TECHNICAL TRANSACTIONS
FUNDAMENTAL SCIENCES

CZASOPISMO TECHNICZNE
NAUKI PODSTAWOWE

ISSUE 1-NP (16)
YEAR 2016 (113)

ZESZYT 1-NP (16)
ROK 2016 (113)

WYDAWNICTWO POLITECHNIKI KRAKOWSKIEJ

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE

FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE

ISSUE 1-NP (16) | ZESZYT 1-NP (16)
YEAR 2016 (113) | ROK 2016 (113)

Basic version of each Technical Transactions magazine is its online version
Pierwotną wersją każdego zeszytu Czasopisma Technicznego jest jego wersja online
www.ejournals.eu/Czasopismo-Techniczne www.technicaltransactions.com www.czasopismotechniczne.pl

Fundamental Sciences Series
1-NP/2016

PHYSICS

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

WIESŁAWA BAŻELA*, MARCIN DUL*, ANDRZEJ SZYTUŁA**,
VOLODYMYR DYAKONOV***

CORRELATION BETWEEN CRYSTAL AND MAGNETIC STRUCTURE OF THE POLYCRYSTALLINE AND NANOPARTICLE $TBMNO_3$ MANGANITE

ZWIĄZEK MIĘDZY STRUKTURĄ KRYSTALICZNĄ I MAGNETYCZNĄ POLIKRYSTALICZNEJ I NANOROZMIAROWYCH PRÓBEK MANGANITU $TBMNO_3$

Abstract

On the basis of neutron diffraction data the Mn−O bond lengths and Mn−O−Mn bond angles for the poly- and nanocrystalline $TbMnO_3$ samples are determined. All the samples crystallize in the orthorhombically distorted perovskite structure (space group *Pnma*) and exhibit antiferromagnetic ordering below 41 K. The Tb atoms and O_1 atoms are in (4)c site, Mn atoms − in 4(b) site and O_2 atoms − in 8(d) site. The Mn−O_2−Mn bond angles for the polycrystalline and nanosize samples are similar, whereas the Mn−O_1−Mn bond angles for the nanoparticle samples are larger. The temperature dependencies of the Mn−O bond lengths and the Mn−O−Mn bond angles, the Jahn-Teller distortion parameter (JT) and MnO_6 − octahedron distortion parameter (delta) for polycrystalline sample exhibit anomalies at T_N temperature for Mn sublattice.

Keywords: *crystal structure, exchange interactions, nanoparticle, grain size, Mn−O bond lengths, Mn−O−Mn bond angles, the Jahn-Teller distortion parameter*

Streszczenie

Na podstawie wyników neutronowej dyfrakcji wyznaczono długości wiązań Mn–O oraz kąty wiązania Mn−O−Mn dla polikrystalicznej oraz nanorozmiarowych próbek manganitu $TbMnO_3$. Wszystkie próbki krystalizują w rombowo zdystorsowanej strukturze perowskitu (grupa przestrzenna *Pnma*) i wykazują antyferromagnetyczne uporządkowanie poniżej 41 K. Atomy Tb i tlenu O_1 zajmują pozycję 4(c), atomy Mn pozycję 4(b), a atomy tlenu O_2 pozycję 4(d). Wartości kątów wiązania Mn–O_2–Mn są zbliżone dla polikrystalicznej i nanorozmiarowych próbek związku $TbMnO_3$, podczas gdy wartości kątów wiązania Mn–O_1–Mn są wyższe dla próbek nanorozmiarowych. Temperaturowe zależności: długości wiązań Mn−O, kątów wiązania Mn–O–Mn, parametru dystorsji Jahna-Tellera (JT) oraz parametru dystorsji oktaedru MnO_6 (delta) wykazują dla próbki polikrystalicznej anomalie w temperaturze Néela dla podsieci Mn.

Słowa kluczowe: *struktura krystaliczna, oddziaływania wymiany, nanocząstki, rozmiar ziarna, długości wiązań Mn−O, kąty wiązania Mn–O−Mn, parametr dystorsji Jahna-Tellera*

DOI: 10.4467/2353737XCT.16.133.5712

* Wiesława Bażela (wbazela@pk.edu.pl), Marcin Dul, Institute of Physics, Cracow University of Technology.
** Andrzej Szytuła, M. Smoluchowski Institute of Physics, Jagiellonian University.
*** Volodymyr Dyakonov, Institute of Physics, PAS, Warsaw.

1. Introduction

The explanation of the complex magnetic interactions and correlation of the magnetic, structural and electron properties of the $REMnO_3$ (RE are the rare – earth ions) manganites is of fundamental interest [1].

$TbMnO_3$ has been attracting a lot of attention in recent years because of its strong coupling between ferroelectricity and magnetism [2].

The main motivation for performed studies was to obtain the data concerning the crystal structure and magnetic properties of the $TbMnO_3$ manganite as a function of the grain size. The model for interpretation of magnetic properties of the nanoparticle compounds is based on the ratio of ideal inner core and nonmagnetic surface, i.e., on the surface/volume ratio [3].

In this paper, we have discussed the influence of the internal structural parameters (Mn–O bond lengths and Mn–O–Mn bond angles) on the magnetic behaviour of the polycrystalline and two nanopowder $TbMnO_3$ samples. The structural distortion parameters, i.e. Jahn-Teller distortion (JT) and MnO_6 – octahedron distortion (delta) were found for all the samples.

2. Experiment and results

The polycrystalline $TbMnO_3$ manganite was prepared by the solid-state reaction. The final sintering treatment was performed at 1150ºC for 15 h. For preparation of the nanosize $TbMnO_3$ manganite the sol-gel method has been used. The two samples of the nanopowders were obtained after annealing at 800 and 850ºC [4]. The crystal structure of the samples was obtained by X-ray powder diffraction at room temperature using the Philips PW-3710 X'PERT diffractometer with CuK_α radiation. The obtained data were analysed with the Rietveld-type refinement soft-ware Fullprof program [5].

The X-ray diffraction data indicate that all the samples studied have orthorhombic crystal structure (space group *Pnma*). In this structure the Tb and O_1 atoms occupy 4(c) site: (x, y, 1/4), O_2 atoms are in 8(d) site: (x, y, z) and Mn atoms are in 4(b) site: (1/2, 0, 0) (Fig.1).

The obtained data indicate that the lattice constants and positional parameters x_i, y_i, z_i slightly change with changing grain size [6]. The data for the nano-samples indicate that the a-constant is smaller and the b and c are larger than ones for the polycrystalline sample. All parameters have minimum at T = 30 K and quickly increase with increasing temperature.

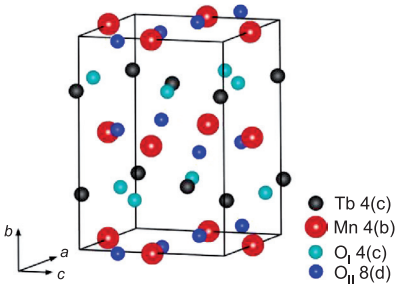

Fig. 1. The orthorhombic crystal unit cell

The grain sizes of nano-samples (800 and 850ºC) were determined using the Scherrer relation $d = \lambda/B\cos\theta_B$, where d is the grain size, λ is the X-rays wavelength, θ_B is the corresponding angle of the Bragg diffraction and B is the difference between half-widths of the Bragg reflex of the nanopowder and the standard sample [7]. The grains sizes were calculated using the experimental X-ray data and the following relation: $B = \beta - \beta_0$, where β is the half-widths of the Bragg reflex of the investigated sample and β_0 the similar value for the standard sample of Si powder with the grain size of 10 μm. The exact method of determination of grain size is described in [8]. The average grain size values determined there are: 60 nm and 45 nm for 850-nano and 800-nano samples, respectively.

In the next step, the grain sizes and strain effects were determined based on the Williamson-Hall method [9]. In this method, the broadening of Bragg peak is a sum of grain size broadening $\beta_d = K\lambda/d\cos\Theta$ and strain broadening $\beta_s = \epsilon\, tg\Theta$, where the shape factor K is close to 1, d is a value of grain size and ϵ is a strain constant.

Thus, the resulting total broadening: $\beta_{total} = \beta_s + \beta_d = \epsilon\, tg\Theta + K\lambda/d\cos\Theta$.

Multiplication of the above equation by $\cos\theta$ leads to

$$\beta_{total}\cos\Theta = \epsilon \sin\Theta + K\Theta/d.$$

Therefore, the grain size d can be determined from the intercept of line fitted with linear regression as applied to the $\beta_{total}\cos\Theta$ versus $\sin\Theta$ data.

The experimental β_{total} values have been determined from the relation:

$$\beta_{total} = [(\beta_\Theta)^2_{sample} - (\beta_\Theta)^2_{Si}]^{1/2},$$

where $(\beta_\Theta)_{sample}$ is a half – width of selected Bragg reflection of the investigated sample, while $(\beta_\Theta)_{Si}$ is a similar value found for the standard sample of Si powder.

The values of the grain size d are equal to 57 nm and 51 nm for 850-nano and 800--nano samples, respectively. Presented data indicate that the value of grain size increases with increasing annealing temperature.

The analysis presented in this paper based on the neutron diffraction powder data collected using the E2 and E6 diffractometers installed at the BERII reactor (Helmholtz--Zentrum Berlin) within the temperature range from 1.6 to 260 K. The data were processed using the program FullProf.

Neutron diffraction data [10] indicate that all the samples have orthorhombic crystal structure. Determined values of the lattice constants and atomic positions parameters are presented in Table I in [10]. Low temperature data indicate that the magnetic ordering of Mn and Tb sublattice for polycrystalline $TbMnO_3$ is sinusoidal modulated described by the propagation vector $\mathbf{k} = (k_x, 0, 0)$. The magnetic moments in Mn sublattice order below 41 K, while in Tb one they order below 9 K.

In the crystal unit cell (space group *Pnma*) the Mn^{3+} and Tb^{3+} sublattices can be described by four modes proposed by Bertaut [11]: one ferromagnetic ordering: $\mathbf{F} = m_1 + m_2 + m_3 + m_4$ and three antiferromagnetic arrangements: $\mathbf{A} = m_1 - m_2 - m_3 + m_4$, $\mathbf{C} = m_1 + m_2 - m_3 - m_4$ and $\mathbf{G} = m_1 - m_2 + m_3 - m_4$.

Below 41 K, neutron diffraction patterns for the polycrystalline sample exhibit additional magnetic peaks connected with the antiferromagnetic modulated ordering with $k_x = 0.28$ in Mn sublattice described by $\mathbf{C_x}$ – mode (see Fig. 1a in [6]).

The Mn magnetic moments, parallel to the *a*-axis, form a collinear incommensurate structure of $\mathbf{C}_x$ – mode. At T = 16 K a noncollinear magnetic structure described by $\mathbf{C}_x\mathbf{A}_z$ – mode with the Mn moment in the *a*–*c* plane was observed (see Fig. 2).

The Tb sublattice exhibits the antiferromagnetic incommensurate ordering of the $\mathbf{F}_y\mathbf{A}_z$ – type at T = 5 K. The Tb magnetic structure is described by propagation vector $\mathbf{k} = (k_x, 0, 0)$ where k_x is equal to 0.423(1) (Fig. 2). At the same temperature, the Mn moments still form the $\mathbf{C}_x\mathbf{A}_z$ structure described by propagation vector $\mathbf{k} = (k_x, 0, 0)$ where k_x is equal to 0.282(1).

The refinement of the magnetic peaks intensities for the nano-800 and nano-850 samples below T_N shows that the Mn moments form a collinear incommensurate magnetic structure of $\mathbf{C}_x$ – type described by the propagation vector $\mathbf{k} = (k_x, 0, 0)$ (see Fig. 3). The corresponding patterns for the nano-800 and nano-850 samples are presented in Figs. 1b and 1c in [6]. At 1.6 K, the additional peaks connected to the Tb moments ordering are visible. The Tb structure can be described by the $\mathbf{A}_z$ – mode with propagation vector $\mathbf{k} = (k_x, 0, 0)$ (see Fig. 3), while for the polycrystalline sample the $\mathbf{F}_y\mathbf{A}_z$ – mode was evidenced.

The Mn magnetic moments values for nano-samples (at 1.6 K μ(Mn) = 2.94(2) μ_B and 3.03(4) μ_B for nano-800 and nano-850, respectively) are smaller than for the polycrystalline sample (at 5 K μ(Mn) = 4.06(2) μ_B), whereas for the nano-samples the k_x components equal to 0.321(2) and 0.328(2) for nano-800 and nano-850, respectively, are larger than in the polycrystalline sample (0.282(1)).

Similar conclusions concern the parameters characterizing the ordering in Tb sublattice. At 1.6 K μ(Tb) = 3.68(11) μ_B and 4.43(7) μ_B for nano-800 and nano-850, respectively.

For polycrystalline sample μ(Tb) is equal to 6.55(4) μ_B at 5 K. The values of k_x component for Tb sublattice are larger for nano-samples (0.443(5) and 0.451(3) for nano-800 and nano--850, respectively.

The T_N Néel temperature connected with the Tb sublattice is lower for nano-samples (6.7 K) in comparison to polycrystalline sample (9 K).

Magnetic structures of the polycrystalline and nanoparticle $TbMnO_3$ compounds are presented in Figs. 2 and 3, respectively. These magnetic structures are incommensurate

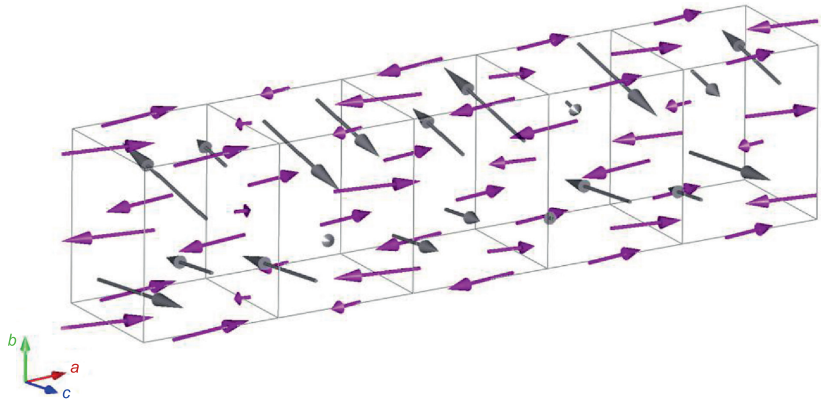

Fig. 2. Sinusoidal magnetic ordering in Mn sublattice – violet ($\mathbf{C}_x\mathbf{A}_z$ – mode, k_x = 0.282(1)) and in Tb sublattice – black ($\mathbf{F}_y\mathbf{A}_z$ – mode, k_x = 0.423(1)) for polycrystalline $TbMnO_3$

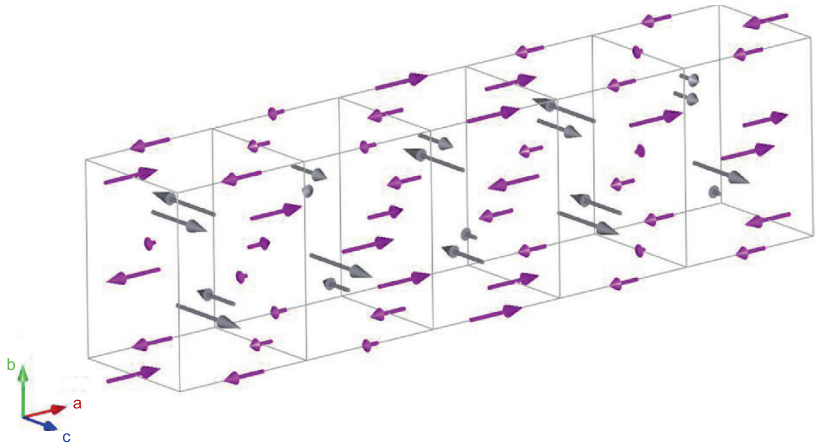

Fig. 3. Sinusoidal magnetic ordering in Mn sublattice – violet ($\mathbf{C}_x$ – mode, $k_x = 0.326(4)$) and in Tb sublattice – black ($\mathbf{A}_z$ – mode, $k_x = 0.443(5)$) for 800-nano $TbMnO_3$

in comparison with the crystal one. The periods of modulation of the magnetic structure are equal to 3.54*a* (Mn sublattice) and 2.36*a* (Tb sublattice) for polycrystalline sample and 3.06*a* (Mn) and 2.25*a* (Tb) for nano-samples, respectively.

In this paper we have focused on the behaviour of the internal structural parameters in the polycrystalline and two nanoparticle samples (Mn–O bond lengths and Mn–O–Mn bond angles) as a function of temperature. In the orthorhombic unit cell there are the three crystallographically independent ($Mn–O_I(4c) = r_1$, $Mn–O_2(8d)_1 = r_2$, $Mn–O_2(8d)_2 = r_3$) bond lengths and the two ($Mn–O_1–Mn = \alpha$, $Mn–O_2–Mn = \beta$) bond angles (Fig. 4). The temperature dependencies of the Mn–O bond lengths and Mn–O–Mn bond angles for the polycrystalline and two nanoparticle $TbMnO_3$ samples are presented in Fig. 5.

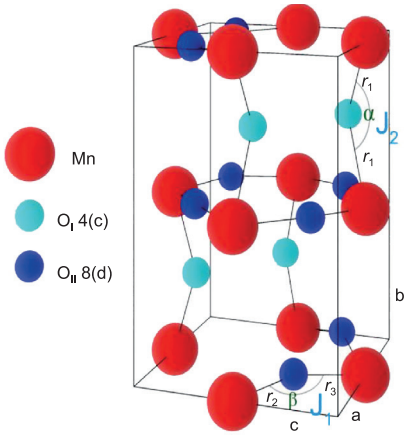

Fig. 4. The orthorhombic crystal unit cell with the marked Mn–O bond lengths (r_1, r_2, r_3) and Mn–O–Mn bond angles (α, β) and the exchange integrals J_1, J_2

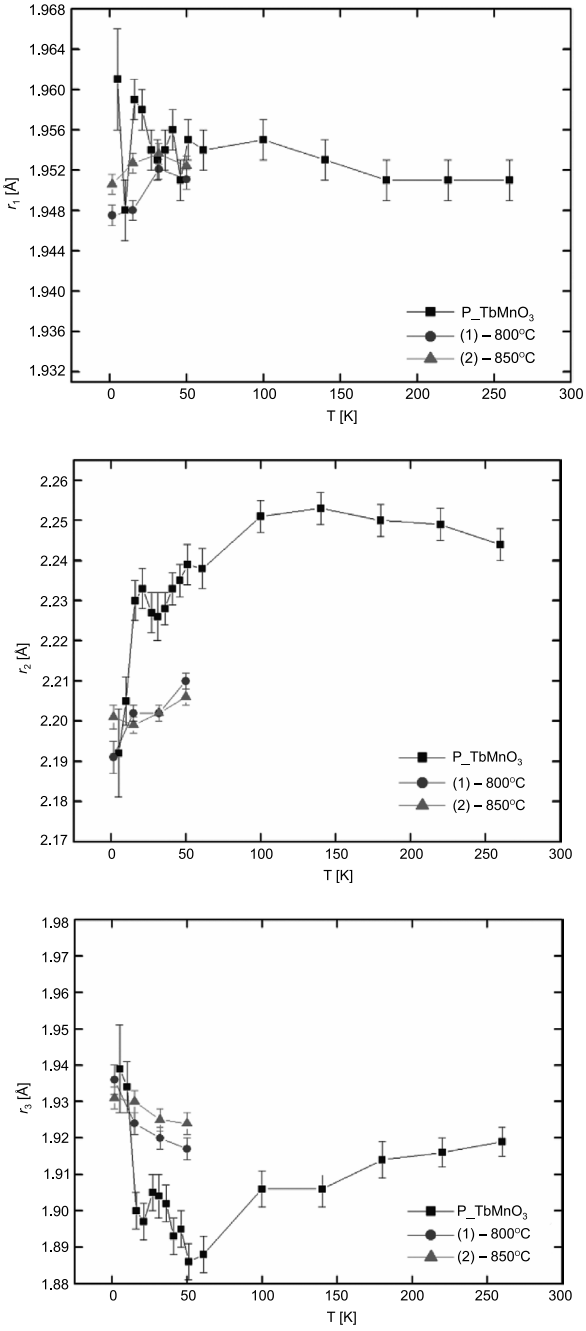

Fig. 5. Mn–O bond lengths (r_1, r_2, r_3) as a function of temperature for polycrystalline and 800 and 850-nano-samples of $TbMnO_3$

The temperature dependencies of the r_1 and r_2 bond lengths show that r_1 and r_2 are smaller for the nanosize samples as compared to the polycrystalline sample (see Fig. 5). This suggests that in these samples there are a greater overlap of *p* and *d* orbitals.

We have observed an increase of the r_1 and r_2 bond lengths for the nanosize samples with approaching to the Néel temperature. For the polycrystalline sample above T = 50 K the stabilization of all three r_1, r_2 and r_3 bond lengths is visible. The dependence of r_3(T) exhibits an inverse behaviour as compared to r_2(T) (see Fig. 5).

Fig. 6 presents a gradual increase of the α bond angle vs temperature for the polycrystalline sample, whereas for the nanosize samples a decrease of α till to the Néel temperature and an increase beyond T_N is observed. The α bond angles are larger for the nanoparticle samples as compared to α for the polycrystalline sample.

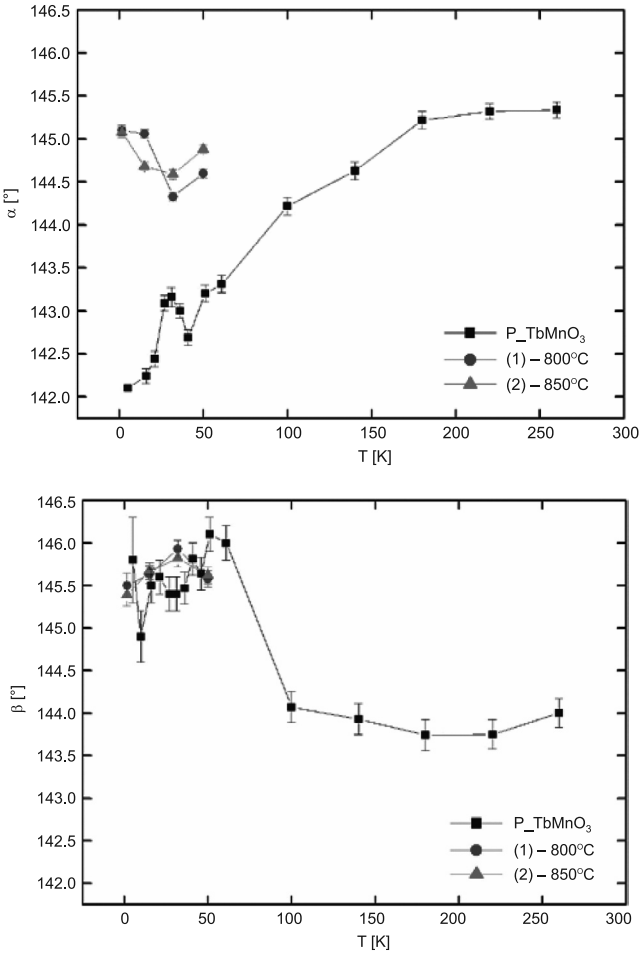

Fig. 6. Mn–O–Mn bond angles (α, β) as a function of temperature for the polycrystalline and 800-nano and 850-nano $TbMnO_3$ samples

This suggests an increase of superexchange interactions along the *b*-axis. Values of β bond angle are similar for the nano- and poly-$TbMnO_3$ samples. For both types of samples an increase of β is observed till the Néel temperature. Beyond this temperature the β bond angle value substantially drops. Using the r_1, r_2 and r_3 bond lengths the Jahn-Teller parameter [12] for the polycrystalline and nanosize samples has been determined according to the formula [13]:

$$JT = \sqrt{\frac{1}{3}\sum_{i=1}^{3}[(r_i) - < r >]^2}$$

where r_i are the experimentally determined values of (Mn–O) interatomic lengths (see Fig. 4) and $< r >$ is the average value of these lengths.

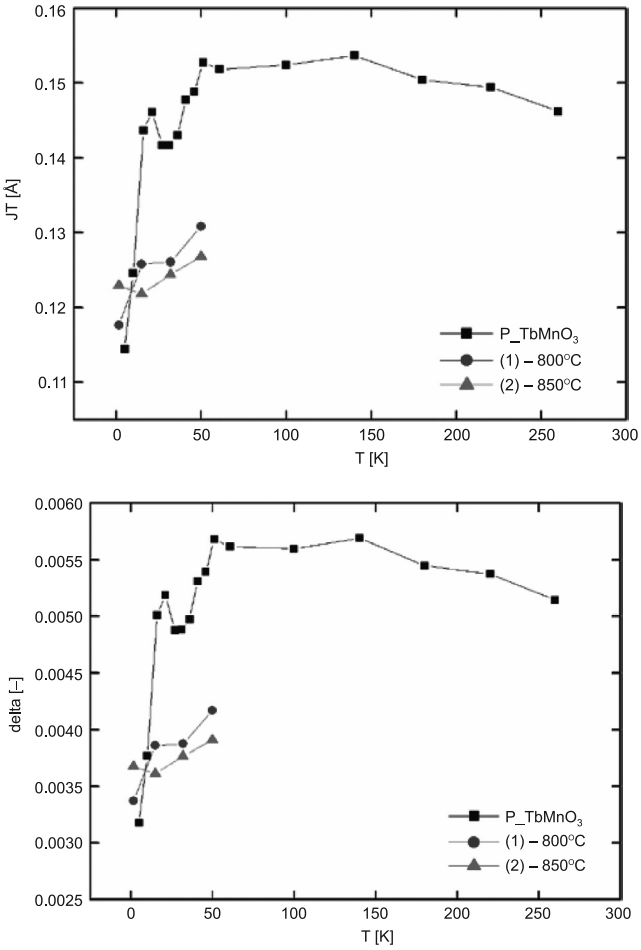

Fig. 7. Temperature dependences of the Jahn-Teller parameter (JT) and the parameter delta for the polycrystalline and nanosize samples of $TbMnO_3$

The parameter delta, which describes the distortion of MnO_6 octahedron is calculated using the formula:

$$\text{delta} = \frac{1}{3}\sum_{i=1}^{3}\left[\frac{r_i - < r >}{< r >}\right]^2$$

Temperature dependences of the Jahn-Teller parameter (JT) and the parameter delta for the polycrystalline and nanosize samples of $TbMnO_3$ are presented in Fig. 7.

The values of both the Jahn-Teller parameter and the delta parameter indicate the MnO_6 octahedron distortion. Distortion is much smaller for the nanocrystalline samples than for polycrystalline one. For polycrystalline sample the Jahn-Teller parameter has anomaly at Néel temperature.

3. Discussion

The data presented in this paper indicate that the magnetic properties of the nanoparticle samples strongly depend on the grain size. This manifests itself in decrease of the value of both magnetic moments in the ordered state and magnetic ordering temperature with decreasing grain size.

The $TbMnO_3$ manganite exhibits a para- antiferromagnetic phase transition at 41 K, where the Mn^{3+} ions develop a sinusoidal incommensurate ordering propagating along the *a* – direction of the unit cell, described by $\mathbf{C_xA_z}$ – mode. Magnetic order in the Mn sublattice is collinear of $\mathbf{C_x}$ – type in the temperature range of 21–41 K. For the investigated nano--samples a magnetic ordering in the Mn sublattice is described by collinear $\mathbf{C_x}$ – mode only.

Observed antiferromagnetic order in the Mn sublattice is result of the superexchange mechanism (cation-anion-cation) which exists in manganites. The superexchange interaction depends on the Mn–O–Mn bond angles (α, β) and is joined with partial overlap of the *p* (O) and *d* (Mn) orbitals. The interactions between Mn moments are based on the exchange integrals discussed by Bertaut [14].

At temperature 1.6 K, the values of α and β bond angles are equal to 142º and 146º for the polycrystalline sample while they are equal to 145º and 145.5º for the nanoparticle samples, respectively.

The obtained values of the Mn–O–Mn bond angles (α, β) are smaller than 180º. This fact indicates the moderate ferro- or antiferromagnetic interaction between magnetic moments of Mn according to the Goodenough-Kanamori rules [15, 16].

Analysis of interactions in the orthorhombic manganites with magnetic structure described by the propagation vector $\mathbf{k} = (k_x, 0, 0)$ gives the following dependence between k_x and exchange integrals: $\cos(\pi k_x) \approx (2J_2 - J_1)$ [17], where J_1 is the exchange integral in the basal *a*–*c* plane [t_{2g}(Mn) – $2p_\pi$ (O) – t_{2g} (Mn)] and J_2 is the exchange integral along the *b*-axis [e_g(Mn) – $2p_\pi$ (O) – e_g (Mn)].

The inelastic neutron scattering for the bulk $TbMnO_3$ yields the positive value of $J_1 \approx 0.15(1)$ meV and negative one of $J_2 \approx -0.31(2)$ meV [18]. This result confirms, that for the $TbMnO_3$ manganite the superexchange interaction between Mn–O_2–Mn spins in the *a*–*c* plane (J_1) is ferromagnetic, while the interaction Mn–O_1–Mn along the *b*-axis

is antiferromagnetic (J_2) (see Fig. 4). An increase of the k_x component observed in the nanoparticle $TbMnO_3$ sample indicates the decrease of the exchange integrals in nano--samples.

The presented results suggest that the nanoparticle size plays an important role in the formation of magnetic properties. The influence of deformation of the MnO_6-octahedron on the magnetic structure of $TbMnO_3$ manganite is observed. The values of Mn–O_2–Mn bond angles in the polycrystalline and nanosize samples are similar and the temperature dependences exhibit anomalies at T_N temperature. The values of the Mn–O_1–Mn bond angles are larger for the nanoparticle samples.

For nano-samples the Jahn-Teller distortion parameter (JT) and MnO_6-octahedron distortion parameter (delta) are lowered in comparison to the polycrystalline sample.

References

[1] Dagotto E., *Nanoscale Phase Separation and Colossal Magnetoresistance*, Springer-Verlag, Berlin 2001.

[2] Kimura T., Goto T., Shintani H., Ishizaka K., Arima T., Tokura Y., *Magnetic control of ferroelectric polarizatoin*, Nature 426, 2003, 55-58.

[3] Lopez-Quintela M.A., Huesco L.E., Rivas J., Rivandulla F., *Intergranular magnetoresistance in nanomanganites*, "Nanotechnology" 14, 2003, 212-219.

[4] Dyakonov V., Szytuła A., Szymczak R., Zubov E., Szewczyk A., Kravchenko Z., Bażela W., Dyakonov K., Zarzycki A., Varyukhin V., Szymczak H., *Phase transitions in $TbMnO_3$*, Low Temperature Physics, 38, 1, 2012.

[5] Rodriguez-Carvajal J., *Recent advances in magnetic structure determination by neutron powder diffraction*, Physica B 192, 1993, 55-69.

[6] Bażela W., Dul M., Dyakonov V., Gondek Ł., Hoser A., Hoffmann J.-U., Penc B., Szytuła A., Kravchenko Z., Nosalev I., Zarzycki A., *Infuence of the grain size on the magnetic properties of $TbMnO_3$*, Acta Physica Polonica A, Vol. 121, No. 4, 2012, 785-788.

[7] Rasberry S.D., *Bureau of Standards Certicate-Standard Reference Material 640b*, 1987.

[8] Dul M., Bażela W., *The determination of crystal structure and grain size of $La_{0.7}Sr_{0.3}MnO_3$*, Czasopismo Techniczne (Technical Transations) 1-NP/2010, Issue 1, Year 107, 2010, p. 71-91.

[9] Williamson G.K., Hall W.H., *X-ray line broadening from filled aluminium and wolfram*, Acta Metallurgica 1, 1953, 22-31.

[10] Bażela W., Dul M., Dyakonov V., Gondek Ł., Hoser A., Hoffmann J.-U., Penc B., Szytuła A., Kravchenko Z., Nosalev I., Zarzycki A., *Magnetic and neutron diffraction studies of the polycrystalline and nanoparticle $TbMnO_3$*, Acta Physica Polonica A 122, 2012, 384-390.

[11] Bertaut E. F., *Spin configuration of ionic structures: theory and practice in: Magnetism*, Vol III, Eds. Rado G.T., Shul H., Academic Press, N.Y. 1963, p.149-209.

[12] Radaelli P.G., Iaonne G., Marezio M., Hwang H.Y., Cheong S.-W., Jorgensen J.D., Argyrion D.V., *Structural effects on the magnetic and transport properties of perovskite $A_{1-x}A'_xMnO_3$*, Physical Review B, 56, 1997, 8265-8276.

[13] Radaelli P.G., Marezio M., Hwang H.Y., Cheong S.-W., Batlogg B., *Charge localization by static and dynamic distortions of the MnO_6 octahedra in perovskite manganites*, Physical Review B, 54, 1996, 8992-8995.

[14] Bertaut E.F., *Representation analysis of magnetic structures*, Acta Crystallographica A24, 217, 1963.

[15] Goodenough J.B., *An interpretation of the magnetic properties of the perovskite – type mixed crystal* $La_{1-x}Sr_xCoO_{3-x}$, Journal of Physics and Chemistry of Solids, 6, 1958, 287-297.

[16] Kanamori J., *Superexchange interaction and symmetry properties of electron orbitals*, Journal of Physics and Chemistry of Solids, 10, 1959, 87-98.

