

ISSN: 1337-6365

© Slovak University of Technology in Bratislava

All rights reserved

**APLIMAT - JOURNAL OF APPLIED MATHEMATICS**

**VOLUME 5 (2012), NUMBER 2**



# **APLIMAT – JOURNAL OF APPLIED MATHEMATICS**

## **VOLUME 5 (2012), NUMBER 2**

**Edited by:** Slovak University of Technology in Bratislava

**Editor - in - Chief:** KOVÁČOVÁ Monika (Slovak Republic)

**Editorial Board:** CARKOVŠ Jevgenijs (Latvia )  
CZANNER Gabriela (Great Britain)  
CZANNER Silvester (Great Britain)  
DOLEŽALOVÁ Jarmila (Czech Republic)  
FEČKAN Michal (Slovak Republic)  
FERREIRA M. A. Martins (Portugal)  
FRANCAVIGLIA Mauro (Italy)  
KARPÍŠEK Zdeněk (Czech Republic)  
KOROTOV Sergey (Finland)  
LORENZI Marcella Giulia (Italy)  
MESIAR Radko (Slovak Republic)  
VELICHOVÁ Daniela (Slovak Republic)

**Editorial Office:** Institute of natural sciences, humanities and social sciences  
Faculty of Mechanical Engineering  
Slovak University of Technology in Bratislava  
Námestie slobody 17  
812 31 Bratislava

**Correspondence concerning subscriptions, claims and distribution:**

F.X. spol s.r.o  
Dúbravská cesta 9  
845 03 Bratislava 45  
journal@aplimat.com

**Frequency:** One volume per year consisting of three issues at price of 120 EUR, per volume, including surface mail shipment abroad.  
Registration number EV 2540/08

**Information and instructions for authors are available on the address:**

<http://www.journal.aplimat.com/>

**Printed by:** FX spol s.r.o, Azalková 21, 821 00 Bratislava

**Copyright © STU 2007-2012, Bratislava**

All rights reserved. No part may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without prior written permission from the Editorial Board. All contributions published in the Journal were reviewed with open and blind review forms with respect to their scientific contents.

# APLIMAT – JOURNAL OF APPLIED MATHEMATICS

## VOLUME 5 (2012), NUMBER 2

### DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS

<b>BAŠTINEC Jaromír, PIDUBNA Ganna:</b> CONTROLLABILITY FOR A CERTAIN CLASS OF LINEAR MATRIX SYSTEMS WITH DELAY	13
<b>CARKOVŠ Jevgenijs, SADURSKIS Karlis:</b> EQUILIBRIUM STOCHASTIC STABILITY OF MARKOV DYNAMICAL SYSTEMS	25
<b>CARKOVŠ Jevgenijs, SLYUSARCHUK Vasyl:</b> ON STABILITY ANALYSIS OF QUASILINEAR DIFFERENCE EQUATIONS IN BANACH SPACE(SPECTRAL THEORY APPROACH)	35
<b>ENSHAEIAN, Alireza, ROFOOEI, Fayaz R.:</b> GEOMETRICALLY NONLINEAR PLATES SUBJECTED TO A MOVING MASS	53
<b>FLOREA Olivia:</b> THE STUDY OF A NON LINEAR SYSTEM IN THE CASE WITH TWO OSCILLATING MASSES	63
<b>HALFAROVÁ Hana, KUKHARENKO Alexandra, ŠMARDÁ Zdeněk:</b> APPLICATION OF HOMOTOPY PERTURBATION METHOD TO SOLVING SINGULAR INITIAL VALUE PROBLEMS	69
<b>HRABALOVÁ Jana:</b> ON STABILITY INTERVALS OF EULER METHODS FOR A DELAY DIFFERENTIAL EQUATION	77
<b>KADERÁBEK Zdeněk:</b> THE AUTONOMOUS SYSTEM DERIVED FROM VAN DER POL-MATHIEU EQUATION	85
<b>KHAN Yasir, ŠMARDÁ Zdeněk:</b> SINGULAR INITIAL VALUE PROBLEM FOR IMPLICIT VOLTERRA INTEGRO-DIFFERENTIAL EQUATIONS	97
<b>KUBJATKOVÁ Martina, OLACH Rudolf, ŠTOBEROVÁ Júlia:</b> EXISTENCE OF NONOSCILLATORY SOLUTIONS OF DELAY DIFFERENTIAL EQUATIONS	103
<b>MURESAN Viorica:</b> DIFFERENTIABILITY WITH RESPECT TO DELAY OF THE SOLUTION OF A CAUCHY PROBLEM	111
<b>ROFOOEI Fayaz R., ENSHAEIAN Alireza:</b> DYNAMIC BEHAVIOR OF LARGE DEFORMABLE RECTANGULAR PLATES SUBJECTED TO A MOVING MASS GOVERNED BY NONLINEAR NON-HOMOGENOUS HILL EQUATION	119

<b>SVOBODA Zdeněk, DIBLÍK Josef, KHUSAINOV Denys:</b> SOME PROPERTIES OF SPECIAL DELAYED MATRIX FUNCTIONS IN THEORY OF SYSTEMS OF LINEAR DIFFERENTIAL EQUATIONS WITH CONSTANT COEFFICIENTS AND WITH SINGLE DELAY	<b>131</b>
<b>VÍTOVEC Jiří:</b> SOME GENERALIZATIONS IN THEORY OF RAPID VARIATION ON TIME SCALES AND ITS APPLICATION IN DYNAMIC EQUATIONS	<b>139</b>
<b>ŽÍDEK Arnošt, VLČEK Jaroslav, KRČEK Jiří:</b> SOLUTION OF DIFFRACTION PROBLEMS BY BOUNDARY INTEGRAL EQUATIONS	<b>147</b>

# APLIMAT – JOURNAL OF APPLIED MATHEMATICS

## VOLUME 5 (2012), NUMBER 2

### ENGINEERING APPLICATIONS AND SCIENTIFIC COMPUTATIONS

<b>BÍMOVÁ Daniela:</b> PARALLEL SOLUTION OF POISSON EQUATION	<b>157</b>
<b>BITTNEROVÁ Daniela:</b> BLOCK-CYCLIC-STRIPED MAPPINGS OF MATRICES IN THE PARALLEL PROGRAMMING	<b>169</b>
<b>ČERNÁ Dana, FINĚK Václav:</b> ON EFFICIENCY OF APPROXIMATE MATRIX-VECTOR MULTIPLICATION IN ADAPTIVE WAVELET METHODS	<b>173</b>
<b>ČERNÁ Dana, FINĚK Václav:</b> DISCRETE WAVELET TRANSFORM FOR FINITE SIGNALS	<b>181</b>
<b>FERDIÁNOVÁ Věra, HURTÍK Petr, KOLCUN Alexej:</b> RECONSTRUCTION OF THE BOREHOLE WALL FROM VIDEO	<b>191</b>
<b>HOZMAN Jiří:</b> DISCONTINUOUS GALERKIN METHOD FOR THE NUMERICAL SOLUTION OF OPTION PRICING	<b>197</b>
<b>POTŮČEK Radovan:</b> ELEMENTARY SOLUTION TO THE THREE VEHICLES JEEP PROBLEM WITH SUPPORT OF THE CAS MAPLE	<b>207</b>
<b>SÝKOROVÁ Irena:</b> ROUNDING ERRORS IN COMPUTER ARITHMETIC	<b>217</b>
<b>ŠIMEČEK Ivan, LANGR Daniel:</b> HETEROGENEOUS CLUSTER FOR ACCELERATION OF LINEAR ALGEBRA COMPUTATIONS	<b>225</b>

# **APLIMAT – JOURNAL OF APPLIED MATHEMATICS**

## **VOLUME 5 (2012), NUMBER 2**

### **FINANCIAL AND ACTUARY MATHEMATICS**

- HLADIKOVA Hana: TERM STRUCTURE MODELLING BY USING  
EXPONENTIAL BASIS FUNCTIONS 233

### **FUZZY MATHEMATICS AND ITS APPLICATIONS**

- HLINĚNÁ Dana, BIBA Vladislav: A NOTE ON SOME CLASSES OF  
GENERATED FUZZY IMPLICATIONS 243

### **MODELING AND SIMULATION**

- FILIPÉ José António, FERREIRA Manuel Alberto M., COELHO M.,  
PEDRO Maria Isabel:** ANTI-COMMONS: FISHERIES PROBLEMS AND  
BUREAUCRACY IN AQUACULTURE 253
- JANČO Roland, KOVÁČOVÁ Monika:** SOLUTION OF TORSION OF  
PRISMATIC BAR USING PROGRAM MATHEMATICA FOR ELLIPTICAL  
CROSS-SECTION AREA 261
- ZEITHAMER, Tomáš, R.:** NON-LINEAR MOTION EQUATION OF  
MONOTONE COMMODITY STATE DEVELOPMENT WITH INFLEXION  
UNDER THE CONDITION OF PERFECT COMPETITION 269

# LIST OF REVIEWERS

<b>Andrade Marina</b> , Professor Auxiliar	University Institute of Lisbon, Lisboa, Portugal
<b>Bartošová Jitka</b> , RNDr., PhD	University of Economics, Jindřichův Hradec, Czech Republic
<b>Baštinec Jaromír</b> , doc. RNDr., CSc.	FEEC, Brno University of Technology, Brno, Czech Republic
<b>Beránek Jaroslav</b> , doc. RNDr., CSc.	Masaryk University, Brno, Czech Republic
<b>Biswas Md. Haider Ali</b> , Associate Professor	Engineering and Technology School, Khulna University, Belize
<b>Bittnerová Daniela</b> , RNDr., CSc.	Technical Univerzity of Liberec, Liberec, Czech Republic
<b>Brabec Marek</b> , Ing., PhD	Academy of Sciences of the Czech Republic, Praha, Czech Republic
<b>Buikis Maris</b> , Prof. Dr.	Riga Technical University, Riga, Latvia
<b>Cyhelský Lubomír</b> , Prof. Ing., DrSc.	Vysoká škola finanční a správní, Praha, Czech Republic
<b>Dorociaková Božena</b> , RNDr., PhD	University of Žilina, Žilina, Slovak Republic
<b>Emanovský Petr</b> , Doc. RNDr., PhD	Palacky University, Olomouc, Czech Republic
<b>Ferreira Manuel Alberto M.</b> , Professor Catedrático	University Institute of Lisbon, Lisboa, Portugal
<b>Filipe José António</b> , Professor Auxiliar	IBS - IUL, ISCTE - IUL, Lisboa , Portugal
<b>Habiballa Hashim</b> , RNDr. PaedDr., PhD	University of Ostrava, Ostrava, Czech Republic
<b>Habiballa Hashim</b> , RNDr. PaedDr., PhD	University of Ostrava, Ostrava, Czech Republic
<b>Hošková-Mayerová Šárka</b> , doc. RNDr., PhD	University of Defence, Brno, Czech Republic

<b>Hošpesová Alena</b> , doc. PhDr., PhD	Jihočeská univerzita, České Budějovice, Czech Republic
<b>Iorfida Vincenzo</b>	Lamezia Terme, Italy
<b>Iveta Stankovičová</b> , PhD	UK Bratislava, Bratislava, Slovak Republic
<b>Jancarík Antonín</b> , PhD	Charles University, Prague, Czech Republic
<b>Jukl Marek</b> , RNDr., PhD	Palacky University, Olomouc, Czech Republic
<b>Kráľ Pavol</b> , RNDr., PhD	Matej Bel University, Banská Bystrica, Slovak Republic
<b>Kunderová Pavla</b> , doc. RNDr., CSc.	Palacky University, Olomouc, Czech Republic
<b>Kvasz Ladislav</b> , Prof.	Charles University, Prague, Czech Republic
<b>Langhamrová Jitka</b> , doc. Ing., CSc	University of Economics in Prague, Prague Czech Republic
<b>Linda Bohdan</b> , doc. RNDr., CSc.	University of Pardubice, Pardubice, Czech Republic
<b>Maroš Bohumil</b> , doc. RNDr., CSc.	University of Technology, Brno, Czech Republic
<b>Matvejevs Andrejs</b> , DrSc., Ing.	Riga Technical university, Riga, Latvia
<b>Mikeš Josef</b> , Prof. RNDr., DrSc.	Palacky University, Olomouc, Czech Republic
<b>Milerová Helena</b> , Bc.	Charles University, Prague, Czech Republic
<b>Miroslav Husek</b>	Charles University, Prague, Czech Republic
<b>Miskolczi Martina</b> , Mgr., Ing.	University of Economics in Prague, Prague, Czech Republic
<b>Morkisz Paweł</b> ,	AGH University of Science and Technology, Krakow, Poland
<b>Mošná František</b> , RNDr., PhD	Czech Univ. of Life Sciences, Praha, Czech Republic
<b>Paláček Radomír</b> , RNDr., PhD	VŠB - Technical University of Ostrava, Ostrava, Czech Republic
<b>Pospíšil Jiří</b> , Prof. Ing., CSc.	Czech Technical University of Prague, Prague, Czech Republic

<b>Potůček Radovan</b> , RNDr., PhD	University of Defence, Brno, Czech Republic
<b>Radova Jarmila</b> , doc. RNDr., PhD	University of Economics, Prague, Czech Republic
<b>Rus Ioan A.</b> , Professor	Babes-Bolyai University of Cluj-Napoca, Cluj-Napoca, Romania
<b>Růžicková Miroslava</b> , doc. RNDr., CSc.	University of Žilina, Žilina, Slovak Republic
<b>Segeth Karel</b> , Prof. RNDr., CSc.	Academy of Sciences of the Czech Republic, Prague , Czech Republic
<b>Slaby Antonin</b> , Prof. RNDr., PhDr., CSc.	University of Hradec Kralove, Hradec Kralove, Czech Republic
<b>Sousa Cristina Alexandra</b> , Master	Universidade Portucalense Infante D. Henrique, Porto, Portugal
<b>Svoboda Zdeněk</b> , RNDr., CSc.	FEEC, Brno University of Technology, Brno, Czech Republic
<b>Šamšula Pavel</b> , doc. PaedDr., CSc	Charles University, Prague, Czech Republic
<b>Torre Matteo</b> , Laurea in Matematica	Scula Secondaria Superiore, Alessandria, Italy
<b>Trojovsky Pavel</b> , RNDr., PhD	University of Hradec Kralove, Hradec Kralove, Czech Republic
<b>Trokanová Katarína</b> , Doc.	Slovak Technical University, Bratislava, Slovak Republic
<b>Ulrychová Eva</b> , RNDr.	University of Finance and Administration, Prague, Czech Republic
<b>Vanžurová Alena</b> , doc. RNDr., CSc.	Palacký University, Olomouc, Czech Republic
<b>Velichová Daniela</b> , doc. RNDr., CSc. mim.prof.	Slovak University of Technology, Bratislava, Slovak Republic
<b>Vítovec Jiří</b> , Mgr., PhD	Brno University of Technology, Brno, Czech Republic
<b>Voicu Nicoleta</b> , Dr.	Transilvania University of Brasov, Romania, Brasov, Romania
<b>Volna Eva</b> , doc. RNDr. PaedDr. PhD	University of Ostrava, Ostrava, Czech Republic

**Wimmer Gejza**, Professor

Slovak Academy of Sciences, Bratislava, Slovak  
Republic

**Zeithamer Tomáš R.**, Ing., PhD

University of Economics, Prague, Czech Republic

## CONTROLLABILITY FOR A CERTAIN CLASS OF LINEAR MATRIX SYSTEMS WITH DELAY

BAŠTINEC Jaromír, (CZ), PIDUBNA Ganna, (CZ)

**Abstract.** In this paper existence of solutions of a certain class of differential linear matrix equations with delay was investigated. The solutions were found in general form. Necessary and sufficient condition for controllability of differential linear matrix equation with delay was defined and control was built. Paper contains calculated examples.

**Key words and phrases.** matrix equation with delay, matrix exponential.

*Mathematics Subject Classification.* Primary 34K20, 34K25; Secondary 34K12.

### 1 Introduction

This paper is devoted to computing of the solution of differential linear matrix equation with delay  $\dot{X}(t) = AX(t) + AX(t - \tau)$ , with help of the special matrix function - matrix exponential. Matrix exponential was used for solving differential equations by Krasovskiy [10], [11] and for solving systems with aftereffects by many authors, e.g. Boichuk, Diblík, Khusainov, Růžicková, Shuklin [3] - [9].

**Definition 1.1** Let  $A$  be a square matrix. Matrix exponential is defined by

$$e^{At} = I + A\frac{t}{1!} + A^2\frac{t^2}{2!} + A^3\frac{t^3}{3!} + \cdots = \sum_{i=0}^{\infty} A^i\frac{t^i}{i!},$$

where  $I$  is the identity matrix.

**Lemma 1.2** Let  $A$  be a square matrix, then holds  $Ae^{At} = e^{At}A$ ,  $e^{At}e^{A\tau} = e^{A(t+\tau)}$ .

## 2 Linear matrix equation with delay

Let we have the equation

$$\dot{X}(t) = AX(t) + AX(t - \tau), \quad (1)$$

with initial condition

$$X(t) = I, \quad -\tau \leq t \leq 0, \quad (2)$$

where  $A$  is square matrix,  $I$  is identity matrix,  $\tau > 0, \tau \in R$  is a constant delay.

**Theorem 2.1** *Let  $A$  is regular. Then the solution of equation (1) with identity initial condition has the recurrence form:*

$$X_{n+1}(t) = e^{A(t-n\tau)} X_n(n\tau) + \int_{n\tau}^t e^{A(t-s)} AX_n(s - \tau) ds, \quad (3)$$

where  $X_n(t)$  is defined on the interval  $(n-1)\tau \leq t \leq n\tau$ .

**Proof:** Theorem 3.1 is a special case of the more general Theorem 3.1, [2].

**Theorem 2.2** *Let  $A$  is regular. Then the solution of equation (1) with identity initial condition has the form:*

$$X_k(t) = \sum_{l=0}^{k-1} 2e^{A(t-l\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(t-l\tau)^p}{p!} + (-1)^k I, \quad (4)$$

where  $X_k(t)$  is defined on the interval  $(k-1)\tau \leq t \leq k\tau$ .

**Proof:** To prove Theorem 2.2, we find the form of the solution for  $k\tau \leq t \leq (k+1)\tau$ . Let  $k\tau \leq t \leq (k+1)\tau$  holds. Then the equation (1) has the form

$$\dot{X}_{k+1}(t) = AX_{k+1}(t) + AX_{k+1}(t - \tau) = AX_{k+1}(t) + AX_k(t - \tau).$$

Then from (3) follows, that for the solution of equation (1) on this interval for  $n = k$

$$X_{k+1}(t) = e^{A(t-k\tau)} X_k(k\tau) + \int_{k\tau}^t e^{A(t-s)} AX_k(s - \tau) ds.$$

After substitution  $X_k(t)$  from (4) we have

$$\begin{aligned} X_{k+1}(t) &= e^{A(t-k\tau)} \left[ \sum_{l=0}^{k-1} 2e^{A(k\tau-l\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(k\tau-l\tau)^p}{p!} + (-1)^k I \right] \\ &+ \int_{k\tau}^t e^{A(t-s)} A \left[ \sum_{l=0}^{k-1} 2e^{A(s-\tau-l\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(s-\tau-l\tau)^p}{p!} + (-1)^k I \right] ds \\ &= \left[ \sum_{l=0}^{k-1} e^{A(t-k\tau)} 2e^{A(k\tau-l\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(k\tau-l\tau)^p}{p!} + (-1)^k e^{A(t-k\tau)} \right] \end{aligned}$$

$$\begin{aligned}
& + \left[ \sum_{l=0}^{k-1} \int_{k\tau}^t e^{A(t-s)} A 2e^{A(s-\tau-l\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(s-\tau-l\tau)^p}{p!} ds + (-1)^k \int_{k\tau}^t e^{A(t-s)} A ds \right] \\
& = \left[ \sum_{l=0}^{k-1} 2e^{A(t-l\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(k\tau-l\tau)^p}{p!} + (-1)^k e^{A(t-k\tau)} \right] \\
& + \left[ \sum_{l=0}^{k-1} \int_{k\tau}^t 2e^{A(t-(l+1)\tau)} \sum_{p=0}^l (-1)^{p+l} A^{p+1} \frac{(s-(l+1)\tau)^p}{p!} ds + (-1)^k \int_{k\tau}^t e^{A(t-s)} A ds \right] \\
& = \sum_{l=0}^{k-1} 2e^{A(t-l\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(k\tau-l\tau)^p}{p!} + (-1)^k 2e^{A(t-k\tau)} \\
& \quad + \sum_{l=0}^{k-1} 2e^{A(t-(l+1)\tau)} \sum_{p=0}^l (-1)^{p+l} A^{p+1} \frac{(t-(l+1)\tau)^{p+1}}{(p+1)!} \\
& \quad - \sum_{l=0}^{k-1} 2e^{A(t-(l+1)\tau)} \sum_{p=0}^l (-1)^{p+l} A^{p+1} \frac{(k\tau-(l+1)\tau)^{p+1}}{(p+1)!} + (-1)^{k+1} I \\
& = \sum_{l=0}^{k-1} 2e^{A(t-l\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(k\tau-l\tau)^p}{p!} + (-1)^k 2e^{A(t-k\tau)} + (-1)^{k+1} I \\
& \quad + \sum_{l=1}^k 2e^{A(t-l\tau)} \sum_{p=1}^l (-1)^{p+l} A^p \frac{(t-l\tau)^p}{p!} - \sum_{l=1}^{k-1} 2e^{A(t-l\tau)} \sum_{p=1}^l (-1)^{p+l} A^p \frac{(k\tau-l\tau)^p}{p!} \\
& = 2e^{At} + \sum_{l=1}^{k-1} 2e^{A(t-l\tau)} (-1)^l + \sum_{l=1}^k 2e^{A(t-l\tau)} \sum_{p=1}^l (-1)^{p+l} A^p \frac{(t-l\tau)^p}{p!} + (-1)^k 2e^{A(t-k\tau)} + (-1)^{k+1} I \\
& = \left( 2e^{At} + \sum_{l=1}^{k-1} 2e^{A(t-l\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(t-l\tau)^p}{p!} \right) + \left( 2e^{A(t-k\tau)} \sum_{p=0}^l (-1)^{p+k} A^p \frac{(t-k\tau)^p}{p!} \right) + (-1)^{k+1} I \\
& = \sum_{l=0}^{k-1} 2e^{A(t-l\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(t-l\tau)^p}{p!} + 2e^{A(t-k\tau)} \sum_{p=0}^l (-1)^{p+k} A^p \frac{(t-k\tau)^p}{p!} + (-1)^{k+1} I.
\end{aligned}$$

Finally we get  $X_{k+1}(t) = \sum_{l=0}^k 2e^{A(t-l\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(t-l\tau)^p}{p!} + (-1)^{k+1} I$ .

And we got the expression (4) for  $k\tau \leq t \leq (k+1)\tau$ .

If we have initial condition in the form

$$X(t) = \varphi(t), \quad -\tau \leq t \leq 0, \quad (5)$$

where  $\varphi(t) \in C^1[-\tau, 0]$ , then we could write the following result.

**Theorem 2.3** [1] *Let  $A$  is regular. Then the solution of equation (1) with initial condition (5) have the form:  $X(t) = X_n(t)\varphi(-\tau) + \int_{-\tau}^0 X_n(t - \tau - s)\varphi'(s)ds$ , where  $X_n(t)$  is the solution of the same equation with identity initial condition, defined in Theorem 2.2.*

Let we have the linear heterogeneous equation with delay

$$\dot{X}(t) = AX(t) + AX(t - \tau) + F(t). \quad (6)$$

**Theorem 2.4** [1] *Let  $A$  is regular. Then the solution  $\overline{X(t)}$  of the heterogeneous equation (6) with zero initial condition, has the form  $\overline{X(t)} = \int_0^t X_n(t - \tau - s)F(s)ds$ ,  $t \geq 0$ , where  $X_n(t)$  is defined in Theorem 2.2.*

**Theorem 2.5** [1] *Let  $A$  is regular. The solution of heterogeneous equation (6) with the initial condition (5) has the form  $X(t) = X_n(t)\varphi(-\tau) + \int_{-\tau}^0 X_n(t - \tau - s)\varphi'(s)ds + \int_0^t X_n(t - \tau - s)F(s)ds$ , where  $X_n(t)$  is defined in Theorem 2.2.*

### 3 Controllability of the linear matrix system with delay

#### 3.1 General terms

Let  $X$  is the space of states of dynamic system;  $U$  is the set of the controlled effects (controls). Let  $x = x(x_0, u, t)$  is the vector that characterizes state of the dynamic system in moment of time  $t$ , by the initial condition  $x_0$ ,  $x_0 \in X$ , ( $x_0 = x|_{t=t_0}$ ) and by the control function  $u$ ,  $u \in U$ .

**Definition 3.1** *The state  $x_0$  is called controllable state in the class  $U$  (controlled state), if there are exist such control  $u(x_0) \in U$  and the number  $T$ ,  $t_0 \leq T$  that  $x(x_0, u(x_0), T) = 0$ .*

**Definition 3.2** *If every state  $x_0 \in X$  of the dynamic system is controllable, then we say that the system is controllable (controlled system).*

Consider the following Cauchy's problem:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Ax(t - \tau) + Bu(t), \quad t \in [0, T], \quad T < \infty, \\ x(0) &= x_0, \quad x(t) = \varphi(t), \quad -\tau \leq t < 0, \end{aligned} \quad (7)$$

where  $x = (x_1, \dots, x_n)^T$  is the vector of phase coordinates,  $x \in X$ ,  $u(t) = (u_1(t), \dots, u_r(t))^T$  is the control function,  $u \in U$ ,  $U$  is the set of piecewise-continuous functions;  $A, B$  are constant matrices of dimensions  $(n \times n)$ ,  $(n \times r)$  respectively,  $\tau$  is the constant delay. Space of states  $Z$  of this system is the set of  $n$ -dimensional functions.

$$\{x(\theta), \quad t - \tau \leq \theta \leq t\} \quad (8)$$

The space of the  $n$ -dimensional vectors  $x$  (phase space  $X$ ) is subspace for  $Z$ . The initial state  $z_0$  of the system (7) is determined by conditions

$$z_0 = \{x_0(\theta), x_0(\theta) = \varphi(\theta), -\tau \leq \theta < 0, x(0) = x_0\}. \quad (9)$$

The state  $z = z(z_0, u, t)$  of the system (7) in the space  $Z$  in moment of time  $t$  is defined by trajectory segment (8) of phase space  $X$ .

Next considered, that the movement system (7) goes ( $t \geq 0$ ) in the space of continuous function. We determined initial state (9) of the function  $\varphi(\theta)$  as piecewise-continuous.

In accordance with specified definitions, state (9) of the system (7) is controllable if there exist such control  $u \in U$  that  $x(t) \equiv 0$ ,  $T - \tau \leq t \leq T$  when  $T < \infty$ .

### 3.2 The construction of control for system with delay

Let we have the control system of differential matrix equation

$$\dot{x}(t) = Ax(t) + Ax(t - \tau) + Bu(t), \quad x(t) \in R^n, \quad t \geq 0, \quad \tau > 0. \quad (10)$$

where  $x(t) = \varphi(t)$ ,  $-\tau \leq t \leq 0$ ,  $A, B$  are square constant matrices.

**Remark 3.3** For convenience purpose, here and further, we say that  $x(t)$  is a vector of length  $n$ . All next statements are proved in the same way for the case when  $x(t) = X(t)$  is a matrix of dimension  $(n \times n)$ .

**Theorem 3.4** For controllability of linear system with delay (10) is necessary and sufficient to next condition to hold:  $t \geq (k-1)\tau$  and  $\text{rank}(S) = n$ , where  $S = \{B \ AB \ A^2B \ \dots \ A^{k-1}B \ \dots\}$ , hence  $S$  is a matrix which was achieved by recording matrices  $B, AB, \dots, A^{k-1}B, \dots$  side by side.

**Proof:** Let system (10) is controllable. Then for any  $\varphi(t)$ ,  $x_1$  and  $t_1$  there exist a control  $u^*(t)$  such that for a system (10) exist solution  $x^*(t)$  which satisfies initial condition  $x(t) = \varphi(t)$ ,  $-\tau \leq t \leq 0$ . The representation of the solution of the Cauchy problem for heterogeneous equation is:

$$x(t) = X_0(t)\varphi(-\tau) + \int_{-\tau}^0 X_0(t - \tau - s)\varphi'(s)ds + \int_0^t X_0(t - \tau - s)Bu(s)ds.$$

When the control is  $u^*(t)$ , then in time moment  $t = t_1$  we get

$$x_1 = X_0(t_1)\varphi(-\tau) + \int_{-\tau}^0 X_0(t_1 - \tau - s)\varphi'(s)ds + \int_0^{t_1} X_0(t_1 - \tau - s)Bu^*(s)ds. \quad (11)$$

Denoted

$$x_1 - X_0(t_1)\varphi(-\tau) - \int_{-\tau}^0 X_0(t_1 - \tau - s)\varphi'(s)ds = \mu. \quad (12)$$

And using the representation of  $X_0(t)$  from (4) we get

$$\begin{aligned}
 & \int_0^{t_1} X_0(t_1 - \tau - s) B u^*(s) ds \\
 &= \int_0^{t_1} \left[ \sum_{l=0}^{k-1} 2e^{A(t_1 - \tau - s - l\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(t_1 - \tau - s - l\tau)^p}{p!} + (-1)^k I \right] B u^*(s) ds \\
 &= \int_0^{t_1} \sum_{l=0}^{k-1} 2e^{A(t_1 - s - (l+1)\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(t_1 - s - (l+1)\tau)^p}{p!} B u^*(s) ds + (-1)^k B \int_0^{t_1} u^*(s) ds \\
 &= \sum_{l=0}^{k-1} \sum_{p=0}^l (-1)^{p+l} 2A^p \int_0^{t_1} e^{A(t_1 - s - (l+1)\tau)} \frac{(t_1 - s - (l+1)\tau)^p}{p!} B u^*(s) ds + (-1)^k B \int_0^{t_1} u^*(s) ds \\
 &= \sum_{l=0}^{k-1} \sum_{p=0}^l (-1)^{p+l} 2A^p \int_0^{t_1} \sum_{m=0}^{\infty} A^m \frac{(t_1 - s - (l+1)\tau)^{m+p}}{m!p!} B u^*(s) ds + (-1)^k B \int_0^{t_1} u^*(s) ds \\
 &= \sum_{l=0}^{k-1} \sum_{p=0}^l \sum_{m=0}^{\infty} (-1)^{p+l} 2A^{p+m} B \int_0^{t_1} \frac{(t_1 - s - (l+1)\tau)^{m+p}}{m!p!} u^*(s) ds + (-1)^k B \int_0^{t_1} u^*(s) ds = (h).
 \end{aligned}$$

Denoted

$$\psi_{l,p,m}(t_1) = (-1)^{p+l} 2 \int_0^{t_1} \frac{(t_1 - s - (l+1)\tau)^{m+p}}{m!p!} u^*(s) ds,$$

then

$$\begin{aligned}
 (h) &= \sum_{l=0}^{k-1} \sum_{p=0}^l \sum_{m=0}^{\infty} A^{p+m} B \psi_{l,p,m}(t_1) + (-1)^k B \int_0^{t_1} u^*(s) ds \\
 &= B \left[ \sum_{l=0}^{k-1} \psi_{l,0,0}(t_1) + (-1)^k \int_0^{t_1} u^*(s) ds \right] \\
 &+ AB \left[ \sum_{l=0}^{k-1} \psi_{l,0,1}(t_1) + \sum_{l=1}^{k-1} \psi_{l,1,0}(t_1) \right] + A^2 B \left[ \sum_{l=0}^{k-1} \psi_{l,0,2}(t_1) + \sum_{l=1}^{k-1} \psi_{l,1,1}(t_1) + \sum_{l=2}^{k-1} \psi_{l,2,0}(t_1) \right] + \dots \\
 &+ A^{k-1} B \left[ \sum_{l=0}^{k-1} \psi_{l,0,k-1}(t_1) + \sum_{l=1}^{k-1} \psi_{l,1,k-2}(t_1) + \dots + \sum_{l=k-2}^{k-1} \psi_{l,k-2,1}(t_1) + \psi_{l,k-1,0} \right] + \dots \\
 &= B f_1(t_1) + AB f_2(t_1) + A^2 B f_3(t_1) + \dots + A^{k-1} B f_k(t_1) + \dots
 \end{aligned}$$

And using (12), correlation (11) get the form

$$B f_1(t_1) + AB f_2(t_1) + A^2 B f_3(t_1) + \dots + A^{k-1} B f_k(t_1) + \dots = \mu.$$

So we got a system with an infinite number of unknown functions  $f_i$  and the vector of constant terms  $\mu$  is length  $n$ . The system will have only one solution if and only if the rank of the matrix  $S = \{B \ AB \ A^2B \ \dots \ A^{k-1}B \ \dots\}$  equals  $n$ . In this case the solution of the system will be the vector  $f$ , that is uniquely determined by the vector of constant terms  $x_1$ . Since the vector of constant terms is defined from any finite state of the system (10), we conclude that system (10) can be moved in any point if the conditions of the theorem is true. It means, that the system (10) is controllable if and only if the rank of matrix  $S$  is  $n$ .

**Theorem 3.5** *Let  $t_1 \geq (k-1)\tau$  and the necessary and sufficient condition for controllability is implemented:  $\text{rank}(S) = \text{rank}(\{B \ AB \ A^2B \ \dots \ A^{k-1}B \ \dots\}) = n$ . Then the control function can be taken as*

$$u(s) = [X_0(t_1 - \tau - s)B]^T \left[ \int_0^{t_1} X_0(t_1 - \tau - s)BB^T[X_0(t_1 - \tau - s)]^T ds \right]^{-1} \mu,$$

where  $\mu = x_1 - X_0(t_1)\varphi(-\tau) - \int_{-\tau}^0 X_0(t_1 - \tau - s)\varphi'(s)ds$ .

**Proof:** Using the result of the Theorem 2.5, we have that the solution of the system (10) with initial conditions  $x_0(t) = \varphi(t)$ ,  $-\tau \leq t \leq 0$  has the form

$$x(t) = X_0(t)\varphi(-\tau) + \int_{-\tau}^0 X_0(t - \tau - s)\varphi'(s)ds + \int_0^t X_0(t - \tau - s)Bu(s)ds \quad (13)$$

Using the notations (12), we obtain: for the system (13) to have a solution  $x(t)$  that satisfies the initial conditions  $x(t) = \varphi(t)$ ,  $-\tau \leq t \leq 0$ ,  $x(t_1) = x_1$ , is necessary and sufficient that the integrated equation

$$\int_0^{t_1} \left( \sum_{l=0}^{k-1} 2e^{A(t_1-s-(l+1)\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(t_1-s-(l+1)\tau)^p}{p!} + (-1)^k I \right) Bu(s)ds = \mu \quad (14)$$

has solution  $u(s)$ ,  $0 \leq s \leq t_1$ . We will search a solution as a linear combination

$$u(s) = \left[ \left( \sum_{l=0}^{k-1} 2e^{A(t_1-s-(l+1)\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(t_1-s-(l+1)\tau)^p}{p!} + (-1)^k I \right) B \right]^T C \quad (15)$$

where  $C = (c_1, c_2, \dots, c_n)^T$ - is unknown vector. After substitution (15) in system (14), we get

$$\left[ \int_0^{t_1} \left( \sum_{l=0}^{k-1} 2e^{A(t_1-s-(l+1)\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(t_1-s-(l+1)\tau)^p}{p!} + (-1)^k I \right) B \right. \\ \left. \times B^T \left( \sum_{l=0}^{k-1} 2e^{A(t_1-s-(l+1)\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(t_1-s-(l+1)\tau)^p}{p!} + (-1)^k I \right)^T ds \right] C = \mu. \quad (16)$$

We will show that system (16) has the only one solution. From prove of previous theorem we know, that  $X_0(t - \tau - s)B$  can be represented as a linear combination with coefficients  $B; AB; \dots; A^k B; \dots$ . Since  $\text{rank}(S) = n$ , then, when  $0 \leq s \leq t_1$ , will be done  $X_0(t - \tau - s)B \neq 0$ . Therefore for any vector  $l = (l_1, l_2, \dots, l_n)^T$  in  $0 \leq s \leq t_1$  will be done  $([X_0(t - \tau - s)B]^T l)^2 \neq 0$ ,  $0 \leq s \leq t_1$ . And for any  $l > 0$

$$\int_0^{t_1} \left( B^T \left[ \sum_{l=0}^{k-1} 2e^{A(t_1-s-(l+1)\tau)} \sum_{p=0}^l (-1)^{p+l} A^p \frac{(t_1-s-(l+1)\tau)^p}{p!} + (-1)^k I \right]^T l \right)^2 ds$$

$$= \left[ \int_0^{t_1} X_0(t_1 - \tau - s) B B^T [X_0(t_1 - \tau - s)]^T ds \right] l^2.$$

Because the matrix  $\int_0^{t_1} X_0(t_1 - \tau - s) B B^T [X_0(t_1 - \tau - s)]^T ds$  is positive definite. Therefore its determinant is nonzero. Solving system (16), we obtain

$$C = \left[ \int_0^{t_1} X_0(t_1 - \tau - s) B B^T [X_0(t_1 - \tau - s)]^T ds \right]^{-1} \mu.$$

## 4 Examples

Let us consider few examples of controllability researches of the linear matrix systems with delay.

### Example 4.1

Let us have the differential equation of 3-th degree with a constant delay:

$$\dot{x}(t) = Ax(t) + Ax(t-1) + Bu(t), \text{ where } A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

As we see  $\tau = 1, n = 3$  and  $A$  is regular. We want to know if this system is controllable so let us check the necessary and sufficient condition. We will find the matrix  $S$ :

$$S = \{B \ AB \ A^2 B \ \dots \ A^{k+1} B \ \dots\} = \begin{pmatrix} 1 & 1 & 0 & 2 & 2 & 0 & 3 & 3 & 0 & k & k & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & \dots & 1 & 1 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

We have,  $\text{rank}(S) = 2$ , so the system is not controllable.

### Example 4.2

Let us have the differential equation of 3-th degree with a constant delay:

$$\dot{x}(t) = Ax(t) + Ax(t-1) + Bu(t), \text{ where } A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

As we see  $\tau = 1, n = 3$  and  $A$  is regular. It is easy to see that the necessary and sufficient condition for controllability is implemented (because of full rank of the matrix  $B$ , matrix  $S$  have full rank too), so the system is controllable.

Let us construct such control function, that move system in time moment  $t_1 = 2$  in point  $x_1 = (1, 1, 1)^T$ , using initial condition  $x_0(t) = \varphi(t) = (0, 0, 0)^T, -1 \leq t \leq 0$ . Using the result of the theorem (3.5) we write:

$$u(t) = [X_0(t_1 - \tau - t)B]^T \left[ \int_0^{t_1} X_0(t_1 - \tau - s)BB^T[X_0(t_1 - \tau - s)]^T ds \right]^{-1} \mu,$$

$$\mu = x_1 - X_0(t_1)\varphi(-\tau) - \int_{-\tau}^0 X_0(t_1 - \tau - s)\varphi'(s)ds.$$

While  $\varphi(t) = (0, 0, 0)^T, -1 \leq t \leq 0$  then  $\mu = (1, 1, 1)^T$ . So, we have

$$u(t) = [X_0(1-t)B]^T \left[ \int_0^2 X_0(1-s)BB^T[X_0(1-s)]^T ds \right]^{-1} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

While  $t_1 = 2$ , then  $k = 2$  and, using (4) we can calculate

$$u(t) = \begin{pmatrix} 2(t+1)e^t + 2(t^2 - t - 1)e^{t-1} + 1 & 2e^t + 2(t-2)e^{t-1} & 0 \\ 2(t+1)e^t + 2(t^2 - t - 1)e^{t-1} + 1 & 2e^t + 2(t-2)e^{t-1} & 0 \\ (t^2 + 2t)e^t + (t^3 - 3t + 2)e^{t-1} & 2te^t + 2(t-1)^2e^{t-1} & 2e^t + 2(t-2)e^{t-1} \end{pmatrix} \begin{pmatrix} 0.05 & -0.13 & 0.09 \\ -0.13 & 0.38 & -0.25 \\ 0.09 & -0.25 & 0.18 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

$$u(t) = 0.01 \begin{pmatrix} 2(t+1)e^t + 2(t^2 - t - 1)e^{t-1} + 1 \\ 2(t+1)e^t + 2(t^2 - t - 1)e^{t-1} + 1 \\ (t^2 + 2t + 4)e^t + (t^3 + t - 6)e^{t-1} \end{pmatrix}.$$

## 5 Conclusion

In this paper a solution of the system in general form was built. The necessary and sufficient condition for controllability of this system was defined and control was built. Two examples were given to illustrate the proposed theory. Getting results analogous to the ones in sections 3 and 4 for equation  $\dot{X}(t) = AX(t) + BX(t - \tau)$ , where  $A, B$  are different matrices, remains an open problem.

## Acknowledgement

This research was supported by the Grant 201/10/1032 of Czech Grant Agency and by Grant FEKT-S-11-2-921 of Faculty of Electrical Engineering and Communication, BUT.

## References

- [1] BAŠTINEC, J., PIDUBNA, G.: *Controllability of stationary linear systems with delay*. 10th International conference APLIMAT. Bratislava, FME STU. 2011, 207 - 216. ISBN 978-80-89313-51-8.
- [2] BAŠTINEC, J., PIDUBNA, G.: *Solution of Matrix Linear Delayed System*. 7th conference of mathematics and physics at the technical universities with international participation. 2011. p. 48 - 57. ISBN 978-80-7231-815-5.
- [3] BOICHUK, A., DIBLÍK, J., KHUSAINOV, D., RŮŽIČKOVÁ, M.: *Boundary Value Problems for Delay Differential Systems*, Advances in Difference Equations, vol. 2010, Article ID 593834, 20 pages, 2010. doi:10.1155/2010/593834
- [4] BOICHUK, A., DIBLÍK, J., KHUSAINOV, D., RŮŽIČKOVÁ, M.: *Fredholm's boundary-value problems for differential systems with a single delay*, Nonlinear Analysis, **72** (2010), 2251–2258. (ISSN 0362-546X)
- [5] BOICHUK, A., DIBLÍK, J., KHUSAINOV, D., RŮŽIČKOVÁ, M.: *Controllability of linear discrete systems with constant coefficients and pure delay*, SIAM Journal on Control and Optimization, **47**, No 3 (2008), 1140–1149. DOI: 10.1137/070689085, url = <http://link.aip.org/link/?SJC/47/1140/1>. (ISSN Electronic: 1095-7138, Print: 0363-0129)
- [6] BOICHUK, A., DIBLÍK, J., KHUSAINOV, D., RŮŽIČKOVÁ, M.: *Controllability of linear discrete systems with constant coefficients and pure delay*, SIAM Journal on Control and Optimization, **47**, No 3 (2008), 1140–1149. DOI: 10.1137/070689085, url = <http://link.aip.org/link/?SJC/47/1140/1>. (ISSN Electronic: 1095-7138, Print: 0363-0129)
- [7] DIBLÍK, J., KHUSAINOV, D., LUKÁČOVÁ, J., RŮŽIČKOVÁ, M.: *Control of oscillating systems with a single delay*, Advances in Difference Equations, Volume 2010 (2010), Article ID 108218, 15 pages, doi:10.1155/2010/108218.
- [8] KHUSAINOV D.Ya., IVANOV A.F., SHUKLIN G.V.: *About a single submission of solution of linear systems with delay*. Differential equations. 2005. No.41,7. 1001-1004. (In Russian)
- [9] KHUSAINOV D.Ya., SHUKLIN G.V.: *About relative controllability of systems with pure delay*. Applied Mechanics. 2005. No.41,2. 118-130. (In Russian)
- [10] KRASOVSKII N.N.: *Inversion of theorems of second Lyapunov's method and stability problems in the first approximation*. Applied Mathematics and Mechanics. 1956. 255-265. (In Russian)
- [11] KRASOVSKII N.N.: *The theory of motion control. Linear systems*. Nauka. 1968. 475. (In Russian)

## Current address

**BAŠTINEC Jaromír, doc., RNDr., CSc.**

Department of Mathematics,  
Faculty of Electrical Engineering and Communication,  
Brno University of Technology,  
Technická 8, 616 00 Brno, Czech Republic

telefon: +420 541 143 222  
email: bastinec@feec.vutbr.cz

**PIDDUBNA Ganna, Mgr.**

Department of Mathematics,  
Faculty of Electrical Engineering and Communication,  
Brno University of Technology,  
Technická 8, 616 00 Brno, Czech Republic  
telefon: +420 541 143 218  
email: xpiddub00@stud.feec.vutbr.cz



# EQUILIBRIUM STOCHASTIC STABILITY OF MARKOV DYNAMICAL SYSTEMS

CARKOVŠ Jevgeņijs, (LV), ŠADURSKIS Kārlis, (LV)

**Abstract.** In first section of paper we will prove that for linear Markov dynamical systems an equilibrium asymptotic stochastic stability is equivalent to exponential p-stability for sufficiently small positive values p. Then we will prove that exponential p-stability of linearized in vicinity of equilibrium Markov dynamical system guarantees equilibrium asymptotic (local) stochastic stability. This result permits to construct such Lyapunov quadratic functional, which one may use for local equilibrium stochastic stability of sufficiently smooth nonlinear Markov dynamical systems.

**Key words and phrases.** Markov dynamical systems; stochastic stability; Lyapunov stability.

*Mathematics Subject Classification.* Primary 37H99; Secondary 34D20.

## 1 Stochastic stability of linear differential equations with Markov coefficients

Let  $y(t)$  be Feller type Markov process on phase space  $\mathbb{Y}$  and with weak infinitesimal operator [Doob]  $\mathcal{L}$ ,  $f(x, y)$  be a continuous mapping  $\mathbb{R}^n \times \mathbb{Y} \rightarrow \mathbb{R}^n$ , and  $f(0, y) \equiv 0$ . The solution of equation

$$\frac{dx(t)}{dt} = f(x(t), y(t)) \quad (1)$$

with initial condition  $x(s) = x, y(s) = y$  we will denote  $x(t, s, x, y)$ . We will say [3] that trivial solution of differential equation (1)

- *locally stable almost sure, if for any  $s \in \mathbb{R}$ ,  $\eta > 0$  and  $\beta > 0$  there exists such  $\delta > 0$  that the inequality*

$$\sup_{\substack{y \in \mathbb{R}^m \\ \xi \in \mathbb{G}}} \mathbb{P}(\sup_{t \geq s} |x(t, s, x, y)| > \eta) < \beta, \quad (2)$$

*follows the condition  $x \in B_\delta(0)$ , where  $B_\delta(0) := \{x \in \mathbb{R}^n : |x| < \delta\}$ ;*

- *locally asymptotically stochastically stable, if it is locally almost sure stable and there exists such  $\gamma > 0$  that the trajectories which do not leave the ball  $B_\gamma$  tend to 0 as  $t \rightarrow \infty$ ;*
- *asymptotically stochastically stable, if it is locally almost sure stable and for any  $x \in \mathbb{R}^n$ ,  $s \in \mathbb{R}$ , and  $c > 0$  the equality*

$$\lim_{T \rightarrow \infty} \sup_{\substack{y \in \mathbb{R}^m \\ \xi \in \mathbb{G}}} \mathbb{P}(\sup_{t > T} |x(t, s, x, y)| > c) = 0 \quad (3)$$

*is fulfilled;*

- *exponentially  $p$ -stable, if there exist such positive numbers  $M$  and  $\gamma$  that for any  $x \in \mathbb{R}^n$ ,  $y \in \mathbb{R}^m$ ,  $\xi \in \mathbb{G}$ ,  $s \in \mathbb{R}$  and  $t > s$  the inequality*

$$\mathbb{E} |x(t, s, x, y)|^p \leq M |x|^p e^{-\gamma(t-s)} \quad (4)$$

*is fulfilled.*

In this section we will deal with linear differential equations in  $\mathbb{R}^n$

$$\frac{dx}{dt} = A(y(t))x, \quad (5)$$

where  $A(y)$  is continuous bounded matrix-valued function and  $y(t)$  is stochastically continuous Feller Markov process with weak infinitesimal operator  $Q$ . The pair  $\{x(t), y(t)\}$  forms [Skorokhod] homogeneous stochastically continuous Markov process with the weak infinitesimal operator  $L_0$  defined by equality

$$L_0 v(x, y) = (A(y)x, \nabla_x) v(x, y) + Q v(x, y). \quad (6)$$

It is clearly that there exists family of the matrix-valued functions  $\{X(t, s, y), t \geq s \geq 0\}$ , defined by equality  $X(t, s, y)x = x(t, s, x, y)$ , where  $x(t, s, x, y)$  is the solution of Cauchy problem  $x(s, s, x, y) = x$  under condition  $y(s) = y$ . The matrices  $X(t, s, y)$  also satisfy the equation (5) for all  $t > s$  and initial condition  $X(s, s, y) = I$ , where  $I$  is matrix unit. This matrix family has the evolution property:

$$X(t, s, y) = X(t, \tau, y(\tau))X(\tau, s, y) \quad (7)$$

for any  $y \in \mathbb{Y}, t \geq \tau \geq s \geq 0$ . Let us define the Lyapunov  $p$ -index of (5) as

$$\lambda^{(p)} = \sup_{x, y} \overline{\lim}_{t \rightarrow \infty} \frac{1}{pt} \ln \mathbb{E} |X(t, s, y)x|^p. \quad (8)$$

Not so difficult to prove that exponential  $p$ -stability of trivial solution of the equation (5) is equivalent to inequality  $\lambda^{(p)} < 0$ . Because

$$(\mathbb{E} |X(t, s, y)x|^{p_1})^{1/p_1} \leq (\mathbb{E} |X(t, s, y)x|^{p_2})^{1/p_2} \quad (9)$$

for any positive  $p_1 < p_2$ , the inequality

$$\lambda^{(p_1)} \leq \lambda^{(p_2)} \quad (10)$$

follows the inequality  $p_1 < p_2$  and  $\lambda^{(p)}$  is monotone decreasing function as  $p$  decreases to 0. It is intuitively clearly, that asymptotic stochastic stability of (5) is equivalent to the condition

$$\exists p_0 > 0, \forall p \in (0, p_0) : \lambda^{(p)} < 0.$$

We will essentially use further this assertion and hence it should be proven.

**Lemma 1.** *If the equation (5) is asymptotically stochastically stable then it is exponentially  $p$ -stable for all sufficiently small positive  $p$ .*

**Proof.** Let us put in definition of almost sure stability  $\eta = 1, \beta = \frac{1}{2}$  and choose so small positive  $\alpha$  that the inequality

$$\sup_{\substack{|x| \leq 2^{-\alpha} \\ y \in \mathbb{Y}}} \mathbb{P}(\sup_{t \geq 0} |X(t, 0, y)x| > 1) < \frac{1}{2}.$$

is fulfilled. Due to a linearity of the equation (5) from the above inequality one may write the inequality

$$\sup_{\substack{|x| \leq 2^{-\alpha(l-1)} \\ y \in \mathbb{Y}}} \mathbb{P}(\sup_{t \geq 0} |X(t, 0, y)x| > 2^{l\alpha}) < \frac{1}{2}$$

for any  $l \in \mathbb{N}$ . Let us denote

$$g_l := \sup_{\substack{|x| \leq 1 \\ y \in \mathbb{Y}}} \mathbb{P}(\sup_{t \geq 0} |X(t, 0, y)x| \geq 2^{l\alpha}).$$

The pair  $\{x(t), y(t)\}$  is stochastically continuous Markov process and it has the Markov property in the moment  $\tau_1(x)$  of exit of the trajectory  $x(t, 0, x, y)$  from the ball  $B_1(0)$  if  $x \in B_1(0)$ . Hence

$$\begin{aligned} g_{l+1} &= \sup_{\substack{|x| \leq 1 \\ y \in \mathbb{Y}}} \mathbb{P}(\sup_{t \geq 0} |X(t, 0, y)x| \geq 2^{(l+1)\alpha}) \\ &= \sup_{\substack{|x| \leq 1 \\ y \in \mathbb{Y}}} \int_{s=0}^{\infty} \int_{\substack{|u|=2^{l\alpha} \\ v \in \mathbb{Y}}} \mathbb{P}_{x,y}(\tau_1(x) \in ds, x(s) \in du, y(s) \in dv) \times \\ &\quad \times \mathbb{P}(\sup_{t \geq 0} |X(t, 0, v)u| > 2^{(l+1)\alpha}) \\ &\leq \sup_{\substack{|x| \leq 2^{l\alpha} \\ y \in \mathbb{Y}}} \mathbb{P}(\sup_{t \geq 0} |X(t, 0, y)x| > 2^{(l+1)\alpha}) \sup_{\substack{|x| \leq 1 \\ y \in \mathbb{Y}}} \times \\ &\quad \times \int_{s=0}^{\infty} \int_{\substack{|u|=2^{l\alpha} \\ v \in \mathbb{Y}}} \mathbb{P}_{x,y}(\tau_1(x) \in ds, x(s) \in du, y(s) \in dv) \\ &\leq \frac{1}{2} \sup_{\substack{|x| \leq 1 \\ y \in \mathbb{Y}}} \mathbb{P}(\sup_{t \geq 0} |X(t, 0, y)x| \geq 2^{l\alpha}) = \frac{1}{2} g_l. \end{aligned}$$

Hence  $g_l \leq \frac{1}{2^l}$  for any  $l \in \mathbb{N}$ . Let us denote

$$\zeta := \sup_{t \geq 0} |x(t, 0, x, y)|^p.$$

It is clearly to see that for all  $p > 0$ ,  $x \in \mathbb{R}^n$ ,  $y \in \mathbb{Y}$  it may be written

$$\begin{aligned} \mathbb{E} \zeta &\leq |x|^p \sup_{|x| \leq 1} \mathbb{E} \zeta \leq \sum_{l=1}^{\infty} 2^{l\alpha p} \mathbb{P}(\sup_{t \geq 0} |x(t, 0, x, y)| \geq 2^{(l-1)\alpha}) \\ &\leq \sum_{l=1}^{\infty} 2^{l\alpha p} 2^{-l} |x|^p := K_1 |x|^p. \end{aligned}$$

Therefore random variable  $\zeta$  has expectation for all  $x \in \mathbb{R}^n$ ,  $y \in \mathbb{Y}$ ,  $p \in (0, \alpha^{-1})$ . According to Lemma's conditions the solution of (5)  $x(t, 0, x, y)$  tends to 0 almost sure as  $t$  tends to  $\infty$  uniformly on  $y \in \mathbb{Y}$  and by the Lebesgue Theorem one can write

$$\lim_{t \rightarrow \infty} \sup_{y \in \mathbb{Y}} \mathbb{E} |x(t + s, s, x, y)|^p = 0$$

for all  $x \in \mathbb{R}^n$ ,  $p \in (0, \alpha^{-1})$ . Besides, not complicatedly to verify that this convergence is uniform on  $x$  in the ball  $B_1(0)$  and  $s \geq 0$ , i.e.

$$\lim_{t \rightarrow \infty} \sup_{\substack{x \in B_1(0) \\ y \in \mathbb{Y}}} \mathbb{E} |x(t + s, s, x, y)|^p = 0.$$

Now we can choose a number  $T$  so large then the inequality

$$\sup_{y \in \mathbb{Y}} \mathbb{E} |x(t + s, s, x, y)|^p \leq |x|^p e^{-1}$$

is fulfilled and further, by using the inequality

$$\begin{aligned} \mathbb{E} |x(lT, 0, x, y)|^p &= \int_{\mathbb{R}^n} \int_{\mathbb{Y}} \mathbb{P}(x, y, (l-1)T, du, dv) \mathbb{E} |x(T, 0, u, v)|^p \\ &\leq e^{-1} \mathbb{E} |x((l-1)T, 0, x, y)|^p, \end{aligned}$$

where  $\mathbb{P}(x, y, t, du, dv)$  is transition probability of homogeneous Markov process  $\{x(t), y(t)\}$ , one can write

$$\mathbb{E} |x(t, 0, x, y)|^p \leq K_1 e^{-[\frac{t}{T}]T} |x|^p,$$

where  $[a]$  is integer of number  $a$ . This inequality completes the proof.

To analyze the behaviour of solutions of (5) one may use well known the Dynkin formula [2]

$$\mathbb{E}_{x,y}^{(u)} v(x(\tau_r(t)), y(\tau_r(t))) = v(x, y) + \mathbb{E}_{x,y}^{(u)} \left\{ \int_u^{\tau_r(t)} (L_0 v)(x(s), y(s)) ds \right\}, \quad (11)$$

where the indexes of expectation denote the condition  $x(u) = x$ ,  $y(u) = y$  and  $\tau_r(t) = \min\{\tau_r, t\}$ ,  $\tau_r = \inf\{t > u : x(t, u, x, y) \notin B_r(0)\}$ . If  $u = 0$ , then upper index will be absent.

If for all  $t \geq u \geq 0$  there exist the expectations  $\mathbb{E}_{x,y}v(x(t), y(t))$  and  $\mathbb{E}_{x,y}(L_0v)(x(t), y(t))$  one can use the Dynkin formula (11) in the more simple form

$$\mathbb{E}_{x,y}^{(u)}v(x(t), y(t)) = v(x, y) + \int_u^t \mathbb{E}_{x,y}^{(u)}(L_0v)(x(s), y(s)) ds. \quad (12)$$

Sometimes it is necessary to use the Lyapunov functions depending also on argument  $t$ . If the function  $v(t, x, y)$  belongs (as the function of arguments  $x$  and  $y$ ) to the region of definition of infinitesimal operator  $L_0$  and has continuous  $t$ -derivative, one may use the Dynkin formula (11) in the form

$$\begin{aligned} \mathbb{E}_{x,y}^{(u)}v(\tau_r(t), x(\tau_r(t)), y(\tau_r(t))) &= \\ &= v(u, x, y) + \mathbb{E}_{x,y}^{(u)} \left\{ \int_u^{\tau_r(t)} \left( \frac{\partial}{\partial s} + L_0 \right) v(s, x(s), y(s)) ds \right\}, \end{aligned}$$

or formula (12) in the form

$$\begin{aligned} \mathbb{E}_{x,y}^{(u)}v(t, x(t), y(t)) &= \\ &= v(u, x, y) + \int_u^t \mathbb{E}_{x,y}^{(u)} \left\{ \left( \frac{\partial}{\partial s} + L_0 \right) v(s, x(s), y(s)) \right\} ds. \end{aligned} \quad (13)$$

Besides Dynkin formula and the Second Lyapunov method one can use also well known the supermartingale inequality [1] for positive supermartingale  $\{\xi(t), \mathfrak{F}^t\}$  with filtration  $\mathfrak{F}^t$  in the form

$$\mathbb{P}(\sup_{t \geq u} \xi(t) \geq c) \leq \frac{1}{c} \mathbb{E} \xi(u). \quad (14)$$

**Lemma 2.** *The trivial solution of equation (5) is exponentially  $p$ -stable if and only if there exists the Lyapunov function  $v(x, y)$ , which satisfies the conditions*

$$c_1|x|^p \leq v(x, y) \leq c_2|x|^p, \quad c_1 > 0 \quad (15)$$

$$L_0v(x, y) \leq -c_3|x|^p, \quad c_3 > 0 \quad (16)$$

for all  $x \in \mathbb{R}^n$ ,  $y \in \mathbb{Y}$  with some positive  $p$ .

**Proof.** Let there exists above mentioned the Lyapunov function. It is clearly to verify that

$$\left( \frac{\partial}{\partial s} + L_0 \right) \left( v(x, y) e^{\frac{c_3}{c_2} t} \right) \leq 0,$$

and then one can write

$$\mathbb{E}_{x,y}v(x(t), y(t)) e^{\frac{c_3}{c_2} t} \leq v(x, y) \leq c_2|x|^p$$

for all  $t > 0$ ,  $x \in \mathbb{R}^n$  and  $y \in \mathbb{Y}$ . Hence

$$\mathbb{E}_{x,y}|x(t)|^p \leq \frac{1}{c_1} e^{-\frac{c_3}{c_2}t} \mathbb{E}_{x,y} v(x(t), y(t)) e^{\frac{c_3}{c_2}t} \leq \frac{c_2}{c_1} e^{-\frac{c_3}{c_2}t} |x|^p$$

and the equation (5) is exponentially  $p$ -stable. By using the solutions  $x(t+s, s, x, y)$  of the equation (5) one can construct for any  $T > 0$  function

$$v(x, y) := \int_0^T \mathbb{E}|x(s+t, s, x, y)|^p dt, \quad (17)$$

which do not depend on  $s$  owing to homogeneity of Markov process  $y(t)$ . It is easily to verify that under conditions that the matrix  $A(y)$  is uniformly bounded, that is,  $\sup_{y \in \mathbb{Y}} \|A(y)\| := a < \infty$  this function satisfies the conditions (15). Let  $L_0$  be the weak infinitesimal operator of the pair  $\{x(t), y(t)\}$ . If the trivial solution of equation (5) is exponentially  $p$ -stable, one can write the inequality

$$\begin{aligned} L_0 v(x, y) &= \lim_{\delta \rightarrow 0} \frac{1}{\delta} \left[ \int_0^T \mathbb{E}_{x,y} \{ \mathbb{E}_{x(\delta), y(\delta)} |x(t)|^p \} dt - \int_0^T \mathbb{E}_{x,y} |x(t)|^p dt \right] \\ &= \lim_{\delta \rightarrow 0} \frac{1}{\delta} \left[ \int_0^T \mathbb{E}_{x,y} |x(t+\delta)|^p dt - \int_0^T \mathbb{E}_{x,y} |x(t)|^p dt \right] \\ &= \mathbb{E}_{x,y} |x(T)|^p - |x|^p \leq (Me^{-\gamma T} - 1)|x|^p, \end{aligned}$$

where  $M$  and  $\gamma$  are constants from definition of exponential  $p$ -stability. Now we can put  $T = (\ln 2 + \ln M)/\gamma$  and proof is complete.

**Corollary 1.** *In the conditions of Lemma 2 the trivial solution of equation (5) is asymptotically stochastically stable.*

**Proof.** Due to formula (16) for  $\bar{v}(t, x, y) = v(x, y)e^{\frac{c_3}{c_2}t}$  one may conclude that random process

$$\xi(t) := v(x(t), y(t))e^{\frac{c_3}{c_2}t}$$

is positive supermartingale. Hence

$$\begin{aligned} \sup_{y \in \mathbb{Y}} \mathbb{P}(\sup_{t \geq 0} |x(t, 0, x, y)| > \varepsilon) &= \sup_{y \in \mathbb{Y}} \mathbb{P}(\sup_{t \geq 0} |x(t, 0, x, y)|^p > \varepsilon^p) \\ &\leq \sup_{y \in \mathbb{Y}} \mathbb{P}_{x,y}(\sup_{t \geq 0} \{ \frac{1}{c_1} v(x(t), y(t)) \} > \varepsilon^p) \\ &= \sup_{y \in \mathbb{Y}} \mathbb{P}_{x,y}(\sup_{t \geq 0} \{ \frac{1}{c_1} \xi(t) e^{-\frac{c_3}{c_2}t} \} > \varepsilon^p) \\ &\leq \sup_{y \in \mathbb{Y}} \mathbb{P}_{x,y}(\sup_{t \geq 0} \xi(t) > \varepsilon^p c_1) \leq \frac{1}{\varepsilon^p c_1} \mathbb{E}_{x,y} \xi(0) \leq \frac{c_2}{\varepsilon^p c_1} |x|^p \end{aligned}$$

and trivial solution of (5) is stochastically stable almost sure. Now to prove asymptotic stochastic stability one can apply the supermartingale inequality (14) and write the inequalities

$$\begin{aligned} \sup_{y \in \mathbb{Y}} \mathbb{P}(\sup_{t \geq u} |x(t, u, x, y)| > c) &= \sup_{y \in \mathbb{Y}} \mathbb{P}(\sup_{t \geq u} |x(t, u, x, y)|^p > c^p) \\ &\leq \sup_{y \in \mathbb{Y}} \mathbb{P}_{x,y}^{(u)}(\sup_{t \geq u} \{\frac{1}{c_1} v(x(t), y(t))\} > c^p) \\ &\leq \sup_{y \in \mathbb{Y}} \mathbb{P}_{x,y}^{(u)}(\sup_{t \geq u} \{\frac{1}{c_1} \xi(t) e^{-\frac{c_3}{c_2} t}\} > c^p) \\ &\leq \sup_{y \in \mathbb{Y}} \mathbb{P}_{x,y}^{(u)}(\sup_{t \geq u} \{\frac{1}{c_1} \xi(t) e^{-\frac{c_3}{c_2} u}\} > c^p) \leq \frac{1}{c^p c_1} \mathbb{E} \xi(u) \leq \frac{c_2}{c^p c_1} |x|^p e^{-\frac{c_3}{c_2} u}. \end{aligned}$$

## 2 Stochastic stability by linear approximation

In this section we will consider the quasilinear equation

$$\frac{d\tilde{x}}{dt} = A(y(t))\tilde{x} + g(\tilde{x}, y(t)), \quad (18)$$

under conditions that the matrix  $A(y)$  and Markov process  $y(t)$  satisfy the conditions of the Section 1, the function  $g(x, y)$  has bounded continuous  $x$ -derivative with conditions  $g(0, y) \equiv 0$ , and for any  $r > 0$  its  $x$ -derivative is uniformly bounded at any ball  $B_r(0)$ , i.e.

$$\sup_{\substack{y \in \mathbb{Y} \\ x \in B_r(0)}} \|D_x g(x, y)\| := g_r < \infty \quad (19)$$

**Theorem 1.** *If the equation (5) is asymptotically stochastically stable and  $\lim_{r \rightarrow 0} g_r = 0$ , then the trivial solution of equation (18) is asymptotically stochastically stable.*

**Proof.** Side by side with the equation (18) we will consider the equation (5) as an equation of its linear approximation. Due to Lemma 1 and Lemma 2 we can construct the Lyapunov function (17) with some small positive  $p$ . Because the matrix-valued function  $D_x x(t, 0, x, y)$  is the Cauchy matrix of the equation (5) it permits the estimation

$$\sup_{y \in \mathbb{Y}} \mathbb{E} \|D_x x(t + s, s, x, y)\|^p \leq h_2 e^{-\gamma t}$$

with some positive constants  $h, \gamma$  for all  $t > 0$ . Therefore the above Lyapunov function satisfies the conditions (15)-(16) and by construction for all  $x \neq 0$  has  $x$ -derivative satisfying the inequalities

$$\begin{aligned} |\nabla_x v(x, y)| &= \left| \int_0^T \mathbb{E} \{ \nabla_x |x(t + s, s, x, y)|^p \} dt \right| \\ &\leq p \int_0^T \mathbb{E} \{ |x(t + s, s, x, y)|^{(p-2)} | \{ D_x x(t + s, s, x, y) \} x(t + s, s, x, y) | \} dt \\ &\leq p |x|^{(p-1)} \int_0^T \sup_{y \in \mathbb{Y}} \mathbb{E} \|D_x x(t + s, s, x, y)\|^p dt \leq c_3 |x|^{(p-1)} \end{aligned}$$

with some positive  $c_3$ . Because the above estimations do not depend on initial time moment  $s$  we will put for simplicity  $s = 0$ . Now one can estimate the function  $Lv(x, y)$  where  $L$  is weak infinitesimal operator of the pair  $\{\tilde{x}(t), y(t)\}$ :

$$\begin{aligned} Lv(x, y) &:= (A(y)x + g(x, y), \nabla_x)v(x, y) + Qv(x, y) \\ &= L_0v(x, y) + (g(x, y), \nabla_x)v(x, y) \\ &\leq -\frac{1}{2}|x|^p + c_3|x|^p|g(x, y)| \leq (g_rc_3 - \frac{1}{2})|x|^p \end{aligned}$$

for all  $y \in \mathbb{Y}$ ,  $x \in B_r(0)$ ,  $r > 0$ . Hence, due to Dynkin formula, we may use inequality

$$\begin{aligned} \mathbb{E}_{x,y}^{(u)}v(\tilde{x}(\tau_r(t)), y(\tau_r(t))) &= v(x, y) + \mathbb{E}_{x,y}^{(u)}\left\{\int_u^{\tau_r(t)} (Lv)(\tilde{x}(s), y(s)) ds\right\} \\ &\leq v(x, y) + (g_rc_3 - \frac{1}{2})\mathbb{E}_{x,y}^{(u)}\left\{\int_u^{\tau_r(t)} |\tilde{x}(s)|^p ds\right\} \end{aligned} \quad (20)$$

for all  $y \in \mathbb{Y}$ ,  $x \in B_r(0)$ ,  $r > 0$ ,  $t \geq u \geq 0$ . If  $r$  is sufficiently small number the second summand in the right hand part of inequality (16) is nonpositive. Hence the stochastic process  $v(\tilde{x}(\tau_r(t)), y(\tau_r(t)))$  is supermartingale and we can write the inequalities

$$\begin{aligned} \mathbb{P}_{x,y}(\sup_{t \geq 0} |\tilde{x}(t)| > \varepsilon) &= \mathbb{P}_{x,y}(\sup_{t \geq 0} |\tilde{x}(t)|^p > \varepsilon^p) \\ &= \mathbb{P}_{x,y}(\sup_{t \geq 0} |\tilde{x}(\tau_r(t))|^p > \varepsilon^p) \leq \mathbb{P}_{x,y}(\sup_{t \geq 0} v(\tilde{x}(\tau_r(t)), y(\tau_r(t)))) \\ &> c_1\varepsilon^p \leq \frac{v(x, y)}{c_1\varepsilon^p} \leq \frac{c_2\delta^p}{c_1\varepsilon^p} \end{aligned} \quad (21)$$

for all  $y \in \mathbb{Y}$ ,  $x \in B_\delta(0)$ ,  $\delta \in (0, \varepsilon)$ ,  $\varepsilon \in (0, r)$  and sufficiently small  $r > 0$ . The local stability almost sure immediately follows from these inequalities. Let us define function

$$h_R(r) = \begin{cases} 1, & \text{for } x \in [0, R) \\ \frac{2R-r}{R}, & \text{for } x \in [R, 2R) \\ 0, & \text{for } x \geq 2R. \end{cases}$$

The differential equation

$$\frac{dx_R}{dt} = A(y(t))x_R + h_R(|x_R(t)|)g(x_R, y(t)) \quad (22)$$

has unique solution of the Cauchy problem  $x_R(0) = x$  because function  $h_R(|x|)g(x, y)$  satisfies the Lipschitz condition with constant  $c_{2R}$ . Hence the pair  $\{x_R(t), y(t)\}$  is Markov process with weak infinitesimal operator  $L_R$  defined by equality

$$\begin{aligned} L_Rv(x, y) &= (A(y)x, \nabla_x)v(x, y) + (h_R(|x|)g(x, y), \nabla_x)v(x, y) + Qv(x, y) \\ &= L_0v(x, y) + (h_R(|x|)g(x, y), \nabla_x)v(x, y) \end{aligned}$$

and choosing  $R$  such small that  $(c_{2R}c_3 - \frac{1}{2}) := -c_4 < 0$  one can write the inequality

$$L_Rv(x, y) \leq -c_4|x|^p.$$

Therefore

$$\begin{aligned}\mathbb{E}_{x,y}^{(u)} v(x_R(t), y(t)) &\leq v(x, y) - c_4 \int_u^t \mathbb{E}_{x,y}^{(u)} |x_R(s)|^p ds \\ &\leq v(x, y) - \frac{c_4}{c_1} \int_u^t \mathbb{E}_{x,y}^{(u)} v(x_R(s), y(s)) ds\end{aligned}\quad (23)$$

for all  $t \geq u \geq 0$ . Hence the stochastic process  $v(x_R(t), y(t))$  is positive supermartingale and one can write

$$\begin{aligned}\mathbb{P}_{x,y}(\sup_{t \geq s} |x_R(t)| > \varepsilon) &= \mathbb{P}_{x,y}(\sup_{t \geq s} |x_R(t)|^p > \varepsilon^p) \\ &\leq \mathbb{P}_{x,y}(\sup_{t \geq s} v(x_R(t), y(t)) > c_1 \varepsilon^p) \leq \frac{1}{c_1 \varepsilon^p} \mathbb{E}_{x,y} v(x_R(s), y(s))\end{aligned}\quad (24)$$

for all  $y \in \mathbb{Y}$ ,  $x \in B_R(0)$ ,  $\varepsilon \in (0, R)$  and sufficiently small  $R > 0$ . It is not complete to get the inequality

$$\mathbb{E}_{x,y} v(x_R(t), y(t)) \leq v(x, y) e^{-\frac{c_4}{c_1} t} \leq c_2 |x|^p e^{-\frac{c_4}{c_1} t}$$

from the inequality(19), and then f it can be written

$$\mathbb{P}_{x,y}(\sup_{t \geq s} |x_R(t)| > \varepsilon) \leq \frac{c_2 |x|^p}{\varepsilon^p c_1} e^{-\frac{c_4}{c_1} s}.$$

Hence all solutions of the equation (22) which have start at  $t = 0$  in the ball  $B_\varepsilon(0)$  with  $\varepsilon \in (0, R)$  and sufficiently small  $R$  tend to 0 with probability one. But up to time of the ball  $B_\varepsilon(0)$  leaving the solutions of the equations (18) and (22) with the same initial conditions in the ball  $B_\varepsilon(0)$  are coincident. So all solutions of (18) which do not leave the ball  $B_\varepsilon(0)$  with sufficiently small  $\varepsilon$  tend to zero with probability one and the proof is complete.

## References

- [1] DOOB, J. L.,: *Stochastic Processes*. John Willey & Sons, New York, 1953
- [2] DYNKIN, E. B.,: *Markov Processes*. Springer-Verlag, Berlin, 1965.
- [3] KHASHMISKY, R. Z.,: *Stochastic Stability of Differential Equations*. Kluwer Academic Pubs., Norwell, MA, 1980.
- [4] SKOROKHOD, A.V.,: *Asymptotic Methods of the Theory of Stochastic Differential Equations*. AMS, Providence, RI, 1989.
- [5] CARKOV, Jevgenijs: Asymptotic methods for stability analysis Markov of impulse dynamical systems, In *Advances of Stability Theory of the End of XXth Century. Stability and Control: Theory, Methods and Applications*. Gordon and Breach Science Publishers, London, 2000. p.p. 251–264.

**Current address**

**Jevgeņijs Čarkovs, professor**

Probability and Statistics Chair, Riga Technical University,  
Kaļķu iela 1, Riga, LV-1658, Latvia, tel. +371 26549111  
e-mail: carkovs@latnet.lv

**Kālis Šadurskis, professor**

Probability and Statistics Chair, Riga Technical University,  
Kaļķu iela 1, Riga, LV-1658, Latvia, tel. +371 67089517  
e-mail: skarlis@latnet.lv

# ON STABILITY ANALYSIS OF QUASILINEAR DIFFERENCE EQUATIONS IN BANACH SPACE (SPECTRAL THEORY APPROACH)

CARKOVŠ Jevgenijs, (LV), SLYUSARCHUK Vasyl, (UA)

**Abstract.** The paper deals with the mappings of Banach space  $\mathcal{E}$  given in a form of quasilinear difference equation

$$x_{n+1} = \mathbf{A}x_n + \mathbf{F}_n(x_n), \quad n \geq 0 \quad (1)$$

where  $\mathbf{A}$  is linear continuous operator,  $\{\mathbf{F}_n : \mathcal{E} \rightarrow \mathcal{E}\}$  are nonlinear bounded operators satisfying identity  $\mathbf{F}_n(0) \equiv 0$ . Side by side with the above equation we consider an equation of the first approximation, that is, the linear difference equation

$$y_{n+1} = \mathbf{A}y_n, \quad n \geq 0 \quad (2)$$

We will discuss the assertions which guarantee local stability or instability for the trivial solution of (1) if (2) to be of this specificity. The proposal paper not only generalizes well known finite dimensional stability analysis results for quasilinear difference equations. Using spectral properties of operator  $\mathbf{A}$  as a basis, our research shows that the infinite dimension of the space  $\mathcal{E}$  not only strongly complicates computations and proofs of relevant theorems on stability analysis by the first approximation but also can have significant influence to statement of these results.

**Key words and phrases.** Quasilinear difference equations; Lyapunov stability; Instability..

*Mathematics Subject Classification.* Primary 65P35; Secondary 39A11.

## 1 Notations, main definitions, and auxiliary assertions

We will follow the giving below classical notations of linear operator theory [12]:

$\mathbb{L}(\mathcal{E})$  – Banach algebra of linear continuous operators with unit  $\mathbf{I}$ ;

$\mathbb{K}(\mathcal{E})$  – subset of compact operators in  $\mathbb{L}(\mathcal{E})$ ;

$\text{Ker}(\mathbf{A})$  – kernel of operator  $\mathbf{A} \in \mathbb{L}(\mathcal{E})$ ;

$\text{Im}(\mathbf{A})$  – image of operator  $\mathbf{A} \in \mathbb{L}(\mathcal{E})$ ;

$\sigma(\mathbf{A})$  – spectrum of operator  $\mathbf{A} \in \mathbb{L}(\mathcal{E})$ , that is,  
 $\lambda \in \sigma(\mathbf{A}) \Leftrightarrow \text{Im}(\mathbf{A} - \lambda \mathbf{I}) \neq \mathcal{E}$ ;

$r(\mathbf{A}) := \max\{|\lambda| : \lambda \in \sigma(\mathbf{A})\}$  – spectral radius of operator  $\mathbf{A} \in \mathbb{L}(\mathcal{E})$ .

The trivial solution  $x_n \equiv 0$  is fixed point of mpping  $\mathbf{A}x + \mathbf{F}(x)$  and we will discuss a behavior of iterations (1) in some neighbourhood of it. The trivial solution of (1) is referred to as

- *stable if for any positive number  $\varepsilon$  and number  $n_0 \in \mathbf{N} \cup \{0\}$  there exists such a number  $\delta = \delta(\varepsilon, n_0) > 0$ , that for any solution  $x_n$  of this equation the inequality  $\sup_{n \geq n_0} \|x_n\| < \varepsilon$  follows inequality  $\|x_{n_0}\| < \delta$ ;*
- *instable if for some  $\varepsilon > 0$ ,  $n_0 \in \mathbf{N} \cup \{0\}$ , and any  $\delta > 0$  there exists such a solution  $x_n$  of this equation that  $\|x_{n_0}\| < \delta$  and  $\sup_{n \geq n_0} \|x_n\| \geq \varepsilon$ ;*
- *asymptotically stable if this solution is stable and for any  $n_0 \in \mathbf{N} \cup \{0\}$  there exists such a number  $\gamma = \gamma(n_0) > 0$  that from inequality  $\|x_{n_0}\| < \gamma$  it follows the equality  $\lim_{n \rightarrow \infty} \|x_n\| = 0$ ;*
- *exponential stable if for any  $n_0 \in \mathbf{N} \cup \{0\}$  there exist such numbers  $M = M(n_0) \geq 1$  and  $q = q(n_0) \in (0, 1)$  that*

$$\forall n \geq n_0 : \|x_n\| \leq Mq^{n-n_0}\|x_{n_0}\| \quad (3)$$

*for any solution of this equation;*

- *local exponential stable if for any  $n_0 \in \mathbf{N} \cup \{0\}$  there exist such numbers  $M = M(n_0) \geq 1$ ,  $q = q(n_0) \in (0, 1)$ , and  $r = r(n_0)$  that for any solution of this equation from inequality  $\|x_{n_0}\| < r$  follows an inequality (3).*

In the subsequent text of this paper we will need some of our previous results citing below.

**Theorem 1.1** *The following assertions are implications:*

- (i) *the trivial solution of (2) is exponential stable;*
- (ii)  $r(\mathbf{A}) < 1$ ;
- (iii) *the series  $\sum_{k=0}^{\infty} \|\mathbf{A}^k\|$  converges.*

**Theorem 1.2** *Let  $\mathcal{E}$  be a complex Banach space, and  $\mu$  is a boundary point of the set  $\sigma(\mathbf{A}) \setminus \{0\}$ . For any  $\delta > 0$  and  $m \in \mathbf{N}$  there exists such a vector  $\xi$  that*

$$(1 - \delta)|\mu|^n \leq \|\mathbf{A}^n \xi\| \leq (1 + \delta)|\mu|^n \|\xi\|$$

*for all  $n = \overline{0, m}$ .*

**Theorem 1.3** *Let  $\mathcal{E}$  be a real Banach space, and  $\mu$  is a boundary point of the set  $\sigma(\mathbf{A}) \setminus \{0\}$ . For any  $\delta > 0$  and  $m \in \mathbf{N}$  there exist such an integer number  $m_0 \geq m$  and vector  $u \in \mathcal{E}$  with norm  $|u| = 1$  that*

$$\|\mathbf{A}^n u\| \leq (\sqrt{2} + \delta) |\mu|^n \text{ for any } n = \overline{0, m_0}$$

and

$$\|\mathbf{A}^n u\| \geq (1 - \delta) |\mu|^{m_0}$$

The proofs of these results one can find in the papers [1] and [7].

## 2 Stability by the first approximation.

This Section is devoted to stability analysis of equation (1) by the first approximation. It seems naturally that the linear approximation equation (1) has to be subjected to condition  $r(\mathbf{A}) < 1$ . But it is wrong to believe that even the assertion  $r(\mathbf{A}) \leq 1$  is necessary in a case  $\dim \mathcal{E} = \infty$ . Corresponding examples we will give in Section 3. But the proposal in this Section results only generalize the well known similar theorems for finite dimensional space  $\mathcal{E}$  and therefore an assertion  $r(\mathbf{A}) < 1$  is present there.

**Theorem 2.1** *Assume that*

- (i) *the trivial solution of linear equation (2) is exponential stable;*
- (ii) *the operators  $\mathbf{F}_n, n \geq 0$  satisfy condition of uniform sufficiently small sublinear growth at zero, that is, for some positive number  $R$  there exists such a positive number  $\nu$  that*

$$\sup_{n \geq 0} \|\mathbf{F}_n x\| \leq \nu \|x\|, \text{ for } \|x\| \leq R$$

and

$$\nu \sum_{k=0}^{\infty} \|\mathbf{A}^k\| < 1 \tag{4}$$

*Then the trivial solution of (1) is local exponential stable.*

**Proof.** On the basis of the Theorem 3.1 and the first condition of the present theorem one may be certain of convergence of series  $\sum_{k=0}^{\infty} \|\mathbf{A}^k\| := M$ . This permits to introduce in the space  $\mathcal{E}$  new norm  $\|x\|_A = \sum_{k=0}^{\infty} \|\mathbf{A}^k x\|$  which satisfies inequality  $\|x\| \leq \|x\|_A \leq M\|x\|$ . Assuming  $\|x_n\| \leq R$  one can estimate the value of difference  $\Delta\|x_n\|_A = \|x_{n+1}\|_A - \|x_n\|_A$  for solution of

equation (1) in a following form:

$$\begin{aligned}
\Delta \|x_n\|_A &= \sum_{k=0}^{\infty} \|\mathbf{A}^k x_{n+1}\| - \sum_{k=0}^{\infty} \|\mathbf{A}^k x_n\| = \\
&= \sum_{k=0}^{\infty} \|\mathbf{A}^{k+1} x_n + \mathbf{A}^k \mathbf{F}_n x_n\| - \sum_{k=0}^{\infty} \|\mathbf{A}^k x_n\| \leq \\
&\leq \sum_{k=0}^{\infty} \|\mathbf{A}^{k+1} x_n\| + \sum_{k=0}^{\infty} \|\mathbf{A}^k \mathbf{F}_n x_n\| - \sum_{k=0}^{\infty} \|\mathbf{A}^k x_n\| = \\
&= -\|x_n\| + \sum_{k=0}^{\infty} \|\mathbf{A}^k \mathbf{F}_n x_n\| \leq -\|x_n\| + M \|\mathbf{F}_n x_n\| \leq \\
&\leq -\|x_n\| + M\nu \|x_n\| = (-1 + M\nu) \|x_n\| \leq \frac{M\nu - 1}{M} \|x_n\|_A \leq \frac{M\nu - 1}{M} \|x_n\|_A
\end{aligned}$$

Therefore under condition  $\|x_n\| \leq R$  one can apply inequality

$$\|x_{n+1}\|_A \leq q \|x_n\|_A$$

where  $q = 1 + \frac{M\nu-1}{M} < 1$  because by assumption of theorem  $M\nu < 1$  and  $M > 1$  by definition. Taking into account the above inequality and inequality  $\|x\| \leq \|x\|_A \leq M\|x\|$  one can may be sure that

$$\|x_n\| \leq Mq^{n-n_0} \|x_{n_0}\|, \quad n \geq n_0$$

for any  $\|x_{n_0}\| \leq \frac{R}{M}$ , where  $n_0$  – any integer number.

**Corollary 2.2** *If the trivial solution of linear equation (2) is exponential stable and*

$$\lim_{\|x\| \rightarrow 0} \frac{\sup_{n \geq 0} \|F_n x\|}{\|x\|} = 0$$

*then the trivial solution of (1) is local exponential stable.*

**Remark 2.3** *If (4) is not to hold then the trivial solution of equation (1) may not be local exponential stable.*

**Example 2.4** *Let us consider scalar difference equation*

$$x_{n+1} = ax_n + \nu|x_n|, \quad n \geq 0, \tag{5}$$

where  $a, \nu \in (0, 1)$ . The formula (4) for this equation has a form  $\nu \sum_{k=0}^{\infty} a^k < 1$  which equivalent to inequality  $a + \nu < 1$ . If  $a + \nu \geq 1$  then  $x_n = (a + \nu)^n x_0$  for each  $n \geq 0$  and trivial solution of (5) is not local exponential stable.

The proof technique of the above theorem may be used for more interesting assertion. Let  $\mathbf{F}_n^{[m]} : \mathcal{E} \longrightarrow \mathcal{E}$ ,  $n \geq m \geq 0$ , be mappings defined by equalities

$$\mathbf{F}_n^{[m]} = \mathbf{F}_n^{[m-1]}(\mathbf{A} + \mathbf{F}_{n-m}), \quad n \geq m \geq 0, \quad \mathbf{F}_n^{[0]} = \mathbf{F}_n$$

where  $\mathbf{A}$  and  $\mathbf{F}_n$ ,  $n \geq 0$  from equation (1).

**Theorem 2.5** Assume that:

(i) the trivial solution of equation (2) is exponential stable;

(ii) operators  $\mathbf{F}_n$ ,  $n \geq 0$  satisfy inequalities

$$\sup_{n \geq 0} \|\mathbf{F}_n x\| \leq \varphi(\|x\|), \quad \text{for } \|x\| \leq R,$$

where  $R > 0$  and  $\varphi : [0, R] \longrightarrow [0, +\infty)$  – is positive continuous definitely increasing function, and  $\varphi(0) = 0$ ;

(iii)  $\sup_{n \geq m} \|\mathbf{F}_n^{[m]} x\| \leq \nu \|x\|$ , for  $\|x\| \leq R$ , and  $\sqrt{\nu} \sum_{k=0}^{\infty} \|\mathbf{A}^k\| < 1$  for some integer  $m$  and positive number  $\nu$ .

Then the trivial solution of equation (1) is local exponential stable.

**Proof.** Further we will apply the same notations as in the proof of 2.1. It is easily seen that the solutions of (1) satisfies inequalities

$$x_{n+1} = \mathbf{A}x_n + \mathbf{F}_n^{[k]}x_{n-k}, \quad n \geq k,$$

for any  $k = \overline{0, m}$ . Under assumption  $\|x_{n-m}\|_A \leq R$  one can write the inequalities

$$\begin{aligned} \Delta \|x_n\|_A &= \sum_{k=0}^{\infty} \|\mathbf{A}^k x_{n+1}\| - \sum_{k=0}^{\infty} \|\mathbf{A}^k x_n\| = \\ &= \sum_{k=0}^{\infty} \|\mathbf{A}^{k+1} x_n + \mathbf{A}^k \mathbf{F}_n x_n\| - \sum_{k=0}^{\infty} \|\mathbf{A}^k x_n\| \leq \\ &\leq \sum_{k=0}^{\infty} \|\mathbf{A}^{k+1} x_n\| + \sum_{k=0}^{\infty} \|\mathbf{A}^k \mathbf{F}_n x_n\| - \sum_{k=0}^{\infty} \|\mathbf{A}^k x_n\| = \\ &= -\|x_n\| + \sum_{k=0}^{\infty} \|\mathbf{A}^k \mathbf{F}_n x_n\| \leq -\|x_n\| + M \|\mathbf{F}_n x_n\| \leq \\ &\leq -\frac{1}{M} \|x_n\|_A + M \|\mathbf{F}_n x_n\| = -\frac{1}{M} \|x_n\|_A + M \|\mathbf{F}_n^{[m]} x_{n-m}\| \leq \\ &\leq -\frac{1}{M} \|x_n\|_A + M \nu \|x_{n-m}\| \leq -\frac{1}{M} \|x_n\|_A + M \nu \|x_{n-m}\|_A \end{aligned}$$

and therefore

$$\|x_{n+1}\|_A \leq \left(1 - \frac{1}{M}\right) \|x_n\|_A + M \nu \|x_{n-m}\|_A \quad (6)$$

Let  $n_0$  be an arbitrary integer and  $\mu \in (0, R)$  is such a number, that for any  $\|x_{n_0}\| \leq \mu$  the solution of (1) with this initial condition satisfies inequality

$$\max_{0 \leq k \leq m} \|x_{n_0+k}\|_A \leq R.$$

It may be done because continuous function  $\varphi : [0, R] \rightarrow [0, +\infty)$  definitely increases and  $\varphi(0) = 0$ . Then based on (6) and inequalities

$$\left(1 - \frac{1}{M}\right) + M\nu < 1 \quad (7)$$

one can write  $\|x_n\|_A \leq R$  for each  $n \geq n_0$ . Now from (6) it is easily conclude that

$$\|x_{n+1}\|_A \leq \left(1 - \frac{1}{M} + M\nu\right) \max\{\|x_n\|_A \dots, \|x_{n-m}\|_A\},$$

for  $\|x_{n-m}\|_A \leq R$ . Then under condition  $\max_{0 \leq k \leq m} \|x_{n_0+k}\|_A \leq R$ , where  $\left[\frac{n-n_0}{m}\right]$  is ineger part of number  $\frac{n-n_0}{m}$  one can write inequalities

$$\|x_n\|_A \leq \left(1 - \frac{1 - M^2\nu}{M}\right)^{\left[\frac{n-n_0}{m}\right]} \max_{0 \leq k \leq m} \|x_{n_0+k}\|_A$$

for any  $n \geq n_0$ . From this and (7) follows that the trivial solution of equation (1) is local exponential stable.

The special case of the above theorem is following assertion.

**Corollary 2.6** *Assume that:*

- (i) *the first and the second assumptions of Theorem 2.4 are fulfilled;*
- (ii) *there exists such a number  $\nu > 0$  that*

$$\sup_{n \geq 1} \|\mathbf{F}_n(\mathbf{A}x + \mathbf{F}_{n-1}x)\| \leq \nu\|x\|, \text{ for } \|x\| \leq R,$$

*and*

$$\sqrt{\nu} \sum_{k=0}^{\infty} \|\mathbf{A}^k\| < 1.$$

*Then the trivial solution of equation (1) is local exponential stable.*

**Remark 2.7** *In Theorem 2.4 and Corollary 2.5 function  $\varphi(t)$  may be also of this a type as*

$$\lim_{t \rightarrow +0} \frac{\varphi(t)}{t} = +\infty. \quad (8)$$

The next example illustrates possibility of application of Theorem 2.4 or Corollary 2.5 when the Theorem 2.1 is unusable.

