ISSN 2782-2427

CONTROL SCIENCES

5/2021

[1]Russian Academy of Sciences.

CONTROL SCIENCES
Scientific Technical Journal

6 issues per year
ISSN 2782-2427
Open access

Published since 2021

Original Russian Edition
Problemy Upravleniya
Published since 2003

FOUNDER AND PUBLISHER
V.A. Trapeznikov
Institute of Control Sciences
of Russian Academy of Sciences

Editor-in-Chief
D.A. Novikov, Corr. Member of RAS

Deputy Editor-in-Chief
F.F. Pashchenko

Executive Editor-in-Chief
N.E. Maximova

Editor
L.V. Petrakova

Editorial address
65 Profsoyuznaya st., office 410,
Moscow 117997, Russia

☎/🖷 +7(495) 198-17-20, ext. 1410

✉ pu@ipu.ru

URL: http://controlsciences.org

Published: October 28, 2021

Registration certificate of
Эл № ФС 77-80482
of 17 February 2021
issued by the Federal Service
for Supervision of Communications,
Information Technology, and Mass
Media

CONTROL SCIENCES

5.2021

CONTENTS

DOI: http://doi.org/10.25728/cs.2021.5.1

ANALYSIS OF INFORMATION INCONSISTENCY IN BELIEF FUNCTION THEORY. PART I: EXTERNAL CONFLICT[1]

A.E. Lepskiy

National Research University Higher School of Economics, Moscow, Russia

✉ alex.lepskiy@gmail.com

Abstract. The analysis results of information inconsistency within belief function theory (the Dempster–Shafer theory of evidence) are reviewed. This theory has been intensively developing over the past 10–15 years. Part I of the survey considers the measure of external conflict between bodies of evidence. The concepts of conflict and non-conflict bodies of evidence and the basic requirements applied to measures of external conflict are discussed. Different axioms of the measure of external conflict are analyzed. The general forms of measures of external conflict that satisfy the system of axioms are given. Different methods for constructing measures of external conflict (metric, algebraic, and structural approaches; evaluation by combining rules) are presented. The robust estimation of external conflict, the relationship between its measure and the metric on the set of bodies of evidence, and the consistency of combining rules and measures of external conflict are discussed. Many illustrative examples are provided.

Keywords: theory of belief functions, combining rules, inconsistency of bodies of evidence, measure of external conflict.

INTRODUCTION

In many problems of data analysis and decision-making, information coming from various sources (experts, analysts, monitoring systems, etc.) can be conveniently described using the so-called bodies of evidence, the main concept in belief function theory (the Dempster–Shafer theory of evidence) [12]. A body of evidence is an aggregate of a set of certain subsets (the so-called focal elements) from a universal set and a mass function defined on these subsets. At present, belief function theory finds numerous applications in data analysis, processing of expert information, decision-making, etc. This is due to the understandable and interpretable foundations of the theory itself, associated with the classical probabilistic approach, and a wide range of tools to model and evaluate important factors such as the uncertainty and inaccuracy of information, the reliability of information sources, the inconsistency of information from different sources, the amount of ignorance, etc. In addition, trust function theory has a well-developed apparatus for combining bodies of evidence (combining rules suggested by Dempster, Yager, Inagaki, and others) considering the factors mentioned.

This paper presents the main results related to analyzing the inconsistency (conflict) of information sources within belief function theory that have been obtained over the past 15 years. The inconsistency of information obtained from different sources is their important a priori characteristic, which should be considered when deciding on the possibility of aggregating these sources or choosing sources from a set for aggregation. For example, one analyst predicts that in a month, the company's stock will have a value within the interval [40, 50]; according to another's opinion, a value within the interval [70, 85] (in conventional units). In this case, we have a large (external) conflict between their evidence.

Although the very concept of conflict (inconsistency) of information sources appeared in the pioneering works on the theory of evidence, the study of the concept of conflict using belief function theory has recently attracted the attention of many researchers, becoming a separate scientific direction. Within this direction, the axiomatics of measures of external and inter-

[1]This work was supported by the Russian Foundation for Basic Research, project no. 20-11-50077.

nal conflict, different methods for estimating the conflict, and the properties of measures of conflict were studied; several practical problems related to estimating the conflict were considered. In particular, estimating the amount of information conflict finds application in conflict management [3], the analysis of expert information and recommendations [4–6], the analysis of voting results [7], the design of trading strategies [8], image detection [9], etc. In addition, the concept of a metric between bodies of evidence, related to conflict, is used to approximate bodies of evidence [10–13] (to replace a "complex" body of evidence with a large number of focal elements by a body of evidence with a smaller number of focal elements), to classify data [14], to cluster data [15], etc. Due to the variety of such problems, this paper does not consider the applied aspects of inconsistency analysis within belief function theory.

The remainder of this paper is organized as follows. Section 1 recalls the fundamentals of belief function theory. In Section 2, we discuss combining rules for bodies of evidence. Section 3 deals with the concepts of conflict and non-conflict bodies of evidence and requirements applied to measures of external conflict. In Section 4, we analyze different axioms for a measure of conflict and present its general form satisfying a certain system of axioms. Section 5 considers different methods for estimating external conflict: metric (Section 5.1), structural (Section 5.2), algebraic (Section 5.3), and combining rules-based (Section 5.4). In Section 6, we describe a robust conflict estimation procedure. In the Conclusions, we summarize some findings of this study.

In addition to the conflict between the bodies of evidence, researchers consider the conflict of information provided by the same evidence (internal conflict). Part II of the survey will be devoted to this concept.

1. FUNDAMENTALS OF THE DEMPSTER–SHAFER THEORY OF EVIDENCE

Let $X=\{x_1,...,x_n\}$ be a finite set, and 2^X be the set of all subsets from *X.* In the Dempster–Shafer theory of evidence, a *basic belief assignment* (BBA, also termed a mass function) is a set function $m: 2^X \to [0,1]$ that satisfies the condition $\sum_{A\in 2^X} m(A)=1.$

As a rule, the normalized BBAs are considered: $m(\varnothing)=0$. Unnormalized BBAs are studied within the so-called *Transferable Belief Model* introduced in the paper [16]. In this case, a value $m(\varnothing)>0$ is interpreted as a measure of belief that the true value $x\notin X$.

A subset $A\subseteq X$ is called a focal element of a BBA m if $m(A)>0$. A pair $F=(\mathcal{A}, m)$ composed of the set of all focal elements $\mathcal{A}=\{A\}$ and a corresponding BBA $m(A)$, $A\in\mathcal{A}$, is called a body of evidence. We denote by $\mathcal{F}(X)$ the set of all bodies of evidence on X and by $\mathcal{P}(X)$ the set of all probability measures on X.

If a body of evidence $F=(\mathcal{A},m)$ is known, the Dempster—Shafer theory operates some set functions. Among them, the most important ones are the *belief function* $Bel(A)=\sum_{B\subseteq A} m(B)$ and its dual called the plausibility function $Pl(A)=1-Bel(A^c)$, where A^c indicates the complement of the set A. A mass function can be uniquely restored by the belief function using the Möbius transform: $m(A)=\sum_{B\subseteq A}(-1)^{|A\setminus B|}Bel(B)$. The inequality $Bel(A)\le Pl(A)$ holds $\forall A\subseteq X$, and the length of the interval $[Bel(A), Pl(A)]$ determines the degree of uncertainty of the event A; see [4]. The belief and plausibility functions will be denoted by Bel_F and Pl_F, respectively, whenever their dependence on the body of evidence $F=(\mathcal{A}, m)$ should be emphasized. The function $Pl(x)=\sum_{A\in\mathcal{A}:x\in A} m(A)$, $x\in X$, is called the contour function of a body of evidence.

A body of evidence $F=(\mathcal{A}, m)$ and the corresponding set functions can be defined in vector form: $\mathbf{m}$ is a $2^{|X|}$-dimensional vector composed of the values of the mass function $m(A)$, $A\in 2^X$, for some ordering of all subsets from 2^X; $\mathbf{Pl}$ is a $2^{|X|}$-dimensional vector composed of the values $Pl(A)$, $A\in 2^X$; and so on.

There are two main interpretations of BBAs and, therefore, two approaches to constructing belief function theory. According to the first interpretation, dating back to the paper [1], $m(A)$ is the probability of a random event (*random set*) A; for details, see [17]. In the statistical sense, a BBA is treated as the relative frequency of the fact that the true alternative belongs to the set *A*: $m(A)=q(A)/N$, where $q(A)$ is the number of observed sets $A\subseteq X$, and N is the total number of observations. In an alternative approach developed in the book [2], belief and plausibility functions are first introduced as set functions that satisfy the weakened axiomatics of a probability measure. These functions are then used to estimate the degree of uncertainty of information sources (evidence). Therefore, the Dempster–Shafer theory is also called the theory of evidence or belief function theory.

As shown in the paper [18],

$$Bel(A)=\inf_{P\in\mathcal{P}_{Bel}} P(A),\ Pl(A)=\sup_{P\in\mathcal{P}_{Bel}} P(A),\ A\subseteq X,$$

where $\mathcal{P}_{Bel}$ denotes the set of probability measures $P\in\mathcal{P}(X)$ agreed with Bel, i.e., $Bel(A)\le P(A)$ $\forall A\subseteq X$. The converse does not hold: the lower bound of a set of probability measures is not necessarily a belief function. The set $\mathcal{P}_{Bel}$ is also called the credal set of the function Bel.

Example 1. Suppose that ten experts predict the prospects for developing three technologies $X=\{a, b, c\}$: three experts spoke in favor of technologies a or b, two in favor of technology b, four in favor of technologies b or c, and one in favor of technology c. Without a known distribution between the alternatives (the uncertainty of information), the mass functions are given by

$$m(\{a, b\})=\frac{3}{10},\ m(\{b, c\})=\frac{4}{10},$$
$$m(\{b\})=\frac{2}{10},\ m(\{c\})=\frac{1}{10}.$$

Thus, we have the body of evidence $F=(\mathcal{A}, m)$ with the set of focal elements $\mathcal{A}=\{\{a, b\}, \{b, c\}, \{b\}, \{c\}\}$. This body of evidence can be used to find the values $Bel(A)$ and $Pl(A)$ $\forall A\subseteq X$. In particular, $Bel(\{a\})=0$, $Pl(\{a\})=0.3$, $Bel(\{b\})=0.2$, $Pl(\{b\})=0.9$, $Bel(\{c\})=0.1$, and $Pl(\{c\})=0.5$. These values can be considered lower and upper bounds for the probability of good prospects for developing particular technologies: $0\le P(\{a\})\le 0.3$, $0.2\le P(\{b\})\le 0.9$, and $0.1\le P(\{c\})\le 0.5$. Note that the greatest uncertainty exists in the forecast of technology b. ♦

The so-called *pignistic probability* $Bet_F\in\mathcal{P}(X)$ can be assigned to each body of evidence $F=(\mathcal{A}, m)$ [19]. This characteristic is the probability of an event $x_i\in X$, $i=1,..., n$, provided that the random variables within the focal elements obey the uniform distribution:

$$Bet_F(x_i)=\sum_{A\in\mathcal{A},\, x_i\in A}\frac{m(A)}{|A|},\ i=1,..., n.$$

These values coincide with the Shapley values

$$Bet_F(x_i)=$$
$$\sum_{A\in\mathcal{A},\, x_i\in A}\frac{(n-|A|)!(|A|-1)!}{n!}\big(Bel(A)-Bel(A\setminus\{x_i\})\big)$$

of the corresponding belief function Bel; see [20]. The pignistic probability of an arbitrary set $B\subseteq X$ is given by $Bet_F(B)=\sum_A\frac{|A\cap B|}{|A|}m(A)$. Note that $Bet_F\in\mathcal{P}_{Bel}$.

Belief function theory is developed not only for finite sets X. For example, the real axis (the so-called continuous belief structures) [21] or fuzzy focal elements [22] are studied.

Various special cases of belief functions and the corresponding bodies of evidence are considered in applications. They reflect simple and widespread structures of propositions on the belonging of a true alternative to a certain set. In particular, a belief function (and the corresponding body of evidence) is said to be:

– *categorical* if it has only one focal element; the corresponding body of evidence is the simplest one and will be denoted by $F_A=(A, 1)$;

– *vacuous* if the entire set X is the only focal element of this function, $F_X=(X, 1)$; such a body of evidence carries no information about the belonging of the true alternative to any subset of the set X (in this case, $Bel(A)=0$ and $Pl(A)=1$ $\forall A\ne\varnothing, X$);

– *consonant* if its focal elements are nested, i.e., $\forall A, B\in\mathcal{A}$: $A\subseteq B$ or $B\subseteq A$; the corresponding body of evidence is "refining";

– *simple* if the BBA has no more than two focal elements and, in the case of two focal elements, X is one of them; this body of evidence consists of the simplest meaningful proposition "$x\in A$" with a certain degree of belief $m(A)$ and a meaningless proposition with a degree of belief $1-m(A)$;

– *dogmatic* if $X\notin\mathcal{A}$ (the body of evidence does not contain a meaningless proposition: $m(X)=0$).

Any body of evidence $F=(\mathcal{A}, m)$ can be represented as $F=\sum_{A\in\mathcal{A}}m(A)F_A$, i.e., as a convex combination of categorical bodies of evidence. In particular, the body of evidence from Example 1 can be written in the form $F=0.2F_{\{b\}}+0.1F_{\{c\}}+0.3F_{\{a,b\}}+0.4F_{\{b,c\}}$. A simple body of evidence can be represented as $F_A^{\omega}=(1-\omega)F_A+\omega F_X$, where $\omega\in[0, 1]$. In particular, $F_A^0=F_A$ and $F_A^1=F_X$. Conversely, a finite sum $F=\sum_i\alpha_iF_{A_i}$, $\alpha_i\ge 0\ \forall i$, $\sum_i\alpha_i=1$, defines a certain body of evidence.

The amount of ignorance in the information contained in a body of evidence F can be estimated using the so-called imprecision indices [23]: the more focal elements of high power and with a greater mass a body of evidence has, the greater value these indices will take. An example of such an index is the generalized Hartley measure [24, 25] $H(F)=\sum_{A\in\mathcal{A}}m(A)\log_2|A|$. Below, we will use the normalized generalized Hartley measure $H_0(F)=$

$H(F)/\log_2|X|$, $H_0:\mathcal{F}(X)\to[0,1]$. Note that $H_0(F)=1\Leftrightarrow F=F_X$, and $H_0(F)=0\Leftrightarrow F=P$, where $P\in\mathcal{P}(X)$.

Despite the popularity and high demand, belief function theory was criticized [26–28], and several studies were initiated to clarify the theory's limits of applicability, interpretability of results, etc. As demonstrated in the paper [18], many critical remarks are due to mixing two points of view on belief functions: the first considers a belief function as a generalization of a probabilistic measure, and the second as a way of representing evidence.

2. COMBINING RULES FOR BODIES OF EVIDENCE AND THE CONCEPT OF CONFLICT

A convenient tool in the theory of evidence is the ability to combine bodies of evidence, i.e., to aggregate information obtained from different sources. A combining rule is understood as a certain operation $\otimes:\mathcal{F}(X)\times\mathcal{F}(X)\to\mathcal{F}(X)$. There are several general schemes for constructing such rules. The most widespread one is the so-called generalized conjunctive combining rule $\otimes_\cap$ with the following scheme [29]: $F=F_1\otimes_\cap F_2$, $F=(\mathcal{A}, m_\cap)$, $F_1=(\mathcal{A}_1, m_1)$ and $F_2=(\mathcal{A}_2, m_2)$, where

$$m_\cap(A)=\sum_{B\cap C=A}\tilde{m}(B, C),$$

and the values of the set function $\tilde{m}:2^X\times 2^X\to[0, 1]$ satisfy the matching conditions with the bodies of evidence F_1 and F_2:

$$\begin{gathered}\sum_{C\in 2^X}\tilde{m}(B, C)=m_1(B),\\ \sum_{B\in 2^X}\tilde{m}(B, C)=m_2(C),\\ B, C\in 2^X.\end{gathered}\qquad(1)$$

This system may have a set of solutions, yielding a set of new bodies of evidence $F=F_1\otimes_\cap F_2$. The latter set will be denoted by $\mathcal{R}_\cap(F_1, F_2)$.

Example 2. Let $X=\{x_1, x_2\}$, $F_1=0.8F_{\{x_1\}}+0.2F_X$, and $F_2=0.7F_{\{x_1\}}+0.3F_{\{x_2\}}$. To find the set $\mathcal{R}_\cap(F_1, F_2)$, we solve the system

$$\sum_C\tilde{m}(\{x_1\}, C)=0.8,\ \sum_C\tilde{m}(X, C)=0.2,$$
$$\sum_B\tilde{m}(B, \{x_1\})=0.7,\ \sum_B\tilde{m}(B, \{x_2\})=0.3,$$
$$\tilde{m}(B, C)\in[0, 1]\ \ \forall B, C\in 2^X.$$

Since $\tilde{m}(\cdot, \varnothing)=\tilde{m}(\varnothing, \cdot)=\tilde{m}(\cdot, X)=\tilde{m}(\{x_2\}, \cdot)=0$, we have a system of four equations and inequalities in four variables. The solution is the numbers $\tilde{m}(\{x_1\}, \{x_1\})=0.5+t$, $\tilde{m}(\{x_1\}, \{x_2\})=0.3-t$, $\tilde{m}(X, \{x_1\})=0.2-t$, and $\tilde{m}(X, \{x_2\})=t$, $t\in[0, 0.2]$. Then $\mathcal{R}_\cap(F_1, F_2)=\{F_1\otimes_\cap F_2=(0.3-t)F_\varnothing+0.7F_{\{x_1\}}+tF_{\{x_2\}}: t\in[0, 0.2]\}$. ♦

In particular, if the information sources are independent, then $\tilde{m}(B, C)=m_1(B)m_2(C)\ \ \forall B, C\in 2^X$, and we obtain the so-called unnormalized Dempster rule $\otimes_{ND}: m_{ND}(A)=\sum_{B\cap C=A}m_1(B)m_2(C)\ \ \forall A\in 2^X$. In this case, it may happen that $K=K(F_1, F_2)=m_{ND}(\varnothing)=\sum_{B\cap C=\varnothing}m_1(B)m_2(C)>0$. The value $K\in[0, 1]$ is called **the canonical measure of conflict**. It characterizes the degree of conflict of information sources described by the bodies of evidence F_1 and F_2: the greater this value is, the more inconsistent information the sources will provide. Uniformly distributing the amount of conflict over all focal elements of the new body of evidence, we get the classical Dempster rule $\otimes_D: m_D(A)=\dfrac{m_{ND}(A)}{1-K}\ \ \forall A\in 2^X\setminus\varnothing$ [1]. If the information sources have complete conflict ($B\cap C=\varnothing$ for all $B\in\mathcal{A}_1$ and $C\in\mathcal{A}_2$), the Dempster combining rule becomes inapplicable: $K=1$. Relating the amount of conflict to the mass of the entire set $m_{ND}(X)$ (multiplying the mass of the meaningless proposition $x\in X$ by the amount of conflict), we get the Yager rule $\otimes_Y: m_Y(A)=m_{ND}(A)\ \ \forall A\in 2^X\setminus\{\varnothing, X\}$, $m_Y(X)=m_{ND}(X)+K$ [30].

The combining operation $\otimes_{ND}$ is associative and hence can be used to combine any finite number of bodies of evidence. Information on several other combining rules was provided in the report [31].

In some sense, the disjunctive consensus rule $\otimes_\cup$ [32, 33] is dual to the Dempster rule:

$$m_\cup(A)=\sum_{B\cup C=A}m_1(B)m_2(C),\ A\in 2^X.$$

(The body of evidence obtained using this rule is denoted by $F=F_1\otimes_\cup F_2=(\mathcal{A}, m_\cup)$.) The conjunctive and disjunctive combining rules satisfy an analog of de Morgan's law [32]:

$$\overline{F_1\otimes_{ND}F}=\overline{F}_1\otimes_\cup\overline{F_2},$$

where the complement $\overline{F}=\left(\overline{\mathcal{A}}, \overline{m}\right)$ of a body of evidence $F=(\mathcal{A}, m)$ is defined by $\overline{\mathcal{A}}=\left\{A^c: A\in\mathcal{A}\right\}$ and $\overline{m}(A)=m\left(A^c\right)\ \ \forall A\in\overline{\mathcal{A}}$.

Example 3. Suppose that two independent expert groups predict the prospects of developing three technologies $X=\{a, b, c\}$. Information from the first group is described by the body of evidence $F_1=0.2F_{\{b\}}+0.1F_{\{c\}}+0.3F_{\{a,b\}}+0.4F_{\{b,c\}}$ (Example 1). Information from the second group is described by the body of evidence

$F_2 = 0.7F_{\{a,c\}} + 0.3F_{\{a,b,c\}}$. The results of combining these bodies of evidence using different conjunctive rules and the disjunctive consensus rule are presented in Table 1. (Only the masses of focal elements are indicated there.)

For these bodies of evidence, the canonical measure of conflict is $K = m_{DN}(\varnothing) = 0.14$. The last row of Table 1 contains the values of the imprecision index H_0 for each combination result. ♦

Remark 1. A conjunctive rule is optimistic in the following sense. Suppose there are two independent information sources. According to the first rule, the true alternative belongs to a set A (a categorical body of evidence F_A); according to the second one, to a set B (a categorical body of evidence F_B). After the conjunctive combination of these bodies, we establish that the true alternative belongs to the set $A \cap B$. Applying the disjunctive consensus rule, we obtain that the true alternative belongs to the set $A \cup B$; for details, see the paper [34]. In this sense, the rule under consideration is pessimistic.

The so-called matrix of evidence $R_F^{\otimes} = \left\{r_F^{\otimes}(A,B)\right\}_{A,B\in 2^X}$ with the elements $r_F^{\otimes}(A,B) = (F \otimes F_B)(A)$, where F_B is a categorical body of evidence, can be assigned to each combining rule and a fixed body of evidence $F = (\mathcal{A}, m)$; see [35, 36]. In particular, the matrix representations $R_F^{\cap}$ and $R_F^{\cup}$ for the unnormalized Dempster rule $\otimes_{ND}$ and the disjunctive consensus rule $\otimes_{\cup}$, respectively, are widespread: due to the remark above, they set completely information about the body of evidence F itself (since $F \otimes_{ND} F_X = F$ and $F \otimes_{\cup} F_{\varnothing} = F$) and about the combination results with other categorical bodies of evidence. For different categorical bodies of evidence, these results are written in the matrix $R_F^{\otimes}$ column-by-column. The matrices $R_F^{\cap}$ and $R_F^{\cup}$ are called the specialization and generalization matrices of the body of evidence F, respectively, because they contain information about all bodies of evidence of the form $F_B^{\cap} = (\mathcal{A}_B^{\cap}, m)$ and $F_B^{\cup} = (\mathcal{A}_B^{\cup}, m)$, respectively, where $\mathcal{A}_B^{\cap} = \{A \cap B : A \in \mathcal{A}\}$ and $\mathcal{A}_B^{\cup} = \{A \cup B : A \in \mathcal{A}\}$. Obviously, $R^{\cap}_{F_1 \otimes_{ND} F_2} = R^{\cap}_{F_1} \cdot R^{\cap}_{F_2}$. The papers [35] and [37] considered matrix operators corresponding to certain parametric families of conjunctive and disjunctive rules (the so-called α-*junctions*). For $\alpha = 1$, they turn into the operators $R_F^{\cap}$ and $R_F^{\cup}$.

Example 4. Consider the bodies of evidence $F_1 = 0.8F_{\{x_1\}} + 0.2F_X$ and $F_2 = 0.7F_{\{x_1\}} + 0.3F_{\{x_2\}}$ on the set $X = \{x_1, x_2\}$ (Example 2). For these bodies, we obtain:

$$R^{\cap}_{F_1} = \begin{pmatrix} 1 & 0 & 0.8 & 0 \\ 0 & 1 & 0 & 0.8 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.2 \end{pmatrix}, \quad R^{\cap}_{F_2} = \begin{pmatrix} 1 & 0.3 & 0.7 & 0 \\ 0 & 0.7 & 0 & 0.7 \\ 0 & 0 & 0.3 & 0.3 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$R^{\cup}_{F_1} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0.8 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.2 & 0.2 & 1 & 1 \end{pmatrix}, \quad R^{\cup}_{F_2} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0.7 & 0.7 & 0 & 0 \\ 0.3 & 0 & 0.3 & 0 \\ 0 & 0.3 & 0.7 & 1 \end{pmatrix}.$$

(The rows and columns correspond to the subsets $\varnothing$, $\{x_1\}$, $\{x_2\}$, and X.) ♦

Table 1

The results of combining bodies of evidence

A	F_1	F_2	$F_1 \otimes_{ND} F_2$	$F_1 \otimes_D F_2$	$F_1 \otimes_Y F_2$	$F_1 \otimes_{\cup} F_2$
$\varnothing$	–	–	0.14	–	–	–
$\{a\}$	–	–	0.21	0.244	0.21	–
$\{b\}$	0.2	–	0.06	0.07	0.06	–
$\{c\}$	0.1	–	0.38	0.442	0.38	–
$\{a, b\}$	0.3	–	0.09	0.104	0.09	–
$\{a, c\}$	–	0.7	–	–	–	0.07
$\{b, c\}$	0.4	–	0.12	0.14	0.12	–
$\{a, b, c\}$	–	0.3	–	–	0.14	0.93
H_0	0.442	0.742	0.132	0.154	0.272	0.974

3. THE CONCEPT OF CONFLICT AND NON-CONFLICT BODIES OF EVIDENCE

In the general case, the measure of external conflict is understood as a certain functional $Con_{\text{ext}} : \mathcal{F}(X) \times \mathcal{F}(X) \to [0, 1]$. Each body of evidence is determined by a set of focal elements and a mass function on these sets. Hence, the conflict between two bodies of evidence depends on the mutual arrangement of their sets of focal elements and the values of their mass functions.

First of all, the measure of conflict should reflect the extreme situations of conflict: be maximum in case of complete conflict of two bodies of evidence and minimum in case of no conflict.

A natural condition for complete conflict is that all pairs of focal elements do not intersect for two bodies of evidence. More precisely, bodies of evidence $F_1 = (\mathcal{A}_1, m_1)$ and $F_2 = (\mathcal{A}_2, m_2)$ are said to be in complete conflict if $A \cap B = \varnothing \;\; \forall A \in \mathcal{A}_1, \; \forall B \in \mathcal{A}_2$.

However, different degrees of non-conflict are possible. For example, the paper [38] considered the following degrees of non-conflict for bodies of evidence $F_1 = (\mathcal{A}_1, m_1)$ and $F_2 = (\mathcal{A}_2, m_2)$:

– strong non-conflict: $\bigcap_{A \in \mathcal{A}_1 \cup \mathcal{A}_1} A \neq \varnothing$,

– (simple) non-conflict: $A \cap B \neq \varnothing \quad \forall A \in \mathcal{A}_1$, $\forall B \in \mathcal{A}_2$,

– weak non-conflict: $\mathcal{P}_{Bel_1} \cap \mathcal{P}_{Bel_2} \neq \varnothing$, where Bel_i is a belief function corresponding to a body of evidence F_i, $i = 1, 2$.

As is known [38], strong non-conflict implies (simple) non-conflict, and non-conflict implies weak non-conflict. For categorical bodies of evidence F_A and F_B, all these concepts of non-conflict reduce to the non-empty intersection of their focal elements: $A \cap B \neq \varnothing$. However, e.g., if $F_1 = F_2 = \alpha F_A + (1-\alpha) F_B$, where $A \cap B = \varnothing$ and $\alpha \in (0,1)$, then such identical bodies of evidence are in weak non-conflict only: they will be neither (simply) non-conflict nor strongly non-conflict.

Due to the conditions of complete conflict and non-conflict, the measure of external conflict Con_{ext} should satisfy the following conditions [4, 38, 39]:

E1: $Con_{\text{ext}}(F_1, F_2) = Con_{\text{ext}}(F_2, F_1) \; \forall F_1, F_2 \in \mathcal{F}(X)$ (symmetry).

E2: $Con_{\text{ext}}(F_1, F_2) = 0$ if F_1 and F_2 are weakly non-conflict.

In particular, any body of evidence F will be weakly non-conflict with itself and the meaningless body of evidence F_X. Therefore, under condition E2, we have the following corollaries:

a) $Con_{\text{ext}}(F, F) = 0 \;\; \forall F \in \mathcal{F}(X)$ (nilpotency).

b) $Con_{\text{ext}}(F_X, F) = 0 \;\; \forall F \in \mathcal{F}(X)$ (non-conflict with ignorance).

E3: $Con_{\text{ext}}(F_1, F_2) = 1$ if $A \cap B = \varnothing \quad \forall A \in \mathcal{A}_1$, $\forall B \in \mathcal{A}_2$ (complete conflict).

The next condition conceptually relates to the specialization of evidence [32].

A body of evidence $F' = (\mathcal{A}', m')$ is called a specialization of a body of evidence $F'' = (\mathcal{A}'', m'')$ (and denoted by $F' \sqsubseteq F''$) if there exists a partition $\mathcal{A}' = \mathcal{A}_1' \cup \ldots \cup \mathcal{A}_k'$, $\mathcal{A}_i' \cap \mathcal{A}_j' = \varnothing \quad \forall i \neq j$, $k = |\mathcal{A}''|$ such that $\bigcup_{A \in \mathcal{A}_i'} A \subseteq B_i$ and $\sum_{A \in \mathcal{A}_i'} m'(A) = m''(B_i)$, $\forall B_i \in \mathcal{A}''$, $i = 1, \ldots, k$.

In other words, a body of evidence F' refines (specializes) a body of evidence F''. The latter body is called a generalization of a body of evidence F'. Note that $F \sqsubseteq F_X \;\; \forall F \in \mathcal{F}(X)$.

If $F' \sqsubseteq F''$, then the imprecision of F' does not exceed that of F'' in terms of any index (e.g., $H(F') \leq H(F'')$ for the generalized Hartley measure).

E4: $Con_{\text{ext}}(F', F) \geq Con_{\text{ext}}(F'', F) \quad \forall F, F', F'' \in \mathcal{F}(X)$, $F' \sqsubseteq F''$ (antimonotonicity with respect to specialization).

Note that the canonical measure of conflict K satisfies all conditions mentioned, except E2 (particularly is not nilpotent). At the same time, it satisfies

E2′: $Con_{\text{ext}}(F_1, F_2) = 0$ if F_1 and F_2 are (simply) non-conflict.

4. AN AXIOMATICS FOR A MEASURE OF EXTERNAL CONFLICT

The paper [40] considered an axiomatics for a measure of conflict based on strengthening conditions E1 – E4:

A1: $Con_{\text{ext}}(F_1, F_2) = Con_{\text{ext}}(F_2, F_1) \; \forall F_1, F_2 \in \mathcal{F}(X)$.

A2: $Con_{\text{ext}}(F_1, F_2) = 0 \Leftrightarrow F_1$ and F_2 are weakly non-conflict.

A3: $Con_{\text{ext}}(F_1, F_2) = 1 \Leftrightarrow A \cap B = \varnothing \quad \forall A \in \mathcal{A}_1$, $\forall B \in \mathcal{A}_2$.

A4: $Con_{\text{ext}}(F', F) \geq Con_{\text{ext}}(F'', F) \quad \forall F, F', F'' \in \mathcal{F}(X)$, $F' \sqsubseteq F''$.

Also, the following additional axiom was introduced therein:

A5: If $Con_{\text{ext}}(F_1, F_2) = a \in [0, 1]$, then $\exists F_i^{(k)} \in \mathcal{F}(X)$, $k = 1, 2$, $i = 1, 2$: $F_i = (1-a) F_i^{(1)} + a F_i^{(2)}$, $i = 1, 2$, $Con_{\text{ext}}(F_1^{(1)}, F_2^{(1)}) = 0$, $Con_{\text{ext}}(F_1^{(2)}, F_2^{(2)}) = 1$. This axiom assumes that information sources can be linearly divided into completely conflict and weakly non-conflict parts proportionally to the amount of conflict estimated by a given measure. This is a rather strong requirement. In particular, the canonical measure of conflict does not satisfy this axiom. (This fact can be easily verified by an appropriate example.)

The following theorem on the measure of conflict is true.

Theorem 1 [40]. *If a functional* $Con_{\text{ext}}: \mathcal{F}(X)\times\mathcal{F}(X)\to[0,1]$ *satisfies axioms* A1 – A5*, it has the form*

$$Con_{\text{ext}}(F_1, F_2) = \inf\{m_\cap(\varnothing): F = (\mathcal{A}, m_\cap) \in \mathcal{R}_\cap(F_1, F_2)\}. \quad (2)$$

Example 5. Consider the bodies of evidence $F_1 = 0.8F_{\{x_1\}} + 0.2F_X$ and $F_2 = 0.7F_{\{x_1\}} + 0.3F_{\{x_2\}}$ on the set $X = \{x_1, x_2\}$. Applying the generalized conjunctive combining rule $\otimes_\cap$, we obtain $m_\cap(\varnothing) = 0.3 - t$, $t \in [0, 0.2]$ (Example 2). Hence, $Con_{\text{ext}}(F_1, F_2) = \min\{0.3 - t : t \in [0, 0,2]\} = 0.1$. Note that for these bodies of evidence, the canonical measure of conflict is $K = 0.24$. ♦

As shown in the paper [40], the measure (2) can be calculated by the formula

$$Con_{\text{ext}}(F_1, F_2) = \inf\{Con_{ext}(P_1, P_2): P_1 \in \mathcal{P}_{Bel_1}, P_2 \in \mathcal{P}_{Bel_2}\},$$

where $\mathcal{P}_{Bel_i}$ denotes the creedal sets corresponding to the bodies of evidence F_i, $i = 1,2$. For probability measures,

$$Con_{\text{ext}}(P_1,P_2) = 1 - \sum_{i=1}^{n} \min\{P_1(x_i), P_2(x_i)\}. \quad (3)$$

Example 6. For the previous example, we have: $\mathcal{P}_{Bel_1} = \{\alpha F_{\{x_1\}} + (1-\alpha)F_{\{x_2\}} : 0.8 \le \alpha \le 1\}$, $\mathcal{P}_{Bel_2} = \{F_2\} = \{0.7F_{\{x_1\}} + 0.3F_{\{x_2\}}\}$. Then $Con_{\text{ext}}(F_1, F_2) = \inf\{Con_{ext}(P_1, P_2): P_1 \in \mathcal{P}_{Bel_1}, P_2 \in \mathcal{P}_{Bel_2}\} = \inf_{0.8\le\alpha\le1}(1 - \min\{\alpha; 0.7\} - \min\{1-\alpha; 0.3\}) = \inf_{0.8\le\alpha\le1}(1-0.7-(1-\alpha)) = 0.1$. ♦

Remark 2. For probability measures $P_1, P_2 \in \mathcal{P}(X)$, the canonical measure of conflict K is given by

$$K(P_1,P_2) = \sum_{i=1}^{n} P_1(x_i) \sum_{j=1, j\ne i}^{n} P_2(x_j) = \sum_{i=1}^{n} P_1(x_i)\big(1 - P_2(x_i)\big) = 1 - \sum_{i=1}^{n} P_1(x_i)P_2(x_i).$$

(Compare with formula (3).) It follows that the canonical measure of conflict is generally not nilpotent since $K(P, P) = 1 - \sum_{i=1}^{n} P^2(x_i)$, $P \in \mathcal{P}(X)$. In particular, for the uniform distribution P on X, we obtain $K(P, P) = 1 - \frac{1}{|X|}$.

The paper [4] considered an axiomatics for the measure of external conflict Con_{ext} on an arbitrary finite set of bodies of evidence $M = \{F_1,..., F_l\}$, $F_i \in \mathcal{F}(X)$, $i = 1,..., l$. Let 2^M be the set of all subsets from the set M. By definition, $Con_{\text{ext}}(B) = 0$ if $|B| = 1$, $B \in 2^M$, and $Con_{\text{ext}}(\varnothing) = 0$. In addition to direct generalizations of axioms A1 – A4 to the set case, the monotonicity axiom was introduced:

A6: $Con_{\text{ext}}(B) \le Con_{\text{ext}}(C)$ if $B \subseteq C$ and $B, C \in 2^M$.

For example, due to associativity, the canonical measure of conflict can be extended to an arbitrary finite set M of bodies of evidence. As is easily demonstrated, the canonical measure of conflict will satisfy the monotonicity axiom A6 on this set.

5. METHODS TO ESTIMATE EXTERNAL CONFLICT

There are several approaches to estimating the conflict of bodies of evidence. They can be conventionally divided into metric, structural, algebraic, and combining rule-based ones.

5.1. Metric methods to estimate conflict

The metric approach is one of the most popular methods to estimate conflict [41, 42]. Metrics can be introduced on the bodies of evidence themselves and their set functions or matrix representations. The definition of metrics on the set $\mathcal{F}(X)$ of bodies of evidence considers, to one degree or another, structural features of bodies of evidence: the power of the focal elements, the degree of intersection for the focal elements of two bodies of evidence, the mutual arrangement of the focal elements, and others. For measuring conflict, a metric d must be normalized so that $d: \mathcal{F}(X)\times\mathcal{F}(X)\to[0,1]$.

Metric between bodies of evidence based on the generalized Euclidean distance. This metric is given by

$$d_Q(F_1,F_2) = \|\mathbf{m}_1 - \mathbf{m}_2\|_Q = \sqrt{\tfrac{1}{2}(\mathbf{m}_1 - \mathbf{m}_2)^{\mathrm{T}} Q(\mathbf{m}_1 - \mathbf{m}_2)}, \quad (4)$$

where $\mathbf{m}_i$ denotes a $2^{|X|}$-dimensional column vector composed of the mass function values $m_i(A)$, $A \in 2^X$, $i = 1, 2$, for some ordering of all subsets from 2^X; $Q = (q_{A,B})_{A,B\in2^X}$ is the matrix of similarity measures between the focal elements with the entries $q_{A,B} = q_{B,A} \in [0,1]$ such that $q_{A,A} = 1$ for $A \in 2^X \setminus \varnothing$, and $q_{A,B} = 0$ if $A \cap B = \varnothing$. Then $d_Q \in [0, 1]$. For example, a popular metric of this type was described in [41]. As a similarity measure, it involves *the Jaccard index* [43]: $J = (jac_{A,B})_{A,B\in2^X}$, where $jac_{A,B} = \frac{|A\cap B|}{|A\cup B|}$ for $A, B \ne \varnothing$ and $d_{\varnothing,\varnothing} = 0$. The positive definiteness of the matrix (the fact that d_J is a metric on the set $\mathcal{F}(X)\times\mathcal{F}(X)$) was proved in the paper [44]. In par-

ticular, for categorical bodies of evidence, we have $d_J(F_A, F_B) = \sqrt{1 - \frac{|A \cap B|}{|A \cup B|}}$. In [45], the metric (4) with the Jaccard similarity measure was generalized to the class of continuous belief structures.

Example 7. Consider the bodies of evidence $F_1 = 0.8F_{\{x_1\}} + 0.2F_X$ and $F_2 = 0.7F_{\{x_1\}} + 0.3F_{\{x_2\}}$ on the set $X = \{x_1, x_2\}$ (Example 2). We have: $J = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{pmatrix}$, $\mathbf{m}_1 = (0.8; 0; 0.2)^{\mathrm{T}}$, $\mathbf{m}_2 = (0.7; 0.3; 0)^{\mathrm{T}}$, and $d_J(F_1, F_2) = \frac{1}{2\sqrt{5}} \approx 0.224$. ♦

Among other similarity measures used in formula (4), we mention the Sørensen and Simpson coefficients: $sor_{A,B} = \frac{2|A \cap B|}{|A| + |B|}$ and $sim_{A,B} = \frac{|A \cap B|}{\min\{|A|, |B|\}}$ [46]. However, the corresponding matrix is not always positive definite, and the distance function is not always a (complete) metric. The paper [47] proposed a metric of the form (4) with a metric space X and the similarity measure $q_{A,B} = \frac{1}{1 + c \cdot d_H(A,B)}$, where $d_H(A, B)$ is the Hausdorff metric between sets A and B of the space X and $c > 0$. This similarity measure[1] does not satisfy the condition $q_{A,B} = 0$ for $A \cap B = \varnothing$. This metric was used in [47] to aggregate information from different sensors under uncertain measurement errors and uncertain noise characteristics.

Example 8. Consider the bodies of evidence $F_1 = 0.5F_{[40,50]} + 0.3F_{[45,55]} + 0.2F_X$ and $F_2 = 0.3F_{[30,60]} + 0.7F_{[40,50]}$ on the set $X = [20, 70]$, which predict the company's stock value. We find the distance between them using the metric (4) with $q_{A,B} = \frac{1}{1 + 0.2 d_H(A,B)}$, where d_H denotes the Hausdorff metric. According to the body of evidence F_1, the stock value will belong to the interval [40, 50] with the degree of belief 0.5, the interval [45, 55] with the degree of belief 0.3, and the interval $X = [20, 70]$ with the degree of belief 0.2 (in conditional units). The body of evidence F_2 is interpreted by analogy. Consider the vector representations of these bodies of evidence on the sets $\mathcal{A}_1 \cup \mathcal{A}_2 = \{[40, 50], [45, 55], [30, 60], [20, 70]\}$ only (with the specified ordering). Then $\mathbf{m}_1 = (0.5; 0.3; 0; 0.2)^{\mathrm{T}}$ and $\mathbf{m}_2 = (0.7; 0; 0.3; 0)^{\mathrm{T}}$. For intervals $[a_1, b_1]$ and $[a_2, b_2]$, the Hausdorff measure is given by $d_H([a_1, b_1], [a_2, b_2]) = \max\{|a_1 - a_2|, |b_1 - b_2|\}$. Therefore, the matrix $Q = \left(q_{[a_i,b_i],[a_j,b_j]}\right)_{i,j=1}^{4}$ has the form

$$Q = \begin{pmatrix} 1 & 1/2 & 1/3 & 1/5 \\ 1/2 & 1 & 1/4 & 1/6 \\ 1/3 & 1/4 & 1 & 1/3 \\ 1/5 & 1/6 & 1/3 & 1 \end{pmatrix},$$

and $d_Q(F_1, F_2) = \sqrt{\frac{1}{2}(\mathbf{m}_1 - \mathbf{m}_2)^{\mathrm{T}} Q (\mathbf{m}_1 - \mathbf{m}_2)} \approx 0.282$. ♦

The metric

$$d(F_1, F_2) = \frac{1}{2|X| - 1} \sum_{A \in 2^X} |A| |m_1(A) - m_2(A)|,$$

suggested in [11], is another one calculated directly between bodies of evidence with information uncertainty. This metric[2] was used therein to approximate complex bodies of evidence by those with a simpler structure of focal elements.

Metric between set functions describing bodies of evidence bijectively. Consider distance functions (metrics, semimetrics, and pseudometrics) between set functions describing bodies of evidence bijectively. Among them, we separate the Minkowski metrics l_p, $1 \le p \le \infty$, between belief functions or plausibility functions:

$$d_p(F_1, F_2) = c\left(\sum_{A \in 2^X} |Pl_1(A) - Pl_2(A)|^p\right)^{\frac{1}{p}} = c\left(\sum_{A \in 2^X} |Bel_1(A) - Bel_2(A)|^p\right)^{\frac{1}{p}}.$$

(Since $|Pl_1(A) - Pl_2(A)| = |Bel_1(A^c) - Bel_2(A^c)|$ $\forall A \in 2^X$, these metrics coincide; see the papers [11, 48].) For $p = \infty$, the sum is replaced by the maximum over all subsets $A \in 2^X$. Below, the normalizing factor $c > 0$ is chosen so that $d \in [0, 1]$. Such metrics are often used to approximate bodies of evidence.

Metrics between set functions describing bodies of evidence not bijectively. One example is the Minkowski metric l_p, $1 \le p \le \infty$, between pignistic probabilities [49, 50]

$$d_{Bet,p}(F_1, F_2) = c\left(\sum_{A \in 2^X} |Bet_{F_1}(A) - Bet_{F_2}(A)|^p\right)^{\frac{1}{p}},$$

between the probabilities of basic set elements

$$d_{Bet,p,x}(F_1, F_2) = c\left(\sum_{x \in X} |Bet_{F_1}(x) - Bet_{F_2}(x)|^p\right)^{\frac{1}{p}},$$

[1] The positive definiteness of the similarity matrix was established in [47] for a particular case.

[2] The paper [11] considered the unnormalized metric.

or between the so-called contour functions of bodies of evidence [51]

$$d_{Pl,p,x}(F_1, F_2) = c\left(\sum_{x\in X}\left|Pl_1(x) - Pl_2(x)\right|^p\right)^{\frac{1}{p}}.$$

However, these functions are only pseudometrics on $\mathcal{F}(X)\times\mathcal{F}(X)$: they satisfy the reflexivity axiom $d(F, F) = 0$ instead of the coincidence axiom $d(F_1, F_2) = 0 \Leftrightarrow F_1 = F_2$.

In particular, the pseudometric $d_{Bet,1,x}$ was used in [14] to train a classifier with classes described by belief functions.

Normalized semimetrics based on the scalar product of vector representations of evidence bodies. (A semimetric satisfies all axioms of a metric, except for the triangle inequality.) An example of such a semimetric is the function introduced in [42]:

$$d(F_1, F_2) = 1 - \frac{\mathbf{Pl}_1^{\mathrm{T}}\cdot\mathbf{Pl}_2}{\|\mathbf{Pl}_1\|\|\mathbf{Pl}_2\|},$$

where $\mathbf{Pl}_i$ is a $2^{|X|}$-dimensional column vector composed of the plausibility function values $Pl_i(A)$, $A\in 2^X$, $i = 1, 2$; $\|\cdot\|$ denotes the Euclidean norm.

Example 9. Consider the bodies of evidence $F_1 = 0.8F_{\{x_1\}} + 0.2F_X$ and $F_2 = 0.7F_{\{x_1\}} + 0.3F_{\{x_2\}}$ on the set $X = \{x_1, x_2\}$ (Example 2). We have: $\mathbf{Pl}_1 = (1;\ 0.2;\ 1)^{\mathrm{T}}$, $\mathbf{Pl}_2 = (0.7;\ 0.3;\ 0)^{\mathrm{T}}$, and $d(F_1, F_2) = 1 - \frac{\mathbf{Pl}_1^{\mathrm{T}}\cdot\mathbf{Pl}_2}{\|\mathbf{Pl}_1\|\|\mathbf{Pl}_2\|} \approx 0.301$. ♦

Metrics based on comparing uncertainty intervals $\{[Bel_1(A), Pl_1(A)]\}_{A\in 2^X}$ and $\{[Bel_2(A),\ Pl_2(A)]\}_{A\in 2^X}$ that describe bodies of evidence F_1 and F_2 bijectively. In this case, a metric on the set of bodies of evidence can be defined by extending the metric on uncertainty intervals:

$$d_{I,p}(F_1, F_2) = c\times$$

$$\left(\sum_{A\in 2^X} d_I^p\left([Bel_1(A),\ Pl_1(A)], [Bel_2(A),\ Pl_2(A)]\right)\right)^{\frac{1}{p}}, \quad (5)$$

$$1\le p\le\infty,$$

where d_I denotes the metric on the intervals. For example, the Hausdorff metric d_H can be used as d_I; see Example 8. The paper [12] considered a metric of the form (5) with $d_I = d_{W,p}([a_1, b_1], [a_2, b_2])$ (the Wasserstein p-metric of two uniform distributions on intervals $[a_1, b_1]$ and $[a_2, b_2]$). This metric was applied to approximate bodies of evidence. In particular, $d_{W,2}([a_1, b_1], [a_2, b_2]) = \sqrt{(mean_1 - mean_2)^2 + \frac{1}{3}(rad_1 - rad_2)^2}$, where $mean_i$ and rad_i are the middle and half-length of an interval $[a_i, b_i]$, $i = 1, 2$. As shown in [12], for $p = 2$, the normalizing factor is $c = 2^{-\frac{|X|-1}{2}}$. This metric estimates the deviation between the information uncertainties of bodies of evidence.

Example 10. Consider the bodies of evidence $F_1 = 0.8F_{\{x_1\}} + 0.2F_X$ and $F_2 = 0.7F_{\{x_1\}} + 0.3F_{\{x_2\}}$ on the set $X = \{x_1, x_2\}$ (Example 2). We find the values of the metrics $d_{W,2}(F_1, F_2)$ and $d_{H,2}(F_1, F_2)$. For $|X| = 2$ and $p = 2$, the normalizing factor is $c = \frac{1}{\sqrt{2}}$. The calculation results are presented in Table 2.