[17] Brinks H.W., Rodrigues-Carvajal J., Fjellvag H., Kjaksus A., Hauback B.C., *Crystal and magnetic structure of orthorhombic* $HoMnO_3$, Physical Review B 63, 094411-094412, 2001.

[18] Senff D., Link P., Hradil K., Hiess A., Regnault L.P., Sidis Y., Aliouane N., Argyrion D.V., Braden M., *Magnetic excitations in multiferroic* $TbMnO_3$*: evidence for a hybridized soft mode*, Physical Review Letters, 98, 137206, 2007.

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

PIOTR ZABAWA*

NAMEDELEMENT REVISITED IN AN ASPECT-ORIENTED APPROACH

NOWE SPOJRZENIE NA NAMEDELEMENT W PODEJŚCIU ZORIENTOWANYM NA ASPEKTY

Abstract

In this paper a novel concept of adding structural responsibilities to meta-model classes for decreasing the meta-model complexity is introduced. This mechanism is supported by a combination of new Context-Driven Meta-Modeling Paradigm (CDMM-P) and its implementation in the form of the Context-Driven Meta-Modeling Framework (CDMM-F) with aspect-oriented paradigm and its AspectJ implementation supporting functionality and structure enrichment. The concept presented in the paper confirms the openness of CDMM-P and CDMM-F on the applicability of the aspect-oriented approach. It is also crucial for the process of generalization of notions introduced into the meta-model when a new modeling language is designed. It also helps to restructure the meta-model from the perspective of reusability. The NamedElement, known from many Object Management Group's (OMG) standards, was chosen.

Keywords: aspect oriented design, *aspect oriented programming*, *model, meta-model*, *meta-meta-model*, *responsibility*, *cross-cutting concern*, *dependency injection*, *inversion of control*

Streszczenie

W artykule wprowadzono nową koncepcję dodawania odpowiedzialności strukturalnych do klas metamodelu służącą zmniejszeniu jego złożoności. Mechanizm ten jest wspierany przez zestawienie nowego paradygmatu Context-Driven Meta-Modeling Paradigm (CDMM-P) i jego implementacji w postaci frameworku Context-Driven Meta-Modeling Framework (CDMM-F) z paradygmatem aspektowym i jego implementacją AspectJ wspierającą wzbogacanie funkcjonalności i struktury. Koncepcja prezentowana w artykule stanowi potwierdzenie otwartości CDMM-P i CDMM-F na możliwość stosowania podejścia aspektowego. Jest ona również kluczowa dla procesu uogólniania pojęć wprowadzanych do metamodelu podczas projektowania języka modelowania. Pomaga ono także w restrukturyzowaniu metamodelu z perspektywy ponownego użycia. Został wybrany NamedElement znany z wielu standardów Object Management Group (OMG).

Słowa kluczowe: projektowanie zorientowane na aspekty, *projektowanie aspektowe*, *programowanie zorientowane na aspekty*, *programowanie aspektowe*, *model*, *metamodel*, *metametamodel*, *odpowiedzialność*, *zobowiązanie przekrojowe*, *wstrzykiwanie zależności*, *odwrócenie sterowania*

DOI: 10.4467/2353737XCT.16.136.5715

* Piotr Zabawa (pzabawa@pk.edu.pl), Department of Physics, Mathematics and Computer Science, Cracow University of Technology.

1. Introduction

The paper is addressed to the NamedElement meta-model or meta-meta-model element, which is common to many well-known (meta-)meta-models. For convenience, the following notions are applied further in the paper:

- (meta-)meta-model or (m)mm denotes meta-meta-model or meta-model respectively
- mm denotes meta-model
- (m)mm denotes meta-meta-model
- (meta-)model or (m)m denotes meta-model or model respectively
- m denotes model
- (m)m is an instance of (m)mm, that is m is an instance of mm and mm is an instance of (m)mm
- s suffix denotes plural number of each notion above

The NamedElement can be met for example in (m)mms, like Meta-Object Facility (MOF) and its different realizations as well as in (m)ms, like Unified Modeling Language (UML) and Business Process Model and Notation (BPMN2). This (m)mm element is specific, because the responsibility it introduces into (m)mm affects many (m)mm elements. So, the nature of such common responsibility can be named cross-cutting structural responsibility or cross-cutting structural concern. Its responsibility is to enrich many (m)mm elements by the name represented in the form of a string. This way for example instances of classes or meta-classes or relationships may store their names in (m)ms.

Traditionally, the NamedElement class is introduced to (m)mms via a generalization relationship. However, this relationship is not supported directly by Context-Driven Meta--Modeling Framework (CDMM-F) [10–12] based on Context-Driven Meta-Modeling Paradigm (CDMM-P) introduced in [9] as the framework is located in data-layer. That is why the paper was introduced, just to explain how this kind of (m)mm elements may be introduced in the context of CDMM-F with the help of an aspect-oriented approach. The way of the element is introduced impacts on the features of the (m)mm as the whole.

The analogy between functional responsibilities and structural responsibilities which is referenced in the paper results from the observation that both responsibilities have a dual nature. Moreover, the problem of functional responsibilities is widely discussed in [1, 2, 7, 8] while the problem of structural responsibilities is almost completely ignored. However, it is crucial for meta-modeling domain as well as to data layer design. This paper is focused on the meta-modeling domain only.

2. Traditional Approach to NamedElement

As mentioned in section 1, the NamedElement is represented in the form of the class that contains one field of type string to store the name of the instance of this class. The NamedElement class is related to other mmm elements via a generalization relationship. If the NamedElement class is not abstract, then, in the case of the mm, the NamedElement instance is the model element which contains the name of model class. In the case of the mmm the NamedElement instance is the mm element which contains the

name of the mm class. Otherwise, if the NamedElement class is abstract, then, in the case of the mm, the NamedElement instance is the instance of the nearest concrete subclass of NamedElement class. This instance constitutes the model element which contains the name of the model class. In the case of the mmm the NamedElement instance is the instance of the nearest concrete subclass of NamedElement class. This instance constitutes the mm element, which contains the name of the mm class. NamedElement is represented in (m)mm by abstract classes.

The main problem with the traditional approach to creating (m)mms is the fact that their elements are interrelated at compile-time. So, the (m)mm graph is created during compilation process and not at run time. Thus, the relationships are static. This kind of interrelating mm elements influences the change introduction ease significantly. Different relationships interrelate classes differently. The weakest static relationship is UML dependency relationship, the stronger relationships are associative relationships, association (the weakest from this set), aggregation and composition (the strongest from this set). Unfortunately, the generalization relationship is the strongest one. And just this relationship is applied not only for NamedElement but for many other (m)mm classes. The popularity of this relationship was originated by knowledge modeling, where generalization is one of the most important relationships – it helps to build generalization hierarchies. However, in the software engineering domain the application of this relationship should be limited. And it is limited in target applications in many ways, like for example by application of design patterns. Nevertheless, in the meta-modeling domain it is promoted.

The approach discussed in the paper is different than the one presented above – it helps to interrelate (m)mm classes with their additional static responsibilities dynamically by injecting static responsibilities to (m)mm classes. The mechanism of injecting this kind of responsibilities is supported by aspect-oriented approach while (m)mm classes are managed by CDMM-F. The injecting concept and its applicability to (m)mms construction is discussed in section 3 while the role the CDMM-F plays in (m)mms definition is explained in section 4.

3. Structural Responsibility Injection

In contrast to the static compile-time concept of interrelating (m)mm classes presented above, this section is focused on application of the concept of interrelating (m)mm classes dynamically. Some consequences of the dynamic injection of relationships into (m)mms are also briefly discussed below.

The static responsibilities help to construct top hierarchies for (m)mms. They can be injected to (m)mms in order to address two modeling language designer needs – to introduce cross-cutting structural responsibilities for many existing (m)mm classes or to perform an mmm restructurization/refactorization process. The first need is typically planned from the very beginning of (m)mm defining process while the second need is usually involved by the observations made during the process of (meta-)modeling language definition. The characteristics of each need and its possible solution is presented in succeeding subsections below.

3.1. Cross-Cutting Structural Responsibility

The notion of cross-cutting concern is well known in Aspect-Oriented Design (AOD) and Aspect-Oriented Programming (AOP) [1–8], but it is related there to the functional concerns and is handled there by pointcuts and advices. However, at the same time AOD and AOP introduce the concept of enriching existing class hierarchies by classes interrelated statically to these hierarchies. Thus, for symmetry, the concept of enriching hierarchies statically can be seen from the perspective of static/structural concerns. Per analogy we have core structural concern, that is (m)mm to be enriched statically and other concerns. As the aspect-oriented approach fits to the concept of inversion of control (IoC) architectural pattern, all concerns both functional and static may be injected to the core concerns both functional and static respectively in the form that does not impact core concern's source code in any form. Moreover, the concerns are orthogonal, which means that one concern does not influence other any concern. In consequence, the static responsibilities can be added to (m)mm independently of each other.

It is worth noticing that there is also a notion of cross-cutting concerns in AOP. The cross--cutting concern is such a concern that crosses core-concern in a significant number of places. The more such places can be encountered, the more useful the IoC architectural pattern is. However, this pattern was applied so far for adding functional concerns, like error handling, system activity logging, auditing and many others. In this paper the same approach is applied to adding cross-cutting structural concerns. The NamedElement is a good example for such the cross-cutting concern as having the name is very common feature of (m)mm elements. Cross-cutting concerns are usually identified well before the meta-modeling process starts. However, they can be also added during this process in the case when they are recognized late. In section 3.2 the last cross-cutting structural responsibility addition is presented.

3.2. Meta-Model Refactorization

This section is focused on the (m)mm refactorization problem. A special case of the refactorization is involved by late recognition of a cross-cutting concern – this problem was characterized in section 3.1. But, usually the scope of the (m)mm refactorization for the purpose of structural responsibility addition is limited. As the consequence, both kinds of refactorization can be handled in the similar way although they have different purposes in the (meta-)modeling language design process. So, the same mechanism of adding structural responsibilities as the one described in section 3.1 can be applied for both forms of refactorization.

A simple example of such refactorization is presented in Figures 1 and 2. The UML class diagram for the mm structure before refactorization is presented in Figure 1 while the model after refactorization is depicted in Figure 2. The refactorization is limited to generalization of the fact, that both classes B1 and B2 have the same data field. The data field is thus moved to the new class T, which is aggregated both in B1 and B2.

Figure 2 presents a conceptual UML diagram, as it is informal for AOP. Nevertheless it reflects the fact of sharing common data field from class T well.

Fig. 1. Sample meta-model before structural refactorization

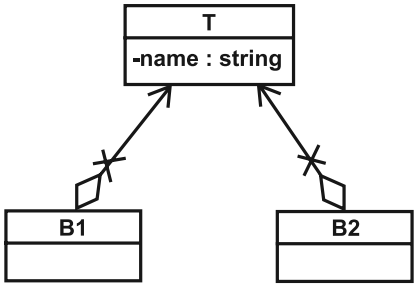

Fig. 2. Sample meta-model after structural refactorization in AOP approach – conceptual UML class diagram

The example from Figures 1–2 can be implemented in Java/AspectJ technologies in the form presented in Listings 1–2 respectively.

```
package pl.edu.pk.pz.aop.mm;

public class B1 {
  String name;
}
public class B2 {
  String name;
}
```

Listing 1. Java implementation of sample meta-model before refactorization

The classes that constitute structural responsibilities are located in the aspects layer (aspects) while the hierarchies to be enriched are placed in the class-object layer (classes). All these constructs are already available in AspectJ in the form of inter-type declarations (for modifying class hierarchies) and aspects (containers for all elements introduced by AspectJ to Java language).

In Listing 2 just the AspectJ AOP implementation was used to inject the T class as the default implementation of its IT interface. And this is a typical approach for this technology – classes are injected in the form of the relationships constructed from @DeclareParents annotation arguments.

One step more may be done in AOP – the aspects layer may be moved to the Spring framework and combined with AspectJ. However, the most important limit in the application of AOP to meta-modeling this way is connected to the fact that aspects are not instantiable (their lifecycle is synchronized with the lifecycle of the appropriate class instance in the best case, so they cannot exist without the class instance). As the result, the relationships represented by aspects do not have their instances. In consequence the relationships cannot

be differentiated, used or re-used between different (m)mms as separate entities. The CDMM approach is different as the relationships may have their instances as they are represented by classes. The core concept of CDMM-F is also based on the same mechanism as the one shown in Listings 1–2. However, the aspects in CDMM-F are used to interrelate (m)mm graph nodes represented by (m)mm entity classes by (m)mm edges represented by (m)mm relationship classes.

```
// meta-model classes
package pl.edu.pk.pz.aop.mm;

public class B1 {}
public class B2 {}

// top hierarchy package
package pl.edu.pk.pz.aop.th;

public interface IT {
  public String getName();
  public void setName(String name);
}

public class T implements IT {
  private String name;
  public String getName(){return name;}
  public void setName(String name){this.name = name;}
}

// aspects layer
package pl.edu.pk.pz.aop.aspect;

import org.aspectj.lang.annotation.Aspect;
import org.aspectj.lang.annotation.DeclareParents;

import pl.edu.pk.pz.aop.th.IT;
import pl.edu.pk.pz.aop.th.T;

@Aspect
public class B1 {
  @DeclareParents(value="pl.edu.pk.pz.aop.mm.B1",defaultImpl=T.class)
  public IT t;
}
@Aspect
public class B2 {
  @DeclareParents(value="pl.edu.pk.pz.aop.mm.B2",defaultImpl=T.class)
    public IT t;
}
```

Listing 2. Java/AspectJ implementation of sample meta-model after refactorization

The NamedElement class can be injected in place of class T to the class layer or to the classes defined in CDMM-F. However, the classes from the example in this section have different names than the ones presented in the context of CDMM in order to underline significant differences between AOP approach and CDMM approach. The specifics of the CDMM approach from the refactorization perspective is explained in section 4.

4. Structural Responsibility Injection in CDMM

It was shown in section 3 that direct application of AOP to the meta-model classes introduces an important limit – relationships are not represented in the form of classes but in the form of aspects. In consequence, (m)mms can be built in this approach from classes located in (m)mm graph nodes interrelated by relationships located in (m)mm edges, which are represented in the form of an aspect. So, this approach just supports the concept of modeling relationships between classes in the form of references. Moreover, this feature introduces asymmetry to this approach. As the result of this asymmetry, (m)mm graph nodes are reusable (classes and their instances) while (m)mm graph edges are not reusable (aspects without instances).

In contrast to the typical AOP approach presented above, relationships in CDMM are represented in the form of classes which are reusable. CDMM approach allows for injecting relationships as classes that represent relationships in place of injecting relationships into classes in the form of aspects. In consequence, the relationships play the role of structural responsibilities of the interrelated classes. This approach is symmetrical and more general than the one based on naive application of AOP paradigm. In the CDMM approach both classes and relationships exist independently of each other and they are interrelated at run--time by Spring application context XML file [12]. So, (m)mm graph node classes as well as (m)mm graph edge classes are subject of reuse between different (m)mms. Thus, the (m)mms constructed according to CDMM approach may be easily customized, changed, designed from scratch and each part of them can be easily reused.

The paper is focused on handling the problem of NamedElement handling in CDMM. It is worth noticing that the structural responsibilities can be injected with the help of aspect oriented approach to the CDMM based (m)mm. The same technique can be used for injecting NamedElement into (m)mm graph. The NamedElement may be seen as just another structural responsibility of the (m)mm graph node or edge classes – the responsibility of (m)mm element name storage.

The concept of introducing NamedElement into CDMM (m)mm graph with the help of aspect orientation is presented below in the form of the example similar to the one presented above. However, this example refers to Spring notions like beans and application context and it is related to CDMM-F (m)mm.

The same (m)mm as the one presented in Figure 1 was chosen to represent the (m)mm before refactorization. The result of the refactorization of this (m)mm is presented in Figure 3.

Both Figure 2 and Figure 3 contain conceptual diagrams – they are not formal as since 2001 the AOP is out of scope of the UML standard. The CDMM approach makes it possible to define any (meta-)modeling language and to generate the self-organizing MDA-like modeling tool for this language. This way the concept of automatic model-driven aspect-oriented software generating can be achieved without standardization of the (meta-)modeling language. This CDMM characteristic feature applies for any technology which is in scope or out of scope of MDA standards.

The most important elements of the source codes for the example from Figure 1 that implement (m)mm in CDMM are presented in Listing 3 and Listing 4. Java source code

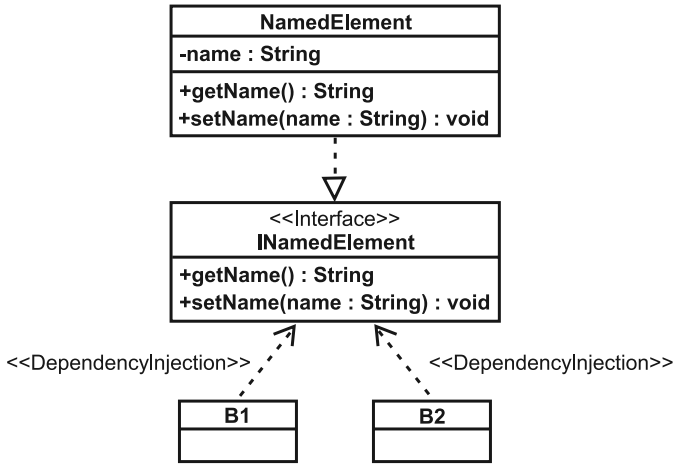

Fig. 3. Sample meta-model after structural refactorization in CDMM approach – conceptual UML class diagram

for two (m)mm classes is shown on Listing 3. The most important part of the CDMM-F's application context file for the (m)mm from Figure 1 is presented on Listing 4.

```
package com.componentcreator.metamodel.coremetamodel.domainsimpl;

public class DB1 extends BaseMetamodelCore implements IDB1 {
  String name;
}
public class DB2 extends BaseMetamodelCore implements IDB2 {
  String name;
}
```

Listing 3. Meta-model elements before refactorization

Now, a new (m)mm element is introduced. As a consequence of its introduction the name field migrates from B1 and B2 CDMM (m)mm classes to the new (m)mm element. This new element is just NamedElement and its definition is presented on Listing 5.

The structural responsibility represented in the paper by NamedElement can be added to some (m)mm elements via inclusion of the application context file presented in Listing 6 in the application context file from Listing 4. It was already stated that each such cross--cutting structural responsibility like NamedElement is orthogonal to the core concern – (m)mm and to other cross-cutting concerns. This orthogonality is represented by independence between responsibilities injected this way and by inclusion of extra application context.

The example presented above shows how to apply aspect oriented approach to enrich (m)mm defined in CDMM-F structurally. This approach to adding new concerns to (m) mm does not impact classes from (m)mm as long as this addition does not result from (m) mm refactorization. However, even in this case, the (m)mm change does not result from addition of new structural responsibility by aspects, but from the nature of the refactorization process itself. In the case of designing (m)mm and defining cross-cutting responsibilities

for it in advance, the addition of these responsibilities may be done in separation from the (m)mm's definition.

```
<beans>
  <!-- Root -->
  <bean class="com.componentcreator.metamodel.coremetamodel.root.RootMetamodelCore"
    id="root" scope="singleton"></bean>
  <!-- Root direct neighbours (collections) -->
  <bean class="com.componentcreator.metamodel.coremetamodel.domainsimpl.DB1"
    id="generalization" scope="prototype"></bean>
  <bean class="com.componentcreator.metamodel.coremetamodel.domainsimpl.DB2"
    id="class" scope="prototype"></bean>
  <!-- Responsibility implementations -->
  <!-- Root direct neighbours (collections of CPoliOMulti type) -->
  <bean class=
    "com.componentcreator.metamodel.coremetamodel.responsibilitiesimpl.RCollectionCPOM"
    id="collectionImplForRoot">
      <constructor-arg>
        <list>
          <value>com.componentcreator.metamodel.coremetamodel.domainsimpl.DB1</value>
          <value>com.componentcreator.metamodel.coremetamodel.domainsimpl.DB2</value>
        </list>
      </constructor-arg>
  </bean>
  <!-- Responsibility injections -->
  <aop:config>
    <aop:aspect id="holderA" ref="holderAAspect">
      <aop:declare-parents
        types-matching=
          "com.componentcreator.metamodel.coremetamodel.root.RootMetamodelCore"
        implement-interface=
          "com.componentcreator.metamodel.coremetamodel.responsibilities.IRCollectionCPOM"
        delegate-ref="collectionImplForRoot"/>
    </aop:aspect>
  </aop:config>
</beans>
```

Listing 4. Meta-model graph before refactorization

```
package com.componentcreator.metamodel.coremetamodel.metaontologies.namedelement;

public interface INamedElement {
  public void setName(String name);
  public String getName();
}
public class NamedElement implements INamedElement {
  private String name;
  @Override
  public void setName(String name) {
    this.name = name;
  }
  @Override
  public String getName() {
    return name;
  }
}
```

Listing 5. New meta-model element to be injected into meta-model graph (NamedElement)

```
<beans>
 <!-- NamedElement responsibility injections -->
 <aop:config>
  <aop:aspect>
   <aop:declare-parents
    default-impl=
     "com.componentcreator.metamodel.coremetamodel.metaontologies.namedelement.NamedElement"
    implement-interface=
    "com.componentcreator.metamodel.coremetamodel.metaontologies.namedelement.INamedElement"
    types-matching="com.componentcreator.metamodel.coremetamodel.domainsimpl.DB1" />
  </aop:aspect>
  <aop:aspect>
   <aop:declare-parents
    default-impl=
     "com.componentcreator.metamodel.coremetamodel.metaontologies.namedelement.NamedElement"
    implement-interface=
     "com.componentcreator.metamodel.coremetamodel.metaontologies.namedelement.INamedElement"
    types-matching="com.componentcreator.metamodel.coremetamodel.domainsimpl.DB2" />
  </aop:aspect>
 </aop:config>
</beans>
```

Listing 6. Meta-model graph after refactorization

5. Conclusions

This paper shows that the CDMM concept may be joined with other concepts applicable in software engineering domain. More specifically, it illustrates how the (meta-)meta--model graph implemented in CDMM-F can be enriched structurally by the application of AOP approach. The examples have shown that the most important advantages of AOP are preserved when the CDMM approach is used for (meta-)meta-model creation. Moreover, the combination of CDMM and AOP can be applied both for (meta-)meta-model refactorization as well as for the initial (meta-)meta-model design decisions.

The fact that AOP-oriented structural responsibilities can be injected into (meta-) meta-model results in very important feature of the presented combination of technologies – it introduces independence of life-cycles. The CDMM based (meta-)meta-model can be changed in large extent independently from changes introduced into AOP based structural responsibilities and vice versa. This feature helps to simplify and manage the process of designing modeling languages (meta-models) or designing languages used to define modeling languages (meta-meta-models).

References

[1] Filman R.E., Elrad T., Clarke S., Aksit M., *Aspect-Oriented Software Development*, 2004

[2] Gradecki J.D., Lesiecki N., *Mastering AspectJ: Aspect-Oriented Programming in Java*, First Edition, Wiley 2003

[3] Huang Sh.Sh., Smaragdakis Y., *Easy Language Extension with Meta-AspectJ*, Proceeding ICSE'06 Proceedings of the 28th International Conference on Software Engineering, p. 865-868, ACM, New York, NY, 2006.

[4] Kiczales G., Hilsdale E., Hugunin J., Kersten M., Palm J., Griswold W.G., *An overview of AspectJ*, In ECOOP'01: Proceedings of the 15th European Conference on Object-Oriented Programming, p. 327-353, London, UK, 2001, Springer-Verlag.

[5] Kiczales G., Lamping J., Menhdhekar A., Maeda C., Lopes C., Loingtier J.-M., Irwin J., *Aspect--oriented programming*, [in:] Akşit M., Matsuoka S., editors, *Proceedings European Conference on Object-Oriented Programming*, volume 1241, p. 220-242, Springer-Verlag, Berlin, Heidelberg and New York, 1997.

[6] Kojarski S., *Third-Party Composition of AOP Mechanisms*, Ph.D. thesis, Graduate School of Northeastern University, ProQuest LLC, 2008.

[7] Laddad R., *AspectJ in Action*, Second Edition, Enterprise AOP With Spring Applications, Manning Publications, Greenwich, 2010.

[8] Miles R., *AspectJ Cookbook*, First Edition, O'Reilly Media, 2004.

[9] Zabawa P., Stanuszek M., *Characteristics of the Context-Driven Meta-Modeling Paradigm (CDMM-P)*, Technical Transactions of Cracow University of Technology, 2014, vol. 111, No. 3-NP, p. 123-134.

[10] P. Zabawa, *Context-Driven Meta-Modeling Framework (CDMM-F) – Internal Structure*, 2016, submitted for publication.

[11] Zabawa P., Nowak K., *Context-Driven Meta-Modeling Framework (CDMM-F) – Simple Horizontal Case-Study*, 2016, submitted for publication.

[12] P. Zabawa, *Context-Driven Meta-Modeling Framework (CDMM-F) – Context Role*, Technical Transaction 1-NP/2015, p. 105-114, DOI: 10.4467/2353737XCT.15.119.4156

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

PIOTR ZABAWA*

THE SCOPE MANAGEMENT PROBLEM IN JAVA ENTERPRISE EDITION FRAMEWORKS

PROBLEM ZARZĄDZANIA ZAKRESEM WE FRAMEWORKACH JAVA ENTERPRISE EDITION

A b s t r a c t

The paper focuses on the problem of managing the scope understood as managing the multiplicity of elements that constitute the application context for Java Enterprise Edition (Java EE) frameworks. The subject of constructing graph modeling languages is the basis for scope management considerations. The problem can be demonstrated while the frameworks are superposed, which is necessary for meta-modeling compliant to the Context-Driven Meta-Modeling (CDMM) approach. The realization of the approach is based on Spring and AspectJ frameworks, which offer incompatible concepts of scope management. As part of the analysis the scope management problem in Java EE frameworks application context was identified, formulated, its area was defined and the sketch of the generalized concept of scope management elaborated and implemented by the author in relation to Java EE frameworks was presented..

Keywords: modeling language, *meta-model*, *graph*, *application context*, *java bean*, *java enterprise framework*, *Spring*, *aspect-oriented programming*, *AspectJ*

S t r e s z c z e n i e

Artykuł ten koncentruje się na problemie zarządzania zakresem rozumianym jako zarządzanie krotnościami elementów składających się na kontekst aplikacji we frameworkach Java Enterprise Edition (Java EE). Punktem odniesienia dla rozważań dotyczących zarządzania zakresem jest zagadnienie konstruowania grafowych języków modelowania. Problem ten ujawnia się przy składaniu ze sobą tych frameworków niezbędnym w meta-modelowaniu zgodnym z podejściem Context-Driven Meta-Modeling (CDMM). Jego realizacja oparta jest na frameworkach Spring i AspectJ, w których koncepcje zarządzania zakresem nie są zgodne. W ramach analizy zidentyfikowano problem zarządzania zakresem w kontekście aplikacji Java EE, sformułowano ten problem, określono jego zakres oraz zaprezentowano zarys opracowanej i zrealizowanej przez autora uogólnionej koncepcji zarządzania zakresem w odniesieniu do frameworków Java EE.

Słowa kluczowe: język modelowania, *metamodel*, *graf*, *kontekst aplikacji*, *ziarno java*, *framework java enterprise*, *Spring*, *programowanie aspektowe*, *AspectJ*

DOI: 10.4467/2353737XCT.16.137.5716

* Piotr Zabawa (pzabawa@pk.edu.pl), Department of Physics, Mathematics and Computer Science, Cracow University of Technology.

1. Introduction

Scope management in a broad sense comprises managing the number of instances during the process of constructing them, that is, at run-time. The conventional approach of programmers associates the responsibility of multiplicity determining mentioned above with a class. This is apparent, among other things, in singleton (anti)pattern. However, in Java EE frameworks this responsibility is moved to the framework. A bean multiplicity in the framework can be specified by the application context. The bean, on the other hand, reflects the way the framework (and in the consequence the software system implemented in the framework) perceives classes. The bean contains more information than the class, among other things just information about multiplicity of the bean. This additional information stored in the bean is specified in the application context file based on which the framework creates bean instances (and thus class instances). A particular class may occur once as the instance of one bean the scope of which is specified as singleton and, at the same time in the same application, the same class may occur multiple times as instances of other bean (associated to the same class) the scope of which is defined as prototype. In contrast to meta--models (modeling languages) constructing this solution turns out not to be sufficient due to the high complexity of graph meta-models. Also relating the scope to the bean only turns out not to be sufficient while applying it to graph modeling languages. That is why the need to enrich the current mechanism occurred.

In scientific papers [12] as well as in the IT industry literature [5] and in industry standards [6, 13, 19] meta-models are created statically – modeling languages are defined at compile time. However, as research results achieved by the author show [20], it is possible to define modeling languages at run-time. The application context mentioned above can be used to specify graph-like interrelations between language elements.

Further in the paper it is shown that when the scope notion is addressed to modeling languages constructed at run-time, this notion should be addressed both to Java EE beans and to classes. Moreover, bean sets as well as sets of classes involved with relationships interrelating particular bean sets play an important role in meta-modeling.

A characteristic feature of the CDMM approach [21] is constructing meta-model graph from elements consisting of meta-model entity classes and meta-model relation classes. The graph is constructed from Java EE beans defined for these classes, thus from entity beans and relation beans. Entity beans are placed in graph nodes while relation beans are placed in graph edges. In such an approach the application context XML file constitutes the definition of the meta-model graph. However, in such approach the correct management of relation instances quantity during relation beans injections into entity beans is an important problem. It is especially evident with reference to N-ary relationships [4, 8, 9, 14, 15, 17], relations that join more than two graph nodes at the same time. In the case of such relations the mechanism of injecting the same relation object (relationship bean instance or relation class instance) to all nodes involved with this relation must be provisioned. It appears, however that the possibilities offered by Java EE frameworks are not sufficient in the area of multiplicity management, as they are focused on management of multiplicities of singular beans only. It is worth noticing that the implementation of N-ary relations and the so called "arity problem" is difficult while constructing graph modeling languages. It is shown by

documented problems visible in Object Management Group (OMG) standards, like Meta--Object Facility (MOF) – the definition of N-ary association was omitted here because of too much difficulty [1, 13], then the implementation of this relationship as a separate notion in Unified Modeling Language (UML) standard was retired [19] (it is represented on the diagramming and not on the modeling level, so the code cannot be generated from UML modeling tools) [7]. The root cause for these problems and limits is the lack of representation for relationships in all sources known from scientific literature, IT industry publications and software modeling tools documentation [3, 10, 11, 18] However, these problems can be solved in CDMM technology as the relations have their representation in it.

It should be pointed out that the scope management problem with reference to the CDMM technology concerns such meta-model elements only which are involved in representing relations, so they play the role of edges of the graph being the representation of a modeling language. Edge (meta-model relation) classes play the role of static responsibilities for node (meta-model entity) classes. These responsibilities are injected to entity classes as default implementations of interfaces of these relation classes with the help of dependency injections (Spring) and with the help of aspect-oriented inter-type declarations (AspectJ).

2. Scope Management in Spring Framework

The Spring framework offers scope management limited to the Spring beans. The bean is the way Spring as well as the Spring-based application (more generally – a software system), sees Java POJO classes. The object model in Spring is enriched in comparison to Java object model by many attributes that can be associated to beans. One of them is the "scope" attribute of a bean. According to the documentation of Spring framework [16] the scope attribute can have one of the following values: "singleton", "prototype". The "singleton" attribute informs Spring that the bean with this attribute value can have exactly one instance – the bean and not the POJO class behind the bean. The "prototype" attribute informs that the bean with this attribute can be multiplied as needed.

Static information about beans is defined in Spring application context XML file. As long as bean instances are created from the application through the Spring Application Programming Interface (API) the solution offered by the framework is sufficient.

When the instance of a particular bean is created from the Spring-based application through the API of Spring the constructor of the class which is behind the bean is called by default. However, Spring offers also another mechanism, which is applied in the approach presented in the paper. The bean instances may be created through factories. This approach is much more flexible and was originally added to Spring to simplify application of creational design patterns.

3. Scope Management Problems in AspectJ with Spring

The situation described in section 2, when Spring is used as the only framework and when bean instances are created from the Spring-based application is simple and does not trigger any problems. However, when the Spring is superposed with other framework and

this additional framework influences or even takes control over bean instances creation process, some problems appear. They result from the fact that the additional framework may take responsibility for bean instance creation from the Spring-based application to the additional framework. Moreover, the additional framework may delegate this responsibility back to Spring and to application context. And this is the case when Spring is superposed with AspectJ [2].

The Spring framework is integrated to Aspect Oriented Programming (AOP) via two Spring sub-projects: SpringAOP [16] and Spring+AspectJ [16]. The first one constitutes a limited implementation of AOP concepts and is not sufficient for the CDMM-F implementation. However, the second project offers full AspectJ functionality and is sufficient for the application of the CDMM concept. The rest of the paper is limited to the full integration of Spring with AspectJ.