**Example 2.8** Let  $\mathbf{H}$  be a nilpotent operator satisfying equalities  $\mathbf{H}^2 \neq 0$  and  $\mathbf{H}^3 = 0$ , and  $\mathbf{F} : \mathcal{E} \rightarrow \mathcal{E}$  is operator defined by equality

$$\mathbf{F}(x) = \begin{cases} 0, & \text{if } x = 0, \\ \|x\|^{-1/2} \mathbf{H}^2 x + \|x\| \mathbf{H} x, & \text{if } x \neq 0. \end{cases}$$

Let us consider equation

$$x_{n+1} = \mathbf{H}x_n + \mathbf{F}(x_n), \quad n \geq 0, \quad (9)$$

It is easily seen that

$$\|\mathbf{F}(x)\| \leq \|\mathbf{H}^2\| \sqrt{\|x\|} + \|\mathbf{H}\| \|x\|^2$$

for each  $x \in \mathcal{E}$ , and

$$\|\|x\|^{-1/2} \mathbf{H}^2 x\| - \|x\| \|\mathbf{H}x\|\| \leq \|\mathbf{F}(x)\| \quad (10)$$

for  $x \in \mathcal{E} \setminus \{0\}$ , (that is the condition of Corollary 2.5 for function  $\varphi(t) = \|\mathbf{H}^2\| \sqrt{t} + \|\mathbf{H}\| t^2$ ). This function satisfies equality (8). This and (10) make it clear that the Theorem 4 may not be in use for stability analysis of (1). It is obviously that the first assumption of Theorem 5 is also fulfilled. Besides

$$\|\mathbf{F}(\mathbf{H}x + \mathbf{F}(x))\| = o(\|x\|) \text{ if } \|x\| \rightarrow 0 \quad (11)$$

because  $\mathbf{H}x + \mathbf{F}(x) \neq 0$  and therefore

$$\begin{aligned} \mathbf{F}(\mathbf{H}x + \mathbf{F}(x)) &= \|\mathbf{H}x + \mathbf{F}(x)\|^{-1/2} \mathbf{H}^2 (\mathbf{H}x + \mathbf{F}(x)) + \|\mathbf{H}x + \mathbf{F}(x)\| \mathbf{H} (\mathbf{H}x + \mathbf{F}(x)) = \\ &= \|\mathbf{H}x + \mathbf{F}(x)\|^{-1/2} \mathbf{H}^2 (\mathbf{H}x + \|x\|^{-1/2} \mathbf{H}^2 x + \|x\| \mathbf{H}x) + \\ &+ \|\mathbf{H}x + \|x\|^{-1/2} \mathbf{H}^2 x + \|x\| \mathbf{H}x\| \mathbf{H} (\mathbf{H}x + \|x\|^{-1/2} \mathbf{H}^2 x + \|x\| \mathbf{H}x) = \\ &= \|\mathbf{H}x + \|x\|^{-1/2} \mathbf{H}^2 x + \|x\| \mathbf{H}x\| (\mathbf{H}^2 x + \|x\| \mathbf{H}^2 x) \end{aligned}$$

Then

$$\|\mathbf{F}(\mathbf{H}x + \mathbf{F}(x))\| \leq \left( \|\mathbf{H}\| \|x\| + \|\mathbf{H}^2\| \sqrt{\|x\|} + \|\mathbf{H}\| \|x\|^2 \right) \|\mathbf{H}^2\| (1 + \|x\|) \|x\|$$

for any  $x \in \mathcal{E}$  and local exponential stability of the trivial solution of equation (9) follows formula (11). It should be mentioned that this assertion is trivial corollary of operator  $\mathbf{H}$  nilpotency because  $x_n = 0$  for any  $n \geq 3$ .

### 3 Conditions of instability by the first approximation.

In this section we will analyze equation (2) under assumption that  $r(\mathbf{A}) > 1$ . As in previous Section the proof in many respects uses Banach space renormalization technique. This permits not only sufficiently easily to prove results, which are similar to corresponding results for finite dimensional space  $\mathcal{E}$ , but also to derive theorems which are specific in a case  $\dim \mathcal{E} = \infty$ .

**Theorem 3.1** Assume that:

- (i)  $r(\mathbf{A}) > 1$ ;

(ii)  $\exists r \in [1, r(\mathbf{A})) : \sigma(\mathbf{A}) \cap \{z \in \mathbf{C} : |z| = r\} = \emptyset$ ;

(iii) there exist such positive numbers  $q$  and  $\rho$  that

$$\sup_{n \geq 0} \|\mathbf{F}_n x\| \leq q \|x\| \text{ for } x \in \{y \in \mathcal{E} : \|y\| \leq \rho\}.$$

Then for sufficiently small  $q$  the trivial solution of equation (1) is unstable.

**Proof.** From the beginning let us assume that  $r = 1$  and  $\sigma(\mathbf{A}) \cap \{z \in \mathbf{C} : |z| < 1\} \neq \emptyset$ . Let  $P_+$  and  $P_-$  be spectral projectors corresponding to spectral sets

$$\sigma_+(\mathbf{A}) = \sigma(\mathbf{A}) \cap \{z \in \mathbf{C} : |z| > 1\}$$

and

$$\sigma_-(\mathbf{A}) = \sigma(\mathbf{A}) \cap \{z \in \mathbf{C} : |z| < 1\}$$

These operators define spectral decomposition of space  $\mathcal{E}$   $\mathcal{E}_+ = P_+ \mathcal{E}$ ,  $\mathcal{E}_- = P_- \mathcal{E}$  and restrictions  $\mathbf{A}|_{\mathcal{E}_+}$  and  $\mathbf{A}|_{\mathcal{E}_-}$  of operator  $\mathbf{A}$  on these subspaces. By definition the spectrum of the above restrictions coincide with sets  $\sigma_+(\mathbf{A})$  and  $\sigma_-(\mathbf{A})$ , besides  $0 \notin \sigma_+(\mathbf{A})$ . Therefore operator  $\mathbf{A}|_{\mathcal{E}_+} : \mathcal{E}_+ \rightarrow \mathcal{E}_+$  is reversible and spectral radii of operators  $\mathbf{A}|_{\mathcal{E}_-}$  and  $(\mathbf{A}|_{\mathcal{E}_+})^{-1}$  less than 1. By the Theorem 1.1 the serieses

$$\sum_{k=1}^{\infty} \|(\mathbf{A}|_{\mathcal{E}_+})^{-k}\| \text{ and } \sum_{k=0}^{\infty} \|(\mathbf{A}|_{\mathcal{E}_-})^k\|.$$

are convergent and this permits to define in the space  $\mathcal{E}$  a new norm

$$\|x\|_A = \sum_{k=1}^{\infty} \|(\mathbf{A}|_{\mathcal{E}_+})^{-k} P_+ x\| + \sum_{k=0}^{\infty} \|(\mathbf{A}|_{\mathcal{E}_-})^k P_- x\|$$

Owing inequality  $m\|x\| \leq \|x\|_A \leq M\|x\|$  for all  $x \in \mathcal{E}$ , where  $m = \min \left\{ 1, \frac{1}{\|\mathbf{A}|_{\mathcal{E}_+}\|^{-1}} \right\} > 0$  and  $M = \sum_{k=1}^{\infty} \|(\mathbf{A}|_{\mathcal{E}_+})^{-k}\| + \sum_{k=0}^{\infty} \|(\mathbf{A}|_{\mathcal{E}_-})^k\| < \infty$  one makes sure that the norms  $\|\cdot\|$  and  $\|\cdot\|_A$  are

equivalent. It follows the inequalities

$$\begin{aligned}
\min \left\{ 1, \frac{1}{\|\mathbf{A}|_{E_+}\| - 1} \right\} \|x\| &\leq \\
&\leq \min \left\{ 1, \frac{1}{\|\mathbf{A}|_{E_+}\| - 1} \right\} \|P_+x\| + \|P_-x\| \leq \\
&\leq \frac{1}{\|\mathbf{A}|_{E_+}\| - 1} \|P_+x\| + \|P_-x\| = \\
&= \sum_{k=1}^{\infty} \|\mathbf{A}|_{E_+}\|^{-k} \|P_+x\| + \|P_-x\| = \\
&= \sum_{k=1}^{\infty} \|\mathbf{A}|_{E_+}\|^{-k} \left\| (\mathbf{A}|_{E_+})^k (\mathbf{A}|_{E_+})^{-k} P_+x \right\| + \|P_-x\| \leq \\
&\leq \sum_{k=1}^{\infty} \|\mathbf{A}|_{E_+}\|^{-k} \|\mathbf{A}|_{E_+}\|^k \left\| (\mathbf{A}|_{E_+})^{-k} P_+x \right\| + \|P_-x\| = \\
&= \sum_{k=1}^{\infty} \left\| (\mathbf{A}|_{E_+})^{-k} P_+x \right\| + \|P_-x\| \leq \\
&\leq \sum_{k=1}^{\infty} \left\| (\mathbf{A}|_{E_+})^{-k} P_+x \right\| + \sum_{k=0}^{\infty} \left\| (\mathbf{A}|_{E_-})^k P_-x \right\| = \\
&= \|x\|_A \leq \left( \sum_{k=1}^{\infty} \left\| (\mathbf{A}|_{E_+})^{-k} \right\| + \sum_{k=0}^{\infty} \left\| (\mathbf{A}|_{E_-})^k \right\| \right) \|x\|
\end{aligned}$$

Now one can apply the above constructed projective operators to solution  $x_n$  of equation (1)

$$\|P_+x_n\|_A = \sum_{k=1}^{\infty} \left\| (\mathbf{A}|_{E_+})^{-k} P_+x_n \right\|, \quad \|P_-x_n\|_A = \sum_{k=0}^{\infty} \left\| (\mathbf{A}|_{E_-})^k P_-x_n \right\|$$

and write a decomposition  $\|x_n\|_A = \|P_+x_n\|_A + \|P_-x_n\|_A$  Let us estimate each item taken separately:

$$\begin{aligned}
\Delta \|P_+x_n\|_A &= \|P_+x_{n+1}\|_A - \|P_+x_n\|_A = \\
&= \|P_+Ax_n + P_+\mathbf{F}_n x_n\|_A - \|P_+x_n\|_A \geq \\
&\geq \|P_+Ax_n\|_A - \|P_+x_n\|_A - \|P_+\mathbf{F}_n x_n\|_A = \\
&= \|P_+x_n\| - \|P_+\mathbf{F}_n x_n\|_A \geq \frac{1}{M} \|P_+x_n\|_A - \|P_+\mathbf{F}_n x_n\|_A
\end{aligned}$$

and  $\Delta \|P_-x_n\|_A \leq -\frac{1}{M} \|P_-x_n\|_A + \|P_-\mathbf{F}_n x_n\|_A$ . Therefore

$$\begin{aligned}
\|P_+x_{n+1}\|_A - \|P_+x_0\|_A &\geq \sum_{k=0}^n \left( \frac{1}{M} \|P_+x_k\|_A - \|P_+\mathbf{F}_k x_k\|_A \right), \\
\|P_-x_{n+1}\|_A - \|P_-x_0\|_A &\leq \sum_{k=0}^n \left( -\frac{1}{M} \|P_-x_k\|_A + \|P_-\mathbf{F}_k x_k\|_A \right)
\end{aligned}$$

and one may write an inequality

$$\begin{aligned}
\|x_{n+1}\|_A &\geq \|P_+x_{n+1}\|_A - \|P_-x_{n+1}\|_A \geq \\
&\geq \|P_+x_0\|_A - \|P_-x_0\|_A + \sum_{k=0}^n \left( \frac{1}{M} \|x_k\|_A - \|\mathbf{F}_k x_k\|_A \right)
\end{aligned}$$

Suppose that  $\|x_k\| < \rho$  for  $k = \overline{0, n}$ . Then  $\|F_n x_n\|_A \leq M \|F_n x_n\| \leq M q \|x_n\| \leq \frac{M}{m} q \|x_n\|_A$  and for initial vector  $x_0 = P_+ x_0$  we can use an inequality

$$\|x_{n+1}\|_A \geq \sum_{k=0}^n \left( \frac{1}{M} - \frac{M}{m} q \right) \|x_k\|_A + \|x_0\|_A \quad (12)$$

Note that for  $0 < q < \frac{m}{M^2}$  value  $\frac{1}{M} - \frac{M}{m} q$  is positive and therefore

$$\|x_n\|_A \geq \left( 1 + \frac{1}{M} - \frac{M}{m} q \right)^n \|x_0\|_A \quad (13)$$

An instability of trivial solution of (1) follows the above inequality because for  $\varepsilon < \rho$  and for any  $\|x_0\| = \|P_+ x_0\|$  there exists such a number  $n \in \mathbb{N}$  that  $\|x_n\| \geq \varepsilon$ . Note that if  $\sigma(\mathbf{A}) \cap \{z \in \mathbf{C} : |z| < 1\} = \emptyset$  the theorem can be proved with the help of norm  $\|x\|_A = \sum_{k=1}^{\infty} \|A^{-k} x\|$  in the same way as for  $\sigma(\mathbf{A}) \cap \{z \in \mathbf{C} : |z| < 1\} \neq \emptyset$ . First we shall show that for solution of (1) under assumption  $\max_{0 \leq k \leq n} \|x_k\| < \rho$  the formula (12) is true. Secondly as it has been done before one can establish the inequality (13) which convinces of instability of trivial solution of (1). Thus we have proved Theorem 3.1 for a case  $r = 1$ .

Let us assume now that  $r \in (1, r(\mathbf{A}))$ . Side by side with (1) we consider equation

$$x_{n+1} = \mathbf{A}x_n + \check{\mathbf{F}}_n x_n, \quad n \geq 0 \quad (14)$$

where

$$\check{\mathbf{F}}_n x = \begin{cases} \mathbf{F}_n x, & \text{for } \|x\| \leq \rho, \\ \frac{\|x\|}{\rho} \mathbf{F}_n \frac{\rho}{\|x\|} x, & \text{for } \|x\| > \rho. \end{cases}$$

Owing inequality  $\|\check{\mathbf{F}}_n x\| \leq q \|x\|$  for each  $x \in \mathcal{E}$  one can easily be certain that the trivial solutions of equations (1) and (2) are stable or instable concurrently. Substituting in (14)

$$x_n = r^n y_n \quad (15)$$

we will have for  $y_n$  equation

$$y_{n+1} = r^{-1} \mathbf{A} y_n + r^{-n-1} \check{\mathbf{F}}_n r^n y_n, \quad n \geq 0 \quad (16)$$

where  $r(r(A))^{-1} > 1$ ,  $\sigma(r^{-1} \mathbf{A}) \cap \{z \in \mathbf{C} : |z| = 1\} = \emptyset$ , and  $\|r^{-n-1} \check{\mathbf{F}}_n r^n x\| \leq r^{-1} q \|x\| \leq q \|x\|$  for any  $x \in \mathcal{E}$ . As it follows from our previous results, this inequality guarantees instability of the trivial solution of equation (16). With regard to equation (15) and inequality  $r > 1$  one may assert that the trivial solution of equation (14) is instable.

Let us remark that for a case  $\dim \mathcal{E} < \infty$  it follows that spectrum set  $\sigma(\mathbf{A})$  consists of finite number of points and therefore one may resign the third assertion of Theorem 3.1. This convinces of the following assertion. But if  $\dim \mathcal{E} = \infty$  the below example makes it clear that on the above this assertion may not be rejected.

**Example 3.2** ([2]). Let  $\mathbf{B} \in \mathbb{L}(\mathcal{E})$ ,  $\sigma(\mathbf{B}) = \{z \in \mathbb{C} : |z| \leq 1\}$ , and  $(\mathbf{B}_m)_{m \geq 0}$  be a sequence of nilpotent operators acting in Banach space  $\mathcal{E}$  satisfying following assumptions:

$$\lim_{n \rightarrow \infty} \|\mathbf{B}_n - \mathbf{B}\| = 0 \quad (17)$$

Applying the results of [10] one can construct the above mentioned operators for example in the spaces  $l_2$  or  $L_2([0, 1])$ . Assume that in (1)  $\mathbf{A} = e^{-\varepsilon \mathbf{I} + \mathbf{B}}$  with  $\varepsilon \in (0, 1)$ , and  $\mathbf{F}^{[m]}x = (e^{-\varepsilon \mathbf{I} + \mathbf{B}_m} - e^{-\varepsilon \mathbf{I} + \mathbf{B}})x$ ,  $x \in \mathcal{E}$ . By definition  $r(\mathbf{A}) > 1$ , and

$$\|\mathbf{F}^{[m]}x\| \leq \|e^{-\varepsilon \mathbf{I} + \mathbf{B}_m} - e^{-\varepsilon \mathbf{I} + \mathbf{B}}\| \|x\|, \quad x \in \mathcal{E}$$

Therefore it follows from (17) that  $\lim_{m \rightarrow \infty} \|\mathbf{F}^{[m]}x\| = 0$  and for any  $q > 0$  due to assumption (17) one can choose such an integer  $m$  that  $\|\mathbf{F}^{[m]}x\| \leq q\|x\|$  for any  $x \in \mathcal{E}$ . Besides equation (1) of our example may be rewritten in a following form

$$x_{n+1} = e^{-\varepsilon \mathbf{I} + \mathbf{B}_m} x_n, \quad n \geq 0,$$

and  $r(e^{-\varepsilon \mathbf{I} + \mathbf{B}_m}) = e^{-\varepsilon} < 1$ ,  $m \geq 1$ . Therefore the trivial solution of defined in our example difference equation (1) is asymptotically stable for whatever positive number  $q$ .

It should be mentioned that if  $\dim \mathcal{E} = \infty$  even under assumptions  $\lim_{\|x\| \rightarrow 0} \frac{\sup_{n \geq 0} \|\mathbf{F}_n x\|}{\|x\|} = 0$  and  $r(\mathbf{A}) > 1$  the trivial solution of (1) may be asymptotically stable. Corresponding example one can find in [5]. To resign the second assertion permits more rigid condition on behaviour of function  $\mathbf{F}_n(x)$  as  $\|x\| \rightarrow 0$ . In our previous paper [1] we have prove a following result.

**Theorem 3.3** Assume that:

- (i)  $r(\mathbf{A}) > 1$ ;
- (ii) there exist such positive number  $a, p$ , and  $\rho$  that  $\sup_{n \geq 0} \|\mathbf{F}_n x\| \leq a\|x\|^{1+p}$  for any  $x \in \{y \in \mathcal{E} : \|y\| \leq \rho\}$ .

Then the trivial solution of (1) is instable.

In this paper we prove more stronger result, weakening the second assertion of the above theorem.

**Theorem 3.4** . Assume that:

- (i)  $r(\mathbf{A}) > 1$ ;
- (ii) there exists such a continuous monotone function  $\{q(y), 0 \leq y \leq \rho\}$  that  $q(0) = 0$  and  $\rho$  that  $\sup_{n \geq 0} \|\mathbf{F}_n x\| \leq q(\|x\|)\|x\|$ ;
- (iii) there exist such a number  $\nu \in (0, \rho]$  and a sequence  $\{f(n), n \in \mathbb{N}\}$  that  $\|\mathbf{A}^n\| \leq f(n)(r(\mathbf{A}))^n$  for any  $n \in \mathbb{N}$  and series  $\sum_{k=1}^{\infty} f(k)q(\nu(r(\mathbf{A}))^{-k})$  converges.

Then the trivial solution of (1) is instable.

**Proof.** At first we assume that  $\mathcal{E}$  is a complex Banach space. Let  $\delta, P$  and  $\gamma$  are such positive numbers that  $1 + \delta < P < 2$ , and

$$\sum_{k=0}^{\infty} f(k)q(\gamma r^{-k}) < \frac{(P-1-\delta)r}{P} < \frac{r}{2} \quad (18)$$

Taking into account (18), a monotony of function  $q(y)$ , and choosing such a number  $\varepsilon \in (0, \gamma P^{-1}r^{-1})$ , and an integer  $n(\varepsilon) > 0$  that

$$\frac{\gamma}{r} \leq \varepsilon P r^{n(\varepsilon)} \leq \gamma \quad (19)$$

one can be certain of the inequality

$$\sum_{k=0}^{n-1} f(n-1-k)q(\varepsilon P r^k) < \frac{(P-1-\delta)r}{P}, \quad (20)$$

for any  $n = \overline{1, n(\varepsilon)}$ . Based on Theorem 1.2 one can find such a vector  $\xi \in \{x \in \mathcal{E} : \|x\| = 1\}$  that

$$(1-\delta)r^n \leq \|\mathbf{A}^n \xi\| \leq (1+\delta)r^n, \quad n = \overline{1, n(\varepsilon)} \quad (21)$$

Let us split the solution  $x_n$  of equation (1) with initial condition  $x_0 = \xi \in \{x \in \mathcal{E} : \|x\| = 1\}$  in a following form

$$x_n = x_{1,n} + x_{2,n}, \quad (22)$$

where  $x_{1,n} = \mathbf{A}^n x_0$ , and  $x_{2,n} = \sum_{k=0}^{n-1} \mathbf{A}^{n-1-k} F_k x_k$ ,  $n \geq 1$

In compliance with (21) and the second asseretion of theorem there exists such an integer  $m \in [0, n(\varepsilon)]$ , that

$$\|x_n\| \leq \varepsilon P r^n \quad (23)$$

for any  $n = \overline{0, m}$ . Therefore

$$\begin{aligned} \|x_{2,n}\| &= \left\| \sum_{k=0}^{n-1} \mathbf{A}^{n-1-k} F_k x_k \right\| \leq \sum_{k=0}^{n-1} \|A^{n-1-k}\| \|F_k x_k\| \leq \\ &\leq \sum_{k=0}^{n-1} f(n-1-k)r^{n-1-k}q(\varepsilon P r^k) \varepsilon P r^k = \\ &= \varepsilon r^n \frac{P}{r} \sum_{k=0}^{n-1} f(n-1-k)q(\varepsilon P r^k) \end{aligned}$$

and  $\|x_{1,n}\| \leq (1+\delta)\varepsilon r^n$  for any  $n \in [1, m]$ . Then

$$\forall n = \overline{1, m} : \|x_n\| \leq \left( 1 + \delta + \frac{P}{r} \sum_{k=0}^{n-1} f(n-1-k)q(\varepsilon P r^k) \right) \varepsilon r^n,$$

and, because from (20) follows inequality

$$1 + \delta + \frac{P}{r} \sum_{k=0}^{n-1} f(n-1-k)q(\varepsilon Pr^k) < P,$$

for  $n = \overline{1, n(\varepsilon)}$ , we may apply (23) for any  $n \in [1, n(\varepsilon)]$ . Applying (19)–(21), we can find lower bound for  $\|x_{n(\varepsilon)}\|$ :

$$\begin{aligned} \|x_{n(\varepsilon)}\| &\geq \|x_{1, n(\varepsilon)}\| - \|x_{2, n(\varepsilon)}\| \geq \\ &\geq (1 - \delta)\varepsilon r^{n(\varepsilon)} - \varepsilon r^{n(\varepsilon)} \frac{P}{r} \sum_{k=0}^{n(\varepsilon)-1} f(n(\varepsilon) - 1 - k)q(\varepsilon Pr^k) = \\ &= \varepsilon r^{n(\varepsilon)} \left( 1 - \delta - \frac{P}{r} \sum_{k=0}^{n(\varepsilon)-1} f(n(\varepsilon) - 1 - k)q(\varepsilon Pr^k) \right) \geq \\ &\geq \varepsilon r^{n(\varepsilon)} \left( 1 - \delta - \frac{P}{r} \cdot \frac{(P - 1 - \delta)r}{P} \right) \geq \frac{\gamma(2 - P)}{Pr} = a > 0 \end{aligned}$$

Therefore  $\|x_{n(\varepsilon)}\| \geq a$  for any arbitrarily small  $\varepsilon = \|x_0\|$  and the proof of theorem for a complex Banach space is completed. Now let  $\mathcal{E}$  be a real Banach space. Like before we can find such positive numbers  $\delta$ ,  $P$  and  $\gamma$  that  $\sqrt{2} + \delta < P < 2$  and

$$\sum_{k=0}^{\infty} f(k)q(\gamma r^{-k}) < \frac{(P - \sqrt{2} - \delta)r}{P} < \frac{r}{2} \quad (24)$$

For any  $\varepsilon \in (0, \gamma P^{-1}r^{-1})$  there exists such an integer  $n(\varepsilon) > 0$ , that  $\frac{\gamma}{r} \leq \varepsilon Pr^{n(\varepsilon)} \leq \gamma$ . Applying Theorem 1.3 one can choose a number  $m_0 \geq n(\varepsilon)$  and a vector  $u \in E$ ,  $\|u\| = 1$ , which permits write inequalities

$$\forall n = \overline{0, m_0} : \|A^n u\| \leq (\sqrt{2} + \delta) |\mu|^n \quad (25)$$

and

$$\|A^n u\| \geq (1 - \delta) |\mu|^{m_0} \quad (26)$$

Besides owing monotony of sequence  $f(n)$  from (24) follows inequality

$$\sum_{k=0}^{n-1} f(n-1-k)q(\varepsilon Pr^k) < \frac{(P - \sqrt{2} - \delta)r}{P} \quad (27)$$

for all  $n = \overline{1, m_0}$ . Let us choose such a number  $\varepsilon_1 \in (0, \varepsilon)$  that

$$\frac{\gamma}{r} \leq \varepsilon_1 Pr^{m_0} \leq \gamma \quad (28)$$

and estimate the solution  $x_n$  of (1) with initial condition  $x_0 = \varepsilon_1 u$ , splitting this in a form (22). Formula (25) and the second assertion of theorem guarantee existence such an integer  $m \in [0, m_0]$  that

$$\|x_n\| \leq \varepsilon_1 Pr^n \quad (29)$$

for any  $n = \overline{0, m}$ . Then

$$\begin{aligned}
 \|x_{2,n}\| &= \left\| \sum_{k=0}^{n-1} \mathbf{A}^{n-1-k} F_k x_k \right\| \leq \sum_{k=0}^{n-1} \|\mathbf{A}^{n-1-k}\| \|\mathbf{F}_k x_k\| \leq \\
 &\leq \sum_{k=0}^{n-1} f(n-1-k) r^{n-1-k} q(\varepsilon_1 P r^k) \varepsilon_1 P r^k = \\
 &= \varepsilon_1 r^n \frac{P}{r} \sum_{k=0}^{n-1} f(n-1-k) q(\varepsilon_1 P r^k) \\
 \|x_{1,n}\| &\leq (\sqrt{2} + \delta) \varepsilon_1 r^n
 \end{aligned}$$

and one can apply inequality

$$\|x_n\| \leq \left( \sqrt{2} + \delta + \frac{P}{r} \sum_{k=0}^{n-1} f(n-1-k) q(\varepsilon_1 P r^k) \right) \varepsilon_1 r^n$$

for each  $n = \overline{1, m}$ . Because from (27) follows formula

$$\sqrt{2} + \delta + \frac{P}{r} \sum_{k=0}^{n-1} f(n-1-k) q(\varepsilon_1 P r^k) < P$$

for any  $n = \overline{1, m_0}$ , we have proved inequality (29) for any integer  $n \in [1, m_0]$ .

To find lower bound of  $\|x_{m_0}\|$  one can use the formulae (25)–(27) and derive inequalities

$$\begin{aligned}
 \|x_{m_0}\| &\geq \|x_{1,m_0}\| - \|x_{2,m_0}\| \geq \\
 &\geq (1 - \delta) \varepsilon_1 r^{m_0} - \varepsilon_1 r^{m_0} \frac{P}{r} \sum_{k=0}^{m_0-1} f(m_0-1-k) q(\varepsilon_1 P r^k) = \\
 &= \varepsilon_1 r^{m_0} \left( 1 - \delta - \frac{P}{r} \sum_{k=0}^{m_0-1} f(m_0-1-k) q(\varepsilon_1 P r^k) \right) \geq \\
 &\geq \varepsilon_1 r^{m_0} \left( 1 - \delta - \frac{P}{r} \cdot \frac{(P - \sqrt{2} - \delta) r}{P} \right) \geq \frac{\gamma (1 + \sqrt{2} - P)}{P r} = a > 0
 \end{aligned}$$

Therefore the value of chosen solution  $\|x_{m_0}\|$  with satisfying equality  $\varepsilon_1 = \|x_0\|$  initial condition remains not less than  $a > 0$  for any arbitrary small number  $\varepsilon_1$ . The proof is completed.

**Example 3.5** Let us consider difference equation

$$x_{n+1} = \mathbf{A} x_n + \begin{cases} (1 - \ln \|x_n\|)^{-2-p} \mathbf{B} x_n, & \text{if } x_n \neq 0, \\ 0, & \text{if } x_n = 0, \end{cases} \quad (30)$$

where  $p > 0$ , operator  $\mathbf{A} \in \mathbb{L}(\mathcal{E})$  satisfies inequality

$$\forall n \in \mathbb{N} : \|\mathbf{A}^n\| \leq M(1 + n)2^n,$$

$\sigma(\mathbf{A}) = \{t : 0 \leq t \leq 2\}$ ,  $\mathbf{B} \in \mathbb{L}(\mathcal{E})$  – nontrivial operator, and  $\mathcal{E}$  – a complex Banach space. Now we choose sequence  $f(n) = M(1+n)$  and function

$$q(y) = \begin{cases} \|\mathbf{B}\| |1 - \ln y|^{-2-p}, & \text{for } y > 0, \\ 0, & \text{if } y = 0, \end{cases}$$

and substitute these in series  $\sum_{k=0}^{\infty} f(k)q(\nu(r(A))^{-k})$  from the third assertion of Theorem 3.4:

$$\sum_{k=0}^{\infty} \frac{M\|\mathbf{B}\|(1+k)}{(1 - \ln \nu + k \ln 2)^{2+p}}$$

Not so difficult to proof that this series converges for any  $\nu \in (0, 1)$ . From the above we can be sure that for equation (30) all assertions of theorem 3.4 are satisfied and therefore the trivial solution of (30) is instable.

**Remark 3.6** Theorem 3.3 is a sequence of Theorem 3.4.

**Proof.** Let us define sequence  $f(n) = \max_{s \in [0, n] \cap (\mathbb{N} \cup \{0\})} \|\mathbf{A}^s\| (r(\mathbf{A}))^{-s}$  and function

$$q(y) = \begin{cases} \left( \hat{f} \left( \frac{1}{\ln r(\mathbf{A})} \ln \frac{1}{y} \right) \right)^{-1} (1 - \ln y)^{-1-p}, & \text{for } y \in (0, 1], \\ 0, & \text{if } y = 0, \end{cases}$$

where  $p > 0$  and  $\hat{f}(t)$  is such a continuous monotony function that restriction  $\hat{f}|_{\mathbb{N} \cup \{0\}}$  onto  $\mathbb{N} \cup \{0\}$  coincide with above defined  $f(n)$ . By definition

$$\hat{f} \left( \frac{1}{\ln r(\mathbf{A})} \ln \frac{1}{y} \right) \geq \hat{f} \left( \frac{1}{\ln r(\mathbf{A})} \ln \frac{v}{y} \right)$$

for each  $y \in (0, v]$  and  $v \in (0, 1]$ . Therefore

$$\begin{aligned} \sum_{k=0}^{\infty} f(k)q(\nu(r(A))^{-k}) &\leq \sum_{k=0}^{\infty} f(k)q((r(A))^{-k}) = \\ &= \sum_{k=0}^{\infty} f(k) \left( \hat{f}(k) \right)^{-1} (1 + k \ln r(A))^{-1-p} = \\ &= \sum_{k=0}^{\infty} (1 + k \ln r(A))^{-1-p} < \infty \end{aligned}$$

To prove that

$$\lim_{y \rightarrow +0} \frac{y^\varepsilon}{q(y)} = 0 \quad \text{for any } \varepsilon > 0 \quad (31)$$

one may apply a substitution  $y = (r(\mathbf{A}))^{-t}$  and rewrite (31) in following form

$$\lim_{t \rightarrow +\infty} \frac{(r(\mathbf{A}))^{-\varepsilon t}}{q((r(\mathbf{A}))^{-t})} = 0 \quad \text{for any } \varepsilon > 0 \quad (32)$$

Taking into account the Gelfand formula  $r(\mathbf{A}) = \lim_{n \rightarrow +\infty} \sqrt[n]{\|\mathbf{A}^n\|}$  and equalities

$$\begin{aligned} \frac{y^\varepsilon}{q(y)} &= \frac{\hat{f}(t)(1+t \ln r(\mathbf{A}))^{1+p}}{(r(\mathbf{A}))^{\varepsilon t}} \leq \frac{f([t+1])(1+t \ln r(\mathbf{A}))^{1+p}}{(r(\mathbf{A}))^{\varepsilon[t]}} = \\ &= \frac{f([t+1])}{\sqrt{\varepsilon} (r(\mathbf{A}))^{\varepsilon[t]}} \cdot \frac{(1+t \ln r(\mathbf{A}))^{1+p}}{\sqrt{\varepsilon} (r(\mathbf{A}))^{\varepsilon[t]}}, \end{aligned}$$

$$\lim_{t \rightarrow +\infty} \frac{f([t+1])}{\sqrt{\varepsilon} (r(\mathbf{A}))^{\varepsilon[t]}} = 0,$$

$$\lim_{t \rightarrow +\infty} \frac{(1+t \ln r(\mathbf{A}))^{1+p}}{\sqrt{\varepsilon} (r(\mathbf{A}))^{\varepsilon[t]}} = 0,$$

we can get formula (32), which is equivalent to (31). Therefore if one may apply Theorem 3.4 then, based on (31), one also may apply Theorem 3.3.

It is well known [12] that the spectrum  $\sigma(\mathbf{A})$  may be presented as a sum of disjoint sets

$$\sigma(\mathbf{A}) = \sigma_p(\mathbf{A}) \cup \sigma_c(\mathbf{A}) \cup \sigma_r(\mathbf{A})$$

where

$$\lambda \in \sigma_p(\mathbf{A}) \Leftrightarrow \{\exists x \neq 0 : (\mathbf{A} - \lambda \mathbf{I})x = 0\};$$

$$\lambda \in \sigma_c(\mathbf{A}) \Leftrightarrow \{\overline{Im(\mathbf{A} - \lambda \mathbf{I})} = \mathcal{E}, \exists x \notin Im(\mathbf{A} - \lambda \mathbf{I})\};$$

$$\lambda \in \sigma_r(\mathbf{A}) \Leftrightarrow \{\overline{Im(\mathbf{A} - \lambda \mathbf{I})} \neq \mathcal{E}\}.$$

Here and further an overline over a metric set denotes a closure of it. Stability analysis of (1) becomes simpler if there exists such a number  $\delta < 1$  that the set  $\{z \in \mathbf{C} : |z| > \delta\}$  contains only eigenvalues of operator  $\mathbf{A}$  (for example,  $\dim \mathcal{E} < \infty$ ,  $\mathbf{A}$  is compact operator). But sometimes, as it has been shown by our research, one can successfully use bound points of  $\sigma(\mathbf{A})$ , eliminating a part of spectrum  $\sigma_{ess.a}(\mathbf{A}) \subset \sigma(\mathbf{A})$  called *essentially approximative spectrum*.

**Definition 3.7** ([6]) *Complex number  $\lambda$  is an essentially approximative spectrum point iff there exists such an essentially divergent sequence  $\{x_n, n \in \mathbf{N}\} \subset \mathcal{E}$  that  $\lim_{n \rightarrow \infty} \|(\mathbf{A} - \lambda \mathbf{I})x_n\| = 0$ .*

In [6] has been proved following results.

**Lemma 3.8** *For any  $\mathbf{A} \in L(\mathcal{E})$*

$$\sigma(\mathbf{A}) \cap \{z \in \mathbf{C} : |z| = r\} \neq \emptyset \Leftrightarrow \sigma(\mathbf{A}) \cap \{z \in \mathbf{C} : |z| = r\} \neq \emptyset \quad (33)$$

**Theorem 3.9** *Let us assume that:*

$$(i) \quad \sigma_{ess.a}(\mathbf{A}) \cap \{z \in \mathbf{C} : |z| > 1\} \neq \emptyset;$$

$$(ii) \quad \text{there exist such a continuous function } \varphi : \mathbb{R}_+ \rightarrow \mathbb{R} \text{ and operator sequence } \mathbf{K}_n \in \mathcal{K}(\mathcal{E}), n \geq 0 \\ \text{that } \varphi(0) = 0 \text{ and } \|\mathbf{F}_n x\| \leq \varphi(\|\mathbf{K}_n x\|) \text{ for all } (n, x) \in \mathbf{N} \times \mathcal{E}.$$

Then the trivial solution of (1) is instable.

Applying the above results, we can reasonably simply generalize Theorem 3.1.

**Theorem 3.10** *Let us assume that:*

- (i)  $r(\mathbf{A}) > 1$ ;
- (ii) *there exists a sequence of compact operators  $\{\mathbf{K}_n, n \in \mathbb{N}\} \subset \mathcal{K}(\mathcal{E})$ , and  $q_0 := \sup_{n \geq 0} \|\mathbf{K}_n\| < \infty$ ;*
- (iii)  $\|\mathbf{F}_n x\| \leq \|\mathbf{K}_n x\|$  for all  $(n, x) \in \mathbb{N} \times \mathcal{E}$ .

Then for sufficiently small  $q_0$  the trivial solution of (1) is instable.

**Proof.** If  $\sigma(\mathbf{A}) \cap \{z \in \mathbb{C} : |z| = r\} = \emptyset$  for some  $r \in [1, r(\mathbf{A}))$  the proof of theorem follows from Theorem 3.1. If  $\sigma(\mathbf{A}) \cap \{z \in \mathbb{C} : |z| > 1\} \neq \emptyset$  then by (33)  $\sigma_{ess.a}(\mathbf{A}) \cap \{z \in \mathbb{C} : |z| > 1\} \neq \emptyset$  and one can apply Theorem 3.4. The proof is completed.

## References

- [1] V.Yu. SLYUSARCHUK and Ye.F. TSARKOV (J.Carkovs), *Difference equations in Banach space*, Latv. Math. Yearbook, 17 (1976), pp. 214–229. (Rus.)
- [2] V.Yu. SLYUSARCHUK, *On instability by the first approximation*, Math. Notes, 23(1978). Nr. 5, pp.721–723.
- [3] V.Yu. SLYUSARCHUK, *On stability theory by the first approximation*, Dopovidi AN Ukraine, Ser. A, 9(1981), pp. 27–30.
- [4] V.Yu. SLYUSARCHUK, *New theorems on instability of difference systems by the first approximation*, Differential Equations, 19(1983), Nr. 5, pp. 906–908.
- [5] V.Yu. SLYUSARCHUK, *On instability by the first approximation*, Differential Equations, 22(1986), Nr. 4, pp. 722–723.
- [6] V.Yu. SLYUSARCHUK, *Essentially instable solutions of difference equations*, Ukr. Math. Journ. 51(1999), Nr. 12, pp. 1659–1672.
- [7] V.Yu. SLYUSARCHUK, *Instability of Solutions of Evolution Equations*, National University of Water Management and Nature Resources Use, Rivne, Ukraine, 2004. – 416 p. (Ukr.)
- [8] V.Yu. SLYUSARCHUK, *Equations with Essentially Instable Solutions*, National University of Water Management and Nature Resources Use, Rivne, Ukraine, 2005. – 217 p. (Ukr.)
- [9] V.Yu. SLYUSARCHUK, *New theorem on instability of difference equations in linear approximation*, Scientific bulletin of Chelm, Section of mathematics and computer science, No. 1, 2007, pp. 145–147.
- [10] P.Halmos, *A Hilbert Space Problem Book* Springer Verlag, NY, 1982 – 387 p.
- [11] Dan HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, Berlin-Heldelberg-New York, 1981.
- [12] Tosio KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin-Heldelberg-New York, 1995.

**Current address**

**Jevgeņijs Čarkovs, professor**

Probability and Statistics Chair, Riga Technical University,  
Kaļķu iela 1, Riga, LV-1658, Latvia, tel. +371 26549111  
e-mail: carkovs@latnet.lv

**Vasyl Slyusarchuk, professor**

National University of Water Management and Nature Resources Use,  
Soborna Str., 11, 33000, Rivne.

## GEOMETRICALLY NONLINEAR PLATES SUBJECTED TO A MOVING MASS

ENSHAEIAN Alireza, (IR), ROFOOEI Fayaz R., (IR)

**Abstract.** The dynamic displacement of a geometrically nonlinear rectangular plate under a moving concentrated mass is evaluated utilizing both perturbation techniques and numerical methods. The governing differential equation of motion for an un-damped large deformable rectangular plate is derived using Lagrange method. While the main differential equation is obtained for a moving mass travelling on an arbitrary trajectory, the multiple scales method is used to find the solution for a moving mass passing over the plate on a straight line parallel to any of the plate's edges. The inertial effect of the moving mass is considered by inclusion of all out-of-plane translational acceleration components. A numerical example is used to evaluate the dynamic response of the nonlinear plate obtained using perturbation method. The numerical results obtained show good agreement with the closed form solution for the case of relatively slow moving mass velocity, for the moving mass weight being less than 20% of the plate's weight.

**Key Words:** Moving Mass, Geometric Nonlinearity, Multiple Scales Method, Dynamic Amplification Factor, Lagrange Method

*Mathematics Subject Classification:* Dynamic equations on time scales

### 1 Introduction

The dynamic behavior of the structures subjected to moving loads has been addressed by many researchers over time. The problem is of central importance in the structural design of bridges as an example, where the nature of loading influences the optimum design substantially. There exist numerous investigations in this regard. The earlier studies were generally based on an integral transformation approach, and the inertial effects of the moving mass were limited only to considering a moving load (vertical component of the mass inertia) ([1, 2]). On the other hand, the inertial effects of the moving mass cannot be ignored especially when the weight of the moving mass is comparable to the weight of the supporting structure. Recent investigations have proved that neglecting the convective acceleration components may lead to significant errors in determining the dynamic response of the system[5].

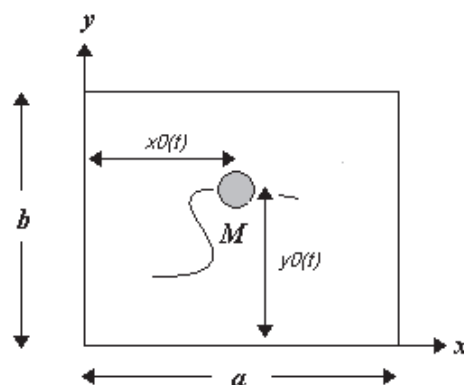
The moving mass problem has been mostly focused on beam models, while the effect of traveling masses on plates has received less attention. Various researches performed on the influence of a moving mass traversing a Kirchhoff plate, have recognized the importance of load inertia [4, 6]. A comprehensive investigation on the dynamic response of plates subjected to moving masses has been performed by Fryba [3]. Surprisingly, not much attention has been paid to study the effect of geometric nonlinearities on the dynamic response of plates under a moving mass with the inclusion of all vertical and convective acceleration components.

In the present work, the geometrical nonlinearities of a simply supported plate under a moving mass are included in the form of dynamic analog of von Karman equations. In this regard, the membrane and bending energies of a Hookean linear plate is evaluated using a vibrational mode shape of a simply supported rectangular plate. Also, the potential energy associated with the moving mass is obtained by considering all inertial components. Having calculated the potential energy terms of the coupled mass-plate system and also the kinetic energy of the excited plate, the governing differential equation of motion is derived through application of the well-known Lagrange method. Observing the significance of the moving mass inertial effects, all out-of-plane translational acceleration components are considered in the formulation of the problem.

The resulting governing ordinary differential equation describes the vibration of Duffing's oscillator with cubic nonlinearity and time varying coefficients. Since derivation of this equation incorporates all inertial component, apart from time-varying mass coefficient, damping term and time-varying stiffness coefficient are also present. The solution to the complete form of the derived differential equation is obtained employing multiple scales method. Besides, the resulting non-linear ODE is numerically solved using the MATLAB program, to investigate the accuracy of the developed closed-form solution to the problem. Since the perturbation solution provides an insight into the dynamic behavior of the system, it is of more scientific value in comparison to the numerical solution. It is shown that the geometric nonlinearity is well captured by the provided solution, especially for low-velocity, low weight moving masses.

## 2 Problem formulation

As it was mentioned before, the dynamic behavior of a plate is considered using the von Karman plate theory. The discrete governing equations are derived by application of Hamilton's principle. A uniform un-damped rectangular plate of length  $a$  and width  $b$ , shown in Fig. 1, with arbitrary boundary condition is considered.



**Fig.1.** Moving mass traversing the plate on a arbitrary trajectory

The mass density of the plate is assumed to be  $\rho$ , with its bending stiffness defined as  $D = \frac{Eh^3}{12(1-\nu^2)}$ , in which  $E$ ,  $h$  and  $\nu$  are plate's modulus of elasticity, thickness and Poisson's ratio, respectively. Also,  $u(x, y, t)$ ,  $v(x, y, t)$  and  $w(x, y, t)$  denote the deflection of the mid-plane of the plate at any point and at any time  $t$ , in directions parallel to  $x$ ,  $y$  and  $z$  axes. The kinetic energy of the plate is equal to:

$$K_{plate} = \frac{1}{2} \int_A \rho h \dot{w}^2 dA \quad (1)$$

On the other hand, assuming Green-Lagrange strains, the strain energy of plate becomes equal to [8]:

$$U_{plate} = \frac{D}{2} \int \left\{ (\nabla^2 w)^2 + \frac{12}{h^2} e_1^2 - 2(1-\nu) \left[ \frac{12}{h^2} e_2 + \frac{\partial^2 w}{\partial x^2} \frac{\partial^2 w}{\partial y^2} - \left( \frac{\partial^2 w}{\partial x \partial y} \right)^2 \right] \right\} dA \quad (2)$$

where,

$$e_1 = \partial u / \partial x + \frac{1}{2} \left( \frac{\partial w}{\partial x} \right)^2 + \partial v / \partial y + \frac{1}{2} \left( \frac{\partial w}{\partial y} \right)^2 \quad (3)$$

and,

$$e_2 = \left( \partial u / \partial x + \frac{1}{2} \left( \frac{\partial w}{\partial x} \right)^2 \right) \left( \partial v / \partial y + \frac{1}{2} \left( \frac{\partial w}{\partial y} \right)^2 \right) - \frac{1}{4} \left( \partial u / \partial y + \partial v / \partial x + \frac{\partial w}{\partial x} \frac{\partial w}{\partial y} \right)^2 \quad (4)$$

Using the Dirac-delta function, the external excitation force due to a moving mass,  $M$ , traveling on an arbitrary trajectory on the plate surface can be described as,

$$f(x, y, t) = M \left( g - \frac{d^2 w_0(t)}{dt^2} \right) \delta(x - x_0(t)) \delta(y - y_0(t)) \quad (5)$$

where  $g$  is the acceleration of gravity. The vertical displacement of the moving mass is shown by  $w_0(t)$ , while  $x_0(t)$  and  $y_0(t)$  describe its trajectory on the plate. Considering all the out-of-plane translational acceleration components, and observing the full contact condition between the moving mass and the plate, Eq. (5) can be expanded as:

$$f(x, y, t) = M \left( g - \frac{d^2 w(t)}{dt^2} \right)_{x=x_0(t), y=y_0(t)} \delta(x - x_0(t)) \delta(y - y_0(t)) = M \left\{ g - \left[ \frac{\partial^2 w}{\partial t^2} + \dot{x}_0^2(t) \frac{\partial^2 w}{\partial x^2} + \dot{y}_0^2(t) \frac{\partial^2 w}{\partial y^2} + 2\dot{x}_0(t)\dot{y}_0(t) \frac{\partial^2 w}{\partial x \partial y} + \dot{x}_0(t) \frac{\partial^2 w}{\partial x \partial t} + \dot{y}_0(t) \frac{\partial^2 w}{\partial y \partial t} + \ddot{x}_0(t) \frac{\partial w}{\partial x} + \ddot{y}_0(t) \frac{\partial w}{\partial y} \right]_{x=x_0(t), y=y_0(t)} \right\} \delta(x - x_0(t)) \delta(y - y_0(t)) \quad (6)$$

Therefore the virtual work done by external force becomes:

$$W = \int_A f(x, y, t) w dA \quad (7)$$

where  $f(x, y, t)$  is evaluated using Eq. (6). The unknown parameters of the plate are  $u(x, y, t)$ ,  $v(x, y, t)$  and  $w(x, y, t)$  that can be discretized using appropriate spatial functions as the following:

$$u(x, y, t) = r(t)\eta(x, y) \quad , \quad v(x, y, t) = s(t)\psi(x, y) \quad , \quad w(x, y, t) = q(t)\phi(x, y) \quad (8)$$

Where the selected spatial functions  $\eta(x, y)$ ,  $\psi(x, y)$  and  $\phi(x, y)$  should satisfy the required boundary conditions. The function associated with the vertical displacement  $\phi(x, y)$  is selected as the linear mode shape of the plate in the vertical direction. Using Eq. (8) and applying Lagrange method leads to,

$$\frac{\partial(K_{plate}-U_{plate})}{\partial r} - \frac{d}{dt} \left( \frac{\partial K_{plate}}{\partial r} \right) = - \frac{\partial W}{\partial r} \quad (9)$$

$$\frac{\partial(K_{plate}-U_{plate})}{\partial s} - \frac{d}{dt} \left( \frac{\partial K_{plate}}{\partial s} \right) = - \frac{\partial W}{\partial s} \quad (10)$$

$$\frac{\partial(K_{plate}-U_{plate})}{\partial q} - \frac{d}{dt} \left( \frac{\partial K_{plate}}{\partial q} \right) = - \frac{\partial W}{\partial q} \quad (11)$$

Performing the mathematical manipulations, equations (9) and (10) can be used to calculate the parameters  $r(t)$  and  $s(t)$  as functions of  $q(t)$ . Therefore, both potential and kinetic energy of the system in Eq. (11) are expressed as functions of  $q(t)$  only. Thus, Eq. (11) reduces to an ordinary differential equation with cubic nonlinearity as the following:

$$\begin{aligned} & \left[ \int_0^A \rho h \phi^2 dA + M \phi(x_0(t), y_0(t)) \phi(x_0(t), y_0(t)) \right] \ddot{q}(t) + \\ & M \phi(x_0(t), y_0(t)) [\dot{x}_0(t) \phi_{,x}(x_0(t), y_0(t)) + \dot{y}_0(t) \phi_{,y}(x_0(t), y_0(t))] \dot{q}(t) + \left\{ \omega_0^2 \left( \int_0^A \rho h \phi^2 dA \right) + \right. \\ & M \phi(x_0(t), y_0(t)) [\dot{x}_0^2(t) \phi_{,xx}(x_0(t), y_0(t)) + \dot{y}_0^2(t) \phi_{,yy}(x_0(t), y_0(t)) + \\ & \ddot{x}_0(t) \phi_{,x}(x_0(t), y_0(t)) + \ddot{y}_0(t) \phi_{,y}(x_0(t), y_0(t)) + 2 \dot{x}_0(t) \dot{y}_0(t) \phi_{,xy}(x_0(t), y_0(t))] \} q(t) + \\ & \frac{D}{2} \Gamma q(t)^3 = M g \phi(x_0(t), y_0(t)) \end{aligned} \quad (12)$$

Selecting  $\phi(x, y)$  as the natural mode shape of the linear system,  $\omega_0$  denotes the related natural frequency of the plate. Assuming a simply supported plate, the general mode shape is,  $\phi(x, y) = \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right)$ , where  $m$  and  $n$  are positive integers. The associative natural frequency is as the following:

$$\omega_0^2 = D \frac{\pi^4}{\rho h} \left( \frac{m^2}{a^2} + \frac{n^2}{b^2} \right)^2 \quad (13)$$

Parameter  $\Gamma$  in Eq. (12) is a constant that depends on the geometric shape of the plate, Poisson's ratio and the vertical spatial function parameters respectively. This parameter originates from the nonlinear strain terms in equations (3) and (4) leading to a nonlinear equation of motion. Eq. (12) includes all the vertical and convective acceleration components.

## 2.1 Equation Solution

Consider a simply supported plate with the general mode shape,  $\phi(x, y) = \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right)$ . A moving mass is traversing the plate with constant velocity  $c$  on a path parallel to  $x$  axis that is coincided on one of the plate's edge. In that case, Eq. (12) becomes:

$$\begin{aligned} & \left[ \frac{1}{4} \rho h a b + M \sin^2(m\omega t) \sin^2\left(\frac{n\pi y_0}{b}\right) \right] \ddot{q}(t) + \\ & M \sin(m\omega t) \sin^2\left(\frac{n\pi y_0}{b}\right) \cos(m\omega t) \dot{x}_0(t) \left(\frac{m\pi}{a}\right) \dot{q}(t) + \\ & \left\{ D \frac{\pi^4 a b}{4} \left( \frac{m^2}{a^2} + \frac{n^2}{b^2} \right)^2 - M \sin^2(m\omega t) \sin^2\left(\frac{n\pi y_0}{b}\right) \dot{x}_0^2(t) \left(\frac{m\pi}{a}\right)^2 \right\} q(t) + \frac{D}{2} \Gamma q(t)^3 = \\ & M g \sin(m\omega t) \sin\left(\frac{n\pi y_0}{b}\right) \end{aligned} \quad (14)$$

where,

$$\frac{m\pi x_0(t)}{a} = \frac{m\pi c t}{a} = m\omega t \quad , \quad \omega = \frac{\pi c}{a} \quad (15)$$

Introducing the perturbation parameter as,

$$\epsilon = \frac{M}{\rho h a b} \quad (16)$$

and substituting in Eq. (14) leads to:

$$\left[1 + 4\epsilon \sin^2(m\omega t) \sin^2\left(\frac{n\pi y_0}{b}\right)\right] \ddot{q}(t) + 4\epsilon \sin(m\omega t) \sin^2\left(\frac{n\pi y_0}{b}\right) \cos(m\omega t) c\left(\frac{m\pi}{a}\right) \dot{q}(t) + \left\{\omega_0^2 - 4\epsilon \sin^2(m\omega t) \sin^2\left(\frac{n\pi y_0}{b}\right) c^2\left(\frac{m\pi}{a}\right)\right\} q(t) + \frac{2D\Gamma}{M} \epsilon q(t)^3 = 4\epsilon g \sin(m\omega t) \sin\left(\frac{n\pi y_0}{b}\right) \quad (17)$$

Introducing the following non- dimensional parameters:

$$\tau = t\omega, X_0(\tau) = \frac{x_0(t)}{a}, Y_0 = \frac{y_0}{b}, Q(\tau) = \frac{q(t)}{a}, \bar{\omega}_0 = \frac{\omega_0}{\omega}, \Omega = \frac{2D\Gamma a^2}{M\omega^2}, G = g \frac{a}{\pi^2 c^2} \quad (18)$$

Equation (17) can be re-written in a non- dimensional form as the following:

$$[1 + 4\epsilon \sin^2(m\tau) \sin^2(n\pi Y_0)] \ddot{Q}(\tau) + 4m\epsilon \sin(m\tau) \sin^2(n\pi Y_0) \cos(m\tau) \dot{Q}(\tau) + \{\bar{\omega}_0^2 - 4m^2\epsilon \sin^2(m\tau) \sin^2(n\pi Y_0)\} Q(\tau) + \Omega \epsilon Q(\tau)^3 = 4\epsilon G \sin(m\tau) \sin(n\pi Y_0) \quad (19)$$

Equation (19) now is in the right form for application of perturbation techniques such as multiple scales method. Defining:

$$T_0 = \tau, \quad T_1 = \epsilon \tau \quad (20)$$

and:

$$D_0 = \frac{\partial}{\partial T_0}, \quad D_1 = \frac{\partial}{\partial T_1} \quad (21)$$

The unknown function  $Q(t)$  can be assumed as:

$$Q(\tau) = Q_0(T_0, T_1) + \epsilon Q_1(T_0, T_1) \quad (22)$$

Substituting equations (20) to (22) in Eq. (19) and separating different orders of  $\epsilon$ , the following two linear ordinary differential equations are obtained:

$$\text{for zero order of } \epsilon: \quad D_0^2 Q_0 + \bar{\omega}_0^2 Q_0 = 0 \quad (23)$$

and :

$$\text{for 1st order of } \epsilon:$$

$$D_0^2 Q_1 + \bar{\omega}_0^2 Q_1 = -2D_0 D_1 Q_0 - 4 \sin^2(m\tau) \sin^2(n\pi Y_0) D_0^2 Q_0 - 4m \sin(m\tau) \sin^2(n\pi Y_0) \cos(m\tau) D_0 Q_0 + 4m^2 \sin^2(m\tau) \sin^2(n\pi Y_0) Q_0 - \Omega Q_0^3 + 4G \sin(m\tau) \sin(n\pi Y_0) \quad (24)$$

The solution to Eq. (23) is as follows:

$$Q_0 = A(T_1) e^{i\bar{\omega}_0 T_0} + cc \quad (25)$$

Where the  $cc$  denote the complex conjugate of the other present terms on the right hand side. To avoid secular terms in solution of  $Q_1$ , the coefficients on the right side of Eq. (24) should be set to zero:

$$2 \sin^2(n\pi Y_0) (\bar{\omega}_0^2 + m^2) A - 3 \Omega A^2 \bar{A} - 2 \frac{\partial A}{\partial T_1} i \bar{\omega}_0 = 0 \quad (26)$$

Thus from equation (24),  $Q_1$  is calculated as:

$$Q_1 = A \sin^2(n\pi Y_0) \left[ \frac{(\bar{\omega}_0^2 + m^2 + m \bar{\omega}_0)}{(\bar{\omega}_0 + 2m)^2 - \bar{\omega}_0^2} e^{i\bar{\omega}_0 T_0 + 2imT_0} + \frac{(\bar{\omega}_0^2 + m^2 - m \bar{\omega}_0)}{(\bar{\omega}_0 - 2m)^2 - \bar{\omega}_0^2} e^{i\bar{\omega}_0 T_0 - 2imT_0} \right] + \Omega A^3 \left( \frac{e^{3i\bar{\omega}_0 T_0}}{8\bar{\omega}_0^2} \right) - 2 A i \frac{\sin(n\pi Y_0)G}{\bar{\omega}_0^2 - m^2} e^{imT_0} + c c \quad (27)$$

Assuming  $A = \frac{1}{2} \alpha(T_1) e^{i\beta(T_1)}$  and using Eq. (26) to evaluate  $A(T_1)$ , one gets:

$$\alpha = \alpha_0, \quad \beta = \eta T_1 + \beta_0 = \left( \frac{3\Omega}{8\bar{\omega}_0} \alpha_0^2 - \bar{\omega}_0 \sin^2(n\pi Y_0) \left( 1 + \left( \frac{m}{\bar{\omega}_0} \right)^2 \right) \right) T_1 + \beta_0 \quad (28)$$

Where  $\alpha_0$  and  $\beta_0$  can be obtained using initial conditions. Having calculated  $A(T_1)$  from Eq. (28),  $Q(\tau)$  is found to be:

$$Q(\tau) = Q_0 + \epsilon Q_1 = \alpha_0 \cos(\bar{\omega}_0 \tau + \eta \epsilon \tau + \beta_0) + \epsilon \left[ \sin^2(n\pi Y_0) \frac{(\bar{\omega}_0^2 + m^2 + m \bar{\omega}_0)}{(\bar{\omega}_0 + 2m)^2 - \bar{\omega}_0^2} \cos(\bar{\omega}_0 \tau + \eta \epsilon \tau + 2m\tau + \beta_0) + \sin^2(n\pi Y_0) \frac{(\bar{\omega}_0^2 + m^2 - m \bar{\omega}_0)}{(\bar{\omega}_0 - 2m)^2 - \bar{\omega}_0^2} \cos(\bar{\omega}_0 \tau + \eta \epsilon \tau - 2m\tau + \beta_0) + \frac{\Omega}{32} \alpha_0^3 \cos(3\bar{\omega}_0 \tau + 3\eta \epsilon \tau + 3\beta_0) + \frac{4G}{\bar{\omega}_0^2 - m^2} \sin(n\pi Y_0) \sin(m\tau) \right] \quad (29)$$

In Eq. (29), the constants  $\alpha_0$  and  $\beta_0$  can be calculated using the initial conditions of the problem. Eq. (29) is valid as long as the moving mass has not left the plate. After the mass traverses the plate completely, the plate vibrates in its free vibration phase described by ([8]),

$$Q(\tau) = \epsilon a \cos(\bar{\omega}' \tau + \beta_0) + O(\epsilon^3) \quad (30)$$

where:

$$\bar{\omega}' = \bar{\omega}_0 \left[ 1 + \frac{3\Omega\epsilon}{8\bar{\omega}_0^2} (\epsilon a)^2 \right] + O(\epsilon^3) \quad (31)$$

The parameters  $\epsilon a$  and  $\beta_0$  are calculated using the initial conditions for the free oscillation phase. Equations (29) and (31) present the closed form solution of the moving mass problem described earlier. The ODE solver of the MATLAB program which is based on the Runge-Kutta method, is utilized to numerically verify the accuracy of the presented closed-form solutions.

### 3 Numerical Example

A simply supported square plate shown in Fig. 2, with a modulus of elasticity,  $E = 7.1 \times 10^{10} Pa$ , mass density:  $\rho = 2700 kg/m^3$ , length: 2 m, thickness: 1 cm, and the Poisson's ratio:  $\nu = 0.33$ , is considered. A straight trajectory is assumed for the moving mass, passing through the center line of the plate as shown in Fig.2. The problem is solved for three moving mass velocities as well as three mass weights. Representing the weight of the moving mass as a fraction of the plate's weight, the mass ratios equal to 0.05, 0.1 and 0.2 are considered in this example.

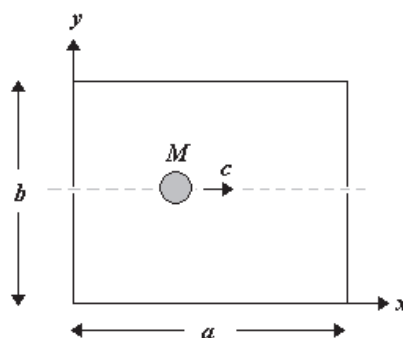


Fig.2. Path of the moving mass

The spatial function  $\phi(x, y)$  is assumed to be the simply supported plate's first modal shape function. The first mode shape and the related natural period of the plate are:

$$\phi(x, y) = \sin\left(\frac{\pi x}{a}\right) \sin\left(\frac{\pi y}{b}\right) \quad (32)$$

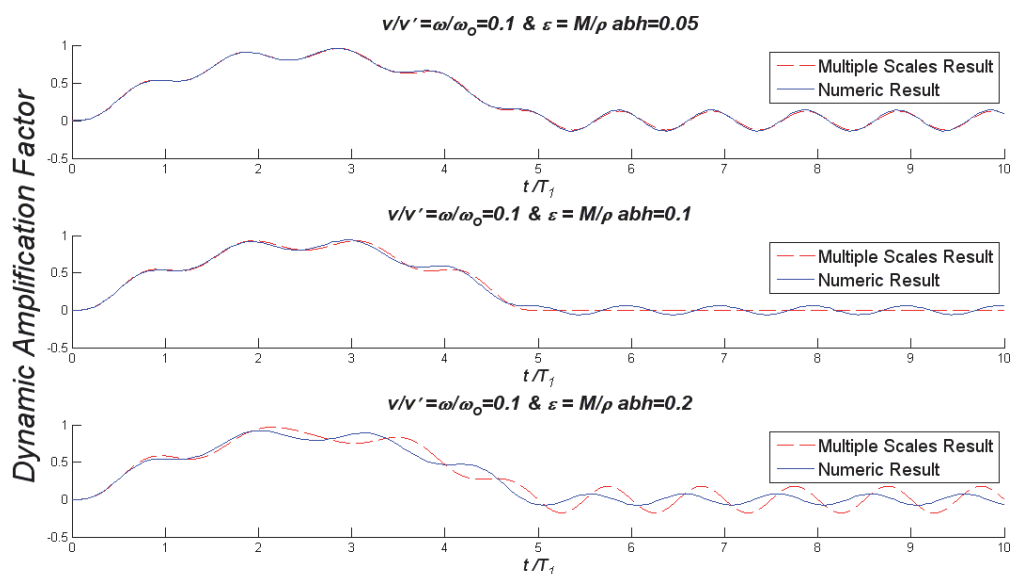
$$T_1 =$$

$$\frac{2}{\left\{ \pi \left[ \left( \frac{1}{a} \right)^2 + \left( \frac{1}{b} \right)^2 \right] \right\}} \sqrt{\frac{\rho h}{D}} \quad (33)$$

As it was mentioned earlier, it is assumed that the plate is demonstrating a geometrically nonlinear behaviour under the applied loading. The moving mass  $M$  is assumed to travel along a linear path over the plate. Under the moving mass excitation, the dynamic response of the plate is made up of a forced vibration part followed by a free vibration, once the moving mass leaves the plate's boundaries. The linear path is defined by the following equation (Fig.2):

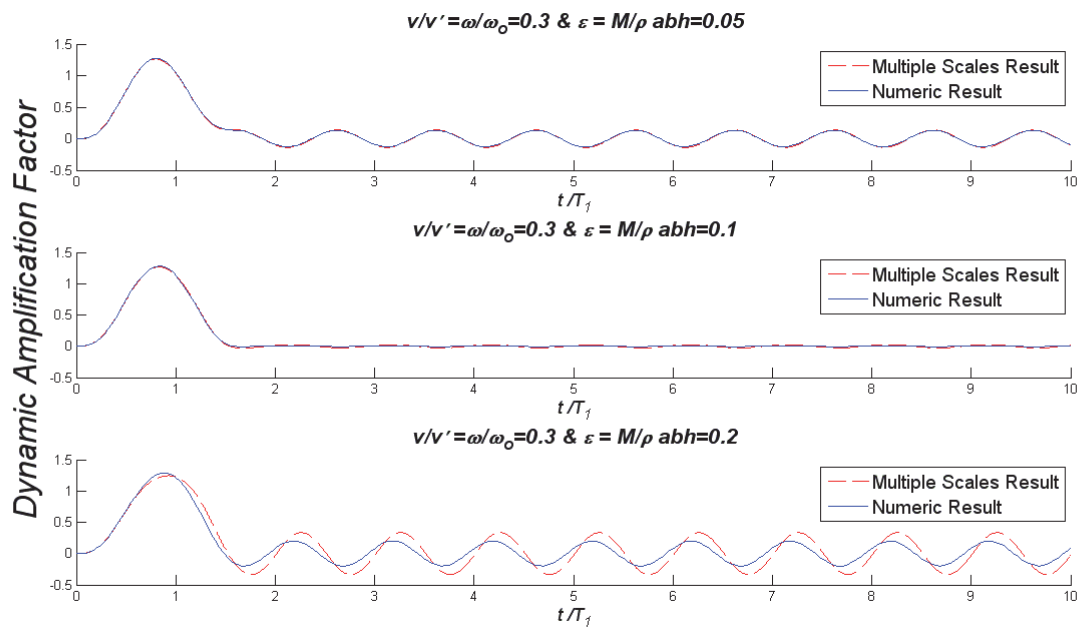
$$y_0(t) = \frac{b}{2}, x_0(t) = ct \quad (34)$$

In which  $c$  is the velocity of the moving mass. Fig.3 shows the dynamic amplification factor (DAF) of the center point of a  $2m \times 2m$  rectangular plate, when the weight of the moving mass is considered to be 0.05, 0.1 and 0.2 of the plate's weight. The velocity of the mass in this figure is  $0.1v'$  where  $v' = \frac{2a}{T_1}$ .

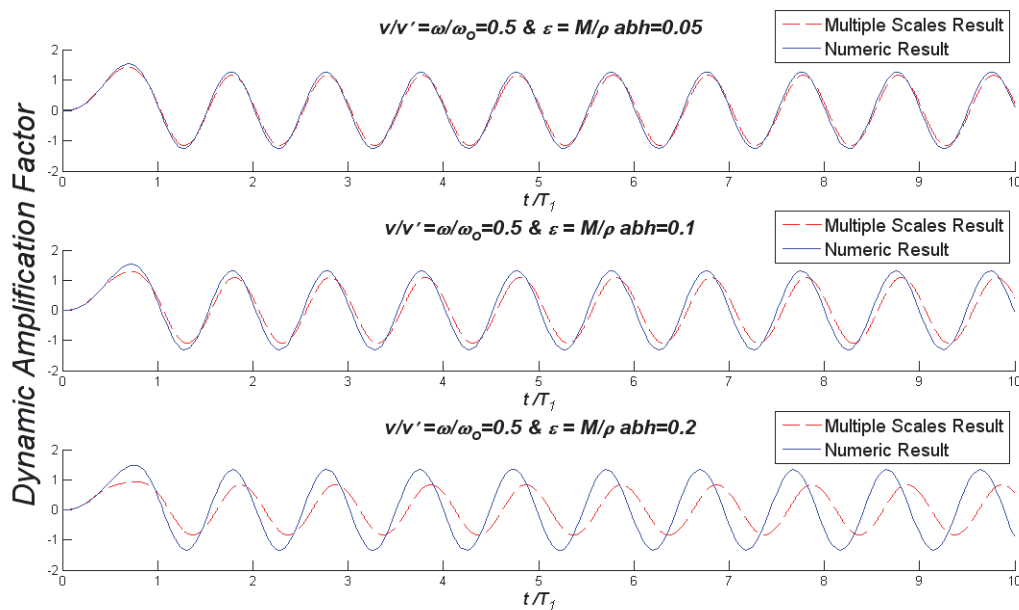


**Fig.3.** Plate's center time history response for  $v = 0.1v'$  and  $\epsilon = 0.05, 0.1$  and  $0.2$

The result for the case with the mass velocity equal to  $0.3v'$  is presented in Fig. 4. Similarly, Fig. 5 shows the output for the mass velocity equal to  $0.5v'$ .



**Fig.4.** Plate's center time history response for  $v = 0.3v'$  and  $\epsilon = 0.05, 0.1$  and  $0.2$



**Fig.5.** Plate's center time history response for  $v = 0.5v'$  and  $\epsilon = 0.05, 0.1$  and  $0.2$

The dynamic amplification factor (DAF) is defined as the ratio of the absolute maximum dynamic deflection of the plate to its maximum static response at the center point. The static deflection of the center point of a square plate under a concentrated mass  $M$ , applied at the same point is equal to  $\Delta_{static} = \frac{0.0116Mga^2}{D}$  [7].

As it can be observed, the accuracy of the solution obtained using perturbation technique is highly dependent on the mass weight and the velocity of the moving mass. Deviation of the analytical results from their numerical counterpart grows as the mass and velocity ratios increase. Since the

mass ratio is considered as the perturbation parameter, this phenomenon makes sense for this case, while the velocity ratio effect originates from the time-varying nature of Eq. (17). As the velocity of the mass increases, the characteristic parameters (natural frequencies) of the governing equation change more rapidly. Therefore the slowly-varying time scales incorporated in the multiple scales method fail to capture this rapid change which leads to considerable errors in their related results.

#### 4 Conclusion

The dynamic response of a geometrically nonlinear rectangular plate under a moving concentrated mass is evaluated utilizing both perturbation techniques and numerical methods. Governing differential equation of motion for von Karman plates subjected to a moving mass was developed based on Lagrange method. This equation was solved for the case of recti-linear mass trajectory. The effects of weight and velocity of the mass on the dynamic response of the system have been investigated. The solutions gained using multiple scales method show good agreement with their pertinent numeric results for cases in which the mass ratios are less than 0.2 and their velocities be a small fraction of the specific value  $v'$ . As the mass velocity and its weight increases, the obtained results start to deviate from the numerical results. For appropriate ranges of mass and velocity ratios, the closed form solution resulted from the application of multiple scale method, captures the real response of the geometrically nonlinear plate quite well.

#### References

- [1] G. G. STOKES 1849 Transactions of Cambridge Philosophical Society 8, 707. Discussion of a differential equation relating to the breaking of railway bridges.
- [2] R. S. AYRE, L. S. JACOBSON and C. S. HSU 1951 Proceedings of the 1st National Congress on Applied Mechanics, 11-16 June, Chicago, 1951. Transverse vibration of one and two-span beams under the action of a moving mass load.
- [3] FRYBA, L. 1999. Vibration of Solids and Structures under Moving Loads. Tomas Telford, London.
- [4] F.R. ROFOOEI , A. NIKKHOO, Application of active piezoelectric patches in controlling the dynamic response of a thin rectangular plate under a moving mass, International Journal of Solids and structures, 46 (2009)2429-2443.
- [5] A. NIKKHOO, F.R. ROFOOEI , M.R. Shadnam ,Dynamic behavior and modal control of beams under moving mass, Journal of Sound and Vibration 306 (3–5)(2007)712–724.
- [6] GBADEYAN, J.A., ONI, S.T., 1995. Dynamic behavior of beams and rectangular plates under moving loads. Journal of Sound and Vibration 182 (5), 677–695.
- [7] TIMOSHENKO, S. WOINOWSKY-KREIGER, 1959. Theory of Plates and Shells, second edition. McGraw-Hill, New York.
- [8] A.H. NAYFEH , D.T. MOOK ,“Nonlinear Oscillations” , 1995

#### Current Addresses

**Alireza Enshaeian, Ph.D Candidate**  
 Civil Engineering Department,  
 Sharif University of Technology  
 P. O. Box: 11155-9313, Azadi Ave.

Tehran, Iran.  
Tel.: (+98)917-317-4013  
Email: enshaiean@mehr.sharif.ir

**Fayaz R. Rofooei, Ph.D., P.E.,**  
Professor,  
Director of the Earthquake Engineering Research Center,  
Civil Engineering Department,  
Sharif University of Technology  
P. O. Box: 11155-9313, Azadi Ave.  
Tehran, Iran.  
Tel. & Fax: (+98)21-6616-4233  
Email: rofooei@sharif.edu  
Internet: [www.sina.sharif.edu/~rofooei](http://www.sina.sharif.edu/~rofooei)

## THE STUDY OF A NONLINEAR SYSTEM IN THE CASE WITH TWO OSCILLATING MASS LAST LINE

FLOREA Olivia, (RO)

**Abstract.** The dynamical system referred in this paper is part of the category of the dynamical systems with geometrical or mechanical variables parameters. Such oscillating parametrical systems are encountered in the case of the pendulum with variable length wire, of the variation of the length or the width of the shaft rotation, of the modification of the rigidity and amortization coefficients. These systems have an important utility in practice in the case of elevators or cranes, in the case of the transporters based on the vibrations. These systems are controlled from the stability, bifurcations and resonances point of view, such type of control leading to the avoidance of the catastrophes.

We consider a fixed system of axis and a mass  $M$  which performs some oscillations on an inclined plane; on this mass is suspended a pendulum of mass  $m$  and length  $l(t)$  and the wire is passing over a pulley.

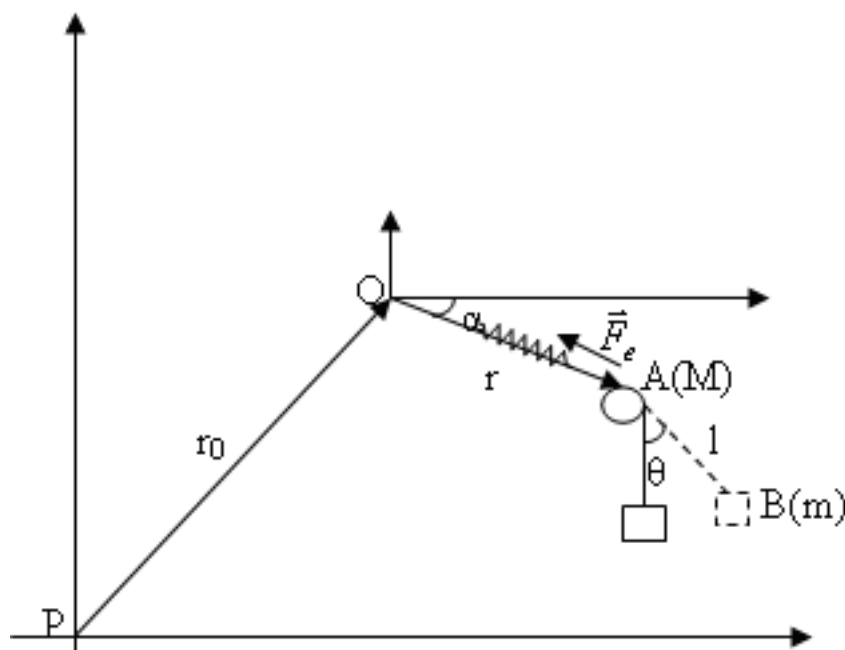
The mathematical model leads to a system with two freedom degrees; the two masses are connected non stationary, hence obtaining some non linear equations with variables coefficients. In our model we have: the displacement of  $M$  on the inclined plane is  $r$ , the oscillating angle of the pendulum is  $\theta$ , the angle of the inclined plane made with the horizontal axis is  $\alpha$  and the wire length of the pendulum which varies linearly or harmonically is  $l(t)$ . The solutions for the Cauchy problem are obtained through the method of the small parameters or the method of successive approximations, but more important is the study of the stability, of the bifurcations and the resonance. Some of equations are of the Hill or Mathieu type. The Ince - Strutt stability chart or the analytical and numerical simulations with averages in the phase's plane are used.

**Key words and phrases.** nonlinear dynamical systems, stability, Ince Strutt diagrams.

*Mathematics Subject Classification.* Primary 34B30, 70H03, 70K20; Secondary 37N05.

# 1 The study of a nonlinear dynamic system of two nonstationary bound bodies with permanent oscillations

Given a dynamic system which consists in a  $M$  mass oscillator operated by an elastic resort  $POA$  on an incline plane (  $POA = r_0 + r$  ,  $P$  and  $O$  fixed, where  $PO = r_0$  is the static position) and a mass  $m$  pendulum suspended in  $A$  trough the  $AB=l$  ( $A(M)$ ,  $B(m)$ ) wire. The wire is operated by affixed pulley  $N$  and passes trough  $NAB$ ; the pulley mobility can give various variations of the wire length  $AB=l(t)$  (linear  $l(t) = l_0 \pm vt$  or harmonic  $l(t) = l(t) \cos \chi t$  ). Given the  $xOy$  axes in the vertical plane (with  $Ox$  vertical and  $Oy$  horizontal) where the forces act; the elastic  $\vec{F}_e = -k(\vec{r} + \vec{r}_0)$  force and the gravity force  $\vec{G} = M\vec{g} = Mg\vec{i}$  act over  $M$  ; and the gravity force  $m\vec{g}\vec{i}$  acts on  $m$ ; with  $k$  the elastic constant. The coordinates of  $A$ ,  $B$  are  $A(x_1, y_1)$ ,  $B(x_2, y_2)$



$$x_1 = r \sin \alpha, y_1 = r \cos \alpha, x_2 = x_1 + l \cos \theta, y_2 = y_1 + l \sin \theta \quad (1)$$

Where  $\alpha$  is the angle of the inclined plane  $POA$  with the horizontal and  $\theta$  is the angle between the  $AB$  wire and the vertical.

The kinetic energy of the system is:  $T = \frac{M}{2} (\dot{x}_1^2 + \dot{y}_1^2) + \frac{m}{2} (\dot{x}_2^2 + \dot{y}_2^2)$  and the potential energy a  $U$  for the  $q_1 = r, q_2 = \theta$  degrees of freedom will be

$$T = \frac{(M + m)\dot{r}^2}{2} + \frac{m}{2} \left( l^2 \dot{\theta}^2 + 2rl\dot{\theta} \cos(\theta + \alpha) \right) + \frac{m}{2} \left( \dot{l}^2 + 2r\dot{l} \sin(\theta + \alpha) \right) \quad (2)$$

$$U = \frac{kr^2}{2} + mgl(1 - \cos \theta)$$

If  $l=ct$  the system becomes stationay  $S$  and if  $l=l(t)$  the system becomes nonstationary  $N$ ; for  $\alpha = 0, \frac{\pi}{2}$  the oscillator is horizontal or vertical.

The Lagrange equations  $\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{q}_k} \right) - \frac{\partial L}{\partial q_k} = 0, k = 1, 2$  with  $L=T-U$  become in the described two situations:

$$\begin{cases} \frac{d}{dt} \left[ (M+m)\dot{r} + ml\dot{\theta} \cos(\theta + \alpha) \right] + kr = 0, & l = ct \\ \frac{d}{dt} \left[ ml^2\dot{\theta} + ml\dot{r} \cos(\theta + \alpha) \right] + mgl \sin \theta = 0 \end{cases} \quad (S) \quad (3)$$

$$\begin{cases} \frac{d}{dt} \left[ (M+m)\dot{r} + ml\dot{\theta} \cos(\theta + \alpha) + m\dot{l} \sin(\theta + \alpha) \right] + kr = 0, & l = l(t) \\ \frac{d}{dt} \left[ ml^2\dot{\theta} + ml\dot{r} \cos(\theta + \alpha) \right] + ml\dot{r} \sin(\theta + \alpha) - m\dot{l} \cos(\theta + \alpha) + mgl \sin \theta = 0 \end{cases} \quad (N) \quad (4)$$

In the paper we'll study the two situations (3), (4) only in the case of the vertical oscillator  $M$  with  $\alpha = \frac{\pi}{2}$ .

## 2 The vertical stationary case (SV) $r = x, \alpha = \frac{\pi}{2}$ horizontal armed crane, the A(x) extremity oscillates vertically

The system (3) becomes:

$$\begin{cases} (M+m)\ddot{x} - ml\ddot{\theta} \sin \theta - ml\dot{\theta}^2 \cos \theta = -kx \\ ml^2\ddot{\theta} - m\dot{x}l \sin \theta = -mgl \sin \theta \end{cases} \quad (5)$$

We'll make the notations:

$$\begin{aligned} \mu &= \frac{m}{m+M}, \omega^2 = \frac{k}{m+M}, x_0 = \frac{(M+m)g}{k}, \varepsilon = \frac{x_0}{l}, \tau = \omega t \\ X &= \frac{x}{l}, \delta = \frac{g}{l\omega^2} \end{aligned} \quad (6)$$

Where:  $\frac{dx}{dt} = \frac{dx}{d\tau} \omega = x' \omega, \frac{d\theta}{dt} = \theta' \omega$  we obtain:

$$\begin{cases} X'' + X = \mu [\theta'^2 \cos \theta + \theta'' \sin \theta] \\ \theta'' - X'' \sin \theta = -\delta \sin \theta \end{cases} \quad (7)$$

We normalize the system:

$$\begin{cases} X'' (1 - \mu \sin^2 \theta) + X = \mu \theta'^2 \cos \theta - \delta \mu \sin^2 \theta \\ \theta'' (1 - \mu \sin^2 \theta) = -\delta \sin \theta + \mu \theta'^2 \sin \theta \cos \theta - X \sin \theta \end{cases} \quad (8)$$

The stability study is made around the equilibrium solutions  $\theta = 0, X = \varepsilon \cos(\tau - \tau_0)$  where, for  $t=0, \tau = \tau_0, X_0 = \frac{x_0}{l} = \varepsilon, \dot{x}_0 = 0$ . The system (8) linearized around equilibrium points has the characteristic polynomial  $P_4 = (r^2 + 1)(r^2 + \delta)$  with pure imaginary roots. In this case the equilibrium point for the linearized system is simple stable (center) but we can't appreciate the nonlinear system stability (8). For this situation we'll appeal the direct study of the system (8), that will lead us to an equation of Mathieu type.

So, with the substitution  $\phi = \theta$  and  $Z = X - \varepsilon \cos(\tau - \tau_0)$  and the development  $\sin \phi = \phi - \frac{\phi^3}{3} + \dots$ , in the first approximation  $\sin \phi \approx \phi$  we obtain from (8) the fundamental equations:

$$\frac{d^2 \phi}{d\tau^2} + (\delta + \varepsilon \cos \tau) \phi = 0 \text{ Mathieu equation} \quad (9)$$

The problem of stability for the solution  $\theta \equiv 0, x = \frac{x_0}{l} \cos \omega t$  leads to the study of the equation (9) of Mathieu type, for which stability studies are made through the Ince-Strutt  $\delta = \delta(\varepsilon)$  diagonals.

In the equation (9) the perturbation function is  $\cos \tau$  with a period of  $2\pi$ ; it can be seen that a solution is  $\phi = \cos \tau$  and  $\phi(-\tau) = -\phi(\tau), \phi(-\tau), -\phi(\tau)$  must be solutions; in this case we'll seek for both even and odd solutions:

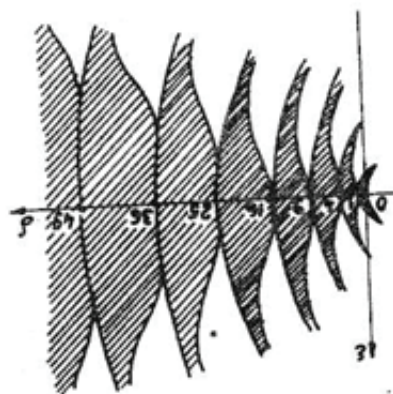
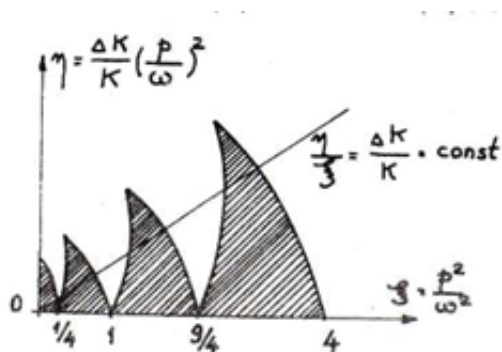
$$\phi = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos k\pi, \quad \phi = \sum_{k=1}^{\infty} b_k \sin k\pi \quad (10)$$

With the period  $T = 4\pi$ ;  $\phi = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos \frac{k\pi}{2}, \quad \phi = \sum_{k=1}^{\infty} b_k \sin \frac{k\pi}{2}$ .

By introducing these solutions into the equation and transforming the products in sums we have trigonometric identifications that lead to the system homogeneous in  $a_k$  which imply that the system determinant is null.

These determinants in the  $(\delta O\varepsilon)$  plane lead to the  $\varepsilon = \varepsilon(\delta)$  graphics. Considering in the Fourier solutions the  $n = 1, 2, 3, \dots$  rank terms, we have:

$$\Delta = \begin{vmatrix} \delta & \varepsilon & 0 & 0 & \dots & 0 \\ \frac{\varepsilon}{2} & \delta - 1 & \frac{\varepsilon}{2} & 0 & \dots & 0 \\ 0 & \frac{\varepsilon}{2} & \delta - 4 & \frac{\varepsilon}{2} & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{\varepsilon}{2} & \delta - n^2 & \dots \end{vmatrix}, \quad \begin{aligned} \Delta_2 &= \delta(\delta - 1) - \frac{\varepsilon^2}{2} = 0 \\ \Delta_3 &= 0, \dots, \Delta_n = 0 \\ \Delta_n &= [\delta - (n-1)^2] \Delta_{n-1} - \frac{\varepsilon^2}{4} \Delta_{n-2} \Rightarrow \\ \delta &= \delta(\varepsilon) \rightarrow \varepsilon = \varepsilon(\delta) \\ \delta &= 1 - \frac{1}{12}\varepsilon^2, \delta = \frac{1}{4} - \frac{\varepsilon}{2}, \delta = \frac{1}{4} + \frac{\varepsilon}{2} \end{aligned} \quad (11)$$



For  $\varepsilon \geq 0$  above the graphics  $\varepsilon = \varepsilon(\delta)$  we have instability and between  $\varepsilon$  and the graphics we have stability. The instability is called parametrical resonance, when  $\varepsilon$  grows above the diagonals. In the inferior half-plane  $\varepsilon < 0$  we have symmetry. So, by taking  $\delta = \frac{g}{l\omega^2} = \frac{k^2}{4}, k = 1, 2, 3, \dots$  and  $\omega_1 = 2\sqrt{\frac{g}{l}}, \omega_2 = \sqrt{\frac{g}{l}}, \dots, \omega_n = \frac{2}{n}\sqrt{\frac{g}{l}}$ . We have parametric resonance around these frequencies  $\omega_n$ . Therefore, the pendulum around  $\theta = 0$  can be perturbed in the vertical plane; the usual pendulum resonance is in the vicinity of  $\omega = \sqrt{\frac{g}{l}}$  and in composition with the oscillator, parametrical resonances appear in the vicinity of  $\omega_n$ . For  $\varepsilon$  with small values,

$\varepsilon = \frac{x_0}{l}$ , we have instability for  $\Delta_2$  with  $\frac{1}{4} - \frac{\varepsilon}{2} < \delta$  and  $\delta < \frac{1}{4} + \frac{\varepsilon}{2}$  (initil positions choosen on the vertical), if:

$$x_0 > \frac{l}{2} - \frac{2g}{\omega^2} \text{ or } x_0 > \frac{2g}{\omega^2} - \frac{l}{2} \quad (12)$$

If besides the oscillator, an damper is mount, the equation becomes:

$$\phi'' + 2h\phi' + (\delta + \varepsilon \cos \tau)\phi = 0, h > 0 \quad (13)$$

In this case the asymptotic stability from  $Z$  is transmitted also in the nonlinear case for  $X$ . With the transformation  $\phi = e^{-h\tau}\psi$  the equation:

$$\psi^2 + (\delta - h^2 + \varepsilon \cos \tau)\psi = 0 \quad (14)$$

is obtained, which is of Mathieu type, if we take  $\delta_1 = \delta - h^2$ . For  $\delta, \varepsilon, h$  given, the stability in  $\psi$  is obtained, meaning asymptotic stability related to  $\phi$ .

### 3 The nonstationary vertical case (NV) $l = l(t), r = x, \alpha = \frac{\pi}{2}$

In this situation, the equations (4) become, considering  $\theta$  small with  $\sin \theta \approx 0, \cos \theta \approx 1$  and the linear bound  $l_1 = l_0 \pm \nu t$ .

$$\begin{cases} \ddot{x} + \omega^2 x = \mu \frac{d^2}{dt^2} \left( \frac{l_1(t)\theta^2}{2} \right) \\ \frac{d^2}{dt^2} (l(t)\theta) + \frac{g - \ddot{x}}{l} (l_1\theta) = 0 \end{cases}, x(0) = a, \dot{x}(0) = b, \omega^2 = \frac{k}{m + M}, \theta(0) = c, \dot{\theta}(0) = d \quad (15)$$

Following the small parameter  $\mu = \frac{m}{m+M}$  we look for solutions of the following type:

$$\begin{aligned} x(t) &= x_0(t) + \mu x_1(t) + \dots \\ \theta(t) &= \theta_0(t) + \mu \theta_1(t) + \dots \end{aligned} \quad (16)$$

Identifying based on the powers of  $\mu$  we obtain the following equations after  $l_1 = l_0 \pm \nu t$  where the sign  $+\nu$  is the raise to power of  $M$  and  $-\nu$  is the lowering of  $M$ .

$$\begin{aligned} \ddot{x}_0 + \omega^2 x_0 &= 0, x_0(0) = a, \dot{x}_0(0) = b \Rightarrow x_0 = a \cos \omega t + \frac{b}{\omega} \sin \omega t \\ \frac{d^2}{dt^2} (l_1 \theta_0) + \frac{g - \ddot{x}}{l} (l_1 \theta_0) &= 0, \theta_0(0) = c, \dot{\theta}_0(0) = d \end{aligned} \quad (17)$$

Considering the solution (17) we obtain for  $\theta_0$  from (18) the Cauchy problem. We look for the solution of this problem shaped as a series of powers with Picard successive approximations where  $0 \leq t \leq \frac{l_0}{\nu} = \chi$ .

$$\theta_0(t) = c + dt + \left[ -\frac{d}{\chi} - \frac{(g + a\omega^2)c}{2\chi\nu} \right] t^2 + \dots \quad (18)$$

The solution  $\theta_0(t)$  is unique in the simple regularity conditions from here. We study the implications of these solutions (17), (19) on the stability. We note with  $E(t)$  the coefficient from

the (18) equation; for a periodic solution  $E > 0$  and stable; we retrieve the boundaries specifying the period, frequency and the movement.

$$E(t) = \frac{g + \omega\sqrt{a^2\omega^2 + b^2} \sin(\omega t + \varphi)}{l_2(t)}, tg\varphi = \frac{a\omega}{b}, 0 \leq t \leq \frac{l_0}{v} = \chi \quad (19)$$

For the case of  $M$  with the speed  $(+v)$ , the length  $l_1$  shortens from  $l_0$  to  $l_1^*$  ( $l_0 > l_1^*$ ) with time  $T = \frac{l_0 - l_1^*}{v}$ , with  $E_{\min} < E < E_{\max}$ , we have for  $E > 0$ ,  $g > \omega\sqrt{a^2\omega^2 + b^2}$  the solution  $\theta_0$  becomes oscillated (harmonic) being able to pass through the zero position multiple times, until an oscillation of  $m$  is done in a period  $T = \frac{2\pi}{\omega}$ .

$$m^* = \min_{t \in [0, T]} E = \frac{g - \omega\sqrt{a^2\omega^2 + b^2}}{l_0}, M^* = \max_{t \in [0, T]} E = \frac{g + \omega\sqrt{a^2\omega^2 + b^2}}{l_0}.$$

During the wire shortening the length  $l$  shortens from  $l_0$  to  $l_1$  in  $T = (l_0 - l_1)/v$  time, the solution  $\theta_0(t)$  is oscillated and the difference between two stagnations is  $\rho$ ,  $\frac{\pi}{\sqrt{M^*}} \leq \rho \leq \frac{\pi}{\sqrt{m^*}}$ . In order for the solution  $\theta_0(t)$  to pass through zero at least  $k$  times it must be that  $T \left( \frac{\pi}{\sqrt{m^*}} \right) \geq k$  meaning

$$\frac{(l_0 - l_1)^2 (g - \omega) \sqrt{a^2\omega^2 + b^2}}{l_0 \pi^2 v^2 k^2} \geq 1 \quad (20)$$

## 4 Conclusions

1. During the pendulum wire's shortening  $l(t)$ , there is the solution that for an oscillation of  $M$  (the oscillator) in a period  $T$ , the  $m$  pendulum can make  $k$  oscillations.
2. At a lengthening of the wire  $l(t)$ , there is a possibility that for each oscillation of the  $m$  pendulum the oscillator  $M$  can perform  $n$  vertical oscillations.
3. For short periods in which  $E < 0$  auto-oscillation phenomenon may appear.

