(|a_1| + |a_2|)^2 + |a_2 l| + \alpha_1^2 + \alpha_2^2 - 1][l^2 + |a_2 l| + M(b^2c^2 - 1)] > a^2 l^2,\end{gathered} \tag{14}$$

we choose the parameters β_1 and β_2 of the functional (12) as

$$\beta_1 = a_2^2 + |a_1 a_2| + |a_2 l| + \varepsilon_1, \quad \beta_2 = M(1 - b^2c^2) - \varepsilon_2.$$

For sufficiently small numbers ε_1, ε_2, and $h_1 > 0$, we have $LV(\psi) \leqslant 0$ in the domain (5) (*but not in the domain* (4)) for any values of the parameters d and e. Hence, besides conditions (6) and (7), condition (8) holds for the functional V (12) in the domain (5).

Based on Theorem 2, under conditions (14), the "partial" equilibrium (11) of system (10) is uniformly y_1-stable in probability for large values of z_{10} and on the whole with respect to z_{20}. The operator $LV(\psi)$ associated with system (10) is *alternating* in the domain (4).

Note that given combinations of the parameters of system (10) can be included in (or excluded from) the stability domain by choosing an appropriate number M. For example, if $l^2 + |a_2 l| = 1$, then the seemingly "natural" choice $M = 1$ in the functional (12) makes the stability domain (14) an empty set for any values of the parameters b, c, d, and e. However, in the same case $l^2 + |a_2 l| = 1$, we may consider the domain (14) by setting $M = 2$.

On the other hand, given a number M, the y_1-stability domain can be changed by modifying the estimates for the quadratic part $(LV)_2$ of the expression defining $LV(\psi)$. Indeed, with the inequality

$$2a_1 l\psi_{y1}(0)\mu_1(-1) \leqslant |a_1 l|[\psi_{y1}^2(0) + \mu_1^2(-1)],$$

in the case of $a_2 = 0$ and a sufficiently small number $h_1 > 0$, the inequality $LV(\psi) \leqslant 0$ will hold in the domain (5) under the conditions

$$\begin{gathered}a_1^2 + \alpha_1^2 + \alpha_2^2 + |a_1 l| - 1 + \beta_1 \leqslant 0, \\ l^2 + |a_1 l| - \beta_2 \leqslant 0, \quad M(b^2c^2 - 1) + \beta_2 \leqslant 0.\end{gathered}$$

In this case, the y_1-stability domain is given by

$$(|a_1| + |l|)^2 + \alpha_1^2 + \alpha_2^2 < 1 + M(1 - b^2c^2), \quad b^2c^2 < 1; \tag{15}$$

in contrast to the domain (14), for $M = l^2 = 1$ the domain (15) is a non-empty set. For illustrative purposes, note that for $M = 1$ and $b^2c^2 = \alpha_1^2 + \alpha_2^2 = 0$, the stability domains (14) and (15) have the form $a_1^2 + l^2 < 1$ and $|a_1| + |l| < \sqrt{2}$, respectively; moreover, the domain (15) covers the case $l^2 = 1$.

In addition, for $\alpha_2 = 0$ and $M = 2$, the domain (12) coincides with that of the uniform y_1-stability in probability of the "partial" equilibrium $y_1(k) = y_2(k) = 0$ of system (10) without aftereffect. Such a system was analyzed in [8] using the Lyapunov function $V(x) = y_1^2 + 2y_2^2z_1^2$ and the auxiliary function $\mu_1 = y_2 z_1$.

As a numerical experiment, we present the results of calculations using the recurrence relations (12) on the interval $k \in [0, 25]$ for $y_i(-1) = y_i(0) = 0.1$ $(i = 1, 2)$, $z_1(-1) = z_1(0) = 1$ and the parameters $a_1 = 1/2$, $b = 3/2$, $a_2 = 0$, $c = 1/3$, and $d = e = l = 1$.

Table

k	$y_1(k)$	$y_2(k)$	$z_1(k)$	$\xi_1(k)$	$\xi_2(k)$	$y_1(k)$	$y_2(k)$	$z_1(k)$
–1	0.1	0.1	1	–	–	0.1	0.1	1
0	0.1	0.1	1	0	0	0.1	0.1	1
1	0.15	0.16	0.4333	–1	0	0.15	0.16	0.4333
2	0.1750	0.2560	0.1877	1	0	0.1251	0.2560	0.1877
3	0.1568	0.4224	0.2094	1	0	0.1735	0.4224	0.2094
4	0.1264	0.4288	0.0954	0	0	0.1926	0.6864	0.0960
5	0.1517	0.7104	0.0468	–1	0	0.1847	1.1487	0.0487
6	0.1168	1.1554	0.0215	–1	0	0.0967	1.9443	0.0256
7	0.0916	1.9084	0.0104	1	0	0.0720	3.2757	0.0133
8	0.0706	3.0854	0.0060	0	0	0.1098	5.2303	0.0057
9	0.0551	4.9107	0.0025	0	0	0.0985	8.2220	0.0023
10	0.0461	7.7127	0.0004	1	0	0.0791	13.236	0.0011
...	...	...	...	...	...	...	...	...
15	0.0050	65.557	5.5×10^{-6}	–1	0	0.0151	127.93	1.1×10^{-5}
...	...	...	...	...	...	...	...	...
20	0.00039	504.13	2.4×10^{-8}	–1	0	0.0010	1005.5	5.4×10^{-8}
...	...	...	...	...	...	...	...	...
25	1.9×10^{-5}	3832.1	4.7×10^{-10}	0	0	2.9×10^{-5}	7649.1	2.2×10^{-10}

For the "undisturbed" case $\xi_{1,2}(k) \equiv 0$, the calculation results are combined on the left-hand side of the table. Under the random disturbances $\xi_1(k)$ and $\xi_2(k)$ whose intensities α_1 and α_2 satisfy condition (14), the sample trajectories are grouped around the undisturbed trajectory focusing along the Oy_2 axis as $k \to \infty$. To assess the effect of random disturbances on the dynamics of system (10), we also provide the calculation results for $\alpha_1 = 1/3$, $\alpha_2 = 0$, and the same parameter values in the case $\xi_2(k) \equiv 0$ and the admissible realization $\xi_1(k)$ on the interval $k \in [0, 25]$ given by the sequence $\{0, -1, 1, 1, 0, -1, -1, 1, 0, 0, 1, 1, -1, 0, 0, -1, 1, 0, -1, 1, -1, 0, 1, -1, 0\}$ (see the right-hand side of the table).

Example 2. Consider the discrete system (1) composed of the equations

$$\begin{gathered} y_1(k+1) = [a_1 + \alpha_1\xi_1(k) + ly_2(k-1)z_1(k-1)]y_1(k) + \alpha_2 y_1(k-1)\xi_2(k), \\ y_2(k+1) = [b + dy_1(k-1)]y_2(k), \\ z_1(k+1) = [c + ey_1(k-1)]z_1(k), \quad z_2(k+1) = Z_2(k, x_k, \xi(k)), \end{gathered} \tag{16}$$

which represent a structural modification of system (10).

To analyze the y_1-stability in probability of the "partial" equilibrium (11) of system (16), we utilize the discrete functional V (12) for $\beta_1 = 0$ and $\beta_2 = M(1 - b^2c^2) - \varepsilon_2$ and the auxiliary discrete function (13).

The quadratic part $[LV(\psi)]_2$ of the averaged difference (increment) $LV(\psi)$ of this functional due to system (16) has the form

$$[LV(\psi)]_2 = (a_1^2 + \alpha_1^2 + \alpha_2^2 - 1)\psi_{y1}^2(0) + (Mb^2c^2 - M + \beta_2)\mu_1^2(0) + (l^2 - \beta_2)\mu_1^2(-1).$$

Therefore, for a sufficiently small number $h_1 > 0$, the inequality $LV(\psi) \leqslant 0$ will hold in the domain (5) for any values of the parameters d, e, l_1, and l_2 if

$$a_1^2 + \alpha_1^2 + \alpha_2^2 < 1, \quad M(b^2c^2 - 1) + \beta_2 < 0, \quad l^2 - \beta^2 < 0.$$

As a result, under the conditions

$$a_1^2 + \alpha_1^2 + \alpha_2^2 < 1, \quad l^2 + M(b^2c^2 - 1) < 1, \tag{17}$$

the "partial" equilibrium (11) of system (17) is uniformly y_1-stable in probability for large values of z_{10} and on the whole with respect to z_{20} based on Theorem 2.

Similar to the analysis of system (10), if $l^2 = 1$, then the "natural" choice $M = 1$ in the functional (12) makes the stability domain (17) an empty set for any values of the parameters b, c, d, and e. However, this set can be expanded as well with a suitably chosen number M, e.g., $M = 2$.

5. CONCLUSIONS

This paper has considered a nonlinear system of stochastic functional-difference equations with finite delay subjected to a discrete "white" noise process. For such a system, problems of stability in probability with respect to a given part of the variables of the "partial" trivial equilibrium have been formulated. The discrete vector function defining the initial state of the system has been supposed deterministic.

Sufficient stability conditions have been established using a discrete-stochastic version of the method of Lyapunov–Krasovskii functionals in an appropriate modification. Along with the main discrete functional V, an additional auxiliary discrete function μ (generally speaking, a vector function) has been considered for correcting the construction domain of the functional V in the discrete function space. The rationale for this approach lies in that the resulting functional V and its averaged difference (increment) due to the system under consideration can be alternating.

APPENDIX

Proof of Theorem 1.

Assume that conditions (6)–(8) are valid for each $k \in \mathbb{Z}_+$ and a sufficiently small number $h_1 > 0$ in the domain (5). Let us take arbitrary numbers ε $(0 < \varepsilon < h_1)$ and k_0 and an initial value x_0 from the domain $D_\varepsilon = \{||y_0|| < \varepsilon, ||z_{10}|| \leqslant L, ||z_{20}|| < \infty\}$. Consider the random process $x(k_0, x_0)$ representing the solution of system (1). We denote by k_ε the "integer" moment when this process first leaves the domain $|y_1| \leqslant \varepsilon : k_\varepsilon = \inf\{k : |y_1(k; k_0, x_0)| > \varepsilon\}$ for $k \geqslant k_0$. Letting $t(k) = \min(k_\varepsilon, k); t(k_0) = k_0$, we have the equalities

$$V\left(t(k), x_{t(k)}(k_0, x_0)\right) - V(k_0, x_0) = \sum_{s=k_0}^{k-1} \Delta V\left(t(s), x_{t(s)}(k_0, x_0)\right);$$

$$\Delta V\left(t(s), x_{t(s)}(k_0, x_0)\right) = \Delta V\left(t(s+1), x_{t(s+1)}(k_0, x_0)\right) - \Delta V\left(t(s), x_{t(s)}(k_0, x_0)\right).$$

Due to these equalities, the sequence $v(k)$ of the random variables $v(k) = V(t(k), x_{t(k)}(k_0, x_0))$ generated by the realizations $x(k, \omega), \xi(k, \omega)$ of the random process $x(k), \xi(k)$ defined by system (1) satisfies the "averaged" relations

$$\begin{gathered} E\left[V\left(t(k), x_{t(k)}(k_0, x_0)\right) - V(k_0, x_0)\right] \\ = EV\left(t(k), x_{t(k)}(k_0, x_0)\right) - V(k_0, x_0) = \sum_{s=k_0}^{k-1} E\Delta V\left(t(s), x_{t(s)}(k_0, x_0)\right). \end{gathered} \tag{A.1}$$

By the rule for calculating the repeated expectation, from (A.1) it follows that

$$\begin{gathered} E\Delta V\left(t(s), x_{t(s)}(k_0, x_0)\right) \\ = E\left\{E_{t(s), x_{t(s)}(k_0, x_0)}\left[V(t(s+1), x_{t(s+1)}(k_0, x_0))\right]\right\} - V\left(t(s), x_{t(s)}(k_0, x_0)\right) \\ = E\left[LV\left(t(s), x_{t(s)}(k_0, x_0)\right)\right], \end{gathered}$$

and we arrive at the discrete-functional version of Dynkin's formula:

$$EV\left(t(k), x_{t(k)}(k_0, x_0)\right) - V(k_0, x_0) = \sum_{s=k_0}^{k-1} E\left[LV\left(t(s), x_{t(s)}(k_0, x_0)\right)\right].$$

Consequently, based on condition (8),

$$EV\left(t(k), x_{t(k)}(k_0, x_0)\right) \leqslant V(k_0, x_0) < \infty. \tag{A.2}$$

If $k > k_\varepsilon$ (in this case, $t(k) = k_\varepsilon$), we have $|y_1(t(k); k_0, x_0)| = |y_1(k_\varepsilon; k_0, x_0)| \geqslant \varepsilon$. If $k < k_\varepsilon$ (in this case, $t(k) = k$), the Chebyshev–Markov inequality and the estimate (A.2) yield

$$\begin{gathered} P[|y_1(k; k_0, x_0)| > \varepsilon] \leqslant a^{-1}(\varepsilon)E[a(|y_1(k; k_0, x_0)|)] \\ \leqslant a^{-1}(\varepsilon)E[a(|y_1(k; k_0, x_0)| + |\mu(k, x(k; k_0, x_0))|)] \\ \leqslant a^{-1}(\varepsilon)E[V(k, x_{t(k)}(k_0, x_0))] \\ = a^{-1}(\varepsilon)E[V(t(k), x_{t(k)}(k_0, x_0))] \leqslant a^{-1}(\varepsilon)V(k_0, x_0). \end{gathered} \tag{A.3}$$

The functional V is continuous for each $k \in \mathbb{Z}_+$, $V(t, 0) \equiv 0$, and conditions (7) hold. Therefore, for all $k_0 \geqslant 0$ and any given number $L > 0$, the limit relation

$$\lim_{||y_0|| \to 0} V(k_0, x_0) = 0 \tag{A.4}$$

is valid for $||z_{10}|| \leqslant L$ uniformly in $||z_{20}|| < \infty$.

Hence, for all $k_0 \geqslant 0$ and any given number $L > 0$, inequalities (A.3) and (A.4) lead to the limit relation

$$\lim_{||y_0|| \to 0} P\left[\sup_{k > k_0} |y_1(k; k_0, x_0)| > \varepsilon\right] = 0,$$

holding for $||z_{10}|| \leqslant L$ uniformly in $||z_{20}|| < \infty$. As a result, for each $k_0 \geqslant 0$, any arbitrarily small numbers $\varepsilon > 0$ and $\gamma > 0$, and any given number $L > 0$, there exists a number $\delta(\varepsilon, \gamma, L, k_0) > 0$ such that inequality (2) will hold for all $k \geqslant k_0$ and $x_0 \in D_\delta$. Thus, the "partial" equilibrium $y_k = 0$ of system (1) is y_1-stable in probability for large values of z_{10} and on the whole with respect to z_{20}.

Proof of Theorem 2.

If conditions (9) are satisfied instead of conditions (7), for any given number $L > 0$ the limit relation (A.4) will hold for $||z_{10}|| \leqslant L$ uniformly in $||z_{20}|| < \infty$ and, moreover, in $k_0 \geqslant 0$. As a result, for each $k_0 \geqslant 0$, any arbitrarily small numbers $\varepsilon > 0$ and $\gamma > 0$, and any given number $L > 0$, there exists a number $\delta(\varepsilon, \gamma, L) > 0$ independent of k_0 such that inequality (2) will hold for all $k \geqslant k_0$ and $x_0 \in D_\delta$. Thus, the "partial" equilibrium $y_k = 0$ of system (1) is uniformly y_1-stable in probability for large values of z_{10} and on the whole with respect to z_{20}.

REFERENCES

1. Halanay, A. and Wexler, D., *Kachestvennaya teoriya impul'snykh sistem* (Qualitative Theory of Impulse Systems), Moscow: Mir, 1971.

2. Pakshin, P.V., *Diskretnye sistemy so sluchainymi parametrami i strukturoi* (Discrete Systems with Random Parameters and Structure), Moscow: Fizmatlit, 1994.

3. Azhmyakov, V.V. and Pyatnitskiy, E.S., Nonlocal Synthesis of Systems for Stabilization of Discrete Stochastic Controllable Objects, *Autom. Remote Control*, 1994, vol. 55, no. 2, pp. 202–210.

4. Barabanov, I.N., Construction of Lyapunov Functions for Discrete Systems with Stochastic Parameters, *Autom. Remote Control*, 1995, vol. 56, no. 11, pp. 1529–1537.

5. Teel, A.R., Hespanha, J.P., and Subbaraman, A., Equivalent Characterizations of Input-to-State Stability for Stochastic Discrete-Time Systems, *IEEE Trans. Autom. Control*, 2014, vol. 59, no. 2, pp. 516–522.

6. Jian, X.S., Tian, S.P., Zhang, T.L., and Zhang, W.H., Stability and Stabilization of Nonlinear Discrete-Time Stochastic Systems, *Int. J. Robust Nonlinear Control*, 2019, vol. 29, no. 18, pp. 6419–6437.

7. Qin, Y., Cao, M., and Anderson, B.D.O., Lyapunov Criterion for Stochastic Systems and Its Applications in Distributed Computation, *IEEE Trans. Autom. Control*, 2020, vol. 65, no. 2, pp. 546–560.

8. Vorotnikov, V.I. and Martyshenko, Yu.G., On the Problem of Partial Stability for Discrete-Time Stochastic Systems, *Autom. Remote Control*, 2021, vol. 82, no. 9, pp. 1554–1567.

9. Shaikhet, L., *Lyapunov Functionals and Stability of Stochastic Difference Equations*, Springer Science & Business Media, 2013.

10. Astrom, K.J. and Wittenmark, B., *Computer Controlled Systems: Theory and Design*, 1984.

11. Fridman, E., *Introduction to Time-Delay Systems: Analysis and Control*, Boston: Birkhauser, 2014.

12. Hetel, L., Daafouz, J., and Iung, C., Equivalence between the Lyapunov–Krasovskii Functionals Approach for Discrete Delay Systems and That of the Stability Conditions for Switched Systems, *Nonlinear Analysis: Hybrid Systems*, 2008, vol. 2, no. 3, pp. 697–705.

13. Rodionov, A.M., Certain Modifications of Theorems of the Second Lyapunov Method for Discrete Equations, *Autom. Remote Control*, 1992, vol. 53, no. 9, pp. 1381–1386.

14. Elaydi, S. and Zhang, S., Stability and Periodicity of Difference Equations with Finite Delay, *Funkcialaj Ekvacioj.*, 1994, vol. 37, no. 3, pp. 401–413.

15. Anashkin, O.V., Lyapunov Functions in Stability Theory of Nonlinear Difference Delay Equations, *Differential Equations*, 2002, vol. 38, pp. 1038–1041.

16. Pepe, P., Pola, G., and Di Benedetto, M.D., On Lyapunov–Krasovskii Characterizations of Stability Notions for Discrete-Time Systems with Uncertain Time-Varying Time Delays, *IEEE Trans. Autom. Control*, 2017, vol. 63, no. 6, pp. 1603–1617.

17. Aleksandrov, A.Y. and Aleksandrova, E.B., Delay-Independent Stability Conditions for a Class of Nonlinear Difference Systems, *J. Franklin Institute*, 2018, vol. 355, no. 7, pp. 3367–3380.

18. Zhou, B., Improved Razumikhin and Krasovskii Approaches for Discrete-Time Time-Varying Time-Delay Systems, *Automatica*, 2018, vol. 91, pp. 256–269.

19. Li, X., Wang, R., Du, S., and Li, T., An Improved Exponential Stability Analysis Method for Discrete-Time Systems with a Time-Varying Delay, *Int. J. Robust Nonlin. Control*, 2022, vol. 32, no. 2, pp. 669–681.

20. Kolmanovskii, V.B. and Shaikhet, L.E., General Method of Lyapunov Functionals Construction for Stability Investigations of Stochastic Difference Equations, in *Dynamical Systems and Applications*, World Scientific, 1995, vol. 4, pp. 397–439.

21. Paternoster, B. and Shaikhet, L., About Stability of Nonlinear Stochastic Difference Equations, *Appl. Math. Lett.*, 2000, vol. 13, no. 5, pp. 27–32.

22. Rodkina, A. and Basin, M., On Delay-Dependent Stability for Vector Nonlinear Stochastic Delay-Difference Equations with Volterra Diffusion Term, *Syst. Control Lett.*, 2007, vol. 56, no. 6, pp. 423–430.

23. Diblik, J., Rodkina, A., and Smarda, Z., On Local Stability of Stochastic Delay Nonlinear Discrete Systems with State-Dependent Noise, *Appl. Math. Comp.*, 2020, vol. 374, art. no. 125019.

24. Shaikhet, L., Stability Investigation of Systems of Nonlinear Stochastic Difference Equations, in *Research Highlights in Mathematics and Computer Science*, Rodiono, L.G., Ed., B P International, 2022, vol. 2, pp. 79–92.

25. Shaikhet, L., Stability of the Exponential Type System of Stochastic Difference Equations, *Mathematics*, 2023, vol. 11, no. 18, art. no. 3975.

26. Vorotnikov, V.I., Partial Stability and Control: the State of the Art and Developing Prospects, *Autom. Remote Control*, 2005, vol. 66, no. 4, pp. 511–561.

27. Zong, X., Lei, D., and Wu, F., Discrete Razumikhin-Type Stability Theorems for Stochastic Discrete-Time Delay Systems, *J. Franklin Institute*, 2018, vol. 355, no. 17, pp. 8245–8265.

28. Ngoc, P.H.A. and Hieu, L.T., A Novel Approach to Exponential Stability in Mean Square of Stochastic Difference Systems with Delays, *Syst. Control Lett.*, 2022, vol. 168, art. no. 105372.

29. Mao, X.R. and Yuan, C.G., *Stochastic Differential Equations with Markovian Switching*, London: Imperial College Press, 2006.

30. Rajpurohit, T. and Haddad, W.M., Partial-State Stabilization and Optimal Feedback Control for Stochastic Dynamical Systems, *J. Dynam. Syst., Measuremen, Control*, 2017, vol. 139, no. 9, art. no. DS-15-1602.

31. Vorotnikov, V.I. and Martyshenko, Y.G., On the Partial Stability in Probability of Nonlinear Stochastic Systems, *Autom. Remote Control*, 2019, vol. 80, no. 5, pp. 856–866.

32. Vorotnikov, V.I., On Partial Stability and Detectability of Functional Differential Systems with Aftereffect, *Autom. Remote Control*, 2020, vol. 81, no. 2, pp. 199–210.

33. Vorotnikov, V.I. and Martyshenko, Yu.G., On the Partial Stability in the Probability of Nonlinear Stochastic Functional-Differential Systems with Aftereffect (Delay), *J. Comput. Syst. Sci. Int.*, 2024, vol. 63, no. 1, pp. 1–13.

34. Ignatyev, A.O., Lyapunov Function Method for Systems of Difference Equations: Stability with Respect to Part of the Variables, *Diff. Equat.*, 2022, vol. 58, no. 3, pp. 405–414.

35. Vorotnikov, V.I. and Martyshenko, Yu.G., Approach to the Stability Analysis of Partial Equilibrium States of Nonlinear Discrete Systems, *J. Comput. Syst. Sci. Int.*, 2022, vol. 61, no. 3, pp. 348–359.

36. Rumyantsev, V.V., On the Stability of Motion with Respect to Part of Variables, *Vestn. Mosk. Gos. Univ. Mat., Mekh., Phiz., Astron., Khim.*, 1957, no. 4, pp. 9–16.

37. Vorotnikov, V.I., *Partial Stability and Control*, Boston: Birkhauser, 1998.

38. Kadiev, R. and Ponosov, A., The W-Transform in Stability Analysis for Stochastic Linear Functional Difference Equations, *J. Math. Anal. Appl.*, 2012, vol. 389, no. 2, pp. 1239–1250.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 8, pp. 773–784.

ROBUST, ADAPTIVE, AND NETWORK CONTROL

A Comparative Analysis of Centrality Measures in Complex Networks

N. Meshcheryakova[*,a] **and S. Shvydun**[*,b]

[*]*HSE University, Moscow, Russia*
e-mail: [a]*natamesc@gmail.com,* [b]*shvydun@hse.ru*

Received December 11, 2023
Revised June 19, 2024
Accepted July 10, 2024

Abstract—Identification of central elements in networks is an ill-defined problem. Hence, a large number of centrality measures have been proposed in the literature. We present a survey of existing axioms, which characterize certain properties of centralities. We also perform a perturbation analysis of centrality measures in real and artificial networks.

Keywords: network analysis, centrality, axioms for centrality, perturbation analysis

DOI: 10.31857/S0005117924080031

1. INTRODUCTION

Network models are essential tools for modeling and studying complex systems. Networks can describe most existing technological, biological, social, and other systems, as well as model various socio-economic, transportation, epidemiological, industrial and other processes.

One of the fundamental challenges in network science is the detection of the most important participants [1]. Unfortunately, the identification of central elements is an ill-defined problem because there is no unique definition of importance. In general, the notion of importance depends on the problem statement and the nature of a network. Recently, a large number of centrality models have been proposed in the literature based on the number of connections of each node, the number of paths between them, individual attributes of the nodes and their influence in the group, etc. [2]. However, since the results of the centrality measures vary significantly, it is important to compare these models and estimate the quality of the obtained results.

Comparison of centrality indices is a complex task due to the subjective nature of importance in networks. In several studies, centrality measures can be compared with ground truth or expert knowledge about the real influence of nodes [3, 4]. In that case, the centrality measure, which best approximates the predefined influence, will be selected for further use in that network (e.g.: to identify central nodes in the same network at the next discrete time or to define the centrality of unlabeled nodes). Unfortunately, ground-truth data are extremely rare and difficult to collect in real-world networks.

Another way to compare centrality measures is to check the properties (called *the axioms of centrality*) that these measures satisfy [5–13]. The axioms provide insight into the performance of centrality indices and help to understand whether a particular set of nodes will be identified in a certain graph topology. Hence, the axioms of centrality assess how well a given centrality measure satisfies a set of logical properties. However, the current list of axioms is neither unified nor exhaustive, leading different studies to focus on various sets of properties.

An important aspect of centrality models is their robustness to small perturbations in a network structure. Since many real-world networks are incomplete or contain incorrect data (errors/missing data), the results of many centrality measures may be biased. There are several studies in the literature that investigate the sensitivity of various centrality models [14–19]. Nevertheless, most of these studies are limited to the analysis of only classical centralities on a small set of networks.

This study aims to provide an overview of the existing axioms that characterize the concept of centrality in network structures. We also analyze the sensitivity of centrality measures to small changes in a network. The results of this research provide a deeper understanding of how incorrect and incomplete data affect the results of existing centrality measures.

The paper is organized as follows. Section 2 describes the centrality models that we study in the paper. Section 3 provides an overview of the existing axioms for centrality measures. In Section 4 we evaluate the robustness of centrality measures to small perturbations in classical and real-world network structures. The final section concludes.

2. CENTRALITY MODELS IN NETWORK STRUCTURES

Consider a graph $G = (V, E)$, where $V = \{1, \ldots, n\}$ is a set of nodes, $|V| = n$, and $E \subseteq V \times V$ is a set of links between nodes. The graph G is *undirected* if $\forall i, j \in V$: $(i, j) \in E \Longrightarrow (j, i) \in E$ and *directed* otherwise. The graph G can be represented by $n \times n$ adjacency matrix $A = [a_{ij}]$, where $a_{ij} = 1$ if $(i, j) \in E$, otherwise $a_{ij} = 0$. The graph G is *weighted* if it can be described by a non-negative weight matrix $W = [w_{ij}]$, where w_{ij} represents the intensity of a link between nodes i and j. Additionally, every node may be associated with a set of attributes $\{w_i^k\}$, where $i \in V$ and k is an attribute number, $k \in K$, and threshold value q_i, which denotes the level when node i becomes affected by other nodes. *Centrality score* of a node i is a certain numerical value $c_i \in \mathbb{R}^+$ that characterizes the i's importance in graph G.

Centrality is one of the key concepts in complex networks. There exist many measures to identify the most central elements in a network [1–3]. We provide a description of the centrality measures that we study in the paper.

2.1. Degree Centralities

The simplest centrality measure of node i in undirected graph is the degree [20], which is the number of links that are incident to node i:

$$c_i = \sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n} a_{ji}. \tag{1}$$

For directed graphs, there exist two variants of degree: in-degree and out-degree. The in-degree of a node is the number of in-coming links to this node, while the out-degree is the number of out-coming links from this node. Degree centrality can also be adapted to weighted networks (w_{ij} substitutes a_{ij}).