The implementation of CDMM-F is based on extensive usage of AOP concept applicable to influencing class hierarchies, thus inter-type declarations, and more specifically, declare-parents construct. This way the relationship classes of a meta-model can be injected as default interface implementations to particular meta-model entity classes as their structural responsibilities (in contrast to dynamic responsibilities, which are more typical). The method for such injections is specified in Spring+AspectJ application context file according to the sample code presented in Listing 1.

```
<!– Meta-Model Entity Beans (Spring) -->
<bean
  class=”com.componentcreator.metamodel.coremetamodel.domains.DEntity”
  id=”entity”
  scope=”prototype”>
</bean>

<!– Meta-Model Relation Beans (Spring) -->
<bean
  class=”com.componentcreator.metamodel.coremetamodel.relations.RRelation”
  id=”relImplForDEntity”>
</bean>

<!– Meta-Model Graph Creation (Spring+AspectJ) -->
<!– Meta-Model Relation Injections to Meta-Model Entities -->
<aop:declare-parents
  types-matching
    =”com.componentcreator.metamodel.coremetamodel.domainsimpl.DEntity”
  implement-interface
    =”com.componentcreator.metamodel.coremetamodel.relations.IRRelation”
  delegate-ref=” relImplForDEntity”/>
```

Listing 1. Meta-model elements defined in Spring and their injections defined in Spring+AspectJ application context file (extract only)

It is clear from the Listing 1 that meta-model entity beans have their scope defined as “prototype” while the attribute is ignored for relationship beans. It is not specified in application context file to underline the fact that AspectJ ignores this attribute for beans it injects.

When the Spring integrated to AspectJ loads an application context file that contains such injections, the default implementations of interfaces are created as Spring beans. This

behavior influences and destroys the original Spring concept of scope management. It is even impossible to change the Spring+AspectJ behavior from the bean "scope" parameter – from its predefined as well as from its user-defined version. The Spring interpretation of the "scope" bean parameter is completely overlapped by AspectJ. But, fortunately, Spring+AspectJ create injected beans of default implementation classes for each such injection. Moreover, the AspectJ mechanism does not overwrite the option of calling factories in place of constructors when a bean is instantiated. It is shown further in the paper that combining both mentioned features helps to take full control over the instantiating process when meta--model is created.

4. Scope Management Problem

This section is focused on two goals – showing how the control over scope management (introduced intuitively before) can be regained in case of overlapping incompatible solutions offered by different Java EE frameworks and presenting the skeleton of the concept of advanced scope management for meta-modeling purposes.

In order to address the two goals mentioned above, the scope management problem should be clearly stated and then its solution can be presented. At the end the correctness of the solution should be verified. All these stages are presented below.

4.1. Problem statement

The scope management problem is the problem of controlling the multiplicity of application elements while their construction process driven by Java EE application context under the assumption that the application context file is interpreted by more than one Java EE framework.

As the consequence we have the following situation – the superposition of frameworks:

$$F = F1 \circ F2 \circ \ldots \circ FN$$

where:

F – the framework created as the result of superposition of other frameworks,
F1 – FN – superposed frameworks.

The problem is at least two-dimensional as it concerns both classes and their beans. The problem of the actual dimension is discussed in section 4.2. The size of the problem does not depend of the number of frameworks *F1 – FN*.

The problem is limited to meta-model relation beans and classes.

The problem can be solved if the following conditions are fulfilled:

– *FN* framework tries to construct application elements whenever needed
– *FN* framework does not eliminate the ability to access factories for application elements construction purpose

Topological aspects only are taken into account in the paper. This means that such problems like cardinalities of meta-model relationships (meta-cardinalities) as well as the problem resulting from the above – the problem of existence of some nodes at the meta-model relation ends are ignored in the paper. The problem of meta-model relationship cardinalities which is new and separate from the scope management problem is intended for future publications.

4.2. Problem solution

It was mentioned before that scope may be addressed to beans and/or classes specified in the application context. Another observation related to Spring scope management is that the concept of scope management is related to the whole application. This means that the particular scope associated to a particular bean defines the multiplicity of the bean instances in the whole application. However, in the meta-modeling problem the range of the scope should be differentiated to such areas like meta-model, context file, constructor.

As the result, in the meta-modeling problem, the following dimensions of the scope management problem should be assumed:

Subject (relationship class, relationship bean)

Scope (meta-model, context file, constructor)

Thus, the name of scope fits better to the true meaning of this notion.

For each combination of the above elements, for each pair (Subject, Scope) the element of the framework F which is responsible for scope management should be identified. So, the divagations should be enriched by the following mapping:

$$(\text{Subject} \times \text{Scope}) \rightarrow \text{ScopeManager}$$

where:

$$\text{ScopeManager} = \{\text{class, bean, context, framework}\} \subset F$$

The communication between framework F and the right ScopeManager is controlled by factories that are called while constructing application elements. The special case is when the factory does not delegate the scope management responsibility to dedicated ScopeManager but takes this responsibility. This assumption was assumed in the rest of the paper for simplification. As the result, the naming convention for factories, which in consequence of this assumption can be predefined in F, can be introduced. The naming convention may be as follows:

`Responsibility<Subject><Scope><Manager>ScopeFactory,`

where, under the above assumption <Manager>=Factory $\subset F$

In consequence, the names of such factories are as the ones contained in Table 1.

Table 1

The names of factory classes which are responsible for managing meta-model relationclasses

Scope \ Subject	**Class**	**Bean**
Metamodel	ResponsibilityClassMetamodel ScopeFactory	ResponsibilityBeanMetamodel ScopeFactory
Context file	ResponsibilityClassFile ScopeFactory	ResponsibilityBeanFile ScopeFactory
Constructor	ResponsibilityClassConstructor ScopeFactory	ResponsibilityBeanConstructor ScopeFactory

The ResponsibilityBeanConstructorScopeFactory class is sufficient to solve the arity problem. That is why the nature, implementation and verification of just this class is discussed further in this section as the illustration of the factories implementation concept.

Two variants are taken into account below to characterize the nature of ResponsibilityBeanConstructorScopeFactory class. The simple case is presented first (one relation for a particular set of entities). Then the complex case (many relations for a particular set of entities) is shown. The problem of number of relations in meta-model has not been identified and has not been investigated before. The name suggested by the author for this problem is meta-cardinality and it is related to the CDMM system of notions. However, this problem is discussed in a separate paper. The two cases mentioned above are defined for:

– a particular relation (for a particular bean of a relation class) joining a set of entity classes – one bean instance is created by the factory
– many relations of the same kind (represented by the same bean of a relation class) joining a set of entity classes – the number of bean instances created is equal to the number of relations.

More generally speaking, for a particular set of constructors of any number of a relation beans (for the same relation class) the number of instances of this bean is equal to the number of beans and not to the number of the bean class injections to the set of entity classes.

The characteristics of ResponsibilityBeanConstructorScopeFactory class can be referenced to Figure 1 and Listing 3 in section 4.2.

In the next research stage all possible combinations of relation construction cases were identified for the meta-modeling application domain. These observations have theoretical nature (all cases were identified for consideration completness).

The following notational system was designed to specify scope in the application context file:

– CDMMFsubject (applied in each bean to determine if the scope is related to the bean or to its class)
– CDMMFscope (applied in each bean to define the scope for CDMMFsubject)
– CDMMFmanager (applied in each bean to define the element responsible for the scope management for this bean)
– The following comments are related to the system of tags introduced above:
– CDMMFmanager may be optional (if we assume that the scope management is default)
– CDMMFmanager may not be required if the right class will be determined by the pair (CDMMFsubject, CDMMFscope)
– as long as any Java EE framework *F* has its notation related to scope management the CDMM prefix is required

The implementation of the ResponsibilityBeanConstructorScopeFactory scope manager is presented in Listing 1.

```
public class ResponsibilityBeanConstructorScopeFactory implements
  IResponsibilityBeanScopesFactory {

  private static Map<String, IResponsibility> relationshipMinimal
    = new HashMap<String, IResponsibility>();

  public IResponsibility getInstanceMinimal(String beanId, String cls,
    List<String> str) throws NoSuchMethodException, SecurityException,
    ClassNotFoundException, InstantiationException, IllegalAccessException,
    IllegalArgumentException, InvocationTargetException {
      // the instance of the beanId was already created
      if (relationshipMinimal.containsKey(beanId)) return
        relationshipMinimal.get(beanId);
      // the beanId has not been created yet
      else {
        // create the instance of cls object passing it str parameters
        // - Java reflection needed here
        relationshipMinimal.put(beanId, (IResponsibility)
          ResponsibilityBeansRegister.get(cls).getConstructor(new Class[]
          {List.class}).newInstance(new Object[] { str }));
        return relationshipMinimal.get(beanId);
      }
  }
}
```

Listing 2. Scope management factory class dedicated to N-ary relationship instance multiplicity handling

The factory implemented in the form presented on the Listing 1 works as follows. The meta-model relation bean (represented by beanId in the source code) is defined in the application context file for Spring Java EE. Then the relation bean is injected by AspectJ framework when the default interface implementation of a relation class is associated to a meta-model entity class. In place of constructor the method getInstanceMinimal() is called with the following parameters: beanId equal to the Id of relation bean, cls equal to the pathname of the relation class, str equal to the list of pathnames of entity classes the relation bean is injected to. The method determines if the object was already constructed for the set of parameters (beanId, cls, str) and creates it or returns the reference to already existing bean instance.

4.3. Verification

The concept of scope management was tested for the superposition of Spring and AspectJ frameworks. This combination of frameworks is sufficient for obtaining the superposition with required features as defined in section 4.1. This superposition of just these frameworks is also good enough for defining sufficiently complex meta-models.

The correctness of the approach presented in the paper was verified in three following stages:

- all factory classes presented in Table 1 were implemented,
- appropriate meta-models were defined to generate all test cases (at least one test case was needed to test each factory class),
- appropriate unit tests were implemented to test each test case resulting from meta-models defined above.

All test case executions confirmed the correctness of both the approach and the implementation of all factories dedicated to support meta-modeling. It is worth noticing that the elaboration of all meta-model concepts required to implement test cases for each factory class was especially demanding and time consuming. This complexity resulted from the fact that in this case the special meta-modeling problems should be invented to check the correctness of the solutions which were foreseen before during theoretical research. This approach was abnormal as usually the problem appears first and the solution comes later.

As the illustration of the use of the factory for a sample meta-model is presented in Figure 1 and then the extract from the application context file is shown.

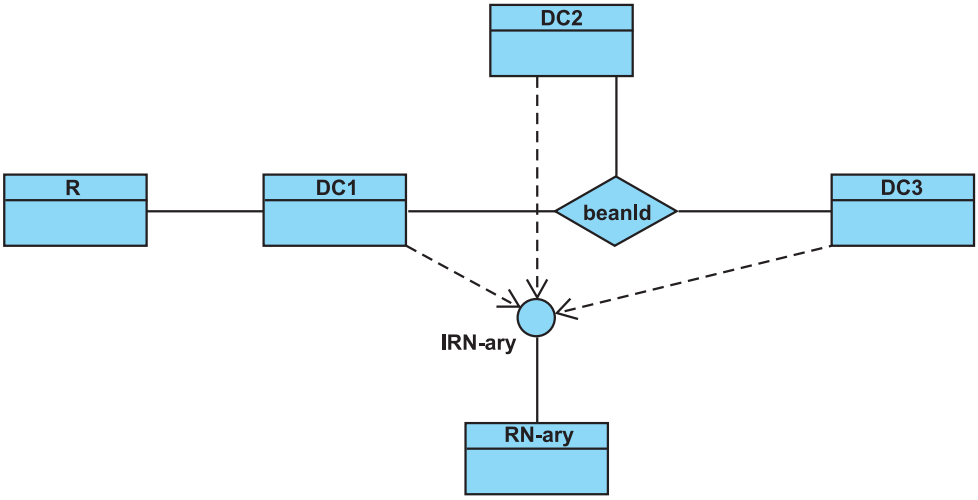

Fig. 1. Sample meta-model for the N-ary relationship

The way the factory is specified in the application context file and how it is associated to the RN-ary bean is clarified in Listing 3.

```
<bean
  class="com.componentcreator.metamodel.corremetamodel.scopefactories
  .ResponsibilityBeanConstructorScopeFactory"
  id=" responsibilityBeanConstructorScopeFactory " scope="singleton"></bean>

<bean class="com.componentcreator.metamodel.corremetamodel.relations.RNary"
  id="naryImpl"
  factory-bean="responsibilityBeanConstructorScopeFactory"
  factory-method="getInstanceMinimal">
    <constructor-arg>
      <value>"naryImpl"</value>
    </constructor-arg>
    <constructor-arg>
      <value>
        "com.componentcreator.metamodel.corremetamodel.relations.RNary"
      </value>
    </constructor-arg>
    <constructor-arg>
      <list>
        <value>
          com.componentcreator.metamodel.corremetamodel.domainsimpl.DC1
        </value>
        <value>
          com.componentcreator.metamodel.corremetamodel.domainsimpl.DC2
        </value>
        <value>
          com.componentcreator.metamodel.corremetamodel.domainsimpl.DC3
        </value>
      </list>
    </constructor-arg>
</bean>
```

Listing 3. Meta-model scope factory and relation beans specification in the application context file

5. Conclusions

The scope management problem was identified for meta-modeling purposes. The meta-modeling application domain as defined by CDMM approach is complex enough to study the problem. The concept of the scope management solution was also implemented in CDMM-F with the help of appropriate factories. Then the solution correctness was verified by appropriate test cases.

The paper initiates further research efforts in the field of scope management by creating solid fundamentals and presenting the skeleton of the solution for the next problems related to scope management. The mentioned problems are named and characterized briefly below.

Several interesting subjects for research are connected to meta-cardinality (the problem of defining the number of relation instances). This problem is very complex and is not supported by currently available technologies.

Another interesting problem which is new for meta-modeling and modeling disciplines is the problem of navigability of meta-model relationships named by the author meta-navigability. This problem is connected to traversing the directed graph of modeling language and impacts CDMM-F API significantly.

Also a complex problem of combining scopes may appear when several application context files that are based on different scope management concepts are used (reused) to constitute

the whole meta-model application context. In the paper a simple case is implemented (see relationshipMinimal), but the concept of relationshipRedundant was also designed (but not verified yet) to support future solution of the scope combining problem.

Other challenging problems are connected to the so-called arity problem. The N-ary relationships can be handled in CDMM-F but in order to gain the full solution of the problem the meta-cardinality and meta-navigability problems must be completely solved and published.

References

[1] Akehurst D., Howells G., McDonald-Maier K., *Implementing associations: UML 2.0 to Java 5*, Softw Syst Model, Springer-Verlag 2006, DOI 10.1007/s10270-006-0020-1

[2] AspectJ framework, https://eclipse.org/aspectj/.

[3] Bildhauer D., *On the relationship between subsetting, redefinition and association specialization*, [in:] Proc. of the 9th Baltic Conference on Databases and Information Systems 2010, Riga, Latvia (07 2010).

[4] Bildhauer D., *Associations as First-class Elements*, Proceedings of the 2011 conference on Databases and Information Systems VI: Selected Papers from the Ninth International Baltic Conference, DB&IS 2010, p. 108-121, IOS Press Amsterdam, The Netherlands, The Netherlands 2011.

[5] Booch G., Rumbaugh J., Jacobson I., *The Unified Modeling Language User Guide*, Addison--Wesley, 2005.

[6] Object Management Group (2011), Business Process Model and Notation 2.0. http://www.omg.org/spec/BPMN/2.0/.

[7] Diskin Z, Easterbrook S., Dingel J., *Engineering Associations: From Models to Code and Back through Semantics*, [in:] *Objects, Components, Models and Patterns*, Volume 11, Lecture Notes in Business Information Processing, Proceedings of 46th International Conference, TOOLS EUROPE 2008, Zurich, Switzerland, June 30–July 4, 2008, p 336-355.

[8] Feinerer I., *A Formal Treatment of UML Class Diagrams as an Efficient Method for Configuration Management*, PhD. dissertation, Vienna, March 2007.

[9] Feinerer I., Salzer G., *Numeric semantics of class diagrams with multiplicity and uniqueness constraints*, Software & Systems Modeling, 13(3), 2014, p. 1167-1187.

[10] Génova, G., Lloréns, J., Martínez, P., *The meaning of multiplicity of N-ary associations in UML*, Software and System Modeling 1(2), 2002, 86-97.

[11] Génova G., Ruiz del Castillo C., Llorens J., *Mapping UML Associations into Java Code*, Journal of Object Technology, Vol. 2, No. 5, September–October 2003.

[12] Kleppe A. G., Warmer J., Bast W., *MDA Explained: The Model Driven Architecture: Practice and Promise*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2003.

[13] Object Management Group (2006), Meta Object Facility (MOF) core specification version 2.0. http://www.omg.org/spec/MOF/2.0/.

[14] Roques P., SysML vs. UML 2: A Detailed Comparison, MoDELS'11 Tutorial, October 16th, Wellington, New Zealand, 2011.

[15] Sergievskiy M., *N-ary Relations of Association in Class Diagrams: Design Patterns*, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016.

[16] Spring framework, https://spring.io/.

[17] Szlenk M., *Formal Semantics and Reasoning about UML Class Diagram*, 2006 International Conference on Dependability of Computer Systems, IEEE, 25–27 May 2006, p. 51-59, DOI: 10.1109/DEPCOS-RELCOMEX.2006.27.

[18] Tan H.B.K., Yang Y., Bian L., *Improving the Use of Multiplicity in UML Association*, Journal of Object Technology, Vol. 5, No. 6, July–August 2006.

[19] Object Management Group (2009), Unified Modeling Language (UML) superstructure version 2.2, http://www.omg.org/spec/UML/2.2/.

[20] Zabawa P., *Context-Driven Meta-Modeling Framework (CDMM-F) – Context Role*, Technical Transactions, 1-NP/2015, p. 105-114, DOI: 10.4467/2353737XCT.15.119.4156.

[21] Zabawa P., Stanuszek M., *Characteristics of Context-Driven Meta-Modeling Paradigm (CDMM-P)*, Technical Transactions of Cracow University of Technology, Fundamental Sciences, 3-NP (111), 2014, p. 123-134.

MATHEMATICS

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

MONIKA HERZOG*

APPROXIMATION THEOREMS FOR SZÁSZ-MIRAKJAN--DURRMEYER TYPE OPERATORS

TWIERDZENIA APROKSYMACYJNE DLA OPERATORÓW TYPU SZÁSZA-MIRAKJANA-DURRMEYERA

A b s t r a c t

In this paper we study an integral modification of Szász-Mirakjan type operators. The modification will be called Szász-Mirakjan-Durrmeyer type operators as in many papers examining this type of operators. We give direct approximation theorems for these operators using the modulus of continuity and the modulus of smoothness for functions belonging to exponential weighted spaces.

Keywords: *linear positive operators*, *Bessel function*, *modulus of continuity*, *degree of approximation*

S t r e s z c z e n i e

W artykule badamy całkową modyfikację operatorów typu Szásza-Mirakjana. Tę modyfikację będziemy nazywać operatorami typu Szász-Mirakjan-Durrmeyera, jak to się czyni w wielu pracach badających tego typu operatory. Podajemy twierdzenia aproksymacyjne wykorzystujące moduł ciągłości i moduł gładkości dla funkcji z wykładniczych przestrzeni wagowych.

Słowa kluczowe: *dodatni operator liniowy*, *funkcja Bessela*, *moduł ciągłości*, *rząd aproksymacji*

DOI: 10.4467/2353737XCT.16.138.5717

* Monika Herzog (mherzog@pk.edu.pl), Institute of Mathematics, Faculty of Physics, Mathematics and Computer Science, Cracow University of Technology.

1. Introduction

In paper [6] we investigated operators of the Szász-Mirakjan type defined as follows

$$L_n^{\nu}(f;x)=\begin{cases}\sum_{k=0}^{\infty} p_{n,k}^{\nu}(x) f\left(\dfrac{2k}{n+q}\right) & x>0;\\ f(0) & x=0\end{cases} \tag{1}$$

where the coefficients

$$p_{n,k}^{\nu}(x)=\frac{1}{I_{\nu}(nx)}\frac{x^{2k+\nu}}{2^{2k+\nu}\,\mathrm{k}!\Gamma(k+\nu+1)} \tag{2}$$

Γ is the gamma function and I_{ν} the modified Bessel function of the first kind defined by the formula ([15], p. 77)

$$I_{\nu}(z)=\sum_{k=0}^{\infty}\frac{z^{2k+\nu}}{2^{2k+\nu}\,\mathrm{k}!\Gamma(k+\nu+1)}$$

This means that we replaced the coefficients of well-known Szász-Mirakjan operators by some terms involving the modified Bessel function I_{ν}.

We studied the approximation properties of these operators in exponential weight spaces

$$E_q=\{f\in C(\mathbb{R}_0): w_q f \quad \text{is uniformly continuous and bounded on } \mathbb{R}_0\},$$

where $C(\mathbb{R}_0)$ denotes the space of all real-valued function continuous on $\mathbb{R}_0=[0;\infty)$ and w_q is the exponential weight function defined as follows

$$w_q(x)=e^{-qx},\quad q\in\mathbb{R}_0 \tag{2}$$

for $x\in\mathbb{R}_0$.

In the spaces we introduced the weighted norm

$$\left\|f_q\right\|=\sup\left\{w_q(x)\left|f(x)\right|: x\in\mathbb{R}_0\right\} \tag{3}$$

and we established ([6], Theorem 2.1) that operators L_n^{ν} are linear, positive, bounded and transform the space E_q into E_q.

In this paper we introduce an integral modification of (1)

$$\tilde{L}_n^{\nu}(f;x)=\begin{cases}\sum_{k=0}^{\infty} p_{n,k}^{\nu}(x)\int_0^{\infty}\tilde{g}_{n,k}^{\nu}(t)dt, & x>0;\\ f(0), & x=0\end{cases} \tag{4}$$

where the coefficients $p_{n,k}^{\nu}$ are defined above and

$$\tilde{g}_{n,k}^{\nu}(t)=\frac{n+q}{\Gamma(2k+\nu+1)}e^{-(n+q)t}((n+q)t)^{2k+\nu}$$

The idea of integral modifications of this kind of operators comes from J.L. Durrmeyer ([2]) who introduced the integral modification of the genuine Bernstein operators. Later on new modifications of other classical operators appeared, for example, M.M. Derriennic ([3]), S.M. Mazhar and V. Totik ([11]), A. Sahai and G. Prasad ([13]), M. Heilmann ([5]). Now the operators are still under consideration [1, 4, 7–10, 12, 14].

The note was inspired by the above results which investigate approximation problems for integral operators and it is a natural continuation of the author's results from paper [7].

Among other things, in the paper we shall prove the theorems giving the degree of approximation of functions from E_q by operators $\tilde{L}_n^{\nu}$. We will estimate the error of approximation using the weighted modulus of continuity of the first and the second order defined as follows

$$\omega_1(f, E_q; t) = \sup\left\{\left\|\Delta_h f\right\|_q : h \in [0,t]\right\}, \quad t > 0 \tag{5}$$

and

$$\omega_2(f, E_q; t) = \sup\left\{\left\|\Delta_h^2 f\right\|_q : h \in [0,t]\right\}, \quad t > 0$$

respectively, where

$$\Delta_h f(x) = f(x+h) - f(x), \quad \Delta_h^2 f(x) = f(x+2h) - 2f(x+h) + f(x)$$

for $x, h \in \mathbb{R}_0$.

It is worth mentioning that Bessel functions are the most important special functions which play a pivotal role in mathematical physics, for example: signal processing, heat conduction, diffusion problems. We hope that the operators examined will have applications to these areas of study.

Remark 1.1

In the paper we shall denote by $M(p, t)$ suitable positive constants depending on the parameters indicated p, t.

2. Auxiliary results

Let us denote

$$e_r(t) = t^r,\ f_r(t) = e_r(t)e^{qt},\ \phi_{x,r}(t) = (t-x)^r,\ \psi_{x,r}(t) = \phi_{x,r}(t)e^{qt}$$

for $r \in \mathbb{N}_0 := \{0\} \cup \mathbb{N}$, $q, x \in \mathbb{R}_0$.

In this section we shall recall preliminary results which are immediately obtained from papers [6, 7] and definition (4).

Remark 2.1

For all $\nu \in \mathbb{R}_0$ and $n, r \in \mathbb{N}$ it holds

$$\tilde{L}_n^{\nu}(e_0; 0) = 1, \quad \tilde{L}_n^{\nu}(f_0; 0) = 1$$

$$\tilde{L}_n^{\nu}(e_r;0)=\tilde{L}_n^{\nu}(\phi_{0,r};0)=\tilde{L}_n^{\nu}(\psi_{0,r};0)=\tilde{L}_n^{\nu}(f_r;0)=0$$

Lemma 2.1 ([6], Lemma 2.1)

For each $\nu\in\mathbb{R}_0$ *there exists a positive constant* $M(\nu)$ *such that for all* $n\in\mathbb{N}$ *and* $x\in\mathbb{R}_0$ *we have*

$$\left|\frac{I_{\nu+1}(nx)}{I_{\nu}(nx)}\right|\le M(\nu),\quad nx\left|\frac{I_{\nu+1}(nx)}{I_{\nu}(nx)}-1\right|\le M(\nu)$$

By elementary calculations and Lemma 2.2. ([6]) we get

Lemma 2.2

For each $n\in\mathbb{N}$, $\nu, q\in\mathbb{R}_0$ and $x\in\mathbb{R}_0$

$$\tilde{L}_n^{\nu}(e_0;x)=L_n^{\nu}(e_0;x)=1,\quad \tilde{L}_n^{\nu}(e_1;x)=L_n^{\nu}(e_1;x)+\frac{\nu+1}{n+q}=\frac{n}{n+q}\left(\frac{xI_{\nu+1}(nx)}{I_{\nu}(nx)}+\frac{\nu+1}{n}\right),$$

$$\begin{aligned}\tilde{L}_n^{\nu}(e_2;x)&=L_n^{\nu}(e_2;x)+\frac{2\nu+3}{n+q}L_n^{\nu}(e_1;x)+\frac{(\nu+1)(\nu+2)}{(n+q)^2}\\&=\left(\frac{n}{n+q}\right)^2\left(\frac{x^2I_{\nu+2}(nx)}{I_{\nu}(nx)}+\frac{(2\nu+5)}{n}\frac{xI_{\nu+1}(nx)}{I_{\nu}(nx)}+\frac{(\nu+1)(\nu+2)}{n^2}\right),\end{aligned}$$

$$\tilde{L}_n^{\nu}(\phi_{x,1};x)=L_n^{\nu}(\phi_{x,1};x)+\frac{\nu+1}{n+q}=x\left(\frac{n}{n+q}\frac{I_{\nu+1}(nx)}{I_{\nu}(nx)}-1\right)+\frac{\nu+1}{n+q},$$

$$\begin{aligned}\tilde{L}_n^{\nu}(\phi_{x,2};x)&=L_n^{\nu}(\phi_{x,2};x)+\frac{2\nu+3}{n+q}L_n^{\nu}(\phi_{x,1};x)+\frac{x}{n+q}\frac{(\nu+1)(\nu+2)}{(n+q)^2}\\&=x^2\left(\left(\frac{n}{n+q}\right)^2\frac{I_{\nu+2}(nx)}{I_{\nu}(nx)}-\frac{2n}{n+q}\frac{I_{\nu+1}(nx)}{I_{\nu}(nx)}+1\right)\\&\quad+\frac{2(\nu+1)x}{n+q}\left(\frac{n}{n+q}\frac{I_{\nu+1}(nx)}{I_{\nu}(nx)}-1\right)+\frac{3nx}{(n+q)^2}\frac{I_{\nu+1}(nx)}{I_{\nu}(nx)}\\&\quad+\frac{(\nu+1)(\nu+2)}{(n+q)^2}.\end{aligned}$$

By Lemmas 2.2 and 2.5 [7] we get

Lemma 2.3 ([7], Lemma 2.6)

For all $\nu, q\in\mathbb{R}_0$ *there exists a positive constant* $M(\nu, q)$ *such that for each* $n\in\mathbb{N}$ *we have*

$$\left\|\tilde{L}_n^{\nu}(f_0;\cdot)\right\|_q\le M(\nu,q).$$

An obvious consequence of the above lemma and definition (4) is

Theorem 2.1 ([7], Theorem 2.1)

For all $\nu, q \in \mathbb{R}_0$ *there exists a positive constant* $M(\nu, q)$ *such that for each* $n \in \mathbb{N}$ *and* $f \in E_q$ *we have*

$$\left\| \tilde{L}_n^{\nu}(f;\cdot) \right\|_q \leq M(\nu,q) \|f\|_q .$$

Note that in the case of the integral modification of our operators we also have the endomorphism E_q into E_q. This is a better result than the one in [8], Theorem 3.1.

Applying Lemma 2.1 and Lemma 2.2 we immediately obtain

Lemma 2.5 ([7], Lemma 3.1)

For all $\nu, q \in \mathbb{R}_0$ *there exists a positive constant* $M(\nu, q)$ *such that for each* $n \in \mathbb{N}$ *and* $x \in \mathbb{R}_0$ *we have*

$$\left| \tilde{L}_n^{\nu}(\phi_{x,2}; x) \right| \leq M(\nu,q) \frac{x(x+1)}{n}.$$

Lemma 2.6 ([7], Lemma 3.3)

For all $\nu, q \in \mathbb{R}_0$ *there exists a positive constant* $M(\nu, q)$ *such that for each* $n \in \mathbb{N}$ *and* $x \in \mathbb{R}_0$ *we have*

$$w_q(x) \left| \tilde{L}_n^{\nu}(\psi_{x,2}; x) \right| \leq M(\nu,q) \frac{x(x+1)}{n}.$$

3. Degree of approximation

The following theorems estimate a weighted error of approximation for functions belonging to the space $E_q^k = \{ f \in E_q : f', f'', \ldots, f^k \in E_q \}$ for $k = 1, 2$.

The proofs of the theorems are analogous to the proofs which are known from the literature but we enclose them for the completeness of the paper.

Remark 3.1

Note that for $x = 0$ in the following lemmas and theorems we get the assertion using Remark 2.1.

Theorem 3.1

For all $\nu, q \in \mathbb{R}_0$ *there exists a positive constant* $M(\nu, q)$ *such that for all* $n \in \mathbb{N}$, $x \in \mathbb{R}_0$ *and* $f \in E_q^1$ *we have*

$$w_q(x) \left| \tilde{L}_n^{\nu}(f;x) - f(x) \right| \leq M(\nu,q) \|f'\|_q \left(\frac{x(x+1)}{n} \right)^{1/2} .$$

Proof. Let $x > 0$. For $f \in E_q^1$ we have

$$f(t) - f(x) = \int_x^t f'(u) du$$

for $t > 0$. By Lemma 2.2 we have $\tilde{L}_n^{\nu}(e_0; x) = 1,$ hence we can write

$$\tilde{L}_n^{\nu}(f; x) - f(x) = \tilde{L}_n^{\nu}\left(\int_x^t f'(u)du; x \right),$$

using the linearity of $\tilde{L}_n^{\nu}$.

Note that

$$\left| \int_x^t f'(u)du \right| \le \|f'\|_q \left| \int_x^t e^{qu}du \right| \le \|f'\|_q (e^{qt} + e^{qx})|t - x|.$$

Therefore, we have

$$w_q \left| \tilde{L}_n^{\nu}(f; x) - f(x) \right| \le w_q \|f'\|_q \tilde{L}_n^{\nu}\left(|\psi_{x,1}|; x \right) + \|f'\|_q \tilde{L}_n^{\nu}\left(|\phi_{x,1}|; x \right). \tag{7}$$

If we apply the Cauchy-Schwarz inequality and Lemma 2.2 we get

$$\tilde{L}_n^{\nu}\left(|\phi_{x,1}|; x \right) \le \left(\tilde{L}_n^{\nu}\left(|\phi_{x,2}|; x \right) \right)^{1/2},$$

$$\tilde{L}_n^{\nu}\left(|\psi_{x,1}|; x \right) \le \left(\tilde{L}_n^{\nu}\left(|\psi_{x,2}|; x \right) \right)^{1/2} (\tilde{L}_n^{\nu}(f_0; x))^{1/2}.$$

Now we can use Lemma 2.3, 2.5 and 2.6 to estimate (7)

$$w_q \left| \tilde{L}_n^{\nu}(f; x) - f(x) \right| \le M(\nu, q) \|f'\|_q \left(\frac{x(x+1)}{n} \right)^{1/2}$$

for $x > 0$ and $n \in \mathbb{N}$.

Theorem 3.2

For all $\nu, q \in \mathbb{R}_0$ *there exists a positive constant* $M(\nu, q)$ *such that for all* $n \in \mathbb{N}$, $x \in \mathbb{R}_0$ *and* $f \in E_q$ *we have*

$$w_q(x) \left| \tilde{L}_n^{\nu}(f; x) - f(x) \right| \le M(\nu, q) \omega_1 \left(f, E_q; \left(\frac{x(x+1)}{n} \right)^{1/2} \right).$$

Proof. Let $x > 0$. As always we denote by f_h the Steklov function of f, this means

$$f_h(x) = \frac{1}{h} \int_0^h f(x+t)dt$$

for $h > 0$. Note that

$$f_h(x) - f(x) = \frac{1}{h} \int_0^h f(x+t) - f(x)dt,$$

$$f_h'(x) = \frac{1}{h}(f(x+h) - f(x))$$

for $h > 0$. Therefore, we immediately conclude that $f_h, f_h' \in E_q$ because $f \in E_q$ and we have the following estimations

$$\|f_h - f\|_q \le \omega_1(f, E_q; h) \tag{8}$$

$$\|f_h'\|_q \le \frac{1}{h}\omega_1(f, E_q; h) \tag{9}$$

for $h > 0$. By the linearity of the operators $\tilde{L}_n^{\nu}$ we get the inequality

$$\begin{aligned} w_q(x)\left|\tilde{L}_n^{\nu}(f;x) - f(x)\right| &\\ \le w_q(x)\left|\tilde{L}_n^{\nu}(f - f_h;x\right| &+ w_q(x)\left|\tilde{L}_n^{\nu}(f_h;x) - f_h(x)\right| \\ + w_q(x)\left|f_h(x) - f(x)\right| & \end{aligned}$$

Taking into account the boundedness of the operators $\tilde{L}_n^{\nu}$ and (8) we obtain

$$w_q(x)\left|\tilde{L}_n^{\nu}(f - f_h;x\right| \le M(\nu, q)\|f_h - f\|_q \le M(\nu, q)\omega_1(f, E_q; h)$$

for $x, h > 0$. From Theorem 3.1 and (9) we have

$$\begin{aligned} w_q(x)\left|\tilde{L}_n^{\nu}(f_h;x) - f_h(x)\right| &\le M(\nu, q)\|f_h'\|_q\left(\frac{x(x+1)}{n}\right)^{1/2} \\ &\le M(\nu, q)\frac{1}{h}\omega_1(f, E_q; h)\left(\frac{x(x+1)}{n}\right)^{1/2} \end{aligned}$$

for $x, h > 0$.