Then $d_{W,2}(F_1, F_2) = \frac{1}{10}\sqrt{\frac{13}{3}} \approx 0.208$ and $d_{H,2}(F_1, F_2) = 0.3$. ♦

The Wasserstein metric. According to the paper [52], a metric d defined on 2^X can be extended to the set $\mathcal{F}(X)$ by solving the Kantorovich problem

$$d_W(F_1, F_2) = \min_{\tilde{m}}\sum_{A\in 2^X}\sum_{A\in 2^X}\tilde{m}(A, B)d(A, B).$$

Here, the minimum is taken over all set functions satisfying the matching conditions (1). The result can be treated as a Wasserstein 1-metric between bodies of evidence. This metric has a good interpretation as a solution of the Kantorovich optimization problem.

Example 11. Consider the bodies of evidence $F_1 = 0.8F_{\{x_1\}} + 0.2F_X$ and $F_2 = 0.7F_{\{x_1\}} + 0.3F_{\{x_2\}}$ on the set $X = \{x_1, x_2\}$ (Example 2). We have: $\tilde{m}(\{x_1\}, \{x_1\}) = 0.5 + t$, $\tilde{m}(\{x_1\}, \{x_2\}) = 0.3 - t$, $\tilde{m}(X, \{x_1\}) = 0.2 - t$, $\tilde{m}(X, \{x_2\}) = t$, $t\in[0, 0.2]$, and $\tilde{m}(A, B) = 0$ for all other pairs

Table 2

The values of the metrics $d_{W,2}$ and $d_{H,2}$ between the uncertainty intervals $[Bel_i(A), Pl_i(A)]$, $i = 1, 2$

A	F_1				F_2				$d^2_{W,2}$	$d^2_{H,2}$
	Bel_1	Pl_1	$mean_1$	rad_1	Bel_2	Pl_2	$mean_2$	rad_2		
$\{x_1\}$	0.8	1	0.9	0.1	0.7	0.7	0.7	0	13/300	0.09
$\{x_2\}$	0	0.2	0.1	0.1	0.3	0.3	0.3	0	13/300	0.09
X	1	1	1	0	1	1	1	0	0	–

$(A, B) \in 2^X \times 2^X$. On the set 2^X, let $d(A, B) = d_J(F_A, F_B) = \sqrt{1 - \frac{|A \cap B|}{|A \cup B|}}$. Then the Wasserstein metric between the bodies of evidence F_1 and F_2 is given by

$$d_W(F_1, F_2) = \min_{0 \le t \le 0.2} \left\{ (0.3 - t) \cdot 1 + (0.2 - t) \frac{1}{\sqrt{2}} + t \frac{1}{\sqrt{2}} \right\} =$$

$$\min_{0 \le t \le 0.2} \left\{ \frac{3 + \sqrt{2}}{10} - t \right\} = \frac{1 + \sqrt{2}}{10} \approx 0.241. \blacklozenge$$

Metrics between the specialization (generalization) matrices of bodies of evidence. The paper [53] introduced and examined the following metrics on the set $\mathcal{F}(X)$:

$$d^{\cap}(F_1, F_2) = c_{\cap} \left\| R^{\cap}_{F_1} - R^{\cap}_{F_2} \right\|,$$

$$d^{\cup}(F_1, F_2) = c_{\cup} \left\| R^{\cup}_{F_1} - R^{\cup}_{F_2} \right\|,$$

where $R^{\cap}_{F_i}$ and $R^{\cup}_{F_i}$ are the specialization and generalization matrices, respectively, of a body of evidence F_i, $i = 1, 2$ (Section 3), and $\| \ \|$ denotes the matrix norm. Clearly, $c_{\cap} = \left(\max_{A, B \in 2^X} \left\| R^{\cap}_{F_A} - R^{\cap}_{F_B} \right\| \right)^{-1}$ and $c_{\cup} = \left(\max_{A, B \in 2^X} \left\| R^{\cup}_{F_A} - R^{\cup}_{F_B} \right\| \right)^{-1}$. The generalizations of these metrics for the operator representations of α-junctions were considered in the paper [54]. It was demonstrated therein that $d^{\cap} = d^{\cup}$. Metrics based on specialization matrices have higher robustness to small changes in bodies of evidence.

Example 12. For the bodies of evidence $F_1 = 0.8 \times F_{\{x_1\}} + 0.2 F_X$ and $F_2 = 0.7 F_{\{x_1\}} + 0.3 F_{\{x_2\}}$ on the set $X = \{x_1, x_2\}$, the specialization matrices $R^{\cap}_{F_1}$ and $R^{\cap}_{F_2}$ have been found in Example 4. For example, for the matrix norm $\|A\|_1 = \max_j \sum_i |a_{ij}|$, we obtain $c_{\cap} = \frac{1}{2}$ and

$$d^{\cap}_1(F_1, F_2) = \frac{1}{2} \left\| R^{\cap}_{F_1} - R^{\cap}_{F_2} \right\|_1 =$$

$$\frac{1}{2} \left\| \begin{pmatrix} 0 & -0.3 & 0.1 & 0 \\ 0 & 0.3 & 0 & 0.1 \\ 0 & 0 & -0.1 & -0.3 \\ 0 & 0 & 0 & 0.2 \end{pmatrix} \right\|_1 = 0.3. \blacklozenge$$

As noted in several publications, using only one metric or functional to characterize external conflict is not enough. In particular, the canonical measure of conflict will take large values for a pair of identical (weakly non-conflict) bodies of evidence "close" to the equiprobable distribution (Remark 2), for which the metric component will be equal to zero. On the other hand, the conflict between bodies of evidence is often not reduced to calculating the metric component only. Indeed, categorical consonant bodies of evidence F_A and F_B, where $A \subseteq B$, are strongly non-conflict, and $K(F_A, F_B) = 0$. However, the metric (4) with the Jacquard index for these bodies of evidence, given by $d_J(F_A, F_B) = \sqrt{1 - \frac{|A|}{|B|}}$, will take large values if $|A| \ll |B|$.

Therefore, some researchers proposed using a set of conflict measures and metrics on the set of evidence bodies $\mathcal{F}(X)$. For example, the paper [50] considered the pair $\left(K(F_1, F_2), d_{Bet,\infty}(F_1, F_2) \right)$, where $d_{Bet,\infty}(F_1, F_2) = \max_{A \subseteq X} \left| Bet_{F_1}(A) - Bet_{F_2}(A) \right|$. Large values of each component in this pair guarantee a large conflict between the bodies of evidence. In particular, a large value of the metric $d_{Bet,\infty}(F_1, F_2)$ indicates that the bodies of evidence are far from probability distributions.

In the general case, the canonical measure of conflict can be represented as [55]

$$K(F_1, F_2) = E_Q(\mathbf{m}_1) + E_Q(\mathbf{m}_2) + \left\| \mathbf{m}_1 - \mathbf{m}_2 \right\|^2_Q - \sum_{B,C} r_{B,C} m_1(B) m_2(C), \quad (6)$$

where

$E_Q(\mathbf{m}_i) = \frac{1}{2} - \left\| \mathbf{m}_i \right\|^2_Q = \frac{1}{2} \sum_B m_i(B) \left(1 - \sum_C q_{B,C} m_i(C) \right)$, $i = 1, 2$, and $r_{B,C} = \begin{cases} 0, & B \cap C = \varnothing \vee B = C, \\ 1 - q_{B,C}, & \text{otherwise.} \end{cases}$

The last term in formula (6) characterizes the interaction of weakly intersecting focal elements. The functional $E_Q : \mathcal{F}(X) \to [0, \frac{1}{2}]$ is close by properties to the entropy functional and describes the amount of internal conflict in a body of evidence. In particular, for completely conflict bodies of evidence, $K(F_1, F_2) = E_I(\mathbf{m}_1) + E_I(\mathbf{m}_2) + \left\| \mathbf{m}_1 - \mathbf{m}_2 \right\|^2_I$, where I denotes an identity matrix of compatible dimensions. The use of entropies and divergences for estimating the uncertainty of bodies of evidence was considered in [56].

5.2. Structural methods to estimate conflict

The paper [39] presented a structural approach to estimating conflict considering the degree to which the focal elements of one body of evidence are included in the focal elements of another body of evidence. The set inclusion index $Inc(A, B) = \begin{cases} 1, A \subseteq B, \\ 0, A \not\subseteq B, \end{cases} \quad A, B \in 2^X$,

was adopted to calculate the degree of inclusion of one set of focal elements into another set:

$$\delta(\mathcal{A}_1, \mathcal{A}_2) = \max\{d(\mathcal{A}_1, \mathcal{A}_2), d(\mathcal{A}_2, \mathcal{A}_1)\},$$

$$d(\mathcal{A}_1, \mathcal{A}_2) = \frac{1}{|\mathcal{A}_1||\mathcal{A}_2|}\sum_{A_1\in\mathcal{A}_1}\sum_{A_2\in\mathcal{A}_2} Inc(A_1, A_2).$$

The measure of conflict was defined by the formula

$$Con_{\text{ext}}(F_1, F_2) = \big(1-\delta(\mathcal{A}_1, \mathcal{A}_2)\big)d_J(F_1, F_2),$$

where d_J is the distance (4) with the Jaccard index. This measure satisfies conditions E1, E2′, and E3, as well as corollaries a) and b).

Remark 3. The set inclusion index $Inc(A,B)$ can be treated as a value of a belief measure η_A corresponding to a categorical body of evidence F_A on a set B. Then an arbitrary belief function Bel corresponding to a body of evidence $F=(\mathcal{A},m)$ can be represented as $Bel=\sum_{A\in\mathcal{A}} m(A)\eta_A$.

5.3. Algebraic methods to estimate conflict

The paper [57] introduced an algebraic approach to measuring conflict based on set similarity measures: $Q=(q_{A,B})$, where $q_{A,B}=q_{A,B}\in[0,1]$, $q_{A,A}=1$ for $A\in 2^X\setminus\varnothing$, and $q_{A,B}=0$ if $A\cap B=\varnothing$.

A functional $Con_Q:\mathcal{F}(X)\times\mathcal{F}(X)\to[0,1]$ is said to be bilinear if $Con_Q(\alpha F_1+\beta F_2, F)=\alpha Con_Q(F_1, F)+\beta Con_Q(F_2, F)$ for all $\alpha,\beta\in[0,1]$, $\alpha+\beta=1$, and $F, F_1, F_2\in\mathcal{F}(X)$. Let us weaken condition E4 (antimonotonicity with respect to specialization) as follows:

E4′: If $F=(\mathcal{A}_0, m)$, $F'=(\mathcal{A}\cup\{A'\}, m')$, $F''=(\mathcal{A}\cup\{A''\}, m'')$, $m'(A)=m''(A)$ $\forall A\in\mathcal{A}$, $m'(A')=m''(A'')$, and $q_{A',B}\le q_{A'',B}$ $\forall B\in\mathcal{A}_0$, then $Con_Q(F', F)\ge Con_Q(F'', F)$.

Theorem 2 [57]. *The functional Con_Q is a bilinear measure of conflict on the set $\mathcal{F}(X)\times\mathcal{F}(X)$ that satisfies conditions* E1, E3, *and* E4′ *if and only if*

$$Con_Q(F_1, F_2)=\sum_{A\in\mathcal{A}_1, B\in\mathcal{A}_2}\gamma(A, B)m_1(A)m_2(B)=K(F_1, F_2)+\sum_{A\cap B\ne\varnothing}\gamma(A, B)m_1(A)m_2(B), \quad (7)$$

where the coefficients $\gamma(A, B)=Con_Q(F_A, F_B)\in[0, 1]$ have the following properties: $\gamma(A, B)=\gamma(B, A)$, $\gamma(A', B)\ge\gamma(A'', B)$ if $q_{A',B}\le q_{A'',B}$, and $\gamma(A, B)=1$ if $A\cap B=\varnothing$.

Due to (7), the canonical measure of conflict $K(F_1, F_2)$ is least among all bilinear measures of conflict: $Con_Q(F_1, F_2)\ge K(F_1, F_2)$.

Remark 4. The coefficients $\gamma(A, B)=Con_Q(F_A, F_B)=\psi(q_{A,B})$, $A, B\ne\varnothing$, satisfy conditions a)–c) of Theorem 2 if ψ is a nonincreasing function such that $\psi(1)=0$, $\psi(0)=1$, and $q_{A,B}=|A\cap B|/\min\{|A|, |B|\}$. In particular, if $q_{A,B}=\begin{cases}1, A\cap B\ne\varnothing,\\ 0, A\cap B=\varnothing\end{cases}$ is a primitive measure of intersection, then $Con_Q(F_1, F_2)=K(F_1, F_2)$.

5.4. Combining rule-based methods to estimate conflict

For two bodies of evidence $F_1=(\mathcal{A}_1, m_1)$ and $F_2=(\mathcal{A}_2, m_2)$, the canonical measure of conflict $K(F_1, F_2)$ is equal to the mass of an empty set obtained by combining, $F_1\otimes_{ND}F_2=(\mathcal{A}_1\wedge\mathcal{A}_2, m_{ND})$, $\mathcal{A}_1\wedge\mathcal{A}_2=\{A\cap B: A\in\mathcal{A}_1, B\in\mathcal{A}_2\}$, using the unnormalized Dempster rule: $K(F_1,F_2)=m_{DN}(\varnothing)=\sum_{B\cap C=\varnothing}m_1(B)m_2(C)$.

The measure of conflict from Theorem 1 is equal to the infimum of the masses of all empty sets yielded by combining using the generalized conjunctive rule $\otimes_\cap$.

These examples show that the measure of conflict must be agreed with combining rules for bodies of evidence. In particular, this issue was discussed in the paper [55].

A direct generalization of the Dempster combining rule is its weight analog: if $F_1=(\mathcal{A}_1, m_1)$ and $F_2=(\mathcal{A}_2, m_2)$, then $F_Q=(\mathcal{A}, m_Q)=F_1\otimes_Q F_2$, where $\mathcal{A}=\mathcal{A}_1\wedge\mathcal{A}_2$,

$$m_Q(C)=\frac{1}{K_Q}\sum_{A\cap B=C}q_{A,B}m_1(A)m_2(B),$$

and $Q=(q_{A,B})$ is a measure of set similarity: $q_{A,B}=q_{A,B}\in[0,1]$, $q_{A,A}=1$ for $A\in 2^X\setminus\varnothing$, and $q_{A,B}=0$ if $A\cap B=\varnothing$. The normalization coefficient is $K_Q=\sum_C\sum_{A\cap B=C}q_{A,B}m_1(A)m_2(B)=\sum_{A,B}q_{A,B}m_1(A)m_2(B)$. This rule was studied, e.g., in the paper [58].

Then $Con_Q(F_1, F_2)=1-K_Q$ (7) with $\gamma(A, B)=1-q_{A,B}$, $A, B\in\mathcal{A}$, is a measure of conflict agreed with the combining rule $\otimes_Q$.

The measure of conflict (7) considers pairs of non-intersecting and "weakly intersecting" focal elements. (In the latter case, the cardinality of intersection is small compared to that of each set.)

Example 13. Consider the bodies of evidence $F_1 = 0.2F_{\{b\}} + 0.1F_{\{c\}} + 0.3F_{\{a,b\}} + 0.4F_{\{b,c\}}$ and $F_2 = 0.7F_{\{a,c\}} + 0.3F_{\{a,b,c\}}$ on the set $X = \{a,b,c\}$ (Example 3). For $\gamma(A, B) = 1 - jac_{A,B} = 1 - \frac{|A \cap B|}{|A \cup B|}$, we have: $Con_J(F_1, F_2) = 59/120$ and $K(F_1, F_2) = 0.14$. Clearly, the amount of conflict increases significantly due to consideration of weakly intersecting pairs of focal elements. ♦

The conflict measure (7) is also convenient if the focal elements are subsets of the real axis $\mathbb{R}$.

Example 14. Consider two bodies of evidence $F_1 = 0.5F_{[40,50]} + 0.3F_{[45,55]} + 0.2F_X$ and $F_2 = 0.3F_{[30,60]} + 0.7F_{[40,50]}$ on the set $X = [20, 70]$, which predict the company's stock value (Example 8). For $\gamma(A, B) = 1 - jac_{A,B} = 1 - \frac{|A \cap B|}{|A \cup B|}$, where $|A|$ is the Lebesgue measure of a set A on the real axis $\mathbb{R}$, we have $Con_J(F_1, F_2) = 0.163$ and $K(F_1, F_2) = 0$. ♦

The measure of conflict (7) can be applied in the case of fuzzy focal elements.

Example 15. Consider Example 14 with $F_1 = 0.5F_{(35,40,50,55)} + 0.3F_{(40,45,55,60)} + 0.2F_X$ and $F_2 = 0.3F_{(20,30,60,70)} + 0.7F_{(35,40,50,55)}$, and let $\tilde{A} = (a_1, a_2, a_3, a_4)$, $a_1 \le a_2 \le a_3 \le a_4$ be a trapezoidal fuzzy number on the set $X = [20, 70]$ with the membership function $\mu_{\tilde{A}}(x) = \max\left\{0, \min\left\{\frac{x - a_1}{a_2 - a_1}, 1, \frac{a_4 - x}{a_4 - a_3}\right\}\right\}$, $a_1 < a_2 \le a_3 < a_4$. In this example, the kernels of all the fuzzy focal elements coincide with the corresponding crisp focal elements from the previous example: $\mathrm{Ker}\tilde{A} = \mathrm{Ker}(a_1, a_2, a_3, a_4) = [a_2, a_3]$. Since $\gamma(\tilde{A}, \tilde{B}) = 1 - \frac{|\tilde{A} \cap \tilde{B}|}{|\tilde{A} \cup \tilde{B}|} = 1 - \int_X \min\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)\}dx / \int_X \max\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)\}dx$, we obtain $Con_J(F_1, F_2) = 0.365$. ♦

6. ROBUST METHODS TO ESTIMATE EXTERNAL CONFLICT

Some methods to calculate measures of conflict (in particular, the canonical measure of conflict) are unstable to "small" changes in bodies of evidence. At the same time, the bodies of evidence can be formed subjectively and depend on the characteristics of information sources. For example, assume that one expert makes a "cautious" forecast about the company's stock value in the interval (30, 40) c.u., whereas another expert makes an "optimistic" forecast (38, 45) c.u. In this case, we have two categorical bodies of evidence: F_A and F_B, where $A = (30, 40)$ and $B = (38, 43)$; the canonical measure of the conflict is $K(F_A, F_B) = 0$. In fact, when refining his forecast, the "cautious" expert meant the body of evidence $F_1 = 0.8F_{(30,38)} + 0.2F_{[38,40)}$; the "optimistic" expert could also refine his forecast as the body of evidence $F_2 = 0.2F_{(35,38)} + 0.7F_{[38,42)} + 0.1F_{[42,44)}$. As a result, $K(F_1, F_2) = 0.7$.

An approach to increase the robustness of conflict estimation involves specialization-generalization procedures. A specialization procedure divides focal elements into smaller sets, simultaneously distributing the mass function over them. Without loss of generality, assume the following. For a body of evidence $F_1 = (\{A_i\}, m_1)$, a specialization is another body of evidence $F_2 = (\{B_{ij}\}, m_2) = S(F_1)$ such that $\cup_j B_{ij} = A_i$ and $\sum_j m_2(B_{ij}) = m_1(A_i)$ $\forall i$. (This fact will be denoted by $F_2 \sqsubseteq F_1$.) For a body of evidence $F_1 = (\{A_i\}, m_1)$, a generalization is another body of evidence $F_3 = (\{C_{ij}\}, m_3) = G(F_1)$ such that $\cap_j C_{ij} = A_i$ and $\sum_j m_3(C_{ij}) = m_1(A_i)$ $\forall i$. (This fact will be denoted by $F_1 \sqsubseteq F_3$.)

Consider the specializations and generalizations that are sufficiently close to an original body of evidence $F = (\mathcal{A}, m)$. The degree of closeness will be measured by an imprecision index $f : \mathcal{F}(X) \to [0, 1]$ [23]. (One example is the normalized generalized Hartley measure $H_0(F) = \frac{1}{\ln|X|} \sum_{A \in \mathcal{A}} m(A) \ln|A|$.) Any specialization does not increase the imprecision index: $f(F_2) \le f(F_1)$ for $F_2 \sqsubseteq F_1$. Similarly, any generalization does not reduce the imprecision index.

We denote by $S_\varepsilon(F) = \{F' \sqsubseteq F : f(F) - f(F') < \varepsilon\}$ and $G_\varepsilon(F) = \{F \sqsubseteq F' : f(F') - f(F) < \varepsilon\}$ the sets of all specializations and generalizations, respectively, of a body of evidence F that belong to its ε-neighborhood with respect to the imprecision index f. Also, let $SG_\varepsilon(F) = S_\varepsilon(F) \cup G_\varepsilon(F)$. A measure of conflict can be defined as $K_\varepsilon(F_1, F_2) = \underset{\tilde{F}_i \in SG_\varepsilon(F_i), i=1,2}{\mathrm{MEAN}} \left(K(\tilde{F}_1, \tilde{F}_2)\right)$, where MEAN indicates some averaging operator. For example, let $SG_\varepsilon(F_1) = \{F_{1,j}\}$ and $SG_\varepsilon(F_2) = \{F_{2,k}\}$, where $F_{i,s} = F_{i,s}(\boldsymbol{\theta})$, $i = 1, 2$, and $\boldsymbol{\theta} \in [0, 1]^N$ is the vector of parameters (mass function values). Then

$K(F_{1,j}, F_{2,k}) = \varphi_{j,k}(\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in D_{j,k} = \{\boldsymbol{\theta}: \left|f(F_{1,j}(\boldsymbol{\theta})) - f(F_1)\right| < \varepsilon, \ \left|f(F_{2,k}(\boldsymbol{\theta})) - f(F_2)\right| < \varepsilon\}$. The MEAN operator can be averaging of the mean integral values:

$$K_\varepsilon(F_1, F_2) = \frac{1}{\left|\{F_{1,j}\}\right| \cdot \left|\{F_{2,k}\}\right|} \sum_{j,k} I_{j,k},$$

where $I_{j,k} = \frac{1}{V(D_{j,k})} \int_{D_{j,k}} \varphi_{j,k}(\boldsymbol{\theta}) d\boldsymbol{\theta}$, and V denotes the Lebesgue measure on the parameter space. The set $SG_\varepsilon(F)$ of specializations and generalizations for a body of evidence F can be formed by other methods.

Example 16 [4]. Let $X = \{x_1, x_2, x_3\}$, $F_1 = F_{\{x_1, x_2\}}$, and $F_2 = F_{\{x_3\}}$. In this case, $K_0(F_1, F_2) = 1$. Consider specializations and generalizations for the bodies of evidence F_1 and F_2. We estimate the imprecision using the normalized generalized Hartley measure H_0. Choosing $c = \frac{\ln 2}{\ln 3}$ gives $H_0(F_1) = c$ and $H_0(F_2) = 0$. For specializations, we have $S(F_1) = \{F_{1,0}, F_{1,1}\}$, where $F_{1,0} = F_1$, $F_{1,1} = \theta F_{\{x_1\}} + (1-\theta) F_{\{x_2\}}$, $H_0(F_{1,1}) = 0$, $S(F_2) = \{F_{2,0}\}$, and $F_{2,0} = F_2$. The generalizations of the bodies of evidence F_1 and F_2 are:

$$G(F_1) = \{F_{1,0}, F_{1,2}\},$$

$$G(F_2) = \{F_{2,0}, F_{2,1}, ..., F_{2,4}\},$$

where $F_{1,2} = \theta F_{\{x_1, x_2\}} + (1-\theta) F_{\{x_1, x_2, x_3\}}$ and $H_0(F_{1,2}) = c\theta + 1 - \theta$;

$F_{2,1} = \theta F_{\{x_1, x_3\}} + (1-\theta) F_{\{x_3\}}$ and $H_0(F_{2,1}) = c\theta$,

$F_{2,2} = \theta F_{\{x_2, x_3\}} + (1-\theta) F_{\{x_3\}}$ and $H_0(F_{2,2}) = c\theta$,

$F_{2,3} = \theta F_{\{x_1, x_3\}} + (1-\theta) F_{\{x_2, x_3\}}$ and $H_0(F_{2,3}) = c$,

$F_{2,4} = \theta F_{\{x_1, x_2, x_3\}} + (1-\theta) F_{\{x_3\}}$ and $H_0(F_{2,4}) = \theta$.

As a result, $SG(F_1) = \{F_{1,0}, F_{1,1}, F_{1,2}\}$ and $SG(F_2) = \{F_{2,0}, F_{2,1}, ..., F_{2,4}\}$.

Letting $\varepsilon < 1 - c$ yields $SG_\varepsilon(F_1) = \{F_{1,0}, F_{1,2}\}$ and $SG_\varepsilon(F_2) = \{F_{2,0}, F_{2,1}, F_{2,2}, F_{2,4}\}$. Now we have: $K(F_{1,0}, F_{2,0}) = 1$, $K(F_{1,0}, F_{2,1}) = K(F_{1,0}, F_{2,2}) = 1 - \theta$ for $0 < c\theta < \varepsilon$, $K(F_{1,0}, F_{2,4}) = 1 - \theta$ for $0 < \theta < \varepsilon$, $K(F_{1,2}, F_{2,0}) = \theta$ for $0 < (1-\theta)(1-c) < \varepsilon$, $K(F_{1,2}, F_{2,1}) = K(F_{1,2}, F_{2,2}) = \theta_1(1-\theta_2)$ for $0 < (1-\theta_1)(1-c) < \varepsilon$ and $0 < c\theta_2 < \varepsilon$, and $K(F_{1,2}, F_{2,4}) = \theta_1(1-\theta_2)$ for $0 < (1-\theta_1) \times (1-c) < \varepsilon$ and $0 < \theta_2 < \varepsilon$.

Now, $K_\varepsilon(F_1, F_2) = \frac{1}{2 \cdot 4} \sum_{j,k} I_{j,k}$, where $I_{j,k} = \frac{1}{V(D_{j,k})} \times \int_{D_{j,k}} \varphi_{j,k}(\boldsymbol{\theta}) d\boldsymbol{\theta}$, that is, $I_{0,0} = 1$, $I_{0,1} = I_{0,2} = \frac{c}{\varepsilon} \int_0^{\frac{\varepsilon}{c}} (1-\theta) d\theta = 1 - \frac{1}{2c}\varepsilon$, $I_{0,4} = \frac{1}{\varepsilon} \int_0^{\varepsilon} (1-\theta) d\theta = 1 - \frac{1}{2}\varepsilon$, $I_{2,0} = \frac{1-c}{\varepsilon} \int_{1-\frac{\varepsilon}{1-c}}^{1} \theta d\theta = 1 - \frac{1}{2(1-c)}\varepsilon$, $I_{2,1} = I_{2,2} = \frac{1-c}{\varepsilon} \cdot \frac{c}{\varepsilon} \int_{1-\frac{\varepsilon}{1-c}}^{1} \int_0^{\frac{\varepsilon}{c}} \theta_1 (1-\theta_2) d\theta_1 d\theta_2 = \left(1 - \frac{1}{2(1-c)}\varepsilon\right)\left(1 - \frac{1}{2c}\varepsilon\right)$, $I_{2,4} = \frac{1-c}{\varepsilon} \frac{1}{\varepsilon} \int_{1-\frac{\varepsilon}{1-c}}^{1} \int_0^{\varepsilon} \theta_1 (1-\theta_2) d\theta_1 d\theta_2 = \left(1 - \frac{1}{2(1-c)}\varepsilon\right)\left(1 - \frac{1}{2}\varepsilon\right)$.

Thus, $K_\varepsilon(F_1, F_2) = 1 - \varepsilon \frac{(2-c)(c+1)}{8c(1-c)} + \varepsilon^2 \frac{(2+c)}{4c(1-c)}$ for $\varepsilon < 1 - c$. ♦

CONCLUSIONS

This paper has reviewed current research on the inconsistency (conflict) of information from several sources within belief function theory. In particular, the following aspects can be highlighted:

- At present, the scientific community has formed certain requirements to measures of external conflict. We mention the most important ones:

– A measure of conflict should reflect different degrees of conflict in bodies of evidence: from complete non-conflict to a particular degree of conflict (weak, simple, or strong).

– A measure of conflict should be antimonotonic with respect to specialization.

- These requirements underlie the axiomatics of a measure of external conflict. A measure of external conflict satisfying a given system of axioms has been found in a general form.
- There are several methods to estimate external conflict (metric, structural, algebraic, and combining rule-based).
- Together with a measure of conflict, belief function theory considers the concept of a distance between bodies of evidence. However, a measure of conflict is not reduced to calculating a distance.
- A measure of conflict can be calculated robustly using the generalization and specialization procedure.

At the same time, the analysis of information inconsistency within belief function theory has several open problems. Among them, we separate the following:

- constructing measures of external conflict that better reflect various degrees of conflict for bodies of evidence bodies and their structural features;
- investigating the agreement between combining rules and measures of conflict when choosing bodies of evidence for combination;
- studying interrelations between the concepts of the logical chain "inconsistency"–"agreement"–"mutual influence" of information sources;
- finding a general form of a measure of conflict that satisfies other systems of axioms;
- examining measures of conflict for bodies of evidence defined in a metric space.

Applied problems of conflict analysis of information sources, conflict management, etc. are topical as well.

REFERENCES

1. Dempster, A.P., Upper and Lower Probabilities Induced by a Multivalued Mapping, *Annals of Mathematical Statistics*, 1967, vol. 38, pp. 325–339.
2. Shafer, G., *A Mathematical Theory of Evidence*, Princeton: Princeton University Press, 1976.
3. Lefevre, E., Colot, O., and Vannoorenberghe, P., Belief Function Combination and Conflict Management, *Inf. Fusion*, 2002, vol. 3, no. 2, pp. 149–162.
4. Bronevich, A., Lepskiy, A., and Penikas, H., The Application of Conflict Measure to Estimating Incoherence of Analyst's Forecasts about the Cost of Shares of Russian Companies, *Procedia Computer Science*, 2015, vol. 55, pp. 1113–1122.
5. Kutynina, E. and Lepskiy, A., Aggregation of Forecasts and Recommendations of Financial Analysts in the Framework of Evidence Theory, in *Advances in Intelligent Systems and Computing*, Kacprzyk, J., Szmidt, E., Zadrożny, S., Atanassov, K., and Krawczak, M., Eds., vol. 642, Cham: Springer, 2018, pp. 370–381.
6. Bronevich, A.G. and Spiridenkova, N.S., Measuring Uncertainty for Interval Belief Structures and Its Application for Analyzing Weather Forecasts, in *Advances in Intelligent Systems and Computing*, Kacprzyk, J., Szmidt, E., Zadrożny, S., Atanassov, K., and Krawczak, M., Eds., vol. 641, Cham: Springer, 2018, pp. 273–285.
7. Lepskiy, A. and Smolev, V., Application of Non-additive Measures and Integrals for Analysis of the Importance of Party Positions for Voting, in *Atlantis Studies in Uncertainty Modelling: Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)*, Atlantis Press, vol. 1, 2019, pp. 321–327.
8. Lepskiy, A. and Suevalov, A., Application of the Belief Function Theory to the Development of Trading Strategies, *Procedia Computer Science*, 2019, vol. 162, pp. 235–242.
9. Rominger, C. and Martin, A., Using the Conflict: An Application to Sonar Image Registration, *Proceedings of the Workshop on the Theory of Belief Functions*, Brest, France, 2010, pp. 1–6.
10. Harmanec, D., Faithful Approximations of Belief Functions, in *Uncertainty in Artificial Intelligence 15 (UAI99)*, Laskey, K.B. and Prade, H., Eds., Stockholm, Sweden, 1999.
11. Denœux, T., Inner and Outer Approximation of Belief Structures Using a Hierarchical Clustering Approach, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2001, vol. 9, pp. 437–460.
12. Han, D., Dezert, J., and Yang, Y., Belief Interval-Based Distance Measures in the Theory of Belief Functions, *IEEE Transactions on Systems, Man and Cybernetics*, 2016, vol. 48, no. 6, pp. 833–850.
13. Bronevich, A. and Lepskiy, A., Measures of Conflict, Basic Axioms and Their Application to the Clusterization of a Body of Evidence, *Fuzzy Sets and Systems*, 2021. https:doi.org/10.1016/j.fss.2021.04.016
14. Denœux, T. A., Neural Network Classifier Based on Dempster–Shafer Theory, *IEEE Transactions on Systems, Man and Cybernetics*, 2000, vol. 30, pp. 131–150.
15. Ye, Q., Wu, X., and Chen, Z., An Approach for Evidence Clustering Using Generalized Distance, *Journal of Electronics*, 2009, vol. 26, pp. 18–23.
16. Smets, P. and Kennes, R., The Transferable Belief Model, *Artificial Intelligence*, 1994, vol. 66, pp. 191–234.
17. Nguyen, H.T., On Random Sets and Belief Functions, *J. Math. Anal. Appl*., 1978, vol. 65, pp. 531–542.
18. Halpern, J.Y. and Fagin, R., Two Views of Belief: Belief as Generalized Probability and Belief as Evidence, *Artificial Intelligence*, 1992, vol. 54, no. 3, pp. 275–317.
19. Smets, P., Decision Making in TBM: The Necessity of the Pignistic Transformation, *International Journal of Approximate Reasoning*, 2005, vol. 38, pp. 133–147.
20. Shapley, L., A Value for N-person Games, in *Contributions to the Theory of Games. II (28), Annals of Mathematics Studies*, Princeton University Press, 1953, pp. 307–317.
21. Smets, P., Belief Functions on Real Numbers, *International Journal of Approximate Reasoning*, 2005, vol. 40, no. 3, pp. 181–223.
22. Yen, J., Generalizing the Dempster–Shafer Theory to Fuzzy Sets, *IEEE Transactions on Systems, Man, and Cybernetics*, 1990, vol. 20, no. 3, pp. 559–570.
23. Bronevich, A. and Lepskiy, A., Imprecision Indices: Axiomatic, Properties and Applications, *Int. J. of General Systems*, 2015, vol. 44, no. 7-8, pp. 812–832.
24. Higashi, M. and Klir, G.J., Measures of Uncertainty and Information Based on Possibility Distributions, *Int. J. General Systems*, 1983, no. 9, pp. 43–58.
25. Dubois, D. and Prade, H., A Note on Measures of Specificity for Fuzzy Sets, *Int. J. of General Systems*, 1985, no. 10, pp. 279–283.
26. Diaconis, P. and Zabell, S.L., Updating Subjective Probability, *Journal of the American Statistical Society*, 1982, vol. 77, no. 380, pp. 822–830.
27. Hunter, D., Dempster–Shafer vs. Probabilistic Logic, *Proceedings of the Third AAAI Uncertainty in Artificial Intelligence Workshop*, 1987, pp. 22–29.
28. Pearl, J., Reasoning with Belief Functions: An Analysis of Compatibility, *International Journal of Approximate Reasoning*, 1990, vol. 4, no. 5, pp. 363–389.
29. Cattaneo, M., Combining Belief Functions Issued from Dependent Sources, *Proceedings of the Third International Symposium on Imprecise Probabilities and Their Application (ISIPTA'03)*, Lugano, Switzerland, 2003, pp. 133–147.
30. Yager, R.R., On the Dempster–Shafer Framework and New Combination Rules, *Information Sciences*, 1987, vol. 41, pp. 93–138.
31. Sentz, K. and Ferson, S., Combination of Evidence in Dempster-Shafer Theory, *Report SAND 2002-0835*, Sandia National Laboratories, 2002.
32. Dubois, D. and Prade, H., A Set-Theoretic View on Belief Functions: Logical Operations and Approximations by Fuzzy Sets, *Int. J. of General Systems*, 1986, no. 12, pp. 193–226.
33. Dubois, D. and Prade, H., On the Combination of Evidence in Various Mathematical Frameworks, in *Reliability Data Collection and Analysis*. *Eurocourses (Reliability and Risk Analysis),* Flamm, J., Luisi, T., vol. 3, Dordrecht: Springer, 1992, pp. 213–241.
34. Lepskiy, A., General Schemes of Combining Rules and the Quality Characteristics of Combining, in *Belief Functions: Theory and Applications (BELIEF 2014), Lecture Notes on Artificial Intelligence*, Cuzzolin, F., Ed., vol. 8764, Springer-Verlag, 2014, pp. 29–38.
35. Smets, P., The Alpha-Junctions: Combination Operators Applicable to Belief Functions, in *The First Int. Joint Conference on Qualitative and Quantitative Practical Reasoning (ECSQUARU-FAPR'97)*, *Lecture Notes in Computer Sciences*, vol. 1244, 1997, Springer, pp. 131–153.

36.Smets, P., The Application of the Matrix Calculus to Belief Functions, *International Journal of Approximate Reasoning*, 2002, vol. 31, no. 1-2, pp.1–30.
37.Pichon, F. and Denœux, T., Interpretation and Computation of A-junctions for Combining Belief Functions, *Proceedings of the 6th Int. Symposium on Imprecise Probability: Theories and Applications (ISIPTA '09)*, Durham, United Kingdom, 2009.
38.Destercke, S. and Burger, T., Toward an Axiomatic Definition of Conflict between Belief Functions, *IEEE Transactions on Cybernetics*, 2013, vol. 43, no. 2, pp. 585–596.
39.Martin, A., About Conflict in the Theory of Belief Functions, in *Belief Functions: Theory and Applications, Advances in Intelligent and Soft Computing*, vol. 164, 2012, pp. 161–168.
40.Bronevich, A. and Rozenberg, I., The Contradiction Between Belief Functions: Its Description, Measurement, and Correction Based on Generalized Credal Sets, *International Journal of Approximate Reasoning*, 2019, vol. 112, pp. 119–139.
41.Jousselme, A.-L., Grenier, D., and Bossé, E., A New Distance between Two Bodies of Evidence, *Information Fusion*, 2001, no. 2, pp. 91–101.
42.Jousselme, A.-L. and Maupin, P., Distances in Evidence Theory: Comprehensive Survey and Generalizations, *International Journal of Approximate Reasoning*, 2012, vol. 53, pp. 118–145.
43.Deza, M.M. and Deza, E., *Encyclopedia of Distances*, Berlin–Heidelberg: Springer, 2009.
44.Bouchard, M., Jousselme, A.-L., and Doré, P.-E., A Proof for the Positive Definiteness of the Jaccard Index Matrix, *International Journal of Approximate Reasoning*, 2013, vol. 54, pp. 615–626.
45.Attiaoui, D., Doré, P.-E., Martin, A., and Ben Yaghlane, B., A Distance between Continuous Belief Functions, in *The Scalable Uncertainty Management (SUM) Conference*, *Lecture Notes on Artificial Intelligence*, vol. 7520, Berlin–Heidelberg: Springer-Verlag, 2012, pp. 194–205.
46.Diaz, J., Rifqi, M., and Bouchon-Meunier, B., A Similarity Measure between Basic Belief Assignments, *Proceedings of the 9th International Conference Information Fusion*, Firenze, Italy, 2006.
47.Sunberg, Z. and Rogers, J., A Belief Function Distance Metric for Orderable Sets, *Information Fusion*, 2013, vol. 14, pp. 361–373.
48.Cuzzolin, F., Consistent Approximations of Belief Functions, *Proceedings of the 6th International Symposium on Imprecise Probability: Theories and Applications*, Durham, United Kingdom, 2009.
49.Tessem, B., Approximations for Efficient Computation in the Theory of Evidence, *Artificial Intelligence*, 1993, vol. 61, pp. 315–329.
50.Liu, W. Analysing the Degree of Conflict among Belief Functions, *Artificial Intelligence*, 2006, vol. 170, pp. 909–924.
51.Mercier, D., Quost, B., and Denœux, T., Refined Modeling of Sensor Reliability in the Belief Function Framework Using Contextual Discounting, *Information Fusion*, 2008, no. 9, pp. 246–258.
52. Bronevich, A. and Rozenberg, I., The Measurement of Relations on Belief Functions Based on the Kantorovich Problem and the Wasserstein Metric, *International Journal of Approximate Reasoning*, 2021, vol. 131, pp. 108–135.
53.Loudahi, M., Klein, J., Vannobel, J-M., and Colot, O., New Distances between Bodies of Evidence Based on Dempsterian Specialization Matrices and Their Consistency with the Conjunctive Combination Rule, *International Journal of Approximate Reasoning*, 2014, vol. 55, no. 5, pp. 1093–1112.
54.Loudahi, M., Klein, J., Vannobel, J-M., and Colot, O., Evidential Matrix Metrics as Distances between Meta-data Dependent Bodies of Evidence, *IEEE Transactions on Systems, Man, and Cybernetics*, 2016, vol. 46, no. 1, pp. 109–122.
55. Lepskiy, A., On the Conflict Measures Agreed with the Combining Rules, in *Belief Functions: Theory and Applications (BELIEF 2018)*, *Lecture Notes in Computer Science*, Destercke, S., Denoeux, T., Cuzzolin, F., and Martin, A., Eds., vol. 11069, Cham: Springer, 2018, pp. 172–180.
56.Bronevich, A.G. and Rozenberg, I.N., Metrical Approach to Measuring Uncertainty, in *Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2020), Communications in Computer and Information Science*, vol. 1238, Cham: Springer, 2020.
57.Lepskiy, A. About Relation between the Measure of Conflict and Decreasing of Ignorance in Theory of Evidence, *Proceedings of the 8th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-13)*, Amsterdam–Beijing–Paris: Atlantis Press, 2013, pp. 355–362.
58.Zhang, L., Representation, Independence and Combination of Evidence in the Dempster-Shafer Theory, in *Advances in the Dempster-Shafer Theory of Evidence*, Yager, R.R., Kacprzyk, J., and Fedrizzi, M., Eds., New York: John Wiley & Sons, 1994, pp. 51–69.

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

Received July 19, 2021, and revised August 27, 2021.
Accepted August 31, 2021.

Author information

Lepskiy, Aleksandr Evgen'evich. Dr. Sci. (Phys.–Math.), National Research University Higher School of Economics, Moscow, Russia
✉ alex.lepskiy@gmail.com

Cite this article

Lepskiy, A.E. Analysis of Information Inconsistency in Belief Function Theory. Part I: External Conflict. *Control Sciences* **5,** 2–16 (2021). http://doi.org/10.25728/cs.2021.5.1

Original Russian Text © Lepskiy, A.E., 2021, published in Problemy Upravleniya, 2021, no. 5, pp. 3–19.

Translated into English by *Alexander Yu. Mazurov*,
Cand. Sci. (Phys.–Math.),
Trapeznikov Institute of Control Sciences,
Russian Academy of Sciences, Moscow, Russia
✉ alexander.mazurov08@gmail.com

DOI: http://doi.org/10.25728/cs.2021.5.2

THE STRUCTURE OF CREATIVE ACTIVITY

M.V. Belov[1] and D.A. Novikov[2]

[1] Skolkovo Institute of Science and Technology, Moscow, Russia

[2] Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia

[1]✉ mbelov59@mail.ru, [2]✉ novikov@ipu.ru

Abstract. The specifics of creative activity are considered. There are three phases of such activity: discovering a new knowledge domain (subject matter) and accumulating basic knowledge, mastering the knowledge domain, and mass productive use. The life cycle of creative activity is analyzed. As shown by the analysis, creativity is concentrated in the stage of goal-setting only. A qualitative model for mastering knowledge (experience) and a graph-theoretic structural model of a knowledge domain are proposed. New models can be developed, and well-known models can be used to describe and study each phase of creative activity, including those introduced by the authors earlier: in the first phase, optimal distribution models for the researcher's efforts between the tested hypotheses and optimal scheduling models for tested hypotheses; in the second phase, mathematical models of experience; in the third phase, structural and algorithmic models and optimization models.

Keywords: creative activity, experience, creativity, knowledge domain, making and testing hypotheses.

INTRODUCTION

Activity is a dynamic interaction of a human with the reality in which he represents an actor (*subject)* purposefully influencing a *subject matter (object)* [**Ошибка! Источник ссылки не найден.**]. Activity is a form of human actions aimed at cognizing and transforming the surrounding world, humans themselves, and the conditions of their existence.

Elementary activity is understood as an activity whose goals, technologies, and result have no internal structure.[1]

In the monograph [2], an activity that is not elementary was called complex. In other words, *complex activity* (CA) is an activity with a nontrivial internal structure, multiple and (or) changing goals, actor, technology, and the subject matter's role in the goal context.

[1] In the case of elementary activity, there is no need to consider the actor and subject matter together with the activity itself: they play the role of an intuitively clear context. During such activity, only the subject matter evolves in accordance with the technology used by the actor.

The monograph [2] proposed a classification of activities and identified, in particular, regular activity and creative activity.

Regular CA is an activity performed using a known technology to obtain a priori specified results. The structure and technology of regular CA are deterministic.

Creative CA is an activity with a partially defined (incompletely known) technology at its beginning. Therefore, the technology of creative CA is developed when implementing this activity. The unknown technology is due to uncertain demand and (or) a priori uncertain specifications of the activity result.

Historically, there have been two paradigms for the definition and study of *creativity*. Within the first (activity) paradigm, creativity is treated as an activity. The most striking examples are *research activity* [3] and *artistic activity* [4].

- Creativity is a human activity that produces new material and nonmaterial values of social significance [5].
- Creativity is any practical or theoretical human activity in which new results (*knowledge*, decisions,

methods of action, or material products) arise (at least for the actor) [6].

• Creativity is an activity resulting in new material or nonmaterial values [7].

Within the second paradigm (*the psychology of creativity*), creativity is interpreted as "an interaction leading to development" [8, 9]. According to Ya.A. Ponomarev [8] and his followers, the subject must admit the influence of the object (environment) on himself; therefore, unlike activity, this interaction implies a cross-action of the object. The mechanism of this cross-action is associated with different categories: intuition, insight, cognitive unconsciousness, defocusing of attention, action's by-product, and others.

The two approaches mentioned are not contradicting but mutually complementing. Really, regardless of the position adopted, it is necessary to introduce certain assumptions to answer the following question: where does the image of a "creative product" appear in the subject's mind (an artist's intention, a researcher's hypothesis, etc.)? The psychology of creativity investigates, in particular, the reflection mechanisms of the surrounding world in the subject's mind considering the latter's experience. Within the activity approach, followed below, the assumption is the existence of a graph that objectively and adequately describes the knowledge domain's structure; see Fig. 4 in Section 5.

Creativity has become a very popular research topic in management and management psychology since the 1980s–1990s. For example, we refer to the surveys in [10–13]. There has been a significant flow of publications on this range of problems; see a scientometric analysis in the paper [14]. However, the results of creativity studies presented therein are qualitative and, at best, at the level of structural models [15].

The general model of activity, proposed in the book [16], describes the activity of subjects considering their active choice and activity in the environment. This model assesses the results of activity and mastering technologies. Moreover, it describes and examines the dynamics of knowledge, experience, and technologies using a set of admissible *structural elements of activity* (SEAs) as a function of time and previous actions, experience, activity result, subject's state, and values of *uncertainty factors* (UFs). However, the general model relations used directly yield no constructive results due to their analytical complexity.

This paper considers the specifics of creative activity and formulates its general structure in terms of the methodology of complex activity [2]. Within the life cycle model of a knowledge domain (subject matter), we identify the key phases of creative activity, showing that creativity is concentrated in the stage of goal-setting only.

The remainder of this paper is organized as follows. Section 1 introduces basic definitions. In Section 2, the life cycle of a knowledge domain is considered. In Section 3, we localize the creative aspects in the life cycle of CA. Section 4 contains a qualitative model for developing knowledge (experience); Section 5, a structural model of a knowledge domain.

1. DEFINITIONS

Based on the concepts of knowledge, experience, cognition, and skill from the dictionary [17], we will formally define knowledge and experience. Within this paper, the concepts of experience and knowledge (individual or collective) are considered equivalent. *Experience and knowledge* are defined as the result of cognizing the reality, reflected in the consciousness of an individual or a group of individuals and in the material forms available to them (documents, etc.) through beliefs, notions, judgments, inferences, theories, and skills to perform definite activities in definite conditions. *Consciousness* is understood as the process and result of creating a world's model for particular purposes [18]. In this sense, creative activity closely relates to consciousness since both involve the creation of new knowledge.

Let us define a *knowledge element* as an assertion about the properties of the external world that is confirmed to be true at some time instant (or period) by observations when executing an SEA or is verifiable by executing an SEA (a system of SEAs).

For example, a knowledge element can be a subset of the Cartesian product for the admissible ranges of environment's parameters (including the subject matter of technologies): it can determine a set of admissible values of parameters, particularly at different time instants. In other words, a knowledge element can describe the sequence of changes in the states of environment's elements or the relation between their parameters, particularly under the effect of the subjects' activities.

Activity, including the capability for independent goal-setting, the choice of states (actions), and reflection, is the basic characteristic of a human (*active element*, further also called an *agent*).

A knowledge element is said to be known to an *active system* (AS, a system containing agents) if the hypothesis on the corresponding assertion is confirmed to be true when executing one or several SEAs.

For each knowledge element, there are *preconditions*: a set of knowledge elements that must be known for the corresponding hypothesis to be tested. At each time instant, a set of knowledge elements with testable

hypotheses can be determined. For such knowledge elements, the preconditions (the current "knowledge front" in Fig. 4 below) satisfy the current state of experience (knowledge).