## References

- [1] ARYA, A.: *Introduction to Classical Mechanics*. Prentice Hall, 1990 (ed. I) si 1998 (ed. II).
- [2] DRAGOS, L.: *The principle of the continuum mechanics*. Ed. Technique, Bucharest, Romania, 1973.
- [3] IESEAN, D.: *The thermoelasticity theory*. Ed. Acad. Rom., 1979
- [4] VOINEA, R.: *Introduction in the dynamical system theory*. Ed. Acad. Rom., 2000
- [5] POPESCU, S.: *Mechanical oscillations, elastical waves and acoustics*. Ed. Matrix Rom, Bucharest, 2003

## Current address

**Olivia Florea, Lect. univ. PhD.**

Transilvania University of Brasov,

Dept. of Mathematics and Informatics, B-dul Iuliu Maniu, no. 50, Brasov, Romania.

e-mail: oaflorea@gmail.com

## APPLICATION OF HOMOTOPY PERTURBATION METHOD TO SOLVING SINGULAR INITIAL VALUE PROBLEMS

HALFAROVÁ Hana, (CZ), KUKHARENKO Alexandra, (UA),  
ŠMARDÁ Zdeněk, (CZ)

**Abstract.** In this paper we present the homotopy perturbation method. We apply the method to solve a class of singular initial value problems for the second-order and the third-order ordinary differential equations. The homotopy perturbation method yields solutions in convergent series forms with easily computable terms, and in presented examples, this method gives exact solutions.

**Key words and phrases.** Singular initial value problem, homotopy perturbation method.

*Mathematics Subject Classification.* Primary 26A33; Secondary 34A08.

### 1 Introduction

The homotopy perturbation method (HPM) was proposed for solving linear and nonlinear differential equations, integral and integro-differential equations first by He [9-11]. The homotopy perturbation method [7-11, 13] is a combination of the classical perturbation technique and homotopy concept as used in topology.

Several techniques including decomposition, spline, finite difference, multi-integral, modified variational iteration and variational iteration have been applied for solving singular equations which arise in several physical phenomena in mathematical physics, astrophysics, theory of stellar structure, thermal behavior of a spherical cloud of gas, isothermal gas spheres and theory of thermionic currents (see [1-5, 12-18]).

In the paper we apply the He's homotopy perturbation method to solving singular initial value problems for the second-order and the third-order ordinary differential equations. Using this method and its modification we obtain exact solutions for certain classes of singular initial value problems.

## 2 Homotopy perturbation method

To illustrate the basic ideas of this method we consider the following equation

$$A(u) - f(r) = 0, \quad r \in \Omega \quad (1)$$

with the boundary conditions

$$B(u, \partial u / \partial n) = 0, \quad r \in \Gamma \quad (2)$$

where  $A$  is a general differential operator,  $B$  a boundary operator,  $f(r)$  a known analytical function and  $\Gamma$  is the boundary domain  $\Omega$ .

The operator  $A$  can be generally divided into two parts of  $L$  and  $N$ , where  $L$  is the linear part, while  $N$  is the nonlinear part. Equation (1) can be rewritten as

$$L(u) + N(u) - f(r) = 0. \quad (3)$$

By the homotopy technique, we construct a homotopy as

$$v(r, p) : \Omega \times [0, 1] \rightarrow R$$

which satisfies

$$H(v, p) = (1 - p)[L(v) - L(u_0)] + p[A(v) - f(r)] = 0, \quad r \in \Omega \quad (4)$$

or

$$H(v, p) = L(v) - L(u_0) + pL(u_0) + p[N(v) - f(r)] = 0 \quad (5)$$

where  $p \in [0, 1]$  is an embedding parameter and  $u_0$  is an initial approximation of (1) which satisfies the boundary conditions. From here we obtain

$$H(v, 0) = L(v) - L(u_0) = 0, \quad (6)$$

$$H(v, 1) = A(v) - f(r) = 0. \quad (7)$$

Changing the variation of  $p$  from 0 to 1 is the same as changing  $H(v, p)$  from  $L(v) - L(u_0)$  to  $A(v) - f(r)$ . In topology, this is called deformation,  $L(v) - L(u_0)$  and  $A(v) - f(r)$  are called homotopic. According to HPM, we can use the embedding parameter  $p$  as a small parameter and assume that a solution of (4) and (5) can be written as a power series

$$v = v_0 + pv_1 + p^2v_2 + \dots \quad (8)$$

The approximate solution of (1) can be obtained as

$$u = \lim_{p \rightarrow 1} v = v_0 + v_1 + \dots \quad (9)$$

The convergence of series (9) has been proved by He[11].

### 3 HPM for Lane-Emden type equations

Consider the Lane-Emden type initial value problem

$$y'' + \frac{2}{x}y' + F(u) = m, \quad y(0) = A, \quad y'(0) = B, \quad (10)$$

where  $A, B$  are constants. We define the homotopy as

$$y'' + \frac{2}{x}y' + pF(u) = 0, \quad (11)$$

where  $p \in [0, 1]$  is the embedding parameter. Let

$$y = y_0 + py_1 + p^2y_2 + p^3y_3 + \dots \quad (12)$$

be the solution of (10). Substituting (12) in (11) we get

$$\sum_{i=0}^{\infty} p^i y_i'' + \frac{2}{x} \left( \sum_{i=0}^{\infty} p^i y_i' \right) + pF \left( \sum_{i=0}^{\infty} p^i y_i \right) = 0. \quad (13)$$

Equating the coefficients of terms of like powers of  $p$  in (13) gives

$$\begin{aligned} p^0 &: y_0'' + \frac{2}{x}y_0' = 0, \quad y_0(0) = A, \quad y_0'(0) = B. \\ p^1 &: y_1'' + \frac{2}{x}y_1' + F(y_0) = 0, \quad y_1(0) = y_1'(0) = 0. \\ p^2 &: y_2'' + \frac{2}{x}y_2' + \frac{d}{dy_0}F(y_0) = 0, \quad y_2(0) = y_2'(0) = 0, \\ p^3 &: y_3'' + \frac{2}{x}y_3' + y_2 \frac{d}{dy_0}F(y_0) + \frac{1}{2}y_1^2 \frac{d^2}{dy_0^2}F(y_0) = 0, \quad y_3(0) = y_3'(0) = 0, \\ &\vdots \end{aligned}$$

**Example 1.** Consider the following singular initial value problem

$$y'' + \frac{2}{x}y' + 4(2e^y + e^{y/2}) = 0, \quad y(0) = y'(0) = 0.$$

From (13) we have

$$\begin{aligned} p^0 &: y_0'' + \frac{2}{x}y_0' = 0, \quad y_0(0) = 0, \quad y_0'(0) = 0. \\ p^1 &: y_1'' + \frac{2}{x}y_1' + 4(2e^{y_0} + e^{y_0/2}) = 0, \quad y_1(0) = y_1'(0) = 0. \\ p^2 &: y_2'' + \frac{2}{x}y_2' + 4y_1(2e^{y_0} + \frac{1}{2}e^{y_0/2}) = 0, \quad y_2(0) = y_2'(0) = 0, \\ p^3 &: y_3'' + \frac{2}{x}y_3' + 4y_2(2e^{y_0} + \frac{1}{2}e^{y_0/2}) + 2y_1^2(2e^{y_0} + \frac{1}{4}e^{y_0/2}) = 0, \quad y_3(0) = y_3'(0) = 0, \\ &\vdots \end{aligned}$$

Corresponding solutions have the form

$$y_0(x) = 0, \quad y_1(x) = -2x^2, \quad y_2(x) = x^4, \quad y_3(x) = -\frac{2}{3}x^6, \quad \dots$$

From here we obtain

$$y(x) = -2x^2 + x^4 - \frac{2}{3}x^6 + \dots = -2\ln(1 + x^2).$$

**Example 2.** Now we consider singular initial value problem for the differential equation of the third order

$$y''' - y'' - \frac{1}{x}y = 0, \quad y(0) = 0, \quad y'(0) = 1, \quad y''(0) = 2. \quad (14)$$

Put  $y_1(x) = y(x)$ ,  $y_2(x) = y''$ ,  $y_3(x) = y''(x)$  then equation (14) is equivalent to the system

$$\begin{aligned} y_1' &= y_2, \\ y_2' &= y_3, \\ y_3' &= \frac{1}{x}y_1 + y_3 \end{aligned} \quad (15)$$

According to HPM we have

$$v_1 = \sum_{i=0}^{\infty} p^i v_{1i}, \quad v_2 = \sum_{i=0}^{\infty} p^i v_{2i}, \quad v_3 = \sum_{i=0}^{\infty} p^i v_{3i} \quad (16)$$

and

$$y_1(x) = \lim_{p \rightarrow 1} v_1(x), \quad y_2(x) = \lim_{p \rightarrow 1} v_2(x), \quad y_3(x) = \lim_{p \rightarrow 1} v_3(x). \quad (17)$$

We can construct a homotopy of system (15) in the form

$$\begin{aligned} (1-p)(v_1' - v_2 - u_{10}') + p(v_1' - v_2) &= 0 \\ (1-p)(v_2' - v_3 - u_{20}') + p(v_2' - v_3) &= 0 \\ (1-p)(v_3' - u_{30}') + p(v_3' - \frac{1}{x}v_1 - v_3) &= 0 \end{aligned} \quad (18)$$

Substituting initial conditions and (16) into (18) and rearranging based on powers of  $p$ -terms, we get

$$\begin{aligned} (v_{10}' - v_{20}) + p(v_{11}' - v_{21})p^2(v_{12}' - v_{22}) + p^3(v_{13}' - v_{23}) + \dots &= 0, \\ (v_{20}' - v_{30}) + p(v_{21}' - v_{31})p^2(v_{22}' - v_{32}) + p^3(v_{23}' - v_{33}) + \dots &= 0, \\ v_{30}' + p(v_{31}' - v_{30} - \frac{1}{x}v_{10}) + p^2(v_{32}' - v_{31} - \frac{1}{x}v_{11}) + p^3(v_{33}' - v_{32} - \frac{1}{x}v_{12}) + \dots &= 0 \end{aligned}$$

Solving the system differential equations at powers of  $p^i$ ,  $i = 0, 1, 2, \dots$ , we obtain

$$\begin{aligned}
 v_{10}(x) &= x + x^2, \\
 v_{11}(x) &= \frac{1}{2}x^3 + \frac{1}{24}x^4, \\
 v_{12}(x) &= \frac{1}{8}x^4 + \frac{1}{60}x^5 + \frac{1}{2880}x^6, \\
 v_{13}(x) &= \frac{1}{40}x^5 + \frac{11}{2880}x^6 + \frac{13}{100800}x^7 + \frac{1}{967680}x^8, \\
 &\vdots \\
 v_{20}(x) &= 1 + 2x, \\
 v_{21}(x) &= \frac{3}{2}x^2 + \frac{1}{6}x^3, \\
 v_{22}(x) &= \frac{1}{2}x^3 + \frac{1}{12}x^4 + \frac{1}{480}x^5, \\
 v_{23}(x) &= \frac{1}{8}x^4 + \frac{11}{480}x^5 + \frac{13}{14400}x^6 + \frac{1}{120960}x^7, \\
 &\vdots \\
 v_{30}(x) &= 2, \\
 v_{31}(x) &= 3x + \frac{1}{2}x^2, \\
 v_{32}(x) &= \frac{3}{2}x^2 + \frac{1}{3}x^3 + \frac{1}{96}x^4, \\
 v_{33}(x) &= \frac{1}{2}x^3 + \frac{11}{96}x^4 + \frac{13}{2400}x^5 + \frac{1}{17280}x^6, \\
 &\vdots
 \end{aligned}$$

From (17) we obtain

$$y_1(x) = x + x^2 + \frac{1}{2!}x^3 + \frac{1}{3!}x^4 + \frac{1}{4!}x^5 + \frac{1}{5!}x^6 + \dots \quad (19)$$

$$y_2(x) = 1 + 2x + \frac{3}{2}x^2 + \frac{2}{3}x^3 + \frac{5}{24}x^4 + \frac{1}{20}x^5 + \frac{1}{180}x^6 + \dots$$

$$y_3(x) = 2 + 3x + 2x^2 + \frac{5}{6}x^3 + \frac{1}{4}x^4 + \frac{1}{30}x^5 + \frac{23}{14400}x^6 + \dots$$

Hence  $y(x) = y_1(x) = xe^x$  is the exact solution of equation (14).

### Acknowledgement

This research has been supported by the Grant FEKT-S-11-2-921 of Faculty of Electrical Engineering and Communication, Brno University of Technology.

## References

- [1] ADOMIAN, G., RACH, R.: *Noise terms in decomposition series solution*, Comput. Math. Appl., 24 (11)1992, 61-69.
- [2] ADOMIAN, G., RACH, R., SHAWAGFEV, N.T.: *On the analytic solution of Lane-Emden equation*, Found. Phus. Lett., 8 (2)1995, 161-166.
- [3] ARCHALOUSOVÁ, O., ŠMARDA, Z.: *Adomian decomposition method for certain singular initial value problems*, Proceedings of the 9th International Conference Aplimat 2010, 2010, 183-188.
- [4] CHANDRASEKHAR, S.: *Introduction to the Study of Stellar Structure*, Dover, New York, 1967.
- [5] CHAWLA, M., MC KEE, S., SHAW, G.: *Order  $h^2$  method for a singular two-point boundary value problems*, BIT, 26, 1985, 318-325.
- [6] GHORBANI, A., NADJFI, J.S.: *Hes homotopy perturbation method for calculating Adomians polynomials*, Int. J. Nonlin. Sci. Num. Simul., 8 (2) 2007, 229-332.
- [7] HE, J.H.: *Recent developments of the homotopy perturbation method*, Top. Meth. Nonlin. Anal., 31, 2008, 205-209.
- [8] HE, J.H.: *Some asymptotic methods for strongly nonlinear equation*, Int. J. Mod. Phys., 20 (10), 2006, 1144-1199.
- [9] HE, J.H.: *Comparison of homotopy perturbation method and homotopy analysis method*, Appl. Math. Comput., 156, 2004, 527-539.
- [10] HE, J.H.: *Homotopy perturbation method for bifurcation of nonlinear problems*, Int. J. Nonlin. Sci. Numer. Simul., 6 (2), 2005, 207-208.
- [11] HE, J.H.: *A coupling method of homotopy technique and perturbation technique for nonlinear problems*, Int. J. Nonlin. Mech., 35 (1), 2000, 115-123.
- [12] YENGAR, S., R., K., JAIN, P.: *Spline difference methods for singular two-point boundary value problem*, Numer. Math., 500, 1987, 363-369.
- [13] MOHYUD-DIN, S.T., NOOR, M.A., NOOR, K.I.: *Solution of singular equations by Hes variational iteration method*, Int. J. Nonlin. Sci. Num. Sim., 2009, 109-121.
- [14] MOHYUD-DIN, S.T., NOOR, M.A.: *Solution of singular and nonsingular initial and boundary value problems by modified variational iteration method*, Math. Prob. Engg. 2008 (2008), Article ID 917407, 23 pages, doi:10.1155/2008/917407.
- [15] RUSSELL, R.D., SHAMPINE, L.F.: *Numerical methods for singular boundary value problems*, SIAM J. Numer. Anal., 12 , 1975, 13-16.
- [16] SHAWAGFEH, N.T.: *Nonperturbative approximate solution for Lane-Emden equation*, J. Math. Phys., 34 (9) 1993, 4364-4369.
- [17] WAZWAZ, A.M.: *A new algorithm for solving differential equations of Lane-Emden type*, Appl. Math. Comput., 111, 2000, 53-62.
- [18] WAZWAZ, A.M.: *A new method for solving initial value problems in the second-order ordinary differential equations*, Appl. Math. Comput., 128, 2002, 45-57.

## Current address

**Mgr. Hana Halfarová**

Department of Mathematics

Faculty of Electrical Engineering and Communication  
Brno University of Technology, Technická 8, 616 00 BRNO  
e-mail: xhalfa06@stud.feec.vutbr.cz

**Aleksandra Kukharenko**

Department of Complex Systems Modelling  
Kiev University, 01033 Kiev, UKRAINE  
e-mail: akukharenko@ukr.net

**Doc. RNDr. Zdeněk Šmarda, CSc.**

Department of Mathematics  
Faculty of Electrical Engineering and Communication  
Brno University of Technology, Technická 8, 616 00 BRNO  
e-mail: smarda@feec.vutbr.cz



## ON STABILITY INTERVALS OF EULER METHODS FOR A DELAY DIFFERENTIAL EQUATION

HRABALOVÁ Jana, (CZ)

**Abstract.** The paper discusses the asymptotic stability regions of Euler discretizations for a linear delay differential equation

$$y'(t) = ay(t - \tau).$$

We compare our results with the asymptotic stability domain for the underlying delay differential equation.

**Key words and phrases.** Delay differential equation, Euler methods, asymptotic stability.

*Mathematics Subject Classification.* Primary 39A12, 65L20; Secondary 39A30.

### 1 Introduction

The aim of this paper is to investigate the asymptotic stability regions of Euler methods for the test delay differential equation

$$y'(t) = ay(t - \tau), \quad a \in \mathbb{R}, \quad t > 0 \quad (1)$$

$$y(t) = g(t), \quad -\tau \leq t \leq 0. \quad (2)$$

We recall that equation (1) is called asymptotically stable if

$$\lim_{t \rightarrow \infty} y(t) = 0$$

for all continuous initial functions  $g(t)$ . It is well known that the asymptotic stability domain  $S_\tau$  of (1) is given by

$$S_\tau = \left\{ a \in \mathbb{R} : 0 > a\tau > -\frac{\pi}{2} \right\}, \quad (3)$$

which yields the necessary and sufficient condition for the asymptotic stability of (1). The aim of this paper is to formulate intervals of asymptotic stability for its Euler discretizations. Moreover, we are going to discuss their mutual relations, as well as their relations with respect to the asymptotic stability domain  $S_\tau$  of (1).

The paper is organized as follows. Section 2 provides discretizations of the equation (1) for both the forward and backward Euler method. In this part we also introduce the notion of stability and recall the Levin-May result, which gives a criterion for the asymptotic stability of a three-term difference equation. In Section 3, we discuss the asymptotic stability region of the backward Euler method and its properties. The asymptotic stability intervals for the forward Euler method are investigated in Section 4. The final section presents other discretizations of the equation (1).

## 2 Preliminaries

We start with discretizations of the equation (1). Let  $h > 0$  be a stepsize given by

$$h = \frac{\tau}{k}, \quad k \in \mathbb{Z}^+. \quad (4)$$

This stepsize restriction is imposed to avoid an interpolation of a delayed term.

When we apply the backward Euler method with the stepsize  $h$  to the equation (1), we obtain the difference equation

$$x_{n+1} = x_n + ahx_{n+1-k}, \quad (5)$$

where  $x_n$  is the numerical solution at the grid points  $t_n = nh$ ,  $n \in \mathbb{Z}^+$ .

The forward Euler method leads to the difference equation of the form

$$x_{n+1} = x_n + ahx_{n-k}. \quad (6)$$

Both equations (5) and (6) are special three-term delay difference equations of the form

$$x_{n+1} = x_n + \alpha x_{n-m}, \quad n = 0, 1, 2, \dots, \quad (7)$$

where  $\alpha \in \mathbb{R}$  and  $m \in \mathbb{Z}^+$  are scalars. We recall that (7) is said to be asymptotically stable if

$$\lim_{n \rightarrow \infty} x_n = 0$$

for any solution  $x_n$  of (7).

One of basic stability notions considered in numerical discretizations of delay differential equations is the notion of  $\tau(0)$ -stability (see [1]). We recall this notion for a general numerical step-by-step method.

**Definition 2.1** *The  $\tau(0)$ -stability region of a numerical step-by-step method for (1) is the set*

$$S_{\tau(0)} = \bigcap_{k \geq 1} S_{\tau,k},$$

where, for given integers  $k$  and  $\tau$ ,  $S_{\tau,k}$  is the set of the real numbers  $a$  such that the discrete numerical solution  $\{x_n\}_{n \geq 0}$  of (1), with a constant step size  $h = \frac{\tau}{k}$ , satisfies  $\lim_{n \rightarrow \infty} x_n = 0$  for all initial functions  $g(t)$ .

**Definition 2.2** A numerical step-by-step method for (1) is  $\tau(0)$ -stable if

$$S_{\tau(0)} \supseteq S_{\tau}.$$

The problem of  $\tau(0)$ -stability for the equation (1) (involving also a non-delayed term  $by(t)$ ) was discussed by Calvo-Grande [2] and Guglielmi [4], who showed that the backward Euler method is  $\tau(0)$ -stable, whereas the forward Euler method does not have this property. In this paper, we aim to find explicit stability intervals for numerical discretizations (5), (6) including their basic properties (for other recent qualitative investigations of delay differential equations we refer, e.g. to [7]).

To analyze these properties, we utilize the following necessary and sufficient condition for asymptotic stability of the difference equation (7), which is due to Levin-May (see [5]).

**Theorem 2.1** Let  $\alpha$  be a real constant and  $m$  be a positive integer. The difference equation (7) is asymptotically stable if and only if

$$0 > \alpha > -2 \cos \frac{m\pi}{2m+1}. \quad (8)$$

In the sequel, we use the criterion (8) in its equivalent form

$$0 > \alpha > -2 \sin \frac{\pi}{4m+2}. \quad (9)$$

We show that the inequalities (8) and (9) are actually equivalent, i.e. it holds

$$\cos \frac{m\pi}{2m+1} = \sin \frac{\pi}{4m+2}. \quad (10)$$

Rewrite the equality (10) as

$$\arccos \left( \cos \frac{m\pi}{2m+1} \right) = \arccos \left( \sin \frac{\pi}{4m+2} \right).$$

Using

$$\arccos x = \frac{\pi}{2} - \arcsin x$$

we arrive at

$$\arccos \left( \sin \frac{\pi}{4m+2} \right) = \frac{\pi}{2} - \arcsin \left( \sin \frac{\pi}{4m+2} \right) = \frac{\pi}{2} - \frac{\pi}{4m+2} = \frac{m\pi}{2m+1} = \arccos \left( \cos \frac{m\pi}{2m+1} \right).$$

Comparing the relevant relations we can verify the equivalency of (8) and (9).

### 3 The backward Euler method

In this section, we will focus on the backward Euler method (5). To our purposes it is convenient to consider the equation (5) in the form

$$x_{n+1} = x_n + \frac{a\tau}{k} x_{n+1-k}. \quad (11)$$

Let  $S_{\tau,k}$  be the set of all real parameters  $a$  such that (11) is asymptotically stable. The direct application of Theorem 2.1 to (11) yields

$$S_{\tau,k} = \left\{ a \in \mathbb{R} : 0 > a\tau > -2k \sin \frac{\pi}{4k-2} \right\}.$$

The following assertion describes some basic properties of  $S_{\tau,k}$ .

**Theorem 3.1** *Let  $k_1, k_2$  be arbitrary positive integers such that  $k_2 > k_1 \geq 2$ . Then*

$$S_{\tau,k_1} \supset S_{\tau,k_2} \supset S_{\tau}.$$

Moreover,

$$\lim_{k \rightarrow \infty} S_{\tau,k} = S_{\tau}.$$

**Proof.** We wish to show that

$$-2k_1 \sin \frac{\pi}{4k_1-2} < -2k_2 \sin \frac{\pi}{4k_2-2}, \quad 2 \leq k_1 < k_2$$

or equivalently,

$$2k_1 \sin \frac{\pi}{4k_1-2} > 2k_2 \sin \frac{\pi}{4k_2-2}, \quad 2 \leq k_1 < k_2. \quad (12)$$

Consider the function

$$f(x) = 2x \sin \frac{\pi}{4x-2}, \quad x \in \mathbb{R}. \quad (13)$$

The inequality (12) is satisfied when  $f(x)$  is decreasing on the interval  $\langle 2, \infty \rangle$ . Since

$$f'(x) = 2 \sin \frac{\pi}{4x-2} - \frac{8\pi x}{(4x-2)^2} \cos \frac{\pi}{4x-2},$$

it is enough to show that

$$2 \sin \frac{\pi}{4x-2} - \frac{8\pi x}{(4x-2)^2} \cos \frac{\pi}{4x-2} < 0 \quad \text{on } \langle 2, \infty \rangle. \quad (14)$$

Obviously,

$$\cos \frac{\pi}{4x-2} > 0, \quad x \in \langle 2, \infty \rangle,$$

hence (14) is equivalent to

$$\tan \frac{\pi}{4x-2} < \frac{4\pi x}{(4x-2)^2}. \quad (15)$$

Both the functions are continuous and decreasing on the interval  $\langle 2, \infty \rangle$ . Moreover, for  $x = 2$

$$\tan \frac{\pi}{6} < \frac{2\pi}{9}.$$

We consider the equality

$$\tan \frac{\pi}{4x-2} = \frac{4\pi x}{(4x-2)^2}. \quad (16)$$

If we substitute

$$s = \frac{\pi}{4x - 2},$$

then  $x = \frac{\pi}{4s} + \frac{1}{2}$  and (16) becomes

$$\tan s = s + \frac{2s^2}{\pi}, \quad s \in \left(0, \frac{\pi}{6}\right).$$

Let  $g_1(s) = \tan s$  and  $g_2 = s + \frac{2s^2}{\pi}$ . Both the functions are increasing on  $\left(0, \frac{\pi}{6}\right)$  and

$$g_1(0) = g_2(0) \quad \text{and} \quad g_1\left(\frac{\pi}{6}\right) < g_2\left(\frac{\pi}{6}\right).$$

We show that  $g_1(s) < g_2(s)$  on  $\left(0, \frac{\pi}{6}\right)$ . Doing this, we first investigate the derivatives

$$g_1'(s) = \frac{1}{\cos^2 s}, \quad g_2'(s) = 1 + \frac{4s}{\pi}.$$

Obviously,

$$g_1'(0) = g_2'(0), \quad g_1'\left(\frac{\pi}{6}\right) < g_2'\left(\frac{\pi}{6}\right)$$

and both  $g_1'(s)$ ,  $g_2'(s)$  are increasing on  $\left(0, \frac{\pi}{6}\right)$ . To discuss the inequality  $g_1'(s) < g_2'(s)$  on  $\left(0, \frac{\pi}{6}\right)$  we consider the second derivatives

$$g_1''(s) = \frac{2 \sin s}{\cos^3 s}, \quad g_2''(s) = \frac{4}{\pi}.$$

It holds

$$g_1''(0) = 0 \quad \text{and} \quad g_1''\left(\frac{\pi}{6}\right) = \frac{8}{3\sqrt{3}} > \frac{4}{\pi}.$$

Since  $g_1''(s)$  is continuous on  $\left(0, \frac{\pi}{6}\right)$ ,  $g_1''(s)$  and  $g_2''(s)$  intersects each other on  $\left(0, \frac{\pi}{6}\right)$ . Moreover,

$$g_1'''(s) = \frac{2 \cos^4 s + 6 \sin^2 s \cos^2 s}{\cos^6 s} > 0, \quad s \in \left(0, \frac{\pi}{6}\right),$$

hence  $g_1''(s)$  is increasing on  $\left(0, \frac{\pi}{6}\right)$  and there exists a unique  $\beta \in \left(0, \frac{\pi}{6}\right)$  such that  $g_1''(\beta) = g_2''(\beta)$ . Since  $g_1''\left(\frac{\pi}{8}\right) < \frac{4}{\pi}$ , we can specify that  $\beta \in \left(\frac{\pi}{8}, \frac{\pi}{6}\right)$ . Consequently,

$$g_1'(s) < g_2'(s), \quad s \in \left(0, \frac{\pi}{8}\right). \quad (17)$$

Discussions on  $g_1''(s)$ ,  $g_2''(s)$  show that there exists at most one root  $\gamma \in \left(\frac{\pi}{8}, \frac{\pi}{6}\right)$  of  $g_1'(s) = g_2'(s)$ . However,

$$g_1'\left(\frac{\pi}{6}\right) < g_2'\left(\frac{\pi}{6}\right)$$

and combining with (17) we have

$$g_1'(s) < g_2'(s), \quad \text{and} \quad s \in \left(0, \frac{\pi}{6}\right).$$

Consequently,

$$g_1(s) < g_2(s), \quad s \in \left(0, \frac{\pi}{6}\right),$$

which yields that  $f(x)$  is decreasing on  $\langle 2; \infty \rangle$ . It ensures the validity of (12), which implies that  $S_{\tau, k_1} \supset S_{\tau, k_2}$ .

Further, we show that  $\lim_{k \rightarrow \infty} S_{\tau, k} = S_\tau$ . We are interested in the limit

$$\lim_{k \rightarrow \infty} -2k \sin \frac{\pi}{4k-2}.$$

Using the L'Hospital rule we have

$$\lim_{k \rightarrow \infty} -2k \sin \frac{\pi}{4k-2} = \lim_{k \rightarrow \infty} \frac{\sin \frac{\pi}{4k-2}}{-\frac{1}{2k}} = \lim_{k \rightarrow \infty} \frac{\frac{-2\pi}{(4k-2)^2} \cos \frac{\pi}{4k-2}}{\frac{1}{4k^2}} = \lim_{k \rightarrow \infty} \frac{-8\pi k^2}{(4k-2)^2} \cos \frac{\pi}{4k-2} = -\frac{\pi}{2}$$

and the required limit property is proved.

**Corollary 3.2** *The backward Euler method is  $\tau(0)$ -stable.*

**Remark 3.3** *The stability regions for  $k=1$  and  $k=2$  are identical.*

**Remark 3.4** *Recall that the stepsize  $h$  is inversely proportional to  $k$  via the relation (4). Hence, we can interpret Theorem 3.1 as a dependence of the stability regions on changing stepsize  $h$ . Theorem 3.1 implies that stability intervals for the backward Euler discretization (5) enlarge with increasing  $h$ . On the contrary, if  $h$  is approaching zero, the stability domain of (5) tends to the stability domain of the corresponding delay differential equation. Note that the above conclusions have been verified only experimentally in [6].*

## 4 The forward Euler method

Now we analyze the asymptotic stability regions for the forward Euler method (6). Similarly as in the previous section, we find useful to represent (6) in the form

$$x_{n+1} = x_n + \frac{a\tau}{k} x_{n-k}. \quad (18)$$

By Levin-May result, the asymptotic stability region  $S_{\tau, k}$  is

$$S_{\tau, k} = \left\{ a \in \mathbb{R} : 0 > a\tau > -2k \sin \frac{\pi}{4k+2} \right\}.$$

**Theorem 4.1** *Let  $k_1, k_2$  be arbitrary positive integers such that  $k_2 > k_1 \geq 1$ . Then*

$$S_{\tau, k_1} \subset S_{\tau, k_2} \subset S_\tau.$$

Moreover,

$$\lim_{k \rightarrow \infty} S_{\tau, k} = S_\tau.$$

**Proof.**

To prove Theorem 4.1, we can use a similar approach as we utilized in the proof of Theorem 3.1. In this case, we aim to show that

$$2k_1 \sin \frac{\pi}{4k_1 + 2} < 2k_2 \sin \frac{\pi}{4k_2 + 2}, \quad 1 \leq k_1 < k_2.$$

i.e. that the function

$$f(x) = 2x \sin \frac{\pi}{4x + 2}, \quad x \in \mathbb{R} \quad (19)$$

is increasing on  $\langle 1, \infty \rangle$ . Obviously  $f'(x) > 0$  on  $\langle 1; \infty \rangle$  if

$$\tan \frac{\pi}{4x + 2} > \frac{4\pi x}{(4x + 2)^2}, \quad x \in \langle 1, \infty \rangle. \quad (20)$$

Let us introduce the substitution

$$s = \frac{\pi}{4x + 2}.$$

The inequality (20) can be transformed to

$$\tan s > s - \frac{2s^2}{\pi}, \quad s \in \left(0, \frac{\pi}{6}\right). \quad (21)$$

Let  $g_1(s) = \tan s$ ,  $g_2(s) = s - \frac{2s^2}{\pi}$ . Obviously,  $g_1(0) = g_2(0) = 0$ . Using expressions for  $g'_1(s)$  and  $g'_2(s)$  we can directly conclude that the inequality (21) holds for all  $s \in (0, \frac{\pi}{6})$ . Consequently, the function  $f(x)$  is increasing on  $\langle 1, \infty \rangle$ . That implies  $S_{\tau, k_1} \subset S_{\tau, k_2}$ . The proof of the limit property of the stability intervals is a simple modification of the technique used in Theorem 3.1.

**Remark 4.2** Analogously as in Section 3, we can interpret the assertion of Theorem 4.1 as a dependence of stability intervals for the forward Euler method on changing stepsize  $h$ . Theorem 4.1 implies that the stability region is enlarging with decreasing stepsize  $h$ . If  $h$  is approaching zero, the corresponding stability region is tending to  $S_\tau$ .

## 5 Final remarks

We presented the analysis of the asymptotic stability intervals for the backward and forward Euler method applied to the delay differential equation (1). These methods are particular cases of a wider class of methods called the  $\theta$ -methods. When we apply the  $\theta$ -method with the stepsize  $h$  to (1), we obtain a difference equation of the form

$$x_{n+1} = x_n + ah(\theta x_{n+1-k} + (1 - \theta)x_{n-k}). \quad (22)$$

If  $\theta = 1$  we get the backward Euler method, while the case  $\theta = 0$  yields the forward Euler method. We note that the general  $\theta$ -method leads to a four-term difference equation. Therefore the Levin-May criterion (Theorem 2.1) cannot be used for its stability analysis. Instead, we

can use the result of Čermák et al. [3], who derived the conditions for the asymptotic stability of four terms difference equations. The analysis of these more advanced discretizations will be a subject of our further research.

### **Acknowledgement**

The work was supported by the grant P201/11/0768 of the Czech Science Foundation and by the project FEKT/FSI-S-11-1 of Brno University of Technology.

### **References**

- [1] BELLEN, A., ZENNARO, M.: *Numerical methods for delay differential equations*. New York (U.S.A.): Oxford University Press, 2005. 395 p. ISBN 0-19-850654-6.
- [2] CALVO, M., GRANDE, T.: *On the asymptotic stability of  $\theta$ -methods for delayed differential equations*. Numer. Math., **54**, pp. 257-269, 1988.
- [3] ČERMÁK, J., JÁNSKÝ J., KUNDRÁT P.: *On necessary and sufficient conditions for the asymptotic stability of higher order linear difference equations*. Journal of Difference Equations and Applications. DOI:10.1080/10236198.2011.595406 (to appear)
- [4] GUGLIELMI, N.: *Delay dependent stability regions of  $\theta$ -methods for delay differential equations*. IMA Journal of Numerical analysis, **18**, pp. 399-418, 1998.
- [5] LEVIN, S.A, MAY, R.: *A note on difference delay equations*. Theor. Popul. Biol. **9**, pp. 178-187, 1976.
- [6] KIPNIS, M.M, LEVITSKAYA, I.S.: *Stability of delay dependent difference and differential equations: similarities and distinctions*. Proceedings of the International Conference on Difference Equations, Special Functions and Applications, Munich (World Scientific), 2005.
- [7] SVOBODA, Z.: *Asymptotic properties of delayed exponential of matrix*. Journal of Applied Mathematics. pp. 167 - 172, 2010.

### **Current address**

**Jana Hrabalová, Ing.**

Institute of Mathematics, Brno University of Technology,  
Technická2, CZ-616 69 Brno, Czech Republic.

E-mail: yhraba05@stud.fme.vutbr.cz.

Telephone number 00420 54114 2553.

## THE AUTONOMOUS SYSTEM DERIVED FROM VAN DER POL-MATHIEU EQUATION

KADEŘÁBEK Zdeněk, (CZ)

**Abstract.** This work is devoted to the autonomous system derived from Van der Pol-Mathieu equation which was applied to the study the dynamics of dusty plasmas in the article [3]. In this work we shall investigate mathematically this autonomous system and shall find how large region of the plane will be attracted to the equilibrium point.

**Key words and phrases.** Autonomous system, Van der Pol-Mathieu equation, attracting set, equilibrium point, ordinary first order differential equation.

*Mathematics Subject Classification.* Primary 34C05, 34C25; Secondary 34D05.

### 1 Introduction

The main inducement for the study of the autonomous system derived from Van der Pol-Mathieu equation, which describes the dynamics of dusty plasmas, was the article [3]. The mathematical analysis of the autonomous system in [3] is mathematically deficient and this article complete this analysis.

F. Veerman and F. Verhulst proved the existence of periodic and quasiperiodic solutions of the Van der Pol-Mathieu equation in [5]. The aim of this work is to extend the article [5], to perform the phase space analysis, describe the asymptotic behavior of trajectories and to find the attracting set of equilibrium point of the investigated autonomous system.

## 2 The autonomous system derived from Van der Pol-Mathieu equation

In article [3], this autonomous system is derived:

$$\frac{da}{dt} = \frac{\alpha}{2}a - \frac{b}{2} \left( \epsilon + \frac{h\omega_0}{2} \right) - \frac{\beta}{8}(a^3 + ab^2), \quad (1)$$

$$\frac{db}{dt} = \frac{\alpha}{2}b + \frac{a}{2} \left( \epsilon - \frac{h\omega_0}{2} \right) - \frac{\beta}{8}(b^3 + a^2b), \quad (2)$$

where  $\alpha, \beta, \omega_0, h \in \mathbb{R}^+$  and  $\epsilon \in \mathbb{R}$ ,  $|\epsilon| \ll 1$ ,  $h \ll 1$ . The real unknowns  $a(t)$  and  $b(t)$  vary slowly with time  $t$  and they occur as the coefficients of the estimated solution of Van der Pol-Mathieu equation in the work [3]. The estimated solution has the form:

$$x(t) = a(t) \cos \left( \omega_0 + \frac{\epsilon}{2} \right) t + b(t) \sin \left( \omega_0 + \frac{\epsilon}{2} \right) t. \quad (3)$$

We will examine the autonomous system of two ordinary first order differential equations (1), (2). From the form of the autonomous system it is evident that the equations (1), (2) are invariant under the transformation  $(a, b) \rightarrow (-a, -b)$ . Thanks to the continuity of right-hand sides of the autonomous system (1), (2) and their first order derivations it follows that solutions of any initial problem for the autonomous system (1), (2) exist and they are uniquely determined by initial conditions.

## 3 The equilibrium points

It is clear that the autonomous system (1), (2) has the equilibrium point  $(a, b) = (0, 0)$ .

1. Assuming

$$|\epsilon| < \frac{h\omega_0}{2} \quad \wedge \quad \alpha > \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}, \quad (4)$$

and solving the system of equation

$$\frac{\alpha}{2}a - \frac{b}{2} \left( \epsilon + \frac{h\omega_0}{2} \right) - \frac{\beta}{8}(a^3 + ab^2) = 0, \quad \frac{\alpha}{2}b + \frac{a}{2} \left( \epsilon - \frac{h\omega_0}{2} \right) - \frac{\beta}{8}(b^3 + a^2b) = 0,$$

we get four further equilibrium points:

$$(a_{11}, b_{11}) = \left( 2\sqrt{\frac{\frac{h\omega_0}{2} + \epsilon}{\beta h\omega_0}} \sqrt{\alpha - \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}}, 2\sqrt{\frac{\frac{h\omega_0}{2} - \epsilon}{\beta h\omega_0}} \sqrt{\alpha - \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}} \right), \quad (5)$$

$$(a_{12}, b_{12}) = \left( -2\sqrt{\frac{\frac{h\omega_0}{2} + \epsilon}{\beta h\omega_0}} \sqrt{\alpha - \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}}, -2\sqrt{\frac{\frac{h\omega_0}{2} - \epsilon}{\beta h\omega_0}} \sqrt{\alpha - \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}} \right), \quad (6)$$

$$(a_{21}, b_{21}) = \left( -2\sqrt{\frac{h\omega_0}{2} + \epsilon} \sqrt{\alpha + \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}}, 2\sqrt{\frac{h\omega_0}{2} - \epsilon} \sqrt{\alpha + \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}} \right), \quad (7)$$

$$(a_{22}, b_{22}) = \left( 2\sqrt{\frac{h\omega_0}{2} + \epsilon} \sqrt{\alpha + \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}}, -2\sqrt{\frac{h\omega_0}{2} - \epsilon} \sqrt{\alpha + \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}} \right). \quad (8)$$

To determine the type of equilibrium points we calculate eigenvalues for these equilibrium points. The eigenvalues are given by the Jacobi matrix of the right-hand sides of (1), (2):

$$J = \begin{pmatrix} \frac{\alpha}{2} - \frac{\beta}{8}(3a^2 + b^2) & -\frac{1}{2}\left(\epsilon + \frac{h\omega_0}{2}\right) - \frac{ab\beta}{4} \\ \frac{1}{2}\left(\epsilon - \frac{h\omega_0}{2}\right) - \frac{ab\beta}{4} & \frac{\alpha}{2} - \frac{\beta}{8}(3b^2 + a^2) \end{pmatrix}. \quad (9)$$

First, we investigate the type of the equilibrium point  $(0, 0)$ . From the Jacobi matrix of the right-hand sides of (1), (2) we get the characteristic equation

$$\lambda^2 - \alpha\lambda + \frac{\alpha^2}{4} + \frac{1}{4}\left(\epsilon^2 - \frac{h^2\omega_0^2}{4}\right) = 0. \quad (10)$$

This quadratic equation has a discriminant  $D = \frac{h^2\omega_0^2}{4} - \epsilon^2$  which is, with respect to (4), always positive. The eigenvalues of the equilibrium point  $(0, 0)$  are

$$\lambda_{01,02} = \frac{\alpha \pm \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}}{2}. \quad (11)$$

The assumption (4) implies that the eigenvalues are positive and the origin is an unstable improper node.

For the equilibrium points  $(a_{11}, b_{11})$  and  $(a_{12}, b_{12})$  we get the characteristic equation

$$\lambda^2 + \left(\alpha - 2\sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}\right)\lambda - \alpha\sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2} + \frac{h^2\omega_0^2}{4} - \epsilon^2 = 0. \quad (12)$$

The discriminant of this quadratic equation is  $D = \alpha^2 > 0$ , so this equation has always real roots. The eigenvalues of the equilibrium points  $(a_{11}, b_{11})$  and  $(a_{12}, b_{12})$  are

$$\lambda_{11} = \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}, \quad \lambda_{12} = -\alpha + \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}. \quad (13)$$

The assumption (4) for the existence of equilibrium points then implies that  $\lambda_1 > 0$  and  $\lambda_2 < 0$  and therefore the points  $(a_{11}, b_{11})$  and  $(a_{12}, b_{12})$  are equilibrium points of saddle type.

If we form the characteristic equation for the points  $(a_{21}, b_{21})$  and  $(a_{22}, b_{22})$ , we obtain

$$\lambda^2 + \left(\alpha + 2\sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}\right)\lambda + \alpha\sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2} + \frac{h^2\omega_0^2}{4} - \epsilon^2 = 0. \quad (14)$$

This quadratic equation has again a positive discriminant  $D = \alpha^2$ , so eigenvalues are always the real values

$$\lambda_{21} = -\sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}, \quad \lambda_{22} = -\alpha - \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}. \quad (15)$$

From assumption (4) it is clear that both eigenvalues are negative and therefore the equilibrium points  $(a_{21}, b_{21})$  and  $(a_{22}, b_{22})$  are the stable improper nodes.

The Figures 1, 2 show vector fields with nullclines of the autonomous system of equations (1), (2) for specific values of the parameters. These vector fields with nullclines confirm the existence of stable nodes in 2nd and 4th quadrant and the saddles in 1st and 3rd quadrant.

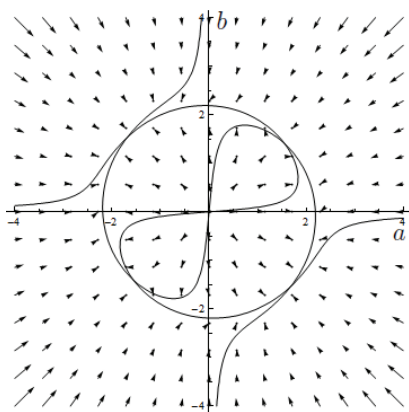


Figure 1: Vector field with nullclines for  $\alpha = 1, 2, \beta = 1, \epsilon = 0,01, h = 0,05, \omega_0 = 4$ .

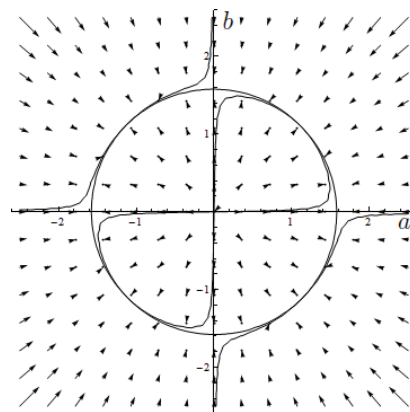


Figure 2: Vector field with nullclines for  $\alpha = 0,5, \beta = 0,8, \epsilon = 0,0001, h = 0,01, \omega_0 = 1,5$ .

**2.** Now we assume that only

$$|\epsilon| < \frac{h\omega_0}{2} \quad (16)$$

is satisfied and the second condition is not met. We see that the system (1), (2) has trivial equilibrium  $(0, 0)$  and two nontrivial equilibria  $(a_{21}, b_{21}), (a_{22}, b_{22})$ . The trivial equilibrium is saddle point for  $\alpha < \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}$  and nonhyperbolic equilibrium having a character of a saddle for  $\alpha = \sqrt{\frac{h^2\omega_0^2}{4} - \epsilon^2}$ . The nontrivial equilibria are stable nodes (positive attractors). The Figure 3 show two stable nodes (nontrivial equilibria) and saddle point  $(0, 0)$ .

**3.** Now we assume that the condition (16) is not met. The system (1), (2) has only trivial equilibrium for  $|\epsilon| > \frac{h\omega_0}{2}$  and this equilibrium is unstable focus (Figure 4). For  $|\epsilon| = \frac{h\omega_0}{2}$  the autonomous system (1), (2) has unstable node in  $(0, 0)$  and two nonhyperbolic equilibria.

The autonomous system (1), (2) with only the trivial equilibrium unstable focus investigated F. Veerman and F. Verhulst in [5]. They demonstrated that solution of Van der Pol-Mathieu equation from [3] exhibits quasiperiodic behavior.

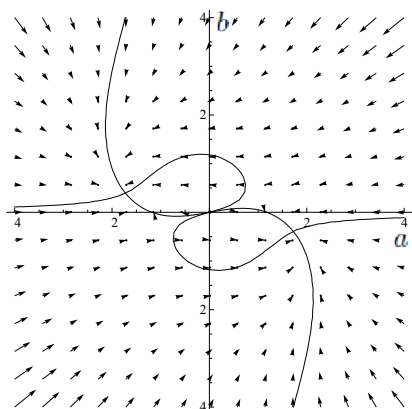


Figure 3: Vector field with nullclines for  $\alpha = 0.35, \beta = 1, \epsilon = 0.9, h = 2, \omega_0 = 1$ .

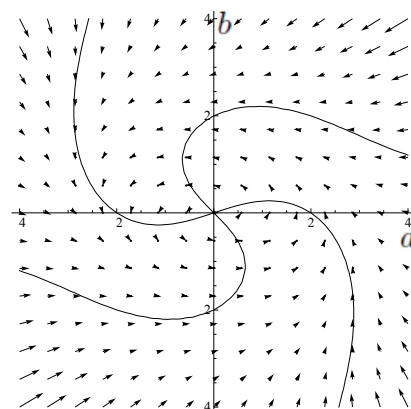


Figure 4: Vector field with nullclines for  $\alpha = 0.1, \beta = 0.1, \epsilon = 0.2, h = 0.2, \omega_0 = 1$ .

#### 4 The attracting set

Now we try to find how large region of the plane will be attracted to the equilibrium point  $(a_{21}, b_{21})$  or  $(a_{22}, b_{22})$ . We assume the condition (16).

**Theorem 4.1** Consider some trajectory  $x(t)$  of the autonomous system (1), (2). If the trajectory  $x(t)$  starts in the set  $X_1 = \{(a, b) : a \geq 0 \wedge b \leq 0\}$ , then the trajectory will not leave the set  $X_1$  and will be attracted to the point  $(a_{22}, b_{22})$ .

**Proof.** The considered set  $X_1$  is the 4th quadrant of the Cartesian coordinate system which is bounded by straight lines  $b = 0$  and  $a = 0$ . Now we show what happens to the trajectory  $x(t)$  which begins near the line  $b = 0$ .

Let  $x(t)$  be an arbitrary trajectory starting at the line  $b = 0$ . Since the trajectory  $x(t)$  corresponds to a solution of the system of equations (1), (2), we obtain, by putting  $b = 0$ , the following equations:

$$\frac{da}{dt} = \frac{\alpha}{2}a - \frac{\beta}{8}a^3 = \frac{a}{2} \left( \alpha - \frac{a^2\beta}{4} \right), \quad (17)$$

$$\frac{db}{dt} = \frac{a}{2} \left( \epsilon - \frac{h\omega_0}{2} \right). \quad (18)$$

The assumptions  $\epsilon < \frac{h\omega_0}{2}$  and  $a > 0$  imply that  $\frac{db}{dt} < 0$ . For this reason we see that the second coordinate of the point of the considered trajectory is decreasing.

Derivation  $\frac{da}{dt}$  is determined by the polynomial of degree 3 with respect to the variable  $a$ . This polynomial equals zero for values  $a = 0$ ,  $a = 2\sqrt{\frac{\alpha}{\beta}}$ ,  $a = -2\sqrt{\frac{\alpha}{\beta}}$ . If  $a \in \left(0, 2\sqrt{\frac{\alpha}{\beta}}\right)$  then the inequality  $\frac{da}{dt} > 0$  holds and for  $a \in \left(2\sqrt{\frac{\alpha}{\beta}}, \infty\right)$  we have  $\frac{da}{dt} < 0$ . It follows that the trajectory  $x(t)$  starting at the points of a half line  $b = 0, a \in (0, \infty)$ , is directed towards the

node  $(a_{22}, b_{22})$ .

Analogously we get:

$$\frac{da}{dt} = -\frac{b}{2} \left( \epsilon + \frac{h\omega_0}{2} \right) > 0, \quad (19)$$

$$\frac{db}{dt} = \frac{\alpha}{2}b - \frac{\beta}{8}b^3 = \frac{b}{2} \left( \alpha - \frac{b^2\beta}{4} \right) \quad (20)$$

for the trajectory near the line  $a = 0$ . Again we get that the trajectory cannot leave the set  $X_1$  because the first coordinates of points on a half line  $a = 0$  for  $b \in (-\infty, 0)$  increase.

For the derivation  $\frac{db}{dt}$  we find the points where the derivation changes sign. For values  $b \in \left(0, -2\sqrt{\frac{\alpha}{\beta}}\right)$  the inequality  $\frac{db}{dt} < 0$  holds whereas for  $a \in \left(-2\sqrt{\frac{\alpha}{\beta}}, -\infty\right)$  we have  $\frac{db}{dt} > 0$ .

This information about the behavior of trajectories near the boundary lines of the set  $X_1$  is shown in Figure 5. One can see that every trajectory which starts near the boundary lines of the 4th quadrant  $a = 0$  and  $b = 0$ , will be directed towards an equilibrium point  $(a_{22}, b_{22})$ .

Now we prove that every trajectory  $x(t)$  starting in the set  $X_1$  approaches the equilibrium

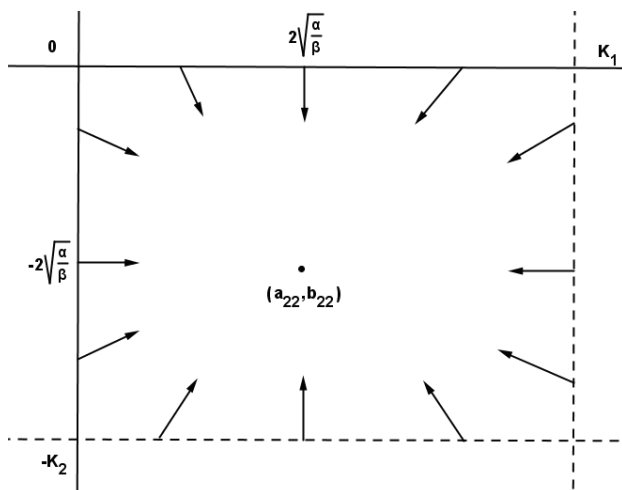


Figure 5: Directional field in the 4th quadrant.

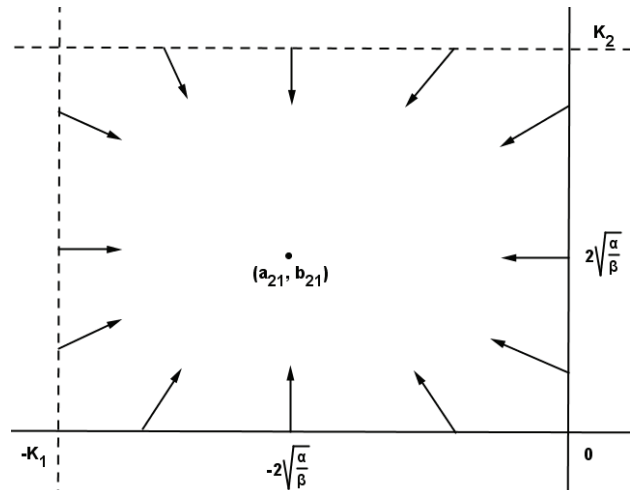


Figure 6: Directional field in the 2nd quadrant.

point  $(a_{22}, b_{22})$ .

We will show that any trajectory of (1), (2) starting in the 4th quadrant is bounded and does not leave this quadrant with increasing  $t$ . Indeed, let  $K_1 > 0$  and  $K_2 > 0$  be sufficiently large real numbers and  $X'_1 = \{(a, b) : 0 \leq a < K_1 \wedge -K_2 \leq b < 0\}$ . If  $K_1$  is substituted into (1) we find that for the points near the line  $a = K_1$  the inequality  $\frac{da}{dt} < 0$  holds, for large  $K_1$ . Similarly for the points near the line  $b = -K_2$  the inequality  $\frac{db}{dt} > 0$  is true. Therefore using the information from Figure 5 we obtain that every trajectory going out from an arbitrary point of the 4th quadrant will remain in  $X'_1$  with increasing time  $t$  for sufficiently large  $K_1, K_2$ .

To complete the proof, we prove that all trajectories of the 4th quadrant come close just to the equilibrium point  $(a_{22}, b_{22})$  with increasing time. Let us have any positive half trajectory  $x(t)$  in the 4th quadrant. Suppose that  $\omega$ -limit set of the trajectory  $x(t)$  contains no equilibrium point. Then, according to the Poincaré - Bendixson's theorem the  $\omega$ -limit set contains a closed

trajectory. Reversely, according to the Dulac's criterion the autonomous system contains no closed trajectory in a simply connected set, if the expression

$$g(a, b) = \frac{\partial}{\partial a} \left[ \frac{1}{ab} \cdot \left( \frac{\alpha}{2}a - \frac{b}{2} \left( \epsilon + \frac{h\omega_0}{2} \right) - \frac{\beta}{8}(a^3 + ab^2) \right) \right] + \quad (21)$$

$$+ \frac{\partial}{\partial b} \left[ \frac{1}{ab} \cdot \left( \frac{\alpha}{2}b + \frac{a}{2} \left( \epsilon - \frac{h\omega_0}{2} \right) - \frac{\beta}{8}(b^3 + a^2b) \right) \right] \quad (22)$$

in this set is still positive or negative. This requirement is true for our autonomous system (1), (2) in the 4th quadrant, since

$$g(a, b) = -\frac{\beta}{4} \left( \frac{a}{b} + \frac{b}{a} \right) - \frac{1}{2b^2} \left( \epsilon - \frac{h\omega_0}{2} \right) + \frac{1}{2a^2} \left( \epsilon + \frac{h\omega_0}{2} \right) \quad (23)$$

is positive in this quadrant. Therefore any closed trajectory cannot exist in the 4th quadrant and the  $\omega$ -limit set of any positive half trajectory going out from any point of the 4th quadrant cannot contain a closed trajectory. The case, that it would contain no equilibrium point, is impossible. Therefore the  $\omega$ -limit set of the positive half trajectory  $x(t)$  include at least one equilibrium point. In the fourth quadrant there are only two equilibrium points  $(0, 0)$  and  $(a_{22}, b_{22})$ , origin is an unstable node and the point  $(a_{22}, b_{22})$  is a stable node. Hence the  $\omega$ -limit set of some positive half trajectory consists of the unique point  $(a_{22}, b_{22})$ . For that reason every trajectory beginning in the fourth quadrant comes with increasing time just to a stable node  $(a_{22}, b_{22})$ .

From the symmetry of the autonomous system (1), (2), the following Theorem 4.2 for equilibrium point  $(a_{21}, b_{21})$  which can be proved analogously as Theorem 4.1 is true. The directional field in the 2nd quadrant is reported in Figure 6.

**Theorem 4.2** Consider some trajectory  $x(t)$  of the autonomous system of the equations (1), (2). If the trajectory  $x(t)$  begins in the set  $X_2 = \{(a, b) : a \leq 0 \wedge b \geq 0\}$ , then this trajectory will not leave the set  $X_2$  and will be attracted to the equilibrium point  $(a_{21}, b_{21})$ .

Before we formulate a main theorem we prove a following lemma about a boundedness of every trajectory  $x(t)$ .

**Lemma 4.3** Consider some trajectory  $x(t)$  of the autonomous system (1), (2). If the trajectory  $x(t)$  begins in the set  $X_3 = \{(a, b) : a \in \langle -M, M \rangle, b \in \langle -kM, kM \rangle\}$ , where  $k = \sqrt{\frac{\frac{h\omega_0}{2} - \epsilon}{\frac{h\omega_0}{2} + \epsilon}}$  and  $M \gg 0$ , then this trajectory will not leave the set  $X_3$ .

**Proof.** The set  $X_3$  is bounded by four straight lines: 1.  $a = M, b \in \langle -kM, kM \rangle$ ; 2.  $a = -M, b \in \langle -kM, kM \rangle$ ; 3.  $b = kM, a \in \langle -M, M \rangle$  and 4.  $b = -kM, a \in \langle -M, M \rangle$ . These lines are shown in Figure 7. We shall prove what happens to the trajectory  $x(t)$  which begins at the point of these lines. We shall examine the signs of  $\frac{da}{dt}$  or  $\frac{db}{dt}$ .

1. For  $a = M, b \in \langle -kM, kM \rangle$  we have

$$\frac{da}{dt} = \frac{\alpha}{2}M - \frac{b}{2} \left( \frac{h\omega_0}{2} + \epsilon \right) - \frac{\beta}{8}(M^3 + Mb^2) \leq \frac{M}{8} \left( 4\alpha + 4k \left( \frac{h\omega_0}{2} + \epsilon \right) - \beta M^2 \right).$$

This expression is still negative for  $M > 2\sqrt{\frac{\alpha + \sqrt{(\frac{h\omega_0}{2} + \epsilon)(\frac{h\omega_0}{2} - \epsilon)}}{\beta}}$ .

2. For  $a = -M, b \in \langle -kM, kM \rangle$  it is true that

$$\frac{da}{dt} = -\frac{\alpha}{2}M - \frac{b}{2}\left(\frac{h\omega_0}{2} + \epsilon\right) - \frac{\beta}{8}(-M^3 - Mb^2) \geq -\frac{M}{8}\left(4\alpha + 4k\left(\frac{h\omega_0}{2} + \epsilon\right) - \beta M^2\right).$$

This expression is still positive for  $M > 2\sqrt{\frac{\alpha + \sqrt{(\frac{h\omega_0}{2} + \epsilon)(\frac{h\omega_0}{2} - \epsilon)}}{\beta}}$ .

3. Using the equation (2) we obtain for  $b = kM, a \in \langle -M, M \rangle$ , that

$$\frac{db}{dt} = \frac{\alpha}{2}kM + \frac{a}{2}\left(\epsilon - \frac{h\omega_0}{2}\right) - \frac{\beta}{8}(k^3M^3 + kMa^2) \leq \frac{M}{8}\left(4\alpha k - 4\left(\epsilon - \frac{h\omega_0}{2}\right) - \beta k^3M^2\right).$$

If  $M > 2\sqrt{\frac{\alpha k - (\epsilon - \frac{h\omega_0}{2})}{\beta k^3}}$  then this expression is still negative.

4. For  $b = -kM, a \in \langle -M, M \rangle$  we have analogously

$$\frac{db}{dt} = -\frac{\alpha}{2}kM + \frac{a}{2}\left(\epsilon - \frac{h\omega_0}{2}\right) - \frac{\beta}{8}(-k^3M^3 + kMa^2) \geq -\frac{M}{8}\left(4\alpha k - 4\left(\epsilon - \frac{h\omega_0}{2}\right) - \beta k^3M^2\right).$$

This expression is still positive for  $M > 2\sqrt{\frac{\alpha k - (\epsilon - \frac{h\omega_0}{2})}{\beta k^3}}$ .

Figure 7 shows directional field in the set  $X_3$ . If the trajectory  $x(t)$  starts at the point of the line  $a = M, b \in \langle -kM, kM \rangle$ , then  $\frac{da}{dt} < 0$  for large  $M$ . First coordinate decrease therefore this trajectory is directed into the set  $X_3$ . For the other boundary lines from the steps 2. – 4. we obtain analogously that the trajectory  $x(t)$  will be still in the set  $X_3$ . If the trajectory  $x(t)$  starts at the point  $(M, kM)$  (similarly for  $(-M, kM)$ ,  $(M, -kM)$  or  $(-M, -kM)$ ) then  $\frac{da}{dt} < 0$  and  $\frac{db}{dt} < 0$ . Therefore this trajectory does not leave the set  $X_3$ .

From these four steps it follows that every trajectory starting in the set  $X_3$  does not leave this set  $X_3$  for large  $M$ .

We want prove more general theorems than Theorem 4.1 and 4.2. Therefore we transform coordinates  $a, b$  to polar coordinates  $\rho, \varphi$ :

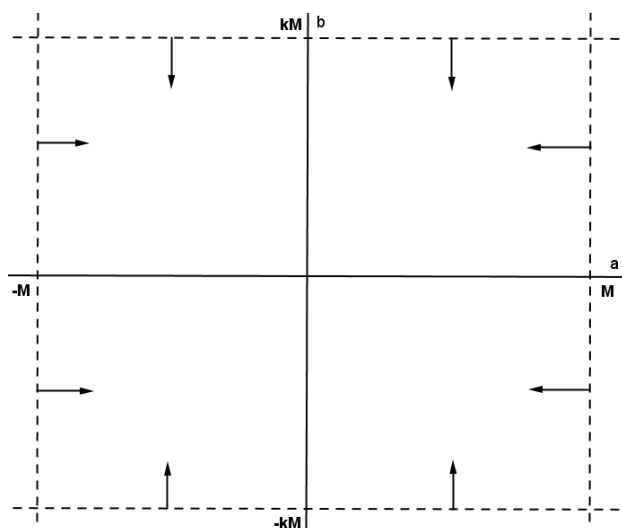
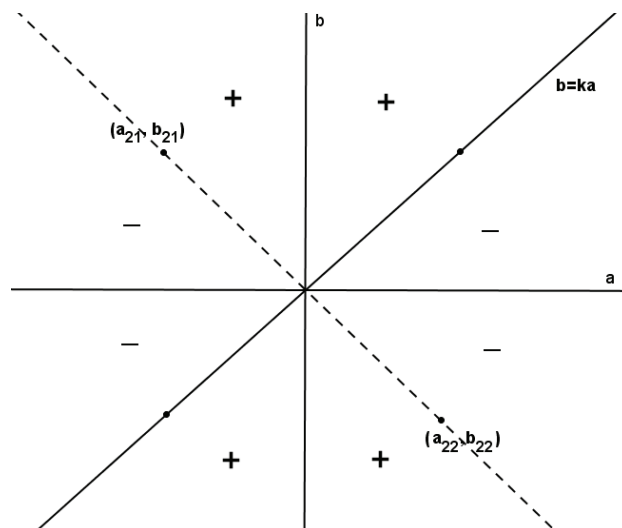
$$\rho(t) = \sqrt{a^2(t) + b^2(t)}, \quad \cos \varphi(t) = \frac{a(t)}{\sqrt{a^2(t) + b^2(t)}}, \quad \sin \varphi(t) = \frac{b(t)}{\sqrt{a^2(t) + b^2(t)}}. \quad (24)$$

Using the formulas (24) we transform the system (1), (2) into a form

$$\frac{d\varphi}{dt} = \frac{1}{2}\left(\epsilon - \frac{h\omega_0}{2}\right)\cos^2\varphi + \frac{1}{2}\left(\epsilon + \frac{h\omega_0}{2}\right)\sin^2\varphi, \quad (25)$$

$$\frac{d\rho}{dt} = \frac{\alpha}{2}\rho - \frac{h\omega_0}{4}\rho\sin(2\varphi) - \frac{\beta}{8}\rho^3. \quad (26)$$

If we go back to the coordinates of equilibrium points (5), (6), (7) and (8) then we find that


Figure 7: Directional field in the set  $X_3$ .

Figure 8: The sets  $Y_1$  and  $Y_2$  with signs of the derivation  $\varphi'$ .

these coordinates comply with  $b = \pm \sqrt{\frac{h\omega_0}{2} - \epsilon} \cdot a$ . All non-trivial equilibrium points belong to these two lines. It follows that the points  $(a_{21}, b_{21})$ ,  $(a_{22}, b_{22})$  belong to the line with a slope  $-k = -\sqrt{\frac{h\omega_0}{2} - \epsilon}$ .

Now we formulate the main theorems of this section. They describe the largest region of the plane which is attracted to the equilibrium point  $(a_{21}, b_{21})$  or  $(a_{22}, b_{22})$ .

**Theorem 4.4** Suppose that  $x(t)$  is an arbitrary trajectory of the system (1), (2). If the trajectory  $x(t)$  starts in the set  $Y_1 = \{(a, b) : b < k \cdot a\}$ , where  $k = \sqrt{\frac{h\omega_0}{2} - \epsilon}$ , then this trajectory will not leave this set and will be attracted to the point  $(a_{22}, b_{22})$ .

**Proof.** Let's assume that  $x(0)$  is located in the set  $Y_1$  and  $M > 0$  is sufficiently large to be valid Lemma 4.3. Suppose an arbitrary trajectory  $x(t) = (\rho(t) \cos t, \varphi(t) \sin t)$  starting in the set  $Y_1$ ,  $\varphi(0) \in (\arctg k - \pi, \arctg k)$ .

The Lemma 4.3 implies that all trajectories from the set  $X_3$  will remain bounded with increasing time in this set. Besides the set  $Y_1$  is bounded by the line  $b = k \cdot a$ . The equilibria  $(a_{11}, b_{11})$ ,  $(a_{12}, b_{12})$  belong to this line and for the polar angle  $\varphi(t)$  of the point of this line  $\frac{d\varphi}{dt} = 0$  is satisfied. It implies that the set  $P_1 = \{(a, b) : b = \sqrt{\frac{h\omega_0}{2} - \epsilon} \cdot a\}$  is invariant set of the autonomous system (1), (2). Therefore these informations imply that all trajectories starting in the set  $Y_1$  do not leave this set.

We shall prove now that every trajectory in the set  $Y_1$  is directed to the line with a slope  $-k = -\sqrt{\frac{h\omega_0}{2} - \epsilon}$ . Obviously, if the polar angle  $\varphi(t)$  approaches the angle  $\arctg(-k)$  then the trajectory  $x(t)$  goes to the line with a slope  $-k$ .

The set  $Y_1$  corresponds the polar angle  $\varphi \in (\arctg k - \pi, \arctg k)$ . Using (25) we have

$\frac{d\varphi}{dt} > 0$  if  $\operatorname{tg}^2 \varphi > \frac{\frac{h\omega_0}{2} - \epsilon}{\frac{h\omega_0}{2} + \epsilon}$ . It follows that  $\frac{d\varphi}{dt} > 0$  is satisfied for  $\varphi \in (\arctg k - \pi, -\arctg k)$ . The polar angle  $\varphi(t)$  of the trajectory  $x(t)$  with  $\varphi(0) \in (\arctg k - \pi, -\arctg k)$  will be increase and this trajectory will be directed to the line with a slope  $-k$  where  $\frac{d\varphi}{dt} = 0$ .

Analogously we have  $\frac{d\varphi}{dt} < 0$  for  $\varphi \in (-\arctg k, \arctg k)$  and this polar angle  $\varphi(t)$  will decrease to  $\varphi = \arctg(-k)$ . The informations which we proved are shown in Figure 8.

Now we know that every trajectory  $x(t)$  starting in the set  $Y_1$  is directed to the line with a slope  $-k$ . It remains to prove that every trajectory beginning in the set  $Y_1$  is attracted to the point  $(a_{22}, b_{22})$ .

From this proof we know that the trajectories from  $Y_1$  will go into the set  $X_1$  with increasing time and will not leave this set  $X_1$ . Therefore the preceding Theorem 4.1 implies that  $\omega$ -limit set consists of only the stable improper node  $(a_{22}, b_{22})$ .

It's clear from this proof that every trajectory  $x(t)$  starting in the set  $Y_1$  remains in this set and it is attracted to the node  $(a_{22}, b_{22})$  with increasing time.

Analogously we can prove the following theorem. The set  $Y_2$  and a sign of a derivation  $\varphi'$  are shown in Figure 8.

**Theorem 4.5** *Assume that  $x(t)$  is an arbitrary trajectory of the system (1), (2). If the trajectory  $x(t)$  starts in the set  $Y_2 = \{(a, b) : b > k \cdot a\}$ , where  $k$  is defined as in Lemma 4.3, then this trajectory will not leave this set and will be attracted to the point  $(a_{21}, b_{21})$ .*

## 5 Conclusion

The previous section supplements the article [5] and shows that all trajectories of (1), (2) in the set  $Y_1$  approach the node  $(a_{22}, b_{22})$ . The trajectories located in the set  $Y_2$  arrive with increasing time to the node  $(a_{21}, b_{21})$ . The trajectories starting at a half line  $b = \sqrt{\frac{\frac{h\omega_0}{2} - \epsilon}{\frac{h\omega_0}{2} + \epsilon}} \cdot a$  for  $a > 0$  go to the point  $(a_{11}, b_{11})$ , which is the type of saddle. The trajectory of the opposite half line, i.e.  $b = \sqrt{\frac{\frac{h\omega_0}{2} - \epsilon}{\frac{h\omega_0}{2} + \epsilon}} \cdot a$  for  $a < 0$ , goes to the saddle  $(a_{12}, b_{12})$ .

The derived results show that the coefficients in the estimated solution (3) stabilize with increasing time  $t$  at values corresponding to the equilibrium points of the autonomous system of equations (1), (2).

## Acknowledgement

The research was supported by the grant MUNI/A/0964/2009 of Masaryk University and by the grant GAP201/11/0769 of the Czech Science Foundation.

## References

- [1] CODDINGTON E.A., LEVINSON N.: *Theory of ordinary differential equations*. Mc-Graw-Hill Book Co., New York, 1955.

- [2] LAWRENCE P. *Differential Equations and Dynamical Systems*. Springer - Verlag Berlin Heidelberg New York, 2nd edition, 1996.
- [3] MOMENI M., KOURAKIS I., MOSLEHI-FRAD M., SHUKLA P. K. *A Van der Pol-Mathieu equation for the dynamics of dust grain charge in dusty plasmas*. Journal of Physics A: Mathematical and Theoretical, Issue 40, 473-481, 2007.
- [4] SANDERS J.A., VERHULST F. *Averaging methods in nonlinear dynamical systems*. Appl Math. Sciences 59, Springer - Verlag, New York, 1985.
- [5] VEERMAN F., VERHULST F. *Quasiperiodic phenomena in the Van der Pol - Mathieu equation*. Journal of Sound and Vibration, Vol. 326, Issues 1-2, 314-320, 2009.
- [6] VERHULST F. *Nonlinear Differential Equations and Dynamical Systems*. Springer - Verlag Berlin Heidelberg New York, 2nd edition, 1996.
- [7] WIGGINS S. *Introduction to applied nonlinear dynamical systems and chaos*. Springer - Verlag Berlin Heidelberg New York, 2nd edition, 2003.
- [8] YOSHIZAWA T. *Stability Theory and the Existence of Periodic Solutions and Almost Periodic Solutions*. Springer - Verlag Berlin Heidelberg New York, Applied Mathematical Sciences, Volume 14, 1975.

#### Current address

**Zdeněk Kadeřábek, Mgr.**

Faculty of Science - Department of Mathematics and Statistics,  
Masaryk university, Brno, Kotlářská 2, 611 37, Czech Republic,  
email: 151353@mail.muni.cz



# SINGULAR INITIAL VALUE PROBLEM FOR IMPLICIT VOLTERRA INTEGRO-DIFFERENTIAL EQUATIONS

KHAN Yasir, (CH), ŠMARDA Zdeněk, (CZ)

**Abstract.** Singular initial value problem for implicit Volterra integro-differential equations depending on a parameter and continuous dependence of solutions are the subject of the paper. The existence and uniqueness of a solution is proved using Banach contraction principle. Obtained results are illustrated with an example.

**Key words and phrases.** Singular initial value problem, Banach contraction principle.

*Mathematics Subject Classification.* Primary 45J05; Secondary 34A08.

## 1 Introduction

In the past three decades, singular initial value problems for differential and integro-differential equations have been studied under various conditions on the nonlinearity and the kernel (see e.g.[1-17]). The fundamental tools used in the existence proofs of all mentioned works are essentially Schauder-Tychonoff's fixed point theorem, Banach contraction principle and Wazewski's topological method.

In cases of integro-differential equations of Fredholm type it is necessary to use Lipschitz constants with weighted exponential functions (see [12], [13]). In the paper we obtain for Volterra type of singular implicit integrodifferential equations similar results as in above mentioned papers but using only Lipschitz constants without weighted exponential functions.

Consider the following singular initial value problem

$$y'(t) = \mathcal{F} \left( t, y(t), y'(t), \int_0^t K(t, s, y(s), y'(s)) ds, \mu \right), \quad y^{(i)}(0^+, \mu) = 0, \quad i = 0, 1, \quad (1)$$

where

(I)  $\mathcal{F} : \Omega \rightarrow \mathbb{R}^n$ ,  $\mathcal{F} \in C^0(\Omega)$ ,  $\Omega = \{(t, u_1, u_2, u_3, \mu) \in J \times (\mathbb{R}^n)^3 \times \mathbb{R} : |u_1| \leq \phi_1(t), |u_2| \leq \phi_2(t), |u_3(t)| \leq \psi(t)\}$ ,  $J = (0, t_0]$ ,  $0 < t_0 < 1$ ,  $0 < \phi_i(t) \in C^0(J)$ ,  $i = 1, 2$ ,  $\phi_1(0^+) = 0$ ,  $0 < \psi(t) \in C^0(J)$ ,  $\int_0^t \phi_2(s)ds \leq \phi_1(t)$ ,  $|\cdot|$  denotes the usual norm in  $\mathbb{R}^n$ ,  $|\mathcal{F}(t, \bar{u}_1, \bar{u}_2, \bar{u}_3, \mu) - \mathcal{F}(t, \bar{\bar{u}}_1, \bar{\bar{u}}_2, \bar{\bar{u}}_3, \mu)| \leq \sum_{i=1}^3 M_i |\bar{u}_i - \bar{\bar{u}}_i|$  for all  $(t, \bar{u}_1, \bar{u}_2, \bar{u}_3, \mu), (t, \bar{\bar{u}}_1, \bar{\bar{u}}_2, \bar{\bar{u}}_3, \mu) \in \Omega$ ,  $M_1, M_3 > 0$ ,  $0 < M_2 < 1$  are constants.

(II)  $K : \Omega^1 \rightarrow \mathbb{R}^n$ ,  $K \in C^0(\Omega^1)$ ,  $\Omega^1 = \{(t, s, v_1, v_2) \in J \times J \times (\mathbb{R}^n)^2 : |v_1| \leq \phi_1(t), |v_2| \leq \phi_2(t)\}$ ,  $|K(t, s, \bar{v}_1, \bar{v}_2) - K(t, s, \bar{\bar{v}}_1, \bar{\bar{v}}_2)| \leq \sum_{j=1}^2 N_j |\bar{v}_j - \bar{\bar{v}}_j|$  for all  $(t, s, \bar{v}_1, \bar{v}_2), (t, s, \bar{\bar{v}}_1, \bar{\bar{v}}_2) \in \Omega_1$ ,  $N_j > 0, j = 1, 2$  are constants. There is a sufficiently large constant  $\lambda > 0$  such that

$$\left( M_2 + \frac{M_1 + M_3 N_2}{\lambda} + \frac{M_3 N_1}{\lambda^2} \right) < 1.$$

## 2 Main results

**Theorem 2.1.** *Let the functions  $\mathcal{F}(t, u_1, u_2, u_3, \mu)$ ,  $K(t, s, v_1, v_2)$  satisfy conditions (I), (II) and furthermore*

$$|\mathcal{F}| \leq \sum_{i=1}^3 g_i(t) |u_i|, \quad 0 < g_i(t) \in C^0(J), \quad i = 1, 2, 3, \quad \int_0^t g_1(s) \phi_2(s) ds \leq \alpha_1 \phi_2(t),$$

$$g_2(t) \phi_2(t) \leq \alpha_2 \phi_2(t) \quad g_3(t) \psi(t) \leq \alpha_3 \phi_2(t), \quad \alpha_j \in \mathbb{R}^+, \quad \sum_{j=1}^3 \alpha_j < 1,$$

then the initial problem (1) has a unique solution  $y(t, \mu)$  for each  $\mu \in \mathbb{R}$ ,  $t \in J$ .

**Proof.** Put

$$y(t) = \int_0^t r(s) ds,$$

where  $r(t) \in C^0(J)$  is an unknown function. Then  $y'(t) = r(t)$  and the system (1) is equivalent to the system of integral equations

$$r(t) = \mathcal{F} \left( t, \int_0^t r(s) ds, r(t), \int_0^t K(t, s, \int_0^s r(\tau) d\tau, r(s)) ds, \mu \right). \quad (2)$$

Denote  $H$  the Banach space of continuous vector-valued functions  $h : J_0 \rightarrow \mathbb{R}^n$ ,  $J_0 = [0, t_0]$ ,  $|h(t)| \leq \phi_2(t)$  for each  $t \in J_0$  with the norm

$$\|h(t)\|_\lambda = \max_{t \in J_0} \{e^{-\lambda t} |h(t)|\},$$

where  $\lambda > 0$  is an arbitrary parameter. Define the operator  $T$  by right- hand side of (2)

$$T(h) = \mathcal{F} \left( t, \int_0^t h(s) ds, h(t), \int_0^t K(t, s, \int_0^s h(\tau) d\tau, h(s)) ds, \mu \right),$$

where  $h \in H$ . Let  $\mu \in R$  be fixed. The transformation  $T$  maps  $H$  continuously into itself because

$$\begin{aligned} |T(h)| &\leq \left| \mathcal{F} \left( t, \int_0^t h(s)ds, h(t), \int_0^t K(t, s, \int_0^s h(\tau)d\tau, h(s))ds, \mu \right) \right| \\ &\leq g_1(t) \left| \int_0^t h(s)ds \right| + g_2(t)|h(t)| + g_3(t) \left| \int_0^t K(t, s, \int_0^s h(\tau)d\tau, h(s))ds \right| \\ &\leq g_1(t) \int_0^t \phi_2(s)ds + \alpha_2 \phi_2(t) + g_3(t)\psi(t) \leq \phi_2(t) \end{aligned}$$

for every  $h \in H$ . Now, we prove that

$$\|T(h_2) - T(h_1)\|_\lambda \leq \left( M_2 + \frac{M_1 + M_3 N_2}{\lambda} + \frac{M_3 N_1}{\lambda^2} \right) \|h_2 - h_1\|_\lambda, \quad (3)$$

for all  $h_1, h_2 \in H$ . Using (I),(II) and the definition  $\|\cdot\|_\lambda$  we have

$$\begin{aligned} |T(h_2) - T(h_1)| &= \\ &\left| \mathcal{F} \left( t, \int_0^t h_2(s)ds, h_2(t), \int_0^t K(t, s, \int_0^s h_2(\tau)d\tau, h_2(s))ds, \mu \right) \right. \\ &\quad \left. - \mathcal{F} \left( t, \int_0^t h_1(s)ds, h_1(t), \int_0^t K(t, s, \int_0^s h_1(\tau)d\tau, h_1(s))ds, \mu \right) \right| \\ &\leq M_1 \int_0^t |h_2(s) - h_1(s)|ds + M_2 |h_2(t) - h_1(t)| \\ &\quad + M_3 \int_0^t \left| K(t, s, \int_0^s h_2(\tau)d\tau, h_2(s)) - K(t, s, \int_0^s h_1(\tau)d\tau, h_1(s)) \right| ds \\ &\leq M_1 \|h_2 - h_1\|_\lambda \int_0^t e^{\lambda s} ds + M_2 e^{\lambda t} \|h_2 - h_1\|_\lambda + M_3 N_1 \int_0^t \int_0^s |h_2(\tau) - h_1(\tau)| d\tau ds \\ &\quad + M_3 N_2 \int_0^t |h_2(s) - h_1(s)| ds \\ &\leq \|h_2 - h_1\|_\lambda \left( M_1 \frac{e^{\lambda t} - 1}{\lambda} + M_2 e^{\lambda t} + M_3 N_1 \left( \frac{e^{\lambda t} - 1}{\lambda^2} - \frac{t}{\lambda} \right) + M_3 N_2 \frac{e^{\lambda t} - 1}{\lambda} \right) \\ &\leq e^{\lambda t} \|h_2 - h_1\|_\lambda \left( M_2 + \frac{M_1 + M_3 N_2}{\lambda} + \frac{M_3 N_1}{\lambda^2} \right) \end{aligned}$$

Thus

$$\|T(h_2) - T(h_1)\|_\lambda \leq q \|h_2 - h_1\|_\lambda,$$

where

$$q := M_2 + \frac{M_1 + M_3 N_2}{\lambda} + \frac{M_3 N_1}{\lambda^2} < 1.$$

Applying the classical Banach contraction principle to  $T$  and the distance function  $\|h_2 - h_1\|_\lambda$  we get the assertion of Theorem 2.1.

**Theorem 2.2.** *Let the assumptions of Theorem 2.1. be satisfied and there exist a constant  $L > 0$  and the integrable function  $\gamma : J_0 \rightarrow J_0$ ,  $J_0 = [0, t_0]$  such that*

$$|\mathcal{F}(t, u_1, u_2, u_3, \mu_2) - \mathcal{F}(t, u_1, u_2, u_3, \mu_1)| \leq \gamma(t)|\mu_2 - \mu_1|,$$

where  $((t, u_1, u_2, u_3, \mu_2), (t, u_1, u_2, u_3, \mu_1)) \in \Omega$  and

$$\max_{t \in J_0} \{e^{-\lambda t} \gamma(t)\} \leq L,$$

then the solution  $y(t, \mu)$  of (1) is continuous with respect to the variables  $(t, \mu) \in J \times \mathbb{R}$ .

**Proof.** Define as above, for  $h \in H$ , the transformation  $T_\mu(h)$  by means of the right-hand side (2). From (3) we obtain

$$\|T_\mu(h) - T_\mu(y)\|_\lambda \leq \left( M_2 + \frac{M_1 + M_3 N_2}{\lambda} + \frac{M_3 N_1}{\lambda^2} \right) \|h - y\|_\lambda.$$

By the assumptions of Theorem 2.2 we get

$$\begin{aligned} e^{-\lambda t} |T_{\mu_2}(h) - T_{\mu_1}(h)| &\leq \\ &\left| \mathcal{F} \left( t, \int_0^t h(s) ds, h(t), \int_0^t K(t, s, \int_0^s h(\tau) d\tau, h(s)) ds, \mu_2 \right) \right. \\ &\quad \left. - \mathcal{F} \left( t, \int_0^t h(s) ds, h(t), \int_0^t K(t, s, \int_0^s h(\tau) d\tau, h(s)) ds, \mu_1 \right) \right| \\ &\leq e^{-\lambda t} \gamma(t) |\mu_2 - \mu_1| \leq L |\mu_2 - \mu_1|. \end{aligned}$$

Hence

$$\|T_{\mu_2}(h) - T_{\mu_1}(h)\|_\lambda \leq L |\mu_2 - \mu_1|.$$

From here and by Theorem 2.1 we obtain

$$\begin{aligned} \|h(t, \mu_2) - h(t, \mu_1)\|_\lambda &= \|T_{\mu_2}[h(t, \mu_2)] - T_{\mu_2}[h(t, \mu_1)] + T_{\mu_2}[h(t, \mu_1)] - T_{\mu_1}[h(t, \mu_1)]\|_\lambda \\ &\leq \|T_{\mu_2}[h(t, \mu_2)] - T_{\mu_2}[h(t, \mu_1)]\|_\lambda + \|T_{\mu_2}[h(t, \mu_1)] - T_{\mu_1}[h(t, \mu_1)]\|_\lambda \\ &\leq \left( M_2 + \frac{M_1 + M_3 N_2}{\lambda} + \frac{M_3 N_1}{\lambda^2} \right) \|h(t, \mu_2) - h(t, \mu_1)\|_\lambda + L |\mu_2 - \mu_1|. \end{aligned}$$

Thus

$$\|h(t, \mu_2) - h(t, \mu_1)\|_\lambda \leq \left( 1 - \left( M_2 + \frac{M_1 + M_3 N_2}{\lambda} + \frac{M_3 N_1}{\lambda^2} \right) \right)^{-1} L |\mu_2 - \mu_1|.$$

Consequently  $y(t, \mu)$  is continuous with respect two variables  $(t, \mu) \in J \times \mathbb{R}$ , which completes the proof.

**Example.** Consider the following initial value problem

$$y'(t) - \frac{t}{3} y(t) + \frac{1}{10} \arctan \frac{1}{t} y'(t) + 2t^2 \int_0^t \sqrt{s} e^{-\frac{\mu}{ts}} (y(s) + 2y'(s)) ds, \quad y^{(i)}(0^+, \mu) = 0, \quad (4)$$

where  $t \in J$ ,  $i = 1, 2$ .

Put

$$M_1 = \frac{1}{3}, \quad M_2 = \frac{\pi}{20}, \quad M_3 = 2, \quad N_1 = 2, \quad N_2 = 4, \quad \lambda = 100,$$

then

$$q = M_2 + \frac{M_1 + M_3 N_2}{\lambda} + \frac{M_3 N_1}{\lambda^2} = \frac{\pi}{20} + \frac{1}{300} + \frac{8}{100} + \frac{4}{10000} < 1$$

and

$$|\mathcal{F}| \leq \frac{t}{3}|u_1| + \frac{1}{10} \arctan \frac{1}{t}|u_2| + |u_3| \Rightarrow g_1(t) = \frac{t}{3}, \quad g_2(t) = \frac{1}{10} \arctan \frac{1}{t}, \quad g_3(t) = 1.$$

Set  $\phi_1(t) = \frac{t^5}{4}$ ,  $\phi_2(t) = t^4$ ,  $\psi(t) = \frac{t^4}{10}$ . Thence

$$\int_0^t g_1(s)\phi_2(s)ds = \frac{t^6}{18} \leq \frac{1}{18}\phi_2(t) \Rightarrow \alpha_1 = \frac{1}{18},$$

$$g_2(t)\phi_2(t) = \frac{1}{10} \arctan \frac{1}{t}t^4 \leq \frac{\pi}{20}\phi_2(t) \Rightarrow \alpha_2 = \frac{\pi}{20},$$

$$g_3(t)\psi(t) = \frac{t^4}{10} = \frac{1}{10}\phi_2(t) \Rightarrow \alpha_3 = \frac{1}{10}.$$

Thus

$$\alpha_1 + \alpha_2 + \alpha_3 = \frac{1}{18} + \frac{\pi}{20} + \frac{1}{10} < 1.$$

By Theorem 2.1. there exists a unique solution of (4) such that

$$|y(t, \mu)| \leq \frac{t^5}{4}, \quad |y'(t, \mu)| \leq t^4.$$

## Acknowledgement

This research has been supported by the Grant FEKT-S-11-2-921 of Faculty of Electrical Engineering and Communication, Brno University of Technology.

## References

- [1] AGARWAL, R.P., O'REGAN, D., ZERNOV, A.E.: *A singular initial value problem for some functional differential equations*, J. Appl. Math. Stochastic Anal., 3, 2005, 261-270.
- [2] DIBLÍK, J., RUŽIČKOVÁ, M.: *Existence of positive solutions of a singular initial problem for a nonlinear system of differential equations*, Rocky Mountain Journal of Mathematics, 34, 2004, 923-944.
- [3] DIBLÍK, J., RUŽIČKOVÁ, M.: *Inequalities for solutions of singular initial problems for Carathodory systems via Ważewski's principle*, Nonlinear Anal., Theory Methods Appl. 69(12), 2008, 657-656.
- [4] DIBLÍK, J., NOWAK, C.: *A nonuniqueness criterion for a singular system of two ordinary differential equations*, Nonlinear Analysis, 64, 2006, 637-656. ISSN 0362-546X.

- [5] DYSHKO, A., KONYUCHOVA, N., SUKOV, A.: *Singular problem for a third-order nonlinear ordinary differential equation arising in fluid dynamics*, Computational Mathematics and Mathematical Physics, 47(7), 2007, 1108-1128. ISSN 0041-5553.
- [6] FILIPPOVA, O., ŠMARDÁ, Z.: *Singular initial value problem for Volterra integrodifferential equations*, In Proceedings of the First International Forum of Young Researchers. Izhevsk: design. Publishing House of ISTU, 2008, 331-336.
- [7] KONJUCHOVA, N.B.: *Singular Cauchy problems for systems of ordinary differential equations*, urnal vy. mat. i fiziki, 5, 1981, 629-645 (in Russian).
- [8] NORKIN, C.K.: *On asymptotic behavior and structure of the integral O-set of certain singular system*, Diff. uravnenija, 22, 1986, 228-236 (in Russian).
- [9] ŠMARDÁ, Z.: *On solutions of an implicit singular system of integrodifferential equations depending on a parameter*, Demonstratio Mathematica, 31(1), 1998, 125-130.
- [10] ŠMARDÁ, Z.: *On an initial value problem for singular integrodifferential equations*, Demonstratio Mathematica, 35(4), 2002, 803-811. ISSN 0420-1213.
- [11] ŠMARDÁ, Z.: *Existence and uniqueness of solutions of nonlinear integrodifferential equations*, Journal of Applied Mathematics, Statistics and Informatics, 1(2), 2005, 73-77. ISSN 1336-9180.
- [12] ŠMARDÁ, Z.: *On singular initial value problem for nonlinear Fredholm integrodifferential equations*, Fasciculi Mathematici, 1(37), 2006, 77-83. ISSN 00444413.
- [13] ŠMARDÁ, Z.: *Implicit singular integrodifferential equations of Fredholm type*, Tatra Mt. Math. Publ. 38, 2007, 255-263.
- [14] ZERNOV, Y.: *Geometric Analysis of one singular Cauchy problem*, Nonlinear Oscillations, vol. 7, No. 1, 2004, 65-77. ISSN 1562-3076.
- [15] ZERNOV, Y., KUZINA, Yu.V.: *Qualitative investigation of the singular Cauchy problem*, Ukr. Math. J. 55(10), 2003, 1419-1424 (in Russian).
- [16] ZERNOV, Y., KUZINA, Yu.V.: *Geometric analysis of one singular Cauchy problem*, Nonlinear Oscil., N.Y. 7(1), 2004, 67-80 (in Russian).
- [17] ZERNOV, Y., CHAICHUK, O.R.: *Asymptotic behavior of solutions of a singular Cauchy problem for a functional-differential equation*, J. Math. Sci., 160(1), 2009, 123-135.

## **Current address**

### **Prof. Yasir Khan**

Department of Mathematics  
Zhejiang University, Hangzhou 310027, China  
e-mail: yasirmath@yahoo.com

### **Doc. RNDr. Zdeněk Šmarda, CSc.**

Department of Mathematics  
Faculty of Electrical Engineering and Communication  
Brno University of Technology, Technická 8, 616 00 BRNO  
e-mail: smarda@feec.vutbr.cz

## EXISTENCE OF NONOSCILLATORY SOLUTIONS OF DELAY DIFFERENTIAL EQUATIONS

KUBJATKOVÁ Martina, (SK), OLACH Rudolf,(SK), ŠTOBEROVÁ Júlia,(SK)

**Abstract.** The article is concerned with the existence of the positive solutions of the delay differential equations. The solutions are bounded by the positive functions. The main results are illustrated with some examples.