By the definition of the norm $\|\cdot\|_q$ and (8) we get

$$w_q(x)\left|(f_h(x) - f(x)\right| \le \|f_h - f\|_q \le \omega_1(f, E_q; h)$$

for $x, h > 0$.

Using above inequalities we estimate the expression

$$w_q(x)\left|\tilde{L}_n^{\nu}(f;x) - f(x)\right| \le \omega_1(f, E_q; h)\left(M(\nu, q) + \frac{M(\nu, q)}{h}\left(\frac{x(x+1)}{n}\right)^{1/2} + 1\right).$$

Now substituting $h = \left(\frac{x(x+1)}{n}\right)^{1/2}$ we get the assertion of our theorem.

Theorem 3.2 implies the following corollary.

Corollary 3.3

If $\nu, q \in \mathbb{R}_0$ *and* $f \in E_q$ *then for all* $x \in \mathbb{R}_0$

$$\lim_{n\to\infty}\{\tilde{L}_n^{\nu}(f;x) - f(x)\} = 0.$$

Moreover, the above convergence is uniform on every compact subset of the interval $[0; \infty)$.

Remark 3.4

We can obtain the above convergence in a different way, see Theorem 3.1 ([7]).

To estimate the error of approximation by the second order modulus of smoothness (5) we define the following linear operators

$$\tilde{H}_n^{\nu}(f;x) = \tilde{L}_n^{\nu}(f;x) - f(\tilde{L}_n^{\nu}(e_1;x)) + f(x) \tag{10}$$

for $\nu, q \in \mathbb{R}_0, f \in E_q$ and $x \in \mathbb{R}_0$.

Note that the operators preserve linear functions, namely

$$\tilde{H}_n^{\nu}(\phi_{x,1};x) = 0. \tag{11}$$

Lemma 3.5

For all $\nu, q \in R_0$ *there exists a positive constant* $M(\nu, q)$ *such that for all* $n \in N$, $x \in R_0$ *and* $g \in E_q^2$ *we have*

$$w_q\left|\tilde{H}_n^{\nu}(g;x) - g(x)\right| \le M(\nu,q)\|g''\|_q \frac{x(x+1)}{n}.$$

Proof. Let $x > 0$ be fixed. By the Taylor formula we can write

$$g(t) - g(x) = (t-x)g'(x) + \int_x^t (t-u)g''(u)du$$

for $t > 0$. Now applying linearity of $\tilde{H}_n^{\nu}$ and (11) we derive

$$\left|\tilde{H}_n^{\nu}(g;x) - g(x)\right| = \left|\tilde{H}_n^{\nu}(g(t) - g(x);x)\right| = \left|\tilde{H}_n^{\nu}\left(\int_x^t (t-u)g''(u)du;x\right)\right|. \tag{12}$$

Further, the definition of $\tilde{H}_n^{\nu}$ implies

$$\tilde{H}_n^{\nu}\left(\int_x^t (t-u)g''(u)du;x\right) = \tilde{L}_n^{\nu}\left(\int_x^t (t-u)g''(u)du;x\right) - \int_x^{\tilde{L}_n^{\nu}(t;x)} (\tilde{L}_n^{\nu}(t;x) - u)g''(u)du.$$

Estimating (12) we can write

$$\left|\tilde{H}_n^{\nu}(g;x) - g(x)\right| \le \tilde{L}_n^{\nu}\left(\left|\int_x^t (t-u)g''(u)du\right|;x\right) + \left|\int_x^{\tilde{L}_n^{\nu}} (\tilde{L}_n^{\nu}(t;x) - u)g''(u)du\right|.$$

Note that

$$\left|\int_x^t (t-u)g''(u)du\right| \le \frac{1}{2}\|g''\|_q (t-x)^2 (e^{qx} + e^{qt})$$

and

$$\left|\int_x^{\tilde{L}_n^{\nu}(e_1;x)}(\tilde{L}_n^{\nu}(e_1;x)-u)g''(u)du\right| \le \frac{1}{2}\|g''\|_q\,(\tilde{L}_n^{\nu}(e_1;x)-x)^2(e^{qx}+e^{q\tilde{L}_n^{\nu}(e_1;x)})$$

$$\le \frac{1}{2}\|g''\|_q\,(\tilde{L}_n^{\nu}(\phi_{x,1};x))^2\,e^{qx}(1+e^{q\tilde{L}_n^{\nu}(\phi_{x,1};x)}).$$

Now we can observe that the expression $e^{q\tilde{L}_n^{\nu}(\phi_{x,1};x)}$ is bounded. We immediately obtain it from Lemma 2.2 and 2.1 as follows

$$e^{q\tilde{L}_n^{\nu}(\phi_{x,1};x)} = e^{qx\left(\frac{n}{n+q}\frac{I_{\nu+1}(nx)}{I_{\nu}(nx)}-1\right)}e^{q\frac{\nu+1}{n+q}} \le e^{nx\left(\frac{I_{\nu+1}(nx)}{I_{\nu}(nx)}-1\right)}e^{q\frac{\nu+1}{1+q}} \le M(\nu).$$

Therefore, we have

$$w_q(x)\left|\tilde{H}_n^{\nu}(g;x)-g(x)\right|$$

$$\le \frac{1}{2}\|g''\|_q\,\tilde{L}_n^{\nu}(\phi_{x,2};x)+\frac{1}{2}\|g''\|_q\,w_q(x)\tilde{L}_n^{\nu}(\psi_{x,2};x)$$

$$+\frac{1}{2}M(\nu)\|g''\|_q\,(\tilde{L}_n^{\nu}(\phi_{x,1};x))^2.$$

Applying the Cauchy-Schwarz inequality to the term $\tilde{L}_n^{\nu}(\phi_{x,1};x)$ and Lemmas 2.5, 2.6 we get the desired estimation.

Theorem 3.6

For all $\nu, q \in R_0$ *there exists a positive constant* $M(\nu, q)$ *such that for all* $n \in N$, $x \in R_0$ *and* $f \in E_q$ *we have*

$$w_q\left|\tilde{L}_n^{\nu}(f;x)-f(x)\right| \le M(\nu,q)\omega_2\left(f,E_q;\left(\frac{x(x+1)}{n}\right)^{1/2}\right)+\omega_1\left(f,E_q;\left|\tilde{L}_n^{\nu}(\phi_{x,1};x)\right|\right).$$

Proof. Let $x > 0$ and $\overline{f}_h$ be the second order Steklov mean of $f \in E_q$, i.e.

$$\overline{f}_h(x) = \frac{4}{h^2}\int_0^{h/2}\int_0^{h/2}\{2f(x+s+t)-f(x+2(s+t))\}dsdt, \quad h,x>0$$

Note that

$$f(x)-\overline{f}_h(x) = \frac{4}{h^2}\int_0^{h/2}\int_0^{h/2}\Delta_{s+t}^2 f(x)dsdt.$$

By definition (6) we get the following estimation

$$\left\|f-\overline{f}_h\right\|_q \le \omega_2(f,E_q;h)$$

and since

$$\overline{f}_h''(x) = \frac{1}{h^2}(8\Delta_{h/2}^2 f(x)-\Delta_h^2 f(x))$$

we have

$$\left\|\overline{f}_h''\right\|_q \le \frac{9}{h^2}\,\omega_2(f, E_q; h).$$

The above inequalities imply that the Steklov mean $\overline{f}_h$ and $\overline{f}_h''$ belong to E_q.

Moreover, by the linearity of $\tilde{L}_n^{\nu}$, $\tilde{H}_n^{\nu}$ and the connection (10) we can write

$$\begin{aligned}
&\left|\tilde{L}_n^{\nu}(f;x) - f(x)\right| \\
&\qquad \le \left|\tilde{H}_n^{\nu}(f - \overline{f}_h; x)\right| + \left|(f(x) - \overline{f}_h(x)\right| + \left|\tilde{H}_n^{\nu}(\overline{f}_h; x) - \overline{f}_h(x)\right| \\
&\qquad + \left|f(\tilde{L}_n^{\nu}(e_1; x) - f(x)\right|.
\end{aligned}$$

By the above, the boundedness of the operators $\tilde{H}_n^{\nu}$ and Lemma 3.5 we conclude that

$$\begin{aligned}
&w_q(x)\left|\tilde{L}_n^{\nu}(f;x) - f(x)\right| \\
&\qquad \le w_q(x)\left|\tilde{H}_n^{\nu}(f - \overline{f}_h; x)\right| + w_q(x)\left|f(x) - \overline{f}_h(x)\right| \\
&\qquad + w_q(x)\left|\tilde{H}_n^{\nu}(\overline{f}_h; x) - \overline{f}_h(x)\right| + w_q(x)\left|f(\tilde{L}_n^{\nu}(e_1; x) - f(x)\right| \\
&\qquad \le M(\nu, q)\left\|f - \overline{f}_h\right\|_q + \left\|f - \overline{f}_h\right\|_q + M(\nu, q)\left\|\overline{f}_h''\right\|_q \frac{x(x+1)}{n} \\
&\qquad + w_q(x)\left|f(\tilde{L}_n^{\nu}(\phi_{x,a}; x) + x) - f(x)\right| \\
&\qquad \le M(\nu, q)\omega_2(f, E_q; h)\left(1 + \frac{1}{h^2}\frac{x(x+1)}{n}\right) + \omega_1\left(f, E_q; \left|\tilde{L}_n^{\nu}(\phi_{x,1}; x)\right|\right).
\end{aligned}$$

where $\tilde{L}_n^{\nu}(\phi_{x,1}; x) = x\left(\frac{n}{n+q}\frac{I_{\nu+1}(nx)}{I_{\nu}(nx)} - 1\right) + \frac{\nu+1}{n+q}$. Substituting $h = \left(\frac{x(x+1)}{n}\right)^{1/2}$ we get the estimation in the theses of Theorem 3.6.

The above theorem shows that one can estimate the weighted error of approximation for positive linear operators reproducing constant functions by the sum of two moduli of continuity.

The author is thankful to the referees for making valuable suggestions leading to the overall improvement of the paper.

References

[1] Cárdenas-Morales D., Garrancho P., Raşa I., *Approximation properties of Bernstein-Durrmeyer type operators*, Appl. Math. Comput., **232**, 2014, 1-8.

[2] Durrmeyer J.L., *Une formule d'inversion de la transformée de Laplace: Applications à la théorie des moments*, Thèse de 3ème cycle. Faculté des Sciences Univ. Paris, 1967.

[3] Derriennic M.M., *Sur l'approximation de fonctions intégrables sur* [0;1] *par des polynômes de Bernstein modifiés*, J. Approx. Theory, **31**, 4, 1981, 325-343.

[4] Gonska H., Păltănea R., *Simultaneous approximation by a class of Bernstein-Durrmeyer operators preserving linear functions*, Czechoslovak Math. J. **60**, 135, 2010, 783-799.

[5] Heilmann M., *Direct and converse results for operators of Baskakov-Durrmeyer type*, Approx. Theory Appl., **5**, 1,1989, 105-127.

[6] Herzog M., *Approximation of functions of two variables from exponential weight spaces*, Technical Transactions, Fundamental Sciences, **1-NP**, 2012, 3-10.

[7] Herzog M., *The Voronovskaja type theorem for positive linear operators*, Ciência e Técnica. Vitivinicola Journal, **31**, 9, 2016, 79-86.

[8] Krech G., *Some approximation results for operators of Szász-Mirakjan-Durrmeyer type*, Math. Slovaca, **66**, 4, 2016, 945-958.

[9] Krech G., Wachnicki E., *Direct estimate for some operators of Durrmeyer type in exponential weighted space*, Demonstr. Math., **47**, 2, 2014, 336-349.

[10] Malejki R., Wachnicki E., *On the Baskakov-Durrmeyer type operators*, Comment. Math., **54**, 1, 2014, 39-49.

[11] Mazhar S.M., Totik V., *Approximation by modified Szász operators*, Acta Sci. Math., **49**, 1–4, 1985, 257-269.

[12] Rempulska L., Graczyk Sz., *On certain class of Szász-Mirakyan operators in exponential weight spaces*, Int. J. Pure Appl. Math., **60**, 3, 2010, 259-267.

[13] Sahai A., Prasad G., *On simultaneous approximation by modified Lupaş operators*, J. Approx. Theory, **45**, 2, 1985, 122-128.

[14] Walczak Z., *Bernstein-Durrmeyer type operators*, Acta Math. Univ. Ostrav., **12**, 1, 2004, 65-72.

[15] Watson G.N., *Theory of Bessel functions*, Cambridge Univ. Press, Cambridge, 1966.

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

NAVNIT JHA*, LESŁAW K. BIENIASZ**

AN $O(h_k^5)$ ACCURATE FINITE DIFFERENCE METHOD FOR THE NUMERICAL SOLUTION OF FOURTH ORDER TWO POINT BOUNDARY VALUE PROBLEMS ON GEOMETRIC MESHES

METODA RÓŻNICOWA O DOKŁADNOŚCI $O(h_k^5)$, DO ROZWIĄZYWANIA DWUPUNKTOWYCH ZAGADNIEŃ BRZEGOWYCH CZWARTEGO RZĘDU NA SIATKACH GEOMETRYCZNYCH

Abstract

Two point boundary value problems for fourth order, nonlinear, singular and non-singular ordinary differential equations occur in various areas of science and technology. A compact, three point finite difference scheme for solving such problems on nonuniform geometric meshes is presented. The scheme achieves a fifth or sixth order of accuracy on geometric and uniform meshes, respectively. The proposed scheme describes the generalization of Numerov-type method of Chawla (IMA J Appl Math 24:35-42, 1979) developed for second order differential equations. The convergence of the scheme is proven using the mean value theorem, irreducibility, and monotone property of the block tridiagonal matrix arising for the scheme. Numerical tests confirm the accuracy, and demonstrate the reliability and efficiency of the scheme. Geometric meshes prove superior to uniform meshes, in the presence of boundary and interior layers.

Keywords: Geometric mesh, finite difference method, compact scheme, singularity, stiff equations, Korteweg-de Vries equation, maximum absolute errors

Streszczenie

Dwupunktowe zagadnienia z warunkami brzegowymi, dla nieliniowych, osobliwych i nieosobliwych równań różniczkowych zwyczajnych czwartego rzędu, występują w różnych obszarach nauki i techniki. Zaprezentowano kompaktowy, trzypunktowy schemat różnicowy do rozwiązywania takich problemów na niejednorodnych siatkach geometrycznych. Schemat ten osiąga dokładność piątego lub szóstego rzędu, odpowiednio na siatkach geometrycznych lub jednorodnych. Proponowany schemat przedstawia uogólnienie metody typu Numerowa, autorstwa Chawli (IMA J Appl Math 24:35-42, 1979), opracowanej dla równań różniczkowych drugiego rzędu. Udowodniono zbieżność schematu, korzystając z twierdzenia o własności średniej, nieredukowalności oraz monotoniczności macierzy blokowo-trójdiagonalnej wynikającej ze schematu. Testy numeryczne potwierdzają dokładność, oraz demonstrują niezawodność i wydajność schematu. Siatki geometryczne wykazują przewagę nad siatkami jednorodnymi, w obecności warstw brzegowych i wewnętrznych.

Słowa kluczowe: Siatka geometryczna, metoda różnic skończonych, schemat kompaktowy, osobliwość, równania sztywne, równanie Kortewega-de-Vriesa, maksymalne błędy bezwzględne

DOI: 10.4467/2353737XCT.16.139.5718

* Navnit Jha (navnitjha@sau.ac.in), Department of Mathematics, South Asian University, Akbar Bhawan, Chanakyapuri, New Delhi, India.

** Lesław K. Bieniasz (nbbienia@cyf-kr.edu.pl), Institute of Network Computing, Faculty of Physics, Mathematics and Computer Science, Cracow University of Technology.

1. Introduction

In this paper we consider a numerical solution of the fourth order ordinary differential equation (ODE):

$$-U^{(4)}(r)+g(r,U(r),U^{(1)}(r),U^{(2)}(r),U^{(3)}(r))=0,-\infty<a<r<b<\infty \tag{1.1}$$

subject to the boundary conditions $U(a)=m_1,U(b)=m_2,U^{(2)}(a)=m_3,U^{(2)}(b)=m_4,$ where m_1, m_2, m_3, m_4 are finite real constants. We assume that $g \in C^{(6)}(a,b)$, with the possibility that $g(.)$ can be singular inside and on the boundaries of the domain $[a, b]$.

Boundary value problems of this kind play an important role in various areas of science and technology. The mathematical formulation of noise removal and edge preservation (Yu--Li and Kaveh [1]), Kirchhoff plates (Zhong [2]), theory of plates and shell (Timoshenko and Krieger [3]), waves on a suspension bridge (Chen and McKenna [4]), geological folding of crock layers (Budd [5]) and hydrodynamics equation (Wasow [6]) are some examples of such problems.

The solvability, existence and uniqueness of the solutions of fourth order boundary value problems have been discussed by O'Regan [7], Agarwal [8] and Atabizadeh [9]. For solving Eq. (1.1) a number of approaches have been proposed, such as differential transform (Momani et. al. [10]), Adomian decomposition (Wazwaz [11]), homotopy perturbation (Din et. al. [12]), variational iteration (Noor et. al. [13]), exponential spline (Zahra [14]) and finite difference approximations (Usmani [15], Schroder [16] and Shanthi [17]).

Possible approaches to solving Eq. (1.1) can be roughly divided into two categories. The first category includes methods which solve Eq. (1.1) as is, either analytically as in [10–13] or numerically as in [14–17]. The second category includes methods in which Eq. (1.1) is first converted to a system of second order ODEs:

$$-U^{(2)}(r)+V(r)=0, \tag{1.2}$$

$$-V^{(2)}(r)+g(r,U(r),U^{(1)}(r),V(r),V^{(1)}(r))=0,-\infty<a<r<b<\infty. \tag{1.3}$$

Subsequently, one solves system (1.2) and (1.3) by a technique appropriate to second order ODEs (see, for example Twizell and Boutayeb [18]).

In the present paper we describe a new method that belongs to the second category. The method uses a fifth order accurate, compact three point finite difference scheme that approximates system (1.2) and (1.3) on a specific nonuniform mesh called a geometric mesh (Jain et. al. [19], Kadalbajoo [20] and Mohanty [21]); in some application areas, like electrochemistry the name "exponentially expanding grid" is also used (Britz [22]). The geometric mesh is defined by the formulae: $a=r_0<\ldots<r_{n+1}=b,\ h_k=r_k-r_{k-1},$ $k=1(1)n+1,\ h_{k+1}=\tau h_k,$ where $\tau>0$ is a constant mesh ratio parameter and $n+2$ is the total number of nodes. Such a mesh is particularly suitable when ODEs such as Eq. (1.1) or (1.2) and (1.3) are singularly perturbed, so that their solutions possess boundary or interior layers (Roos [23], Farrell et. al. [24]). The compact, three point character of the scheme makes it particularly convenient. This is because in the process of the numerical solution of the resulting nonlinear algebraic equation systems (for example, by the Newton method)

one obtains linear algebraic systems with block tridiagonal matrices. Such systems are easy to solve, using standard algorithms, for example the generalized Thomas algorithm (Thomas [25], Bieniasz [26]). In contrast, higher order discretizations associated with non-compact stencils lead to the increase of the bandwidth of the resultant coefficient matrix, which implies a larger number of arithmetic operations.

There exists an ample literature devoted to the development of compact schemes for solving two point boundary value problems for single second order ODEs. In particular, we mention here the various improvements of the classical Numerov scheme (Numerov [27], Agarwal [28]) and the arithmetic average schemes, obtained by (Chawla [29, 30], Wang [31], Bieniasz [32], Mohanty [33], Zhang [34] and Jha [35, 36]). The new scheme proposed in the present work, can be regarded as an extension, and adaptation to the nonuniform mesh, of the sixth order compact scheme of Chawla [30]. Minor modifications of the scheme are required for the singular problems.

The paper is organized as follows: In section 2, we develop the higher order finite difference scheme on the geometric mesh. The convergence analysis is contained in section 3. In section 4, some computational experiments are described that show the reliability of the algorithm. In the last section, the findings are summarized.

2. Formulation of the $O(h_k^5)$ finite difference scheme on the geometric mesh

Let U_k, V_k be the exact solution values and u_k, v_k be the approximate values of $U(r)$ and $V(r)$ at the mesh node r_k respectively. With the help of finite Taylor's expansions, we first obtain the following relation that approximates the second order derivative at r_k using geometric meshes:

$$h_k^2 c_0 U_k^{(2)} = -U_{k+1} + (1+\tau)U_k - \tau U_{k-1} - h_k^2 (c_1 U_{k+1}^{(2)} + c_2 U_{k-1}^{(2)} + c_3 U_{k+1/2}^{(2)} + c_4 U_{k-1/2}^{(2)}) + O(h_k^7), \tag{2.1}$$

where:

$$c_0 = -(1+\tau)(3\tau^2 + 7\tau + 3)/60,$$

$$c_1 = -(2\tau^3 + \tau^2 - \tau + 1)/[60(1+2\tau)],\ c_3 = -2(1+\tau)(2\tau^2 + 2\tau - 1)/[15(2+\tau)],$$

$$c_2 = -\tau(\tau^3 - \tau^2 + \tau + 2)/[60(2+\tau)],\ c_4 = 2\tau(1+\tau)(\tau^2 - 2\tau - 2)/[15(1+2\tau)]$$

As Eq. (1.3) involves first solution derivatives, we need certain approximations to these derivatives. Consider the following geometric mesh approximations to $U^{(1)}$:

$$\tilde{U}_k^{(1)} = [U_{k+1} - (1-\tau^2)U_k - \tau^2 U_{k-1}]/[h_k \tau(1+\tau)], \tag{2.2}$$

$$\tilde{U}_{k+1}^{(1)} = [(1+2\tau)U_{k+1} - (1+\tau)^2 U_k + \tau^2 U_{k-1}]/[h_k \tau(1+\tau)], \tag{2.3}$$

$$\tilde{U}_{k-1}^{(1)} = [-U_{k+1} + (1+\tau)^2 U_k - \tau(2+\tau)U_{k-1}]/[h_k \tau(1+\tau)], \tag{2.4}$$

In a similar manner, we can obtain approximations $\tilde{V}_k^{(1)}$ and $\tilde{V}_{k\pm1}^{(1)}$ to $V^{(1)}$. We denote

$$\tilde{G}_{k+\theta} = g(r_{k+\theta}, U_{k+\theta}, \tilde{U}_{k+\theta}^{(1)}, V_{k+\theta}, \tilde{V}_{k+\theta}^{(1)}), \theta = 0, \pm 1. \tag{2.5}$$

With the help of Eqs. (2.2)–(2.5), we obtain

$$\begin{aligned}\tilde{G}_k &= g_k + h_k^2\tau(A_kU_k^{(3)} + D_kV_k^{(3)})/6 + h_k^3\tau(\tau-1)(A_kU_k^{(4)} + D_kV_k^{(4)})/24\\ &\quad + h_k^4\tau^2[B_k(U_k^{(3)})^2 + 2C_kU_k^{(3)}V_k^{(3)} + E_k(V_k^{(3)})^2]/72\\ &\quad + h_k^4\tau(\tau^2-\tau+1)(A_kU_k^{(5)} + D_kV_k^{(5)})/120 + O(h_k^5),\end{aligned} \tag{2.6}$$

$$\begin{aligned}\tilde{G}_{k+1} &= g_{k+1} - h_k^2\tau(1+\tau)[A_kU_k^{(3)} + D_kV_k^{(3)} + h_k\tau(A_k^{(1)}U_k^{(3)} + D_k^{(1)}V_k^{(3)})]/6\\ &\quad - h_k^3\tau(\tau+1)(2\tau-1)[A_kU_k^{(4)} + D_kV_k^{(4)} + \tau(A_k^{(1)}U_k^{(4)} + D_k^{(1)}V_k^{(4)})]/24\\ &\quad - h_k^4\tau(\tau+1)[(3\tau^2-2\tau+1)(A_kU_k^{(5)} + D_kV_k^{(5)}) + 10\tau^2(A_k^{(2)}U_k^{(3)} + D_k^{(2)}V_k^{(3)})]/120\\ &\quad + h_k^4\tau^2(\tau+1)^2[\mathrm{B}_k(U_k^{(3)})^2 + 2C_kU_k^{(3)}V_k^{(3)} + E_k(V_k^{(3)})^2]/72 + O(h_k^5),\end{aligned} \tag{2.7}$$

$$\begin{aligned}\tilde{G}_{k-1} &= g_{k-1} - h_k^2(1+\tau)[A_kU_k^{(3)} + D_kV_k^{(3)} - h_k(A_k^{(1)}U_k^{(3)} + D_k^{(1)}V_k^{(3)})]/6\\ &\quad - h_k^3(\tau^2-\tau-2)[A_kU_k^{(4)} + D_kV_k^{(4)} - h_k(A_k^{(1)}U_k^{(4)} + D_k^{(1)}V_k^{(4)})]/24\\ &\quad - h_k^4(\tau+1)[(\tau^2-2\tau+3)(A_kU_k^{(5)} + D_kV_k^{(5)}) + 10(A_k^{(2)}U_k^{(3)} + D_k^{(2)}V_k^{(3)})]/120\\ &\quad + h_k^4(\tau+1)^2[B_k(U_k^{(3)})^2 + 2C_kU_k^{(3)}V_k^{(3)} + E_k(V_k^{(3)})^2]/72 + O(h_k^5),\end{aligned} \tag{2.8}$$

where:

$$A_k = (\partial g/\partial U^{(1)})_{r_k},\ B_k = (\partial^2 g/\partial U^{(1)^2})_{r_k},\ C_k = (\partial^2 g/\partial U^{(1)}\partial V^{(1)})_{r_k},$$

$$D_k = (\partial g/\partial V^{(1)})_{r_k} \text{ and } E_k = (\partial^2 g/\partial V^{(1)^2})_{r_k}.$$

By using $\tilde{G}_k$ and $\tilde{G}_{k\pm1}$, one can look for the approximations to the solution values and derivatives;

$$\begin{aligned}&[\hat{U}_{k+1/2}, \hat{U}_{k-1/2}, \hat{U}_{k+1}^{(1)}, \hat{U}_{k-1}^{(1)}, \hat{U}_{k+1/2}^{(1)}, \hat{U}_{k-1/2}^{(1)}]^T =\\ &\qquad \begin{bmatrix} a_{11} & a_{12} & a_{13}\\ \vdots & \vdots & \vdots\\ a_{61} & a_{62} & a_{63}\end{bmatrix}\begin{bmatrix} U_{k-1}\\ U_k\\ U_{k+1}\end{bmatrix} + h_k^2\begin{bmatrix} a_{14} & a_{15} & a_{16}\\ \vdots & \vdots & \vdots\\ a_{64} & a_{65} & a_{66}\end{bmatrix}\begin{bmatrix} V_{k-1}\\ V_k\\ V_{k+1}\end{bmatrix},\end{aligned} \tag{2.9}$$

$$\begin{aligned}&[\hat{V}_{k+1/2}, \hat{V}_{k-1/2}, \hat{V}_{k+1}^{(1)}, \hat{V}_{k-1}^{(1)}, \hat{V}_{k+1/2}^{(1)}, \hat{V}_{k-1/2}^{(1)}]^T =\\ &\qquad \begin{bmatrix} b_{11} & b_{12} & b_{13}\\ \vdots & \vdots & \vdots\\ b_{61} & b_{62} & b_{63}\end{bmatrix}\begin{bmatrix} V_{k-1}\\ V_k\\ V_{k+1}\end{bmatrix} + h_k^2\begin{bmatrix} b_{14} & b_{15} & b_{16}\\ \vdots & \vdots & \vdots\\ b_{64} & b_{65} & b_{66}\end{bmatrix}\begin{bmatrix} \tilde{G}_{k-1}\\ \tilde{G}_k\\ \tilde{G}_{k+1}\end{bmatrix},\end{aligned} \tag{2.10}$$

where $a_{lm}, b_{lm}, l, m = 1(1)6$ are free parameters to be determined in such a way that we can achieve the following high order approximations

$$\hat{U}_{k\pm1/2} - U_{k\pm1/2} = O(h_k^5), \quad \hat{V}_{k\pm1/2} - V_{k\pm1/2} = O(h_k^5), \tag{2.11}$$

$$\hat{U}^{(1)}_{k+\theta} - U^{(1)}_{k+\theta} = O(h_k^4), \quad \hat{V}^{(1)}_{k+\theta} - V^{(1)}_{k+\theta} = O(h_k^4), \quad \theta = \pm1, \pm1/2. \tag{2.12}$$

With the help of algebraic calculations using MAPLE (see Ref. [37]), explicit expressions for the free parameters were obtained and they are shown in Table 1, where we have denoted $\sigma = \tau^2 + 3\tau + 1$ and $\rho = \tau^2 + \tau + 1$. Consequently,

$$\hat{U}^{(1)}_{k+1} = U^{(1)}_{k+1} + h_k^4 \tau^2 (1+\tau)^3 (4+\tau) U_k^{(5)} / (360\sigma) + O(h_k^5), \tag{2.13}$$

$$\hat{U}^{(1)}_{k-1} = U^{(1)}_{k-1} + h_k^4 (1+\tau)^3 (1+4\tau) U_k^{(5)} / (360\sigma) + O(h_k^5), \tag{2.14}$$

$$\hat{U}^{(1)}_{k+1/2} = U^{(1)}_{k+1/2} - h_k^4 \tau^2 (4+\tau)(7\tau^3 + 9\tau^2 - 5\tau - 4) U_k^{(5)} / (5760\sigma) + O(h_k^5), \tag{2.15}$$

$$\hat{U}^{(1)}_{k-1/2} = U^{(1)}_{k-1/2} + h_k^4 (1+4\tau)(4\tau_k^3 + 5\tau_k^2 - 9\tau_k - 7) U_k^{(5)} / (5760\sigma) + O(h_k^5), \tag{2.16}$$

$$\begin{aligned} \hat{V}^{(1)}_{k+1} = V^{(1)}_{k+1} - h_k^4 \tau^2 (1+\tau)^2 [(2\tau^2 + 2\tau - 1)\{10 A_k^{(1)} U_k^{(3)} + 10 D_k^{(1)} V_k^{(3)} \\ + 5(A_k U_k^{(4)} + D_k V_k^{(4)})/2\} + (5\tau^2 + 5\tau - 4) V_k^{(5)}] / (360\rho) + O(h_k^5), \end{aligned} \tag{2.17}$$

$$\begin{aligned} \hat{V}^{(1)}_{k-1} = V^{(1)}_{k-1} + h_k^4 (1+\tau)^2 [(\tau^2 - 2\tau - 2)\{10 A_k^{(1)} U_k^{(3)} + 10 D_k^{(1)} V_k^{(3)} \\ + 5(A_k U_k^{(4)} + D_k V_k^{(4)})/2\} - (4\tau^2 - 5\tau - 5) V_k^{(5)}] / (360\rho) + O(h_k^5), \end{aligned} \tag{2.18}$$

$$\begin{aligned} \hat{V}^{(1)}_{k+1/2} = V^{(1)}_{k+1/2} + h_k^4 \tau^2 [(\tau^4 + 3\tau^3 + 2\tau^2 - 2\tau - 1)\{80(A_k^{(1)} U_k^{(3)} + D_k^{(1)} V_k^{(3)}) \\ + 20(A_k U_k^{(4)} + D_k V_k^{(4)})\} - (23\tau^4 + 63\tau^3 + 31\tau^2 \\ - 64\tau - 32) V_k^{(5)}] / (5760\rho) + O(h_k^5), \end{aligned} \tag{2.19}$$

$$\begin{aligned} \hat{V}^{(1)}_{k-1/2} = V^{(1)}_{k-1/2} - h_k^4 [(\tau^4 + 2\tau^3 - 2\tau^2 - 3\tau - 1)\{80(A_k^{(1)} U_k^{(3)} + D_k^{(1)} V_k^{(3)}) \\ + 20(A_k U_k^{(4)} + D_k V_k^{(4)})\} - (32\tau^4 + 64\tau^3 - 31\tau^2 \\ - 63\tau - 23) V_k^{(5)}] / (5760\rho) + O(h_k^5). \end{aligned} \tag{2.20}$$

Further, we define

$$\hat{G}_{k\pm1} = g(r_{k\pm1}, U_{k\pm1}, \hat{U}^{(1)}_{k\pm1}, V_{k\pm1}, \hat{V}^{(1)}_{k\pm1}), \tag{2.21}$$

$$\hat{G}_{k\pm1/2} = g(r_{k\pm1/2}, \hat{U}_{k\pm1/2}, \hat{U}^{(1)}_{k\pm1/2}, \hat{V}_{k\pm1/2}, \hat{V}^{(1)}_{k\pm1/2}). \tag{2.22}$$

With the help of the above approximations (2.13)–(2.20), we obtain

$$\begin{aligned} \hat{G}_{k+1} = g_{k+1} - h_k^4 \tau^2 (1+\tau)^2 [(2\tau^2 + 2\tau - 1)(720 D_k^{(1)} (A_k^{(1)} U_k^{(3)} + D_k^{(1)} V_k^{(3)}) \\ + 180 D_k (A_k U_k^{(4)} + D_k V_k^{(4)})) + 72((\tau^2 + 5\tau + 4)\rho A_k U_k^{(5)} / \sigma \\ + (5\tau^2 + 5\tau - 4) D_k V_k^{(5)})] / (25920\rho) + O(h_k^5), \end{aligned} \tag{2.23}$$

$$\hat{G}_{k-1} = g_{k-1} + h_k^4(1+\tau)^2[(\tau^2-2\tau-2)(720D_k(A_k^{(1)}U_k^{(3)} + D_k^{(1)}V_k^{(3)}) + 180D_k(A_kU_k^{(4)} + D_kV_k^{(4)})) + 72((4\tau^2+5\tau+1)\rho A_kU_k^{(5)}/\sigma - (4\tau^2-5\tau-5)D_kV_k^{(5)})]/(25920\rho) + O(h_k^5), \tag{2.24}$$

$$\hat{G}_{k+1/2} = g_{k+1/2} + h_k^4\tau^2[20(\tau^4+3\tau^3+2\tau^2-2\tau-1)D_k(4A_k^{(1)}U_k^{(3)} + 4D_k^{(1)}V_k^{(3)} + A_kU_k^{(4)} + D_kV_k^{(4)}) - (\tau+4)(7\tau^3+9\tau^2-5\tau-4)\rho A_kU_k^{(5)}/\sigma - (23\tau^4+63\tau^3+31\tau^2-64\tau-32)D_kV_k^{(5)}]/(5760\rho) + O(h_k^5), \tag{2.25}$$

$$\hat{G}_{k-1/2} = g_{k-1/2} - h_k^4[(\tau^4+2\tau^3-2\tau^2-3\tau-1)20D_k(4A_k^{(1)}U_k^{(3)} + 4D_k^{(1)}V_k^{(3)} + A_kU_k^{(4)} + D_kV_k^{(4)}) - (4\tau+1)(4\tau^3+5\tau^2-9\tau-7)\rho A_kU_k^{(5)}/\sigma - (32\tau^4+64\tau^3-31\tau^2-63\tau-23)D_kV_k^{(5)}]/(5760\rho) + O(h_k^5). \tag{2.26}$$

We define additional approximations to the first derivatives:

$$\breve{U}_k^{(1)} = \tilde{U}_k^{(1)} + h_k(t_0V_k + t_1V_{k+1} + t_2V_{k-1}) + h_k^3t_3\tilde{G}_{k-1}, \tag{2.27}$$

$$\breve{V}_k^{(1)} = \tilde{V}_k^{(1)} + h_k(z_1\tilde{G}_{k+1} + z_2\tilde{G}_{k-1} + z_3\hat{G}_{k+1} + z_4\hat{G}_{k-1} + z_5\tilde{G}_{k+1/2} + z_6\tilde{G}_{k-1/2}), \tag{2.28}$$

where t_k's and z_k's are unknown coefficients to be determined so as to achieve the following final approximations:

$$U_{k+1} - (1+\tau)U_k + \tau U_{k-1} + h_k^2(c_0V_k + c_1V_{k+1} + c_2V_{k-1} + c_3\hat{V}_{k+1/2} + c_4\hat{V}_{k-1/2}) = O(h_k^7), \tag{2.29}$$

$$V_{k+1} - (1+\tau)V_k + \tau V_{k-1} + h_k^2(c_0\breve{G}_k + c_1\hat{G}_{k+1} + c_2\hat{G}_{k-1} + c_3\hat{G}_{k+1/2} + c_4\hat{G}_{k-1/2}) = O(h_k^7), \tag{2.30}$$

where $k = 1(1)n$ and $\breve{G}_k$ is an extra approximation to G_k, to be determined.