Then each knowledge element at any time instant has one of the following states with respect to the AS:

– not available for hypothesis testing (the preconditions are not satisfied);

– available for hypothesis testing, but the hypothesis has not been tested;

– known (the hypothesis has been tested).

Let us formalize the agent's experience (knowledge) accumulated by the current time instant using a set algebra for the set of his currently known knowledge elements.

According to [2], a *technology* is a system of conditions, criteria, forms, methods, and means of consistently achieving a given goal. Following this definition, we will consider a technology consisting of two components: technological knowledge and objects (*means* of activity). In this case, technological knowledge is a subset of experience (knowledge) as a whole, and the object part of the technology (means) can be treated as an environment's component (activity resources).

As shown by historical practice, the evolution of knowledge (experience) of humankind has a "spasmodic" character: short periods (*scientific revolutions* according to T. Kuhn [19]) of forming new paradigms (new areas of knowledge) are replaced by relatively long periods of the so-called *normal development* (mastering and productive use of knowledge). This process naturally "selects" knowledge domains, i.e., subsets of sets of knowledge elements, possibly conceptually close and interconnected with each other. Let us consider their life cycles.

2. THE LIFE CYCLE OF A KNOWLEDGE DOMAIN

Based on the public-historical practice, we identify three phases[2] of the life cycle of a knowledge domain; see Figs. 1 and 2.

Phase I (*discovering a new knowledge domain and accumulating basic knowledge*). In this phase, a set of SEAs is implemented sequentially and (or) in parallel to gain knowledge: to test the hypotheses that make up knowledge (experience) elements. When implementing the SEAs for testing hypotheses, the significant conditions of the assertions are set (selected) by the subject (individual or collective). The corresponding mathematical models were considered in subsection 5.2 of the book [16].

Each such test has a binary a priori unknown (!) result. Therefore, true uncertainty is realized in each SEA [20]; the hypothesis is either rejected or confirmed depending on the result. These SEAs can be implemented in the form of activity over different objects: material (e.g., physical experiments), informational (e.g., mathematical modeling), and imaginable (thought experiments).

Assume that the first phase continues while potentially useful applications of a given knowledge domain are unknown, potentially useful goals are not formulated, and technologies for achieving them are not developed.

Each SEA and the content of this phase are intended to gain knowledge of the environment.

Thus, basic knowledge (experience) is accumulated: knowledge of the UF properties (their values and dynamic laws) and the CA technologies executed under a certain set of UF values are acquired. In the first phase, CA technologies aim to gain new knowledge (constructing a model of the surrounding reality) rather than obtain a useful result.

Phase II (*mastering the knowledge domain*) includes technology development and single productive and experimental use. A sign of the transition between phases I and II is the emerging hypotheses about a potentially useful application of knowledge (the formulation of new useful goals). In this phase, the goals of SEAs are to form technologies for obtaining useful results based on the UF model yielded by phase I. The subject does not choose the UF values: they are realized by the environment's "natural choice." In other words, some significant conditions of the assertions are determined by the UF values and are not set by the subject (unlike phase I). Note that different values of the UFs can be realized: the already known ones (new knowledge elements are not formed) or the ones not encountered before (new knowledge elements are formed). With multiple repetitions, a known technology either confirms operability (the desired productive result of CA) under all or most of the UF values or identifies new UF values. The models presented in [21, 22] adequately describe this process.

In each knowledge domain and their combination, a finite number of "reasonable," "rational," and "optimal" technologies can be created. (For example, the best electric motor, steam engine, or airplane design from the currently available materials.)

[2] The phase boundaries and the beginning of the life cycle as a whole are conditional. In most cases, we can hardly indicate a single event (the time when it occurs) corresponding to the beginning of a particular phase or the life cycle. However, this is not required to build formal models.

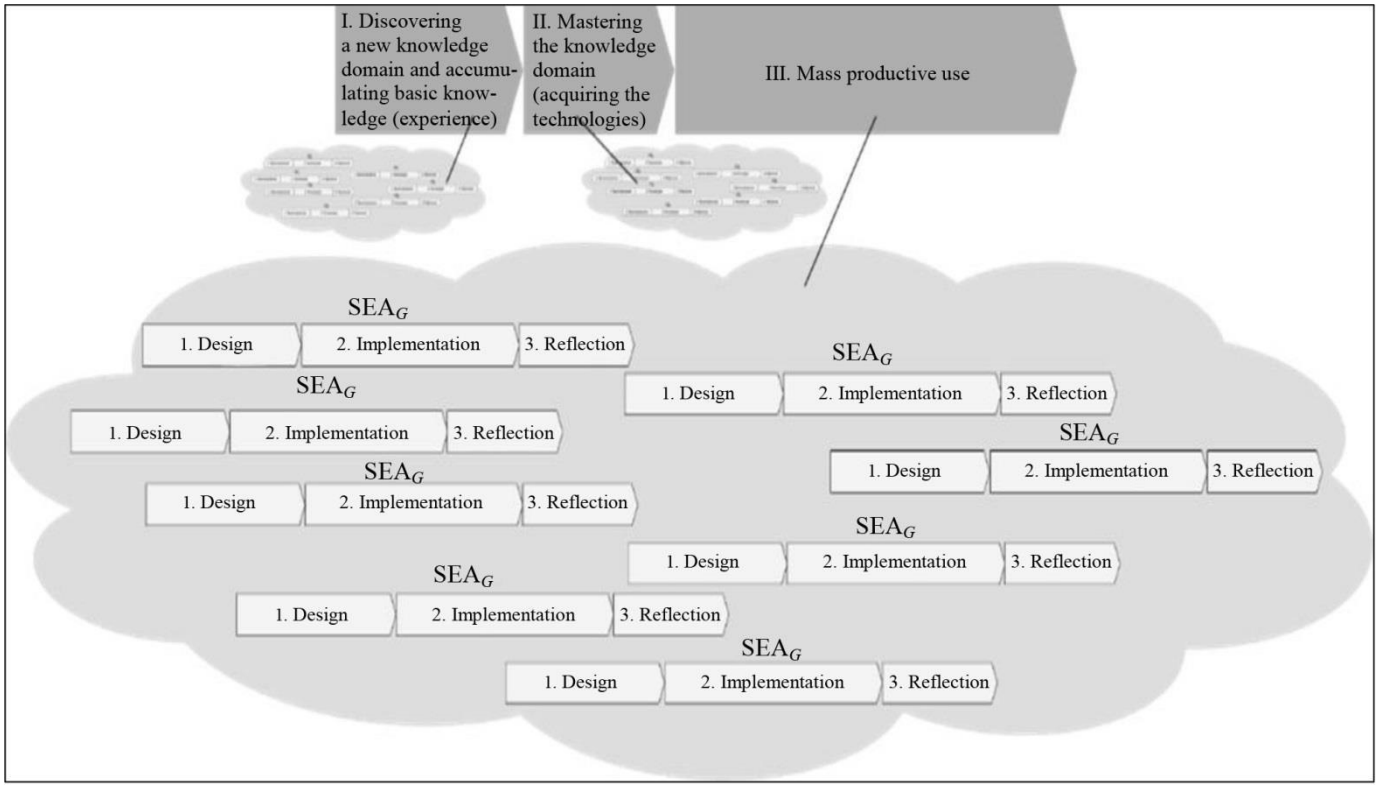


Fig. 1. The life cycle model of a knowledge domain.

Phase III (*mass productive use*). A sign of the transition between phases II and III is the massive (not single) use of known technologies. A set of SEAs with an already known technology (phase II) is being implemented to obtain a productive result (the ultimate goal of this technology). Note that different values of the UFs can be realized: the already known ones (the desired productive result is obtained) or the ones not encountered before (a new experience is formed, and the productive result can be lost). In the case of a new UF value, a return to phase I or II within the same knowledge domain occurs, or a new knowledge domain appears.

New SEAs formed in phases I and II correspond to the hypotheses about the knowledge elements; new SEAs formed in phase III, to the needs for the productive CA to obtain a useful result.

3. CREATIVE ASPECTS IN THE LIFE CYCLE OF COMPLEX ACTIVITY

Which elements (procedural, internal, or external) are "responsible" for creative activity? In other words, where is creativity concentrated? To answer these questions, we will analyze *the CA's life cycle (LC)*; see Table 10 of the monograph [2] and Table 1 below.

Stage II of the CA's life cycle (Table 1) is *goal-setting* (forming the structure of goals) and checking whether technologies exist for all subgoals. If not, the goal is the development of an appropriate technology. Note that sometimes goals—an anticipated image of a future result—can be formulated by the subject unconsciously, leading to "unexpected" results. (This is often the case for creativity; from the subject's viewpoint, goal-setting is "absent.")

According to Table 1, ***creativity is concentrated in the goal-setting stage only***! Really, technology formation and implementation (stages III, IV, and the subsequent ones in Table 1) are always (!) performed by known means and methods: we cannot assure the result without reliable means and methods (technology). In this stage, the subject sets a goal not achieved before (moreover, it is unknown whether the goal can be achieved). The corresponding CA has no technology, and the subject decomposes the goal into subgoals (this is a heuristic, creative activity) and checks the presence of a known technology for each subgoal. In the absence of an appropriate technology, he decomposes the subgoal further. This procedure continues until known technologies are found for all subgoals of the goal. Implementing a set of known technologies to achieve the obtained structure of (sub)goals is a regular CA. The subsequent phase (reflection, checking the compliance between the result and the original goal) is a regular CA as well.

Table 1

Phases, stages and steps of the life cycle of a complex activity element (A-SEA) and their content

Phase	Stage	no.	Step	Content
Design	I. Fixing demand and understanding needs	1	Fixing demand and understanding needs	A superior U-SEA or environment forms the demand for the results of CA element. The subject (actor) fixes the demand, understands the needs and decides to perform activity.
	II. Setting goals, structuring goals and tasks	2	Creating logical model	The need is structured and checked whether it is known or not (in the former case, the activity is regular). If CA is regular, this step comes to extracting information about the logical model from an information store. Otherwise, the structure of goals is formed. The goals are formulated in terms of the expected characteristics of the results of CA elements; see Section 6.1 for a detailed discussion of the result of CA. Consistency is checked/the structure of goals is modified. With each goal of A-SEA the role of the subject and technology is associated (this has been done for the result earlier); in other words, the characteristics of the subjects and technologies are specified. The result of this step is a logical model, i.e., the structure of A-SEA in the form of a set of subordinate SEAs (L-SEAs) and elementary operations (L-Op).
	III. Selecting and developing technology	3	Checking the readiness of technology and the sufficiency of resources	The presence of already known components of the A-SEA's technology is checked: the causal model of A-SEA, the technologies of all L-SEAs and the technologies of all L-Ops. The logical consistency of A-SEA and resource pools is checked: the availability and sufficiency of resources for assigning the subjects of U-SEAs and supporting the technologies of L-Ops, taking into account the use of these resources in parallel when implementing other SEAs. The result of this step is confirmation of the readiness of the technology, confirmation of the availability of necessary resources and transition to step 7, or the implementation of steps 4, 5 or 6, respectively.
		4	Creating cause-effect model	The causal relationships between the goals/results of subordinate elements (L-SEAs and L-Op) are determined and described. Possible events of uncertainty and the response rules for them are described (SEAs to be performed, or escalation to a higher level). The result of this step is the cause-effect model of A-SEA.
		5	Creating technology of lower-level elements	For an elementary operation, due to its specificity and absence of internal structure, the process of designing and describing technology elements is specific and therefore has no general description. For all subordinate L-SEAs without ready-made technologies, steps 1–6 of their life cycles are implemented recursively. The result of this step is the technologies of subordinate elementary operations (L-Ops) and the technologies of subordinate L-SEAs.
		6	Forming/modernizing resources	In the absence of necessary resources, goals responsible for their generation are set; SEAs ensuring the creation or modernization of resource pools are implemented. The result of this step is resource pools required.

Table 1 (continued)

Phase	Stage	no.	Step	Content
Design	III. Selecting and developing technology	7	Calendar-network scheduling and resource planning	A calendar-network schedule is being formed. The consistency of key deadlines of needs is checked. The temporal consistency of the calendar-network schedule and resource pool is checked, taking into account the use of resources by other elements of CA. In case of inconsistency, a return to steps 2-4 is carried out or the impossibility to meet the deadlines is escalated to the subject of a upper SEA. The result of this step is a calendar-network schedule for the use of resources.
		8	Performing optimization	The dynamics of resources use is optimized, taking into account the possibility of using these resources for other CAs implemented in parallel. The result of this step is an optimal calendar-network schedule for the use of resources.
		9	Assigning actors and defining responsibilities	The responsibility matrix is fixed, which describes a correspondence between the subjects of SEAs and personnel. In fact, the assignment of subjects means the formation of demand for the results of lower SEAs and, hence, the recursive implementation of the life cycle of L-SEAs: all steps of the Design phase are carried out. The result of this step is the responsibility matrix, which together with the structure of A-SEA determines its organizational structure.
		10	Allocating resources	In accordance with the technologies of elementary operations, the resources required for the implementation of technologies are request and allocated. The result of this step is the resource allocation matrix of elementary operations.
Implementation	IV. Performing actions and obtaining results	11	Performing actions and obtaining results	In accordance with the causal model, the preconditions for the start of actions of elementary operations (L-Ops) and L-SEAs are repeatedly and constantly checked and they are launched. The elementary operations (L-Ops) are performed. The execution of subordinate L-SEAs is started. The result of this stage is the execution of actions by A-SEA and also the result of its activity.
Reflection	V. Assessing results and reflecting	12	Assessing results and reflecting	Comparison of the characteristics of the result with the required ones. Comparison of the volumes of resources with the given ones. Design of the requirements to the corrections of goals, technology, etc.

The general scheme of the life cycles of CA and a knowledge domain describes well research CA, practical CA (including engineering CA), and artistic CA.

Research CA. Nowadays, well-established paradigms and research approaches (methods of study and presentation of results) have already been developed in many branches of knowledge. Alternative approaches are perceived with a priori suspicion: researchers tend to follow the well-known (regular) technology involving SEAs for hypotheses testing, which are also regular! The exception is the periods of scientific revolutions. Some examples are provided in Table 2.

Table 2

The LC of knowledge domains: some examples of phases

LC phase	UF values	Electricity examples	Atomic energy examples
Phase I. Discovering a new knowledge domain and accumulating basic experience	Are generated or chosen by the subject and are realized in parallel	– the experiments of the ancients with amber and wool, observation of electric eels, – Gilbert's work and apparatus, – Franklin's experiments with a kite, – Galvani's and Volt's experiments and devices, – Faraday's generator, – others	– ideas and hypotheses of the ancients about the structure of matter, – Dalton's theory, Mendeleev's discovery, – X-rays, – the Bohr and Rutherford atom models, – Becquerel's observation, the research works of Curie, Flerov, and Petrzhak, – others
Phase II. Mastering the knowledge domain. The goal is to develop and master technologies	Are realized, whereas the subject fixes the new UF values	– the use of electricity for lighting, industrial drives, transport, etc., – industrial devices designed by Tesla, Edison, Dolivo–Dobrovolsky, etc.	– the Manhattan Project (the first reactor and bomb, enrichment technologies), – nuclear weapons, – the first nuclear power plants
Phase III. Mass productive use. The goal is to use the developed technologies for obtaining new results	Are realized	Batch production and mass use	

Artistic CA [4] is "arranged" in a similar way. Some examples of projects implemented by many people are:

– filming,

– setting up a theatrical performance,

– creating a monument.

Forming an idea (plot, or meaning) as *a holistic image* of a book or picture (or their elements if the subject divides the object of creativity into elements) is an elementary but creative activity. After that, the subject expresses the idea until liking the result, i.e., hypothesis testing (or rethinking of the idea) takes place as well. At the same time, the technology of applying paint, processing marble, or typing text (formulas) is regular. Such technology can also be the object of the first phase; see item C below. However, once developed, it becomes regular.

Thus, for artistic activity, we have the following:

– The idea of a work of art is a hypothesis (hypotheses) formed.

– Attempts to express the idea and checking whether the result fits the desires are tests of the hypotheses.

– Artistic technique (gouache, oil, clay, bronze, or brushstroke), which reflects the idea up to the author's individuality, is regular and "imported" from the industry.

Stages I and II (and stage V) of CA (see Table 1) are always implemented for (and in terms of) an information model of the subject matter. The other stages, III and IV, may require a CA associated with a physical object.

The goals or hypotheses of CA (stage II in Table 1) can describe any subsets of elements (their interconnections) in the body of knowledge, regardless of the representation model. In particular, hypotheses can describe new technology components.

The generalized scheme of a single creative SEA and the life cycle of a knowledge domain coincide with *the scheme of research activity*. It includes the following:

A. understanding of existing knowledge domains;

B. forming goals (in the case of creative CA, this is the hypothesis about achievable goals since at the time of goal-setting, the technology is unknown, and the possibility of achieving the goal is also unknown; in the case of research CA, the goal is to acquire new knowledge, i.e., directly test hypotheses about the laws of the researcher's environment;

C. testing the hypotheses;

D. generalizing and forming new laws (technologies);

E. passing to item A.

Items A and B correspond to goal-setting (forming the structure of goals) and checking whether technologies exist for all subgoals. We emphasize again: creativity is concentrated right here.

Hypothesis testing (item C) is always (!) performed

using a known technology: we cannot assure the result without a reliable technology (means, methods, and techniques to obtain the result). If the testing technology is unknown, the hypothesis is decomposed, and the causal structure of the lower-level hypothesis testing is determined, followed by the aggregation of the intermediate results. Decomposition, formation of a causal structure, and aggregation are well-known operations: proven components of the system-wide technology for achieving complex goals (testing complex hypotheses). Decomposition is performed until a known technology is found for all goals (hypotheses). It yields a fractal set of SEAs. Next, the actions of the SEAs are executed according to the causal structures. After that, the original hypothesis is either confirmed or rejected.

Hypotheses (item B) can describe any subsets of elements (and their interconnections) in the body of knowledge, regardless of the representation model. As noted above, hypotheses can describe new technology components.

Goals are always formed to satisfy the needs of some interested parties and (or) solve their problems (equivalent). In a particular case, such an interested party is the subject itself. (In research and artistic activities, the researcher or artist himself.)

Well, the execution of complex activities is always regular: all stages after goal-setting and structuring of goals (see Table 1) are implemented using a technology known at the beginning of the action. When the goal is structured, and the structure of the SEAs is created with a verified and known technology leading to the required result, the implementation of CA becomes regular.

4. A QUALITATIVE MODEL OF KNOWLEDGE (EXPERIENCE) EVOLUTION

In any AS, agents exist and operate in the following way:

a) An active system as a complex entity "permanently" implements a set of regular SEAs.

b) Events of true uncertainty occur.

c) The agents of the AS perform reflection, comprehending the factual occurrence of these events.

d) The structure of goals is (re)formed, yielding the structure of SEAs and the structure of complex subjects.

e) For a new structure of goals, a new technology is developed or reduced by decomposition to known ones.

f) The implementation of a regular, albeit different, CA continues, and a return to item a) takes place.

The sequence a)–f) is implemented for all life cycle stages of knowledge domains; see Fig. 2.

The events of true uncertainty (b) and the events of re-forming the structures of goals (d) occur asynchronously. They are "connected" through the process of *reflection*, which has uncertain duration and result. Reforming the goals is a manifestation of the subject's true uncertainty.

Thus, the subject's uncertainty has two forms:

– deciding to carry out the activity (or refuse),

– forming the structure of goals.

Generally speaking, the life cycle of knowledge domains includes all three phases. In some cases, however, the development of a knowledge domain cannot lead (yet) to its productive use, and the life cycle is interrupted at the first or second phase. (The corresponding graphic images are shown at the top of Fig. 2.)

Reflection is an assessment of the existing experience and the environment, including the events of true uncertainty. On the one hand, reflection precedes goal-setting and is its source: this is how hypotheses are generated. On the other hand, reflection fixes the experience: this is how the hypotheses are confirmed or rejected.

The "general model" (see the Introduction and the book [16]) describes the evolution of knowledge (experience) using the dynamics of the sets of admissible SEAs and their dependence on the history. However, as noted above, the "general model" relations used directly yield no constructive results.

Therefore, let us concretize the "general model of ASs" [16] to investigate the development of knowledge (experience) analytically as the process of discovering new knowledge domains and accumulating "basic knowledge." For this purpose, we will:

i. abstract from the multiplicity of agents;

ii. discard the set functions describing the SEAs in favor of another representation of the evolution process.

Consider the implementation features of the life cycle of experience (knowledge) with goals i and ii; see Table 3. For each agent, the set of admissible actions consists of SEAs attributed to one of the characteristic subsets for different phases of the life cycle of knowledge domain:

- Phase I, SEAs for testing hypotheses available at the current level of experience;
- Phase II, SEAs for acquiring and mastering technologies;
- Phase III, SEAs for productive use of mastered technologies.

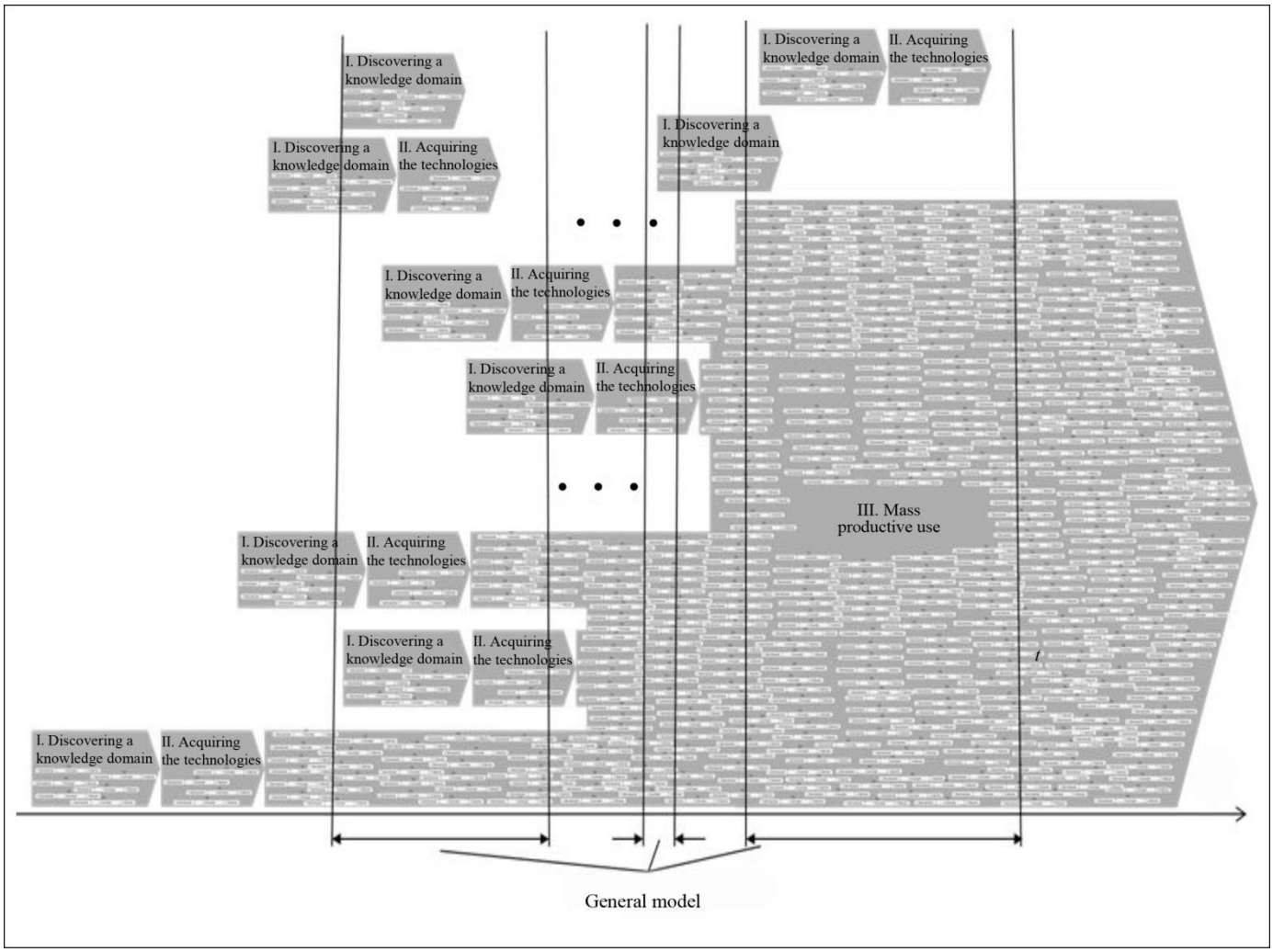


Fig. 2. The general model of knowledge (experience) evolution in an active system.

When implementing the SEAs of subset I, the agent chooses parameter values for the technology and environment, testing the hypothesis under precisely this combination of the values. True uncertainty manifests itself through the CA result, which is a priori unknown to the agent. If the environment's true uncertainty manifests itself so that the UFs take values differing from the required ones, then the CA result characterizes the test of another hypothesis not coinciding with the original one. Upon completion of the hypothesis testing, the sets of UF values and available technologies can be transformed; see Table 3 and Figs. 3 and 4.

When implementing the SEAs of subsets II and III, the agents choose the number of the CA element being executed (the technology parameters). Note that the environment's parameters (the number of the UF state) are implemented independently of the agent and are a priori uncertain for him. In this case, the CA result depends on the parameters values of the technology and the environment.

5. A STRUCTURAL MODEL OF A KNOWLEDGE DOMAIN

Consider a connected circuit-free digraph, i.e., a *network* $G = (N, E)$ with proper numbering. (No edges connect a greater-number vertex to a smaller-number one.) The network vertices correspond to knowledge domains (the sets of hypotheses and assertions), and the edge set $E \subseteq N \times N$ reflects the logical interconnections of vertices; see Fig. 4.

We denote by $N_i = \{j \in N \mid (j; i) \in E\}$ the set of immediate *predecessors* of vertex i in the network G, $i \in N$. Let the network G have a set $N_0 \subseteq N$ of *inputs* (vertices without predecessors, which reflect axioms and (or) facts of recognized common knowledge).

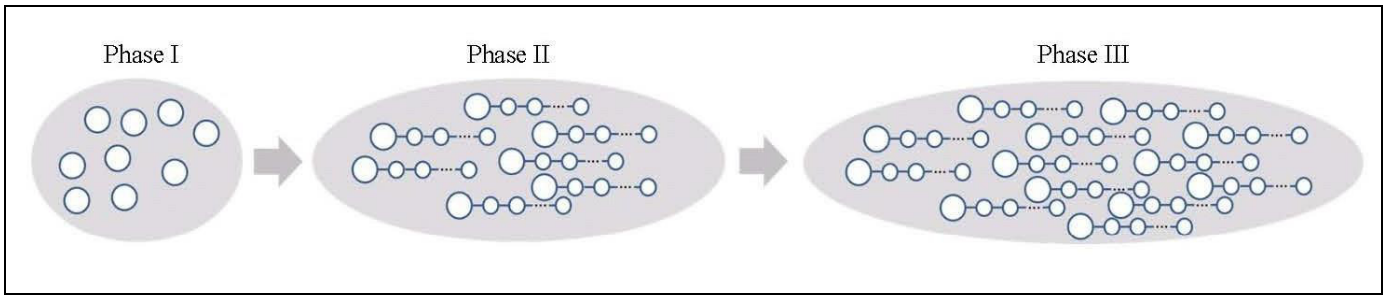


Fig. 3. The life cycle of a knowledge domain.

Table 3

The phases of creative CA

Agent's choice	UF value	Outcome	Consequence
Hypothesis testing, phase I	Coincides with the required one	The hypothesis is tested	The set of admissible technologies is transformed. Possibly, the set of UF values is transformed
	Differs from the required one	Another hypothesis is tested	
Mastering the technology, phase II	Known	The hypothesis is tested	The level of mastering remains the same
	Unknown		The level of mastering increases
Using the technology, phase III	Known	The productive SEA is executed	The expected useful result is obtained
	Unknown		The level of mastering increases, but the expected useful result may be not obtained

Assume that each vertex of the network G has a *precondition*, i.e., a Boolean predicate $\pi_i(\cdot)$ defined on the set of $|N_i| + 1$ *inputs* of two types: the initial facts $z_{N_i} = \{z_j, j \in N_i\}$ and *external conditions* $\omega_i(\cdot) \in \Omega_i$. This predicate calculates a binary *output* (new fact), i.e., a logical variable $z_i = \pi_i(z_{N_i}, \Omega_i)$, which is determinate if the output has the same value under any admissible external conditions, and is uncertain otherwise.

Consequently, a hypothesis with known initial facts is tested by finding the output value for some value(s) of the external conditions (phase I of creative CA). In phase II of creative CA, the invariability of the output value is checked for different (all admissible) values of the external conditions (UFs).

Thus, vertex i of the graph G is given by the tuple $(N_i, \Omega_i, \pi_i(\cdot))$, which includes the initial facts, external conditions, and a logical predicate.

We denote by G_t a subgraph of the graph G that is reliably known to the researchers at a time instant t. (No matter how many subjects in parallel test hypotheses, exchanging their results.) For example, the graph G_t in Fig. 4 is shaded.

Within the structural model of a knowledge domain, a *hypothesis* assumes that some assertion or a combination of some assertions is true. A hypothesis is a subgraph or vertex in which the incoming arcs of all vertices are either contained in it or originate from the graph G_t.

Confirming or *rejecting a hypothesis* is testing the definiteness (truth) of a corresponding assertion under all external conditions figuring in it. In a special case, the predicate is known, and it is necessary to find the maximum set of external conditions under which its value is definite.

The hypothesis testing model was considered in subsection 5.2 of the book [16]. In particular, the following problems were posed and solved therein: the optimal distribution of the researcher's efforts between the testing of various hypotheses and the optimal scheduling of the hypotheses.

As noted, in phase I of the knowledge domain's life cycle, each knowledge element (hypothesis described by a vertex in Fig. 4) has one of the following states:

– not available for hypothesis testing (the preconditions are not satisfied; see the dotted line);

– available for hypothesis testing, but the hypothesis has not been tested (see the thin line);

– known (the hypothesis has been tested; see the thick line).

The set of known knowledge elements (the vertices indicated in Fig. 4 by thick lines) is *the current amount of knowledge*. The set of hypotheses available for testing (the vertices indicated in Fig. 4 by thin lines) is *the*

current horizon of cognition. The current amount of knowledge and horizon of cognition form the subgraph G_t of the graph G known to the researcher at the current time instant. Each vertex indicated by a thin line represents one hypothesis or a set of independently tested hypotheses (sequentially or in parallel).

There are two types of agent's actions at each time instant: *algorithmic* and *creative*. The former actions consist in fully automatic generation and conceptual analysis of all logically possible consequences from the existing body of knowledge G_t. The result is a graph $\hat{G}_t$, conditionally called a "logical closure" of the graph G_t. The latter actions are the advancement and confirmation or rejection of hypotheses, i.e., new subgraphs G_t^h of the graph $G \setminus \hat{G}_t$. We denote by G_t^{h+} the set of confirmed hypotheses. Then $G_{t+1} = \hat{G}_t \cup G_t^{h+}$. The advancement (generation) of hypotheses is an essentially creative and non-formalized stage. Therefore, in modeling, it is advisable to describe the occurrence of hypotheses and the duration of their testing in stochastic terms.

Thus, new hypotheses can be made automatically (algorithmically) or creatively in phase I and algorithmically or creatively in response to the events of true uncertainty in phases II and III of creative CA (see arrows 6 and 7, respectively, in Fig. 4).

The hypotheses testing process (confirming or rejecting hypotheses) can be formalized using the models below.

- In phase I:

– The transition to a new horizon of cognition occurs during scientific revolutions [19]. How and why does this happen? We will not attempt to answer, simply supposing that the graph G is given. This as sumption is the essential one for the models of creative CA under consideration.

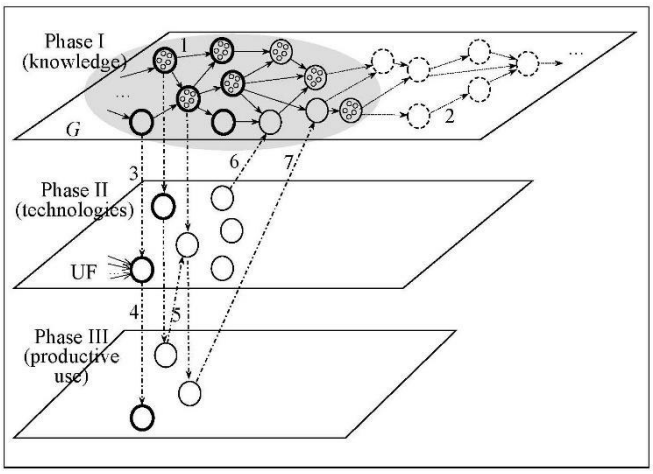


Fig. 4. The phases of creative activity (the LC of a knowledge domain).

– Hypothesis testing is described by the model given in subsection 5.2 of the book [16].

– The arrows of types 1 and 2 correspond to specifying or increasing (decreasing) the dimension (analysis and synthesis, decomposition, and generalization).

- In phase II, technology development is described by "experience models" [21, 22].
- In phase III, practical activity is described by the general schemes of CA given in the monograph [2].

The chain of arrows 3 and 4 describes the "fundamental research => technology development => production" life cycle.

Arrows 5–7 show that the problems arising in phase II or III (situations of true uncertainty) may require a return to the previous phase(s) with the advancement and confirmation of new hypotheses and (or) the development of appropriate technologies.

CONCLUSIONS

This paper has identified three phases of creative activity (the life cycle of a knowledge domain):

– phase I: discovering a new knowledge domain and accumulating basic knowledge (generating and testing hypotheses);

– phase II: developing mastering the knowledge domain;

– phase III: mass productive use.

As shown, creativity is concentrated in the stage of goal-setting only (in the case of research or artistic activity, in the generation of hypotheses). New models can be developed, and well-known models can be used to describe and study each phase of creative activity, including those introduced by the authors earlier:

– in the first phase, optimal distribution models for the researcher's efforts between the tested hypotheses and choice models for an optimal sequence of tested hypotheses [16];

– in the second phase, mathematical models of experience [21, 22];

– in the third phase, structural and algorithmic models [2] and optimization models [23].

REFERENCES

1. Novikov, A.M. and Novikov, D.A., *Metodologiya* (Methodology), Moscow: Sinteg, 2007. (In Russian.)
2. Belov, M.V. and Novikov, D.A., *Methodology of Complex Activity: Foundations of Understanding and Modelling*, Cham: Springer, 2020.
3. Novikov, A.M. and Novikov, D.A., *Research Methodology: From Philosophy of Science to Research Design*, CRC Press, 2019.
4. Novikov, A.M., *Metodologiya khudozhestvennoi deyatel'nosti* (Methodology of Artistic Activity), Moscow: Egves, 2008. (In Russian.)

5. Rubinshtein, S.L., *Osnovy obshchei psikhologii* (Foundations of General Psychology), St. Petersburg: Piter, 2000. (In Russian.)
6. *Psikhologicheskii slovar'* (Psychological Dictionary), Zinchenko, V.P. and Meshcheryakov, B.G., Eds., 2nd ed., Moscow: Pedagogika-Press, 1999. (In Russian.)
7. *Kratkii psikhologicheskii slovar'* (A Brief Psychological Dictionary), Karpenko, L.A., Petrovskii, A.V., and Yaroshevskii, M.G., Eds., Moscow: Feniks, 1998. (In Russian.)
8. Ponomarev, Ya.A., *Psikhologiya tvorchestva* (Psychology of Creativity), Moscow: Institute of Psychology USSR AS, 1976. (In Russian.)
9. *Psikhologiya tvorchestva: shkola Ya.A. Ponomareva* (Psychology of Creativity: Ya.A. Ponomarev's School), Ushakov, D.V., Ed., Moscow: Institute of Psychology RAS, 2006. (In Russian.)
10. Amabile, T. and Pratt, M., The Dynamic Componential Model of Creativity and Innovation in Organizations, *Research in Organizational Behavior*, 2016, vol. 36, pp. 157–183.
11. *Handbook of Organizational Creativity*, Mumford, M., Ed., New York: Academic Press, 2011.
12. *Handbook of the Management of Creativity and Innovation*, Tangand, M. and Werner, C., Eds., Singapore: World Scientific, 2017.
13. *The Cambridge Handbook of Creativity*, Kaufman, J. and Sternberg, R., Eds., Cambridge: Cambridge University Press, 2010.
14. Cai, W., Khapovs, S.N., Bossnik, B., and Lysova, E.I., Optimizing Employee Creativity in the Digital Era: Uncovering the Interactional Effects of Abilities, Motivations, and Opportunities, *Int. J. Environ. Res. Public Health*, 2020, vol. 17, pp. 1038–1057.
15. Amabile, T., Componential Theory of Creativity, *Working paper no. 12-096*, Harvard: Harvard Business School, 2012.
16. Belov, M.V. and Novikov, D.A., *Modeli deyatel'nosti* (Models of Activity), Moscow: Lenand, 2021. (In Russian.)
17. Platonov, K.K., *Kratkii slovar' sistemy psikhologicheskikh ponyatii* (A Brief Dictionary of Psychological Concepts), Moscow: Vysshaya Shkola, 1984. (In Russian.)
18. Michio, K., *The Future of the Mind: The Scientific Quest to Understand, Enhance, and Empower the Mind*, New York: Doubleday, 2014.
19. Kuhn, T.S., *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press, 1962.
20. Knight, F., Risk, Uncertainty and Profit, in *Hart Schaffner and Marx Prize Essays*, no. 31, Boston and New York: Houghton Mifflin, 1921.
21. Belov, M.V. and Novikov, D.A., Models of Experience, *Control Sciences*, 2021, no. 1, pp. 37–52.
22. Belov, M.V. and Novikov, D.A., *Models of Technologies*, Cham: Springer Nature, 2020.
23. Belov, M.V. and Novikov, D.A., *Upravlenie zhiznennymi tsiklami organizatsionno-tekhnicheskikh system* (Management of Life Cycles of Organizational and Technical Systems), Moscow: Lenand, 2020. (In Russian.)

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

Received May 16, 2021, and revised July 31, 2021.
Accepted August 2, 2021.

Author information

Belov, Mikhail Valentinovich. Dr. Sci. (Eng.), Skolkovo Institute of Science and Technology, Moscow, Russia
✉ mbelov59@mail.ru

Novikov, Dmitry Aleksandrovich. Corresponding Member, Russian Academy of Sciences; Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ novikov@ipu.ru

Cite this article
Belov, M.V., Novikov, D.A. The Structure of Creative Activity. *Control Sciences* **5,** 17–28 (2021).
http://doi.org/10.25728/cs.2021.5.2

Original Russian Text © Belov, M.V., Novikov, D.A., 2021, published in Problemy Upravleniya, 2021, no. 5, pp. 20–33.

Translated into English by *Alexander Yu. Mazurov*,
Cand. Sci. (Phys.–Math.),
Trapeznikov Institute of Control Sciences,
Russian Academy of Sciences, Moscow, Russia
✉ alexander.mazurov08@gmail.com

DOI: http://doi.org/10.25728/cs.2021.5.3

ADAPTIVE NEURAL-NETWORK-BASED CONTROL OF NONLINEAR UNDERACTUATED PLANTS: AN EXAMPLE OF A TWO-WHEELED BALANCING ROBOT[1]

A.I. Glushchenko[1], V.A. Petrov[2], and K.A. Lastochkin[1]

[1]Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
[2]Stary Oskol Technological Institute, National University of Science and Technology MISIS, Stary Oskol, Russia

✉ strondutt@mail.ru, ✉ petrov.va@misis.ru, ✉ lastconst@yandex.ru

Abstract. This paper proposes a new method to control nonlinear underactuated plants for eliminating unmatched parametric uncertainties. The method is based on a model reference adaptive control. The controller consists of a basic LQ one and an adaptive compensator reducing the uncertainty norm under certain assumptions. The compensator involves a multilayer neural network due to its universal approximation properties. The network is trained online. The equations to tune the compensator's neural network parameters are derived using Lyapunov's second method and the backpropagation algorithm. The asymptotic convergence of the tracking error (the difference between the plant's and reference model's outputs) to a given domain is proved. The theoretical results are validated by numerical experiments with the developed control system for the mathematical model of a balancing LEGO EV3 robot in MATLAB.

Keywords: model reference adaptive control, balancing robot, suppression of unmatched parametric uncertainties, neural networks, online training, stability.

INTRODUCTION

In modern automatic control practice, the problems of controlling technical systems (plants) with a significant effect of parametric uncertainties are becoming increasingly important. Classical examples include manipulators [1, 2], unmanned and manned aerial vehicles in special operating modes [3, 4], industrial electric drives [5], and technological processes in chemical [6] and metallurgical [7, 8] industries.

Most modern methods for constructing control systems for plants with parametric uncertainties can be divided into robust approaches and model reference adaptive control methods. Robust systems are designed so that the performance criteria of the closed-loop control system (usually the phase and gain margins) satisfy a priori requirements in the worst operating conditions of the plant. On the other hand, adaptive systems estimate the uncertainty online and then form a control action to minimize the plant's deviation from a reference [3]. Compared to robust approaches, model reference adaptive systems need no a priori knowledge of the range of variations in the plant's parameters (the maximum value of parametric uncertainty); with a sufficient power margin for the control action, they yield a reference performance instead of a compromise one.

All methods of constructing model reference adaptive control systems can be divided into direct, indirect, and composite ones; see [3, 10, 11]. In the first case, the parameters of a preselected-structure controller are directly tuned; in the second case, the parameters of the plant and (or) parametric uncertainty are estimated, and this information is used to calculate the controller's parameters. Composite adaptive control systems combine direct and indirect adaptive control approaches to improve the adaptation process.

[1]This work was partially supported by the Russian Foundation for Basic Research, project no. 18-47-310003-r_a.

The main problems of all three groups of model reference adaptive control methods are as follows [3, 10]:

– The tuned parameters converge to the ideal values only when satisfying the regressor's persistent excitation requirement, which is rather restrictive.

– From the practice viewpoint, the standard tuning loop yields an unsatisfactory quality of transients for tunable parameters, control, and tracking error (especially when the number of tunable parameters increases).

– The gain matrix of the adaptation loop is selected experimentally (manually).

– It is necessary to know the sign of the plant's gain matrix.

In recent years, much effort has been applied by domestic and foreign researchers to solve these problems. Among the last significant results, we mention the publications [12–15].

However, there is a fifth (more fundamental and less developed) problem in the model reference adaptive control theory: the assumption about the plant's adaptability [3, 10]. According to this assumption, model reference adaptive control methods in the general statement can be applied only if the plant's parametric uncertainty is matched with the control signal. It means the theoretical possibility of fully compensating the uncertainty by direct subtraction of the generated control. If the adaptability condition does not hold, direct compensation becomes impossible: the uncertainty is unmatched, and special methods are required for designing the control law and its tuning.

Generally speaking, there are two main classes of uncertainties unmatched with the control signal. The first class includes disturbances in the plant's description by autonomous differential equations. The second class includes disturbances arising in the plant's nonautonomous equations with a deficit of control channels (the so-called *underactuated systems with unmatched uncertainties*).

For a long time, unmatched uncertainties of the first class have been compensated using adaptive backstepping methods [16] and indirect methods based on tuning functions [17]. The disadvantages of these approaches include a high dynamic order of the control law and its tuning and higher complexity of the design process when increasing the plant's order. New methods have recently been proposed [18–20] to solve these problems—consider and compensate the effect of unmatched parametric uncertainties—in a different way. These solutions directly combine the theory of adaptive [10] and robust [9] control. In particular, the following procedure was proposed in [18, 20]. First, indirect model reference adaptive control methods were used to estimate the unmatched uncertainties. Then the resulting information was adopted to recalculate the parameters of the controller and the reference model using LMIs synthesis. With such an approach, the robustness of the closed loop system to arbitrary unmatched uncertainties is adaptively maintained, and the effect of matched uncertainties is compensated.

The literature suggests few model reference adaptive control methods to compensate the effect of unmatched uncertainties from the second group. The main difficulty here is that the deficit of control channels leads to the presence of one control signal in several equations. In the general case, this leads to a nontrivial problem of compensating control design. Various methods of changing coordinates [21–25] are well known [21] in geometric and nonlinear control theory to solve this problem. These methods allow passing from a plant's model with a deficit of control actions to an equivalent normal-form model. As a result, an appropriate control law can be chosen by the feedback linearization method [21]. The disadvantages of such methods are the complexity (or even impossibility) of calculating the exact transformation for nonlinear high-dimensional plants and the dependence of the transformation itself on the plant's parameters. (Hence, it should have adaptation.) Therefore, the problem under consideration is still not completely solved.

This paper proposes a new approach to compensate an unmatched parametric uncertainty for an underactuated plant. It rests on the assumed existence of an ideal compensating control for reducing the unmatched uncertainty by a norm. With this assumption, we obtain compensating control by solving an optimization problem. Most real underactuated plants are described by nonlinear differential equations. Hence, the unmatched uncertainty is nonlinear, and the optimization problem turns out to be nonlinearly parameterized and, in the general statement, rather difficult to solve. To solve it in the general statement, we use artificial neural networks with universal approximating properties to form a compensating control action [26]. In this case, the parameter tuning laws of the compensator's neural network are designed by combining Lyapunov's second method and the backpropagation method.

The main result is a new neural-network-based compensating control law and an online algorithm to tune its parameters, which ensure the asymptotic con-

vergence of the tracking error of a nonlinear underactuated plant to a given domain with a chosen reference model.

A two-wheeled balancing robot is a classical example of a nonlinear underactuated plant. Therefore, we construct an adaptive neural-network-based control for such a robot without loss of generality as an illustrative example.

This paper uses the following notations: $(.)(i,j)$ or $(.)_{i,j}$ is an element standing at the junction of the ith row and jth column of a given matrix; $\det(.)$ is the matrix determinant; $\mathrm{tr}(.)$ is the matrix trace (the sum of all elements on the principal diagonal of a given matrix); $\mathrm{vec}(.)$ is the matrix vectorization (stacking the columns of a given matrix); $\mathrm{diag}\{a \quad b \quad \dots \quad c\}$ is a diagonal matrix with elements $a,b,\dots,c$ on the principal diagonal; $\|.\|_\infty$ is the L_∞ norm of a given matrix; $\|.\|$ is the Euclidean vector norm or Frobenius matrix norm, depending on the context; $(.)_{n\times n}$ is a matrix of dimensions $n\times n$.

1. THE MATHEMATICAL MODEL OF A TWO-WHEELED BALANCING ROBOT

The differential equations describing the dynamics of a two-wheeled balancing robot are derived using the Euler–Lagrange second method [27]. After reducing to the Cauchy form, they are given by

$$\begin{cases} \dot{x}_1 = x_3, \\ \dot{x}_2 = x_4, \\ \dot{x}_3 = -2(\beta + f_w)\left[E^{-1}(1,1) - E^{-1}(1,2)\right]x_3 - \\ 2\beta\left[E^{-1}(1,2) - E^{-1}(1,1)\right]x_4 + \\ \left[E^{-1}(1,1)MLRx_4^{\,2} + E^{-1}(1,2)MgL\right]\sin(x_2) + \\ \alpha\left[E^{-1}(1,1) - E^{-1}(1,2)\right](u_1 + u_2), \\ \dot{x}_4 = -2(\beta + f_w)\left[E^{-1}(2,1) - E^{-1}(2,2)\right]x_3 - \\ 2\beta\left[E^{-1}(2,2) - E^{-1}(2,1)\right]x_4 + \\ \left[E^{-1}(2,1)MLRx_4^{\,2} + E^{-1}(2,2)MgL\right]\sin(x_2) + \\ \alpha\left[E^{-1}(2,1) - E^{-1}(2,2)\right](u_1 + u_2), \end{cases} \tag{1.1}$$

where

$$E^{-1} = \begin{bmatrix} \dfrac{J + ML^2 + 2n^2J_m}{\det(\mathrm{E})} & -\dfrac{MLR\cdot\cos(x_2) - 2n^2J_m}{\det(\mathrm{E})} \\ -\dfrac{MLR\cdot\cos(x_2) - 2n^2J_m}{\det(\mathrm{E})} & \dfrac{2J_w + (2m_w + M)R^2 + 2n^2J_m}{\det(\mathrm{E})} \end{bmatrix},$$

$$\det(E) = \left(2J_w + (2m_w + M)R^2 + 2n^2J_m\right)\times$$
$$\left(J + ML^2 + 2n^2J_m\right) - \left(MLR\cdot\cos(x_2) - 2n^2J_m\right)^2,$$
$$\beta = \frac{nK_tK_b}{R_m} + f_m, \text{ and } \alpha = \frac{nK_t}{R_m}.$$

The model (1.1) is obtained by assuming the structural and parametric identity of the actuating motors. The parameters of the model (1.1) have the following physical interpretation: J_w is the wheel's moment of inertia; m_w is the wheel's mass; M is the robot's mass; R is the wheel's radius; n is the gear ratio of the gearbox; J_m is the motor's moment of inertia; L is the distance between the center of mass and the wheel axis; K_t is the motor torque constant; R_m is the resistance of the motor's armature circuit; K_b is the back emf constant; f_m is the coefficient of friction between the robot's body (further called the body) and the motor shaft; f_w is the coefficient of friction between the wheel and the motion surface; J is the body's moment of inertia; g is the acceleration of gravity. The robot's state variables are the average angle of rotation of the wheels (x_1), the body's angle of deflection from the normal (x_2), the wheel turning rate (x_3), and the body's rate of deflection from the normal (x_4). The voltages u_1 and u_2 applied to the left and right motors, respectively, are the control action.