2.2. Spectral Centralities

Degree centralities are local measures that do not take into account indirect connections between nodes. The generalization of a degree centrality is the eigenvector centrality [20] that considers both direct and indirect connections between nodes. The eigenvector centrality $\vec{c}$ assigns the relative importance of node i as follows: more important neighbors should contribute more to the centrality of node i than less important neighbors. The calculation of the eigenvector centrality reduces to the eigenvector problem of the adjacency matrix A:

$$A\vec{c} = \lambda\vec{c}, \tag{2}$$

where λ is the principal eigenvalue of matrix A. In practice, the eigenvector centrality is used for undirected graphs because the estimation of the eigenvector for directed graphs may result in complex eigenvalues.

Another example of spectral centrality measure is PageRank [21], which evaluates the probability of visiting each node by the random walker:

$$c_i = \alpha \sum_{j=1}^{n} \frac{c_j}{\sum_{k=1}^{n} a_{jk}} a_{ji} + \frac{1-\alpha}{n}, \tag{3}$$

where α is a teleportation coefficient, $0 \leqslant \alpha \leqslant 1$. PageRank can be applied to undirected and directed graphs. Other centrality measures based on the eigenvector calculation also include Katz centrality [22], the HITS algorithm (hubs and authorities) [23], the subgraph centrality [24].

2.3. Centralities That Are Based on the Shortest Paths

Another class of centrality measures is based on the shortest paths between nodes. The most common centrality is a closeness centrality [20, 25] that estimates how close node i is located to other nodes in a graph:

$$c_i = \frac{1}{\sum_{j=1}^{n} d_{ij}}, \tag{4}$$

where d_{ij} – the length of the shortest path from node i to node j.

The betweenness centrality also considers the shortest paths between nodes [26, 27]. This measure reflects the importance of a node i as an intermediary between all pairs of other nodes:

$$c_i = \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}, \tag{5}$$

where σ_{jk} is the number of the shortest paths from node j to node k and $\sigma_{jk}(i)$ is the number of the shortest paths from j to k that include node i.

2.4. Short- and Long-Range Interaction Centralities

Many centrality measures do not consider individual attributes of nodes and/or the possibility of their group influence. Moreover, many existing methods do not assess the intensity of distant connections or take into account insignificant connections between nodes. The Long-Range Interaction Centrality (LRIC) is proposed in [3, 28] to account for these features.

Denote by N_j the set of neighbors of node j. The group of nodes $\Omega(j) \subseteq N_j$ is called *critical* for node j if $\sum_{i \in \Omega(j)} w_{ij} \geqslant q_j$. Node $l \in \Omega(j)$ is called *pivotal* if its exclusion from the critical group $\Omega(j)$ makes this group non-critical. All pivotal nodes in group $\Omega(j)$ is denoted by $\Omega_p(j)$. Then the influence of node i on node j is defined based on the critical group $i \in \Omega_p(j)$, in which node i has the maximum influence, i.e.

$$c_{ij} = \begin{cases} \max\limits_{\Omega_p(j): i \in \Omega_p(j)} \dfrac{w_{ij}}{\sum\limits_{k \in \Omega_p(j)} w_{kj}}, & \text{if } \exists \Omega_p(j)\text{: } i \in \Omega_p(j), \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Value c_{ij} is the element of an $n \times n$ matrix $C = [c_{ij}]$, which can be further used to determine indirect influence between nodes. The LRIC index has several modifications. In particular, there exists the Short-Range Interaction Centrality (SRIC) index [29] as well as the bundle index [30]. LRIC demonstrated its effectiveness for many real networks including foreign claims, international migration, world trade, global food network, the citation network of economic journals and the networks of interactions between terrorist groups [3, 31, 32].

3. PROPERTIES OF CENTRALITY INDICES: LITERATURE REVIEW

Several studies compare centrality measures with respect some axioms, that determine a set of logical properties that a given centrality measure should satisfy [5–11]. Section 3 discusses the axioms of centrality and provides a list of centrality indices for which these axioms have been studied.

3.1. Anonymity Axiom

The anonymity axiom is proposed in [5]. Anonymity axiom is satisfied if the centrality of node i only depends on its position in a network, i.e.

$$c_i(V,E) = c_{f(i)}(V, \{(f(i), f(j))\}\colon (i,j) \in E), \tag{7}$$

where $c_i(V,E)$ is the centrality score of node i in graph $G = (V,E)$ and f is a bijection on a set of nodes $f\colon V \to V$. In other words, the centrality of a node remains the same as the centrality of the corresponding node in the isomorphic graph. We note that all centrality indices that do not use nodes attributes satisfy the anonymity axiom.

3.2. Endpoint Increase Axiom

The endpoint increase axiom [5] is satisfied if the addition of a link between two nodes does not decrease their centralities, i.e. $\forall G,\ \forall i,j \in V$

$$\begin{cases} c_i(V, E \cup (i,j)) \geqslant c_i(V,E), \\ c_j(V, E \cup (i,j)) \geqslant c_j(V,E), \end{cases} \tag{8}$$

where $c_i(V, E \cup (i,j))$ is a centrality value of node i in graph G' which is formed by adding a link (i,j) to the graph $G = (V,E)$. Thus, any pair of non-adjacent nodes has an incentive to establish a connection between them.

3.3. Monotonicity Axiom

The monotonicity axiom [5] is satisfied if the addition of link (i,j) does not decrease the centrality of any node, i.e. $\forall G,\ \forall i,j,k \in V$

$$c_k(V, E \cup (i,j)) \geqslant c_k(V,E). \tag{9}$$

If a centrality measure involves the normalization at the last step, then the monotonicity axiom will most likely not be satisfied, as an increase in the centrality of one node leads to the decrease in the relative centrality of other nodes. Furthermore, if the endpoint increase axiom is not satisfied for a given centrality measure, then the monotonicity axiom also does not hold.

3.4. Top Node Axiom

According to the top node axiom [5], if a node i has the highest centrality in the initial graph G then adding an incident link will not change its position, i.e.

$$c_i(V,E) \geqslant c_k(V,E)\ \forall k \in V \implies \forall j,k \in V\ c_i(V, E \cup (i,j)) \geqslant c_k(V, E \cup (i,j)). \tag{10}$$

Thus, the top node axiom is not satisfied if the addition of incident links decreases the centrality of the most central node, which contradicts the intuitive concept of a centrality in a network.

3.5. Fairness Axiom

The fairness axiom [6] is satisfied if the addition of a link (i, j) changes the centrality of incident nodes by the same value, i.e. $\forall G,\ \forall i, j \in V$

$$c_i(V, E \cup (i, j)) - c_i(V, E) = c_j(V, E \cup (i, j)) - c_j(V, E). \tag{11}$$

Thus, the contribution of a new connection should be the same for each of the incident nodes.

3.6. Balanced Contributions Axiom

The balance of contributions axiom is proposed in [7] and is axiomatized in [9]. A centrality measure satisfies the axiom if removing all links E_i that are incident to node i affects the centrality of node j in the same way as removing all links E_j affects the centrality of node i, i.e. $\forall i, j \in V$

$$c_i(V, E) - c_i(V, E \backslash E_j) = c_j(V, E) - c_j(V, E \backslash E_i), \tag{12}$$

where E_i is a set of links that are incident to node i.

3.7. Add Edge Distance Axiom

The edge distance axiom is proposed in [9]. According to this axiom, the addition of a new link between nodes i and j, which are located at the same distance from node k, does not affect the centrality of node k, i.e. $\forall i, j, k \in V$: $d_{ik} = d_{jk}$

$$c_k(V, E) = c_k(V, E \cup (i, j)). \tag{13}$$

Thus, the addition of a link between nodes should not affect the centrality of other nodes that are at the same distance from these nodes.

3.8. Bridge Axiom

According to the bridge axiom [9], if a link (i, j) is the only connection between two disconnected components of the graph (i.e., (i, j) is a bridge), then the centrality of a node i should be higher than the centrality of a node j if the component that contains node i is larger than the component with node j.

3.9. Cut-Vertex Additivity Axiom

According to the cut-vertex additivity axiom [10], if two graphs G and G', which share only one common node i, are merged, then the centrality of node i equals the sum of its centrality in graphs G and G', i.e. $\forall G = (V, E),\ G' = (V', E')$: $V \cap V' = \{i\}$

$$c_i(V \cup V', E \cup E') = C_i(V, E) + C_i(V', E'). \tag{14}$$

Alternatively, this centrality can be formulated as follows:

$$\frac{1}{c_i(V \cup V', E \cup E')} = \frac{1}{c_i(V, E)} + \frac{1}{c_i(V', E')}. \tag{15}$$

The next two axioms are formulated in the context of specific graph structures.

3.10. Density Axiom

Consider the graph $G_{k,p}$, which consists of a complete graph with k nodes connected by a single link (i, j) to a cycle of length p $(p \geqslant 3)$. Suppose that node i is from the complete graph and node j is from the cycle. The density axiom [8] is satisfied if for $k = p$ the centrality of node i is greater than the centrality of node j.

3.11. Size Axiom

Consider the graph $G_{k,p}$, which consists of a complete graph with k nodes connected by a single edge (i, j) to a cycle of length p $(p \geqslant 3)$. The size axiom [8] is satisfied if for any k there exists a value $p < k$ such that the centrality of the nodes in the cycle is higher than the centrality of the nodes in the complete graph, and conversely, for any k there exists a value $p \geqslant k$ such that the centrality of the nodes in the cycle is lower than the centrality of the nodes in the complete graph.

Many axioms have already been studied for certain centrality models (see Table 1) [33]. More details about the properties of centrality measures from Section 2 are discussed in [8, 10, 11, 34]. We remark that even classical centrality measures have been studied only partially.

Table 1 shows that only degree centrality satisfies most axioms. This observation is due to the fact that degree centrality is a local measure (it considers only direct connections between nodes). Nevertheless, most of the axioms focus on the global structure of a network and assess the impact of indirect connections between nodes. Since changes in links do not affect the degree of distant nodes, degree centrality satisfies most axiomatic properties. Other centrality models do not satisfy most axioms. All measures, which do not consider individual attributes of nodes, satisfy the anonymity axiom. The endpoint increase and monotonicity axioms are only satisfied by closeness and Katz centralities while the density axiom is satisfied by eigenvector, Katz, PageRank and closeness centralities.

The presented list of axioms is not exhaustive. Other axioms can be also found in [35–38]. We note that many axioms from other related scientific fields could be adjusted to centrality metrics. For instance, consider the *Concordance axiom* for the choice functions [39]. According to this axiom, an alternative chosen in two subsets should also be chosen on their union. In choice theory, this condition characterizes the rational choice of the set of alternatives. Consequently, we can

Table 1. Properties of centrality measures ("+" – the axiom is satisfied, "−" – the axiom is not satisfied, "?" – the axiom have not been studied in the literature yet)

No.	Models of centrality	Anonymity	Endpoint Increase	Monotonicity	Top node	Fairness	Balanced Contribution	Add Edge Distance	Bridge	Cut-vertex	Density	Size
1	Degree centralities	+	+	+	+	+	+	+	−	+	+	−
2	Eigenvector	+	−	−	−	−	−	−	−	−	+	−
3	PageRank	+	?	−	?	−	−	−	−	−	+	−
4	Katz centrality	+	+	+	?	−	−	−	−	−	+	−
5	HITS	+	?	?	?	−	−	?	?	?	?	?
6	Subgraph centrality	+	?	?	?	−	−	?	?	?	?	?
7	Closeness centrality	+	+	+	−	−	−	+	+	−	+	−
8	Betweenness centrality	+	−	−	−	−	−	−	−	−	−	−
9	LRIC	−	?	−	?	−	−	?	?	?	?	?

extend the cut-vertex additivity axiom by assuming that if two graphs G and G' share a common node i, which has the highest centrality score in G and G', then node i should also be the most central node in their union.

We would like to emphasize that axioms of centrality helps to understand the properties and the behaviour of the centrality measure. However, if a specific centrality model does not satisfy certain axioms, we cannot conclude that this centrality measure is defective.

4. ROBUSTNESS OF CENTRALITY MEASURES

The robustness of centrality measures is one of the most important criteria for selecting an appropriate centrality model in the context of real-world systems. Unfortunately, many existing networks are incomplete or contain errors in data. For instance, criminal networks, which are discussed in [40, 41], are incomplete (due to the nature of the network), contain errors (unintentional data collection errors and intentional deception by criminals) and inconsistent information (misleading information from different sources). The network of foreign bank claims, which is analyzed in [31], covers about 94% of total foreign claims, as some countries do not report their statistics. Meshcheryakova [42] investigated the global food trade network under asymmetry that occurs due to different commodity classification systems, trade volume assessment and time lag. The results of centrality measures in these networks require a careful examination because many measures may be sensitive to small changes in the graph structure.

4.1. Robustness of Centrality Measures in Classical Graph Structures

We evaluate the sensitivity of centrality measures to link removal or addition to undirected unweighted networks. We perform the experiments on classical graph structures with $n = 100$ nodes (see Fig. 1):

1. Erdos-Renyi graph (ER graph): a random graph model $G(n, p)$ where n in a number of nodes and p is a probability for edge creation.

2. Barabasi-Albert graph (BA graph): a random graph model where nodes are sequentially connected to m existing nodes with a probability that is proportional to the degree of the existing nodes.

3. Small-world (Watts-Strogatz) graph model: a random graph model where a regular ring graph n vertices is created and then some links are randomly rewired with probability p.

4. Square lattice: a graph with n nodes whose nodes are the points in the Cartesian plane. Nodes are connected if the distance between them is 1.

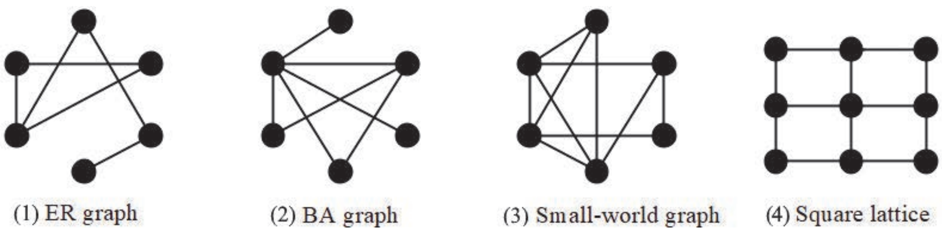

Fig. 1. Classical graph structures.

The computational experiments are performed on classical graph structures. First, $N = 100$ graphs of a given structure are generated ($N = 1$ graph for a square lattice). We consider the parameter $p = 0.05$ for random graph models and $m = 4$ for BA graphs[1]. For each graph, we

[1] The authors also considered other values of p (from 0.01 to 0.1 with a step of 0.01) and m (from 2 to 5). We observe that the relative position of the centrality measures in terms of their robustness does not change significantly for other parameter values.

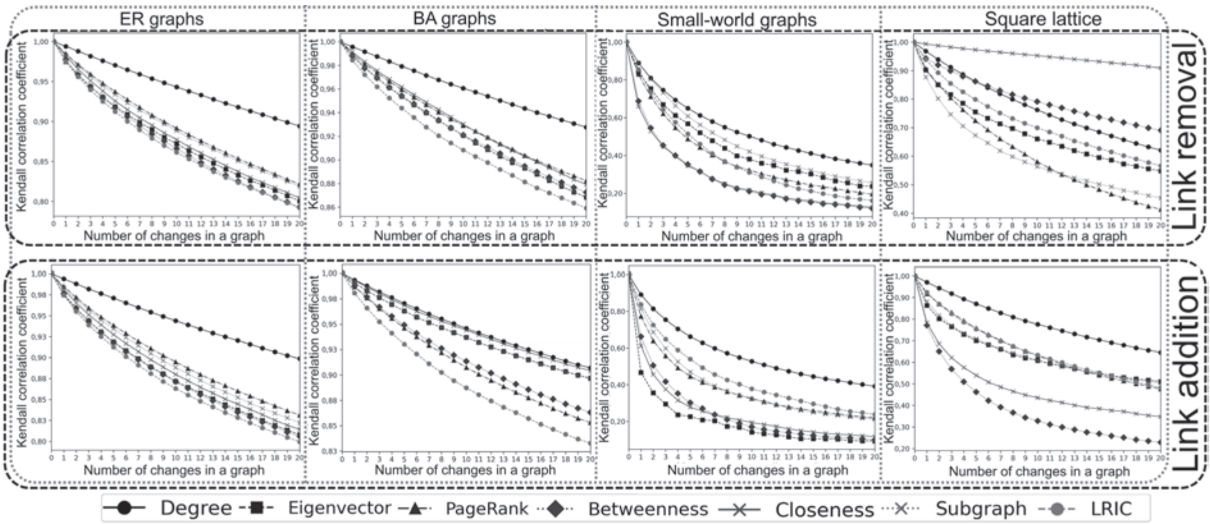

Fig. 2. Stability of centrality measures on classical graph structures.

examine $K = 100$ ways ($K = 10\,000$ for a square lattice) to change the graph structure. Next, we assess the centrality of the nodes in original and modified graphs and compare the centrality scores with respect to the Kendall correlation coefficient. The sensitivity of the robustness of a centrality measure is defined as the arithmetic mean across all 10 000 experiments. We remark that the increase in the number of experiments does not change the results significantly.

The computational experiments are carried out for the following centrality models: degree centrality, eigenvector centrality, PageRank, closeness centrality, betweenness centrality and LRIC. The results are provided in Fig. 2.

Figure 2 shows that the centrality measures are very robust in ER and BA graphs: the Kendall correlation coefficient remains high (> 0.8) after deleting/adding 20 links. However, small perturbations in small-world graphs and a square lattice dramatically change the centrality scores of the nodes. Thus, the centrality measures on these networks require a careful examination as most of them are very unstable. The exceptions are the degree centrality (link addition) and the closeness centrality (link removal in the square lattice graph).

Among the centrality measures, which consider the global structure of the network and have a low sensitivity to graph perturbations, we remark PageRank and subgraph centrality (ER, BA and small-world graphs) as well as LRIC (square lattice). The eigenvector and centralities, which are based on the shortest paths, are the most susceptible to random changes in most network structures. In general, the choice of the most appropriate centrality model depends on the network structure under consideration.

4.2. Robustness of Centrality Measures in Real Networks

In many applications, the network data is complete. Consequently, the need to assess the reliability and representativeness of the results obtained occurs in real networks. This section discusses the sensitivity of centrality models in criminal and international trade networks, where the structure is only partially observed.

Criminal network

The criminal network describes Sicilian Mafia meeting interconnections, which is derived from Court reports in 2007 based on the results of the anti-mafia "Montagna Operation." The mafia meetings network is an undirected graph with 101 nodes and 256 links [41]. We assume that

Table 2. Robustness of centrality measures in criminal network

Models of centrality	Scenario 1		Scenario 2	
	Correlation	TOP-5	Correlation	TOP-5
Degree centrality	0.91	0.98	0.92	0.98
Eigenvector centrality	0.89	0.85	0.89	0.84
PageRank	0.78	0.94	0.85	0.93
Katz centrality	0.91	0.95	0.9	0.94
Subgraph centrality	0.92	0.91	0.92	0.88
Closeness centrality	0.81	0.88	0.83	0.87
Betweenness centrality	0.66	0.84	0.71	0.85
LRIC	0.74	0.8	0.79	0.81

Table 3. Robustness of centrality measures in global food trade network

Models of centrality	Scenrio 1		Scenrio 2	
	Correlation	TOP-10	Correlation	TOP-10
In-degree centrality (total import)	0.99	0.99	0.38	0.22
Out-degree centrality (total export)	0.99	1	0.79	0.92
Degree difference (net export)	0.99	1	0.14	0.81
Hubs	0.99	1	0.78	0.72
Authorities	0.99	1	0.26	0.22
PageRank	0.99	1	0.43	0.26
LRIC	0.99	0.98	0.71	0.5

approximately 10% of the total number of recorded connections are missing in the network and consider two strategies for adding new links: random addition (scenario 1) and the addition of new links with a probability that is proportional to the distance between nodes (scenario 2). The sensitivity of centrality measures is assessed using the average Kendall correlation coefficient between node rankings and the average Jaccard coefficient between the sets of TOP-5 nodes in original and modified networks. The results are depicted in Table 2.

Table 2 demonstrates that random changes in the graph structure (scenario 1) have a greater impact on the ranking of nodes compared to changes based on the distance between nodes (scenario 2). The degree centrality shows the most robust results, which is consistent with the findings in Section 4.1. Katz centrality and subgraph centrality, which are global measures, also provide a robust ranking of nodes. In contrast, the results of LRIC and shortest path-based centralities are very sensitive to missing links in a criminal network.

Global food trade network

The global food trade network represents trade flows between 222 countries in 2020 [43]. Since the network is directed and weighted (the largest trade flow is between the USA and Canada – about 5.5 billion dollars), we focus on centrality measures that are usually applied to trade networks. We consider the following strategies to change the graph structure:

(1) Scenario 1: modifying the recorded trade flows by up to 5% to account for possible errors in estimating trade flows;

(2) Scenario 2: adding links with a total value of 1% of the total world trade to account for missing flows due to the lack of reporting by some countries.

The sensitivity of centrality measures is assessed using the average Kendall correlation coefficient between node rankings and the average Jaccard coefficient between the sets of TOP-10 countries in original and modified networks. The results of the performed analysis are provided in Table 3.

Table 3 shows that errors in trade flows within 5% (scenario 1) have virtually no effect on the ranking of countries by all centrality indices. However, the lack of trade flow data between some countries (scenario 2) can significantly affect the ranking of countries. In particular, the highest correlation coefficient (> 0.71) is observed for out-degree centrality, LRIC, and hubs (part of the HITS algorithm). The low correlation coefficient can be explained by significant changes in the ranking of countries that were not the most important in the original network. Nevertheless, the set of the most influential countries (TOP-10) according to out-degree centrality, degree difference and hubs remains quite stable.

5. DISCUSSION

The concept of centrality in networks lacks a unique definition, leading to the development of many different centrality models. In this regard, selecting the most appropriate centrality measure often involves using an axiomatic approach and analyzing the sensitivity of models to changes in network structure.

Our study provided an overview of the existing axioms of centrality and current results on verifying these axioms for classical centrality measures. Many properties of centrality measures remain unexplored. Hence, we urge other researchers to examine the axiomatic properties of these measures. Studying the axioms could also be valuable for the development of new centrality models.

We also compared the sensitivity of centrality measures to small perturbations in a network structure. As most existing networks change over time and their structure is only partially observed, the question arises about the representativeness and reliability of the resulting set of central elements. We demonstrated the robustness of classical centrality models to changes in the structure of both artificial and real networks. The results of the models that are highly sensitive to changes in network structure require a careful examination, especially for partially observed networks.

Our study expands the theoretical knowledge of the concept of centrality in network structures. Our research findings can help choose the most suitable centrality index, especially in cases where the observed network is incomplete or inaccurate. Finally, the results of the study can be utilized to develop new centrality measures.

FUNDING

This work was supported by grant for state support of young Russian scientists – candidates of sciences (grant no. MK-3867.2022.1.6.)

REFERENCES

1. Newman, M.E.J., *Networks: Second Edition*, Oxford: Oxford University Press, 2018.
2. Centiserver: The most comprehensive centrality resource and web application for centrality measures calculation. [online] available at: https://www.centiserver.org/centrality/list/ [accessed 7.12.2023].
3. Aleskerov, F.T., Shvydun, S., and Meshcheryakova, N., *New Centrality Measures in Networks: How to Take into Account the Parameters of the Nodes and Group Influence of Nodes to Nodes*, Boca Raton: CRC Press, 2022.
4. Shvydun, S., The Impact of COVID-19 on the Air Transportation Network, *Adv. Intell. Syst. Comput.*, 2023, vol. 1435, pp. 94–107.
5. Sabidussi, G., The Centrality Index of a Graph, *Psychometrika*, 1966, vol. 31, no. 4, pp. 581–603.
6. Myerson, R.B., Graphs and Cooperation in Games, *Math. Oper. Res.*, 1977, vol. 2, pp. 225–229.
7. Myerson, R.B., Conference Structures and Fair Allocation Rules, *Int. J. Game Theory*, 1980, vol. 9, pp. 169–182.

8. Boldi, P. and Vigna, S., Axioms for Centrality, *Internet Math.*, 2014, vol. 10, pp. 222–262.

9. Skibski, O., Michalak, T.P., and Rahwan, T., Axiomatic Characterization of Game-Theoretic Centrality, *J. Artif. Int. Res.*, 2018, vol. 62, no. 1, pp. 33–68.

10. Skibski, O. and Sosnowska, J., Axioms for Distance-Based Centralities, *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1218–1225.

11. Chebotarev, P., Selection of Centrality Measures using Self-Consistency and Bridge Axioms, *J. Complex Networks*, 2023, vol. 11, no. 5, pp. 1–23.

12. Khitraya, V.A. and Mazalov, V.V., Teoretiko-igrovaya Tsentralnost Vershin Orientirovannogo Grafa (Game Theoretic Centrality of a Directed Graph Vertices), *Math. Game Theory Appl.*, 2013, vol. 15, no. 3, pp. 64–87.

13. Shcherbakova, N.G., Aksiomatika Tsentralnosti v Kompleksnykh Setyakh (Axioms of Centrality in Complex Networks), *Problems Inform.*, 2015, vol. 3, no. 28, pp. 3–14.

14. Borgatti, S., Carley, K., and Krackhardt, D., On the Robustness of Centrality Measures under Conditions of Imperfect Data, *Soc. Networks*, 2006, vol. 28, no. 2, pp. 124–136.

15. Frantz, T.L., Cataldo, M., and Carley, K.M., Robustness of Centrality Measures under Uncertainty: Examining the Role of Network Topology, *Comput. Math. Organ. Theor.*, 2009, vol. 15, no. 4, pp. 303–328.

16. Segarra, S. and Ribeiro, A., Stability and Continuity of Centrality Measures in Weighted Graphs, *IEEE Trans. Signal Process.*, 2016, vol. 64, no. 3, pp. 543–555.

17. Martin, C. and Niemeyer, P., Influence of Measurement Errors on Networks: Estimating the Robustness of Centrality Measures, *Network Sci.*, 2019, vol. 7, no. 2, pp. 180–195.

18. Murai, S. and Yoshida, Y., Sensitivity Analysis of Centralities on Unweighted Networks, *WWW'19*, 2019, pp. 1332–1342.

19. Meshcheryakova, N. and Shvydun, S., Perturbation Analysis of Centrality Measures, *2023 IEEE/ACM International Conf. on Advances in Social Networks Analysis and Mining (ASONAM)*, 2023, pp. 407–414.

20. Freeman, L.C., Centrality in Social Networks: Conceptual Clarification, *Soc. Networks*, 1979, vol. 1, no. 3, pp. 215–239.

21. Brin, S. and Page, L., The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Comput. Netw.*, 1998, vol. 30, pp. 107–117.

22. Katz, L., A New Status Index Derived from Sociometric Index, *Psychometrika*, 1953, vol. 18, pp. 39–43.

23. Kleinberg, J.M., Authoritative Sources in a Hyperlinked Environment, *J. ACM*, 1999, vol. 46, no. 5, pp. 604–632.

24. Estrada, E. and Rodriguez-Velazquez, J.A., Subgraph Centrality in Complex Networks, *Phys. Rev. E.*, 2005, vol. 71, no. 5, pp. 1–9.

25. Bavelas, A., Communication Patterns in Task–Oriented Groups, *J. Acoust. Soc. Amer.*, 1950, vol. 22, no. 6, pp. 725–730.

26. Anthonisse, J.M., *The Rush in a Directed Graph*, Technical Report BN 9/71, Stichting Mathematisch Centrum, Amsterdam, 1971.

27. Freeman, L.C., A Set of Measures of Centrality based upon Betweenness, *Sociometry*, 1977, vol. 40, no. 1, pp. 35–41.

28. Aleskerov, F.T., Meshcheryakova, N.G., and Shvydun, S.V., Power in Network Structures, *Springer Proc. Math. Stat.*, 2017, vol. 197, pp. 79–85.

29. Aleskerov, F.T., Andrievskaya, I.K., and Permjakova, E.E., *Key Borrowers Detected by the Intensities of their Short-Range Interactions*, Working papers by NRU Higher School of Economics. Series FE "Financial Economics," 2014, no. WP BRP 33/FE/2014.

30. Aleskerov, F. and Yakuba, V., *Matrix-vector Approach to Construct Generalized Centrality Indices in Networks*, NRU Higher School of Economics, Series WP7 "Mathematical methods for decision making in economics, business and politics," 2020, no. 2323.

31. Aleskerov, F., Andrievskaya, I., Nikitina, A., and Shvydun, S., Key Borrowers Detected by the Intensities of Their Interactions, *Handbook of Financial Econometrics, Mathematics, Statistics, and Machine Learning (In 4 Volumes), World Scientific*, 2020, pp. 355–389.