The explicit expressions for the unknown coefficients are given in Table 2, where we have denoted $\delta = 3\tau^2 + 7\tau + 3$. From Eqs. (2.7), (2.8) and (2.23)–(2.26), we obtain

$$\breve{U}_k^{(1)} = U_k^{(1)} + h_k(t_0+t_1+t_2)U_k^{(2)} + h_k^3[(1+12t_1)\tau^2 + 12t_2 + 24t_3 - \tau]U_k^{(4)}/24 + h_k^2[(6t_1+1)\tau - 6t_2]U_k^{(3)}/6 + h_k^4[(1+20t_1)\tau^3 - 20t_2 - 120t_3 - \tau^2 + \tau]U_k^{(5)}/120 + O(h_k^5), \tag{2.31}$$

$$\breve{V}_k^{(1)} = V_k^{(1)} + h_k(z_1+z_2+z_3+z_4+z_5+z_6)U_k^{(4)} + h_k^2[\tau(1+6z_1+6z_3+3z_5) - 3(2z_2+2z_4+z_6)]U_k^{(5)}/6 + h_k^3(1+\tau)(z_1\tau+z_2)(A_kU_k^{(3)} - D_kV_k^{(3)})/6 + h_k^3[\tau^2(1+3z_5+12z_3+12z_1) + 3z_6 + 12z_2 + 12z_4 - \tau]U_k^{(6)}/24 - h_k^4(1+\tau)(\tau(2\tau-1)z_1 + (\tau-2)z_2)(A_kU_k^{(4)} + D_kV_k^{(4)})/24$$

$$-h_k^4(1+\tau)(z_1\tau^2 - z_2)(A_k^{(1)}U_k^{(3)} + D_k^{(1)}V_k^{(3)})/6 + h_k^4[2(\tau^3 - \tau^2 + \tau)$$
$$+40(\tau^3(z_1+z_3) - z_2 - z_4)/240 + 5(z_5\tau^3 - z_6)]V_k^{(5)} + O(h_k^5). \tag{2.32}$$

Finally, by using Eqs. (2.27) and (2.28), we define

$$\breve{G}_k = g(r_k, U_k, \breve{U}_k^{(1)}, V_k, \breve{V}_k^{(1)}). \tag{2.33}$$

Hence, we have obtained the final geometric mesh finite difference scheme (2.29) and (2.30), which is compact and applicable to the numerical solution of the boundary value problem (1.1) or (1.2) and (1.3). A more detailed analysis reveals that the local truncation error of the scheme is $(\tau - 1)O(h_k^7) + O(h_k^8)$ and hence in the case of a uniform mesh ($\tau = 1$), the proposed method is sixth order accurate.

The scheme needs an amendment in the vicinity of a singularity, which arises when, for example, our domain of integration is $[0, 1]$ and we need to evaluate the terms like r_{k-1}^{-1} at $k = 1$. This leads to the division by zero and hence in order to avoid such situations, we need to incorporate the Taylor's approximations $r_{k-1}^{-1} = \sum_{l=0(1)4} h_r^l r_k^{-(1+l)} + O(h_k^5)$, into Eqs. (2.29) and (2.30). The resulting scheme is applicable to singular ODEs such as ODEs involving the Laplacian operator in cylindrical and spherical coordinates. For practical implementations, one replaces the exact values U_k and V_k present in Eqs. (2.29) and (2.30) by approximate values u_k and v_k, and one omits the residual terms $O(h_k^7)$. The resulting system of algebraic equations for u_k and v_k must be extended with boundary conditions.

3. Convergence analysis

In this section, we discuss the convergence property of the proposed finite difference scheme (2.29) and (2.30) for the numerical solution of the two point boundary value problem (1.1). At $r = r_k$, $k = 1(1)n$, Eq. (1.1) can be written as

$$U_k^{(2)} = V_k, V_k^{(2)} = g(r_k, U_k, U_k^{(1)}, V_k, V_k^{(1)}) \equiv G_k, k = 1(1)n. \tag{3.1}$$

Then, the geometric mesh finite difference method (2.29)–(2.30) is given by

$$\begin{cases} \phi_k(U_{k-1}, U_k, U_{k+1}, V_{k-1}, V_k, V_{k+1}) + L_k(h_k) = 0, \\ \varphi_k(U_{k-1}, U_k, U_{k+1}, V_{k-1}, V_k, V_{k+1}) + M_k(h_k) = 0, k = 1(1)n, \end{cases} \tag{3.2}$$

where

$$\phi_k = -U_{k+1} + (1+\tau)U_k - \tau U_{k-1}$$
$$- h_k^2(c_0 V_k + c_1 V_{k+1} + c_2 V_{k-1} + c_3 \hat{V}_{k+1/2} + c_4 \hat{V}_{k-1/2}),$$

$$\varphi_k = -V_{k+1} + (1+\tau)V_k - \tau V_{k-1}$$
$$- h_k^2(c_0 \breve{G}_k + c_1 \hat{G}_{k+1} + c_2 \hat{G}_{k-1} + c_3 \hat{G}_{k+1/2} + c_4 \hat{G}_{k-1/2}),$$

$$L_k(h_k) = O(h_k^7) \quad \text{and} \quad M_k(h_k) = O(h_k^7).$$

The scheme (3.2) in the matrix/vector notation is written as

$$\begin{cases} \boldsymbol{\phi}(\boldsymbol{U},\boldsymbol{V}) + \boldsymbol{L} = \boldsymbol{0} \\ \boldsymbol{\varphi}(\boldsymbol{U},\boldsymbol{V}) + \boldsymbol{M} = \boldsymbol{0}, \end{cases} \tag{3.3}$$

where

$$\boldsymbol{U} = \begin{bmatrix} U_1 \\ \vdots \\ U_n \end{bmatrix}, \quad \boldsymbol{V} = \begin{bmatrix} V_1 \\ \vdots \\ V_n \end{bmatrix}, \quad \boldsymbol{L} = \begin{bmatrix} L_1 \\ \vdots \\ L_n \end{bmatrix}, \quad \boldsymbol{M} = \begin{bmatrix} M_1 \\ \vdots \\ M_n \end{bmatrix}.$$

We wish to find the approximations $\boldsymbol{u}$ and $\boldsymbol{v}$ for $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively, which are determined by solving $2n \times 2n$ systems

$$\begin{cases} \boldsymbol{\phi}(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{0} \\ \boldsymbol{\varphi}(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{0}. \end{cases} \tag{3.4}$$

From (3.3) and (3.4), we obtain

$$\begin{cases} \boldsymbol{\phi}(\boldsymbol{u},\boldsymbol{v}) - \boldsymbol{\phi}(\boldsymbol{U},\boldsymbol{V}) = \boldsymbol{L} \\ \boldsymbol{\varphi}(\boldsymbol{u},\boldsymbol{v}) - \boldsymbol{\varphi}(\boldsymbol{U},\boldsymbol{V}) = \boldsymbol{M}. \end{cases} \tag{3.5}$$

Let $\varepsilon_k = u_k - U_k$, $\eta_k = v_k - V_k$, $k = 1(1)n$ be the discretization errors and $\boldsymbol{\varepsilon} = \boldsymbol{u} - \boldsymbol{U}$, $\boldsymbol{\eta} = \boldsymbol{v} - \boldsymbol{V}$ be the vectors of these errors. Let us denote

$$\tilde{g}_{k+\theta} = g(r_{k+\theta}, u_{k+\theta}, \tilde{u}^{(1)}_{k+\theta}, v_{k+\theta}, \tilde{v}^{(1)}_{k+\theta}) \simeq \tilde{G}_{k+\theta},\ \theta = 0, \pm 1,$$

$$\hat{g}_{k\pm1} = g(r_{k\pm1}, u_{k\pm1}, \hat{u}^{(1)}_{k\pm1}, v_{k\pm1}, \hat{v}^{(1)}_{k\pm1}) \simeq \hat{G}_{k\pm1},$$

$$\hat{g}_{k\pm1/2} = g(r_{k\pm1/2}, \hat{u}_{k\pm1/2}, \hat{u}^{(1)}_{k\pm1/2}, \hat{v}_{k\pm1/2}, \hat{v}^{(1)}_{k\pm1/2}) \simeq \hat{G}_{k\pm1/2}$$

$$\breve{g}_k = g(r_k, u_k, \breve{u}^{(1)}_k, v_k, \breve{v}^{(1)}_k) \simeq \breve{G}_k,$$

$$\tilde{E}_{k+\theta} = \tilde{g}_{k+\theta} - \tilde{G}_{k+\theta},\ \theta = 0, \pm 1,$$

$$\hat{E}_{k\pm\theta} = \hat{g}_{k\pm\theta} - \hat{G}_{k\pm\theta},\ \theta = 1, 1/2,$$

$$\breve{E}_k = \breve{g}_k - \breve{G}_k,$$

$$\tilde{\varepsilon}^{(1)}_{k+\theta} = \tilde{u}^{(1)}_{k+\theta} - \tilde{U}^{(1)}_{k+\theta},\ \tilde{\eta}^{(1)}_{k+\theta} = \tilde{v}^{(1)}_{k+\theta} - \tilde{V}^{(1)}_{k+\theta},\ \theta = 0, \pm 1,$$

$$\hat{\varepsilon}_{k\pm1/2} = \hat{u}_{k\pm1/2} - \hat{U}_{k\pm1/2},\ \hat{\eta}_{k\pm1/2} = \hat{v}_{k\pm1/2} - \hat{V}_{k\pm1/2},$$

$$\hat{\varepsilon}^{(1)}_{k\pm\theta} = \hat{u}^{(1)}_{k\pm\theta} - \hat{U}^{(1)}_{k\pm\theta},\ \hat{\eta}^{(1)}_{k\pm\theta} = \hat{v}^{(1)}_{k\pm\theta} - \hat{V}^{(1)}_{k\pm\theta},\ \theta = 1, 1/2,$$

$$\breve{\varepsilon}^{(1)}_k = \breve{u}^{(1)}_k - \breve{U}^{(1)}_k,\ \breve{\eta}^{(1)}_k = \breve{v}^{(1)}_k - \breve{V}^{(1)}_k,$$

$$\tilde{\xi}^{(1)}_k = [\xi_{k+1} - (1-\tau^2)\xi_k - \tau^2\xi_{k-1}] / [h_k\tau(1+\tau)],\ \xi \in \{\varepsilon, \eta\},$$

$$\tilde{\xi}^{(1)}_{k+1} = [(1+2\tau)\xi_{k+1} - (1+\tau)^2\xi_k + \tau^2\xi_{k-1}] / [h_k\tau(1+\tau)],$$

$$\tilde{\xi}_{k-1}^{(1)} = [-\xi_{k+1} + (1+\tau)^2 \xi_k - \tau(2+\tau)\xi_{k-1}] / [h_k \tau(1+\tau)].$$

By applying the mean value theorem, one obtains:

$$\tilde{E}_{k+\theta} = \alpha_{k+\theta}\tilde{\varepsilon}_{k+\theta}^{(1)} + \beta_{k+\theta}\varepsilon_{k+\theta} + \gamma_{k+\theta}\tilde{\eta}_{k+\theta}^{(1)} + \delta_{k+\theta}\eta_{k+\theta},\ \theta = 0, \pm 1, \tag{3.6}$$

where

$$\alpha_l = \left.\frac{\partial g}{\partial u^{(1)}}\right|_{r=r_1},\quad \beta_l = \left.\frac{\partial g}{\partial u}\right|_{r=r_1},\quad \gamma_l = \left.\frac{\partial g}{\partial v^{(1)}}\right|_{r=r_1},\quad \delta_l = \left.\frac{\partial g}{\partial v}\right|_{r=r_1},\quad l = k, \boldsymbol{k} \pm 1, k \pm 1/2.$$

Let us define:

$$[\hat{\varepsilon}_{k+1/2}, \hat{\varepsilon}_{k-1/2}, \hat{\varepsilon}_{k+1}^{(1)}, \hat{\varepsilon}_{k-1}^{(1)}, \hat{\varepsilon}_{k+1/2}^{(1)}, \hat{\varepsilon}_{k-1/2}^{(1)}]^T = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ \vdots & \vdots & \vdots \\ a_{61} & a_{62} & a_{63} \end{bmatrix} \begin{bmatrix} \varepsilon_{k-1} \\ \varepsilon_k \\ \varepsilon_{k+1} \end{bmatrix} + h_k^2 \begin{bmatrix} a_{14} & a_{15} & a_{16} \\ \vdots & \vdots & \vdots \\ a_{64} & a_{65} & a_{66} \end{bmatrix} \begin{bmatrix} \eta_{k-1} \\ \eta_k \\ \eta_{k+1} \end{bmatrix}, \tag{3.7}$$

$$[\hat{\eta}_{k+1/2}, \hat{\eta}_{k-1/2}, \hat{\eta}_{k+1}^{(1)}, \hat{\eta}_{k-1}^{(1)}, \hat{\eta}_{k+1/2}^{(1)}, \hat{\eta}_{k-1/2}^{(1)}]^T = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ \vdots & \vdots & \vdots \\ b_{61} & b_{62} & b_{63} \end{bmatrix} \begin{bmatrix} \eta_{k-1} \\ \eta_k \\ \eta_{k+1} \end{bmatrix} + h_k^2 \begin{bmatrix} b_{14} & b_{15} & b_{16} \\ \vdots & \vdots & \vdots \\ b_{64} & b_{65} & b_{66} \end{bmatrix} \begin{bmatrix} \tilde{E}_{k-1} \\ \tilde{E}_k \\ \tilde{E}_{k+1} \end{bmatrix}, \tag{3.8}$$

where are coefficients given in Table 1 and 2, and

$$\hat{E}_{k\pm1} = \alpha_{k\pm1}\hat{\varepsilon}_{k\pm1}^{(1)} + \beta_{k\pm1}\varepsilon_{k\pm1} + \gamma_{k\pm1}\hat{\eta}_{k\pm1}^{(1)} + \delta_{k\pm1}\eta_{k\pm1}, \tag{3.9}$$

$$\hat{E}_{k\pm1/2} = \alpha_{k\pm1/2}\hat{\varepsilon}_{k\pm1/2}^{(1)} + \beta_{k\pm1/2}\varepsilon_{k\pm1/2} + \gamma_{k\pm1/2}\hat{\eta}_{k\pm1/2}^{(1)} + \delta_{k\pm1/2}\eta_{k\pm1/2}, \tag{3.10}$$

$$\breve{\varepsilon}_k^{(1)} = \tilde{\varepsilon}_k^{(1)} + h_k(t_0\eta_k + t_1\eta_{k+1} + t_2\eta_{k-1}) + h_k^3 t_3 \tilde{E}_{k-1}, \tag{3.11}$$

$$\breve{\eta}_k^{(1)} = \tilde{\eta}_k^{(1)} + h_k(z_1\tilde{E}_{k+1} + z_2\tilde{E}_{k-1} + z_3\hat{E}_{k+1} + z_4\hat{E}_{k-1} + z_5\hat{E}_{k+1/2} + z_6\hat{E}_{k-1/2}), \tag{3.12}$$

$$\breve{E}_k = \alpha_k\breve{\varepsilon}_k^{(1)} + \beta_k\varepsilon_k + \gamma_k\breve{\eta}_k^{(1)} + \delta_k\eta_k. \tag{3.13}$$

In view of the Eq. (3.5), we obtain

$$\begin{aligned} R_k &\equiv \phi_k(u_{k-1}, u_k, u_{k+1}, v_{k-1}, v_k, v_{k+1}) - \phi_k(U_{k-1}, U_k, U_{k+1}, V_{k-1}, V_k, V_{\boldsymbol{k}+1}) \\ &= -\varepsilon_{k+1} + (1+\tau)\varepsilon_k - \tau\varepsilon_{k-1} - h_k^2(c_0\eta_k + c_1\eta_{k+1} + c_2\eta_{k-1} + c_3\hat{\eta}_{k+1/2} + c_4\hat{\eta}_{k-1/2}), \\ S_k &\equiv \varphi_k(u_{k-1}, u_k, u_{k+1}, v_{k-1}, v_k, v_{k+1}) - \varphi_k(U_{k-1}, U_k, U_{k+1}, V_{k-1}, V_k, V_{\boldsymbol{k}+1}) \\ &= -\eta_{k+1} + (1+\tau)\eta_k - \tau\eta_{k-1} - h_k^2(c_0\breve{E}_k + c_1\hat{E}_{k+1} + c_2\hat{E}_{k-1} + c_3\hat{E}_{k+1/2} + c_4\hat{E}_{k-1/2}). \end{aligned}$$

Equivalently, in the matrix notation

$$\begin{bmatrix} \boldsymbol{\phi}(\boldsymbol{u}, \boldsymbol{v}) - \boldsymbol{\phi}(\boldsymbol{U}, \boldsymbol{V}) \\ \boldsymbol{\varphi}(\boldsymbol{u}, \boldsymbol{v}) - \boldsymbol{\varphi}(\boldsymbol{U}, \boldsymbol{V}) \end{bmatrix} = \boldsymbol{P} \begin{bmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\eta} \end{bmatrix}, \tag{3.14}$$

where

$\boldsymbol{P} =$

$$\text{tridiag}\left(\begin{bmatrix} C(R_k,\varepsilon_{k-1}) & C(R_k,\eta_{k-1}) \\ C(S_k,\varepsilon_{k-1}) & C(S_k,\eta_{k-1}) \end{bmatrix}, \begin{bmatrix} C(R_k,\varepsilon_k) & C(R_k,\eta_k) \\ C(S_k,\varepsilon_k) & C(S_k,\eta_k) \end{bmatrix}, \begin{bmatrix} C(R_k,\varepsilon_{k+1}) & C(R_k,\eta_{k+1}) \\ C(S_k,\varepsilon_{k+1}) & C(S_k,\eta_{k+1}) \end{bmatrix}\right)$$

is a block tridiagonal matrix and $C(R_k, \eta_k)$ = Coefficient of η_k in R_k etc.

From (3.5) and (3.14), one obtains

$$\boldsymbol{P}\xi = \boldsymbol{T}, \quad \boldsymbol{T} = [\boldsymbol{L} \quad \boldsymbol{M}]^T, \quad \xi = [\varepsilon \quad \eta]^T. \tag{3.15}$$

In the limiting case of small h_k, matrix $\boldsymbol{P}$ takes the form

$$\lim_{h_k \to 0} \boldsymbol{P} = \text{tridiag}\left(\begin{bmatrix} -\tau & 0 \\ 0 & -\tau \end{bmatrix}, \begin{bmatrix} 1+\tau & 0 \\ 0 & 1+\tau \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}\right).$$

Thus, the lower, upper and main diagonal blocks are non-zero, since $\tau > 0$. Hence the graph $G(\boldsymbol{P})$ of the matrix $\boldsymbol{P}$ is strongly connected and consequently, the matrix $\boldsymbol{P}$ is irreducible (Varga [38]).

Let

$$\alpha = \min_k \{\alpha_k, \alpha_{k\pm1}, \alpha_{k\pm1/2}\},\ \beta = \min_k \{\beta_k, \beta_{k\pm1}, \beta_{k\pm1/2}\},$$
$$\gamma = \min_k \{\gamma_k, \gamma_{k\pm1}, \gamma_{k\pm1/2}\},\ \delta = \min_k \{\delta_k, \delta_{k\pm1}, \delta_{k\pm1/2}\}.$$

Further, let $\sum_l$ be the sum of the l^{th} row elements of the matrix $\boldsymbol{P}$, then

For $l = 1$: $\quad \sum_l \geq \tau + O(h_l^2), \sum_{l+1} \geq \tau + O(h_l).$

For $l = 3(2)2n-2$: $\quad \sum_l \geq \frac{h_l^2}{2}\tau(1+\tau) + O(h_l^4), \sum_{l+1} \geq \frac{h_l^2}{2}\tau(1+\tau)(\beta+\delta) + O(h_l^3).$

For $l = 2n-1$: $\quad \sum_l \geq 1 + O(h_l^2), \sum_{l+1} \geq 1 + O(h_l).$

This implies that for sufficiently small value of h_k, i.e. in the limiting case of $h_k \to 0$,

$$\sum_l \geq \tau > 0, l = 1, 2; \quad \sum_l \geq 0,\ l = 3(1)2n-2; \quad \sum_l \geq 1 > 0, l = 2n-1, 2n.$$

Hence, $\boldsymbol{P}$ is monotone (Henrici [39], Young [40]). Consequently $\boldsymbol{P}^{-1}$ exists and is non-negative. Let $P_{i,l}^{-1}$ be the $(i,l)^{th}$ element of $\boldsymbol{P}^{-1}$, and define

$$\left\|\boldsymbol{P}^{-1}\right\|_\infty = \max_{1\leq i\leq 2n} \sum_{l=1}^{2n} \left|P_{i,l}^{-1}\right|, \quad \|T\| = \max_{1\leq l\leq 2n} \sum_{l=1}^{2n} \left|L_l(h_l) + M_l(h_l)\right| = O(h_l^7).$$

From the obvious identity, $\boldsymbol{P}^{-1} = (\boldsymbol{PJ}) = \boldsymbol{J}$, where $\boldsymbol{J} = [1, 1, \ldots, 1]^T$, we obtain

$$\sum_{l=1}^{2n} P_{i,l}^{-1} \sum_l = 1,\ i = 1(1)2n. \tag{3.16}$$

Thus, the following bounds can be estimated by using Taylor series expansions

For $l = 1$:

$$P_{i,l}^{-1} \leq \sum_l^{-1} = \frac{1}{\tau} + O(h_l^2),$$

$$P_{i,l+1}^{-1} \le \sum_{l+1}^{-1} \le \frac{1}{\tau} + O(h_l).$$

For $l = 3(2)2n - 2$:

$$P_{i,l}^{-1} \le \min_l \sum_l^{-1} \le \frac{2}{\tau(1+\tau)h_l^2} + O(h_l^v),\ v \ge 0,$$

$$P_{i,l+1}^{-1} \le \min_l \sum_{l+1}^{-1} \le \frac{2}{\tau(1+\tau)(\beta+\delta)h_l^2} + O(h_l^v),\ v \ge 0.$$

For $l = 2n - 1$:

$$P_{i,l}^{-1} \le \sum_l^{-1} = 1 + O(h_l^2),$$

$$P_{i,l+1}^{-1} \le \sum_{l+1}^{-1} \le 1 + O(h_l).$$

As a result, from Eqs. (3.15) and (3.16), we obtain the following error estimates:

$$\|\xi\| \le \left\|\boldsymbol{P}^{-1}\right\|_\infty .\ \|\boldsymbol{T}\| \le O(h_l^5),\ \text{provided that } \beta + \delta \ne 0. \tag{3.17}$$

This proves the fifth order convergence of the proposed method. Another result is that the coefficients c_k, $k = 0(1)4$ in Eq. (2.1) are negative if $(\sqrt{3}-1)/2 < \tau$ and hence we obtain a lower bound on τ, whereas the upper bound on τ is less than 1.5, otherwise the grid will be too non-uniform to be practical. Thus, we summarise the above result in the following theorem:

Theorem 3.1. The geometric mesh finite difference method (2.29) and (2.30) for the numerical solution of differential equation (1.1) or (1.2) and (1.3) with sufficiently small h_k and $(\sqrt{3}-1)/2 < \tau < 1.5,\ \tau \ne 1,$ gives a fifth order of convergent solution provided that $\dfrac{\partial g}{\partial U} + \dfrac{\partial g}{\partial V} \ne 0.$

4. Computational experiment

To verify the theoretical predictions, we have solved several linear and nonlinear problems. We defined the geometric mesh as follows

$$r_0 = a, h_1 = \begin{cases} (b-a)(1-\tau)/(1-\tau^{n+1}),\ \tau < 1 \\ (b-a)(\tau-1)/(\tau^{n+1}-1),\ \tau > 1 \end{cases}$$

Hence, $h_{k+1} = \tau h_k$, $k = 1(1)n$. If a boundary value problem exhibits a boundary layer at the left boundary, choosing $\tau > 1$ is appropriate. If the layer occurs at the right boundary, we choose $\tau < 1$. If the layer occurs in the interior region, then the mesh can be arranged by choosing $\tau > 1$ in the first half of the interval and $\tau < 1$ in the second half.

The numerical accuracy of the results is expressed using maximum absolute errors $(\varepsilon_{u^{(m)}}^{(\infty)})$ and computational orders of convergence (Θ_m) for m^{th} order derivatives of $u(r)$.

$$\varepsilon_{u^{(m)}}^{(\infty)} = \max_{1\le k\le n}\left|u_k^{(m)} - U_k^{(m)}\right|,\ \Theta_m = \log_2\left(\frac{\varepsilon_{u^{(m)}}^{(2)}\Big|_{n\,\text{grids}}}{\varepsilon_{u^{(m)}}^{(2)}\Big|_{2n\,\text{grids}}}\right).$$

Numerical computations were performed using long double arithmetic extended precision variables having 80 bits and 18 digits precision. The code was written in *C* and run under Linux operating system. For solving linear or nonlinear algebric equations resulting from the discretisation, the Newton method and the Thomas algorithm were used, with the error tolerance being $\le 10^{-15}$.

Example 4.1 (Conte [41]) The fourth order two point boundary value problem

$$U^{(4)}(r) - (1+\lambda)U^{(2)}(r) + \lambda U(r) = \frac{\lambda}{2}r^2 + 1, 0 < r < 1,$$

$$U(0) = 1,\ U(1) = \frac{3}{2} + \sinh(1),\ U^{(2)}(0) = 1,\ U^{(2)}(1) = 1 + \sinh(1),$$

possesses analytical solution $U(r) = 1 + \dfrac{r^2}{2} + \sinh(r)$. We know that ± 1 and $\pm\lambda$ are the eigenvalues of this equation and hence the problem is stiff for large values of λ. We have solved the problem for small as well as for large values of λ. The solution is found accurate for $\lambda < 10^8$ both in the case of uniform and geometric meshes. Table 3 presents errors of the approximate solutions and computational orders of convergence obtained for $\lambda = 10^8$, in the case of uniform meshes ($\tau = 1$) and geometric meshes ($\tau \ne 1$). It is evident that the geometric mesh technique is superior to the uniform mesh.

Example 4.2 (Mohanty [33]) The fourth order singular linear problem in polar coordinates

$$\nabla^4 U(r) \equiv \left(\frac{d^2}{dr^2} + \frac{\lambda}{r}\frac{d}{dr}\right)^2 U(r) = \left(1 + \frac{2\lambda}{r} + \frac{\lambda(\lambda-2)}{r^2} - \frac{\lambda(\lambda-2)}{r^3}\right)e^r,\ 0 < r < 1,$$

$$U(0) = U^{(2)}(0) = 1,\ U(1) = U^{(2)}(1) = e,$$

possesses analytical solution $U(r) = e^r$. The choice of $\lambda = 0,1$ and 2, corresponds to Cartesian, cylindrical and spherical coordinates respectively.The errors for the various values of n and λ are reported in Table 4.

Example 4.3 (Elcrat [42]) The nonlinear boundary value problem arising from a model of the axisymmetric flow of an incompressible fluid contained between infinite disks is:

$$U^{(4)}(r) = \lambda U(r)U^{(2)}(r) - \lambda(r^2-1)(1+4r+r^2)e^{2r} - (11+8r+r^2)e^r,\ 0 < r < 1,$$

$$U(0) = 1,\ U(1) = 0,\ U^{(2)}(0) = -1,\ U^{(2)}(1) = -6e.$$

The analytical solution is $U(r) = (1 - r^2)e^r$. The errors obtained are given in Table 5, for various values of n, and for $\lambda = 10^3$.

Example 4.4 (Takaoka [43]) The boundary value problem arising from the steady state form of the Korteweg-de Vries equation of fifth order is:

$$U^{(4)}(r) = \lambda U^{(2)}(r) + \frac{1}{2}U(r)^2 - U(r)$$
$$+ \frac{\lambda}{2}\sin(10\pi r)[2 + 200\pi^2(\lambda + 100\pi^2) - \lambda\sin(10\pi r)],$$
$$U(0) = U(1) = U^{(2)}(0) = U^{(2)}(1) = 0,\ 0 < r < 1.$$

The analytical solution is $U(r) = \lambda\sin(10\pi r)$. The maximum absolute errors obtained for $\lambda = 4$ are given in Table 6 for various values of n.

5. Conclusion and remarks

A compact, three point finite difference scheme using geometric mesh has been designed to obtain accurate numerical solutions of fourth order two point regular and singular boundary value problems for nonlinear ordinary differential equations. The theoretical order of accuracy is 5 (or 6 in the limit of uniform meshes).The scheme is shown theoretically to be convergent when the grid ratio τ is $(\sqrt{3}-1)/2 < \tau < 1.5$.

Computational tests confirm that the scheme converges and is applicable both to singular and non singular differential equations. Numerical solutions obtained using geometric meshes prove more accurate than those corresponding to uniform meshes, when local layers are present. The scheme can be effectively combined with the Newton-method and Thomas algorithm for solving block-tridiagonal linear algebraic systems arising in the calculations.