For convenience, let us introduce the following additional notations for the model (1.1):

$$\begin{gathered} f_3(x_2, x_3, x_4) = \\ -2(\beta + f_w)\left[E^{-1}(1,1) - E^{-1}(1,2)\right]x_3 - \\ 2\beta\left[E^{-1}(1,2) - E^{-1}(1,1)\right]x_4 + \\ \left[E^{-1}(1,1)MLRx_4^{\,2} + E^{-1}(1,2)MgL\right]\sin(x_2), \\ g_{31}(x_2) = g_{32}(x_2) = \alpha\left[E^{-1}(1,1) - E^{-1}(1,2)\right], \\ f_4(x_2, x_3, x_4) = \\ -2(\beta + f_w)\left[E^{-1}(2,1) - E^{-1}(2,2)\right]x_3 - \\ 2\beta\left[E^{-1}(2,2) - E^{-1}(2,1)\right]x_4 + \\ \left[E^{-1}(2,1)MLRx_4^{\,2} + E^{-1}(2,2)MgL\right]\sin(x_2), \\ \text{and } g_{41}(x_2) = g_{42}(x_2) = \alpha\left[E^{-1}(2,1) - E^{-1}(2,2)\right]. \end{gathered} \tag{1.2}$$

2. PROBLEM STATEMENT

Using the notations (1.2), we write the model (1.1) in the state space:

$$\dot{x} = A_0 x + B_3 f_3(x_2, x_3, x_4) + B_4 f_4(x_2, x_3, x_4) + g(x)u,$$
$$A_0 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}; \; B_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}; \tag{2.1}$$
$$B_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}; \; g(x_2) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ g_{31}(x_2) & g_{32}(x_2) \\ g_{41}(x_2) & g_{42}(x_2) \end{pmatrix}.$$

Here $x \in R^4$ is the measured state vector of the robot; $u = [u_1; u_2] \in R^2$ is the vector of voltages applied to the left and right motors; $f_3(x_2, x_3, x_4)$, $f_4(x_2, x_3, x_4)$, and $g(x_2)$ are the nonlinear functions given by (1.2), which satisfy the Lipschitz smoothness conditions.

Assumption 1. *The control action u is such that* $u_1 = u_2$.

This assumption is classical when the robot's rotation about the center of mass in the horizontal plane is not required to control.

Due to Assumption 1, the model (1.1) with two control actions can be reduced to an equivalent model with one control action $v = u_1$:

$$\dot{x} = A_0 x + B_3 f_3(x_2, x_3, x_4) + B_4 f_4(x_2, x_3, x_4) + Bv,$$
$$B = \begin{bmatrix} 0 & 0 & g_3(x_2) & g_4(x_2) \end{bmatrix}^{\mathrm{T}}; \tag{2.2}$$
$$\begin{cases} g_3(x_2) = g_{31}(x_2) + g_{32}(x_2), \\ g_4(x_2) = g_{41}(x_2) + g_{42}(x_2). \end{cases}$$

In addition, we will consider an auxiliary linear model obtained from (2.2) by linearizing the functions $f_3(x_2, x_3, x_4)$ and $f_4(x_2, x_3, x_4)$ and the nonlinear elements of the vector B in the neighborhood of the unstable equilibrium $x_2 = 0$:

$$\dot{x} = A_0 x + B_3 \bar{f}_3(x_2, x_3, x_4) + B_4 \bar{f}_4(x_2, x_3, x_4) + \bar{B}v = \bar{A}x + \bar{B}v. \tag{2.3}$$

Remark 1. To derive the model (2.3) from the model (1.1), we set some values of the physical and geometric parameters of the robot model (1.1) and use the equalities $\lim\limits_{x_2 \to 0} \sin(x_2) = x_2$, $\lim\limits_{x_2 \to 0} \cos(x_2) = 1$, and $x_4^2 = 0$ holding in the neighborhood of the linearization point $x_2 = 0$. When constructing the linearized model (2.3), the robot's parameters can be uncertain.

Based on the linear model (2.3), the LQ-optimal control law can be calculated by

$$v_{\mathrm{LQ}} = K_{\mathrm{LQ}}(r - x) = K_{\mathrm{LQ}} e, \tag{2.4}$$

where $r \in R^4$ is the vector of reference signals for the robot's state variables, and the matrix K_{LQ} is found by minimizing the criterion

$$J = \frac{1}{2} \int_0^\infty x^{\mathrm{T}} Q_{\mathrm{LQ}} x + R_{\mathrm{LQ}} v^2 \, d\tau \tag{2.5}$$

with positive definite diagonal matrices $Q_{\mathrm{LQ}} \in R^{4\mathrm{x}4}$ and $R_{\mathrm{LQ}} \in R$.

The desired control performance for the nonlinear plant (2.2) is given by the system (2.3) with the controller (2.4):

$$\dot{x}_{ref} = A_{ref} x_{ref} + B_{ref} r,$$
$$A_{ref} = \bar{A} - \bar{B} K_{\mathrm{LQ}}; \; B_{ref} = \bar{B} K_{\mathrm{LQ}}. \tag{2.6}$$

To obtain an error equation between the nonlinear plant equations (2.2) and its linear reference model (2.6), we introduce the relations:

$$f_3(x_2, x_3, x_4) = \bar{f}_3(x_2, x_3, x_4) + \Delta_{f3},$$
$$f_4(x_2, x_3, x_4) = \bar{f}_4(x_2, x_3, x_4) + \Delta_{f4}, \tag{2.7}$$
$$B = \bar{B} + \Delta_B,$$

where Δ_{f3}, Δ_{f4}, and Δ_B are unknown Lipschitz smooth functions due to the parametric uncertainties and the nonlinearities for $x_2 \gg 0$.

Substituting the relations (2.7) into the model (2.2), we have

$$\dot{x} = A_0 x + B_3 \left[\bar{f}_3(x_2, x_3, x_4) + \Delta_{f3} \right] + B_4 \left[\bar{f}_4(x_2, x_3, x_4) + \Delta_{f4} \right] + \left(\bar{B} + \Delta_B \right) v. \tag{2.8}$$

Considering the expression (2.8), we choose the control law v in the form

$$v = v_{LQ} - v_{ad}. \tag{2.9}$$

Due to the expressions (2.3), (2.6), (2.7), and (2.9), equation (2.8) reduces to

$$\dot{x} = A_{ref} x + B_{ref} r + \underbrace{B_3 \Delta_{f3} + B_4 \Delta_{f4} + \Delta_B v}_{\Lambda(z)} - \bar{B} v_{ad}, \tag{2.10}$$

where $\Lambda(z)$ is an unknown Lipschitz smooth function that describes the effect of parametric uncertainties and nonlinearities on the control performance of the balancing robot, and $z = \begin{bmatrix} x_2 & x_3 & x_4 & v \end{bmatrix} \in D \subset R^4$ is the

variable of the function $\Lambda(z)$ defined in a compact domain D of the space R^4.

The error equation between (2.10) and (2.6) has the form

$$\dot{e}_{ref} = A_{ref} e_{ref} + \Lambda(z) - \bar{B} v_{ad}, \qquad (2.11)$$

where $e_{ref} = x - x_{ref}$ is the tracking error of the plant (2.10) with the reference model (2.6).

As is easily checked, the vector $\bar{B}$ has no Moore–Penrose pseudoinverse matrix such that $\bar{B}^{\dagger}\bar{B} = I$. Hence, $\Lambda(z)$ is a disturbance unmatched with the control signal. To design the control law v_{ad} in the adaptive problem statement, we accept the following assumption regarding compensation.

Assumption 2. *There exists a compensating signal* v_{ad}^* *of the variable z such that*

$$\begin{gathered} \left\| \Lambda(z) - \bar{B} v_{ad}^* \right\| \le \varepsilon_{\Lambda} < \Lambda_{\max}, \\ v_{ad}^* = \arg \left[\min_{v_{ad}^*} \left\{ \sup \left\| \Lambda(z) - \bar{B} v_{ad}^* \right\| \right\} \right], \end{gathered} \qquad (2.12)$$

where $\Lambda_{\max} = \sup_{\forall z \in L_{\infty}} \left\| \Lambda(z) \right\|$, *and* ε_{Λ} *is the approximation error of* $\Lambda(z)$ *with the compensating signal* v_{ad}^*.

Remark 2. If Assumption 2 is not satisfied, then the adaptive compensation of the disturbance $\Lambda(z)$ is a fundamentally unsolvable problem in the class of smooth functions.

For clarity, we provide an illustrative example of when Assumption 2 holds. Let the difference $\Lambda(z) - \bar{B} v_{ad}^*$ have the form

$$\Lambda(z) - \bar{B} v_{ad}^* = \begin{bmatrix} 0 \\ 0 \\ a_{32} x_2 + a_{33} x_3 + a_{34} x_4 \\ a_{42} x_2 + a_{43} x_3 + a_{44} x_4 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ b_3 \\ b_4 \end{bmatrix} v_{ad}^*.$$

In this case, Assumption 2 is satisfied under at least one of the following relations:

$$\frac{a_{32}}{a_{42}} \propto \frac{b_3}{b_4}; \ \frac{a_{33}}{a_{43}} \propto \frac{b_3}{b_4}; \ \frac{a_{34}}{a_{44}} \propto \frac{b_3}{b_4}.$$

The part of the uncertainty $\Lambda(z)$ unmatched with the control signal can be compensated.

Thus, without any information about the structure of the function $\Lambda(z)$ and the impossibility to compensate it fully, we pose the following problem: minimize the error (2.12) and ensure the asymptotic convergence of the tracking error (2.11) to a bounded set:

$$\lim_{t \to \infty} \left\| e_{ref}(t) \right\| \le \bar{\varepsilon}_{e_{ref}}, \qquad (2.13)$$

where $\bar{\varepsilon}_{e_{ref}}$ is the limit tracking error.

Before presenting the main result, we introduce and justify a constraint on $\bar{\varepsilon}_{e_{ref}}$. For this purpose, we estimate the minorant and majorant of the tracking error e_{ref}. Letting $v_{ad} = v_{ad}^*$ and $v_{ad} = 0$ in equation (2.11) yields lower and upper bounds on the derivative $\dot{e}_{ref}$:

$$A_{ref} e_{ref} + \Lambda(z) - \bar{B} v_{ad}^* \le \dot{e}_{ref} < A_{ref} e_{ref} + \Lambda(z). \qquad (2.14)$$

For calculating the minorant and majorant of the error e_{ref}, consider the quadratic form

$$\begin{gathered} L = e_{ref}^{\mathrm{T}} P e_{ref}, \\ \lambda_{\min}(P) \left\| e_{ref} \right\|^2 \le L \le \lambda_{\max}(P) \left\| e_{ref} \right\|^2, \end{gathered} \qquad (2.15)$$

where P is the solution of the Lyapunov equation $A_{ref}^{\mathrm{T}} P + P A_{ref} = -Q, \ Q > 0$.

Due to (2.14), the derivative of the quadratic form (2.15) satisfies the two-sided inequality

$$\begin{gathered} e_{ref}^{\mathrm{T}} \left(A_{ref}^{\mathrm{T}} P + P A_{ref} \right) e_{ref} + \\ 2 e_{ref}^{\mathrm{T}} P \left[\Lambda(z) - \bar{B} v_{ad}^* \right] \le \dot{L} < \\ e_{ref}^{\mathrm{T}} \left(A_{ref}^{\mathrm{T}} P + P A_{ref} \right) e_{ref} + 2 e_{ref}^{\mathrm{T}} P \Lambda(z). \end{gathered} \qquad (2.16)$$

According to Assumption 2, from (2.16) we obtain:

$$\begin{gathered} -\lambda_{\min}(Q) \left\| e_{ref} \right\|^2 + 2 \varepsilon_{\Lambda} \lambda_{\max}(P) \left\| e_{ref} \right\| \le \\ \dot{L} < -\lambda_{\min}(Q) \left\| e_{ref} \right\|^2 + 2 \lambda_{\max}(P) \Lambda_{\max} \left\| e_{ref} \right\|. \end{gathered} \qquad (2.17)$$

For any $a > 0$ and $b > 0$,

$$\begin{gathered} -a^2 + ab = \frac{1}{2} \left[-a^2 - (a-b)^2 + b^2 \right] \le \\ -\frac{1}{2} a^2 + \frac{1}{2} b^2. \end{gathered} \qquad (2.18)$$

Hence, the relations (2.17) imply

$$\begin{gathered} -\frac{\lambda_{\min}(Q)}{2\lambda_{\max}(P)} L + \frac{2\varepsilon_{\Lambda}^2 \lambda_{\max}^2(P)}{\lambda_{\min}(Q)} \le \dot{L} < \\ -\frac{\lambda_{\min}(Q)}{2\lambda_{\max}(P)} L + \frac{2\lambda_{\max}^2(P) \Lambda_{\max}^2}{\lambda_{\min}(Q)}. \end{gathered}$$

Considering (2.15), we have the following minorant and majorant of the tracking error:

$$\begin{gathered} \frac{1}{\lambda_{\min}(P)} e^{-\frac{\lambda_{\min}(Q)}{2\lambda_{\max}(P)} t} \left\| e_{ref}(0) \right\|^2 + \\ \frac{4\varepsilon_{\Lambda}^2 \lambda_{\max}^3(P)}{\lambda_{\min}(P) \lambda_{\min}^2(Q)} \le \left\| e_{ref} \right\|^2 < \\ \frac{1}{\lambda_{\min}(P)} e^{-\frac{\lambda_{\min}(Q)}{2\lambda_{\max}(P)} t} \left\| e_{ref}(0) \right\|^2 + \frac{4\lambda_{\max}^3(P) \Lambda_{\max}^2}{\lambda_{\min}(P) \lambda_{\min}^2(Q)}. \end{gathered} \qquad (2.19)$$

Letting $t \to \infty$ in (2.19), we finally estimate the limit tracking error as

$$2\varepsilon_\Lambda \frac{\lambda_{\max}(P)}{\lambda_{\min}(Q)}\sqrt{\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}} \le \bar{\varepsilon}_{e_{ref}} < 2\Lambda_{\max}\frac{\lambda_{\max}(P)}{\lambda_{\min}(Q)}\sqrt{\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}}. \quad (2.20)$$

Well, according to the expressions (2.13) and (2.20), it is required to construct a compensating law v_{ad} that ensures the asymptotic convergence of the tracking error e_{ref} to a given set with the boundary $\bar{\varepsilon}_{e_{ref}}$.

3. AUXILIARY RESULTS FROM THE THEORY OF NEURAL-NETWORK-BASED CONTROL

To achieve this goal under an unknown structure of the nonlinear function $\Lambda(z)$ and a nonlinear parametrization of the optimization problem (2.12), we will construct the compensating control v_{ad} using neural networks with their universal approximation properties [26]. This section provides auxiliary results from the theory of neural-network-based control necessary for further considerations.

Proposition 1 [26]. *Any continuous function $f(z)$: $R^n \to R$ can be uniformly approximated in a compact domain $D \subset R^n$ using a neural network with one hidden layer with a sigmoidal activation function: for all $\bar{\varepsilon}_{NN} > 0$ and $z \in D$, there exist matrices V and W and values b^1 and b^2 such that*

$$\left\| f(z) - f_{NN}(z) \right\|_\infty = \left\| f(z) - W^{\mathrm{T}}\sigma\left(V^{\mathrm{T}}\bar{z}\right) \right\|_\infty \le \bar{\varepsilon}_{NN},$$

where

$$\bar{z} = \begin{bmatrix} b^1 & z \end{bmatrix}^{\mathrm{T}}, \ \sigma(V\cdot\bar{z}) = \begin{bmatrix} b^2 & \sigma_1 & \sigma_2 & \cdots & \sigma_{N_2} \end{bmatrix}^{\mathrm{T}},$$

$$V = \begin{bmatrix} \theta_1^V & \cdots & \theta_{N_2}^V \\ v_{1,1} & \cdots & v_{1,N_2} \\ \vdots & \ddots & \vdots \\ v_{N_1,1} & \cdots & v_{N_1,N_2} \end{bmatrix}^{\mathrm{T}}, \text{ and } W = \begin{bmatrix} \theta_1^W & \cdots & \theta_{N_3}^W \\ w_{1,1} & \cdots & w_{1,N_3} \\ \vdots & \ddots & \vdots \\ w_{N_2,1} & \cdots & w_{N_2,N_3} \end{bmatrix}^{\mathrm{T}}.$$

The matrices $V \in R^{N_1 \times N_2}$ and $W \in R^{N_2 \times N_3}$ are the weight matrices of the hidden and output layers, respectively; b^1 and b^2 are the biases of the hidden and output layers, respectively; σ is the sigmoidal activation function of the hidden layer.

In other words, a neural network with a sigmoidal activation function approximates any continuous function of the variable z in a compact domain $D \in R^n$ with the error $\bar{\varepsilon}_{NN} = \sup_{z\in D}\left\|\varepsilon(z)\right\|$:

$$f(z) = W^{\mathrm{T}}\sigma\left(V^{\mathrm{T}}\bar{z}\right) + \varepsilon(z). \quad (3.1)$$

In this case, the error $\bar{\varepsilon}_{NN}$ can be made arbitrarily small by increasing the number of neurons N_2 in the hidden layer.

Proposition 1 establishes the existence of ideal parameters of a neural network, not determining their values. Therefore, equation (3.1) is interpreted as describing the ideal output of a neural network. It is used to introduce the concept of a neural network with the parameters tuned by training:

$$\hat{f} = \hat{W}^{\mathrm{T}}\sigma\left(\hat{V}^{\mathrm{T}}\bar{z}\right). \quad (3.2)$$

The error between the current (3.2) and ideal (3.1) outputs of the neural network has the form

$$e = W^{\mathrm{T}}\sigma\left(V^{\mathrm{T}}\bar{z}\right) - \hat{W}^{\mathrm{T}}\sigma\left(\hat{V}^{\mathrm{T}}\bar{z}\right) + \varepsilon(z). \quad (3.3)$$

Hence, the ideal parameters V and W of the neural network can be found by optimizing the error function (3.3) with respect to the tuned parameters:

$$(V,\ W) = arg\left[\min_{(\hat{V},\hat{W})}\left\{\sup_{z\in D}|e|\right\}\right]. \quad (3.4)$$

Assumption 3. *The ideal weights of the neural network are bounded in a compact domain D:*

$$\|V\| \le V_M,\ \|W\| \le W_M.$$

The optimization problem (3.4) is nonlinearly parameterized due to the nonlinear activation function of the hidden layer. The error (3.3) is therefore rewritten in an approximate linearly parameterized form by expanding the activation function of the hidden layer into the Taylor series.

Proposition 2 [28]. *The linearly parameterized error e_{lin} of the neural-network-based approximation is given by*

$$e_{lin} = e = \tilde{W}^{\mathrm{T}}\left(\sigma\left(\hat{V}^{\mathrm{T}}\bar{z}\right) - \sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right)\hat{V}^{\mathrm{T}}\bar{z}\right) + \hat{W}^{\mathrm{T}}\sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right)\tilde{V}^{\mathrm{T}}\bar{z} + \varepsilon(z) - d, \quad (3.5)$$

where $\sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right) = \mathrm{diag}\left\{0 \quad \sigma_1' \quad \sigma_2' \quad \cdots \quad \sigma_{N_2}'\right\}$ denotes the derivative of the activation function of the hidden layer; $\tilde{V} = V - \hat{V}$ is the parametric error of the hidden layer of the neural network; $\tilde{W} = W - \hat{W}$ is the parametric error of the output layer of the neural network;

d is the residual term. The difference $(\varepsilon(z) - d)$ *in equation* (3.5) *is bounded* [28, 29] *due to the condition*

$$\left\|\varepsilon(z)-d\right\| \le \alpha_1 \left\|\tilde{Z}\right\| + \alpha_2,$$
$$\alpha_1 > 0,\ \alpha_2 > 0,\ \tilde{Z} = \begin{bmatrix} \tilde{W} & 0 \\ 0 & \tilde{V} \end{bmatrix}. \tag{3.6}$$

Then problem (3.4) is equivalent to the linearly parameterized problem

$$(V,W) = \arg\left[\min_{(\tilde{V},\tilde{W})}\left\{\sup_{z\in D}\left|e_{lin}\right|\right\}\right]. \tag{3.7}$$

The laws to tune the current parameters of the neural network [28] are obtained by solving (3.7):

$$\begin{cases} \dot{\hat{W}} = \Gamma_W\left(\sigma\left(\hat{V}^{\mathrm{T}}\bar{z}\right) - \sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right)\hat{V}^{\mathrm{T}}\bar{z}\right)e, \\ \hat{W}(0) = 0_{N_2\times N_3}, \\ \dot{\hat{V}} = \Gamma_V \bar{z} e \hat{W}^{\mathrm{T}}\sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right),\ \hat{V}(0) = 0_{N_1\times N_2}. \end{cases} \tag{3.8}$$

These auxiliary results from the theory of neural-network-based control will be used below to obtain the compensating control v_{ad}.

4. THE MAIN RESULT

Under Assumption 2, adding and subtracting the value $\bar{B}v_{ad}^*$ from equation (2.11) yield

$$\dot{e}_{ref} = A_{ref}e_{ref} + \Lambda(z) - \bar{B}v_{ad} \pm \bar{B}v_{ad}^* = A_{ref}e_{ref} + \bar{B}\left[v_{ad}^* - v_{ad}\right] + \Lambda(z) - \bar{B}v_{ad}^*. \tag{4.1}$$

By the problem statement, $z \in D \subset R^4$, and v_{ad}^* is a function of the variable z (Assumption 2). According to Proposition 1, the function v_{ad}^* can be approximated using an artificial neural network:

$$v_{ad}^* = W^{\mathrm{T}}\sigma\left(V^{\mathrm{T}}\bar{z}\right) + \varepsilon(z),$$
$$\bar{z} = \begin{bmatrix} b^1 & z \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} b^1 & x_2 & x_3 & x_4 & v \end{bmatrix}^{\mathrm{T}}. \tag{4.2}$$

Therefore, we choose v_{ad} in the form

$$v_{ad} = \hat{W}^{\mathrm{T}}\sigma\left(\hat{V}^{\mathrm{T}}\bar{z}\right). \tag{4.3}$$

Due to the expressions (4.3), (4.2), and (3.5), equation (4.1) reduces to

$$\dot{e}_{ref} = A_{ref}e_{ref} + \bar{B}\left[\tilde{W}^{\mathrm{T}}\left(\sigma\left(\hat{V}^{\mathrm{T}}\bar{z}\right) - \sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right)\hat{V}^{\mathrm{T}}\bar{z}\right) + \hat{W}^{\mathrm{T}}\sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right)\tilde{V}^{\mathrm{T}}\bar{z} + \varepsilon(z) - d\right] + \Lambda(z) - \bar{B}v_{ad}^*. \tag{4.4}$$

Based on the laws (3.8), we introduce the following tuning laws for the weights of the hidden and output layers of the compensating neural network (4.3):

$$\begin{cases} \dot{\hat{W}} = \Gamma_W\left[\left(\sigma\left(\hat{V}^{\mathrm{T}}\bar{z}\right) - \sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right)\hat{V}^{\mathrm{T}}\bar{z}\right)\times \right. \\ \left. e_{ref}^{\mathrm{T}}P\bar{B} - \sigma_W\hat{W}\right],\ \hat{W}(0) = 0_{N_2\times N_3}, \\ \dot{\hat{V}} = \Gamma_V\left[\bar{z}e_{ref}^{\mathrm{T}}P\bar{B}\hat{W}^{\mathrm{T}}\sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right) - \sigma_V\hat{V}\right], \\ \hat{V}(0) = 0_{N_1\times N_2}, \end{cases} \tag{4.5}$$

where $\sigma_W > 0$ and $\sigma_V > 0$ are the coefficients of the sigma modifications [10].

Remark 3. Contrary to popular belief, the compensating neural network (4.3) needs no preliminary autonomous training: it can be tuned by formulas (4.5), starting from the zero parameters of the layers, directly during the plant's operation.

Based on equation (4.4), we introduce the generalized error vector $\zeta = \left[e_{ref}^{\mathrm{T}} \quad \mathrm{vec}^{\mathrm{T}}\left(\tilde{W}\right) \quad \mathrm{vec}^{\mathrm{T}}\left(\tilde{V}\right)\right]^{\mathrm{T}}$ and study its properties.

Theorem 1. *Let the compensating law* v_{ad} *be given by* (4.3), *and let its parameters be tuned by formulas* (4.5). *Then the generalized error* ζ *is uniformly and ultimately bounded. Moreover, the steady-state tracking error* e_{ref} *can be reduced to satisfy inequalities* (2.13) *and* (2.20) *by increasing the number of neurons* N_2 *in the hidden layer and decreasing the values of the coefficients* σ_V *and* σ_W.

The proof of Theorem 1 is postponed to the Appendix.

Thus, we ensure the asymptotic convergence of the tracking error e_{ref} to a given domain using the neural-network-based compensating law (4.3) and tuning its parameters by formulas (4.5).

Remark 4. These recommendations for increasing the number of neurons N_2 in the hidden layer and decreasing the values of the coefficients σ_V and σ_W have rather simple interpretations:

– Increasing the number of neurons N_2 in equation (4.4) allows satisfying the inequality $\Lambda(z) + \bar{B}\left[-v_{ad}^* + \varepsilon(z) - d\right] \le \left\|\varepsilon_\Lambda + \bar{B}\left[\varepsilon(z) - d\right]\right\| < \Lambda_{\max}$. In other words, the uncertainty $\bar{B}\left[\varepsilon(z) - d\right]$ introduced by the neural network into the closed loop does not increase the system uncertainty after compensating ε_Λ compared to the initial value $\Lambda_{\max}$.

– Choosing small values of the coefficients σ_V and σ_W allows the tunable neural network to better approximate the ideal compensating signal $W^{\mathrm{T}}\sigma\left(V^{\mathrm{T}}\bar{z}\right)$, thereby reducing more the error $\tilde{W}^{\mathrm{T}}\left(\sigma\left(\hat{V}^{\mathrm{T}}\bar{z}\right)-\sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right)\hat{V}^{\mathrm{T}}\bar{z}\right)+\hat{W}^{\mathrm{T}}\sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right)\tilde{V}^{\mathrm{T}}\bar{z}$. However, decreasing the values of the coefficients σ_W and σ_V also reduces the robustness of the tuning laws (4.5) to the uncompensated uncertainty $\left\|\varepsilon_\Lambda+\bar{B}\left[\varepsilon(z)-d\right]\right\|$ (under σ_V and σ_W close to 0, the value γ_1 in (A.7) can be negative). Hence, the values σ_V and σ_W should be assigned by the classical tradeoff between the quality of tracking the ideal trajectory x_{ref} and system robustness to the uncompensated uncertainty $\left\|\varepsilon_\Lambda+\bar{B}\left[\varepsilon(z)-d\right]\right\|$.

5. EXPERIMENTAL VALIDATION OF THE RESULTS

The developed control system was applied during experiments to the mathematical model of a LEGO EV3 balancing robot in Matlab/Simulink. The nominal values of the robot's parameters in the model (1.1) are given below.

Nominal values of robot's parameters

Parameter	Value	Parameter	Value
J_w, kg·m^2	$8.75\cdot10^{-6}$	K_b, V·s/rad	0.468
m_w, kg	0.024	R_m, Ω	6.69
R, m	0.027	f_m	0.0022
n	1	f_w	0
J_m, kg·m^2	10^{-5}	g, m/s^2	9.81
L, m	0.105	M, kg	0.8
K_t, N·m/A	0.317		

The body's moment of inertia J was calculated by the formula $J = ML^2/3 = 0.0029$ kg · m^2. The gain matrix K_{LQ} of the LQ controller was obtained using the robot's linearized model (2.3) with the nominal parameter values (see the table) by optimizing the criterion (2.5) with the matrices $Q = I$ and $R = 1$:

$$K_{\mathrm{LQ}} = \left[-0.7071 \quad -77.0619 \quad -1.5816 \quad -9.3949\right].$$

The experiments involved a neural network with a sigmoidal activation function with four neurons on the input layer ($N_1 = 5$), forty neurons on the hidden layer ($N_2 = 40$), and one neuron on the output layer ($N_3 = 1$). In all experiments, the variable parameters of the neural network's tuning loop were as follows:

$\Gamma_W = 10^5 I_{N_2\times N_2}$, $\Gamma_V = 10^{-3} I_{N_2\times N_2}$, $\sigma_W = 0.1$, $\sigma_V = 0.001$. For switching from the control v back to control u, the relation $v = u_1 = u_2$ was used in the experiments (see Assumption 1).

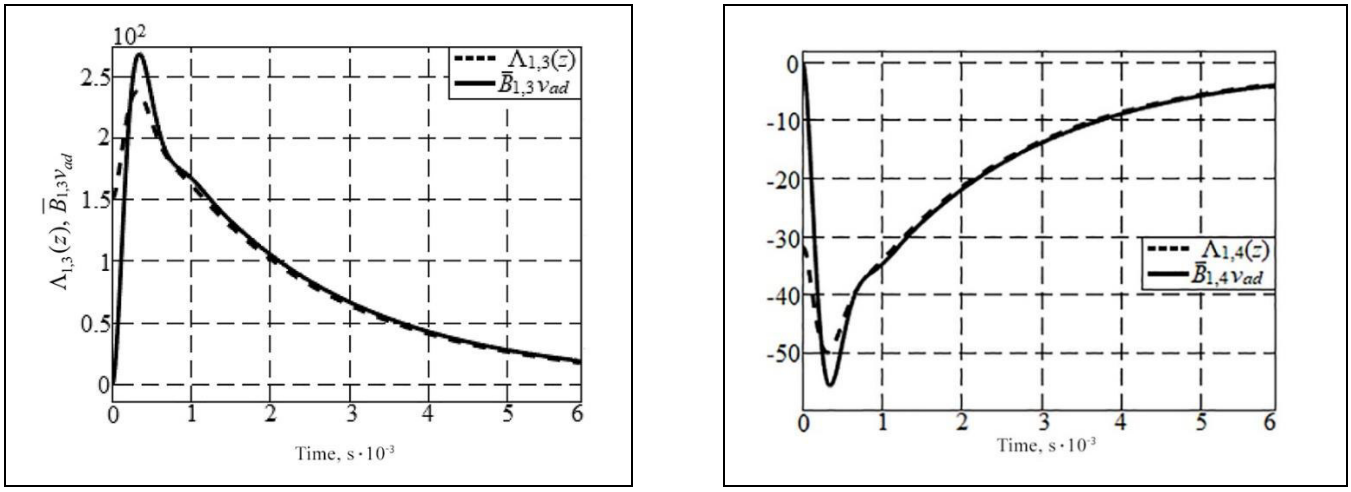
Two experiments were carried out in total. The first experiment was intended to check the compensator (4.3) when approximating the uncertainties caused by changes in the robot's parameters during operation in the neighborhood of the linearization point $x_2 = 0$. The second experiment was intended to check the compensator (4.3) when approximating the uncertainties caused by changes in the robot's parameters and nonlinearities during operation in a domain out of the neighborhood mentioned. The initial conditions of the plant (2.1) and the reference model (2.6) were the same in all experiments, and the zero vector was used as the reference signal r (the stabilization mode of the balancing robot).

In the first experiment, the models (2.1) and (2.6) began to move from the state-space point $x(0) = \left[0 \quad 0.01 \quad 0 \quad 0\right]^{\mathrm{T}}$, and the function $\Lambda(z)$ was caused by doubling the robot's nominal mass M. Figure 1 shows the elements of the vector v_{ad} and function $\Lambda(z)$ in this experiment.

The transients in Fig. 1 demonstrate the high accuracy and sufficiently fast approximation of the disturbance $\Lambda(z)$ using the neural network.

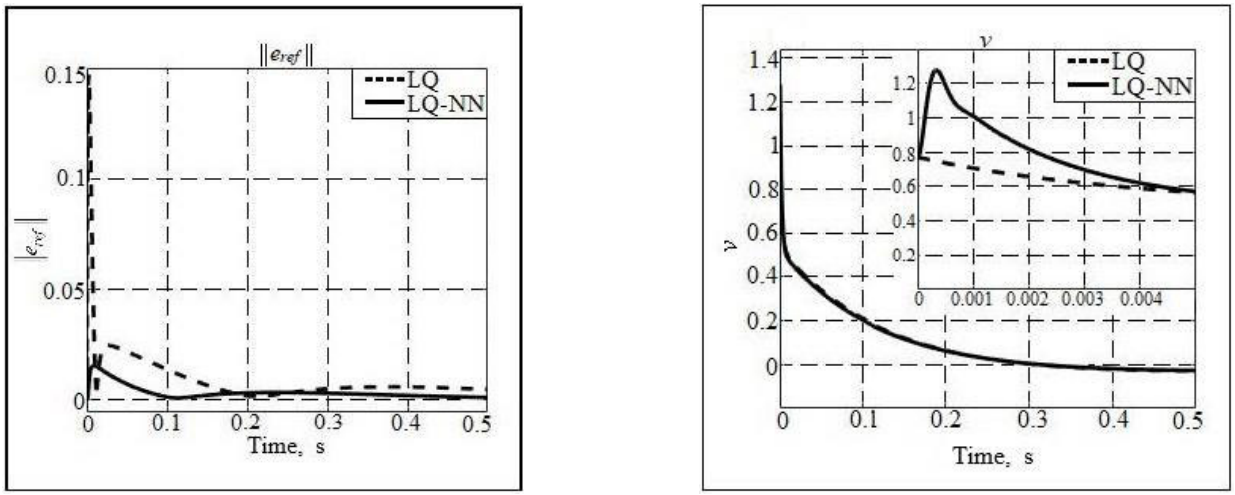
Figures 2*a* and 2*b* show the norms of the tracking errors e_{ref} and the control actions v, respectively, obtained in the first experiment using the control system with the neural network (LQ-NN) and without it (LQ).

The upper bound on the target set (2.13), (2.20) is given by the upper bound on the trajectory of the closed loop system with the LQ controller (for $v_{ad} = 0$). Hence, Fig. 2*a* confirms the uniform and ultimate boundedness of the tracking error e_{ref} by the target set (2.20). This result validates the conclusions of Theorem 1. According to Fig. 2*b*, the costs of the total control action v to compensate the uncertainty $\Lambda(z)$ are not significant.



(a) (b)

Fig. 1. The results of neural-network-based approximation: (*a*) $\Lambda_{13}(z)$ and (*b*) $\Lambda_{14}(z)$.



(a) (b)

Fig. 2. (*a*) Norms of the tracking errors e_{ref} in systems LQ-NN and LQ, (*b*) control actions v in systems LQ-NN and LQ.

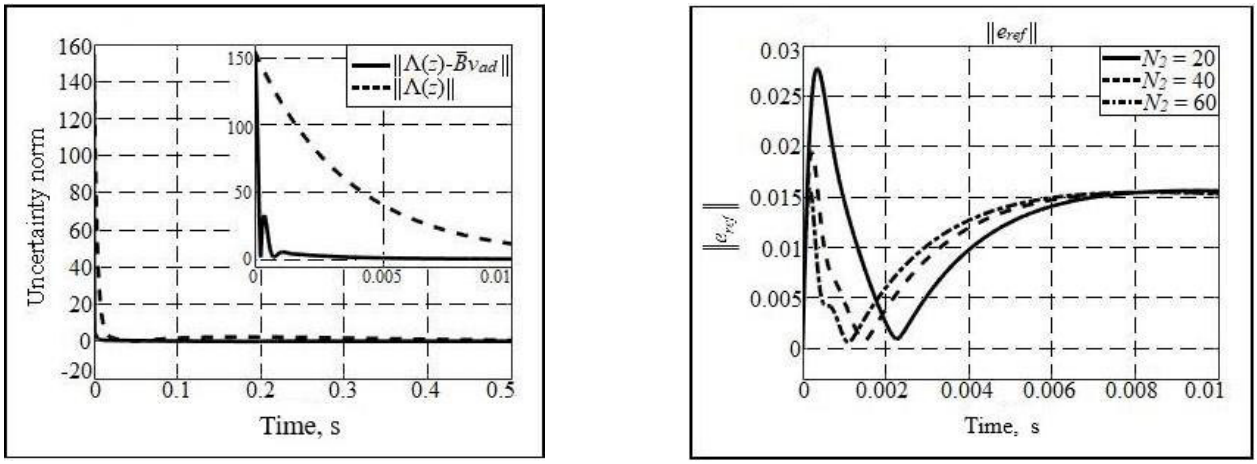
Figure 3*a* shows the uncertainty $\|\Lambda(z)\|$ in the control system with the LQ controller and the uncertainty $\|\Lambda(z)-\bar{B}v_{ad}\|$ after compensation in the control system with the neural-network-based compensator LQ-NN.

According to Fig. 3*a*, during the entire experiment, the area under the system uncertainty curve after compensation, $\|\Lambda(z)-\bar{B}v_{ad}\|$, is smaller than the area under the system uncertainty curve without compensation, $\|\Lambda(z)\|$. Hence, the following inequality holds:

$$\forall t>0:\ \dot{e}_{ref} \le A_{ref}e_{ref}+\Lambda(z)-\bar{B}v_{ad} < A_{ref}e_{ref}+\Lambda(z).$$

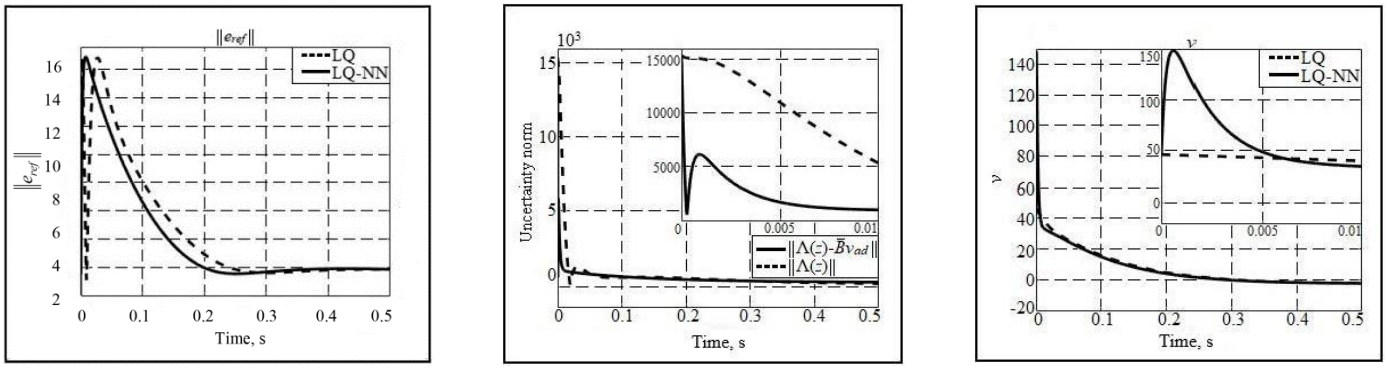
Using this result and considerations similar to (2.15)–(2.20), we validate the convergence of the tracking error e_{ref} to the domain specified by inequality (2.20); see Theorem 1. Figure 3*b* confirms the possibility of further reducing the tracking error e_{ref} by increasing the number of neurons N_2 in the hidden layer.

In the second experiment, the robot's mass was also doubled, but the robot (2.1) started moving from the initial state $x(0)=\begin{bmatrix}0 & 0.8 & 0 & 0\end{bmatrix}^{\mathrm{T}}$. Therefore, the uncertainty $\Lambda(z)$ was caused by the robot's nonstationary parameters and the nonlinearities. Figure 4 shows the norm of the error $(\Lambda(z)-\bar{B}v_{ad})$ and the norms of the error e_{ref} and control actions obtained in the second experiment using the control system with the neural network (LQ-NN) and without it (LQ).



(a) (b)

Fig. 3. (*a*) Uncertainties $\|\Lambda(z)\|$ and $\|\Lambda(z)-\bar{B}v_{ad}\|$, (*b*) norms of the tracking errors e_{ref} under different values N_2.



(a) (b) (c)

Fig. 4. (*a*) Norms of the tracking errors e_{ref} in systems LQ-NN and LQ, (*b*) uncertainties $\|\Lambda(z)\|$ and $\|\Lambda(z)-\bar{B}v_{ad}\|$, (*c*) control actions v in systems LQ-NN and LQ.

According to Fig. 4, the developed system can compensate the disturbance $\Lambda(z)$ caused by parametric uncertainty and robot's nonlinearities. Moreover, since the upper bound on the target set (2.13), (2.20) is given by the trajectories of the closed loop system with the LQ controller (for $v_{ad} = 0$), Fig. 4*a* confirms the convergence of the tracking error e_{ref} to the given domain in the case of nonlinear uncertainty $\Lambda(z)$. Comparing the costs of the control action v for compensating the linear (Fig. 2*b*) and nonlinear uncertainties (Fig. 4*b*), we arrive at the following result: the costs of the control action v grow proportionally with increasing the complexity of the function $\Lambda(z)$.

CONCLUSIONS

This paper has proposed an adaptive neural-network-based control system for a two-wheeled balancing robot with a rigorously proved stability of a closed control loop and a neural network compensator trained online.

The new procedure for designing an adaptive neural-network-based control system can be applied not only to a two-wheeled balancing robot but also to other nonlinear underactuated plants (e.g., industrial cranes [30, 31], manipulators [32], underwater vehicles [33], vertical and/or short take-off and landing (V/STOL) aircrafts [34, 35], and other mechanical systems [35, 36]).

APPENDIX

P r o o f of Theorem 1.

Consider the bounded quadratic form

$$V = e_{ref}^{\mathrm{T}} P e_{ref} + \mathrm{tr}\left(\tilde{W}^{\mathrm{T}} \Gamma_W^{-1} \tilde{W}\right) + \mathrm{tr}\left(\tilde{V}^{\mathrm{T}} \Gamma_V^{-1} \tilde{V}\right),$$
$$\lambda_m \|\zeta\|^2 \le V\left(\|\zeta\|\right) \le \lambda_M \|\zeta\|^2, \tag{A.1}$$
$$\lambda_{\mathrm{m}} = \min\left\{\lambda_{\min}(P), \lambda_{\min}\left(\Gamma_W^{-1}\right), \lambda_{\min}\left(\Gamma_V^{-1}\right)\right\},\ \lambda_M = \max\left\{\lambda_{\max}(P), \lambda_{\max}\left(\Gamma_W^{-1}\right), \lambda_{\max}\left(\Gamma_V^{-1}\right)\right\}.$$

We calculate the derivative of (A.1) taking into account (4.4) and (4.5):

$$\begin{gathered}\dot{V} = -e_{ref}^{\mathrm{T}} Q e_{ref} + 2e_{ref}^{\mathrm{T}} P\bar{B}\left[\tilde{W}^{\mathrm{T}}\left(\sigma\left(\hat{V}^{\mathrm{T}}\bar{z}\right) - \sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right)\hat{V}^{\mathrm{T}}\bar{z}\right) + \hat{W}^{\mathrm{T}}\sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right)\tilde{V}^{\mathrm{T}}\bar{z} + \varepsilon(z) - d\right] + \\ 2e_{ref}^{\mathrm{T}} P\left[\Lambda(z) - \bar{B}v_{ad}^{*}\right] - 2\mathrm{tr}\left(\tilde{W}^{\mathrm{T}}\Gamma_W^{-1}\Gamma_W\left[\left(\sigma\left(\hat{V}^{\mathrm{T}}\bar{z}\right) - \sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right)\hat{V}^{\mathrm{T}}\bar{z}\right)e_{ref}^{\mathrm{T}} P\bar{B} - \sigma_W\hat{W}\right]\right) - \\ 2\mathrm{tr}\left(\tilde{V}^{\mathrm{T}}\Gamma_V^{-1}\Gamma_V\left[\bar{z}e_{ref}^{\mathrm{T}} P\bar{B}\hat{W}^{\mathrm{T}}\sigma'\left(\hat{V}^{\mathrm{T}}\bar{z}\right) - \sigma_V\hat{V}\right]\right) = \\ -e_{ref}^{\mathrm{T}} Q e_{ref} + 2e_{ref}^{\mathrm{T}} P\bar{B}\left[\varepsilon(z) - d\right] + 2e_{ref}^{\mathrm{T}} P\left[\Lambda(z) - \bar{B}v_{ad}^{*}\right] + 2\mathrm{tr}\left(\tilde{W}^{\mathrm{T}}\sigma_W\hat{W}\right) + 2\mathrm{tr}\left(\tilde{V}^{\mathrm{T}}\sigma_V\hat{V}\right).\end{gathered} \tag{A.2}$$

Due to the expression (2.12) и (3.6), an upper bound on the derivative (A.2) is given by

$$\begin{gathered}\dot{V} \le -\lambda_{\min}(Q)\|e_{ref}\|^2 + 2\|e_{ref}\|\lambda_{\max}(P)\|\bar{B}\|\left(\alpha_1\|\tilde{Z}\| + \alpha_2\right) + 2\|e_{ref}\|\lambda_{\max}(P)\varepsilon_\Lambda + \\ 2\sigma_W\|\tilde{W}\|\|\hat{W}\| + 2\sigma_V\|\tilde{V}\|\|\hat{V}\| = \\ -\lambda_{\min}(Q)\|e_{ref}\|^2 + 2\|e_{ref}\|\lambda_{\max}(P)\|\bar{B}\|\left(\alpha_1\|\tilde{Z}\| + \alpha_2 + \alpha_3\right) + 2\sigma_W\|\tilde{W}\|\|\hat{W}\| + 2\sigma_V\|\tilde{V}\|\|\hat{V}\|,\end{gathered} \tag{A.3}$$

where $\alpha_3 = \dfrac{\varepsilon_\Lambda}{\|\bar{B}\|}$.