32. Aleskerov, F., Gavrilenkova, I., Shvydun, S., and Yakuba, V., Power Distribution in the Networks of Terrorist Groups: 2001–2018, *Group. Decis. Negot.*, 2020, vol. 29, pp. 399–424.

33. Centrality Measures: from Theory to Applications. [online] available at: https://centrality.mimuw.edu.pl/ [accessed 7.12.2023].

34. Skibski, O., Vitality Indices are Equivalent to Induced Game-Theoretic Centralities, *IJCAI-21*, 2021, pp. 398–404.

35. Nieminen, U.J., On the Centrality in a Directed Graph, *Soc. Sci. Res.*, 1973, vol. 2, no. 4, pp. 371–378.

36. Garg, M., Axiomatic Foundations of Centrality in Networks, *Soc. Sci. Res. Network*, 2009, no. 1372441.

37. Landherr, A., Friedl, B., and Heidemann, J., A Critical Review of Centrality Measures in Social Networks, *Bus. Inf. Syst. Eng.*, 2010, vol. 2, pp. 371–385.

38. Bloch, F., Jackson, M.O., and Tebaldi, P., Centrality Measures in Networks, *Soc. Choice Welf.*, 2023, vol. 61, pp. 413–453.

39. Aizerman, M.A. and Aleskerov, F.T., *Theory of Choice*, Amsterdam: North-Holland. 1995.

40. Ficara, A., Cavallaro, L., Curreri, F., Fiumara, G., De Meo, P., Bagdasar, O., Song, W., and Liotta, A., Criminal Networks Analysis in Missing Data Scenarios through Graph Distances, *PLOS ONE*, 2021, vol. 16, no. 8, e0255067.

41. Ficara, A., Cavallaro, L., De Meo, P., Fiumara, G., Catanese, S., Bagdasar, O., and Liotta, A., Social Network Analysis of Sicilian Mafia Interconnections, *Studies Comput. Int.*, 2020, vol. 882, pp. 440–451.

42. Meshcheryakova, N., Network Analysis of Bilateral Trade Data Under Asymmetry, *2020 IEEE/ACM International Conf. on Advances in Social Networks Analysis and Mining (ASONAM)*, 2020, pp. 379–383.

43. The World Integrated Trade Solution. [online] available at: https://wits.worldbank.org/ [accessed 7.12.2023].

This paper was recommended for publication by P. Chebotarev, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 8, pp. 785–801.

The Study of the Strategic Consequences of a Scoring Model Disclosure

G. M. Kryukov[*,a] **and M. S. Sandomirskaia**[**,b]

[*] *New Economic School, Moscow, Russia*
[**] *HSE University, Moscow, St. Petersburg, Russia*
e-mail: [a] *gkryukov@nes.ru,* [b] *msandomirskaya@hse.ru*

Received January 25, 2024
Revised July 1, 2024
Accepted July 10, 2024

Abstract—In this paper, the disclosure of information about the scoring model is investigated. Some of the company's customers find out their internal rating in the company. Such customers can change their behavior to increase their internal rating. The customers who are aware of the leakage are represented as players who can choose a strategy: whether to increase their internal rating and, if so, how much. The main goal is to find the Bayesian-Nash equilibrium in this game and find out how it depends on various parameters, such as the scale of the leakage, the distribution of ratings.

Keywords: scoring model, Bayesian game, manipulation, information disclosure

DOI: 10.31857/S0005117924080046

1. INTRODUCTION

Machine learning methods are being implemented by companies around the world to solve business problems. In particular, companies often use Data Science methods to calculate internal user parameters (for example, the probability of repaying a loan for a bank or attractiveness for an online dating service). A scoring model assigns a certain rating (score) to each user. This rating can be either a discrete random variable (for example, dividing users into several clusters) or continuous (a neural network often returns real values from the interval [0,1]). These models can be useful for solving a wide variety of business problems, such as: "select m users with the greatest propensity for spontaneous purchases to participate in a promotion" or "select m users from a cluster the company needs and offer them a trial version of a new product."

The problem is that information leakages happen from time to time. As a result of such leakages, users can find out what their internal rating is or what internal cluster they belong to. For example, in 2016, the internal rating of users in the dating app Tinder leaked [1]. In addition to the leaked final rating, users can find out how exactly the machine learning model works, as well as the ratings/cluster distribution of other users. Then, users may be motivated to cheat the algorithm: users can change their behavior so that the machine learning algorithm improves their internal rating or assigns them to a better cluster. However, changing behavior and other characteristics related to the algorithm is not free of charge: the user needs to spend time and sometimes money to change their rating (e.g., visit certain pages on the site, apply for a certain product). In addition, the costs of changing the algorithm's assessment may vary for different users. Accordingly, each user decides individually whether to change their behavior to improve their internal rating or leave everything as is.

This paper addresses the question of how users will behave when information about the company's internal user ranking leaks. How does their behavior depend on the values of the scoring model parameters? To do this, a market is modeled where the company wants to select m clients from n for a certain activity (issuing loans, participating in a profitable promotion, testing a new product, etc.). It is assumed that all clients value this activity equally and that taking part in it is better than not taking. Some users receive information about how the machine learning algorithm works, which ranks/distributes clients into clusters. These users are faced with a strategic choice. On the one hand, they can incur certain costs in order to increase their rating or move to a better cluster, thereby increasing their chances of getting into the m clients who will receive a certain bonus. On the other hand, they can consider their current internal rating as acceptable and not change anything. Thus, clients who have gained access to internal information about the company's machine learning models become players who choose the optimal strategy for themselves. This situation is represented as a Bayesian game, and its equilibrium is further investigated. The relationship between strategic consequences and the type of cost distribution function can be used to analyze the robustness of the scoring model to data leakages. Different scoring models can include parameters that are more or less susceptible to manipulation by an informed client, so that a firm, when choosing a specific scoring model, can predict in advance the potential scale of scoring errors in the case of small or large leakages. This can improve the security and robustness of the scoring models used.

1.1. About Scoring Models in Literature

The focus of this paper is on a machine learning model that scores customers. Customer scoring models are widely used in the industry and are actively evolving. The canonical example is credit scoring models that assess the creditworthiness of customers. Data science methods are actively used to assess credit ratings, including interpretable deep learning [2] and genetic hierarchical networks [3]. The authors in [4] found that after implementing the advanced methods mentioned above, in which customers are not aware of the detailed mechanism of the scoring model, they become less sophisticated in the steps they should take to improve their credit rating. This paper suggests that if the information about the credit scoring model is leaked, customers will obtain a better understanding of how to influence their credit rating, and therefore customers may be willing to manipulate their rating.

It should be noted that companies can evaluate users not only for credit scoring. Advanced machine learning methods are actively used to predict whether a user will perform a certain action, such as buying a certain product [5]. This data are also used to make business decisions about promotion strategies and distribution of personal discounts. Therefore, information about scoring models can also be used by customers to gain additional benefits.

It is worth noting that machine learning algorithms are also used to cluster users into categories [6]. In this case, users also have incentives to manipulate the model when it is disclosed in order to get into a more attractive cluster for themselves.

1.2. About the Game-Theoretic Context in the Model

To model the situation of leakage of the information about the scoring model, a game-theoretic approach will be used. Its effectiveness for studying empirical problems is noted, for example, in [7]. The game component arises from the fact that clients, i.e. players in the model, compete for a finite amount of the good, so that an increase in the rating of an individual client reduces the chances of being selected for others. Having an idea of the scoring model mechanism, the client is able to calculate her own rating, but the exact rating of other clients is unknown to her, and she can only rely on some generally known distribution of ratings in the population. Also, clients act independently and do not observe the actions taken by others, so their interaction can be considered as simultaneous. Thus, a Bayesian game [8] is constructed.

It is worth noting that the process of strategic manipulation is elaborated in game-theoretic models, including optimal control models. For example, [9–12] study optimal control mechanisms in systems with active elements. The most complete review of their ideas is presented in [13]. This direction elaborates the problem of mechanism design, in which the key idea is to disclose information about the type of agents and adequately take into account the agent's behavior in the objective function of the center. Most of the works are devoted to the manipulation problem and to the identification of preference classes for which the control procedure is resistant to a manipulation. Manipulation is understood as the distortion of agent declared preferences, and this distortion does not usually require explicit costs from agents. In this paper, the attention is focused on the associated costs of rating distortion, which is more typical, for example, for works on reputation manipulation [14, 15]. Another important feature of this work is that as a result of the leakage, a certain random subsample of agents becomes active, but not all agents, and inside this subsample, due to heterogeneity in costs, there will be those who prefer to manipulate, and those who prefer to leave their true characteristics. Thus, even under an exogenous narrowing of the set of active agents, the problem of manipulability does not disappear, and its analysis requires joint consideration of the behavior of both active agents and those who are formally present in the system with fixed characteristics.

Since the goal of the scoring model is to select a limited set of customers, the analysis will yield structures typical of works in the field of finite markets [16]. [17] studies the price competition in markets with a rare product and private information about buyers' valuation of the product. The authors consider a market with two sellers, each has one unit of an identical product. Sellers simultaneously choose prices, then buyers choose which seller to go to for the product or not to anyone. The conditions derived when solving the buyer's problem are similar to those that arise in this work when analyzing a game with a single candidate selected by the scoring model (the winner). For a model with several winners, the formulas with a binomial coefficient obtained here are conceptually similar to the results in [18] on Bertrand oligopoly with constraints on the production capacities of firms.

2. MODEL

Let's formalize a game model in which users who have received information about the scoring algorithm decide whether to use the received information to increase their rating.

Let a firm produces a certain product (e.g., a bank that issues credit cards). Let there be n customers who apply for the product (n customers who want to get credit cards). The firm has a limit, e.g., on the amount of plastic for cards, so they are willing to sell only $m < n$ units of the product. Assume that all customers value the product equally. Owning the product brings utility 1.

The firm decides to use a scoring model to classify users. In this article, it is assumed that the scoring model classifies customers into only two categories: good (one) and bad (zero). The scoring model's mechanism is based on some machine learning methods and is essentially a "black box" due to the complexity of its work, but the firm knows the input parameters of this box and, if necessary, can estimate their weight in the final classification result by running the model on a sufficiently wide sample of its customers whose characteristics are known to the firm. Also, using this sample, the firm can approximate the distribution of customer costs to change some of their individual parameters important for the scoring model to values sufficient to be classified as "good."

The game considers a situation where a leakage of information about a scoring model happens. The user who has access to this information learns what class the model assigns her to and what parameters it uses. She also learns the distribution of the rating on the set of all clients (there are many clients, the rating is determined for everyone, not just for n who apply for a credit card or another reward). However, the user does not know for sure what characteristics other users who learn about the leakage have, i.e. the strategic user makes further decisions under the incomplete information.

Let a customer is classified as good by the scoring algorithm with probability p. The customer finds the disclosed data with probability α. If the customer learns that the model classifies her as a good type, she has no incentive to change her usual behavior. In other words, changing the behavior so that the algorithm classifies her as bad is a weakly dominated strategy and therefore will not be used in the further analysis.

If a customer learns that the algorithm classifies her as a bad type, she has a strategy to change her behavior, bear the costs c_i and mimic a good one. We assume that c_i can be different for different players within the support of the distribution function, each player knows her own value of costs c_i. Indeed, someone needs to do very little to fulfill the conditions of the algorithm for a "good consumer," while other needs to change their behavior greatly and, accordingly, bear large costs. Let the distribution function of the costs of "mimicry" for users whom the model classifies as a bad type be equal to $F(x) = P[c_i \leqslant x]$; for consistency with the normalization of utility to 1, we also assume that the distribution function of costs is defined on $[0; 1]$. If some customers had costs of mimicry higher than the utility from owning the firm's product, then they do not mimic with certainty and can be excluded from consideration. Suppose that $F(x)$ is known to all strategic players. It is appropriate to consider the value of costs as a relative value, i.e. the share of utility from owning the product. Then all further calculations in the model are linearly scaled by the required size of the "prize." As a possible example of a scoring model parameter, whose costs of changing are high for most clients, we can consider the presence and size of a mortgage loan for a borrower (in this example it is easy to see that the costs may exceed the size of the "prize," so such a parameter in the model is unlikely to be subject to distortion). On the contrary, an example of a parameter whose distortion is low-cost is the presence of a completely filled out open profile in a social network.

Note that the detailed knowledge of the internal scoring mechanism is not important for the current analysis. It is sufficient to know only the probability p and the function $F(x)$, which relate both to the scoring mechanism and to the characteristics of the customer population to which scoring is to be applied.

With probability $1 - \alpha$, the customer knows nothing about the leakage. She is not a strategic player, and the information leakage does not affect her internal rating and her behavior. Thus, the only strategic players in this game are agents who are aware of the leakage, which are assigned to the bad class by the scoring model. In this case, the Bayesian type of player i is her costs c_i.

The payoff of a strategic agent from the strategy "NOT to mimic" is equal to the probability of getting the product if the scoring model classifies this player as bad. This is possible only in the situation where there are more products than good customers determined by the algorithm, so that the remaining ones from the bad category must be randomly selected. The payoff of a strategic agent from the strategy "to mimic" is equal to the difference between the probability of getting the product if the model classifies the player as good and the costs of the mimicry. Note that all strategies are chosen by the players simultaneously and independently.

The last issue is how the firm determines exactly m winners. Let k agents be classified by the scoring model as good (they can be either initially good or strategic agents who decided to change their data and cheat the model). If $k > m$, the winners are random m out of k good clients according to the model. If $k \leqslant m$, all good clients according to the model are winners, and $m - k$ winners are additionally randomly selected from the bad ones.

Thus, in the resulting simultaneous game, each client who learned about the leakage and was classified as bad makes a binary choice and decides whether she is ready to invest in improving the rating in accordance with her costs, or leave everything as is. When a "bad" client, in accordance with the original work of the scoring algorithm, decides to mimic a "good" client, this distorts the results of the model classification and can lead to an incorrect choice of the winner by the firm. The problem is to find the equilibrium strategies for the agents who learned about the leakage, i.e. to determine for each client the optimal choice of the strategy "to mimic" or not, depending on the number of "prizes," the selectivity of the scoring model p, the parameters it uses and, as a result, the distribution of the costs of mimicry of these parameters in the population, the scale of the leakage, and the value of her own costs of improving her type, which the client knows precisely. This choice is not obvious, since an individual client cannot rely solely on its own strategic activity, but must correctly predict the behavior of other active agents that will affect the final probability of being selected. In other words, being among the winners is probabilistic, while the costs of mimicry are deterministic and non-refundable.

By examining the equilibrium strategy of all clients, the firm can predict what proportion of informed clients will decide to mimic and, consequently, how this will distort the scoring results. Since a scoring model can include various parameters, the cost of manipulating which may be different for users, then under the same scale of data leakage, different models may show a greater or lesser degree of the distortion. Accordingly, at the stage of choosing a scoring model, the firm can take into account the further risks from such leakages.

3. NASH EQUILIBRIUM IN A SINGLE-WINNER GAME

Let's consider a special case of the model at $m = 1$. For example, a firm can choose one client who will be the face of a new product. Some of the clients (bad according to the algorithm) will definitely not suit the firm. Among the good ones, the firm can choose arbitrary, since everyone is suitable enough.

We are looking for a symmetric Bayes–Nash equilibrium. Let's denote by y the probability that a bad client will mimic a good one in equilibrium. Note that this probability in the future will consist of the fact that agents of some costs types will deterministically mimic, while other types will not. However, since an individual player does not know the actual type of other agents (for her, it is a random variable), their behavior will also look like random, or probabilistic. Define q as the a posteriori probability that the player will be classified by the algorithm as a good type, taking into account the strategic mimicry of the proportion of consumers who learned about the leakage, then

$$q = p + (1 - p)\alpha y.$$

If the player does not mimic, then they have a chance to win only if everyone else also turns out to be bad *ex-post*. The expected payoff from the "do not mimic" strategy is given by:

$$u_i(0) = \frac{1}{n}(1 - q)^{n-1}.$$

Let's conduct the formula for the expected payoff when choosing the "mimic" strategy. Regardless of the outcome, the player will have to pay c_i. If exactly k of competitors are classified as good, the probability of winning is $\frac{1}{k+1}$. The random variable "the number of competitors to be classified as good" has a binomial distribution: $\mathrm{Bin}(n - 1, q)$. In total, the expected gain from the "mimic" strategy is expressed by the formula

$$u_i(1) = -c_i + \sum_{k=0}^{n-1} \frac{1}{k+1} \binom{n-1}{k} q^k (1 - q)^{n-1-k}.$$

Let's use the following identity proved in [17]:

$$\sum_{k=0}^{n-1} \frac{1}{k+1} \binom{n-1}{k} q^k (1-q)^{n-1-k} = \frac{1-(1-q)^n}{nq}.$$

The expected payoff from the "mimic" strategy is rewritten as follows:

$$u_i(1) = -c_i + \frac{1-(1-q)^n}{nq}.$$

The player will choose the one of the two strategies that will bring her the greatest expected payoff. Note that the expected payoff from the "do not mimic" strategy does not depend on c_i, while the expected payoff from the "mimic" strategy decreases with an increase in c_i. This means that the optimal strategy of player i is monotonic (threshold). In other words, there is such a threshold value of costs c^* that for all $c_i < c^*$ the player i mimics and for all $c_i > c^*$ i the player does NOT mimic. With $c = c^*$, the expected payoffs from both strategies are the same.

Then the probability that the bad type will mimic the good one in equilibrium is equal to the probability that the costs will be lower than the threshold level, and is expressed in terms of c^* as

$$y = P[c_i \leqslant c^*] = F(c^*).$$

The condition for c^* is determined from the equality of expected payoffs from the two pure strategies:

$$c^* = -\frac{1}{n}(1-q)^{n-1} + \frac{1-(1-q)^n}{nq}.$$
$$c^* = \frac{1-(1-q)^{n-1}}{nq}.$$

It is important to clarify that this formula is an implicit expression, because q depends on y, which is uniquely defined through c^*:

$$q(c^*) = p + (1-p)\alpha y = p + (1-p)\alpha F(c^*).$$

Let's introduce the function $f(q)$:

$$f(q) = \frac{1-(1-q)^{n-1}}{nq}.$$

Then the condition for c^* is written as $c^* = f(q(c^*))$.

Theorem 1. *If the costs distribution function $F(c)$ is continuous, then the threshold value c^* exists and is the only one in the optimal monotonic strategy.*

Proof. First, note that $f(q)$ decreases monotonously with the growth of q at $n > 2$ and is condtant at $n = 2$. Let's prove it. Consider the partial derivative

$$\frac{\partial f}{\partial q} = \frac{-1 + (1-q)^{n-2}((n-2)q+1)}{nq^2}.$$

The denominator is positive. Consider the numerator. Note that for $n = 2$ the numerator is zero.

The following inequality holds:

$$(1-q)^n(nq+1) > (1-q)^{n+1}((n+1)q+1).$$

Indeed,

$$(1-q)^n(nq+1) - (1-q)^{n+1}((n+1)q+1) = (n+1)q^2(1-q)^n > 0.$$

The monotonicity follows from this statement: the numerator is $-1+(1-q)^{n-2}((n-2)q+1)$ and decreases by n, while it is equal to 0 for $n=2$. It follows that $-1+(1-q)^{n-2}((n-2)q+1) < 0$ for $n > 2$, and from this the partial derivative is negative.

$$\begin{cases} \dfrac{\partial f}{\partial q} = 0, & \text{if } n = 2, \\ \dfrac{\partial f}{\partial q} < 0, & \text{if } n > 2. \end{cases}$$

Now let's prove that $\frac{1}{n} \leqslant c^* \leqslant \frac{n-1}{n}$. To do this, it is enough to show that $\frac{1}{n} \leqslant f(q) \leqslant \frac{n-1}{n}$ for $q \in (0,1)$. Due to the monotonicity, it is enough to calculate the values of the function at the points $q=0$ and $q=1$.

To calculate the limit at the point $q=0$, use the Lopital rule.

$$\lim_{q\to 0} f = \lim_{q\to 0} \frac{1-(1-q)^{n-1}}{nq} = \lim_{q\to 0} \frac{(n-1)(1-q)^{n-2}}{n} = \frac{n-1}{n}.$$

Also calculate the value at the point $q=1$: $f(1) = \frac{1}{n}$.

Then both the function $f(q)$ and the equilibrium value c^* lie in the desired interval.

Finally, it remains to consider the following function:

$$g(c) = c - f(q(c)).$$

The solution of the equation $g(c) = 0$ will be the solution of the original equation by c^*. It follows from the properties of the function f that $g(c)$ increases monotonously with c. Also $g(c)$ is continuous.

At the boundaries of the interval one has $g(0) \leqslant -\frac{n-1}{n} < 0$ and $g(1) \geqslant 1 - \frac{1}{n} > 0$. Then, according to the intermediate value theorem, there is a point where $g = 0$. The theorem has been proved.

Analysis of the Results in a Single-Winner Game

Let's study the properties of the equilibrium in the game with $m = 1$.

Corollary 1. *In a two-player game, $c^* = \frac{1}{2}$ holds for any cost distribution function $F(c)$.*

Indeed,

$$c^* = \frac{1-(1-q)^{2-1}}{2q} = \frac{1}{2}.$$

This means that for $n = 2$, a player's decision does not depend on the proportion of good types, the proportion of strategic players, and the distribution of mimicry costs.

In the proof Theorem 1, it was additionally obtained the following.

Corollary 2. *For any values of the model parameters, $\frac{1}{n} \leqslant c^* \leqslant \frac{n-1}{n}$.*

It follows from this that for any values of the model parameters, there will be types of agents for whom it is optimal to mimic and those who will not mimic. If the costs are small $c_i < \frac{1}{n}$, then it is optimal for the player to mimic regardless, for example, of the scale of the leakage or the cost distribution function of other players. Conversely, at high costs $c_i > \frac{n-1}{n}$ the player never mimics.

Statement 1. *The threshold equilibrium value c^* decreases monotonously with the growth of p, α at $n > 2$. Also, c^* has a monotonic dependence on the distribution parameter λ if the distribution function F monotonically depends on λ.*

Proof. Since c^* is not expressed analytically, we will use the implicit function theorem. Consider this equation:

$$g = c^* - f(q(c^*)) = 0.$$

To analyze the dependence of the threshold value on the fraction of p, we determine the sign of the derivative:

$$\frac{\partial c^*}{\partial p} = -\frac{\frac{\partial g}{\partial p}}{\frac{\partial g}{\partial c^*}}.$$

The numerator is:

$$\frac{\partial g}{\partial p} = -\frac{\partial f}{\partial p} = -\frac{\partial f}{\partial q}\frac{\partial q}{\partial p} > 0,$$

since it was previously shown that $\frac{\partial f}{\partial q} < 0$ if $n > 2$, and $\frac{\partial q}{\partial p} = 1 - \alpha F(c^*) > 0$.

The denominator is:

$$\frac{\partial g}{\partial c^*} = 1 - \frac{\partial f}{\partial c^*} = 1 - \frac{\partial f}{\partial q}\frac{\partial q}{\partial c^*} \geqslant 1 > 0,$$

since $\frac{\partial f}{\partial q} < 0$ if $n > 2$, and $\frac{\partial F}{\partial c^*} \geqslant 0$ by definition of the cumulative distribution function, so $\frac{\partial q}{\partial c^*} = (1-p)\alpha\frac{\partial F}{\partial c^*} \geqslant 0$.

It is also true that $\frac{\partial c^*}{\partial p} < 0$.

This result has a natural interpretation. The more likely it is that agents who are really strong according to the model will appear, the more difficult it is for weak, albeit strategic, players to compete with them. This means that fewer bad agents will try to mimic.

Similarly, we prove the dependence on α for $n > 2$.

$$\frac{\partial c^*}{\partial \alpha} = -\frac{\frac{\partial g}{\partial \alpha}}{\frac{\partial g}{\partial c^*}} < 0$$

due to the fact that

$$\frac{\partial g}{\partial \alpha} = -\frac{\partial f}{\partial \alpha} = -\frac{\partial f}{\partial q}\frac{\partial q}{\partial \alpha} = -\frac{\partial f}{\partial q}((1-p)F(c^*)) > 0.$$

This result stems from a similar logic of strategic behavior of agents. The more likely it is that a weak agent who knows about information leakages will appear, the more likely it will be necessary to compete with them too. This means that fewer weak agents, according to the initial scoring model, will try to mimic.

Finally, consider the parametric family of distribution functions $F(c, \lambda)$. Let's determine the possible nature of the dependence on the parameter λ using the implicit function theorem:

$$\frac{\partial c^*}{\partial \lambda} = -\frac{\frac{\partial g}{\partial \lambda}}{\frac{\partial g}{\partial c^*}}.$$

Transform the numerator

$$\frac{\partial g}{\partial \lambda} = -\frac{\partial f}{\partial \lambda} = -\frac{\partial f}{\partial q}\frac{\partial q}{\partial \lambda} = -\frac{\partial f}{\partial q}(1-p)\alpha\frac{\partial F}{\partial \lambda}.$$

It follows that $\operatorname{sgn}(\frac{\partial g}{\partial \lambda}) = \operatorname{sgn}(\frac{\partial F}{\partial \lambda})$, which means

$$\operatorname{sgn}\left(\frac{\partial c^*}{\partial \lambda}\right) = -\operatorname{sgn}\left(\frac{\partial F}{\partial \lambda}\right).$$

In particular, if F monotonously increases (decreases) with the growth of the parameter λ, then the threshold value of c^* monotonously decreases (increases) with the growth of the parameter λ. Thus, the growth of the parameter λ in the cost distribution function means that the cost of mimicry is reduced in society (for example, the new scoring model uses more obvious parameters that can be more easily distorted), and this leads to a decrease in the threshold value at which the player stops trying to mimic.

Statement 1 has been proven.

It became clear how the threshold value changes with a change in the λ parameter, and this result does not seem intuitive until it is found out how the proportion of those strategic agents who decide to mimic changes. This value corresponds to the value of $F(c^*)$. We have a distribution function $F(c, \lambda)$. Denote by $F^{(1,0)}$, $F^{(0,1)}$ the derivatives of F in the first and second variables, respectively. By definition of the distribution function $F^{(1,0)} \geqslant 0$. Let the distribution function depend monotonically on the parameter λ, i.e. the sign $F^{(0,1)}$ is the same for all parameter values. Let's examine the sign of $\frac{\partial F(c^*(\lambda),\lambda)}{\partial \lambda}$. Make some calculations

$$\frac{\partial F(c^*(\lambda),\lambda)}{\partial \lambda} = \frac{\partial c^*}{\partial \lambda}F^{(1,0)} + F^{(0,1)} = -\frac{-\frac{\partial f}{\partial q}\frac{\partial q}{\partial F}F^{(0,1)}}{1-\frac{\partial f}{\partial q}\frac{\partial q}{\partial F}F^{(1,0)}}F^{(1,0)} + F^{(0,1)},$$

that gives

$$\frac{\partial F(c^*(\lambda),\lambda)}{\partial \lambda} = \left(\frac{\frac{\partial f}{\partial q}\frac{\partial q}{\partial F}F^{(1,0)}}{1-\frac{\partial f}{\partial q}\frac{\partial q}{\partial F}F^{(1,0)}} + 1\right)F^{(0,1)}.$$

It was shown above that $\frac{\partial f}{\partial q} < 0$, $\frac{\partial q}{\partial F} > 0$, $F^{(1,0)} \geqslant 0$ for $n > 2$. It is also easy to see that $\frac{x}{1-x} > -1$ for $x \leqslant 0$. It follows from these statements that the expression in large brackets is positive. From here we conclude:

$$\operatorname{sgn}\left(\frac{\partial F(c^*(\lambda),\lambda)}{\partial \lambda}\right) = \operatorname{sgn}(F^{(0,1)}).$$

Let's look at this effect in more details using the example of an exponential distribution. Let's fix $n = 3$, $m = 1$, $p = 0.2$, $\alpha = 0.5$, $F(c, \lambda) = 1 - \exp(-\lambda_i c)$, where $\lambda_1 = 0.5$, $\lambda_2 = 2$.

We solve the equation $c^* - f(q(c^*)) = 0$ numerically and find c^* with an accuracy of four decimal places.

$$\begin{gathered} c^*(\lambda_1) = 0.5671, \\ c^*(\lambda_2) = 0.5143, \\ F(c^*(\lambda_1), \lambda_1) = 0.2469, \\ F(c^*(\lambda_2), \lambda_2) = 0.6425. \end{gathered}$$

In the example considered, an increase in λ entailed a slight decrease in the threshold value, but significantly increased the proportion of mimics.