The authors would like to thank Indian National Science Academy and Polish Academy of Sciences for the support of this research work which was funded by the grant: Intl/PAS/2014/2608 received by the first author.

Table 1

Expressions for the coefficients a_{lm}, b_{lm}, l, m = 1(1)6 in Eqs. (2.9) and (2.10)

$a_{11} = -\tau^3(5\tau+12)/[16\sigma(1+\tau)]$	$b_{11} = -3\tau^4/[16\rho(1+\tau)]$
$a_{12} = (\tau+2)(5\tau^2+10\tau+4)/(16\sigma)$	$b_{12} = (\tau+2)(3\tau^2+2\tau+4)/(16\rho)$
$a_{13} = (\tau+2)(3\tau^2+14\tau+4)/[16\sigma(1+\tau)]$	$b_{13} = (\tau+2)(5\tau^2+6\tau+4)/[16\rho(1+\tau)]$
$a_{14} = (\tau+2)(4\tau+3)\tau^3/[96\sigma(1+\tau)]$	$b_{14} = \tau^4(\tau+2)^2/[96\rho(1+\tau)]$
$a_{15} = 0$	$b_{15} = -\tau^2(\tau+2)(\tau^2+2\tau+3)/(96\rho)$
$a_{16} = -\tau^2(\tau+2)(\tau^2+6\tau+6)/[96\sigma(1+\tau)]$	$b_{16} = -\tau^2(\tau+2)(2\tau^2+4\tau+3)/[96\rho(\tau+1)]$
$a_{21} = (2\tau+1)(4\tau^2+14\tau+3)/[16\sigma(\tau+1)]$	$b_{21} = (2\tau+1)(4\tau^2+6\tau+5)/[16\rho(\tau+1)]$
$a_{22} = (2\tau+1)(4\tau^2+10\tau+5)/(16\sigma\tau)$	$b_{22} = (2\tau+1)(4\tau^2+2\tau+3)/(16\rho\tau)$
$a_{23} = -(12\tau+5)/[16\sigma(1+\tau)\tau]$	$b_{23} = -3/[16\rho(\tau+1)\tau]$
$a_{24} = -(2\tau+1)(6\tau^2+6\tau+1)/[96\sigma(1+\tau)]$	$b_{24} = -(2\tau+1)(3\tau^2+4\tau+2)/[96\rho(\tau+1)]$
$a_{25} = 0$	$b_{25} = (2\tau+1)(3\tau^2+2\tau+1)/(96\rho\tau)$
$a_{26} = (3\tau+4)(2\tau+1)/[96\sigma(1+\tau)]$	$b_{26} = (2\tau+1)^2/[96\rho(\tau+1)\tau]$
$a_{31} = (\tau+2)\tau^2/[h_k\sigma(1+\tau)]$	$b_{31} = -\tau^2(\tau+2)/[h_k\rho(\tau+1)]$
$a_{32} = -(\tau+1)^2/(h_k\sigma\tau)$	$b_{32} = (\tau-1)(\tau+1)^2/(h_k\rho\tau)$
$a_{33} = (2\tau^3+6\tau^2+4\tau+1)/[h_k\sigma(1+\tau)\tau]$	$b_{33} = (2\tau+1)/[h_k\rho(\tau+1)\tau]$
$a_{34} = -(\tau+1)\tau^2/(6h_k\sigma)$	$b_{34} = \tau^2(1-\tau^2)/(6h_k\rho)$
$a_{35} = 0$	$b_{35} = \tau(2+\tau)(1+\tau)^2/(6h_k\rho)$
$a_{36} = \tau(\tau+3)(\tau+1)/(6h_k\sigma)$	$b_{36} = \tau(1+\tau)(1+2\tau)/(6h_k\rho)$
$a_{41} = -(\tau^3+4\tau^2+6\tau+2)/[h_k\sigma(1+\tau)]$	$b_{41} = -\tau^2(\tau+2)/[h_k\rho(1+\tau)]$
$a_{42} = (\tau+1)^3/(h_k\sigma\tau)$	$b_{42} = (\tau-1)(\tau+1)^2/(h_k\rho\tau)$
$a_{43} = -(2\tau+1)/[h_k\sigma(1+\tau)\tau]$	$b_{43} = (2\tau+1)/[h_k\rho(\tau+1)\tau]$
$a_{44} = -(\tau+1)(3\tau+1)/(6h_k\sigma)$	$b_{44} = -(\tau+2)(\tau+1)/(6h_k\rho)$
$a_{45} = 0$	$b_{45} = -(2\tau+1)(\tau+1)^2/(6h_k\rho\tau)$
$a_{46} = (\tau+1)/(6h_k\sigma)$	$b_{46} = (1-\tau^2)/(6h_k\rho)$

$a_{51} = \tau^2 / [2h_k\sigma(1+\tau)]$	$b_{51} = \tau^2(\tau+2) / [2h_k\rho(1+\tau)]$
$a_{52} = -(3\tau^2+6\tau+2) / (2h_k\sigma\tau)$	$b_{52} = -(\tau^3+4\tau^2+2\tau+2) / (2h_k\rho\tau)$
$a_{53} = (\tau^2+4\tau+2)(2\tau+1) / [2h_k\sigma(1+\tau)\tau]$	$b_{53} = (3\tau^3+6\tau^2+4\tau+2) / [2h_k\rho(\tau+1)\tau]$
$a_{54} = (\tau^2-\tau-1)\tau^2 / [24h_k\sigma(1+\tau)]$	$b_{54} = \tau^2(\tau^2-\tau-1)(\tau+2) / [24h_k\rho(1+\tau)]$
$a_{55} = 0$	$b_{55} = -\tau(\tau^3+4\tau^2+6\tau+1) / (24h_k\rho)$
$a_{56} = -(\tau_k^2+5\tau_k+5)\tau_k^2 / [24h_k\sigma_k(1+\tau_k)]$	$b_{56} = -\tau(2\tau^3+5\tau^2+3\tau-1) / [24h_k\rho(1+\tau)]$
$a_{61} = -(2\tau^2+4\tau+1)(\tau+2) / [2h_k\sigma(1+\tau)]$	$b_{61} = -(2\tau^3+4\tau^2+6\tau+3) / [2h_k\rho(1+\tau)]$
$a_{62} = (2\tau^2+6\tau+3) / (2h_k\sigma)$	$b_{62} = (2\tau^3+2\tau^2+4\tau+1) / (2h_k\rho\tau)$
$a_{63} = -1 / [2h_k\sigma(1+\tau)]$	$b_{63} = -(1+2\tau) / [2h_k\rho(1+\tau)\tau]$
$a_{64} = (5\tau^2+5\tau+1) / [24h_k\sigma(1+\tau)]$	$b_{64} = -(\tau^3-3\tau^2-5\tau-2) / [24h_k\rho(1+\tau)]$
$a_{65} = 0$	$b_{65} = (\tau^3+6\tau^2+4\tau+1) / (2h_k\rho)$
$a_{66} = (\tau^2+\tau-1) / [24h_k\sigma(1+\tau)]$	$b_{66} = (\tau^2+\tau-1)(1+2\tau) / [24h_k\rho(1+\tau)\tau]$

Table 2

Expressions for the coefficients t_i, i = 0(1)3, z_j j = 1(1)6 in Eqs. (2.27) and (2.28)

$t_0 = -(1+\tau)(27\tau^5+133\tau^4+155\tau^3-10\tau^2-62\tau-18) / [60\delta\sigma(2+\tau)]$
$t_1 = -(3\tau^6+60\tau^5+302\tau^4+555\tau^3+422\tau^2+140\tau+18) / [60\sigma(2+\tau)(1+\tau)\delta]$
$t_2 = \tau(27\tau^6+190\tau^5+508\tau^4+735\tau^3+628\tau^2+270\tau+42) / [60\delta\sigma(2+\tau)(1+\tau)]$
$t_3 = -\tau(12\tau^6+65\tau^5+103\tau^4+90\tau^3+103\tau^2+65\tau+12) / [120\delta\sigma(2+\tau)]$
$z_1 = (6\tau^6+15\tau^5-\tau^4-28\tau^3-\tau^2+15\tau+6) / [6\delta\rho(1+\tau)^2]$
$z_2 = -\tau(6\tau^6+15\tau^5-\tau^4-28\tau^3-\tau^2+15\tau+6) / [6\delta\rho(1+\tau)^2]$
$z_3 = -(27\tau^7+70\tau^6+20\tau^5-52\tau^4+83\tau^3+100\tau^2+25\tau-3) / [30\delta\rho(1+2\tau)(1+\tau)^2]$
$z_4 = -\tau(3\tau^7-25\tau^6-100\tau^5-83\tau^4+52\tau^3-20\tau^2-70\tau-27) / [30\delta\rho(2+\tau)(1+\tau)^2]$
$z_5 = -(48\tau^6+157\tau^5+133\tau^4-21\tau^3+83\tau^2+107\tau+33) / [15\delta\rho(2+\tau)(1+\tau)]$
$z_6 = \tau(33\tau^6+107\tau^5+83\tau^4-21\tau^3+133\tau^2+157\tau+48) / [15\delta\rho(1+2\tau)(1+\tau)]$

Table 3

Solution errors obtained for example 1*

n	λ	$\varepsilon_u^{(\infty)}$	$\varepsilon_{u^{(2)}}^{(\infty)}$	Θ_0	Θ_2	τ	$\varepsilon_u^{(\infty)}$	$\varepsilon_{u^{(2)}}^{(\infty)}$
8	1e08	2.40e-11	2.40e-11	---	---	0.9980	4.52e-12	4.52e-12
16	1e08	5.32e-13	5.32e-13	5.5	5.5	0.9991	6.00e-14	5.98e-14
32	1e08	1.58e-14	1.58e-14	5.1	5.1	0.9997	9.70e-16	9.71e-16

Table 4

Solution errors obtained for example 2*

n	λ	$\varepsilon_u^{(\infty)}$	$\varepsilon_{u^{(2)}}^{(\infty)}$	Θ_0	Θ_2	τ	$\varepsilon_u^{(\infty)}$	$\varepsilon_{u^{(2)}}^{(\infty)}$
8	0	1.97e-07	8.09e-08	---	---	0.985	1.90e-08	3.12e-08
16	0	4.34e-09	1.78e-09	5.5	5.5	0.991	8.82e-10	9.61e-10
32	0	8.12e-11	3.34e-11	5.7	5.7	0.996	8.52e-12	1.17e-11
8	1	7.56e-05	1.45e-03	---	---	1.160	1.67e-05	2.75e-04
16	1	7.80e-06	3.81e-04	3.3	2.0	1.110	3.46e-07	5.25e-05
32	1	7.50e-07	8.86e-05	3.4	2.1	1.040	6.70e-08	1.84e-05
8	2	5.64e-05	5.23e-04	---	---	0.910	1.22e-05	8.53e-04
16	2	3.94e-06	3.75e-05	3.8	3.9	0.960	1.21e-06	1.68e-05
32	2	2.65e-07	2.49e-06	3.9	3.8	0.790	8.67e-08	1.98e-06

Table 5

Solution errors obtained for example 3*

n	λ	$\varepsilon_u^{(\infty)}$	$\varepsilon_{u^{(2)}}^{(\infty)}$	Θ_0	Θ_2	τ	$\varepsilon_u^{(\infty)}$	$\varepsilon_{u^{(2)}}^{(\infty)}$
8	1e03	1.53e-09	8.31e-08	---	---	0.96	1.05e-10	2.72e-08
16	1e03	2.77e-11	1.51e-09	5.8	5.9	0.98	2.29e-12	3.79e-10
32	1e03	4.69e-13	2.53e-11	5.9	5.9	0.99	4.42e-14	5.84e-12

Table 6

Solution errors obtained for example 4*

n	λ	$\varepsilon_u^{(\infty)}$	$\varepsilon_{u^{(2)}}^{(\infty)}$	Θ_0	Θ_2	τ	$\varepsilon_u^{(\infty)}$	$\varepsilon_{u^{(2)}}^{(\infty)}$
8	4	2.40e-10	3.90e-09	---	---	0.995	2.99e-11	4.66e-09
16	4	5.32e-12	8.57e-11	5.5	5.5	0.997	6.49e-13	1.05e-10
32	4	1.06e-13	1.60e-12	5.7	5.8	0998	2.55e-14	2.07e-12

* Column 3–6 refer to uniform meshes, column 7–9 refer to geometric meshes.

References

[1] Yu-Li Y., Kaveh M., *Fourth order partial differential equations for noise removal*, IEEE Trans. Image Process. 9, 2000, 1723-1730.

[2] Zhong H., *Spline-based differential quadrature for fourth order differential equations and its application to Kirchhoff plates*, Appl. Math. Model, 28, 2004, 353-366.

[3] Timoshenko S., Krieger, S.W., *Theory of plates and shells*, McGraw Hill, 1987.

[4] Chen Y., McKenna P.J., *Traveling waves in a nonlinearly suspended beam: theoretical results and numerical observations*, J. Differ. Equ., 136, 1991, 325-335.

[5] Budd C.J., Hunt G.W., Peletier, M.A., *Self-similar fold evolution under prescribed end shortening*, Math. Geol., 31, 1999, 989-1005.

[6] Wasow W., *The complex asymptotic theory of a fourth order differential equation of hydrodynamics*, Ann. Math., 46, 1948, 852-871.

[7] O'Regan D., *Solvability of some fourth (and higher) order singular boundary value problems*, J. Math. Anal. Appl., 161, 1991, 78-116.

[8] Agrawal R.P., Chow, Y.M., *Iterative methods for a fourth order boundary value problem*, J. Comput. Appl. Math., 10, 1984, 203-217.

[9] Aftabizadeh A.R., *Existence and uniqueness theorems for fourth-order boundary value problems*, J. Math. Anal. Appl., 116, 1986, 415-426.

[10] Momani S., Noor M.A., *Numerical comparison of methods for solving a special fourth-order boundary value problem*, Appl. Math. Comput., 191, 2007, 218-224.

[11] Wazwaz A.M., *The numerical solution of spacial fourth-order boundary value problem by the modified decomposition method*, Int. J. Comput. Math., 79, 2002, 345-356.

[12] Mohyud-Din S.T., Noor M.A., *Homotopy perturbation method for solving fourth order boundary value problems*, Math. Probl. Eng., 98602, 2007, 1-15.

[13] Noor M.A., Mohyud-Din S.T., *An efficient method for fourth-order boundary value problems*, Comput. Math. Appl., 54, 2007, 1101-1111.

[14] Zahra W.K., *A smooth approximation based on exponential spline solutions for nonlinear fourth order two point boundary value problems*, Appl. Math. Comput., 217, 2011, 8447-8457.

[15] Usmani R.A., Taylor P.J., *Finite difference methods for solving* $(p(x)y'')'' + q(x)y = r(x)$, Int. J. Comput. Math., 14, 1983, 277-293.

[16] Schroder J., *Numerical error bounds for fourth order boundary problems, simultaneous estimation of* $u(x)$ *and* $u''(x)$, Numer. Math., 44, 1984, 233-245.

[17] Shanthi V., Ramanujam, N., *A numerical method for boundary value problems for singularly perturbed fourth-order ordinary differential equations*, Appl. Math. Comput., 129, 2002, 269-294.

[18] Twizell E.H., Boutayeb A., *Numerical methods for the solution of special and general sixth-order boundary-value problems, with applications to Benard layer eigenvalue problems*, Proc. Royal Soc. Lond. A Mat., 431, 1990, 433-450.

[19] Jain M.K., Iyengar, S.R.K., Subramanyam, G.S., *Variable mesh method for the numerical solution of two point singular perturbation problems*, Comput. Meth. Appl. Mech. Eng., 42, 1984, 273-286.

[20] Kadalbajoo M.K., Kumar D., *Geometric mesh FDM for self-adjoint singular perturbation boundary value problems*, Appl. Math. Comput., 190, 2007, 1646-1656.

[21] Mohanty R.K., *A class of non-uniform mesh three point arithmetic average discretization for* $y'' = f(x, y, y')$ *and the estimates of* y, Appl. Math. Comput., 183, 2006, 477-485.

[22] Britz D., *Digital simulation in electrochemistry*, Springer, Berlin 2005.

[23] Roos H.G., Stynes M., Tobiska L., *Numerical methods for singularly perturbed differential equations convection diffusion and flow problems*, Springer, Berlin 1996.
[24] Farrell P.A., Hegarty, A.F., Miller, J.J.H., O'Riordan, E., Shishkin, G.I., *Robust Computational Techniques for Boundary Layers*, Chapman & Hall/CRC, Boca Raton, 2000.
[25] Thomas L.H., *Elliptic problems in linear difference equations over a network Watson Scientific Computing Laboratory Report*, Columbia University, New York 1949.
[26] Bieniasz L.K., *Extension of the Thomas algorithm to a class of algebraic linear equation systems involving quasi-block-tridiagonal matrices with isolated block pentadiagonal rows, assuming variable block dimension*, Computing. 67, 2001, 269-285 (With erratum in Computing 70, 2003, 275).
[27] Numerov B.V., *A method of extrapolation of perturbation*, Royal Astron. Soc. Mon. Notices. 84, 1924, 592-601.
[28] Agarwal R.P., *Some recent developments of Numerov's method*, Comput. Math. Appl., 42, 2001, 561-592.
[29] Chawla M.M., *High accuracy tridiagonal finite difference approximations for non linear two point boundary value problems*, J. Inst. Maths. Appl., 22, 1978, 203-209.
[30] Chawla M.M., *A sixth-order tridiagonal finite difference method for general non-linera two--point boundary value problems*, IMA J. Appl. Math., 24, 1979, 35-42.
[31] Wang, Y.M.. *Numerov's method for strongly nonlinear two-point boundary value problems*, Comput. Math. Appl., 45, 2003, 759-763.
[32] Bieniasz L.K., *Two new compact finite difference schemes for the solution of boundary value problems in second order nonlinear ordinary differential equations using non-uniform grids*, J. Comput. Math. Sci. Eng., 8, 2008, 3-18.
[33] Mohanty R.K., Jha, N., Chauhan, V., *Arithmetic average geometric mesh discretizations for fourth and sixth order nonlinear two point boundary value problems*, Neural Parallel Sci. Comput., 21, 2013, 393-410.
[34] Zhang X.Y., Fang, Q., *A sixth order numerical method for a class of nonlinear two-point boundary value problems*, Numer. Algebra. Contr. Optim., 2, 2012, 31-43.
[35] Jha N., *A fifth order accurate geometric mesh finite difference method for general nonlinear two point boundary value problems*, Appl. Math. Comput., 219, 2013, 8425-8434.
[36] Jha N., Mohanty R.K., Chauhan, V., *Geometric mesh three point discretization for fourth order nonlinear singular differential equations in polar system*, Adv. Numer. Anal., 614508, 2013, 1-10.
[37] http://www.maplesoft.com/solutions/education/solutions/matheducation.aspx (access: 30.01.2015).
[38] Varga R.S., *Matrix iterative analysis*, Springer Series in Computational Mathematics, Springer, Berlin, 2000.
[39] Henrici P., *Discrete variable methods in ordinary differential equations*, Wiley, New York, 1962.
[40] Young D.M., *Iterative solution of large linear systems*, Academic Press, New York 1971.
[41] Conte S.D., *The numerical solution of linear boundary value problems*, SIAM Rev., 8, 1966, 309-321.
[42] Elcrat A.R., *On the radial, flow of a viscous fluid between porous disks*, Arch. Ration. Mech. Anal., 61, 1976, 91-96.
[43] Takaoka M., *Pole distribution and steady pulse solution of the fifth order Korteweg-de Vries equation*, J. Phys. Soc. Jpn., 58, 1989, 73-81.

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

BEATA KOCEL-CYNK*

HAUSDORFF LIMITS OF ONE PARAMETER FAMILIES OF DEFINABLE SETS IN *O*-MINIMAL STRUCTURES

GRANICE HAUSDORFFA JEDNOPARAMETROWYCH RODZIN ZBIORÓW DEFINIOWALNYCH W STRUKTURACH *O*-MINIMALNYCH

A b s t r a c t

We give an elementary proof of the following theorem on definability of Hausdorff limits of one parameter families of definable sets: let $A \subset \mathbb{R} \times \mathbb{R}^n$ be a bounded definable subset in *o*-minimal structure on $(\mathbb{R},+,\cdot)$ such that for any $y \in (0,c)$, $c>0$, the fibre $A_y := \{x \in \mathbb{R}^n : (y,x) \in A\}$ is a Lipschitz cell with constant L independent of y. Then the Hausdorff limit $\lim_{y \to 0} \overline{A}_y$ exists and is definable.

Keywords: Hausdorff limit, *definable sets*, *o-minimal structure*

S t r e s z c z e n i e

W prezentowanej pracy przedstawiamy elementarny dowód następującego twierdzenia o definiowalności granicy Hausdorffa jednoparametrowej rodziny zbiorów definiowalnych: niech $A \subset \mathbb{R} \times \mathbb{R}^n$ będzie ograniczonym zbiorem definiowalnym w strukturze *o*-minimalnej typu $(\mathbb{R},+,\cdot)$ takim, że dla dowolnego $y \in (0,c)$, $c>0$, wókno $A_y := \{x \in \mathbb{R}^n : (y,x) \in A\}$ jest komórką Lipschitza ze staą L niezależną od y. Wtedy granica Hausdorffa $\lim_{y \to 0} \overline{A}_y$ istnieje i jest definiowalna.

Słowa kluczowe: granica Hausdorffa, *zbiory definiowalne*, *struktury o-minimalne*

DOI: 10.4467/2353737XCT.16.140.5751

* Beata Kocel-Cynk (bkocel@pk.edu.pl), Institute of Mathematics, Faculty of Physics, Mathematics and Computer Science, Cracow University of Technology.

1. Introduction

In [1] Bröcker proved that for any family of semialgebraic sets A_y and any convergent sequence y_v of parameters the Hausdorff limit of A_{y_v} exists and is semialgebraic. In [3] a short geometric proof of the generalization of Bröcker's result to the case of sets definable in an *o*-minimal structure was given.

The aim of this paper is to present an elementary proof of the following one-parameter case of this result

Theorem 1. *Let* $A \subset \mathbb{R} \times \mathbb{R}^n$ *be a definable subset in an o-minimal structure on* $(\mathbb{R}, +, \cdot)$ *such that for any* $y \in (0, c),\ c > 0,$ *the fibre* $A_y := \{x \in \mathbb{R}^n : (y, x) \in A\}$ *is a bounded Lipschitz cell with constant L independent of y. Then the Hausdorff limit* $\lim_{y \to 0} \overline{A}_y$ *exists and is definable.*

For the convenience of the reader we present in Section 2 results on Hausdorff distance and *o*-minimal structure that we use in the proof of the main result.

2. Preliminaries

2.1. Hausdorff distance.

Let (X, d) be a complete metric space, denote by $\mathcal{C}(X)$ the space of all non-empty compact subsets in X.

Definition 1. *For any two sets* $Y_1, Y_2 \in \mathcal{C}(X)$ *we define Hausdorff distance as*

$$d_H(Y_1, Y_2) = \max\{\max_{x \in Y_1} \min_{y \in Y_2} d(x, y), \max_{y \in Y_2} \min_{x \in Y_1} d(x, y)\}$$

Remark 1. *Hausdorff distance of two sets is the infimum of positive numbers* $\varepsilon > 0$ *such that each of them is contained in the* ε*-envelope of the other, i.e.*

$$d_H(Y_1, Y_2) = \inf\{\varepsilon > 0;\ Y_2 \subseteq B(Y_1, \varepsilon) \text{ and } Y_1 \subseteq B(Y_2, \varepsilon)\}$$

where

$$B(Z, \varepsilon) = \bigcup_{z \in Z} B(z, \varepsilon)$$

for any $Z \in \mathcal{C}(X)$ and $\varepsilon > 0$.

Remark 2. *Observe that the function* $\tilde{d} : \mathcal{C}(X) \times \mathcal{C}(X) \to \mathbb{R}_+$ *defined by the following formula*

$$\tilde{d}(Y_1, Y_2) := \max\{\tilde{d}(x, Y_2) : x \in Y_1\}, \quad \text{for} \quad Y_1, Y_2 \in \mathcal{C}(X)$$

where

$$\tilde{d}(x, Y) := \min\{d(x, y) : y \in Y\}, \quad \text{for} \quad x \in X, Y \in \mathcal{C}(X)$$

cannot be used to define a metric on $\mathcal{C}(X)$ as in general the function $\tilde{d}$ is not symmetric, we have only the following

$$d_H(Y_1,Y_2)=\max\{\tilde{d}(Y_1,Y_2),\ \tilde{d}(Y_2,Y_1)\} \quad \text{for} \quad Y_1Y_2\in\mathcal{C}(X).$$

Example 2. *Let* $Y_1=(0,15)$ *and* Y_2:=$[8,112]\times\{0\}$, *then*

$$\tilde{d}(Y_1,Y_2)=17=113=\tilde{d}(Y_2,Y_1).$$

By definition, in this example we have $d_H(Y_1, Y_2)$=113.

We end this section with the following characterization of convergence in Hausdorff metric.

Theorem 3. *Let X be a compact metric space,* $A,A_v\in\mathcal{C}(X),\ \nu=1,2,3,\ldots$. *Then the sequence* A_v *converges to A in Hausdorff metric* $(A_\nu\longrightarrow A)$ *iff the following two conditions hold*

1) $(x_{\nu_k}\in A_{\nu_k}, x_{\nu_k}\longrightarrow x_0, \nu_1<\nu_2<\nu_3<\ldots)\Rightarrow x_0\in A,$

2) $x_0\in A\Rightarrow\exists x_\nu\in A_\nu$ *such that* $x_\nu\longrightarrow x_0$.

Proof. First we shall prove that conditions 1) and 2) are necessary for the convergence in Hausdorff metric.

Assume that $A_\nu\longrightarrow A,$ since X is a compact set we can find a sequence $x_{\nu_k}\in A_{\nu_k}$ (with $\nu_1<\nu_2<\nu_3<\ldots$) such that $x_{\nu_k}\longrightarrow x_0$ for some poin $x_0\in X$. We want to show that $x_0\in A$. Since the set A is compact and $x_{\nu_k}\in A_{\nu_k}$ there exists $y_{\nu_k}\in A$ such that

$$d(x_{\nu_k},y_{\nu_k})=\tilde{d}(x_{\nu_k},A)\le d_H(A_{\nu_k},A)\to 0$$

Therefore $d(x_{\nu_k},y_{\nu_k})\longrightarrow 0.$ We shall show that $\tilde{d}(x_0,A)=0.$ Observe that

$$\tilde{d}(x_0,A)\le d(x_0,y_{\nu_k})$$

As $y_{\nu_k}\in A$ and consequently

$$d(x_0,y_{\nu_k})\le d(x_0,x_{\nu_k})+d(x_{\nu_k},y_{\nu_k}).$$

Therefore $\tilde{d}(x_0,A)=0$ and $x_0\in\overline{A}=A.$

Assume that $A_\nu\longrightarrow A$ and $x_0\in A$. To prove that condition 2) is necessary fix a point $x_\nu\in A_\nu$ for $v=1,2,\ldots$ such that $d(x_0,x_\nu)=\tilde{d}(x_0,A_\nu).$ Then

$$0\le d(x_0,x_\nu)=\tilde{d}(x_0,A_\nu)\le\tilde{d}(x_0,A_\nu)\le d_H(A,A_\nu)\longrightarrow 0$$

implies $d(x_0,x_\nu)\to 0$.

Now, we shall prove the opposite implication. Assume to the contrary that conditions 1) and 2) hold while the sequence (A_ν) does not converge to A. Then there exists $\varepsilon > 0$ such that $d_H(A_\nu, A) > \varepsilon$ for infinitely many ν. Consequently at least one of the inequalities

$$\tilde{d}(A_\nu, A) > \varepsilon \quad \text{or} \quad \tilde{d}(A, A_\nu) > \varepsilon$$

holds for infinitely many ν.

In the first case there exist $\nu_1 < \nu_2 < \dots$ and $x_{\nu_k} \in A$ such that $\tilde{d}(x_{\nu_k}, A) > \varepsilon$, since X is compact replacing x_{ν_k} by a subsequence we can also assume that x_{ν_k} converges to a point $x_0 \in X$. From condition 1) we get $x_0 \in A$ which contradicts $\tilde{d}(x_{\nu_k}, A) > \varepsilon$.

In the second case for infinitely many ν there exists $y_\nu \in A$ such that $\tilde{d}(y_\nu, A_\nu) > \varepsilon$, by compactness of A there exists a sequence $\nu_1 < \nu_2 < \dots$ such that $\tilde{d}(y_{\nu_k}, A_{\nu_k}) > \varepsilon$ and $y_{\nu_k} \longrightarrow x_0$ for some $x_0 \in A$. By condition 2) there exists $x_{\nu_k} \in A_{\nu_k}$ such that $x_{\nu_k} \longrightarrow x_0$. In this situation we have

$$\varepsilon < \tilde{d}(y_{\nu_k}, A_{\nu_k}) \le d(y_{\nu_k}, x_{\nu_k}) \le d(y_{\nu_k}, x_0) + d(x_0, x_{\nu_k}) \longrightarrow 0$$

which is a contradiction.

□

Remark 3. The above theorem does not hold without the assumption that X is a compact space.

Example 4. Let X be any non-compact complete space, fix $x_0 \in X$, let $x_\nu \in X$ be a sequence that does not contain any convergent subsequence. Put $A := \{x_0\}$, $A_\nu = \{x_0, x_\nu\}$. Then conditions 1) and 2) hold true but the sequence A_ν does not converge in Hausdoff metric.

2.2. ***o***-minimal structures.

We shall collect here the basic definitions and properties of ***o***-minimal structures that are crucial for our further considerations. For a detailed exposition of ***o***-minimal structures we refer the reader to [2].

Definition 2. A structure $\mathcal{S}$ on $\mathbb{R}$ consists of a collection $\mathcal{S}_n$ of subsets of $\mathbb{R}^n$, for each $n \in \mathbb{N}$, such that

1. $\mathcal{S}_n$ is a boolean algebra of subsets of $\mathbb{R}^n$,
2. $\mathcal{S}_n$ contains the diagonals $d(x_0, x\{(x_1, \dots x_n) \in \mathbb{R}^n : x_i = x_i\}$ for $1 \le i < j \le n$,
3. if $A \in \mathcal{S}_{n+1}$, then $A \times \mathbb{R}$ and $\mathbb{R} \times A$ belong to $\mathcal{S}_{n+1}$,
4. if $A \in \mathcal{S}_{n+1}$, then $\pi(A) \in \mathcal{S}_n$, where $\pi: \mathbb{R}^{n+1} \to \mathbb{R}^n$ is the projection on the first n coordinates.

We say that a set $A \subset \mathbb{R}^n$ is *definable* if and only if $A \in \mathcal{S}_n$. A function $f: A \to \mathbb{R}^m$ with $A \subset \mathbb{R}^n$ is called *definable* if and only if its graph is definable.

Definition 3. A structure $\mathcal{S}$ on $\mathbb{R}$ is *o-minimal* if and only if

1. $\{(x,y): x<y\} \in \mathcal{S}_2$ and $\{a\} \in \mathcal{S}_1$ for each $a \in \mathbb{R}$,

2. each set in is a finite union of intervals (a,b), $-\infty \le a < b \le +\infty$, and points $\{a\}$.

A *structure on* $(\mathbb{R},+,\cdot)$ is a structure on $\mathbb{R}$ containing the graphs of both addition and multiplication.

The main technical tool used in the studies of geometry of sets definable in *o*-minimal structures is the cell decomposition. The notions of a cell and that of a cell decomposition are defined inductively.

Definition 4. The *cells* in $\mathbb{R}^1$ exactly are points and open intervals.

A definable set $C \subset \mathbb{R}^n$, where $n > 1$, is a *cell* if its image $\pi(C) \subset \mathbb{R}^{n-1}$ by the projection $\pi: \mathbb{R}^n \ni (x_1,\ldots,x_{n-1},x_n) \longrightarrow (x_1,\ldots,x_{n-1}) \in \mathbb{R}^{n-1}$ is a cell and C is one of the following two types:

either

$$C = \Gamma(f) = \{(x', x_n) \in \pi(C) \times \mathbb{R} : x_n = f(x')\}$$

(and then C is called a *graph*)

or

$$C = (g_1, g_2) := \{(x', x_n) \in \pi(C) \times \mathbb{R} : g_1(x') < x_n < g_2(x')\}$$

(and then C is called a *band*),

where f: $\pi(C) \to \mathbb{R}$ is a continuous definable function (resp. $g_1, g_2 : \pi(C) \to \overline{\mathbb{R}}$ are functions such that $g_1 < g_2$ on $\pi(C)$ and, for each $i \in \{1,2\}$, g_i is either a continuous definable function g_i: $\pi(C) \to \mathbb{R}$ or g_i is identically equal to $-\infty$, or else g_i is identically equal to $+\infty$).

A cell C is called a $\mathcal{C}^k$-*cell* (where $k \in \mathbb{N} \cup \{\infty\}$), if $\pi(C)$ is a $\mathcal{C}^k$-cell and f (resp. g_i, $i = 1,2$ if finite) is a $\mathcal{C}^k$-function. Notice that every $\mathcal{C}^k$-cell is a $\mathcal{C}^k$-submanifold of $\mathbb{R}^n$.

Definition 5. A *cell decomposition* of $\mathbb{R}^1$ is a finite collection of open intervals and points of the following form:

$$\{(-\infty, a_1), (a_1, a_2), \ldots, (a_k, +\infty), \{a_1\}, \ldots, \{a_k\}\},$$

where $a_1 < a_2 < \ldots < a_k$ are real numbers.