According to inequality (2.18), the terms in (A.3) satisfy the following upper bounds:

$$\begin{gathered}-\lambda_{\min}(Q)\|e_{ref}\|^2 + 2\|e_{ref}\|\lambda_{\max}(P)\|\bar{B}\|\left(\alpha_1\|\tilde{Z}\| + \alpha_2 + \alpha_3\right) \le \\ \frac{1}{2}\left[-\lambda_{\min}(Q)\|e_{ref}\|^2 - \left(\sqrt{\lambda_{\min}(Q)}\|e_{ref}\| - \frac{2\lambda_{\max}(P)\|\bar{B}\|}{\sqrt{\lambda_{\min}(Q)}}\left(\alpha_1\|\tilde{Z}\| + \alpha_2 + \alpha_3\right)\right)^2 + \right. \\ \left. \frac{4\lambda_{\max}^2(P)\|\bar{B}\|^2\left(\alpha_1\|\tilde{Z}\| + \alpha_2 + \alpha_3\right)^2}{\lambda_{\min}(Q)}\right] \le \frac{-\lambda_{\min}(Q)\|e_{ref}\|^2}{2} + \frac{2\lambda_{\max}^2(P)\|\bar{B}\|^2}{\lambda_{\min}(Q)}\left(\alpha_1\|\tilde{Z}\| + \alpha_2 + \alpha_3\right)^2, \\ \|\tilde{W}\|\|\hat{W}\| = -\|\tilde{W}\|^2 + \|\tilde{W}\|\|W\| \le -\frac{1}{2}\|\tilde{W}\|^2 + \frac{1}{2}W_M^2,\ \|\tilde{V}\|\|\hat{V}\| = -\|\tilde{V}\|^2 + \|\tilde{V}\|\|V\| \le -\frac{1}{2}\|\tilde{V}\|^2 + \frac{1}{2}V_M^2.\end{gathered} \tag{A.4}$$

Considering (A.4), the upper bound (A.3) takes the form

$$\dot{V} \le \frac{-\lambda_{\min}(Q)\|e_{ref}\|^2}{2} - \sigma_W\|\tilde{W}\|^2 - \sigma_V\|\tilde{V}\|^2 + \frac{2\lambda_{\max}^2(P)\|\bar{B}\|^2}{\lambda_{\min}(Q)}\left(\alpha_1\|\tilde{Z}\| + \alpha_2 + \alpha_3\right)^2 + \sigma_W W_M^2 + \sigma_V V_M^2. \tag{A.5}$$

Since

$$\sigma_W \left\|\tilde{W}\right\|^2 + \sigma_V \left\|\tilde{V}\right\|^2 \geq \underbrace{\min\{\sigma_W, \sigma_V\}}_{\sigma_Z} \left\|\tilde{Z}\right\|^2,$$

the expression (A.5) reduces to

$$\dot{V} \leq \frac{-\lambda_{\min}(Q)\left\|e_{ref}\right\|^2}{2} - \sigma_Z \left\|\tilde{Z}\right\|^2 + \frac{2\lambda_{\max}^2(P)\left\|\bar{B}\right\|^2}{\lambda_{\min}(Q)}\left[\alpha_1^2 \left\|\tilde{Z}\right\|^2 + 2\alpha_1(\alpha_2+\alpha_3)\left\|\tilde{Z}\right\| + (\alpha_2+\alpha_3)^2\right] +$$

$$\sigma_W W_M^2 + \sigma_V V_M^2 \leq \frac{-\lambda_{\min}(Q)\left\|e_{ref}\right\|^2}{2} - \left(\sigma_Z - \frac{2\lambda_{\max}^2(P)\left\|\bar{B}\right\|^2}{\lambda_{\min}(Q)}\alpha_1^2\right)\left\|\tilde{Z}\right\|^2 + \tag{A.6}$$

$$\frac{2\lambda_{\max}^2(P)\left\|\bar{B}\right\|^2}{\lambda_{\min}(Q)}\left(2\alpha_1(\alpha_2+\alpha_3)\left\|\tilde{Z}\right\| + (\alpha_2+\alpha_3)^2\right) + \sigma_W W_M^2 + \sigma_V V_M^2.$$

By analogy with (A.4), completing the square for the terms containing $\left\|\tilde{Z}\right\|$ in (A.6), we write

$$\dot{V} \leq \frac{-\lambda_{\min}(Q)\left\|e_{ref}\right\|^2}{2} - \frac{1}{2}\underbrace{\left(\sigma_Z - \frac{2\lambda_{\max}^2(P)\left\|\bar{B}\right\|^2}{\lambda_{\min}(Q)}\alpha_1^2\right)}_{\gamma_1}\left\|\tilde{Z}\right\|^2 +$$

$$\underbrace{\frac{2\lambda_{\max}^4(P)\left\|\bar{B}\right\|^4}{\lambda_{\min}^2(Q)}\gamma_1^{-1}\left(2\alpha_1(\alpha_2+\alpha_3)\right)^2 + \frac{2\lambda_{\max}^2(P)\left\|\bar{B}\right\|^2}{\lambda_{\min}(Q)}(\alpha_2+\alpha_3)^2 + \sigma_W W_M^2 + \sigma_V V_M^2}_{\gamma_2} = \tag{A.7}$$

$$-\frac{1}{2}\left(\lambda_{\min}(Q)\left\|e_{ref}\right\|^2 + \gamma_1\left\|\tilde{W}\right\|^2 + \gamma_1\left\|\tilde{V}\right\|^2\right) + \gamma_2 \leq -\kappa V + \gamma_2,\ \kappa = \frac{1}{2\lambda_M}\min\{\lambda_{\min}(Q), \gamma_1\}.$$

Here γ_1 and γ_2 are special notations for the compact form of (A.7); see above.

Applying the comparison lemma [21], we obtain a solution of inequality (A.7):

$$\left\|\zeta(t)\right\| \leq \sqrt{\frac{\lambda_M}{\lambda_m}e^{-\kappa t}\left\|\zeta(0)\right\| + \frac{\gamma_2}{\kappa\lambda_m}}. \tag{A.8}$$

Hence, the error ζ is uniformly and ultimately bounded [3, 10, 21].

Let us prove the asymptotic convergence of the tracking error e_{ref} to a given domain from inequality (A.8). Letting $t \to \infty$ and using the value γ_2, we arrive at the following limit estimate of the tracking error:

$$\left\|e_{ref}\right\| \leq \sqrt{\frac{2\lambda_{\max}(P)}{\lambda_{\min}(Q)\lambda_{\min}(P)}\gamma_2} \leq 2\varepsilon_\Lambda \frac{\lambda_{\max}(P)}{\lambda_{\min}(Q)}\sqrt{\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}} +$$

$$\underbrace{\sqrt{\frac{2\lambda_{\max}(P)}{\lambda_{\min}(Q)\lambda_{\min}(P)}}\left[\frac{\sqrt{2}\lambda_{\max}^2(P)\left\|\bar{B}\right\|^2}{\lambda_{\min}(Q)}\gamma_1^{-\frac{1}{2}}\left(2\alpha_1(\alpha_2+\alpha_3)\right) + \frac{\sqrt{2}\lambda_{\max}(P)\left\|\bar{B}\right\|}{\lambda_{\min}^{\frac{1}{2}}(Q)}\alpha_2 + \sqrt{\sigma_W}W_M + \sqrt{\sigma_V}V_M\right]}_{\alpha_4}. \tag{A.9}$$

Substituting this expression into (2.20), we check the inequality

$$2\varepsilon_\Lambda \frac{\lambda_{\max}(P)}{\lambda_{\min}(Q)}\sqrt{\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}} \leq 2\varepsilon_\Lambda \frac{\lambda_{\max}(P)}{\lambda_{\min}(Q)}\sqrt{\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}} + \alpha_4 < 2\left\|\Lambda(z)\right\|\frac{\lambda_{\max}(P)}{\lambda_{\min}(Q)}\sqrt{\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}}. \tag{A.10}$$

Inequality (A.10) holds if and only if the value α_4 is sufficiently small:

$$0 \le \alpha_4 < 2\left[\left\|\Lambda(z)\right\| - \varepsilon_\Lambda\right]\frac{\lambda_{\max}(P)}{\lambda_{\min}(Q)}\sqrt{\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}}.$$

By the definition (A.9), the value of the coefficient α_4 depends on those of the coefficients σ_V, σ_W, α_1, and α_2. In turn, the coefficients α_1 and α_2 (see formulas (3.1) and (3.6)) are inversely proportional to the number of neurons N_2 in the hidden layer. Therefore, the value of the coefficient α_4 can be reduced (thereby, ensuring inequality (A.10) and the asymptotic convergence of the tracking error e_{ref} to the given domain (2.13)) by increasing the number N_2 and decreasing the values of the coefficients σ_W and σ_V. The proof of Theorem 1 is complete.

REFERENCES

1. Spong, M.W., Hutchinson, S., and Vidyasagar, M., *Robot Modeling and Control*, New York: John Wiley & Sons, 2020.
2. Zhang, D. and Wei, B., A Review on Model Reference Adaptive Control of Robotic Manipulators, *Annual Reviews in Control*, 2017, vol. 43, pp. 188–198.
3. Nguyen, N.T., *Model-Reference Adaptive Control*, Cham: Springer, 2018.
4. Wise, K.A., Lavretsky, E., and Hovakimyan, N., Adaptive Control of Flight: Theory, Applications, and Open Problems, *Proceedings of the American Control Conference*, 2006, pp. 5966–5971.
5. Korzonek, M., Tarchala, G., and Orlowska-Kowalska, T.A., Review on MRAS-Type Speed Estimators for Reliable and Efficient Induction Motor Drives, *ISA Transactions*, 2019, vol. 93, pp. 1–13.
6. Cartes, D. and Wu, L., Experimental Evaluation of Adaptive Three-Tank Level Control, *ISA Transactions*, 2005, vol. 44, no. 2, pp. 283–293.
7. Lim, Y., Ulsoy, A.G., and Venugopal, R., *Process Control for Sheet-Metal Stamping*, New York: Springer, 2013.
8. Eremenko, Yu.I., Poleshchenko, D.A., Glushchenko, A.I., and Solodov, S.V., The Appliance Efficiency Estimation of PID-Regulator Parameters of Neural Optimizer for Solving of Control Problem of Metallurgical Heating Plants, *Izvestiya Ferrous Metallurgy*, 2014, vol. 57, no. 7, pp. 61–65. (In Russian.)
9. Zhou, K. and Doyle, J.C., *Essentials of Robust Control*, Upper Saddle River, NJ: Prentice Hall, 1998.
10. Narendra, K.S. and Annaswamy, A.M., *Stable Adaptive Systems*, North Chelmsford, MA: Courier Corporation, 2012.
11. Lavretsky, E., Combined/Composite Model Reference Adaptive Control, *IEEE Transactions on Automatic Control*, 2009, vol. 54, no. 11, pp. 2692–2697.
12. Chowdhary, G., Yucelen, T., Muhlegg, M., and Johnson, E.N., Concurrent Learning Adaptive Control of Linear Systems with Exponentially Convergent Bounds, *International Journal of Adaptive Control and Signal Processing*, 2013, vol. 27, no. 4, pp. 280–301.
13. Cao, C. and Hovakimyan, N., Design and Analysis of a Novel L1 Adaptive Control Architecture with Guaranteed Transient Performance, *IEEE Transactions on Automatic Control*, 2008, vol. 53, no. 2, pp. 586–591.
14. Aranovskiy, S., Bobtsov, A., Ortega, R., and Pyrkin, A., Performance Enhancement of Parameter Estimators via Dynamic Regressor Extension and Mixing, *IEEE Transactions on Automatic Control*, 2016, vol. 62, no. 7, pp. 3546–3550.
15. Gerasimov, D.N., Ortega, R., and Nikiforov, V.O., Relaxing the High-Frequency Gain Sign Assumption in Direct Model Reference Adaptive Control, *European Journal of Control*, 2018, vol. 43, pp. 12–19.
16. Kanellakopoulos, I., Kokotovic, P.V., and Morse, A.S., Systematic Design of Adaptive Controllers for Feedback Linearizable Systems, *American Control Conference*, IEEE, 1991, pp. 649–654.
17. Nikiforov, V.O. and Voronov, K.V. Adaptive Backstepping with a High-Order Tuner, *Automatica*, 2001, vol. 37, no. 12, pp. 1953–1960.
18. Quindlen, J.F., Chowdhary, G., and How, J.P., Hybrid Model Reference Adaptive Control for Unmatched Uncertainties, *Proceedings of the American Control Conference*, 2015, pp. 1125–1130.
19. Yayla, M. and Kutay, A.T., Adaptive Control Algorithm for Linear Systems with Matched and Unmatched Uncertainties, *Proceedings of the 55th Conference on Decision and Control (CDC)*, 2016, pp. 2975–2980.
20. Joshi, G. and Chowdhary, G., Hybrid Direct-Indirect Adaptive Control of Nonlinear System with Unmatched Uncertainty, *Proceedings of the 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, 2019, pp. 127–132.
21. Khalil, H.K., *Nonlinear Systems*, 3rd ed., Upper Saddle River: Prentice Hall, 2002.
22. Olfati-Saber, R. and Megretski, A., Controller Design for a Class of Underactuated Nonlinear Systems, *Proceedings of the 37th IEEE Conference on Decision and Control*, 1998, vol. 4, pp. 4182–4187.
23. Olfati-Saber, R., Normal Forms for Underactuated Mechanical Systems with Symmetry, *IEEE Transactions on Automatic Control*, 2002, vol. 47, no. 2, pp. 305–308.
24. Hauser, J., Sastry, S., and Kokotovic, P., Nonlinear Control via Approximate Input-Output Linearization: The Ball and Beam Example, *IEEE Transactions on Automatic Control*, 1992, vol. 37, no. 3, pp. 392–398.
25. Spong, M.W., Partial Feedback Linearization of Underactuated Mechanical Systems, *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'94)*, 1994. vol. 1, pp. 314–321.
26. Funahashi, K.I., On the Approximate Realization of Continuous Mappings by Neural Networks, *Neural Networks*, 1989, vol. 2, no. 3, pp. 183–192.
27. Yamamoto, Y., NXTway-GS Model-Based Design-Control of Self-Balancing Two-Wheeled Robot Built with Lego Mindstorms NXT, Cybernet Systems Co., Ltd., 2008.
28. Lewis, F.L., Yesildirek, A., and Liu, K., Multilayer Neural-Net Robot Controller with Guaranteed Tracking Performance, *IEEE Transactions on Neural Networks*, 1996, vol. 7, no. 2, pp. 388–399.
    Hovakimyan, N., Nardi, F., Calise, A., and Kim, N., Adaptive Output Feedback Control of Uncertain Nonlinear Systems Using Single-Hidden-Layer Neural Networks, *IEEE Transactions on Neural Networks*, 2002, vol. 13, no. 6, pp. 1420–1431.

29. Zhao, Y. and Gao, H., Fuzzy-Model-Based Control of an Overhead Crane with Input Delay and Actuator Saturation, *IEEE Transactions on Fuzzy Systems*, 2011, vol. 20, no. 1, pp. 181–186.
30. He, W. and Ge, S., Cooperative Control of a Nonuniform Gantry Crane with Constrained Tension, *Automatica*, 2016, vol. 66, pp. 146–154.
31. De Luca, A., Iannitti, S., Mattone, R., and Oriolo, G., Control Problems in Underactuated Manipulators, *Proceedings of the IEEE International Conference on Advanced Intelligent Mechatronics*, 2001, vol. 2, pp. 855–861.
32. Wang, J., Wang, C., Wei, Y., and Zhang, C., Command Filter Based Adaptive Neural Trajectory Tracking Control of an Underactuated Underwater Vehicle in Three-Dimensional Space, *Ocean Engineering*, 2019, vol. 180, pp. 175–186.
33. Hauser, J., Sastry, S., and Meyer, G., Nonlinear Control Design for Slightly Non-minimum Phase Systems: Application to V/STOL Aircraft, *Automatica*, 1992, vol. 28, no. 4, pp. 665–679.
34. Liu, Y. and Yu, H., A Survey of Underactuated Mechanical Systems, *IET Control Theory & Applications*, 2013, vol. 7, no. 7, pp. 921–935.
36. Spong, M., Underactuated Mechanical Systems, *Control Problems in Robotics and Automation*, 1998, pp. 135–150.

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

Received March 15, 2021, and revised August 17, 2021.
Accepted August 24, 2021.

Author information

Glushchenko, Anton Igorevich. Dr. Sci. (Eng.), Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ strondutt@mail.ru

Petrov, Vladislav Anatol'evich. Cand. Sci. (Eng.), Stary Oskol Technological Institute, National University of Science and Technology MISIS, Stary Oskol, Russia
✉ petrov.va@misis.ru

Lastochkin, Konstantin Andreevich. Engineer, Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ lastconst@yandex.ru

Cite this article
Glushchenko, A.I., Petrov, V.A., Lastochkin, K.A. Adaptive Neural-Network-Based Control of Nonlinear Underactuated Plants: An Example of a Two-Wheeled Balancing Robot. *Control Sciences* **5**, 29–42 (2021). http://doi.org/10.25728/cs.2021.5.3

Original Russian Text © Glushchenko, A.I., Petrov, V.A., Lastochkin, K.A., 2021, published in *Problemy Upravleniya*, 2021, no. 5, pp. 34–47.

Translated into English by *Alexander Yu. Mazurov*, Cand. Sci. (Phys.–Math.), Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ alexander.mazurov08@gmail.com

DOI: http://doi.org/10.25728/cs.2021.5.4

A BLOCK APPROACH TO CSTR CONTROL UNDER UNCERTAINTY, STATE-SPACE AND CONTROL CONSTRAINTS[1]

S.I. Gulyukina[1] and V.A. Utkin[2]

Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia

[1]✉ gulyukina.s.i@mail.ru, [2]✉vicutkin@ipu.ru

Abstract. This paper designs a control law to maintain the temperature in the jacket of a continuous stirred tank reactor (CSTR). The standard mathematical model describing the reactor operation is extended by introducing the actuator's dynamics. The state-space and control constraints are considered by a nonlinear change of the variables of the plant's initial model using linear sat functions. In the transformed system, these constraints are considered by feedback control law design. The block approach allows linearizing the feedback control law by sequentially solving the first-order design subproblems. Under incomplete information on the state vector and the effect of exogenous disturbances, an observer of the state vector and disturbances is constructed to estimate the unknown signals with a given accuracy. The effectiveness of the proposed approach is illustrated by simulating the CSTR–DC motor system in MATLAB.

Keywords: CSTR, tracking problem, block approach, observer of state vector and disturbances, state-space and control constraints.

INTRODUCTION

The continuous stirred tank reactor (CSTR) is widespread in the chemical industry. The CSTR dynamics are usually described by two nonlinear first-order differential equations [1, 2], a reference model for applying and testing new control algorithms.

Nowadays, improving CSTR control is a topical problem that attracts the attention of many control theorists and practitioners [3–8].

Much research in this field involves the sliding mode approach [9–16]: it ensures robust properties of closed loop systems and invariance to exogenous disturbances acting in control channels. Note that within this approach, control laws are often represented by the plant's variables that physically cannot be discontinuous functions, e.g., the flow rate of the coolant in the CSTR jacket. Hence, the practical importance of the sliding mode approach in the automation of various technological processes is significantly reduced.

State observers based on sliding modes and systems with deep feedback [17–21] are widely used to obtain information about the state vector and disturbances. Note that under disturbances, all these vectors can be estimated only within such an approach.

The problem of considering physical constraints on the state vector and control is underinvestigated in control theory. For example, only control constraints were taken into account in [22–24].

This paper proposes a complex solution for CSTR control that develops the original control law design method [25] for mechanical systems with constraints. It is methodologically based on a block approach to control [26], which decomposes high-dimensional problems into independently solvable subproblems of lower dimensions when designing feedback control laws and state observers. Treating the state variables as fictitious control actions, first of all, we satisfy the matching conditions for the disturbances in each subproblem. (In other words, the disturbance belongs to the control space.) In addition, using sat functions in local feedback law design, we ensure the bounded components of the state vector and controls.

[1]This work was supported by the Russian Foundation for Basic Research, project no. 20-01-00363 A.

This paper is organized as follows. Section 1 briefly describes the operation principle of a CSTR and its mathematical model and states the problem. In Section 2, we construct an observer of the state vector and disturbances with discontinuous and continuous corrections with large gains. In Section 3, feedback design algorithms are developed by combining local feedback laws with continuous sat functions and discontinuous control of the armature voltage of a DC motor. Section 4 illustrates the effectiveness of the proposed algorithms by simulation modeling in MATLAB.

1. PLANT'S MATHEMATICAL MODEL. PROBLEM STATEMENT

A CSTR is a key component of equipment needed to complete chemical reactions in many chemical and biochemical industries. A complex chemical reaction occurs in a CSTR, e.g., converting a hazardous chemical waste (reagent) into an acceptable chemical product. The schematic diagram of the reactor is shown in Fig. 1.

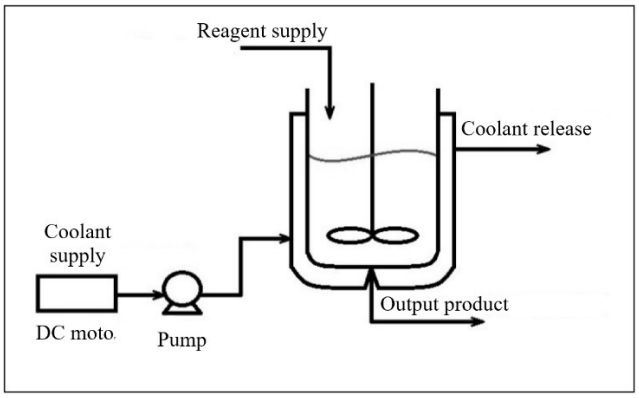


Fig. 1. The schematic diagram of a CSTR.

An irreversible first-order exothermic reaction $A \rightarrow B + \Delta H$ occurs in the tank, where A is a reagent, B is a product, and ΔH is the thermal effect of the chemical reaction (enthalpy).

The CSTR volume is equal to V. A reagent having a concentration C_{Af}, a temperature T_f, and a density ρ is supplied to the reactor input at a flow rate q.

The temperature in the reactor is maintained at a certain level through setting a coolant temperature in the jacket, T_c, by controlling the flow rate of the coolant. The reactor's output product is characterized by a temperature T, a concentration C_A, and a flow rate q under the invariable substance volume in the reactor.

Assuming the invariable substance volume in the reactor, ideal mixing, and the invariable substance density in the reactor, the laws of conservation of mass and energy yield the following dynamic model of CSTR [16]:

$$\dot{C}_A(t) = \frac{q}{V}(C_{Af}(t) - C_A(t)) - k_0 C_A(t) e^{-E/RT(t)},$$

$$\dot{T}(t) = \frac{q}{V}(T_f(t) - T(t)) + \frac{(-\Delta H)}{\rho C_p} k_0 C_A(t) e^{-E/RT(t)} + \frac{UA}{\rho V C_p}(T_c - T(t)). \quad (1)$$

The parameters of the model (1) are described in Table 1.

Table 1

The parameters of the CSTR model: values and units of measurement

Parameter	Value	Unit of measurement
C_A	The concentration of product B	kmol/m^3
T	Temperature in the reactor and product temperature	K
q	Reagent flow rate	m^3/min
V	Reactor volume	m^3
C_{af}	Reagent concentration	kmol/m^3
k_0	First-order reaction rate constant	min^{-1}
E	Activation energy	J/mol
R	Universal gas constant	J/(mol·K)
T_f	Reagent temperature	K
ΔH	Reaction enthalpy	cal/kmol
ρ	Reagent density	g/m^3
T_c	Coolant temperature	K
C_p	Reagent specific heat	cal/(K·g)
U	Heat-transfer coefficient	W/(m^2·K)
A	Heat delivery surface	m^2

The goal of CSTR control is to tune the coolant temperature T_c by changing the coolant flow rate so that the temperature T in the reactor corresponds to the desired values.

The coolant is supplied to the reactor by a pump with a DC motor. According to [27], the operation of this motor is described by the system of equations

$$\dot{x}_3(t) = a_{21}(g x_4(t) - m_L(t)),$$
$$\dot{x}_4(t) = a_{32}(u_2(t) - g x_3(t) - a_{31} x_4(t)), \quad (2)$$

where $x_3(t)$ is the motor shaft rotation frequency; $x_4(t)$ is the armature current; $u_2(t)$ is the armature voltage; $g = \text{const}$ is a magnetic flux; $m_L(t)$ is the

load moment; $a_{ij} = \text{const} > 0$ are the motor's design parameters.

The problem is to track a given coolant temperature in the jacket, $T_d(t)$, by the output variable $T(t)$:

$$e_2(t) = T(t) - T_d(t) \to 0.$$

Assume that the pump load is $m_L = mx_3^2(t)$, $m = \text{const}$, and the temperature in the jacket is proportional to the coolant flow rate: $T_c(t) - T(t) = x_3(T_{c0} - T(t))$, where $T_{c0} = \text{const} > 0$ denotes the coolant temperature at the jacket input. Under these assumptions, we write the plant's model (1), (2) in new variables:

$$\begin{aligned} \dot{x}_1(t) &= -ax_1(t) - f_1(x_2)x_1(t) + \xi_1(t), \\ \dot{e}_2(t) &= -ae_2(t) + bf_1(x_2)x_1(t) + \\ &\beta x_3(T_{c0} - T_d(t) - e_2) + \xi_2(t), \\ \dot{x}_3(t) &= a_{21}(gx_4(t) - mx_3^2(t)), \\ \dot{x}_4(t) &= a_{32}(u_2(t) - gx_3(t) - a_{31}x_4(t)), \end{aligned} \tag{3}$$

где $x_1(t) = C_A(t)$, $x_2(t) = T(t)$, $e_2(t) = x_2(t) - T_d(t)$, $\xi_1(t) = aC_{Af}(t)$, $\xi_2(t) = a(T_f - T_d) - \dot{T}_d$, $f_1(x_2) = k_0 e^{-\gamma/x_2(t)}$, $a = \frac{q}{V}$, $\gamma = \frac{E}{R}$, $b = \frac{\Delta H}{C_p \rho}$, and $\beta = \frac{UA}{\rho V C_p}$.

For the control of (3), let the output variable $y(t) = e_2(t)$ and the armature current x_4 be measured.

The disturbance $\xi_1(t)$ depends on the input reagent concentration $C_{Af}(t)$ and is difficult to measure in real-time. On the contrary, the input reagent temperature $T_f(t)$ is rather easy to measure. In the sequel, assume that the disturbance $\xi_1(t)$ is unmeasured, whereas the signal $\xi_2(t)$ is measured.

In view of the technological process features, the system variables should satisfy the constraints

$$\begin{aligned} &x_i \in [0, X_i],\ i = 2, 3;\ |x_4| \le X_4; \\ &|u_2| \le U,\ X_i, U = \text{const} > 0. \end{aligned} \tag{4}$$

The problem of considering state-space and control constraints is currently underinvestigated in control theory. This paper proposes to take them into account in the plant's mathematical model by introducing a change of variables—a linear sat function. For details, see Section 3.

Control should maintain given values for the output product's temperature and concentration. In this case, the voltage u_2 at the DC motor armature is the control action in the temperature loop, and the reagent flow rate q is the control action in the concentration loop. In what follows, we will design the temperature control loop in the reactor, assuming the value $a = \frac{q}{V}$ to be known. The next section deals with estimating the state vector and disturbances in the system (3).

2. DESIGNING AN OBSERVER OF THE STATE VECTOR AND DISTURBANCES

Consider the problem of estimating the state vector and disturbances in the system (3) from the measured output variables e_2 and x_4 under the assumption that the signals q and ξ_2 are measured.

The observability of the system (3) in the output variables e_2 and x_4 is established based on the following considerations:

- The DC motor model (the last two equations in (3)) does not depend on the variables of the reactor model (the first two equations in (3)), and its observability in the variable x_4 is obvious due to the block observability form (BOF) [28]. The variable x_3 is estimated by an observer designed using the DC motor model.
- Since the variables e_2, x_3, and ξ_2 are measured (see below), the reactor model structurally coincides with the block observability form considering the disturbances [21]. Hence, it is also observable [21].

Note that the desired observer will be constructed designed in the sequence indicated above. Now we formulate some theoretical results on systems with deep feedback, which will be used in the further presentation.

Lemma. *Consider a first-order system of the form*

$$\dot{\varepsilon}(t) = u + \xi(t), \tag{5}$$

where $\varepsilon(t), u(t), \xi(t) \in R$ *denote the state variable, control, and disturbance, respectively, such that* $|\xi(t)| \le E = \text{const}$ *and* $|\dot{\xi}(t)| \le \bar{E} = \text{const}, \forall t \ge 0$.

Then there exists a control law $u(t) = -\alpha\varepsilon(t)$, $\alpha = \text{const} > 0$, *such that the relations:*

1) $|\varepsilon(t)| \le \Delta_0$,
2) $|\dot{\varepsilon}(t)| < \bar{\Delta}_0$,
3) $|\alpha\varepsilon - \xi| < \bar{\Delta}_0$

hold for any given constants $\Delta_0, \bar{\Delta}_0, \alpha, \bar{\alpha} = \text{const} > 0$ *in a finite time* $t_0 > 0$.

Proof

1) The convergence of the state variable ε of the system (5) to a given neighborhood of zero, $|\varepsilon| \le \Delta_0$, is ensured by choosing a candidate Lyapunov function $V = 0.5\varepsilon^2$. We define the gain as $\alpha_0 = \frac{E}{\Delta_0}$.

Then the condition $\dot{V} = \varepsilon\dot{\varepsilon} = \varepsilon(-\alpha_0\varepsilon + \xi) \le |\varepsilon|(-\alpha_0|\varepsilon| + E) < 0 \Rightarrow -\alpha_0|\varepsilon| + E < 0$ holds out of the domain

$|\varepsilon| \le \frac{E}{\alpha_0} = \Delta_0$. In the case $|\varepsilon(0)| \le \Delta_0$, the variable does not leave the given neighborhood $|\varepsilon(t)| \le \Delta_0,\ \forall t > 0$; in the case $|\varepsilon(0)| > \Delta_0$, it can approach this neighborhood from outside without limit. Hence, choosing $\alpha > \alpha_0$ ensures the convergence of the state variable to the given neighborhood $|\varepsilon(t)| \le \Delta_0$ in a finite time. Really, since $\Delta = \frac{E}{\alpha} < \Delta_0 = \frac{E}{\alpha_0}$, the variable will stay in the new neighborhood $|\varepsilon(t)| \le \Delta,\ \forall t > 0$, if $|\varepsilon(0)| \le \Delta$, or approach the new neighborhood without limit if $|\varepsilon(0)| > \Delta$, reaching the neighborhood $\Delta_0 \le \frac{E}{\alpha_0}$ in a finite time. The solution of (5) satisfies the bound $|\varepsilon(t)| \le \left|\varepsilon(0)e^{-\alpha t}\right| + \left|e^{-\alpha t}\int_0^t e^{\alpha\tau}\xi\ (\tau)\ d\tau\right| \le |\varepsilon(0)|e^{-\alpha t} + \frac{E}{\alpha}(1-e^{-\alpha t})$, and the time t_0 of reaching the domain $|\varepsilon| \le \Delta_0,\ \forall t \ge t_0$, can be estimated as: $(|\varepsilon(0)| - \Delta)e^{-\alpha t} + \Delta = \Delta_0 \Rightarrow t_0 = \frac{1}{\alpha}\ln\left(\frac{|\varepsilon(0)| - \Delta}{\Delta_0 - \Delta}\right),\ |\varepsilon(0)| > \Delta_0$.

2) A similar result applies to the derivative of the state variable in the system $\ddot{\varepsilon} = -\alpha\dot{\varepsilon} + \dot{\xi}$ obtained by differentiating both sides of equation (5). Let $|\dot{\varepsilon}| \le \bar{\Delta}_0 = \frac{\bar{E}}{\bar{\alpha}_0}$ be a given neighborhood of zero. We choose $\alpha > \bar{\alpha}_0$ and denote $\bar{\Delta} = \frac{\bar{E}}{\alpha} < \bar{\Delta}_0 = \frac{\bar{E}}{\bar{\alpha}_0}$. Then the relation $|\dot{\varepsilon}| \le \bar{\Delta}_0,\ t \ge \bar{t}_0$, where $\bar{t}_0 = \frac{1}{\bar{\Delta}_0}\ln\left(\frac{|\dot{\varepsilon}(0)| - \bar{\Delta}}{\bar{\Delta}_0 - \bar{\Delta}}\right)$ and $|\dot{\varepsilon}(0)| > \bar{\Delta}_0$, will hold in a finite time.

3) Due to equality (5) and item 2), we have $|\alpha\varepsilon(t) - \xi(t)| \le \bar{\Delta}_0,\ \forall t \ge \bar{t}_0$. Hence, the disturbance can be estimated with a given accuracy: $\alpha\varepsilon(t) = \xi(t) + \delta(t)$, $\delta(t) \le \bar{\Delta}_0,\ \forall t \ge \bar{t}_0$. Note that for constant disturbances, the system $\ddot{\varepsilon} = -\alpha\dot{\varepsilon} + \dot{\xi},\ \dot{\xi} = 0$, has the solution $\ddot{\varepsilon} = \dot{\varepsilon}(0)e^{-\alpha t}$. Therefore, the disturbance estimate is asymptotically convergent: $\alpha\varepsilon \to \xi,\ t \to \infty$.

Choosing the parameter α based on the condition $\alpha \ge \max\{E/\Delta_0\ , \bar{E}/\bar{\Delta}_0\}$ ensures the desired convergence of the variable $\varepsilon(t)$ and its derivative $\dot{\varepsilon}(t)$ to the given domains: 1) $|\varepsilon(t)| \le \Delta_0$, 2) $|\dot{\varepsilon}(t)| \le \bar{\Delta}_0$, 3) $|\alpha\varepsilon(t) - \xi(t)| \le \bar{\Delta}_0$, $\forall t \ge \max\{t_0, \bar{t}_0\}$. The proof of this lemma is complete. ♦

For estimating the state vector of the system (3), we design an observer of the form

$$\begin{aligned} \dot{z}_1 &= -[a + f_1(x_2)]z_1 + \nu_1, \\ \dot{z}_2 &= -az_2 + \beta z_3(T_{c0} - T_d - e_2) + \xi_2 + \nu_2, \\ \dot{z}_3 &= a_{21}gx_4 + \nu_3, \\ \dot{z}_4 &= a_{32}(u_2 - a_{31}z_4) + \nu_4, \end{aligned} \tag{6}$$

where ν_i are the observer corrections determined below.

Using formulas (3) and (6), we rewrite the system in the residues $\varepsilon_i = x_i - z_i,\ i = 1, 3, 4$, and $\varepsilon_2 = e_2 - z_2$:

$$\begin{aligned} \dot{\varepsilon}_1 &= -(a + f_1(x_2))\varepsilon_1 + \xi_1(t) - \nu_1, \\ \dot{\varepsilon}_2 &= -a\varepsilon_2 + \beta(T_{c0} - T_d - e_2)\varepsilon_3 + bf_1(x_2)x_1(t) - \nu_2, \\ \dot{\varepsilon}_3 &= -a_{21}mx_3^2 - \nu_3, \\ \dot{\varepsilon}_4 &= -a_{32}a_{31}\varepsilon_4 - a_{32}gx_3 - \nu_4. \end{aligned} \tag{7}$$

Within the cascade approach [28], a design procedure for the observer (7) of the state vector and disturbances includes the following steps:

1. We choose an appropriate correction for the last subsystem of the system (7), i.e., the discontinuous function $\nu_4 = l_4\text{sign}(\varepsilon_4)$, where $l_4 = \text{const} > |a_{32}gx_3|$, to ensure the occurrence of a sliding mode on the line $\varepsilon_4 = 0$. The average (equivalent) value of the discontinuous signal is $\nu_{4eq} = -a_{32}gx_3$. In practice, the equivalent value of the discontinuous control can be obtained using the first-order filter $\mu\dot{\tau} = -\tau + \nu_4$, where $\mu > 0$ and $\nu_{4eq} \approx \tau$ [28].

2. Using the equivalent value and $x_3 = -\nu_{4eq}/(a_{32}g)$, we construct the correction $\nu_3 = -a_{21}m \times (\nu_{4eq}/(a_{32}g))^2 + l_3[-\nu_{4eq}/(a_{32}g) - z_3]$ for the third subsystem of the system (7). As a result, the third subsystem of (7) takes the form $\dot{\varepsilon}_3 = -l_3\varepsilon_3$, and the variable ε_3 asymptotically vanishes with an appropriately assigned coefficient $l_3 > 0$: $\varepsilon_3 \to 0 \Rightarrow z_3 \to x_3$.

3. Under the assumption $\varepsilon_3 \to 0 \Rightarrow z_3 \to x_3$, the second equation of the system (7) takes the form

$$\dot{\varepsilon}_2 = -a\varepsilon_2 + bf_1(x_2)x_1 - \nu_2.$$

We choose the correction $\nu_2 = l_2\text{sign}(\varepsilon_2)$, $l_2 = \text{const} > bf_1(x_2)x_1$, to ensure the occurrence of a sliding mode on the plane $\varepsilon_2 = 0$. The average value of the discontinuous signal is $\nu_{2eq} = bf_1(x_2)x_1$.

4. In the last step, we choose $\nu_1 = (-f_1(x_2) + l_1)\left(\frac{\nu_{2eq}}{bf_1(x_2)} - z_1\right) = (-f_1(x_2) + l_1)\varepsilon_1$.

Then the first equation of (7) takes the form $\dot{\varepsilon}_1 = -(a+l_1)\varepsilon_1 + \xi_1$. Under the assumptions $|\xi_1(t)| \le E_1$ and $|\dot{\xi}_1(t)| \le \bar{E}$, the relations $|\varepsilon_1| \le \frac{E_1}{a+l_1} = \Delta_1$ and $|\dot{\varepsilon}_1| \le \frac{\bar{E}_1}{a+l_1} = \bar{\Delta}_1$ hold by the lemma. Thus, assigning an appropriate coefficient $l_1 > 0$, we ensure the desired stabilization accuracy $|\varepsilon_1| \le \Delta_1$ in a finite time, thereby estimating the variable $x_1(t) = z_1(t) + \delta_1(t)$ and the disturbance $\xi_1(t) = (a+l_1)\varepsilon_1(t) + \bar{\delta}_1(t)$ with the desired accuracy: $|\delta_1(t)| \le \Delta_1$ and $|\bar{\delta}_1(t)| \le \bar{\Delta}_1$. Note that $\delta_1,\ \bar{\delta}_1 \to 0$ as $l_1 \to \infty$. (The estimation accuracy grows infinitely in the case of large gains.)

Thus, the state vector of the system (3) and the disturbance $\xi_1(t)$ have been estimated.

This design procedure for an observer of the state vector and disturbances involves discontinuous and continuous corrections. Clearly, the design based on sliding modes is easier from a computational viewpoint. However, the average (equivalent) values of corrections are produced by first-order filters [28], which dynamically extends the state space of the original model. The technique for linearizing the residual equations (7) with continuous corrections (Step 4 of the observer design procedure) can be used in all other steps by analogy. In this case, the cascade approach [28] allows decomposing the observer design procedure into one-dimensional subproblems successively solved with a predetermined accuracy of the resulting estimates without extending the state space of the closed loop system.

The next section considers a control law design procedure for the system (7) under complete information (the available estimates of the state vector and disturbances).

3. FEEDBACK LAW DESIGN

Using the block approach, we present the general solution under complete information about the state vector and disturbances provided by the observer (6). In the case of complete information and the state-space and control constraints (4), the feedback law design procedure based on the block approach [21] includes the following steps.

Step 1. We rewrite the second equation of the system (3) as

$$\dot{e}_2 = -ae_2(t) + \beta d(t)x_3 + \bar{\xi}_2(t), \qquad (8)$$

where $\bar{\xi}_2 = bf_1(x_2)x_1 + \xi_2$ and $d(t) = T_{c0} - T_d(t) - e_2(t) < 0$, $|d| \le D = \text{const}$.

In the system (8), the control action is the motor shaft rotation frequency x_3. Under pump loading, this frequency is positive. We introduce the change of variable

$$e_3 = \beta d(t)x_3 + M_3 \text{sat}^+(s_3), \qquad (9)$$

where $s_3 = k_2 e_2 + \bar{\xi}_2$. Stabilizing the variable $e_3 = \beta d(t)x_3 + M_3 \text{sat}^+(s_3) = 0$, we consider the constraint (4) on the variable $x_3 \in [0, X_3]$ by choosing $0 < M_3 < \beta D X_3$.

Definition. Let $M = \text{const} > 0$ and $b = \text{const}$. Then $M\text{sat}(s) = \min(M, |s|)\ \text{sign}(s)$ and $M\text{sat}^+(s) = M\text{sat}(s)0.5[1 + \text{sign}(s)]$.

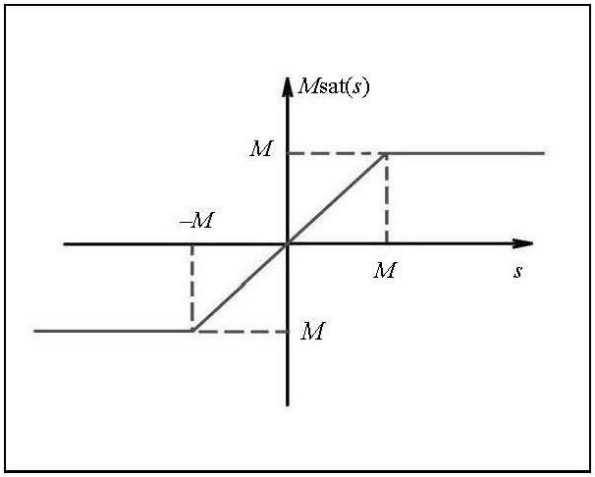


Fig. 2. The graph of *M*sat(*t*).

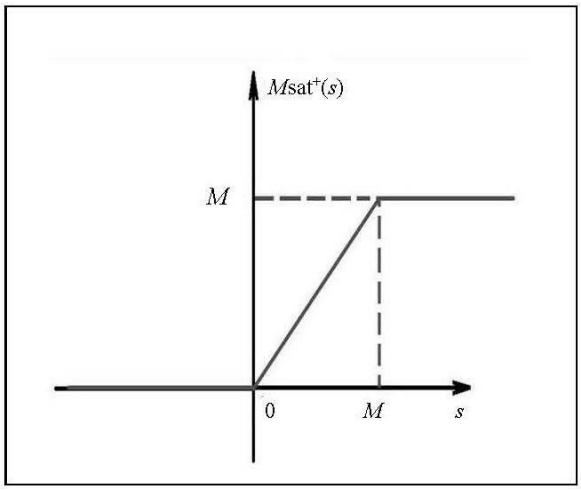


Fig. 3. The graph of *M*sat⁺(*t*).

Substituting the sum (9) into formula (8), we transform the second equation to

$$\dot{e}_2 = -ae_2 + e_3 - M_3\text{sat}^+(s_3) + \bar{\xi}_2. \qquad (10)$$

If the variable s_3 falls into the linear zone $0 < s_3 < M_3$, equation (10) reduces to

$$\dot{e}_2 = -(a + k_2)e_2 + e_3,$$

where the parameter $k_2 > 0$ determines the convergence of e_2 to a given neighborhood of zero. According to the lemma, this neighborhood is described by $|e_2| \le \Delta_3 / (a + k_2)$ under the condition $|e_3| \le \Delta_3 = \text{const}$.

Falling into the linear zone depends on the amplitude M_3 due to the relation $s_3\dot{s}_3 < 0$ holding in the nonlinear zone $s_3 \notin (0, M_3)$; see Fig. 3. We write the equation for the variable s_3: $\dot{s}_3 = -(as_3 - \bar{\xi}_2) + \dot{\bar{\xi}}_2 + k_2[-M_3\text{sat}^+(s_3) + e_3 + \bar{\xi}_2]$.

The parameters M_3 and k_2 of the function $M_3\text{sat}^+(s_3)$ are assigned from the following considerations.

- For $s_3 \in [M_3, \infty]$ (which implies $-M_3\text{sat}^+(s_3) = -M_3$), the inequality $\dot{s}_3 = -a(s_3 - \bar{\xi}_2) + \dot{\bar{\xi}}_2 + k_2[-M_3 + e_3 + \bar{\xi}_2] < 0$ holds. Hence, the amplitude M_3 should satisfy the constraint $M_3 > \frac{1}{k_2}[a\bar{\xi}_2 + \dot{\bar{\xi}}_2] + e_3 + \bar{\xi}_2$.

- For $s_3 \in (-\infty, 0)$ (which implies $M_3\text{sat}^+(s_3) = 0$ and $M_3\text{sat}^+(s_3) = 0$), the inequality $\dot{s}_3 = -a(s_3 - \bar{\xi}_2) + \dot{\bar{\xi}}_2 + k_2[e_3 + \bar{\xi}_2] > 0$ holds. Hence, under the conditions $e_3 + \bar{\xi}_2 > 0$ and $|e_3| \le \Delta_3 = \text{const}$, the coefficient k_2 should satisfy the constraint $k_2 > \frac{|a\bar{\xi}_2 + \dot{\bar{\xi}}_2|}{e_3 + \bar{\xi}_2}$. Note that due to $bf_1(x_2)x_1 + a(T_f(t) - T_d) > 0$, the requirement $\bar{\xi}_2 + e_3 = bf_1(x_2)x_1 + a(T_f(t) - T_d) - \dot{T}_d + e_3 > 0$ restricts the desired stabilization accuracy $|e_3| \le \Delta_3$ (see the next step) and the rate of change of the given neighborhood: $\Delta_3 < bf_1(x_2)x_1 + a(T_f(t) - T_d(t))$ and $|\dot{T}_d(t)| < bf_1(x_2)x_1 + a(T_f(t) - T_d(t)) - \Delta_3$. In the physical sense, the latter inequality limits the rate of change of the given neighborhood to an acceptable level when the tracking problem becomes solvable.

Step 2. We ensure that the variable e_3 from equation (10) falls into the neighborhood of zero: $|e_3| \le \Delta_3$. According to (9), the dynamics of the variable e_3 are described by

$$\dot{e}_3 = \beta d(t)a_{21}gx_4 + \xi_3. \qquad (11)$$

Treating the variable x_4 in the system (11) as a fictitious control action, we make it equal to

$$e_4 = \beta d(t)a_{21}gx_4 + M_4\text{sat}(s_4), \qquad (12)$$

where $M_4\text{sat}(s_4) = \min(M_4, |s_4|)\text{sign}(s_4)$ (Fig. 2) and $s_4 = k_3e_3 + \xi_3$. Stabilizing the variable $e_4 \to 0$, we consider the constraint $|x_4| \le X_4$ by choosing an amplitude $M_4 < \beta Da_{21}gX_4$.

Equation (11) with the local feedback law (12) takes the form

$$\dot{e}_3 = e_4 - M_4\text{sat}(s_4) + \xi_3.$$

In the linear zone ($|s_4| < M_4 \Rightarrow M_4\text{sat}(s_4) = s_4$), it reduces to

$$\dot{e}_3 = -k_3e_3 + e_4.$$

The amplitude M_4 under which the variable s_4 falls into the linear zone is found using Lyapunov's second method.

We write the derivative of the function s_4 as

$$\dot{s}_4 = k_3[-M_4\text{sat}(s_4) + \varphi_4(.)] + \dot{\xi}_3,$$

where $\dot{\xi}_3 = 2\beta da_{21}mx_3\dot{x}_3 + \frac{d^2}{dt^2}[M_3\text{sat}^+(s_3)]$ and $\varphi_4 = e_4 + \beta da_{21}mx_3^2 + \frac{d}{dt}[M_3\text{sat}^+(s_3)]$.

We choose a candidate Lyapunov function of the form $V = 0.5s_4^2$. Then the requirement $\dot{V} = s_4\dot{s}_4 < 0$ outside the linear zone ($|s_4| \ge M_4$) yields the derivative $\dot{V} = s_4\{k_3[-M_4\text{sign}(s_4) + \varphi_4(.)] + \dot{\xi}_3\} < 0$. Hence, the amplitude should be assigned from the condition $M_4 > \Phi_4 + \bar{E}_3 / k_3$, where $|\dot{\xi}_3| \le \bar{E}_3 = \text{const}$ and $|\varphi_4| \le \Phi_4 = \text{const}$.

Step 3. The last step is to stabilize the variable (12) described by

$$\dot{e}_4 = \beta da_{21}ga_{32}u_2 + \xi_4, \qquad (13)$$

where $\xi_4 = \beta\dot{d}(t)a_{21}gx_4 - \beta d(t)a_{21}ga_{32}(gx_3 + a_{31}x_4) + \frac{d}{dt}[M_4\text{sat}(s_4)]$.

We choose the discontinuous control

$$u_2 = M_2\text{sign}(e_4). \qquad (14)$$

In the system (13), a sliding mode occurs on the plane $e_4 = 0$ in a finite time under the existence condition $M_2 > \left| \frac{\xi_4}{\beta d a_{21} g a_{32}} \right|$.

For clarity, we write the dynamic equation of the system (5) in this sliding mode:

$$\begin{aligned} \dot{x}_1 &= -ax_1 - f_1(\mathrm{T_d})x_1 + \xi_1(t), \\ \dot{e}_2 &= -(a+k_2)e_2 + e_3,\ \dot{e}_3 = -k_3 e_3,\ e_4 = 0. \end{aligned} \tag{15}$$

In the system (15), the variables e_3 and e_2 asymptotically vanish: $e_4 = 0 \Rightarrow e_3 \to 0 \Rightarrow e_2 \to 0$. At the same time, the first subsystem is an equation of zero dynamics. Due to $f_1(T_\mathrm{d}) > 0$, this equation is stable.

Thus, the discontinuous control (14) ensures a sliding motion on the plane $e_4 = 0$ in a finite time, described by the stable system (15) of linear differential equations.

Note that if the subsystem (13) is stabilized using the continuous feedback law

$$\beta d a_{21} g a_{32} u_2 + \xi_4 = -k_4 e_4 \Rightarrow u_2 = -\frac{\xi_4 + k_4 e_4}{\beta d a_{21} g a_{32}},$$

then the last equation of the system (15) will take the form $\dot{e}_4 = -k_4 e_4$. Hence, the closed loop system will be stable. Among additional requirements for implementing this continuous control, we mention information about the variable ξ_4 and a pulse-width modulation device to control the DC motor voltage inverter. For implementing the discontinuous control (14), we need only an upper bound on this variable: $|\xi_4| \le E_4 = \mathrm{const}$.

The first subsystem of (15), treated as an equation of zero dynamics, has a bounded solution: $|x_1| \le \frac{E_1}{a + f_1(T_d)}$ for $|\xi_1(t)| = aC_{Af} \le E_1 = \mathrm{const}$; see the lemma. Particularly for $\xi_1 = E_1 = \mathrm{const}$ and $x_2 = T_d = \mathrm{const}$, the lemma implies $x_1 \to \frac{qC_{Af}}{q + Vf_1(T_d)}$. (Recall that $a = \frac{q}{V}$.) Choosing the parameter q as the control action in the product concentration loop, we have the following limit relations: $q \to \infty \Rightarrow C_A \to C_{Af}$ and $q \to 0 \Rightarrow C_A \to 0$. Therefore, we can maintain the product concentration within a reasonable range $C_A \in [C_{A1}, C_{A2}]$ by tuning the reagent flow rate into the reactor.