**Key words and phrases.** Delay differential equation, existence, positive solution, Banach space.

*Mathematics Subject Classification.* Primary 34K15; Secondary 34K12.

### 1 Introduction

The article deals with the existence of the nonoscillatory solutions of the delay differential equations of the form

$$\dot{x}(t) + p(t)x(t) + q(t)x(\tau(t)) = 0, \quad t \geq t_0. \quad (1)$$

With respect to Eq. (1) throughout we will assume the following conditions:

- (i)  $\tau, q \in C([t_0, \infty), [0, \infty))$ ,  $q(t) \not\equiv 0$ ,  $p \in C([t_0, \infty), R)$ ,
- (ii)  $\tau$  is increasing,  $\tau(t) < t$  and  $\lim_{t \rightarrow \infty} \tau(t) = \infty$ .

A solution of Eq. (1) is called oscillatory if it has arbitrarily large zeros and otherwise it is nonoscillatory.

The problem of the existence of nonoscillatory solutions of the delay differential equations received less attention as oscillation problem. It is due mainly to the technical difficulties arising in analysis of the problem. For the similar problems we refer the readers to [1–9] and the references cited therein. In this article we are interested in the study of the existence of solutions of Eq. (1) which are bounded by positive functions.

The following fixed point theorem will be used to prove the main results in the next section.

**Lemma 1.1 ([4,5]Schauder's Fixed Point Theorem)** *Let  $\Omega$  be a closed, convex and non-empty subset of a Banach space  $X$ . Let  $S : \Omega \rightarrow \Omega$  be a continuous mapping such that  $S\Omega$  is a relatively compact subset of  $X$ . Then  $S$  has at least one fixed point in  $\Omega$ . That is, there exists an  $x \in \Omega$  such that  $Sx = x$ .*

## 2 Existence of Nonoscillatory Solutions

In this section we will study the existence of nonoscillatory solutions for Eq.(1), which are bounded by positive functions. The main result is in the next theorem.

**Theorem 2.1** *Suppose that there exist functions  $k_1, k_2 \in C([t_0, \infty), (0, \infty))$  such that*

$$\begin{aligned} k_1(t) &\leq k_2(t), \quad t \geq t_0, \\ p(t) + k_1(t)q(t) &\geq 0, \quad t \geq t_0, \\ \ln k_1(t) &\leq \int_{\tau(t)}^t [p(s) + k_1(s)q(s)] ds \quad \text{and} \quad \int_{\tau(t)}^t [p(s) + k_2(s)q(s)] ds \leq \ln k_2(t), \quad t \geq t_0. \end{aligned} \quad (2)$$

*Then Eq. (1) has a solution which is bounded by positive functions.*

**Proof.** We choose  $T \geq t_0 + \tau(T)$  and set

$$u(t) = \exp \left( - \int_T^t [p(s) + k_2(s)q(s)] ds \right), \quad v(t) = \exp \left( - \int_T^t [p(s) + k_1(s)q(s)] ds \right), \quad t \geq T.$$

Let  $C([t_0, \infty), R)$  be the set of all continuous functions with the norm  $\|x\| = \sup_{t \geq t_0} |x(t)| < \infty$ . Then  $C([t_0, \infty), R)$  is a Banach space. We now define a close, bounded and convex subset  $\Omega$  of  $C(t_0, \infty), R)$  as follows:

$$\begin{aligned} \Omega = \{ & u(t) \leq x(t) \leq v(t), \quad t \geq T, \\ & x(\tau(t)) \leq k_2(t)x(t), \quad t \geq T, \\ & x(\tau(t)) \geq k_1(t)x(t), \quad t \geq T, \\ & x(t) = 1, \quad \tau(T) \leq t \leq T \}. \end{aligned}$$

Define the map  $S : \Omega \rightarrow C([t_0, \infty), R)$  as follows

$$(Sx)(t) = \begin{cases} \exp \left( - \int_T^t [p(s) + q(s) \frac{x(\tau(s))}{x(s)}] ds \right), & t \geq T, \\ 1, & \tau(t) \leq t \leq T. \end{cases}$$

We will show that for any  $x \in \Omega$  we obtain  $Sx \in \Omega$ . For every  $x \in \Omega$  and  $t \geq T$  we get

$$(Sx)(t) \leq \exp \left( - \int_T^t [p(s) + k_1(s)q(s)] ds \right) = v(t).$$

Furthermore for  $t \geq T$  and  $x \in \Omega$  we obtain

$$(Sx)(t) \geq \exp \left( - \int_T^t [p(s) + k_2(s)q(s)] ds \right) = u(t).$$

For  $t \in [\tau(T), T]$  we have  $(Sx)(t) = 1$ , that is  $(Sx)(t) \in \Omega$ . Further for every  $x \in \Omega$  and  $\tau(t) \geq T$  we get

$$\begin{aligned} (Sx)(\tau(t)) &= \exp \left( - \int_T^{\tau(t)} \left[ p(s) + q(s) \frac{x(\tau(s))}{x(s)} \right] ds \right) \\ &= (Sx)(t) \exp \left( \int_{\tau(t)}^t \left[ p(s) + q(s) \frac{x(\tau(s))}{x(s)} \right] ds \right). \end{aligned} \quad (3)$$

With regard to (2) and (3) it follows

$$\begin{aligned} (Sx)(\tau(t)) &\leq (Sx)(t) \exp \left( \int_{\tau(t)}^t [p(s) + k_2(s)q(s)] ds \right) \\ &\leq k_2(t)(Sx)(t), \quad \tau(t) \geq T, \end{aligned}$$

$$\begin{aligned} (Sx)(\tau(t)) &\geq (Sx)(t) \exp \left( \int_{\tau(t)}^t [p(s) + k_1(s)q(s)] ds \right) \\ &\geq k_1(t)(Sx)(t), \quad \tau(t) \geq T, \end{aligned}$$

For  $\tau(T) \leq \tau(t) \leq T$  we have  $(Sx)(\tau(t)) = 1$ , that is  $(Sx)(\tau(t)) \in \Omega$ . Thus we have proved that  $Sx \in \Omega$  for any  $x \in \Omega$ .

We now show that  $S$  is continuous. Let  $x_i = x_i(t) \in \Omega$  be such that  $x_i(t) \rightarrow x(t)$  as  $i \rightarrow \infty$ . Because  $\Omega$  is closed,  $x = x(t) \in \Omega$ . For  $t \geq T$  we get

$$\begin{aligned} &|(Sx_i)(t) - (Sx)(t)| \\ &= \left| \exp \left( - \int_T^t \left[ p(s) + q(s) \frac{x_i(\tau(s))}{x_i(s)} \right] ds \right) - \exp \left( - \int_T^t \left[ p(s) + q(s) \frac{x(\tau(s))}{x(s)} \right] ds \right) \right|. \end{aligned}$$

By applying the Lebesgue dominated convergence theorem we obtain that

$$\lim_{i \rightarrow \infty} \|(Sx_i)(t) - (Sx)(t)\| = 0.$$

For  $t \in [\tau(T), T]$  the relation above is also valid. We conclude that  $S$  is continuous.

The family of functions  $\{Sx : x \in \Omega\}$  is uniformly bounded on  $[\tau(T), \infty)$ . It follows from the definition of  $\Omega$ . This family is also equicontinuous on  $[\tau(T), \infty)$ , (cf. [5], p. 45). Then by Arzela-Ascoli theorem the  $S\Omega$  is relatively compact subset of  $C([t_0, \infty), R)$ . By Lemma 1.1 there is an  $x_0 \in \Omega$  such that  $Sx_0 = x_0$ . We see that  $x_0(t)$  is a positive solution of Eq. (1) which is bounded by the functions  $u, v$ . The proof is complete.

### 3 Corollaries

In this section we will present some corollaries.

**Corollary 3.1** Suppose that  $0 < k_1 \leq k_2$ ,

$$p(t) + k_1 q(t) \geq 0, \quad t \geq t_0,$$

$$\ln k_1 \leq \int_{\tau(t)}^t [p(s) + k_1 q(s)] ds \quad \text{and} \quad \int_{\tau(t)}^t [p(s) + k_2 q(s)] ds \leq \ln k_2, \quad t \geq t_0.$$

Then Eq. (1) has a solution which is bounded by positive functions.

**Proof.** We put  $k_1(t) = k_1$ ,  $k_2(t) = k_2$  and apply the Theorem 2.1.

**Corollary 3.2** Suppose that there exists a function  $k \in C([t_0, \infty), (0, \infty))$  such that

$$p(t) + k(t)q(t) \geq 0, \quad t \geq t_0,$$

$$\int_{\tau(t)}^t [p(s) + k(s)q(s)] ds = \ln k(t), \quad t \geq t_0.$$

Then Eq. (1) has a solution

$$x(t) = \exp \left( - \int_T^t [p(s) + k(s)q(s)] ds \right), \quad t \geq T.$$

**Proof.** We put  $k_1(t) = k_2(t) = k(t)$  and apply the Theorem 2.1.

**Corollary 3.3** Suppose that  $k > 0$  and

$$p(t) + kq(t) \geq 0, \quad t \geq t_0,$$

$$\int_{\tau(t)}^t [p(s) + kq(s)] ds = \ln k, \quad t \geq t_0.$$

Then Eq. (1) has a solution

$$x(t) = \exp \left( - \int_T^t [p(s) + kq(s)] ds \right), \quad t \geq T.$$

**Proof.** We put  $k(t) = k$  and apply the Corollary 3.2.

**Corollary 3.4** Suppose that there exist functions  $k_1, k_2 \in C([t_0, \infty), (0, \infty))$ ,  $\alpha \in C([t_0, \infty), [0, \infty))$  such that

$$\alpha(t) \leq k_1(t) \leq k_2(t), \quad t \geq t_0,$$

$$p(t) + \alpha(t)q(t) = 0, \quad t \geq t_0,$$

$$\ln k_1(t) \leq \int_{\tau(t)}^t [k_1(s) - \alpha(s)]q(s) ds, \quad \int_{\tau(t)}^t [k_2(s) - \alpha(s)]q(s) ds \leq \ln k_2(t), \quad t \geq t_0.$$

Then Eq. (1) has a solution which is bounded by positive functions.

**Proof.** We put  $p(t) = -\alpha(t)q(t)$  into (2) and apply the Theorem 2.1.

**Corollary 3.5** Suppose that  $0 \leq \alpha < k_1 \leq k_2$  and

$$p(t) + \alpha q(t) = 0, \quad t \geq t_0,$$

$$\frac{\ln k_1}{k_1 - \alpha} \leq \int_{\tau(t)}^t q(s) ds \leq \frac{\ln k_2}{k_2 - \alpha}, \quad t \geq t_0.$$

Then Eq. (1) has a solution which is bounded by positive functions.

**Proof.** We put  $\alpha(t) = \alpha$ ,  $k_1(t) = k_1$ ,  $k_2(t) = k_2$  and apply the Corollary 3.4.

## 4 Examples

**Example 4.1** Consider the delay differential equation

$$\dot{x}(t) + \frac{1}{5}x(t) + \frac{1}{10}x(t-1) = 0, \quad t \geq 1. \quad (4)$$

If we take  $k_1 = 1$ ,  $k_2 = 2$ , then all conditions of Corollary 3.1 are satisfied and Eq. (4) has a solution which is bounded with the functions

$$u(t) = \exp(-0.4(t-1)), \quad v(t) = \exp(-0.3(t-1)), \quad t \geq 1.$$

**Example 4.2** Consider the delay differential equation

$$\dot{x}(t) + \frac{1}{2}x(t) + \frac{1}{2}e^{-t}x(t-2) = 0, \quad t \geq 2. \quad (5)$$

If we set  $k_1(t) = 2$ ,  $k_2(t) = e^t$ , then all conditions of Theorem 2.1 are satisfied and Eq. (5) has a solution which is bounded by the functions

$$u(t) = \exp(2-t), \quad v(t) = \exp(1 - e^{-2} + e^{-t} - 0.5t), \quad t \geq 2.$$

**Example 4.3** Consider the delay differential equation

$$\dot{x}(t) + \frac{1}{2t}x(t) + \frac{1}{5t}x(0.5t) = 0, \quad t \geq 1. \quad (6)$$

If we take  $k_1 = 1.5$ ,  $k_2 = 2$ , then all conditions of Corollary 3.1 are satisfied and Eq. (6) has a solution which is bounded by the functions

$$u(t) = \left(\frac{2}{t}\right)^{0.9}, \quad v(t) = \left(\frac{2}{t}\right)^{0.8}, \quad t \geq 2.$$

**Example 4.4** Consider the delay differential equation

$$\dot{x}(t) + \frac{1}{2t}x(t) + \frac{1}{3t}x\left(\frac{2}{3}t\right) = 0, \quad t \geq 1. \quad (7)$$

If we take  $k = 1.5$ , then all conditions of Corollary 3.3 are satisfied and Eq. (7) has the solution

$$x(t) = \frac{2}{t}, \quad t \geq 2.$$

**Example 4.5** Consider the delay differential equation

$$\dot{x}(t) + \frac{1}{t}x(t) + \frac{1}{t^2}x(\sqrt{t}) = 0, \quad t \geq 1. \quad (8)$$

If we set  $k(t) = t$ , then all conditions of Corollary 3.2 are satisfied and Eq. (8) has the solution

$$x(t) = \frac{4}{t^2}, \quad t \geq 2.$$

### Acknowledgement

The research was supported by the grants 1/0090/09 and 1/1260/12 of the Scientific Grant Agency of the Ministry of Education of the Slovak Republic.

### References

- [1] DIBLÍK, J., KÚDELČÍKOVÁ, M.: *Two classes of asymptotically different positive solutions of the equation  $\dot{y}(t) = -f(t, y_t)$* . Nonlinear Analysis, Vol. 70, pp. 3702-3714, 2009.
- [2] DIBLÍK, J., RŮŽIČKOVÁ, M.: *Existence of positive solutions of a singular initial problem for a nonlinear system of differential equations*. Rocky Mountain J. Math., Vol. 3, pp. 923-944, 2004.
- [3] DOROCIAKOVÁ, B., OLACH, R.: *Existence of positive solutions of delay differential equations*. Tatra Mt. Math. Publ., Vol. 43, pp. 63-70, 2009.
- [4] ERBE, L. H., KONG, Q. K., ZHANG, B. G.: *Oscillation Theory for Functional Differential Equations*. Marcel Dekker, New York, 1994.
- [5] GYÖRI, I., LADAS, G.: *Oscillation Theory of Delay Differential Equations*. Clarendon Press, Oxford, 1991.
- [6] LADAS, G., SFICAS, Y. G., STAVROULAKIS, I. P.: *Nonoscillatory functional differential equations*. Pacific J. Math., Vol. 115, pp. 391-398, 1984.
- [7] YU, Y. H., WANG, H. Z.: *Nonoscillatory solutions of second-order nonlinear neutral delay equations*. J. Math. Anal. Appl., Vol. 311, pp. 445-456.
- [8] ZHOU, Y.: *Existence for nonoscillatory solutions of second-order nonlinear differential equations*. J. Math. Anal. Appl., Vol. 331, pp. 91-96, 2007.
- [9] ZHOU, Y., ZHANG, B. G.: *Existence of nonoscillatory solutions of neutral differential equations with positive and negative coefficients*. Appl. Math. Lett., Vol. 15, pp. 867-874, 2002.

**Current address**

**Martina Kubjatková, Mgr.**

Katedra matematiky, Žilinská univerzita, Univerzitná 1, 010 26 Žilina,  
e-mail: martina.kubjatkova@fhv.uniza.sk

**Rudolf Olach, doc.**

Katedra matematiky, Žilinská univerzita, Univerzitná 1, 010 26 Žilina,  
e-mail: rudolf.olach@fpv.uniza.sk

**Júlia Štoberová, Mgr.**

Katedra krízového manažmentu, Žilinská univerzita, ul. 1.mája 32, 010 26 Žilina,  
e-mail: julia.stoberova@gmail.com



## DIFFERENTIABILITY WITH RESPECT TO DELAY OF THE SOLUTION OF A CAUCHY PROBLEM

MUREȘAN Viorica, (RO)

**Abstract.** In this paper we use fiber contraction principle (see Rus I. A. [24]), to study the differentiability with respect to delay of the solution of a Cauchy problem for a differential equation with linear modification of the argument.

**Key words and phrases.** fixed point, functional differential equation, differentiability of the solution.

*Mathematics Subject Classification.* 34K05, 34K15, 47H10.

### 1 Introduction

The theory of functional differential equations and of functional integral equations are both active fields in mathematics.

Many problems from physics, chemistry, astronomy, biology, engineering, social sciences lead to mathematical models described by functional differential equations. The theory of these equations has developed very much.

The differential equations with linear modification of the argument are a special class of functional differential equations. The pantograph equation and its generalization have been studied very much (see [ 9] - [11 ], [13], [17], [18],...).

For the monographs in the field of functional differential equations we quote here ([1] - [6], [12], [14], [15], [19],...).

All these monographs and papers contain a lot of techniques, ideas and applications.

In the conditions of an existence and uniqueness theorem, we study the differentiability with respect to delay of the solution of a Cauchy problem for a differential equation with linear modification of the argument. We apply fiber contraction principle (Theorem 2.3.).

## 2 Needed notions from Picard operators' theory

Here, we present some notions and results from Picard operators' theory.

Let  $(X, d)$  be a metric space and  $A : X \longrightarrow X$  an operator.

We denote by  $A^0 := 1_X$ ,  $A^1 := A$ , ...,  $A^{n+1} := A \circ A^n$ ,  $n \in \mathbb{N}$ , the iterate operators of the operator  $A$ . Also:

$$\begin{aligned} P(X) &:= \{Y \subset X / Y \neq \emptyset\}, \\ I(A) &:= \{Y \in P(X) / A(Y) \subset Y\}, \end{aligned}$$

the family of all nonempty invariant subsets of  $A$ ,

$$F_A = \{x \in X / A(x) = x\},$$

the fixed point set of the operator  $A$ .

Following Rus I.A. [19] - [23], [25] - [27], we have:

**Definition 2.1.** *A is a Picard operator if there exists  $x^* \in X$  such that*

- 1)  $F_A = \{x^*\}$ ;
- 2) *the successive approximation sequence  $(A^n(x_0))_{n \in \mathbb{N}}$  converges to  $x^*$ , for all  $x_0 \in X$ .*

**Definition 2.2.** *A is a weakly Picard operator if the sequence  $(A^n(x_0))_{n \in \mathbb{N}}$  converges for all  $x_0 \in X$  and the limit (which generally depends on  $x_0$ ) is a fixed point of  $A$ .*

**Definition 2.3.** *For an weakly Picard operator  $A : X \rightarrow X$  we define the operator  $A^\infty$  as follows:*

$$A^\infty : X \rightarrow X, \quad A^\infty(x) := \lim_{n \rightarrow \infty} A^n(x), \quad \text{for all } x \in X.$$

**Remark 2.1.**  $A^\infty(X) = F_A$ .

**Theorem 2.1.** (**Contraction principle**). *Let  $(X, d)$  be a complete metric space and  $A : X \rightarrow X$  a contraction. Then  $A$  is a Picard operator.*

**Theorem 2.2.** (**data dependence theorem**). *Let  $(X, d)$  be a complete metric space and  $A, B : X \longrightarrow X$  two operators. We suppose that:*

- (i) *A is an  $\alpha$ -contraction and let  $F_A = \{x_A^*\}$ ;*
- (ii)  *$F_B \neq \emptyset$  and let  $x_B^* \in F_B$ ;*
- (iii) *there exists  $\delta > 0$ , such that  $d(A(x), B(x)) \leq \delta$ , for all  $x \in X$ .*

*Then*

$$d(x_A^*, x_B^*) \leq \frac{\delta}{1 - \alpha}.$$

The following theorem can be used for proving solution of operatorial equations to be differentiable:

**Theorem 2.3. (fiber contraction principle)** (Hirsch-Pugh [8], Rus [24]) *Let  $(X, d)$  be a metric space,  $(Y, \rho)$  be a complete metric space and  $T : X \times Y \rightarrow X \times Y$  a continuous operator. We suppose that:*

- (i)  $T(x, y) = (T_1(x), T_2(x, y))$ ;
- (ii)  $T_1 : X \rightarrow X$  is a Picard operator;
- (iii) there exists  $0 < c < 1$  such that

$$\rho(T_2(x, y), T_2(x, z)) \leq c \rho(y, z), \text{ for all } x \in X \text{ and all } y, z \in Y.$$

*Then the operator  $T$  is a Picard operator.*

### 3 Mathematical models which contain differential equations with linear modification of the argument

**Example 3.1.** (A problem of the pantograph) As it was shown by Ockendon and Tayler in [18] (1971) and Ockendon in [17] (1980), the dynamics of a current collection system for an electric locomotive (of a pantograph) with some imposed physical conditions, gives a mathematical model. This model is a system of equations with linear modification of the argument of the form:

$$Y'(t) = AY(\lambda t) + BY(t), \quad t > 0, \quad 0 < \lambda < 1,$$

where  $A$  and  $B$  are constant matrices.

**Remark 3.1.** The equation

$$y'(t) = y(\lambda t), \quad t > 0, \quad 0 < \lambda < 1,$$

appeared for the first time in the paper of Mahler [13] (1940) in relationship with a problem from the theory of numbers.

**Remark 3.2.** The equation

$$y'(t) = y\left(\frac{t}{2}\right), \quad t \geq 0$$

was mentioned by Harari and Palmer in [7] (1977) in relationship with a problem from the theory of graphs.

**Example 3.2.** (A problem from the geometry of curves)

Consider  $(C) : y = y(x)$ ,  $x \in I \subseteq \mathbb{R}$  and  $M(x, y(x)) \in (C)$ . We have to determine all the curves  $y = y(x)$  for which the tangent vector in every point  $M(x, y(x))$  is parallel with the vector determined by  $O(0, 0)$  and  $P(1, y(\lambda x))$ , where  $\lambda \in \mathbb{R}$ . So, we obtain the following equation with linear modification of the argument:

$$y'(x) = y(\lambda x), \quad x \in I, \quad \lambda \in \mathbb{R}.$$

In 1971 appeared a very important paper [11] in which Kato and McLeod studied the asymptotical properties for the solutions of the following problem for the pantograph equation:

$$\begin{aligned} y'(x) &= ay(\lambda x) + by(x), \quad x > 0, \quad \lambda > 0, \quad \lambda \neq 1 \\ y(0) &= 1, \end{aligned}$$

in the cases  $0 < \lambda < 1$  and  $\lambda > 1$ , where  $a, b \in \mathbb{R}$ .

In the next years appeared many papers on the subject of pantograph equation and its generalizations ([9], [10], [14], [16], [17], [18], [30],...). So, the subject of "pantograph equation" is an old but very actual subject.

#### 4 Differentiability with respect to delay

Consider the following problem

$$y'(x) = f(x, y(x), y(\lambda x)), \quad x \in [0, b], \quad 0 < \lambda < 1 \quad (4.1)$$

$$y(0) = \tilde{y}_0, \quad (4.2)$$

where  $f \in C([0, b] \times \mathbb{R} \times \mathbb{R})$  and  $\tilde{y}_0 \in \mathbb{R}$ .

As it is well known, this problem is equivalent with the following functional integral equation:

$$y(x) = \tilde{y}_0 + \int_0^x f(s, y(s), y(\lambda s)) ds, \quad x \in [0, b], \quad 0 < \lambda < 1. \quad (4.3)$$

Consider the following Bielecki norm  $\|\cdot\|_B$  on  $C[0, b]$ , defined by

$$\|y\|_B = \max_{x \in [0, b]} (|y(x)|e^{-\tau x}), \quad \text{where } \tau > 0.$$

By using this norm and by applying Contraction principle to the operator

$$\begin{aligned} A &: (C[0, b], \|\cdot\|_B) \rightarrow (C[0, b], \|\cdot\|_B), \\ (A(y))(x) &: = \tilde{y}_0 + \int_0^x f(s, y(s), y(\lambda s)) ds, \quad x \in [0, b], \quad 0 < \lambda < 1, \end{aligned}$$

we obtain

**Theorem 4.1.** (Theorem 3.1. [16]) *Suppose that the following conditions are satisfied:*

- (i)  $f \in C([0, b] \times \mathbb{R} \times \mathbb{R})$  and  $\tilde{y}_0 \in \mathbb{R}$ ;
- (ii) *there exists  $L > 0$  such that*

$$|f(x, u, v) - f(x, \bar{u}, \bar{v})| \leq L(|u - \bar{u}| + |v - \bar{v}|),$$

for all  $x \in [0, b]$  and  $u, \bar{u}, v, \bar{v} \in \mathbb{R}$ .

*Then the problem (4.1)+(4.2) has in  $C[0, b]$  a unique solution  $y^*$  and this solution can be obtained by the successive approximation method starting from any element of  $C[0, b]$ .*

Let us consider the following problem:

$$y'(x) = g(x, y(x), y(\lambda x)), \quad x \in [0, b], 0 < \lambda < 1 \quad (4.4)$$

$$y(0) = \tilde{y}_0, \quad (4.5)$$

where  $g \in C([0, b] \times \mathbb{R} \times \mathbb{R})$  and  $\tilde{y}_0, \lambda$  are the same as in the problem (4.1)+(4.2).

By using data dependence theorem (Theorem 2.2.), we obtain

**Theorem 4.2.** (Theorem 3.2. [16]) *Suppose that:*

(i) *the conditions in Theorem 4.1 are satisfied and  $y^* \in C[0, b]$  is the unique solution of the problem (4.1)+(4.2);*

(ii) *there exists  $\eta > 0$  such that*

$$|f(x, u, v) - g(x, u, v)| \leq \eta, \text{ for all } x \in [0, b] \text{ and } u, v \in \mathbb{R}.$$

Then

$$d(y^*, w^*) \leq \frac{\eta b}{1 - L_A}, \text{ where } L_A = \frac{L(1 + \frac{1}{\lambda})}{\tau},$$

for all  $w^*$  solutions of (4.4)+(4.5).

In [14], we have studied Cauchy problems for differential equations with linear modification of the argument. We have given existence, uniqueness and data dependence results. The aim of this paper is to study the differentiability with respect to delay of the solution of a Cauchy problem, by applying fiber contraction principle (Theorem 2.3.).

So, in what follows we consider the following integral equation:

$$y(x, \lambda) = \tilde{y}_0 + \int_0^x f(s, y(s, \lambda), y(\lambda s, \lambda)) ds, \quad x \in [0, b], \lambda \in [0, 1], \quad (4.6)$$

where  $f \in C([0, b] \times \mathbb{R} \times \mathbb{R})$  and  $\tilde{y}_0 \in \mathbb{R}$ .

We are looking for the solution of this equation in

$$X = (C([0, b] \times [0, 1]), \|\cdot\|_B) \text{ with } \|y\|_B = \max_{\substack{x \in [0, b] \\ \lambda \in [0, 1]}} |y(x, \lambda)| e^{-\tau x}, \text{ where } \tau > 0.$$

We have

**Theorem 4.3.** *We suppose that:*

(i)  $f \in C^1([0, b] \times \mathbb{R} \times \mathbb{R})$ ;

(ii) *there exists  $M_i > 0$  such that*

$$\left| \frac{\partial f}{\partial u_i}(s, u_1, u_2) \right| \leq M_i, \quad i = 1, 2.$$

Then

(a) *the equation (4.6) has in  $C([0, b] \times [0, 1])$  a unique solution  $y^*$ ;*

(b) for all  $y_0 \in C([0, b] \times [0, 1])$ , the sequence  $(y_n)_{n \in \mathbb{N}}$ , defined by

$$y_{n+1}(x, \lambda) := \tilde{y}_0 + \int_0^x f(s, y_n(s, \lambda), y_n(\lambda s, \lambda)) ds, \quad x \in [0, b], \quad \lambda \in [0, 1],$$

converges uniformly to  $y^*$ ;

(c)  $y^* \in C^1([0, b] \times [0, 1])$ .

**Proof.** By the same proof as of the Theorem 4.1. we obtain (a), (b) and the continuity of  $y^*$ . We remark that  $\frac{\partial y^*}{\partial x} \in C([0, b] \times [0, 1])$ .

Let  $y^*$  be the solution of the equation (4.6). So, we have

$$y^*(x, \lambda) := \tilde{y}_0 + \int_0^x f(s, y^*(s, \lambda), y^*(\lambda s, \lambda)) ds, \quad x \in [0, b], \quad \lambda \in [0, 1]. \quad (4.7)$$

Let us prove that there exists  $\frac{\partial y^*}{\partial \lambda}(x, \lambda)$  and  $\frac{\partial y^*}{\partial \lambda} \in C([0, b] \times [0, 1])$ .

We consider a subset  $X_1 \subset X$ ,  $X_1 := \{y \in X / \frac{\partial y}{\partial x}(\cdot, \lambda) \in C[0, b]\}$ . We remark that for all fixed  $\lambda \in [0, 1]$ ,  $y^* \in X_1$  and  $T_1(X_1) \subset X_1$ , where  $T_1 : (X_1, \|\cdot\|_B) \rightarrow (X_1, \|\cdot\|_B)$  is defined by

$$(T_1(y))(x, \lambda) := \tilde{y}_0 + \int_0^x f(s, y(s, \lambda), y(\lambda s, \lambda)) ds, \quad x \in [0, b], \quad \lambda \in [0, 1]$$

and  $T_1$  is a Picard operator. We shall use the following heuristic argument.

We suppose that there exists  $\frac{\partial y^*}{\partial \lambda}$ . Then from (4.7) we obtain

$$\begin{aligned} \frac{\partial y^*}{\partial \lambda}(x, \lambda) &= \int_0^x \frac{\partial f}{\partial u_1}(s, y^*(s, \lambda), y^*(\lambda s, \lambda)) \frac{\partial y^*}{\partial \lambda}(s, \lambda) ds + \\ &+ \int_0^x \frac{\partial f}{\partial u_2}(s, y^*(s, \lambda), y^*(\lambda s, \lambda)) \frac{\partial y^*}{\partial \lambda}(\lambda s, \lambda) ds + \\ &+ \int_0^x \frac{\partial f}{\partial u_2}(s, y^*(s, \lambda), y^*(\lambda s, \lambda)) s \frac{\partial y^*}{\partial u}(u, \lambda)|_{u=\lambda s} ds, \end{aligned}$$

for  $x \in [0, b]$ , where  $f = f(s, u_1, u_2)$ .

This relationship suggests us to consider the operator  $T_2 : X_1 \times X \rightarrow X$ ,  $(y, z) \rightarrow T_2(y, z)$ , defined by

$$\begin{aligned} (T_2(y, z))(x, \lambda) &:= \int_0^x \frac{\partial f}{\partial u_1}(s, y(s, \lambda), y(\lambda s, \lambda)) z(s, \lambda) ds + \\ &+ \int_0^x \frac{\partial f}{\partial u_2}(s, y(s, \lambda), y(\lambda s, \lambda)) z(\lambda s, \lambda) ds + \\ &+ \int_0^x s \frac{\partial f}{\partial u_2}(s, y(s, \lambda), y(\lambda s, \lambda)) \frac{\partial y}{\partial u}(u, \lambda)|_{u=\lambda s} ds, \end{aligned}$$

for  $x \in [0, b]$  and  $\lambda \in [0, 1]$ .

By using (ii), we obtain

$$\|T_2(y, z) - T_2(y, w)\|_B \leq (M_1 + \frac{M_2}{\lambda}) \tau^{-1} \|z - w\|_B,$$

for all  $y \in X_1$  and all  $z, w \in X$ .

Choosing  $\tau = M_1 + \frac{M_2}{\lambda} + 1$ , we have that  $T_2$  is a contraction with respect to its last argument. If we take the operator  $T = (T_1, T_2)$ , then we are in the conditions of the fiber contraction principle (Theorem 2.3). It follows from this theorem that  $T : X_1 \times X \rightarrow X_1 \times X$ ,  $T(y, z) = (T_1(y), T_2(y, z))$  is a Picard operator. So, the sequences  $(y_n)_{n \in \mathbb{N}}, (z_n)_{n \in \mathbb{N}}$ , defined by  $(y_{n+1}, z_{n+1}) = T(y_n, z_n)$ , where

$$y_{n+1}(x, \lambda) := \tilde{y}_0 + \int_0^x f(s, y_n(s, \lambda), y_n(\lambda s, \lambda)) ds, \quad x \in [0, b], \quad \lambda \in [0, 1],$$

respectively

$$\begin{aligned} z_{n+1}(x, \lambda) : &= \int_0^x \frac{\partial f}{\partial u_1}(s, y_n(s, \lambda), y_n(\lambda s, \lambda)) z_n(s, \lambda) ds + \\ &+ \int_0^x \frac{\partial f}{\partial u_2}(s, y_n(s, \lambda), y_n(\lambda s, \lambda)) z_n(\lambda s, \lambda) ds + \\ &+ \int_0^x s \frac{\partial f}{\partial u_2}(s, y_n(s, \lambda), y_n(\lambda s, \lambda)) \frac{\partial y_n}{\partial u}(u, \lambda)|_{u=\lambda s} ds, \end{aligned}$$

converge uniformly on  $[0, b] \times [0, 1]$  to  $(y^*, z^*)$ , for all  $y_0 \in X_1$ ,  $z_0 \in X$ , and  $(y^*, z^*)$  is the unique fixed point of the operator  $T$ . But for fixed  $y_0 \in X_1$  and  $z_0 \in X$  such that  $z_0 = \frac{\partial y_0}{\partial \lambda}$  we have that  $z_1 = \frac{\partial y_1}{\partial \lambda}$ . By induction, we can prove that  $z_n = \frac{\partial y_n}{\partial \lambda}$ . Thus  $(y_n)_{n \in \mathbb{N}}$  converges uniformly to  $y^*$  and  $(\frac{\partial y_n}{\partial \lambda})_{n \in \mathbb{N}}$  converges uniformly to  $z^*$ . By using a Weierstrass argument, we conclude that  $\frac{\partial y^*}{\partial \lambda}$  exists and  $\frac{\partial y^*}{\partial \lambda} = z^*$ . These imply that  $\frac{\partial y^*}{\partial \lambda}$  is a continuous function.

## References

- [1] AZBELEV N.V., MAKSIMOV V. P. and RAHMATULINA L. F., *Introduction to functional - differential equations theory*, MIR, Moscow, 1991 (In Russian)
- [2] BELLMAN R. E. and COOKE K. L., *Differential - difference equations*, Acad. Press, New York, 1963
- [3] BERNFELD S.R. and LAKSHMIKANTHAM V., *An introduction to nonlinear boundary value problems*, Acad. Press, New York, 1974
- [4] ELSGOLTZ L. F. and NORKIN S. B., *Introduction to the theory of differential equations with deviating arguments*, MIR, Moscow, 1971 (In Russian)
- [5] HALE J. K., *Theory of functional - differential equations*, Springer Verlag, 1977
- [6] HALE J. K. and SJOERD M. VERDUYN Lunel, *Introduction to functional - differential equations*, Springer Verlag, New York, 1993
- [7] HARARI F. and PALMER E., *The theory of graphs*, MIR, Moscow, 1977 (in Russian)
- [8] HIRSCH M. W. and PUGH C. C., *Stable manifolds and hyperbolic sets*, Proc. Symp. in Pure Math, 14 (1970), 133 - 163
- [9] ISERLES A., *On the generalized pantograph functional - differential equation*, European J. Appl. Math. 4 (1992), 1 - 38
- [10] ISERLES A., *Exact and discretized stability of the pantograph equation*, Appl. Numer. Math. 24 (1997), No.2-3, 295-308

- [11] KATO T., Mc LEOD J. B., *The functional - differential equation  $y'(x) = ay(\lambda x) + by(x)$* , Bull. Amer. Math. Soc. , 77, 6 (1971), 891-937
- [12] LAKSHMIKANTHAM V.(ed), *Trends in the theory and practice of nonlinear differential equations*, Marcel Dekker, New York, 1984
- [13] MAHLER K., *On a special functional equation*, J. London Math. Soc.,15 (1940), 115-123
- [14] MUREȘAN V., *Differential equations with affine modification of the argument*, Transilvania Press, Cluj-Napoca, 1997 (in Romanian)
- [15] MUREȘAN V., *Functional - integral equations*, Ed. Mediamira, Cluj - Napoca, 2003
- [16] MUREȘAN V., *From the pantograph equation to the theory of functional differential equations with linear modification of the argument*, Proceedings of the 6-th International Conference APLIMAT 2007, 245-256
- [17] OCKENDON J. R., *Differential equations and industry*, The Math. Scientist, 5(1980), No.1, 1 - 12
- [18] OCKENDON J. R., TAYLER A. B., *The dynamics of a current collection system for an electric locomotive*, Proc. Roy. Soc. London, A 322, 1971, 447 - 468
- [19] RUS I. A., *Principles and applications of the fixed point theory*, Ed. Dacia, Cluj - Napoca, 1979 (In Romanian)
- [20] RUS I. A., *Picard mappings: results and problems*, Babeș - Bolyai University, Cluj - Napoca, Seminar on fixed point theory, Preprint 6 (1987), 55-64
- [21] RUS I. A., *Weakly Picard mappings*, Comment. Math. Univ. Carolinae, 34, 4 (1993), 769 - 77
- [22] RUS I. A., *Picard operators and applications*, Babeș - Bolyai University of Cluj - Napoca, Preprint 3 (1996)
- [23] RUS I. A., *Differential equations, integral equations and dynamical systems*, Transilvania Press, Cluj - Napoca, 1996 (In Romanian)
- [24] RUS I. A., *A fiber generalized contraction theorem and applications*, Mathematica, Tome 41(64), No.1 (1999), 85-90
- [25] RUS I. A., *Weakly Picard operators and applications*, Babeș - Bolyai University of Cluj - Napoca, Seminar on fixed point theory, 2 (2001), 41 - 58
- [26] RUS I. A., *Picard operators and applications*, Scientiae Mathematicae Japonicae, 58 (2003), No.1, 191 - 219
- [27] RUS I. A., *Some nonlinear functional differential and integral equations, via weakly Picard operators theory: a survey*, Carpathian J. Math., 26 (2010), No.2, 230-258
- [28] RUS I. A., Petrușel A. and Petrușel G., *Fixed Point Theory: 1950 - 2000 Romanian Contributions*, House of the Book Science, Cluj - Napoca, 2002
- [29] SOTOMAYOR I., *Smooth dependence of solution of differential equation on initial data: a simple proof*, Bol. Soc. Brasil, 4, 1 (1973), 55-59
- [30] TERJÉKI J., *Representation of the solutions to linear pantograph equation*, Acta Sci. Math. (Szeged), 60 (1995), 705-713

## **Current address**

### **Mureșan Viorica, Professor**

Department of Mathematics, Faculty of Computer Science and Automation,  
Technical University of Cluj-Napoca, Romania, e-mail: vmuresan@math.utcluj.ro

**DYNAMIC BEHAVIOR  
OF LARGE DEFORMABLE RECTANGULAR PLATES  
SUBJECTED TO A MOVING MASS GOVERNED  
BY NONLINEAR NON-HOMOGENOUS HILL EQUATION**

**ROFOOEI Fayaz, R. (IR), ENSHAEIAN Alireza, (IR)**

**Abstract.** In the present study the dynamic oscillations of a large deformable rectangular plate caused by a concentrated moving mass is investigated using modified Homotopy method. The displacement parameter is assumed to be the product of a time dependent weighting function by the related mode shape of the plate depending on the given boundary conditions. Due to the trigonometric modal shapes assumed for the spatial functions and also the moving nature of the loading, the governing differential equation possesses periodic coefficients. Fourier expansion of these coefficients, leads to a nonlinear non-homogenous Hill's equation. In the present study, the analytical solution for the resulting nonlinear Hill's differential equation is presented, using the modified Homotopy technique. Finally, the obtained results are compared with the numerical solutions to the problem using an example. The comparison shows good agreement between analytical and numerical results for a relatively wide range of moving mass weights and velocities.

**Key Words:** Modified Homotopy Method, Moving Mass, Geometric Nonlinearity, Floquet Theory, Dynamic Amplification Factor

*Mathematics Subject Classification:* Dynamic equations on time scales

## **1 Introduction**

The dynamic vibrations induced in structural members by a moving object have been an interesting issue for many researchers in the last decades. For the first time, this subject was raised by bridge engineers, who found that ignoring vertical dynamic deformations in the bridge structural elements caused by moving traffic load may lead to catastrophic events. Afterwards, mathematical models were developed using different methods to describe this phenomenon. The earliest models were generally based on integral transformations which provided an algebraic version of the differential equations ([1, 2]). Also finite element and finite difference methods were utilized to find a solution for this issue ([2]). In most of these methods, the moving object is assumed as a moving force rather

than a moving mass. That resulted in a differential equation with constant coefficients. On the other hand, recent investigations have proved that neglecting the convective acceleration components of the moving object may cause significant errors in determining the dynamic response of the system ([4]). Therefore, numerical and semi-numerical methods are applied to attack the moving mass problem with complete terms. The moving mass problem has been mostly scrutinized for beam structural models, while the effect of traveling masses on plates being received less attention. Meanwhile, some researchers have studied the dynamic influence of a moving mass traversing a Kirchhoff plate, recognizing the importance of load inertia ([4]). Herein, the moving mass problem for large deformable plates is addressed using analytical methods. Therefore some mathematical relations are developed so as to determine the dynamic response of a rectangular plate under a concentrated moving mass.

As already mentioned, the dynamic deformations are assumed to follow a related modal shape of the plate, depending on the assumed boundary conditions, multiplied by a time-dependent weighting function. Based on these spatial functions, the kinetic and potential energies of the plate and the moving mass are determined. The Lagrange method is then employed to derive the main governing differential equation of the problem. Since in the calculation of the strain energy of the plate, Green-Lagrange strain relations are considered so as to include the effect of large deformations, that resulted in a cubic nonlinearity term in the main differential equation of the dynamic system.

Because of utilized trigonometric modal shapes, the governing nonlinear differential equation possesses periodic coefficients. Eventually, the final nonlinear non-homogenous Hill's equation is obtained through Fourier expansion of the periodic coefficients. The nonlinearity of this equation is overcome analytically using modified Homotopy technique. Since the Homotopy method applied herein is based on the linear solution of the governing differential equation, therefore the linear non-homogenous Hill's equation is primarily investigated utilizing Floquet theory. Eventually, the final analytical results are compared with the numerical solutions using an example. The comparisons made show good agreement between the analytical and the numerical results for a relatively wide range of moving mass weights and velocities.

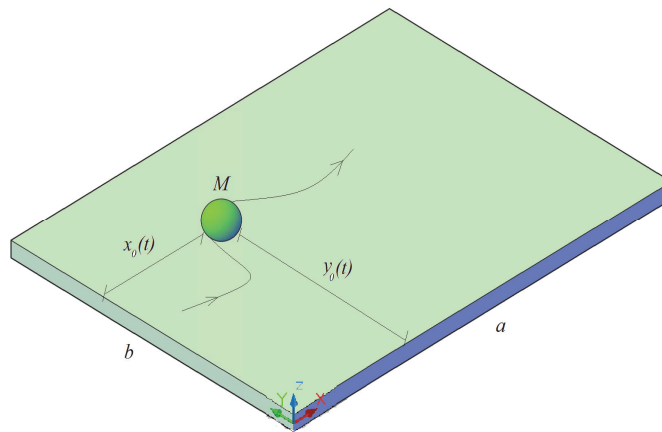
## 2 Problem formulation

The dynamic deformation of a large deformable rectangular plate is expressed using the von Karman theory for plates. In this regard, a uniform un-damped rectangular plate of length **a** and width **b** with arbitrary boundary condition is considered. The mass density of the plate is  $\rho$ , and its bending stiffness is designated by  $D = Eh^3 / (12(1 - \nu^2))$ , in which  $E$ ,  $h$  and  $\nu$  are the plate's modulus of elasticity, thickness and Poisson's ratio respectively (Fig.1). Mid-plane deformations of the assumed plate in directions parallel to  $x$ ,  $y$  and  $z$  axes are denoted by  $u(x, y, t)$ ,  $v(x, y, t)$  and  $w(x, y, t)$ . Ignoring the in-plane velocity components based on the von Karman theory, the kinetic energy of the plate is as the following:

$$K_{plate} = \frac{1}{2} \int_A \rho h \dot{w}^2 dA \quad (1)$$

Also the relevant strain energy of the plate is derived using Green-Lagrange strains relations ([7]) as the following:

$$U_{plate} = \frac{D}{2} \int \left\{ (\nabla^2 w)^2 + \frac{12}{h^2} e_1^2 - 2(1 - \nu) \left[ \frac{12}{h^2} e_2 + \frac{\partial^2 w}{\partial x^2} \frac{\partial^2 w}{\partial y^2} - \left( \frac{\partial^2 w}{\partial x \partial y} \right)^2 \right] \right\} dA \quad (2)$$



**Fig.1.** A mass traversing the plate on a arbitrary trajectory

where,

$$e_1 = \partial u / \partial x + \frac{1}{2} \left( \frac{\partial w}{\partial x} \right)^2 + \partial v / \partial y + \frac{1}{2} \left( \frac{\partial w}{\partial y} \right)^2 \quad (3)$$

and,

$$e_2 = \left( \partial u / \partial x + \frac{1}{2} \left( \frac{\partial w}{\partial x} \right)^2 \right) \left( \partial v / \partial y + \frac{1}{2} \left( \frac{\partial w}{\partial y} \right)^2 \right) - \frac{1}{4} \left( \partial u / \partial y + \partial v / \partial x + \frac{\partial w}{\partial x} \frac{\partial w}{\partial y} \right)^2 \quad (4)$$

As the moving mass  $M$ , traverses the plate on a trajectory described by  $x_0(t)$  and  $y_0(t)$ , it experiences a vertical displacement of  $w_0(t)$ . Therefore, assuming full contact between the mass and the plate, the external excitation force of the moving object on the plate underneath will be calculated as the following:

$$f(x, y, t) = M \left( g - \frac{d^2 w(t)}{dt^2} \right)_{x=x_0(t), y=y_0(t)} \delta(x - x_0(t)) \delta(y - y_0(t)) \quad (5)$$

Including all translational acceleration components of the moving mass, Eq. (5) can be expanded as:

$$f(x, y, t) = M \left\{ g - \left[ \frac{\partial^2 w}{\partial t^2} + \dot{x}_0^2(t) \frac{\partial^2 w}{\partial x^2} + \dot{y}_0^2(t) \frac{\partial^2 w}{\partial y^2} + 2\dot{x}_0(t)\dot{y}_0(t) \frac{\partial^2 w}{\partial x \partial y} + \dot{x}_0(t) \frac{\partial^2 w}{\partial x \partial t} + \dot{y}_0(t) \frac{\partial^2 w}{\partial y \partial t} + \ddot{x}_0(t) \frac{\partial w}{\partial x} + \ddot{y}_0(t) \frac{\partial w}{\partial y} \right]_{x=x_0(t), y=y_0(t)} \right\} \delta(x - x_0(t)) \delta(y - y_0(t)) \quad (6)$$

So, the external work imposed by external force on the plates turns out to be:

$$W = \int_A f(x, y, t) w \, dA \quad (7)$$

where  $f(x, y, t)$  is calculated in Eq. (6). The plate's unknown displacement parameters,  $u(x, y, t)$ ,  $v(x, y, t)$  and  $w(x, y, t)$ , are considered as the product of a time-dependant weighting function by appropriate spatial functions as:

$$u(x, y, t) = r(t)\eta(x, y) \quad , \quad v(x, y, t) = s(t)\psi(x, y) \quad , \quad w(x, y, t) = q(t)\phi(x, y) \quad (8)$$

where the spatial functions  $\eta(x, y)$ ,  $\psi(x, y)$  and  $\phi(x, y)$  are selected as the vibrational modal functions which satisfy the essential boundary conditions. Substituting Eq. (8), into relations (1), (2) and (7) and applying the Lagrange method, an ordinary differential equation describing the vertical vibrations of the plate will be resulted:

$$\begin{aligned}
& \left[ \int_0^A \rho h \phi^2 dA + M \phi(x_0(t), y_0(t)) \phi(x_0(t), y_0(t)) \right] \ddot{q}(t) \\
& + M \phi(x_0(t), y_0(t)) [\dot{x}_0(t) \phi_{,x}(x_0(t), y_0(t)) + \dot{y}_0(t) \phi_{,y}(x_0(t), y_0(t))] \dot{q}(t) \\
& + \left\{ \omega_0^2 \left( \int_0^A \rho h \phi^2 dA \right) + M \phi(x_0(t), y_0(t)) [\dot{x}_0^2(t) \phi_{,xx}(x_0(t), y_0(t)) + \dot{y}_0^2(t) \phi_{,yy}(x_0(t), y_0(t)) + \right. \\
& \left. \ddot{x}_0(t) \phi_{,x}(x_0(t), y_0(t)) + \ddot{y}_0(t) \phi_{,y}(x_0(t), y_0(t)) + 2 \dot{x}_0(t) \dot{y}_0(t) \phi_{,xy}(x_0(t), y_0(t))] \right\} q(t) + \\
& \frac{D}{2} \Gamma q(t)^3 = M g \phi(x_0(t), y_0(t))
\end{aligned} \quad (9)$$

As described, the spatial function  $\phi(x, y)$  is chosen as the natural vibrational mode shape of the plate according to:

$$\phi(x, y) = \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right) \quad (10)$$

where  $m$  and  $n$  are positive integers. Thus, the associative plate's natural frequency will be equal to:

$$\omega_0 = \left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right) \pi^2 \sqrt{\frac{D}{\rho h}} \quad (11)$$

Parameter  $\Gamma$  in Eq. (9) is a constant that depends on the geometric shape of the plate, Poisson's ratio and the vertical and horizontal spatial functions. This parameter originates from the nonlinear strain terms in equations (3) and (4) leading to a non-homogenous nonlinear equation of motion with time-dependent coefficients. The general Eq. (9) is now simplified for a specific simply supported plate with the general mode shape given by Eq. (10). In this case, the moving mass traverses the plate with constant velocity  $c$  on a straight path parallel to  $x$  axis with constant  $y_0$ . Introducing the following non-dimensional parameters:

$$\begin{aligned}
\omega &= \frac{\pi c}{a}, X_0(\tau) = \frac{x_0(t)}{a}, Y_0 = \frac{y_0}{b}, Q(\tau) = \frac{q(t)}{a}, \bar{\epsilon} = \frac{4M}{\rho h a b} \sin^2(n\pi Y_0), \\
\bar{\Lambda} &= \frac{2\Gamma a b h^2}{\pi^4 \left(\frac{a}{b} m^2 + \frac{b}{a} n^2\right)^2}, \tau = t\omega, \bar{\omega}_0 = \frac{\omega_0}{\omega}, G = g \frac{a}{\pi^2 c^2}
\end{aligned} \quad (12)$$

and substituting Eq. (10) into Eq. (9), the following differential equation is derived:

$$\begin{aligned}
\ddot{Q}(\tau) + \frac{m\bar{\epsilon} \sin(m\tau) \cos(m\tau)}{1+\bar{\epsilon} \sin^2(m\tau)} \dot{Q}(\tau) + \frac{\bar{\omega}_0^2 - m^2 \bar{\epsilon} \sin^2(m\tau)}{1+\bar{\epsilon} \sin^2(m\tau)} Q(\tau) + \frac{\bar{\Lambda} \bar{\omega}_0^2}{1+\bar{\epsilon} \sin^2(m\tau)} \left(\frac{a}{h}\right)^2 Q(\tau)^3 = \\
\frac{\bar{\epsilon} G}{\sin(n\pi Y_0)} \frac{\sin(m\tau)}{1+\bar{\epsilon} \sin^2(m\tau)}
\end{aligned} \quad (13)$$

Contrary to the moving force case, Eq. (13), which describes the moving mass problem, includes some damping terms originated from the convective acceleration components of the mass inertia. Using a change of variable according to:

$$Q(\tau) = \frac{x(\tau)}{(1+\bar{\epsilon} \sin^2(m\tau))^{\frac{1}{4}}} \quad (14)$$

Equation (13) turns into a simplified form with no distinct damping terms as the following:

$$\ddot{x}(\tau) + \left( \frac{\bar{\omega}_0^2 - m^2 \bar{\epsilon}/2}{1+\bar{\epsilon} \sin^2(m\tau)} + \frac{3}{16} \left( \frac{m\bar{\epsilon} \sin(2m\tau)}{1+\bar{\epsilon} \sin^2(m\tau)} \right)^2 \right) x(\tau) + \frac{\bar{\Lambda} \bar{\omega}_0^2}{(1+\bar{\epsilon} \sin^2(m\tau))^{3/2}} \left(\frac{a}{h}\right)^2 x(\tau)^3 = \frac{\bar{\epsilon} G}{\sin(n\pi Y_0)} \frac{\sin(m\tau)}{(1+\bar{\epsilon} \sin^2(m\tau))^{3/4}} \quad (15)$$

which is considered as the fundamental equation whose solution is discussed in the rest of the work.

## 2.1 Solution to the Derived Equation

Observing the periodic coefficients in Eq. (15), it can be written as a non-homogenous nonlinear Hill's equation. To do so, these coefficients are expanded using Fourier series relations. Thus, the Eq. (15) can be written as:

$$\ddot{x}(\tau) + (a_0 + \sum_{j=1}^{\infty} a_j \cos(2jm\tau))x(\tau) + (b_0 + \sum_{j=1}^{\infty} b_j \cos(2jm\tau))x(\tau)^3 = \sum_{j=1}^{\infty} c_j \sin(jm\tau) \quad (16)$$

To solve Eq. (16) three steps are considered. First, the general solution for the linear version of the equation is discussed using the Floquet theory. Then, the pertinent particular solution is obtained using the general solution through the method of variation of constants. Finally, calculating the linear solution, its nonlinear version is addressed via modified Homotopy technique.

### 2.1.1 General Solution

Observing evenness and periodicity of the coefficients in the left hand side of Eq. (16) which have a period of  $\pi$ , the related general solution is considered as:

$$x_g = Ax_1(\tau) + Bx_2(\tau) = Ax_1(\tau) + Bx_1(-\tau) = Ae^{\mu\tau}\Phi(\tau) + Be^{-\mu\tau}\Phi(-\tau) \quad (17)$$

Exponential Fourier expansion of periodic function  $\Phi(\tau)$ , which it also has a period of  $\pi$ ,  $x_1(\tau)$  is written as:

$$x_1(\tau) = e^{\mu\tau}\Phi(\tau) = e^{\mu\tau} \sum_{r=-\infty}^{\infty} \alpha_r e^{2ri\tau} \quad (18)$$

Substituting Eq. (18) into Eq. (16) and assuming  $b_j$ s and  $c_j$ s to be zero, one arrives at a set of infinite homogenous equations which can be presented in a matrix form as the following for the case of  $m = 1$ :

$$[H]\{\alpha\} = \begin{bmatrix} 1 & \frac{\frac{a_1}{2}}{a_0 + (\mu - 4i)^2} & \frac{\frac{a_2}{2}}{a_0 + (\mu - 4i)^2} & \frac{\frac{a_3}{2}}{a_0 + (\mu - 4i)^2} & \frac{\frac{a_4}{2}}{a_0 + (\mu - 4i)^2} \\ \frac{\frac{a_1}{2}}{a_0 + (\mu - 2i)^2} & 1 & \frac{\frac{a_1}{2}}{a_0 + (\mu - 2i)^2} & \frac{\frac{a_2}{2}}{a_0 + (\mu - 2i)^2} & \frac{\frac{a_3}{2}}{a_0 + (\mu - 2i)^2} \\ \frac{\frac{a_2}{2}}{a_0 + \mu^2} & \frac{\frac{a_1}{2}}{a_0 + \mu^2} & 1 & \frac{\frac{a_1}{2}}{a_0 + \mu^2} & \frac{\frac{a_2}{2}}{a_0 + \mu^2} \\ \frac{\frac{a_3}{2}}{a_0 + (\mu + 2i)^2} & \frac{\frac{a_2}{2}}{a_0 + (\mu + 2i)^2} & \frac{\frac{a_1}{2}}{a_0 + (\mu + 2i)^2} & 1 & \frac{\frac{a_1}{2}}{a_0 + (\mu + 2i)^2} \\ \frac{\frac{a_4}{2}}{a_0 + (\mu + 4i)^2} & \frac{\frac{a_3}{2}}{a_0 + (\mu + 4i)^2} & \frac{\frac{a_2}{2}}{a_0 + (\mu + 4i)^2} & \frac{\frac{a_1}{2}}{a_0 + (\mu + 4i)^2} & 1 \end{bmatrix} \begin{Bmatrix} \vdots \\ \alpha_{-2} \\ \alpha_{-1} \\ \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \end{Bmatrix} = \begin{Bmatrix} \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{Bmatrix} \quad (19)$$

To avoid the trivial solution for  $\alpha_r$ s, the determinant of matrix  $[H]$  should become zero. This determinant is the well-known Hill's infinite determinant. Setting  $\det(H) = 0$ , the value of  $\mu$  is calculated considering enough rows and columns of the determinant. In this regard, the following relation is invaluable helpful ([3]):

$$\det(H) = \det(H)|_{\mu=0} - \left( \frac{\sin(\frac{\pi}{2}i\mu)}{\sin(\frac{\pi}{2}\sqrt{a_0})} \right)^2 \quad (20)$$

According to Floquet theory and noting Eq. (18), stability of the general solution is guaranteed if  $\mu$  becomes an imaginary number with zero real part. Besides, elimination of the central row and column of matrix in Eq. (19) helps to evaluate  $\alpha_r|_{r \neq 0}$  in terms of  $\alpha_0$  according to:

$$x_1(\tau) = \alpha_0 (e^{\mu\tau} \sum_{r=-\infty}^{\infty} \bar{\alpha}_r e^{2ri\tau}) \quad (21)$$

Therefore, having obtained the value of  $\mu$ , all  $\bar{\alpha}_r$  parameters in Eq. (21) are known. Since  $\alpha_0$  in Eq. (21) is an arbitrary constant, the general solution of Eq. (16) will be equal to equations (17) and (18), in which  $\alpha_r$  parameters are substituted by known  $\bar{\alpha}_r$  values.

### 2.1.2 Particular Solution to the Derived Equation

If the right hand side of linear Hill's equation is nonzero, one can obtain its particular solution using general functions in Eq. (17) through the method of variation of constants. Therefore the particular solution of the Eq. (16), assuming  $b_j$ s to be zeros, will be equal to:

$$x_p(\tau) = x_2(\tau) \int \frac{x_1(\tau) \left( \sum_{j=1}^{\infty} c_j \sin(jm\tau) \right)}{\delta} d\tau - x_1(\tau) \int \frac{x_2(\tau) \left( \sum_{j=1}^{\infty} c_j \sin(jm\tau) \right)}{\delta} d\tau \quad (22)$$

where  $\delta$  is the Wronskian determinate equal to:

$$\delta = x_1(\tau)x_2'(\tau) - x_2(\tau)x_1'(\tau) = \text{Constant} \neq 0 \quad (23)$$

Taking  $\tau = 0$ , this constant determinate is calculated as:

$$\tau = 0 \Rightarrow \delta = x_1(0)x_2'(0) - x_2(0)x_1'(0) = 2 \left( \sum_{s=-\infty}^{\infty} \alpha_s \right) \left( \sum_{r=-\infty}^{\infty} \alpha_r (\mu + 2ri) \right) \quad (24)$$

Performing the required manipulations, the particular solution to the linear form of Eq. (16) becomes equal to:

$$x_p(\tau) = \frac{1}{\delta} \sum_{j=1}^{\infty} \left[ c_j \sum_{s=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} \alpha_r \alpha_s \left( \frac{1}{2ri+jmi+\mu} - \frac{1}{2si+jmi-\mu} \right) \sin([2(r+s)+jm]\tau) \right] \quad (25)$$

Equations (17) and (25) are sufficient to calculate the solution for the linear non-homogenous Hill's equation. In the following, the modified Homotopy method will be used to evaluate the nonlinear response of the system.

### 2.1.3 Nonlinear Solution (Homotopy)

Introducing an embedment parameter  $\xi$ , the nonlinear operator  $\Psi(\tau; \xi)$  is defined as:

$$\begin{aligned} \Psi(\tau; \xi) = & \ddot{\psi}(\tau; \xi) + (a_0 + \sum_{j=1}^{\infty} a_j \cos(2jm\tau)) \psi(\tau; \xi) + f(\xi) (b_0 + \sum_{j=1}^{\infty} b_j \cos(2jm\tau)) \psi(\tau; \xi)^3 - \\ & \sum_{j=1}^{\infty} c_j \sin(jm\tau) \end{aligned} \quad (26)$$

to create a transformation. In this transformation,  $\psi(\tau; \xi)$  is a function which equals the particular linear solution of Eq. (16) when  $\xi = 0$  and is the corresponding nonlinear solution when  $\xi = 1$ . Therefore, as the value of  $\xi$  starts from zero towards unity,  $f(\xi)$  shall approach unity from zero and also the solution of  $\psi(\tau; \xi)$  deviates from the linear solution of Eq. (16) towards its nonlinear solution. The transformation parameter,  $\psi(\tau; \xi)$  can be expanded around  $\xi = 0$  using Taylor series as the following:

$$\psi(\tau; \xi) = \sum_{k=0}^{\infty} \xi^k X_k(\tau) \quad , \quad X_k(\tau) = \frac{1}{k!} \frac{\partial^k \psi(\tau; \xi)}{\partial \xi^k} \quad (27)$$

Substituting  $\xi = 1$  in Eq. (27), the particular solution of the nonlinear equation (16) is obtained as:

$$x(\tau) = \psi(\tau; 1) = \sum_{k=0}^{\infty} X_k(\tau) \quad (28)$$

To obtain the functions  $X_k(\tau)$  in Eq. (28), one can substitute  $\xi = 0$  in Eq. (26) according to:

$$\xi = 0 : \ddot{X}_0(\tau) + (a_0 + \sum_{j=1}^{\infty} a_j \cos(2jm\tau)) X_0(\tau) = \sum_{j=1}^{\infty} c_j \sin(jm\tau) \quad (29)$$

which yields the function  $X_0$  that is also the particular solution of the linear equation (16), obtained in Eq. (25). Based on Eq. (27),  $\psi(\tau; \xi)$  is differentiable to the required order with respect to

parameter  $\xi$ . Therefore  $\frac{\partial^k \Psi(\tau; \xi)}{\partial \xi^k}$  is analytical for integer values of  $k$ . Differentiating Eq. (26) with respect to  $\xi$  and letting  $\xi = 0$ , leads to:

$$\left. \frac{\partial \Psi(\tau; \xi)}{\partial \xi} \right|_{\xi=0} : \ddot{X}_1(\tau) + (a_0 + \sum_{j=1}^{\infty} a_j \cos(2jm\tau))X_1(\tau) + f'(0)(b_0 + \sum_{j=1}^{\infty} b_j \cos(2jm\tau))X_0(\tau)^3 = 0 \quad (30)$$

which is called the first order deformation equation. Since, function  $X_0$  can be determined from Eq. (29), Eq. (30) is a linear non-homogenous Hill's equation which its solution can be analytically calculated applying Eq. (25). Differentiating Eq. (26) to higher orders of  $\xi$  and letting  $\xi = 0$ , higher order terms of Eq. (28) can be obtained using higher order deformation equations. For instance:

$$\left. \frac{\partial^2 \Psi(\tau; \xi)}{\partial \xi^2} \right|_{\xi=0} : \ddot{X}_2(\tau) + (a_0 + \sum_{j=1}^{\infty} a_j \cos(2jm\tau))X_2(\tau) + \frac{f''(0)}{2}(b_0 + \sum_{j=1}^{\infty} b_j \cos(2jm\tau))X_0(\tau)^3 + 3(b_0 + \sum_{j=1}^{\infty} b_j \cos(2jm\tau))X_0(\tau)^2 X_1(\tau) = 0 \quad (31)$$

leads to a new linear non-homogenous Hill's equation to obtain  $X_2(\tau)$ . Imposing the existence of the required derivatives of  $f(\xi)$  according to the order of deformation equations, and also observing that  $f(0) = 0$  and  $f(1) = 1$ , one can assume  $f(\xi)$  as the following polynomial series:

$$f(\xi) = \sum_{n=1}^N \frac{1}{N} \xi^n \quad (32)$$

where  $N$  shows the highest order of the deformation equation. Adding the calculated functions  $X_0(\tau)$ ,  $X_1(\tau)$ , ... the particular nonlinear solution of Eq. (26) is obtained. The general solution of Eq. (17) is then added to the final solution and the unknown constants A and B are then determined based on the given initial conditions. After the mass leaves the plate completely, the plate vibrates in its free oscillation phase. The free oscillation of the plate can be described by ([7]):

$$Q(\tau) = \gamma \cos \left( \bar{\omega}_0 \left[ 1 + \frac{3\bar{\lambda}}{8} \left( \frac{a}{h} \right)^2 \gamma^2 \right] \tau + \beta_0 \right) \quad (33)$$

In which the parameters  $\gamma$  and  $\beta_0$  can be calculated using the initial conditions. Thus the analytical solution of the moving mass problem is obtained completely. The accuracy of the obtained analytical results is verified using a numerical example. The MATLAB ODE solver which is based on the Runge-Kutta method is used for numerical analysis.

### 3 Numerical Example

A simply supported square plate shown in Fig. 2, with a modulus of elasticity,  $E = 7.1 \times 10^{10} Pa$ , mass density:  $\rho = 2700 kg/m^3$ , length: 2 m, thickness: 1 cm, and the Poisson's ratio:  $\nu = 0.33$ , is considered. The moving mass traverses the plate with constant velocity  $c$  on a straight trajectory passing through the center line, according to Fig.2.

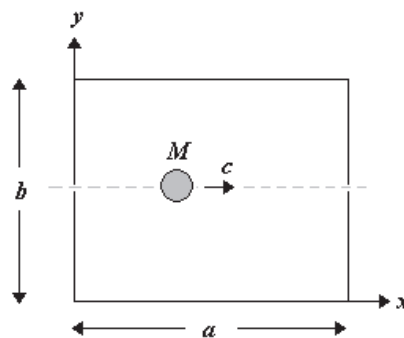
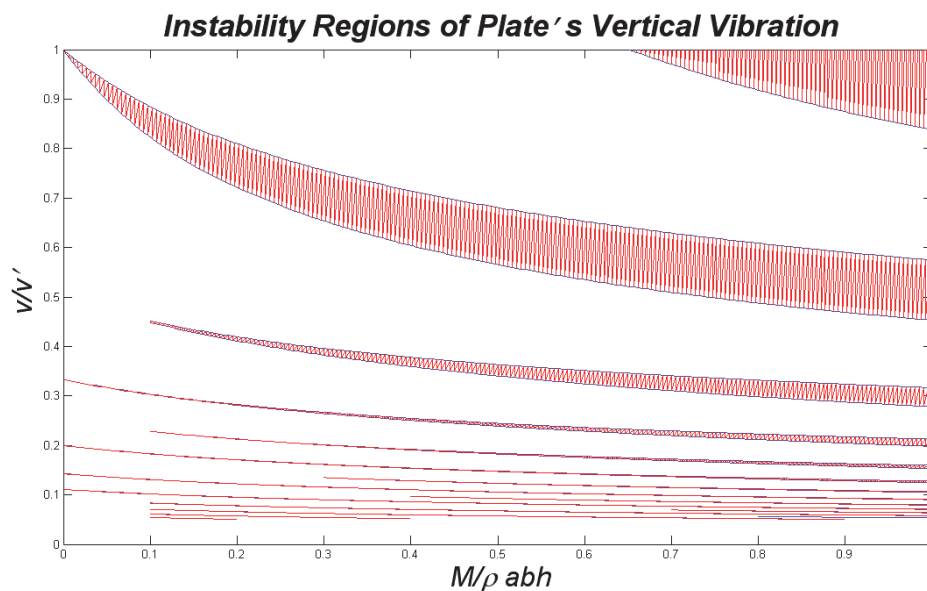


Fig.2. Straight path of the moving mass

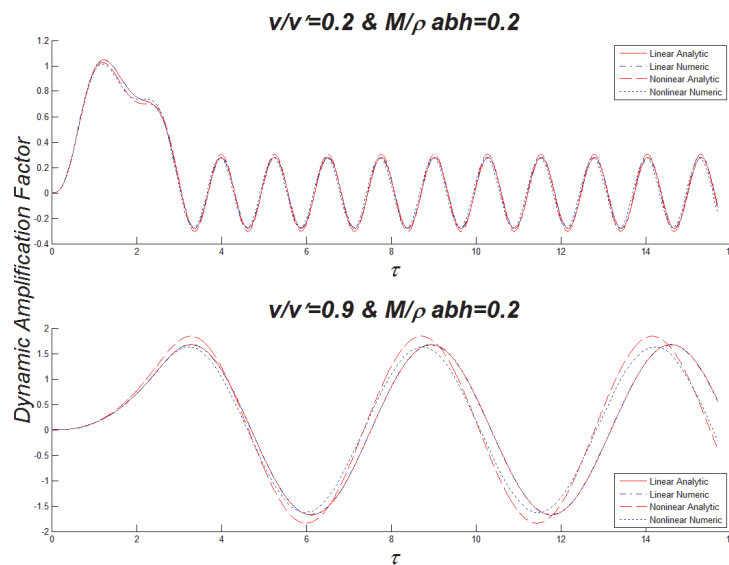
The first general vibrational modal shape of the plate is considered in this example i.e.,  $m$  and  $n$  are equal to one in Eq. (10). Prior to solving the moving mass equation, the instability of plate's vibrations with respect to different moving mass weights and velocities is presented.

Evaluation of the infinite Hill's determinant, specifies the range of the mass weights and velocities for which the parameter  $\mu$  in Eq. (17) to have nonzero real part, leading to instability of the general solution. The regions of instability for the given example are shown by the hatched areas in Fig. 3. The parameter  $v'$  is defined as  $v' = a\omega_0/\pi$ . Based on Fig. 3, some different stable parameters for mass and velocity are chosen to evaluate the related dynamic responses in time domain. Figures 4-6 show the dynamic amplification factors that are presented with respect to non-dimensionalized time,  $\tau$ , for some selected moving mass weights and velocities which cover relatively wide range of possible masses and velocities. The dynamic amplification factor (DAF) is defined as the ratio of the absolute maximum dynamic deflection of the plate to its maximum static response at the center point. The static deflection of the center point of a simply supported square plate under a concentrated mass  $M$ , applied at the same point is equal to  $\Delta_{\text{static}} = 0.0116Mga^2/D$  ([6]).

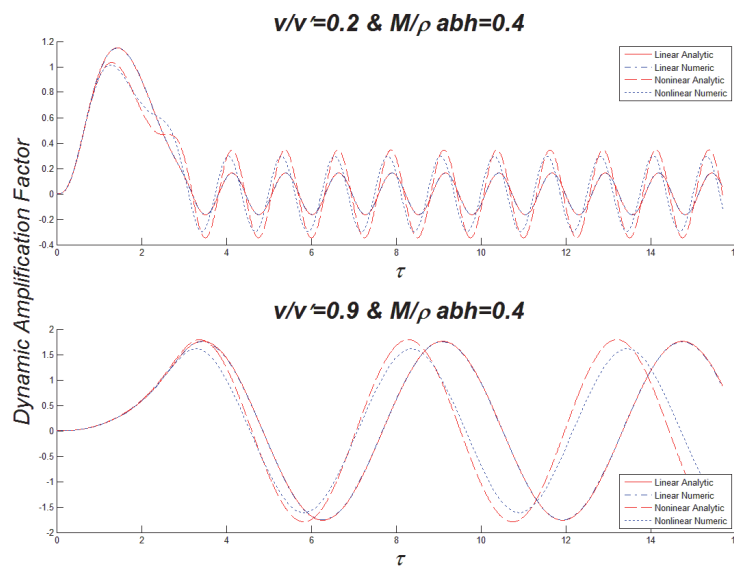


**Fig.3.** Hatched area show the instability regions for different mass and velocity of the moving object

In each of the figures 4-6, the DAF is depicted for two distinct moving mass velocities  $v = 0.2v'$  and  $v = 0.9v'$ . In Fig. 4, the moving object has a mass equal to 0.2 of the plate's mass. Apart from the mass ratio of 0.2 shown in Fig.4, two other mass ratios of 0.4 and 0.6 are selected to generate the Fig. 5 and Fig. 6. The deformation equations up to 2nd order of the Homotopy method are used to produce the nonlinear analytic results shown in Fig. 4 to Fig. 6. Also, it is clearly shown that the analytical solution for the linear version of the main equation completely matches the corresponding numerical result.



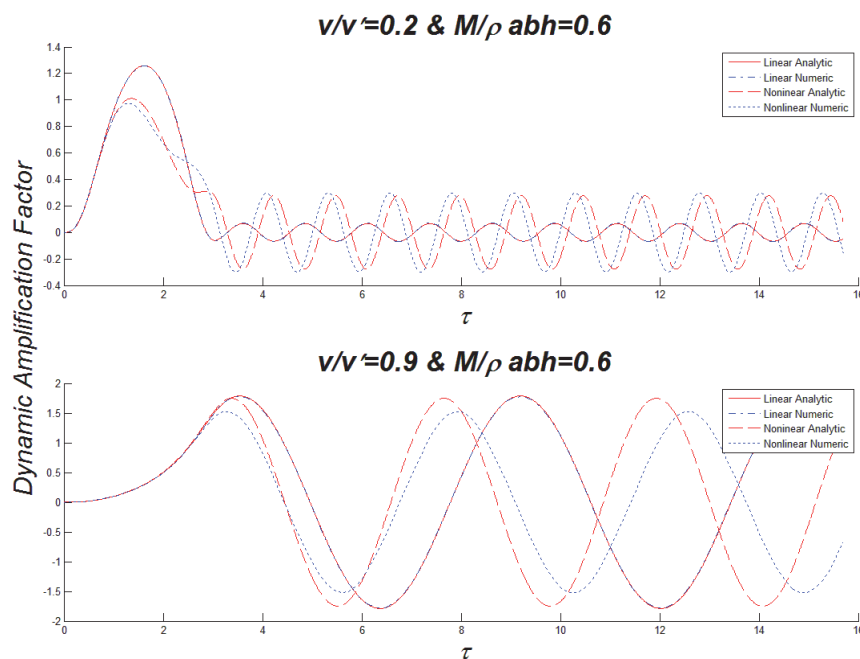
**Fig.4.** Dynamic Amplification Factor in case  $M/\rho abh = 0.2$  for  $v = 0.2 v'$  and  $v = 0.9v'$



**Fig.5.** Dynamic Amplification Factor in case  $M/\rho abh = 0.4$  for  $v = 0.2 v'$  and  $v = 0.9v'$

## 4 Conclusion

The governing differential equation of vibrations of a rectangular plate subjected to a moving mass was derived applying Lagrange method, taking into account the large deformation. This nonlinear equation was analytically solved using the modified Homotopy method. The effect of the velocity and weight of the moving mass on the plate's dynamic response was investigated. The gained results show that the solution for the linear form of the differential equation completely matches the numerical results. When the nonlinearities are involved, the obtained analytical solutions show good agreement with the numerical results for different velocities and weights of the moving mass. It is notable that based on the Homotopy procedure used in this study, the nonlinear analytical results approach the exact nonlinear solutions starting from the linear ones.



**Fig.6.** Dynamic Amplification Factor in case  $M/\rho abh = 0.6$  for  $v = 0.2 v'$  and  $v = 0.9v'$

## References

- [1] G. G. STOKES 1849 Transactions of Cambridge Philosophical Society 8, 707. Discussion of a differential equation relating to the breaking of railway bridges.
- [2] FRYBA, L. 1999. Vibration of Solids and Structures under Moving Loads. Tomas Telford, London.
- [3] WHITTAKER, E.T. and WATSON, G.N., Course of Modern Analysis, 1990 - Cambridge University Press
- [4] F.R. ROFOOEI, A. NIKKHOO, Application of active piezoelectric patches in controlling the dynamic response of a thin rectangular plate under a moving mass, International Journal of Solids and structures, 46 (2009)2429-2443.
- [5] LIAO. S., Beyond Perturbation, Introduction to the Homotopy analysis method, 2004
- [6] TIMOSHENKO, S. WOINOWSKY-KREIGER, 1959. Theory of Plates and Shells, second edition. McGraw-Hill, New York.
- [7] A.H. NAYFEH, D.T. MOOK, "Nonlinear Oscillations", 1995
- [8] V. MARINCA, N. HERIŞANU, Forced Duffing Oscillator With Slight Viscous Damping And Hardening Non-Linearity, Mechanics, Automatic Control and Robotics Vol. 4, No 17, 2005, pp. 245 - 255

## Current Addresses:

**Fayaz R. Rofooei, Ph.D., P.E.,**  
Professor,  
Director of the Earthquake Engineering Research Center,  
Civil Engineering Department,

Sharif University of Technology  
P. O. Box: 11155-9313, Azadi Ave. Tehran, Iran.  
Tel. & Fax: (+98)21-6616-4233  
Email: rofooei@sharif.edu  
Internet: www.sina.sharif.edu/~rofooei

**Alireza Enshaiean, Ph.D Candidate,**  
Civil Engineering Department,  
Sharif University of Technology  
P. O. Box: 11155-9313, Azadi Ave. Tehran, Iran.  
Tel.: (+98)917-317-4013  
Email: enshaiean@mehr.sharif.ir



## SOME PROPERTIES OF SPECIAL DELAYED MATRIX FUNCTIONS IN THEORY OF SYSTEMS OF LINEAR DIFFERENTIAL EQUATIONS WITH CONSTANT COEFFICIENTS AND WITH SINGLE DELAY

Svoboda Zdeněk, (CZ), Josef Diblík, (CZ), Denys Khusainov, (UA)

**Abstract.** The well-known “step by step” method is one of basic concepts for investigation of linear differential equations and systems with delay. A special matrix function - so called delayed matrix exponential - is used for application of this method to linear systems of first order with single constant delay and with constant matrix of linear terms. Special delayed matrix functions are defined on intervals  $(k-1)\tau \leq t < k\tau$ ,  $k = 0, 1, \dots$  (where  $\tau$  is a positive delay) as matrix polynomials, continuous at nets  $t = k\tau$ . This circumstance complicates asymptotic analysis of delayed matrix functions. This contribution deals with asymptotic properties of delayed matrix functions.

**Key words and phrases.** Linear differential equation, delay, matrix function, characteristic equation.

*Mathematics Subject Classification.* Primary 34K06; Secondary 34K25.

### 1 Introduction

Recently, new representations of solutions of linear differential systems with constant coefficients and single constant delay were derived using special delayed matrix functions. E.g., investigation of the structure of solutions of the second-order linear differential systems with constant single delay and with a constant matrix is based on the concepts of so-called delayed matrix cosine and delayed matrix sine. Such special matrix functions are defined in [1, 2]. Analogous results for systems of linear differential equations with single constant delay and a constant matrix are derived using the delayed matrix exponential, for more details we refer, e.g., to [3] and [4]. Analogies of continuous special delayed matrix functions are considered for difference systems as well and relevant results can be found, e.g., in [5]-[8]. This contribution deals with the asymptotic properties of delayed matrix functions.

## 2 Linear Systems of First-Order

Let  $A$  and  $B$  be  $n \times n$  constant matrices,  $\Theta$  be  $n \times n$  null matrix,  $I$  be  $n \times n$  unit matrix and let  $\tau > 0$  be a constant. The delayed matrix exponential of the matrix  $B$  is  $n \times n$  matrix, defined as follows:

$$e_{\tau}^{Bt} = \begin{cases} \Theta, & -\infty < t < -\tau, \\ I, & -\tau \leq t < 0, \\ I + B \frac{t}{1!}, & 0 \leq t < \tau, \\ \dots & \\ I + B \frac{t}{1!} + B^2 \frac{(t-\tau)^2}{2!} + \dots + B^k \frac{(t-(k-1)\tau)^k}{k!} & (k-1)\tau \leq t < k\tau, \\ \dots & \end{cases}$$

where  $k = 0, 1, \dots$ . The main property of the delayed matrix exponential  $e_{\tau}^{Bt}$  is following:

$$(e_{\tau}^{Bt})' = B e_{\tau}^{B(t-\tau)},$$

and the matrix

$$Y(t) = e_{\tau}^{Bt}$$

solves the initial problem for matrix differential system with single delay

$$Y'(t) = BY(t-\tau), \quad t \in [-\tau, \infty),$$

$$Y(t) = I, \quad t \in [-\tau, 0].$$

Let  $\varphi: [-\tau, 0] \rightarrow \mathbb{R}^n$  be continuously differentiable vector-function. Then the solution of the initial-value problem

$$y'(t) = By(t-\tau), \quad t \in [-\tau, \infty),$$

$$y(t) = \varphi(t), \quad t \in [-\tau, 0]$$

is

$$y(t) = e_{\tau}^{Bt} \varphi(-\tau) + \int_{-\tau}^0 e^{B(t-\tau-s)} \varphi'(s) ds.$$

Let matrices  $A$ ,  $B$  commute, i.e.,  $AB = BA$  and let  $A$  be a regular matrix. Then the solution of the initial-value problem

$$y'(t) = Ay(t) + By(t-\tau), \quad t \in [-\tau, \infty),$$

$$y(t) = \varphi(t), \quad t \in [-\tau, 0]$$

can be expressed as

$$y(t) = e^{A(t+\tau)} e_{\tau}^{B_1(t-\tau)} \varphi(-\tau) + \int_{-\tau}^0 e^{A(t-\tau-s)} e_{\tau}^{B_1(t-\tau-s)} e^{A\tau} [\varphi'(s) - A\varphi(s)] ds$$

where

$$B_1 = e^{-A\tau} B.$$

Except this, in [3] the representations of solutions of the non-homogeneous system

$$x'(t) = Ax(t) + Bx(t - \tau) + f(t),$$

where  $f: [0, \infty) \rightarrow \mathbb{R}^n$ , through delayed matrix exponential is studied.

In [9] and [10] is proved that for regular matrix  $B$  there exists a constant matrix  $C$  such that the exponential of the matrix  $e^{Ct}$  has the similar asymptotic properties as the matrix  $e_\tau^{Bt}$ . For such two matrices we have

$$\lim_{t \rightarrow \infty} (e^{Ct} - e_\tau^{Bt}) = 0$$

and

$$\lim_{t \rightarrow \infty} \frac{d}{dt} (e^{Ct} - e_\tau^{Bt}) = 0.$$

If there exists the limit

$$\lim_{n \rightarrow \infty} e_\tau^{B(n+1)\tau} (e_\tau^{Bn\tau})^{-1} = e^{C\tau} \quad (1)$$

and the constant matrix  $C$  has at least one eigenvalue with positive real part, then the exponential of matrix  $C$ , i.e., the matrix  $e^{Ct}$ , is the solution

$$Y(t) = e^{Ct}$$

of the matrix equation

$$Y'(t) = BY(t - \tau)$$

and the matrix  $C$  is a solution of the matrix equation

$$C = Be^{-C\tau}. \quad (2)$$

For integers  $s = 0, 1, \dots$  we have

$$e_\tau^{Bs\tau} = \sum_{k=0}^s \frac{(s+1-k)^k}{k!} B^k \tau^k = B_s(B\tau)$$

where  $B_s$  are polynomials with respect to the matrix  $B$  and the delay  $\tau$ . Let Jordan canonical form of the matrix  $B$  be a diagonal matrix and let modules of the eigenvalues  $\lambda_j$ ,  $j = 1, 2, \dots, n$  of  $B$  be such that the inequalities  $e|\lambda_j|\tau < 1$  are valid. Then there exists a nonsingular matrix  $P$  such that

$$\begin{aligned} e^{-C\tau} &= \lim_{s \rightarrow \infty} e_\tau^{Bs\tau} (e_\tau^{B\tau(s+1)})^{-1} \\ &= \lim_{s \rightarrow \infty} P^{-1} \times \text{diag} \left( \frac{B_s(\lambda_1\tau)}{B_{s+1}(\lambda_1\tau)}, \dots, \frac{B_s(\lambda_n\tau)}{B_{s+1}(\lambda_n\tau)} \right) \times P = \lim_{s \rightarrow \infty} P^{-1} \times \\ &\text{diag} \left( \sum_{k=1}^{s+2} \frac{(-k)^{k-1}}{k!} (\lambda_1\tau)^{k-1} + O((\lambda_1\tau)^{s+3}), \dots, \sum_{k=1}^{s+2} \frac{(-k)^{k-1}}{k!} (\lambda_n\tau)^{k-1} + O((\lambda_n\tau)^{s+3}) \right) \times P \\ &= P^{-1} \times \text{diag} \left( \sum_{k=1}^{\infty} \frac{(-k)^{k-1}}{k!} (\lambda_1\tau)^{k-1}, \dots, \sum_{k=1}^{\infty} \frac{(-k)^{k-1}}{k!} (\lambda_n\tau)^{k-1} \right) \times P \\ &= P^{-1} \times \text{diag} \left( \frac{W_0(\lambda_1\tau)}{\lambda_1\tau}, \dots, \frac{W_0(\lambda_n\tau)}{\lambda_n\tau} \right) \times P \end{aligned}$$

where  $W_0(x)$  is the principal branch of the well-known Lambert  $W$ -function. For the reader's convenience we give its explanation in the following part.

### 3 Lambert function

The Lambert  $W$ -function is useful tool for description of the set of solutions of characteristic equation (2) in the scalar case, i.e., in the case

$$\lambda = be^{-\tau\lambda},$$

which is equivalent to the equation

$$\lambda e^{\tau\lambda} = b.$$

Lambert function (named after Johann Heinrich Lambert, see [11]) is the inverse function to the function

$$f(w) = we^w.$$

The function satisfying

$$z = W(z)e^{W(z)}$$

is a multi-valued (except at  $z = 0$ ). For real arguments  $z = x$ ,  $x > -1/e$  and real  $w$  ( $w > -1$ ) the equation above defines a single-valued function  $W_0(x)$ . The Taylor series of  $W_0(x)$  around 0 is given by

$$W_0(x) = \sum_{n=1}^{\infty} \frac{(-n)^{n-1}}{n!} x^n = x - x^2 + \frac{3}{2}x^3 - \frac{8}{3}x^4 + \frac{125}{24}x^5 - \dots,$$

which has radius  $r$  of convergence  $r = 1/e$ . The Lambert  $W$ -function cannot be expressed in terms of elementary functions.

From the fact that

$$u + iv = W(z) \Rightarrow (u^2 + v^2)e^{2u} = |z|^2$$

follows that for any constant  $|z|$  and any couple of values  $(u_k, v_k), (u_l, v_l)$  such that

$$u_k + iv_k = W_k(z),$$

$$u_l + iv_l = W_l(z),$$

the implication

$$v_k^2 < v_l^2 \Rightarrow u_k > u_l$$

holds and so, the inequality

$$\Re W_k(z) > \Re W_l(z)$$

holds, too.  $\Re W_0(z)$  is the greatest real part of all real values  $W(z)$ . If the matrix  $B$  has the diagonal Jordan canonical form with eigenvalues  $\lambda_j$ ,  $j = 1, 2, \dots, n$  satisfying the inequalities  $e\lambda_j\tau < 1$ , then the matrix function  $e^{Ct}$ , where the matrix  $C$  is defined by (1), bounds exponentials  $e^{\bar{C}t}$  of other matrices  $\bar{C}$ . For more details see [12].

The specification of the set of complex numbers such that values of real part of the Lambert function are non-positive yields possibility to determine asymptotic properties of above cited

delayed matrix functions. The set of complex numbers for which the real part of the Lambert function equals zero ( $u = 0$ ) is defined in the parametric form

$$\begin{aligned}x &= -v \sin v, \\y &= v \cos v.\end{aligned}$$

Then

$$(0 + iv) \exp(0 + iv) = iv \exp(iv) = x + iy$$

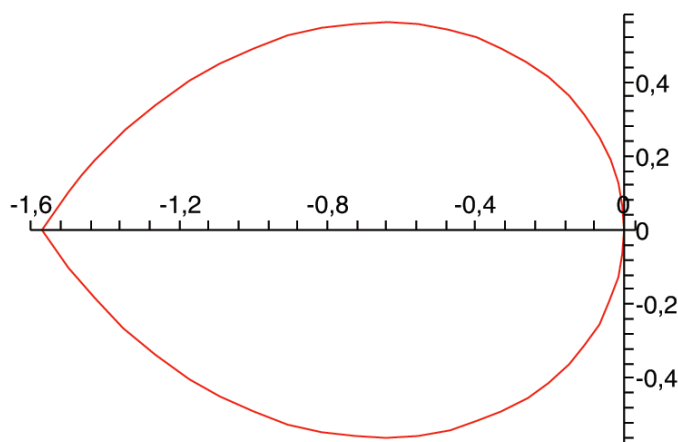
and

$$\Re W(x + iy) = 0.$$

The part of this curve corresponding the principal branch  $W_0(x + iy)$  is a simple closed curve for a parameter  $v \in [-\pi/2, \pi/2]$  and for numbers from the interior of this curve the value of the real part of the principal part of the Lambert function is negative. This set is specified as solution of the inequality:

$$\sqrt{x^2 + y^2} < -\arctan\left(\frac{x}{|y|}\right)$$

(see the figure below).



For more details see [11].

## 4 Main result

In this part we give an estimation of the Euclidean norm  $\|e_\tau^{Bt}\|$  of the delayed matrix exponential of the matrix  $B$ . We assume that the matrix  $B$  has eigenvalues with coinciding geometric and algebraic multiplicities only.

**Theorem 4.1** Assume that the matrix  $B$  has eigenvalues

$$\lambda_s = x_s + iy_s, \quad s = 1, \dots, n$$

with coinciding geometric and algebraic multiplicities only. Let moreover inequalities

1.  $e\tau\sqrt{x_s^2 + y_s^2} < 1, \quad s = 1, \dots, n,$
2.  $\tau\sqrt{x_s^2 + y_s^2} < -\arctan\left(\frac{x_s}{|y_s|}\right), \quad s = 1, \dots, n$

are valid. Then for the Euclidean norm of the delayed exponential of matrix we have

$$\lim_{t \rightarrow \infty} \|e_\tau^{Bt}\| = 0.$$

**Proof.** The first inequality implies the existence the of the constant matrix  $C$  such that

$$P^{-1} \times \text{diag}\left(\frac{W_0(\lambda_1\tau)}{\lambda_1\tau}, \dots, \frac{W_0(\lambda_n\tau)}{\lambda_n\tau}\right) \times P = e^{-C\tau} = \lim_{n \rightarrow \infty} e_\tau^{B\tau n} (e_\tau^{B\tau(n+1)})^{-1}.$$

This limit equality we may for any  $\varepsilon > 0$ , any positive integer  $k$  and sufficiently large  $n$  rewrite in the form:

$$\|e_\tau^{B\tau(n+k)} - e_\tau^{B\tau n} e^{Ck\tau}\| < \varepsilon$$

where  $\|\cdot\|$  is Euclidean norm. Therefore the matrix  $C$  is a matrix solution of the characteristic equation (2). We obtain

$$C = Be^{-C\tau} = P^{-1} \times \text{diag}\left(\frac{W_0(\lambda_1\tau)}{\tau}, \dots, \frac{W_0(\lambda_n\tau)}{\tau}\right) \times P$$

The second inequality says that real parts of eigenvalues of the matrix  $C$  are negative and the Euclidean norm of the matrix satisfies inequality  $\|e^{C\tau}\| < 1$ . The proof is complete.

## 5 Conclusion

In this contribution some properties of special matrix delayed functions were considered. Such functions can successfully be used to represent solutions of linear differential equations and systems with the single delay. It was shown, that in some cases there exists non-delayed matrix functions having the same asymptotic properties as delayed matrix functions. Investigation in this way can lead to further interesting applications in the field of linear differential equations with a single delay. We refer, e.g., to recent papers [13]-[15].

## Acknowledgement

The research was supported by the Grants P201/10/1032 and P201/11/0768 of the Czech Grant Agency (Prague), and by the Grant FEKT-S-11-2-921 of Faculty of Electrical Engineering and Communication, Brno University of Technology.