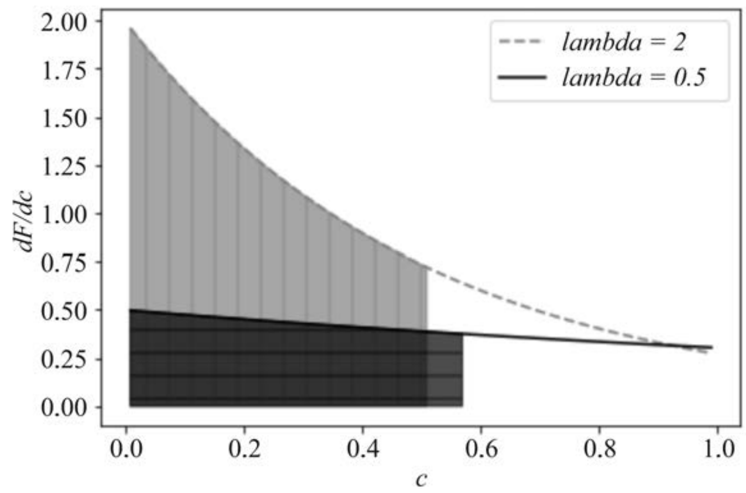

Fig. 1. The equilibrium threshold value and the proportion of mimicking players with exponential cost distribution.

Figure 1 shows a graph with distribution densities at $\lambda_1 = 0.5$, $\lambda_2 = 2$ and the values of the proportion of mimics in equilibrium as the area under the graphs at $c \in [0, c^*]$.

Let's summarize the results. Suppose the cost distribution has changed in such a way that $F(c)$ has become larger for all values of c. This is equivalent to simplifying the scoring model and increasing its vulnerability. Then the threshold value in the equilibrium of c^* will decrease, but not so much: the probability that the consumer will mimic will increase, as well as the proportion of mimics. In addition, as it was found out earlier, c^* decreases with the growth of p and α. Since these parameters do not affect the distribution, the probability that the consumer will mimic also decreases with the growth of p and α.

4. NASH EQUILIBRIUM IN A GAME WITH m WINNERS

Now let $m \leqslant n$ good players get access to the product according to the scoring model. We will look for a symmetric Bayesian–Nash equilibrium. Let y be the probability that the bad one will mimic the good one in equilibrium. Similarly to the case of a single winner, we define q as the probability that the player will be classified by the algorithm as a good type. This is calculated using the full probability formula:

$$q = p + (1 - p)\alpha y.$$

Let's find the expected utilities of the ith player, who is a strategic agent, from the "mimic" and "do not mimic" strategies. Let k competitors be classified as a good type. As in the case of a single winner, the random variable "the number of competitors who will be classified as good" has a binomial distribution: $\mathrm{Bin}(n - 1, q)$.

If k competitors are classified as a good type and the ith player chooses the "do not mimic" strategy, then if $m \leqslant k$ all places will be occupied by players who will be classified as good. If $m > k$, then there remains $m - k$ of products for bad players according to the model, the probability of getting them is $\frac{m-k}{n-k}$.

If k competitors are classified as a good type and the ith player chooses the "mimic" strategy, then if $m > k$ all the players who are classified as good by the model are guaranteed to receive the goods. This means that the ith player will receive the product with probability 1. If $m \leqslant k$, then $k + 1$ good players according to the model will compete for m places. Therefore, the probability of receiving the product is $\frac{m}{k+1}$.

Therefore, the expected gain from the "do not mimic" strategy is

$$\sum_{k=0}^{n-1} \max\left(0, \frac{m-k}{n-k}\right)\binom{n-1}{k} q^k (1-q)^{n-1-k}.$$

The expected gain from the "mimic" strategy is

$$-c_i + \sum_{k=0}^{n-1} \min\left(1, \frac{m}{k+1}\right)\binom{n-1}{k} q^k (1-q)^{n-1-k}.$$

Note that the expected gain from the "do not mimic" strategy does not depend on c_i, while the expected gain from the "mimic" strategy decreases with an increase in c_i. This means that the optimal strategy of the player i is monotonous in the same meaning as in Section 3: there is a c^* such that for all $c_i < c^*$ ith player mimics and for all $c_i > c^*$ ith player does not mimic. With $c = c^*$, the expected gains from both strategies are the same.

The probability that a bad type will mimic a good one in equilibrium is expressed in terms of c^*:

$$y = P[c_i \leqslant c^*] = F(c^*).$$

By equating the expected gains from pure strategies, we obtain a condition for the threshold value of c^*:

$$c^* = \sum_{k=0}^{n-1} \left(\min\left(1, \frac{m}{k+1}\right) - \max\left(0, \frac{m-k}{n-k}\right)\right)\binom{n-1}{k} q^k (1-q)^{n-1-k},$$

where

$$q = p + (1-p)\alpha F(c^*).$$

Let's transform the resulting expression. If $k < m$, then

$$\min\left(1, \frac{m}{k+1}\right) - \max\left(0, \frac{m-k}{n-k}\right) = 1 - \frac{m-k}{n-k} = \frac{n-m}{n-k}.$$

If $k \geqslant m$, then

$$\min\left(1, \frac{m}{k+1}\right) - \max\left(0, \frac{m-k}{n-k}\right) = \frac{m}{k+1} - 0 = \frac{m}{k+1}.$$

Then we get the following expression for the threshold value:

$$c^* = \sum_{k=0}^{m-1} \frac{n-m}{n-k}\binom{n-1}{k} q^k (1-q)^{n-1-k} + \sum_{k=m}^{n-1} \frac{m}{k+1}\binom{n-1}{k} q^k (1-q)^{n-1-k}.$$

For the convenience of the equilibrium analysis, we introduce an additional function

$$\tilde{f}(m, q) = \sum_{k=0}^{m-1} \frac{n-m}{n-k}\binom{n-1}{k} q^k (1-q)^{n-1-k} + \sum_{k=m}^{n-1} \frac{m}{k+1}\binom{n-1}{k} q^k (1-q)^{n-1-k}.$$

Then the condition is rewritten as

$$c^* = \tilde{f}(m, q(c^*)),$$
$$q(c^*) = p + (1-p)\alpha F(c^*).$$

Analysis of the Results of the Game with m Winners

Let's start by analyzing the dependence of the equilibrium threshold on q. This dependence reflects the marginal increase in the equilibrium threshold with the a posteriori probability of mimic. Technically, this is an important object, since the effect of other parameters on the equilibrium ultimately depends on the sign of df/dq.

Statement 2. *The function $\tilde{f}(m, q)$ is monotone in q if and only if $m = 1$ or $m = n - 1$.*

Proof. First of all, note that $\tilde{f}(m = 1, q) = \tilde{f}(m = n - 1, 1 - q)$.

As it was proved in the analysis of the game with one winner, $\tilde{f}(m = 1, q)$ strictly decreases with the growth of q. Then it follows that $\tilde{f}(m = n - 1, q)$ strictly increases with the growth of q.

Now we show that for $1 < m < n - 1$, the function $\tilde{f}(m, q)$ is not monotonic with respect to q. To do this, calculate the limits of the derivative at $q \to 0$ and $q \to 1$.

$$\frac{\partial \tilde{f}}{\partial q} = \sum_{k=0}^{m-1} \frac{n-m}{n-k} \binom{n-1}{k} q^{k-1}(1-q)^{n-2-k}(k-(n-1)q)$$
$$+ \sum_{k=m}^{n-1} \frac{m}{k+1} \binom{n-1}{k} q^{k-1}(1-q)^{n-2-k}(k-(n-1)q).$$

Let's use the fact that $\lim\limits_{q \to 0} q^a(1-b)^b = 1$ for $a = 0,\ b > 0$.

$$\lim_{q \to 0} \frac{\partial \tilde{f}}{\partial q} = \begin{cases} -(n-1)\dfrac{n-m}{n} + \dfrac{m}{2}\dbinom{n-1}{1}, & \text{if } m = 1, \\[2ex] -(n-1)\dfrac{n-m}{n} + \dfrac{n-m}{n-1}\dbinom{n-1}{1}, & \text{if } m > 1, \end{cases}$$
$$= \begin{cases} -\dfrac{(n-1)(n-2)}{2n}, & \text{if } m = 1, \\[2ex] \dfrac{n-m}{n}, & \text{if } m > 1. \end{cases}$$

$$\lim_{q \to 1} \frac{\partial \tilde{f}}{\partial q} = \begin{cases} -\dfrac{n-m}{n-(n-2)}(n-1) + \dfrac{m}{n}(n-1), & \text{if } m = n-1, \\[2ex] -\dfrac{m}{n-1}(n-1) + \dfrac{m}{n}(n-1), & \text{if } m < n-1, \end{cases}$$
$$= \begin{cases} \dfrac{(n-1)(n-2)}{2n}, & \text{if } m = n-1, \\[2ex] -\dfrac{m}{n}, & \text{if } m < n-1. \end{cases}$$

Thus, three cases are possible:

1. $m = 1$. In this case, $\frac{\partial \tilde{f}}{\partial q} < 0$.

2. $m = n - 1$. In this case, $\frac{\partial \tilde{f}}{\partial q} > 0$.

3. $1 < m < n - 1$. In this case, $\lim\limits_{q \to 0} \frac{\partial \tilde{f}}{\partial q} > 0$ and $\lim\limits_{q \to 1} \frac{\partial \tilde{f}}{\partial q} < 0$. Then it follows from the continuity of $\frac{\partial \tilde{f}}{\partial q}$ that the function reaches a maximum at $q \in [0, 1]$ at $q^* \in (0, 1)$.

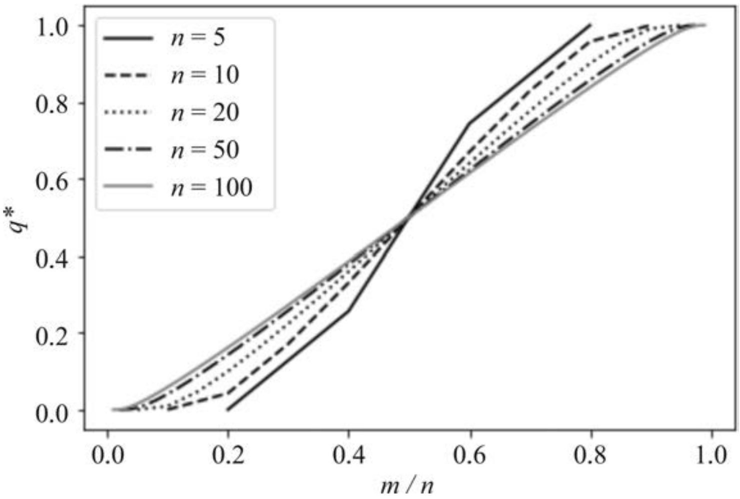

Fig. 2. The maximum point of $\tilde{f}(q)$ depending on n and $\frac{m}{n}$.

So, it is shown that if $1 < m < n-1$, then $\tilde{f}(m,q)$ is not monotonic in q. Apparently, the sums used to describe $\tilde{f}(m,q)$ are too complex and cannot be simplified in the form of elementary functions. Therefore, it is not possible to analytically express q^*, which maximizes $\tilde{f}(m,q)$ at a fixed m.

Statement 2 has been proven.

Let's consider a remarkable special case $n = 2m$, $m > 1$. Let's suppose $k \geqslant m$. Then, for $n = 2m$, $n-1-k < m$ holds. Also note that for $n = 2m$,

$$\frac{n-m}{n-(n-1-k)}\binom{n-1}{n-1-k} = \frac{m}{k+1}\binom{n-1}{k}.$$

In addition, when $q = \frac{1}{2}$, it is true that $q^{k-1}(1-q)^{n-2-k} = 2^{3-n}$.

These facts allow us to calculate the partial derivative at $q = \frac{1}{2}$.

$$\frac{\partial \tilde{f}}{\partial q}\left(n = 2m, q = \frac{1}{2}\right) = \sum_{k=m}^{n-1} \frac{m}{k+1}\binom{2m-1}{k} 2^{3-2m}(k + (2m-k-1) - (2m-1)) = 0.$$

In general, it is not easy to show that the given point is the maximum point. For example, consider $n = 4, m = 2$. The second-order condition is:

$$\frac{\partial^2 \tilde{f}}{\partial q^2}(n = 4, m = 2) = -1.$$

Thus, $\tilde{f}$ is a concave function with respect to q at $n = 4,\ m = 2$, which means that at the point $q = 0.5$ it reaches a global maximum.

An interesting question is what the maximum point of the function $\tilde{f}$ tends to (denote by $q^*(n,m)$) for large m. To do this, we find q^* for fixed n, m by numerical methods.

We note that the larger n, the more the dependence $q^*(m)$ is similar to a linear one (already at $n = 100$ we note $R^2 > 99.9\%$ for regression $q^* = \beta_0 + \beta_1 m$; R^2 grows with the growth of n). In addition, with the growth of n $q^*(m = 2)$ tends to 0, and $q^*(m = n-2)$ tends to 1. This allows us to hypothesize that for large n and $1 < m < n-1$, the maximum point of $\tilde{f}(q)$ tends to $q^* = \frac{m-2}{n-4}$ (Fig. 2) The verification of this hypothesis is interesting for the further research; in the case of truth, the intuition of the result is interesting.

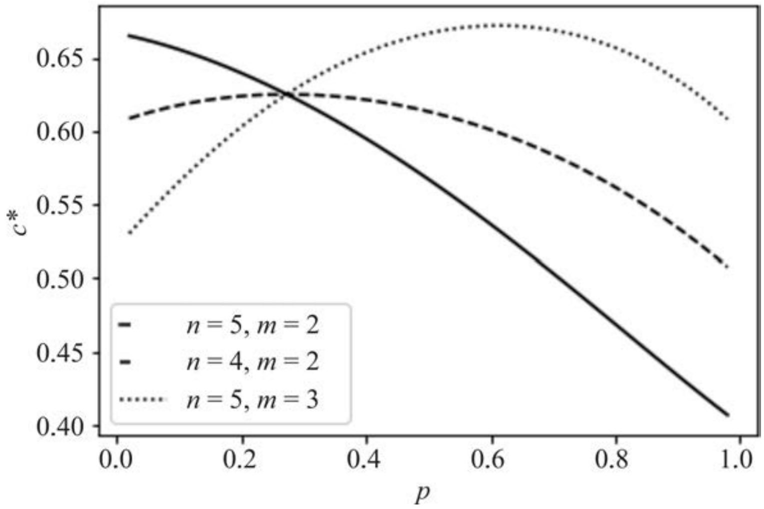

Fig. 3. Dependence of c^* on p at $\alpha = 0.5$, $F(c) = c$.

Statement 3. *In general, the threshold value of c^* is non-monotonic with respect to p and α.*

Proof. Let's calculate the derivatives of the function given implicitly, as was done earlier in Statement 2.

$$\tilde{g} = c^* - \tilde{f}(q(c^*)) = 0.$$

$$\operatorname{sgn}\left(\frac{\partial c^*}{\partial p}\right) = \operatorname{sgn}\left(-\frac{\frac{\partial \tilde{g}}{\partial p}}{\frac{\partial \tilde{g}}{\partial c^*}}\right) = \operatorname{sgn}\left(\frac{\partial \tilde{f}}{\partial q}\right).$$

$$\operatorname{sgn}\left(\frac{\partial c^*}{\partial \alpha}\right) = \operatorname{sgn}\left(-\frac{\frac{\partial \tilde{g}}{\partial \alpha}}{\frac{\partial \tilde{g}}{\partial c^*}}\right) = \operatorname{sgn}\left(\frac{\partial \tilde{f}}{\partial q}\right).$$

Partial differential transformations are similar to the transformations in statement 1, when these derivatives were calculated for the case of $m = 1$. In Statement 2, the nonmonotonicity of $\tilde{f}$ by q was proved in the general case. It follows that $\operatorname{sgn}\left(\frac{\partial \tilde{f}}{\partial q}\right)$ is not a constant value. But then $\operatorname{sgn}\left(\frac{\partial c^*}{\partial p}\right), \operatorname{sgn}\left(\frac{\partial c^*}{\partial \alpha}\right)$ are not constant values for $1 < m < n - 1$. The case of $m = 1$ was considered earlier, for $m = n - 1$ all conclusions are opposite to the case of $m = 1$ (i.e. $\frac{\partial c^*}{\partial p} > 0$, $\frac{\partial c^*}{\partial \alpha} > 0$ for $m = n - 1$). Moreover, it is obvious that the maxima for p and for q are also uniquely related.

Statement 3 has been proven.

Let's demonstrate nonmonotonicity with examples. Numerically calculate the value of c^* for fixed parameter values from the conditions $c^* = \tilde{f}(m, q(c^*))$, $q = p + (1 - p)\alpha F(c^*)$. Let's fix that the costs are uniformly distributed over the interval $[0, 1]$, and find several solutions to the equations for different parameter values with an accuracy of up to the fourth decimal place.

$$c^*(n = 4, m = 2, p = 0.1, \alpha = 0.5) = 0.6175,$$
$$c^*(n = 4, m = 2, p = 0.3, \alpha = 0.5) = 0.6248,$$
$$c^*(n = 4, m = 2, p = 0.5, \alpha = 0.5) = 0.6132.$$

Thus, there is generally no monotonicity of c^* by p (Fig. 3). Intuitively, we explain this fact: with increasing probability p, the initial number of good agents increases, so that under a small proportion of them, it makes sense to actively mimic and fight for a relatively large number of

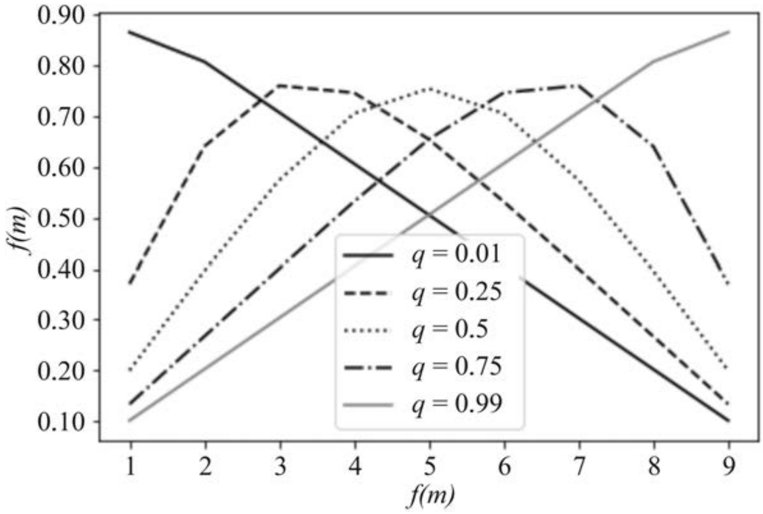

Fig. 4. Dependence of c^* on m at $n = 20$, $\alpha = 0.2$, $F(c) = c$.

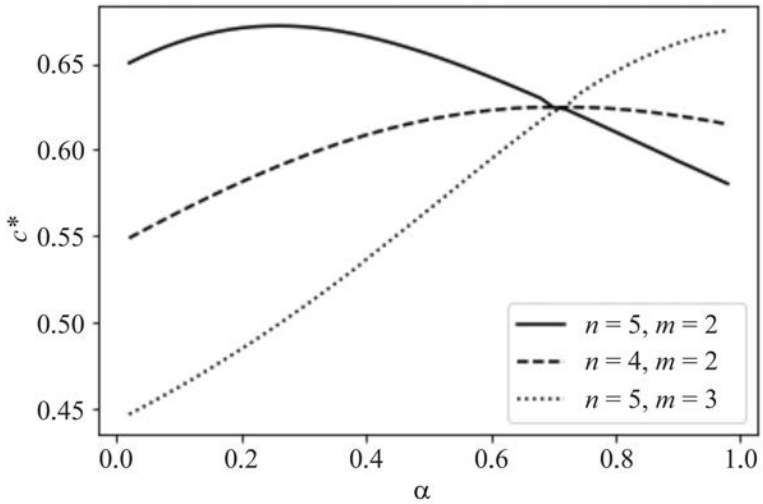

Fig. 5. Dependence of c^* on α at $p = 0, 1$, $F(c) = c$.

"prizes." However, if their share is already high, then it is difficult to compete with them, the probability of winning is too low, so the proportion of mimicking players (and the corresponding threshold value) falls.

Similarly, let's look at the dependence on the scale of the leakage α.

$$c^*(n = 4, m = 2, p = 0.1, \alpha = 0.5) = 0.6175,$$
$$c^*(n = 4, m = 2, p = 0.1, \alpha = 0.7) = 0.6248,$$
$$c^*(n = 4, m = 2, p = 0.1, \alpha = 0.9) = 0.6198.$$

Thus, there is generally no monotonicity of c^* by α (Fig. 4). A similar logic applies here: with increasing leakage, more and more weak people will potentially want to mimic the strong ones and up to some level it will be beneficial for them. However, the further competition will become too tough and therefore the desire to mimic will decrease.

Statement 4. *The threshold value c^* is non-monotonic with respect to m in the general case.*

As in Statement 3, we calculate c^* for various parameter values with an accuracy of up to four decimal places and show the non-monotonicity of the threshold value in equilibrium by the number of winners.

$$c^*(n = 4, m = 1, p = 0.5, \alpha = 0.5) = 0.3910,$$
$$c^*(n = 4, m = 2, p = 0.5, \alpha = 0.5) = 0.6132,$$
$$c^*(n = 4, m = 3, p = 0.5, \alpha = 0.5) = 0.5045.$$

In this case, the explanation is that with the number of winners growth close to the number of all participants, it may become pointless to try and spend money on mimicry, since even without this, the probability of entering the number of m winners is high. However, as can be seen from Fig. 5, for special scoring models that generate a very high or very low probability of classifying a player as good, there is a monotonous dependence of the proportion of mimics on the number of prizes.

5. CONCLUSION

The article models a situation in which some of the firm's clients learn their internal rating in the company and can change their behavior to increase their internal rating. In the considered framework, the scoring model splits users into "good" and "bad" (binary classification of agent types).

It is proven that equilibrium exists, is unique and is a profile of monotonic strategies. The presence of an internal threshold value indicates that not all agents, obtaining access to the mechanism of the scoring model, use this information for manipulation.

We believe that there is great potential for further research. In the case of a binary distribution of types, it is interesting to study the discovered non-monotonic dependencies in more details, in particular, to find the extreme points analytically depending on the parameter values. It may be useful to extend the binary classification to a discrete one, since some scoring models distribute users across several clusters, and to a continuous one, since many models return the rating as a real number, often from the interval $[0, 1]$.

In conclusion, it should be noted that the considered problem of behavior manipulation is an example of the manifestation of the property of natural intelligence to continuously adapt and search for complex and non-trivial strategies to improve its utility. In this sense, automated models are unlikely to be able to fully resist all the possibilities potentially available to humans. However, if one predicts the strategic behavior of users in advance, she can more effectively assess both the stability and prospects of the original model, which can affect the optimal choice of the model, or formulate more adequate criteria for improving the model. Another related problem is the manipulation of behavior during model training, which will certainly affect its further functioning. This leads to the proposal that the optimal trajectory of the development of the machine learning and artificial intelligence should include an adequate game-theoretic element, taking into account that the data for these models are generated by strategic people.

FUNDING

This work was supported by HSE University (Basic Research Program).

REFERENCES

1. Carr, A., I found out my secret internal tinder rating and now I wish I hadn't, https://www.fastcompany.com/3054871/whats-your-tinder-score-inside-the-apps-internal-ranking-system (Accessed: 22.01.2024)

2. Albanesi, S. and Vamossy, D., Predicting consumer default: A deep learning approach, *National Bureau of Econom. Res.*, 2019, no. w26165, 72 p.

3. Pławiak, P., Abdar, M., Plawiak, J., et al., DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring, *Inform. Sci.*, 2020, vol. 516, pp. 401–418.

4. Hurley, M. and Adebayo, J., Credit scoring in the era of big data, *Yale JL & Tech.*, 2016, vol. 18, no. 148.

5. Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., and Haltmeier, M., A machine learning framework for customer purchase prediction in the non-contractual setting, *Eur. J. Oper. Res.*, 2020, vol. 281, no. 3, pp. 588–596.

6. Syakur, M.A. et al., Integration k-means clustering method and elbow method for identification of the best customer profile cluster, *IOP conference series: materials science and engineering*, IOP Publishing, 2018, vol. 336, no. 012017.

7. Roth, A., Game theory as a part of empirical economics, *The Econ. J.*, 1991, vol. 101, no. 404, pp. 107–114.

8. Harsanyi, J., Games with incomplete information played by "bayesian" players, I–III part I. The basic model, *Management Sci.*, 1967, vol. 14, no. 3, pp. 159–182.

9. Burkov, V.N., *Osnovy matematicheskoi teorii aktivnyh sistem* (Foundations for mathematical theory of active systems), Moscow: Nauka, 1977.

10. Burkov, V.N. and Novikov, D.A., Teorii aktivnyh system 50 let: istoriia razvitiia (50 years to the theory of active systems), *Matherials of international scientific-practical conference "Theory of active systems – 50 years"*, Moscow: ICS, 2019, pp. 10–57.

11. Enaleev, A.K., The optimality of correlated mechanisms of funtioning in active systems, *Upravlenie bolshimi sistemami*, 2011, no. 33, pp. 143–166.

12. Enaleev, A.K., The optimal correlated mechanism in the system with several active agents, *Problemy upravleniia*, 2015, no. 3, pp. 20–28.

13. Burkov, V.N., Enaleev, A.K., and Korgin, N.A., Incentive Compatibility and Strategy-Proofness of Mechanisms of Organizational Behavior Control: Retrospective, State of the Art, and Prospects of Theoretical Research, *Autom. Remote Controle*, 2021, no. 7, pp. 5–37.

14. Dellarocas, C., Strategic manipulation of internet opinion forums: Implications for consumers and firms, *Management Sci.*, 2006, vol. 52, no. 10, pp. 1577–1593.

15. Dini, F. and Spagnolo, G. Buying reputation on eBay: Do recent changes help?, *Int. J. Electron. Business*, 2009, vol. 7, no. 6, pp. 581–598.

16. Wright, R. et al., Directed search and competitive search equilibrium: A guided tour, *J. Econ. Lit.*, 2021, vol. 59, no. 1, pp. 90–148.

17. Sandomirskaia, M. and Shavshin, R., Price Competition in Finite Markets with a Rare Good and Private Consumer Valuations, *Higher School Econom. Res. Paper*, 2021, vol. 248, 29 p.

18. Peters, M., Bertrand equilibrium with capacity constraints and restricted mobility, *Econometrica*, 1984, vol. 52, no. 5, pp. 1117–1127.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 8, pp. 802–808.

Beckmann's Bertrand–Edgeworth Duopoly Model: New Pure Strategy Equilibria

A. S. Shvedov

National Research University Higher School of Economics, Moscow, Russia
e-mail: ashvedov@hse.ru

Received December 27, 2023
Revised May 27, 2024
Accepted June 27, 2024

Abstract—In this paper, the concept of weak active equilibrium, which was introduced in the eighties of the twentieth century by Smol'yakov, is used for the first time for a study of Bertrand–Edgeworth oligopoly (that is, competition among firms when firms' strategies are prices and firms' production capacities are limited). All symmetric weak active equilibria are found for the basic model of Bertrand–Edgeworth duopoly, which was analysed in the sixties of the twentieth century by Beckmann. However, this model is highly simplified. Also, the question of existence of nonsymmetric weak active equilibria is studied.

Keywords: Bertrand–Edgeworth duopoly, noncooperative game, weak extremal profile of actions, weak active equilibrium

DOI: 10.31857/S0005117924080059

1. INTRODUCTION

For its simplicity, the Beckmann's model [1, 2] is far enough from reality. However, this model has served as a starting point for many subsequent studies (e.g., [3–15]). In the classical setting for Bertrand oligopoly, it is assumed that, after the announcement of prices, firms can produce as much as necessary to meet demands. In this setting, firms' strategies are prices and the firm's goal is to maximize profits. The assumption of bounded capacities of firms (in this case the Bertrand oligopoly is called the Bertrand–Edgeworth oligopoly) significantly changes the answer to the question what strategies should be chosen by firms in this game. For the model in question, Beckmann proved that Nash equilibria in pure strategies exist for some firms' capacities and do not exist for other firms' capacities. For the latter case, Beckmann found Nash equilibria in mixed strategies. This led to the fact that considerable attention is given to Nash equilibria in mixed strategies in many subsequent studies of Bertrand–Edgeworth oligopolies. Despite the fact that, in studies of oligopolies, Nash equilibria in pure strategies are more important than Nash equilibria in mixed strategies for practical purposes.

The present paper shows that equilibria in pure strategies exist for all firms' capacities in the Beckmann's model if weak active equilibria [16, 17] are used instead of Nash equilibria. (The Nash equilibrium is a special case of the weak active equilibrium.) So, there is no need for equilibria in mixed strategies when studying this duopoly model. In this paper, it is shown which symmetric profiles of actions are weak active equilibria and which ones are not. Also, the question of existence of nonsymmetric weak active equilibria is investigated. In more detail, Cournot oligopoly and Bertrand oligopoly are described, for instance, in [18].