Clearly, increasing (decreasing) the reagent flow rate into the reactor, we decrease (increase, respectively) the product concentration.

Consider three sets of parameters in which the values $C_{Af} = 0.9$ and $q = 0.9$ are fixed, whereas the desired temperature varies: 1) $T_d = 350$, 2) $T_d = 380$, 3) $T_d = 400$. According to the relation $x_1 \to \frac{qC_{Af}}{q + Vf_1(T_d)}$, we obtain: 1) $x_1 \to 0.3251$, 2) $x_1 \to 0.3214$, 3) $x_1 \to 0.3192$.

Thus, the product concentration decreases as the temperature in the reactor increases, and this conclusion agrees with the simulation results; see Fig. 10 in Section 4.

4. SIMULATION RESULTS

The effectiveness of the proposed approach was verified by numerical simulations of the CSTR–DC motor system in MATLAB. The parameters for the system (3), observer (6), and control (14) were selected from Table 2.

Table 2

Model parameters

Group of parameters	Parameter values
CSTR parameters	$q = 0.9,\ V = 1,\ \beta = 0.003,\ \gamma = 80, b = 5,$ $k_0 = 2,\ T_{c0} = 300,$ $C_{Af} = 0.9 + 0.005\sin(0.03\pi t),$ $T_f = 395 + 0.01\sin(0.05\pi t).$
DC motor parameters	$a_{21} = 0.8,\ g = 0.7,\ m = 0.0001,$ $a_{31} = 12.5, a_{32} = 2$
Initial conditions and reference	$x_1(0) = 0.3,\ x_2(0) = 400,\ x_3(0) = 100,$ $x_4(0) = 20$
Simulation scenario	$T_d(0) = 350$ for $t \in [0,\ t_1],$ $T_d(t_1) = 380$ for $t \in (t_1, \infty], t_1 = 75$
Observer parameters	$z_i(0) = 0,\ i = \overline{1,4},$ $l_1 = 100,\ l_2 = 3,\ l_3 = 100,\ l_4 = 500$
Controller parameters	$M_2 = 400, M_3 = 50, M_4 = 170,$ $k_2 = 0.5,\ k_3 = 0.1$
Physical constraints on state variables	$C_A \in [0,1],\ T \in [0, 400],\ x_3 \in [0, 100],$ $\lvert x_4 \rvert < 150, u_2 = \pm 400$

The temperature in the reactor jacket, x_2, the armature current x_4, the reagent flow rate q, and its temperature T_f were assumed measurable in the plant. The state observer (6) was constructed to obtain complete information about the state variables of the CSTR–DC motor system and the exogenous disturbances (to estimate the unknown signals with a given accuracy).

The observation errors $\varepsilon_i = x_i - z_i, i = \overline{1,4}$, are presented in Figs. 4–7.

Figure 8 shows the graph of the exogenous disturbance $\xi_1 = 0.9(0.9 + 0.005\sin(0.03\pi t))$ reconstructed using the observer (6). The dashed line corresponds to the real disturbance values and the solid line to the restored ones. Figure 9 shows the graph of temperature variations in the reactor jacket.

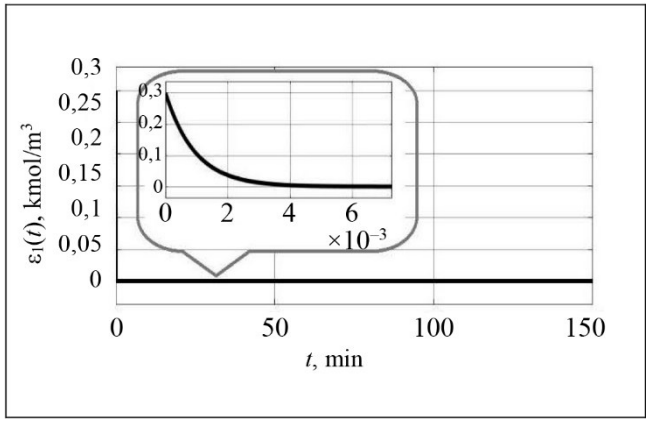


Fig. 4. The observation error $\varepsilon_1(t)$.

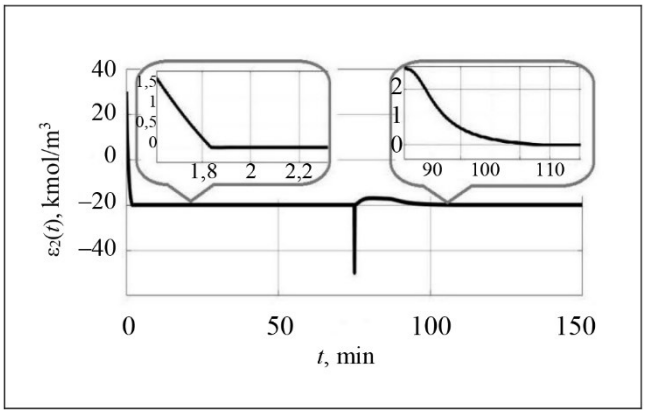


Fig. 5. The observation error $\varepsilon_2(t)$.

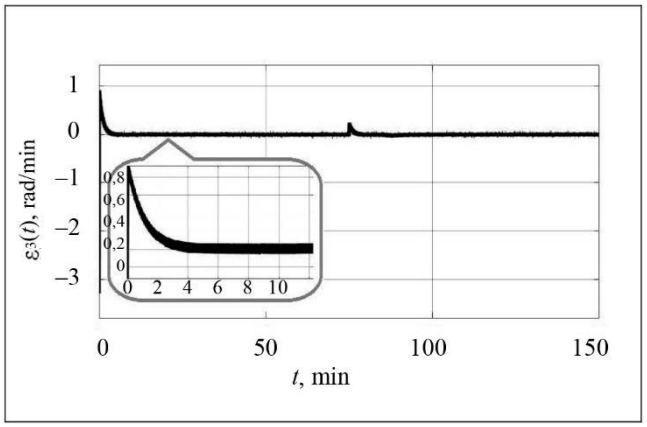


Fig. 6. The observation error $\varepsilon_3(t)$.

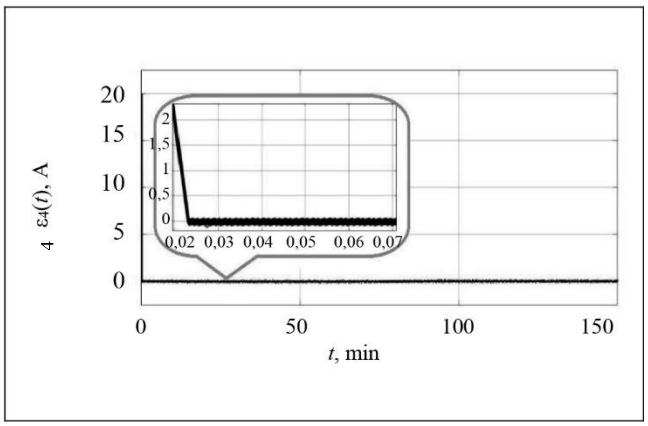


Fig. 7. The observation error $\varepsilon_4(t)$.

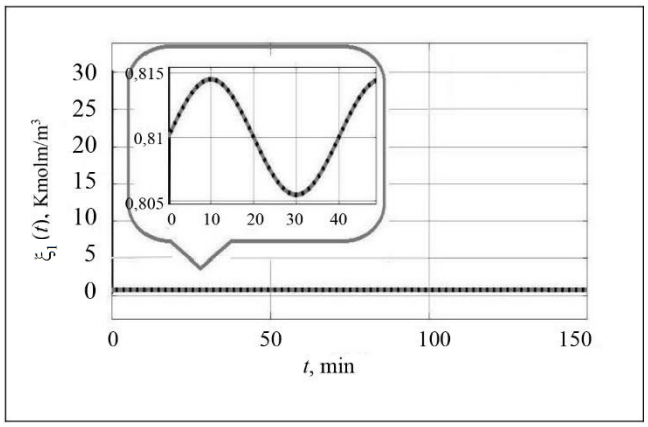


Fig. 8. The exogenous disturbance $\xi_1(t)$.

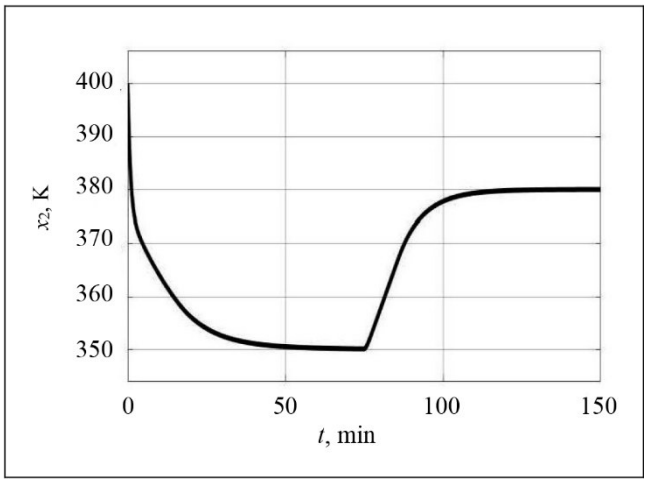


Fig. 9. Temperature in the jacket, $x_2(t)$.

Figures 10–13 show the graphs of the product concentration $x_1(t)$, the motor shaft rotation frequency $x_3(t)$, the armature current $x_4(t)$, and the armature voltage $u_2(t)$, respectively.

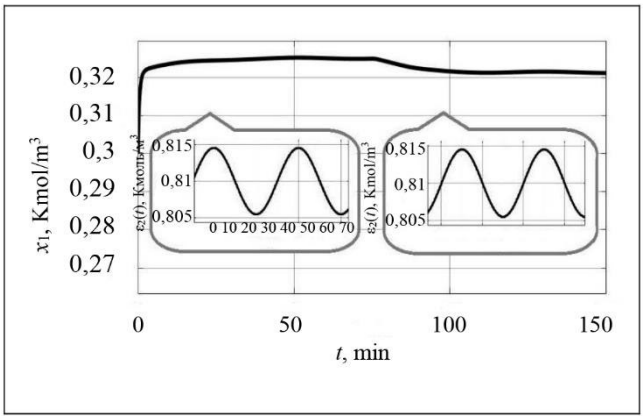


Fig. 10. The product concentration $x_1(t)$.

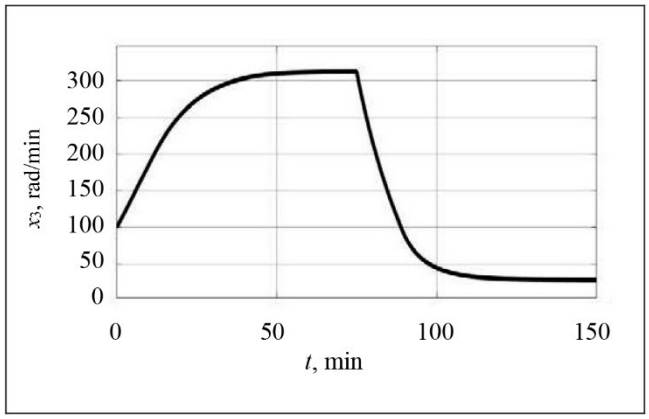


Fig. 11. The motor shaft rotation frequency $x_3(t)$.

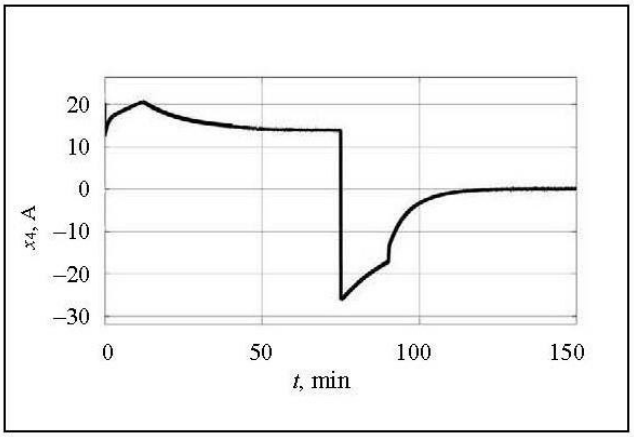


Fig. 12. The armature current $x_4(t)$.

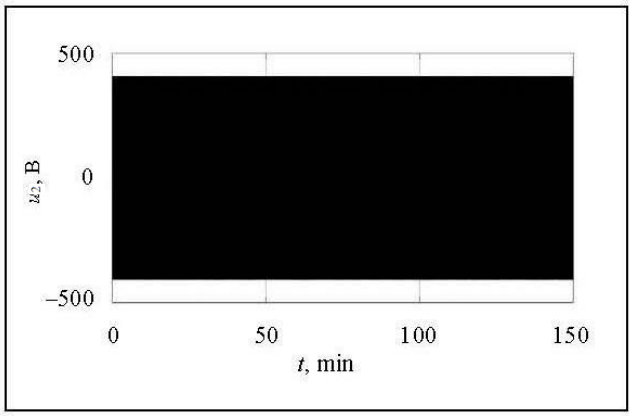


Fig. 13. The armature voltage $u_2(t)$.

CONCLUSIONS

The mathematical model of a continuous stirred tank reactor has been extended by introducing the dynamics of an actuator (DC motor) to apply the theory of sliding modes by (de)activating the switches of the voltage inverter.

An observer with mixed corrections (discontinuous and continuous) has been designed to obtain information about the unmeasured state variables and disturbances.

In the proposed feedback law design procedure, an observer estimates the real signals with a given accuracy both in a real sliding mode and when using deep feedback. A key feature of this work is the feedback law designed within the block approach to consider state-space and control constraints for the CSTR.

The effectiveness of this control algorithm has been validated analytically and illustrated numerically by simulations in MATLAB.

REFERENCES

1. Saravanathamizhan, R., Paranthaman, R., and Balasubramanian, N., Tanks in Series Model for Continuous Stirred Tank Electrochemical Reactor, *Industrial & Engineering Chemistry Research*, 2018, no. 47(9), pp. 2976–2984.
2. Flores-Tlacuahuac, A., and Grossmann, I.E., Simultaneous Cyclic Scheduling and Control of a Multiproduct CSTR, *Industrial & Engineering Chemistry Research*, 2006, no. 45(20), pp. 6698–6712.
3. Li, S., Sun, H., Yang, J., and Yu, X., Continuous Finite-Time Output Regulation for Disturbed Systems under Mismatching Condition, *IEEE Trans. Autom. Control*, 2015, no. 60(1), pp. 277–282.
4. Lagos, B., and Cipriano, A., Performance Evaluation of a Distributed MPC Strategy Applied to the Continuous Stirred Tank Reactor, *IEEE Latin America Transactions*, 2015, no. 13(6), pp. 1921–1926.

5. Delbari, M., Salahshoor, K., and Moshiri, B., Adaptive Generalized Predictive Control and Model Reference Adaptive Control for CSTR Reactor, *Proceedings of the 2010 International Conference on Intelligent Control and Information Processing (ICICIP)*, Dalian, China, 2010, pp. 165–169.
6. Al-Jehani, N.A., Nounou, H.N., and Nounou, M.N., Fuzzy Control of a CSTR Process, *Mechatronics and Its Applications*, 2012, pp. 1–6.
7. Antonelli, R., and Astolfi, A., Continuous Stirred Tank Reactors: Easy to Stabilise, *Automatica*, 2003, no. 39(10), pp. 1817–1827.
8. Graichen, K., Hagenmeyer, V., and Zeitz, M., Design of Adaptive Feedforward Control under Input Constraints for a Benchmark CSTR Based on a BVP Solver, *Computers and Chemical Engineering*, 2009, no. 33(2), pp. 473–483.
9. Chiu, C.S., Derivative and Integral Terminal Sliding Mode Control for a Class of MIMO Nonlinear Systems, *Automatica*, 2012, no. 48(2), pp. 316–326.
10. Zhao, D., Zhu, Q., and Dubbeldam, J., Terminal Sliding Mode Control for Continuous Stirred Tank Reactor, *Chemical Engineering Research and Design*, 2015, no. 94, pp. 266–274.
11. Ma, H., Wu, J., and Xiong, Z., A Novel Exponential Reaching Law of Discrete-Time Sliding-Mode Control, *IEEE Transactions on Industrial Electronics*, 2017, no. 64(5), pp. 3840–3850.
12. Ma, H., Wu, J., and Xiong, Z., Discrete-Time Sliding-Mode Control with Improved Quasi-Sliding-Mode Domain, *IEEE Transactions on Industrial Electronics*, 2016, no. 63(10), pp. 6292–6304.
13. Yan, X.-G., Spurgeon, S.K., and Edwards, C., State and Parameter Estimation for Nonlinear Delay Systems Using Sliding Mode Techniques, *IEEE Transactions on Automatic Control*, 2013, no. 58(4), pp. 1023–1029.
14. Boudjellal, M., and Illoul, R., High-Order Sliding Mode and High-Gain Observers for State Estimation and Fault Reconstruction for a Nonlinear CSTR, *Proceedings of the 6th International Conference on Systems and Control (ICSC)*, Batna, Algeria, 2017, pp. 231–236.
15. Ignaciuk, P., Nonlinear Inventory Control with Discrete Sliding Modes in Systems with Uncertain Delay, *IEEE Transactions on Industrial Informatics*, 2014, no. 10(1), pp. 559–568.
16. Ma, L., Zhao, D, and Spurgeon, S.K., Disturbance Observer Based Discrete Time Sliding Mode Control for a Continuous Stirred Tank Reactor, *Proceedings of the 15th Workshop on Variable Structure Systems*, 2018, pp. 372–377.
17. Chen, W.H., Yang, J., Guo, L., and Li, S., Disturbance-Observer-Based Control and Related Methods: An Overview, *IEEE Transactions on Industrial Electronics*, 2016, no. 63(2), pp. 1083–1095.
18. Rios, H., Efimov, D., Moreno, J.A., and Perruquetti, W., Time-Varying Parameter Identification Algorithms: Finite and Fixed-Time Convergence, *IEEE Transactions on Automatic Control*, 2017, no. 62(7), pp. 3671–3678.
19. Utkin, V.A. and Utkin, A.V., Problem of Tracking in Linear Systems with Parametric Uncertainties under Unstable Zero Dynamics, *Autom. Remote Control*, 2014, vol. 75, no. 9, pp. 1577–1592.
20. Krasnova, S.A. and Utkin, A.V., Sigma Function in Observer Design for States and Perturbations, *Automation and Remote Control*, 2016, no. 77(9), pp. 1676–1688.
21. Krasnova, S.A., Utkin, V.A., and Utkin, A.V., Block Approach to Analysis and Design of the Invariant Nonlinear Tracking Systems, *Autom. Remote Control*, 2017, vol. 78, no. 12, pp. 2120–2140.
22. Han, J.S., Bahn, W., and Kim, T.I., Decoupled Disturbance Compensation under Control Saturation with Discrete-Time Variable Structure Control Method in Industrial Servo Systems, *Proceedings of the 16th International Conference on Control, Automation and Systems*, Gyeongju, Korea, 2016, pp. 1453–1457.
23. Wu, W., Nonlinear Bounded Control of a Nonisothermal CSTR, *Industrial and Engineering Chemistry Research*, 2000, no. 39(10), pp. 3789–3798.
24. Utkin, A.V. and Utkin, V.A., The Synthesis of Stabilization Systems under One-Sided Restrictions on Control Actions, *Control Sciences*, 2020, no. 3, pp. 3–14. (In Russian.)
25. Antipov, A.S. and Krasnova, S.A., Block-Based Synthesis of a Tracking System for a Twin-Rotor Electromechanical System with Constraints on State Variables, *Mech. Solids*, 2021, vol. 56, no. 7, pp. 43–56.
26. Drakunov, S.V., Izosimov, D.B., Luk'yanov, A.G., et al., The Block Control Principle. I, *Automation and Remote Control*, 1990, vol. 51, no. 5, pp. 601–609.
27. Chilikin, M.G., Klyuchev, V.I., and Sandler, A.S., *Teoriya avtomatizirovannogo elektroprivoda* (The Theory of Automatic Electric Drives), Moscow: Energiya, 1979. (In Russian).
28. Krasnova, S.A. and Utkin, V.A., *Kaskadnyi sintez nablyudatelei sostoyaniya dinamicheskikh sistem* (Cascade Design of State Observers for Dynamic Systems), Moscow: Nauka, 2006. (In Russian.)

This paper was recommended for publication by V.N. Afanas'ev, a member of the Editorial Board.

Received May 24, 2021, and revised July 28, 2021.
Accepted August 24, 2021.

Author information

Gulyukina, Svetlana Igorevna. Junior Researcher, Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ gulyukina.s.i@mail.ru

Utkin, Viktor Anatol'evich. Dr. Sci. (Eng.), Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ vicutkin@ipu.ru

Cite this article
Gulyukina, S.I., Utkin, V.A. A Block Approach to CSTR Control under Uncertainty, State-Space and Control Constraints. *Control Sciences* **5,** 43–52 (2021). http://doi.org/10.25728/cs.2021.5.4

Original Russian Text © Gulyukina, S.I., Utkin, V.A., 2021, published in *Problemy Upravleniya*, 2021, no. 5, pp. 48–59.

Translated into English by *Alexander Yu. Mazurov*,
Cand. Sci. (Phys.–Math.),
Trapeznikov Institute of Control Sciences,
Russian Academy of Sciences, Moscow, Russia
✉ alexander.mazurov08@gmail.com

DOI: http://doi.org/10.25728/cs.2021.5.5

ESTIMATING THE INFLUENCE OF ENVIRONMENTAL FACTORS ON MORTALITY IN ELDER AGE GROUPS: AN EXAMPLE OF KRASNOYARSK[1]

O.V. Taseiko[1,2] and D.A. Chernykh[1,2]

[1]Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia,
[2]Siberian Branch, Russian Academy of Sciences, Krasnoyarsk, Russia

✉ taseiko@gmail.com, ✉ dachernykh93@gmail.com

Abstract. This paper analyzes and forecasts the influence of climatic parameters and air quality on the mortality of the Krasnoyarsk industrial agglomeration. The long-term climatic data for the city of Krasnoyarsk are studied. The diseases sensitive to climate change are considered for the period from 2011 to 2014. The relationship between climatic parameters, air pollution, and the number of deaths is established using multivariate statistical analysis. The factors with the greatest contribution to mortality for climate-dependent diseases are identified. The mortality causes associated with negative environmental factors are examined. The age groups most affected by climatic parameters and air pollution are determined. As shown below, the nonlinear Poisson regression model predicts population mortality quite close to the factual data.

Keywords: air quality, climatic parameters, generalized linear model with Poisson regression, population mortality, social-natural-technogenic (S-N-T) system.

INTRODUCTION

The joint influence of anthropogenic activity and natural factors on the population's health increases the sustainable development risks of territorial entities. In addition, the urbanization trends of the environment create new types of technological and natural threats. Due to the economic and social conditions emerging in Russia, the main object of urbanization in the nearest future will be the territories of Siberia and the Arctic. Considering the uniqueness of natural systems and the global significance of these regions for the country's sustainable development, a particularly important scientific problem is studying natural and technological threats and development risks using new-generation information technologies. Sustainable development of territories is associated with a quantitative assessment of complex security and regional management using a risk-based approach [1–4].

The territory of an industrial region is a single complex social-natural-technogenic (S-N-T) system. The elements of this system mutually influence each other and have interconnected types of risks [5]. The operation of the sociosphere as an element of the S-N-T system is associated with individual strategic risks of life and health loss. The most significant factors forming individual strategic risks include environmental pollution (primarily air pollution) and the globally changing climate. Specific meteorological phenomena affecting the health of the population are short-term; nevertheless, their frequency is determined by the climatic characteristics of the territory. Therefore, this formulation involves the terms "meteorological conditions" and "climatic parameters."

A relationship between mortality rates and climatic factors (temperature waves) was established in several studies carried out for the cities of Arkhangelsk, Yakutsk, Astrakhan, Krasnoyarsk, Moscow, and others

[1]The reported study was funded by Russian Foundation for Basic Research, Government of Krasnoyarsk Territory, Krasnoyarsk Regional Fund of Science, project number 19-413-240013 "Risk assessment methodology caused by environmental factors on population health and mortality in industrial agglomerations".

[6–11]. For the city of Krasnoyarsk, the relative risk of mortality from the negative influence of cold and heat waves was assessed using the Poisson regression model; for details, see [7].

According to the 1999–2008 data, periods of intense heat in Arkhangelsk (a temperature threshold of 21°C) were associated with an increase in mortality from cardiovascular diseases and all-natural causes among the over-65s (the people aged 65 and elder) and from all external causes among the over-30s. During that period, 110 additional deaths were recorded due to heat waves and 179 additional deaths due to cold waves [6].

Temperature waves were first identified and studied in 1881–1884 and were described as an area of high or low temperatures of atmospheric air not staying in one place for a long time [12]. Temperature waves are associated with serious environmental and health problems and may even incur significant economic damage. High temperatures can lead to thermal stress and deteriorate air quality, causing adverse health effects, especially for more vulnerable groups of population (including those with cardiovascular or respiratory diseases and the elderly) [13].

Extremely low and high temperatures in winter and summer, respectively, are characteristic of the climate of Central Siberia, Yakutia, and other regions. They are associated with sharply continental weather. With these unique climatic characteristics of industrial agglomerations, there is a need to analyze the synergistic effect of environmental factors on mortality. This problem can be partially solved by developing a methodology for assessing the risks of a multifactorial influence on the population's mortality using statistical regression models and elaborating measures to protect the population from the negative influence of environmental factors. The methodology should consider the structural characteristics of the population and the region.

Methods for assessing public health risks have been developed for three decades. The negative influence of high-level air pollution on the population's morbidity and mortality has been demonstrated by many studies. Children and adolescents are most susceptible to the influence of chemical pollution of the atmospheric air [14–16]. The child's respiratory system is a prime target for air pollutants. They cause a wide range of acute and chronic effects, either as a single risk factor or, most often, in combination with other external agents and (or) characteristics of the child's susceptibility. Age plays an important role during exposure to inhaled pollutants [17–19]. Infants are more susceptible to lung damage from the same-type toxins than adults, even at doses below the latter's optimal value [20, 21].

According to the population data for the city of Krasnoyarsk for the period 2000–2018, the total mortality rate decreased until 2009 (by 17%) and increased until 2018 (by 9%) with a population growth of more than 20%; see Fig. 1. (The data are provided by the Krasnoyarsk Regional Body of the Federal State Statistics Service.)

From 2006 to 2015, mortality from cerebrovascular diseases decreased by 9.9% (Fig. 2). By contrast, the proportion of deaths from ischemic heart disease increased from 18% in 2000 to 30% in 2018.

The Poisson regression model is recommended to identify the relationship between mortality and environmental factors [22]. In foreign epidemiological studies, the application of the Poisson model goes back to the second half of the 20th century. This model was

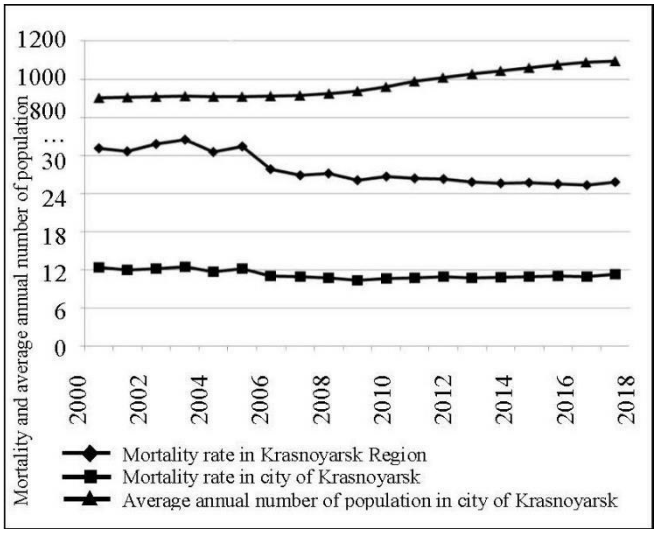


Fig. 1. Dynamics of mortality and average annual number of population: city of Krasnoyarsk and Krasnoyarsk Region.

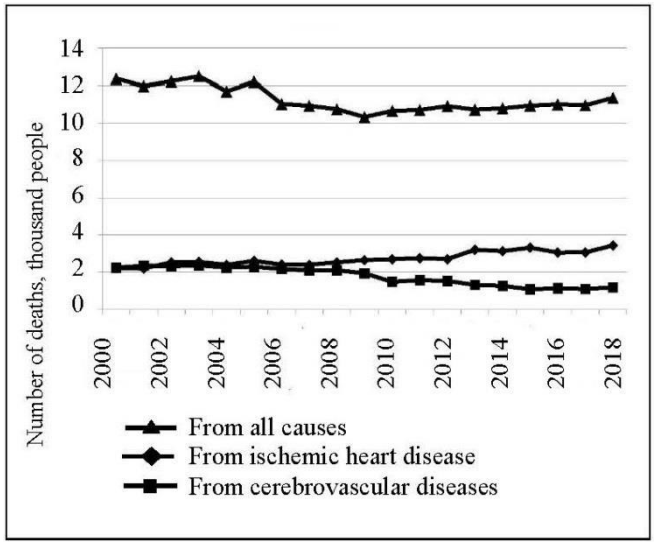


Fig. 2. Dynamics of mortality from ischemic heart and cerebrovascular diseases: city of Krasnoyarsk.

adopted to study the relationship between mortality and different environmental factors: air pollution [23], average temperature [24], and temperature waves [25–27].

Some studies identified the relationship between cancer incidence and polluted air and bad habits [28, 29]. The research where this model was applied to Moscow included three influence factors: the average daily air temperature, ozone concentration, and the concentration of suspended particles PM_{10}. Higher mortality from all causes, except for external ones (accidents, conditions, and circumstances due to an injury, poisoning, and other adverse effects) was discovered under increasing the number of days with abnormally low and high temperatures [30].

This paper assesses the relative risks of the sociosphere of the S-N-T system for the city of Krasnoyarsk. We consider the characteristics of climatic parameters and levels of air pollution as influence factors. We predict the risk of population's premature mortality from the combined influence of climatic and environmental factors. The main problem includes estimating the relationship between climatic parameters, air pollution, and population mortality using multivariate statistical analysis methods. In addition, we analyze the significance of the contribution of each factor mentioned to mortality rates for the main groups of climate-dependent diseases.

1. DATA AND METHODS

We studied mortality rates in the city of Krasnoyarsk from **the initial causes** with the greatest sensitivity to climatic factors, according to the International Classification of Diseases (ICD-10) [22] in two age groups (60–74 years old and the over-75s), divided by sex for the period from 2011 to 2014:

– ischemic heart disease (I20–I25),

– cerebrovascular diseases (I60–I69),

– respiratory diseases (J00–J22, J30, J40–J44, and J45).

The climatic characteristics were determined and estimated using data from state observation networks of meteorological stations [31, 32]. To normalize the time series, we divided them by the standard deviation [33]. Missed values of these factors were restored using the moving average method.

For constructing a nonlinear Poisson regression model describing the dependence of mortality on environmental factors, we adopted the daily mortality data for the city of Krasnoyarsk for the period from 2011 to 2014. (The data are provided by the Krasnoyarsk Regional Body of the Federal State Statistics Service [22].) The model has the form

$$\log(\mu_t) = \beta_0 + \beta_1 X_{1,t-L} + \ldots + \beta_k X_{k,t-L} + \beta_{k+1} Z,$$

$$\mu_t = e^{\beta_0} e^{\beta_1 X_{1,t-L}} \ldots e^{\beta_k X_{k,t-L}} e^{\beta_k Z},$$

where μ_t is the predicted parameter, i.e., the Poisson independent (output) variable; $X_{1,t}, \ldots, X_{k,t}$ are the explanatory (input) variables; Z is the short-term time trend by day of the week; L denotes the negative effect delay (lag); $\beta_1, \ldots, \beta_{k+1}$ are the model parameters; β_0 is the free term.

Epidemiological studies determine the relationship between the concentration of air pollutants in one day and the health effects on some lag days. Lag means time delay: when exposed to a negative phenomenon, its effect may appear only after some period [22]. In this study, we chose a lag of 1 to 15 days.

For interpreting the regression coefficient, we used *the relative risk* (*RR*). This indicator measures the relationship between an independent variable (e.g., the concentration of air pollutants) and the risk of a particular result (e.g., the number of people with respiratory injury). It shows how many times the population's mortality increases due to a negative environmental factor relative to background mortality (mortality of people not exposed to the negative influence). For the Poisson regression, the relative risk is defined as follows [13, 34]:

$$RR = \exp(\beta_i),$$

where β_i denotes the regression coefficient.

Table 1 presents the dynamics of the studied environmental factors from 2000 to 2018. For the city of Krasnoyarsk, the daily concentration data on air pollutants are provided by the Central Siberian Department for Hydrometeorology and Environmental Monitoring.

The pollutants were chosen based on the evidence of their influence on the morbidity and mortality rates of the population; see Table 2. Ozone (O_3) is one of the most dangerous air pollutants. This substance belongs to the first hazard class and represents the main component of photochemical smog. An increased level of ground-level ozone is observed only in hot sunny weather. The inhalation effect of the substance on the body is accompanied by irritation of the respiratory system, a reduction in lung functions, the development of asthma and allergies, and a significant decrease in immunity to infections [22, 35–37]. Ozone concentrations are not measured within the state observation network of the Krasnoyarsk Region. Therefore, we used the concentrations of nitrogen dioxide and formaldehyde as precursors of ozone.

When processing the data and performing a numerical experiment, we analyzed the collinearity or multicollinearity of the factors and correct short-term time trends (by days of the week) by adding another coefficient into the model.

Table 1

Dynamics of environmental factors

Year	Hot waves				Cold waves				Temperature, °C	Relative humidity, %	Extreme temperature drops during the day			Average concentration of pollutants		
	Number of waves	Duration of waves, days	Amplitude of waves, °C		Number of waves	Duration of waves, days	Amplitude of waves, °C				Number of events	Range, °C		Suspended substances, mg/m^3	Nitrogen dioxide (NO_2), mg/m^3	Formaldehyde (F), mg/m^3
			min	max			min	max				min	max			
2000	6	37	7.2	20.2	5	52	–38	–14	0.8	73	49	10	16.6	0.254	0.024	–
2003	4	29	6.4	20.6	3	30	–14	–10	2.5	71	73	10	31	0.265	0.042	0.008
2006	2	24	13.6	20.4	4	49	–24	–10	0.8	71	15	10.1	20	0.198	0.051	0.0043
2009	3	19	8.3	21	9	82	–24	–7.7	0.2	73	15	10.1	22.2	0.173	0.062	0.0094
2012	4	31	8.3	21.1	4	78	–21	–9	0.6	71	14	10.1	30.3	0.228	0.051	0.0185
2015	7	42	7.5	20.8	1	14	–11	–11	3.8	66	16	10	13.7	0.1380	0.0360	0.0120
2018	4	40	7.7	20.7	8	60	–20	–11	1.3	70	26	10.1	19.4	0.1063	0.0379	0.0168

Table 2

The effect of exposure to pollutants on the body [38]

Substance	Hazard class	Critical organ	Critical effect
Suspended substances	3	Respiratory system	– An increase in total mortality, – Mortality from diseases of the cardiovascular system and respiratory system, – Frequent symptoms from the upper and lower respiratory tract, – Consulting a doctor for respiratory diseases, – Frequent exacerbation of bronchial asthma
Ozone (O_3) [22, 35]	1	Respiratory system	– An increase in total mortality, – Irritation of the respiratory system, – Decreased lung function, – Development of asthma and allergies, – A significant decrease in immunity to infections
Nitrogen dioxide (NO_2)	2	Respiratory system, blood and hemopoietic organs	– Increased incidence and duration of diseases of the upper and lower respiratory tract, – An increased number of lower respiratory tract diseases among children,
Formaldehyde (F)	2	Respiratory system, organs of vision, and immune system	– An increased incidence of diseases of the upper and lower respiratory tract, – Inflammatory processes in the lungs, – Diseases of the immune system, including allergic reactions, – Diseases of the organs of vision

2. RESULTS OF THE STUDY

For a particular factor, the level of influence on the mortality rate was determined using the regression coefficient. It shows by how many units the result will change when a given factor changes by one unit [39].

Among females aged 60–74 years, mortality from respiratory diseases was most influenced by temperature waves on the ninth day of exposure (lag equaled 9). The influence of pollutants was represented by the effect of formaldehyde:

$$\ln(\mu_t) = -7.4 + 0.9F_t - 4.9Temp_{t-2} - 3.3Temp_{t-12} + 0.9Diff_{t-1} + 0.9Diff_{t-12} + 1.3Hum_{t-11} + 2.8W_{t-9}, \quad (1)$$

where F is the concentration of formaldehyde in the atmospheric air, mg/m^3; *Temp* denotes temperature, °C; *Diff* is the temperature drop, °C; *Hum* indicates the relative humidity, %; W gives the number of temperature waves.

Since the free term β_0 has a small value compared to the total value of the factor weights, the factors not included in this model are insignificant. According to the expression (1), the same influence factors can negatively affect different lags. The influence of the average daily temperature manifested on the 2nd and 12th days after the inhalation exposure; the temperature drop during the day, on the 1st and 12th days.

Figure 3 shows the distribution of factual mortality and mortality calculated using the Poisson model. Clearly, for an example of mortality from respiratory diseases among females aged 60–74, the calculated mortality rate describes well the factual one with a correlation coefficient of 0.8.

We processed the available statistical dataset for 2011 to 2014 for two age groups (from 60 to 74 years old and over 75 years old) to obtain 12 mortality prediction models for three main immediate causes belonging to the climate-dependent category. The relative risk of mortality from exposure to environmental factors, the lags, and the correlation indices for each model are presented in Table 3.

3. DISCUSSION OF THE RESULTS

The analysis of the population mortality from the negative influence of environmental factors using the Poisson regression model has yielded the following outcomes.

- According to the lag distribution, for males, most of the negative effects from ischemic heart disease and cerebrovascular diseases manifest within six days after exposure to the combination of factors; from respiratory diseases, after the 7th day. In contrast, for females, negative health effects accumulate more smoothly within two weeks; see Fig. 4.

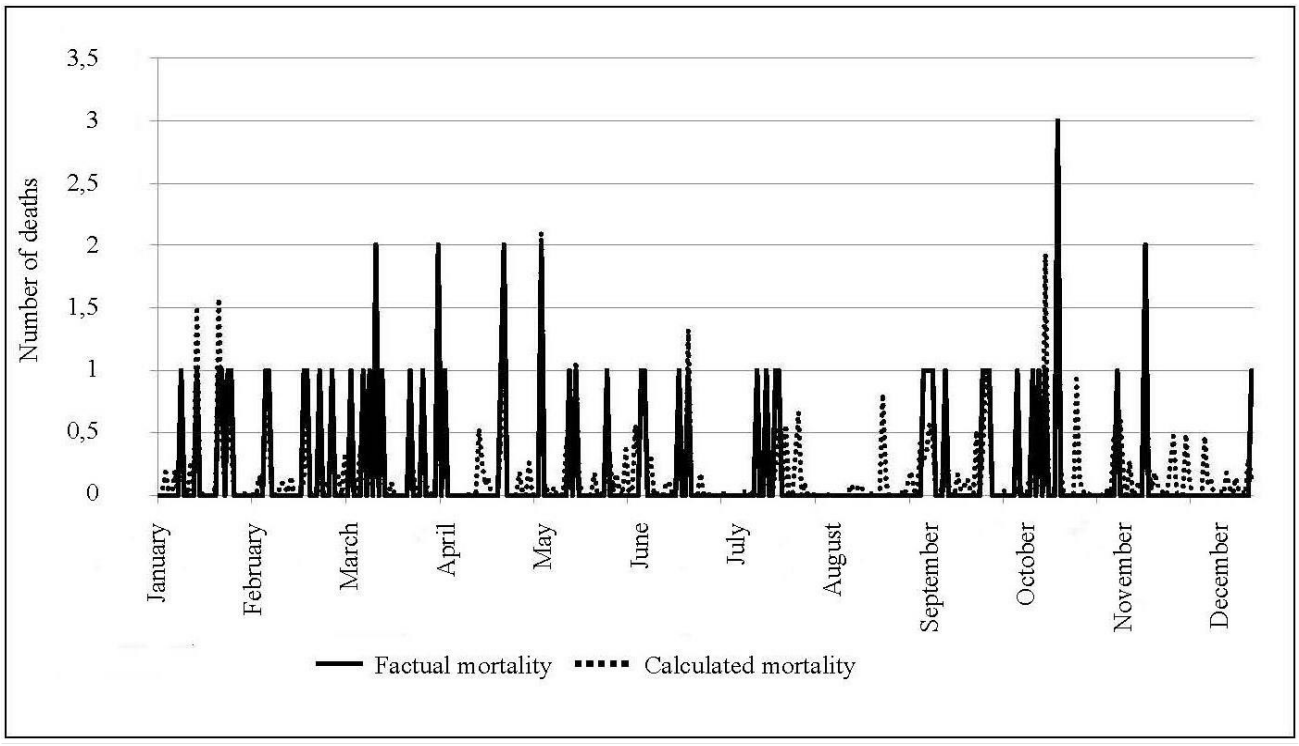


Fig. 3. Factual and calculated mortality from respiratory diseases in 2014: females aged 60–74 years.

Table 3

The relative risk of population mortality

Influence factors	Sex	Ischemic heart disease	Cerebrovascular diseases	Respiratory system diseases	Ischemic heart disease	Cerebrovascular diseases	Respiratory system diseases
		Age group					
		60–74 years			75 years and elder		
		Value of correlation coefficient for factual and calculated mortality					
	F	0.69	0.66	0.8	0.63	0.53	0.66
	M	0.6	0.63	0.6	0.55	0.64	0.66
Relative risk of mortality							
Nitrogen dioxide	F	1.2 / L = 5	1.3 / L = 15	–	–	–	1.3
	M	1.1 / L = 5	1.2 / L = 13	1.4	–	–	1.4 / L = 14
Formaldehyde	F	1.2 / L = 9	–	2.5	–	1.2	1.4 / L = 11
	M	–	1.5 / L = 4	–	0.7 / L = 14	0.7 / L = 2	1.6 / L = 3
Suspended substances	F	1.2 / L = 5	1.3 / L = 8	–	–	0.8	–
	M	1.2 / L = 15	1.3 / L = 5	–	1.2 / L = 3	–	1.4 / L = 8
Temperature	F	0.4 / L = 11	–	0.04 / L = 12	0.7 / L = 4	–	3.7 / L = 5
	M	0.6 / L = 15	–	–	–	0.4 / L = 12	3.3 / L = 10
Relative humidity	F	1.3 / L = 6	–	3.6 / L = 11	0.8 / L = 11		1.6 / L = 2
	M	1.1	–	–	1.2 / L = 5	1.5 / L = 4	–
Extreme temperature drops	F	–	–	2.5 / L = 12	–	–	–
	M	–	–	1.4 / L = 9	–	0.8 / L = 14	1.4 / L = 1
Temperature waves	F	1.8 / L = 2	–	16.4 / L = 9	1.5 / L = 13	–	0.5 / L = 2
	M	–	0.4 / L = 10	–	–	–	–

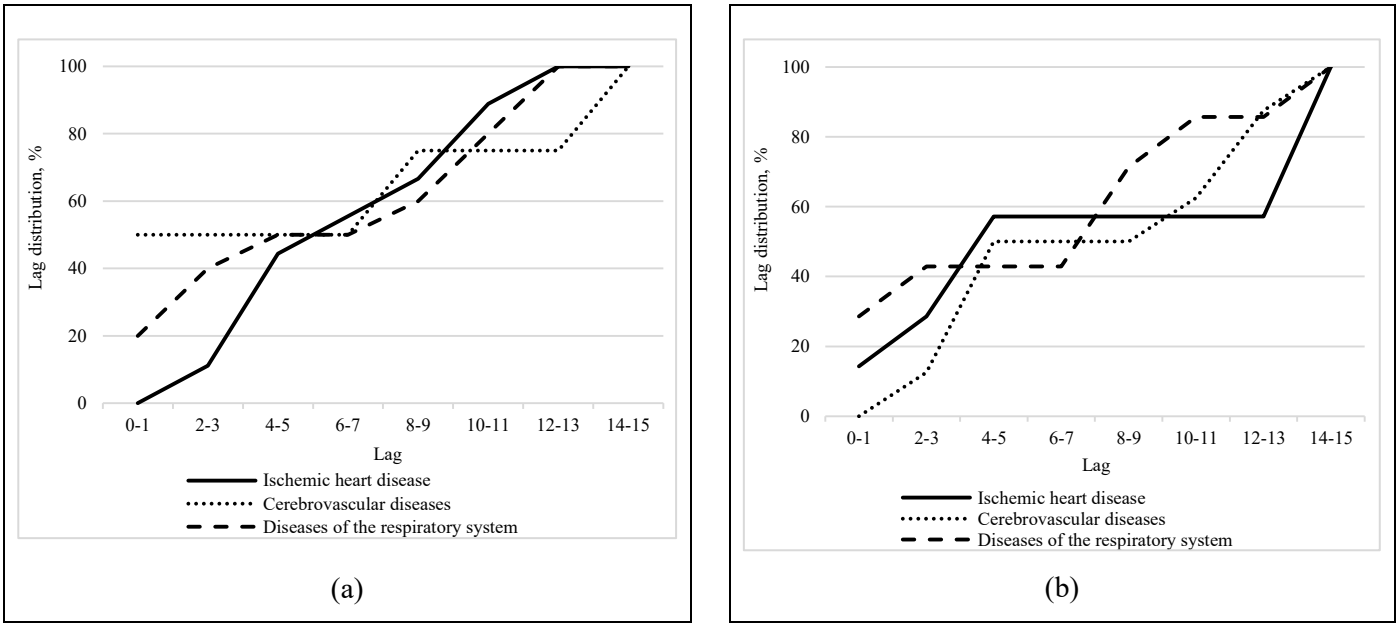


Fig. 4. Lag distribution of the negative effect: (a) females, (b) males.

- The climatic factors have the greatest contribution to population mortality in all age and sex groups: the ratio of influence is 58.1% (climatic factors) to 41.9% (polluted atmospheric air). In addition, females are most exposed to the negative effects of environmental factors.
- Among the pollutants considered, formaldehyde has the greatest relative contribution to mortality. Among the climatic factors considered, the greatest harm is caused by temperature waves.
- The highest correlation coefficients between the predicted and factual mortality rates correspond to mortality from respiratory diseases.
- Among the age groups considered, the people aged 60–74 years are most sensitive to the effect of all factors.

CONCLUSIONS

As demonstrated in the paper, the nonlinear Poisson regression model predicts population mortality from environmental factors quite close to the factual data. Hence, this approach can be used to estimate the combined influence of climatic parameters and air quality on mortality. The approach to risk assessment will be further developed by excluding insignificant factors and considering in detail the influence of significant factors by groups, depending on the direction of the identified relationships.