A *cell decomposition* of $\mathbb{R}^n$ $(n > 1)$ is a finite partition $\mathcal{C}$ of $\mathbb{R}^n$ into cells such that the set of all projections $\{\pi(C)$: $C \in \mathcal{C}\}$ is a cell decomposition of $\mathbb{R}^{n-1}$, where π: $\mathbb{R}^n \to \mathbb{R}^{n-1}$ is the projection on the first $n-1$ coordinates as in Definition 4.

Theorem 5. *Let (X,d) be a compact metric space, f_n: $X \to \mathbb{R}$ be a sequence of Lipschitz continuous functions with a common Lipschitz constant $M > 0$. Then the sequence (f_n) converges uniformly to a function f_0 if and only if their graphs converge to the graph of f_0 in Hausdorff metric.*

Moreover, $f_0 = \lim\limits_{n\to\infty} f_n$ is a Lipschitz function with the Lipschitz constant M.

Proof. Let us notice that if $f_n \rightrightarrows f_0$ then f_0 is a Lipschitz function with constant M.

$$|f_0(x)-f_0(y)| = \lim_{n\to\infty} |f_n(x)-f_n(y)| \le \lim_{n\to\infty} M\cdot d(x,y) = M\cdot d(x,y).$$

We will prove that

$$d_H(graph\ f_0,\ graph\ f_n) \le \|f_n - f_0\| \le (M+1)\cdot d_H(graph\ f_0,\ graph\ f_n).$$

First we shall show the first of the inequalities:

$$d_H(graph\ f_0,\ graph\ f_n) \le \|f_n - f_0\|.$$

$$d_H(graph\ f_0,\ graph\ f_n) = \max\{\tilde{d}(graph\ f_0,\ graph\ f_n),\ \tilde{d}(graph\ f_n,\ graph\ f_0)\}$$

As the inequality is symmetric with respect to f_0 and f_n, we may assume that $\tilde{d}(graph\ f_0,\ graph\ f_n) \ge \tilde{d}(graph\ f_n,\ graph\ f_0)\}$ and then

$$\begin{aligned} d_H(graph\ f_0,\ graph\ f_n) &= \tilde{d}(graph\ f_0,\ graph\ f_n) = \\ &= \max\{x\in X : \tilde{d}((x,f_0(x)),\ graph\ f_n)\} \le \\ &\le \max\{x\in X : d((x,f_0(x)),(x,f_n(x))\} = \\ &= \max\{x\in X : |f_0(x)-f_n(x)|\} = \|f_0 - f_n\| \end{aligned}$$

Now we shall show that

$$\|f_n - f_0\| \le (M+1)\cdot d_H(graph\ f_0,\ graph\ f_n)$$

Fix $x\in X$ and let $y\in X$ such that

$$\begin{aligned} d_H(graph\ f_0,\ graph\ f_n) &\ge \tilde{d}((x,f_0(x)),(y,f_n(y))) = \\ &= d(x,y) + |f_0(x)-f_n(y)| \ge \tilde{d}((x,f_0(x)),\ graph\ f_n) \end{aligned}$$

Consequently

$$\begin{aligned} |f_n(x)-f_0(x)| &\le |f_n(x)-f_n(y)| + |f_n(y)-f_0(x)| \le \\ &\le M\cdot d(x,y) + d_H(graph\ hf_0,\ graph\ f_n) \le \\ &\le M\cdot d_H(graph\ f_0,\ graph\ f_n) + d_H(graph\ f_0,\ graph\ f_n) = \\ &= (M+1)\cdot d_H(graph\ f_0,\ graph\ f_n) \end{aligned}$$

and taking the limits we conclude the proof.

□

3. Proof of the main result

Let us start with some technical results on extending Lipschitz functions

Lemma 6. *Let* $F:(0,1)\times\mathbb{R}^n \to \mathbb{R}$ *be a bounded definable map such that for any* $y\in(0,1)$ *the restriction* $F_y:\mathbb{R}^n \ni x \longrightarrow F(y,x)\in\mathbb{R}$ *satisfies the Lipschitz condition with*

a constant independent of y. Then for any $a \in \mathbb{R}^n$ *the limit* $\lim_{(y,x)\to(0,a)} F(y,x)$ *exists and defines a definable extension of F to a function* $\tilde{F}:[0,1)\times\mathbb{R}^n \to \mathbb{R}$.

Proof. For any $a \in \mathbb{R}^n$ the function $(0,1) \ni y \longrightarrow F(y,a)$ is definable, so there exists the limit $\tilde{F}(0,a) := \lim_{y\to 0} F(y,a)$. Now, $\left|F(y,x)-\tilde{F}(0,a)\right| \le \left|F(y,x)-F(y,a)\right| + \left|F(y,a)-\tilde{F}(0,a)\right| \le L\left|x-a\right| + \left|F(y,a)-\tilde{F}(0,a)\right|$, hence the limit in question exists. Since, the graph of $\tilde{F}$ is the closure of *graph* (*F*), the function $\tilde{F}$ is definable.

□

Lemma 7 (Banach–McShane–Whitney extension theorem, [6]). *Let* $f: S \to \mathbb{R}$ *be L-lipschitz function on the subset S in a metric space X. Then the formula*

$$F(x) := \sup\{f(x') - L\cdot d(x,x') : x' \in S\}$$

For $x \in X$ defines the extension of the function *f* such that $F: X \to \mathbb{R}$ is L-lipschitz.

Now, we are in a position to give the proof of our main result

Proof of Theorem 1. Induction with respect to *n*. For *n* = 0 it is obvious. Let A_1 be the projection of *A* onto $\mathbb{R} \times \mathbb{R}^{n-1}$, by the inductive hypothesis the limit $A_0 := \lim_{y\to 0} \overline{(A_1)_y}$ exists and is definable. Without loss of generality we may assume that $\dim(A_1)_y$ and $\dim(A_y)$ is constant for $y \in (0,c)$, so all cells A_y are of the same type (a graph or a band).

If all fibres are graphs, there exists a definable function $F: A_1 \to \mathbb{R}$ such that $A = graph\,(F)$, for any $y \in (0,c)$, the function F_y is Lipschitz with a constant *L* independent of *y*. Using lemmas 6 and 7 we can extend this function to a definable function $\tilde{F}:[0,c)\times\mathbb{R}^n \to \mathbb{R}$, set $\tilde{F}_0(x) := \tilde{F}(0,x)$, for $x \in \mathbb{R}^n$.

Let $C := graph(\tilde{F}_0 \,|\, A_0)$, we shall show $\lim_{y\leftrightarrow 0} A_y = C$. Let $y_\nu \in (0,c)$ be a sequence such that $y_\nu \longrightarrow 0$, let $x_\nu \in A_{y_\nu}$, $x_\nu \longrightarrow x_0$ be a convergent sequence, we shall prove that $x \in C$. Let $x_\nu = (x'_\nu, x_n^\nu)$ and $x_0 = (x'_0, x_n^0)$. We have $(y_\nu, x'_\nu) \in (A_1)_{y_\nu}$, so $x'_0 \in A_0$. By the definition $\tilde{F}_0(x'_0) = \lim_{\nu\to\infty} F(y_\nu, x'_\nu) = \lim_{\nu\to\infty} x_n^\nu = x_n^0$, hence $x \in C$.

Now, let $x \in C$ and $y_\nu \in (0,c)$ be a sequence such that $y_\nu \longrightarrow 0$. Since $x'_0 \in A_0$, $x_n^0 = \tilde{F}_0(x'_0)$ there is $x'_\nu \in (A_1)_{y_\nu}$ such that $x'_\nu \longrightarrow x'_0$. Put $x_n^\nu = F(y_n, x_{\nu'})$, we get $x_\nu \in A_{y_\nu}$ and $x_n^\nu = F(y_\nu, x'_\nu) \longrightarrow \tilde{F}(0,x'_0) = \tilde{F}_0(x'_0) = x_n^0$. Consequently we have $x_\nu \longrightarrow x_0$ which proves $\lim_{y\to 0} A_y = C$.

If is a band for $y \in (0,c)$ proceeding in a similar way, we have $A = (G,H)$, where $G, H: A_1 \longrightarrow \mathbb{R}$ and define $\tilde{G}_0$, $\tilde{H}_0$. We shall show that

$$C : \{x \in \mathbb{R}^n : x' \in A_0, \tilde{G}_0(x') \leq x_n \leq \tilde{H}_0(x')\}$$

is the Hausdorff limit of A_y as $y \longrightarrow 0,\ y \in (0,c)$.

Let $y_\nu \in (0,c)$ be a sequence such that $y_\nu \longrightarrow 0$, let $x_\nu \in A_{y_\nu}, x_\nu \longrightarrow x_0$. Let $x_\nu = (x'_\nu, x_n^\nu)$ and $x_0 = (x'_0, x_n^0)$. We have $(y_\nu, x'_\nu) \in (A_1)_{y_\nu}$, so $x'_0 \in A_0$. By the definition $\tilde{G}_0(x'_0) = \lim_{\nu\to\infty} G(y_\nu, x'_\nu)$, $\tilde{G}_0(x'_0) = \lim_{\nu\to\infty} G(y_\nu, x'_\nu)$ so

$$\tilde{G}_0(x'_0) \leq x_n^0 \leq \tilde{H}_0(x'_0)$$

and hence $x_0 \in C$.

Now, fix $x_0 \in C$ and $y_\nu \in (0,c)$ such that $y_\nu \longrightarrow 0$. We have $x'_0 \in A_0$ and $\tilde{G}_0(x'_0) \leq x_n^0 \leq \tilde{H}_0(x'_0)$. There exists $x'_\nu \in (A_1)_{y_\nu}$ such that $x'_\nu \longrightarrow x'_0$.

If $\tilde{G}_0(x') = \tilde{H}_0(x')$ put $x_n^\nu = \frac{1}{2}(G(y_\nu, x'_\nu), H(y_\nu, x'_\nu))$. If $\tilde{G}_0(x') < \tilde{H}_0(x')$ put

$$x_n^\nu = \frac{x_n^0 - \tilde{G}_0(x'_0)}{(\tilde{H}_0(x'_0) - \tilde{G}_0(x'_0))}(H(y_\nu, x'_\nu) - G(y_\nu, x_\nu)) + G(y_\nu, x_\nu).$$

Then $x_\nu \in A_{y_\nu}$ and $x_\nu \longrightarrow x_0$.

□

References

[1] Bröcker L., *Families of semialgebraic sets and limits*, [in:] *Real algebraic* (Rennes, 1991), volume 1524 of Lecture Notes in Math., 145-162, Springer 1992.

[2] van den Dries L., *Tame topology and o-minimal structures*, Mathematical Society Lectures Notes, **248**, Cambridge University Press, London 1998.

[3] Kocel-Cynk B., Pawłucki W., Valette A., *A short geometric proof that Hausdorff limits are definable in any o-minimal structure*, Adv. Geom., 14, no. 1, 2014, 49-58.

[4] Lion J.-M., Speissegger P., *A geometric proof of the definability of Hausdorff limits*, Selecta Math. (N.S.), 10, no. 3, 2004, 377-390.

[5] Łojasiewicz S., *Ensembles semi-analytiques*, IHES, 1965.

[6] McShane E.J., *Extension of range of functions*, Bull. Amer. Math. Soc., 40, 1934, 837-842.

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

PIOTR KOT*

PEAK SET ON THE UNIT DISC

ZBIÓR SZCZYTOWY DLA DYSKU JEDNOSTKOWEGO

Abstract

Abstract: We show that any compact subset K in the boundary of the unit disc D with a zero measure is a peak set for $A(\mathsf{D})$.

Keywords:

Streszczenie

Pokażemy, że dowolny podzbiór zwarty K miary zero w brzegu dysku jednostkowego jest zbiorem szczytowym dla $A(\mathsf{D})$.

Słowa kluczowe:

DOI: 10.4467/2353737XCT.16.141.5752

* Piotr Kot (pkot@pk.edu.pl), Institute of Mathematics, Faculty of Physic, Mathematics and Computer Sciences, Cracow University of Technology.

1. Schwarz integral

The goal of this paper is to consider some properties of one-dimensional holomorphic functions in the unit disc. We focus our attention on such boundary properties of these functions which imply their uniqueness. In this aspect Luzin-Privalov theorem [4–6] seems to be crucial. This theorem refers to a meromorphic function $f(z)$ of the complex variable z in a simply-connected domain D with rectifiable boundary Γ. If $f(z)$ takes angular boundary values zero on a set $E \subset \Gamma$ of positive Lebesgue measure on Γ, then $f(z) = 0$ in D. There is no function meromorphic in D that has infinite angular boundary values on a set $E \subset \Gamma$ of positive measure.

We are going to construct some examples of a holomorphic non-constant function f for a given E set of measure zero with $f = 1$ on E.

It will turn out that this E set is a peak set for a proper algebra of holomorphic functions.

We say that a compact set K is a peak set for $A(\mathbb{D})$ if there exists $f \in A(\mathbb{D})$ such that $|f| < 1$ on $\overline{\mathbb{D}} \setminus K$ and $f = 1$ on K. Stensönes Henriksen has proved [2] that every strictly pseudoconvex domain with C^∞ boundary in C^d has a peak set with a Hausdorff dimension $2d - 1$.

In this paper we give an alternative, even stronger construction for the unit disc. In the context of the Luzin-Privalov theorem we give the optimal construction for algebra $A(\mathbb{D})$.

Main tool in our construction is the Schwarz kernel.

Let us consider a natural measure σ on boundary of the unit circle $\partial\mathbb{D}$. For a given u which satisfies a Hölder condition we use Schwarz integral (see [7, 8]):

$$Su(z) := \frac{1}{2\pi i} \int_{\partial\mathbb{D}} u(t) \frac{t+z}{t-z} \frac{dt}{t}.$$

We can easily observe that $Su \in O(\mathbb{D})$.

Then the Schwarz integral formula Su defining an analytic function, the boundary values of whose real part coincide with u. Additionally, the real part of Su is a continuous harmonic function on $\overline{\mathbb{D}}$ (see [1, The Basic Lemma].

There exists a harmonic function v on $\mathbb{D}$ so that $Su = u + iv$.

However when applying the above integral formula, a very important and more difficult problem arises concerning the existence and the expression of the boundary values of the imaginary part v and of the complete function Su by the given boundary values of the real part u. Still, in some cases we have complete information about v.

If a given function u satisfies a Hölder condition, then the corresponding values of imaginary part v on $\partial\mathbb{D}$ are expressed by the Hilbert formula (see [3, 1, pp. 45-49]):

$$v(\phi) = -\frac{1}{2\pi} \int_0^{2\pi} u(t) \cot\left(\frac{t-\varphi}{2}\right) dt.$$

The above formula is a singular integral and exists in the Cauchy principal-value sense.

Moreover, if u satisfies a Hölder condition then the values of v exist on all $\phi \in \partial\mathbb{D}$ and satisfy the same Hölder condition as u. Now we can recover Su using v in the following way:

$$Su(z) := \frac{1}{2\pi}\int_{\partial \mathbb{D}} v(t)\frac{t+z}{t-z}\frac{dt}{t} + c_1.$$

But now the imaginary part of Su is continuous on $\bar{\mathbb{D}}$, so $Su \in A(\mathbb{D})$ if u satisfies a Hölder condition.

2. Peak sets

Lemma 1. *Assume that K, D are distinct compact sets in $\partial\mathbb{D}$. Then there exists a function $u \in C^\infty(\partial\mathbb{D})$ so that $u = 0$ on D, $u = 1$ on K and $0 \le u \le 1$ on $\partial\mathbb{D}$.*

Proof. There exist open arcs $I_i : \{e^{2\pi i t} : a_i < t < b_i\}$ such that $K \subset \bigcup_{i=1}^{n} I_i$ and $\bar{I}_i \cap D = \varnothing$. In fact we can assume that $\bar{I}_i \cap \bar{I}_j = \varnothing$ for $i \neq j$. Now there exist functions $u_i : \partial\mathbb{D} \to [0,1] \in C^\infty(\partial\mathbb{D})$ so that $u_i = 1$ on I_i, and $u_i = 0$ on D but with distinct supports. It is enough to define $u = \sum_{k=1}^{n} u_k$.

Theorem 2. *Let K be a compact subset of $\partial\mathbb{D}$ measure zero ($\sigma(K) = 0$). There exists a function $f \in A(\mathbb{D})$ such that $|f| < 1$ on $\bar{\mathbb{D}} \setminus K$ and $f = 1$ on K.*

Proof. Let us choose $\varepsilon > 0$ and define

$$D_\varepsilon := \{z \in \partial\mathbb{D} : \inf_{w \in K} |z - w| \ge \varepsilon\}$$

There exists $u_\varepsilon \in C^\infty(\partial\mathbb{D})$ such that $0 \le u_\varepsilon \le 1$, $u_\varepsilon(z) = 0$ if $z \in D_\varepsilon$ and $u_\varepsilon(z) = 1$ if $z \in K$. In particular $Su_\varepsilon \in A(\mathbb{D})$ and $0 \le \Re Su_\varepsilon \le 1$.

Let us choose $z \in \bar{\mathbb{D}} \setminus K$ and define $\delta(z,\varepsilon) := \inf_{w \in \partial\mathbb{D}\setminus D_\varepsilon} |z - w|$. We can estimate

$$|Su_\varepsilon(z)| \le \left|\frac{1}{2\pi}\int_{\partial\mathbb{D}\setminus D_\varepsilon} \frac{t+z}{t-z}\frac{dt}{t}\right| \le \frac{\sigma(\partial\mathbb{D}\setminus D_\varepsilon)}{2\pi}\max_{t \in U(\varepsilon)}\left|\frac{t+z}{t-z}\right| \le \frac{\sigma(\partial\mathbb{D}\setminus D_\varepsilon)}{\delta(z,\varepsilon)}.$$

Let us consider the following compact set:

$$T_n : \{z \in \bar{\mathbb{D}} : \inf_{w \in K} |z - w| \ge 2^{-n} + 2^{-2n}\}$$

There exists $\varepsilon_n \in (0, 2^{-2n})$ such that $\sigma(\partial\mathbb{D} \setminus D_{\varepsilon_n}) < 2^{-2n}$. Now let $g_n := Su_{\varepsilon_n} \in A(\mathbb{D})$.

Obviously $\Re g_n = 1$ on K and $0 \le \Re g_n \le 1$.

Moreover if $z \in T_n$ then

$$|g_n(z)| \le \frac{\sigma(\partial\mathbb{D}\setminus D_{\varepsilon_n})}{\delta(z,\varepsilon_n)} \le \frac{2^{-2n}}{2^{-n} + 2^{-2n} - 2^{-2n}} = 2^{-n}.$$

Now we are able to define $g := 1 + \sum_{n \in N} g_n$.

Since $\bigcup_{n\in N} T_n = \overline{\mathsf{D}} \setminus K$ we can observe that $g \in O(\mathsf{D}) \cap C(\overline{\mathsf{D}} \setminus K)$. As $0 \le \Re g_n \le 1$ and $\Re g_n = 1$ on K we have $\lim_{z\to w} \Re g_n(z) = \infty$ for $w \in K$.

Now we choose $f := \exp\left(-\frac{1}{g}\right)$. Obviously $f \in O(\mathsf{D}) \cap C(\overline{\mathsf{D}} \setminus K)$.

Since $\Re \frac{1}{g} = \frac{\Re \overline{g}}{|g|^2} = \frac{\Re g}{|g|^2} > 0$ on $\bar{\Omega} \setminus K$ we may easily observe that $0 < |f| < 1$ on $\bar{\Omega} \setminus K$.

Additionally due to $\lim_{z\to w} \frac{1}{|g(z)|} = 0$ for $w \in K$ we have $f = 1$ on K and $f \in A(\Omega)$.

Example 3. *There exists $K \subset \partial\mathsf{D}$, a compact set with Hausdorff dimension equal one which is also a peak set for $A(\mathsf{D})$.*

Let us consider a sequence of closed distinct intervals $I_n := [2^{-2n-1}, 2^{-2n}]$. There exists Cantor set $C_n \subset I_n$ with Hausdorf dimension equal $\frac{n}{n+1}$. Now we define a compact set

$$K := \{1\} \cup \bigcup_{n=1}^{\infty} \{e^{2\pi i t} : t \in C_n\}$$

in $\partial\mathsf{D}$ with Hausdorff dimension one and due to Theorem 2 we conclude that K is a peak set for $A(\mathsf{D})$.

References

[1] Gakhov F.D., *Boundary value problems*, Pergamon, 1966 (Translated from Russian).

[2] Stensönes Henriksen: *A peak sets of Hausdorff dimension* $2n - 1$ *for the algebra* $A(D)$ in the *boundary of a domain D with* C^∞ *-boundary in* C^n, Math. Ann., 259, 1982, 271-277.

[3] Hilbert singular integral, B.V. Khvedelidze (originator), Encyclopedia of Mathematics: http://www.encyclopediaofmath.org/index.php?title=Hilbert_singular_integral&oldid=11933 [access: 19.12.2016].

[4] Luzin-Privalov theorems. Encyclopedia of Mathematics: http://www.encyclopediaofmath.org/index.php?title=Luzin-Privalov_theorems&oldid=27205 [access: 19.12.2016].

[5] Lusin N.N., Priwaloff I.I., Sur l'unicité et la multiplicité des fonctions analytiques, Ann. Sci. Ecole Norm. Sup. (3), 42, 1925, pp. 143-191.

[6] Priwalow I.I., *Randeigenschaften analytischer Funktionen*, Deutsch. Verlag Wissenschaft, 1956 (Translated from Russian).

[7] Schwarz integral. Encyclopedia of Mathematics: http://www.encyclopediaofmath.org/index.php?title=Schwarz_integral&oldid=31192 [access: 19.12.2016].

[8] Schwarz H.A., Gesamm. math. Abhandl., 2, Springer, 1890.

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

GRAŻYNA KRECH*, RENATA MALEJKI*

ON THE BIVARIATE BASKAKOV-DURRMEYER TYPE OPERATORS

O OPERATORACH DWÓCH ZMIENNYCH TYPU BASKAKOWA-DURRMEYERA

A b s t r a c t

In this paper we introduce some linear positive operators of the Baskakov-Durrmeyer type in the space of continuous functions of two variables. The theorems on convergence and the degree of approximation are established.

Keywords: Baskakov-Durrmeyer type operators, *linear operators*, *approximation order*

S t r e s z c z e n i e

W artykule definiuje się dodatnie operatory liniowe typu Baskakowa-Durrmeyera w przestrzeni ciągłych funkcji dwóch zmiennych. Formułuje się i dowodzi twierdzenia dotyczące zbieżności oraz rzędu zbieżności.

Słowa kluczowe: operatory typu Baskakowa-Durrmeyera, *operatory liniowe*, *rząd aproksymacji*

DOI: 10.4467/2353737XCT.16.142.5753

* Grażyna Krech (gkrech@up.krakow.pl), Renata Malejki, Institute of Mathematics, Pedagogical University of Cracow.

1. Introduction

In recent years, several researchers have studied various modifications of the Baskakov--Durrmeyer operators. The approximation properties of these operators in many different spaces were considered, for example, in [4, 8, 10, 11, 18, 19].

A large amount of literature is available on approximation of function of one variable, but the corresponding problem for bivariate functions has received less attention. The bivariate Bernstein operator was first introduced by Dhingas [3] and it was also considered by Lorentz [9] and Stancu [14]. Recently, some positive linear operators for function of two variables and their approximation properties were investigated in a series of research articles (e.g. [2, 5, 6, 7, 12, 13, 15, 17, 20, 21].

In this paper, we will introduce the Baskakov-Durrmeyer type operators in the space of continuous functions of two variables. This is an extension of the paper [10] for a bivariate case.

Let $\mathbb{R}_0^+ = [0,\infty)$ and $\mathbb{R}_+^2 = \mathbb{R}_0^+ \times \mathbb{R}_0^+$. We denote by $C(\mathbb{R}_+^2)$ the space of all real-valued functions continuous on $\mathbb{R}_+^2$ and by $C_B(\mathbb{R}_+^2)$ – the space of functions continuous and bounded on $\mathbb{R}_+^2$. The norm on $C_B(\mathbb{R}_+^2)$ is defined by

$$\|f\|_{C_B(\mathbb{R}_+^2)} = \sup_{(x,y)\in\mathbb{R}_+^2} |f(x,y)|.$$

Let

$$W_{n,k}^a(x) = e^{-\frac{ax}{1+x}} \sum_{i=0}^{k} \binom{k}{i} (n)_i \, a^{k-i} \frac{x^k}{k!(1+x)^{n+k}},$$

where $a \in \mathbb{R}_0^+$, $(n)_0 = 1$, $(n)_i = n(n+1)\ldots(n+i-1)$, $i \geq 1$.

We consider the class of operators $M_{n,m}^{\alpha,\beta,a,b}$ given by the formula

$$M_{n,m}^{\alpha,\beta,a,b}(f;x,y) = mn \sum_{k,l=0}^{\infty} W_{n,k}^a(x) W_{m,l}^b(y) \frac{1}{\Gamma(\alpha+k+1)} \frac{1}{\Gamma(\beta+l+1)}$$
$$\times \int_0^\infty \int_0^\infty e^{-ns} (ns)^{\alpha+k} e^{-mz} (mz)^{\beta+l} f(s,z)\,ds\,dz$$

for $(x,y) \in \mathbb{R}_+^2$, where $m,n \in \mathbb{N}$, $a,b \in \mathbb{R}_0^+$, $\alpha,\beta > -1$. It is clear that the operator $M_{n,m}^{\alpha,\beta,a,b}$ is linear and positive on $\mathbb{R}_+^2$. In this paper we study some approximation properties of $M_{n,m}^{\alpha,\beta,a,b}$ in the space of continuous functions of two variables on a compact set. We find the order of this approximation using full and partial modulus of continuity.

Observe that if $f(s,z) = f_1(s) f_2(z)$, then

$$M_{n,m}^{\alpha,\beta,a,b}(f;x,y) = M_n^{\alpha,a}(f_1;x) M_m^{\beta,b}(f_2;y), \tag{1.1}$$

where

$$M_n^{\alpha,a}(f_1;x)=n\sum_{k=0}^{\infty}W_{n,k}^{a}(x)\frac{1}{\Gamma(\alpha+k+1)}\int_0^{\infty}e^{-ns}(ns)^{\alpha+k}f_1(s)ds.$$

Some properties of the operator $M_n^{\alpha,a}$, in particular, an estimation of the rate of convergence, were studied in [10].

Let $(x,y)\in\mathbb{R}_+^2$ and

$$e^{i,j}(s,z)=s^i z^j,\ \phi_{x,y}^{i,j}(s,z)=(s-x)^i(z-y)^j,\ i,j=0,1,2,4,\ (s,z)\in\mathbb{R}_+^2.$$

Now, we give some lemmas which will be useful in the future proofs of the main results. The following lemmas are simple consequences of the above definitions and the results obtained in [10, Lemma 2.2, Lemma 2.3].

Lemma 1. *Let* $m,n\in\mathbb{N},\ a,b\in\mathbb{R}_0^+,\ \alpha,\beta>-1$. For $(x,y)\in\mathbb{R}_+^2$ *we get*

$$M_{n,m}^{\alpha,\beta,a,b}(e^{0,0};x,y)=1, \tag{1.2}$$

$$M_{n,m}^{\alpha,\beta,a,b}(e^{1,0};x,y)=\frac{\alpha+1}{n}+x+\frac{ax}{n(1+x)}, \tag{1.3}$$

$$M_{n,m}^{\alpha,\beta,a,b}(e^{0,1};x,y)=\frac{\beta+1}{m}+y+\frac{by}{m(1+y)}, \tag{1.4}$$

$$\begin{aligned}M_{n,m}^{\alpha,\beta,a,b}(e^{2,0};x,y)&=\frac{(\alpha+1)(\alpha+2)}{n^2}+\frac{2(\alpha+2)x+x^2}{n}+x^2+\frac{a^2x^2}{n^2(1+x)^2}\\&\quad+\frac{2ax^2}{n(1+x)}+\frac{2(\alpha+2)ax}{n^2(1+x)},\end{aligned} \tag{1.5}$$

$$\begin{aligned}M_{n,m}^{\alpha,\beta,a,b}(e^{0,2};x,y)&=\frac{(\beta+1)(\beta+2)}{m^2}+\frac{2(\beta+2)y+y^2}{m}+y^2+\frac{b^2y^2}{m^2(1+y)^2}\\&\quad+\frac{2by^2}{m(1+y)}+\frac{2(\beta+2)by}{m^2(1+y)}.\end{aligned} \tag{1.6}$$

Lemma 2. *Let* $m,n\in\mathbb{N},\ a,b\in\mathbb{R}_0^+,\ \alpha,\beta>-1$. For $(x,y)\in\mathbb{R}_+^2$ *we get*

$$M_{n,m}^{\alpha,\beta,a,b}(\phi_{x,y}^{1,0};x,y)=\frac{\alpha+1}{n}+\frac{ax}{n(1+x)},$$

$$M_{n,m}^{\alpha,\beta,a,b}(\phi_{x,y}^{0,1};x,y)=\frac{\beta+1}{m}+\frac{by}{m(1+y)},$$

$$M_{n,m}^{\alpha,\beta,a,b}(\phi_{x,y}^{1,1};x,y)=\frac{(\alpha+1)(\beta+1)}{nm}+\frac{(\alpha+1)by}{nm(1+y)}+\frac{(\beta+1)ax}{nm(1+x)}+\frac{abxy}{nm(1+x)(1+y)},$$

$$M_{n,m}^{\alpha,\beta,a,b}(\phi_{x,y}^{2,0};x,y)=\frac{(\alpha+1)(\alpha+2)}{n^2}+\frac{2x+x^2}{n}+\frac{a^2x^2}{n^2(1+x)^2}+\frac{2(\alpha+2)ax}{n^2(1+x)},$$

$$M_{n,m}^{\alpha,\beta,a,b}(\phi_{x,y}^{0,2};x,y)=\frac{(\beta+1)(\beta+2)}{m^2}+\frac{2y+y^2}{m}+\frac{b^2y^2}{m^2(1+y)^2}+\frac{2(\beta+2)by}{m^2(1+y)}.$$

Theorem 1. *For each* $f\in C_B(\mathbb{R}_+^2)$, *we have*

$$\left\|M_{n,m}^{\alpha,\beta,a,b}(f)\right\|_{C_B(\mathbb{R}_+^2)}\le\|f\|_{C_B(\mathbb{R}_+^2)}$$

for all $n,m\in\mathbb{N}$.

Proof. Using the definition $M_{n,m}^{\alpha,\beta,a,b}$, we obtain

$$\begin{aligned}\left|M_{n,m}^{\alpha,\beta,a,b}(f;x,y)\right|&\le mn\sum_{k,l=0}^{\infty}W_{n,k}^{a}(x)W_{m,l}^{b}(y)\frac{1}{\Gamma(\alpha+k+1)}\frac{1}{\Gamma(\beta+l+1)}\\&\times\int_0^\infty\int_0^\infty e^{-ns}(ns)^{\alpha+k}e^{-mz}(mz)^{\beta+l}\left|f(s,z)\right|dsdz\\&\le\sup_{(s,z)\in\mathbb{R}_+^2}\left|f(s,z)\right|mn\sum_{k,l=0}^{\infty}W_{n,k}^{a}(x)W_{m,l}^{b}(y)\frac{1}{\Gamma(\alpha+k+1)}\frac{1}{\Gamma(\beta+l+1)}\\&\times\int_0^\infty\int_0^\infty e^{-ns}(ns)^{\alpha+k}e^{-mz}(mz)^{\beta+l}dsdz\\&=\sup_{(s,z)\in\mathbb{R}_+^2}\left|f(s,z)\right|M_{n,m}^{\alpha,\beta,a,b}(e^{0,0};x,y)=\sup_{(s,z)\in\mathbb{R}_+^2}\left|f(s,z)\right|=\|f\|,\end{aligned}$$

which gives the result. □

Theorem 2 [22]. *Let* I_1 *and* I_2 *be compact intervals of the real line. Let* $n,m\in\mathbb{N}$ *and* $T_{n,m}:C(I_1\times I_2)\to C(I_1\times I_2)$ *be linear positive operators. If*

$$\lim_{n,m\to\infty}T_{n,m}(e^{i,j})=e^{i,j},\quad(i,j)\in\{(0,0),(1,0),(0,1)\}$$

and

$$\lim_{n,m\to\infty}T_{n,m}(e^{2,0}+e^{0,2})=e^{2,0}+e^{0,2},$$

uniformly on $I_1\times I_2$, *then the sequence* $(T_{n,m}f)$ *converges to f uniformly on* $I_1\times I_2$, *for any* $f\in C(I_1\times I_2)$.

Let $A,B>0$. Throughout the rest of this paper we will denote $\mathbb{R}_{AB}^2=[0,A]\times[0,B]$.

Theorem 3. *Let* $(x,y)\in\mathbb{R}_{AB}^2$ *are fixed. If* $f\in C(\mathbb{R}_{AB}^2)$, *then*

$$\lim_{n,m\to\infty}M_{n,m}^{\alpha,\beta,a,b}(f;x,y)=f(x,y).$$

Moreover, this convergence is uniform on $\mathbb{R}^2_{AB}$.