Suppose that each of the indices i and j takes the values 1 and 2, $i \neq j$. The player i can choose an action x_i from a set D_i. By $f_i(x_i, x_j)$ denote the payoff of the player i for the profile (x_i, x_j) of actions.

Definition 1. The profile $(\bar{x}_i, \bar{x}_j)$ of actions is called weak extremal for the player i if for any action $x_i \in D_i$ there exists an action $x_j \in D_j$ such that

$$f_i(x_i, x_j) \leqslant f_i(\bar{x}_i, \bar{x}_j).$$

Definition 2. The profile $(\bar{x}_i, \bar{x}_j)$ of actions is called a weak active equilibrium if this profile is weak extremal for any player.

The concepts of the weak extremal profile and the weak active equilibrium were introduced in the eighties of the twentieth century by Smol'yakov.

Definition 3. The profile (x_i^*, x_j^*) of actions is called a Nash equilibrium if

$$f_i(x_i^*, x_j^*) = \max_{x_i \in D_i} f_i(x_i, x_j^*)$$

for any i.

It is not difficult to see that any Nash equilibrium is a weak active equilibrium. In general, the converse is not true.

In Section 2, Bertrand–Edgeworth duopoly is described. In Section 3, a necessary and sufficient condition for a profile to be weak extremal is given for small production capacities. The same condition is given for big production capacities in Section 4. In Section 5, weak active equilibria are presented and practical recommendations are discussed.

2. BERTRAND–EDGEWORTH DUOPOLY

Two firms that produce a homogeneous good are under consideration. Production costs are assumed to be zero. Production capacities of the firms are equal. A quantity of goods that a firm can produce is denoted by c, $0 < c < 1$. The firms simultaneously announce prices $x_1 \in D_1$ and $x_2 \in D_2$ at which the product will be sold by the firm, $D_1 = D_2 = [0, 1]$. The demand function is assumed to be linear:

$$q = 1 - x. \tag{1}$$

This means that if the firms announced the same price x, then the quantity of sold goods is determined by formula (1). Then the income of each firm is equal to $0.5x(1 - x)$. (It is assumed that the firms share the market equally in this case.) If the firms announced different prices, then the firm that announced a lower price sells the good first; by x denote this price. Suppose that $x > 1 - c$. Then it follows from (1) that $q < c$. Thus, the quantity of goods that is sold by the firm is equal to $1 - x$. In this case, the other firm sells nothing and the income of the other firm is zero. Suppose that $x \leqslant 1 - c$. Then it follows from (1) that $q \geqslant c$. The quantity of goods that is sold by the firm is equal to c due to limited production capacities. By y denote the price announced by the other firm. The residual demand function for the other firm has the form

$$\left(1 - \frac{c}{1 - x}\right)(1 - y). \tag{2}$$

This means that the quantity of goods that can be sold by the other firm is determined by formula (2). Economic meaning of the residual demand function is discussed, for instance, in [5]. Also, [5] compares (2) and another form of the residual demand function. The goal of the firm is maximize its income which is calculated as the product of the price and the quantity of the good that is sold. It can be assumed that the quantity of the good that is produced by each firm is equal to c since the production costs are zero.

Thus, the income of the firm 1 is calculated by the following formula

$$f(x_1, x_2) = \begin{cases} cx_1 & \text{for} \quad x_1 < x_2,\ x_1 \leqslant 1-c, \\ x_1(1-x_1) & \text{for} \quad x_1 < x_2,\ x_1 > 1-c, \\ cx_1 & \text{for} \quad x_1 = x_2,\ x_2 \leqslant 1-2c, \\ 0.5x_1(1-x_1) & \text{for} \quad x_1 = x_2,\ x_2 > 1-2c, \\ x_1 \min\left(c, \left(1 - \dfrac{c}{1-x_2}\right)(1-x_1)\right) & \text{for} \quad x_1 > x_2,\ x_2 \leqslant 1-c, \\ 0 & \text{for} \quad x_1 > x_2,\ x_2 > 1-c. \end{cases} \tag{3}$$

The income of the firm 2 is equal to $f(x_2, x_1)$. Zero prices are allowed to be announced. But it turns out that such case is of little interest for this model in contrast to classical Bertrand oligopoly.

It is not difficult to show that the profile $(1-2c, 1-2c)$ is a Nash equilibrium if $0 < c \leqslant 0.25$ ([1, 2]). In accordance with (1), this means that demand is satisfied as much as firms' production capacities permit. Also, it is not difficult to show that the profile $(1-2c, 1-2c)$ is not a Nash equilibrium for $0.25 < c < 0.5$ ([1, 2]).

3. WEAK EXTREMAL PROFILES FOR SMALL PRODUCTION CAPACITIES

Denote

$$\mu = (1-2c)c.$$

Lemma 1. *Assume that* $0 < c \leqslant \frac{1}{3}$. *The profile* $(\bar{x}_1, \bar{x}_2)$ *is weak extremal for the firm 1 if and only if* $f(\bar{x}_1, \bar{x}_2) \geqslant \mu$.

Proof. Suppose that $f(\bar{x}_1, \bar{x}_2) \geqslant \mu$. Assume that $0 \leqslant x_1 \leqslant 1-2c$. Then for any x_2 we have

$$f(x_1, x_2) = cx_1 \leqslant (1-2c)c \leqslant f(\bar{x}_1, \bar{x}_2).$$

Assume that $1-2c < x_1 \leqslant 1-c$. If we consider the limit on the left we find

$$\lim_{x_2 \to x_1 - 0} \left(1 - \frac{c}{1-x_2}\right) x_1(1-x_1) = \left(1 - \frac{c}{1-x_1}\right) x_1(1-x_1) = (1-c-x_1)x_1.$$

Consequently,

$$\lim_{x_2 \to x_1 - 0} f(x_1, x_2) = (1-c-x_1)x_1$$

because of the inequality $1-c-x_1 < c$. Since the function $\left(1 - \frac{c}{1-x_2}\right)$ is continuous and monotone decreasing for $x_2 < x_1$, for any $\varepsilon > 0$ there exists $x_2 < x_1$ such that

$$f(x_1, x_2) \leqslant (1-c-x_1)x_1 + \varepsilon.$$

Value of the function $(1-c-x_1)x_1$ is equal to $(1-2c)c$ for $x_1 = 1-2c$. The function $(1-c-x_1)x_1$ is monotone decreasing for $x_1 > 1-2c$ because of the condition $c \leqslant \frac{1}{3}$. Therefore $(1-c-x_1)x_1 < \mu$. Consequently, there exists ε such that $f(x_1, x_2) \leqslant \mu$ for some $x_2 < x_1$.

Assume that $1-c < x_1 \leqslant 1$. Then $f(x_1, 1-c) = 0 < \mu$.

Suppose that $f(\bar{x}_1, \bar{x}_2) < \mu$. The profile $(\bar{x}_1, \bar{x}_2)$ is not weak extremal for the firm 1 since $f(1-2c, x_2) = \mu$ for any x_2.

This completes the proof of Lemma 1.

4. WEAK EXTREMAL PROFILES FOR BIG PRODUCTION CAPACITIES

Denote
$$\nu = \left(\frac{1-c}{2}\right)^2.$$

Lemma 2. *Assume that $\frac{1}{3} < c < 1$. The profile $(\bar{x}_1, \bar{x}_2)$ is weak extremal for the firm 1 if and only if $f(\bar{x}_1, \bar{x}_2) > \nu$.*

Proof. Suppose that $f(\bar{x}_1, \bar{x}_2) > \nu$. Assume that $0 \leqslant x_1 \leqslant \max(0, 1-2c)$. Then
$$f(x_1, x_2) = cx_1 \leqslant c\max(0, 1-2c) \leqslant \left(\frac{1-c}{2}\right)^2 < f(\bar{x}_1, \bar{x}_2)$$
for $x_2 > x_1$.

Assume that $\max(0, 1-2c) < x_1 \leqslant 1-c$. Denote $\varepsilon = f(\bar{x}_1, \bar{x}_2) - \nu$. We have
$$\lim_{x_2 \to x_1 - 0} \left(1 - \frac{c}{1-x_2}\right) x_1(1-x_1) = (1-c-x_1)x_1.$$
Consequently,
$$\lim_{x_2 \to x_1 - 0} f(x_1, x_2) = (1-c-x_1)x_1$$
since $1-c-x_1 < c$. The function $(1-c-x_1)x_1$ reaches the maximum at the point $x_1 = \frac{1-c}{2}$ (an interior point of the segment $[\max(0, 1-2c), 1-c]$); the maximum is $\left(\frac{1-c}{2}\right)^2$. Thus, for any $x_1 \in (\max(0, 1-2c), 1-c]$ there exists $x_2 < x_1$ such that
$$f(x_1, x_2) < \left(\frac{1-c}{2}\right)^2 + \frac{\varepsilon}{2} < f(\bar{x}_1, \bar{x}_2).$$

Assume that $1-c < x_1 \leqslant 1$. Then $f(x_1, 1-c) = 0 \leqslant \nu$.

We now prove that the profile $(\bar{x}_1, \bar{x}_2)$ is not weak extremal for the firm 1 if $f(\bar{x}_1, \bar{x}_2) \leqslant \nu$. It is sufficient to show that $f\left(\frac{1-c}{2}, x_2\right) > \nu$ for any x_2. Note that $\max(0, 1-2c) < x_1 \leqslant 1-c$ for $x_1 = \frac{1-c}{2}$ since $c > \frac{1}{3}$.

Assume that $x_2 < x_1$. The function $f(x_1, x_2)$ considered as a function of the argument x_2 is monotone nonincreasing for $x_2 \in [0, x_1)$. As was established above in proving this lemma, the function $f(x_1, x_2)$ coincides with the function $\left(1 - \frac{c}{1-x_2}\right) x_1(1-x_1)$ near the point x_1. Therefore the function is monotone decreasing. Thus,
$$f(x_1, x_2) > (1-c-x_1)x_1 = \left(\frac{1-c}{2}\right)^2.$$
Assume that $x_2 = x_1$. Then
$$f(x_1, x_2) = 0.5x_1(1-x_1) = \frac{1-c}{2} \times \frac{1+c}{4} > \left(\frac{1-c}{2}\right)^2,$$
since $c > \frac{1}{3}$. Assume that $x_2 > x_1$. Then
$$f(x_1, x_2) = c\,\frac{1-c}{2} > \left(\frac{1-c}{2}\right)^2,$$
since $c > \frac{1}{3}$.

This completes the proof of Lemma 2.

5. WEAK ACTIVE EQUILIBRIA

By (3) it follows that if $0 < c \leqslant \frac{1}{4}$, then $f(x,x) = \mu$ for $x = 1-2c$ and $f(x,x) < \mu$ for $x \neq 1-2c$. Thus, by Lemma 1, $(1-2c, 1-2c)$ is the unique profile (x,x), which is a weak active equilibrium, for $0 < c \leqslant \frac{1}{4}$. Also, this profile is a Nash equilibrium.

Proposition 1. *Assume that $\frac{1}{4} < c \leqslant \frac{1}{3}$. A profile (x,x) is a weak active equilibrium if and only if $1-2c \leqslant x \leqslant 2c$.*

Proof. The proof is reduced to use of Lemma 1 and check of condition $0.5x(1-x) \geqslant \mu$, which is equivalent to $1-2c \leqslant x \leqslant 2c$.

This completes the proof of Proposition 1.

The symmetric pure strategy equilibrium $(1-2c, 1-2c)$ was found in [15], Proposition 4.3, for $\frac{1}{4} < c \leqslant \frac{1}{3}$. Also, it is established in [15] that there are no other equilibria in secure strategies for this game. However, by (3) it follows that the income of the firm is higher for the profile $(0.5, 0.5)$, which is a weak active equilibrium, than for the profile $(1-2c, 1-2c)$ for $\frac{1}{4} < c \leqslant \frac{1}{3}$.

Proposition 2. *Assume that $\frac{1}{3} < c < 1$. A profile (x,x) is a weak active equilibrium if and only if*

$$\frac{1}{2} - \frac{1}{2}\sqrt{1-2(1-c)^2} < x < \frac{1}{2} + \frac{1}{2}\sqrt{1-2(1-c)^2}.$$

Proof. The proof is reduced to use of Lemma 2 and check of condition $0.5x(1-x) > \nu$, which is equivalent to what $x \in (a_1, a_2)$, where a_1 and a_2 are the roots of the quadratic equation $0.5x(1-x) = \left(\frac{1-c}{2}\right)^2$.

This completes the proof of Proposition 2.

These results can be interpreted in the following way with respect to the real behaviour of firms. When firms' capacities are small, $c \leqslant \frac{1}{4}$, the firms need to focus on the upper part of the demand curve and put the price $1-2c$. This follows from the symmetric pure strategy Nash equilibrium $(1-2c, 1-2c)$. When firms' capacities are big, $c > \frac{1}{4}$, firms should put the price 0.5. This follows from Propositions 1 and 2.

In Propositions 1 and 2, symmetric equilibria are considered when both firms put the same price x. However, Lemmas 1 and 2 can be used in order to answer the question about existence of nonsymmetric weak active equilibria. Do profiles $(\bar{x}_1, \bar{x}_2)$, $\bar{x}_1 \neq \bar{x}_2$ that are weak active equilibria exist?

Proposition 3. *Assume that $0 < c \leqslant \frac{1}{4}$. Then there is no profile (x_1, x_2) of actions such that $x_1 \neq x_2$, $f(x_1, x_2) \geqslant \mu$, $f(x_2, x_1) \geqslant \mu$.*

Proof. Suppose that the function f is defined by (3). However, the domain of the function is not the unit square $0 \leqslant x_1 \leqslant 1$, $0 \leqslant x_2 \leqslant 1$ but the unit square with the carved segment $x_1 = x_2$. This means that the values of the function f are ignored for $x_1 = x_2$.

Note that the inequality

$$c < \left(1 - \frac{c}{1-x_2}\right)(1-x_1)$$

holds for $x_1 < 1-2c$, $x_2 < x_1$. Then it follows from (3) that we have the inequality $f(x_1, x_2) < \mu$ for the income of the firm 1 for $x_1 < 1-2c$, $x_2 \neq x_1$. Similarly, we have the inequality $f(x_2, x_1) < \mu$ for the income of the firm 2 for $x_2 < 1-2c$, $x_1 \neq x_2$. So in the future in proving the theorem, we consider only profiles (x_1, x_2) such that $x_1 \geqslant 1-2c$, $x_2 \geqslant 1-2c$.

Note that the inequality

$$c > \left(1 - \frac{c}{1-x_2}\right)(1-x_1) \tag{4}$$

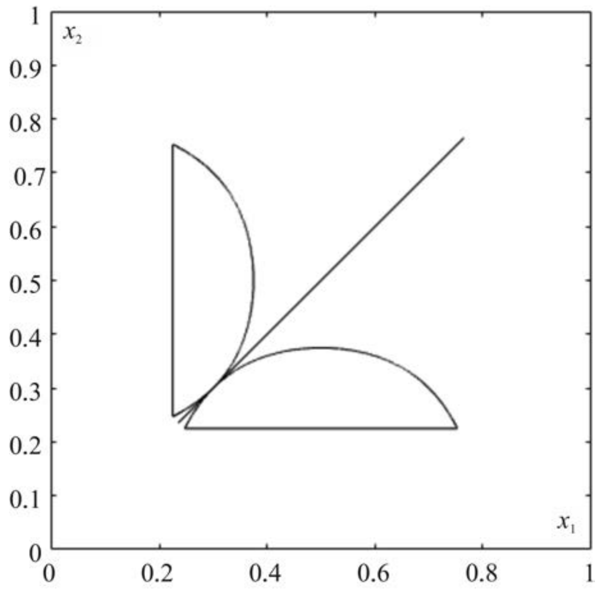

Fig. 1. Weak active equilibria for $c = 0.4$.

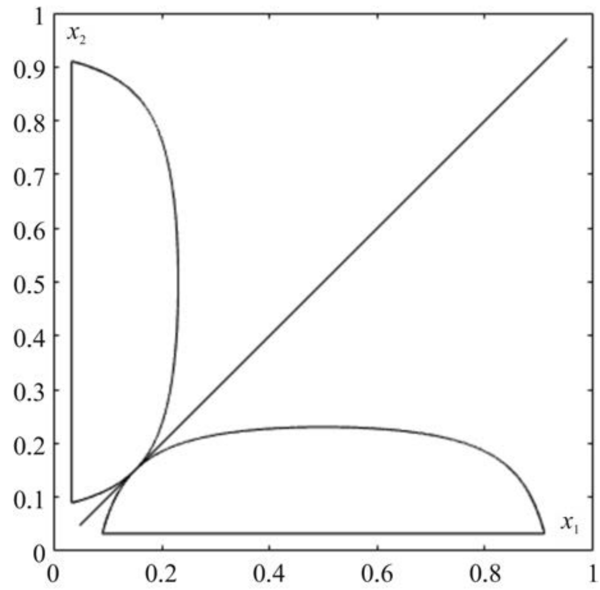

Fig. 2. Weak active equilibria for $c = 0.7$.

holds for $x_1 > 1 - 2c$, $x_2 = 1 - 2c$. Thus, we have $f(x_1, x_2) = 0.5x_1(1 - x_1)$ in this case. Hence, $f(x_1, x_2) < \mu$. Also, inequality (4) holds for $1 - 2c < x_2 \leqslant 1 - c$, $x_1 > x_2$. Thus, $f(x_1, x_2) < 0.5x_1(1 - x_1) < \mu$. Similarly, $f(x_2, x_1) < \mu$ for $1 - 2c \leqslant x_1 \leqslant 1 - c$, $x_2 > x_1$.

Finally, it follows from (3) that $f(x_1, x_2) = 0$ for $x_2 > 1 - c$, $x_1 > x_2$. Similarly, $f(x_2, x_1) = 0$ for $x_1 > 1 - c$, $x_2 > x_1$.

This completes the proof of Proposition 3.

It follows from Lemma 1 and Proposition 3 that there are no nonsymmetric weak active equilibria for $0 < c \leqslant \frac{1}{4}$. Calculations show that nonsymmetric weak active equilibria exist for $c > \frac{1}{4}$. By γ denote the ratio of the number of profiles (x_1, x_2), $x_1 > x_2$ that are weak active equilibria to the total number of profiles (x_1, x_2), $x_1 > x_2$. (The fine grid is used.) Values γ for some c are given in the table.

Values γ for different c

c	0.3	0.4	0.5	0.6	0.7	0.8	0.9
γ	0.0063	0.1113	0.2059	0.2586	0.2684	0.2349	0.1506

It is seen from Figs. 1 and 2 that the set of profiles (x_1, x_2) that are weak active equilibria consists of a part of the diagonal of the unit square (symmetric weak active equilibria) and two lunes that are adjacent to this diagonal (nonsymmetric weak active equilibria).

CONCLUSION

A study of Bertrand–Edgeworth oligopoly is of great practical importance. This is the approach used to analyse competition policy by authorities in a number of countries ([13]). In this paper, the concept of the weak active equilibrium is being used for the first time to study Bertrand–Edgeworth oligopoly. Its use makes it possible to consider only pure strategy equilibria and not consider mixed strategy equilibria like one have to do when using Nash equilibria. All symmetric weak active equilibria for the Beckmann's Bertrand–Edgeworth duopoly model are found. The question of the existence of nonsymmetric weak active equilibria is investigated.

The weak active equilibrium is not unique except for the cases when the weak active equilibrium coincides with the pure strategy Nash equilibrium. In this connection there is an interesting question about the narrowing of the concept of the weak active equilibrium, indication of some properties that allow to select one profile from the set of weak active equilibria. For example, the narrowing of the concept of the weak active equilibrium is the concept of the strong active equilibrium, which is introduced in [17]. But most likely, the latter concept is not suitable for Bertrand–Edgeworth oligopoly.

FUNDING

This work was supported in part by the Academician Nikolai Fedorenko International Scientific Foundation of Economic Research, project no. 2022-139.

REFERENCES

1. Beckmann, M.J. (with the assistance of D. Hochstädter), Edgeworth–Bertrand Duopoly Revisited, *Oper. Res. – Verfahren III* (R. Henn, Ed.) Meisenheim: Verlag Anton Hein, 1965, pp. 55–68.
2. Cheviakov, A.F. and Hartwick, J.M., Beckmann's Edgeworth–Bertrand Duopoly Example Revisited, *Int. Game Theory Rev.*, 2005, vol. 7, pp. 461–472.
3. Levitan, R.E. and Shubik, M., Price Duopoly and Capacity Constraints, *Int. Econom. Rev.*, 1972, vol. 13, pp. 111–122.
4. Kreps, D.M. and Scheinkman, J.A., Quantity Precommitment and Bertrand Competition Yield Cournot Outcomes, *Bell J. Econom.*, 1983, vol. 14, pp. 326–337.
5. Davidson, C. and Deneckere, R. Long-run Competition in Capacity, Short-run Competition in Price, and the Cournot Model, *Rand J. Econom.*, 1986, vol. 17, pp. 404–415.
6. Tasnádi, A., Existence of Pure Strategy Nash Equilibrium in Bertrand–Edgeworth Oligopolies, *Econom. Lett.*, 1999, vol. 63, pp. 201–206.
7. Lepore, J.J., Cournot Outcomes under Bertrand–Edgeworth Competition with Demand Uncertainty, *J. Mathem. Econom.*, 2012, vol. 48, pp. 177–186.
8. Sun, C.-J., Dynamic Price Dispersion in Bertrand–Edgeworth Competition, *Int. J. Game Theory.*, 2017, vol. 46, pp. 235–261.
9. Chen, Z. and Li, G., Horizontal Mergers in the Presence of Capacity Constraints, *Econom. Inquiry*, 2018, vol. 56, pp. 1346–1356.
10. Grüb, J.T., Can Mergers Lead to Partial Collusion? Introducing Heterogeneous Discount Factors to a Bertrand–Edgeworth Model, *Eur. Compet. J.*, 2020, vol. 16, pp. 512–530.
11. Somogyi, R., Bertrand–Edgeworth Competition with Substantial Horizontal Product Differentiation, *Math. Soc. Sci.*, 2020, vol. 108, pp. 27–37.
12. Bos, I. and Vermeulen, D., On Pure Strategy Nash Equilibria in Price-quantity Games, *J. Mathem. Econom.*, 2021, vol. 96, art. no. 102501.
13. Edwards, R.A. and Routledge, R.R., Information, Bertrand–Edgeworth Competition and the Law of One Price, *J. Mathem. Econom.*, 2022, vol. 101, art. no. 102658.
14. Bos, I. and Marini, M.A., Oligopoly Pricing: The role of Firm Size and Number, *Games*, 2023, vol. 14, art. no. 3.
15. Iskakov, A.B. and Iskakov, M.B., Equilibrium in Secure Strategies in the Bertrand–Edgeworth Duopoly, *Autom. Remote Control*, 2016, vol. 77, pp. 2239–2248.
16. Smol'yakov, E.R., Extremality and Equilibrium for Specified Functions and Game Models. I. Weak Equilibria, *Autom. Remote Control*, 1983, vol. 44, pp. 358–363.
17. Smol'yakov, E.R., Extremality and Equilibrium for Specified Functions and Game Models. II. Strong Equilibria, *Autom. Remote Control*, 1983, vol. 44, pp. 440–446.
18. Shvedov, A.S., Cournot Oligopoly: Strategy Choice under Uncertainty and Other Problems, *Control Sciences*, 2023, no. 5, pp. 18–31.

This paper was recommended for publication by F.T. Aleskerov, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 8, pp. 809–818.

OPTIMIZATION, SYSTEM ANALYSIS, AND OPERATIONS RESEARCH

Criteria Convolutions When Combining the Solutions of the Multicriteria Axial Assignment Problem

L. G. Afraimovich[*,a] **and M. D. Emelin**[*,b]

Lobachevsky Nizhny Novgorod State University, Nizhny Novgorod, Russia
e-mail: [a]*levafraimovich@gmail.com*, [b]*makcum888e@mail.ru*

Received March 4, 2024
Revised May 21, 2024
Accepted June 27, 2024

Abstract—This paper is devoted to a classical NP-hard problem, known as the three-index axial assignment problem. Within the corresponding framework, the problem of combining feasible solutions is posed as an assignment problem on the set of solutions containing only the components of selected feasible solutions. The issues of combining solutions for the multicriteria problem with different criteria convolutions are studied. In the general case, the combination problem turns out to be NP-hard. Polynomial solvability conditions are obtained for the combination problem.

Keywords: axial assignment problem, multi-index problems, combining solutions, polynomial solvability, NP-hardness

DOI: 10.31857/S0005117924080064

1. INTRODUCTION

There is a wide class of applied problems formalized by multi-index axial assignment problems [1–4]. In the general case (no additional constraints), the class of multi-index axial assignment problems is NP-hard [5]. Particular polynomially solvable subclasses and subclasses with polynomial approximation algorithms are known [1, 6, 7]. Axial assignment problems with a special structure of the multi-index cost matrix were studied in [6–10]; the branch-and-bound method for solving the axial assignment problem was described in [11]; a parallel implementation of the solution algorithm was discussed in [12]; a genetic algorithm was developed in [13]; a lower bound for the problem was considered in [14]; asymptotically optimal solutions were investigated in [15]. Multicriteria formulations of multi-index assignment problems were addressed in [16–19]. The problem of combining solutions of the axial assignment problem was investigated in [20–22].

This paper is devoted to the multicriteria three-index axial assignment problem. Within the corresponding framework, we formulate the problem of combining feasible solutions as a multicriteria assignment problem on the set of solutions obtained by combining the components of given feasible solutions. The convolution of criteria is considered as a compromise scheme for solving multicriteria problems. We consider linear, minimax, and lexicographic convolutions.

For these types of convolutions, the problems of combining N feasible solutions are studied. The following results are established below:

—In the linear convolution case, the combination problem is polynomially solvable for $N = 2$ and NP-hard for $N \geqslant 4$.

—In the lexicographic convolution case, the combination problem is polynomially solvable for $N = 2$ and NP-hard for $N \geqslant 4$.

—In the minimax convolution case, the combination problem is NP-hard for $N \geqslant 2$.

This research continues the series of papers [20–22], where combination was studied when solving the (single-criterion) axial assignment problem. Here, combining solutions means constructing a solution that contains only assignments from selected feasible solutions. In turn, solving the combination problem involves searching for an optimal solution, in terms of a selected criterion or compromise scheme, on the set of solutions obtained by combining the selected feasible solutions.

Note that the original multicriteria three-index axial assignment problem with the considered types of convolutions as a compromise scheme is NP-hard. Thus, the combination algorithms proposed below for the polynomially solvable cases can be applied as a complement to heuristic or approximate algorithms for solving the original NP-hard problems. Many heuristic or approximate algorithms for solving assignment problems are based on constructing a series of feasible solutions with further record selection among them; for example, see [1, 3, 6]. As an alternative to the conventional approach of selecting a record among the feasible solutions found, we use a step of combining the constructed feasible solutions. In this case, the solution obtained by combination is surely not worse than the record in terms of the optimization criterion: the resulting set of combination-based solutions also contains all feasible solutions found initially.

The remainder of the paper is organized as follows. Section 2 presents the formal statement of the multicriteria three-index axial assignment problem and describes the types of criteria convolutions used. In Section 3, the problem of combining feasible solutions is posed. In Sections 4, 5, and 6, we investigate combination problems for the linear, lexicographic, and minimax criteria convolutions, respectively. Section 7 provides the results of computational experiments.

2. THE MULTICRITERIA THREE-INDEX AXIAL ASSIGNMENT PROBLEM

Let I, J, and K be disjoint index sets, $I \cap J = \emptyset$, $I \cap K = \emptyset$, $J \cap K = \emptyset$, and $|I| = |J| = |K| = n$; M is a fixed number of problem criteria; c^u_{ijk}, where $i \in I$, $j \in J$, $k \in K$, and $u = \overline{1, M}$, are three-index cost matrices; x_{ijk}, where $i \in I$, $j \in J$, and $k \in K$, is the three-index matrix of the variables. Then the multicriteria three-index axial assignment problem is formulated as the following integer linear programming problem:

$$\sum_{i\in I}\sum_{j\in J} x_{ijk} = 1, \quad k \in K, \tag{1}$$

$$\sum_{i\in I}\sum_{k\in K} x_{ijk} = 1, \quad j \in J, \tag{2}$$

$$\sum_{j\in J}\sum_{k\in K} x_{ijk} = 1, \quad i \in I, \tag{3}$$

$$x_{ijk} \in \{0, 1\}, \quad i \in I, \quad j \in J, \quad k \in K, \tag{4}$$

$$\sum_{i\in I}\sum_{j\in J}\sum_{k\in K} c^u_{ijk} x_{ijk} \to \min, \quad u = \overline{1, M}. \tag{5}$$

For the sake of convenience, let Z_M denote the multicriteria problem (1)–(5). As is known, in the single-criterion formulation (for $M = 1$), the assignment problem Z_M is NP-hard [5].