REFERENCES

1. *Bezopasnost' Rossii. Pravovye, sotsial'no-ekonomicheskie i nauchno-tekhnicheskie aspekty. Regional'nye problemy bezopasnosti. Krasnoyarskii krai* (Security of Russia. Legal, Socio-economic, and Scientific and Technical Aspects. Regional Security Problems. The Krasnoyarsk Region), Shabanov, V.F., Ed., Moscow: Znanie, 2001. (In Russian.)
2. Vinokurov, Yu.I., Lepikhin, A.M., Moskvichev, V.V., et al., *Geoinformatsionnye tekhnologii i matematicheskie modeli dlya monitoringa i upravleniya ekologicheskimi i sotsial'no-ekonomicheskimi sistemami* (Geoinformation Technologies and Mathematical Models for Monitoring and Management of Environmental and Socio-economic Systems), Shokin, Yu.I., Ed., Barnaul: Pyat' Plyus, 2011.(In Russian.)
3. Levkevich, V.E., Lepikhin, A.M., Moskvichev, V.V., et al., *Bezopasnost' i riski ustoichivogo razvitiya territorii* (Security and Risks of Sustainable Development of Territories), Krasnoyarsk: Siberian Federal University, 2014. (In Russian.)
4. Akimov, V.A., Bykov, A.A., Shevchenko, A.V., and Strelko, S.V., Development of Scientific and Methodological Foundations of Public Administration Using Indicators of Strategic Risks, Including the Risks of Emergency Situations (on an Example of the Ulyanovsk Region), *Strateg. Grazhdansk. Zashchity: Probl. Issledovan.*, 2013, vol. 3, no. 2, pp. 736–780. (In Russian.)
5. Moskvichev, V.V., Bychkov, I.V., Potapov, V.P., et al., An Information System for the Territorial Management of Development Risks and Security, *Vestn. Ross. Akad. Nauk*, 2017, vol. 87, no. 8, pp. 696–705. (In Russian.)
6. URL: https://www.who.int/publications/i/item/protecting-health-from-climate-change-a-seven-country-initiative.
7. Revich, B.A. and Shaposhnikov, D.A., Features of the Impact of Cold and Heat Waves on Mortality in Cities with a Sharply Continental Climate, *Siberian Medical Review*, 2017, no. 2, pp. 84–90. (In Russian.)
8. Revich, B.A. and Shaposhnikov, D.A., Cold Waves in Southern Cities of European Russia and Premature Mortality, *Studies on Russian Economic Development*, 2016, vol. 27, no. 2, pp. 210–215. DOI:10.1134/S107570071602012X.
9. Revich, B.A., Khar'kova, T.L., and Podol'naya, M.A., Mortality Dynamics and Life Expectancy of Population of Arctic/Subarctic Region of the Russian Federation in 1999–2014, *Ekologiya Cheloveka (Human Ecology)*, 2017, no. 9, pp. 48–58. (In Russian.)
10. Shaposhnikov, D.A. and Revich, B.A., On Some Approaches to Calculation of Health Risks Caused by Temperature Waves, *Health Risk Analysis*, 2018, no. 1, pp. 22–31. (In Russian.)
11. Kvasha, E.A., Revich, B.A., and Khar'kova, T.L., Similarities and Differences of Mortality in Four Russian Megalopolises, *Byull. Natsion. Nauchn.-Issled. Inst. Obshchestv. Zdorov. im. N.A. Semashko*, 2017, no. 4, pp. 69–75. (In Russian.)
12. URL: http://dic.academic.ru/dic.nsf/brokgauz_efron/110731/Холода.
13. Tadano, Y.S., Ugaya, C.M.L., and Franco, A.T., Methodology to Assess Air Pollution Impact on Human Health Using the Generalized Linear Model with Poisson Regression, in *Air Pollution - Monitoring, Modelling and Health*, Khare, M., Ed., Rijeka, Croatia: IntechOpen, 2012, pp. 281–304. DOI: 10.5772/33385.
14. Kolesnikova, L.I., Dolgikh, V.V., Astakhova, T.A., et al., Evaluation of Health Impairment and Malformations of Development of Children, *The Siberian Scientific Medical Journal*, 2008, vol. 28, no. 1(129), pp. 26–29. (In Russian.)
15. Pinegin, B.V., Baimukanova, G.P., and Pechurkina, N.S., Environmental Immunodeficiency: Immunogenetic Aspects of Its Development and Correction, *Vestn. Ross. Akad. Medits. Nauk*, 1994, no. 4, pp. 20–28. (In Russian.)
16. Savilov, E.D. and Il'ina, O.V., *Infektsionnaya patologiya v usloviyakh tekhnogennogo zagryazneniya okruzhayushchei sredy* (Infectious Pathology under Technogenic Pollution of the Environment), Novosibirsk: Nauka, 2010. (In Russian)
17. Dezateux, C., Lum, S., Hoo, A., et al., Low Birth Weight for Gestation and Airway Function in Infancy: Exploring the Fetal Origins Hypothesis, *Thorax*, 2004, vol. 59, pp. 60–66.
18. Hoo, A.F., Dezateux, C., Henschen, M., et al., Development of Airway Function in Infancy after Preterm Delivery, *Journal of Pediatrics*, 2002, vol. 141, pp. 652–658.
19. Dezateux, C., Stocks, J., Dundas, I., and Fletcher, M.E., Impaired Airway Function and Wheezing in Infancy: The Influence of Maternal Smoking and a Genetic Predisposition to Asthma, *American Journal of Respiratory and Critical Care Medicine*, 1999, vol. 159, pp. 403–410.
20. Fanucchi, M.V., Pulmonary Developmental Responses to Toxicants, in *Comprehensive Toxicology: Toxicology of the Respiratory System*, Oxford: Pergamon Press, 1997, pp. 203–220.
21. Smiley-Jewell, S.M., Liu, F.J., Weir, A.J., and Plopper, C.G., Acute Injury to Differentiating Clara Cells in Neonatal Rabbits Results in Age-Related Failure of Bronchiolar Epithelial, *Toxicologic Pathology*, 2000, vol. 28, pp. 267–276.
22. Methodical Recommendations MR 2.1.10.0057-12.2.1.10. *The Health Status of the Population due to the State of the Environment and Living Conditions of the Population. Assessing the Risk and Damage from Climate Changes Increasing the Levels*

of Morbidity and Mortality in Age Groups at Higher Risk. Approved January 17, 2012. (In Russian.)
23. Tang, G., Zhao, P., Wang, Y., Gao, W. Mortality and Air Pollution in Beijing: The Long-Term Relationship, Atmospheric Environment, 2017, no. 150, pp. 238–243.
24. Leone, M., D'Ippoliti, D., De Sario, M., Analitis, A., et al., A Time Series Study on the Effects of Heat on Mortality and Evaluation of Heterogeneity into European and Eastern-Southern Mediterranean Cities: Results of EU CIRCE Project, *Environmental Health*, 2013, no. 12:55, pp. 1–12.
25. D'Ippoliti, D., Michelozzi, P., Marino, C., de'Donato, F., et al., The Impact of Heat Waves on Mortality in 9 European Cities: Results from the EuroHEAT Project, *Environmental Health*, 2010, no. 9:37, pp. 1–9.
26. Almeida, S.P., Casimiro, E., and Calheiros, J.M., Effects of Apparent Temperature on Daily Mortality in Lisbon and Oporto, Portugal, *Environmental Health*, 2010, no. 9:12, pp. 1–7.
27. Heo, S., Lee, E., Kwon, B.Y., et al., Long-Term Changes in the Heat-Mortality Relationship according to Heterogeneous Regional Climate: A Time-Series Study in South Korea, *BMJ Open*, 2016, no. 6, pp. 1–10.
28. Eitan, O., Yuval, B.M., et al., Spatial Analysis of Air Pollution and Cancer Incidence Rates in Haifa Bay, Israel, *Science of the Total Environment*, 2010, no. 408, pp. 4429–4439.
29. Frome, E.L., The Analysis of Rates Using Poisson Regression Models, *Biometrics*, 1983, no. 39, pp. 665–674.
30. Revich, B.A., Shaposhnikov, D.A., and Semutnikova, E.G., Climatic Conditions and Air Quality as Risk Factors for Mortality in Moscow, *Russian Journal of Occupational Health and Industrial Ecology*, 2008, no. 7, pp. 29–35. (In Russian.)
31. URL: http://meteo.ru/
32. URL: https://rp5.ru/Weather_in_the_world
33. Basovskii, L.E., *Istoriya i metodologiya ekonomicheskoi nauki* (History and Methodology of Economics), Moscow: INFRA-M, 2017. (In Russian.)
34. Guangyong, Z., A Modified Poisson Regression Approach to Prospective Studies with Binary Data, *American Journal of Epidemiology*, 2004, vol. 159, no. 7, pp. 702–706. DOI: 10.1093/aje/kwh090.
35. Milyaev, V.A. and Kotel'nikov, S.N., Poisonous Ozone: Another Environmental Threat to Russia, *Ekolog. Zhizn'*, 2008, no. 2(75), pp. 52–56. (In Russian.)
36. Zhang, J.(J.), Wei, Y., and Fang, Z., Ozone Pollution: A Major Health Hazard Worldwide, *Frontiers in Immunology*, 2019, 10:2518.
37. Balajee, K.L., Babu, S., Suliankatchi, R.A., and Meena, S., Characteristics of the Ozone Pollution and Its Health Effects in India, *International Journal of Medicine and Public Health*, 2017, vol. 7(1), pp. 56-60.
38. URL:https://cfpub.epa.gov/ncea/iris/search/index.cfm?keyword=ozone.
39. Nashchenko, N.I., *Lektsii po ekonometrike* (Lectures on Econometrics), Ulyanovsk: Ulyanovsk State Technical University, 2008. (In Russian.)

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

Received June 17, 2021, and revised July 19, 2021.
Accepted August 24, 2021.

Author information

Taseiko, Olga Viktorovna. Cand. Sci. (Phys.-Math.), Reshetnev Siberian State University of Science and Technology; Siberian Branch, Russian Academy of Sciences, Krasnoyarsk, Russia
✉ taseiko@gmail.com

Chernykh, Dar'ya Aleksandrovna. Siberian Branch, Russian Academy of Sciences; Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia
✉ dachernykh93@gmail.com

Cite this article

Taseiko, O.V., Chernykh, D.A. Assessing the Impact of Environmental Factors on Mortality in Elder Age Groups: An Example of Krasnoyarsk, *Control Sciences* **5**, 53–60 (2021).
http://doi.org/10.25728/cs.2021.5.5

Original Russian Text © Taseiko, O.V., Chernykh, D.A., 2021, published in *Problemy Upravleniya*, 2021, no. 5, pp. 60–69.

Translated into English by *Alexander Yu. Mazurov*,
Cand. Sci. (Phys.–Math.),
Trapeznikov Institute of Control Sciences,
Russian Academy of Sciences, Moscow, Russia
✉ alexander.mazurov08@gmail.com

DOI: http://doi.org/10.25728/cs.2021.5.6

NON-BLOCKING FAULT-TOLERANT DUAL PHOTON SWITCHES WITH HIGH SCALABILITY

V.S. Podlazov

Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia

✉ podlazov@ipu.ru

Abstract. This paper continues the construction of a fundamentally new class of system area networks (dual photon networks) with the following features: non-blocking property and static self-routing, high scalability with the maximum achievable speed and a small complexity compared to a full switch, and balancing the scalability-speed and complexity-speed ratios. These networks are implemented in an extended circuit basis consisting of dual photon switches and separate photon multiplexers and demultiplexers. We propose a method for constructing a fault-tolerant dual network with the indicated properties based on networks with the quasi-complete graph and quasi-complete digraph topologies and the invariant extension method with internal parallelization. Also, we propose a method for extending the two-stage dual network designed previously into four-stage and eight-stage dual networks with high scalability while maintaining the original network period and reducing its exponential complexity.

Keywords: photon switch, dual switch, photon multiplexers and demultiplexers, multistage switch, conflict-free self-routing, non-blocking switch, static self-routing, quasi-complete digraph, quasi-complete graph, invariant extension of networks, switching properties, direct channels, scalability and speed.

INTRODUCTION

This paper develops a method for constructing a fundamentally new class of system area networks [1–6], the so-called dual photon networks. They are non-blocking networks with static self-routing [1–3, 5] and can have a given degree of channel fault tolerance [6].

In what follows, we propose a method for constructing non-blocking self-routing photon networks with high scalability. These are dual networks based on a non-blocking dual $p \times p$ switch with a signal period of p cycles [1–3]. For resolving signal conflicts, the dual switch combines the bus method (separation of conflicting signals to different cycles in one channel) and the switch method (separation of conflicting signals to different channels). The dual switch is a non-blocking switch on any input traffic if data bits are transmitted with a signal period of p cycles. The dual switch was developed by the author's colleagues [1, 2] and then applied and named in the joint publications [3–5]. This switch turned out to be a prerequisite for constructing non-blocking networks with high scalability and acceptable complexity.

The dual photon switch transmits signal and control information in parallel at different frequencies for each data bit. This method eliminates the problem of synchronizing signals from different channels.

The photon specifics of such networks consist in using in-bit channel virtualization with feedback links through delay lines with a duration of one cycle and control signals at different frequencies to route individual bits. The separation of information signals to different cycles is accompanied by the separation of the corresponding control information to the same cycles. It is used to route the bits by moving them between different channels without changing the established cycle numbers.

Throughout the paper, the terms "dual switch" and "dual network based on dual switches" imply using bits with a period of p cycles in them. These bits en-

sure the non-blocking property in the first stage of the network; in the other stages, they remain by "inertia" without using the bus method of conflict resolution.

The scalability of dual networks is provided using networks with the quasi-complete graph or digraph topology [4], which are implemented in an extended circuit basis consisting of dual photon switches and separate photon multiplexers and demultiplexers. In [1–3, 5], high scalability was achieved using the invariant extension method of system area networks with many additional demultiplexers and multiplexers.

In the papers [5, 6], a new method for extending dual networks by their internal parallelization without additional devices was developed and applied for the first time. Particularly in [5], a two-stage non-blocking network was constructed. It consists of networks with the quasi-complete digraph topology with $N = p^2$ channels on each stage, whereas the two-stage non-blocking network has N^2 channels in total. On the other hand, a two-stage non-blocking network with $(\sigma - 1)$-channel fault tolerance was constructed in [6]. It consists of networks with the quasi-complete graph topology with $N = p(p-1)/\sigma + 1$ channels, whereas the two-stage fault-tolerant non-blocking network has N^2 channels.

Below, we construct four-stage and eight-stage fault-tolerant networks by developing and applying the generalized method of internal parallelization. In this case, the same degree of network scalability is achieved as in the invariant method, but without using external demultiplexers and multiplexers, and the resulting networks have a significantly smaller complexity.

This paper is organized as follows. Section 1 briefly considers the non-blocking property and channel fault tolerance in modern system area networks. In Section 2, following [6], we introduce the notions of p-permutations (crucial for proving the non-blocking property of four- and eight-stage networks) and repeat the proofs of the non-blocking property for two-stage networks. Section 3 presents four-stage non-blocking self-routing switches with one-channel and two-channel fault tolerance and their performance characteristics. The method of internal parallelization from [5, 6] is generalized for four-stage switches.

Section 4 compares the performance characteristics of four-stage non-blocking self-routing switches based on switches with the dual quasi-complete graph and digraph topologies. Finally, in Section 5, we construct eight-stage non-blocking self-routing switches based on switches with the dual quasi-complete graph and digraph topologies. Moreover, the method from Section 3 is generalized for eight-stage switches. Section 6 discusses the properties of these networks compared to other non-blocking networks (particularly their disadvantages and possible ways to overcome them).

In the Conclusions, we analyze the generalized method of internal parallelization, which is the core for constructing dual non-blocking networks with high scalability and low specific complexity. There are three main components of the proposed methodology: a non-blocking dual switch, a switch with the quasi-complete graph or digraph topology based on a dual switch, and the method of internal parallelization.

1. NON-BLOCKING PROPERTY AND FAULT TOLERANCE IN SYSTEM AREA NETWORKS

The problem of constructing non-blocking fault-tolerant system area networks of supercomputers has not been completed solved so far.

A system area network is non-blocking if for any packet permutation, conflict-free paths from sources to sinks can be built in it. A system area network is self-routing if conflict-free paths can be built locally over network nodes without their interaction based on routing information in packets only. Finally, self-routing is static if any source can independently choose conflict-free paths to its sink without interacting with other sources.

The existence of non-blocking networks was proved by Clos [7, 8]. Self-routing procedures for non-blocking Clos networks have not yet been developed. However, these networks can be a qualitative measure for other non-blocking networks.

A network in the form of a two-dimensional generalized hypercube with the quasi-complete digraph topology is non-blocking, e.g., in the YARK and ROSETTA switches used in several networks of different structure: a reconfigurable Clos network [9], a three-dimensional torus [10], and a hierarchy of complete and quasi-complete digraphs [11–13]. Unfortunately, a quasi-complete digraph has a small number of channels $N = p^2$, where p is the degree of internal switches, and a high switching complexity $S \geq N^2$, exceeding the complexities of a complete digraph and a non-blocking Clos network (in the latter case, considerably).

Modern literature widely describes system area networks with the fat tree structure (particularly reconfigurable Clos networks), the generalized hypercube structure, the multidimensional torus structure, and system area networks with a hierarchy of complete and quasi-complete digraphs.

Fat-tree networks are reconfigurable networks [9, 14, 15] with conflict-free transmission only according to predetermined schedules for specific packet permutations. In the case of arbitrary permutations, these networks turn out to be blocking; permutations in them are implemented in several jumps between network nodes. The maximum number of such jumps determines the network diameter. In reconfigurable Close networks, the diameter equals the number of network stages.

Networks with the generalized hypercube structure [16–19] are not even reconfigurable [20, 21]. They can be made such by increasing the number of channels in some dimensions. Generalized cubes have a diameter equal to the number of dimensions or less by 1 in the extended hypercube [17, 18]. Generalized hypercubes with a doubled number of channels in each dimension are reconfigurable networks for two permutations simultaneously. Note that an attempt to use a generalized hypercube as a non-blocking network for a photon computer [22] seems to be a very dubious venture.

For arbitrary permutations, networks with the multidimensional torus structure cannot transmit packets over direct channels [11, 23–25]. They implement permutations in several jumps between network nodes. Multidimensional tori are the simplest, albeit slowest, networks due to their large diameters. For example, the networks considered in [11, 23–25] have diameters measured in tens of jumps.

On the contrary, networks with a hierarchy of complete or quasi-complete digraphs [10, 12, 26] have the smallest diameter of three jumps. Many networks with small diameters have appeared recently [27–32]. All of them have serious problems with balancing network load under channel faults.

Channel fault tolerance is the network's ability to preserve full availability under channel failures while maintaining its original performance characteristics (the non-blocking property, transmission delays, or network diameter).

Apparently, only networks with the quasi-complete graph topology possess channel fault-tolerance in pure form. These networks are isomorphic to such a mathematical object as an incomplete balanced symmetric block design [33]. These networks have an element base of $p\times p$ switches, $1\times p$ demultiplexers, and $p\times 1$ multiplexers and are non-blocking networks with static self-routing. They have direct channels between $N = p(p-1)/\sigma + 1$ network users and σ different channels between any two users [4].

In other networks, the restoration of full network availability under channel faults is accompanied in one way or another by an increase in network transmission delays. For example, under channel faults in a reconfigurable Clos network, the load on the remaining channels grows, increasing the number of conflicts and delays in the transmission of some packets.

The TOFY network with the three-dimensional torus structure [25] uses three more dimensions to create redundant channels. If some network rings fail, their integrity is restored by increasing the network diameter by 1.

Generalized hypercubes with doubled channels in each dimension are one-fault-tolerant networks with a constant diameter [19].

In networks with a hierarchy of complete or quasi-complete digraphs [10, 12, 26], if some of the channels fail, the network's full availability is restored using bypass paths with a duration of five jumps. In other words, the transmission delay increases by a factor of 5/3.

2. DUAL QUASI-COMPLETE GRAPH, *A*PERMUTATIONS, AND DUAL TWO-STAGE SWITCH

The dual switch SFN_1 with the quasi-complete graph topology, SQG(N_1, p, σ), consists of $N_1 = p(p-1)/\sigma + 1$ dual $p\times p$ switches SSp, N_1 input $1\times p$ demultiplexers, and N_1 output $p\times 1$ multiplexers [6]. They are interconnected using the combinatorial method [4]: there are σ different paths through different dual switches SSp. Figure 1 shows the circuit of an SF4 switch as an SQG (4, 3, 2) graph with one-channel fault tolerance. Two paths are highlighted, connecting two randomly selected inputs and outputs, (2, 4) and (3, 3).

Any dual switch SFN_1 has the same signal period T_1 as the dual switch SSp included in it. For the switch SFN_1, the following performance characteristics are calculated: the switching complexity S_1, expressed in the number of switching points, and the channel complexity, expressed in the number of channels. They are

written exponentially through the number of channels and are called exponential complexities[1]; see Table 1.

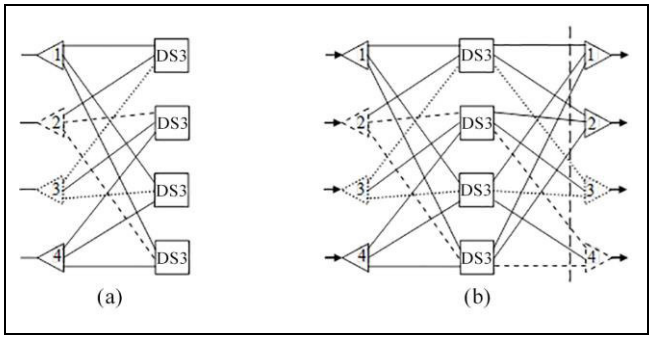


Fig. 1. Dual quasi-complete switch SF4 with signal period of three cycles represented by graph SQG(4, 3, 2): (a) original form with duplex channels, (b) application with simplex channels. Dashed lines and dots indicate different paths between selected inputs and outputs.

Table 1

Performance characteristics of dual switches SFN_1 with one-channel fault tolerance

p	N_1	T_1	S_1	L_1
2	2	2	$24 = N_1^{4.58}$	$8 = N_1^{3}$
4	7	4	$280 = N_1^{2.9}$	$56 = N_1^{2.07}$
6	15	6	$1\,260 = N_1^{2.64}$	$180 = N_1^{1.92}$
8	27	8	$3\,888 = N_1^{2.51}$	$432 = N_1^{1.84}$

For the dual switch SFN_1 with the quasi-complete *digraph* topology [5], the performance characteristics are given in Table 2. They are better than their counterparts from Table 1 but without channel fault tolerance.

Table 2

Performance characteristics of dual switches SFN_1 based on the quasi-complete digraph topology

p	N_1	T_1	S_1	L_1
2	4	2	$48 = N_1^{2.79}$	$16 = N_1^{2}$
4	16	4	$640 = N_1^{2.33}$	$128 = N_1^{1.75}$
6	36	6	$3\,024 = N_1^{2.24}$	$432 = N_1^{1.69}$
8	64	8	$9\,216 = N_1^{2.19}$	$1\,024 = N_1^{1.67}$

We introduce the notion of p-partitions of packets transmitted through some cross-section of a network at the multiplexer inputs. All packets are divided into groups of variable composition, each containing at most p packets. For a common permutation of packets, a 1-partition occurs at the input and output of the switch. A transmission in which a 1-partition occurs at the network input and a p-partition on a given cross-section will be called a p-permutation.

For the dual switch SQG(N_1, p, σ), this cross-section is through the inputs of the output multiplexers and is called the output cross-section. In Fig. 1, the output cross-section is indicated by the vertical dashed line. According to the property of the dual switch DSp, a p-partition occurs on the output cross-section of the dual switch SQG(N_1, p, σ) for any traffic.

Lemma 1 [6]**.** *The dual switch* SFN_1 *with a signal period of p cycles is a non-blocking switch with static self-routing under any common permutation and has* (σ – 1)*-channel fault-tolerance.*

The papers [5, 6] developed a method for constructing two-stage non-blocking switches with N_1^2 channels. Consider this method on an example of the dual switch SF2 with the dual quasi-complete graph topology SQG (2, 2, 2) (Fig. 2).

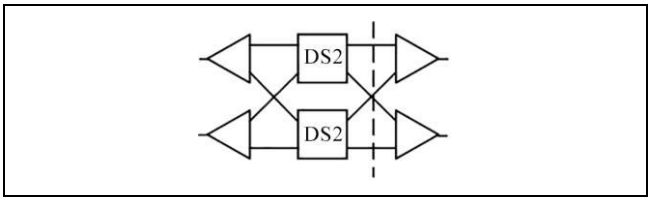


Fig. 2. Dual non-blocking switch SF2 with one-channel fault tolerance.

On its basis, a four-channel two-stage network $N_2$4 is constructed. This network contains two SF2 switches on each stage, connected by exchange links (Fig. 3). The network is blocking on multiplexers of the first stage, highlighted in grey, and loses channel fault tolerance on them.

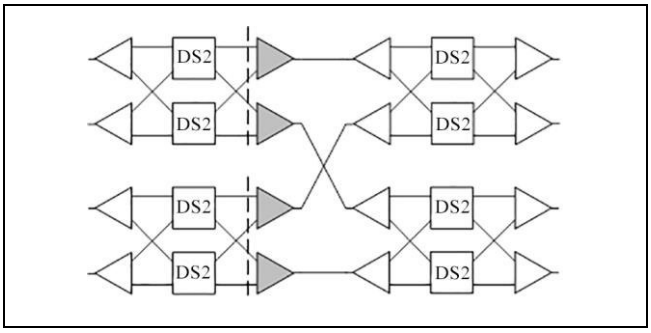


Fig. 3. Dual two-stage blocking network $N_2$4 with exchange links.

[1] This term was introduced by the author.

The network $N_2 4$ can be transformed into a non-blocking switch $S_2 4$ with one-channel fault tolerance by internal parallelization [5, 6]. The second stage of this switch uses two copies of the second stage of the $N_2 4$ network. On the first stage, multiplexers on the cross-sections are eliminated, and their inputs are connected to the inputs of the second stage copies: odd to the first copy and even to the second copy. These links preserve the order of connecting the channels located on the second stage in the network $N_2 4$. The cut-out multiplexers are moved to connect the outputs of the second stage copies, forming the output multiplexers of the switch $S_2 4$ (Fig. 4), which becomes a non-blocking switch with static self-routing and one-channel fault tolerance.

In the general case ($p \geq 2$), the dual switch $SF N_1$ has the dual quasi-complete graph topology $SQG(N_1, p, \sigma)$ with a signal period of p cycles. On its basis, a two-stage blocking network $N_2 N_2$ ($N_2 = N_1^2$) is constructed, where each stage contains N_1 switches $SF N_1$ with exchange links between the stages. For internal parallelization, p copies of the second stage of the network $N_2 N_2$ are formed, and the first stage multiplexers are used to combine the same-name outputs of the second stage copies.

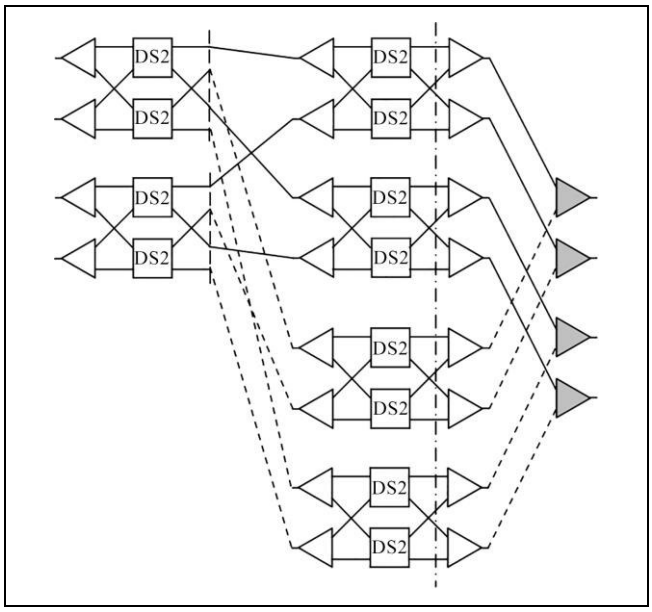


Fig. 4. Dual non-blocking self-routing switch $S_2 4$ with one-channel fault tolerance.

On the cross-sections of the first stage, there are pN_2 inputs to the multiplexers. They are renumbered top-to-bottom by I ($1 \leq I \leq pN_2$), and the inputs $i = I(\mathrm{mod} p)+1$ are connected to the same-name inputs of the ith copy of the second stage, maintaining the same arrangement of the switches $S_1 N_1$ as in the network $N_2 N_2$.

Lemma 2. *The dual switch $S_2 N_2$ has a p-permutation on the indicated cross-section. It is a non-blocking switch with static routing on any common permutation and has $(\sigma - 1)$-channel fault tolerance.*

The switching and channel complexities of the switch $S_2 N_2$ are given by the recursive formulas $S_2 = N_1 S_1 + pN_1 S_1$ and $L_2 = N_1 L_1 + pN_1 L_1$, respectively. The performance characteristics of the switches $S_2 N_2$ for $\sigma = 2$ are presented in Table 3. Note that the exponential complexities decrease compared to Table 1.

Table 3

Performance characteristics of dual switches $S_2 N_2$ with one-channel fault tolerance

p	N_1	$N_2 = N_1^2$	$T_2 = p$	S_2	L_2
2	2	4	2	$N_2^{3.58}$	$N_2^{2.9}$
4	7	49	4	$N_2^{2.37}$	$N_2^{1.97}$
6	15	225	6	$N_2^{2.18}$	$N_2^{1.84}$
8	27	729	8	$N_2^{2.09}$	$N_2^{1.77}$

The performance characteristics of a dual switch $S_2 N_2$ based on the quasi-complete digraph topology [5] are combined in Table 4. They are significantly better than in Table 2, but without channel fault tolerance.

Table 4

Performance characteristics of switches $S_2 N_2$ based on the quasi-complete digraph topology

p	N_1	$N_2 = N_1^2$	$T_2 = p$	S_2	L_2
2	4	16	2	$N_2^{2.29}$	$N_2^{1.95}$
4	16	256	4	$N_2^{1.96}$	$N_2^{1.68}$
6	36	1 296	6	$N_2^{1.89}$	$N_2^{1.63}$
8	64	4 096	8	$N_2^{1.86}$	$N_2^{1.6}$

Also, note that the dual switch $S_2 N_2$ has two stages of output multiplexers containing pN_2 and N_2 multiplexers, respectively. For the purposes of Section 3, we cut the switch $S_2 N_2$ through the inputs of the first stage of multiplexers; see the dash-and-dot line in Fig. 4.

3. FOUR-STAGE FAULT-TOLERANT NON-BLOCKING SELF-ROUTING SWITCH WITH TWO-DIMENSIONAL INTERNAL PARALLELIZATION

In the papers [1–3, 5], the number of channels of a non-blocking switch was further increased using the invariant extension method with external parallelization. This method does not change the signal period. However, it is of little use for fault-tolerant non-blocking switches [6].

In this section, we increase the number of channels without changing the signal period using the generalized method of internal parallelization of the network by constructing four-stage switches S_4N_4 from two-stage switches S_2N_2 with the number of channels $N_4 = N_2^2$ and the signal period $T_4 = T_2 = p$.

The network is constructed using switches $S_2$4 as an example (Fig. 4). First, the two-stage network $N_4$16 is created. Each stage in this network consists of four copies of the $S_2$4 switch, and the stages are interconnected by exchange links (Fig. 5).

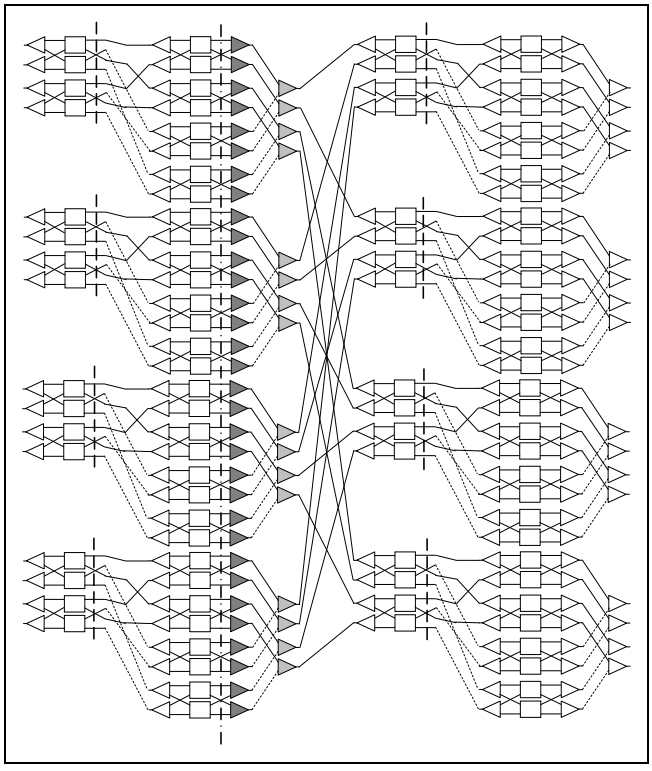


Fig. 5. Blocking dual network $N_4$16 without channel fault tolerance.

In reality, the network $N_4$16 consists of four stages SQG(2, 2, 2), which explains the subscript in the notation. The $N_4$16 network is blocking due to possible signal conflicts on the two stages of the output multiplexers M2 (highlighted in grey). There are two layers of such multiplexers, $W_4 = 48$ in total. In addition, channel fault tolerance is violated on them. The dash-and-dot cut is made through the inputs to the first stage (Fig. 5). For this cut, the notion of a *p*-permutation has been formulated in Section 2.

Then the network $\underline{N}_4$16 is created. It contains the first stage of the network $N_4$16 and two copies of the second stage of the network $N_4$16. One parallel circuit of the first dimension is created in the network $\underline{N}_4$16 (Fig. 6). For this, the outer layer highlighted in pale grey is first cut out with $W_2^* = 16$ multiplexers M2 in total. They remain unconnected for now. Then the multiplexers M2 of the inner layer highlighted in dark grey are cut out, and their odd inputs are routed to the inputs of two second-stage copies of the network $\underline{N}_4$16. In this case, the $W_{4,1} = 16$ cut multiplexers combine the outputs of these two copies.

The remaining $W_{4,2} = 16$ multiplexers M2 highlighted in dark grey are used to create the second parallel circuit of the first dimension in the same way (Fig. 7). Their even inputs are routed to the inputs of two additional copies of the second stage of the network $N_4$16.

Looking ahead, note that Figs. 6 and 7 show the new connections of the multiplexers of the first and second layers. They define the combination of the first dimension circuits into the second dimension circuit.

As a result, two circuits of the first dimension are constructed, each consisting of two switches $S_2$4 connected in parallel. (The connections of the second circuit of the first dimension are not shown in Fig. 8.) The two circuits of the first dimension form a two-dimensional circuit. The outputs of the two-dimensional circuit combine $W_2^* = 16$ multiplexers highlighted in pale grey, forming the outputs of the switch $S_4$16. In Fig. 8, the latter connections are indicated by dotted lines and are shown completely in one copy only due to the lack of space. (They can be found in Figs. 6 and 7.)

The resulting sixteen-channel network consists of 16 copies of the switch $S_2$4 connected in parallel. Their inputs receive sparse alternative direct *p*-permutations implemented without conflict according to a single static schedule; see Lemma 2. (Alternatives *p*-permutations intersect neither in inputs nor outputs.) The paths between sources and sinks in switches $S_2$4 follow two subpaths through different circuits of the first dimension. Therefore, the switch $S_2$4 has one-channel fault tolerance since $p = \sigma = 2$.

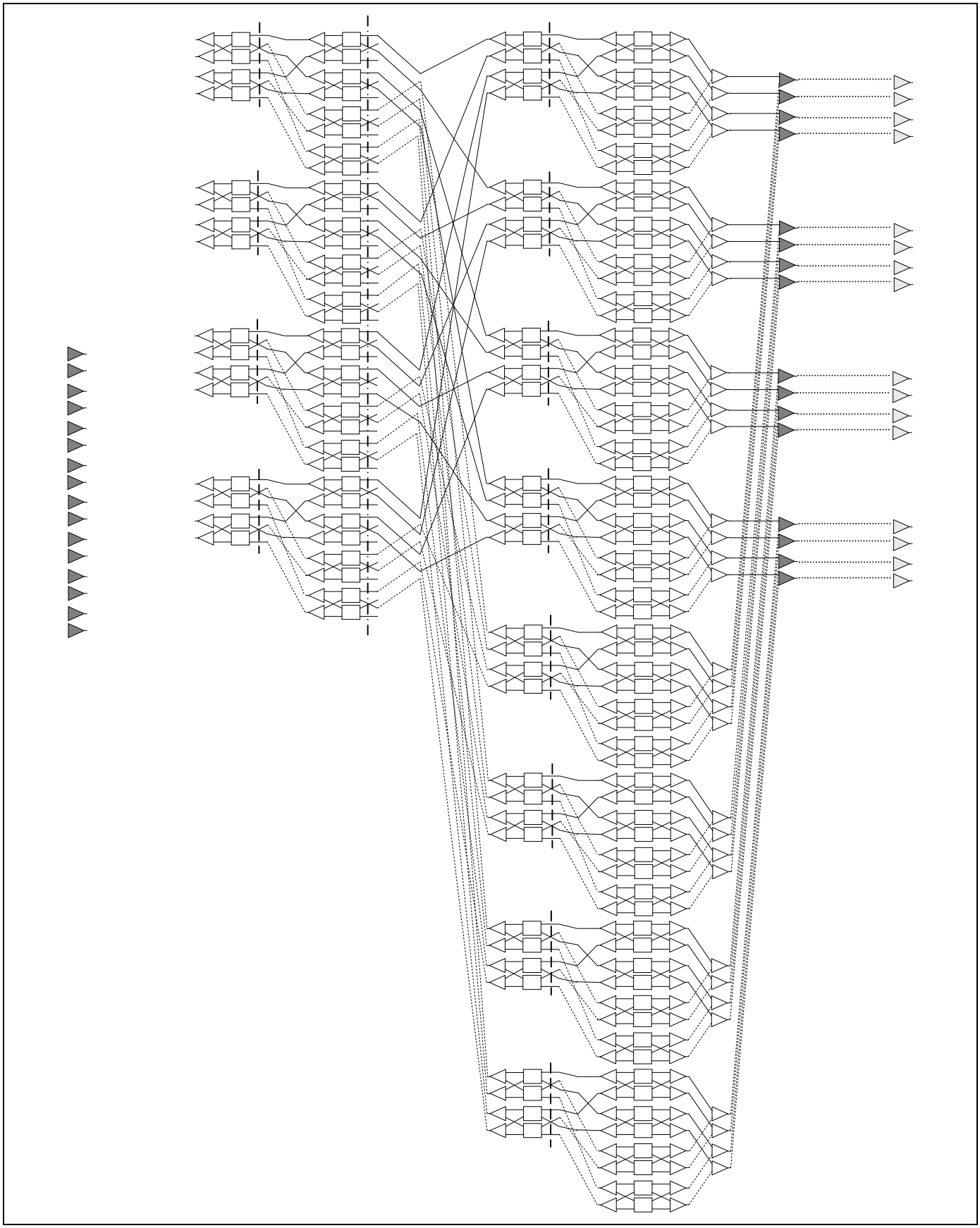


Fig. 6. Constructing the first circuit of the first dimension. Multiplexers of network $N_4 16$ not used in this circuit are shown on the left.

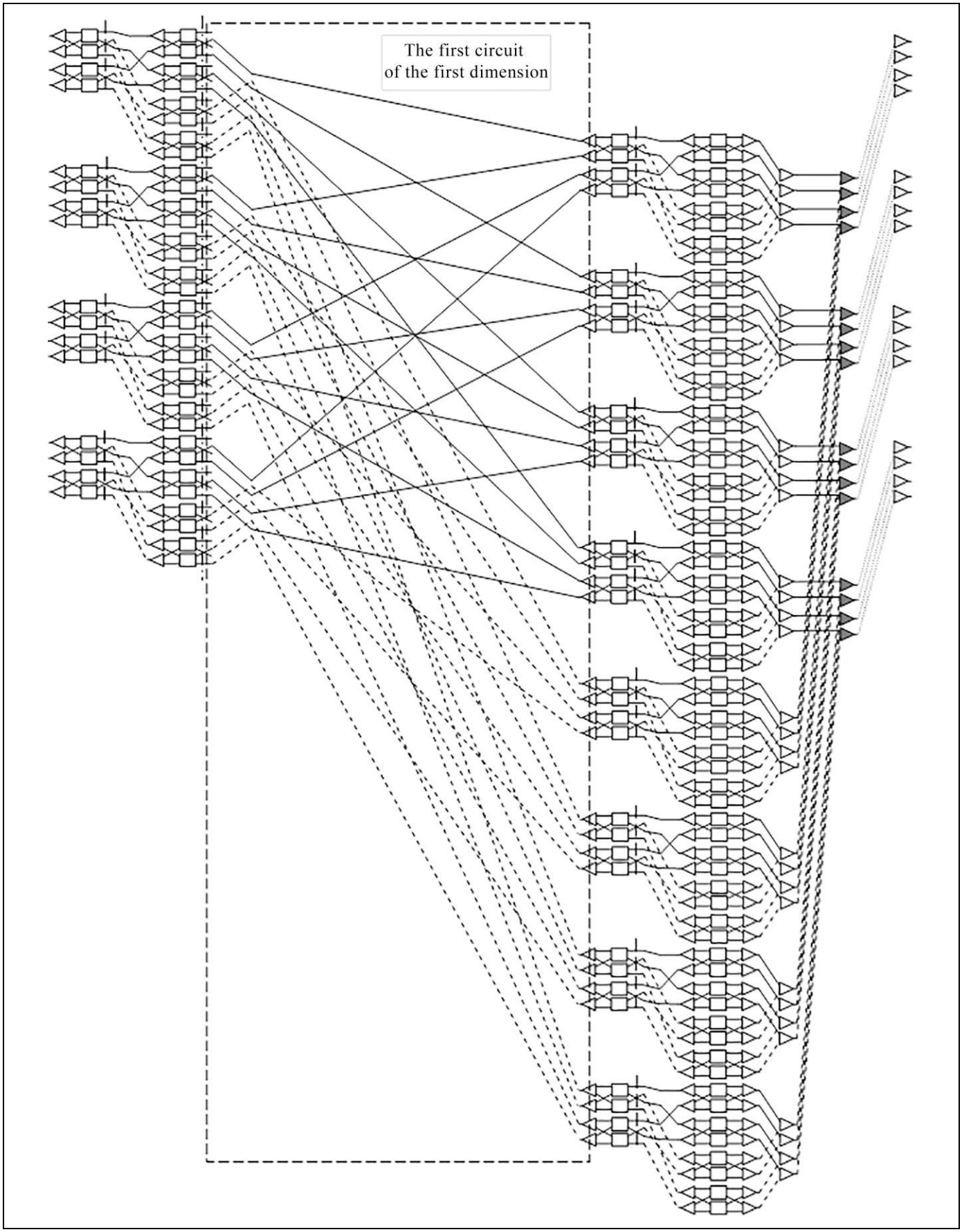


Fig. 7. Constructing the second circuit of the first dimension using the multiplexers not included in the first circuit of the first dimension.

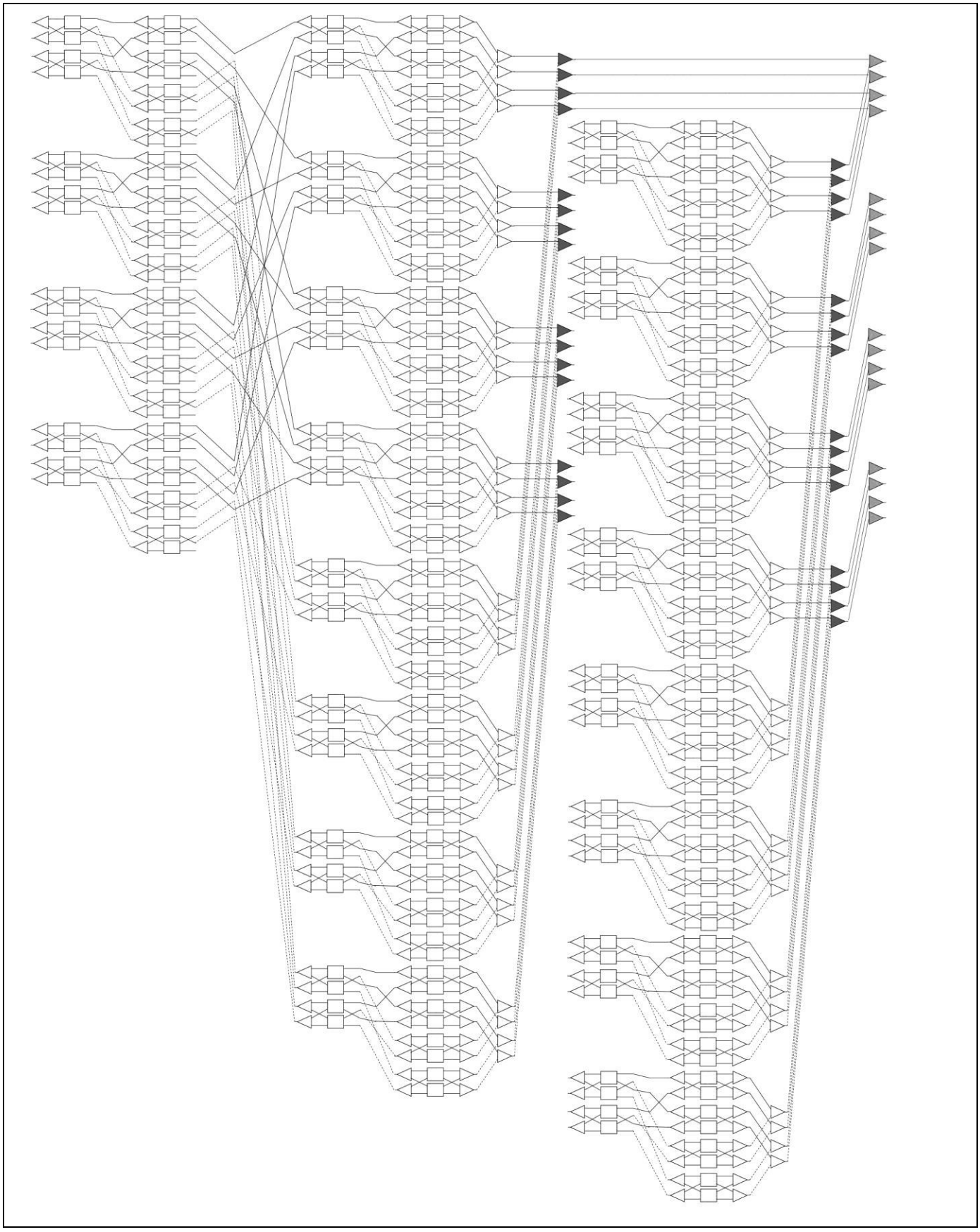


Fig. 8. Non-blocking dual switch $S_4$16 with one-channel fault tolerance.

In the general case ($p > 2$ and $\sigma \geq 2$), the network N_4N_4 is first constructed. It consists of two stages with N_2 switches S_2N_2 in each, connected by exchange links. This network has $N_4 = N_2^2$ channels, and signals can conflict at output multiplexers Mp. Therefore, it is a blocking network without channel fault tolerance.

The first stage of the network N_4N_4 has two layers of output multiplexers Mp, $W_4 = N_2V_2 = N_4(p+1)$ in total. The first inner layer of multiplexers Mp contains $W_{4,1} = pN_4$ multiplexers Mp, which together have p^2N_4 inputs.

Then the network $\underline{N}_4N_4$ is created. It contains the first stage of the network N_4N_4 and p^2 copies of the second stage of the network N_4N_4. The network $\underline{N}_4N_4$ is constructed with the following structure: copies of the second stage of the network N_4N_4 form p circuits of the first dimension, and all together, they form the circuit of the second dimension.

The network $\underline{N}_4N_4$ contains p^2N_2 switches S_2N_2, which have p^2N_4 inputs in total. Further, both layers of multiplexers Mp in the first stage of the network N_4N_4 are cut out, and the inputs of the first layer of multiplexers Mp are connected to the inputs of switches S_2N_2. This is possible since the latter and former have the same number of inputs.

We divide the p^2 copies of the second stage of the network N_4N_4 in the network $\underline{N}_4N_4$ into p groups, denoting by G ($1 \leq G \leq p$) the group number and by I ($1 \leq I \leq p$) the copy number. In fact, G is the circuit number of the first dimension, and I is the copy number in the first dimension circuit.

In addition, we introduce the following notations: J ($1 \leq J \leq N_2$) is the number of the switch S_2N_2 in the first dimension circuit, and K ($1 \leq K \leq N_2$) is the input number of each such switch S_2N_2. Thus, the input of any switch S_2N_2 is given by the composite number G, I, J, K.

Also, we denote by i ($0 \leq i \leq p-1$) the mod$p$ input number of each multiplexer Mp in their first layer. Each switch in the first stage of the network N_4N_4 contains pN_2 such multiplexers Mp. We divide them into N_2 groups, denoting by g ($1 \leq g \leq N_2$) the group number and by j ($1 \leq j \leq p$) the number of the multiplexer Mp in the group. Let k ($1 \leq k \leq N_2$) denote the number of the switch S_2N_2 in the first stage of the network N_4N_4. Thus, the input of any multiplexer Mp in the first layer is given by the composite number i, g, j, k.

An arbitrary input of the multiplexer Mp with the number i, g, j, k is connected to the input of the switch S_2N_2 with the number G, I, J, K, where $G = i + 1$, $I = j$, $J = g$, and $K = k$. As a result, p^2N_2 switches S_2N_2 are connected in parallel, and their inputs receive sparse disjoint direct p-permutations.

The cut-out multiplexers Mp of the first layer combine the outputs of the first dimension circuits. The cut-out multiplexers Mp of the second layer combine the outputs of the p circuits of the first dimension, forming the outputs of the second-level circuit (the outputs of the switch S_4N_4).

Switches S_4N_4 have the following property.

Lemma 3. *The dual switch* S_4N_4 *is a non-blocking switch with static routing on any common permutation for any p. It has* $(\sigma - 1)$*-channel fault tolerance.*

P r o o f. The first statement is based on using the switch S_2N_2 and Lemma 2. The second statement is based on the non-blocking property of the switch S_2N_2 and the fact that the p-permutation on the cross-section consists of sparse 1-permutations separated to different channels and cycles.