# References

- [1] J. DIBLÍK, D. YA. KHUSAINOV, M. RŮŽIČKOVÁ, J. LUKÁČOVÁ: *Control of Oscillating Systems with a Single Delay*, Advances in Difference Equations, Volume 2010, Article ID 108218, 15 pages, 2010.
- [2] D. YA. KHUSAINOV, J. DIBLÍK, J. LUKÁČOVÁ M. RŮŽIČKOVÁ: *Representation of a solution of the Cauchy problem for an oscillating system with pure delay*, Nonlinear Oscillations **11**, No 2, pp. 276–285, 2008.
- [3] D. YA. KHUSAINOV, G. V. SHUKLIN: *Linear autonomous time-delay system with permutation matrices solving*, Studies of the University of Žilina, Mathematical Series **16**, 1-8, 2003.
- [4] A. BOICHUK, J. DIBLÍK, D. KHUSAINOV, M. RŮŽIČKOVÁ: *Fredholm's boundary-value problems for differential systems with a single delay*. Nonlinear Analysis **72**, pp. 2251–2258, 2010.
- [5] J. DIBLÍK, D. YA. KHUSAINOV, M. RŮŽIČKOVÁ: *Controllability of linear discrete systems with constant coefficients and pure delay* SIAM J. Control Optim. **47**, No. 3, pp. 1140–1149, 2008.
- [6] J. DIBLÍK, D. KHUSAINOV: *Representation of solutions of discrete delayed system  $x(k+1) = Ax(k) + Bx(k-m) + f(k)$  with commutative matrices*, Journal of Mathematical Analysis and Applications **318**, No 1, 63–76, 2006.
- [7] J. DIBLÍK, D. KHUSAINOV: *Representation of solutions of linear discrete systems with constant coefficients and pure delay*, Advances in Difference Equations, Art. ID 80825, DOI 10.1155/ADE/2006/80825, pp. 1–13, 2006.
- [8] A. BOICHUK, J. DIBLÍK, D. KHUSAINOV, M. RŮŽIČKOVÁ: *Boundary value problems for delay differential systems*, Advances in Difference Equations, Volume 2010, Article ID 593834, 20 pages.
- [9] Z. SVOBODA: *Asymptotic properties of delayed exponential of matrix*, Journal of Applied Mathematics, Slovak University of Technology in Bratislava, 167–172, 2010.
- [10] Z. SVOBODA: *The system of linear differential equations with constant coefficients and constant delay*, XXVIII International Colloquium on the Management of Educational Process, Proceedings, Brno, 247–251, 2010.
- [11] J. H. LAMBERT: *Observationes variae in mathesin puram*. Acta Helveticae physicomathematico- anatomico-botanico-medica, Band III, pp. 128–168, 1758.
- [12] R. M. CORLESS, G. H. GONNET, D. E. G. HARE, D. E. KNUTH: *On the Lambert W Function* Advances in Computational Mathematics, **Vol 5** , 329–359, 1996.
- [13] M. MEDVEĎ, M. POSPÍŠIL, L. ŠKRIPKOVÁ: *Stability and the nonexistence of blowing-up solutions of nonlinear delay systems with linear parts defined by permutable matrices*, Nonlinear Analysis, **74** (2011), 3903–3911.
- [14] J. DIBLÍK, O. KHUKHARENKO, D. KHUSAINOV: *Representation of solution of the first boundary value problem for delay systems*, Bull. Kiev University, Series: Physics & Mathematics, No 1 (2011), 59–62.
- [15] A. BOICHUK, J. DIBLÍK, D. KHUSAINOV, M. RŮŽIČKOVÁ: *Boundary-value problems for weakly nonlinear delay differential systems*, Abstract and Applied Analysis, **vol. 2011**, Article ID 631412, 19 pages, 2011. doi:10.1155/2011/631412.

**Current address**

**Zdeněk Svoboda, RNDr. CSc.**

Department of Mathematics, Faculty of Electrical Engineering and Communication, Brno University of Technology, Technická 3058/10, 61600 Brno, Česká republika, tel. 420541142830  
svobodaz@feec.vutbr.cz

**Josef Diblík, prof., RNDr. DrSc.**

Department of Mathematics and Descriptive Geometry, Faculty of Civil Engineering, Brno University of Technology, Veverí 331/95, 60200 Brno, Česká republika, tel. 420541147601  
diblik.j@fce.vutbr.cz,

Department of Mathematics, Faculty of Electrical Engineering and Communication, Brno University of Technology, Technická 3058/10, 61600 Brno Česká republika, tel. 420541143130  
diblik@feec.vutbr.cz

**Denys Khusainov, prof., DrSc.**

Department of Complex Systems Modeling, Kiev University, 01033 Kiev, Ukraine  
khusainov@unicyb.kiev.ua

# SOME GENERALIZATIONS IN THEORY OF RAPID VARIATION ON TIME SCALES AND ITS APPLICATION IN DYNAMIC EQUATIONS

VÍTOVEC Jiří, (CZ)

**Abstract.** In this paper we introduce a new definition of rapidly varying function on time scales. Unlike the recently studied concept of rapid variation, this new concept is more general and naturally extends and complements the already established class of rapidly varying functions. We prove some of its properties and show the relation between this new type of definition and recently introduced “classical” Karamata type of definition of rapid variation on time scales. Note that the theory of rapid variation on time scales unifies the existing theories from continuous and discrete cases. As an application, we establish necessary and sufficient conditions for all positive solutions of the second order half-linear dynamic equations on time scales to be rapidly varying.

**Key words and phrases.** Rapidly varying function, regularly varying function, regularly bounded function, time scale, half-linear dynamic equation.

*Mathematics Subject Classification.* Primary 26A12, 26A99, 26E70; Secondary 34N05.

## 1 Introduction

Recall that a measurable function  $f : [a, \infty) \rightarrow (0, \infty)$  of a real variable is said to be *rapidly varying of index  $\infty$* , resp. *of index  $-\infty$*  if it satisfies

$$\lim_{x \rightarrow \infty} \frac{f(\lambda x)}{f(x)} = \begin{cases} \infty \text{ resp. } 0 & \text{for } \lambda > 1, \\ 0 \text{ resp. } \infty & \text{for } 0 < \lambda < 1; \end{cases} \quad (1)$$

we write  $f \in \mathcal{RPV}_{\mathbb{R}}(\infty)$ , resp.  $f \in \mathcal{RPV}_{\mathbb{R}}(-\infty)$ . For more information about the rapid variation on  $\mathbb{R}$ , see for example [1] and references therein.

In [8], the concept of rapidly varying sequences was introduced in the following way. Let  $[u]$  denote the integer part of  $u$ . A positive sequence  $\{f_k\}$ ,  $k \in \{a, a+1, \dots\} \subset \mathbb{Z}$  is said to be *rapidly varying of index  $\infty$* , resp. *of index  $-\infty$*  if it satisfies

$$\lim_{k \rightarrow \infty} \frac{f_{[\lambda k]}}{f_k} = \begin{cases} \infty \text{ resp. } 0 & \text{for } \lambda > 1, \\ 0 \text{ resp. } \infty & \text{for } 0 < \lambda < 1; \end{cases} \quad (2)$$

we write  $f \in \mathcal{RPV}_{\mathbb{Z}}(\infty)$ , resp.  $f \in \mathcal{RPV}_{\mathbb{Z}}(-\infty)$ . Note that these types of definitions of rapidly varying functions (1) and rapidly varying sequences (2), which include a parameter  $\lambda$ , correspond to the classical Karamata type definition of regularly varying functions, see [1, 6, 7, 15] and references therein. In [8] it was shown that if a positive sequence  $\{f_k\}$  has the property that  $\Delta f_k$  increases, then  $f \in \mathcal{RPV}_{\mathbb{Z}}(-\infty)$  if and only if

$$\lim_{k \rightarrow \infty} \frac{k \Delta f_k}{f_k} = -\infty. \quad (3)$$

This result shows that under certain conditions there exists an alternative (in some cases more practical) possibility, how to define rapidly varying sequences (resp. functions). For further reading of rapid and regular variation in discrete case we refer, e.g., to [3, 4, 8, 9, 10] and the references therein.

In [16], we introduced the concept of rapidly varying functions on time scales (i.e., considered functions are defined on nonempty closed subsets of  $\mathbb{R}$ ) in these two following ways:

**Definition 1.1** *A measurable function  $f : \mathbb{T} \rightarrow (0, \infty)$  is said to be rapidly varying of index  $\infty$ , resp. of index  $-\infty$  if there exist bounded (from below and above the positive constants) function or regularly varying function  $\varphi : \mathbb{T} \rightarrow (0, \infty)$  and a positive rd-continuously  $\Delta$ -differentiable function  $\omega$  such that  $f(t) = \varphi(t)\omega(t)$  and*

$$\lim_{t \rightarrow \infty} \frac{t\omega^{\Delta}(t)}{\omega(t)} = \infty, \quad \text{resp.} \quad \lim_{t \rightarrow \infty} \frac{t\omega^{\Delta}(t)}{\omega(t)} = -\infty; \quad (4)$$

*we write  $f \in \mathcal{RPV}_{\mathbb{T}}(\infty)$ , resp.  $f \in \mathcal{RPV}_{\mathbb{T}}(-\infty)$ . Moreover, the function  $\omega$  is said to be normalized rapidly varying of index  $\infty$ , resp. normalized rapidly varying of index  $-\infty$ ; we write  $\omega \in \mathcal{NRPV}_{\mathbb{T}}(\infty)$ , resp.  $\omega \in \mathcal{NRPV}_{\mathbb{T}}(-\infty)$ .*

**Definition 1.2 (Karamata type definition)** *Let  $\tau : \mathbb{R} \rightarrow \mathbb{T}$  be defined as  $\tau(t) = \max\{s \in \mathbb{T} : s \leq t\}$ . A measurable function  $f : \mathbb{T} \rightarrow (0, \infty)$  satisfying*

$$\lim_{t \rightarrow \infty} \frac{f(\tau(\lambda t))}{f(t)} = \begin{cases} \infty \text{ resp. } 0 & \text{for } \lambda > 1, \\ 0 \text{ resp. } \infty & \text{for } 0 < \lambda < 1 \end{cases} \quad (5)$$

*is said to be rapidly varying of index  $\infty$ , resp. of index  $-\infty$  in the sense of Karamata. We write  $f \in \mathcal{KRPV}_{\mathbb{T}}(\infty)$ , resp.  $f \in \mathcal{KRPV}_{\mathbb{T}}(-\infty)$ .*

We studied properties of these two definitions, where the earlier one was motivated by (3).

In this paper we improve and generalize the Definition 1.1 to get a “wider” class of rapidly varying functions. Then we show their representation, properties and relation between this

new definition and Karamata type definition. The theory of rapid variation on time scales automatically holds for the continuous and discrete cases, moreover, at the same time, the theory works also on other time scales which may be different from the “classical” ones. Finally, note that the theory of rapid variation on time scales naturally extends and completes our knowledge concerning the theory of regular variation on time scales, which was earlier studied in [12, 14].

As an application, we study asymptotic properties of solutions of the second order half-linear dynamic equation

$$[\Phi(x^\Delta)]^\Delta - p(t)\Phi(x^\sigma) = 0 \quad (6)$$

on a time scale, where  $p > 0$  is an rd-continuous function, and  $\Phi(x) = |x|^{\alpha-1} \operatorname{sgn} x$ ,  $\alpha > 1$ . We show that the central theorem mentioned in [16] is also fulfilled for the newly defined class of rapidly varying functions.

In this paper, the time scale  $\mathbb{T}$  is assumed to be unbounded above,  $\min \mathbb{T} = a$  (with  $a > 0$ ) and the graininess satisfies  $\mu(t) = o(t)$ . This condition was precisely discussed in [16]. As was shown there, if we want to obtain a reasonable theory, we cannot omit this additional requirement on the graininess.

## 2 Preliminaries

We assume that the reader is familiar with the notion of time scales. Thus note just that  $\mathbb{T}$ ,  $\sigma$ ,  $\rho$ ,  $f^\sigma$ ,  $\mu$ ,  $f^\Delta$ ,  $\int_a^b f^\Delta(s) \Delta s$ ,  $\mathcal{R}^+$  and  $e_p(t, a)$  stand for the time scale, forward jump operator, backward jump operator,  $f \circ \sigma$ , graininess,  $\Delta$ -derivative of  $f$ ,  $\Delta$ -integral of  $f$  from  $a$  to  $b$ , class of positive regressive function and generalized exponential function, respectively. See [5], which is the initiating paper of the time scale theory, and [2] containing a lot of information on time scale calculus.

In [12], the concept of regular variation on  $\mathbb{T}$  was introduced in the following way. A measurable function  $f : \mathbb{T} \rightarrow (0, \infty)$  is said to be *regularly varying of index*  $\vartheta$ ,  $\vartheta \in \mathbb{R}$ , if there exists a positive rd-continuously  $\Delta$ -differentiable function  $g$  satisfying

$$f(t) \sim Cg(t) \quad \text{and} \quad \lim_{t \rightarrow \infty} \frac{tg^\Delta(t)}{g(t)} = \vartheta, \quad (7)$$

$C$  being a positive constant; we write  $f \in \mathcal{RV}_\mathbb{T}(\vartheta)$ . If  $\vartheta = 0$ , then  $f$  is said to be *slowly varying*; we write  $f \in \mathcal{SV}_\mathbb{T}$ . Moreover, the function  $g$  is said to be *normalized regularly varying of index*  $\vartheta$ ; we write  $g \in \mathcal{N}\mathcal{RV}_\mathbb{T}(\vartheta)$ . If  $\vartheta = 0$ , then  $g$  is said to be *normalized slowly varying*; we write  $g \in \mathcal{N}\mathcal{SV}_\mathbb{T}$ . In [14], we introduced a Karamata type definition of regularly varying function on time scales and developed and enriched the existing theory with new statements (the embedding theorem, a relation between previous and Karamata type definition, etc.). Here is the Karamata type definition. Let  $f : \mathbb{T} \rightarrow (0, \infty)$  be a measurable function satisfying

$$\lim_{t \rightarrow \infty} \frac{f(\tau(\lambda t))}{f(t)} = \lambda^\vartheta \quad (8)$$

uniformly on each compact  $\lambda$ -set in  $(0, \infty)$ , where  $\tau$  is defined as in Definition 1.2. Then  $f$  is said to be *regularly varying of index*  $\vartheta$  ( $\vartheta \in \mathbb{R}$ ) *in the sense of Karamata*; we write  $f \in \mathcal{K}\mathcal{RV}_\mathbb{T}(\vartheta)$ . If  $\vartheta = 0$ , then  $f$  is said to be *slowly varying in the sense of Karamata*; we write  $f \in \mathcal{K}\mathcal{SV}_\mathbb{T}$ . For further information about theory of regular variation on  $\mathbb{T}$  see, e.g., [12, 13, 14].

### 3 Theory of rapid variation on time scales

Before we introduce a new definition of rapidly varying function, we established (analogously as in discrete and continuous case) a concept of regularly bounded function on time scale. This concept can be viewed as a generalization of regular variation on time scales in the sense that the limit in (8) may not exist, but the expression in them still exhibit a “moderate” behavior. Recall that throughout the paper,  $\mathbb{T}$  is assumed to be unbounded above,  $\min \mathbb{T} = a$  (with  $a > 0$ ) and  $\mu(t) = o(t)$ .

**Definition 3.1** A measurable function  $f : \mathbb{T} \rightarrow (0, \infty)$  is said to be regularly bounded if it satisfies

$$0 < \liminf_{t \rightarrow \infty} \frac{f(\tau(\lambda t))}{f(t)} \leq \limsup_{t \rightarrow \infty} \frac{f(\tau(\lambda t))}{f(t)} < \infty \quad \text{for all } \lambda > 0,$$

where  $\tau$  is defined as in Definition 1.2; we write  $f \in \mathcal{RB}_{\mathbb{T}}$ .

It is clear that  $\cup_{\vartheta \in \mathbb{R}} \mathcal{KR}\mathcal{V}_{\mathbb{T}}(\vartheta) \subset \mathcal{RB}_{\mathbb{T}}$ . Moreover, it is not difficult to show (see [14, proof of Theorem 2]) that rd-continuously  $\Delta$ -differentiable function  $f$  satisfying relation

$$-\infty < \liminf_{t \rightarrow \infty} \frac{tf^{\Delta}(t)}{f(t)} \leq \limsup_{t \rightarrow \infty} \frac{tf^{\Delta}(t)}{f(t)} < \infty$$

is regularly bounded. Now we are ready to introduce a new definition of rapidly varying function (compare with Definition 1.1).

**Definition 3.2** A measurable function  $f : \mathbb{T} \rightarrow (0, \infty)$  is said to be rapidly varying of index  $\infty$ , resp. of index  $-\infty$  if there exist regularly bounded function  $\varphi : \mathbb{T} \rightarrow (0, \infty)$  and a positive rd-continuously  $\Delta$ -differentiable function  $\omega$  such that  $f(t) = \varphi(t)\omega(t)$  and

$$\lim_{t \rightarrow \infty} \frac{t\omega^{\Delta}(t)}{\omega(t)} = \infty, \quad \text{resp.} \quad \lim_{t \rightarrow \infty} \frac{t\omega^{\Delta}(t)}{\omega(t)} = -\infty; \quad (9)$$

we write  $f \in \mathcal{RPV}_{\mathbb{T}}(\infty)$ , resp.  $f \in \mathcal{RPV}_{\mathbb{T}}(-\infty)$ . Moreover, the function  $\omega$  is said to be normalized rapidly varying of index  $\infty$ , resp. normalized rapidly varying of index  $-\infty$ ; we write  $\omega \in \mathcal{NRPV}_{\mathbb{T}}(\infty)$ , resp.  $\omega \in \mathcal{NRPV}_{\mathbb{T}}(-\infty)$ .

From the above definition it is easy to see that the function  $f(t) = a^t$  with  $a > 1$  is a typical representative of the class  $\mathcal{RPV}_{\mathbb{T}}(\infty)$ , while the function  $f(t) = a^t$  with  $a \in (0, 1)$  is a typical representative of the class  $\mathcal{RPV}_{\mathbb{T}}(-\infty)$ . Of course, as we can see also from the following theorem, these classes are much wider. Using elementary properties of linear first order dynamic equations and generalized exponential functions  $e_{\delta}(t, s)$ , we establish the following representation.

**Theorem 3.3 (Representation theorem)** A positive function  $f \in C_{\text{rd}}(\mathbb{T})$  belongs to the class  $\mathcal{RPV}_{\mathbb{T}}(\infty)$ , resp. to the class  $\mathcal{RPV}_{\mathbb{T}}(-\infty)$  if and only if it has a representation

$$f(t) = \psi(t)e_{\delta}(t, a), \quad (10)$$

where function  $\psi : \mathbb{T} \rightarrow (0, \infty)$  is regularly bounded and  $\delta \in \mathcal{R}^+(\mathbb{T})$  satisfies  $\lim_{t \rightarrow \infty} t\delta(t) = \infty$ , resp.  $\lim_{t \rightarrow \infty} t\delta(t) = -\infty$ .

**Proof.** “Only if”: Let  $f \in \mathcal{RPV}_{\mathbb{T}}(\infty)$ , resp.  $f \in \mathcal{RPV}_{\mathbb{T}}(-\infty)$ . Then there is  $\delta \in C_{\text{rd}}(\mathbb{T})$  such that  $\delta = \omega^{\Delta}/\omega$  and  $\lim_{t \rightarrow \infty} t\delta(t) = \infty$ , resp.  $\lim_{t \rightarrow \infty} t\delta(t) = -\infty$ . Moreover,  $\omega$  satisfies the first order linear dynamic equation  $\omega^{\Delta} = \delta(t)\omega$  and hence it has the form  $\omega(t) = \omega_0 e_{\delta}(t, a)$  with  $\omega_0 > 0$ . Since  $\omega$  is positive, we have for every right-scattered  $t$  (for right-dense  $t$  the following inequality holds trivially)

$$1 + \mu(t)\delta(t) = 1 + \mu(t)\frac{\omega^{\Delta}(t)}{\omega(t)} = 1 + \frac{\omega^{\sigma}(t) - \omega(t)}{\omega(t)} = \frac{\omega^{\sigma}(t)}{\omega(t)} > 0$$

and hence  $\delta \in \mathcal{R}^+(\mathbb{T})$ . In view of  $f(t) = \varphi(t)\omega(t) = \omega_0\varphi(t)e_{\delta}(t, a)$  and of the fact that  $\psi$  is defined as  $\varphi$ , (10) holds.

“If”: Let (10) hold with  $\delta \in \mathcal{R}^+(\mathbb{T})$  and  $\lim_{t \rightarrow \infty} t\delta(t) = \infty$ , resp.  $\lim_{t \rightarrow \infty} t\delta(t) = -\infty$ . Put  $\omega(t) = e_{\delta}(t, a)$ . Then  $\omega$  is positive function such that  $\lim_{t \rightarrow \infty} t\omega^{\Delta}(t)/\omega(t) = \lim_{t \rightarrow \infty} t\delta(t) = \infty$ , resp.  $\lim_{t \rightarrow \infty} t\omega^{\Delta}(t)/\omega(t) = \lim_{t \rightarrow \infty} t\delta(t) = -\infty$ . Since  $f(t) = \psi(t)\omega(t)$ ,  $f \in \mathcal{RPV}_{\mathbb{T}}(\infty)$ , resp.  $f \in \mathcal{RPV}_{\mathbb{T}}(-\infty)$ .

**Proposition 3.4** (i) It holds  $f \in \mathcal{RPV}_{\mathbb{T}}(\infty)$  if and only if  $1/f \in \mathcal{RPV}_{\mathbb{T}}(-\infty)$ .

(ii) Let  $f \in \mathcal{NRPV}_{\mathbb{T}}(\infty)$ . Then for every  $\vartheta \in [0, \infty)$  the function  $f(t)/t^{\vartheta}$  is increasing for large  $t$  and  $\lim_{t \rightarrow \infty} f(t)/t^{\vartheta} = \infty$ .

(iii) Let  $f \in \mathcal{NRPV}_{\mathbb{T}}(-\infty)$ . Then for every  $\vartheta \in [0, \infty)$  the function  $f(t)t^{\vartheta}$  is decreasing for large  $t$  and  $\lim_{t \rightarrow \infty} f(t)t^{\vartheta} = 0$ .

(iv)  $f \in \mathcal{NRPV}_{\mathbb{T}}(\infty)$  implies  $f^{\Delta}(t) > 0$  for large  $t$  and  $f(t)$  is increasing for large  $t$ , moreover  $f$  and  $f^{\Delta}$  are tending to  $\infty$ .

(v)  $f \in \mathcal{NRPV}_{\mathbb{T}}(-\infty)$  implies  $f^{\Delta}(t) < 0$  for large  $t$  and  $f(t)$  is decreasing for large  $t$ , moreover  $f$  is tending to 0. If  $f$  is convex for large  $t$  or if there exists  $h > 0$  such that  $\mu(t) > h$  for large  $t$ , then  $f^{\Delta}$  is tending to 0.

**Proof.** (i) Let  $f \in \mathcal{RPV}_{\mathbb{T}}(\infty)$ ,  $f = \varphi\omega$ . First, we show that  $\omega \in \mathcal{NRPV}_{\mathbb{T}}(\infty) \Leftrightarrow 1/\omega \in \mathcal{NRPV}_{\mathbb{T}}(-\infty)$ . Due to (9),  $\omega^{\Delta}(t) > 0$  for large  $t$ . Therefore,

$$\begin{aligned} \omega \in \mathcal{NRPV}_{\mathbb{T}}(\infty) &\Leftrightarrow \lim_{t \rightarrow \infty} \frac{\omega(t)}{t\omega^{\Delta}(t)} = 0 \Leftrightarrow \lim_{t \rightarrow \infty} \frac{\omega^{\sigma}(t) - \mu(t)\omega^{\Delta}(t)}{t\omega^{\Delta}(t)} = 0 \\ &\Leftrightarrow \lim_{t \rightarrow \infty} \left( \frac{\omega^{\sigma}(t)}{t\omega^{\Delta}(t)} - \frac{\mu(t)}{t} \right) = 0 \Leftrightarrow \lim_{t \rightarrow \infty} \frac{\omega^{\sigma}(t)}{t\omega^{\Delta}(t)} = 0 \\ &\Leftrightarrow \lim_{t \rightarrow \infty} \frac{t\omega^{\Delta}(t)}{\omega^{\sigma}(t)} = \infty \Leftrightarrow \lim_{t \rightarrow \infty} \left( \frac{t}{1/\omega(t)} \cdot \frac{-\omega^{\Delta}(t)}{\omega(t)\omega^{\sigma}(t)} \right) = -\infty \\ &\Leftrightarrow \lim_{t \rightarrow \infty} \frac{t(1/\omega(t))^{\Delta}}{1/\omega(t)} = -\infty \Leftrightarrow \frac{1}{\omega} \in \mathcal{NRPV}_{\mathbb{T}}(-\infty). \end{aligned}$$

Now, since  $\varphi \in \mathcal{RB}_{\mathbb{T}}$ , we have

$$0 < \liminf_{t \rightarrow \infty} \frac{\varphi(t)}{\varphi(\tau(\lambda t))} \leq \limsup_{t \rightarrow \infty} \frac{\varphi(t)}{\varphi(\tau(\lambda t))} < \infty$$

and hence  $1/\varphi \in \mathcal{RB}_{\mathbb{T}}$ . Therefore  $1/f \in \mathcal{RPV}_{\mathbb{T}}(-\infty)$ . Similarly,  $1/f \in \mathcal{RPV}_{\mathbb{T}}(-\infty)$  implies  $f \in \mathcal{RPV}_{\mathbb{T}}(\infty)$ .

The proof of part (ii) - (v) we can find in [16].

The following proposition summarizes some important properties of rapidly varying functions in the sense of Karamata. The proof of this proposition we can find in [16].

**Proposition 3.5** (I)  $f \in \mathcal{KRPV}_{\mathbb{T}}(\infty)$  if and only if  $1/f \in \mathcal{KRPV}_{\mathbb{T}}(-\infty)$ .

(II) Let  $f : \mathbb{T} \rightarrow (0, \infty)$  be a measurable function, monotone for large  $t$ . Then

- (i)  $f \in \mathcal{KRPV}_{\mathbb{T}}(\infty)$  implies  $f$  is increasing for large  $t$  and  $\lim_{t \rightarrow \infty} f(t) = \infty$ .
- (ii)  $f \in \mathcal{KRPV}_{\mathbb{T}}(-\infty)$  implies  $f$  is decreasing for large  $t$  and  $\lim_{t \rightarrow \infty} f(t) = 0$ .
- (iii)  $\lim_{t \rightarrow \infty} \frac{f(\tau(\lambda t))}{f(t)} = \infty$  ( $\lambda > 1$ ) implies  $f \in \mathcal{KRPV}_{\mathbb{T}}(\infty)$ .
- (iv)  $\lim_{t \rightarrow \infty} \frac{f(\tau(\lambda t))}{f(t)} = 0$  ( $\lambda > 1$ ) implies  $f \in \mathcal{KRPV}_{\mathbb{T}}(-\infty)$ .

In the end of this section we show that the Karamata type definition is under certain conditions equivalent to Definition 3.2.

**Lemma 3.6** Let  $f$  be a positive rd-continuously differentiable function and let  $f^{\Delta}(t)$  be increasing for large  $t$ . Then

- (i)  $f \in \mathcal{KRPV}_{\mathbb{T}}(\infty)$  iff  $f \in \mathcal{RPV}_{\mathbb{T}}(\infty)$  iff  $f \in \mathcal{NRPV}_{\mathbb{T}}(\infty)$ .
- (ii)  $f \in \mathcal{KRPV}_{\mathbb{T}}(-\infty)$  iff  $f \in \mathcal{RPV}_{\mathbb{T}}(-\infty)$  iff  $f \in \mathcal{NRPV}_{\mathbb{T}}(-\infty)$ .

Moreover,  $f^{\Delta}(t)$  be increasing for large  $t$  is not to be assumed in all if parts.

**Proof.** (i) We will proceed in the following way:

$$f \in \mathcal{KRPV}_{\mathbb{T}}(\infty) \Rightarrow f \in \mathcal{NRPV}_{\mathbb{T}}(\infty) \Rightarrow f \in \mathcal{RPV}_{\mathbb{T}}(\infty) \Rightarrow f \in \mathcal{KRPV}_{\mathbb{T}}(\infty).$$

First implication we can find in [16]. The second implication is trivial. Now we show the third implication. Let  $f \in \mathcal{RPV}_{\mathbb{T}}(\infty)$  and take  $\lambda > 1$ . Then, by Definition 3.2

$$\lim_{t \rightarrow \infty} \frac{f(\tau(\lambda t))}{f(t)} = \lim_{t \rightarrow \infty} \frac{\varphi(\tau(\lambda t))}{\varphi(t)} \cdot \frac{\omega(\tau(\lambda t))}{\omega(t)} = \lim_{t \rightarrow \infty} h_{\lambda}(t) \frac{\omega(\tau(\lambda t))}{\omega(t)}. \quad (11)$$

Let  $\varphi \in \mathcal{RB}_{\mathbb{T}}$ . Therefore,  $h_{\lambda}(t)$  is bounded both above and below by the positive constants. Due to  $\omega \in \mathcal{NRPV}_{\mathbb{T}}(\infty)$ ,  $\omega(t)$  is increasing for large  $t$  (thanks to Proposition 3.4). Now, for all  $\lambda > 1$ , we have

$$\begin{aligned} \omega(\tau(\lambda t)) &\geq \omega(\tau(\lambda t)) - \omega(t) = \int_t^{\tau(\lambda t)} \omega^{\Delta}(s) \Delta s \geq \omega^{\Delta}(t) [\tau(\lambda t) - t] \\ &\geq \omega^{\Delta}(t) [\lambda t - \mu(\tau(\lambda t)) - t] = \omega^{\Delta}(t) [t(\lambda - 1) - \mu(\tau(\lambda t))]. \end{aligned}$$

Hence,

$$\frac{\omega(\tau(\lambda t))}{\omega(t)} \geq \frac{\omega^\Delta(t)[t(\lambda - 1) - \mu(\tau(\lambda t))]}{\omega(t)}. \quad (12)$$

Since  $\lambda > 1$ , in view of  $\mu(\tau(\lambda t))/\omega(t) \rightarrow 0$  as  $t \rightarrow \infty$ , from (11) and (12) we have

$$\lim_{t \rightarrow \infty} \frac{f(\tau(\lambda t))}{f(t)} \geq \lim_{t \rightarrow \infty} h_\lambda(t) \frac{t\omega^\Delta(t)(\lambda - 1)}{\omega(t)} = \infty \quad (\lambda > 1),$$

and thus (thanks to Proposition 3.5)  $f \in \mathcal{KR}\mathcal{PV}_\mathbb{T}(\infty)$ .

(ii) We will proceed analogically as in case (i) and get the series of three implications (where  $\infty$  is replaced by  $-\infty$ ). First implication we can find in [16]. The second implication is trivial. Now we show the third implication. Let  $f \in \mathcal{RPV}_\mathbb{T}(-\infty)$ . By using Proposition 3.4, part (i) of this lemma and Proposition 3.5 we can successively write:

$$f \in \mathcal{RPV}_\mathbb{T}(-\infty) \Rightarrow \frac{1}{f} \in \mathcal{RPV}_\mathbb{T}(\infty) \Rightarrow \frac{1}{f} \in \mathcal{KR}\mathcal{PV}_\mathbb{T}(\infty) \Rightarrow f \in \mathcal{KR}\mathcal{PV}_\mathbb{T}(-\infty).$$

## 4 Applications to half-linear dynamic equations

As an application of the theory of rapid variation, we show that the asymptotic behavior of solutions of half-linear dynamic equation in the form (6) is the same as in the case, where the class of all rapidly varying functions is “thinner”, see Definition 1.1 and [16, Theorem 1]. In [11] the reader can find many useful information about half-linear dynamic equations. In view of the structure of equation (6), it is not difficult to see that every positive solution  $y$  of (6) satisfies  $y^{\Delta\Delta} > 0$ , i.e.,  $y$  is convex and  $y^\Delta$  is increasing.

**Theorem 4.1** *Equation (6) has solutions  $u \in \mathcal{RPV}_\mathbb{T}(-\infty)$  and  $v \in \mathcal{RPV}_\mathbb{T}(\infty)$  if and only if for all  $\lambda > 1$*

$$\lim_{t \rightarrow \infty} t^{\alpha-1} \int_t^{\tau(\lambda t)} p(s) \Delta s = \infty. \quad (13)$$

*Moreover, all positive decreasing solutions of (6) belong to  $\mathcal{NRPV}_\mathbb{T}(-\infty)$  and all positive increasing solutions of (6) belong to  $\mathcal{NRPV}_\mathbb{T}(\infty)$ .*

**Proof.** The proof of this theorem is analogous as a proof in [16, Theorem 1], but some of its steps follow from the Proposition 3.4, Proposition 3.5 and Lemma 3.6.

In the end note, that more and precise information about equation (6) and its Karamata solutions (i.e., solutions, which are slowly, regularly or rapidly varying) reader can find in [16].

## Acknowledgement

The paper was supported by the Grant P201/10/1032 of the Czech Grant Agency.

## References

- [1] BINGHAM, N.H., GOLDIE, C.M., TEUGELS, J.L.: *Regular Variation*, Encyclopedia of Mathematics and its Applications, Vol. 27, Cambridge Univ. Press, 1987.
- [2] BOHNER, M., PETERSON, A. C.: *Dynamic Equations on Time Scales: An Introduction with Applications*. Birkhäuser, Boston, 2001.
- [3] DJURČIĆ, D., KOČINAC, L.D.R., ŽIŽOVIĆ, M.R.: *Some properties of rapidly varying sequences*, J. Math. Anal. Appl. **327** (2007), 1297–1306.
- [4] GALAMBOS, J. SENETA, E.: *Regularly varying sequences*, Proc. Amer. Math. Soc. **41** (1973), 110–116.
- [5] HILGER, S.: *Ein Maßkettenkalkül mit Anwendung auf Zentrumsmannigfaltigkeiten.*, Ph.D. dissertation, Universität of Würzburg, 1988.
- [6] KARAMATA, J.: *Sur certain “Tauberian theorems” de M. M. Hardy et Littlewood*, Mathematica Cluj **3** (1930), 33–48.
- [7] MARIĆ, V.: *Regular Variation and Differential Equations*, Lecture Notes in Mathematics 1726, Springer-Verlag, Berlin-Heidelberg-New York, 2000.
- [8] MATUCCI, S., ŘEHÁK, P.: *Rapidly varying decreasing solutions of half-linear difference equations*, Comput. Modelling **49** (2009), 1692–1699.
- [9] MATUCCI, S., ŘEHÁK, P.: *Regularly varying sequences and second-order difference equations*, J. Difference Equ. Appl. **14** (2008), 17–30.
- [10] MATUCCI, S., ŘEHÁK, P.: *Second order linear difference equations and Karamata sequences*, Int. J. Difference Equ. **3** (2008), 277–288.
- [11] ŘEHÁK, P.: *Half-linear dynamic equations on time scales: IVP and oscillatory properties*, J. Nonl. Funct. Anal. Appl. **7** (2002), 361–404.
- [12] ŘEHÁK, P.: *Regular variation on time scales and dynamic equations*, Aust. J. Math. Anal. Appl. **5** (2008), 1–10.
- [13] ŘEHÁK, P., VÍTOVEC, J.: *q-regular variation and q-difference equations*, J. Phys. A: Math. Theor. **41** (2008) 495203, 1–10.
- [14] ŘEHÁK, P., VÍTOVEC, J.: *Regular variation on measure chains*, Nonlinear Analysis TMA, **72** (2010), 439–448.
- [15] SENETA, E.: *Regularly Varying Functions*, Lecture Notes in Mathematics 508, Springer-Verlag, Berlin-Heidelberg-New York, 1976.
- [16] VÍTOVEC, J.: *Theory of rapid variation on time scales with applications to dynamic equations*, Arch. Math. (Brno) **46** (2010), 263–284.

## Current address

**Jiří Vítovec, Mgr., PhD.**

Department of Mathematics

Faculty of Electrical Engineering and Communication

Brno University of Technology

Technická 8

616 00 Brno, Czech Republic

tel.: +420-541143134, email: vitovec@feec.vutbr.cz

## SOLUTION OF DIFFRACTION PROBLEMS BY BOUNDARY INTEGRAL EQUATIONS

ŽÍDEK Arnošt, (CZ), VLČEK Jaroslav, (CZ), KRČEK Jiří, (CZ)

**Abstract.** Optical diffraction belongs to fewer exploited applications of boundary integral equations. In this paper, we describe theoretical background of their use to numerical modeling of diffraction in periodic structures including some needed theorems and their proofs.

**Key words and phrases.** optical diffraction, grating, boundary integral method.

*Mathematics Subject Classification.* Primary 60A05, 08A72; Secondary 28E10.

### 1 Introduction

Development of optical micro- and nanostructures with periodical ordering takes important place in many branches. Besides less or more complicated experiments, theoretical studies are carried out including mathematical models of electromagnetic wave interaction with geometrically or material-wise modulated media. In the last three decades, there were published numerous monographs and articles treating of optical scattering, especially diffraction in periodic structures (e.g. [1]-[3] and references therein). The various implementation of Rigorous Coupled Waves Algorithm (RCWA) or differential method became mostly used [4]-[5]. One of relatively new approaches is based on Boundary Integral Equations (BIE). Purely theoretical background of this method is referred in [6]-[8] in framework of optical diffraction, some important applications can be found among others in the papers [9]-[11].

In this article, we aim to show the basic features of BIE in diffractive optics. We describe theoretical background of obtained algorithm and some needed theorems and their proofs.

## 2 Boundary integral equations

### 2.1 Formulation of problem

We consider two semi-infinite homogenous isotropic dielectrics with relative permittivities  $\varepsilon^{(1)}, \varepsilon^{(2)}$  and permeabilities  $\mu^{(1)} = \mu^{(2)} = 1$  divided by boundary, which is smooth and periodically modulated in coordinate  $x_1$  with period  $\Lambda$  and uniform in the  $x_2$  direction - see Fig. 1. In any medium, we introduce wave numbers  $k^{(\kappa)} = 2\pi\sqrt{\varepsilon^{(\kappa)}\mu^{(\kappa)}}/\lambda = k_0 n^{(\kappa)}$  and propagations  $\alpha = k^{(1)} \sin \theta$ ,  $\gamma^{(\kappa)} = \sqrt{(k^{(\kappa)})^2 - \alpha^2}$ , where  $n^{(\kappa)} = \sqrt{\varepsilon^{(\kappa)}\mu^{(\kappa)}}$ ,  $\kappa = 1, 2$  are refractive indices of superstrate and substrate, respectively. Incident beam of wavelength  $\lambda$  propagates in the plane  $x_2 = 0$  under incidence angle  $\theta$  with respect to the  $x_3$  axis.

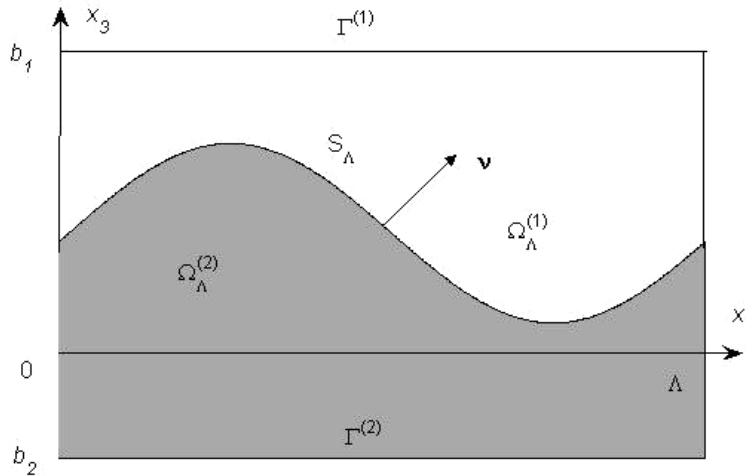


Figure 1: Scheme to the problem formulation.

As usual, we formulate the problem for basic polarizations of planar monochromatic incident beam - normal to the plane of incidence (TE polarization) and parallel to this one (TM polarization). Thus, the spatially dependent intensities of electro-magnetic field have the form  $\mathbf{E} = (0, E_2, 0)$ ,  $\mathbf{H} = (H_1, 0, H_3)$  in the first case or  $\mathbf{E} = (E_1, 0, E_3)$ ,  $\mathbf{H} = (0, H_2, 0)$  for TM field.

Optical diffraction as the interaction of light electromagnetic wave with material medium is generally modeled by Maxwell equations with appropriate boundary conditions. Described situation allows us to consider the boundary problem for a pair of Helmholtz equation. Invariance of the interface in  $x_2$  coordinate let us to write the problem as two-dimensional, where we denote

$$u(\mathbf{x}) = u(x_1, x_3) = \begin{cases} E_2(x_1, x_3) & \text{for TE polarization,} \\ H_2(x_1, x_3) & \text{for TM polarization.} \end{cases} \quad (1)$$

In agreement with physical principle of diffraction this function must be pseudoperiodic in  $x_1$  coordinate to fulfil the relation  $u(\Lambda, x_3) = e^{i\alpha\Lambda}u(0, x_3)$ . Expected solution represent correspon-

ding field component of reflected and transmitted beam in an arbitrary point  $\mathbf{x} = (x_1, x_3)$ ,

$$u(\mathbf{x}) = \begin{cases} u^{(1)}(\mathbf{x}) + u_{\text{in}}(\mathbf{x}), & \mathbf{x} \in \Omega^{(1)}, \\ u^{(2)}(\mathbf{x}), & \mathbf{x} \in \Omega^{(2)}. \end{cases} \quad (2)$$

Incident wave given by the function  $u_{\text{in}} = e^{i(\alpha x_1 + \gamma_1^{(1+)} x_3)}$  on zero diffraction order fulfils Helmholtz equation in  $\Omega^{(1)}$ . Upper index (+) denotes forward wave, while (−) will be used for backward one. Described situation allows us to write problem on the one period by the pair of Helmholtz equation

$$\Delta u^{(\kappa)} + (k^{(\kappa)})^2 u^{(\kappa)} = 0 \quad \text{in } \Omega_{\Lambda}^{(\kappa)}, \quad \kappa = 1, 2. \quad (3)$$

The transient boundary conditions

$$u^{(1)} = u^{(2)}, \quad c \frac{\partial u^{(1)}}{\partial \mathbf{n}} = \frac{\partial u^{(2)}}{\partial \mathbf{n}} \quad (4)$$

express the continuity of tangential field components on common boundary  $S : x_3 = f(x_1)$ ,  $x_1 \in [0, \Lambda]$ . The vector  $\mathbf{n} = (n_1, n_3)$  denotes inner normal for the domain  $\Omega_{\Lambda}^{(1)}$ ,  $c = 1$  for TE and  $c = \varepsilon^{(2)}/\varepsilon^{(1)}$  for TM polarization. The Sommerfeld radiation conditions are applied for far fields on fictitious boundaries  $\Gamma^{(\kappa)} : x_3 = b_{\kappa}$ :

$$\lim_{x \rightarrow \infty} \sqrt{x} \left( \frac{\partial u^{(1)}}{\partial x} - i k^{(1)} u^{(1)} \right) = 0, \quad \lim_{x \rightarrow -\infty} \sqrt{x} \left( \frac{\partial u^{(2)}}{\partial x} - i k^{(2)} u^{(2)} \right) = 0, \quad x = \|\mathbf{x}\|. \quad (5)$$

## 2.2 Boundary integral equations

To formulate the scattering problem with the help of BIE, we start with the standard integral representation of the function  $u(\mathbf{x})$  in an arbitrary point  $\mathbf{x} \in \Omega$  by integrals along the boundary of the domain [12]. Because of the periodicity and Sommerfeld conditions the computation is reduced only on boundary  $S$ :

$$u(\mathbf{x}) = \int_S u(\boldsymbol{\eta}) \frac{\partial G(\mathbf{x}, \boldsymbol{\eta})}{\partial \mathbf{n}_{\boldsymbol{\eta}}} d\ell_{\boldsymbol{\eta}} - \int_S \frac{\partial u(\boldsymbol{\eta})}{\partial \mathbf{n}_{\boldsymbol{\eta}}} G(\mathbf{x}, \boldsymbol{\eta}) d\ell_{\boldsymbol{\eta}}, \quad \boldsymbol{\eta} \in S, \quad (6)$$

where  $G$  denotes fundamental solution of Helmholtz equation. This formula presents the background of integral equations method in the potential form, since the first integral is the double layer potential with the density  $u(\mathbf{x})$ , the second one is the single layer potential with the density  $\partial u(\boldsymbol{\eta})/\partial \mathbf{n}_{\boldsymbol{\eta}}$ . While the second term is continuous on the boundary, potential of double layer has the jump of the size  $\pm \frac{1}{2}u(\boldsymbol{\xi})$  for  $\mathbf{x} = \boldsymbol{\xi} \in S$ , the sign of which corresponds to the normal orientation. Denoting  $\partial u^{(\kappa)}(\boldsymbol{\eta})/\partial \mathbf{n}_{\boldsymbol{\eta}} = v^{(\kappa)}$ , we can create system of integral equation for both regions in potential form

$$u^{(1)}(\mathbf{x}) = - \int_S u^{(1)}(\boldsymbol{\eta}) \frac{\partial G^{(1)}(\mathbf{x}, \boldsymbol{\eta})}{\partial \mathbf{n}_{\boldsymbol{\eta}}} d\ell_{\boldsymbol{\eta}} + \int_S v^{(1)}(\boldsymbol{\eta}) G^{(1)}(\mathbf{x}, \boldsymbol{\eta}) d\ell_{\boldsymbol{\eta}} - u_{\text{in}}(\mathbf{x}), \quad (7)$$

$$u^{(2)}(\mathbf{x}) = \int_S u^{(2)}(\boldsymbol{\eta}) \frac{\partial G^{(2)}(\mathbf{x}, \boldsymbol{\eta})}{\partial \mathbf{n}_\eta} d\ell_\eta - \int_S v^{(2)}(\boldsymbol{\eta}) G^{(2)}(\mathbf{x}, \boldsymbol{\eta}) d\ell_\eta. \quad (8)$$

Obtained system allows us to find the solution  $u^{(\kappa)}$  in an arbitrary point of the domain  $\Omega^{(\kappa)}$ , if we know solution and its gradient on the boundary  $S$ . Therefore, we need to realize the limit transition  $\mathbf{x} \in \Omega^{(\kappa)} \rightarrow \boldsymbol{\xi} \in S$  according to direction of the normal. Since we aim to solve the system for unknown functions in the domain  $\Omega^{(1)}$ , we denote  $u^{(1)} = u, v^{(1)} = v$  and apply boundary conditions (4). This step leads to boundary integral equation in the form

$$u(\boldsymbol{\xi}) = -2 \int_S u(\boldsymbol{\eta}) \frac{\partial G^{(1)}(\boldsymbol{\xi}, \boldsymbol{\eta})}{\partial \mathbf{n}_\eta} d\ell_\eta + 2 \int_S v(\boldsymbol{\eta}) G^{(1)}(\boldsymbol{\xi}, \boldsymbol{\eta}) d\ell_\eta - 2u_{\text{in}}(\boldsymbol{\xi}), \quad (9)$$

$$u(\boldsymbol{\xi}) = 2 \int_S u(\boldsymbol{\eta}) \frac{\partial G^{(2)}(\boldsymbol{\xi}, \boldsymbol{\eta})}{\partial \mathbf{n}_\eta} d\ell_\eta - 2c \int_S v(\boldsymbol{\eta}) G^{(2)}(\boldsymbol{\xi}, \boldsymbol{\eta}) d\ell_\eta. \quad (10)$$

Finally, we write boundary integral operators as the potentials of single and double layer

$$\mathcal{W}^{(\kappa)} u = 2 \int_S u(\boldsymbol{\eta}) \frac{\partial G^{(\kappa)}(\boldsymbol{\xi}, \boldsymbol{\eta})}{\partial \mathbf{n}_\eta} d\ell_\eta, \quad \mathcal{V}^{(\kappa)} u = 2 \int_S v(\boldsymbol{\eta}) G^{(\kappa)}(\boldsymbol{\xi}, \boldsymbol{\eta}) d\ell_\eta \quad (11)$$

to obtain the system (9)-(10) in matrix operator form

$$\begin{bmatrix} \mathcal{W}^{(1)} + \mathcal{I} & -\mathcal{V}^{(1)} \\ \mathcal{W}^{(2)} - \mathcal{I} & -c\mathcal{V}^{(2)} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} -2u_{\text{in}} \\ 0 \end{bmatrix}, \quad (12)$$

where  $\mathcal{I}$  is identity operator.

### 3 Fundamental solution of Helmholtz equation

#### 3.1 General case

Fundamental solution of Helmholtz equation in  $\mathbb{R}^2$  is Hankel function of the first kind zero order  $G(\mathbf{x}, \mathbf{y}) = \frac{i}{4} H_0^1(k\|\mathbf{x} - \mathbf{y}\|)$  that fulfils equation  $\Delta G + k^2 G = \delta(\mathbf{x} - \mathbf{y})$ . Such function satisfies Sommerfeld radiation conditions. Further, we can denote  $z = k\|\mathbf{x} - \mathbf{y}\|$  and prove following theorem.

**Theorem 1.** *Let  $H_0^1(z)$  is Hankel function of the first kind zero order. Then function  $\frac{i}{4} H_0^1(z) - \frac{1}{2\pi} \ln \frac{z}{2}$  is smooth for each  $z \in \mathbb{C}$ .*

**Proof.** Hankel function of the first kind zero order can be decomposed as a sum of Bessel function  $J_0(z)$  and Neumann function  $N_0(z)$  as  $H_0^1(z) = J_0(z) + iN_0(z)$ , where we can express Neumann function in the form of convergent infinite series

$$N_0(z) = \frac{2}{\pi} J_0(z) \left( \ln \frac{z}{2} + \gamma \right) - \frac{2}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{(k!)^2} \left( \frac{z}{2} \right)^{2k} \sum_{j=1}^k \frac{1}{j}, \quad (13)$$

where  $\gamma$  is Euler constant. The formula can be written in the form

$$H_0^1(z) = \frac{2i}{\pi} J_0(z) \ln z + J_0(z) \left[ 1 + \frac{2i}{\pi} (\gamma - \ln 2) \right] - \frac{2i}{\pi} \sigma(z), \quad (14)$$

where the term  $\sigma(z) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(k!)^2} \left(\frac{z}{2}\right)^{2k} \sum_{j=1}^k \frac{1}{j}$  must be smooth being an uniform convergent power series. Therefore, only terms that include logarithms can have some singularities, so we take into consideration only the first term of (14). It is sufficient to prove that the term  $(1 - J_0(z)) \ln z$  has finite limit at  $z = 0$ . We use recurrent relations between Bessel functions of the zero, first and second order, their derivatives and their values for  $z = 0$ ,

$$\lim_{z \rightarrow 0} (1 - J_0(z)) \ln z = -2 \lim_{z \rightarrow 0} \frac{J_1^2(z) + (1 - J_0(z)) J_1'(z)}{J_0(z) - z J_1(z)} = 0. \quad (15)$$

We have just proved continuity of investigated function. Furthermore, we need to prove continuity of its first derivative. The singularity can still be only in the term  $(1 - J_0(z)) \ln z$ , respectively its derivative  $[(1 - J_0(z)) \ln z]' = J_1(z) \ln z + \frac{1}{z} (1 - J_0(z))$ . It is obvious that the second term of the derivative converges to zero and we need to calculate only the limit of the first one. Using analogical approach as in the first problem we get

$$\lim_{z \rightarrow 0} J_1(z) \ln z = - \lim_{z \rightarrow 0} \frac{2 J_1(z) J_1'(z)}{-J_1'(z) + J_0(z) + z J_0'(z)} = 0 \quad (16)$$

and prove the whole theorem.

### 3.2 Periodic case

If we search for solution of Helmholtz equation  $G(\mathbf{x}, \mathbf{y}) = G(\mathbf{x} - \mathbf{y})$  in a strip of the width  $\Lambda$ , for which the term  $e^{-i\alpha x_1} G_\Lambda$  is periodic in  $x_1$  coordinate with period  $\Lambda$ , the resulting function must satisfy the equation [13]

$$\Delta G_\Lambda + k^2 G_\Lambda = \sum_{m \in \mathbb{Z}} \delta(x_1 - y_1 - m\Lambda, x_3 - y_3) e^{i\alpha m \Lambda}. \quad (17)$$

For clarity, the upper index ( $\kappa$ ) will be suppressed hereafter. This requirement is fulfilled among others for the function [3]

$$G_\Lambda(\mathbf{x}, \mathbf{y}) = \frac{1}{2i\Lambda} \sum_{m \in \mathbb{Z}} \frac{1}{\gamma_m} \exp \{ i [\alpha_m (x_1 - y_1) + \gamma_m |x_3 - y_3|] \}, \quad (18)$$

where  $\alpha_m = \alpha + 2\pi m/\Lambda$ ,  $\gamma_m^2 = k^2 - \alpha_m^2$ . Fundamental solution (18) also satisfies Sommerfeld radiation conditions (5), but for  $\mathbf{x} - \mathbf{y} = \mathbf{o}$  the series diverges. Nevertheless, we are able to prove [7] that this singularity is also of a logarithmical type, so that the difference  $G_\Lambda(\mathbf{x} - \mathbf{y}) - \frac{1}{2\pi} \ln \frac{1}{\|\mathbf{x} - \mathbf{y}\|}$  is continuous for all  $\mathbf{x}, \mathbf{y}$ . In the following considerations the normal derivative of fundamental solution

$$\frac{\partial G_\Lambda(\mathbf{x}, \mathbf{y})}{\partial \mathbf{n}_y} = -\frac{1}{2\Lambda} \sum_{m \in \mathbb{Z}} \frac{1}{\gamma_m} [\alpha_m n_1 + \gamma_m \operatorname{sgn}(x_3 - y_3) n_2] e^{i[\alpha_m (x_1 - y_1) + \gamma_m |x_3 - y_3|]} \quad (19)$$

plays the important role, as we have already shown in section 2.2.

### 3.3 Parametrization

Let  $\mathbf{p}(t) = (p(t), q(t)), t \in [0, 2\pi]$  be a parametrization of the boundary  $S$  with following properties:

$$\begin{aligned} p(0) &= 0, & p(2\pi) &= \Lambda, & q(0) &= q(2\pi), \\ p(t+2\pi) &= p(t) + \Lambda, & q(t+2\pi) &= q(t). \end{aligned} \quad (20)$$

For the boundary points we have  $\boldsymbol{\xi} = \mathbf{p}(s), \boldsymbol{\eta} = \mathbf{p}(t), s, t \in [0, 2\pi]$  with corresponding normal vector

$$\mathbf{n}(t) = (n_1(t), n_3(t)) = (-\dot{q}(t), \dot{p}(t)), \quad \|\mathbf{n}\| = n(t) = \sqrt{\dot{p}^2(t) + \dot{q}^2(t)}. \quad (21)$$

The kernel of the operators  $\mathcal{V}^{(\kappa)}$  is periodical fundamental solution (18), which can be written by above parametrization as

$$\begin{aligned} G_\Lambda(s, t) &= \sum_{m \in \mathbb{Z}} G_{\Lambda, m}(s, t), \\ G_{\Lambda, m}(s, t) &= \frac{1}{2i\Lambda} \frac{1}{\gamma_m} \exp \{i[\alpha_m(p(s) - p(t)) + \gamma_m(q(s) - q(t))]\} \end{aligned} \quad (22)$$

with logarithmical singularity that can be transformed using following theorem.

**Theorem 2.** *Let  $\mathbf{p} : [0, 2\pi] \rightarrow \mathbb{R}^2$  is above parametrization. Then the logarithmical singularity fulfils*

$$\ln \|\mathbf{p}(s) - \mathbf{p}(t)\| = - \sum_{m \neq 0} \frac{e^{-im(s-t)}}{2|m|}. \quad (23)$$

**Proof.** Using above parametrization we consider two near points of the boundary  $\boldsymbol{\xi} = \mathbf{p}(s), \boldsymbol{\eta} = \mathbf{p}(t), s, t \in [0, 2\pi]$ . In the neighbourhood of a singularity point  $s = t$  we can replace periodized boundary  $S$  by the arc of unit circle with a centre  $(p_0, q_0)$ , that lies out of  $S$ , i.e.  $p(t) = p_0 + \cos t, q(t) = q_0 + \sin t$ , and, modify the singularity

$$\ln \|\mathbf{p}(s) - \mathbf{p}(t)\| = \ln \sqrt{2 - 2\cos(s-t)} = \ln \left| 2 \sin \frac{s-t}{2} \right|. \quad (24)$$

Now, we can use Euler formula to transform trigonometric functions into functions of complex variable

$$\ln \|\mathbf{p}(s) - \mathbf{p}(t)\| = \ln \left( \left| e^{i\frac{s-t}{2}} \right| |1 - e^{-i(s-t)}| \right). \quad (25)$$

The first absolute value is equal to 1. Finally, we can use expansion of Taylor series of the function  $\ln(1-z)$  on unit circle  $|z| < 1$ ,  $\ln(1-z) = - \sum_{n=1}^{\infty} \frac{z^n}{n}$  to prove (23).

### 3.4 Splitting of periodical fundamental solution

For the numerical solution it is necessary to split off the obtained term (23) from the kernel  $G_\Lambda$ . We denote

$$G_{\Lambda, m}(s, t) = \frac{1}{2i\Lambda\gamma_m} e^{i(\alpha_m(p(s)-p(t)) + \gamma_m(q(s)-q(t)))} \quad (26)$$

and split the kernel in the following way

$$G_{\Lambda}(s, t) = G_{\Lambda,0}(s, t) + \sum_{m \neq 0} \left\{ G_{\Lambda,m}(s, t) - \frac{1}{2\pi} \frac{e^{-im(s-t)}}{2|m|} \right\} + \frac{1}{2\pi} \sum_{m \neq 0} \frac{e^{-im(s-t)}}{2|m|}. \quad (27)$$

While the first two terms on the right hand side are kernels of the compact operators, the third one generates a singular kernel in the single layer potential, which needs to be treated separately during the implementation. While the compactness of the first term is obvious, for the second one we need to prove following theorem.

**Theorem 3.** *The series*

$$\sum_{m \neq 0} \left\{ G_{\Lambda,m}(s, t) - \frac{1}{2\pi} \frac{e^{-im(s-t)}}{2|m|} \right\} \quad (28)$$

*is absolutely convergent for all  $m \in \mathbb{Z}, m \neq 0$ .*

**Proof.** To prove the convergence of series (28) we focus on modification of its  $m$ -th member. For simplicity we can denote

$$G_{\Lambda,m}(s, t) = \frac{1}{2i\Lambda\gamma_m} e^{i\Omega_m}, \quad (29)$$

where  $\Omega_m = \alpha_m(p(s) - p(t)) + \gamma_m|q(s) - q(t)|$ . We have already used propagation in the direction of  $x_1$  axis  $\alpha_m = \alpha + 2\pi m/\Lambda$ ,  $\alpha = 2\pi\tilde{\varepsilon}/\lambda$ , where we can denote  $\tilde{\varepsilon} = \sqrt{\varepsilon^{(1)}} \sin \theta$ . Permittivity  $\varepsilon^{(1)}$  is used for incident media,  $\varepsilon$  for any other consequent layer. Furthermore, we denote  $\beta = \Lambda/\lambda$  the ratio of period of the boundary and wavelength of incidental wave. Now, we can modify the term (29)

$$\begin{aligned} 2i\Lambda\gamma_m &= 2i\Lambda \sqrt{\left(\frac{2\pi}{\lambda}\right)^2 \varepsilon - \left(\frac{2\pi}{\lambda}\tilde{\varepsilon} + \frac{2\pi m}{\Lambda}\right)^2} = \\ &= 4\pi \sqrt{(m + \beta\tilde{\varepsilon})^2 - \beta^2\varepsilon}. \end{aligned} \quad (30)$$

We can also denote the second term in the difference (28)  $B_m$  and now we can finally modify the  $m$ -th member of the series:

$$\begin{aligned} G_{\Lambda,m} - B_m &= \frac{1}{4\pi} \left[ \frac{e^{i\Omega_m}}{\sqrt{(m + \beta\tilde{\varepsilon})^2 - \beta^2\varepsilon}} - \frac{e^{-im(s-t)}}{|m|} \right] = \\ &= \frac{1}{4\pi} \frac{|m|e^{i\Omega_m} - e^{-im(s-t)}\sqrt{(m + \beta\tilde{\varepsilon})^2 - \beta^2\varepsilon}}{|m|\sqrt{(m + \beta\tilde{\varepsilon})^2 - \beta^2\varepsilon}}. \end{aligned} \quad (31)$$

Let us consider the singular point  $s = t$ , i.e.  $\Omega_m = 0$ , where for the difference holds

$$\begin{aligned} G_{\Lambda,m} - B_m &= \frac{1}{4\pi} \frac{|m| - \sqrt{(m + \beta\tilde{\varepsilon})^2 - \beta^2\varepsilon}}{|m|\sqrt{(m + \beta\tilde{\varepsilon})^2 - \beta^2\varepsilon}} = \\ &= \frac{1}{4\pi} \frac{m^2 - (m + \beta\tilde{\varepsilon})^2 + \beta^2\varepsilon}{|m|\sqrt{(m + \beta\tilde{\varepsilon})^2 - \beta^2\varepsilon}(|m| + \sqrt{(m + \beta\tilde{\varepsilon})^2 - \beta^2\varepsilon})} = \\ &= \frac{1}{4\pi} \frac{\beta^2(\varepsilon - \tilde{\varepsilon}^2) - 2m\beta\tilde{\varepsilon}}{m^2\sqrt{(m + \beta\tilde{\varepsilon})^2 - \beta^2\varepsilon} + |m|((m + \beta\tilde{\varepsilon})^2 - \beta^2\varepsilon)}. \end{aligned} \quad (32)$$

Considering that power of the variable  $m$  is one in the numerator, but three in the denominator of the fraction, we can simply use integral criteria of convergence to prove the theorem.

## 4 Conclusion

Presented results were applied in the mathematical model of optical diffraction on periodical boundary and implemented in MATLAB computational code. Obtained numerical results will be referred in future work.

## Acknowledgement

This work has been partially supported under the IT4 Innovations Centre of Excellence project, reg. no. CZ.1.05/1.1.00/02.0700.

## References

- [1] *Electromagnetic theory of gratings*, ed. by R. PETIT, Springer, Berlin, 1980.
- [2] NEVIÈRE, M., POPOV, E.: *Light propagation in periodic media: Differential theory and design*. Marcel Dekker, 2002.
- [3] *Mathematical modeling in optical science*, eds. G. BAO, L. COWSAR and W. MASTERS, SIAM, Philadelphia, 2001.
- [4] LI, L.: *Formulation and comparison of two recursive matrix algorithms for modeling layered diffraction gratings*. J. Opt. Soc. Am. A 13, pp. 1024-1035, 1996.
- [5] GRANET, G., GUIZAL, B.: *Efficient implementation of the coupled wave method for metallic lamellar gratings in TM polarization*. J. Opt. Soc. Am. A 17, pp. 1019-1023, 2000.
- [6] COSTABEL, M., STEPHAN, E.: *A direct boundary integral equation method for transmission problems*. J. Math. Anal. Appl. 106, pp. 367-413, 1985.
- [7] NEDELEC, J.C., STARLING, F.: *Integral equation methods in a quasi-periodic diffraction problem for the time-harmonic Maxwell's equations*. SIAM, J. Math. Anal. 22, pp. 1679-1701, 1991.
- [8] ELSCHNER, J., SCHMIDT, G.: *Diffraction in periodic structures and optimal design of binary gratings. Part I: Direct problems and gradient formulas*. Math. Meth. Appl. Sci. 21, pp. 1297-1342, 1998.
- [9] PRATHER, D.W., MIROTZNIK, M.S., MAIT J.N.: *Boundary integral methods applied to the analysis of diffractive optical elements*. J. Opt. Soc. Am. A 14, pp. 34-43, 1997.
- [10] BENDICKSON, J.M., GLYTSIS, E.N., GAYLORD, T.K., PETERSON, A.F.: *Modeling considerations of rigorous boundary element method analysis of diffractive optical elements*. J. Opt. Soc. Am. A 18, pp. 1495-1506, 2001.
- [11] MAGATH, T., SEREBRYANNIKOV, A.E.: *Fast iterative, coupled-integral-equation technique for inhomogeneous profiled and periodic slabs*. J. Opt. Soc. Am. A 12, pp. 2405-2018, 2005.
- [12] PARÍS, F., CAÑAS, J.: *Boundary element method. Fundamentals and applications*. Oxford, University Press, 1997.
- [13] LINTON, C.M.: *The Green's function for the two-dimensional Helmholtz equation in periodic domains*, J. Eng. Math. Vol. 33, pp. 377-402, 1998.

**Current address**

**Mgr. Arnošt Žídek**

Department of mathematics and descriptive geometry, Technical University of Ostrava,  
17. listopadu 15, 708 33, Ostrava-Poruba, Czech Republic, tel: +420 59 732 4177,  
arnost.zidek@vsb.cz

**Doc. RNDr. Jaroslav Vlček, CSc.**

Nanotechnology centre, Technical University of Ostrava, 17. listopadu 15, 708 33,  
Ostrava-Poruba, Czech Republic, tel: +420 59 732 4176, jaroslav.vlcek@vsb.cz

**Mgr. Jiří Krčák**

Department of mathematics and descriptive geometry, Technical University of Ostrava,  
17. listopadu 15, 708 33, Ostrava-Poruba, Czech Republic, tel: +420 59 732 4179,  
jiri.krcek@vsb.cz



## PARALLEL SOLUTION OF POISSON EQUATION

**BÍMOVÁ Daniela, (CZ)**

**Abstract.** The paper is devoted to the parallel solution of Poisson equation. We distinguish two cases of the problem - one- and two-dimensional ones. By the usage of parallel approach to the linear algebra representation we create the parallel algorithm for computing a numerical solution of the Poisson equation. We compare calculation times of computing the approximate solution of the system of (linear) difference equations for different sizes of the system matrix by the numerical method of steepest descent on eight-processor cluster for 2, 4, and 8 processors, respectively.

**Key words and phrases.** Poisson equation, finite difference method, parallel linear algebra, parallel mapping of a matrix, method of steepest descent, numerical experiment, parallel algorithm, condition number, cluster

*Mathematics Subject Classification:* 35J05, 65M06, 65Y05, 65F10, 65N22

### 1 Introduction

Our model problem is represented by the Poisson equation, which discretization via the finite difference method leads to the system of linear algebraic equations.

There are many methods which we can use for solving the system of linear algebraic equations. Some of them are based on the principles of direct solvers (such as e.g. Gaussian elimination, Gauss-Jordan elimination, and the use of inverse matrix), the other ones are built on the principles of iterative methods (we can name for example the method of steepest descent, the conjugate gradient method, etc.).

We choose the numerical method of steepest descent for solving the system of linear algebraic equations in this paper and we try to apply this method in parallel algorithm implemented in Fortran. The aim of this paper is to find how we can save the calculating time computing the system of linear algebraic equations on two, four, and eight processors of the cluster instead of on one processor. We perform our calculation in numerical experiment in which we compute the vector of the approximate solution of the system of difference equations by the method of steepest descent.

## 2 Setting of a problem

We consider the one- and two-dimensional boundary value problems in this chapter: Find function  $u: \Omega \rightarrow \mathbf{R}$  fulfilling the equation

$$-\varepsilon \Delta u = f \text{ in } \Omega = \langle 0, 1 \rangle^n, \text{ where } \varepsilon = 1 \text{ and where } n = \{1, 2\} \quad (1)$$

$$u|_{\delta\Omega} = g \text{ on } \delta\Omega, \quad (2)$$

where  $f: \Omega \rightarrow \mathbf{R}$  is the given function and where  $g: \delta\Omega \rightarrow \mathbf{R}$  represents Dirichlet boundary condition.

### 2.1 One-dimensional problem

For  $m \in \mathbf{N}$  we bring up  $(m+1)$  equidistant line-segment points  $X_i = i \cdot h$ , where  $i = 0, 1, \dots, m$ , and  $h = \frac{1}{m+1}$ . The symbol  $U_i$  stands for the approximate solution at the points  $X_i$ , i. e.  $U_i = u(X_i)$ , and we set  $F_i = f(X_i)$ , consequently.

We use three-point stencil for approximation the second derivatives in every regular knot  $X_i$ , i.e.

$$\frac{\delta^2 u}{\delta x^2}(x_i) \approx \frac{U_{i-1} - 2U_i + U_{i+1}}{h^2}.$$

Using this we can rewrite (1) as

$$-\frac{U_{i-1} - 2U_i + U_{i+1}}{h^2} = F_i, \quad i = 1, 2, \dots, m-1.$$

After adjustment we are able to write the difference equations for the regular knots  $X_i$ , e.g.

$$-U_{i-1} + 2U_i - U_{i+1} = h^2 \cdot F_i, \quad i = 1, 2, \dots, m-1.$$

This is the standard three-point scheme. The approximate values of the sought function  $u$  in knots of the given grid are represented by the numerical solution of the system of (linear) algebraic equations

$$\mathbf{A}\mathbf{U} = \mathbf{F}, \quad (3)$$

where

$$\mathbf{U} = (U_1, U_2, \dots, U_{m-1})^T \in \mathbf{R}^{(m-1)}$$

is an unknown vector,

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & & \vdots \\ 0 & 0 & -1 & 2 & \ddots & 0 \\ \vdots & \vdots & & \ddots & \ddots & -1 \\ 0 & 0 & \cdots & 0 & -1 & 2 \end{pmatrix}$$

is a matrix of the system of linear algebraic equations, and

$$\mathbf{F} = \begin{pmatrix} h^2 \cdot F_1 + U_0 \\ h^2 \cdot F_2 \\ \vdots \\ h^2 \cdot F_{m-2} \\ h^2 \cdot F_{m-1} + U_m \end{pmatrix}$$

is a right-hand side of the system of (linear) algebraic equations.

The Dirichlet boundary conditions

$$U_0 = g(X_0), U_m = g(X_m)$$

result from (2).

## 2.2 Two-dimensional problem

For  $m, n \in \mathbf{N}$  we bring up  $(m+1) \times (n+1)$  of evenly spaced points  $X_{ij} = [x_i, y_j] = [a + i \cdot h, c + j \cdot h]$ , where  $i = 0, 1, \dots, m$ ,  $j = 0, 1, \dots, n$  and where  $h$  is the spatial step. We denote  $U_{ij}$  the approximate solution at the points  $X_{ij}$ , i. e.  $U_{ij} = u(x_i, y_j) = u(X_{ij})$ , and we put  $F_{ij} = f(x_i, y_j) = f(X_{ij})$ .

For every regular knot  $X_{ij}$  we use three-point stencil for approximation the second derivatives, i.e.

$$\frac{\delta^2 u}{\delta x^2}(x_i, y_j) \approx \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2}, \quad \frac{\delta^2 u}{\delta y^2}(x_i, y_j) \approx \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{h^2},$$

then (1) can be reformulated into

$$-\frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} - \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{h^2} = F_{ij}, \quad i = 1, 2, \dots, m-1, \\ j = 1, 2, \dots, n-1.$$

After modification we obtain the difference equations for the regular knots  $X_{ij} = [x_i, y_j]$ , it means the equations of the form

$$-U_{i-1,j} - U_{i,j-1} + 4U_{i,j} - U_{i+1,j} - U_{i,j+1} = h^2 \cdot F_{ij}, \quad i = 1, 2, \dots, m-1, \quad j = 1, 2, \dots, n-1.$$

This is the standard five-point scheme. The approximate values of the sought function  $u$  in knots of the given grid are represented by the numerical solution of the system of (linear) algebraic equations (3) with an unknown vector

$$U = (U_{11}, \dots, U_{1,n-1}, U_{21}, \dots, U_{2,n-1}, \dots, U_{n-1,1}, \dots, U_{n-1,n-1})^T \in \mathbf{R}^{(m-1) \times (n-1)}$$

and a matrix  $\mathbf{A}$ , which we can write in the following way

$$\mathbf{A} = \begin{pmatrix} \overbrace{4 \quad -1 \quad \cdots \quad 0}^{U_{11}, \dots, U_{1,n-1}} & \overbrace{-1 \quad 0 \quad \cdots \quad 0}^{U_{21}, \dots, U_{2,n-1}} & \overbrace{0 \quad 0 \quad \cdots \quad 0}^{U_{m-1,1}, \dots, U_{m-1,n-1}} \\ -1 \quad 4 \quad \ddots \quad \vdots & 0 \quad -1 \quad \ddots \quad \vdots & 0 \quad 0 \quad \ddots \quad \vdots \\ \vdots \quad \ddots \quad \ddots \quad -1 & \vdots \quad \ddots \quad \ddots \quad 0 & \cdots \quad \vdots \quad \ddots \quad \ddots \quad 0 \\ 0 \quad \cdots \quad -1 \quad 4 & 0 \quad \cdots \quad 0 \quad -1 & 0 \quad \cdots \quad 0 \quad 0 \\ -1 \quad 0 \quad \cdots \quad 0 & 4 \quad -1 \quad \cdots \quad 0 & 0 \quad 0 \quad \cdots \quad 0 \\ 0 \quad -1 \quad \ddots \quad \vdots & -1 \quad 4 \quad \ddots \quad \vdots & 0 \quad 0 \quad \ddots \quad \vdots \\ \vdots \quad \ddots \quad \ddots \quad 0 & \vdots \quad \ddots \quad \ddots \quad -1 & \cdots \quad \vdots \quad \ddots \quad \ddots \quad 0 \\ 0 \quad \cdots \quad 0 \quad -1 & 0 \quad \cdots \quad -1 \quad 4 & 0 \quad \cdots \quad 0 \quad 0 \\ & \vdots & \ddots & \vdots \\ 0 \quad 0 \quad \cdots \quad 0 & 0 \quad 0 \quad \cdots \quad 0 & 4 \quad -1 \quad \cdots \quad 0 \\ 0 \quad 0 \quad \ddots \quad \vdots & 0 \quad 0 \quad \ddots \quad \vdots & -1 \quad 4 \quad \ddots \quad \vdots \\ \vdots \quad \ddots \quad \ddots \quad 0 & \vdots \quad \ddots \quad \ddots \quad 0 & \cdots \quad \vdots \quad \ddots \quad \ddots \quad -1 \\ 0 \quad \cdots \quad 0 \quad 0 & 0 \quad \cdots \quad 0 \quad 0 & 0 \quad \cdots \quad -1 \quad 4 \end{pmatrix}.$$

as well as the vector of the right side of the mentioned system of (linear) difference equations we can write as follows

$$\mathbf{F} = \begin{pmatrix} h^2 F_{11} + U_{0,1} + U_{1,0} \\ h^2 F_{12} + U_{0,2} \\ \vdots \\ h^2 F_{1,n-1} + U_{0,n-1} + U_{1,n} \\ h^2 F_{21} + U_{1,0} \\ h^2 F_{22} \\ \vdots \\ h^2 F_{2,n-1} + U_{2,n} \\ \vdots \\ h^2 F_{m-1,1} + U_{0,1} + U_{1,0} \\ h^2 F_{m-1,2} + U_{0,2} \\ \vdots \\ h^2 F_{m-1,n-1} + U_{m,n-1} + U_{m-1,n} \end{pmatrix}.$$

It is also possible to write the system of (linear) difference equations by blocks

$$\mathbf{A} = \begin{pmatrix} \mathbf{B} & -\mathbf{I} & \cdots & \mathbf{0} \\ -\mathbf{I} & \mathbf{B} & \ddots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B} \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} h^2 F^1 \\ h^2 F^2 \\ \vdots \\ h^2 F^{m-1} \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} U^1 \\ U^2 \\ \vdots \\ U^{m-1} \end{pmatrix},$$

where

$$\mathbf{B} = \begin{pmatrix} 4 & -1 & \cdots & 0 \\ -1 & 4 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ 0 & \cdots & -1 & 4 \end{pmatrix}, \mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}, \mathbf{0} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix}$$

are the matrices of the order  $(n-1) \times (n-1)$  and where

$$F^i = (F_{i1}, F_{i2}, \dots, F_{i, n-1})^T, U^i = (U_{i1}, U_{i2}, \dots, U_{i, n-1})^T \in \mathbf{R}^{n-1}.$$

According to (2) we consider the following Dirichlet boundary conditions:

$$\begin{aligned} U_{i,0} &= g(X_{i,0}), U_{i,n} = g(X_{i,n}) \quad \forall i = 1, 2, \dots, m-1, \\ U_{0,j} &= g(X_{0,j}), U_{m,j} = g(X_{m,j}) \quad \forall j = 1, 2, \dots, n-1. \end{aligned}$$

### 3 Parallel approach to the solution of the system of linear algebraic equations

#### 3.1 Basic ways of matrices mapping

The way, in which the matrix is mapped on the net of processors, determines efficiency and elegance of algorithm in most the cases. There are two basic kinds of mapping:

- a) striped mapping
- b) checkerboard mapping

These two mappings divide further on several examples. E.g. there are 6 variants of the striped mapping that can arise from 2 and 3 possibilities:

- 1) by rows or by columns
- 2) by blocks, cyclically or cyclically by blocks.

#### 3.2 The striped mapping of the matrix cyclically by rows

In this paragraph we describe how we map a matrix using MPI in parallel algorithms.

Let us assume, without loss of generality, that there is given the matrix  $\mathbf{A}$  of the mentioned shape and of the type  $(8, 8)$ . Further, we assume that we work on two processors. (In the parallel algorithm we work generally on  $n$  processors). The first task is to map the matrix  $\mathbf{A}$  onto the particular processors. The matrix  $\mathbf{A}$  will be mapped by the striped mapping cyclically by rows onto the particular processors. The striped mapping assumes that the processors are connected into the linear virtual array and that they are numbered  $0, 1, 2, \dots, p-1$ , where  $p$  is the number of used processors, in general. The matrix  $\mathbf{A}$  (denoted  $\mathbf{A}_{\text{global}}$ ) is generated and known only by the master processor. The master processor distributes data, it means the row vectors of the matrix  $\mathbf{A}$ , into the particular processors. We obtain new matrices (denoted  $\mathbf{A}_{\text{local}}$ ) of the type  $(4, 8)$  by the data distribution in our illustration case (of the type  $(\lceil n/p \rceil, n)$ , where  $n$  is the number of rows of original matrix and where  $p$  is the number of used processors, in general) that consist of the appropriate rows of the matrix  $\mathbf{A}$ .

Figure n. 1 illustrates the striped mapping of the matrix  $A$  of the type  $(8, 8)$  cyclically by rows onto two processors.

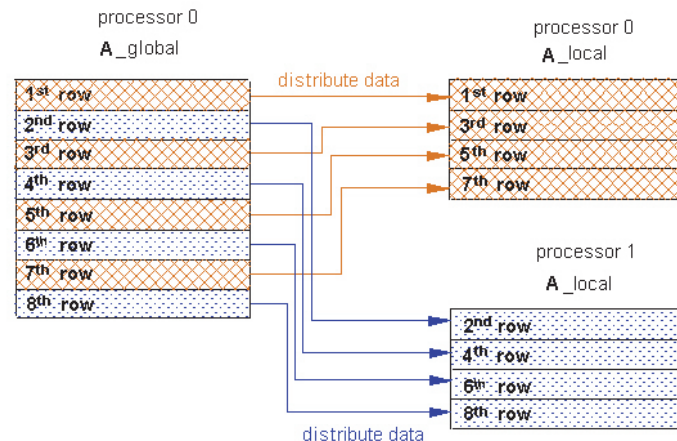


Figure n. 1 – Striped mapping of a matrix cyclically by rows on two processors

The analogous situation is true for mapping the vector  $F$  (of the right side of the system of linear algebraic equations). There is one small difference, only one element (not the whole row) is sent in every step of the distribution.

### 3.3 Product of a matrix and of a vector in parallel algorithm

We use the iterative methods for solving the system of linear algebraic equations. Part of every iterative method is the product of a matrix and of a vector. That is why we describe product of a matrix  $A$  and of a vector  $U$  from the matrix equation

$$AU = F$$

using MPI in parallel algorithms in this paragraph.

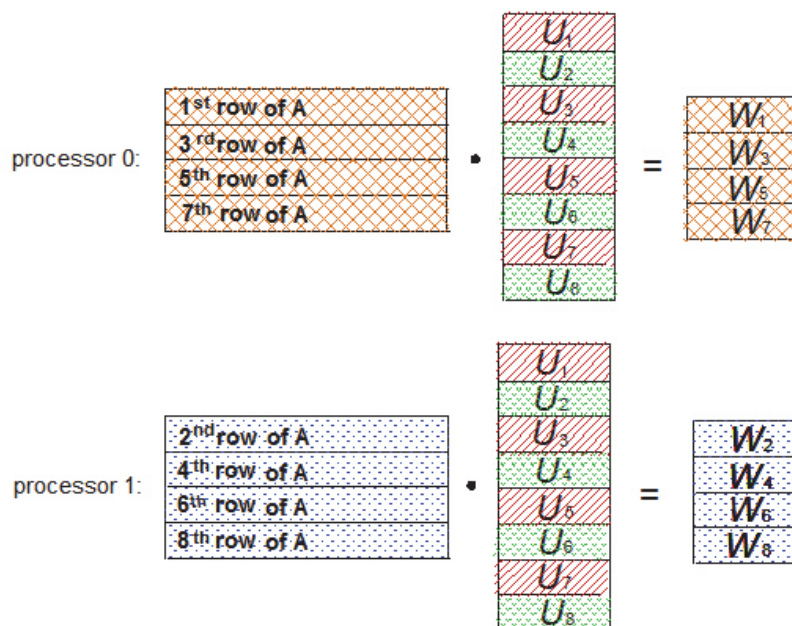


Figure n. 2 – Matrix - vector product in parallel algorithm

Every processor knows only a part of the vector  $\mathbf{U}$ , with whom we want to multiply the matrix  $\mathbf{A}$ , on the basis of the previous mapping of a vector onto the particular processors. (We assume that the given vector  $\mathbf{U}$  is mapped in the same way in which was mapped the vector  $\mathbf{F}$ ). That is why every processor has to find the rest part of the vector  $\mathbf{U}$ , which it does not know. Every processor can call the command “MPI\_ALLGATHER” by which it can find the missing information about the rest parts of the vector  $\mathbf{U}$ .

If all the processors know all the elements of the vector  $\mathbf{U}$ , we can perform dot product and calculate the appropriate element of the result vector (denoted e.g.  $\mathbf{W}$ ).

Figure n. 2 illustrates, without loss of generality, the way of calculating the product of the matrix  $\mathbf{A}$  of the type (8, 8) and of the vector  $\mathbf{U}$  of the type (8, 1) in parallel algorithm computed on two processors.

### 3.4 Scalar product of two vectors in parallel algorithm

Except of the product of a matrix and a vector the scalar product of two vectors occurs in the iterative methods, too. In this paragraph we describe process of computing a scalar product of two vectors in parallel algorithms.

Every processor knows only a part of the vector  $\mathbf{U}$  as well as of the vector  $\mathbf{V}$  (which we want to multiply scalarly) on the basis of the previous mapping (the striped mapping cyclically by rows) of a vector onto the particular processors. That is why we firstly multiply scalarly the appropriate elements of the vectors  $\mathbf{U}$ ,  $\mathbf{V}$ . Secondly, we call the command “MPI\_ALLREDUCE” that sums over all processors and distributes the resulting sum onto all the processors. All the processors (processors P0 and P1 in our illustration case) finally know value of the scalar product of the vectors  $\mathbf{U}$ ,  $\mathbf{V}$ .

Figure n. 3 illustrates, without loss of generality, calculation of the scalar product of the vectors  $\mathbf{U}$  and  $\mathbf{V}$  of the types (8, 1), that were mapped by striped mapping cyclically by rows, in parallel algorithm computed on two processors.

$$\begin{aligned} \text{processor P0 } & ( \boxed{U_1} \boxed{U_3} \boxed{U_5} \boxed{U_7}, \boxed{V_1} \boxed{V_3} \boxed{V_5} \boxed{V_7} ) = \sum_{k=1}^4 U_{2k-1} \cdot V_{2k-1} \\ \text{processor P1 } & ( \boxed{U_2} \boxed{U_4} \boxed{U_6} \boxed{U_8}, \boxed{V_2} \boxed{V_4} \boxed{V_6} \boxed{V_8} ) = \sum_{k=1}^4 U_{2k} \cdot V_{2k} \end{aligned}$$

Figure n. 3 – Scalar product of two vectors in parallel algorithm

## 4 Parallel approach to the solution of the system of linear algebraic equations

It is true that the method of steepest descent converges for any choice of initial approximation to the exact solution of the equation (3).

We will use the method of steepest descent for solving the mentioned system of difference equations. For that reason we describe the method of steepest descent step by step:

- 1) We choose the initial approximation. E. g. we can take the trivial initial solution vector  $\mathbf{U}_0$  of the type (n, 1).
- 2) We calculate residue

$$\mathbf{r}_0 = \mathbf{F} - \mathbf{A} \cdot \mathbf{U}_0.$$

Vector  $\mathbf{r}_0$  determines the direction of steepest descent function  $G(\mathbf{U}) = \frac{1}{2} \mathbf{A} \mathbf{U}^2 - \mathbf{F} \mathbf{U}$  at the point  $\mathbf{U}_0$ , or we can say that it determines the value  $-\mathbf{grad} G(\mathbf{U}_0)$ .

- 3) We determine the value  $a_0$  for which the function  $G(\mathbf{U}_0 - a \cdot \mathbf{r}_0)$  takes its minimum. It is true that

$$a_0 = -\frac{(\mathbf{r}_0, \mathbf{r}_0)}{(\mathbf{r}_0, \mathbf{A} \mathbf{r}_0)}.$$

- 4) We calculate the new vector

$$\mathbf{U}_1 = \mathbf{U}_0 - a_0 \cdot \mathbf{r}_0.$$

Above process is repeated until the norm of residue achieves the prescribed precision.

## 5 Numerical experiment

We apply the above described theory in our numerical experiment. We prepare the parallel algorithm in FORTRAN 90 syntax with MPI implementation MPICH separately for one-dimensional problem and specifically for the two-dimensional problem.

### 5.1 Numerical experiment of one-dimensional problem

We consider

$$-u'' = f \text{ in } \Omega = [0, 1], \quad (4)$$

$$u(0) = 0, u(1) = 0 \quad (5)$$

We cover unit interval by the grid of knots with the equidistant step  $h$ . We choose the right/hand side  $f$  in the problem (4) in such a way that the function

$$u(x) = 4x \cdot (1 - x) \cdot e^x$$

is the exact solution.

The approximate values of the solution  $u$  in all the inner regular knots of the given grid are the numerical solution of the problem (4) – (5). We find this numerical solution by the method of steepest descent with prescribed tolerance  $10^{-5}$  for the norm of residue. Table 1 shows the development of calculation times according to the number of used processors and mesh step, together with the number of iterations and condition number of iterative matrices.

Step	Number of knots	Number of iterations	Condition number	2 processors	4 processors	8 processors
$h = 1/161$	160	16 257	10 505	8 min 5 s	5 min 10 s	4 min 30 s
$h = 1/321$	320	43 027	41 760	27 min 21 s	10 min 49 s	20 min 1 s
$h = 1/641$	640	85 835	166 523	2 h 17 min 56 s	56 min 6 s	1 h 16 min 27 s

Table n. 1 – 1D problem: Calculation times needed for computing the vector  $\mathbf{U}$  of the approximate solution for different mesh sizes of the grid on two, four, and eight processors.

## 5.2 Numerical experiment of two-dimensional problem

We consider

$$-\Delta u = f \text{ in } \Omega = [0, 1]^2, \quad (6)$$

$$u|_{\partial\Omega} = 0 \text{ on } \partial\Omega. \quad (7)$$

We cover the domain  $\Omega$  by the grid of knots with the spatial step  $h$  in the direction of both the coordinate axes  $x$  and  $y$ . We choose function  $f$  in the problem (6) so that the function

$$u(x, y) = 16 \cdot x \cdot y \cdot (1 - x) \cdot (1 - y) \cdot e^{x+y}$$

is the exact solution.

The approximate values of the solution  $u$  in all the inner regular knots of the given grid are the numerical solution of the problem (6) – (7). We find this numerical solution by the method of steepest descent with prescribed tolerance  $10^{-5}$  for the norm of residue. Analogously as in 1D case, Table 2 illustrates the development of calculation times according to the number of used processors with respect to spatial step.

Step	Number of knots	Number of iterations	Condition number	2 processors	4 processors	8 processors
$h = 1/21$	400	1158	178	1 min 45 s	1 min 01 s	41 s
$h = 1/41$	1600	4228	680	18 min 32 s	7 min 4 s	8 min 58 s
$h = 1/81$	6400	15 652	2652	6 h 9 min 50 s	2 h 42 min 43 s	2 h 24 min 26 s

Table n. 2 – 2D problem: Calculation times needed for computing the vector  $\mathbf{U}$  of the approximate solution for different mesh sizes of the grid on two, four, and eight processors.

There is shown the graph of the exact solution of the problem (6) – (7) on the Figure n. 4a. Furthermore, there is drawn the graph of the approximate solution of the problem (6) – (7) for mesh size  $h = \frac{1}{9}$  on the Figure n. 4b, for mesh size  $h = \frac{1}{21}$  on the Figure n. 4c, and for mesh size  $h = \frac{1}{41}$  on the Figure n. 4d.

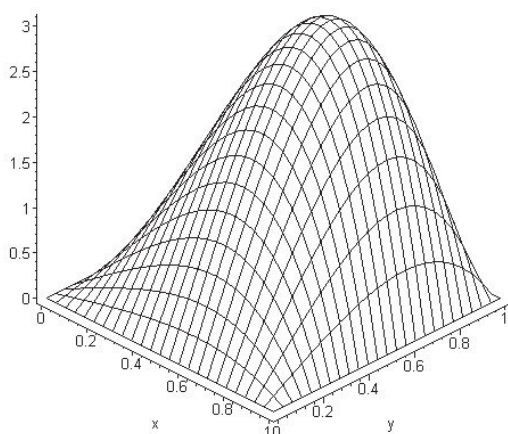


Figure n. 4a  
Exact solution

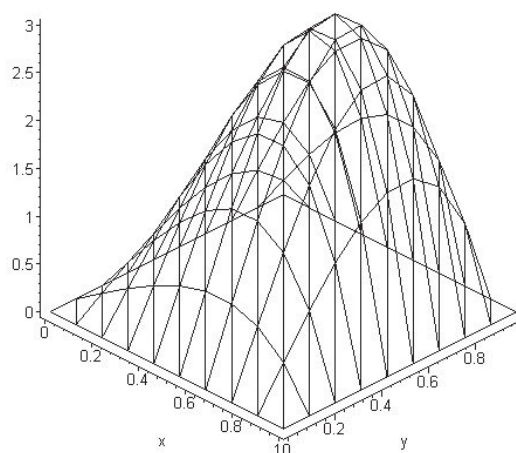


Figure n. 4b  
Approximate solution, mesh size  $h = \frac{1}{9}$

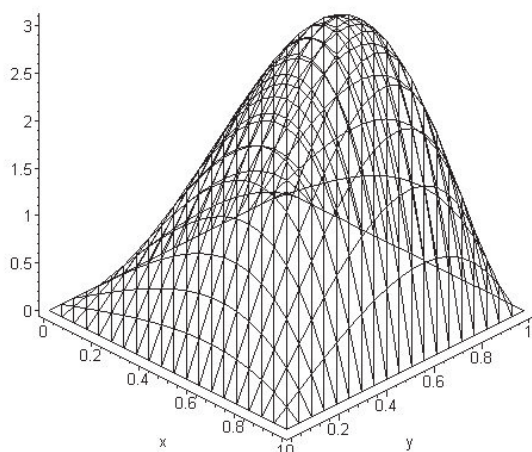


Figure n. 4c

Approximate solution, mesh size  $h = \frac{1}{21}$

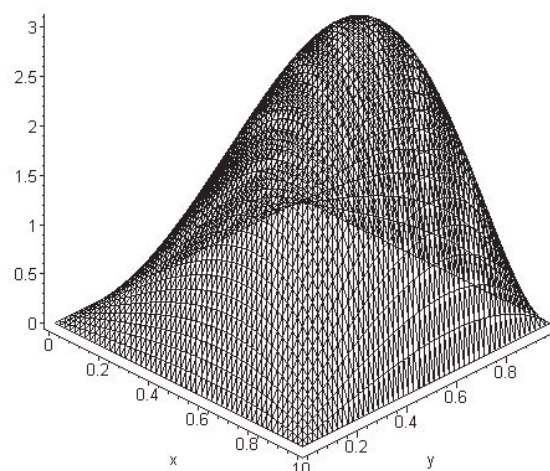


Figure n. 4d

Approximate solution, mesh size  $h = \frac{1}{41}$

## 6 Conclusion

We discussed the basic ways of mapping the matrices in parallel programming. We especially described the striped mapping of a matrix cyclically by rows. We showed the possibilities of applying the basic principles of linear algebra – product of a matrix and of a vector, scalar product of two vectors – in parallel algorithm. We described the method of steepest descent as well as the problem coming out of the Poisson equation. Finally, we commented the numerical experiments in which the mentioned theory is applied in practice.