In the multicriteria formulation, we will consider the convolution of criteria as a compromise scheme. The linear, lexicographic, and minimax convolutions will be studied below.

Given weights α_u, $u = \overline{1, M}$, the linear convolution of criteria has the form

$$\sum_{u=1}^{M} \alpha_u \sum_{i\in I}\sum_{j\in J}\sum_{k\in K} c^u_{ijk} x_{ijk} \to \min. \tag{6}$$

For the sake of convenience, let Z_L denote problem (1)–(4), (6). Obviously, the NP-hard (single-criterion) three-index axial assignment problem is polynomially reducible to the problem Z_L. Indeed, consider the three-index axial assignment problem with a cost matrix c_{ijk}, where $i \in I$, $j \in J$, and $k \in K$. In the corresponding problem Z_L, we define $c^1_{ijk} = c_{ijk}$, $c^u_{ijk} = 0$, $\alpha_1 = 1$, $\alpha_u = 0$, where $i \in I$, $j \in J$, $k \in K$, and $u = \overline{2, M}$. Then the following result is true.

Proposition 1. *The problem Z_L is NP-hard.*

Assume that the order of significance of the criteria (5) coincides with their original indexing. We introduce a preference relation on the set of feasible solutions of the assignment problem. Let P denote the set of feasible solutions of the system of constraints (1)–(4). For $x^1, x^2 \in P$, we write $x^1 \preccurlyeq x^2$ if and only if there exists $y \in \{1, \ldots, M\}$ such that

$$\sum_{i\in I}\sum_{j\in J}\sum_{k\in K} c^u_{ijk} x^1_{ijk} = \sum_{i\in I}\sum_{j\in J}\sum_{k\in K} c^u_{ijk} x^2_{ijk}, \quad u \in \{1, \ldots, y\}, \tag{7}$$

$$\sum_{i\in I}\sum_{j\in J}\sum_{k\in K} c^u_{ijk} x^1_{ijk} < \sum_{i\in I}\sum_{j\in J}\sum_{k\in K} c^u_{ijk} x^2_{ijk}, \quad u \in \{1, \ldots, M\} \cap \{y+1\} \tag{8}$$

(i.e., conditions (7) and (8) hold). As a result, the multicriteria problem with the lexicographic convolution of criteria is to find x^* satisfying the system of constraints

$$x^* \in P, \tag{9}$$

$$x^* \preccurlyeq x, \quad x \in P. \tag{10}$$

For the sake of convenience, let $Z_{\preccurlyeq}$ denote problem (9), (10). Obviously, the NP-hard three-index axial assignment problem is polynomially reducible to the problem $Z_{\preccurlyeq}$. In the corresponding problem $Z_{\preccurlyeq}$, we define $c^1_{ijk} = c_{ijk}$ and $c^u_{ijk} = 0$, where $i \in I$, $j \in J$, $k \in K$, and $u = \overline{2, M}$. Then the following result is true.

Proposition 2. *The problem $Z_{\preccurlyeq}$ is NP-hard.*

The minimax convolution of criteria has the form

$$\max_{u\in\{1,\ldots,M\}} \left(\sum_{i\in I}\sum_{j\in J}\sum_{k\in K} c^u_{ijk} x_{ijk} \right) \to \min\,. \tag{11}$$

For the sake of convenience, let Z_{minmax} denote problem (1)–(4), (11). Obviously, the NP-hard three-index axial assignment problem is polynomially reducible to the problem Z_{minmax}. In the corresponding problem Z_{minmax}, we define $c^u_{ijk} = c_{ijk}$, where $i \in I$, $j \in J$, $k \in K$, and $u = \overline{1, M}$. Then the following result is true.

Proposition 3. *The problem Z_{minmax} is NP-hard.*

Thus, the multicriteria axial assignment problem with the linear, lexicographical, or minimax convolution of criteria as a compromise scheme is NP-hard. This raises the question of combining feasible solutions as a complement to heuristic or approximate algorithms for solving these NP-hard problems.

3. THE PROBLEM OF COMBINING SOLUTIONS

Let a given set $W \subseteq I \times J \times K$ define a subset of allowed assignments. Consider problem (1)–(4), (12), (5) with

$$x_{ijk} = 0, \quad (i, j, k) \notin W. \tag{12}$$

For the sake of convenience, let $Z_M(W)$ denote the multicriteria problem (1)–(4), (12), (5) with a given set W. Obviously, problem (1)–(5) corresponds to the problem $Z_M(I \times J \times K)$.

Next, we consider the multicriteria problem $Z_M(W)$ with different criteria convolutions as a compromise scheme.

In the linear convolution case, the corresponding problem takes the form (1)–(4), (12), (6), further denoted by $Z_L(W)$.

In the lexicographic convolution case, the corresponding problem takes the form (9), (10), where the set P is the set of feasible solutions for the system of constraints (1)–(4), (12). It will be denoted by $Z_{\preccurlyeq}(W)$.

In the minimax convolution case, the corresponding problem takes the form (1)–(4), (12), (11), further denoted by $Z_{\text{minmax}}(W)$.

Checking the consistency of the system of constraints in the problem $Z_M(W)$ with an arbitrary set W, i.e., system (1)–(4), (12), is an NP-complete problem [1]. We will consider the sets W corresponding to the assignments of a given subset of feasible solutions.

Let x_{ijk}, where $i \in I$, $j \in J$, and $k \in K$, be a feasible solution of the system of constraints (1)–(4). Then $W(x)$ will denote the following set of allowed assignments:

$$W(x) = \{(i, j, k) | x_{ijk} = 1, i \in I, j \in J, k \in K\}.$$

Let $x^1_{ijk}, \dots, x^r_{ijk}$, where $i \in I$, $j \in J$, and $k \in K$, be arbitrary r feasible solutions of the system of constraints (1)–(4). Then

$$W(x^1, \dots, x^r) = W(x^1) \cup \dots \cup W(x^r).$$

Below, we will investigate the problems

$$Z_L(W(x^1, \dots, x^r)), \quad Z_{\preccurlyeq}(W(x^1, \dots, x^r)), \quad Z_{\text{minmax}}(W(x^1, \dots, x^r)).$$

They are combination problems, i.e., optimization ones on the set of allowed assignments $W(x^1, \dots, x^r)$, which is built by combining the factual assignments of the selected feasible solutions $x^1, \dots, x^r$. Thus, the solution of combination problems corresponds to the solution containing only the assignments of the selected feasible solutions.

4. THE LINEAR CONVOLUTION OF CRITERIA

Consider the multicriteria assignment problem in the case of the linear convolution of criteria. Obviously, the problem $Z_L(W(x^1, \dots, x^r))$ is equivalent to the single-criterion three-index axial problem with the cost matrix

$$c_{ijk} = \sum_{u \in \{1,\dots,M\}} \alpha_u c^u_{ijk}, \quad i \in I, \quad j \in J, \quad k \in K.$$

Thus, according to [20], the problem $Z_L(W(x^1, x^2))$ is polynomially solvable. The combination algorithm proposed in [20] requires $O(n)$ computational operations and can be applied to solve the problem $Z_L(W(x^1, x^2))$. Based on [22], the following result is true.

Proposition 4. *For $r \geqslant 4$, the class of problems $Z_L(W(x^1, \dots, x^r))$ is NP-hard.*

For the time being, the complexity status of the class of problems $Z_L(W(x^1, x^2, x^3))$ remains unknown.

5. THE LEXICOGRAPHIC CONVOLUTION OF CRITERIA

We present an algorithm for solving the multicriteria assignment problem in the case of the lexicographic convolution of criteria when combining two feasible solutions.

Algorithm 1 (solution of the problem $Z_{\preccurlyeq}(W(x^1, x^2))$).

Step 1. Construct a graph $G = (V, A)$, where

$$V = \{I \cup J \cup K\}, \quad A = \Big\{(i,j), (i,k), (j,k) | (i,j,k) \in W(x^1, x^2)\Big\}.$$

Step 2. Find the connected components V_l, $l = \overline{1,q}$, of the graph G and construct the subgraphs $G_l = (V_l, A_l)$, $l = \overline{1,q}$, generated by the corresponding connected components.

Step 3. Construct the sets

$$D_l^1 = \Big\{(i,j,k) | (i,j,k) \in W(x^1), (i,j), (i,k), (j,k) \in A_l\Big\}, \quad l = \overline{1,q},$$
$$D_l^2 = \Big\{(i,j,k) | (i,j,k) \in W(x^2), (i,j), (i,k), (j,k) \in A_l\Big\}, \quad l = \overline{1,q}.$$

Step 4. Let

$$P_l^1 = \left\{ p | p = \operatorname*{argmin}_{p \in \{1,2\}} \sum_{(i,j,k) \in D_l^p} c_{ijk}^1 \right\},$$

$$P_l^u = \left\{ p | p = \operatorname*{argmin}_{p \in P_l^{u-1}} \sum_{(i,j,k) \in D_l^p} c_{ijk}^u \right\}, \quad u = \overline{2, M-1},$$

$$p_l^* = \operatorname*{argmin}_{p \in P_l^{M-1}} \sum_{(i,j,k) \in D_l^p} c_{ijk}^M, \quad l = \overline{1,q}.$$

Step 5. Determine the solution of the problem $Z_{\preccurlyeq}(W(x^1, x^2))$ using the following algorithm. Let $x_{ijk}^* := 0$, where $i \in I$, $j \in J$, and $k \in K$. For each $l = \overline{1,q}$, execute

$$x_{ijk}^* := 1, \quad (i,j,k) \in D_l^{p_l^*}.$$

The value of the criteria (5) on a solution x^* is given by

$$\sum_{l=1}^{q} \sum_{(i,j,k) \in D_l^{p_l^*}} c_{ijk}^u, \quad u = \overline{1,M}.$$

Theorem 1. *Algorithm 1 outputs the solution of the problem* $Z_{\preccurlyeq}(W(x^1, x^2))$.

Proof. Assume on the contrary that the output of Algorithm 1 is not a solution of the problem $Z_{\preccurlyeq}(W(x^1, x^2))$. Then there exists a feasible solution x of the problem $Z_{\preccurlyeq}(W(x^1, x^2))$ such that the condition $x^* \preccurlyeq x$ is false.

In each connected component V_l, $l = \overline{1,q}$, any feasible solution of the problem $Z_{\preccurlyeq}(W(x^1, x^2))$ may contain assignments consisting of triplets of only the first or only the second solution; see the proof of [20, Theorem 1].

By the assumption above, there exists a connected component V_l in which the assignments of the solution x do not coincide with the assignments of the solution x^*. Then we construct the solution x' as follows:

Step 1. $x^0 = x,\ l = 1.$

Step 2. $x_{ijk}^t = \begin{cases} 1 \ \text{ if } \ (i,j,k) \in D_l^{p_l^*} \\ 0 \ \text{ if } \ (i,j,k) \in D_l^{3-p_l^*} \\ x_{ijk}^{t-1} \ \text{ otherwise,} \end{cases} \qquad i \in I,\ j \in J,\ k \in K,$

Step 3. If $l = q$, then stop; otherwise, $l = l + 1$ and return to Step 2.

By the construction procedure, $x^{l+1} \preccurlyeq x^l$, $l = \overline{0, q-1}$. In addition, $x^q = x^*$ and $x^0 = x$. Hence, we arrive at the contradiction $x^* \preccurlyeq x$, and the proof of Theorem 1 is complete.

According to [20, Theorem 2], the following result is true.

Proposition 5. *Algorithm 1 requires $O(n)$ computational operations.*

Obviously, the class of three-index axial problems with the set of allowed assignments $W(x^1, x^2, x^3, x^4)$ is polynomially reducible to the class of problems $Z_{\preccurlyeq}(W(x^1, x^2, x^3, x^4))$. As was established in [22], the former class is NP-hard. Indeed, consider the three-index axial assignment problem with the cost matrix c_{ijk}, where $i \in I$, $j \in J$, and $k \in K$. When performing reduction in the corresponding problem $Z_{\preccurlyeq}(W(x^1, x^2, x^3, x^4))$, we define $c_{ijk}^1 = c_{ijk}$ and $c_{ijk}^u = 0$, where $i \in I$, $j \in J$, $k \in K$, and $u = \overline{2, M}$. Then the following result is true.

Proposition 6. *For $r \geqslant 4$, the class of problems $Z_{\preccurlyeq}(W(x^1, \ldots, x^r))$ is NP-hard.*

For the time being, the complexity status of the class of problems $Z_{\preccurlyeq}(W(x^1, x^2, x^3))$ remains unknown.

6. THE MINIMAX CONVOLUTION OF CRITERIA

Consider the multicriteria assignment problem in the case of the minimax convolution of criteria.

Lemma 1. *The optimization problem*

$$x'_i \in \{0, 1\}, \quad i = \overline{1, n}, \tag{13}$$

$$\max\left(\sum_{i=1}^n a_i x'_i, \sum_{i=1}^n b_i(1 - x'_i)\right) \to \min \tag{14}$$

is NP-hard.

Proof. To prove this lemma, we polynomially reduce the classical NP-complete PARTITION problem [5] to problem (13), (14). Consider the PARTITION problem with the initial parameters w_i, $i = \overline{1, m}$, letting $n = m$ and $a_i = b_i = w_i$, $i = \overline{1, n}$. The optimal value of the criterion of problem (13), (14) is $\frac{1}{2}\sum_{i=1}^m w_i$ if and only if the PARTITION problem has a solution. Therefore, problem (13), (14) is NP-hard. The proof of Lemma 1 is complete.

Theorem 2. *The class of problems $Z_{\text{minmax}}(W(x^1, x^2))$ is NP-hard.*

Proof. We show the polynomial reducibility of the NP-hard problem (13), (14) to the class of problems $Z_{\text{minmax}}(W(x^1, x^2))$.

Let $N = 2n$, $I = J = K = \{1, \ldots, N\}$, and $M = 2$. We define the three-index cost matrices c_{ijk}^1 and c_{ijk}^2, where $i \in I$, $j \in J$, and $k \in K$, as follows:

$$c_{ijk}^1 = \begin{cases} a_q \text{ if } \exists q \in \{1, \ldots, n\} \text{ such that } i = j = k = 2q - 1 \\ 0 \quad \text{otherwise,} \end{cases} \qquad i \in I,\ j \in J,\ k \in K,$$

$$c_{ijk}^2 = \begin{cases} b_q \text{ if } \exists q \in \{1, \ldots, n\} \text{ such that } i = j = k + 1 = 2q \\ 0 \quad \text{otherwise,} \end{cases} \qquad i \in I,\ j \in J,\ k \in K.$$

We construct two subsets $P_1, P_2 \subseteq I \times J \times K$ determining two feasible solutions of system (1)–(4):

$$P_1 = \{(i,i,i) | i = \overline{1,N}\},$$
$$P_2 = \{(2i-1, 2i-1, 2i), (2i, 2i, 2i-1) | i = \overline{1,n}\}.$$

We define the corresponding two feasible solutions x^1, x^2 of system (1)–(4) by

$$x^t_{ijk} = \begin{cases} 1 & \text{if } (i,j,k) \in P_t \\ 0 & \text{otherwise,} \end{cases} \quad i \in I,\ j \in J,\ k \in K,\ t \in \{1,2\}.$$

Next, consider the corresponding combination problem $Z_{\text{minmax}}(W(x^1, x^2))$. It is necessary to demonstrate that the optimal value of the criterion of problem (13), (14) coincides with that of problem $Z_{\text{minmax}}(W(x^1, x^2))$.

1. Let x'^* be the optimal solution of problem (13), (14). We construct $P(x'^*)$ using the following algorithm:

Step 1. $P(x'^*) = \emptyset$.

Step 2. For each $i = \overline{1,n}$:
if $x'^*_i = 1$, then $P(x'^*) = P(x'^*) \cup \{(2i-1, 2i-1, 2i-1), (2i, 2i, 2i)\}$;
otherwise, $P(x'^*) = P(x'^*) \cup \{(2i-1, 2i-1, 2i), (2i, 2i, 2i-1)\}$.

As is easily verified, the value of the criterion of the problem $Z_{\text{minmax}}(W(x^1, x^2))$ on the solution corresponding to $P(x'^*)$ coincides with that of problem (13), (14). Assume on the contrary that the solution corresponding to $P(x'^*)$ is not optimal in the problem $Z_{\text{minmax}}(W(x^1, x^2))$. Let x^* be the optimal solution of the problem $Z_{\text{minmax}}(W(x^1, x^2))$. Then the value of the criterion of the problem $Z_{\text{minmax}}(W(x^1, x^2))$ on the solution x^* is strictly smaller than that of problem (13), (14) on the solution x'^*. Then we construct a feasible solution x' of problem (13), (14) as follows:

$$x'_i = \begin{cases} 1 & \text{if } x^*_{iii} = 1 \\ 0 & \text{otherwise,} \end{cases} \quad i = \overline{1,n}.$$

By the construction procedure, the value of the criterion of problem (13), (14) on the solution x' coincides with that of the problem $Z_{\text{minmax}}(W(x^1, x^2))$ on the solution x^*. Consequently, x'^* is not an optimal solution of problem (13), (14). This contradiction proves that the value of the criterion of the problem $Z_{\text{minmax}}(W(x^1, x^2))$ on the solution corresponding to $P(x'^*)$ will coincide with that of the problem (13), (14).

2. Let x^* be the optimal solution to the problem $Z_{\text{minmax}}(W(x^1, x^2))$. Then we construct the optimal solution of problems (13), (14) as follows:

$$x'_i = \begin{cases} 1 & \text{if } x^*_{iii} = 1 \\ 0 & \text{otherwise,} \end{cases} \quad i = \overline{1,n}.$$

As is easily verified, the value of the criterion of problem (13), (14) on the solution x' coincides with the optimal criterion of the problem $Z_{\text{minmax}}(W(x^1, x^2))$.

Assume on the contrary that the solution x' is not optimal in problem (13), (14). Let x'^* be the optimal solution of problem (13), (14). Then the value of the criterion of problem (13), (14) on the solution x'^* is strictly smaller than that of the problem $Z_{\text{minmax}}(W(x^1, x^2))$ on the solution x^*. We construct the solution of the problem $Z_{\text{minmax}}(W(x^1, x^2))$ using the following algorithm:

Step 1. $P(x'^*) = \emptyset$.

Step 2. For each $i = \overline{1,n}$:
if $x'^*_i = 1$, then $P(x'^*) = P(x'^*) \cup \{(2i-1, 2i-1, 2i-1), (2i, 2i, 2i)\}$;
otherwise, $P(x'^*) = P(x'^*) \cup \{(2i-1, 2i-1, 2i), (2i, 2i, 2i-1)\}$.

By the construction procedure, the value of the criterion of the problem $Z_{\mathrm{minmax}}(W(x^1, x^2))$ on the solution corresponding to $P(x'^*)$ coincide with that of problem (13), (14) on the solution x'^*. Therefore, x^* is not the optimal solution of the problem $Z_{\mathrm{minmax}}(W(x^1, x^2))$. This contradiction proves that the value of the criterion of problem (13), (14) on the solution x' will coincide with the optimal value of the criterion of the problem $Z_{\mathrm{minmax}}(W(x^1, x^2))$.

Thus, problem (13), (14) is polynomially reducible to the class of problems $Z_{\mathrm{minmax}}(W(x^1, x^2))$. Consequently, the class of problems $Z_{\mathrm{minmax}}(W(x^1, x^2))$ is NP-hard. The proof of Theorem 2 is complete.

7. A COMPUTATIONAL EXPERIMENT

Consider the problem $Z_{\preccurlyeq}$ for $M = 2$. By analogy with [12], we construct a test set with three-index cost matrices whose elements are random integer values with the uniform distribution on the interval [0, 300]. Let us conduct a series of experiments for fixed values of n, denoting by K the number of problems in the series. We design two heuristic algorithms for solving the problem $Z_{\preccurlyeq}$. The first algorithm is based on constructing subsets of feasible solutions and selecting a record among them; the second one is similar to the first, but the record selection step is replaced by the sequential combination of solutions using the first algorithm. Note that by Proposition 5, the first algorithm has a complexity of $O(n)$, which coincides with the complexity of record selection.

The first heuristic algorithm. Construct $N = n^3$ random solutions, and apply the local optimization algorithm [13] to each of them. Denote by x'_t, $t = \overline{1, N}$, the resulting feasible solutions of the problem $Z_{\preccurlyeq}$. Among them, choose the record $x^* : x^* \preccurlyeq x'_t$, $t = \overline{1, N}$, as the output of the algorithm.

The second heuristic algorithm. Replace the record selection step of the first heuristic algorithm with the sequential combination of the pairs of solutions as follows. Let x''_1 be the solution of the problem $Z_{\preccurlyeq}(W(x'_1, x'_2))$. Denote by x''_t the solution of the problem $Z_{\preccurlyeq}(W(x''_{t-1}, x'_{t+1}))$, $t = \overline{2, N-1}$. Choose x''_{N-1} as the output of the algorithm.

Denoting by $C^1(x)$ and $C^2(x)$ the values of the criteria (5) on the solution x, we will compare the average deviation of the criteria values under combination (the second heuristic algorithm) and record choice (the first heuristic algorithm) in the series. The results are presented in Table 1.

Table 1

n	K	$100\% \dfrac{C^1(x^*) - C^1(x''_{N-1})}{C^1(x^*)}$	$100\% \dfrac{C^2(x^*) - C^2(x''_{N-1})}{C^2(x^*)}$
10	10	3.09%	1.98%
11	10	5.3%	−6.23%
12	10	5.31%	6.58%
13	10	2.05%	0.46%
14	10	0%	0%
15	10	1.26%	−1.85%
16	10	2.5%	2.14%
17	10	2.74%	−4.92%
18	10	3.19%	−0.7%
19	10	7.08%	2.3%

Thus, the average deviations over all series are 3.94% and 0.7% for the first and second criteria, respectively. This demonstrates the effectiveness of the combination strategy instead of the generally accepted record selection strategy.

The linear convolution case is equivalent to the single-criterion problem with the linear criterion and was considered in [20, 21], including the results of a computational experiment. The minimax convolution case is not considered in the computational experiment here: by Theorem 2, the combination problem with the minimax convolution is NP-hard even when combining two solutions.

8. CONCLUSIONS

Solution algorithms for the NP-hard multicriteria three-index axial assignment problem with different types of criteria convolutions have been studied. The problem of combining feasible solutions of this problem as a compromise scheme has been formulated. Solutions can be combined as a complement to known heuristics or approximate algorithms for post-processing the obtained approximate solutions of the assignment problem instead of the generally accepted practice of record selection.

The complexity status of combining solutions has been investigated. It has been shown that, in the case of the linear or lexicographic convolution, the pairs of solutions can be completely combined in a time of $O(n)$; the class of combination problems for four or more solutions is NP-hard; the complexity status of combining three solutions is an open issue. In the minimax convolution case, the class of combination problems for two or more solutions is NP-hard. For better clarity, the outcomes of this paper are presented in Table 2.

Table 2

	$Z_L(W(x^1,\dots,x^r))$	$Z_{\preccurlyeq}(W(x^1,\dots,x^r))$	$Z_{\text{minmax}}(W(x^1,\dots,x^r))$
$r=2$	$O(n)$	$O(n)$	NP-hard
$r=3$	?	?	
$r\geqslant 4$	NP-hard	NP-hard	

In addition, an algorithm for combining solution pairs has been designed in the lexicographic convolution case. According to the computational experiment, the combination strategy allows decreasing the deviation from the optimum as compared to the record selection strategy.

Further research will address the open cases of combining three solutions and combining solutions when constructing Pareto-optimal solutions of the multicriteria assignment problem.

REFERENCES

1. Spieksma, F.C.R., Multi Index Assignment Problems. Complexity, Approximation, Applications, in *Nonlinear Assignment Problems: Algorithms and Applications*, Pardalos, P.M. and Pitsoulis, L.S., Eds., Dordrecht: Kluwer Acad. Publishers, 2000, pp. 1–11.
2. Burkard, R., Dell'Amico, M., and Martello, S., *Assignment Problems*, PA: SIAM, 2012.
3. Kuroki, Y. and Matsui, T., An Approximation Algorithm for Multidimensional Assignment Problems Minimizing the Sum of Squared Errors, *Discret. Appl. Math.*, 2009, vol. 157, no. 9, pp. 2124–2135.
4. Poore, A.B., Multidimensional Assignment Problems Arising in Multitarget and Multisensor Tracking, in *Nonlinear Assignment Problems: Algorithms and Applications*, Pardalos, P.M. and Pitsoulis, L.S., Eds., Dordrecht: Kluwer Acad. Publishers, 2000, pp. 13–38.
5. Garey, M.R. and Johnson, D.S., *Computers and Intractability. A Guide to the Theory of NP-completeness*, San Francisco: Freeman, 1979.
6. Crama, Y. and Spieksma, F.C.R., Approximation Algorithms for Three-Dimensional Assignment Problems with Triangle Inequalities, *Eur. J. Oper. Res.*, 1992, vol. 60, pp. 273–279.
7. Bandelt, H.J., Crama, Y., and Spieksma, F.C.R., Approximation Algorithms for Multidimensional Assignment Problems with Decomposable Costs, *Discret. Appl. Math.*, 1994, vol. 49, pp. 25–50.

8. Burkard, R.E., Rudolf, R., and Woeginger, G.J., Three-Dimensional Axial Assignment Problems with Decomposable Cost Coefficients, *Discret Appl. Math.*, 1996, vol. 65, pp. 123–139.
9. Spieksma, F. and Woeginger, G., Geometric Three-Dimensional Assignment Problems, *Eur. J. Oper. Res.*, 1996, vol. 91, pp. 611–618.
10. Custic, A., Klinz, B., and Woeginger, G.J., Geometric Versions of the Three-Dimensional Assignment Problem under General Norms, *Discret. Optim.*, 2015, vol. 18, pp. 38–55.
11. Balas, E. and Saltzman, M.J., An Algorithm for the Three-Index Assignment Problem, *Oper. Res.*, 1991, vol. 39, no. 1, pp. 150–161.
12. Natu, S., Date, K., and Nagi, R., GPU-Accelerated Lagrangian Heuristic for Multidimensional Assignment Problems with Decomposable Costs, *Parallel Comput.*, 2020, vol. 97, art. no. 102666.
13. Huang, G. and Lim, A., A Hybrid Genetic Algorithm for the Three-Index Assignment Problem, *Eur. J. Oper. Res.*, 2006, vol. 172, pp. 249–257.
14. Kim, B.J., Hightower, W.L., Hahn, P.M., Zhu, Y.R., and Sun, L., Lower Bounds for the Axial Three-Index Assignment Problem, *Eur. J. Oper.*, 2010, vol. 202, pp. 654–668.
15. Dichkovskaya, S.A. and Kravtsov, M.K., Investigation of Polynomial Algorithms for Solving the Three-Index Planar Assignment Problem, *Comput. Math. and Math. Phys.*, 2006, vol. 46, no. 2, pp. 212–217.
16. Dichkovskaya, S.A. and Kravtsov, M.K., Investigation of Polynomial Algorithms for Solving the Multicriteria Three-Index Planar Assignment Problem, *Comput. Math. and Math. Phys.*, 2007, vol. 47, no. 6, pp. 1029–1038.
17. Emelichev, V.A. and Perepelitsa, V.A., Complexity of Discrete Multicriteria Problems, *Discrete Math. Appl.*, 1994, vol. 4, no. 2, pp. 89–118.
18. Prilutskii, M.Kh., Multicriterial Multi-index Resource Scheduling Problems, *J. Comput. Syst. Sci. Int.*, 2007, vol. 46, no. 1, pp. 78–82.
19. Prilutskij, M.Kh., Multicriteria Distribution of a Homogeneous Resource in Hierarchical Systems, *Autom. Remote Control*, 1996, vol. 57, no. 2, part 2, pp. 266–271.
20. Afraimovich, L.G. and Emelin, M.D., Combining Solutions of the Axial Assignment Problem, *Autom. Remote Control*, 2021, vol. 82, no. 8, pp. 1418–1425.
21. Afraimovich, L.G. and Emelin, M.D., Heuristic Strategies for Combining Solutions of the Three-Index Axial Assignment Problem, *Autom. Remote Control*, 2021, vol. 82, no. 10, pp. 1635–1640.
22. Afraimovich, L.G. and Emelin, M.D., Complexity of Solutions Combination for the Three-Index Axial Assignment Problem, *Mathematics*, 2022, vol. 10, no. 7, art. no. 1062.