According to the assignment $G = i+1$, different inputs of one multiplexer in the first layer are connected to different one-dimensional circuits (different copies of the second stage of the network N_4N_4), and the inputs of different multiplexers are connected to the inputs of different switches S_2N_2 in the second stage of the network N_4N_4. In other words, there are σ different paths of switches S_2N_2 along p different paths in the switch S_4N_4. Due to $p \geq \sigma$, the latter switch is therefore $(\sigma - 1)$-channel fault-tolerant. ♦

As a result, the non-blocking self-routing switch S_4N_4 with $N_4 = N_2^2$ channels and $(\sigma - 1)$-channel fault tolerance has the performance characteristics shown in Tables 5 and 6. They are calculated using the recursive formulas $S_4 = N_2S_2 + p^2N_2S_2$ and $L_4 = N_2L_2 + p^2N_2L_2$. The switch S_4N_4 has four layers of output multiplexers Mp, $V_4 = N_4(p^4 - 1)/(p - 1)$ in total.

Note the further decrease in the exponential switching and channel complexities of the switch S_4N_4 (Table 4) compared to the switch S_2N_2 (Table 2).

Abandoning the requirement of channel fault tolerance, we can construct a non-blocking self-routing switch S_4N_4 based on the switch with the quasi-complete digraph topology. Its performance characteristics are presented in Table 7. Compared to the fault-tolerant modifications, it has more channels and smaller complexity. In addition, the entire set of switches has switching and channel complexities less than those of a two-stage switch based on a switch with the complete graph topology (Table 3) and less than those of a switch with the complete graph topology (*switchboard*).

4. EIGHT-STAGE NON-BLOCKING SWITCHES WITH FOUR-DIMENSIONAL INTERNAL PARALLELIZATION BASED ON THE GRAPH AND DIGRAPH TOPOLOGIES

The method of extending two-stage switches S_2N_2 into four-stage switches S_4N_4 can be generalized to construct eight-stage switches S_8N_8 from four-stage switches S_4N_4.

First, the network N_8N_8 is constructed. It consists of two stages with N_4 switches S_4N_4 in each, connected by exchange links. This network has $N_8 = N_4^2$ channels, and signals can conflict at output multiplexers Mp of the first stage. Therefore, it is a blocking network without channel fault tolerance.

The first stage of the network N_8N_8 has four layers of output multiplexers Mp, $W_8 = N_4V_4 = N_8(p^4 - 1)/(p - 1)$ in total. The first inner layer of multiplexers Mp contains $W_{8,1} = p^3N_8$ multiplexers Mp, which together have p^4N_8 inputs.

Then the network $\underline{N}_8N_8$ is created. It contains p^4 copies of the second stage of the network N_8N_8. The network $\underline{N}_8N_8$ is created with the following structure.

The second stage copies of the network N_8N_8, p in total, form the circuit of the first dimension, and p circuits of the first dimension form the second dimension circuit. Similarly, the third dimension circuit consists of p second dimension circuits, and the fourth dimension circuit consists of p third dimension circuits.

The network $\underline{N}_8N_8$ contains p^4N_4 switches S_4N_4, which have p^4N_8 inputs in total. Further, all four layers of multiplexers Mp in the first stage of the network N_8N_8 are cut out, and the inputs of the first layer of multiplexers Mp are connected to the inputs of switches S_4N_4. This is possible since the latter and former have the same number of inputs.

We divide the p^4 copies of the second stage of the network N_8N_8 in the network $\underline{N}_8N_8$ into p^3 groups, denoting by G ($1 \le G \le p^3$) the group number and by I ($1 \le I \le p$) the number in the group. In fact, G is the circuit number of the first dimension, and I is the copy number of S_4N_4 in the first dimension circuit.

Table 5

Performance characteristics of dual switches S_4N_4 with one-channel fault tolerance

p	N_1	$N_4 = N_1^4$	$T_4 = p$	S_4	L_4
2	2	16	2	2 720 = $N_4^{2.85}$	1 120 = $N_4^{2.53}$
3	4	256	3	238 080 = $N_4^{2.23}$	69 120 = $N_4^{2.01}$
4	7	2 401	4	8 000 132 = $N_4^{2.04}$	1 795 948 = $N_4^{1.85}$
5	11	14 641	5	1.35E+08 = $N_4^{1.95}$	24 743 290 = $N_4^{1.77}$
6	15	65 536	6	1.41E+09 = $N_4^{1.9}$	2.18E+08 = $N_4^{1.73}$
7	21	194 481	7	8.64E+09 = $N_4^{1.88}$	1.16E+09 = $N_4^{1.71}$
8	27	531 441	8	4.45E+10 = $N_4^{1.86}$	5.25E+09 = $N_4^{1.7}$

Table 6

Performance characteristics of dual switches S_4N_4 with two-channel fault tolerance

p	N_1	$N_4 = N_1^4$	$T_4 = p$	S_4	L_4
3	3	81	3	75 330 = $N_4^{2.56}$	21 870 = $N_4^{2.27}$
4	5	625	4	2 082 500 = $N_4^{2.26}$	467 500 = $N_4^{2.03}$
5	7	2 401	5	22 161 230 = $N_4^{2.17}$	4 057 690 = $N_4^{1.95}$
6	11	14 641	6	3.15E+08 = $N_4^{2.04}$	48 754 530 = $N_4^{1.85}$
7	15	50 625	7	2.25E+09 = $N_4^{1.99}$	3.01E+08 = $N_4^{1.8}$
8	19	130 321	8	1.09E+10 = $N_4^{1.96}$	1.29E+09 = $N_4^{1.78}$

Table 7

Performance characteristics of dual switches S_4N_4 based on the digraph topology

p	N_1	$N_2 = N_1^2 = p^4$	$N_4 = N_2^4 = p^8$	$T_4 = p$	S_4	L_4
2	4	16	256	2	43 520 = $N_4^{1.93}$	17 920 = $N_4^{1.77}$
3	9	81	6 561	3	6 101 730 = $N_4^{1.78}$	1 771 470 = $N_4^{1.64}$
4	16	256	65 536	4	2.18E+08 = $N_4^{1.73}$	49 020 928 = $N_4^{1.60}$
5	25	625	390 625	5	3.61E+09 = $N_4^{1.71}$	6.6E+08 = $N_4^{1.58}$
6	36	1 296	1 679 616	6	3.62E+10 = $N_4^{1.70}$	5.59E+09 = $N_4^{1.57}$
7	49	2 401	5 764 801	7	2.56E+11 = $N_4^{1.69}$	3.43E+10 = $N_4^{1.56}$
8	64	4 096	16 777 216	8	1.4E+12 = $N_4^{1.68}$	1.66E+11 = $N_4^{1.55}$

In addition, we introduce the following notations: J ($1 \le J \le N_4$) is the number of the switch S_4N_4 in the first dimension circuit, and K ($1 \le K \le N_4$) is the input number of each such switch S_4N_4. Thus, the input of any switch S_4N_4 is given by the composite number *G, I, J, K*.

Also, we denote by i ($0 \le i \le p-1$) the mod$p$ input number of each multiplexer Mp in their first layer. Each switch in the first stage of the network N_8N_8 contains p^3N_4 such multiplexers Mp. We divide them into N_4 groups, denoting by g ($1 \le g \le N_4$) the group number and by j ($1 \le j \le p$) the number of the multiplexer Mp in the group.

Let k ($1 \le k \le N_4$) denote the number of the switch S_4N_4 in the first stage of the network N_8N_8. Thus, the input of any multiplexer Mp in the first layer is given by the composite number *i, g, j, k.* For this purpose, all layers of multiplexers Mp in the first stage of the network N_8N_8 are cut out. The first inner layer contains $W_{8,1} = N_4p^3$ multiplexers Mp, and their inputs are connected to the inputs of switches S_4N_4 in the network $\underline{N_8}N_8$.

An arbitrary input of the first layer multiplexer Mp with the number *i, g, j, k* is connected to the input of the switch S_4N_4 with the number *G, I, J, K,* where $G = i + 1$, $I = j$, $J = g$, and $K = k$. As a result, p^4N_4 switches S_4N_4 are connected in parallel, and their inputs receive sparse disjoint direct p-permutations.

The cut-out multiplexers Mp of the first layer combine the outputs of the p^3 first dimension circuits. The cut-out multiplexers Mp of the second layer combine the outputs of the first dimension circuits with the same number G, forming the outputs of the second-level circuits with this number. The cut-out multiplexers Mp of the third layer combine the outputs of the second dimension circuits with the same numbers G, forming the outputs of the third-level circuits with this number. The cut-out multiplexers Mp of the fourth layer combine the outputs of the third dimension circuits, forming the outputs of the switch S_8N_8.

Switches S_8N_8 have the following property.

Lemma 4. *The dual switch* S_8N_8 *is a non-blocking switch with static routing on any common permutation for any p. It has* $(\sigma - 1)$*-channel fault tolerance.*

P r o o f. The first statement is based on using the switch S_4N_4 and Lemma 3. The second statement is based on the non-blocking property of the switch S_4N_4 and the fact that the p-permutation on the cross-section consists of sparse 1-permutations separated to different channels and cycles.

Channel fault-tolerance holds since the paths between sources and sinks in the switch S_4N_4 are through different circuits of each dimension and $p \ge \sigma$.

The resulting non-blocking self-routing switch S_8N_8 includes $N_8 = N_4^2$ channels and is $(\sigma - 1)$-channel fault-tolerant. ♦

The switch S_8N_8 has eight layers of output multiplexers Mp, $V_8 = N_8(p^8 - 1)/(p - 1)$ in total.

Tables 8–10 present the performance characteristics of the fastest modifications of the switch S_8N_8 for $\sigma = 2$ and $\sigma = 3$. The switching and channel complexities are calculated by the recursive formulas $S_8 = N_4S_4 + p^4N_4S_4$ and $L_8 = N_4L_4 + p^4N_4L_4$, respectively. Note that the switching and channel complexities of the switch S_8N_8 based on a digraph can be made significantly less than those of a lumped switch with the complete graph topology.

Table 8

Performance characteristics of dual switches S_8N_8 with one-channel fault tolerance

p	N_1	$N_8 = N_1^8$	$T_8 = p$	S_8	L_8
2	2	256	2	739 840 = $N_8^{2.44}$	304 640 = $N_8^{2.28}$
3	4	65 536	3	4.998E+09 = $N_8^{2.01}$	1.451E+09 = $N_8^{1.9}$
4	7	5 764 801	4	4.937E+12 = $N_8^{1.88}$	1.108E+12 = $N_8^{1.78}$

Table 9

Performance characteristics of dual switches S_8N_8 with two-channel fault tolerance

p	N_1	$N_8 = N_1^8$	$T_8 = p$	S_8	L_8
3	3	6 *561*	3	500 341 860 = $N_8^{2.28}$	145 260 540 = $N_8^{2.14}$
4	5	390 625	4	3.345E+11 = $N_8^{2.06}$	7.509E+10 = $N_8^{1.94}$
5	7	5 764 801	5	3.331E+13 = N_8^{2}	6.099E+12 = $N_8^{1.89}$

Table 10

Performance characteristics of dual switches S_8N_8 based on the digraph topology

p	N_1	$N_8 = N_1^8$	$T_8 = p$	S_8	L_8
2	4	65 536	2	189 399 040 = $N_8^{1.72}$	77 987 840 = $N_8^{1.64}$
3	9	5 764 801	3	3.283E+12 = $N_8^{1.64}$	9.531E+11 = $N_8^{1.57}$
4	16	4.29E+09	4	3.678E+15 = $N_8^{1.62}$	8.256E+14 = $N_8^{1.55}$

5. ANALYSIS OF THE RESULTS. PRACTICAL DEVELOPMENT OF THE CONSTRUCTED NETWORKS

This paper has proposed a method for constructing a new class of non-blocking self-routing photon networks with high scalability. These are the so-called dual networks based on a non-blocking dual $p\times p$ switch with a signal period of p cycles.

The dual switch is used as an integral part of the non-blocking self-routing $N_1\times N_1$ switch S_1N_1 with the quasi-complete graph or digraph topology. In the former case, the number of channels is $N_1 = p(p - 1)/\sigma + 1$, and $(\sigma - 1)$-channel fault tolerance can be provided. In the latter case, the number of channels is $N_1 = p^2$, which can be even increased. The switch with the quasi-complete (di)graph topology consists of N_1 dual $p\times p$ switches together with N_1 $1\times p$ demultiplexers Dp and $p\times 1$ multiplexers Mp without delay lines. The switching complexity of S_1N_1 is given by $S_1 = N_1(S_0 + 2p)$. The signal period T_1 of the switch S_1N_1 equals that of the dual switch: $T_1 = p$.

Switches S_1N_1 form two stages to construct a blocking $N_2\times N_2$ network N_2N_2 with $N_2 = N_1^2$ channels. Each stage consists of N_1 $N_1\times N_1$ switches, and the channels between the stages are built using exchange links. The network N_2N_2 is transformed into a non-blocking self-routing two-stage switch S_2N_2 by one-dimensional internal parallelization.

If the $N_1\times N_1$ switch is based on a quasi-complete graph, the switch S_2N_2 has $(\sigma - 1)$-channel fault tolerance since $p \geq \sigma$. The switching and channel complexities of the switch S_2N_2 are given by the recursive formulas $S_2 = N_1S_1 + pN_1S_1$ and $L_2 = N_1L_1 + pN_1L_1$, respectively. By construction, the signal period T_2 of the switch S_2N_2 equals that of the switch S_1N_1: $T_2 = T_1 = p$.

If the $N_1\times N_1$ switch is based on a quasi-complete graph, the four-stage switch S_4N_4 has $(\sigma - 1)$-channel fault tolerance since $p \geq \sigma$. The switching and channel complexities of the switch S_4N_4 are given by the recursive formulas $S_4 = N_2S_2 + p^2N_2S_2$ and $L_4 = N_2L_2 + p^2N_2L_2$, respectively. By construction, the signal period T_4 of the switch S_4N_4 equals that of the switch S_2N_2: $T_4 = T_2 = p$.

Similarly, switches S_4N_4 form two stages to construct a blocking $N_8\times N_8$ network N_8N_8 with $N_8 = N_4^2 = N_1^8$ channels. Each stage consists of N_4 switches S_4N_4, and the channels between the stages are built using exchange links.

The network N_8N_8 is transformed into a non-blocking self-routing two-stage switch S_8N_8 by four-dimensional internal parallelization.

If the $N_1\times N_1$ switch is based on a quasi-complete graph, then the eight-stage switch S_8N_8 has $(\sigma - 1)$-channel fault tolerance as well. The switching and channel complexities of the switch S_8N_8 are given by the recursive formulas $S_8 = N_4S_4 + p^4N_4S_4$ and $L_8 = N_4L_4 + p^4N_4L_4$, respectively. By construction, the signal period T_8 of the switch S_8N_8 equals that of the switch S_4N_4: $T_8 = T_4 = p$.

The performance characteristics of the switches S_2N_2, S_4N_4, and S_8N_8 have several degrees of freedom. First of all, the number of channels grows with increasing the base p, and the speed decreases. In addition, the exponential complexity decreases with increasing the base p, and the speed can be traded for complexity. Also, more channels due to increasing the number of stages reduce the exponential complexity.

The proposed method allows constructing non-blocking self-routing networks with a self-similar structure. The switch S_2N_2 consists of dual switches S_1N_1 with the dual graph or digraph topology and uses one-dimensional internal parallelization. In turn, the switch S_4N_4 consists of switches S_2N_2 and uses two-dimensional internal parallelization. Finally, the switch S_8N_8 is composed of switches S_4N_4 and uses four-dimensional internal parallelization. All these switches inherit the basic properties of the switch S_1N_1, such as the non-blocking property under static self-routing and channel fault tolerance (if necessary), but with significantly less complexity.

The high scalability of non-blocking switches can also be achieved by repeated application of the invariant extension method to the switch S_1N_1 with the digraph topology based on a conventional $p\times p$ switch. Such extended switches have a signal period of one cycle but increased complexity. Table 11 compares the switching complexity of dual switches S_4N_4 and S_8N_8 and extended switches S_1N_1. Clearly, the switching complexity of dual switches is by several orders of magnitude lower.

Note that the dual switch resolves conflicts by the bus method only in the first stage of the switch S_1N_1. All other conflicts in all stages are prevented using internal parallelization, and the dual switches in them are used as common $p\times p$ switches. Therefore, it seems reasonable to use the dual switch in its original form [1–3] (the multiplexer–demultiplexer pair): its switching complexity is p times less. This approach will reduce the switching complexity of the dual switches S_4N_4 and S_8N_8 by several times (1.5–4.5).

Note that for small p, the complexity of the fault-tolerant switches S_2N_2 and S_4N_4 is greater than that of the complete graph; for large p, it is smaller. In this

Table 11

Complexity analysis: dual switches vs. extended digraphs

Switching complexities of non-blocking four-stage switches ($S_{4,\,D}$) and extended switches based on the quasi-complete digraph topology (S_{PO})				
p	N_4	Dual switch S_4N_4 $S_{4,\,D}$	Extended switch S_1N_1 S_{PO}	Ratio $S_{PO}/S_{4,\,D}$
2	256	46 080 = $N_4^{1.94}$	261 120 = $N_4^{2.25}$	5.67
3	6 561	6 298 560 = $N_4^{1.78}$	129 120 480 = $N_4^{2.12}$	20.5
4	65 536	22 282 400 = $N_4^{1.74}$	11 453 071 360 = $N_4^{2.09}$	514
Switching complexities of non-blocking eight-stage switches ($S_{8,\,D}$) and extended switches based on the quasi-complete digraph topology (S_{PO})				
p	N_8	Dual switch S_8N_8 $S_{8,\,D}$	Extended switch S_1N_1 S_{PO}	Ratio $S_{PO}/S_{8,\,D}$
2	65 536	18 939 9040 = $N_8^{1.72}$	17 179 607 040 = $N_8^{2.12}$	85.7
3	43 046 721	3.283E+12 = $N_8^{1.64}$	5.55822E+15 = $N_8^{2.06}$	1 640
4	4.29E+09	3.678E+15 = $N_8^{1.62}$	4.91906E+19 = $N_8^{2.04}$	13 107

case, the complexity of switches S_8N_8 is significantly less than that of the complete graph for any p. We emphasize the performance characteristics of the switch S_8N_8 for $p = 2$ and $\sigma = 1$. With $N_8 = 65\,536$ channels and half the speed, its switching complexity is comparable to that of a five-stage non-blocking Clos network based on a 64-channel YARC router [9] with $N = 32\,768$ channels constructed as a non-blocking network [7, 8]. The complexity of this non-blocking Clos network is estimated as $S = N^{1.73}$. However, this network has no parallel static or dynamic self-routing procedures. The other switches S_8N_8 with $p > 2$ and $\sigma = 1$ have even lower switching complexity and higher scalability but with lower speed.

The p-times reduced speed of the switches S_2N_2, S_4N_4, and S_8N_8 can be compensated by different protocols. It is possible to choose processors with p independent ports, divide packets into p parts and transmit them in parallel. The high scalability of these switches supports such an operation mode, albeit by reducing the number of users by p times and increasing the network complexity. An alternative is to apply the parallel-serial method for transmitting packets over p lines, as in the PCI Express protocol, without reducing the number of users.

A shortcoming of the switches S_1N_1, S_2N_2, S_4N_4, and S_8N_8 is their optimization for the conflict-free implementation of arbitrary permutations. What will be their behavior on arbitrary traffic? To determine it, we can assign the one-channel property to the multiplexers in the output stages. When receiving several input packets, such a multiplexer passes only one packet and blocks the others. The blocked packets not acknowledged by sinks are re-transmitted by the sources.

A considerable disadvantage of the proposed method is the need for parallel transmission of signal and control information, which significantly increases the required bandwidth. This disadvantage is not fatal for photon switches since an optical cable can simultaneously carry hundreds of different frequencies. However, this disadvantage can be generally eliminated with bit synchronization of signals from different channels. This can be done using the method [34, 35], locating the mutual arrangement of sources and sinks and the corresponding transmission delays from the sources. In this case, control information for dual switches and demultiplexers can be transmitted, as usual, in the form of sets of bits in the packet header.

A pleasant bonus of bit synchronization is the ability to construct an arithmetic logic unit (ALU) in the channel at each sink's input using network means. Such ALUs were developed for computing in the common channel [36]. For implementing them, it is necessary to transmit through the channel the digit values in the two-signal form: along two lines with active signals for values 0 and 1 in each. In the network ALU, an operation is performed over the number arriving through the channel and the number at the sink; the result is formed in the channel after the ALU. In the channel, it is possible to perform addition, multiplication, and any bitwise logical operations, including finding the maximum (minimum).

CONCLUSIONS

This paper has proposed a method for constructing non-blocking fault-tolerant photon networks with high scalability, considered in [5], but with much less complexity. This method is based on three main components:

- A p-channel dual switch with a signal period of p cycles, which turns out to be non-blocking on any input traffic (a prerequisite for constructing more complex non-blocking networks).
- A switch with the quasi-complete graph or digraph topology and a dual switch inside. As a result, the non-blocking property is maintained, and channel fault tolerance and higher scalability during cascading are provided compared to a pure dual switch.
- Internal parallelization to maintain the non-blocking property by preventing conflicts and maintaining fault tolerance, which provides high scalability when cascading non-blocking networks.

In the paper [5], scaling was implemented by cascade application of the invariant extension method with additional external multiplexers and demultiplexers. In this paper, scaling has been implemented by cascading smaller non-blocking networks and applying the generalized internal parallelization method at each cascading step.

The cascading of a non-blocking network with N channels is performed by constructing a blocking network with N^2 channels. This network consists of two stages with exchange links with N original non-blocking networks in each. Interlocks in this two-stage network occur at the output multiplexers of the first stage. These interlocks are prevented by separating the conflicting channels to multiple copies of the second stage and moving the multiplexers to the outputs by the second stage part responsible for packet routing. No conflicts occur in this part of the network since it consists of copies of non-blocking subnetworks that route sparse permutations. Sparse permutations are united into a complete permutation on a network with N^2 channels by the moved stages of multiplexers without conflict.

During the first cascading [5], internal parallelization is performed using p second stage copies and a one-layer stage of output multiplexers. When constructing a non-blocking network with N^4 channels, the second cascading is performed using p^2 second stage copies and a two-layer stage of the output multiplexers. When constructing a non-blocking network with N^8 channels, the third cascading is performed using p^4 second stage copies and a four-layer stage of the output multiplexers. Thus, we have designed non-blocking two-, four-, and eight-stage networks with stages consisting of non-blocking dual networks with the quasi-complete graph or digraph topology.

During each cascading, internal parallelization maintains the signal period and reduces the specific complexity of the non-blocking network. In particular, we have constructed non-blocking networks with a specific complexity not exceeding that of the theoretical non-blocking Clos network.

This method can be a fundamental base for constructing practical non-blocking switches with high scalability, static self-routing, and channel fault tolerance.

REFERENCES

1. Barabanova, E.A., Vytovtov, K.A., and Podlazov, V.S., Multistage Switches for Optical and Electronic Supercomputer Systems, *Proceedings of the 8th National Supercomputer Forum* (*NSCF–2019*), *Pereslavl-Zalessky*, 2019. URL: http://2019.nscf.ru/TesisAll/02_Apparatura/037_BarabanovaEA.pdf. (In Russian.)
2. Barabanova, E.A., Vytovtov, K.A., Vishnevsky, V.M., and Podlazov, V.S., The New Principle for the Construction of Optical Information Processing Devices for Information-Measuring Systems, *Sensors and Systems*, 2019, no. 9, pp. 3–9. (In Russian.)
3. Barabanova, E., Vytovtov, K., Podlazov, V., and Vishnevskiy, V. Model of Optical Non-blocking Information Processing System for Next-generation Telecommunication Networks, *Proceedings of the 22nd International Conference on Distributed Computer and Communication Networks: Control, Computation, Communications* (*DCCN–2019*), Moscow, 2019. Communications in Computer and Information Science, vol. 1141, Cham: Springer, pp. 188–198. DOI: 10.1007/978-3-030-36625-4_16.
4. Karavai, M.F. and Podlazov, V.S., An Invariant Extension Method for System Area Networks of Multicore Computational Systems. An Ideal System Network, *Automation Remote Control*, 2010, vol. 71, no. 12, pp. 2644–2654.
5. Barabanova, E.A., Vytovtov, K.A., and Podlazov, V.S., Two-Stage Dual Photon Switches in an Extended Scheme Basis, *Control Sciences*, 2021, no. 1, pp. 69—81.
6. Barabanova, E.A., Vytovtov, K.A., Podlazov, V.S., Non-blocking Fault-Tolerant Two-Stage Dual Photon Switches, *Control Sciences*, 2021, no. 4, pp. 67–76.
7. Clos, C., A Study of Non-locking Switching Networks, *Bell System Tech. J.*, 1953, vol. 32, pp. 406–424.
8. Benes, V.E., *Mathematical Theory of Connecting Networks and Telephone Traffic*, New York: Academic Press, 1965.
9. Scott, S., Abts, D., Kim, J. and Dally, W., The Black Widow High-radix Clos Network, *Proc. of the 33rd International Symposium on Computer Architecture* (*ISCA'2006*), Boston, 2006. https://www.researchgate.net/publication/4244660_The_Black Widow_High-Radix_Clos_Network.
10. De Sensi, D., Di Girolamo, S., McMahon, K.H., Roweth, D., and Hoefler, T., An In-Depth Analysis of the Slingshot Interconnect, *arXiv: 2008.08886v1*, August 20, 2020. https://www.researchgate.net/publication/343786515_An_In-Depth_Analysis_of_the_Slingshot_Interconnect.
11. Alverson, R., Roweth, D., and Kaplan, L., The Gemini System Interconnect, *Proceedings of the 18th IEEE Symposium on*

High Performance Interconnects, Santa Clara, CA, 2009, pp. 83–87.
12. Alverson, R., Roweth, D., Kaplan, L., and Roweth, D., Cray XC® Series Network. http://www.cray.com/Assets/PDF/products/xc/CrayXC30Networking.pdf.
13. Kim, J., Dally, W. J., Scott, S., and Abts, D., Technology-Driven, Highly-Scalable Dragonfly Topology, *Proceedings of the 35th Annual International Symposium on Computer Architecture (ISCA'2008)*, Beijing, 2008, pp. 77–88. http://users.ece.gatech.edu/~sudha/academic/class/Networks/Lectures/4%20-%20Topologies/papers/dragonfly.pdf.
14. *Mellanox OFED for Linux User Manual. Rev 2.3-1.0.1*, Mellanox Technologies, 2014. https://dlcdnets.asus.com/pub/ASUS/mb/accessory/PEM-FDR/Manual/Mellanox_OFED_Linux_User_Manual_v2_3-1_0_1.pdf.
15. Pipenger, N., On Rearrangeable and Non-blocking Switching Networks, *J. Comput. Syst. Sci.*, 1978, vol. 17, pp. 307–311.
16. Bhuyan, L.N. and Agrawal, D.P., Generalized Hypercube and Hyperbus Structures for a Computer Network, *IEEE Trans. on Computers*, 1984, vol. C-33, no. 4, pp. 323–333.
17. Tzeng, N. and Wei, S., Enhanced Hypercubes, *IEEE Trans. Computers*, 1991, vol. 40, no. 3, pp. 284–294.
18. Efe, K., A Variation on the Hypercube with Lower Diameter, *IEEE Trans. Computers*, 1991, vol. 40, no. 11, pp. 1312–1316.
19. Kim, J. and Dally, W.J., Flattened Butterfly Topology for On-Chip Networks, *IEEE Computer Architecture Letters*, 2007, vol. 6, no. 2, pp. 37–40.
20. Gu, Q.P., and Tamaki, H., Routing a Permutation in Hypercube by Two Sets of Edge-Disjoint Paths, *J. of Parallel and Distributed Comput.*, 1997, vol. 44, no. 2, pp. 147–152.
21. Lubiw, A. Counterexample to a Conjecture of Szymanski on Hypercube Routing, *Inform. Proc. Let*., 1990, vol. 35(2), pp. 57–61.
22. Stepanenko, S., Structure and Implementation Principles of a Photonic Computer, *EPJ Web of Conferences*, vol. 224, 2019. DOI: https://doi.org/10.1051/epjconf/201922404002.
23. Zhabin, I.A., Makagon, D.V., Polyakov, D.A., Simonov, A.S., Syromyatnikov, E.L., and Shcherbak, A.N., First Generation of Angara High-Speed Interconnection Network, *Science Intensive Technologies*, 2014, no. 1, pp. 21–27. (In Russian.)
24. Stegailov, V., Agarkov, A., Biryukov, V., et al., Early Performance Evaluation of the Hybrid Cluster with Torus Interconnect Aimed at Molecular Dynamics Simulations, *Proceedings of the International Conference on Parallel Processing and Applied Mathematics*, Cham: Springer, 2017, pp. 327–336.
25. Ajima, Y., Inoue, T., Hiramoto, S., and Shimiz, T., Tofu: Interconnect for the K Computer, *Fujitsu Scientific & Technical Journal*, 2021, vol. 48, no. 3, pp. 280–285. https://www.researchgate.net/publication/265227674_Tofu_Interconnect_for_the_K_computer.
26. Arimili, B., Arimilli, A., Chung, V., et al., The PERCS High-Performance Interconnect, *Proceedings of the 18th IEEE Symposium on High Performance Interconnects*, New York, 2009, pp. 75–82.
27. Kathareios, G., Minkenberg, C., Prisacari, B., et al., Cost-Effective Diameter-Two Topologies: Analysis and Evaluation, *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'15)*, 2015, pp. 1–11.
28. Besta, M. and Hoefler, T., Slim Fly: A Cost Effective Low-Diameter Network Topology, *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'14)*, 2014, pp. 348–359
29. Flajslik, M., Borch, E., and Parker, M.A., Megafly: A Topology for Exascale Systems, in *High Performance Computing*, Yokota, R., Weiland, M., Keyes, D., and Trinitis, C., Eds., Cham: Springer, 2018, pp. 289–310.
30. Ahn, J.H., Binkert, N., Davis, A., et al., Hyperx: Topology, Routing, and Packaging of Efficient Large-Scale Networks, *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'09)*, 2009, pp. 1–11.
31. Domke, J., Matsuoka, S., Ivanov, I.R., et al., Hyperx Topology: First At-scale Implementation and Comparison to the Fat-Tree, *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'19)*, New York, Association for Computing Machinery, 2019.
32. Singla, A., Hong, C.-Y., Popa, L., and Godfrey, P.B., Jellyfish: Networking Data Centers Randomly, *The 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, 2012, San Jose, CA, USENIX, pp. 225–238.
33. Hall, M., *Combinatorial Theory*, Waltham: Blaisdell Publishing Company, 1967.
34. Stetsyura, G.G., Computer Network with the Fast Distributed Reorganization of Its Structure and Data Processing During Their Transmission, *Control Sciences*, 2017, no. 1, pp. 47–56. http://pu.mtas.ru/archive/Stetsyura_117.pdf (In Russian.)
35. Stetsyura, G.G., The Computer Clusters with Fast Synchronization of Messages and with Fast Distributed Computing by the Network Hardware, *Control Sciences*, 2020, no. 4, pp. 61–69. (In Russian.)
36. Prangishvili, I.V., Podlazov, V.S., and Stetsyura, G.G., *Lokal'nye mikroprocessornye vychislitel'nye seti. Glava 6* (Local Microprocessor Computing Networks. Chapter 6), Moscow: Nauka, 1984. (In Russian.)

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

Received March 25, 2021, and revised August 12, 2021.
Accepted August 24, 2021.

Author information

Podlazov, Viktor Sergeevich. Dr. Sci. (Eng.), Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ podlazov@ipu.ru

Cite this article
Podlazov, V.S. Non-blocking Fault-Tolerant Dual Photon Switches with High Scalability. *Control Sciences* **5,** 61–76 (2021). http://doi.org/10.25728/cs.2021.5.6

Original Russian Text © Podlazov, V.S., 2021, published in *Problemy Upravleniya*, 2021, no. 5, pp. 70–87.

Translated into English by *Alexander Yu. Mazurov*,
Cand. Sci. (Phys.–Math.),
Trapeznikov Institute of Control Sciences,
Russian Academy of Sciences, Moscow, Russia
✉ alexander.mazurov08@gmail.com

DOI: http://doi.org/10.25728/cs.2021.5.7

SCENARIO METHODS TO IMPROVE THE EFFICIENCY OF IMPLEMENTING THE LIFE CYCLE OF PROGRAM-TARGET MANAGEMENT: A CONCEPTUAL ANALYSIS

I.V. Chernov

Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia

✉ ichernov@gmail.com

Abstract. This paper is devoted to methodological and applied problems of improving the efficiency of program-target planning and management in socio-economic systems based on a scenario approach. The main features of the methodology and mechanisms of program-target management are considered. The idea proposed below is to manage and control the implementation of target programs using a scenario approach and life cycle models. A scenario W-shaped life cycle model is developed to improve the efficiency of program-target planning and management. This model involves the end-to-end methodology of proactive management to implement long-term socio-economic programs under uncertainty and risk. The model's methodological core consists of the following mechanisms: scenario analysis, simulation, forecasting, planning, and group management. The new approach is most effective under increasing uncertainty due to its focus on anticipating future conditions and alternatives for the development of socio-economic systems.

Keywords: planning, management, efficiency, target program, scenario approach, life cycle, uncertainty.

INTRODUCTION

Strengthening of negative processes in world development in the 2010s, the deepening of the global economic crisis caused by the COVID-19 pandemic, the destruction of international economic ties, the strengthening of the anti-Russian sanctions policy of the Western countries, the deterioration of budget indicators related to internal problems, and the decline in production have resulted in significant growth of various risks and uncertainties. All these factors tighten the requirements for the quality and efficiency of management of socio-economic systems (SESs), considerably narrowing the scope of conventional approaches and methods to plan and manage the sustainable development of such systems.

In the current unstable conditions, scenario analysis and simulation methods based on the program-target approach increase their role and importance. These methods reduce uncertainty and consider a wide range of external and internal threats to achieving the goals by assessing the most probable and reasonable trends in the development of dynamic processes in the sectors under consideration, their stability, and other desired and undesired properties based on information about their structural features. Note that the sectors are characterized by a complex spatial, administrative, and managerial organization of SESs.

1. IMPROVING THE MECHANISMS OF PROGRAM-TARGET PLANNING AND MANAGEMENT

Program-target planning and management is a special public administration method intended primarily for the critical development areas of national SESs that combines the interests of society, the state, and economic entities of various forms of ownership. This methodology allows solving large-scale and complex problems of SESs development by structuring and interconnecting their components and organizing mutually beneficial partnerships between the state, business, science, education, and civil society [1].

To a large extent, program-target planning relies on the hierarchy analysis methodology [2, 3] to achieve the main goal by decomposing it into a system of intermediate goals (subgoals). The subgoals are achieved by implementing a set of interconnected measures (programs) by a group of executors.

The complexity of increasing the efficiency of program-target planning and management is due to different methods and approaches to solving management problems at particular stages of the process. In addition, the implementation of any large-scale, long-term project is associated with numerous variable and newly emerging risks and external and internal threats of various natures to be considered during program-target planning and management. These risks and uncertainties are different and significantly affect the problem specifics at different stages of program-target management.

Large-scale projects and programs involve many economic entities of various forms of ownership from different industries. They directly participate in the measures of state or federal target programs or (directly or indirectly) work to achieve the main goal on their initiative based on the state's interest in particular problems of socio-economic development. (In the latter case, they expect the economic demand for the results of such work.) The economic entities possibly use market mechanisms of economic management [4].

In this situation, when using conventional approaches to management, there is a threat of a reduced socio-economic effect from implementing particular projects and programs due to possible errors in the process of analysis, forecasting, and goal-setting, determining the real needs for resources, improper coordination in the activities of the managing substructures and executors of programs.

A way to improve the efficiency of program-target planning and management is to develop and implement a methodology based on the process and scenario approaches [5]. This methodology should reduce uncertainty and consider a wide range of external and internal threats to achieving the goals at different stages of designing and implementing target programs and projects. In addition, this methodology should effectively coordinate and control the activities of numerous economic entities for executing program measures while solving strategic development problems characterized by a complex spatial, administrative, and managerial organization of socio-economic systems.

2. ANALYSIS OF MANAGEMENT PROCESSES FOR IMPLEMENTING TARGET PROGRAMS BASED ON LIFE CYCLE MODELS

Any target program goes through several life cycle stages: identifying problems that need to be solved, forming a system of goals and efficiency criteria, planning and developing mechanisms to achieve the goals under uncertainty (hard-to-predict changes in the internal and external environments), managing the implementation of program measures (particularly by exception), assessing the results, and completing the program (the last stage).

Nowadays, the term "life cycle" (LC) is widely used in natural sciences, engineering, humanities, and other sciences. We note the structural and, roughly speaking, conceptual similarity of the LC definitions in application domains [6]. A systems engineering definition of the complete LC model was given in the book [7]: "The complete life cycle model of an individual object is a description of the sequence of all phases and stages of its existence from conception and emergence ("birth") to disappearance ("extinction")." Consequently, the complete life cycle model is characterized by structural invariance: the set of the LC phases does not significantly depend on the object described. Due to this invariance, the model becomes quite universal and widely applicable since, in the general case, almost any control object (a program or project) goes through all the main LC phases: design, implementation and development, and completion. The most common LC parameters are the time instants of beginning and end, the total duration, the duration of each phase (stage, step, task, etc.), the sequence of stages and steps, the type and form of intraphase changes, and the set of indicators characterizing the object's state (by phases, stages, or steps). Each LC stage has particular managerial tasks and features.

At its core, the LC model is an applied organizational management tool based on a systems approach that allows:

- ensuring compliance with the goals and an adequate understanding of management processes by representing the final results as a cumulative outcome of intermediate tasks that are easier to understand and assess;
- decomposing the management process into relatively independent and, at the same time, logically interconnected temporal blocks;

- structuring the management process;
- determining step by step the content of managerial tasks, the technology for their solution, partial efficiency criteria, and resource-time constraints, considering the goals and the object's dynamics;
- as a consequence, improving the methodological base and organizational principles of the management system.

Various LC models are currently used for solving a wide class of problems. Among them, we mention the cascade, incremental, spiral, V-shaped, and other models. Each model is focused on particular application domains and, accordingly, has some advantages and disadvantages. Perhaps, only the V-shaped model agrees with the tasks and features of program-target management [6].

Note that program-target management has the following specifics: for a long life cycle of programs (3–5 years or more), a complex set of interconnected but different problems should be solved under uncertainty, risk, and continuous changes in the external and internal environments. (For example, this set includes goal-setting, risk analysis, forecasting, design of program measures, financial and resource support, operational management of the implementation of measures, monitoring and assessment of the program's result and the efficiency of using allocated resources, to name a few.) Moreover, ineffective decisions at any stage will entail a whole "bunch" of problems at the subsequent ones (in the worst case, return to the initial stages of goal-setting and planning to make significant corrections affecting the entire management cycle).

At present, most organizational management systems of various levels and designated purposes focus on the object's internal development processes. As a result, they solve mainly operational and partially tactical tasks with maximum success. For the same reason, as the input data they use difficult-to-predict and, in many cases, significant changes in the internal environment and almost all events outside the system (in the external environment).

For such management systems, the "negative" input information requiring an immediate response is mainly the results of direct (external or internal) influence on the object when, in most practical cases, the damage has been done (in the broadest sense of this term). If the management system is limited to counteracting the consequences of unfavorable situations, it cannot guarantee an efficient solution for strategic and medium-term tasks of managing the socio-economic development of the state and society. All these circumstances are the main source of reducing the efficiency of management processes for developing socio-economic systems.

In these conditions, the role and importance of the scenario approach in program-target management increase. This approach yields a predictive assessment of the quality of decisions made at various stages of the life cycle of large projects and programs. Moreover, it reduces uncertainty by considering a wide range of external and internal threats to achieving goals. Uncertainty is understood as a situation when information about the structure and possible states of the SES and (or) its external environment is partially or completely absent. This concept is one of the key methodological concepts of the scenario approach. The construction of scenarios pursues two goals: reducing the original uncertainty as much as possible within this approach and describing the residual uncertainty using some scenarios. Thus, the uncertainty in situation development is subsequently decreased to find the best solutions of the arising problems.

Scenario analysis, forecasting, planning, and group management are technologies to assess comprehensively the efficiency and consistency of a set of managerial decisions when selecting and implementing the development programs of SESs and to consider and coordinate the goals and activities of the economic entities involved in the implementation of program measures. They independently plan and execute their tactical actions in the middle- and short-term horizons within common long-term goals, rules, and obligations regulated by the relevant target programs and projects.

For increasing the efficiency of program-target planning and management, we modify the V-shaped life cycle model by adding a branch that reflects:

– the methodological basis for ensuring the sustainable progressive socio-economic development of the state and society under uncertainty requiring a response to unpredictable changes in the internal environment,

– the influence of negative phenomena and processes occurring in the external environment; see Fig. 1.

In the classical V-shaped life cycle model, special importance is attached to actions and procedures to verify and attest the results of planning and project management. Along with traditional analytical technologies, the W-shaped model yields a predictive assessment of the final goals and most significant intermediate results for them. Hence, we pass from the passive consideration of already accomplished events to proactive management of the implementation of target programs.

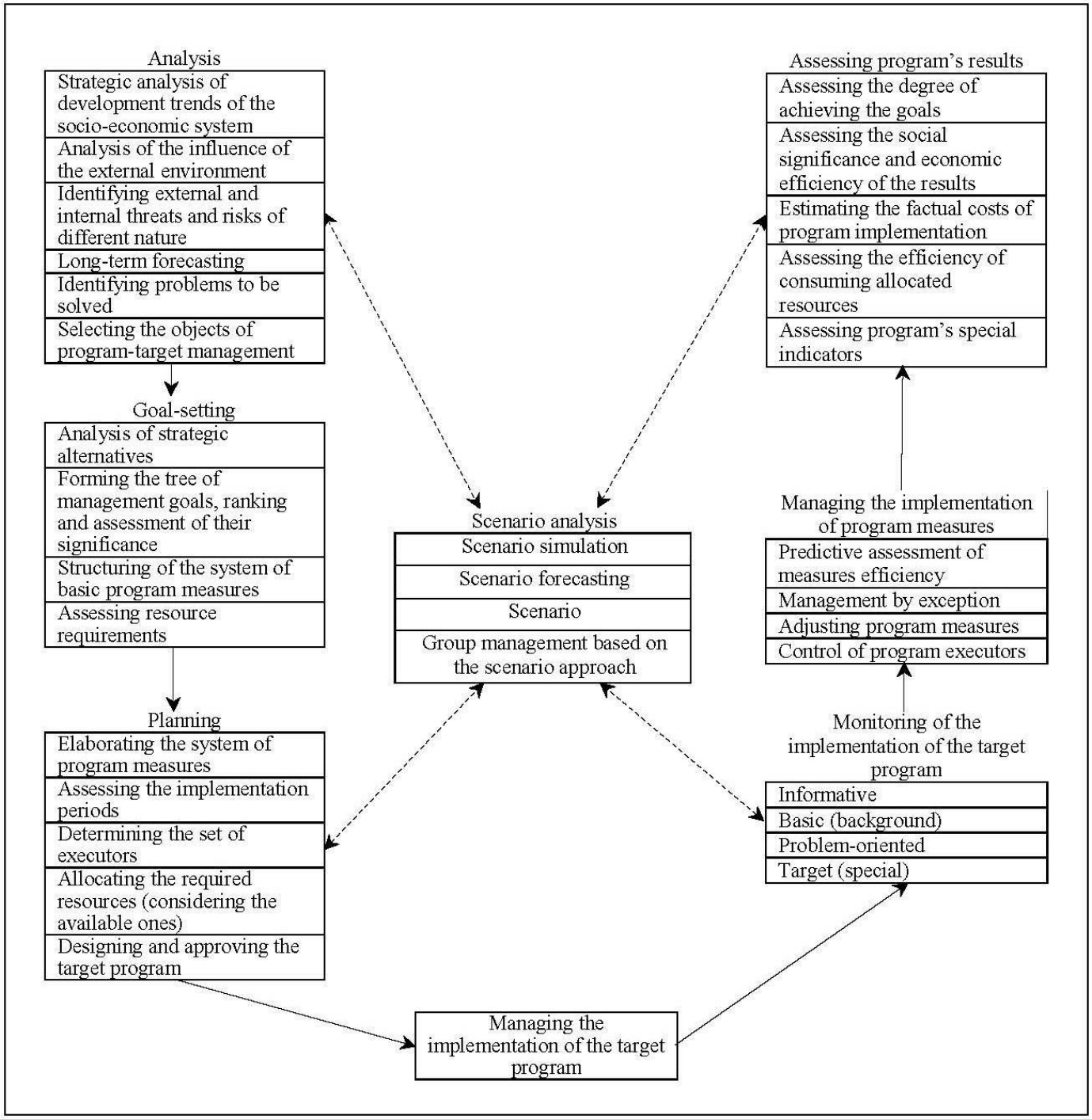


Fig. 1. The life cycle of program-target planning and management: W-shaped model.

The end-to-end methodology of scenario planning and group management of complex systems under uncertainty increases the efficiency of managing groups of objects within complex SESs (economic entities) by forming and analyzing a wide range of scenarios for the development of these systems and selected sets of their segments (including objects) that consider the influence of the external environment and associated risks of various nature.

The new approach is most efficient in current conditions of growing uncertainty, which forms the need for tools and mechanisms of scenario analysis, simulation, forecasting, planning, and group management. To a large extent, they allow foreseeing future conditions

and corresponding alternatives for the development of SESs on different (mostly long) time horizons, identifying external and internal threats and risks, and exploring possible paths of situation development in the internal and external environments.

When elaborating managerial decisions, the methodology under consideration reduces the object of analysis to a finite number of alternatives (including the most probable ones) that reflect the most significant risks and threats to the goals at any stage of the target program life cycle. Moreover, this methodology allows developing optimistic and pessimistic scenarios and making forecasts for the development of given segments of SESs at different management levels, adjusting the strategic vision of the situation and highlighting the desired directions of its development, and assessing the level of coordination between the decisions in different stages of the life cycle of target programs and large projects.

CONCLUSIONS

Currently, considerable experience has been accumulated in developing a formal methodology for the analysis and design of models and methods that implement the scenario approach and in solving applied problems in the field of analysis, forecasting, planning, and management of the development of different segments of socio-economic systems. However, most of the solutions focus on the specifics of particular tasks and are limited by their scope. At the same time, there are almost no publications on the integrated application of the scenario approach:

– to increase the efficiency of program-target management for developing and implementing large-scale spatially distributed projects throughout their life cycle,

– to improve the quality and coordination of managerial decisions under uncertainty and risks of different nature.

Further fundamental and applied research in this area will solve a wide range of tasks to improve the efficiency of program-target planning and management processes and the implementation of priority national projects and socio-economic programs.

REFERENCES

1. Kulba, V.V., Shultz, V.L., Shelkov, A.B., and Chernov, I.V., The Efficiency Management Methods for the Implementation of Social and Economic Target Programs, *Trends and Management*, 2013, no. 4, pp. 4–28. (In Russian.)
2. Saaty, T.L., *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*, New York: McGraw-Hill, 1980.
3. Saaty, T.L. and Kearns, K.P., *Analytical Planning: The Organization of Systems*, Oxford: Pergamon, 1985.
4. *Upravlenie i kontrol' realizatsii sotsial'no-ekonomicheskikh tselevykh programm* (Implementation of Socio-economic Target Programs: Management and Control), Kul'ba, V.V. and Kovaleskii, S.S., Eds., Moscow: LIBROKOM, 2009. (In Russian.)
5.*Modeli i metody analiza i sinteza stsenariev razvitiya sotsial'no-ekonomicheskikh system* (Models and Methods to Analyze and Design the Development Scenarios of Socio-economic Systems), Shultz, V.L. and Kul'ba, V.V., Eds., Moscow: Nauka, 2012. (In Russian.)
6. Berg, D.B., Ul'yanova, E.A., and Dobryak, P.V., *Modeli zhiznennogo tsikla* (Life Cycle Models), Yekaterinburg: Ural Federal University, 2014. (In Russian.)
7. Burkov, V.N. and Irikov, V.A., *Modeli i metody upravleniya organizatsionnymi sistemami* (Models and Methods of Organizational Systems Control), Moscow: Nauka, 1994. (In Russian.)

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

Received August 16, 2021.
Accepted August 31, 2021.

Author information

Chernov, Igor Viktorovich. Cand. Sci. (Eng.), Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ ichernov@gmail.com

Cite this article

Chernov, I.V. Scenario Methods to Improve the Efficiency of Implementing the Life Cycle of Program-Target Management: A Conceptual Analysis, *Control Sciences* **5,** 77–81 (2021). http://doi.org/10.25728/cs.2021.5.7

Original Russian Text © Chernov, I.V., 2021, published in *Problemy Upravleniya*, 2021, no. 5, pp. 88–93.

Translated into English by *Alexander Yu. Mazurov*,
Cand. Sci. (Phys.–Math.),
Trapeznikov Institute of Control Sciences,
Russian Academy of Sciences, Moscow, Russia
✉ alexander.mazurov08@gmail.com