We can summarize on the basis of the numerical experiment results that computing the approximate solution of the system of difference equations is two times quicker, if we calculate the system of linear equations, on four processors instead of on two processors. Computations on eight processors are not completely optimal. They are influenced by the hierarchy of the cluster. The cluster consists of eight processors and if the computation runs on all the eight processors then the resulting time is distorted by other applications running. On the other hand, the calculations for a larger number of knots were running faster in two-dimensional problem than the computations for a smaller number of knots in one-dimensional problem (see Table 1 and Table 2). This occurs due to condition of iteration matrices.

For the future work it would be interesting to compare the results e.g. with computing the same problem using the other method, e.g. the conjugate gradient method.

## Acknowledgement

The paper was supported by the project ESF of no. CZ.1.07/2.3.00/09.0155 with title "Constitution and improvement of a team for demanding technical computations on parallel computers at TU Liberec".

## References

- [1] DOLEJŠÍ, V. – KNOBLOCH, P. – KUČERA, V. – VLASÁK, M.: *Finite element methods: theory, applications and implementation*. TUL, Liberec 2011. ISBN 978-80-7372-728-4.
- [2] FEISTAUER, M.: *Diskrétní metody řešení diferenciálních rovnic*. SPN, Praha 1981.
- [3] MACKOVÁ, K.: *Gradientní metody řešení soustav lineárních rovnic*. [Diplomová práce.] Univerzita Palackého v Olomouci, Olomouc 2008.

## Current address

**Mgr. Daniela Bímová, Ph.D.**

Department of Mathematics and Didactics of Mathematics

Technical University of Liberec

Studentská 2

461 17 Liberec 1

Czech Republic

Phone number: +420 48 53 52 308

E-mail: daniela.bimova@tul.cz



## BLOCK-CYCLIC-STRIPED MAPPINGS OF MATRICES IN THE PARALLEL PROGRAMMING

BITTNEROVÁ Daniela, (CZ)

**Abstract.** The paper presents one method of the parallel programming – the block-cyclic-striped mapping of a matrix into two processors. The matrix is divided into row-blocks and these blocks are mapped into two processors alternately. The algorithm of it and also an application of the calculation of the matrix-vector product by using the way are shown.

**Key words.** Parallel programming, block-cyclic-striped mapping, product of a matrix and a vector.

*Mathematics Subject Classification:* Primary 65F05, 65F10; Secondary 65K05.

### 1 Introduction

Matrices and vectors play a very important role in scientific and technical computations, especially in the economy and the engineering. The computational technology is developing permanently. The progress of the hardware makes the advantage possible to distribute calculations across multiple processors and by that way to accelerate these calculations. Therefore it is necessary to develop also the software. We need new technologies for mappings of a large quantum of input dates, mostly in a form of large matrices and vectors. These dates are coefficients of algebraic equations usually, which were produced from mathematical numerical models solving some technical problems, for example by differential or integral equations. In last years, new computer architectures are developed. One of them is called the parallel programming. Solving a mathematical problem we can use the so-called computer cluster, which is a block of linked one-processor computers working deeply together.

In our faculty, the cluster consists of two servers – the central server (parallel1) and the secondary server (parallel2). We use the programming language FORTRAN 90 and the operating system LINUX. As a specification of the library functions for FORTRAN 90, we have the standard MPI (Message Passing Interface).

## 2 Mapping of Matrices to Two Processors

Technical calculations depend on the efficiency of matrix operations often, especially on the matrix-vector product. Using the parallel programming, a matrix can be mapped into processors by two ways – by the striped mapping or the checkerboard mapping. Each of these ways we choose one type from. Now we deal with the so-called block-cyclic-striped mapping where the matrix is divided into blocks of rows and these blocks are mapped into processors alternatively.

Let us suppose that we have two processors –  $P_1$  and  $P_2$ . Let  $\mathbf{A} = (a_{i,j})$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, n$ , be a matrix of order  $n$ . The matrix  $\mathbf{A}$ , which is generated by the master processor  $P_0$  only, should be mapped into  $P_1$  and  $P_2$  so that each of  $k$  rows,  $1 < k < n$ , will be in one block of one from both processors. If the number of rows  $n$  is not divided by  $k$ , then a remainder of rows is in the last block. Elements of the matrix  $\mathbf{A}$  will be map into the processors so that the first  $k$  rows go into the  $P_1$ , the second  $k$  rows into the  $P_2$ , the following  $k$  rows into the  $P_1$ , and so on – see Fig. 1.

$P_1$	$k$ rows
$P_2$	$k$ rows
$P_1$	$k$ rows
$P_2$	$k$ rows
$\dots$	

Fig. 1: The Mapping of the Matrix  $\mathbf{A}$  to Blocks into the Processor  $P_1$ , and  $P_2$ .

The algorithm where the  $i$ -th row is mapped into the  $p$ -th processor,  $p = 1$  or  $p = 2$ , will be given by the following relation:

$$i = 2ak + (p-1)k + 1 + b, \quad (1)$$

where

$$a = 0, 1, \dots, \frac{n}{2k} - 1, \quad b = 0, 1, \dots, k - 1. \quad (2)$$

It means that for the processor  $P_1$ , it is

$$p = 1, \quad i = 2ak + 1 + b. \quad (3)$$

For the processor  $P_2$ , it is

$$p = 2, \quad i = 2ak + k + 1 + b. \quad (4)$$

**Example:** The mapping of the matrix  $\mathbf{A}$  of order  $n = 40$  to blocks on the processor  $P_1$ , and  $P_2$  is presented in Fig. 2, where, for the specific constants, the numbers of rows are shown into the corresponding processors.

$P$	$p$	$a$	$b$	$i$ -th row
$P_1$	1	0	0,1,2,3,4	1, ..., 5
		1	0,1,2,3,4	11, ..., 15
		2	0,1,2,3,4	21, ..., 25
		3	0,1,2,3,4	31, ..., 35
$P_2$	2	0	0,1,2,3,4	6, ..., 10
		1	0,1,2,3,4	16, ..., 20
		2	0,1,2,3,4	26, ..., 30
		3	0,1,2,3,4	36, ..., 40

Fig. 2: The mapping of the matrix  $A$  of order  $n = 40$  to blocks into the processor  $P_1$ , and  $P_2$ .

### 3 The Matrix-Vector Multiplication

Let  $A$  be a matrix of order  $n$ ,  $u = (u_1, u_2, \dots, u_n)^T$  a column vector, and  $v = (v_1, v_2, \dots, v_n)^T$  a column vector, which is calculated as a product

$$v = Au. \quad (5)$$

There exist some possibilities how to solve that system by using two parallel processors. The elements of the matrix  $A$  are distributed from the master processor to corresponding processors  $P_1$ , and  $P_2$ , and the numbers  $u_j$ ,  $j = 1, \dots, n$ , to both processors in the  $j^{\text{th}}$  column. All products of the blocks can now be formed in one step. Many applications of parallel computers are conversions of existing sequential programs. The basic (naïve) parallel algorithm is also based on the sequential programs, but now each processor calculates corresponding product by using the following process (in the block-cyclic-striped mapping case, the processor  $P_1$  calculates the rows of the numbers  $1, \dots, k, 2k+1, \dots, 3k, 4k+1, \dots, 5k, \dots$  and the processor  $P_2$  calculates the rows of the numbers  $k+1, \dots, 2k, 3k+1, \dots, 4k, \dots$ , till the last block of rows):

1. Initialize the starting point  $V(1) = 0$ ,  $POM1 = 1$ ,  $POM2 = K + 1$ ,  $NK1 = K$ ,  $NK2 = 2K$ .
2. Calculate in  $P_1$ , respective  $P_2$ :

```

DO 2 J=1,N
    DO 1 I=POM1,NK1           respective DO 1 I=POM2,NK2
        V(I) = V(I) + A(I,J) * U(J)
    1 CONTINUE
2 CONTINUE

```

3. Repeat for  $POM1 = POM1 + 2K$ ,  $NK1 = NK1 + 2K$ , respective  
 $POM2 = POM2 + 2K$ ,  $NK2 = NK2 + 2K$ .

If the matrix  $A$  is sparse, we choose some of special algorithms depended on a work with non-zero elements of the matrix and the vector. Mapping a sparse matrix, we must map together with non-zero elements also the column and row indexes of these elements. Methods differ by the type of mapping of indicators to the beginning of rows and columns each other.

In MPI, the data systems are sent to particular processors through the commands MPI\_SEND and MPI\_RECV, and or MPI\_BCAST, which is a combination of the first two. If we map a small number of various types of data, it is the simplest solution to use the commands MPI\_SEND or MPI\_RECV, and or MPI\_BCAST more times, for large quantum of similar data, the advantageous solution is to use an auxiliary vector. For large quantum of similar or larger number of various types of data systems, it is the best solution to use the commands MPI\_PACK and MPI\_UNPACK – see [4].

#### **4 Conclusion**

The goal of the paper was to present one of the specific methods for a matrices mapping by using the parallel programming – the block-cyclic-striped mapping, which is not so often use. The method could be applied in solving of technical and science method where differential and integral equations are discretized into a set of linear equations.

#### **Acknowledgement**

The paper was supported by the project ESF, no. CZ.1.07/2.3.00/09.0155, „Constitution and Improvement of a Team for Demanding Technical Computations on Parallel Computers at TU Liberec.“

#### **References**

- [1] BITTNEROVÁ, D.: *The Boston Matrix and the Checkerboard Mapping*. In ICPM'11, TUL, Liberec, pp. 29-33, 2011.
- [2] HOZMAN, J.: *Study materials*. [http://kmd.fp.tul.cz/lide/hozman/M3A/M3A-10\\_12-FS.pdf](http://kmd.fp.tul.cz/lide/hozman/M3A/M3A-10_12-FS.pdf)
- [3] JENNKINGS, A. – McKEOWN, J. J.: *Matrix Computation*. Wiley, Chichester, 1992. 2nd Edition.
- [4] RAUBER, T. – RÜNGER, G.: *Parallel Programming: For Multicore and cluster systems*. Springer Verlag 2010. ISBN 978-3-642-04817-3.
- [5] [http://physics.ujep.cz/~mlisal/par\\_prog/pprg-web.pdf](http://physics.ujep.cz/~mlisal/par_prog/pprg-web.pdf)

#### **Current address**

**Daniela Bittnerová, RNDr., CSc.**

Technical Univerzity of Liberec

Studentská 2, 461 17 Liberec, Czech Republic

Tel. +420 485352310

Email: [daniela.bittnerova@tul.cz](mailto:daniela.bittnerova@tul.cz)

## ON EFFICIENCY OF APPROXIMATE MATRIX-VECTOR MULTIPLICATION IN ADAPTIVE WAVELET METHODS

ČERNÁ Dana, (CZ), FINĚK Václav, (CZ)

**Abstract.** In recent years, wavelets have been successfully used for the numerical solution of operator equations. Among most important advantages of wavelets belong the following two properties. They allow to characterize various function spaces such as Sobolev or Besov spaces by weighted sequence norms of the corresponding wavelet coefficient and they have cancellation properties. It means that the inner product of a smooth function and a wavelet vanishes or decreases fast as the scale of the wavelet increases. These two properties of wavelets can be exploited considerably in numerical solution of differential equations. Due to the cancellation properties of wavelets, a representation of functions as well as representation of differential operators in wavelet coordinates is sparse or quasi sparse. And further, a consequence of the equivalences between function norms and weighted sequence norms is efficient diagonal preconditioning for stiffness matrices. To be able to solve realistic problems, it is necessary to use adaptive methods with highly nonuniform meshes to keep the number of unknowns at a reasonable level. Key ingredients are a posteriori error estimators and adaptive refinement strategies. As reliable a posteriori estimators serve wavelet expansions of the current residual. This is the consequence of the above mentioned norm equivalences. One of the basic adaptive refinement strategies is based on iterations in the infinite-dimensional space which are carefully approximated by choosing accuracies in numerical subroutines for an approximation of the right-hand side and an approximation of matrix-vector multiplications. New elements of an unknown solution are then generated by increasing accuracy in both subroutines. The most time consuming part of this approach is the matrix-vector multiplication and therefore it is necessary to perform it in the most efficient way. In our contribution, we compare different approaches for approximate matrix-vector multiplication.

**Key words and phrases.** Wavelet, adaptive methods, matrix-vector multiplication.

*Mathematics Subject Classification.* 65T60, 65F99, 65N99.

## 1 Introduction

In [4, 5], automatically adaptive and asymptotically optimal wavelet based methods were proposed. They consists from the following three steps:

- To transform a variational formulation into the well-conditioned infinite-dimensional  $l^2$  problem.
- To find a convergent iteration process for the  $l^2$  problem which works with infinite vectors, the exact right hand side and exact matrix-vector multiplication.
- To derive a finite dimensional version of above idealized iteration process with an inexact right hand side and approximate matrix-vector multiplication. The algorithm should provide an approximation of the unknown solution up to a given target accuracy  $\epsilon$ , a convergence rate should match the rate of the best  $N$ -term approximation, and the associated computational work should be proportional to the number of unknowns.

Efficient approximate matrix-vector multiplication is enabled by a fast off-diagonal decay of entries of the stiffness matrix and a fast decay of the load vector in wavelet coordinates. In [4], a numerical routine **APPLY** was proposed which approximates the exact matrix-vector product with the desired tolerance  $\epsilon$  and that has linear computational complexity, up to sorting operations. In [6], binning and approximate sorting was used to eliminate sorting costs and then an algorithm with linear complexity was obtained. An optimized version of the approach from [4] was proposed in [7]. Authors optimize estimated number of matrix-vector multiplication subject to estimated multiplication error. To better utilize actual decay of matrix entries, a modified approach was proposed in [2]. Vector entries are not sorted with respect to their size but instead an actual decay of matrix entries is measured. Consequently in dependence on this decay, the multiplication is performed. Also this approach is asymptotically optimal. At the end, we use above mentioned matrix-vector multiplication techniques to solve adaptively Poisson equation and compare their efficiency. In numerical experiments, we use wavelet bases proposed in [1].

## 2 Discretization

Let  $H$  be a real Hilbert space with the inner product  $(\cdot, \cdot)_H$  and the induced norm  $\|\cdot\|_H$ . Let  $A : H \rightarrow H'$  be the selfadjoint and  $H$ - elliptic differential operator, i.e.

$$a(v, w) := (Av, w) \lesssim \|v\|_H \|w\|_H \quad \text{and} \quad a(v, v) \sim \|v\|_H^2.$$

Then, there exist positive constants  $c_A$  and  $C_A$  such that

$$c_A \|v\|_H \leq \|Av\|_{H'} \leq C_A \|v\|_H, \quad v \in H$$

and the equation  $Au = f$  has for any  $f \in H'$  a unique solution. Further we assume that  $\mathbf{D}^{-1}\Psi$ ,  $\Psi = \{\psi_\lambda, \lambda \in I\}$ , is a suitable wavelet (Riesz) basis in the energy space  $H$  and  $I$  an index set. Then, there exist positive constants  $c_\psi$  and  $C_\psi$  such that

$$c_\psi \|\mathbf{v}\|_2 \leq \|\mathbf{v}^T \mathbf{D}^{-1} \Psi\|_H \leq C_\psi \|\mathbf{v}\|_2, \quad \mathbf{v} \in l^2(I) \quad (1)$$

and consequently

$$Au = f \quad \Leftrightarrow \quad \mathbf{A}\mathbf{u} = \mathbf{f},$$

where  $\mathbf{D} := \text{diag}(\omega_\lambda)_{\lambda \in I}$ ,  $\omega_\lambda = \sqrt{(A\psi_\lambda, \psi_\lambda)}$ ,  $\mathbf{A} = \mathbf{D}^{-1}(A\Psi, \Psi)\mathbf{D}^{-1}$  is a biinfinite diagonally preconditioned stiffness matrix,  $u = \mathbf{u}^T \mathbf{D}^{-1} \Psi$  and  $\mathbf{f} = \mathbf{D}^{-1}(f, \Psi)$ . The condition number of matrix  $\mathbf{A}$  satisfies

$$\kappa(\mathbf{A}) \leq \frac{C_\psi^2 C_A}{c_\psi^2 c_A} < +\infty \quad (2)$$

and the same holds (matrix  $\mathbf{A}$  is positive definite) for all finite sections

$$\mathbf{A}_\Lambda := \mathbf{D}^{-1}(A\Psi_\Lambda, \Psi_\Lambda)\mathbf{D}^{-1}, \quad \Psi_\Lambda := \{\psi_\lambda, \lambda \in \Lambda\}, \quad \Lambda \subset I.$$

To solve the above mentioned system of equations, we use the steepest descent scheme

$$\mathbf{u}^0 := \mathbf{0}, \quad \mathbf{u}^{n+1} := \mathbf{u}^n + \frac{(\mathbf{r}^n, \mathbf{r}^n)}{(\mathbf{r}^n, \mathbf{A}\mathbf{r}^n)} \mathbf{r}^n, \quad \mathbf{r}^n = \mathbf{f} - \mathbf{A}\mathbf{u}^n, \quad n = 0, 1, \dots, \quad (3)$$

which has the error reduction parameter equal to

$$\rho = \frac{\kappa(\mathbf{A}) - 1}{\kappa(\mathbf{A}) + 1} = 1 - \frac{2}{\kappa(\mathbf{A}) + 1} < 1.$$

And then it is a convergent process.

### 3 Approximate Matrix-Vector Multiplications

In [4], authors exploited an off-diagonal decay of entries of the wavelet stiffness matrices and a decay of entries of the load vector in wavelet coordinates to design a numerical routine **APPLY** which approximates the exact matrix-vector product with the desired tolerance  $\epsilon$  and that has linear computational complexity, up to sorting operations. An example of the decay of matrix entries can be found in [3]. The idea of **APPLY** for one dimensional problems is the following: To truncate  $\mathbf{A}$  in scale by zeroing  $a_{i,j}$  whenever  $\delta(i, j) > k$  ( $\delta$  represents the level difference of two functions in the wavelet expansion) and denote resulting matrix by  $\mathbf{A}_k$ . At the same time, vector entries  $\mathbf{v}$  are sorted with respect to the size of their absolute values. One obtains  $\mathbf{v}_k$  by retaining  $2^k$  biggest coefficients in absolute values of  $\mathbf{v}$  and setting all other equal to zero. The maximum value of  $k$  is determined in a similar way as in [7] to reach a desired accuracy by using upper bounds for errors of matrix approximations. Then one computes an approximation of  $\mathbf{A}\mathbf{v}$  by

$$\mathbf{w} := \mathbf{A}_k \mathbf{v}_0 + \mathbf{A}_{k-1}(\mathbf{v}_1 - \mathbf{v}_0) + \dots + \mathbf{A}_0(\mathbf{v}_k - \mathbf{v}_{k-1}) \quad (4)$$

with the aim to balance both accuracy and computational complexity.

In [6], binning and approximate sorting was used to eliminate sorting costs and then an algorithm with linear complexity was obtained. The idea is following: Reorder the elements of  $\mathbf{v}$  into the sets  $V_0, \dots, V_q$ , where  $v_\lambda \in V_i$  if and only if

$$2^{-i-1} \|\mathbf{v}\|_2 < |v_\lambda| < 2^{-i} \|\mathbf{v}\|_2, \quad 0 \leq i < q.$$

And then to generate vectors  $\mathbf{v}_k$  by successively extracting  $2^k$  elements from  $\bigcup_i V_i$ , starting from  $V_0$  and when it is empty continuing with  $V_1$  and so forth. Finally the scheme (4) is applied.

An optimized version of the above approach was proposed in [7]. The indices of  $\mathbf{v}$  are stored in buckets, depending on the modulus of the corresponding wavelet coefficients in this way:

$$v_\lambda \in \mathbf{v}_k \iff 2^{-(k+1)/2} \|\mathbf{v}\|_\infty < |v_\lambda| \leq 2^{-k/2} \|\mathbf{v}\|_\infty.$$

Then  $\forall \mathbf{v} \in l_2(\Lambda)$ , we compute the approximate matrix vector product by  $\sum_{k=0} \mathbf{A}_{j_k} \mathbf{v}_k$ , where  $j_k \in \mathbf{N}_0$  are solutions of

$$\sum_{k=0} c_{j_k} \# \mathbf{v}_k \longrightarrow \min!, \quad \sum_{k=0} e_{j_k} \|\mathbf{v}_k\| \leq \epsilon - \delta, \quad (5)$$

$$\text{and } \delta = \|\mathbf{A}\| \left\| v - \sum_{k=0} \mathbf{v}_k \right\| \leq \epsilon/2, \quad (6)$$

where  $\mathbf{A}_j$  and  $e_j$  are matrices and constants such that  $\|\mathbf{A} - \mathbf{A}_j\| \leq e_j$  and  $c_j$  are upper bounds for the number of non-zero entries in each column of  $\mathbf{A}_j$ . So, they try to optimize number of arithmetic operations.

To better utilize the actual decay of matrix and vector entries, in [2] a different approach was designed. We are not searching for  $2^k$  biggest vector entries in absolute value but instead we trace actual decay of matrix and vector entries and then the actual number of entries in  $\mathbf{v}_k$  depends on these decays. Let us denote  $S_{A_k} := \max\{|a_{i,j}|, \delta(i,j) = k\}$ . Then, we multiply matrix  $\mathbf{A}_0$  with vector entries which are greater than given tolerance  $\epsilon_k$ , matrix  $\mathbf{A}_1 - \mathbf{A}_0$  with vector entries which are greater than  $\epsilon_k/S_{A_1}$ , ..., and matrix  $\mathbf{A}_k - \mathbf{A}_{k-1}$  with vector entries which are greater than  $\epsilon_k/S_{A_k}$ . The value of  $\epsilon_k$  is determined to reach a desired accuracy of approximation. In [2], an asymptotic optimality of this multiplication algorithm was proved.

## 4 Numerical examples

To obtain a computable version of the iteration process (3), an inexact right hand side and an approximate matrix-vector multiplication have to be used. An approximate wavelet expansion of a right-hand side  $\mathbf{f}$  in the dual basis can be computed up to any given accuracy. Its realization consists of a projection of  $\mathbf{f}$  onto a fine multiresolution space, followed by a thresholding.

We employ the finite version of the ideal iteration (3) with a gradually increasing accuracy in the inexact right hand side and in the approximate matrix-vector multiplication. First we compute an approximation of the right-hand side. Then, we enlarge the set of active coefficients by coefficients generated by the inexact right hand side and by an approximate matrix-vector multiplication (outer iteration). Consequently, we apply several (inner) iterations of steepest descent scheme with fixed precision in both subroutines to reduce the residual efficiently. Finally, the accuracy is increased and this iterative procedure is repeated until desired accuracy has been reached.

At the end, we present numerical comparison of different approximate matrix-vector multiplication techniques proposed in [4], [7], and in [2]. In numerical experiments, we employ the quadratic wavelet basis (3,3) and cubic basis (4,4) both proposed in [1] to solve the one dimensional Poisson equation:

$$-u'' = f, \quad \text{in } \Omega = (-1, 1), \quad u(-1) = u(1) = 0,$$

OI	<i>CDV</i>				<i>DSS</i>			
	II	#	$\ \mathbf{u} - \mathbf{u}_i\ _{L_2}$	DOF	II	#	$\ \mathbf{u} - \mathbf{u}_i\ _{L_2}$	DOF
1	15	2020	0.067644	53	15	561	0.049276	12
2	8	4272	0.018469	80	9	2088	0.016542	18
3	8	9247	0.005069	110	14	5997	0.003061	34
4	9	29948	0.001205	153	9	21540	0.000872	57
5	8	63069	0.000336	234	6	57996	0.000347	106
6	9	130006	0.000085	354	10	135859	0.000086	209
7	9	229305	0.000034	556	9	222653	0.000035	415

Table 1: Results for the quadratic basis (3,3).

OI	<i>CF<sub>1</sub></i>				<i>CF<sub>2</sub></i>			
	II	#	$\ \mathbf{u} - \mathbf{u}_i\ _{L_2}$	DOF	II	#	$\ \mathbf{u} - \mathbf{u}_i\ _{L_2}$	DOF
1	17	465	0.059498	18	10	66	0.161959	8
2	13	1001	0.012127	26	10	176	0.038192	14
3	9	3181	0.003241	36	6	737	0.012804	23
4	11	11490	0.000745	64	6	1960	0.004668	43
5	5	25882	0.000313	148	8	6371	0.001269	92
6	8	52932	0.000088	294	9	16127	0.000299	191
7	9	107703	0.000034	571	9	40577	0.000076	453

Table 2: Results for the quadratic basis (3,3).

with the solution  $u$  given by

$$u(x) = (x^2 - 1)e^x \quad \forall x \in \Omega.$$

The maximum wavelet decomposition level was set equal to ten for the quadratic basis and nine for the cubic basis. The required precision was set equal to  $4^{-OI}$ , where OI represents the order of outer iteration. Columns denoted by  $CF_1$  contain results obtained by the method proposed in [2] using the upper bounds of errors in the similar way as in  $CDV$ , columns denoted by  $CF_2$  contain results obtained by the method proposed in [2] using an error approximation, columns denoted by  $DSS$  contain results obtained by approach (5, 6) proposed in [7], and finally columns denoted by  $CDV$  contain results obtained by approach proposed in [4]. The error approximation in  $CF_2$  is computed in this way: We compute an approximate matrix-vector multiplication and at the same time we compute also an error indicator as difference between approximate matrix-vector multiplications with bins  $\mathbf{v}_k$  containing all elements greater than  $\epsilon_k/S_{A_k}$ , and with bins  $\mathbf{v}_k$  containing all elements greater than  $2\epsilon_k/S_{A_k}$ , respectively. Consequently, until the error indicator is greater than the required error, we recompute an approximate matrix-vector multiplication with slightly moved bins by decreasing  $\epsilon_k$ . In all tables, II denotes the number of inner iterations in the given outer iteration, # represents the mean number of element by element multiplications in the given outer iteration,  $\|\mathbf{u} - \mathbf{u}_i\|_{L_2}$  is the  $L_2$  norm of the difference between the exact and an approximate solution, and DOF denotes the number of active coefficients at the end of the given outer iteration.

## 5 Conclusion

Presented results show that the approximate matrix-vector multiplication technique proposed in [2] is more efficient than methods proposed in [4, 7] because it is less computationally demanding and in the case of cubic basis, it even produced substantially sparser approximate solutions.

## Acknowledgement

The authors have been supported by the project ESF "Constitution and improvement of a team for demanding technical computations on parallel computers at TU Liberec" No. CZ.1.07/2.3.00/09.0155.

OI	<i>CDV</i>				<i>DSS</i>			
	II	#	$\ \mathbf{u} - \mathbf{u}_i\ _{L_2}$	DOF	II	#	$\ \mathbf{u} - \mathbf{u}_i\ _{L_2}$	DOF
1	35	4508	0.199827	98	29	1812	0.196595	30
2	32	9088	0.052809	140	24	3446	0.041889	46
3	34	15400	0.012869	214	28	4179	0.009039	57
4	32	27072	0.003422	264	27	7423	0.002856	84
5	34	48261	0.000848	317	31	14881	0.000802	145
6	34	86740	0.000214	392	32	30938	0.000218	262
7	34	97001	0.000055	508	34	54531	0.000055	450
8	35	173778	0.000014	635	34	87197	0.000014	681

Table 3: Results for the cubic basis (4,4).

OI	<i>CF<sub>1</sub></i>				<i>CF<sub>2</sub></i>			
	II	#	$\ \mathbf{u} - \mathbf{u}_i\ _{L_2}$	DOF	II	#	$\ \mathbf{u} - \mathbf{u}_i\ _{L_2}$	DOF
1	35	1405	0.191655	42	11	242	0.460329	15
2	33	2421	0.047619	55	21	569	0.116968	31
3	32	3507	0.012201	69	11	1844	0.008002	61
4	30	5442	0.003007	79	3	4015	0.006956	88
5	26	8909	0.000728	94	15	5102	0.002579	115
6	24	16804	0.000187	142	31	7360	0.000711	148
7	24	25041	0.000051	215	29	13408	0.000221	193
8	26	44208	0.000013	340	35	21152	0.000054	280

Table 4: Results for the cubic basis (4,4).

## References

- [1] ČERNÁ; D., FINĚK, V.: *Construction of Optimally Conditioned Cubic Spline Wavelets on the Interval*. Adv. Comput. Math., vol. 34, pp. 519-552, 2011.

- [2] ČERNÁ, D.; FINĚK, V.: *Approximate Multiplication in Adaptive Wavelet Methods*. Submitted, 2010.
- [3] ČERNÁ, D.; FINĚK, V.: *Some Aspects of Adaptive Wavelet Methods*. In: TCP 2011, Praha, 2011.
- [4] COHEN, A., DAHMEN, W., DEVORE, R.: *Adaptive Wavelet Schemes for Elliptic Operator Equations - Convergence Rates*. Math. Comput. 70, no. 233, pp. 27-75, 2001.
- [5] COHEN, A., DAHMEN, W., DEVORE, R.: *Adaptive wavelet methods II - beyond the elliptic case*. Found. Math., Vol. 2, pp. 203-245, 2002.
- [6] STEVENSON, R.: *Adaptive solution of operator equations using wavelet frames*, SIAM J. Numer. Anal. 41, no. 3, 1074 - 1100, 2003.
- [7] DIJKEMA, T. J., SCHWAB, Ch., STEVENSON, R.: *An adaptive wavelet method for solving high-dimensional elliptic PDEs*, Constructive approximation, 30, no. 3, 423-455, 2009.

### Current address

#### **Dana Černá, Mgr. Ph.D.**

Department of Mathematics and Didactics of Mathematics, Technical University in Liberec, Studentská 2, Liberec, 46117, Czech Republic, dana.cerna@tul.cz

#### **Václav Finěk, RNDr. Ph.D.**

Department of Mathematics and Didactics of Mathematics, Technical University in Liberec, Studentská 2, Liberec, 46117, Czech Republic, vaclav.finek@tul.cz



## DISCRETE WAVELET TRANSFORM FOR FINITE SIGNALS

ČERNÁ Dana, (CZ), FINĚK Václav, (CZ)

**Abstract.** Originally, the discrete wavelet transform was designed for an infinite input data. Therefore, applying the discrete wavelet transform to finite signals directly leads to the artifacts near the boundary. There are several methods to handle this problem. They consist in padding of a signal and applying the discrete wavelet transform to the extended signal or using periodized wavelet filters. An alternative approach consists of using the discrete wavelet transform with special filters near the boundary. The boundary filters are derived from wavelet bases on the interval. The desired property is the small condition number of a wavelet basis. The condition number guaranties the stability of the computation and it affects the constants in error estimates. The second important issue is the size of filter coefficients. In the paper, we propose a method which enables to optimize the condition number of transform matrices as well as the size of boundary filter coefficients.

**Key words and phrases.** Wavelet, interval, discrete wavelet transform, finite signal.

*Mathematics Subject Classification.* Primary 65T60, 65D30; Secondary 65D07.

## 1 Introduction

Discrete wavelet transform (DWT) uses two pairs of filters: a low pass filter  $\{h_k\}$  and a high pass filter  $\{g_k\}$  for a decomposition and a low pass filter  $\{\tilde{h}_k\}$  and a high pass filter  $\{\tilde{g}_k\}$  for a reconstruction. The one level of the discrete wavelet transforms maps the vector  $\mathbf{c}^j = (c_1^j, \dots, c_{2m}^j)$  to vectors  $\mathbf{c}^{j+1} = (c_1^{j+1}, \dots, c_m^{j+1})$  and  $\mathbf{d}^{j+1} = (d_1^{j+1}, \dots, d_m^{j+1})$ . It is given by the formula

$$c_k^{j+1} = \sum_{l \in \mathbb{Z}} h_l c_{2k+l}^j, \quad d_k^{j+1} = \sum_{l \in \mathbb{Z}} g_l c_{2k+l}^j, \quad k = 1, \dots, m. \quad (1)$$

As can be seen from this formula, applying DWT near the boundary of the signal requires the values  $c_l^j$  also for  $l < 1$  or  $l > 2m$ , which are not defined. There are several methods to handle this problem. They consist in padding of the signal and applying discrete wavelet

transform to the extended signal. As an example, consider the finite-length input signal 1 2 3 4 5. The signal can be extended by the following methods:

- Zero-padding: ... 0 0 0 0 **1 2 3 4 5** 0 0 0 0 ...
- Symmetrization: ... 4 3 2 1 **1 2 3 4 5** 5 4 3 2 ... (half point),  
... 5 4 3 2 **1 2 3 4 5** 4 3 2 1 ... (whole point)
- Antisymmetric padding: ... -4 -3 -2 -1 **1 2 3 4 5** -5 -4 -3 -2 ...
- Smooth padding of order 1: ... -3 -2 -1 0 **1 2 3 4 5** 6 7 8 9 ...
- Smooth padding of order 0: ... 1 1 1 1 **1 2 3 4 5** 5 5 5 5 ...
- Periodic-padding: ... 2 3 4 5 **1 2 3 4 5** 1 2 3 4 ...

The most popular method is a symmetrization, e.g. in wavelet based image compression standard JPEG2000, the half point symmetrization is performed for a filter of even length and the whole point symmetrization is performed for a filter of odd length [8].

An alternative approach consists of using the discrete wavelet transform with special filters near the boundary. The boundary filters are derived from wavelet bases on the interval. Orthonormal wavelet bases on the interval were constructed in [5]. More general biorthogonal wavelet bases were adapted to the interval in [1, 6, 7]. However, the condition numbers of some of the constructed bases are large. It can cause problems in applications. Better conditioned biorthogonal wavelet bases on the interval were constructed in [2, 3, 9]. The second important aspect is the size of filter coefficients. It plays role e.g. in image compression, where a quantization is performed, because a quantization of large coefficients can cause large errors. In the paper, we propose a method which enables to optimize the condition number of refinement matrices and the size of the filter coefficients.

## 2 Wavelet bases in $L^2(\mathbb{R})$

First, we shortly describe the concept of a wavelet basis for  $L^2(\mathbb{R})$ . Let  $\langle \cdot, \cdot \rangle$  be an inner product and  $\|\cdot\|$  be a norm in  $L^2(\mathbb{R})$ . Let  $l^2$  be a space of  $\mathbf{v} := \{v_{j,k}\}_{j,k \in \mathbb{Z}}$  satisfying

$$\|\mathbf{v}\|_{l^2} := \sum_{j,k \in \mathbb{Z}} |v_{j,k}|^2 < \infty. \quad (2)$$

**Definition 2.1** A function  $\psi \in L^2(\mathbb{R})$  is called a wavelet if the family  $\Psi := \{\psi_{j,k}\}_{j,k \in \mathbb{Z}}$ , where  $\psi_{j,k} := 2^{j/2} \psi(2^j \cdot -k)$ , is a Riesz basis in  $L^2(\mathbb{R})$ , i.e.  $\Psi$  is complete in  $L^2(\mathbb{R})$  and there exist constants  $c, C \in (0, \infty)$  such that

$$c \|\mathbf{v}\|_{l^2} \leq \left\| \sum_{j,k \in \mathbb{Z}} v_{j,k} \psi_{j,k} \right\| \leq C \|\mathbf{v}\|_{l^2}, \quad \mathbf{v} \in l^2. \quad (3)$$

The functions  $\psi_{j,k}$  are also called wavelets.

Let  $V_j$  be the closure of the span of the set  $\{\psi_{j,k}, j, k \in \mathbb{Z}\}$  and let us suppose that there exists a function  $\phi$  such that  $\Phi_j := \{\phi_{j,k}, k \in \mathbb{Z}\}$ ,  $\phi_{j,k} := \phi(2^j \cdot -k)$ , is a Riesz basis of  $V_j$ . Functions  $\phi$  and  $\phi_{j,k}$  are called *scaling functions*.

Then there exists a sequence  $\{h_k\}_{k \in \mathbb{Z}}$  such that

$$\phi(x) = \sum_{k \in \mathbb{Z}} h_k \phi(2x - k) \quad \text{for all } x \in \mathbb{R}. \quad (4)$$

This equation is called a *refinement* or a *scaling equation* and the coefficients  $h_k$  are known as *scaling* or *refinement coefficients*.

By the Riesz representation theorem, there exists a unique family  $\tilde{\Psi} = \{\psi_{j,k}\}_{j,k \in \mathbb{Z}}$  biorthogonal to  $\Psi$ , i.e.

$$\langle \psi_{i,k}, \tilde{\psi}_{j,l} \rangle = \delta_{i,j} \delta_{k,l}, \quad i, j, k, l \in \mathbb{Z}, \quad (5)$$

where  $\delta_{i,j}$  denotes a Kronecker delta. The family  $\tilde{\Psi}$  is also a Riesz basis for  $L^2(\mathbb{R})$ . The basis  $\Psi$  is called a *primal* wavelet basis,  $\tilde{\Psi}$  is called a *dual* wavelet basis. *Dual scaling basis*  $\tilde{\Phi}$ , *dual scaling functions*  $\tilde{\phi}$  and  $\tilde{\phi}_{j,k}$ , and *dual refinement coefficients*  $\tilde{h}_k$  are defined in a similar way.

We define *wavelet coefficients* as

$$g_n = (-1)^n \tilde{h}_{1-n}, \quad \tilde{g}_n = (-1)^n h_{1-n}. \quad (6)$$

Wavelets are then given by

$$\psi(x) = \sum_{n \in \mathbb{Z}} g_n \phi(2x - n), \quad \tilde{\psi}(x) = \sum_{n \in \mathbb{Z}} \tilde{g}_n \tilde{\phi}(2x - n) \quad \text{for all } x \in \mathbb{R}. \quad (7)$$

We say that the wavelet  $\psi$  has  $n$  vanishing moments, if

$$\int_{\mathbb{R}} x^l \phi(x) dx = 0, \quad \text{for } l = 0, \dots, n-1. \quad (8)$$

It is equivalent with *the polynomial exactness* of order  $n$  of the dual scaling functions. It means that any polynomial up to order  $n-1$  lies in  $\tilde{V}_j$ .

The coefficients  $\{h_k\}$  and  $\{g_k\}$  are used for the discrete wavelet transform defined by (1). Then the vector  $\mathbf{c}^{j+1}$  represents coarse approximation of the vector  $\mathbf{c}^j$  and the vector  $\mathbf{d}^{j+1}$  represents details. Then the vector  $\mathbf{c}^{j+1}$  is transformed to vectors  $\mathbf{c}^{j+2}$  and  $\mathbf{d}^{j+2}$ . After  $n$  steps we obtain vector  $(\mathbf{c}^n, \mathbf{d}^n, \dots, \mathbf{d}^1)$ . The inverse discrete wavelet transform is an inverse process and it is given by formula

$$c_k^j = \sum_{n=1}^m \left( \tilde{h}_{k-2n} c_n^{j+1} + \tilde{g}_{k-2n} d_n^{j+1} \right), \quad k = 1, \dots, 2m. \quad (9)$$

### 3 Wavelet bases on the interval

The concept of wavelet bases on the interval is similar. Some scaling functions and wavelets on the interval are just the restrictions of scaling functions and wavelets on the real line, only at the boundaries special functions have to be constructed.

We consider the space  $L^2(I)$ , where  $I = [0, 1]$ , and we denote the  $L^2(I)$ -inner product by  $\langle \cdot, \cdot \rangle_I$  and the  $L^2(I)$ -norm by  $\|\cdot\|_I$ , respectively. Let  $\mathcal{J}$  be some index set and let each index  $\lambda \in \mathcal{J}$  take the form  $\lambda = (j, k)$ , where  $|\lambda| := j \in \mathbb{Z}$  is a *scale* or a *level*. Let

$$l^2(\mathcal{J}) := \left\{ \mathbf{v} : \mathcal{J} \rightarrow \mathbb{R}, \sum_{\lambda \in \mathcal{J}} |\mathbf{v}_\lambda|^2 < \infty \right\}. \quad (10)$$

A family  $\Psi := \{\psi_\lambda \in \mathcal{J}\} \subset L^2(I)$  is called a *wavelet basis* of  $L^2(I)$ , if

- i)  $\Psi$  is a *Riesz basis* for  $L^2(I)$ , i.e.  $\Psi$  is complete in  $L^2(I)$  and there exist constants  $c, C \in (0, \infty)$  such that

$$c \|\mathbf{b}\|_{l^2(\mathcal{J})} \leq \left\| \sum_{\lambda \in \mathcal{J}} b_\lambda \psi_\lambda \right\|_I \leq C \|\mathbf{b}\|_{l^2(\mathcal{J})}, \quad \mathbf{b} := \{b_\lambda\}_{\lambda \in \mathcal{J}} \in l^2(\mathcal{J}). \quad (11)$$

Constants  $c_\psi := \sup \{c : c \text{ satisfies (11)}\}$ ,  $C_\psi := \inf \{C : C \text{ satisfies (11)}\}$  are called *Riesz bounds* and  $\text{cond } \Psi := C_\psi / c_\psi$  is called the *condition* of  $\Psi$ .

- ii) The functions are *local* in the sense that  $\text{diam}(\Omega_\lambda) \leq C 2^{-|\lambda|}$  for all  $\lambda \in \mathcal{J}$ , where  $\Omega_\lambda$  is the support of  $\psi_\lambda$ , and at a given level  $j$  the supports of only finitely many wavelets overlap in any point  $x \in I$ .

By the Riesz representation theorem, there exists a unique family

$$\tilde{\Psi} = \{\tilde{\psi}_\lambda, \lambda \in \mathcal{J}\} \subset L^2(I) \quad (12)$$

biorthogonal to  $\Psi$ , i.e.

$$\langle \psi_{i,k}, \tilde{\psi}_{j,l} \rangle_I = \delta_{i,j} \delta_{k,l}, \quad \text{for all } (i,k), (j,l) \in \mathcal{J}. \quad (13)$$

This family is also a Riesz basis for  $L^2(I)$ . The basis  $\Psi$  is called a *primal* wavelet basis,  $\tilde{\Psi}$  is called a *dual* wavelet basis.

In many cases, the wavelet system  $\Psi$  is constructed with the aid of a multiresolution analysis. A sequence  $S = \{S_j\}_{j \geq j_0}$ , of closed linear subspaces  $S_j \subset L^2(I)$  is called a *multiresolution* or *multiscale analysis*, if

$$S_{j_0} \subset S_{j_0+1} \subset \dots \subset S_j \subset S_{j+1} \subset \dots \subset L^2(I) \quad (14)$$

and  $\cup_{j \geq j_0} S_j$  is complete in  $L^2(I)$ . The dual wavelet system  $\tilde{\Psi}$  generates a *dual* multiresolution analysis  $\tilde{S} = \{\tilde{S}_j\}_{j \geq j_0}$ .

The nestedness and the closedness of the multiresolution analysis implies the existence of the *complement spaces*  $W_j$  such that

$$S_{j+1} = S_j \oplus W_j, \quad W_j \perp \tilde{S}_j. \quad (15)$$

We now assume that  $S_j$  and  $W_j$  are spanned by sets of basis functions

$$\Phi_j := \{\phi_{j,k}, k \in \mathcal{I}_j\}, \quad \Psi_j := \{\psi_{j,k}, k \in \mathcal{J}_j\}, \quad (16)$$

where  $\mathcal{I}_j, \mathcal{J}_j$  are finite or at most countable index sets. We refer to  $\phi_{j,k}$  as *scaling functions* and  $\psi_{j,k}$  as *wavelets*. The multiscale basis is given by

$$\Psi_{j_0,s} = \Phi_{j_0} \cup \bigcup_{j=j_0}^{j_0+s-1} \Psi_j \quad (17)$$

and the wavelet basis of  $L^2(I)$  is obtained by

$$\Psi = \Phi_{j_0} \cup \bigcup_{j \geq j_0} \Psi_j \quad (18)$$

Dual scaling basis  $\tilde{\Phi}_{j_0}$  and dual wavelet basis  $\tilde{\Psi}_j$  are defined in a similar way.

*Polynomial exactness* of order  $N \in \mathbb{N}$  for the primal scaling basis and of order  $\tilde{N} \in \mathbb{N}$  for the dual scaling basis is another desired property of wavelet bases. It means that

$$\mathbb{P}_{N-1}(I) \subset S_j, \quad \mathbb{P}_{\tilde{N}-1}(I) \subset \tilde{S}_j, \quad j \geq j_0, \quad (19)$$

where  $\mathbb{P}_m(I)$  is the space of all algebraic polynomials on  $I$  of degree less or equal to  $m$ .

By Taylor theorem, the polynomial exactness of order  $\tilde{N}$  on the dual side is equivalent to  $\tilde{N}$  vanishing wavelet moments on the primal side, i.e.

$$\int_0^1 P(x) \psi_\lambda(x) dx = 0, \quad P \in \mathbb{P}_{\tilde{N}-1}(I), \quad \psi_\lambda \in \bigcup_{j \geq j_0} \Psi_j. \quad (20)$$

#### 4 Refinement matrices

From the definition of spaces  $S_j$  and  $W_j$  it follows that there exist *refinement matrices*  $\mathbf{M}_{j,0}$ ,  $\tilde{\mathbf{M}}_{j,0}$ ,  $\mathbf{M}_{j,1}$ , and  $\tilde{\mathbf{M}}_{j,1}$  such that

$$\Phi_j = \mathbf{M}_{j,0}^T \Phi_{j+1}, \quad \tilde{\Phi}_j = \tilde{\mathbf{M}}_{j,0}^T \tilde{\Phi}_{j+1}, \quad \Psi_j = \mathbf{M}_{j,1}^T \Phi_{j+1}, \quad \tilde{\Psi}_j = \tilde{\mathbf{M}}_{j,1}^T \tilde{\Phi}_{j+1}. \quad (21)$$

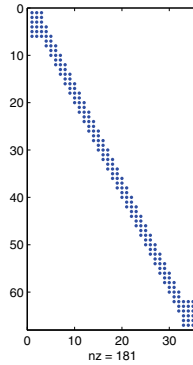
The *discrete wavelet transform* consists of applying  $\mathbf{M}_j^T = (\mathbf{M}_{j,0}^T \mathbf{M}_{j,1}^T)^T$ .

The structure of the matrix  $\mathbf{M}_{j,0}$  for wavelet bases from [3] is shown in Figure 1. Matrices  $\tilde{\mathbf{M}}_{j,0}$ ,  $\mathbf{M}_{j,1}$ , and  $\tilde{\mathbf{M}}_{j,1}$  have similar structures. The inner part is formed by coefficients corresponding to the inner functions. They are the same as filter coefficients for wavelet bases on the real line. For some constructions of wavelet bases, the condition number of a refinement matrix is large. It can cause problems in applications. We can improve the condition number of the refinement matrix by the method from [10]. It is based on the following Theorem.

**Theorem 4.1** *Let  $\mathbf{M}_j, \mathbf{M}_j^T = (\mathbf{M}_{j,0}^T, \mathbf{M}_{j,1}^T)$ , be a refinement matrix for a wavelet basis  $\Psi$ , then there exists a wavelet basis  $\Psi^a$  such that the matrix  $\mathbf{M}_j^a$  of the form  $(\mathbf{M}_j^a)^T = (\mathbf{M}_{j,0}^T, a \mathbf{M}_{j,1}^T)$  is a refinement matrix for a wavelet basis  $\Psi^a$ .*

The proof can be found in [10]. In the following, we propose a method for an improvement of the condition number of the refinement matrix and the size of boundary filters.

Figure 1: The structure of the matrix  $\mathbf{M}_{j,0}$



**Theorem 4.2** Let  $\mathbf{M}_j$  of the form  $\mathbf{M}_j^T = (\mathbf{M}_{j,0}^T, \mathbf{M}_{j,1}^T)$  be a refinement matrix for a wavelet basis  $\Psi$ , then there exists a wavelet basis  $\Psi^a$  such that the matrix

$$\mathbf{M}_j^a = \begin{cases} a m_{k,l}^j & \text{if } l = p, k \in \mathcal{I}_{j+1} \setminus \{p\}, \\ m_{k,l}^j/a & \text{if } l \in \mathcal{I}_{j+1} \setminus \{p\}, k = p, \\ m_{k,l}^j & \text{otherwise,} \end{cases} \quad (22)$$

where  $p$  is an index of some column of the matrix  $\mathbf{M}_{j,0}$  and  $m_{k,l}^j$  are entries of the matrix  $\mathbf{M}_j$ , is a refinement matrix for a wavelet basis  $\Psi^a$ .

**Proof.** As mentioned above a wavelet basis  $\Psi$  has a structure

$$\Psi = \Phi_{j_0} \cup \bigcup_{j \geq j_0} \Psi_j. \quad (23)$$

Let us define

$$\Psi^a = \Phi_{j_0}^a \cup \bigcup_{j \geq j_0} \Psi_j, \quad (24)$$

for  $\Phi_j^a$  containing functions

$$\phi_{j,k}^a = \begin{cases} a\phi_{j,k} & \text{if } k = p \\ \phi_{j,k} & \text{otherwise,} \end{cases} \quad (25)$$

where  $\phi_{j,k}$  are functions in  $\Phi_j$ . Thus, scaling functions and wavelets in  $\Psi^a$  are the same as scaling functions and wavelets in  $\Psi$  up to functions  $\phi_{j,p}^a$  which equals  $a \phi_{j,p}$ .

Due to Lemma 2.1 in [6] the Riesz basis property (11) is a consequence of the relation

$$\|\phi_{j,k}\|_I \leq C, \quad \|\tilde{\phi}_{j,k}\|_I \leq C, \quad j \geq j_0, \quad (26)$$

where  $C$  is a constant independent of  $j$ , the polynomial exactness of primal scaling functions and the smoothness of dual basis functions. It is easy to see that if these properties hold for  $\Psi$ , then they hold for  $\Psi^a$  as well. The supports of all basis function remain unchanged, hence functions in  $\Phi^a$  are local. Therefore, a set  $\Psi^a$  is indeed a Riesz basis of the space  $L^2(I)$ .

Let us now study the structure of the refinement matrix  $\mathbf{M}_j^a$  corresponding to  $\Psi^a$ . We denote by  $m_{k,l}^{j,0}$  and  $m_{k,l}^{j,1}$  the entries of matrices  $\mathbf{M}_{j,0}$  and  $\mathbf{M}_{j,1}$ , respectively. For  $k \neq p$  we have

$$\phi_{j,k}^a = \phi_{j,k} = \sum_{l \in \mathcal{I}_{j+1}} m_{k,l}^{j,0} \phi_{j+1,l} = \sum_{l \in \mathcal{I}_{j+1} \setminus \{p\}} m_{k,l}^{j,0} \phi_{j+1,l}^a + \frac{m_{k,p}^{j,0}}{a} \phi_{j+1,p}^a, \quad (27)$$

and

$$\psi_{j,k}^a = \psi_{j,k} = \sum_{l \in \mathcal{I}_{j+1}} m_{k,l}^{j,1} \phi_{j+1,l} = \sum_{l \in \mathcal{I}_{j+1} \setminus \{p\}} m_{k,l}^{j,1} \phi_{j+1,l}^a + \frac{m_{k,p}^{j,1}}{a} \phi_{j+1,p}^a. \quad (28)$$

For  $k = p$  we obtain

$$\phi_{j,p}^a = \sum_{l \in \mathcal{I}_{j+1}} a m_{p,l}^{j,0} \phi_{j+1,l} = \sum_{l \in \mathcal{I}_{j+1} \setminus \{p\}} a m_{p,l}^{j,0} \phi_{j+1,l} + m_{p,p}^{j,0} \phi_{j+1,p}^a. \quad (29)$$

It follows that the refinement matrix for  $\Psi^a$  is the matrix defined by (22).

**Theorem 4.3** *Let  $\mathbf{M}_j$  of the form  $\mathbf{M}_j^T = (\mathbf{M}_{j,0}^T, \mathbf{M}_{j,1}^T)$  be a refinement matrix for a wavelet basis  $\Psi$ , then there exists a wavelet basis  $\Psi^a$  such that the matrix  $\mathbf{M}_j^a$  of the form  $(\mathbf{M}_j^a)^T = (\mathbf{M}_{j,0}^T, (\mathbf{M}_{j,1}^a)^T)$ ,*

$$\mathbf{M}_{j,1}^a = \begin{cases} a m_{k,l}^{j,1} & \text{if } l = p, k \in \mathcal{I}_{j+1}, \\ m_{k,l}^{j,1} & \text{otherwise,} \end{cases}$$

where  $p$  is an index of some column of the matrix  $\mathbf{M}_{j,1}$  and  $m_{k,l}^j$  are entries of the matrix  $\mathbf{M}_j$ , is a refinement matrix for a wavelet basis  $\Psi^a$ .

**Proof.** A wavelet basis  $\Psi$  has a structure

$$\Psi = \Phi_{j_0} \cup \bigcup_{j \geq j_0} \Psi_j. \quad (30)$$

Let us define

$$\Psi^a = \Phi_{j_0} \cup \bigcup_{j \geq j_0} \Psi_j^a, \quad (31)$$

for  $\Psi_j^a$  containing functions

$$\psi_{j,k}^a = \begin{cases} a \psi_{j,k} & \text{if } k = p \\ \psi_{j,k} & \text{otherwise,} \end{cases} \quad (32)$$

where  $\psi_{j,k}$  are functions in  $\Psi_j$ . By the same argument as in the proof of Theorem 4.2,  $\Psi_a$  is a Riesz basis for the space  $L^2(I)$ . Since we have

$$\psi_{j,k}^a = \sum_{l \in \mathcal{I}_{j+1}} a m_{k,l}^{j,1} \phi_{j+1,l} \quad (33)$$

and other refinements relations remain unchanged, the refinement matrix for  $\Psi^a$  is the matrix defined by (4.3).

By a convenient choice of the constants for all boundary filters, we can influence the size of boundary filter coefficients as well as the condition number of the refinement matrices.

## 5 Conclusion

We propose a method for an improvement of the discrete wavelet transform with boundary filters. The method can be applied to general wavelet bases on the interval, e.g. the wavelet bases from [1, 2, 3, 6, 9]. Our future work is using these principles for a construction of boundary filters for concrete wavelets suitable for a given problem, e.g. boundary filters for CDF 9/7 wavelets suitable to an image compression.

## Acknowledgement

The authors have been supported by the project ESF "Constitution and improvement of a team for demanding technical computations on parallel computers at TU Liberec" No. CZ.1.07/2.3.00/09.0155 .

## References

- [1] ANDERSSON, L.; HALL, N.; JAWERTH, B.; PETERS, G. *Wavelets on Closed Subsets of the Real Line*. In: Topics in the Theory and Applications of Wavelets, Academic Press, Boston, 1994, pp. 1–61.
- [2] ČERNÁ, D., FINĚK, V.: *Optimized construction of biorthogonal spline-wavelets*. In: ICNAAM 2008 (Simos T.E. et al., eds.), AIP Conference Proceedings 1048, American Institute of Physics, New York, pp. 134–137, 2008.
- [3] ČERNÁ, D., FINĚK, V. *Construction of Optimally Conditioned Cubic Spline Wavelets on the Interval*. Adv. Comput. Math., vol. 34, 2011, pp. 519–552.
- [4] COHEN, A., DAUBECHIES, I., FEAUVEAU, J. C.: *Biorthogonal bases of compactly supported wavelets*. Comm. Pure and Appl. Math., Vol. 45, pp. 485–560, 1992.
- [5] COHEN, A.; DAUBECHIES, I.; VIAL, P.: *Wavelets on the interval and fast wavelet transforms*, Appl. Comp. Harm. Anal., Vol. 1, pp. 54–81, 1993.
- [6] DAHMEN, W., KUNOTH, A., URBAN, K.: *Biorthogonal spline wavelets on the interval - stability and moment conditions*. Appl. Comp. Harm. Anal., Vol. 6, pp. 132–196, 1999.
- [7] DAHMEN, W., KUNOTH, A., URBAN, K.: *Wavelets in numerical analysis and their quantitative properties*. In: Surface fitting and multiresolution methods (Le Méhauté, A., Rabut, C., Schumaker, L., eds.), Vol. 2, pp. 93–130, 1997.
- [8] ISO/IEC 15444-1, *Information technology-JPEG2000 image coding system, Part 2: Extensions*, 2000.
- [9] PRIMBS, M.: *New stable biorthogonal spline-wavelets on the interval*. Results in Mathematics, Vol. 57, pp. 121–162, 2010.
- [10] TURCAJOVÁ, R.: *Numerical condition of discrete wavelet transforms*. SIAM J. Matrix Anal. Appl., Vol. 18, pp. 981–999, 1997.

## Current address

**Dana Černá, Mgr. Ph.D.**

Department of Mathematics and Didactics of Mathematics, Technical University in Liberec, Studentská 2, Liberec, 46117, Czech Republic, dana.cerna@tul.cz

**Václav Finěk, RNDr. Ph.D.**

Department of Mathematics and Didactics of Mathematics, Technical University in Liberec,  
Studentská 2, Liberec, 46117, Czech Republic, [vaclav.finek@tul.cz](mailto:vaclav.finek@tul.cz)



## RECONSTRUCTION OF THE BOREHOLE WALL USING VIDEO RECORDS

FERDIÁNOVÁ Věra, (CZ), HURTÍK Petr, (CZ), KOLCUN Alexej, (CZ)

**Abstract** The presented paper describes possibilities of video processing of the camera operating in the borehole. As the motion of the camera is hand-controlled without any stabilization elements, resulting sliding motion is affected by strong instability. The purpose of this research is to eliminate the above-mentioned factors to reconstruct the borehole walls.

**Key words and phrases.** image processing, video stabilization, borehole

*Mathematics Subject Classification.* Primary 15A66, 65D17; Secondary 68U99.

### 1 Introduction

The borehole surface monitoring represents one of methods, which enables to assess the mechanical state of massif. It is realized by a probe provided with a camera and a light source. The probe is attached to the end of a conductive, stiff steel wire. The wire is hand-controlled, thus the attached probe is lowered down/drawn up from the borehole. Unfortunately, this activity causes a great instability of the obtained video record. For the purposes of comparison of the state of the borehole surface, before and after the realization of various experiments, it is advisable to keep the image information of the borehole in a standardized form - unfolded covering of the borehole surface. For this purpose it is necessary to stabilize the image. Similar problem is solved e.g. in [3,8], where the images of the video records are stabilized by hardware. The contribution of this paper consists in description of the SW image stabilization method.

### 2 Definition of partial sub-problems

We suppose a simple model of camera – central projection with projection plane  $z = 0$  and focus  $F = (0, 0, -f)$ .

For unrolling the cylindric surface of the borehole into the rectangle, we need to know what is the result of the camera projection considering the system of parallel circles from the borehole, perpendicular to the borehole axis. Due to the fact that the radius  $r$  of camera is smaller than the borehole radius  $R$ ,  $r < R$  we can distinguish:

- *the camera axis movement*; this means the camera axis may be shifted (due to the borehole axis) or the camera axis may oscillate like a pendulum in the plane perpendicular to the camera axis,
- *rotation of the camera*; the camera can rotate freely around its axis, i.e. in the plane orthogonal to the axis.

So as the motion of the camera in the borehole is hand-controlled,

- *videorecord of the borehole surface is irregular*; in many cases the probe may be temporarily immovable or it may even move backwards.

### 3 Image stabilization

The problem of image stabilization may be solved as follows.

#### 3.1 Elimination of the camera axis movement

We can distinguish three cases of the mutual positions of the borehole and the camera axis.

1. Both axes are identical.
2. The axes are non identical but parallel.
3. The axes are skew.

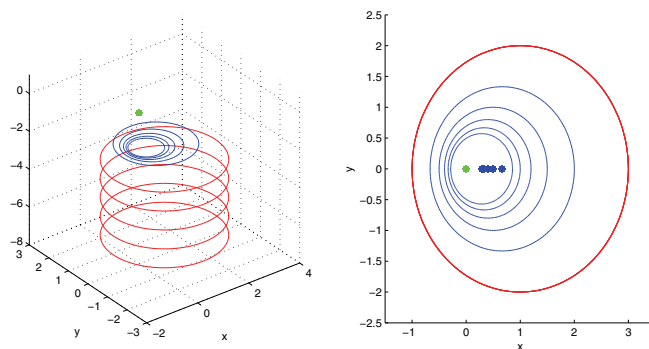
The first case is rather rare; the parallel circles from the borehole, orthogonal to the borehole axis, are projected to the concentric circles. Moreover, this phenomenon can be supposed as a special event of the second case of mutual positions of the axes. So we shall concentrate our attention to parallel and non identical axes of the camera and the borehole.

Let us consider the circles on the borehole surface  $C(S(s_x, s_y, s_z), R)$  with centres  $S = (s_x, s_y, s_z)$ , and radius  $R$  where  $s_z = h$ ,  $h_1 \leq h \leq h_2$ :

$$\begin{aligned}x &= s_x + R \cos \alpha, \\y &= s_y + R \sin \alpha, \\z &= h, \\h_1 &\leq h \leq h_2, \quad 0 \leq \alpha \leq 2\pi.\end{aligned}$$

It is easy to show, that the projection of these circles are the circles

$$C\left(S\left(-\frac{s_x f}{h}, -\frac{s_y f}{h}, 0\right), \frac{Rf}{h}\right).$$



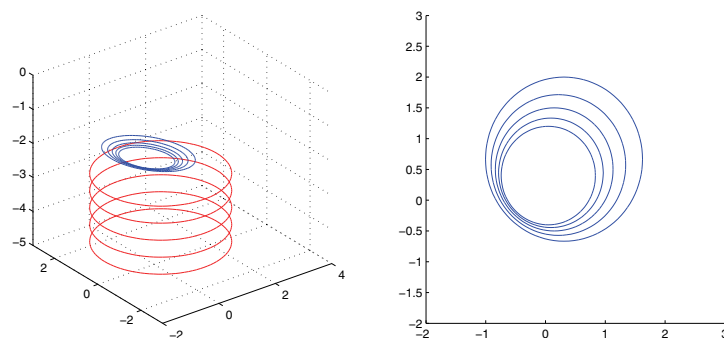
Obrázek 1: Set of borehole circles (left) and their projections (right) – second case of axes.

It means that the parallel borehole circles are projected as a non-centric circles – see Fig. 1.

It can be shown that for skew axes the resulting projections of the borehole circles are non-centric ellipses, Fig. 2.

$$\begin{aligned}x &= -\frac{f}{h}(r \cos \phi \cos \alpha - s_x \cos \phi - h \sin \phi) \\y &= -\frac{f}{h}(r \sin \alpha - s_y) \\z &= -\frac{f}{h}(r \cos \alpha \sin \phi + s_x \sin \phi + h \cos \phi)\end{aligned}$$

However, due to the ratio of  $R - r$  and the length of the probe, the angle of possible skew is too small (less than  $10^\circ$ ) and we can neglect this case of mutual position of axes.

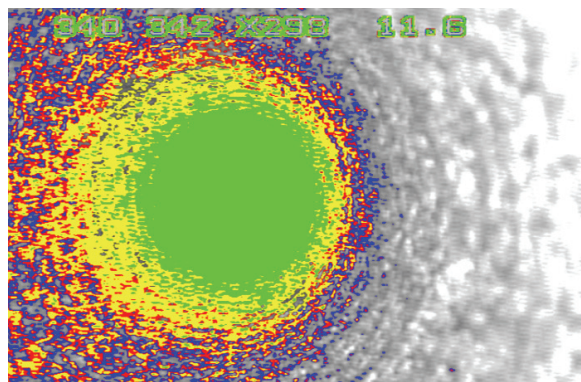


Obrázek 2: Set of borehole circles (left) and their projections (right) – 3rd case of axes.

The rate of the non-centricity is derived from the ideas described below.

1. Intensity of light source decreases with the square of the distance between the source and the surface.
2. Image of the part of the surface which is far away from the camera is blurred. So its color is homogeneous gray.

So the parameters for set of non-centric projections can be obtained using various thresholds and recognizing the circles near the centre of the image which are darker than these thresholds – Fig. 3 (left).



Obrázek 3: Two thresholds in the borehole.

### 3.2 Elimination of camera rotating around its axis

Stabilization of the torsional movement is based on the following steps.

1. Pair of neighbour snaps from videosequence is analysed.
2. For both snaps the area bounded by two non-centric circles is transformed into rectangle (unfolded covering of an image), using linear interpolation along the radial lines – Fig. 3 (right).
3. The maximal similarity in overlap of neighbour coverings  $a$ ,  $b$  is searched.

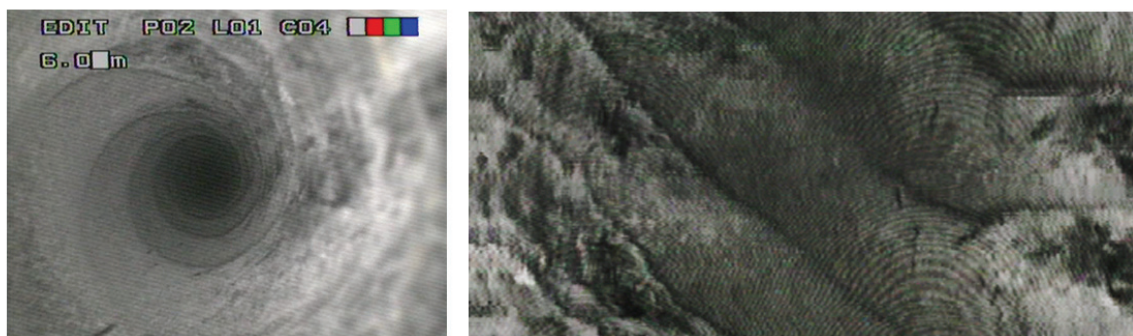
### 3.3 Elimination of the irregularity of the camera movement

This part of the image stabilization is based on similar idea as the method mentioned above – similar parts of the pair of neighbour images are removed from the sequence of video snapshots.

So, the similarity for torsional movement and for irregularity of the camera movement can be analyzed in one step. We search minimal value

$$D(\pi, \tau) = \sum_{j=1}^d \sum_{i=1}^m r(a_{i+\pi, j+\tau}, b_{i,j}) = \min_{0 \leq p \leq h, 0 \leq t \leq d} \sum_{j=1}^d \sum_{i=1}^m r(a_{i+p, j+t}, b_{i,j}).$$

where  $r(x, y) = (x-y)^2$  represents the measure of the difference of scalar values. The overlapping area is characterized with value  $\pi$ , value  $\tau$  represents the angle of the camera rotation between analysed pair of snapshots.



Obrázek 4: Final image after processing

## 4 Practical results

Fig.4 shows one of the videosequence snaps (left) and resulting unfolded covering (right).

## 5 Conclusions

We have presented stabilization method which is used for the unfolded covering procedure from the videosequence of snapshots of the camera probe. The obtained results show that for our configuration of the ratio of the borehole and probe radii, and the length of the probe we can neglect the pendulum-like movement of the probe.

Following research will be focused on more complex models, which involve a greater variability of the probe dimension with respect to the radius of the borehole, as well as more exact model of light (e.g. a ring model instead of the spotlight model).

## 6 Acknowledgements

Research is supported by the projects:

- SGS-2012 projects at University of Ostrava: Geometric mechanics and optimalization, Image processing with artefacts detection by soft computing methods,
- VG20102014034 – Grant of Ministry of the Interior of the Czech Republic.

## Reference

- [1] DOBEŠ, M.: *Zpracování obrazu a algoritmy v C*. 1. vyd. Praha : BEN-Technická literatura, 2008. 143 s. ISBN 978-80-7300-2.
- [2] HURTÍK, P., *Stabilizace obrazu kamery ovlivněné torzním pohybem*, Diplomová práce. Ostravská univerzita v Ostravě, 2011.
- [3] KANNALA, J; BRANDT, SAMI S.: *Measuring the Shape of Sewer Pipes from Video*. Proc. of Conference on Machine Vision Applications. Japan : Tsukuba Science City, 2005. s. 237-240.

- [4] KARKANIS, S. A., et al.: *Detection of lesions in endoscopic video using textural descriptors on wavelet domain supported by artificial*. TEI of Lamia [online]. [cit. 2011- 04-18]. Dostupný z WWW: <http://www.inf.teilam.gr/OLD/staff/KARKANIS/4815.pdf>.
- [5] KITTLER, J.; ILLINGWORTH, J.: *Minimum error thresholding*, Pattern Recognition, vol. 19, pp. 41- 47, 1986
- [6] RUŽICKÝ, E.; FERKO A.: *Počítačová grafika a zpracování obrazu*. Sapientia, 1995.
- [7] Otsu's method. In Wikipedia : the free encyclopedia [online]. St. Petersburg (Florida) : Wikipedia Foundation, 25 May 2005, last modified on 14 March 2011 [cit. 2011-04-16]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Otsu>
- [8] ZHANG, Y.; HARTLEY, R.; MASHFORD, J.; WANG, L.; BURN S: *Pipeline Reconstruction from Fisheye Images*. In Journal of WSCG: Volume 19 (Václav Skala ed.), Number 1-3, 2011. Plzeň : Union Agency, 2011. s. 49-57. ISSN 1213-6972.
- [9] ŽÁRA, J. a kol.: *Moderní počítačová grafika*. Nakladatelství Computer Press, Brno 2004. ISBN 80-251-0454-0.

#### **Current address**

##### **Věra Ferdiánová, Mgr.**

University of Ostrava, Faculty of Science, Department of Mathematics  
30. dubna 22, 702 00 OSTRAVA, CZECH REPUBLIC  
E-mail: Vera.Ferdianova@osu.cz, tel.: +420 597 092142

##### **Petr Hurtík, Mgr.**

University of Ostrava, Faculty of Science, Department of Mathematics  
30. dubna 22, 702 00 OSTRAVA, CZECH REPUBLIC  
E-mail: Petr.Hurtik@osu.cz, tel.: +420 597 092134

##### **Alexej Kolcun, Mgr. Csc.**

University of Ostrava, Faculty of Science, Department of Informatics and Computers  
30. dubna 22, 702 00 OSTRAVA, CZECH REPUBLIC  
E-mail: Alexej.Kolcun@osu.cz, tel.: +420 597 092176  
Institute of Geonics AS CR, v.v.i., Department of Applied Mathematics and Computer Science  
Studentska 1768, 708 00 OSTRAVA, CZECH REPUBLIC  
E-mail: Alexej.Kolcun@ugn.cas.cz, tel.: +420 596 979216

## DISCONTINUOUS GALERKIN METHOD FOR THE NUMERICAL SOLUTION OF OPTION PRICING

HOZMAN Jiří, (CZ)

**Abstract.** The paper is devoted to the use of the discontinuous Galerkin (DG) method for standard option pricing models. As a model application, we consider one-dimensional Black-Scholes partial differential equation for the pricing of European plain vanilla options. The investigated system represents a scalar nonstationary linear convection-diffusion-reaction equation. For this case, we start with a variational formulation, then a DG space semi-discretization is combined with first-order time discretization by backward Euler method. Consequently, the fully discrete scheme applied to the numerical solution of a preliminary test examples is presented.

**Key words and phrases.** Black-Scholes model, discontinuous Galerkin method, convection-diffusion-reaction equation, space semidiscretization, backward Euler method.

*Mathematics Subject Classification.* 91G80, 65M60, 65M12, 65L06.

### 1 Introduction

In this article, we are concerned with the development of sufficiently robust, accurate and efficient numerical method for the solution of option pricing models. Our model convection-diffusion-reaction equation arise in mathematical finance from the well-known *Black-Scholes equation*, see [8], [11]. A brief review of notation and terminology of financial options together with fundamentals of classical market models is introduced in Section 2.

The *discontinuous Galerkin* (DG) method seems to be a promising technique for the solution of such problems. DG space semidiscretization uses higher order piecewise polynomial discontinuous approximation on arbitrary meshes, without any requirement on interelement continuity, for a survey, see [2], [3], [10]. Among several variants of DG methods we prefer the so-called *interior penalty Galerkin* (IPG) discretizations. We deal with three variants of IPG, namely *nonsymmetric* (NIPG), *symmetric* (SIPG) and *incomplete* interior penalty Galerkin (IIPG) techniques, see [1].

The discretization in time coordinate is performed with the aid of the *backward Euler method*, sidetracking the time step restriction well-known from the explicit schemes. Consequently, the fully discrete problem is represented by the system of algebraic equations, see Section 4.

Within this paper we present the derivation of the whole scheme, from a continuous problem to the discrete one, and append the preliminary numerical experiments.

## 2 Financial background

In order to better understand the whole model problem comprehensively, it is necessary to start with a brief review of modelling tools for financial options. Financial *derivatives* are instruments to assist and regulate agreements on transactions of the future. They represent financial contracts whose values are based on the value of an *underlying asset*, e.g. stock or a parcel of shares of a company. Further, financial *option* is a special case of derivative, it is a contract between two parties about trading the asset at a certain future time. The basic option scheme can be written in the simple schedule

$$\mathbf{writer} \xrightarrow{\text{sells}} \begin{matrix} \mathbf{premium} \\ \text{(option value)} \end{matrix} \xrightarrow{\text{purchases}} \mathbf{holder},$$

where writer (e.g. bank) fixes the terms of the option contract and sells the option. On the other hand, holder purchases the option for market price, which is called a premium. Each option has a limited life time, given by expiration date  $T$  fixing the time horizon.

There are two basic types of options: The *call* option gives the holder the right (no obligation) to buy the underlying for an agreed price  $K$  by the date  $T$ . On the other hand, the *put* option gives the holder the right to sell the underlying for an agreed price  $K$  by the date  $T$ . The previously agreed price  $K$  is called a strike price.

From another point of view, the options can be divided into standard and non-standard (exotic) types. In this contribution, we focus only on standard options represented by European plain vanilla option. For European option is typical that its exercise is only permitted at expiration  $T$ .

In our further considerations, let underlying asset be stock price  $S = S(\tau)$  depending on actual time  $\tau$ . The symbol  $V$  stands for the value of certain type of option, i.e. the value  $V = V(S, t)$  is driven according to the stock price, time to expiration  $t = T - \tau$  (reversal actual time) and model of classical market.

For modelling financial options, it is widely used the famous Black-Scholes equation, derived under series of assumptions, for more details see [11]. The price  $V(S, t)$  of option satisfies the following partial differential equation

$$-\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} = rV \quad (1)$$

where  $t$ ,  $0 \leq t \leq T$  is time to expiration, i.e.  $T - t$  is a current time,  $S = S(t)$  price of stock at time  $t$ , i.e.  $0 \leq S < +\infty$ . Market parameters affecting the price are the risk-free interest rate  $r$  and volatility  $\sigma$  of the price  $S$ . In real markets, values  $r$  and  $\sigma$  vary with time, but to keep the model and analysis simple, we assume  $r$  and  $\sigma$  to be constant.

To correctly define the initial-boundary value problem (1), it is necessary to equip equation (1) with initial condition and set of two boundary conditions valid in the endpoints of underlying asset. The initial condition (in reversal time) arises from the terminal condition (in continuous time) which is given by *payoff* function

$$V(S, 0) = V^0(S) := \begin{cases} \max(S - K, 0), & \text{if } V \text{ is a call,} \\ \max(K - S, 0), & \text{if } V \text{ is a put,} \end{cases} \quad (2)$$

where  $K$  is the strike price. From [11], the both types of European options have the following asymptotic behaviour corresponding to the put-call parity of options, i.e.

$$V^{BC}(t) = \begin{cases} 0, & \text{for } S = 0 \\ S - Ke^{-rt}, & \text{for } S \rightarrow +\infty \end{cases} \quad (\text{call}), \quad (3)$$

$$V^{BC}(t) = \begin{cases} 0, & \text{for } S \rightarrow +\infty \\ Ke^{-rt} - S, & \text{for } S \approx 0 \end{cases} \quad (\text{put}). \quad (4)$$

Let us note that in practical considerations, relation  $S \rightarrow +\infty$  is replaced by sufficiently large  $S_{max} > 0$  (maximal possible stock price).

### 3 Continuous problem

In what follows, we focus only on the European call option. According to (1)-(3), we consider the following *unsteady linear convection-diffusion-reaction* problem: Let  $\Omega \equiv (0, S_{max})$  be a bounded open interval and  $T > 0$ . We seek a function  $u : Q_T = \Omega \times (0, T) \rightarrow \mathbb{R}$  such that

$$\begin{aligned} (a) \quad & \frac{\partial u}{\partial t} - \frac{1}{2}\sigma^2 x^2 \frac{\partial^2 u}{\partial x^2} - rx \frac{\partial u}{\partial x} + ru = 0 \quad \text{in } Q_T, \\ (b) \quad & u(0, t) = u_D^L(t) = 0 \quad \text{and} \quad u(S_{max}, t) = u_D^U(t) = S_{max} - Ke^{-rt}, \\ (c) \quad & u(x, 0) = u^0(x) = V^0(x), \quad x \in \Omega, \end{aligned} \quad (5)$$

where  $\sigma$  and  $r$  are positive constants,  $u_D^L, u_D^U : (0, T) \rightarrow \mathbb{R}$  are Dirichlet boundary conditions and  $u^0 : \Omega \rightarrow \mathbb{R}$  is the initial condition.

The convection-diffusion-reaction equation (5a) is equipped with the initial condition (5c) given by (2) and the set of two Dirichlet boundary conditions (5b) prescribed at the endpoints of interval  $(0, S_{max})$ .

Further, we shall introduce standard notation for function spaces and their norms  $\|\cdot\|$  and seminorms  $|\cdot|$ . Let  $k \geq 0$  be a integer and  $p \in [1, \infty]$ . We use the well-known Lebesgue and Sobolev spaces  $L^p(\Omega)$ ,  $H^k(\Omega)$ , Bochner spaces  $L^p(0, T; X)$  of functions defined in  $(0, T)$  with values in Banach space  $X$  and the spaces  $C^k([0, T]; X)$  of  $k$ -times continuously differentiable mappings of the interval  $[0, T]$  with values in  $X$ .

In order to obtain a variational formulation of (5) we use a concept of *weighted Sobolev spaces*, for survey see, e.g. [9]. Let us introduce the space

$$V = V(\Omega) := \{v \in L^2(\Omega) : x \cdot v'(x) \in L^2(\Omega)\} \quad (6)$$

with a scalar product

$$(u, v)_V = (u, v) + \int_{\Omega} x u'(x) \cdot x v'(x) dx \quad u, v \in V, \quad (7)$$

where  $(\cdot, \cdot)$  denotes the scalar product of  $L^2(\Omega)$ . Consequently, space  $V$  becomes a Hilbert space with norm  $\|\cdot\|_V := (\cdot, \cdot)_V^{1/2}$ . In order to fulfil the boundary conditions it is appropriate to define the space

$$V_0 := \overline{C_0^\infty(\Omega)}^{\|\cdot\|_V} \equiv \{v \in V : v(0) = v(S_{max}) = 0\} \quad (8)$$

with seminorm  $|\cdot|_V := \|x \frac{d}{dx} \cdot\|_{\Omega}^2$  as a norm on  $V_0$ . Moreover,  $H^1(\Omega) \subset V_0$ .

Further, we introduce the following bilinear forms  $\hat{a}(\cdot, \cdot)$  and  $\hat{b}(\cdot, \cdot)$  representing the diffusion and convection terms, respectively.

$$\hat{a}(u, v) = \frac{1}{2} \sigma^2 \int_{\Omega} x^2 u'(x) v'(x) dx, \quad (9)$$

$$\hat{b}(u, v) = (\sigma^2 - r) \int_{\Omega} x u'(x) v(x) dx \quad (10)$$

In order to simplify the notation, we set a new bilinear form

$$\mathcal{A}(u, v) = \hat{a}(u, v) + \hat{b}(u, v) + r(u, v), \quad (11)$$

then we are ready to define a weak solution  $u$  of the problem (5).

**Definition 3.1** *We say that  $u$  is a weak solution of problem (5), if the following conditions are satisfied*

- (a)  $u \in L^2(0, T; V_0), \quad u \in L^\infty(Q_T)$
- (b)  $\frac{d}{dt}(u(t), v) + \mathcal{A}(u(t), v) = 0 \quad \forall v \in V_0 \text{ in sense of distributions on } (0, T),$  (12)
- (c)  $u(0) = u^0 \text{ in } \Omega, \quad u^0 \in L^2(\Omega).$

By  $u(t)$  we denote the function on  $\Omega$  such that  $u(t)(x)$ ,  $x \in \Omega$ .

It can be easily checked that bilinear form  $\mathcal{A}(\cdot, \cdot)$  is continuous and coercive on  $V_0$ , i.e.

$$\exists \gamma > 0 : |\mathcal{A}(u, v)| \leq \gamma |u|_V |v|_V \quad \forall u, v \in V_0 \quad (13)$$

$$\exists \alpha > 0, \lambda \in \mathbb{R} : |\mathcal{A}(v, v)| + \lambda \|v\|_{\Omega}^2 \geq \alpha |u|_V^2 \quad \forall v \in V_0 \quad (14)$$

and hence problem (5) has a unique solution  $u$ , for more details see [8].

## 4 Discretization

Let  $\mathcal{T}_h$  ( $h > 0$ ) be a family of the partitions of the closure  $\overline{\Omega} = [0, S_{max}]$  of the domain  $\Omega$  into  $N$  closed mutually disjoint subintervals  $I_k = [x_{k-1}, x_k]$  with length  $h_k := x_k - x_{k-1}$  and the symbol  $\mathcal{J}$  stands for an index set  $\{1, \dots, N\}$ . Then we call  $\mathcal{T}_h = \{I_k, k \in \mathcal{J}\}$  a *triangulation*

with spatial step  $h := \max_{k \in \mathcal{J}}(h_k)$  and interval  $I_k$  an *element*. By  $\mathcal{E}_h$  we denote the smallest possible set of all endpoints of all subintervals  $I_k$ , i.e.  $\mathcal{E}_h = \{x_0 = 0, x_1, \dots, x_{N-1}, x_N = S_{max}\}$ . Further, we label by  $\mathcal{E}_h^I$  the set of all inner nodes. Obviously,  $\mathcal{E}_h = \mathcal{E}_h^I \cup \{0, S_{max}\}$ .

DG method allows to treat with different polynomial degrees over elements. Therefore, we assign a *local Sobolev index*  $s_k \in \mathbb{N}$  and *local polynomial degree*  $p_k \in \mathbb{N}$  to each  $I_k \in \mathcal{T}_h$ . Then we set the vectors

$$\mathbf{s} \equiv \{s_k, I_k \in \mathcal{T}_h\}, \quad \mathbf{p} \equiv \{p_k, I_k \in \mathcal{T}_h\}. \quad (15)$$

Over the triangulation  $\mathcal{T}_h$  we define the so-called *broken Sobolev space* corresponding to the vector  $\mathbf{s}$

$$H^{\mathbf{s}}(\Omega, \mathcal{T}_h) \equiv \{v; v|_{I_k} \in H^{s_k}(I_k) \forall I_k \in \mathcal{T}_h\} \quad (16)$$

with the seminorm  $|v|_{H^{\mathbf{s}}(\Omega, \mathcal{T}_h)} \equiv \left( \sum_{I_k \in \mathcal{T}_h} |v|_{H^{s_k}(I_k)}^2 \right)^{1/2}$ , where  $|\cdot|_{H^{s_k}(I_k)}$  denotes the standard seminorm on the Sobolev space  $H^{s_k}(I_k)$ ,  $I_k \in \mathcal{T}_h$ . Moreover, we introduce *broken weighted Sobolev space*

$$H^{\mathbf{w}}(\Omega, \mathcal{T}_h) \equiv \{v; v|_{I_k} \in V(I_k) \forall I_k \in \mathcal{T}_h\} \quad (17)$$

with the seminorm  $|v|_{H^{\mathbf{w}}(\Omega, \mathcal{T}_h)} \equiv \left( \sum_{I_k \in \mathcal{T}_h} |v|_{V(I_k)}^2 \right)^{1/2}$ .

Finally, the approximate solution is sought in a space of discontinuous piecewise polynomial functions associated with the vector  $\mathbf{p}$  by

$$S_{h\mathbf{p}} \equiv S_{h\mathbf{p}}(\Omega, \mathcal{T}_h) \equiv \{v; v \in L^2(\Omega), v|_{I_k} \in P_{p_k}(I_k) \forall I_k \in \mathcal{T}_h\}, \quad (18)$$

where  $P_{p_k}(I_k)$  denotes the space of all polynomials of degree  $\leq p_k$  on  $I_k$ ,  $I_k \in \mathcal{T}_h$ . Obviously,

$$S_{h\mathbf{p}}([0, S_{max}], \mathcal{T}_h) \subset H^1([0, S_{max}], \mathcal{T}_h) \subset H^{\mathbf{w}}([0, S_{max}], \mathcal{T}_h). \quad (19)$$

For each  $x \in \mathcal{E}_h^I$  there exist two elements  $I_k, I_{k+1} \in \mathcal{T}_h$  such that  $I_k \cap I_{k+1} = \{x\}$ . Let us denote

$$v(x^+) = \lim_{\varepsilon \rightarrow 0+} v(x + \varepsilon) \quad \text{and} \quad v(x^-) = \lim_{\varepsilon \rightarrow 0+} v(x - \varepsilon) \quad (20)$$

the *traces* of  $v$  at inner points of  $\Omega$ . Moreover,

$$[v(x)] = v(x^-) - v(x^+), \quad \langle v(x) \rangle = \frac{1}{2} (v(x^-) + v(x^+)), \quad (21)$$

denote the *jump* and *mean value* of function  $v$  at points  $x \in \mathcal{E}_h^I$ , respectively. By convention, we also extend the definition of jump and mean value for endpoints of domain  $\Omega$ , i.e.

$$[v(x_0)] = -v(x_0^+), \quad \langle v(x_0) \rangle = v(x_0^+), \quad [v(x_N)] = v(x_N^-), \quad \langle v(x_N) \rangle = v(x_N^-) \quad (22)$$

#### 4.1 Space semidiscretization

Firstly, we introduce the *semi-discrete problem*, which is obtained with the aid of the *method of lines*, i.e. the semi-discrete problem is discretized only in space coordinates and time is treated

continuously. We recall the space semi-discrete DG scheme presented in [4] and [6]. To this end we introduce the following bilinear/linear forms

$$a_h^\Theta(u, v) = \sum_{k \in \mathcal{J}} \int_{I_k} \frac{1}{2} \sigma^2 x^2 \cdot \frac{\partial u(x, t)}{\partial x} \cdot v'(x) dx - \sum_{x \in \mathcal{E}_h} \left\langle \frac{1}{2} \sigma^2 x^2 \cdot \frac{\partial u(x, t)}{\partial x} \right\rangle [v(x)] \quad (23)$$

$$+ \Theta \sum_{x \in \mathcal{E}_h} \left\langle \frac{1}{2} \sigma^2 x^2 \cdot v'(x) \right\rangle [u(x, t)],$$

$$b_h(u, v) = - \sum_{k \in \mathcal{J}} \int_{I_k} (\sigma^2 - r) x \cdot u(x, t) \cdot v'(x) dx + \sum_{x \in \mathcal{E}_h^I} H(u(x^-, t), u(x^+, t)) [v(x)] \quad (24)$$

$$- H(u_D^L(t), u(x^+)) \cdot v(x_0^+) + H(u(x^-), u_D^U(t)) \cdot v(x_N^-)$$

$$J_h^\omega(u, v) = \sum_{x \in \mathcal{E}_h^I} \omega(x) [u(x, t)] [v(x)] + \omega(x_0) \cdot u(x_0^+, t) \cdot v(x_0^+) + \omega(x_N) \cdot u(x_N^-, t) \cdot v(x_N^-) \quad (25)$$

$$l_h^\Theta(v)(t) = -\Theta \frac{1}{2} \sigma^2 x_0^2 \cdot v'(x_0^+) \cdot u_D^L(t) + \Theta \frac{1}{2} \sigma^2 x_N^2 \cdot v'(x_N^-) \cdot u_D^U(t) \quad (26)$$

$$+ \alpha \omega(x_0) \cdot u_D^L(t) \cdot v(x_0^+) + \alpha \omega(x_N) \cdot u_D^U(t) \cdot v(x_N^-).$$

The crucial item of the DG formulation is the treatment of the linear convection and diffusion terms.

For the convection form  $b_h$  we treat boundary terms similarly as in the finite volume method, i.e. they are approximated with the aid of the following *numerical flux*  $H(\cdot, \cdot)$  through node  $x \in \mathcal{E}_h$  in the positive direction (i.e. outer normal is equal to one):

$$H(u(x^-), u(x^+)) = \begin{cases} (\sigma^2 - r)x \cdot u(x^-), & \text{if } A > 0 \\ (\sigma^2 - r)x \cdot u(x^+), & \text{if } A \leq 0 \end{cases}, \quad \text{where } A = (\sigma^2 - r)x \quad (27)$$

which is based on the concept of *upwinding*, see [7]. The choice of  $u(x^-), u(x^+)$  for boundary points  $\{0, S_{max}\}$  is necessary to specify. Here we use:

$$u(x_0^-) = u(0^-) = u_D^L \quad \text{and} \quad u(x_N^+) = u(S_{max}^+) = u_D^U. \quad (28)$$

The diffusion form  $a_h^\Theta$  includes *stabilization* terms which are artificially added to the formulation of the semi-discrete problem in order to guarantee the stability of the resulting numerical scheme. According to value of parameter  $\Theta$ , we speak of *symmetric* (SIPG,  $\Theta = -1$ ), *incomplete* (IIPG,  $\Theta = 0$ ) or *nonsymmetric* (NIPG,  $\Theta = 1$ ) variants of stabilization of DG method, i.e., we generally consider three variants of the diffusion form  $a_h^\Theta$  and right-hand side form  $l_h^\Theta$ , arisen from Dirichlet boundary conditions.

Furthermore, in order to replace the inter-element discontinuities, the semi-discrete scheme is completed with *penalty* vanishing for the continuous solution. Penalty terms are represented by  $J_h^\omega$  and the penalty parameter function  $\omega : \mathcal{E}_h \rightarrow \mathbb{R}$  in (25) and (26) is defined in spirit of [5] as

$$\omega(x) = \frac{C_W}{d(x)} \quad \text{with} \quad d(x) = \begin{cases} h_1/p_1^2 & , \quad x = x_0 = 0, \\ \min(h_k/p_k^2, h_{k+1}/p_{k+1}^2) & , \quad x \in \mathcal{E}_h^I \wedge \{x\} = I_k \cap I_{k+1}, \\ h_N/p_N^2 & , \quad x = x_N = S_{max}, \end{cases} \quad (29)$$

where  $C_W > 0$  is a suitable constant depending on the used variant of scheme and on the degree of polynomial approximation generally.

In what follows we define the form

$$\mathcal{B}_h^\Theta(u, v) := a_h^\Theta(u, v) + b_h(u, v) + \alpha J_h^\omega(u, v) + (2r - \sigma^2)(u, v) \quad u, v \in H^s(\Omega, \mathcal{T}_h), \quad (30)$$

where the forms  $a_h^\Theta(\cdot, \cdot)$ ,  $b_h(\cdot, \cdot)$  and  $J_h^\omega(\cdot, \cdot)$  are given by (23), (23) and (26), respectively. The value of multiplicative constant  $\alpha$  before the penalty form  $J_h^\omega$  and in the right-hand side form  $l_h^\Theta$  depends on the properties of diffusion term  $-\frac{1}{2}\sigma^2 x^2 \frac{\partial^2 u}{\partial x^2}$ , for survey, see [5].

Now we are ready to introduce the whole concept of semi-discrete solution  $u_h$  of problem (5).

**Definition 4.1** Let  $u_h^0 \in S_{h\mathbf{p}}$  be the  $L^2(\Omega)$ -projection of the initial condition  $u^0$  into  $S_{h\mathbf{p}}$ , i.e. a function defined by

$$(u_h^0 - u^0, v_h) = 0 \quad \forall v_h \in S_{h\mathbf{p}}. \quad (31)$$

We say that  $u_h$  is a semi-discrete solution of problem (5), if the following conditions are satisfied

$$\begin{aligned} (a) \quad & u_h \in C^1([0, T]; S_{h\mathbf{p}}), \\ (b) \quad & \left( \frac{\partial u_h(t)}{\partial t}, v_h \right) + \mathcal{B}_h^\Theta(u_h(t), v_h) = l_h^\Theta(v_h)(t) \quad \forall v_h \in S_{h\mathbf{p}}, \quad \forall t \in (0, T), \\ (c) \quad & u_h(0) = u_h^0. \end{aligned} \quad (32)$$

In fact, the identity (32b) corresponds to the partial differential equation with self-adjoint operators in diffusive and convective terms, i.e.

$$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x} \left( K(x) \frac{\partial u}{\partial x} \right) + \frac{\partial}{\partial x} f(u) + \beta u = 0 \quad \text{in } Q_T \quad (33)$$

with  $K(x) = \frac{1}{2}\sigma^2 x^2$ ,  $f(u) = (\sigma^2 - r)xu$  and  $\beta = 2r - \sigma^2$ . Obviously, the equation (33) is equivalent with the original Black-Scholes equation (5b).

The problem (32) exhibits a system of ordinary differential equations (ODE) equipped with initial condition for unknown  $u_h(t)$  which has to be discretized in time by a suitable method.