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

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 8, pp. 819–834.

State Observer-Based Iterative Learning Control Design for Discrete Systems Using the Heavy Ball Method

P. Pakshin[*,a], **J. Emelianova**[*,b], **and E. Rogers**[**,c]

[*] *Arzamas Polytechnic Institute of R.E. Alekseev Nizhny Novgorod State Technical University, Arzamas, Russia*
[**] *School of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom*
e-mail: [a]*pakshinpv@gmail.com*, [b]*EmelianovaJulia@gmail.com*, [c]*etar@ecs.soton.ac.uk*

Received April 6, 2024
Revised June 27, 2024
Accepted July 1, 2024

Abstract—The paper considers a state observer-based iterative learning control design problem for discrete linear systems. To accelerate the convergence of the learning error, a combination of the heavy ball method from optimization theory and the vector Lyapunov function method for a class of two-dimensional systems known as repetitive processes is used to develop a new design. A supporting numerical example is given, including a comparison with an existing design.

Keywords: iterative learning control, repetitive processes, stability, convergence, state observer, heavy ball method, vector Lyapunov function, linear matrix inequalities

DOI: 10.31857/S0005117924080071

1. INTRODUCTION

Iterative learning control (ILC) arose from research aimed at increasing the accuracy of repetitive operations performed by robots [1], such as pick-and-place operations. Early results in this actively developing area can be found, e.g., in the survey papers [2, 3], with more recent results in, e.g., [4]. Currently, ILC designs are effectively used in additive manufacturing, in particular in high-precision multilayer laser deposition installation [5, 6], in medical robots for the rehabilitation of patients who have suffered a stroke [7, 8], in ventricular support devices [9], and in numerous other applications.

The class of systems that ILC can be applied to are those that iteratively execute the same finite-duration task. An example is a robot operating in a pick-and-place mode, where it collects a payload from a specified location, transfers it over a finite duration, deposits it, and then returns to the starting location to repeat this sequence of tasks. Each execution, or trial, is of a finite duration, termed the trial length. Once a trial is complete, all the information generated over the trial length is available to update the subsequent trial's control signal, highlighting the iterative nature of ILC. Batch processing is extensively used in chemical processing, where the same finite-duration operation is applied to a sequence of objects with a time gap between the end of processing one batch and the start of processing the next. Examples in this last area, generally known as batch processing, include manufacturing chemicals, pharmaceuticals, consumer products, and bioproducts. Since batch processes are repetitive, ILC can be applied, e.g., [10].

Consider an application where a reference trajectory is known. In the pick-and-place operation, this trajectory represents the ideal path for the robot to follow in each trial. The error between this trajectory and the trial's output forms a sequence of trial errors. The control design problem involves the construction of a sequence of trial inputs. These inputs are designed to guide the robot's movement so that the error sequence converges (in the trial number) to zero or within a specified tolerance. The response over the trial length can be assessed based on knowledge of the application under consideration on a particular trial.

As is well known, two-step methods such as the heavy ball method and the conjugate gradient method can significantly speed up the convergence of gradient descent algorithms in optimization problems [11–13]. In this paper, the problem investigated is the speed-up of the trial-to-trial error convergence through a control law where, on any trial, the input is an explicit function of the input used on the previous trial and other data from this trial. The analysis uses the heavy ball method [11], and the key idea here is the similarity of the structures of gradient descent and ILC.

Other strategies for using data from previous trials have been considered in the literature to speed up learning error convergence. In particular, there has been research on ILC design, where information from a finite number, greater than one, trials is used to generate the input for a subsequent trial; for example, see [14–20]. ILC laws that use control inputs from two or more previous trials to compute the input for the subsequent trial are termed higher-order laws. In these earlier studies, information from past trials is included using heuristic approaches. In general, these approaches proceed by first defining the structure of the ILC update law. Then, the parameters of this law are found using a trial-to-trial error convergence condition. Selecting the number of previous trials for the most significant effect is still an open research problem.

In [21], an ILC law based on Nesterov's method was described, but the convergence rate achieved was lower than that established in [22]. One reason for this is the use of a causal ILC law. In ILC systems, all data generated during a current trial are available to compute the control law for the subsequent trial. As one example, for discrete dynamics at a sample instant p, data generated on all previous trials at sample $p+\lambda$ can be used. Moreover, as is known [2], neglecting such information may impact the benefits of using an ILC law. The advantages of using higher-order laws are still an open question.

Throughout this paper, the notation for variables is of the form $h_k(p),\ 0 \leqslant p \leqslant N-1$, where h is the vector or scalar-valued variable under consideration, the integer $k \geqslant 0$ is the trial number, and N denotes the (finite) number of samples along the trial (N times the sampling period gives the trial length). Also, $\succ 0$, $\prec 0$, $\succeq 0$, and $\preceq 0$ denote, respectively, a symmetric positive define matrix, a symmetric negative definite matrix, a symmetric positive semi-definite matrix, and a symmetric negative semi-definite matrix.

2. PROBLEM FORMULATION

The systems considered are described by the following state-space model in the ILC setting on trial k :

$$\begin{aligned} x_k(p+1) &= Ax_k(p) + Bu_k(p), \\ y_k(p) &= Cx_k(p), \quad 0 \leqslant p \leqslant N-1, k \geqslant 0, \end{aligned} \tag{2.1}$$

where $x_k(p) \in \mathbb{R}^n$ is the state vector, and $u_k(p) \in \mathbb{R}$ and $y_k(p) \in \mathbb{R}$ are the input and output scalar variables, respectively. In the class of problems under consideration, the variable k is usually termed the trial number (trial for short), and $y_k(p)$ is the trial profile. The matrices A, B, and C are known and, by assumption, $CB \neq 0$. This assumption can be removed, e.g., using the result of [23] but is retained here for ease of presentation. The boundary conditions have the form $x_k(0) = x_0,\ k \geqslant 0$, where x_0 is given (on completing a trial, the system resets to the original state); $u_0(p)$ is given and bounded, and for $k \geqslant 1$, it is calculated using an updating law described below. Also, the generalization to multiple input and output systems with the same input and output dimensions is straightforward.

Let the system be completely controllable and observable, and let the variable $y_k(p)$ at each trial k be available for measurement. The state vector is estimated using a full-order asymptotic observer

$$\widehat{x}_k(p+1) = A\widehat{x}_k(p) + Bu_k(p) + G(y_k(p) - C\widehat{x}_k(p)),\ \widehat{x}_k(0) = x_0,\ 0 \leqslant p \leqslant N-1, k \geqslant 0, \quad (2.2)$$

where the matrix G is chosen by any known method such that the estimation error satisfies the condition

$$\widetilde{x}_k(p) = x_k(p) - \widehat{x}_k(p) \to 0$$

as $p \to \infty$.

Remark 1. Under the specified initial conditions, the observer accurately reconstructs the state vector. If x_0 is not available to the observer and or changes after a particular trial, the response of the estimation error will lead to a decrease in accuracy. This case requires further investigation and is left as an open research problem.

Let $y_{ref}(p),\ 0 \leqslant p \leqslant N-1,\ y_{ref}(0) = Cx_0$, denote a desired reference trajectory for (2.1). Then the error on trial k is

$$e_k(p) = y_{ref}(p) - y_k(p). \quad (2.3)$$

The ILC design problem is to construct a sequence of trial inputs $\{u_k\}$ such that for $0 \leqslant p \leqslant N-1$, the following conditions on the error $e_k(p)$ and input $u_k(p)$ hold:

$$|e_k(p)| \leqslant \varkappa\rho^k,\ \varkappa > 0,\ 0 < \rho < 1, \quad (2.4)$$

$$\lim_{k\to\infty} |u_k(p)| = |u_\infty(p)| < \infty, \quad (2.5)$$

where, in some parts of the literature, $u_\infty(p)$ is termed the learned control. In ILC applications, the trial length is finite; hence, it is possible to achieve trial-to-trial error convergence (in k) for unstable examples (in p). Two options then exist: a) design a stabilizing feedback control law and then apply the ILC law to the resulting controlled dynamics, or b) design a control law that simultaneously enforces trial-to-trial error convergence and regulates the dynamics along the trials. In the latter class of designs, the route is to treat ILC as a 2D system where the directions of information propagation are from trial to trial (k) and along the trial (p), respectively.

3. CONVERGENCE ANALYSIS AND DESIGN

The lifted model approach of ILC design for discrete-time systems involves the use of supervectors [14]. As the trial length is finite, the values of a variable over the trial length can be assembled into a column vector, e.g., for the input $\mathbf{U}_k = [u_k(0) \dots u_k(N-1)]^T$. As a result, trial-to-trial updating of variables can be expressed in standard difference equations. This approach can lead to matrices of very large dimensions. Moreover, feedback control must be applied in many cases to regulate the dynamics along a trial. An alternative represents the dynamics as a discrete repetitive process [24], a particular form of 2D systems. In this setting, it is possible to simultaneously design for trial-to-trial error convergence along the trial performance. This paper uses this latter setting for analysis and design, using the incremental variables

$$\begin{aligned}&\eta_{k+1}(p+1) = x_{k+1}(p) - x_k(p),\ \widehat{\eta}_{k+1}(p+1) = \widehat{x}_{k+1}(p) - \widehat{x}_k(p),\\ &\widetilde{\eta}_{k+1}(p+1) = \widetilde{x}_{k+1}(p) - \widetilde{x}_k(p),\ 0 \leqslant p \leqslant N-1, k \geqslant 0, \dots.\end{aligned} \quad (3.1)$$

Using (2.1), (2.2), (2.3) and these incremental variables gives the following description of the dynamics:

$$\begin{aligned} \widetilde{\eta}_{k+1}(p+1) &= (A-GC)\widetilde{\eta}_{k+1}(p), \\ \widehat{\eta}_{k+1}(p+1) &= GC\widetilde{\eta}_{k+1}(p) + A\widehat{\eta}_{k+1}(p) + B\Delta u_{k+1}(p-1), \\ e_{k+1}(p) &= -CGC\widetilde{\eta}_{k+1}(p) - CA\widehat{\eta}_{k+1}(p) - CB\Delta u_{k+1}(p-1) + e_k(p), \end{aligned} \tag{3.2}$$

where $\Delta u_{k+1}(p-1) = u_{k+1}(p-1) - u_k(p-1)$. In ILC design, the most common starting point is to specify that the input for the subsequent trial is the sum of the previous trial input plus a correction term (update law), where previous trial information is used. At the end of any trial, all information from all previous trials is available for use in constructing the control input for the subsequent trial. Many designs use only previous trial data, but the question is: would the use of a finite number, greater than one, of previous trials, be beneficial? This paper considers a control law that uses the control signals from the previous two trials and an optimization procedure to speed up the trial-to-trial error convergence.

The ILC law for (3.2) has the following structure:

$$u_{k+1}(p-1) = u_k(p-1) + \Delta u_{k+1}(p-1). \tag{3.3}$$

Hence, the problem is to find an update law $\Delta u_{k+1}(p-1)$ such that the convergence conditions (2.4) and (2.5) hold. Let this law be defined by

$$\Delta u_{k+1}(p-1) = \alpha\Delta u_k(p-1) + \beta\nabla\left(\frac{1}{2}e_{k+1}(p)^2\right) + \gamma[\Delta u_k(p-1) - \Delta u_{k-1}(p-1)], \tag{3.4}$$

where

$$\nabla\left(\frac{1}{2}e_{k+1}(p)^2\right) = \frac{\partial}{\partial\Delta u_{k+1}(p-1)}\left(\frac{1}{2}e_{k+1}(p)^2\right). \tag{3.5}$$

The update law (3.4) corresponds to the heavy ball method [11] in optimization theory. This simplest two-step method accelerates the convergence of gradient descent and, as will be shown, the same effect is achieved when the ILC law with the update term (3.4) is applied. Moreover, it is one option for using data beyond the previous trial in an ILC law. Other possibilities are discussed later in the paper.

Remark 2. In the original paper [11] $\alpha = 1$, but in the application considered (3.4) will be combined with the system dynamics (3.2) and, in contrast with [11], the argument $\Delta u_{k+1}(p-1)$ is parameterized by the variable p. Hence, the parameters α, β, and γ ensuring convergence may differ from those recommended in [11].

Due to the last equation of (3.2), the expression under the derivative sign in (3.5) can be written as

$$\begin{aligned} \frac{1}{2}e_{k+1}(p)^2 &= [-CGC\widetilde{\eta}_{k+1}(p) - CA\widehat{\eta}_{k+1}(p) - CB\Delta u_{k+1}(p-1) + e_k(p)]^T \\ &\times [-CGC\widetilde{\eta}_{k+1}(p) - CA\widehat{\eta}_{k+1}(p) - CB\Delta u_{k+1}(p-1) + e_k(p)]. \end{aligned}$$

Calculating the derivative of the right-hand side with respect to $\Delta u_{k+1}(p-1)$ gives

$$\nabla\left(\frac{1}{2}e_{k+1}(p)^2\right) = (CB)^2\Delta u_{k+1}(p-1) + CBCA\widehat{\eta}_{k+1}(p) + CBCGC\widetilde{\eta}_{k+1}(p) - CBe_k(p).$$

Substituting this expression into (3.4) after simple transformations, yields

$$\begin{aligned}\Delta u_{k+1}(p-1) = K_1\Delta u_k(p-1) + K_2CGC\widetilde{\eta}_{k+1}(p) + K_2CA\widehat{\eta}_{k+1}(p)\\ -K_2e_k(p) - K_3\Delta u_{k-1}(p-1),\end{aligned} \tag{3.6}$$

where

$$K_1 = \frac{\alpha+\gamma}{1-\beta(CB)^2},\ K_2 = \frac{\beta CB}{1-\beta(CB)^2},\ K_3 = \frac{\gamma}{1-\beta(CB)^2}.$$

Further, to apply the technique of linear matrix inequalities (LMIs), K_1, K_2, and K_3 will be used instead of the initial parameters α, β and γ. Introducing

$$\zeta_k(p) = \widetilde{\eta}_{k+1}(p),\ \xi_k(p) = \widehat{\eta}_{k+1}(p),\ \Delta\bar{u}_k(p) = \Delta u_k(p-1)$$

and substituting (3.6) into (3.2) give

$$\begin{aligned}
\zeta_k(p+1) &= (A-GC)\zeta_k(p),\\
\xi_k(p+1) &= (I+BK_2C)GC\zeta_k(p) + (I+BK_2C)A\xi_k(p)\\
&\quad + BK_1\Delta\bar{u}_k(p) - BK_3\Delta\bar{v}_k(p) - BK_2e_k(p),\\
\Delta\bar{u}_{k+1}(p) &= K_2CA\xi_k(p) + K_1\Delta\bar{u}_k(p) - K_3\Delta\bar{v}_k(p) - K_2e_k(p),\\
\Delta\bar{v}_{k+1}(p) &= \Delta\bar{u}_k(p),\\
e_{k+1}(p) &= -C(I+BK_2C)GC\zeta_k(p) - C(I+BK_2C)A\xi_k(p) - CBK_1\Delta\bar{u}_k(p)\\
&\quad + CBK_3\Delta\bar{v}_k(p) + (1+CBK_2)e_k(p).
\end{aligned} \tag{3.7}$$

In (3.7), the vectors $\zeta_k(p)$ and $\xi_k(p)$ evolve in p, the along-the-trial variable, and $\Delta\tilde{u}_k(p)$, $\Delta\bar{v}_k(p)$, and $e_k(p)$ evolve in k, the trial-to-trial variable. Systems with this type of dynamics are known as discrete linear repetitive processes [24]. These processes make a series of sweeps through dynamics defined over a finite duration. Once each sweep, termed a trial, is complete, the system resets to the starting location and is ready for the start of the subsequent trial. The unique feature is that the output on any trial contributes to the dynamics of the next one, which can result in oscillations that increase from trial to trial and cannot be controlled by standard control action. A stability theory for the case of linear dynamics has been developed [24] and applied to ILC design with experimental validation, e.g., [25].

As in stability analysis and control law design for standard linear systems theory, one approach to the same design problem for the dynamics considered in this paper would be to use a suitably constructed Lyapunov function, which should be constructed from the sum of terms in the trial-to-trial propagation and the along-the-trial dynamics, respectively. But the two equations in (3.7) are coupled and hence the gradient of the Lyapunov function can only be found if the solutions to these equations are available, which is a very stringent requirement, and instead a vector Lyapunov function must be used and the gradient replaced by the divergence, see [26] for the required background. (In the case of linear dynamics, the stability analysis methods in [24] and [26] are equivalent.)

In this paper, the vector Lyapunov function for dynamics described by (3.7) has the form

$$V(\bar{\xi}_k(p), \bar{e}_k(p)) = \begin{bmatrix} V_1(\bar{\xi}_k(p)) \\ V_2(\bar{e}_k(p)) \end{bmatrix}, \tag{3.8}$$

where $\bar{\xi}_k(p) = [\zeta_k(p)^T \; \xi_k(p)^T]^T$, $\bar{e}_k(p) = \begin{bmatrix} \Delta\bar{u}_k(p) & \Delta\bar{v}_k(p) & e_k(p) \end{bmatrix}^T$, $V_1(\bar{\xi}_k(p)) > 0$, $\bar{\xi}_k(p)) \neq 0$, $V_2(\bar{e}_k(p)) > 0$, $\bar{e}_k(p) \neq 0$, $V_1(0) = 0$, $V_2(0) = 0$. Also, let a discrete counterpart of the divergence operator of this function along the trajectories of system (3.7) be defined as

$$\mathcal{D}V(\bar{\xi}_k(p), \bar{e}_k(p)) = V_1(\bar{\xi}_k(p+1)) - V_1(\bar{\xi}_k(p)) + V_2(\bar{e}_{k+1}(p)) - V_2(\bar{e}_k(p)). \tag{3.9}$$

The following result can be established.

Theorem 1. *Assume that there exists a vector Lyapunov function* (3.8) *and positive scalars* c_1, c_2, *and* c_3 *such that the following inequalities are valid on the trajectories of the system* (3.7) (*where* $||\cdot||$ *denotes the Euclidean norm*):

$$c_1||\bar{\xi}_k(p)||^2 \leqslant V_1(\bar{\xi}_k(p)) \leqslant c_2||\bar{\xi}_k(p)||^2, \tag{3.10}$$

$$c_1||\bar{e}_k(p)||^2 \leqslant V_2(\bar{e}_k(p)) \leqslant c_2||\bar{e}_k(p)||^2, \tag{3.11}$$

$$\mathcal{D}V(\bar{\xi}_k(p), e_k(p)) \leqslant -c_3(||\bar{\xi}_k(p)||^2 + ||\bar{e}_k(p)||^2). \tag{3.12}$$

Then conditions (2.4) *and* (2.5) *hold.*

Proof. In [26] (Theorem 1) it was shown that if (3.10), (3.11), and (3.12) are satisfied, then

$$||\bar{\xi}_k(p)||^2 + ||\bar{e}_k(p)||^2 \leqslant \bar{\varkappa}\varrho^k, \; 0 \leqslant p \leqslant N-1, \; k \geqslant 0, \tag{3.13}$$

where $\bar{\varkappa}$ and $0 < \varrho < 1$ depend on c_1, c_2, and c_3. Also it follows from (3.13) and the definition of $\bar{e}_k(p)$ that

$$|e_k(p)|^2 \leqslant \bar{\varkappa}\varrho^k, \quad |\Delta\bar{u}_k(p)|^2 \leqslant \bar{\varkappa}\varrho^k. \tag{3.14}$$

Due to the first inequality in (3.14), condition (2.4) holds with $\varkappa = \bar{\varkappa}^{\frac{1}{2}}$ and $\rho = \varrho^{\frac{1}{2}}$. The boundedness of control (2.5) follows from the obvious relation

$$|u_k(p)| \leqslant |u_{k-1}(p)| + |\Delta\bar{u}_k(p)|, \quad k = 1, 2, \ldots,$$

and the boundedness of $u_0(p)$ and the second inequality in (3.14).

Introduce the following matrices and vectors:

$$\bar{A} = \begin{bmatrix} A - GC & 0 & 0 & 0 & 0 \\ GC & A & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ -CGC & -CA & 0 & 0 & 1 \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} 0 \\ B \\ 1 \\ 0 \\ -CB \end{bmatrix}, \quad \bar{C} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ CGC & CA & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 \end{bmatrix}.$$

Then the following result enables LMI based ILC design.

Theorem 2. *Suppose that for some matrix* $Q \succ 0$ *and a positive scalar* R,

$$\begin{bmatrix} X & (\bar{A}X + \bar{B}Y\bar{C})^T & X & (Y\bar{C})^T \\ \bar{A}X + \bar{B}Y\bar{C} & X & 0 & 0 \\ X & 0 & Q^{-1} & 0 \\ Y\bar{C} & 0 & 0 & R^{-1} \end{bmatrix} \succeq 0, \; \bar{C}X = Z\bar{C}, \tag{3.15}$$

is feasible for $X = \mathrm{diag}[X_1 \; X_2]$, Y, Z, *where* X_1 *and* X_2 *are positive definite matrices of dimensions* $2n \times 2n$ *and* $3n \times 3n$, *respectively. Then the ILC law*

$$\begin{aligned} u_{k+1}(p) &= u_k(p) + \Delta u_{k+1}(p), \\ \Delta u_{k+1}(p) &= K_1\Delta u_k(p) + K_2CA[\widehat{x}_{k+1}(p) - \widehat{x}_k(p)] \\ &\quad + K_2CG[(y_{k+1}(p) - y_k(p)) - (\widehat{y}_{k+1}(p) - \widehat{y}_k(p)] \\ &\quad - K_2[y_{ref}(p+1) - Cx_k(p+1)] - K_3\Delta u_{k-1}(p), \end{aligned} \tag{3.16}$$

where

$$K = \begin{bmatrix} K_1 & K_2 & K_3 \end{bmatrix} = YZ^{-1},$$

ensures that the required conditions (2.4) *and* (2.5) *hold.*

Proof. Applying the Schur's complement formula to inequality (3.15) gives

$$X - (\bar{A}X + \bar{B}Y\bar{C})^T X^{-1}(\bar{A}X + \bar{B}Y\bar{C}) - XQX - (Y\bar{C})^T R(Y\bar{C}) \succeq 0.$$

Further, using $Y = KZ$ and $CX = ZC$, this last inequality can be written in the form

$$X - (\bar{A}X + \bar{B}K\bar{C}X)^T X^{-1}(\bar{A}X + \bar{B}K\bar{C}X) - XQX - (K\bar{C}X)^T R(K\bar{C}X) \succeq 0.$$

Multiplying the last inequality by $P = X^{-1}$ on the left and right, respectively, and then rearranging the terms give

$$(\bar{A} + \bar{B}K\bar{C})^T P(\bar{A} + \bar{B}K\bar{C}) - P + Q + (K\bar{C})^T R(K\bar{C}) \preceq 0. \tag{3.17}$$

Let the entries in the vector Lyapunov function (3.8) be chosen as the quadratic forms

$$\begin{aligned} V_1(\bar{\xi}_k(p)) &= \bar{\xi}_k^T(p)P_1\bar{\xi}_k(p), \\ V_2(\bar{e}_k(p)) &= \bar{e}_k^T(p)P_2\bar{e}_k(p), \end{aligned} \tag{3.18}$$

where $P_1 \succ 0$ and $P_2 \succ 0$. Then calculating the divergence (3.9) along the trajectories of system (3.7) gives that conditions (3.10), (3.11), and (3.12) will be satisfied if

$$(\bar{A} + \bar{B}K\bar{C})^T P(\bar{A} + \bar{B}K\bar{C}) - P \prec 0. \tag{3.19}$$

Finally, (3.19) holds due to (3.17) and (3.16) follows from (3.6) on using (3.1) and (2.3).

Remark 3. The system (3.15) is similar in structure to the relations of the linear-quadratic regulator (LQR) problem in LMI formulation. The matrix $K = \begin{bmatrix} K_1 & K_2 & K_3 \end{bmatrix}$ depends on Q and R, which are selected based on achieving the desired rate of convergence of the learning error. As in the LQR problem, there is no complete formalization of such a choice and a heuristic approach based on knowledge of the dynamics of the system under consideration is recommended.

For further comparative analysis, consider the most commonly used update law, see, for example, [27]:

$$\Delta u_k(p) = \hat{K}_1\hat{\eta}_k(p) + \hat{K}_2 e_{k-1}(p+1). \tag{3.20}$$

Substituting (3.20) in (3.2) and applying Theorem 1 with the entries in (3.8) chosen as

$$\begin{aligned} V_1(\bar{\xi}_k(p)) &= \bar{\xi}_k^T(p)\hat{P}_1\bar{\xi}_k(p), \\ V_2(e_k(p)) &= e_k^T(p)\hat{P}_2 e_k(p), \end{aligned} \tag{3.21}$$

give the divergence

$$\mathcal{D}V(\bar{\xi}_k(p), e_k(p)) = [\bar{\xi}_k^T(p)\ e_k(p)][(\hat{A} + \hat{B}K\hat{C})^T\hat{P}(\hat{A} + \hat{B}K\hat{C}) - \hat{P}][\bar{\xi}_k^T(p)\ e_k(p)]^T, \tag{3.22}$$

where $\hat{P} = \operatorname{diag}[\hat{P}_1\ \hat{P}_2]$, and

$$\hat{A} = \begin{bmatrix} A - GC & 0 & 0 \\ GC & A & 0 \\ -CGC & -CA & 1 \end{bmatrix}, \ \hat{B} = \begin{bmatrix} 0 \\ B \\ -CB \end{bmatrix} \ \hat{C} = \begin{bmatrix} 0 & I & 0 \\ 0 & 0 & 1 \end{bmatrix}, \ \hat{K} = [\hat{K}_1\ \hat{K}_2].$$

Repeating the derivations from the proof of Theorem 2 gives that if

$$\begin{bmatrix} \hat{X} & (\hat{A}\hat{X}+\hat{B}\hat{Y}\hat{C})^T & \hat{X} & (\hat{Y}\hat{C})^T \\ \hat{A}\hat{X}+\hat{B}\hat{Y}\hat{C} & \hat{X} & 0 & 0 \\ \hat{X} & 0 & \hat{Q}^{-1} & 0 \\ \hat{Y}\hat{C} & 0 & 0 & \hat{R}^{-1} \end{bmatrix} \succeq 0,\ \hat{C}\hat{X} = \hat{Z}\hat{C} \tag{3.23}$$

has a solution $\hat{X} = \mathrm{diag}[\hat{X}_1\ \hat{X}_2],\ \hat{Y},\ \hat{Z}$, where $\hat{X}_1 \succ 0$ is a $2n \times 2n$ matrix and $\hat{X}_2$ is a positive scalar, then the ILC law with the update term (3.20) and the matrix $\hat{K} = \hat{Y}\hat{Z}^{-1}$ ensures conditions (2.4) and (2.5).

Remark 4. The approach based on the heavy ball method leads to a new structure of the control law (3.16), which differs from that in the known literature. If the parameter values ensure conditions (2.4) and (2.5), then the update term, in accordance with (3.4), changes along the gradient of the quadratic function of this term. In the case of the update law (3.20), this cannot be guaranteed. Apparently, this property provides faster convergence of the learning error when using the control law (3.16), but rigorous proof has not yet been obtained.

4. NUMERICAL CASE STUDY

In, e.g., [25], a gantry robot, specifically designed and constructed to emulate the pick-and-place operation to which ILC is applicable, has been used to verify ILC designs experimentally. This system has three mutually perpendicular axes, one of which X is directed along the conveyor belt, and the second Y is perpendicular to this axis in the conveyor plane. The third Z is perpendicular to the XY plane. The system is designed such that the interaction between axes can be neglected, and models for control design have been obtained from frequency response tests.

Figure 1 shows the 3D reference signal used in experimental tests; the trial length is 2 s.

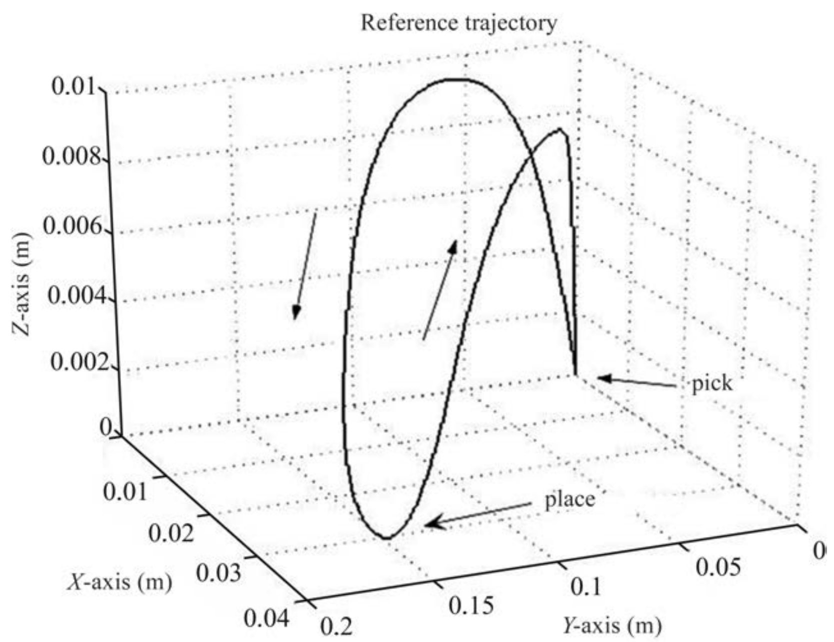

Fig. 1. The 3D reference trajectory.

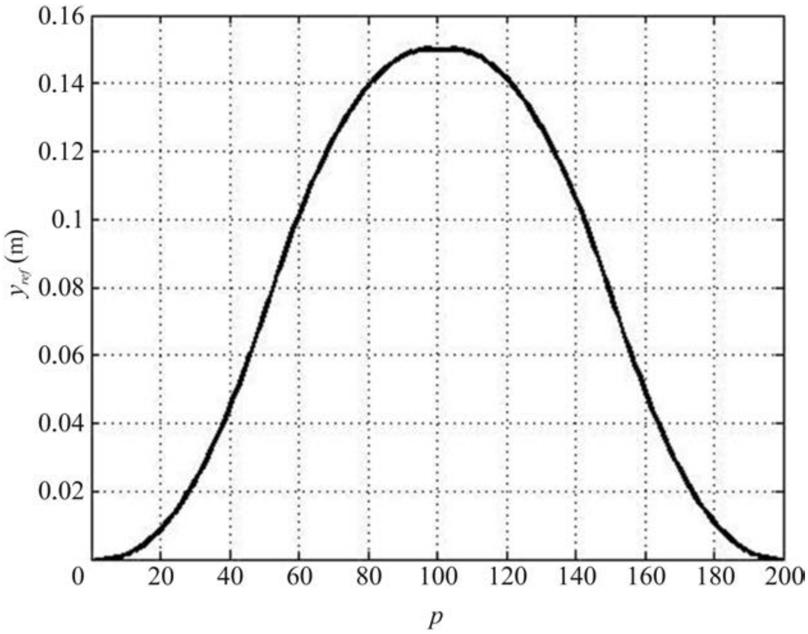

Fig. 2. The reference trajectory for the Y axis.

The transfer function for the Y axis is

$$G_Y(s) = \frac{23.736(s + 661.2)}{s(s^2 + 426.7s + 1.744 \times 10^5)}, \tag{4.1}$$

and the reference trajectory for this axis is shown in Fig. 2. The sampling period for discrete implementation is $T_s = 0.01$ s. The discrete model has been constructed using the standard functions `ss` and `c2d` of MATLAB. The corresponding matrices are

$$A = \begin{bmatrix} -0.0762 & 0.0487 & 0 \\ -0.0732 & -0.1372 & 0 \\ 0.0033 & 0.0026 & 1.0000 \end{bmatrix}, \ B = \begin{bmatrix} -0.0006 \\ 0.0134 \\ 0.0001 \end{bmatrix}, \ C = [0 \ \ 0.0116 \ \ 7.6631].$$

Evaluation of the performance of an ILC design must consider trial-to-trial error convergence and performance along the trials. The latter of these can be evaluated using standard measures for each trial. For trial-to-trial error convergence, one commonly used measure is the root mean square learning error on each trial plotted against the trial number, where on trial k,

$$\mathrm{E}(k) = \sqrt{\frac{1}{N}\sum_{p=0}^{N-1} |e_k(p)|^2}. \tag{4.2}$$

The observer parameters have been found using the results of modal control theory, implemented using the standard function `place` of MATLAB. Assigning the eigenvalues $\lambda_1 = 0.9$ and $\lambda_{2,3} = -0.1 \pm 0.01i$ for the matrix $(A - GC)$ and applying this function give

$$G = [-0.3070 \ \ 2.6427 \ \ 0.0073]^T.$$

According to the problem formulation, under the assumptions made, any observer matrix G for which the eigenvalues of $(A - GC)$ are inside the unit circle accurately reconstructs the state vector. Therefore, the eigenvalues have been assigned to make the observer gains (the entries in the matrix G) relatively small.

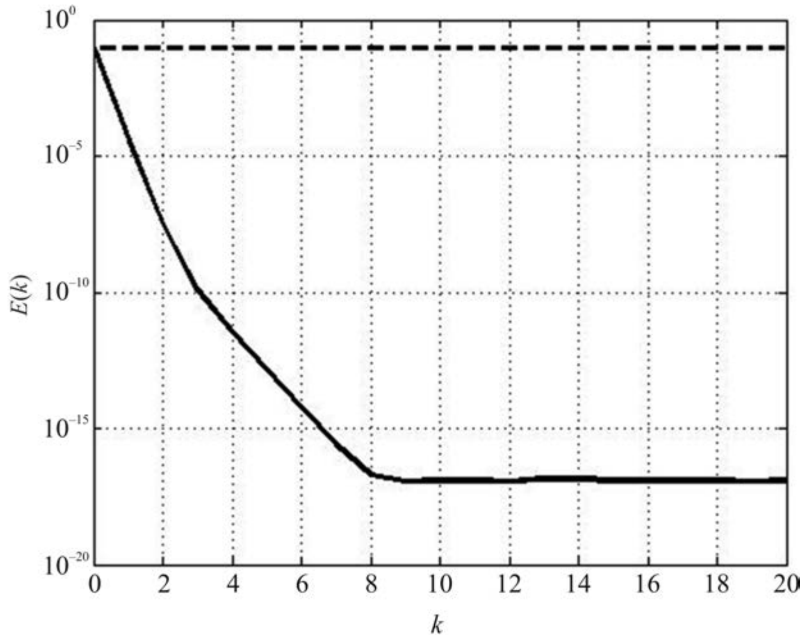

Fig. 3. Progression of $E(k)$ for the Y axis with $Q = I$ and $R = 1$ for (3.16) (solid line) and (3.20) (dashed line).

Selecting $Q = I$, i.e., the identity matrix of compatible dimensions and $R = 1$ and applying Theorem 2 give

$$K_1 = -1.09 \times 10^{-7}, \quad K_2 = -1196.0, \quad K_3 = -1.61 \times 10^{-8}.$$

For this case, the ILC law (3.16) reduces the learning error by 10^5 times in one trial, and after 8 trials, this error has reached numerical zero. For comparison, the ILC algorithm with the update law (3.20) gives an error reduction of 1% for 70 trials (Fig. 3).

In this case, the control law parameters K_1 and K_3 are very small, and numerical investigations demonstrate that a minimal change will occur if they are set to zero. Hence, the ILC law with the update (3.20) in which $\hat{K}_1 = K_2CA$ and $\hat{K}_2 = -K_2$ is obtained. Thus, applying the heavy ball method imposes restrictions on $\hat{K}_1$ and $\hat{K}_2$, under which a high trial-to-trial error convergence rate is ensured. It is, however, difficult to use this feature to provide rules for choosing Q and R for a given example. Establishing a connection between the convergence rate of the learning error and Q and R is an open problem. (It is possible to select matrices providing approximately a 10 times lower convergence rate but at the cost of a significant computation time.) Considering this feature, reducing the convergence rate by decreasing K_2 and checking the LMI (3.19) is possible. In practice, this may be important if actuator saturation is possible, leading to the need for a compromise solution.

For $K_2 = -50$, the convergence rate decreases by approximately 80 times relative to the result obtained for unit Q and R in (3.15), however, this speed remains higher than for $Q = I$ and $R = 1$ in (3.23) (Fig. 4); the control law matrices in the case of $K_2 = -50$ and $K_1 = K_3 = 0$ are

$$\hat{K}_1 = [-1.24 \ -0.93 \ -383.2], \quad \hat{K}_2 = 50.0, \tag{4.3}$$

and, in the case of ILC with the update law (3.20) with $Q = I$ and $R = 1$

$$\hat{K}_1 = [-21.8 \ -18.4 \ -7465.4], \quad \hat{K}_2 = 1.9. \tag{4.4}$$

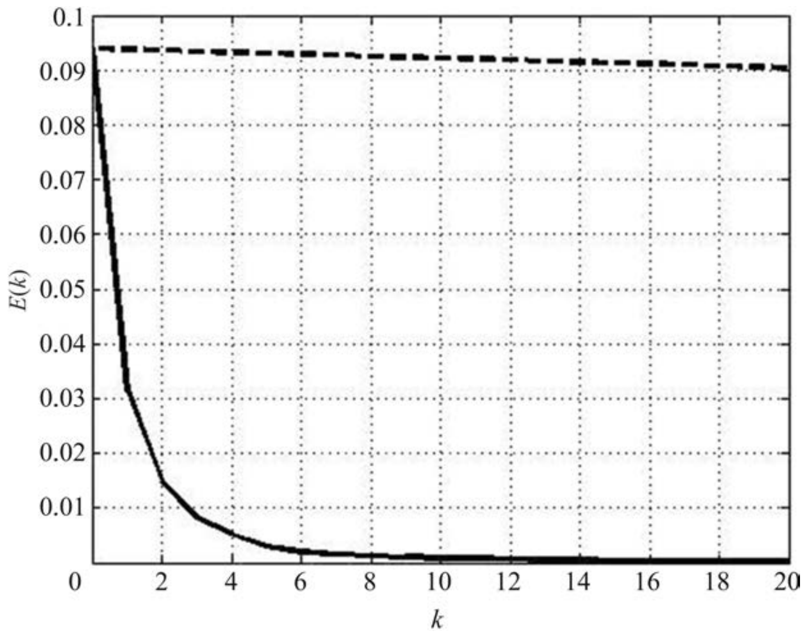

Fig. 4. Progression of $E(k)$ for the Y axis with $Q = I$ and $R = 1$ for (3.16) and $K_2 = -50$ (solid line) and (3.20) (dashed line).

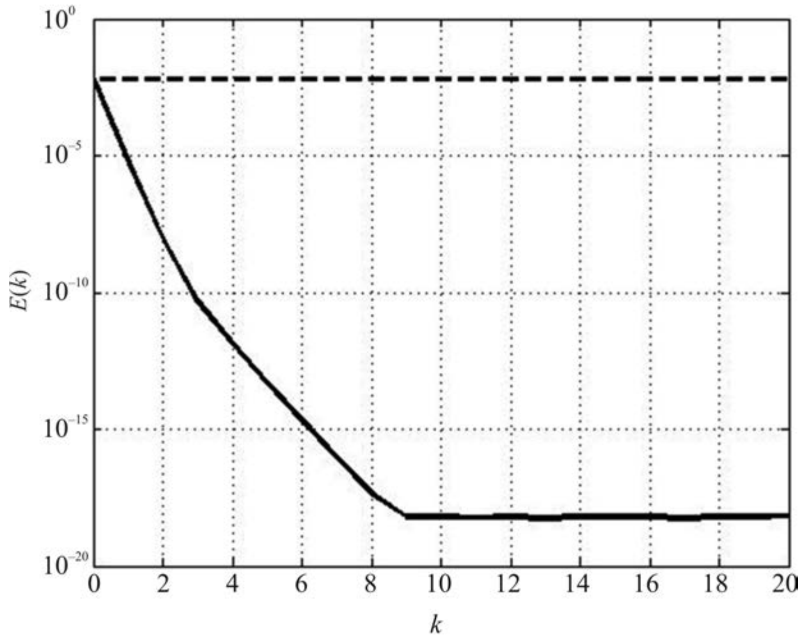

Fig. 5. Progression of $E(k)$ for the Z axis with $Q = I$ and $R = 1$ f (3.16) (solid line) and (3.20) (dashed line).

The linear matrix equations and inequalities have been solved as before but use is also made of the `YALMIP` parser. Analysis of (4.3) and (4.4) shows that, in the absence of structural constraints on $\hat{K}_1$ and $\hat{K}_2$, `SeDuMi`'s internal algorithm finds a solution in a domain characterized by a low convergence rate; at the same time, the constraints imposed by the heavy ball method direct the solution to a domain with accelerated convergence.

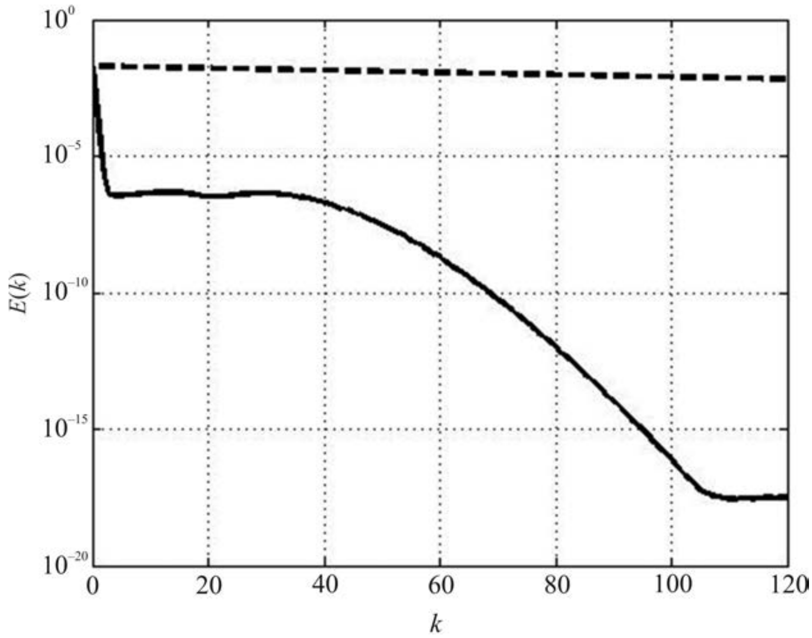

Fig. 6. Progression of $E(k)$ for the X-axis with Q and R given by (4.5) for (3.16) (solid line) and (3.20) (dashed line).

Now consider the ILC design along the Z and X axes. The transfer functions, for these axes are

$$G_Z(s) = \frac{15.8869(s+661.2)}{s(s^2+707.6s+3.377\times10^5)},$$

and

$$G_X(s) = \frac{13077183.4436(s+113.4)(s^2+227.9s+5.467\times10^4)}{s(s^2+61.57s+1.125\times10^4)(s^2+227.9s+5.467\times10^4)(s^2+466.1s+6.142\times10^5)}.$$

To construct an observer along the Z axis, the same approach as in the previous case is used, with the observer poles taken as $\lambda_1 = 0.9,\ \lambda_{2,3} = -0.1\pm0.01i$. Applying the `place` function of MATLAB yields

$$G = [-0.405 \quad -28.01 \quad 0.078]^T.$$

Theorem 2 with unit $Q = I$ and $R = 1$ gives

$$K_1 = -6.33\times10^{-8},\ K_2 = -2.673\times10^3,\ K_3 = -8.33\times10^{-9}.$$

The simulation results are given in Fig. 5.

The observer poles along the X axis are selected as

$$\lambda_1 = 0.95,\quad \lambda_2 = -0.4,\quad \lambda_3 = 0.4,\quad \lambda_{4.5} = 0.4\left(\cos\frac{\pi}{3}\pm\sin\frac{\pi}{3}\right),\quad \lambda_{6,7} = 0.4\left(\cos\frac{\pi}{3}\pm\sin\frac{\pi}{3}\right).$$

Applying the `place` function of MATLAB yields

$$G = [38.1 \quad 40.161 \quad -44.093 \quad 15.913 \quad -16.428 \quad -12.67 \quad 0.309]^T.$$

In this case, Q and R are selected as

$$Q = \mathrm{diag}[I \ 10I \ 1 \ 1 \ 1], \quad R = 1. \tag{4.5}$$

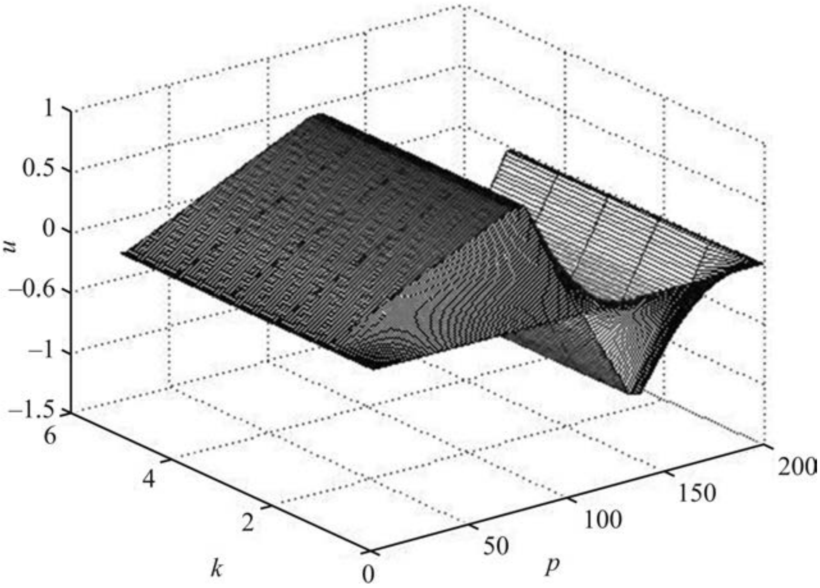

Fig. 7. Control progression along the X axis using (3.16).

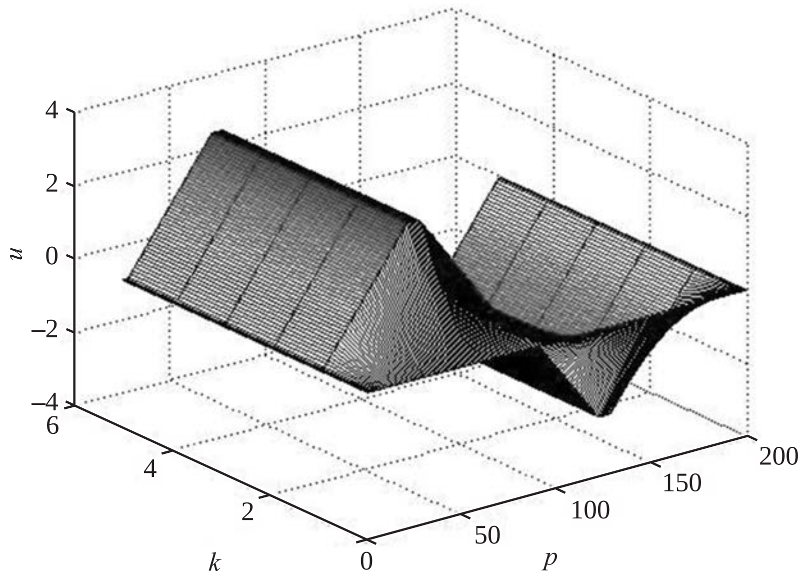

Fig. 8. Control progression along the Y axis with the update law (3.16).

Theorem 2 gives

$$K_1 = -5.715 \times 10^{-9}, \quad K_2 = -1308.2, \quad K_3 = -6.868 \times 10^{-10}.$$

The simulation results are given in Fig. 6. The control inputs for the X, Y, and Z axes for (3.16) are given in Figs. 7, 8, and 9.Note that the learned control on all axes is achieved rather quickly: after three or four trials, the controls on the axes do not change depending on k.

According to Fig. 5, the convergence of the learning error to numerical zero is achieved after several trials. A different picture is observed for the X axis (Fig. 6): $E(k)$ decreases by 10^7 times in four trials, then the convergence rate decreases, and numerical zero is achieved only after 107 trials; at the same time, this rate significantly exceeds the convergence rate of the alternative algorithm.

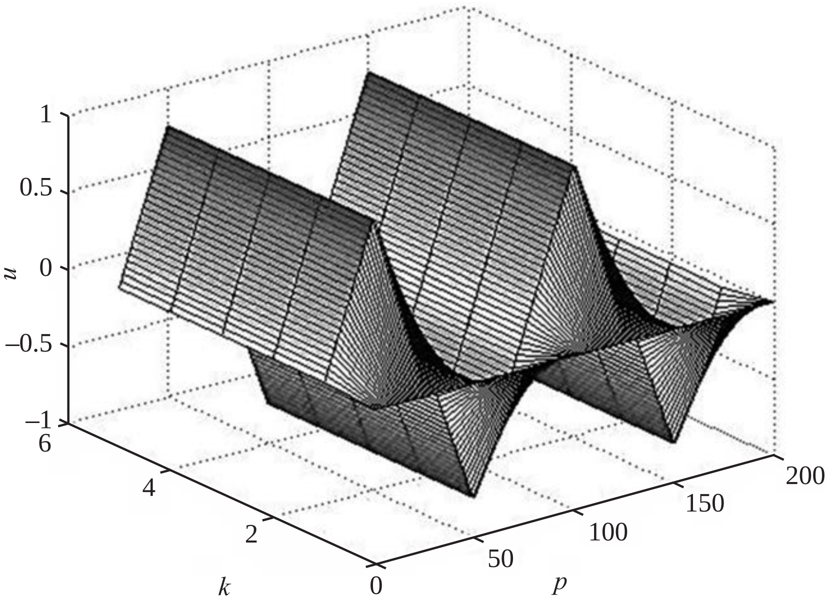

Fig. 9. Control progression along the Z axis with the update law (3.16).

Thus, in this example, the ILC design using the heavy ball method provides a high convergence rate that cannot be achieved by the alternative method. Note that for all three axes, K_1 and K_3 have quite small entries. Whether this is a general property of the new design, or is it related to the peculiarities of the dynamics of the system under consideration, remains unclear and requires additional research.

5. CONCLUSIONS

The ILC law (3.16) differs from those known in the literature in the special structure of the gain matrices. Judging by the simulation results, this structure is the main factor accelerating convergence. In the case of unstructured matrices, such as (3.20), it is probably possible to achieve high convergence rates, but this will demand a time-consuming heuristic procedure for selecting Q and R. Thus, it can be preliminarily concluded that using the heavy ball method gives an ILC law with a high convergence rate. The final conclusion requires a strict justification, which has not been obtained in this paper and will be the subject of further research. In the near future, the use of conventional gradient descent and Newton's method will be considered in ILC design; perhaps, the resulting ILC structures will provide the key to a general proof of accelerated convergence. In the longer term, it is planned to study the use of the conjugate gradient method and Nesterov's method.

Overall, there is much room for further research in this general area.

ACKNOWLEDGMENT

The authors dedicate this paper to the memory of Boris Polyak, the author of the heavy ball method and a number of other very significant results in optimization theory and control theory.

FUNDING

This work was partially supported by the Russian Science Foundation, project no. 23-71-01044, https://rscf.ru/en/project/23-71-01044/.

REFERENCES

1. Arimoto, S., Kawamura, S., and Miyazaki, F., Bettering Operation of Robots by Learning, *J. Robot. Syst.*, 1984, vol. 1, pp. 123–140.
2. Bristow, D.A., Tharayil, M., and Alleyne, A.G., A Survey of Iterative Learning Control: A Learning-Based Method for High-Performance Tracking Control *IEEE Control Syst. Magaz.*, 2006, vol. 26, no. 3, pp. 96–114.
3. Ahn, H-S., Chen, Y.Q., and Moore, K.L., Iterative Learning Control: Survey and Categorization, *IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev.*, 2007, vol. 37, no. 6, pp. 1099–1121.
4. Rogers, E., Chu, B., Freeman, C., and Lewin, P., *Iterative Learning Control Algorithms and Experimental Benchmarking*, Chichester: John Wiley & Sons, 2023.
5. Lim, I., Hoelzle, D.J., and Barton, K.L., A multi-objective iterative learning control approach for additive manufacturing applications, *Control Engineer. Practice*, 2017, vol. 64, pp. 74–87.
6. Sammons, P.M., Gegel, M.L., Bristow, D.A., and Landers, R.G., Repetitive Process Control of Additive Manufacturing with Application to Laser Metal Deposition, *IEEE Transact. Control Syst. Technol.*, 2019, vol. 27, no. 2, pp. 566–577.
7. Freeman, C.T., Rogers, E., Hughes, A.-M., Burridge, J.H., and Meadmore, K.L., Iterative learning control in health care: electrical stimulation and robotic-assisted upper-limb stroke rehabilitation, *IEEE Control Syst. Magaz.*, 2012, vol. 47, pp. 70–80.
8. Meadmore, K.L., Exell, T.A., Hallewell, E., Hughes, A.-M., Freeman, C.T., Kutlu, M., Benson, V., Rogers, E., and Burridge, J.H., The application of precisely controlled functional electrical stimulation to the shoulder, elbow and wrist for upper limb stroke rehabilitation: a feasibility study, *J. NeuroEngin. Rehabilitat.*, 2014, pp. 11–105.
9. Ketelhut, M., Stemmler, S., Gesenhues, J., Hein, M., and Abel, D., Iterative learning control of ventricular assist devices with variable cycle durations, *Control Engin. Practice*, 2019, vol. 83, pp. 33–44.
10. Wang, Y., Gao, F., and Doyle, F.J., Survey on iterative learning control, repetitive control, and run-to-run control, *J. Proc. Control*, 2009, vol. 19, pp. 1589–1600.
11. Polyak, B.T., Some methods of speeding up the convergence of iteration methods, *USSR Computational Mathematics and Mathematical Physics*, 1964, vol. 4, no. 5, pp. 1–17.
12. Polyak, B.T., *Introduction to Optimization, Optimization Software*, New York, 1987.
13. d'Aspremont, A., Scieur, D., and Taylor, A., Acceleration Methods. arXiv:2101.09545v3. 2021.
14. Ahn, H.S., Moore, K.L., and Chen, Y.Q., Iterative Learning Control. Robustness and Monotonic Convergence for Interval Systems, *Lecture Notes in Control and Information Sciences*, London: Springer-Verlag, 2007.
15. Bien, Z. and Huh, K.M., Higher-order iterative learning control algorithm, *IEE Proc. D-Control Theory Appl.*, 1989, vol. 136, pp. 105–112.
16. Chen, Y., Gong, Z., and Wen, C., Analysis of a high-order iterative learning control algorithm for uncertain nonlinear systems with state delays, *Automatica*, 1998, vol. 34, pp. 345–353.
17. Norrlöf, M. and Gunnarsson, S., A frequency domain analysis of a second order iterative learning control algorithm, *Proc. 38th IEEE Conf. Decis. Control*, 1999, vol. 2, pp. 1587–1592.
18. Bu, X., Yu, F., Fu, Z., and Wang, F., Stability analysis of high-order iterative learning control for a class of nonlinear switched systems, *Abstract Appl. Anal.*, 2013, pp. 1–13.
19. Wei, Y.-S. and Li, X.-D., Robust higher-order ILC for non-linear discrete-time systems with varying trail lengths and random initial state shifts, *IET Control Theory Appl.*, 2017, vol. 11, pp. 2440–2447.
20. Wang, X., Chu, B., and Rogers, E., Higher-order Iterative Learning Control Law Design using Linear Repetitive Process Theory: Convergence and Robustness, *IFAC PapersOnLine*, 2017, vol. 50-1, pp. 3123–3128.
21. Gu, P., Tian, S., and Chen, Y., Iterative learning control based on Nesterov accelerated gradient method, *IEEE Access*, 2019, vol. 7, pp. 115836–115842.
22. Nesterov, Y., A method of solving a convex programming problem with convergence rate $O(\frac{1}{k^2})$, *Sov. Math. Doklady*, 1983, vol. 269, no. 3, pp. 372–376.

23. Paszke, W., Rogers, E., and Galkowski, K., Experimentally verified generalized KYP lemma based iterative learning control design, *Control Engin. Practice*, 2016, vol. 33, pp. 57–67.

24. Rogers, E., Gałkowski, K., and Owens, D.H., Control Systems Theory and Applications for Linear Repetitive Processes, *Lecture Notes in Control and Information Sciences*, vol. 349, Berlin: Springer-Verlag, 2007.

25. Hladowski, L., Galkowski, K., Cai, Z., Rogers, E., Freeman, C.T., and Lewin, P.L., Experimentally supported 2D systems based iterative learning control law design for error convergence and performance, *Control Engin. Practice*, 2010, vol. 18, pp. 339–348.

26. Pakshin, P., Emelianova, J., Emelianov, M., Gałkowski, K., and Rogers, E., Dissipativity and stabilization of nonlinear repetitive processes, *Syst. Control Lett.*, 2016, vol. 91, pp. 14–20.

27. Pakshin, P. and Emelianova, J., Iterative learning control design for discrete-time stochastic switched systems, *Automat. Remote Control*, 2020, vol. 81, no. 11, pp. 2011–2025.

This paper was recommended for publication by P.S. Shcherbakov, a member of the Editorial Board