## 4.2 Fully time-space discretization

There exists a wide range of approaches for time discretization of ODE systems resulting from DG semidiscretization. In practical computations, the simplest time discretization is via *explicit scheme* (e.g. Euler forward scheme and Runge-Kutta methods). Since the problem (32) belongs to the class of stiff problems, the explicit schemes suffer from a strong limitation on the time step. In order to avoid the strong time step restriction of explicit DG schemes, it is suitable to use an *implicit* time discretization.

In our case, the proposed implicit approach via the *backward Euler method* is very suitable due to a linearity of discrete form  $\mathcal{B}_h^\Theta(\cdot, \cdot)$  in both arguments and due to an independence of right-hand side form  $l_h^\Theta(\cdot)$  on semi-discrete solution which allows us the implicit treatment of  $u_h$  in (32b), consequently.

The fully discrete solution of problem (32) is defined in following way.

**Definition 4.2** Let  $0 = t_0 < t_1 < \dots < t_r = T$  be a partition of the interval  $[0, T]$  and  $\tau_l \equiv t_{l+1} - t_l$ ,  $l = 0, 1, \dots, r-1$ . We define the approximate solution of problem (5) as functions  $u_h^k \approx u_h(t_k)$ ,  $t \in [0, T]$ ,  $l = 0, \dots, r-1$ , satisfying the conditions

$$\begin{aligned} (a) \quad & u_h^{l+1} \in S_{hp}, \\ (b) \quad & \frac{1}{\tau_l} (u_h^{l+1} - u_h^l, v_h) + \mathcal{B}_h^\Theta(u_h^{l+1}, v_h) = l_h^\Theta(v_h)(t_{k+1}) \quad \forall v_h \in S_{hp}, \quad l = 0, \dots, r-1, \\ (c) \quad & u_h^0 \text{ is } S_{hp} \text{ approximation of } u^0, \end{aligned} \quad (34)$$

The discrete problem (34) is equivalent to a system of linear algebraic equations for each  $t_l \in [0, T]$ , which can be solved by a suitable solver, e.g. GMRES.

## 5 Numerical examples

In this section, we presented a simple numerical example illustrating the potency of derived numerical scheme for solution of European option pricing model. The whole algorithm is implemented in FORTRAN90 and uses piecewise linear, quadratic and cubic approximations on partition of  $\Omega$  with constant mesh size  $h$  and time step  $\tau$ .

The numerical example represents the case of European call option with expiration date  $T = 1.0$  (e.g. 1 year) and strike price  $K = 13.0$ . The computational domain was set as  $\Omega = [0, 15]$  and Black-Scholes market model parameters were the risk-free interest rate  $r = 0.15y^{-1}$  and volatility  $\sigma = 0.01y^{-1}$ .

The initial and boundary conditions are given according to (5c) and (5b), respectively. The mesh size  $h = 0.01$  and the time step  $\tau = 0.01$ . We carried out computations by piecewise cubic approximations and set  $\Theta = 0$  (incomplete variant). In order to guarantee the stability of the discrete scheme (34) with respect to penalty parameter  $\omega$ , the parameter  $C_W$  from (29) is chosen according to [6, Table 6.3].

Figure 1 shows the trimestrial development of approximation solution  $u_h^l$ , e.g. the value of call option with maturity 1/4 (top left), 2/4 (top right), 3/4 (bottom left) and 4/4 of year (bottom right), respectively. Since  $\sigma^2 \ll r$ , the convection term is large compared to the diffusive term and the problem is said to be convection dominated and partial differential equation exhibits a hyperbolic behaviour, i.e. the first-order hyperbolic term involving  $\frac{\partial u}{\partial x}$  propagates information from the right to the left of the  $x$ -axis. In financial terms it represents the increase in value of the option generated by the deterministic increase in the stock price due to the drift term.

## 6 Conclusion

We have dealt with the numerical solution of the standard option pricing models, represented by the linear convection-diffusion-reaction equation. We have derived the above used numerical scheme: from the weak solution, over the semi-discrete one to the fully discrete one. The whole method is based on the semidiscretization by the discontinuous Galerkin method in space and on the implicit backward Euler method used for discretization in time. Presented numerical examples illustrated the potency of the resulting scheme for convection-dominated problems.

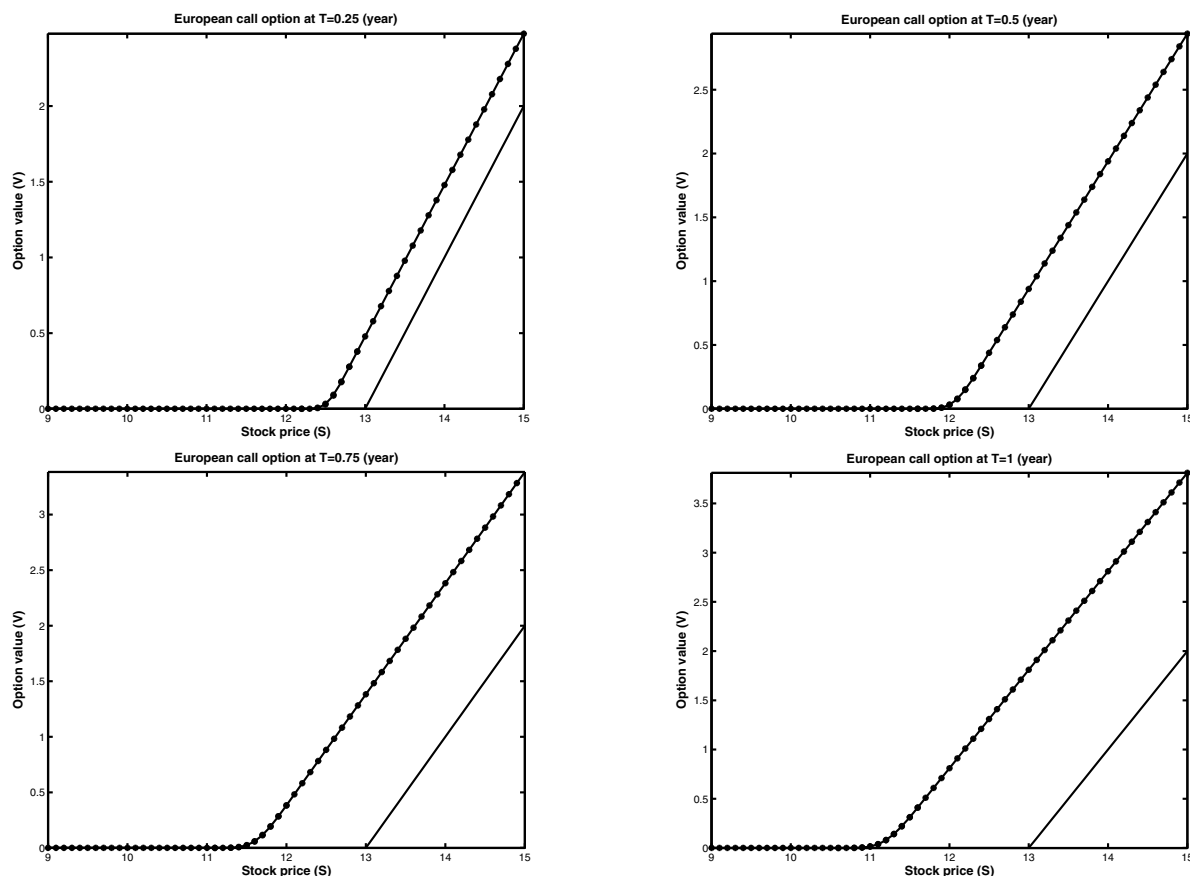


Figure 1: Values of European call options in trimestral time instants (star-line), payoff function (solid line).

For the future work, we intend to extend this method to American type of options and exotic options with barrier or discontinuous payoff function, moreover depending on a basket of several underlying assets, i.e. multivariate Black-Scholes equation.

## Acknowledgement

The paper was supported by the ESF Project No. CZ.1.07/2.3.00/09.0155 “Constitution and improvement of a team for demanding technical computations on parallel computers at TU Liberec”.

## References

- [1] ARNOLD, D. N., BREZZI, F., COCKBURN, B., MARINI, L. D.: *Unified analysis of discontinuous Galerkin methods for elliptic problems*. SIAM J. Numer. Anal. 39(5), pp. 1749–1779, 2002.
- [2] Cockburn, B.: Discontinuous Galerkin methods for convection dominated problems. In: Barth, T. J., Deconinck, H. (eds.) *High-Order Methods for Computational Physics, Lecture Notes in Computational Science and Engineering* 9, pp. 69-224. Springer, Berlin (1999)

- [3] COCKBURN, B., KARNIADAKIS, G. E., SHU, C.-W. (eds.): *Discontinuous Galerkin Methods*. Springer, Berlin, 2000.
- [4] DOLEJŠÍ, V., FEISTAUER M., HOZMAN J.: *Analysis of semi-implicit DGFEM for nonlinear convection-diffusion problems*. Comput. Methods Appl. Mech. Engrg., 196: pp.2813–2827, 2007.
- [5] DOLEJŠÍ, V., HOZMAN, J.: *A priori error estimates for DGFEM applied to nonstationary nonlinear convectiondiffusion equation*. Proceedings of ENUMATH 2009 conference, G. Kreiss et. Al. Eds., Springer, pp. 459-468, 2010.
- [6] HOZMAN, J.: *Discontinuous Galerkin method for convection-diffusion problems*. PhD thesis, Charles University Prague, Faculty of Mathematics and Physics, 2009.
- [7] FEISTAUER M., FELCMAN J., STRAŠKRABA I.: *Mathematical and Computational Methods for Compressible Flow*. Oxford University Press, Oxford, 2003.
- [8] FUSAI G., SANFELICI S., TAGLIANI A.: *Practical Problems in the Numerical Solution of PDE's in Finance*. Rend. Studi Econ. Quant 2001, pp. 105-132, 2002.
- [9] KUFNER, A.: *Weighted Sobolev spaces*. Teubner-Texte zur Mathematik, 31, BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1980.
- [10] RIVIÉRE B.: *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation*. Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 2008.
- [11] SEYDEL R.: *Tools for Computational Finance: 4<sup>th</sup> edition*. Springer, Berlin, 2008.

#### **Current address**

##### **RNDr. Jiří Hozman, Ph.D.**

Technical University of Liberec,  
Faculty of Science, Humanities and Education,  
Studentská 2,  
461 17 Liberec CZ,  
tel. number: +420 485 352 240  
e-mail: [jiri.hozman@tul.cz](mailto:jiri.hozman@tul.cz).

## ELEMENTARY SOLUTION TO THE THREE VEHICLES JEEP PROBLEM WITH SUPPORT OF THE CAS MAPLE

POTUČEK Radovan, (CZ)

**Abstract.** The jeep problem is a well known logistics problem. A jeep must cross a desert wider than it can travel on one tank of fuel with the help of optimal arrangement of fuel dumps along the route. The available resources refer, especially, to solutions of two basic variants – the single jeep problem and the convoy of jeeps problem. This contribution deals with one modification of the jeep problem with 3 vehicles and  $n$  cans of fuel ( $n > 3$ ). Elementary solutions to this problem are detailed derived for small amounts  $n$  of cans of fuel. A general solution, given by the formula for determination a distance which can be reached with  $n$  cans of fuel, is stated. Some numerical results computed by two procedures, written in the computer algebra system Maple, are presented in a form of the tables.

**Key words and phrases.** jeep problem, convoy of jeeps problem, harmonic numbers, computer algebra system Maple.

*Mathematics Subject Classification.* Primary 90B06; Secondary 40A99.

### 1 Introduction and history of the jeep problem

The jeep problem, also called desert crossing problem or exploration problem, and its modification were first mentioned in the late 19th century and in the early 20th century in books [1], [2] containing problems of mathematical recreations. The problem was first solved in 1947 by N. J. Fine in his paper [3]. Shortly thereafter, C. G. Phipps generalized the problem in [4], and solved it by arguing that the single jeep problem is equivalent to a problem involving a convoy of jeeps which travel together, some being used to refuel others, with only one jeep required to cross, the others abandoned along the way.

The jeep problem may have application to arctic expeditions in present and to interplanetary travel in future. A related problem is to determine the range of a fleet of  $n$  aircrafts with some fuel capacities and with some fuel consumption. It is assumed that the aircraft may share fuel in flight and that any of the aircraft may be abandoned at any stage. The range is defined to

be the greatest distance which can be attained in this way. This fleet range problem was solved in 1960 by J. N. Franklin in [5] using dynamic programming.

This little problem can have also a practical application in wartime situations. It achieved a great deal of attention during World War II, especially in strategy used in the pacific theatre in World War II by bombing missions including the atomic bombing missions at the end of this war. The jeep problem is still a modern and topical problem – see e.g. [6], [7].

This contribution is a free follow-up to the papers [8], [9] and [10], where elementary solutions to the single jeep problem, to the convoy of jeeps problem and to the jeep problem with two vehicles were derived and illustrated for small amounts of units of fuel and for small number of vehicles forming the convoy. Numerical solutions for some of the outstanding amounts of fuel and numbers of vehicles were computed by using the computer algebra system Maple and its basic programming language and were presented in form of tables.

## 2 Formulation and notation of the jeep problem

The original jeep problem is formulated as follows. Given a jeep that can carry one tankload (one unit) of fuel and can travel one distance unit per tankload (the jeep's fuel consumption is assumed to be constant). The jeep is required to cross a desert wider than it can travel on one tank of fuel. To do so, it may make depots of fuel in the desert. We assume that in the beginning of the jeep's mission there are  $n$  tanks (cans, units) of fuel at the border of the desert at a fixed base and that the jeep makes  $n$  trips to maximize the distance it can travel.

Further, we will study the case of three jeeps – two supporting jeeps and the chief jeep, which is supposed to reach the greatest distance.

We will use the following notation:

- $n$  – the number of cans, i.e. units of fuel, which three jeeps have available for their mission, whereas we assume that the volume of a can is equal to the volume of the jeep's fuel tank,
- $\lfloor x \rfloor$  – the floor function  $\lfloor x \rfloor$ , also called the greatest integer function, gives the largest integer lesser or equal than  $x$ ;  
for example  $\lfloor 4 \rfloor = 4$ ,  $\lfloor 4.5 \rfloor = 4$ ,
- $t$  – the number of partial trips (stages, ways) covered by two supporting jeeps and at the end of the mission by all three jeeps by using  $n$  cans of fuel,
- $p$  – the number of stretches of the road, i.e. the number of partial trips, covered by two supporting jeeps and at the end of the mission by all three jeeps by using  $n$  cans of fuel,
- $B$  – the base – the starting point, where  $n$  units of fuel are saved and where the supporting jeeps must return at the end of their every trip, except their last (final) trip, when all three jeeps travel as a convoy as far as they can before running out the fuel,
- $T$  – the target point – the farthest point, which the chief jeep can reach by support of two other jeeps and by successive using  $n$  units of fuel on its final trip, and the end of the chief jeep's journey and mission,

- $B_i$  – the list of fuel dumps established at various points along the way for temporary storage of fuel by using  $n$  cans ( $i = 1, 2, \dots, t+1$ );  
at the last fuel dump  $B_t$  but one the first supporting jeep remains standing and far, towards the last fuel dump  $B_{t+1}$ , continue the second supporting jeep together with the chief jeep, at the last fuel dump  $B_{t+1}$  the second supporting jeep remains standing and far, towards to the target  $T$ , continues the chief jeep alone,
- $d_i$  – the length of the  $i$ -th stretch ( $i = 1, 2, \dots, t+2$ ) of the road, i.e. the distance  $|B_{i-1}B_i|$  of the fuel dumps  $B_{i-1}$  and  $B_i$  on the route  $B \rightarrow B_1 \rightarrow B_2 \rightarrow \dots \rightarrow B_{t+1} \rightarrow T$ , whereas  $d_1 = |BB_1|$ ;  
for maximizing the distance the chief jeep can travel, we assume that it will start with a full fuel tank on the last stage  $B_{t+1} \rightarrow T$  of the final trip, so that it holds  $d_n = |B_{t+1}T| = 1$ ,
- $s_i$  – the amount of fuel leaved and stored at the  $i$ -th fuel dump  $B_i$  by  $i$ -th subsequent trip, where  $0 < s_i < 1$  ( $i = 1, 2, \dots, t+1$ ),
- $D(3, n)$  – the maximum total distance travelled by 3 jeeps by using  $n$  cans:

$$D(3, n) = \sum_{i=1}^{t+2} d_i = d_1 + d_2 + \dots + d_{t+1} + 1.$$

### 3 The number of trips and stretches

The number  $t$  of partial trips performed by 3 vehicles (1 chief jeep and 2 supporting jeeps), which have  $n$  cans of fuel available, is determined by the following decomposition. The number of  $n$  cans can be divided into the three parts:

- 1) 3 cans for the last trip performed by the convoy of 3 jeeps, so it remains  $n - 3$  cans of fuel, which are available for all other previous missions,
- 2)  $2k$  cans consumed at  $k$  trips by 2 supporting jeeps, where  $k = \lfloor (n-3)/2 \rfloor$ , so then it remains  $\ell = n - 3 - 2 \cdot \lfloor (n-3)/2 \rfloor$  cans of fuel,
- 3)  $\ell$  cans used up at eventual one preparatory trip performed by 2 supporting jeeps; for  $n$  odd is  $\ell = 0$  and for  $n$  even is  $\ell = 1$ .

For example, we get the following decompositions:

for  $n = 11$  cans we have decomposition  $11 = 4 \cdot 2 + 1 \cdot 3$ , so we have 4 trips of 2 supporting jeeps and 1 trip of 3 jeeps convoy, so that  $t = 5$ ,

for  $n = 12$  cans we have decomposition  $12 = 1 \cdot 1 + 4 \cdot 2 + 1 \cdot 3$ , so we have 1 preparatory trip of 2 supporting jeeps, 4 trips of 2 supporting jeeps and 1 trip of 3 jeeps convoy, so that  $t = 6$ .

Generally, we have

$$t = \left\lfloor \frac{n}{2} \right\rfloor.$$

The number  $p$  of stretches of the road is obviously  $t + 2$ , i.e. the number of partial trips  $t$  increased about 2 stretches of the last trip performed by the convoy of all 3 jeeps, so that the number of stretches is generally

$$p = \left\lfloor \frac{n}{2} \right\rfloor + 2.$$

#### 4 Solutions to the main basic variants of the jeep problem

A general solution maximizing the distance travelled by the single jeep with  $n$  cans of fuel was derived, among others, in the paper [8]. It has a form

$$D(1, n) = \frac{1}{2n-1} + \frac{1}{2n-3} + \cdots + \frac{1}{5} + \frac{1}{3} + 1 = \sum_{i=1}^n \frac{1}{2i-1}. \quad (1)$$

A general solution to the convoy of jeeps problem was derived, in the paper [9], too. For given  $n$  cans of fuel and for  $n$  jeeps, each with capacity of 1 can, the maximum distance which the chief jeep can travel is

$$D(n, n) = \frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{3} + \frac{1}{2} + 1 = \sum_{i=1}^n \frac{1}{i} = H_n, \quad (2)$$

where  $H_n$  represents the  $n$ -th partial sum of the harmonic series, i.e. the sum from one up to the  $n$ -th harmonic number.

Remark, that in the paper [10] was derived a general solution to the jeep problem with two vehicles. The maximum distance which can two jeeps travel with  $n > 2$  cans of fuel is

$$D(2, n) = \frac{1}{2(n-1)} + \cdots + \frac{1}{4} + \frac{1}{2} + 1 = \frac{1}{2} \sum_{i=1}^{n-1} \frac{1}{i} + 1 = \frac{1}{2} H_{n-1} + 1. \quad (3)$$

It is obvious that it holds  $D(n, n) > D(2, n) > D(1, n)$  for  $n > 2$  and

$$D(n, n) > D(1, n) \quad \text{for } n \geq 2. \quad (4)$$

#### 5 Elementary solution to the jeep problem with two vehicles

Because, as we stated above in (4), for  $n \geq 2$  a convoy of  $n$  jeeps with  $n$  cans of fuel overcomes greater distance than a single jeep with  $n$  cans, we will consider and study, for maximizing the distance overcoming by 3 jeeps with  $n$  cans, where  $n > 3$ , the following strategy:

For  $n$  even 2 supporting jeeps (fulled with  $1/2$  volume of their fuel tanks) perform one preparatory trip and establish the first fuel dump  $B_1$ . Other fuel dumps for temporary storage of fuel in the total number  $\lfloor n/2 \rfloor + 1$  are established, step by step, during all  $t = \lfloor n/2 \rfloor$  trips of 2 supporting jeeps. At the last  $\lfloor n/2 \rfloor$ -th trip, which is performed by the convoy of 3 jeeps, are created two last fuel dumps  $B_t$  and  $B_{t+1}$ , where the supporting jeeps remain standing. The chief jeep continues from the fuel dump  $B_{t+1}$  alone on its journey towards to the target  $T$ .

Now, we describe in full details, for example, the solution for  $n = 10$  cans of fuel and then symbolically and briefly describe solutions of other four cases – for  $n = 4, 5, 6$  and  $7$  cans.

For  $n = 10$  cans we have  $t = \lfloor 10/2 \rfloor = \lfloor 5 \rfloor = 5$  trips,  $t+1 = 6$  fuel dumps and  $p = t+2 = 7$  stages of the road. Because we get a decomposition  $10 = 1 \cdot 1 + 3 \cdot 2 + 1 \cdot 3$ , we have ( $n$  is even number) 1 preparatory trip of 2 supporting jeeps, 3 trips of 2 supporting jeeps and 1 (the last and 5th) trip of 3 jeeps convoy. The supporting jeeps step by step create fuel dumps  $B_1, B_2, B_3, B_4, B_5, B_6$  at bottoms of their trips.

On the 1st preparatory trip the both supporting jeeps (each of the jeeps is filled only with  $1/2$  volume of its fuel tank) go from the base  $B$  to the fuel dump  $B_1$  and back to  $B$ . The supporting jeeps further go from  $B$  subsequently to the fuel dumps  $B_2, B_3, B_4$  and back, so through the fuel dump  $B_1$  they go during these all four trips 8-times and then together with the chief jeep once at the last 5th trip. For the unknown distance  $d_1 = |BB_1|$  thus we get the equation  $8d_1 + 3d_1/2 = 1/2$ , hence  $d_1 = 1/19$ , so the amount of fuel leaved and stored by each of the supporting jeep at the 1st fuel dump  $B_1$  by 1st subsequent and preparatory trip is  $s_1 = 1/2 - 2d_1 = 15/38$ .

For the distance  $d_2 = |B_1B_2|$  from the equation  $6d_2 + 3d_2/2 = 1$  it follows  $d_2 = 2/15$ , so the amount of fuel leaved and stored by each of the supporting jeep at the 2nd fuel dump  $B_2$  is  $s_2 = 1 - 2d_2 = 11/15$ . For the distance  $d_3 = |B_2B_3|$  from the equation  $4d_3 + 3d_3/2 = 1$  it follows we have  $d_3 = 2/11$ , so  $s_3 = 1 - 2d_3 = 7/11$ , and further for the distance  $d_4 = |B_3B_4|$  we get  $2d_4 + 3d_4/2 = 1$ , so  $d_4 = 2/7$  and  $s_4 = 1 - 2d_4 = 3/7$ .

Because the maximal distance performed by three jeeps convoy is  $D(3, 3) = 1/3 + 1/2 + 1$ , so  $d_5 = 1/3$  and  $s_5 = 1 - d_5 = 2/3$ . This amount of fuel – the rest volume of the 1st supporting jeep's tank – is in the fuel dump  $B_5$  equally overdrawn to the chief jeep and to the 2nd supporting jeep, while the 1st supporting jeep remains standing in  $B_5$ . The chief jeep and the 2nd supporting jeep continue on their journey towards to the target  $T$ . Further,  $d_6 = 1/2$ ,  $s_6 = 1 - d_6 = 1/2$  (remaining  $1/2$  volume of the 2nd supporting jeep's fuel tank is overdrawn to the chief jeep) and  $d_7 = 1$ . Totally, we have

$$D(3, 10) = \sum_{i=1}^7 d_i = \frac{1}{19} + \frac{2}{15} + \frac{2}{11} + \frac{2}{7} + \frac{1}{3} + \frac{1}{2} + 1 = \frac{327441}{131670} \doteq 2.486831.$$

This case is illustrated on the following picture:

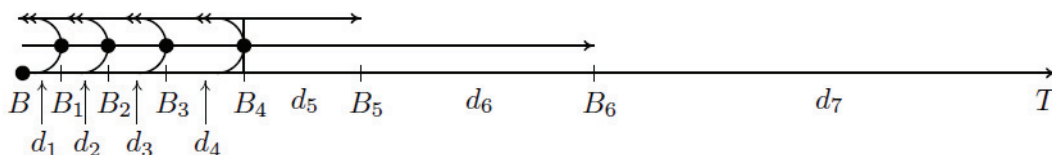


Figure 1: The case of 3 jeeps and 10 cans of fuel

Solutions to four other cases in brief:

▷  $n = 4$  cans:  $t = \lfloor 4/2 \rfloor = 2$  trips, decomposition is  $4 = 1 \cdot 1 + 1 \cdot 3$ , the supporting jeeps go through  $B_1$  once themselves on the preparatory trip and then once together with the chief jeep, therefore we have the equation  $2d_1 + 3d_1/2 = 1/2$ , hence  $d_1 = 1/7$ , and thus  $D(3, 4) = 1/7 + 1/3 + 1/2 + 1 = 83/42 \doteq 1.976190$ ,

▷  $n = 5$  cans:  $t = \lfloor 5/2 \rfloor = 2$  trips, decomposition is  $5 = 1 \cdot 2 + 1 \cdot 3$ , the supporting jeeps go through  $B_1$  once and all 3 jeeps also once, therefore  $2d_1 + 3d_1/2 = 1$ , hence  $d_1 = 2/7$ , thus  $D(3, 5) = 2/7 + 1/3 + 1/2 + 1 = 89/42 \doteq 2.119048$ ,

▷  $n = 6$  cans:  $t = 3$ ,  $6 = 1 \cdot 1 + 1 \cdot 2 + 1 \cdot 3$ , the supporting jeeps go through  $B_1$  twice (once on the preparatory trip) and once with the chief jeep, therefore  $4d_1 + 3d_1/2 = 1/2$ , hence  $d_1 = 1/11$ , the supporting jeeps go through  $B_2$  once and once with the chief jeep, therefore  $2d_2 + 3d_2/2 = 1$ , hence  $d_2 = 2/7$ , thus  $D(3, 6) = 1/11 + 2/7 + 1/3 + 1/2 + 1 = 1021/462 \doteq 2.209957$ ,

▷  $n = 7$  cans:  $t = 3$ ,  $7 = 2 \cdot 2 + 1 \cdot 3$ , the supporting jeeps go through  $B_1$  twice and once with the chief jeep, therefore we have  $4d_1 + 3d_1/2 = 1$ , hence  $d_1 = 2/11$ , the supporting jeeps go through  $B_2$  once and once with the chief jeep, therefore we get  $2d_2 + 3d_2/2 = 1$ , hence  $d_2 = 2/7$ , thus  $D(3, 7) = 2/11 + 2/7 + 1/3 + 1/2 + 1 = 1063/462 \doteq 2.300866$ .

Clearly, we get two cases – for  $n$  even, i.e.  $n = 2k$ , and for  $n$  odd, i.e.  $n = 2k - 1$ , where  $k \geq 2$  is integer:

$$D(3, 2k) = \frac{1}{3 + 4[k - 1]} + \frac{2}{3 + 4([k - 1] - 1)} + \cdots + \frac{2}{11} + \frac{2}{7} + \frac{1}{3} + \frac{1}{2} + 1,$$

$$D(3, 2k - 1) = \frac{2}{3 + 4[(2k - 3)/2]} + \frac{2}{3 + 4([ (2k - 3)/2 ] - 1)} + \cdots + \frac{2}{11} + \frac{2}{7} + \frac{1}{3} + \frac{1}{2} + 1.$$

Because we can write  $[k - 1] = [(2k - 2)/2]$  and  $[(2k - 3)/2] = [((2k - 1) - 2)/2]$ , we get for  $n \geq 3$  a **general formula**

$$D(3, n) = 2 \sum_{i=1}^{\lfloor (n-2)/2 \rfloor} \frac{1}{3 + 4i} - \frac{(n - 1) \bmod 2}{3 + 4\lfloor (n - 2)/2 \rfloor} + \frac{11}{6} \quad (5)$$

(note, that for  $n = 3$  we get, in agree with (2),  $D(3, n) = 11/6 = 1/3 + 1/2 + 1 = D(3, 3)$ ).

Now, let us consider a trivial case of  $n = 4$  cans and note that another strategy does not give better result than the strategy S described below and above in the begin of this section.

▷ Strategy S: The supporting jeep establishes depot of fuel  $B_1$  at its first trip. The second trip is performed by the convoy of 2 jeeps, whereat the supporting jeep establishes the fuel dump  $B_2$ , overdraws the rest volume of its tank to the chief jeep and remains standing in  $B_2$ , whereas the chief jeep continues on its journey towards to the target  $T$  – see Fig. 2. Using the formula (5) we get the distance  $D(2, 3) = 1/4 + 1/2 + 1 = 7/4 = 1.75$ .

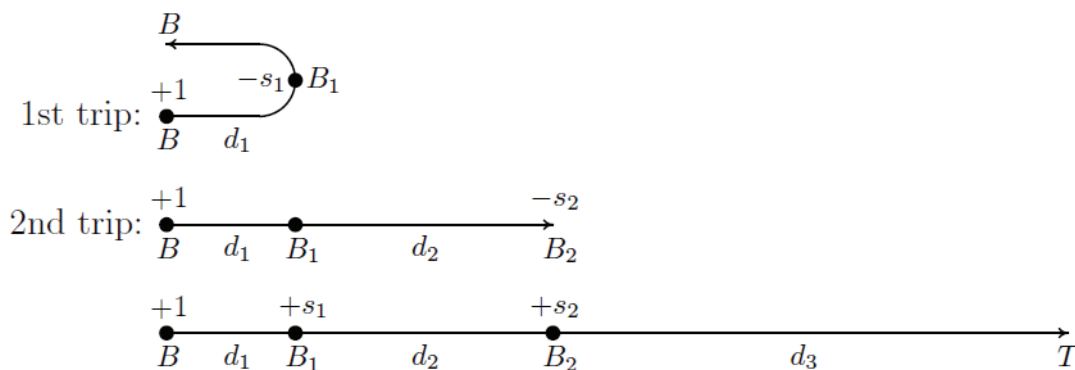


Figure 2: Strategy S for 2 jeeps with 3 cans

▷ Strategy A: This strategy consists of a trip of the chief jeep from  $B$  to the fuel dump  $B_1$ , where it stores  $s_1$  units of fuel, and back, a trip of the supporting jeep from  $B$  through  $B_1$  to the fuel dump  $B_2$ , where it stores  $s_2$  units of fuel and remains standing, and then a trip of the chief jeep from  $B$  to  $B_1$ , where it tanks  $s_1$  units of fuel and continues with a full fuel tank to  $B_2$ ,

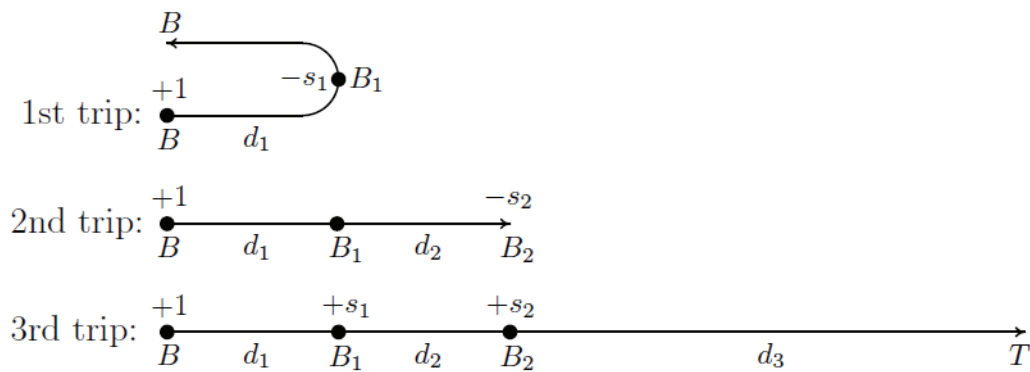


Figure 3: Strategy A for 2 jeeps with 3 cans

where it tanks  $s_2$  units of fuel and continues again with the full fuel tank to the target  $T$  (see Fig. 3), so the length of the stage  $B_2T$  is  $d_3 = 1$ .

The chief jeep goes 3-times through  $B_1$ , so we have  $3d_1 = 1$ , hence  $d_1 = s_1 = 1/3$ . For the supporting jeep in  $B_2$  we get  $d_1 + d_2 + s_2 = 1$ , i.e.  $d_2 + s_2 = 2/3$ . Because for the stage  $B_1B_2$  it holds  $d_2 = s_2$ , we have  $d_2 = s_2 = 1/3$ . The maximum travelled distance is  $D'(2, 3) = 1/3 + 1/3 + 1 = 5/3 \doteq 1.67$ .

▷ Strategy B: Let us suppose that the supporting jeep does not participate in the mission. Then we have the variant of the single jeep (see Fig. 4, where  $d_1 = 1/5$ ,  $d_2 = 1/3$ ,  $d_3 = 1$ ,  $s_1 = 3/5$ ,  $s_2 = 1/3$ ). The maximum travelled distance is  $D(1, 3) = 1/5 + 1/3 + 1 = 23/15 \doteq 1.53$ .

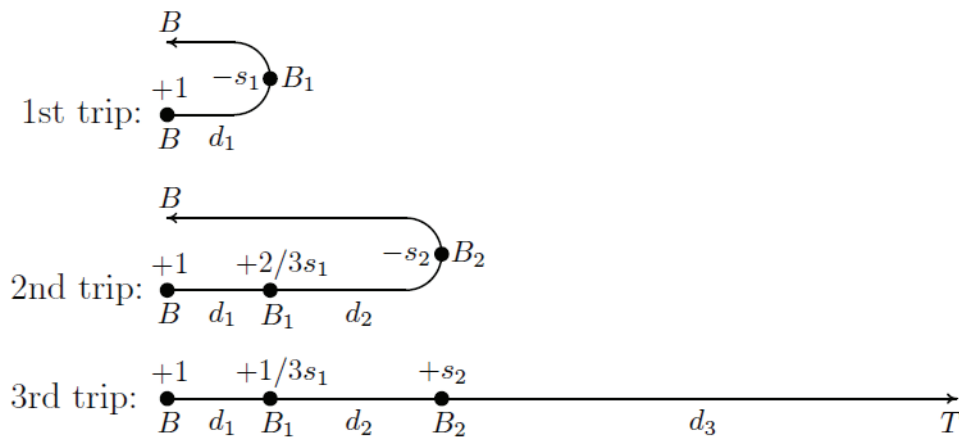


Figure 4: Strategy B for 2 jeeps with 3 cans

## 6 Some numerical solutions to the jeep problem with three vehicles

For computing some solutions to the jeep problem with three vehicles, i.e. distances  $D(3, n)$  for  $n$  cans of fuel, according to the general formula (5), was used the following simple procedure `jp3na` written in the basic programming language of the computer algebra system Maple:

```

jp3na:= proc(n)
    local i, s;
    s:= 11/6;
    for i from 1 to n-3 do
        s:= s+1/(3+4*floor((i+1)/2));
    end do;
    s:= evalf[10](s);
    print("distance D(3,n) for", n, "cans is", s);
end proc:

```

Some solutions, i.e. distances,  $D(3, n)$ , all expressed in 3 decimals, for the outstanding amounts  $n$  of fuel are presented in the table Tab. 1:

$n$	10	20	30	40	50	60	70	80	90
$D(3, n)$	2.487	2.848	3.055	3.201	3.314	3.406	3.484	3.551	3.610
$n$	100	200	1000	2000	10000	20000	70000	80000	1000000
$D(3, n)$	3.663	4.011	4.817	5.163	5.968	6.315	6.941	7.008	8.271

Tab. 1: Some amounts  $n$  of fuel and corresponding distances  $D(3, n)$

The following procedure `jp3nb` based on the general formula (5) and written also in the computer algebra system Maple, contrary to preceding procedure `jp3na`, computes corresponding amounts  $n$  of the fuel consumed by three vehicles necessary for overcoming (or crossing) distances greater than  $d = 2, 3, \dots, 10$ :

```

jp3nb:= proc(n)
    local d, i, s;
    s:= 11/6;
    d:= 2;
    for i from 1 to n-3 do
        s:= evalf[10](s+1/(3+4*floor((i+1)/2));
        if s>d then
            print("for overcoming the distance greater than", d,
                (more exactly", s, ") it is needed to use", i+3, "cans");
            d:= d+1;
        end if;
    end do;
end proc:

```

Relevant amounts  $n$  of fuel, computed by the procedure `jp3nb`, needed for overcoming distances greater than  $d = 2, 3, \dots, 10$  are stated in the table Tab. 2:

$d$	2	3	4	5	6	7	8	9	10
$n$	5	27	196	1443	10658	78745	581842	4299257	31768295

Tab. 2: Needed amounts  $n$  of fuel for overcoming distances greater than  $d = 2, 3, \dots, 10$

## 7 Example with solutions

In the following Example 1 are solved two basic problems – determination a maximal distance which can be reached with fixed amount of fuel and determination an amount of fuel needed for overcoming given units of distance.

### Example 1:

For a group of 3 jeeps, using the strategy  $S$  above, determine:

- 1) a maximal distance which can be reached with 30 cans of fuel,
- 2) a number of cans of fuel needed for overcoming 4 units of distance.

*Solution:*

- 1) For determination the distance  $D(3, 30)$  we use the general formula (5). In this way we have

$$D(3,30) = 2 \sum_{i=1}^{\lfloor (30-2)/2 \rfloor} \frac{1}{3+4i} - \frac{(30-1) \bmod 2}{3+4\lfloor (30-2)/2 \rfloor} + \frac{11}{6} = 2 \sum_{i=1}^{14} \frac{1}{3+4i} - \frac{1}{3+4 \cdot 14} + \frac{11}{6} =$$

$$= 2 \left( \frac{1}{7} + \frac{1}{11} + \frac{1}{15} + \frac{1}{19} + \frac{1}{23} + \frac{1}{27} + \frac{1}{31} + \frac{1}{35} + \frac{1}{39} + \frac{1}{43} + \frac{1}{47} + \frac{1}{51} + \frac{1}{55} + \frac{1}{59} \right) - \frac{1}{59} + \frac{11}{6}.$$

After a short calculation we get the searched result:  $D(3, 30) \doteq 3.055027$ .

- 2) We use direct the results of the procedure `jp3nb` above, where is, among other, stated that for overcoming 4 (more exactly 4.000799) units of distance it is necessary to have 196 cans of fuel available.

## 8 Conclusion

Elementary solutions to one of the interesting logistics problems – the jeep problem with three vehicles – are detailed derived for small amounts of fuel. A general solution, given by the formula for determination a distance  $D(3, n)$  which can be reached with  $n$  cans of fuel, is stated. Two simple procedures written in the computer algebra system Maple – `jp3na` for determination a maximal distance which can be reached with fixed amount of fuel and `jp3nb` for determination an amount of fuel needed for overcoming given units of distance – are presented together with some numerical results arranged in a form of the tables. This paper can be an inspiration for teachers of mathematics or as a subject matter for work with talented students.

## References

- [1] ROUSE BALL, W. W.: Mathematical Recreations and Essays. Macmillan and Co., Ltd., 4th Edition, 1905 (1st Edition, 1892), p. 26. Reprinted as eBook by the Project Gutenberg License, 2008. [online], [cit. 2011-03-01]. Available from WWW: <<http://www.gutenberg.org/files/26839/26839-pdf.pdf>>.
- [2] DUDENEY, H. E.: Amusements in Mathematics. Thomas Nelson & Sons, 1917. Reprinted by Dover Publications, 1958.
- [3] FINE, N. J.: The jeep problem. *The American Mathematical Monthly*, Vol. 54, No. 1 (1947), pp. 24-31.

- [4] PHIPPS, C. G.: The Jeep Problem: A more general solution. *The American Mathematical Monthly*, Vol. 54, s. 458-462, 1947.
- [5] FRANKLIN, J. N.: The Range of a Fleet of Aircraft. *Journal of the Society for Industrial and Applied Mathematics*, Vol. 8, No. 3 (1960), pp. 541-548.
- [6] ZHAO, F., BEN-ISRAEL, A.: A dynamic programming solution of the Jeep problem [online]. 1995, page created: 27.11.1995 [cit. 2011-03-01]. Available from WWW: <<http://benisrael.net/JEEP.pdf>>.
- [7] BONDT de, M.: An ode to Phipps jeep convoys [online]. 2010, page created: 13.9.2010 [cit. 2011-03-01]. Available from WWW: <[http://arxiv.org/PS\\_cache/arxiv/pdf/1009/1009.2937v1.pdf](http://arxiv.org/PS_cache/arxiv/pdf/1009/1009.2937v1.pdf)>.
- [8] POTŮČEK, R.: Elementary solution to the jeep problem with one vehicle. In *Proceedings of XXIXth International Colloquium on the Management of Educational Process*, Brno, FEM UO, 2011, 8 pp. ISBN 978-80-7231-779-0.
- [9] POTŮČEK, R.: Elementary solution to the convoy of jeeps problem and two different computational algorithms. In *Zborník vedeckých prác "Teoretická a edukačná transformácia matematického vzdelávania 2011"*, Fakulta ekonomiky a manažmentu, Slovenská poľnohospodárska univerzita v Nitre, 2011, 6 pp. ISBN 978-80-552-0604-2.
- [10] POTŮČEK, R.: Elementary solution to the jeep problem with two vehicles. In *Proceedings of the 7th Conference on Mathematics and Physics at the Technical Universities*, Brno, FVT UO, 2011, 7 pp. ISBN 978-80-7231-815-5.

#### **Current address**

**RNDr. Radovan Potůček, Ph.D.**

Department of Mathematics and Physics,  
University of Defence,  
Faculty of Military Technology,  
Kounicova 65, 662 10 Brno, Czech Republic,  
tel. 0042 973 443 056,  
e-mail: Radovan.Potucek@unob.cz

## HETEROGENEOUS CLUSTER FOR ACCELERATION OF LINEAR ALGEBRA COMPUTATIONS

ŠIMEČEK Ivan, (CZ), LANGR Daniel, (CZ)

**Abstract.** Plenty of numerical algebra libraries have been developed in recent years. These libraries are tuned for the given CPU and its memory architecture, fully utilize its memory hierarchy and inner pipelines and achieve impressive computation power. There is also a new trend in the high-performance computing: GPU computing.

This paper deals with a new concept of the heterogeneous grid for acceleration of the numerical linear algebra computing. We design this grid with respect to maximal ratio between cost and computational power. It allows a parallelization of scientific codes with minimal programming effort. We also optimize grid concept to be less sensitive to network parameters.

**Key words and phrases.** General purpose GPU computing, grid computing, remote calls, library for numerical linear algebra.

*Mathematics Subject Classification.* Primary 68W10, 68M14, 65Y05; Secondary 68-04.

### 1 Introduction

Time is very often the limiting factor in scientific codes. These codes can be accelerated by parallel executing on special distributed systems (clusters or grids). Parallelization of code is a very difficult task only for experts. Contributions of this paper is a new approach for parallelization of scientific codes by converting local numerical library calls into remote grid calls.

#### 1.1 Libraries for CPU computing

The standard code of these routines have good performance due to high cache hit ratio only for small sizes of order of matrix. For good performance even for larger values, it must be

modified. In numerical algebra packages, this is achieved by explicit loop restructuring[1, 2]. It includes loop unrolling-and-jam which increase the FPU pipeline utilization in the innermost loop, loop blocking and loop interchange to maximize a cache hit ratio. After application of these transformations, these codes are divided into two parts. Outer loops are "out-cache", inner loops are "in-cache". Codes have almost the same performance independently on the amount of data. Normal user just use some libraries (like BLAS [3]), where codes are error-free and also optimized for given architecture.

## **1.2 Computing on GPUs**

### **1.2.1 History of computing on GPUs**

The computing on Graphics Processing Units (GPUs) is a trend caused by the surprising fact that the most powerful part of modern Intel PCs is not the CPU, but the GPU. Modern graphic cards overcome modern CPUs in the memory bandwidth, the number of computational units and possibilities of the vector execution, which results in their surprising floating point performance.

First papers about this GPU computation phenomena were published in 2001, when GPUs (more accurately: their shader units) became programmable. And many papers were published in recent years [4, 5, 6], because the newest GPUs have ability for floating-point computation.

### **1.2.2 Nowadays GPU computing**

Trend of accelerate computations by means of GPU in high-performance computing still grows. This trend recently emerged into a new research area called General-Purpose Computing on Graphics Processing Units (shortly GPGPU). The GPGPU programming is simplified by several existing APIs (Application Programming Interfaces), the most popular and well-established ones are CUDA[8, 9] and OpenCL[10]. Thanks these APIs the GPGPU computations are widespread and has been commonly used in many scientific projects.

The computational abilities of single GPU are very impressive, but some problems, especially with large memory requirements, are still hard to solve. Although the amount of memory on GPUs is increasing rapidly, it is still much less than we need and this leads to the limited application of GPGPU in many scientific problems. Possible solution to that problem could be to connect graphic cards into a GPGPU cluster to distribute computing and memory demands across all available GPU. The benefit of this approach is that it allows us to interconnect GPUs from various vendors but naturally there arise a new problem known as GPU's load balancing that we have to face to retain high computational performance.

### **1.2.3 Compute United Device Architecture (CUDA)**

Compute United Device Architecture (for details see [7]) is a proprietary hardware and software solution for data-intensive computing from NVIDIA. Completely built from ground up, the CUDA represents a new generation of future graphics cards. The biggest difference from the

previous generations resides in the simplified architecture, which however allows the processors to run on higher clock speed. Great progress has been made in design of the processing units, which now allows synchronization of the threads and their cooperation using shared on-chip memory. The graphical pipeline has been enhanced by adding new programmable stage called geometry shader and new data stream output. CUDA breaks down the traditional concepts of the general purpose programming on GPUs by enabling techniques such as scattering and inter-thread communication.

The united architecture comes with the G80 graphics cards series. These cards are marked as CUDA-enabled and their possibilities is still growing. One possible motivation for GPU computation is that the powerful six-core Intel CPU at 4 GHz, has got peak performance about 96 GFlops while the comparable-in-price Nvidia GPU Geforce 590 has peak performance about 2.50 TFlops!

## 2 Grid concept

There are a lot of grids differ in their sizes, capabilities and purposes. We want to design the grid with the maximal ratio between cost and computational power. To achieve this goal with limited budget, we must maximize GPU usage for the computation.

### 2.1 Grid architecture

We assume that:

- The **service** is application (on its server side) for computing results of remote calls. The service also controls and schedules the grid.
- The whole **grid** (system) consists of one or more clusters. They are connected by Internet network.
- Each **cluster** consists of one or more computers (nodes). For the communication among the nodes inside one cluster we assume some type of local network and the MPI (Message Passing Interface) library.
- Each **node** has some number of GPUs (not necessarily of the same type).
- Each cluster has exactly one server of service. Server of cluster will manage and schedule other (slave) parts (CPUs and GPUs) and monitor their workload.

Since in this architecture new and old GPUs are often mixed, this requires good load-balancing strategy. Clusters allow hybrid computations: it means that some parts of one problem are computed on CPU, some parts on GPU.

### 2.2 Numerical linear algebra computations

We will use third-party routines for numerical linear algebra computations:

- for single-node computations (multi-threaded for shared memory):
  - ATLAS for optimized BLAS and LAPACK routines,
  - PARDISO for sparse systems of linear equations,
- for cluster distributed memory computations:
  - ScaLAPACK as library of high-performance linear algebra routines for distributed-memory message-passing MIMD computers,
  - SuperLU for sparse systems of linear equations,
- for single GPU computations:
  - CUBLAS as efficient implementations of Basic Linear Algebra Subroutines on Nvidia GPUs,
  - CUFFT for Fast Fourier Transformation,
  - CUSP, CUSPARSE and MAGMA for iterative methods for solution of sparse systems of linear equations,

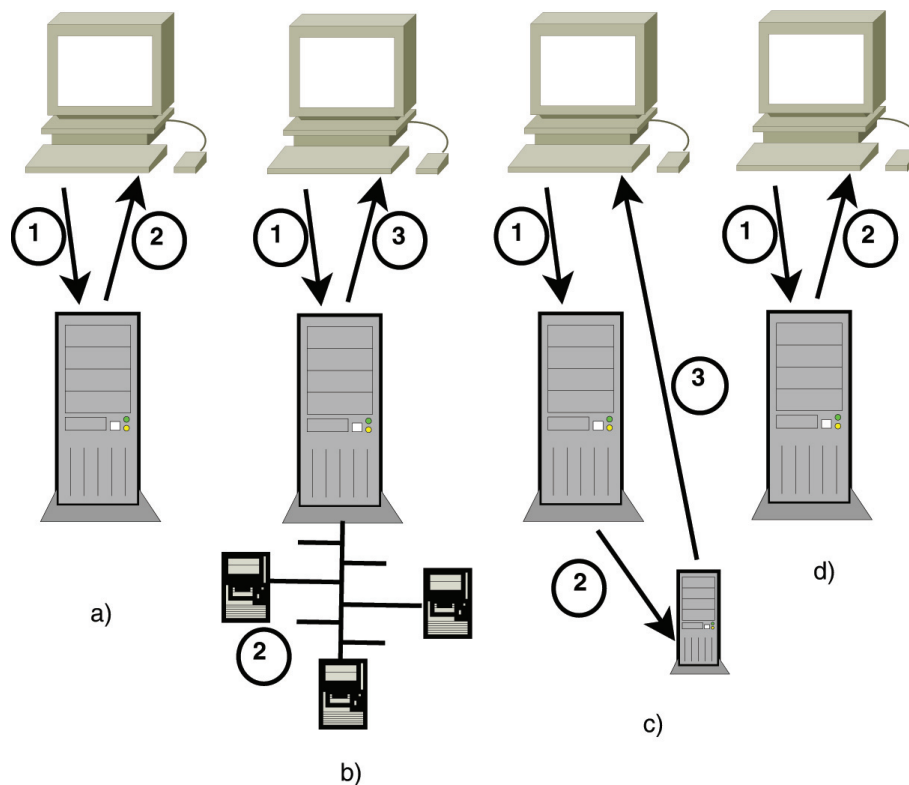


Figure 1: Possible responses from the server of service

### 2.3 Idea of remote grid calls

Usually, only special variants of codes are executed on the grid. This approach has serious drawback that code must be modified for the grid computing. We want to overcome this limitation and extend the utilization of the grid. To do this, we rewrite interface for some routines from numerical linear algebra packages (like BLAS or LAPACK). So, most of codes can utilize the computational power of the grid without any additional modifications.

### 2.4 Example of remote grid calls

The following piece of code represent typical call of BLAS routine (DGEMM = matrix-matrix multiplication).

```
for (int i = 0; i < n; i++)
    for (int j = 0; j < n; j++)
        A(i, j) = rand()/maxr;
for (int i = 0; i < n; i++)
    for (int j = 0; j < n; j++)
        B(i, j) = rand()/maxr;
double alpha = 1.;
double beta = 0.;
cblas_dgemm(CblasColMajor, CblasNoTrans, CblasNoTrans, n, n, n, \
    alpha, &*A.begin(), n, &*B.begin(), n, beta, &*C.begin(), n);
```

The normal BLAS interface invoke DGEMM routine kernel and proceeds computation locally. With modified interface to BLAS routine, the behaviour is different. During initialization of the grid, every routine is benchmarked. During call of DGEMM routine, a heuristic on client side estimate execution time for local computations and time for computations using the grid. Heuristics are needed because time can be precisely predicted for some routine (e.g. matrix-matrix multiplication), but some not for some others (e.g. iterative solvers). In our example, the heuristic for DGEMM enumerates these expressions:

$$t_1 = k_1 n^3 \quad t_2 = k_2 n^3 + n_2 n^2 + l_2,$$

where:

- $k_1$  denotes the performance of local system (evaluated during the benchmark stage).
- $k_2$  denotes the performance of cluster performance (evaluated during the benchmark stage).
- $n_2$  denotes the network throughput (evaluated periodically).
- $l_2$  denotes the network latency (evaluated periodically).
- $t_1$  denotes the expected time for local computations.
- $t_2$  denotes the expected time for grid computations (remote call).

So, the heuristic decides if it will be faster to compute this routine locally or send it to the grid for the execution. Other requirements of routine are also considered (for example amount of global memory). If client's heuristic decide to use the grid for computation, the client do a remote call of this routine by sending a demand to any server of grid. The server considers this demand and chooses one of following options (see Figure 1) depending on current workload:

- a) It computes this demand by itself (the master node of the cluster is used).
- b) It computes this demand by its cluster (one or all nodes of the cluster are used).
- c) It forwards this demand to some other server (nodes of different cluster are used).
- d) It refuses this demand (the grid is overloaded). The client is forced to do the local computation.

After the remote grid call is executed, results are send back to the client and the routine is finished.

#### **2.4.1 Optimization of the grid**

To increase utilization of the grid and increase the level of parallelization, we propose these optimizations:

- We store all structures on the grid. It allows us to reuse these structures without sending them repeatedly through the network. The deallocation of structures is controlled by user.
- The ordering of operations doesn't depends strictly on the time of arrival, but it depends on data dependencies. So, the operations can be performed in out-of-order fashion as soon as once they have access to all input operands. The problem of synchronization of these accesses is known as the Readers/Writers Problem. Our implementation controls data dependencies on the master of the cluster by using IPC mechanisms (mutexes and semaphores).

#### **2.4.2 Discussion**

The proposed concept based on remote grid calls have some advantages:

1. Time: program can be executed faster because most time-consuming parts of the code are executed on more powerful platform.
2. Implementation: some parts of program can be executed in parallel without any additional modifications.
3. Administration: all mathematical libraries (only the newest versions providing the best performance) can be installed on the servers of service.
4. Financial: the proposed grid is not very expensive, but it provides very good ratio between performance and cost. The grid can be used for different programs.

But this concept have also some drawbacks:

1. The server of service must have a good connectivity. Fast and reliable connections to other servers of the grid are also required.
2. Network latency and bandwidth must be taken in account. They are periodically measured by every server of service, but this measurement causes additional network traffic.
3. The service is suitable only for some algorithms (the most time-consuming parts are numerical linear algebra calls, without GUI, input parameters should be given by the command line).
4. Algorithms must have computational demands greater than the communication overhead (matrix-matrix multiplication, matrix factorization, and iterative eigensolvers are good examples).

### 3 Conclusions

We propose the concept of a the new distributed system for numerical linear algebra computations. This concept is based on grid usage and remote grid calls and allows the parallel execution of many codes without any additional modifications.

### 4 Future works

- Dynamic reconfiguration of the grid. Nodes can be dynamically connected or disconnected from the grid. This is great advantage because for example classroom computer can join the grid.
- Support for another libraries like GMP, PETSc and so on.
- Compression of the communication mainly for sparse vectors and matrices.
- Current version of the grid is focused on numerical linear algebra computations, so it supports only scalars, vectors and matrices. It would be useful to support for example graph data structures and graph algorithms.

### Acknowledgement

This research has been supported by MŠMT under research program MSM6840770014, by CESNET Development Fund (project 390/2010), and by Prague CUDA Teaching Centre(PCTC).

## References

- [1] M. WOLFE: *High-Performance Compilers for Parallel Computing*. Addison-Wesley, Reading, Massachusetts, USA, 1995.
- [2] K. R. WADLEIGH and I. L. CRAWFORD: *Software optimization for high performance computing*. Hewlett-Packard professional books, 2000.
- [3] J. J. DONGARRA, J. D. CROZ, S. HAMMARLING, and I. DUFF: *A set of level 3 Basic Linear Algebra Subprograms*. ACM Transactions on Mathematical Software, 16(1):1-17, Mar. 1990.
- [4] I. BUCK, T. FOLEY, D. HORN, J. SUGERMAN, K. FATAHALIAN, M. HOUSTON, and P. HANRAHAN: *Brook for gpus: stream computing on graphics hardware*. ACM Trans. Graph., 23(3):777-786, 2004.
- [5] J. KRUGER and R. WESTERMANN: *Linear algebra operators for gpu implementation of numerical algorithms*. ACM Trans. Graph., 22(3):908-916, 2003.
- [6] M. HARRIS: *Gpgpu: General-purpose computation on gpus*. NVIDIA, 2004.
- [7] NVIDIA Corporation: *Nvidia geforce 8800 gpu architecture overview*. 2006.
- [8] J. SANDERS and E. KANDROT: *CUDA by Example: An Introduction to General-Purpose GPU Programming*. Addison-Wesley Professional, 2010.
- [9] M. PHARR and R. FERNANDO: *GPU Gems: Programming Techniques, Tips, and Tricks for Real-Time Graphics*. Addison-Wesley Professional, March 2004.
- [10] Ryoji TSUCHIYAMA, Takashi NAKAMURA, Takuro IIZUKA, Akihiro ASAHARA and Satoshi MIKI: *The OpenCL Programming Book*. Fixstars Corporation, 2010.

## Current address

**Ivan Šimeček, Ing. PhD.**

Czech Technical University in Prague, Faculty of Information Technology,  
Thákurova 9, 160 00 Praha 6, Czech Republic,  
xsimecek@fit.cvut.cz

**Daniel Langr, Ing.**

Czech Technical University in Prague, Faculty of Information Technology,  
Thákurova 9, 160 00 Praha 6, Czech Republic,  
langrd@fit.cvut.cz

## ROUNDING ERRORS IN COMPUTER ARITHMETIC

SÝKOROVÁ Irena, (CZ)

**Abstract.** The aim of this paper is to show the standard estimate of rounding errors in computation with real computer numbers and give examples of basic computer operations in which the estimates for rounding errors are nearly reached in the computer.

**Key words and phrases.** Computer representation of real numbers, computer operation, rounding, rounding error, relative error.

*Mathematics Subject Classification.* Primary 65G50; Secondary 97N20.

### 1 Introduction

Numerical computing is important in physics, chemistry, biology or engineering. Real numbers are mostly used in these branches. Numerical methods using computation with real numbers introduce rounding errors. The knowledge of real number representation and real number arithmetic is important. The loss of precision in floating point computation can have unexpected consequences and cause some failures or accidents, see e.g. [1].

There is a question what is the best way to represent numbers in the computer. We express numbers in the decimal positional system developed in India hundreds of years ago. Although decimal representation is convenient for people, it is not particularly convenient for use in computers. The binary system is much more useful. Every number is represented as a string of bits, each of which is either 0 or 1.

Each number can be stored in the computer only to a finite number of digits. In this paper, it will be shown how real numbers are represented in the computer. The use of numbers of this type causes the necessity of some way of rounding off when real numbers are input into the computer and during performing arithmetic operations. The question of representing numbers in various number systems and dealing with them in the computer are of fundamental importance.

## 2 Floating point representation of real numbers

The real numbers are mostly stored in the computer in *floating point* representation which is based on exponential notation.

The following theorem on a general *g-adic* expansion of numbers is thus the basis for the computer representation of real numbers.

**Theorem 2.1** *Let  $g$  be an integer,  $g \geq 2$ , and let  $x$  be a real number,  $x \neq 0$ . Then  $x$  can always be represented in the form*

$$x = \operatorname{sgn}(x) g^e \sum_{i=0}^{\infty} a_i g^{-i}, \quad (1)$$

where  $e \in \mathbb{Z}$  and  $a_i \in \{0, 1, \dots, g-1\}$ . Moreover, there is a unique representation of this type with  $a_0 \neq 0$  and with the property that for every  $n_0 \in \mathbb{N}$ , there exists  $i \geq n_0$  such that

$$a_i \neq g-1. \quad (2)$$

**Proof.** It can be easily proved by a small modification of the proof in [2].

**Remark 2.2** *The number  $x$  expressed in the form (1) is called the normalized number and the number  $g$  is called the basis of the number system.*

**Example 2.3** *In a decimal positional system ( $g=10$ ) the number  $x = 423.051$  has the normalized form  $x = 4.23051 \cdot 10^2$ .*

As the binary system is often used in computer arithmetic, the basis of the number system is  $g = 2$ , so  $a_0 = 1$  and the formula (1) can be written as

$$x = \operatorname{sgn}(x) 2^e \sum_{i=0}^{\infty} a_i 2^{-i} = \operatorname{sgn}(x) \cdot 2^e \left( 1 + \sum_{i=1}^{\infty} a_i 2^{-i} \right), \quad \text{where } e \in \mathbb{Z}, a_i \in \{0, 1\}. \quad (3)$$

The set of all numbers which can be represented in the computer is finite and it is described in the next definition.

**Definition 2.4** *Let  $g, t, E_1, E_2 \in \mathbb{N}, g \geq 2, t \geq 2, E_1 \geq 1, E_2 \geq 2$ . The set of all numbers  $x \in \mathbb{R}, x \neq 0$ , which can be written in the form*

$$x = g^e \sum_{i=0}^t a_i g^{-i} \quad (4)$$

where  $e \in \mathbb{Z}, -E_1 \leq e \leq E_2$  and  $a_i \in \{0, 1, \dots, g-1\}$ , and  $a_0 \neq 0$  will be denoted by  $S^+$  and called the set of positive computer numbers. We denote by  $S^-$  the set of all numbers  $x$  such that  $-x \in S^+$  and call it the set of negative computer numbers. The set  $S = S^+ \cup S^- \cup \{0\}$  is called the set of computer numbers.

**Remark 2.5** *This type of representation is called floating point representation of real numbers.*

The expression  $\sum_{i=0}^t a_i g^{-i}$  is called *the mantissa* of the computer number  $x$ , the number  $t$  is called *the length of the mantissa*. We call  $g^e$  *exponential part* of  $x$  and the number  $e$  is *the exponent* of the number  $x$ .

Floating point number  $x$  can be stored exactly in the computer, its binary form is

$$x = \text{sgn}(x) \cdot 2^e \left( 1 + \sum_{i=1}^t a_i 2^{-i} \right), \quad \text{where } -E_1 \leq e \leq E_2, t \in \mathbb{N}, a_i \in \{0, 1\}. \quad (5)$$

To store normalized numbers, we divide the computer word into three fields as follows: one bit for the sign (0 for positive numbers, 1 for negative numbers), one field of bits for the exponent, and one field of bits for the mantissa. Because the exponent field is limited, only exponents  $e$  between  $E_1$  and  $E_2$  can be represented.

If a number  $x$  is not a floating point number, it must be rounded before it can be stored in the computer, the computer number is then an approximation of the given number.

**Remark 2.6** Zero cannot be expressed according to formula (4) with  $a_0 \neq 0$ . Zero is represented in the computer with a zero mantissa. The exponent of zero can be arbitrary, but in computers a certain particular exponent is used.

### 3 Precision

The present PC's use the IEEE Standard 754 – 1985 for computer representation of real numbers. There are three basic types: single precision, double precision, double-extended precision.

Real numbers in floating point arithmetic are stored in the computer in binary form,  $g = 2$  is used in all these types. The computer word is divided into three parts, the first of them is for the sign (1 bit), the next part is for the exponent and the remaining bits are used for the mantissa. The scheme is

sign	exponent	mantissa
------	----------	----------

The assumption  $a_0 \neq 0$  in Theorem 2.1 implies that  $a_0 = 1$  for  $g = 2$ , see the formula (4). The first binary digit  $a_0 = 1$  is not represented in the computer, therefore the computer representation contains only the digits  $a_i$ ,  $i \geq 1$ .

The following table shows numbers of bits in the representation of real numbers.

precision	total lenght	exponent	mantissa	$E_1$	$E_2$
single	32	8	23	126	127
double	64	11	52	1022	1023
double-extended	80	15	64	16382	16383

Except these floating point normalized numbers there exist other numbers called *special values* – zero, “infinity” or NaN (not a number) – the number which doesn't exist, i.e. error, see [4].

#### 4 Rounding

It is necessary to guarantee that the results of arithmetic operations are the computer numbers. The IEEE standard recommends to do it by using rounding. In this paper we realize it by methods based on definitions that describe the situation in an adequate way. First we define the rounding of a real number to a computer number.

**Definition 4.1** Let  $g, t, E_1, E_2 \in \mathbb{N}$ , such that  $g \geq 2, t \geq 2, E_1 \geq 1, E_2 \geq 2$ , and  $x \in \mathbb{R}$ , suppose  $x \neq 0$  has the representation  $x = \operatorname{sgn}(x) g^e \sum_{i=0}^{\infty} a_i g^{-i}$ , where  $e \in \mathbb{Z}, -E_1 \leq e \leq E_2$  and  $a_i \in \{0, 1, \dots, g-1\}, a_0 \neq 0$ , and with the property (2). Then we define

$$[x]_R^t = \operatorname{sgn}(x) g^e \sum_{i=0}^t a_i g^{-i} \quad \text{for } a_{t+1} < \frac{1}{2}g, \quad (6)$$

$$[x]_R^t = \operatorname{sgn}(x) g^e (g^{-t} + \sum_{i=0}^t a_i g^{-i}) \quad \text{for } a_{t+1} \geq \frac{1}{2}g. \quad (7)$$

We put  $[0]_R^t = 0$ .  $[x]_R^t$  is called the value of  $x$  rounded to  $t$  digits.

It is easy to see that applied to the decimal system, this definition reduces to the usual rounding process of arithmetic.

**Remark 4.2** Cutting out was used instead of rounding in old computers. This means that the number

$$x = \operatorname{sgn}(x) g^e \sum_{i=0}^{\infty} a_i g^{-i}$$

was represented in the computer as

$$[x]_C^t = \operatorname{sgn}(x) g^e \sum_{i=0}^t a_i g^{-i}.$$

#### 5 Absolute error, relative error

There are two ways of measuring errors of an approximation: the absolute error and the relative error, see [5].

**Definition 5.1** Let  $x, [x]_* \in \mathbb{R}$ , where  $[x]_*$  is an approximation to  $x$ . Then

$$|x - [x]_*|$$

is called the absolute error. If  $x \neq 0$ , then

$$\frac{x - [x]_*}{x}$$

is called the relative error.

The next theorem gives the estimate for the relative error of rounding.

**Theorem 5.2** Let  $g, t, E_1, E_2 \in \mathbb{N}$ , such that  $g \geq 2, t \geq 2, E_1 \geq 1, E_2 \geq 2$ , and  $x \in \mathbb{R}$ , suppose  $x \neq 0$  has the representation  $x = \text{sgn}(x) g^e \sum_{i=0}^{\infty} a_i g^{-i}$ , where  $e \in \mathbb{Z}, -E_1 \leq e \leq E_2$  and  $a_i \in \{0, 1, \dots, g-1\}, a_0 \neq 0$ , and with the property (2). Then the relative error satisfies

$$\left| \frac{[x]_R^t - x}{x} \right| \leq \frac{1}{2} g^{-t}.$$

**Proof.** For  $a_{t+1} < \frac{1}{2}g$ , we have

$$\left| \frac{[x]_R^t - x}{x} \right| = \frac{g^e \sum_{i=t+1}^{\infty} a_i g^{-i}}{g^e \sum_{i=0}^{\infty} a_i g^{-i}} < \frac{\frac{1}{2} g^{-t}}{1} = \frac{1}{2} g^{-t}.$$

On the other hand, if  $a_{t+1} \geq \frac{1}{2}g$ , then under the formula (7)

$$\begin{aligned} \left| \frac{[x]_R^t - x}{x} \right| &= \frac{g^e (g^{-t} - \sum_{i=t+1}^{\infty} a_i g^{-i})}{g^e \sum_{i=0}^{\infty} a_i g^{-i}} = \frac{g^{-t} - a_{t+1} g^{-t-1} - \sum_{i=t+2}^{\infty} a_i g^{-i}}{\sum_{i=0}^{\infty} a_i g^{-i}} \leq \\ &\leq \frac{g^{-t-1} (g - a_{t+1}) - \sum_{i=t+2}^{\infty} a_i g^{-i}}{1} \leq g^{-t-1} \frac{1}{2} g = \frac{1}{2} g^{-t}. \end{aligned}$$

**Theorem 5.3** Let  $x \in R, x \neq 0$  and  $[x]_R^t$  be its value rounded to  $t$  digits. Then

$$[x]_R^t = x(1 + \varepsilon), \quad (8)$$

where  $|\varepsilon| \leq \frac{1}{2} g^{-t}$ .

**Proof.** It is evident from Theorem 5.2, because  $\varepsilon = \frac{[x]_R^t - x}{x}$ . The formula (8) holds also for  $x = 0$  with any  $\varepsilon$ .

## 6 Computer operations

Now we consider computer operations. We denote the basic computer operations by the symbols  $\oplus, \ominus, \otimes, \odot$ . We use the symbol  $*$  to denote any of the arithmetic operations  $+, -, \times, :$ , and the symbol  $\circledast$  for any basic computer operations. We denote  $M, m$  the maximal and minimal positive computer number. We want the computer operation  $\circledast$  to be defined for all  $x, y \in S$  for which

$$m \leq |x * y| \leq M \quad \text{or} \quad x * y = 0.$$

Division by zero is not defined.

**Definition 6.1** Let  $x, y \in S$ . Let  $m \leq |x * y| \leq M$  or  $x * y = 0$ . Then we define

$$x \circledast y = [x * y]_R^t \quad (9)$$

This definition describes the way how the computer performs arithmetic operations. The following theorem gives the estimate of the corresponding rounding errors.

**Theorem 6.2** *Let  $x, y \in S$ . If the computer operation  $x \circledast y$  is defined, then there exists a number  $\varepsilon$ ,  $|\varepsilon| \leq \frac{1}{2}g^{-t}$  such that*

$$x \circledast y = (x * y)(1 + \varepsilon).$$

**Proof.** It is a direct consequence of Definition 6.1 and Theorem 5.3.

**Remark 6.3** *The expression  $1 + \varepsilon$  is called the correcting factor.*

Nowadays, the correcting factor is standard for the description and study of the influence of rounding errors. Classical results were obtained for problems of linear algebra by Wilkinson, see [6].

**Remark 6.4** *The number  $\frac{1}{2}g^{-t}$  will be denoted by  $\gamma$  and called rounding unit. It has the property*

$$\gamma = \min(\varepsilon \mid \varepsilon > 0 \wedge 1 \oplus \varepsilon > 1).$$

*It is thus the least number that increases unity when computer addition is performed. Let us underline that  $1 \oplus \gamma \neq 1 + \gamma$ .*

The rounding unit for single precision is

$$\gamma = 2^{-24} \doteq 5.960464477539063 \cdot 10^{-8}.$$

It holds  $|\varepsilon| \leq \gamma$  in Theorem 6.2. There is the question, if it is possible to have  $|\varepsilon| = \gamma$  or how close we can approach the estimate  $\gamma$ .

Critical number in the following examples is such a number whose relative error is maximal. It is for instance the number  $C = 1 + 2^{-24}$ .

The following table compares the mantissa of the number  $C = 1 + 2^{-24}$  stored in a single precision and double precision.

precision	mantissa representation of $C = 1 + 2^{-24}$
single	00000000 00000000 00000001
double	00000000 00000000 00000000 10000000 00000000 00000000 00000000

From this we can see that the number  $C$  has the maximal relative error.

## 7 Examples

The following examples for addition and subtraction are very simple. The numerical values were obtained by computation in double precision.

**Example 7.1** *Addition:*

$$\begin{aligned}x &= 1, \quad y = 2^{-24} = 5.960464477539063 \cdot 10^{-8}, \\x + y &= 1 + 2^{-24} = 1.000000059604645, \quad x \oplus y = 1 + 2^{-23} = 1.00000011920929, \\x \oplus y &= (x + y)(1 + 5.960464122267716 \cdot 10^{-8}), \\ \varepsilon &= 5.960464122267716 \cdot 10^{-8}.\end{aligned}$$

**Example 7.2** *Subtraction:*

$$\begin{aligned}x &= 1 + 2^{-23} = 1.00000011920929, \quad y = 2^{-24} = 5.960464477539063 \cdot 10^{-8}, \\x - y &= 1 + 2^{-23} - 2^{-24} = 1 + 2^{-24} = 1.000000059604645, \\x \ominus y &= 1 + 2^{-23} = 1.00000011920929, \\x \ominus y &= (x - y)(1 + 5.960464122267716 \cdot 10^{-8}), \\ \varepsilon &= 5.960464122267716 \cdot 10^{-8}.\end{aligned}$$

**Example 7.3** *Multiplication:*

We look for two computer numbers and we want their product to be the number  $C$ . This can be obtained by splitting

$$1 + 2^{-24} = (1 + 2^{-8})(1 - 2^{-8} + 2^{-16}).$$

$$\begin{aligned}x &= 1 + 2^{-8} = 1.00390625, \quad y = 1 - 2^{-8} + 2^{-16} = 0.9961090087890625, \\x \cdot y &= 1 + 2^{-24} = 1.000000059604645, \quad x \otimes y = 1 + 2^{-23} = 1.00000011920929, \\x \otimes y &= (x \cdot y)(1 + 5.960464122267716 \cdot 10^{-8}), \\ \varepsilon &= 5.960464122267716 \cdot 10^{-8}.\end{aligned}$$

In the case of division, we apply the theory of continued fractions to find suitable operands for the operation of division with possibly maximum rounding error, see e.g. [3]. It is namely known that the convergents of a continued fraction give very good approximations.

**Example 7.4** *Division:* We express the number  $2^{-1}(1 + 2^{-24}) = 2^{-1} + 2^{-25}$  as

$$\cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{2^{23} - 1 + 2^{-1}}}}$$

and approximate it with

$$\cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{2^{23} - 1}}} = \cfrac{1}{1 + \cfrac{2^{23} - 1}{2^{23}}} = \cfrac{2^{23}}{2^{24} - 1} = \cfrac{2^{-1}}{1 - 2^{-24}}.$$

Now, both the dividend and the divisor are computer numbers. We have

$$x = 2^{-1} = 0.5, \quad y = 1 - 2^{-24} = 0.999999940395355,$$

$$\frac{x}{y} = \frac{2^{-1}}{1 - 2^{-24}} \doteq 2^{-1}(1 + 2^{-24}) = 0.500000029802322,$$

$$x \odot y = 2^{-1}(1 + 2^{-23}) = 0.500000059604645,$$

$$x \oslash y = \frac{x}{y}(1 + 5.960463766996338 \cdot 10^{-8}),$$

$$\varepsilon = 5.960463766996338 \cdot 10^{-8}.$$

The value of  $\varepsilon$  is the same for the addition, subtraction and multiplication. Its exact value is

$$\varepsilon = \frac{1 + 2^{-23}}{1 + 2^{-24}} - 1 = \frac{2^{-24}}{1 + 2^{-24}}.$$

The ratio  $\frac{\varepsilon}{\gamma} = \frac{1}{1+2^{-24}}$ . In case of division we have  $\frac{\varepsilon}{\gamma} = 1 - 2^{-23}$ .

We have shown that the absolute value of the correcting factor is less than the rounding unit but it is very close to the estimate for all the four basic arithmetic operations for suitable chosen numbers. We have studied operations only with two numbers. Since it is known the accumulation of rounding errors can be dangerous in the computation it will be interesting to study how realistic the estimates of the correcting factors are in a sequence of more operations than only two.

## References

- [1] GAO Report: *Patriot Missile Defense*.  
<http://www.fas.org/starwars/gao/im92026.htm>
- [2] HÄMERLIN, G., HOFFMANN, K. H.: *Numerical Mathematics*, Springer-Verlag, Inc., New York, 1991.
- [3] CHINČIN, A. J.: *Řetězové zlomky*, Přírodovědecké vydavatelství Praha, 1952.
- [4] PRÁGER, M., SÝKOROVÁ, I.: *Jak počítače počítají*, Pokroky matematiky, fyziky a astronomie, Vol. 49 (2004), No. 1, 32–45.
- [5] RALSTON, A.: *A first course in numerical analysis*, McGraw-Hill, New York, 1965.
- [6] WILKINSON, J. H.: *Rounding Errors in Algebraic Processess*, Her Majesty's Stationery Office, London, 1963.

## Current address

**Irena Sýkorová, RNDr.**

Department of Mathematics, University of Economics, Ekonomická 957, 148 00 Praha 4,  
e-mail: sykorova@vse.cz

## TERM STRUCTURE MODELLING BY USING EXPONENTIAL BASIS FUNCTIONS

HLADÍKOVÁ Hana, (CZ)

**Abstract.** The term structure of interest rates is defined as the relationship between the yields of default-free pure discount (zero-coupon) bonds and their time to maturity. The term structure is not always directly observable. If we deal with approximations of empirical data to create yield curves it is necessary to choose suitable mathematical functions. We explore functions, which are a linear combination of the exponential basis functions. The mathematical apparatus employed for this kind of approximation is outlined. This theoretical background is applied to an estimation of the zero-coupon yield curve derived from the Czech coupon bond market.

**Keywords.** Yield curve estimation, exponential model, least squares methods.

*Mathematics Subject Classification:* Primary 91G60, 91G30; Secondary 97M30.

### 1 Introduction

The term structure of interest rates is defined as the relationship between the yields of default-free pure discount (zero-coupon) bonds and their time to maturity. The term structure is not always directly observable because, with the exception of short-term treasury-bills, most of the substitutes for default-free bonds (government bonds) are not pure discount bonds. In the exponential spline model, as introduced in Li et al. (2001), the theoretical discount function  $d(t)$  is modelled as a linear combination of exponential basis functions. The exponential model is used at the Bank of Canada. Bolder and Gusba (2002), Marciniak (2006), Lin (2002) provide an extensive review and comparison of a number of estimation algorithms.

The mathematical apparatus employed for this kind of approximation is outlined. This theoretical background is applied to an estimation of the zero-coupon yield curve derived from the Czech coupon bond market. The exponential model is a so-called function-based model. In Hladíková (2010) an alternative approach employing Fourier basis function tested in the same setting can be found. As to the Czech coupon bond market, the function-based construction of yield curve has not yet been satisfactorily explored.

In Section 2 we define the exponential model and propose a method to solve arising least squares problem. In Section 3 the data sample from the Czech coupon bond market is described. In Section 4 numerical experiments on these data are performed.

## 2 The exponential model

There are three equivalent descriptions of the term structure of interest rates (Malek, 2005):

- the **discount function** which specifies zero-coupon bond (with a par value \$1) prices as a function of maturity,
- the **spot yield curve** which specifies zero-coupon bond yields (spot rates) as a function of maturity,
- the **forward yield curve** which specifies zero-coupon bond forward yields (forward rates) as a function of maturity.

We will use the following notation:

- $t$         time to payment (measured in years),  
 $T$         time to maturity  
 $d(t, T)$    the discount function, that is the present value of a unit payment due in time  $t$ ,  
 $z(t, T)$    spot rate of maturity  $t$ , expressed as the continuously compounded annual rate.  
 $f(t, T)$    continuously compounded instantaneous forward rate at time  $t$ .

The spot rates  $z$  are related to the discount function  $d$  by the equation:

$$d(t, T) = e^{-(T-t)z(t, T)} \quad (1)$$

$$z(t, T) = \frac{-\ln(d(t, T))}{T - t} \quad (2)$$

The forward rates  $f$  are related to the spot rate  $z$  by the equation:

$$f(t, T) = \frac{\partial}{\partial T} \ln(-d(t, T)) = z(t, T) + (T - t)z'(t, T). \quad (3)$$

The spot rates  $z$  are related to the forward rate  $f$  by the equation:

$$z(t, T) = \frac{\int_t^T f(t, u) du}{T - t}. \quad (4)$$

We define:

- $N$  - number of bonds,  
 $P_{iA}$  - price (offer),  $P_{iB}$  - price (ask)  
 $P_i$  - theoretical price of  $i$ -th bond  
 $m_i$  - number of the payments of the  $i$ -th bond  
 $\bar{P}_i$  - observed price of  $i$ -th bond

$t_{ij}$  the time when the  $j$ -th payment of the  $i$ -th bond occurs

$$t_i = [t_{i1}, \dots, t_{im_i}],$$

$c_{ij}$  - the  $j$ -th payment of the  $i$ -th bond,  $c_i = [c_{i1}, \dots, c_{im_i}]^T$ ,

$$d(t_i) = [d(t_{i1}), \dots, d(t_{im_i})]^T$$

$D_i$  - duration of the  $i$ -th bond

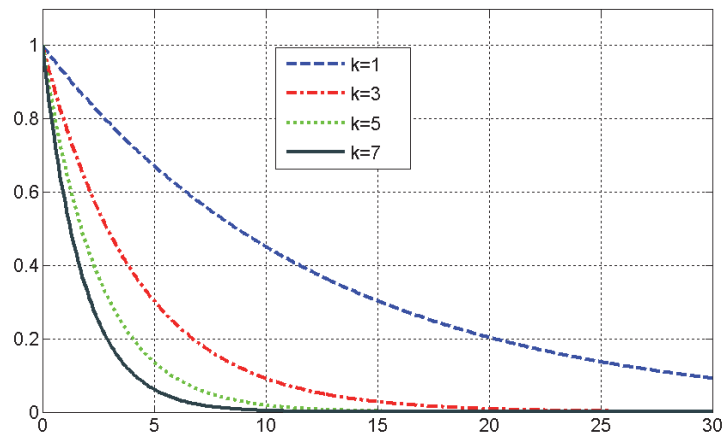
Then we can express:

$$P_i = \sum_{j=1}^{m_i} c_{ij} d(t_{ij}) = c_i^T d(t_i) \quad (5)$$

We examine the following exponential-series basis

$$B = \{f_k(t) = e^{-\alpha k t}, \quad k = 1, 2, \dots\} \quad (6)$$

Figure 1 shows  $f_k(t) = e^{-\alpha k t}$ ,  $k \in \{1, 3, 5, 7\}$  for  $\alpha = 0,08$ .



**Figure 1** Exponential series basis functions.

The discount function  $d$  is given as:

$$d(t) = \sum_{k=1}^L a_k f_k(t), \quad f_k(t) \in B, \quad (7)$$

The exponential model employs a linear combination of basis functions,  $L$  is the number of basis functions. The  $a_k$  and  $\alpha$  are unknown parameters for  $k = 1, \dots, L$  that must be estimated. An interpretation of  $\alpha$  is a long-term instantaneous forward rate.

Since  $e^{-\alpha \cdot k \cdot 0} = 1$  and  $d(0) = \sum_{k=1}^L a_k = 1$ , the number of unknown parameters is reduced by one. We

choose the number of basis functions to be  $L=4$  (cf. the value of 9 on the Canadian market). To get a more accurate fit, a higher number of basis functions is desirable. On the other hand, as  $L$  increases, the matrices used in the computations are more likely to become poorly conditioned.

Exponentials are strongly related to the discount function (see Vasicek and Fong, 1981).

Note that if one interest rate, say the instantaneous forward rate, is constant (say  $b$ ), then all other types of interest rates will be the same constant.

The true discount function is  $d(t) = e^{-bt}$ , which agrees with the linear combination (7):

$$d(t) = \sum_{k=1}^L a_k e^{-b_k t} = e^{-bt}, a_1 = 1, a_2 = 0 = a_3 = \dots = a_L = 0 \quad (8)$$

Further, parameter  $\alpha$  represents a long-term instantaneous forward rate:

$$\lim_{t \rightarrow \infty} \frac{d(t)}{a_1 \cdot e^{-\alpha t}} = \lim_{t \rightarrow \infty} \left( \frac{\sum_{k=1}^L a_k e^{-\alpha \cdot t \cdot k}}{a_1 \cdot e^{-\alpha t}} \right) = \lim_{t \rightarrow \infty} \left( 1 + \sum_{k=2}^L \frac{a_k}{a_1} e^{-\alpha \cdot t \cdot (k-1)} \right) = 1 \quad (9)$$

Discount function is asymptotic to  $a_1 \cdot e^{-\alpha t}$ . For large values of  $t$  the discount function is approximately given by  $e^{-\alpha t}$ .

The theoretical price of bond number  $j$  is given by the sum of the discounted values of its cash flows (5). From Equation (7) it follows:

$$P_i = \sum_{j=1}^{m_i} c_{ij} d(t_{ij}) = \sum_{j=1}^{m_i} c_{ij} \sum_{k=1}^L a_k f_k(t_{ij}) = \sum_{j=1}^{m_i} c_{ij} \sum_{k=1}^L a_k \cdot e^{-\alpha \cdot k \cdot t_{ij}}, \quad (10)$$

The final step is to actually estimate the parameters  $a_k$ . A natural requirement is to find these parameters such that the theoretical (=computed) prices  $P_i$  are as close as possible to the observed prices  $\bar{P}_i$ . Thus, in the sense of the least squares method we want to find a set of parameters  $a_k$  that minimizes the function  $H(P)$  given as

$$H(P) := L^2(P) := \sum_{i=0}^N w_i (P_i - \bar{P}_i)^2 \quad (10)$$

where  $w_i$  is weight of the  $i$ -th bond. We obtain the parameters  $a_k$  as the solution of the following linear system of equations:

$$\begin{aligned} \frac{\partial H(P)}{\partial a_i} &= 2 \sum_{i=0}^N w_i (P_i - \bar{P}_i) \frac{\partial P_i}{\partial a_i} = \\ &2 \sum_{i=0}^N w_i \left( \sum_{j=1}^{m_i} c_{ij} \sum_{k=1}^L a_k \cdot e^{-\alpha \cdot k \cdot t_{ij}} - \bar{P}_i \right) \frac{\partial}{\partial a_i} \left( \sum_{j=1}^{m_i} c_{ij} \sum_{k=1}^L a_k \cdot e^{-\alpha \cdot k \cdot t_{ij}} \right) = \\ &2 \sum_{i=0}^N w_i \left( \sum_{j=1}^{m_i} c_{ij} \sum_{k=1}^L a_k \cdot e^{-\alpha \cdot k \cdot t_{ij}} - \bar{P}_i \right) \left( \sum_{j=1}^{m_i} c_{ij} e^{-\alpha \cdot k \cdot t_{ij}} \right) = 0 \end{aligned} \quad (11)$$

Having found  $a_k$ 's the question arises how to choose an appropriate value of parameter  $\alpha$ . Li et al. (2001) proposed the following algorithm. Since  $\alpha$  can be interpreted as a long-term instantaneous forward rate (cf. Equation 9) its value is restricted to an interval of 5 to 9 per cent and the one-dimensional minimization problem is solved

$$\min_{\alpha} R(\alpha) \quad \alpha \in \langle 0,05, 0,09 \rangle \quad R(\alpha) = \sqrt{\frac{(P_i - \bar{P}_i)^2}{N}}$$

We employ in our numerical experiments the  $l_2$  norm as a measure of the error of observed and computed prices and solve the minimization problem:

$$\min_{\alpha} R(\alpha) \quad \alpha \in \langle \alpha_{\min}, \alpha_{\max} \rangle \quad R(\alpha) = \sqrt{\sum (P_i - \bar{P}_i)^2}$$

with the values of  $\alpha$  from a larger interval:  $\alpha_{\min} = 0,005$ ,  $\alpha_{\max} = 0,15$ . Unlike in the case when B spline basis functions are used the exponential basis functions are sufficiently smooth, and thus none of the smoothing techniques is required.

### 3 Data from the Czech coupon bond market

The Czech market is small and not as liquid as other developed markets. The original life of the Czech government bond is from 3 to 50 years. The government issued bonds with annual coupon payments. We consider here data for a selected day as given in Table 1.

We exclude two bonds with less than three months to maturity, since the yields on these securities often seem to behave oddly and one bond with more than forty-seven years to maturity, since price of bond will evidently include also another risk premium.

**Table 1** Government coupon bonds 22.2.2010 (Prague stock exchange).

ISIN	Price (ask)	Price (offer)	Issue date	Maturity	ISIN	Price (ask)	Price (offer)	Issue date	Maturity
CZ0001002158	101,4	101,67	28.1.2008	11.4.2011	CZ0001000749	120,25	120,85	26.1.2001	26.1.2016
CZ0001000764	105,03	105,37	5.10.2001	5.10.2011	CZ0001001903	106,38	106,98	30.4.2007	11.4.2017
CZ0001001887	103,4	103,73	16.4.2007	18.10.2012	CZ0001000822	110,1	110,7	18.8.2003	18.8.2018
CZ0001000814	104,23	104,67	16.6.2003	16.6.2013	CZ0001002471	111,9	112,5	23.3.2009	11.4.2019
CZ0001002729	101,52	101,88	1.2.2010	16.9.2013	CZ0001001317	101,77	102,57	12.9.2005	12.9.2020
CZ0001001143	104,82	105,28	11.4.2005	11.4.2015	CZ0001001796	102,05	106,05	4.12.2006	4.12.2036
CZ0001002737	102,7	103,15	1.3.2010	1.9.2015					

### 4 Numerical experiments

On the data set described in Section 4 we performed a series of numerical experiments mainly focusing on the following issues:

- Selection of weights  $w_i$  associated with each bond,
- Finding the value of parameter  $\alpha$ ,

- Imposing the initial condition  $d(0) = 1$  on the obtained solution. We simply eliminate one of the unknown parameters  $a_k$  using in Section 2 discussed condition

$$d(0) = \sum_{k=1}^L a_k = 1.$$

As to the weights associated with each bond, general idea is that higher weights should be placed on bonds that we believe to have observed prices that are more accurate estimates of their true prices. Many authors use the reciprocal of the modified duration  $D_i$  (see Table 2, weights labelled by 1 and 11). We tried to find a measure that would reflect the liquidity of the bond. Considering the data available from the market we propose a reciprocal of the difference between  $P_{iA}$  and  $P_{iB}$  ( $P_{iA}$  - price (offer),  $P_{iB}$  - price (ask)). It is believed that this measure reflects to some extent bond's liquidity (see Table 2, weights labelled by 12 and 13).

**Table 2** Weights  $w_i$  associated with bonds (labelled by numbers).

Weight	Description	Weight	Description
<b>0</b>	$w_i = 1,$	<b>1</b>	$w_i = \frac{1}{D_i},$
<b>10</b>	$w_i = \frac{1}{N}$	<b>11</b>	$w_i = \frac{\frac{1}{D_i}}{\sum_{j=1}^N \frac{1}{D_j}},$
Weight	Description	Weight	Description
<b>12</b>	$w_i = \frac{\frac{1}{(P_{iA} - P_{iB})^2}}{\sum_{j=1}^N \frac{1}{(P_{jA} - P_{jB})^2}},$	<b>13</b>	$w_i = \frac{\left( \frac{1}{D_i} + \frac{1}{(P_{iA} - P_{iB})^2} \right)}{\sum_{j=1}^N \left( \frac{1}{D_j} + \frac{1}{(P_{jA} - P_{jB})^2} \right)}$

The tested methods are evaluated according to various criteria. The most important criterion is the goodness of fit. It is a measure of the difference of observed and theoretical (=computed) values. We compare errors of observed prices  $\bar{P}_i$  and theoretical prices  $P_i$  in accordance with the minimization problem (10). Moreover, in place of prices the yields to maturity (YTM) are employed. The criteria are summarized in the following:

$L2_P = \sum_{i=1}^N (\bar{P}_i - P_i)^2,$ $L2_{YTM} = \sum_{i=1}^N (\bar{YTM}_i - YTM_i)^2$	$RMSE_P = \sqrt{\sum_{i=1}^N \frac{(\bar{P}_i - P_i)^2}{N}}$ $RMSE_{YTM} = \sqrt{\sum_{i=1}^N \frac{(\bar{YTM}_i - YTM_i)^2}{N}}$
--	---

$L2W_P = \sum_{i=1}^N (\bar{P}_i - P_i)^2 w_i,$ $L2W_{YTM} = \sum_{i=1}^N (\bar{YTM}_i - YTM_i)^2 w_i$	$MAE_P = \sum_{i=1}^N \frac{ \bar{P}_i - P_i }{N}$ $MAE_{YTM} = \sum_{i=1}^N \frac{ \bar{YTM}_i - YTM_i }{N}$
$HR_P = \frac{card(\bar{P}_i, P_i^O \leq \bar{P}_i \leq P_i^B)}{N}$	$HR_{YTM} = \frac{card(\bar{YTM}_i, YTM_i^O \leq \bar{YTM}_i \leq YTM_i^B)}{N}$

Another criterion is a smoothness of the obtained solution. Two measures of maximum smoothness of a curve  $y = g(x)$  between  $a$  and  $b$  are used:

$s(g) = \int_a^b \sqrt{1 + [g'(x)]^2} dx$	$h(g) = \int_a^b g''(x)^2 dx$
---	-------------------------------

**Table 3** Basic characteristics of the methods considered for comparison.

Method	Weight no.	No. of basis functions	Eliminated for d(0)=1	Method	Weight no.	No. of basis functions	Eliminated for d(0)=1
E0-1	0	4	a <sub>1</sub>	E11-4	11	4	a <sub>4</sub>
E0-4	0	4	a <sub>4</sub>	E12-1	12	4	a <sub>1</sub>
E10-1	10	4	a <sub>1</sub>	E12-4	12	4	a <sub>4</sub>
E10-4	10	4	a <sub>4</sub>	E13-1	13	4	a <sub>1</sub>
E1-1	1	4	a <sub>1</sub>	E13-4	13	4	a <sub>4</sub>
E11-1	11	4	a <sub>1</sub>	E1-4	1	4	a <sub>4</sub>

**Table 4** Ranking of the methods according to separate criteria.

Metody	Cena					YTM			Délka			Křivost			Sum1	Sum2
	MAE	L2	RMSE	L2W	HR	MAE	L2	RMSE	Disc.	Spot	Forw.	Disc.	Spot	Forw.		
<b>E10-4</b>	1	1	1	7	1	3	11	11	11	1	7	1	1	1	58	47
<b>E0-4</b>	1	1	1	11	1	3	11	11	11	1	7	1	1	1	62	51
<b>E10-1</b>	3	3	3	8	1	6	9	9	9	3	11	3	3	3	74	58
<b>E0-1</b>	3	3	3	12	1	6	9	9	9	3	11	3	3	3	78	62
<b>11-4</b>	5	5	5	5	5	9	6	6	7	7	9	6	6	6	87	66,5
<b>E12-1</b>	7	12	12	2	5	1	2	2	1	12	1	12	12	12	93	68
<b>E13-1</b>	8	11	11	3	5	5	1	1	3	11	3	11	11	11	95	70
<b>E1-4</b>	5	5	5	9	5	9	6	6	7	7	9	6	6	6	91	70,5
<b>E12-4</b>	12	10	10	1	10	2	5	5	2	6	2	8	8	8	89	72
<b>E13-4</b>	11	7	7	4	5	8	8	8	6	5	6	5	5	5	90	74
<b>E11-1</b>	9	8	8	6	10	11	3	3	4	9	4	9	9	9	102	80
<b>E1-1</b>	9	8	8	10	10	11	3	3	4	9	4	9	9	9	106	84

The tested alternatives of the exponential model are summarized in Table 3.

Method E10-4 shows the best performance if the criteria of the least error of the observed and theoretical prices are considered. The least error in terms of YTM (Yield To Maturity) reaches method E12-1. The accuracy of the approximation in prices decreases with the use of weights. Elimination of the last parameter  $a_4$  is the most appropriate. The minimum length and smoothness of the computed yield curves was obtained for method E10-4. Ranking of the methods according to separate criteria is summarized in Table 4. Sum1 is a sum of rankings according to all criteria. Sum2 is a weighted sum of rankings according to all criteria where less weight is put on the smoothness and length of the computed yield curves. The winner is in the both cases method E10-4.

The last criterion is the stability of the solution. We measure here how the results change if one bond is excluded from the set of bonds. The less sensitivity of the solution to this change in data the methods exhibits the better stability of the method is. The best stability in prices shows method E10-4 and in YTM's method E12-1 which is also best overall (see Table 5).

**Table 5** Ranking of the methods according to stability of the solution.

Methods	Price errors				YTM errors			Sum1
	MAE	L2	RMSE	L2W	MAE	L2	RMSE	
<b>E10-4</b>	1	1	1	7	11	11	11	43
<b>E0-1</b>	3	3	3	12	9	9	9	48
<b>E0-4</b>	1	1	1	11	11	11	11	47
<b>E10-1</b>	3	3	3	8	9	9	9	44
<b>E1-1</b>	9	9	9	10	3	3	3	46
<b>E11-1</b>	9	9	9	6	3	3	3	42
<b>E11-4</b>	5	6	6	5	5	7	7	41
<b>E12-1</b>	8	11	11	2	1	1	1	35
<b>E12-4</b>	11	8	8	1	8	5	5	46
<b>E13-1</b>	12	12	12	3	2	2	2	45
<b>E13-4</b>	7	5	5	4	7	6	6	40
<b>E1-4</b>	5	6	6	9	5	7	7	45

Method E10-4 was found as the overall winner. The computed yield curves for the selected methods are plotted in Figures 2 and 3. On the left-hand side the discount function  $d$  is depicted (horizontal axis represents time in years and the vertical axis the price of zero coupon bond with the nominal value of 1) and on the right-hand side the forward yield curve  $f$  (dashed line) and the spot yield curve  $z$  (solid line) are depicted.

Parameter  $\alpha$  can be interpreted as a long-term instantaneous forward rate (cf. Equation 9). From the mathematical point of view the best approximations were obtained for values below the recommended 5 percent. Thus we sought for optimum  $\alpha$  in interval 0.005 to 0.15 (i.e. 0.5 to 15 percent). For method E10-4 the optimum value of  $\alpha = 0.03$  and for E12-1  $\alpha = 0.038$  were obtained.

The computations were performed using our own program.

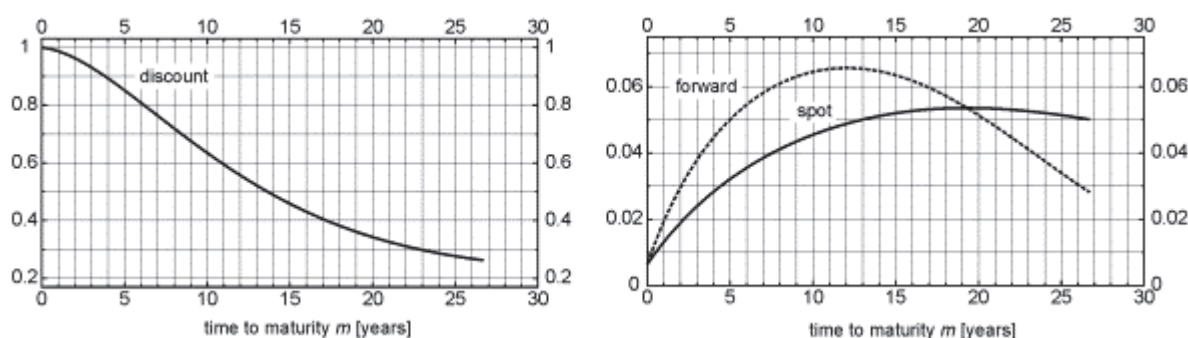


Figure 2 Computed discount function, spot and forward rates vs. time [years], method E10-4.

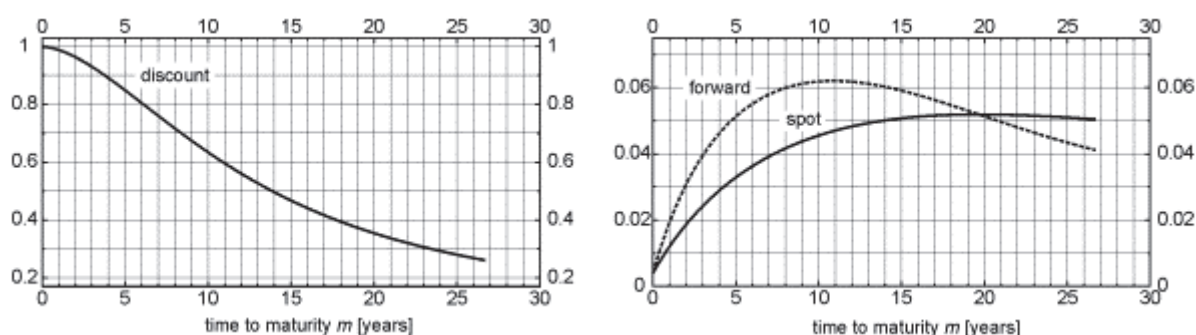


Figure 3 Computed discount function, spot and forward rates vs. time [years], method E12-1.

## 5 Conclusions

Results presented in this paper were based on interest rate estimates from the Czech coupon bond market, which is characterized by a relatively low number of bonds, by moderate liquidity and periodically reduced efficiency. We explored function-based models using the exponential basis functions to create yield curves. This approach produced a reasonably looking spot and forward yield curves. Our attempt to assign weights to each bond reflecting its liquidity was not successful. After substantial experimentation, however, we found the approach employing the exponential basis functions to be a stable and potentially useful. This must be clarified in our subsequent work when compared to other methods (methods using B-splines, Nelson-Siegel method, Svensson method, ...) and on larger set of data than just one day.

## References

- [1] BOLDER, D.J., GUSBA, S.: *Exponentials, Polynomials and Fourier Series: More Yield Curve Modelling at the Bank of Canada*, Bank of Canada Working Paper, no.2002-29, (2002).
- [2] BING-HUEI L.: *B-splines: the case of Taiwanese Government bonds*, Applied Financial Economics, (2002, 12), pp. 57-75
- [3] EILERS, P. H., B. D. MARX : *Flexible Smoothing with B-splines and Penalties*, Statistical Science, 11, (1996) pp.89-102.
- [4] FISHER, M., NYCHKA, D., ZERVOS, D.: *Fitting the Term Structure of Interest Rates with Smoothing Splines*, U.S. Federal Reserve Board Working Paper, (1994).
- [5] HLADÍKOVÁ, H. *Model for the Estimation of Yield Curve Derived From The Czech Coupon Bond Market*. In: *Aplimat*, (2011), p. 949–956..

- [6] HLADÍKOVÁ, H., *Term Structure Modelling Using Fourier Model*. Mundus Symbolicus, 2010 (18), p.33–44.
- [7] HLADÍKOVÁ, H., *Term Structure Modelling Using FNZ Spline Model*. Mundus Symbolicus, 2010 (18), p.33–44.
- [8] LI, B., E. DeWETERING, G. LUCAS, R. BRENNER, and A. SHAPIRO.. *Merrill Lynch Exponential Spline Model*., Merrill Lynch Working Paper,(2001)
- [9] McCULLOCH, J. H.: *Measuring the Term Structure of Interest Rates*, Journal of Business, 44,19{31, (1971).
- [10] MÁLEK, J., *Dynamika úvěrových měr a úrokové deriváty*. Praha:Ekopress (2005).
- [11] MÁLEK, J., Radová, J.-Šterba F., *Konstrukce výnosové křivky pomocí vládních dluhopisů v České republice*, Politická ekonomie 6 (2007), pp792-827.
- [12] MARCINIAK, M., *Yield Curve Estimation at the National Bank of Poland* ,Bank i kredyt (2006).
- [13] SLAVÍK, M., *Odhad časové struktury úrokových sazeb z cen domácích dluhopisů*, Finance a úvěr 11 (2001), pp. 591-606.

**Electronic funds:** [www.cnb.cz](http://www.cnb.cz); [www.bankofcanada.com](http://www.bankofcanada.com), [www.twitter.com](http://www.twitter.com), [www.pse.cz](http://www.pse.cz)

#### **Current address**

**RNDr. Hana Hladíková**

University of Economics,

W. Churchill Sq. 4, 130 67 Prague 3,

Czech Republic

e-mail: [Hana.Hladikova@vse.cz](mailto:Hana.Hladikova@vse.cz)

## A NOTE ON SOME CLASSES OF GENERATED FUZZY IMPLICATIONS

BIBA Vladislav, (CZ), HLINĚNÁ Dana, (CZ)

**Abstract.** A fuzzy implication is a hybrid monotonous extension of the classical implication to the unit interval. There are two well-known families of fuzzy implications— $(S, N)$ -implications and  $R$ -implications. There are also several classes of fuzzy implications generated using appropriate monotonous functions. In this paper we investigate the properties of one interesting class of generated fuzzy implications. Moreover, we study connections between them and families of  $(S, N)$ -implications and  $R$ -implications.

**Key words and phrases.** Fuzzy implication,  $(S, N)$ -implications,  $R$ -implications  
*Mathematics Subject Classification.* Primary 60A05, 08A72; Secondary 28E10.

## 1 Preliminaries

First we recall notations and basic definitions used in the paper. We also briefly mention some important properties and results in order to make this work self-contained. We start with the basic logic connectives.

**Definition 1.1** (see e.g. in [4], Definition 1.1) A decreasing function  $N : [0, 1] \rightarrow [0, 1]$  is called a fuzzy negation if for each  $a, b \in [0, 1]$  it satisfies the following conditions

- (i)  $a < b \Rightarrow N(b) \leq N(a)$ ,
- (ii)  $N(0) = 1, N(1) = 0$ .

**Remark 1.2** A fuzzy negation  $N$  is called strict if  $N$  is strictly decreasing and continuous for arbitrary  $x, y \in [0, 1]$ . In a classical logic we have that  $(A')' = A$ . In multivalued logic this equality is not satisfied for each fuzzy negation. The fuzzy negations with this equality are called involutive negations. The strict fuzzy negation is called strong if and only if it is involutive. A dual fuzzy negation based on a fuzzy negation  $N$  is given by  $N^d(x) = 1 - N(1 - x)$ .

Some examples of strict and/or strong fuzzy negations are included in the following example. More examples of fuzzy negations can be found in [4].

**Example 1.3** *The next functions are fuzzy negations on  $[0, 1]$ .*

- $N_s(a) = 1 - a$  *strong fuzzy negation, standard negation,*
- $N(a) = 1 - a^2$  *strict, but not strong fuzzy negation,*
- $N(a) = \sqrt{1 - a^2}$  *strong fuzzy negation.*

**Definition 1.4** *An increasing mapping  $C : [0, 1]^2 \rightarrow [0, 1]$  is called a fuzzy conjunction if*

1.  $C(x, y) = 0$  whenever  $x = 0$  or  $y = 0$ , and
2.  $C(1, 1) = 1$ .

**Remark 1.5** *Note that the dual operator to a fuzzy conjunction  $C$ , defined by  $D(x, y) = 1 - C(1 - x, 1 - y)$ , is called a fuzzy disjunction.*

Commonly used fuzzy conjunctions in fuzzy logic are the triangular norms.

**Definition 1.6** ([7], Definition 1.1) *A triangular norm (t-norm for short) is a binary operation on the unit interval  $[0, 1]$ , i.e., a function  $T : [0, 1]^2 \rightarrow [0, 1]$  such that for all  $x, y, z \in [0, 1]$ , the following four axioms are satisfied:*

- (T1) Commutativity  $T(x, y) = T(y, x)$ ,
- (T2) Associativity  $T(x, T(y, z)) = T(T(x, y), z)$ ,
- (T3) Monotonicity  $T(x, y) \leq T(x, z)$  whenever  $y \leq z$ ,
- (T4) Boundary Condition  $T(x, 1) = x$ .

**Remark 1.7** *Commonly used fuzzy disjunctions in fuzzy logic are the triangular conorms. A triangular conorm (also called a t-conorm) is a binary operation  $S$  on the unit interval  $[0, 1]$  which, for all  $x, y, z \in [0, 1]$ , satisfies (T1) – (T3) and (S4)  $S(x, 0) = x$ . For more information, see [7].*

Three most common continuous t-norms and their dual t-conorms are:

- *Minimum t-norm*  $T_M(x, y) = \min(x, y)$ ,  
*Maximum t-conorm*  $S_M(x, y) = \max(x, y)$ ,
- *Product t-norm*  $T_P(x, y) = x \cdot y$ ,  
*Probabilistic sum*  $S_P(x, y) = x + y - x \cdot y$ ,
- *Lukasiewicz t-norm*  $T_L(x, y) = \max(0, x + y - 1)$ ,  
*Lukasiewicz t-conorm*  $S_L(x, y) = \min(1, x + y)$ .

In the literature, we can find several different definitions of fuzzy implications. In this paper we will use the following one, which is equivalent with the definition introduced by Fodor and Roubens in [4]. The readers can find more details in [1], [9].

**Definition 1.8** A function  $I : [0, 1]^2 \rightarrow [0, 1]$  is called a fuzzy implication if it satisfies the following conditions:

- (I1)  $I$  is decreasing in its first variable,
- (I2)  $I$  is increasing in its second variable,
- (I3)  $I(1, 0) = 0$ ,  $I(0, 0) = I(1, 1) = 1$ .

We recall definitions of some important properties of fuzzy implications which we will investigate.

**Definition 1.9** A fuzzy implication  $I : [0, 1]^2 \rightarrow [0, 1]$  satisfies:

(NP) the left neutrality property if

$$I(1, y) = y; \quad y \in [0, 1],$$

(EP) the exchange principle if

$$I(x, I(y, z)) = I(y, I(x, z)) \text{ for all } x, y, z \in [0, 1],$$

(IP) the identity principle if

$$I(x, x) = 1; \quad x \in [0, 1],$$

(OP) the ordering property if

$$x \leq y \iff I(x, y) = 1; \quad x, y \in [0, 1],$$

(CP) the contrapositive symmetry with respect to a given fuzzy negation  $N$  if

$$I(x, y) = I(N(y), N(x)); \quad x, y \in [0, 1],$$

(LI) the law of importation with a  $t$ -norm  $T$  if

$$I(T(x, y), z) = I(x, I(y, z)); \quad x, y \in [0, 1],$$

**Definition 1.10** Let  $I : [0, 1]^2 \rightarrow [0, 1]$  be a fuzzy implication. The function  $N_I$  defined by  $N_I(x) = I(x, 0)$  for all  $x \in [0, 1]$ , is called the natural negation of  $I$ .

$(S, N)$ -implications which are based on  $t$ -conorms and fuzzy negations form one of the well-known classes of fuzzy implications.

**Definition 1.11** A function  $I : [0, 1]^2 \rightarrow [0, 1]$  is called an  $(S, N)$ -implication if there exist a  $t$ -conorm  $S$  and a fuzzy negation  $N$  such that

$$I(x, y) = S(N(x), y), \quad x, y \in [0, 1].$$

If  $N$  is a strong negation then  $I$  is called a strong implication.

The following characterization of  $(S, N)$ -implications is from [1].

**Theorem 1.12** (*Baczyński and Jayaram [1], Theorem 5.1*) For a function  $I : [0, 1]^2 \rightarrow [0, 1]$ , the following statements are equivalent:

- $I$  is an  $(S, N)$ -implication generated from some  $t$ -conorm and some continuous (strict, strong) fuzzy negation  $N$ .
- $I$  satisfies (I2), (EP) and  $N_I$  is a continuous (strict, strong) fuzzy negation.

Another way of extending the classical binary implication to the unit interval  $[0, 1]$  is based on the residuation operator with respect to a left-continuous triangular norm  $T$

$$I_T(x, y) = \max\{z \in [0, 1]; T(x, z) \leq y\}.$$

Elements of this class are known as  $R$ -implications. The following characterization of  $R$ -implications is from [4].

**Theorem 1.13** (*Fodor and Roubens [4], Theorem 1.14*) For a function  $I : [0, 1]^2 \rightarrow [0, 1]$ , the following statements are equivalent:

- $I$  is an  $R$ -implication based on some left-continuous  $t$ -norm  $T$ .
- $I$  satisfies (I2), (OP), (EP), and  $I(x, \cdot)$  is a right-continuous for any  $x \in [0, 1]$ .

Our constructions of fuzzy implications will use extensions of the classical inverse of a function. It can be extended as follows.

**Definition 1.14** ([7] Corollary 3.3) Let  $\varphi : [0, 1] \rightarrow [0, \infty]$  be an increasing and non-constant function. The function  $\varphi^{(-1)}$  defined by

$$\varphi^{(-1)}(x) = \sup\{z \in [0, 1]; \varphi(z) < x\}$$

is called the pseudo-inverse of  $\varphi$ , with the convention  $\sup \emptyset = 0$ .

**Definition 1.15** ([7] Corollary 3.3) Let  $f : [0, 1] \rightarrow [0, \infty]$  be a decreasing and non-constant function. The function  $f^{(-1)}$  defined by

$$f^{(-1)}(x) = \sup\{z \in [0, 1]; f(z) > x\}$$

is called the pseudo-inverse of  $f$ , with the convention  $\sup \emptyset = 0$ .

**Lemma 1.16** ([5]) Let  $c$  be a positive real number. Then the pseudo-inverse of a positive multiple of any monotone function  $f : [0, 1] \rightarrow [0, \infty]$  satisfies

$$(c \cdot f)^{(-1)}(x) = f^{(-1)}\left(\frac{x}{c}\right).$$

**Lemma 1.17** [3] Let  $N : [0, 1] \rightarrow [0, 1]$  be a fuzzy negation. Then  $N^{(-1)}$  is a fuzzy negation if and only if

$$N(x) = 0 \quad \Leftrightarrow \quad x = 1. \tag{1}$$

## 2 Generated fuzzy implications

It is well-known that it is possible to generate t-norms from one variable functions. It means it is enough to consider one variable function instead of two variable function. Therefore the question whether something similar is possible in the case of fuzzy implications is very interesting. Yager ([12]) introduced two new classes of fuzzy implications:  $f$ -implications and  $g$ -implications where their generators  $f$  are continuous additive generators of continuous Archimedean t-norms and generators  $g$  are continuous additive generators of continuous Archimedean t-conorms. Smutná ([11]) and Biba, Hliněná ([5]) presented an alternative approach where fuzzy implications are generated using appropriate strictly decreasing or strictly increasing functions and studied basic properties of proposed generated fuzzy implications.

In this section we recall some already known classes of generated fuzzy implications which were proposed in various papers. The first class was introduced in [11] and studied in [5]. These fuzzy implications are based on strictly decreasing functions  $f$ .

**Theorem 2.1** [5] *Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing function such that  $f(1) = 0$ . Then the function  $I_f(x, y) : [0, 1]^2 \rightarrow [0, 1]$  which is given by*

$$I_f(x, y) = \begin{cases} 1 & \text{if } x \leq y, \\ f^{(-1)}(f(y^+) - f(x)) & \text{otherwise,} \end{cases} \quad (2)$$

where  $f(y^+) = \lim_{y \rightarrow y^+} f(y)$  and  $f(1^+) = f(1)$ , is a fuzzy implication.

On the other hand, for strictly increasing functions  $g$ , fuzzy implications  $I^g$  have been introduced in [11].

**Theorem 2.2** [11] *Let  $g : [0, 1] \rightarrow [0, \infty]$  be a strictly increasing function such that  $g(0) = 0$ . Then the function  $I^g(x, y) : [0, 1]^2 \rightarrow [0, 1]$  which is given by*

$$I^g(x, y) = g^{(-1)}(g(1 - x) + g(y)), \quad (3)$$

is a fuzzy implication.

The fuzzy implication  $I^g$  can be generalized. This generalization is based on replacing the standard negation by arbitrary one.

**Theorem 2.3** [11] *Let  $g : [0, 1] \rightarrow [0, \infty]$  be a strictly increasing function such that  $g(0) = 0$  and  $N$  be a fuzzy negation. Then the function  $I_N^g$ :*

$$I_N^g(x, y) = g^{(-1)}(g(N(x)) + g(y)), \quad (4)$$

is a fuzzy implication.

If we compose a strictly decreasing function  $f$  with a fuzzy negation  $N$  then  $g(x) = f(N(x))$  is again an increasing function (though not necessarily strictly increasing). We can apply such a function  $g$  to (4) and have another possibility how to generate fuzzy implications.

**Theorem 2.4** ([3] without proof) Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing function with  $f(1) = 0$  and  $N : [0, 1] \rightarrow [0, 1]$  be a fuzzy negation. Then the function  $I_f^N : [0, 1]^2 \rightarrow [0, 1]$  defined by

$$I_f^N(x, y) = N(f^{(-1)}(f(x) + f(N(y)))) , \quad (5)$$

is a fuzzy implication.

**Proof.** Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing function such that  $f(1) = 0$  and let  $N$  be a fuzzy negation. We will proceed by points of Definition 1.8.

- Let  $x_1 < x_2$ , then  $f(x_1) + f(N(y)) \geq f(x_2) + f(N(y))$ . Pseudo-inverse  $f^{(-1)}$  is a decreasing function (not necessarily strictly decreasing), which means that  $f^{(-1)}(f(x_1) + f(N(y))) \leq f^{(-1)}(f(x_2) + f(N(y)))$ . Since  $N$  is a fuzzy negation, it is a decreasing function and therefore  $I_f^N(x_1, y) \geq I_f^N(x_2, y)$ .
- Let  $y_1 < y_2$ , then  $N(y_1) \geq N(y_2)$  and  $f(x) + f(N(y_1)) \leq f(x) + f(N(y_2))$ . Since  $f^{(-1)}$  is decreasing, it holds that  $f^{(-1)}(f(x) + f(N(y_1))) \geq f^{(-1)}(f(x) + f(N(y_2)))$  and consequently  $I_f^N(x, y_1) \leq I_f^N(x, y_2)$ .

$$\begin{aligned} I_f^N(1, 0) &= N(f^{(-1)}(f(1) + f(N(0)))) = N(f^{(-1)}(0 + 0)) = N(1) = 0, \\ I_f^N(0, 0) &= N(f^{(-1)}(f(0) + f(N(0)))) = N(f^{(-1)}(f(0))) = N(0) = 1, \\ I_f^N(1, 1) &= N(f^{(-1)}(f(1) + f(N(1)))) = N(f^{(-1)}(f(0))) = N(0) = 1. \end{aligned}$$

This concludes the proof.

### 3 Properties of generated fuzzy implications

In this part we investigate the properties of generated fuzzy implications which are mentioned in Theorem 2.4. First we turn our attention to the following examples of fuzzy implication  $I_f^N$ .

**Example 3.1** Let  $f_1(x) = 1 - x$ ,  $f_2(x) = -\ln x$ , and  $N_1(x) = 1 - x$ ,  $N_2(x) = \sqrt{1 - x^2}$ . Then the functions  $f_1^{(-1)}$  and  $f_2^{(-1)}$  are given by  $f_1^{(-1)}(x) = \max(1 - x, 0)$  and  $f_2^{(-1)}(x) = e^{-x}$ . The fuzzy implications  $I_f^N$  are given by

$$I_{f_1}^{N_1}(x, y) = \min(1 - x + y, 1),$$

$$I_{f_2}^{N_1}(x, y) = 1 - x + x \cdot y,$$

$$I_{f_2}^{N_2}(x, y) = \sqrt{1 - x^2 + x^2 \cdot y^2}.$$

Note, that  $I_{f_1}^{N_1}$  and  $I_{f_2}^{N_1}$  are the well-known Lukasiewicz and Reichenbach implication, respectively. Also note, that for all fuzzy implications it holds that  $I(x, 0) = N(x)$ .

We are able to generalize the property from Example 3.1 for all  $I_f^N$  implications and  $N_{I_f^N}(x)$  negations.

**Proposition 3.2** *Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing function such that  $f(1) = 0$  and  $N$  be an arbitrary fuzzy negation. Then the natural negation  $N_I$  given by  $I_f^N$  is  $N_{I_f^N}(x) = N(x)$ .*

**Proof.** Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing function such that  $f(1) = 0$  and  $N$  be an arbitrary fuzzy negation. For  $N_{I_f^N}(x)$  we have

$$N_{I_f^N}(x) = I_f^N(x, 0) = N(f^{(-1)}(f(x) + f(N(0)))) = N(f^{(-1)}(f(x) + f(1))) = N(f^{(-1)}(f(x))).$$

Since the function  $f$  is strictly decreasing, its pseudo-inverse is continuous, and therefore  $f^{(-1)} \circ f(x) = x$ . And for natural negation we get

$$N_{I_f^N}(x) = N(f^{(-1)}(f(x))) = N(x).$$

The above mentioned property of a strictly decreasing function and its pseudo-inverse is again important for fulfilment of (NP). Therefore the proof is similar to the previous and we can omit it.

**Proposition 3.3** *Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing function such that  $f(1) = 0$ . Then the fuzzy implication  $I_f^N$  satisfies (NP) if and only if  $N$  is an involutive negation.*

**Proposition 3.4** *Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing function such that  $f(1) = 0$ . Then the fuzzy implication  $I_f^N$  satisfies (CP) with respect to  $N$  if and only if  $N$  is an involutive negation.*

**Proof.** Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing function such that  $f(1) = 0$ . For fuzzy implications  $I_f^N(x, y)$  and  $I_f^N(N(y), N(x))$  we get

$$I_f^N(x, y) = N(f^{(-1)}(f(x) + f(N(y)))) ,$$

$$I_f^N(N(y), N(x)) = N(f^{(-1)}(f(N(y)) + f(N(N(x))))) .$$

It is obvious that  $I_f^N(x, y) = I_f^N(N(y), N(x))$  if and only if  $N(N(x)) = x$  for all  $x \in [0, 1]$ .

**Proposition 3.5** *Let  $f : [0, 1] \rightarrow [0, \infty]$  be a continuous and bounded strictly decreasing function such that  $f(1) = 0$  and  $N(x) = f^{-1}(f(0) - f(x))$ . Then the fuzzy implication  $I_f^N$  satisfies (OP).*

**Proof.** Let  $f : [0, 1] \rightarrow [0, c]$  be a function described in the proposition, and  $c$  be a positive real number, it is obvious that  $N(x) = f^{-1}(f(0) - f(x))$  is fuzzy negation. Since  $f$  is a strictly decreasing and continuous function, it holds

$$I_f^N(x, y) = N(f^{(-1)}(f(x) + f(N(y)))) = N(f^{(-1)}(f(0) + f(x) - f(y))) .$$

Now we need to distinguish two cases:

- Let  $x \leq y$ , then  $f(x) - f(y) \geq 0$  and  $f(0) + f(x) - f(y) \geq f(0)$ , i.e

$$I_f^N(x, y) = N(f^{(-1)}(f(0))) = N(0) = 1.$$

- Let  $x > y$ , then  $f(0) + f(x) - f(y) < f(0)$  and consequently  $f^{(-1)}(f(0) + f(x) - f(y)) > 0$ , i.e

$$I_f^N(x, y) < N(0) = 1.$$

Summarizing the previous facts we get that  $I_f^N(x, y) = 1$  if and only if  $x \leq y$ .

**Remark 3.6** Let  $f : [0, 1] \rightarrow [0, \infty]$  be a not bounded function. Then the fuzzy implication  $I_f^N$  does not hold (OP). This follows from the fact that for all  $x, y \in ]0, 1[$  we get  $f(x) + f(N(y)) < f(0)$  and consequently  $I_f^N(x, y) < 1$ .

**Proposition 3.7** Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing and continuous function such that  $f(1) = 0$ . Let  $N : [0, 1] \rightarrow [0, 1]$  be a strong negation. Then the fuzzy implication  $I_f^N$  satisfies (EP).

**Proof.** Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing continuous function such that  $f(1) = 0$  and  $N$  be a strong negation. Then

$$I_f^N(x, I_f^N(y, z)) = I_f^N(x, N(f^{(-1)}(f(y) + f(N(z))))).$$

Since  $f$  is a strictly decreasing and continuous function, the following equality is satisfied

$$f^{(-1)}(f(y) + f(N(z))) = \begin{cases} 0 & f(y) + f(N(z)) \geq f(0), \\ f^{-1}(f(y) + f(N(z))) & \text{otherwise.} \end{cases}$$

Now we apply the fact that  $N$  is a strong negation and we get

$$I_f^N(x, I_f^N(y, z)) = \begin{cases} N(f^{(-1)}(f(x) + f(0))) & \text{if } f(y) + f(N(z)) \geq f(0), \\ N(f^{(-1)}(f(x) + f(y) + f(N(z)))) & \text{otherwise.} \end{cases}$$

And for  $I_f^N(y, I_f^N(x, z))$  we have

$$I_f^N(y, I_f^N(x, z)) = \begin{cases} N(f^{(-1)}(f(y) + f(0))) & \text{if } f(x) + f(N(z)) \geq f(0), \\ N(f^{(-1)}(f(x) + f(y) + f(N(z)))) & \text{otherwise.} \end{cases}$$

Since  $N(f^{(-1)}(f(x) + f(0))) = N(f^{(-1)}(f(y) + f(0))) = 1$ , we can write

$$I_f^N(x, I_f^N(y, z)) = \begin{cases} 1 & \text{if } f(y) + f(N(z)) \geq f(0), \\ N(f^{(-1)}(f(x) + f(y) + f(N(z)))) & \text{otherwise.} \end{cases}$$

$$I_f^N(y, I_f^N(x, z)) = \begin{cases} 1 & \text{if } f(x) + f(N(z)) \geq f(0), \\ N(f^{(-1)}(f(x) + f(y) + f(N(z)))) & \text{otherwise.} \end{cases}$$

If  $f(y) + f(N(z)) \geq f(0)$ , then also  $f(x) + f(y) + f(N(z)) \geq f(0)$ , which means that  $I_f^N(y, I_f^N(x, z)) = 1$ . And, on the contrary, if  $f(x) + f(N(z)) \geq f(0)$ , then  $I_f^N(x, I_f^N(y, z)) = 1$ .

The following theorem describes the relationship between the generated fuzzy implications  $I_N^f$  and  $(S, N)$ - or  $R$ -implications.

**Theorem 3.8** Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing and continuous function such that  $f(1) = 0$ . Let  $N : [0, 1] \rightarrow [0, 1]$  be a strong negation. Then  $I_f^N$  is  $(S, N)$ -implication. Moreover, if  $f$  is bounded function and  $N(x) = f^{-1}(f(0) - f(x))$ , then  $I_f^N$  is an  $R$ -implication as well.

Some relation between these generated implications and t-norms is described in the next proposition.

**Proposition 3.9** Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing and continuous function such that  $f(1) = 0$ . Let  $N : [0, 1] \rightarrow [0, 1]$  be a strong negation. Then the fuzzy implication  $I_f^N$  satisfies (LI) with a t-norm  $T(x, y) = f^{(-1)}(f(x) + f(y))$ .

**Proof.** Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing continuous function such that  $f(1) = 0$ ,  $N$  be a strong negation, and  $T : [0, 1]^2 \rightarrow [0, 1]$  be a t-norm given by  $T(x, y) = f^{(-1)}(f(x) + f(y))$ . Then

$$I_f^N(T(x, y), z) = \begin{cases} N(f^{(-1)}(f(0) + f(N(z)))) & \text{if } f(x) + f(y) \geq f(0), \\ N(f^{(-1)}(f(x) + f(y) + f(N(z)))) & \text{otherwise,} \end{cases}$$

and from the previous proof we get for  $I_f^N(x, I_f^N(y, z))$  the following formula

$$I_f^N(x, I_f^N(y, z)) = \begin{cases} 1 & \text{if } f(y) + f(N(z)) \geq f(0), \\ N(f^{(-1)}(f(x) + f(y) + f(N(z)))) & \text{otherwise.} \end{cases}$$

It is obvious that  $N(f^{(-1)}(f(0) + f(N(z)))) = 1$  and by similar method as we have used in previous proof we get that  $I_f^N(T(x, y), z) = I_f^N(x, I_f^N(y, z))$ .

It is well known that generators of continuous Archimedean t-norms are unique up to a positive multiplicative constant, and this is also valid for the  $f$  generators of  $I_N^f$  implications. This follows from Lemma 1.16.

**Proposition 3.10** Let  $c$  be a positive constant and  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing function such that  $f(1) = 0$ . Then the implications  $I_N^f$  and  $I_N^{c \cdot f}$  which are based on functions  $f$  and  $c \cdot f$ , respectively, are identical.

## 4 Conclusions

Note, that mentioned fuzzy implications are not the only generalizations of fuzzy implications  $I_N^g$ . Considering Formula (5) and Lemma 1.17, we can see that  $N$  might be replaced by  $N^{(-1)}$  if it is a fuzzy negation. Still, there are at least two fuzzy negations (in general different from  $N$ ) which are related to  $N$ . Namely,  $N^{(-1)}$  and  $N^d$ . Hence we have the following two additional possibilities how to generate fuzzy implications.

**Theorem 4.1** ([3]) Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing function with  $f(1) = 0$ , and  $N : [0, 1] \rightarrow [0, 1]$  be a fuzzy negation such that (1) is fulfilled for  $N$ . Then the function  $I_f^{(N, N^{(-1)})} : [0, 1]^2 \rightarrow [0, 1]$  defined by

$$I_f^{(N, N^{(-1)})}(x, y) = N^{(-1)}(f^{(-1)}(f(x) + f(N(y)))) \quad (6)$$

is a fuzzy implication.

**Theorem 4.2** ([3]) Let  $f : [0, 1] \rightarrow [0, \infty]$  be a strictly decreasing function with  $f(1) = 0$  and  $N : [0, 1] \rightarrow [0, 1]$  be a fuzzy negation. Then function  $I_f^{(N, N^d)} : [0, 1]^2 \rightarrow [0, 1]$  defined by

$$I_f^{(N, N^d)}(x, y) = N^d \left( f^{(-1)} (f(x) + f(N(y))) \right), \quad (7)$$

is a fuzzy implication.

### Acknowledgement

Dana Hliněná has been supported by Project MSM0021630529 of the Ministry of Education and project FEKT-S-11-2(921).

### References

- [1] BACZYŃSKI, M., BALASUBRAMANIAM, J.: *Fuzzy implications* (Studies in Fuzziness and Soft Computing, Vol. 231), Springer, Berlin (2008)
- [2] BIBA, V., HLINĚNÁ, D., KALINA, M., KRÁL', P.: *Generated fuzzy implications and known classes of implications* (to appear)
- [3] BIBA, V., HLINĚNÁ, D., KALINA, M., KRÁL', P.: *I-fuzzy equivalences and I-partitions* 17th East-West Fuzzy Colloquium Conference Proceedings, Zittau 2010
- [4] FODOR, J. C., ROUBENS, M.: *Fuzzy Preference Modeling and Multicriteria Decision Support* Kluwer Academic Publishers, Dordrecht 1994.
- [5] HLINĚNÁ, D., BIBA, V.: *Generated fuzzy implications and known classes of implications* In: Acta Universitatis Matthiae Belii ser. Mathematics. 1. Banská Bystrica: Matej Bel University, 2010, 25–34.
- [6] KLEMENT, E. P., MESIAR, R.: *Logical, algebraic, analytic and probabilistic aspects of triangular norms*, Elsevier, Amsterdam 2005, ISBN 0-444-51814-2
- [7] KLEMENT, E. P., MESIAR, R., PAP, E.: *Triangular Norms* Kluwer, Dordrecht 2000.
- [8] MASSANET, S., TORRENS, J.: *Fuzzy implications and Weak law of importation* IFSA-EUSFLAT, 2009
- [9] MAS, M., MONSERRAT, M., TORRENS, J., TRILLAS, E.: *A survey on fuzzy implication functions* IEEE Transactions on Fuzzy Systems 15 (6) 1107–1121, (2007).
- [10] SCHWEIZER, B., SKLAR, A.: *Probabilistic Metric Spaces* North Holland, New York 1983
- [11] SMUTNÁ, D.: *On many valued conjunctions and implications* Journal of Electrical Engineering, 10/s, vol. 50, 1999, 8–10
- [12] YAGER, R. R.: On some new classes of implication operators and their role in approximate reasoning. *Information Sciences*, 167(1-4):193 – 216, 2004.

### Current address

#### Hliněná Dana, RNDr., PhD

Department of Mathematics Faculty of Electrical Engineering and Communication  
 Brno University of Technology  
 Technick 8, Brno, Czech Republic  
 e-mail: hlinena@feec.vutbr.cz

## **ANTI-COMMONS: FISHERIES PROBLEMS AND BUREAUCRACY IN AQUACULTURE**

**FILIPE, José António (P), FERREIRA, Manuel Alberto M. (P),  
COELHO, M. (P), PEDRO, Maria Isabel (P)**

**Abstract:** Anti-Commons and bureaucracy have been linked since the study of Buchanan & Yoon (2000). Bureaucracy involves a set of agents that have a deciding power. Conflicting interests, the decision makers inertia or the inertia of the system itself, excessive administrative procedures or excessive administrative circuits push too late decisions, or for non-rational decisions in terms of value creation for economic agents. Property Rights Theory explains new concerns. Considering that an “anti-commons” problem arises when there are multiple rights to exclude, the problem of decision process in aquaculture projects makes sense at this level. However, little attention has been given to the setting where more than one person is assigned with exclusion rights, which may be exercised. “Anti-commons” problem is analyzed in situations in which resources are inefficiently under-utilized rather than over-utilized as in the familiar commons setting. In this study, fisheries problems are studied and some ways to deal with the problem are presented.

**Keywords:** Anti-Commons Theory, Property Rights, Fisheries

### **Introduction**

“Anti-Commons” theory has appeared representing the idea of an excessive partition of property rights. This theory has appeared in the 80’s of last century, introduced by Michelman (1982). In the last years of the 20<sup>th</sup> Century several ideas about this new problem around property rights have emerged in which too many rights of exclusion and a reduced level of utilization of the resource are present. Michelman (1982), when presented the notion of “anti-commons”, defined it as “a type of property in which everyone always has rights respecting the objects in the regime, and no one, consequently, is ever privileged to use any of them except as particularly authorized by others”. This definition of “anti-commons” makes evidence of the lack of efficiency in several situations in which each one of several owners with property rights over a given resource has no effective rights to use the resource (and consequently, each one has the right to exclude other agents from the utilization of the resource).

**Property rights discussion. The underuse of resources under a situation of “anticommons”**

The discussion involving the definition of property rights is old. The types of property rights demand that the limits of these concepts are hardly analyzed. The commons problems are discussed since the middle of last century, involving the idea that commons reflect usually the overexploitation of resources. The lack of property rights implies that no one may exclude others to access to a given resource. The existence of many agents to use a given resource, in these conditions, causes an inefficient level for the resource use and causes a special motivation for agents over-use the resource. The real level of use for the resource will take place at a higher level compared with the optimal level for the society as a whole.

On the opposite side, when several owners of a resource have, each one, the right to exclude others from the use of a scarce resource and no one has the full privilege to use it, this resource may have a very limited and unsatisfactory use. This is the problem of the “tragedy of the anti-commons”: the resource may be prone to under-use.

After an “anti-commons” emerge, its particular passage to an efficient process in the private property sphere is long and extremely slow, due to the properties inherent to “anti-commons” and to the difficulties existing for overcoming the “tragedy of anti-commons”.

As a consequence of all this, it is necessary to make an important reflection about the definition of property rights to overcome several important aspects when resources are exploited. Indeed, we can see that not just the commons lead to the tragedy. When there are too many property rights and too many rights of exclusion, tragedy seems to be the last result, as well. Too many owners have the right to exclude others but, in fact, no one of them, have the privilege to use it suitably. An insufficient use is the corollary for this situation.

### **An “anti-commons” view**

Along this section, it will be taken as a fundamental basis, the work of Schulz, Parisi and Depoorter (2003). They present a general model which permits to distinguish between the simultaneous cases and the sequential cases in “anti-commons” tragedies. As the authors say, the reality may present situations that combine characteristics of the two categories. Anyway, it is important to consider the two situations separately.

For the first of these two cases, they consider that exclusion rights are exercised at the same time, independently. This involves several agents linked in a coincident relationship, such as multiple co-owners with cross-veto powers on the other members’ use of the common resource.

In the sequential case, exclusion rights are exercised in consecutive stages, at different levels of the value chain. The several owners of the exclusion rights exercise their own rights in a succession way. Each agent may be involved in a hierarchy and each one may exercise its own exclusion right or veto power over a given proposition (see some examples about simultaneous and sequential anti-commons tragedies in Schulz, Parisi and Depoorter, 2003).

In that work, a dual model of property, where commons and anticommons problems are the consequence of a lack of conformity between use and exclusion rights, is extended to consider the different equilibria obtained under vertical and horizontal cases of property fragmentation.

Horizontal anticommons cases are the ones where situations of exclusion rights exercised simultaneously and independently are present. This involves situations in which are two agents linked in a horizontal relationship. Both agents contribute on the same level of a value chain.

Situations of a vertical anticommons can be expressed as the situations in which the exclusion rights are exercised sequentially by the various right holders. This involves multiple parties in a hierarchy

each of whom can exercise an exclusion or veto power over a given proposition. Examples involving bureaucracy (for example, situations where multiple permits need to be acquired in order to exercise a given activity) or a production process where a given producer purchases one essential input from a monopolistic seller (see Parisi, Schulz and Depoorter, 2005).

It is shown in Parisi, Schulz and Depoorter (2005) that the symmetrical features of commons and anticommons cases result from a misalignment of the private and social incentives of multiple owners in the use of a common resource. The misalignment is due to externalities not captured in the calculus of interests of the users (commons situations) and excluders (anticommons situations).

The problem of anti-commons is based on a positive externality, when considering it in terms of efficiency<sup>1</sup>.

In anti-commons case  $x_i$ , the activities of agents, shows the extent to which agent  $i$  grants agent  $j$  permission to use the common property. An activity  $x_1$  of agent 1 exerts a positive impact on the productivity of agent 2's activity  $x_2$ .

If  $V_i(x_i, x_j)$  is the value of the common resource of agent  $i$ , agent  $i$  grants agent  $j$  the right to use the common resource. Agent  $j$  owns a complementary right to exclude agent  $i$  from the use of the common resource. These grants are respectively  $x_1$  and  $x_2$ . So,  $V_i(x_i, x_j)$  denotes the profit agent  $i$  takes from this joint project. The positive externality that agent  $j$  exerts on agent  $i$  is given by:

$$\frac{\partial V_i}{\partial x_j}(x_i, x_j) > 0$$

Assuming that the exclusion rights are exercised simultaneously and independently by the various rights' holders, multiple owners exercise their veto power on equal terms. If both agents are in a perfect symmetric situation,

$$V_i(x_i, x_j) = V_j(x_j, x_i),$$

agent  $i$  will be choosing the value of  $x_i$  which maximizes  $V_i(x_i, x_j)$ ; and the resulting Nash equilibrium considering agents 1 and 2 is given by

$$\frac{\partial V_1}{\partial x_1}(x_1, x_2) = 0 \text{ and } \frac{\partial V_2}{\partial x_2}(x_2, x_1) = 0.$$

These two conditions are the best response functions of the two agents.

Assuming also that  $V_i$  is concave in  $x_i$ , the equilibrium exists for mild assumptions on activities  $x_i$ . Given the symmetry assumption, a symmetric equilibrium is expected:

$$x_1 = x^e = x_2.$$

There are uncoordinated choices for the 2 agents which can now be compared to the efficient choices of  $x_i$ , which maximize  $V_1 + V_2$  and are characterized by the following first order conditions:

$$\frac{\partial V_1}{\partial x_1}(x_1, x_2) + \frac{\partial V_2}{\partial x_1}(x_2, x_1) = 0 \text{ and } \frac{\partial V_2}{\partial x_2}(x_2, x_1) + \frac{\partial V_1}{\partial x_2}(x_1, x_2) = 0.$$

Given the symmetric assumption, a symmetric optimum is expected. Assuming  $V_1 + V_2$  concave and that there is a symmetric solution, the efficient choices are equal:

$$x_1 = x^s = x_2.$$

<sup>1</sup> Commons problem is based on a negative externality.

As can be easily concluded,  $x^s > x^c$ . This is, the uncoordinated choices lead to the underutilization of the common resource.

If horizontal and vertical anticommons are introduced, both being obviously the consequence of the existence of non-conformity between use and exclusion rights, the problem of underutilization of resources is exacerbated if the right is fragmented into more than two exclusion rights, with more than two agents deciding independently on their activity or prize (see Schulz, 2000 and Parisi, Schulz and Depoorter (2005)).

### **Tragedy of Fishing Commons. Looking for Solutions**

The classical theory of the “tragedy of the commons” explains the reasons why sea fisheries are prone to over-exploitation (see Hardin, 1968). Gordon (1954) had already examined the problems of common resources in the 50s.

The emergence of individual transferable quotas (ITQs) would potentially allow overcoming, even partially, the serious problem of over-fishing for several species in several parts of the world. Anyway, the enormous enthusiasm around ITQs has given place to the appearance of the problem of the “anti-commons”.

Some nations have tried to avoid the “tragedy of the commons” in sea fisheries by regulating the activity, decade over decade. Reducing fishing seasons, restrict open areas to fisheries, limiting the use of gears or reducing the power and tons for ships were just some measures to avoid tragedies to species. The truth is that these practices many times did not reduce over-fishing for the species that governments intended to protect. Fishing race and massive discards, frequently, continued.

TACs (total allowable catches) are fixed each year by fishing management authorities. Each fisher has a part of the TAC to fish, representing his own individual quota. Theoretically, with this procedure, each fisher may use his quota when he likes and fishing race may come to an end (see Leal, 2002a). Meanwhile, quotas may be transferable and this procedure may lead to situations in which the quotas owners can adjust the dimension of his fishing operations buying or selling quotas or even just leaving the fishery and move the quota from the market.

Many nations have been using this kind of measures (programs of individual transferable quotas) to manage fishing resources in their waters. These programs contributed to improve fishers' rents, to improve the quality of products, to reduce the excess of quotas and to eliminate eventual catches that exceed TACs (see Alesi, 1998; see Repetto, 2001 and Wilen and Homans, 2000, as well).

### **The emergence of anti-commons. Alaska's Halibut example**

Alaska's Halibut allows to study the effect of the existence of ITQs and, additionally, to study the consequences in terms of the “anti-commons” (see Leal, 2002b). In fact, this specie got overexploited and authorities implemented several measures to reduce catches. First, fishing seasons were reduced. At the beginning of 90s, fisheries were opened just for two or three short periods of about 24 hours per year. Consequently, fishing race became the solution for fishers, who tried to get the maximum fish as possible, throughout the available time for fishing. In fact, results got different than the expected ones for worse. However, after the implementation of individual quotas in 1995, fishing seasons became larger and fishers could exploit this resource for around 8 months, per year. Sales increased and prices got higher (see GAO, 2002). Meanwhile, catches got smaller than TACs and fleets excesses were reduced.

Nevertheless, individual quotas excess may lead to sub-exploitation of the resource and Alaska's Halibut is, in fact, a case that must be studied to see the consequences of too many existing fishing quotas. Authorities have implemented rules to protect small fishers. They did not authorized fishers to sell their quotas if they were very small. Consequently, these quotas got unexploited, because they were not profitable for their owners. Halibut got underexploited. Authorities had to change rules for this fishery. It can be seen now that ITQ promote important solutions for the "tragedy of the commons" but they may create conditions for a new situation of "anti-commons".

Some other examples may be presented. For example, Leal (2002b) shows how Alaska's crab may be prone to under-use if fishers are forced to sell their catches to a little number of companies, as it was the case. Low prices lead to situations in which fishers under-use their quotas because they got unprofitable. As a consequence crab got under-exploited.

### **The aquaculture case in Portugal**

The example of aquaculture in Portugal can be presented. There are too many entities to analyze projects of aquaculture. Rules and procedures are so many that projects are approved with long delays<sup>2</sup>. As a consequence, resources get under-used.

The aquaculture sector in Portugal is studied and allows to evaluate the possibility of using the hypothesis suggested by Buchanan & Yoon (2000) that bureaucracy can be studied with the help of the anti-commons conceptualization.

In this context, some questions are posed about live resources exploitation, particularly in fisheries and aquaculture projects and raised the legal problems in the Portuguese case. An economic analysis allows to show how this problem of anti-commons can originate an important loss of value. It is seen as anti-commons tragedies appear in such situations in the aquaculture problem.

The suggestion of Buchanan and Yoon (2000) that the anti-commons construction offers an analytical means of isolating a central feature of "sometimes disparate institutional structures" shows the problems arisen from bureaucracy in this context. The persistence of bureaucratic circuits of approval and implementation of projects can difficult the entrepreneurship activities and it diminishes the potential of regional and coastal development.

The responsible Department for Aquaculture is specifically DGPA (Direcção Geral das Pescas e Aquicultura), which is responsible for supervising and controlling the activity of aquaculture sector<sup>3</sup>. There are an enormous set of initial steps for a project's approval (Decreto Regulamentar nº 14/2000) and there are many entities deciding (see Filipe *et al*, 2011a). This leads to projects rejection or a very delayed approval. Besides it shows how bureaucracy is involved in worsening the conditions of exploitation of the resources. The "disparate institutional structures" get evident and a problem of anti-commons is the obvious result.

---

<sup>2</sup> Other countries have similar problems in this kind of projects.

<sup>3</sup> The aquaculture problem is fitted under the control and supervision of Ministério da Agricultura, do Desenvolvimento Rural e das Pescas (see Decreto Regulamentar nº 14/2000 – September, 21<sup>st</sup>, 2000), that is the Ministry for Agriculture, Fishing and Aquaculture Sector. This Decree specifies the requisites and conditions needed to install and exploit a plant on this area. The Decreto Regulamentar nº 9/2008 (March, 18<sup>th</sup>, 2008) defines a set of rules specifically for installations offshore.

## Concluding Remarks

The anti-commons are a very interesting issue to be studied, with an important scope of analysis. It is shown in this study how anti-commons are associated to negative externalities and to a underuse of resources.

Some examples in wild fisheries are presented; and the aquaculture case for projects in Portugal is studied. Aquaculture contributes for fish production and being, as often they are, ecologically sustainable, projects in this area will contribute for solving fisheries' dilemmas about sea fishing resources exploitation.

In Portugal an excessive number of regulators (some of them with veto power) analyze the projects. They spent too much time to overpass all the steps and when the process is ready for implementation it may be too late (and sometimes, the project is refused).

Too many resources are spent on projects and they simply get unviable. A project may create value for the investor and for the community but all the time wasted in bureaucratic analysis makes the project unviable.

## References

- [1] BUCHANAN, J. M. & YOON, Y. J. (2000). Symmetric tragedies; commons and anticommons. *Journal of Law and Economics* 43.
- [2] COUNCIL, N. P. F. M. (2002). Summary of the Bering Sea and Aleutian Islands crab rationalization program, *A Report by the North Pacific Fishery management Council, Anchorage, Alaska*. [www.fakr.noaa.gov/npfmc](http://www.fakr.noaa.gov/npfmc), 8/07/2005.
- [3] COURNOT (1927). *Researches into the mathematical principles of the theory of Wealth (1838)*. Translated by Nathaniel Bacon. New York: Mcmillan.
- [4] de ALESI, M. (1998). *Fishing for Solutions*, London: Institute of Economics Affairs.
- [5] DEPOORTER, B. & PARISI, F. (2000). Commodification in property law: anticommons fragmentation in servitude law. *CASLE 5*. Working Paper Series.
- [6] DINNEFORD, E., IVERSON, K., MUSE, B. & SCHELLE, K. (1999). Changes under Alaska's Halibut IFQ program, 1995 to 1998. *A Report Published by the Alaska Commercial Fisheries Entry Commission*. <http://www.cfec.state.ak.us/research/h98ts/h98ts.htm>, 8/07/2005.
- [7] EISENBERG, R. (1989). Patents and the progress of science: exclusive rights and experimental use. *University of Chicago Law Review* 56.
- [8] EISENBERG, R. (2001). Bargaining over the transfer of proprietary research tools: is this market failing or emerging?. In R. D. et al (Eds), *Expanding the bounds of intellectual property: innovation policy for the knowledge society*. New York: Oxford University Press.
- [9] FILIPE, J. A., FERREIRA, M. A. M. e COELHO, M. (2011a), "Utilization of Resources: An ethical issue. The anti-commons and the aquaculture case" in Costa, G. J. M. (Ed.), *Ethical Issues and Social Dilemmas in Knowledge Management: Organizational Innovation*, Hershey, USA: IGI Global.
- [10] FILIPE, J. A., FERREIRA, M. A. M., COELHO, M. e YORDANOVA, D. (2011b) "Aquaculture Procedures in Portugal and Bulgaria. Anti-commons and Bureaucracy". *International Journal of Academic Research*, 3
- [11] GAO (2002). *Individual Fishing Quotas: Better Information Could Improve Program Management*. General Accounting Office. GAO-03-159, Washington, DC.

- [12] GORDON, H. S. (1954). The economic theory of a common property resource: the fishery. *Journal of Political Economy* 62.
- [13] GREER, L. A. & Bjornstad, D. J. (2004). Licensing complementary patents, the anticommons and public policy. Technical report. Joint Institute for Energy and Environment.
- [14] HARDIN, G. (1968). The tragedy of the commons. *Science* 162.
- [15] HELLER, M. A. (1998). The tragedy of the anticommons: property in the transition from Marx to markets. *Harvard Law Review* 111.
- [16] HELLER, M. A. (1999). The boundaries of private property, *Yale Law Review* 108.
- [17] HELLER, M. & EISENBERG, R. (1998). Can patents deter innovation? The anticommons in biomedical research. *Science* 280.
- [18] KAMPARI, S. (2004). *Tragedy of digital anti-commons*. S-38.042 Seminar on Networking Business. Helsinki University of Technology, Networking Laboratory.
- [19] LEAL, D. R. (2002a). *Fencing the fishery: a primer on ending the race for fish*. PERC REPORTS in [http://www.perc.org/pdf/guide\\_fish.pdf](http://www.perc.org/pdf/guide_fish.pdf).
- [20] LEAL, D. R. (2002b). *A new fishing tragedy? The "anticommons" leads to underuse*. PERC REPORTS in <http://www.perc.org/publications/percreports/sept2004/fishing.php>.
- [21] MATULICH, S. C., MITTELHAMMER, R. C. & ROBERTE, C. (1996). Toward a more complete model of individual transferable fishing quotas: Implications of incorporating the processing sector. *Journal of Environmental Economics and Management* 31.
- [22] MICHELMAN, F. I. (1982). Ethics, economics and the law of property. In J. R. Pennock & J. W. Chapman (Eds), *Nomos XXIV: Ethics, Economics and the Law*. New York: New York University Press.
- [23] PARISI, F., SCHULZ, N. & DEPOORTER, B. (2003). Simultaneous and Sequential Anticommons. George Mason University of Virginia, School of Law. *Law and Economics Working Paper Series 03-11*
- [24] PARISI, F., SCHULZ, N. & DEPOORTER, B. (2005). Duality in Property: Commons and Anticommons. *International Review of Law and Economics*, 25, 578-591
- [25] REPETTO, R. (2001). A natural experiment in fisheries management regimes. *Marine Policy* 25.
- [26] SCHULZ, N. (2000). Thoughts on the Nature of Vetoes When Bargaining on Public Projects, *Wurzburg Economic Papers*, 00-17.
- [27] SCHULZ, N., PARISI, F. & DEPOORTER, B. (2002). Fragmentation in Property: Towards a General Model. *Journal of Institutional and Theoretical Economics* 158.
- [28] STEWART, S. & BJORNSTAD, D. J. (2002). An experimental investigation of predictions and symmetries in the tragedies of the commons and anticommons. Technical report. Joint Institute for Energy and Environment.
- [29] WILEN, J. E. & HOMANS, F. R. (2000). *Unraveling rent losses in modern fisheries: Production, market, or regulatory inefficiencies?* Paper presented at Western Economics Association 74th International Conference, Vancouver, BC.

**Current address**

**José António Filipe, Professor Auxiliar**

ISCTE - IUL – Lisbon University Institute

UNIDE – IUL

Av. Forças Armadas 1649-026 Lisboa, Portugal

Tel.+351 217 903 000

E-mail: [jose.filipe@iscte.pt](mailto:jose.filipe@iscte.pt)

**Manuel Alberto M. Ferreira, Professor Catedrático**

ISCTE – IUL – Lisbon University Institute

UNIDE – IUL

Av. Forças Armadas 1649-026 Lisboa, Portugal

Tel. +351 217 903 000

E-mail: [manuel.ferreira@iscte.pt](mailto:manuel.ferreira@iscte.pt)

**Manuel Coelho, Professor Auxiliar**

SOCIUS/ISEG - Portugal

Rua do Quelhas, 6, 1200-781

LISBOA, Portugal

Phone: +(351) 213925800.

Fax: +(351) 213922808.

Email: [coelho@iseg.utl.pt](mailto:coelho@iseg.utl.pt)

**Maria Isabel Pedro, Professor Auxiliar**

CEGIST/IST - Portugal

Av. Rovisco Pais

LISBOA, Portugal

Phone: +(351) 214233267.

Fax: +(351) 218417979.

Email: [ipedro@ist.utl.pt](mailto:ipedro@ist.utl.pt)

## SOLUTION OF TORSION OF PRISMATIC BAR USING PROGRAM *MATHEMATICA* FOR ELLIPTICAL CROSS-SECTION AREA

JANČO Roland, (SK), KOVÁČOVÁ Monika, (SK)

**Abstract.** In real design of bar and beam which is load by torque we need properties of cross section area. No all time you have circular cross section area in real problems. For solution of non-circular cross section area we used Saint-Venant's principle. In this paper is short introduction how to used Saint-Venant's principle to solution of elliptical cross section area. Theoretical solutions for elliptical cross section are compared by numerical solution solved in program Mathematica with package Structural Mechanics.

**Key words and phrases.** Torsion, Saint-Venant's principle, elliptical cross section.

*Mathematics Subject Classification.* Primary 74A10, 74B05, 74G50 ; Secondary 74K10.

### 1 Introduction

Because many engineering structures, such as beams, shafts, and airplane wings, are subjected to torsional forces, the torsional problem has been of practical importance in structural analysis for a long time. Saint-Venant (1885) was the first to provide the correct solution to the problem of torsion of bars subjected to moment couples at the ends. He made certain assumptions about the deformation of the twisted bar, and then showed that his solutions satisfied the equations of equilibrium and the boundary conditions. From the uniqueness of solutions of the elasticity equations, it follows that the assumed forms for the displacements are the exact solutions to the torsional problem. The Saint-Venant principle is adopted in **Structural Mechanics** packages.

In this paper is contains of theoretical background of solution elliptical cross-section area properties for torsional problems using Saint-Venant principle and comparison of theoretical solution with solution from **Structural Mechanics** package in *MATHEMATICA*.

## 2 Theoretical background

If bar is loaded by equal and opposite torques  $T$  on its ends, we anticipate that the relative rigid-body displacement of initially plane section will consist of rotation, leading to a twist per unit length  $\vartheta$ . These sections may also deform out of plane, but this deformation must be same for all values of  $z$ . These kinematic considerations lead to the candidate displacement field

$$u_x = -\vartheta z y; \quad u_y = \vartheta z x; \quad u_z = \vartheta f(x, y), \quad (1)$$

where  $f$  is an unknown function of  $x, y$  describing the out-of-plane deformation.

Substituting these kinematic consideration into the strain-displacement relations  $e_{ij} = \frac{1}{2} \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right)$  yields

$$e_{xy} = 0; \quad e_{zx} = \frac{\vartheta}{2} \left( \frac{\partial f}{\partial x} - y \right); \quad e_{zy} = \left( \frac{\partial f}{\partial y} + x \right) \quad (2)$$

and it follow from Hooke's law in form  $\sigma_{ij} = \lambda e_{mm} \delta_{ij} + 2\mu e_{ij}$  [1] that

$$\sigma_{xx} = \sigma_{yy} = \sigma_{zz} = 0 \quad (3)$$

and

$$\sigma_{xy} = 0; \quad \sigma_{zx} = \mu\vartheta \left( \frac{\partial f}{\partial x} - y \right); \quad \sigma_{zy} = \mu\vartheta \left( \frac{\partial f}{\partial y} + x \right). \quad (4)$$

There are no body forces, so substitution into the equilibrium equations  $\frac{\partial \sigma_{ij}}{\partial x_j} + P_i = 0$  from [1] yields

$$\nabla^2 f = 0. \quad (5)$$

The torsion problem is therefore reduced to the determination of harmonic function  $f$  such that the stresses (4) satisfy the traction-free condition on the curved surfaces of the bar. The twist per unit length  $\vartheta$  can be determined by evaluating the torque on the cross-section  $\Omega$

$$T = \int \int_{\Omega} (x \sigma_{zy} - y \sigma_{zx}) dx dy. \quad (6)$$

### 2.1 The elliptical bar

For solution of the rectangular bar we used Prandtl's stress function defined by

$$\boldsymbol{\tau} \equiv \mathbf{i} \sigma_{zx} + \mathbf{j} \sigma_{zy} = \text{curl } \mathbf{k} \phi \quad (7)$$

or

$$\sigma_{zx} = \frac{\partial \phi}{\partial y}; \quad \sigma_{zy} = -\frac{\partial \phi}{\partial x}. \quad (8)$$

With this representation, the traction-free boundary condition can be written

$$\boldsymbol{\tau} \cdot \mathbf{n} = \sigma_{zn} = \frac{\partial \phi}{\partial t} = 0 \quad (9)$$

where  $\mathbf{n}$  is the local normal to the boundary of  $\Omega$  and  $n, t$  are a corresponding set of local orthogonal coordinates respectively normal and tangential to the boundary, Thus  $\phi$  must be constant around the boundary and for simply-connected bodies this constant can be taken as zero without loss of generality giving the simple condition

$$\phi = 0 \quad (10)$$

on the boundary and from this boundary condition we obtain

$$\nabla^2 \phi = -2\mu \vartheta. \quad (11)$$

We consider the bar of elliptical cross-section defined by the boundary

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 = 0, \quad (12)$$

loaded by the a torque  $T$ . The quadratic function

$$\phi = C \left( \frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 \right) \quad (13)$$

clearly satisfies the boundary condition (10). Substituting into (11), we obtain

$$\nabla^2 \phi = C \left( \frac{2}{a^2} + \frac{2}{b^2} \right) = -2\mu \vartheta, \quad (14)$$

which will be satisfied for all  $x, y$  if

$$C = -\frac{\mu \vartheta a^2 b^2}{a^2 + b^2}. \quad (15)$$

The torque  $T$  is obtained from (6) as

$$T = 2C \int_{-b}^b \int_{-a(1-y^2/b^2)^{1/2}}^{a(1-y^2/b^2)^{1/2}} \left( \frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 \right) dx dy = -\pi ab C \quad (16)$$

and hence

$$C = -\frac{T}{\pi ab}; \quad \vartheta = \frac{T(a^2 + b^2)}{\pi \mu a^3 b^3}. \quad (17)$$

The stresses are then obtained from (8) as

$$\sigma_{zx} = \frac{\partial \phi}{\partial y} = \frac{2Cy}{b^2} = -\frac{2Ty}{\pi ab^3} \quad (18)$$

$$\sigma_{zy} = -\frac{\partial \phi}{\partial x} = -\frac{2Cx}{a^2} = \frac{2Tx}{\pi a^3 b}. \quad (19)$$

The torsional rigidity of section  $K$ , defined such that

$$T = \mu K \vartheta, \quad (20)$$

$$K = \frac{\pi a^3 b^3}{(a^2 + b^2)}. \quad (21)$$

## 2.2 Equation for solution the elliptical bar in Mechanics of Materials

In mechanics of material [3] we used for solution of maximum shearing stress the equation

$$\tau_{max} = \frac{2T}{\pi ab^2} \quad (22)$$

and angle of twist is defined by

$$\varphi = \frac{(a^2 + b^2) T \ell}{\pi a^3 b^3 G}, \quad (23)$$

where  $\ell$  is the length of bar,  $G$  is the shear modulus of elasticity for the material.

The torsional rigidity of the section  $K$  is generally defined such that

$$T = \mu \vartheta K \quad (24)$$

## 3 Solution of torsion in program *MATHEMATICA*

For solution of torsional problems in program *MATHEMATICA* we were used the package "Structural Mechanics", which consist of solution the following types of cross sections:

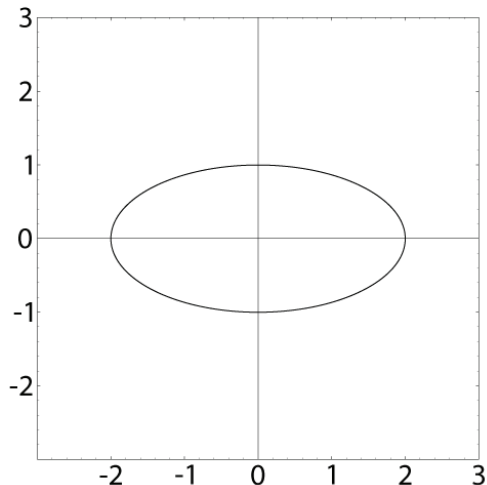
- circular cross sections
- elliptical cross sections
- rectangular cross sections
- equilateral-triangular cross sections
- sectorial-type cross sections
- semicircular cross sections

### 3.1 Elliptical Cross Section - Graphically

Consider a shaft with an elliptical cross section with the major axis  $2a = 4$  and the minor axis  $2b = 2$ . Using the function `CrossSectionPlot` with the first argument `EllipticalSection`, you can obtain the drawing of the elliptical cross section at the next figure.

Definition of cross section area.

```
In[26]:= crsec = CrossSectionPlot [Domain [EllipticalSection , {2, 1}],
    AspectRatio → 1,
    PlotRange → {3 {-1, 1}, 3 {-1, 1}},
    Axes → True ,
    Frame → True ];
```



You plot a three-dimensional drawing of the twisted elliptical bar by using the function `TorsionPlot` with the cross-section object `EllipticalSection`. The argument list of this function indicates the shaft size and applied twist. The grids on the three background faces of the plot box are included with the option `FaceGrids` in the next figure.

Deformation of elliptical bar

In the Fig.3.1 is shown how the section at  $z = 1$ . The line from the origin of the axis system to the ellipse boundary indicates the orientation of the root cross section.

Twisting of cross section area

As in the case of the circular cross sections, you can obtain the twist per unit length and the torsional rigidity coefficient using the domain name `EllipticalSection`.

We computed the torsional rigidity for `EllipticalSection` by the command.

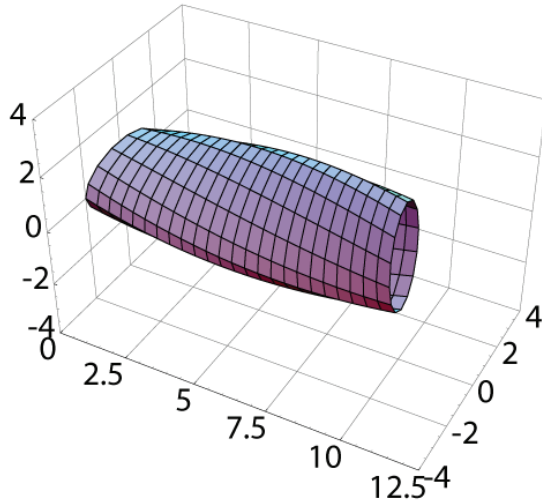
By replacing the twist in the displacements, you represent the displacement components in terms of the geometry, the modulus of rigidity of the shaft material, and the applied moment couple in the next terms.

At  $z = 1$ , you calculate the  $x$  and  $y$  components of the displacement vector  $(u_x, u_y)$  in the polar coordinates in the expression.

We generated the displacement field for the major radius  $a = 2$  units and minor radius  $b = 1$  unit axes with next expression.

Using the function `ListPlotVectorField` from the standard package `Graphics`PlotFields`, you should obtain a graphical representation of the displacement field.

```
In[27]:= TorsionPlot [Domain [EllipticalSection
PlotPoints → {15, 25},
Axes → True,
Boxed → False,
FaceGrids → {{-1, 0, 0}, {0, 1, 0}, {0, 0, -1}},
PlotRange → {{0, 12.5}, {-4, 4}, {-4, 4}}];
```



```
In[56]:= Clear [a, b, T, G, θ];
tw = Twist [EllipticalSection, {a, b}, T, G]
```

```
Out[57]= 
$$\frac{(a^2 + b^2) G}{a^3 b^3 \pi T}$$

```

```
In[58]:= tw /. {a → r, b → r}
```

```
Out[58]= 
$$\frac{2 G}{\pi r^4 T}$$

```

```
In[59]:= to = TorsionalRigidity [EllipticalSection, {a, b}, G]
```

```
Out[59]= 
$$\frac{a^3 b^3 G \pi}{a^2 + b^2}$$

```

```
In[62]:= disp = TorsionalDisplacements [EllipticalSection, {a, b}, θ, {x, y, z}]
```

```
Out[62]= 
$$\left\{ -y z \theta, x z \theta, -\frac{(a^2 - b^2) x y \theta}{a^2 + b^2} \right\}$$

```

```
In[63]:= disp /. θ → tw
```

```
Out[63]= 
$$\left\{ -\frac{(a^2 + b^2) G y z}{a^3 b^3 \pi T}, \frac{(a^2 + b^2) G x z}{a^3 b^3 \pi T}, -\frac{(a^2 - b^2) G x y}{a^3 b^3 \pi T} \right\}$$

```

```
In[64]:= {ux, uy} = Take [disp, 2]
```

```
Out[64]= 
$$\{-y z \theta, x z \theta\}$$

```

Using the command `TorsionalStresses`, you can generate the stress components in the bar due to the torsional load in the next term.

By replacing the twist  $\theta$ , you get the displacement field in closed form in terms of the applied

```

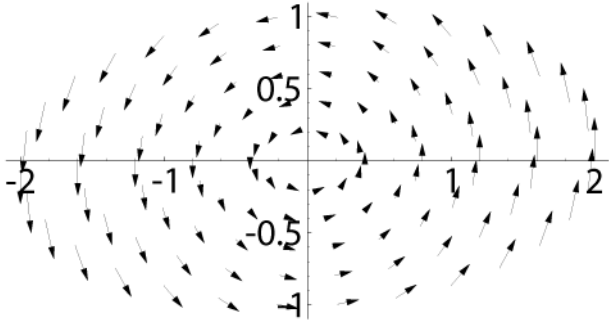
In[65]:= dfl = {ux, uy} /. {x -> r Cos[t], y -> 1/2 r Sin[t], z -> 1} /.  $\theta \rightarrow tw$  /. {a -> 2, b -> 1, T -> G}

Out[65]:=  $\left\{-\frac{5 r \sin [t]}{16 \pi}, \frac{5 r \cos [t]}{8 \pi}\right\}$ 

In[66]:= lx = N[Table[{r Cos[t], .5 r Sin[t]}, dfl],
                  {r, 0, 2, 2/5}, {t, 0, 2  $\pi$ , 2  $\pi$ /(10 (r + 1.))}]];

In[67]:= vecf = ListPlotVectorField[
  Partition[Partition[Flatten[lx], 2], 2], ScaleFunction -> (.75 #1 &), Axes -> True];

```



```

In[71]:= str = TorsionalStresses[EllipticalSection, {a, b}, G,  $\theta$ , {x, y}]

Out[71]:=  $\left\{0, 0, 0, 0, -\frac{2 a^2 G y \theta}{a^2 + b^2}, \frac{2 b^2 G x \theta}{a^2 + b^2}\right\}$ 

```

Figure 1:

torque.

```

In[72]:= str /.  $\theta \rightarrow tw$ 

Out[72]:=  $\left\{0, 0, 0, 0, -\frac{2 G^2 y}{a b^3 \pi T}, \frac{2 G^2 x}{a^3 b \pi T}\right\}$ 

```

## 4 Conclusion

In this paper is presented the theoretical solution of torsion properties for rectangular cross section area of bar loading by torque. This properties was derived using Saint-Venant's principle. In program *MATHEMATICA* was implemented Saint-Venant's principle in package *Structural Mechanics*. This package include two way solution of non-circular cross section area, first way is graphically and second way is solution of analytically. Both this way is described in this paper. When we compare theoretical solution derived in this paper by numerical solution using package *Structural Mechanics*, the results are identically.

## Acknowledgement

The paper was supported by grant from Grant Agency of APVV no. SK-CZ-0028-11.

## **References**

- [1] HOLZAPFEL, G.A.: *Nonlinear Solid Mechanics: A Continuum Approach for Engineering*. Wiley, 2004.
- [2] BARBER, J.R.: *Elasticity*, Second Edition, Series: Solid Mechanics and Its Applications, Vol. 107 Springer Verlag, Berlin, 2003.
- [3] HIBBELER, R.C.: *Mechanics of Materials*, SI Second Edition, Prentice Hall, Singapore, 2005.
- [4] Jančo, R., Kováčová, M.: *Solution of torsion of prismatic bar using program MATHEMATICA*. In: Aplimat - Journal of Applied Mathematics. - ISSN 1337-6365. - Vol. 1, No. 2 (2008), s. 241-248

## **Current address**

### **Assoc. Prof. MSc. Roland Jančo, PhD. ING-PAED IGIP**

Institute of Applied Mechanics and Mechatronics, Section of Strength of Material, Faculty of Mechanical Engineering, Slovak University of Technology Bratislava, Nám. slobody 17, 812 31 Bratislava, Slovak Republic, e-mail: roland.janco@stuba.sk .

### **Mgr. Monika Kováčová, PhD.**

Institute of natural sciences, humanities and social sciences, Faculty of Mechanical Engineering, Slovak University of Technology Bratislava, Nám. slobody 17, 812 31 Bratislava, Slovensk Republika, e-mail: monika.kovacova@stuba.sk .

# NON-LINEAR MOTION EQUATION OF MONOTONE COMMODITY STATE DEVELOPMENT WITH INFLEXION UNDER THE CONDITION OF PERFECT COMPETITION

ZEITHAMER Tomáš R. (CZ)

**Abstract.** A method of modelling commodity depreciation, based on the methodology of theoretical physics, is used to derive a deterministic linear motion equation of the second order to describe the degressive and progressive development of the instantaneous relative depreciation and price of a commodity over time in a model of market structure with perfect competition. The same approach is used to derive a non-linear motion equation of the second order for instantaneous relative depreciation with degressive/progressive development over time.

**Key words.** Depreciation, differential equation, econophysics, equation of motion.

*Mathematics Subject Classification:* Primary 00B02; Secondary 00A06

## 1 Introduction

Let us assume that instantaneous commodity depreciation  $w$  at every time  $t$  throughout the entire lifetime of the commodity is composed of the instantaneous commodity physical depreciation  $w_{PD}$  and the instantaneous commodity external depreciation  $w_{ED}$ . The law of internal composition of both types of depreciation is designated as  $\Delta$ , so the instantaneous commodity depreciation  $w$  is  $w = w_{PD} \Delta w_{ED}$ . The law of composition of magnitudes of instantaneous commodity physical and external depreciation is also designated by the symbol  $\Delta$ , so that  $w(t) = w_{PD}(t) \Delta w_{ED}(t)$ . In further considerations we presume (in linear approximation) that the law of composition of magnitudes of instantaneous commodity physical and external depreciation is algebraic addition, thus  $w(t) = w_{PD}(t) + w_{ED}(t)$ , so that the law of internal composition  $\Delta$  of instantaneous commodity physical and external depreciation is designated „+“ and so  $w = w_{PD} \Delta w_{ED} = w_{PD} + w_{ED}$ . For simplicity, we assume that the law of internal composition  $\Delta$ , or law of composition of magnitudes  $\Delta$  respectively, does not change over time for both kinds of depreciation. We further assume in the linear approximation that the instantaneous commodity depreciation  $w$ , the instantaneous commodity physical depreciation  $w_{PD}$  and the instantaneous commodity external

depreciation  $w_{ED}$  are continuous real functions at interval  $\langle t_0, t_e \rangle$  ( $t_0$  is the initial time of monitoring of the instantaneous commodity state and  $t_e$  is the time at which we cease monitoring the instantaneous commodity state i.e. level of the instantaneous commodity depreciation). The instantaneous commodity physical depreciation  $w_{PD}$  is defined as the permanent adverse change in the surface or dimensions of bodies of various states, induced by the interaction of functional surfaces or a functional surface and medium which causes wear [4]. The instantaneous commodity external depreciation  $w_{ED}$  is defined as a supplement to the instantaneous commodity physical depreciation i.e. the instantaneous commodity external depreciation is the permanent adverse or favorable change in market value of a commodity, which is not caused by instantaneous commodity physical depreciation (damage).

In a market structure with perfect competition<sup>1</sup>, the instantaneous commodity relative depreciation  $RD$  is defined by the magnitudes of instantaneous commodity relative depreciation in accordance with relation [1; 7]

$$RD(t) = \frac{w(t) - w(t_0)}{w(t_0)}, \quad (1)$$

where  $w(t_0) = w_0$  is the magnitude of instantaneous commodity depreciation at the initial time  $t_0$  and  $w(t)$  is the magnitude of instantaneous commodity depreciation at time  $t (t \geq t_0)$ . In addition to instantaneous commodity relative depreciation  $RD$ , the instantaneous commodity relative price  $RP$  is also defined under the condition of perfect competition by the magnitudes  $RP(t)$  at time  $t$  in accordance with the relationship [1; 7]

$$RP(t) = \frac{p(t_0) - p(t)}{p(t_0)}, \quad (2)$$

where  $p(t_0) = p_0$  is the magnitude of instantaneous commodity price  $p$  at the initial time  $t_0$  of monitoring the instantaneous commodity price on a select model market and  $p(t)$  is the magnitude of instantaneous commodity price at time  $t \geq t_0$ .

## 2 Linear motion equation of commodity state without inflexion

Instantaneous commodity depreciation  $w$  is a real composite function of time, i.e.  $w(t) = w(p(t))$ , where  $w(p)$  is the continuous decreasing real function of instantaneous commodity price  $p$  and instantaneous commodity price  $p$  is a continuous decreasing real function of time  $t$ . If we monitor the development of instantaneous commodity depreciation at time interval  $\langle t_0, t_e \rangle$ , then for the first derivation of functions  $w(p)$  and  $p(t)$  it holds that  $\frac{dw}{dp}(p) < 0$  for  $p \in \langle p(t_e), p(t_0) \rangle$  and  $\frac{dp}{dt}(t) < 0$  for  $t \in \langle t_0, t_e \rangle$ . It directly follows from these relationships that for the interval  $\langle t_0, t_e \rangle$ ,  $\frac{dw}{dt}(t) = \frac{dw}{dp}(p(t)) \cdot \frac{dp}{dt}(t) > 0$ . This means that instantaneous commodity depreciation  $w$  is a continuous increasing real function of time  $t$ , which corresponds to trends for

---

<sup>1</sup> In the model of a market structure with perfect competition we assume the following conditions are met: a) in each market there are a large number of buyers and sellers, none of which are strong enough to influence the price or output of a sector; b) all goods are homogeneous; c) there is free entry to and exit from all markets; d) all manufacturers and consumers have perfect information about prices and quantities traded on the market; e) companies attempt to maximize profit and consumers attempt to maximize utility; f) companies have free access to information about technologies [2; 3].

common commodities over time. Then, instantaneous commodity relative depreciation  $RD$  is also a continuous real function at interval  $\langle t_0, t_e \rangle$  and  $\frac{dRD}{dt}(t) > 0$  for every time  $t \in (t_0, t_e)$ .

The magnitude of instantaneous commodity relative depreciation  $RD$  over time  $t$  increases with acceleration and the acceleration of instantaneous commodity relative depreciation increases in direct proportion to the instantaneous speed of change of instantaneous commodity relative depreciation at time  $t$ . The motion equation of instantaneous commodity relative depreciation is thus [7]

$$\frac{d^2RD}{dt^2}(t) = B \frac{dRD}{dt}(t), \quad (3)$$

where  $B$  is the constant of proportionality,  $B > 0$ . In addition, let initial conditions be met where  $RD(t_0) = RD_0 > 0$ ,  $\frac{dRD}{dt}(t_0) = v_0 > 0$ , so that the solution of differential equation (3) at interval  $\langle t_0, t_e \rangle$  is then

$$RD(t) = RD_0 - \frac{v_0}{B} + \frac{v_0}{B} e^{B(t-t_0)}. \quad (4)$$

From here it directly follows that instantaneous commodity relative depreciation  $RD$  is a purely convex function at interval  $\langle t_0, t_e \rangle$ . This means that the increase in instantaneous commodity relative depreciation at interval  $\langle t_0, t_e \rangle$  is progressive.

Instantaneous commodity relative depreciation  $RD$  increases with acceleration at time  $t$  again and the acceleration of instantaneous commodity relative depreciation increases in direct proportion to the speed of change of relative depreciation at time  $t$  while the constant of proportionality is negative. The motion equation of instantaneous commodity relative depreciation is then [6; 7]

$$\frac{d^2RD}{dt^2}(t) = -B \frac{dRD}{dt}(t), \quad (5)$$

where  $(-B)$  is the constant of proportionality,  $B > 0$ . In addition, let initial conditions be met where  $RD(t_0) = RD_0 > 0$ ,  $\frac{dRD}{dt}(t_0) = v_0 > 0$ , so that the solution of the differential equation (5) at interval  $\langle t_0, t_e \rangle$  is then

$$RD(t) = RD_0 + \frac{v_0}{B} - \frac{v_0}{B} e^{-B(t-t_0)}. \quad (6)$$

From here it directly follows that instantaneous commodity relative depreciation  $RD$  is a purely concave function at interval  $\langle t_0, t_e \rangle$ . This means that the increase in instantaneous commodity relative depreciation at interval  $\langle t_0, t_e \rangle$  is degressive. The progressive increase of instantaneous commodity relative depreciation is characteristic, for example, of certain types of food goods, while degressive increase of relative depreciation may be seen in certain commodities in the automotive industry.

Specific types of commodities are not listed here as the breakdown of all commodities under the condition of perfect competition into individual disjoint classes of commodities is the subject of a separate investigation.

Motion equations (3) and (5) for instantaneous commodity relative depreciation  $RD$  yield a deterministic differential equation for instantaneous commodity price  $p$  while a commodity is an element of one of the disjoint classes of the set of all commodities. For each commodity class found, it will be necessary to determine the functional relationship between instantaneous commodity depreciation  $w$  and the instantaneous commodity price  $p$  at interval  $\langle t_0, t_e \rangle$ . Assume

that we have selected a single specific class of commodity from the set of all commodities. For each commodity of this particular class let  $w(t) = D(p(t_0) - p(t))$ , so that, in accordance with equation (1),  $RD(t) = \frac{[D(p_0 - p(t)) - w_0]}{w_0}$  at interval  $\langle t_0, t_e \rangle$ . A constant  $D$  ( $D > 0$ ) is given in such units to ensure that the same units are found on both sides of the equation  $w(t) = D(p(t_0) - p(t))$ . Directly following from deterministic differential equation (3) for instantaneous commodity relative depreciation  $RD$  is the deterministic differential equation for instantaneous commodity price  $p$  at interval  $\langle t_0, t_e \rangle$ , which is  $\frac{d^2 p}{dt^2}(t) = B \frac{dp}{dt}(t)$  with initial conditions  $p(t_0) = p_0 > 0$ ,  $\frac{dp}{dt}(t_0) = r_0 < 0$ , where  $\frac{dp}{dt}(t) < 0$  for  $t \in (t_0, t_e)$ . The solution of this differential equation for a purely concave drop in the instantaneous commodity price may be written as  $p(t) = p_0 - \frac{r_0}{B}(1 - e^{B(t-t_0)})$ . Deterministic differential equation (5) for instantaneous commodity relative depreciation  $RD$  yields a deterministic differential equation for instantaneous commodity price  $p$  at interval  $\langle t_0, t_e \rangle$  which is  $\frac{d^2 p}{dt^2}(t) = -B \frac{dp}{dt}(t)$  with initial conditions  $p(t_0) = p_0 > 0$ ,  $\frac{dp}{dt}(t_0) = r_0 < 0$ , where  $\frac{dp}{dt}(t) < 0$  for  $t \in (t_0, t_e)$ . The solution of this differential equation for a purely convex drop in the instantaneous commodity price may be written as  $p(t) = p_0 + \frac{r_0}{B}(1 - e^{-B(t-t_0)})$ .

### 3 Non-linear motion equation of commodity state with inflexion

In this section of our work we again presume the following conditions to be met: (1) the commodity is on one of the markets of the model of market structure with perfect competition at initial time  $t_0$ ; (2) at time  $t_0$  the commodity is found in its initial state, which is uniquely determined by the magnitude of instantaneous commodity depreciation  $w(t_0) = w_0$ .

Let the acceleration of  $\frac{d^2 RD}{dt^2}$  of the instantaneous commodity relative depreciation be the sum of two components, i.e.

$$\frac{d^2 RD}{dt^2} = \left(\frac{d^2 RD}{dt^2}\right)_1 + \left(\frac{d^2 RD}{dt^2}\right)_2. \quad (7)$$

The first component of acceleration is a consequence of physical and chemical processes (including also social/psychological processes in physico-chemical approximation), which cause the first component of the instantaneous acceleration to increase in direct proportion to the magnitudes of rate of change of the instantaneous commodity relative depreciation, i.e.

$$\left(\frac{d^2 RD}{dt^2}(t)\right)_1 = B \frac{dRD}{dt}(t), \quad (8)$$

where  $B$  is the proportionality constant,  $B > 0$  and  $t \in \langle t_0, t_e \rangle$ . The second component of acceleration results from physical and chemical processes (including also social/psychological processes in physico-chemical approximation), which cause the second component of the instantaneous acceleration to be directly proportional to the product of the magnitude of rate of change of the instantaneous commodity relative depreciation  $\frac{dRD}{dt}(t)$  and the magnitude of instantaneous commodity relative depreciation  $RD(t)$ , while the proportionality constant is negative, thus

$$\left(\frac{d^2 RD}{dt^2}(t)\right)_2 = -A \frac{dRD}{dt}(t) RD(t), \quad (9)$$

where  $(-A)$  is the proportionality constant,  $A > 0$ ,  $t \in \langle t_0, t_e \rangle$ .

By substituting relations (8) and (9) into equation (7), we obtain the following motion equation for the acceleration of instantaneous commodity relative depreciation

$$\frac{d^2 RD}{dt^2}(t) = B \frac{dRD}{dt}(t) - A \frac{dRD}{dt}(t) RD(t), \quad (10)$$

where  $A > 0, B > 0, t \in \langle t_0, t_e \rangle$ .

One of the subsets of the set of solutions for motion equation (10) is given by

$$RD(t) = \frac{y_2 + y_1 e^{\sqrt{D}(t+C_2)}}{1 + e^{\sqrt{D}(t+C_2)}},$$

where for constants  $D, y_1, y_2, C_2$  it follows that  $D = B^2 + 2AC_1$ ,  $y_1 = \frac{B+\sqrt{D}}{A}$ ,  $y_2 = \frac{B-\sqrt{D}}{A}$ ,

$0 < |y_2| < y_1, y_2 < 0$ ,  $-\frac{B^2}{2A} < C_1 < 0$ ,  $C_2 = \frac{1}{\sqrt{D}} \ln \left( \left| \frac{y_2}{y_1} \right| \right) - t_p$ . At time  $|t_p|$  the value of instantaneous commodity relative depreciation is zero. The given subset of the solutions of motion equation (10) shows the progressive – degressive increase of instantaneous commodity relative depreciation with an inflexion point at time  $t = -C_2$  and a limit at  $\lim_{t \rightarrow +\infty} RD(t) = y_1$ .

## 4 Conclusions

Assuming that the market value of the commodity at time  $t$  is fully determined exclusively by the value of the instantaneous commodity price  $p(t)$ , methodological procedures taken from theoretical physics were used to construct motion equations for instantaneous commodity relative depreciation  $RD$ . Motion equations (3) and (5) for the progressive and degressive increase of instantaneous commodity relative depreciation are linear differential equations of the second order with constant coefficients assuming market structures with perfect competition. Motion equation (10) of instantaneous commodity relative depreciation for the progressive/degressive growth of depreciation is a non-linear differential equation of the second order with constant coefficients. Motion equation (10) was also derived for instantaneous commodity relative depreciation on a market with perfect competition. In the solutions set for motion equation (10), there is the subset of solutions which model progressive/degressive growth of the magnitudes of instantaneous commodity relative depreciation with a single inflexion point.

## Acknowledgement

The author is grateful to Mrs. Pavla Jará and the National Technical Library for their great effort and excellent work, which was indispensable in the completion of a large portion of this work. This paper is dedicated to Mrs. Věra Ruml – Zeithamer and Mrs. Anna Ruml and Mr. František Ruml.

## References

- [1] DROZEN, F.: *Modelling of price dynamics and appreciation. Ekonomický časopis* (Journal of Economics), Vol. 56, No. 10, pp. 1033-1044, (2008), ISSN 0013-3055.
- [2] GOODWIN, N., NELSON, J., A., ACKERMAN, F., WEISSKOPF, T.: *Microeconomics in context*, 2nd ed., M. E. Sharpe, Inc., Armonk, New York, (2009) ISBN 978-0-7656-2301-0
- [3] NICHOLSON, W., SNYDER, Ch.: *Microeconomic Theory-Basic principles and extensions*, 10th ed., South- Western College Pub., (2008), ISBN 978-0324-42162-0.
- [4] POŠTA, J., VESELÝ, P., DVOŘÁK, M.: *Degradace strojních součástí, (Degradation of machine parts)* (1st ed.), ČZU, Praha, (2002), ISBN 80-213-0967-9 (in Czech).
- [6] ZEITHAMER, R., T.: *The Approach of Physics to Economic Phenomena*. 10th International conference Aplimat 2011, Proceedings, pp. 1303 – 1308, Bratislava, 2011. ISBN 978-80-89313-51-8.
- [7] ZEITHAMER, R., T.: *On the Possibility of Econophysical Approach to Commodity Valuation Theory*, 7. Konference o matematice a fyzice na vysokých školách technických 2011, Proceedings, pp. 142 - 150, Brno, 2011. ISBN 978-80-7231-816-2.

## Current address

**Tomáš R. Zeithamer, Ing., Ph.D.,**  
Faculty of Informatics and Statistics,  
Department of Mathematics, University of Economics, Prague,  
Ekonomická 957, 148 00 Prague 4, The Czech Republic  
e-mail: [zeith@vse.cz](mailto:zeith@vse.cz)