Proof. Using (1.2)–(1.6), we have

$$\lim_{n,m\to\infty} M_{n,m}^{\alpha,\beta,a,b}(e^{i,j};x,y) = e^{i,j}(x,y), \quad (i,j)\in\{(0,0),(1,0),(0,1)\}$$

and

$$\lim_{n,m\to\infty} M_{n,m}^{\alpha,\beta,a,b}(e^{2,0}+e^{0,2};x,y) = e^{2,0}(x,y)+e^{0,2}(x,y)$$

uniformly on $\mathbb{R}^2_{AB}$. Applying Theorem 2, the proof of the theorem is completed. □

2. Local approximation results

In this section we will investigate the degree of approximation for functions of two variables by operators $M_{n,m}^{\alpha,\beta,a,b}$ in terms of the modulus of continuity on a compact set.

Let $f\in C(\mathbb{R}^2_{AB})$ and $\delta>0$. The full continuity modulus of the function f is defined as (see [1], [16])

$$\omega(f;\delta) = \sup_{\substack{(s,z),(x,y)\in\mathbb{R}^2_{AB}\\ (s-x)^2+(z-y)^2\le\delta^2}} \left|f(s,z)-f(x,y)\right|$$

and its partial continuity moduli are given by

$$\omega^{(1)}(f;\delta) = \sup_{\substack{0\le z\le B\\ |s-x|\le\delta}} \left|f(s,z)-f(x,z)\right|,$$

$$\omega^{(2)}(f;\delta) = \sup_{\substack{0\le s\le A\\ |z-y|\le\delta}} \left|f(s,z)-f(s,y)\right|.$$

It is known that $\lim_{\delta\to 0}\omega(f;\delta)=0$, $\omega(f;\delta_1)\le\omega(f;\delta_2)$ for $0<\delta_1\le\delta_2$ and for any $\lambda>0$, $\omega(f;\lambda\delta)\le(1+\lambda)\omega(f;\delta)$. The same properties are satisfied by partial continuity moduli. The details of the modulus of continuity for the bivariate case can be found in [1].

Theorem 4. *Let* $f\in C(\mathbb{R}^2_{AB})$. *For* $(x,y)\in\mathbb{R}^2_{AB}$, *we have*

$$\left|M_{n,m}^{\alpha,\beta,a,b}(f;x,y)-f(x,y)\right| \le 2\omega(f;\delta),$$

where

$$\delta = \left(\frac{(\alpha+1)(\alpha+2)}{n^2}+\frac{2x+x^2}{n}+\frac{a^2x^2}{n^2(1+x)^2}+\frac{2(\alpha+2)ax}{n^2(1+x)}\right.$$
$$\left.+\frac{(\beta+1)(\beta+2)}{m^2}+\frac{2y+y^2}{m}+\frac{b^2y^2}{m^2(1+y)^2}+\frac{2(\beta+2)by}{m^2(1+y)}\right)^{1/2}.$$

Proof. Let $\delta > 0$. If $\sqrt{(s-x)^2+(z-y)^2} \le \delta$, then $|f(s,z)-f(x,y)| \le \omega(f;\delta)$. If $\sqrt{(s-x)^2+(z-y)^2} > \delta$, then

$$\frac{(s-x)^2+(z-y)^2}{\delta^2} > \frac{\sqrt{(s-x)^2+(z-y)^2}}{\delta} > 1.$$

Therefore, we obtain

$$\begin{aligned}|f(s,z)-f(x,y)| &\le \omega\left(f;\sqrt{(s-x)^2+(z-y)^2}\right)\\ &\le \left(1+\frac{\sqrt{(s-x)^2+(z-y)^2}}{\delta}\right)\omega(f;\delta) \le \left(1+\frac{(s-x)^2+(z-y)^2}{\delta^2}\right)\omega(f;\delta).\end{aligned}$$

The operator $M_{n,m}^{\alpha,\beta,a,b}$ is positive and linear, so

$$\begin{aligned}\left|M_{n,m}^{\alpha,\beta,a,b}(f;x,y)-f(x,y)\right| &\le M_{n,m}^{\alpha,\beta,a,b}\left(|f-f(x,y)|;x,y\right)\\ &\le \omega(f;\delta)\left(M_{n,m}^{\alpha,\beta,a,b}(e^{0,0};x,y)+\frac{1}{\delta^2}M_{n,m}^{\alpha,\beta,a,b}(\phi_{x,y}^{2,0}+\phi_{x,y}^{0,2};x,y)\right).\end{aligned}$$

From Lemma 2 we obtain

$$\begin{aligned}\left|M_{n,m}^{\alpha,\beta,a,b}(f;x,y)-f(x,y)\right| &\le M_{n,m}^{\alpha,\beta,a,b}\left(|f-f(x,y)|;x,y\right)\\ &\le \omega(f;\delta)\left\{1+\frac{1}{\delta^2}\left(\frac{(\alpha+1)(\alpha+2)}{n^2}+\frac{2x+x^2}{n}+\frac{a^2x^2}{n^2(1+x)^2}+\frac{2(\alpha+2)ax}{n^2(1+x)}\right.\right.\\ &\left.\left.+\frac{(\beta+1)(\beta+2)}{m^2}+\frac{2y+y^2}{m}+\frac{b^2y^2}{m^2(1+y)^2}+\frac{2(\beta+2)by}{m^2(1+y)}\right)\right\},\end{aligned}$$

which ends the proof. □

Theorem 5. *If* $f \in C(\mathbb{R}_{AB}^2)$, *then for all* $(x,y) \in \mathbb{R}_{AB}^2$, *we have*

$$\begin{aligned}&\left|M_{n,m}^{\alpha,\beta,a,b}(f;x,y)-f(x,y)\right|\\ &\le \left(1+\frac{(\alpha+1)(\alpha+2)}{n}+2x+x^2+\frac{a^2x^2}{n(1+x)^2}+\frac{2(\alpha+2)ax}{n(1+x)}\right)\omega^{(1)}\left(f;\frac{1}{\sqrt{n}}\right)\\ &+\left(1+\frac{(\beta+1)(\beta+2)}{m}+2y+y^2+\frac{b^2y^2}{m(1+y)^2}+\frac{2(\beta+2)by}{m(1+y)}\right)\omega^{(2)}\left(f;\frac{1}{\sqrt{m}}\right).\end{aligned}$$

Proof. Let $f \in C(\mathbb{R}^2_{AB})$. Observe that

$$\begin{aligned}\left|M_{n,m}^{\alpha,\beta,a,b}(f;x,y)-f(x,y)\right| &\le mn\sum_{k,l=0}^{\infty} W_{n,k}^{a}(x)W_{m,l}^{b}(y)\frac{1}{\Gamma(\alpha+k+1)}\frac{1}{\Gamma(\beta+l+1)}\\ &\times\int_0^{\infty}\int_0^{\infty} e^{-ns}(ns)^{\alpha+k}e^{-mz}(mz)^{\beta+l}\left|f(s,z)-f(x,z)\right|dsdz\\ &+mn\sum_{k,l=0}^{\infty} W_{n,k}^{a}(x)W_{m,l}^{b}(y)\frac{1}{\Gamma(\alpha+k+1)}\frac{1}{\Gamma(\beta+l+1)}\\ &\times\int_0^{\infty}\int_0^{\infty} e^{-ns}(ns)^{\alpha+k}e^{-mz}(mz)^{\beta+l}\left|f(x,z)-f(x,y)\right|dsdz\\ &= J_1+J_2.\end{aligned}$$

Using the properties of the modulus of continuity and (1.5), we have

$$\begin{aligned}J_1 &= mn\sum_{k,l=0}^{\infty} W_{n,k}^{a}(x)W_{m,l}^{b}(y)\frac{1}{\Gamma(\alpha+k+1)}\frac{1}{\Gamma(\beta+l+1)}\\ &\times\int_0^{\infty}\int_0^{\infty} e^{-ns}(ns)^{\alpha+k}e^{-mz}(mz)^{\beta+l}\left|f(s,z)-f(x,z)\right|dsdz\\ &\le \omega^{(1)}(f;\delta_n)\left\{1+\frac{1}{\delta_n^2}M_{n,m}^{\alpha,\beta,a,b}(\phi_{x,y}^{2,0};x,y)\right\}\\ &\le\left(1+\frac{(\alpha+1)(\alpha+2)}{n}+2x+x^2+\frac{a^2x^2}{n(1+x)^2}+\frac{2(\alpha+2)ax}{n(1+x)}\right)\omega^{(1)}\left(f;\frac{1}{\sqrt{n}}\right),\end{aligned}$$

where $\delta_n=\frac{1}{\sqrt{n}}$. Similarly, we obtain

$$J_2\le\left(1+\frac{(\beta+1)(\beta+2)}{m}+2y+y^2+\frac{b^2y^2}{m(1+y)^2}+\frac{2(\beta+2)by}{m(1+y)}\right)\omega^{(2)}\left(f;\frac{1}{\sqrt{m}}\right).$$

Hence, the proof is completed. □

Now, we consider the mixed modulus of smoothness and the modulus of smoothness (see [16]). Let $\delta_j>0, j=1,2$.

The mixed modulus of smoothness is defined as

$$\omega_{\text{mix}}(f;\delta_1,\delta_2)=\sup_{\substack{|s-x|\le\delta_1,|z-y|\le\delta_2\\(x,y),(s,z)\in\mathbb{R}^2_{AB}}}\left|f(s,z)-f(x,z)-f(s,y)+f(x,y)\right|$$

and the modulus of smoothness of the first and the second order are given by

$$\omega_1(f;\delta_1,\delta_2)=\sup_{\substack{0\le h\le\delta_1,\,0\le k\le\delta_2\\(x,y),(x+h,y+k)\in\mathbb{R}^2_{AB}}}\left|f(x+h,y+k)-f(x,y)\right|,$$

$$\omega_2(f;\delta_1,\delta_2)=\sup_{\substack{0\le h\le\delta_1,\,0\le k\le\delta_2\\(x,y),(x+2h,y+2k)\in\mathbb{R}^2_{AB}}}\left|f(x+2h,y+2k)-2f(x+h,y+k)+f(x,y)\right|,$$

respectively.

Theorem 6. *Let* $f\in C(\mathbb{R}^2_{AB})$ *and*

$$H^{\alpha,\beta,a,b}_{n,m}(f;x,y)=M^{\alpha,\beta,a,b}_{n,m}(f;x,y)-f\left(\frac{\alpha+1}{n}+x+\frac{ax}{n(1+x)},\frac{\beta+1}{m}+y+\frac{by}{m(1+y)}\right)+f(x,y).$$

There exists a positive constant C such that, for all $(x,y)\in\mathbb{R}^2_{AB}$, *we have*

$$\left|H^{\alpha,\beta,a,b}_{n,m}(g;x,y)-g(x,y)\right|\le C\left\{\frac{1}{n}\left\|\frac{\partial^2 g}{\partial u^2}\right\|_{C(\mathbb{R}^2_{AB})}+\frac{1}{m}\left\|\frac{\partial^2 g}{\partial v^2}\right\|_{C(\mathbb{R}^2_{AB})}+\frac{1}{nm}\left\|\frac{\partial^2 g}{\partial u\partial v}\right\|_{C(\mathbb{R}^2_{AB})}\right\}$$

for any function g, such that $g,\ \frac{\partial^i g}{\partial x^i},\ \frac{\partial^i g}{\partial y^i},\ \frac{\partial^2 g}{\partial x\partial y}$ $(i=1,2)$ *belong to* $C(\mathbb{R}^2_{AB})$.

Proof. Let $(x,y)\in\mathbb{R}^2_{AB}$. Observe that

$$g(s,z)-g(x,y)=(s-x)\frac{\partial g(x,y)}{\partial x}+(z-y)\frac{\partial g(x,y)}{\partial y}$$
$$+\int_x^s(s-u)\frac{\partial^2 g(u,y)}{\partial u^2}du+\int_y^z(z-v)\frac{\partial^2 g(x,v)}{\partial v^2}dv+\int_x^s\int_y^z\frac{\partial^2 g(u,v)}{\partial u\partial v}dvdu.$$

We have

$$H^{\alpha,\beta,a,b}_{n,m}(e^{0,0};x,y)=1,\quad H^{\alpha,\beta,a,b}_{n,m}(\phi^{1,0}_{x,y};x,y)=H^{\alpha,\beta,a,b}_{n,m}(\phi^{0,1}_{x,y};x,y)=0.$$

Let

$$\xi^{i,j}_g(s,z)=\left(\int_x^s(s-u)\frac{\partial^2 g(u,y)}{\partial u^2}du\right)^i\left(\int_y^z(z-v)\frac{\partial^2 g(x,v)}{\partial v^2}dv\right)^j,\quad i,j=0,1$$

and

$$\xi_g(s,z)=\int_x^s\int_y^z\frac{\partial^2 g(u,v)}{\partial u\partial v}dudv.$$

Hence

$$H^{\alpha,\beta,a,b}_{n,m}(g;x,y)-g(x,y)=H^{\alpha,\beta,a,b}_{n,m}(\xi^{1,0}_g;x,y)+H^{\alpha,\beta,a,b}_{n,m}(\xi^{0,1}_g;x,y)+H^{\alpha,\beta,a,b}_{n,m}(\xi_g;x,y).$$

Using the definition of $H_{n,m}^{\alpha,\beta,a,b}$, we can write

$$\begin{aligned}\left|H_{n,m}^{\alpha,\beta,a,b}(\xi_g^{1,0};x,y)\right| &= \left|M_{n,m}^{\alpha,\beta,a,b}(\xi_g^{1,0};x,y)\right.\\ &\left.-\int_x^{\frac{\alpha+1}{n}+x+\frac{ax}{n(1+x)}}\left(\frac{\alpha+1}{n}+x+\frac{ax}{n(1+x)}-u\right)\frac{\partial^2 g(u,y)}{\partial u^2}\,du\right|\\ &\le \left|M_{n,m}^{\alpha,\beta,a,b}(\xi_g^{1,0};x,y)\right|\\ &+\left|\int_x^{\frac{\alpha+1}{n}+x+\frac{ax}{n(1+x)}}\left(\frac{\alpha+1}{n}+x+\frac{ax}{n(1+x)}-u\right)\frac{\partial^2 g(u,y)}{\partial u^2}\,du\right|\\ &\le \frac{1}{2}\sup_{(u,v)\in\mathbb{R}^2_{AB}}\left|\frac{\partial^2 g(u,y)}{\partial u^2}\right|M_{n,m}^{\alpha,\beta,a,b}(\phi_{x,y}^{2,0};x,y)\\ &+\frac{1}{2}\sup_{(u,v)\in\mathbb{R}^2_{AB}}\left|\frac{\partial^2 g(u,y)}{\partial u^2}\right|\left(\frac{\alpha+1}{n}+\frac{ax}{n(1+x)}\right)^2\\ &\le C_1\frac{1}{n}\left\|\frac{\partial^2 g}{\partial u^2}\right\|_{C(\mathbb{R}^2_{AB})}\end{aligned}$$

and similarly, we get

$$\left|H_{n,m}^{\alpha,\beta,a,b}(\xi_g^{0,1};x,y)\right| \le C_2\frac{1}{m}\left\|\frac{\partial^2 g}{\partial v^2}\right\|_{C(\mathbb{R}^2_{AB})},$$

$$\left|H_{n,m}^{\alpha,\beta,a,b}(\xi_g;x,y)\right| \le C_3\frac{1}{nm}\left\|\frac{\partial^2 g}{\partial u\partial v}\right\|_{C(\mathbb{R}^2_{AB})},$$

where C_1, C_2, C_3 are positive constants. Hence

$$\left|H_{n,m}^{\alpha,\beta,a,b}(g;x,y)-g(x,y)\right| \le C\left\{\frac{1}{n}\left\|\frac{\partial^2 g}{\partial x^2}\right\|_{C(\mathbb{R}^2_{AB})}+\frac{1}{m}\left\|\frac{\partial^2 g}{\partial v^2}\right\|_{C(\mathbb{R}^2_{AB})}+\frac{1}{nm}\left\|\frac{\partial^2 g}{\partial u\partial v}\right\|_{C(\mathbb{R}^2_{AB})}\right\}$$

for some $C>0$ and the theorem is proved. □

Theorem 7. *If* $f\in C(\mathbb{R}^2_{AB})$, *then*

$$\begin{aligned}\left|M_{n,m}^{\alpha,\beta,a,b}(f;x,y)-f(x,y)\right| \le C\left\{\omega_2\left(f;\sqrt{\frac{1}{n}},\sqrt{\frac{1}{m}}\right)+\omega_{mix}\left(f;\sqrt{\frac{1}{n}},\sqrt{\frac{1}{m}}\right)\right.\\ \left.+\omega_1\left(f;\frac{1}{n}\left(\alpha+1+\frac{ax}{1+x}\right),\frac{1}{m}\left(\beta+1+\frac{by}{1+y}\right)\right)\right\},\end{aligned}$$

where $C>0$, $(x,y)\in\mathbb{R}^2_{AB}$.

Proof. Let $f \in C(\mathbb{R}^2_{AB})$ and $\delta_j > 0, j = 1,2$. We shall use the Steklov function of second order defined by

$$f_{\delta_1\delta_2}(x,y) = \frac{16}{\delta_1^2\delta_2^2}\int_0^{\frac{\delta_2}{2}}\int_0^{\frac{\delta_2}{2}}\int_0^{\frac{\delta_1}{2}}\int_0^{\frac{\delta_1}{2}} 2f(x+s_1+s_2, y+z_1+z_2) - f(x+2(s_1+s_2), y+2(z_1+z_2))ds_1ds_2dz_1dz_2.$$

Observe that

$$\left|f_{\delta_1\delta_2}(x,y) - f(x,y)\right| \le \omega_2(f;\delta_1,\delta_2)$$

and

$$\begin{aligned} f_{\delta_1\delta_2}(x,y) &= \frac{32}{\delta_1^2\delta_2^2}\int_0^{\frac{\delta_2}{2}}\int_0^{\frac{\delta_2}{2}}\int_x^{x+\frac{\delta_1}{2}}\int_u^{u+\frac{\delta_1}{2}} f(s, y+z_1+z_2)dsdudz_1dz_2 \\ &\quad -\frac{4}{\delta_1^2\delta_2^2}\int_0^{\frac{\delta_2}{2}}\int_0^{\frac{\delta_2}{2}}\int_x^{x+\delta_1}\int_u^{u+\delta_1} f(s, y+2(z_1+z_2))dsdudz_1dz_2 \\ &= \frac{32}{\delta_1^2\delta_2^2}\int_y^{y+\frac{\delta_2}{2}}\int_v^{v+\frac{\delta_2}{2}}\int_0^{\frac{\delta_1}{2}}\int_0^{\frac{\delta_1}{2}} f(x+s_1+s_2, w)ds_1ds_2dwdv \\ &\quad -\frac{4}{\delta_1^2\delta_2^2}\int_y^{y+\delta_2}\int_v^{v+\delta_2}\int_0^{\frac{\delta_1}{2}}\int_0^{\frac{\delta_1}{2}} f(x+2s_1+2s_2, w)ds_1ds_2dwdv \\ &= \frac{32}{\delta_1^2\delta_2^2}\int_0^{\frac{\delta_2}{2}}\int_y^{y+\frac{\delta_2}{2}}\int_0^{\frac{\delta_1}{2}}\int_x^{x+\frac{\delta_1}{2}} f(u+s_2, v+z_2)duds_2dvdz_2 \\ &\quad -\frac{4}{\delta_1^2\delta_2^2}\int_0^{\frac{\delta_2}{2}}\int_y^{y+\delta_2}\int_0^{\frac{\delta_1}{2}}\int_x^{x+\delta_1} f(u+2s_2, v+2z_2)duds_2dvdz_2. \end{aligned}$$

Hence

$$\begin{aligned} \frac{\partial^2}{\partial x^2} f_{\delta_1\delta_2}(x,y) &= \frac{32}{\delta_1^2\delta_2^2}\int_0^{\frac{\delta_2}{2}}\int_0^{\frac{\delta_2}{2}}\Big[f(x+\delta_1, y+z_1+z_2) \\ &\quad -2f\left(x+\frac{\delta_1}{2}, y+z_1+z_2\right)+f(x, y+z_1+z_2)\Big]dz_1dz_2 \\ &\quad -\frac{4}{\delta_1^2\delta_2^2}\int_0^{\frac{\delta_2}{2}}\int_0^{\frac{\delta_2}{2}}\Big[f(x+2\delta_1, y+2(z_1+z_2)) \\ &\quad -2f(x+\delta_1, y+2(z_1+z_2))+f(x, y+2(z_1+z_2))\Big]dz_1dz_2 \end{aligned}$$

and $$\left|\frac{\partial^2}{\partial x^2} f_{\delta_1\delta_2}(x,y)\right| \le \frac{8}{\delta_1^2}\omega_2\left(f;\frac{\delta_1}{2},\frac{\delta_2}{2}\right)+\frac{1}{\delta_1^2}\omega_2(f;\delta_1,\delta_2) \le \frac{9}{\delta_1^2}\omega_2(f;\delta_1,\delta_2).$$

Similarly, we get

$$\left|\frac{\partial^2}{\partial y^2} f_{\delta_1\delta_2}(x,y)\right| \le \frac{9}{\delta_2^2}\omega_2(f;\delta_1,\delta_2),$$

$$\left|\frac{\partial^2}{\partial x\partial y} f_{\delta_1\delta_2}(x,y)\right| \le \frac{9}{\delta_1\delta_2}\omega_{\text{mix}}(f;\delta_1,\delta_2), \quad (x,y)\in\mathbb{R}^2_{AB}.$$

From the above and by Theorem 6, we obtain

$$\begin{aligned}
&\left|M_{n,m}^{\alpha,\beta,a,b}(f;x,y)-f(x,y)\right| \\
&\quad\le H_{n,m}^{\alpha,\beta,a,b}\left(\left|f-f_{\delta_1\delta_2}\right|;x,y\right)+\left|H_{n,m}^{\alpha,\beta,a,b}(f_{\delta_1\delta_2};x,y)-f_{\delta_1\delta_2}(x,y)\right|+\left|f_{\delta_1\delta_2}(x,y)-f(x,y)\right| \\
&\quad+\left|f\left(\frac{\alpha+1}{n}+x+\frac{ax}{n(1+x)},\frac{\beta+1}{m}+y+\frac{by}{m(1+y)}\right)-f(x,y)\right| \\
&\quad\le C\left\{\omega_2(f;\delta_1,\delta_2)+\omega_{\text{mix}}(f;\delta_1,\delta_2)+\omega_1\left(f;\frac{\alpha+1}{n}+\frac{ax}{n(1+x)},\frac{\beta+1}{m}+\frac{by}{m(1+y)}\right)\right\},
\end{aligned}$$

where C is a positive constant. This completes the proof. □

The authors would like to thank the referees for their helpful remarks which improved the exposition of the paper.

Reference

[1] Anastassiou G.A., Gal S.G., *Approximation theory: moduli of continuity and global smoothness preservation*, Birkhauser, Boston 2000.

[2] Atakut Ç., Büyükyazıcı İ., Serenbay S., *Approximation properties of Baskakov-Balazs type operators for functions of two variables*, "Miskolc Math. Notes" 16.2/2015, 667–678.

[3] Dhingas A., *Über einige Identitäten vom Bernstein Types*, "Norske Vid. Selsk. Trodheim" 24/1951, 96–97.

[4] Erençin A., *Durrmeyer type modification of generalized Baskakov operators*, "Appl. Math. Comput." 218/2011, 4384–4390.

[5] Gurdek M., Rempulska L., Skorupka M., *The Baskakov operators for functions of two variables*, "Collect. Math." 50.3/1999, 289–302.

[6] İzgi A., *Order of approximation of functions of two variables by new type gamma operators*, "General Mathematics" 17.1/2009, 23–32.

[7] Kajla A., Ispir N., Agraval P.N., Goyal M., *Q-Bernstein-Schurer-Durrmeyer type operators for functions of one and two variables*, "Appl. Math. Comput." 275/2016, 372–385.

[8] Krech G., Wachnicki E., *Direct estimate for some operators of Durrmeyer type in exponential weighted space*, "Demonstratio Math." 47.2/2014, 336–349.

[9] Lorenz G.G., *Bernstein Polynomials*, Univ. of Toronto Press, Toronto 1953.

[10] Malejki R., Wachnicki E., *On the Baskakov-Durrmeyer type operators*, "Comment. Math." 54.1/2014, 39–49.

[11] Miheşan V., *Uniform approximation with positive linear operators generalized Baskakov method*, "Automat. Comput. Appl. Math." 7.1/1998, 34–37.
[12] Rempulska L., Graczyk S., *On generalized Szász-Mirakjan operators of functions of two variables*, "Math. Slovaca" 62.1/2012, 87–98.
[13] Skorupka M., *On approximation of functions of two variables by some linear positive operators*, "Le Matematiche" 50.2/1995, 323–336.
[14] Stancu D.D., *On certain polynomials of two variables of Bernstein type and some applications of them*, "Dokl. Akad. Nauk SSSR" 134.1/1960, 221–233.
[15] Taşdelen F., Olgun A., Başcanbaz-Tunca G., *Approximation of functions of two variables by certain linear positive operators*, "Proceedings of the Indian Academy of Sciences: Mathematical Sciences" 117.3/2007, 387–399.
[16] Timan A.F., *Theory of Approximation of Functions of a Real Variable*, Moscow 1960 [in Russian].
[17] Wachnicki E., *Approximation by bivariate Mazhar-Totik operators*, "Comment. Math." 50.2/2010, 141–153.
[18] Wafi, A., Khatoon S., *Direct and inverse theorems for generalized Baskokov operators in polynomial weight space*, "An. Ştiint. Univ. Al. I. Cuza Iaşi. Mat. (N.S.)" 50.1/2004, 159–173.
[19] Wafi, A., Khatoon S., *On the order of approximation of functions by generalized Baskakov operators*, "Indian J. Pure Appl. Math." 35.3/2004, 347–358.
[20] Walczak Z., *On certain modified Szász-Mirakyan operators for functions of two variables*, "Demonstratio Math." 33.1/2000, 91–100.
[21] Walczak Z., *Approximation by some linear positive operators of functions of two variables*, "Saitama Math. J." 21/2003, 23–31.
[22] Volkov V.I., *On the convergence of sequences of linear positive operators in the space of two variables*, "Dokl. Akad. Nauk. SSSR" 115/1957, 17–19.

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

IRENEUSZ KRECH*, RENATA MALEJKI*

APPROXIMATION OF FUNCTIONS OF SEVERAL VARIABLES BY THE BASKAKOV-DURRMEYER TYPE OPERATORS

APROKSYMACJA FUNKCJI WIELU ZMIENNYCH OPERATORAMI TYPU BASKAKOWA-DURRMEYERA

A b s t r a c t

In this paper we introduce some linear positive operators of the Baskakov-Durrmeyer type in the space of uniformly continuous and bounded functions of several variables. The theorem on the degree of the convergence is established. Moreover, we give the Voronovskaya type formula for these operators.

Keywords: Baskakov-Durrmeyer type operator, *linear operators*, *approximation order*, *Voronovskaya type theorem*

S t r e s z c z e n i e

W artykule rozważa się operatory typu Baskakowa-Durrmeyera w przestrzeni ograniczonych i jednostajnie ciągłych funkcji wielu zmiennych. Formułuje się i dowodzi twierdzenia dotyczące rzędu zbieżności, jak również twierdzenie typu Woronowskiej dla tych operatorów.

Słowa kluczowe: operatory typu Baskakowa-Durrmeyera, *operatory liniowe*, *rząd aproksymacji*, *twierdzenie typu Woronowskiej*

DOI: 10.4467/2353737XCT.16.142.5753

* Ireneusz Krech, Renata Malejki (rmalejki@o2.pl), Institute of Mathematics, Pedagogical University of Cracow.

1. Introduction

Let $\mathbb{R}_0^+ = [0,\infty)$, $\mathbb{N} = \{1,2\ldots\}$, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ and for every fixed $m \in \mathbb{N}$ let

$$\mathbb{N}^m = \{\boldsymbol{n} = (n_1,\ldots,n_m) : n_k \in \mathbb{N} \quad \text{for} \quad 1 \le k \le m\},$$
$$\mathbb{R}_+^m = \{\boldsymbol{x} = (x_1,\ldots,x_m) : x_k \in \mathbb{R}_0^+ \quad \text{for} \quad 1 \le k \le m\}.$$

Analogously we define $\mathbb{R}^m$. We denote $\overline{\boldsymbol{\lambda}} = (\lambda,\lambda,\ldots,\lambda) \in \mathbb{R}^m$. For $\boldsymbol{n} \in \mathbb{N}^m$ we write $\boldsymbol{n} \to \overline{\infty}$ if and only if $n_k \to +\infty$ for $k = 1,2\ldots m$. Moreover, for a fixed $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_+^m$, we will use the notation

$$\int_{\boldsymbol{x}}^{\boldsymbol{y}} f(\boldsymbol{s})d\boldsymbol{s} = \int_{x_1}^{y_1} \ldots \int_{x_m}^{y_m} f(s_1,\ldots s_m)ds_1 \ldots ds_m.$$

We denote by $C_B(\mathbb{R}_+^m)$ the space of all real-valued functions f uniformly continuous and bounded on $\mathbb{R}_+^m$. The norm on $C_B(\mathbb{R}_+^m)$ is defined by $\|f\|_{C_B(\mathbb{R}_+^m)} = \sup_{\boldsymbol{x} \in \mathbb{R}_+^m} |f(\boldsymbol{x})|$. Let

$$W_{n,k}^a(x) = e^{-\frac{ax}{1+x}} \sum_{i=0}^{k} \binom{k}{i} (n)_i \, a^{k-i} \frac{x^k}{k!(1+x)^{n+k}},$$

where $a \in \mathbb{R}_0^+$, $(n)_0 = 1$, $(n)_i = n(n+1)\ldots(n+i-1)$, $\quad i \ge 1$.

For a real-valued function f defined on the interval $[0,\infty)$, the generalized Baskakov--Durrmeyer type operators is defined by (see [11])

$$M_n^{\alpha,a}(f;x) = n\sum_{k=0}^{\infty} W_{n,k}^a(x) \frac{1}{\Gamma(\alpha+k+1)} \int_0^{\infty} e^{-ns}(ns)^{\alpha+k} f(s)ds, \quad \alpha > -1. \tag{1.1}$$

In the present paper, inspired by operator (1.1), we introduce the following class of operators $\boldsymbol{M_n^{\alpha,a}}$ given by the formula

$$\boldsymbol{M_n^{\alpha,a}}(f;x) = \sum_{k_1,\ldots,k_m=0}^{\infty} \int_{\boldsymbol{0}}^{\overline{\infty}} \prod_{j=1}^{m} W_{n_j,k_j}^{a_j}(x_j) \frac{n_j}{\Gamma(\alpha_j + k_j + 1)} e^{-n_j s_j} (n_j s_j)^{\alpha_j + k_j} f(\boldsymbol{s})d\boldsymbol{s} \tag{1.2}$$

for $x \in \mathbb{R}_+^m$, where $\boldsymbol{n} \in \mathbb{N}^m$, $\boldsymbol{a} \in \mathbb{R}_+^m$, $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_m)$, $\alpha_k > -1$ for $k = 1,2,\ldots,m$. It is obvious that the operator $\boldsymbol{M_n^{\alpha,a}}$ is linear and positive on $\mathbb{R}_+^m$. Basic facts on positive linear operators, their generalizations and applications, can be found in [3], [4].

Observe that if $f(\boldsymbol{s}) = f_1(s_1)\cdot\ldots\cdot f_m(s_m)$ for $\boldsymbol{s} \in \mathbb{R}_+^m$, then

$$\boldsymbol{M_n^{\alpha,a}}(f;x) = \prod_{j=1}^{m} M_{n_j}^{\alpha_j,a_j}(f_j;\boldsymbol{x}_j),$$

where

$$M_{n_j}^{\alpha_j,a_j}(f_j;x_j)=n_j\sum_{k_j=0}^{\infty}W_{n_j,k_j}^{a_j}(x_j)\frac{1}{\Gamma(\alpha_j+k_j+1)}\int_0^{\infty}e^{-n_js_j}(n_js_j)^{\alpha_j+k_j}f_j(s_j)ds_j.$$

Some properties of the operator defined by (1.1) in particular, an estimation of the rate of convergence, were studied in [11].

Lemma 1 [11]. *Let* $\varphi^r(t)=t^r$, $t\in\mathbb{R}_0^+$, $r\in\mathbb{N}_0$. For $x\geq 0$, $\alpha>-1$ and $a\geq 0$, *we have*

$$M_n^{\alpha,a}(\varphi^0;x)=1,$$

$$M_n^{\alpha,a}(\varphi^1;x)=\frac{\alpha+1}{n}+x+\frac{ax}{n(1+x)},$$

$$M_n^{\alpha,a}(\varphi^1-x;x)=\frac{\alpha+1}{n}+\frac{ax}{n(1+x)},$$

$$M_n^{\alpha,a}((\varphi^1-x)^2;x)=\frac{(\alpha+1)(\alpha+2)}{n^2}+\frac{2x+x^2}{n}+\frac{a^2x^2}{n^2(1+x)^2}+\frac{2(\alpha+2)ax}{n^2(1+x)},$$

$$\lim_{n\to 0}n^2M_n^{\alpha,a}((\varphi^1-x)^4;x)=12x^2+12x^3+3x^4-3ax^3+\frac{(3x-5)ax^3}{1+x}-\frac{[12(\alpha+2)+4]a^2x^3}{(1+x)^2}.$$

Using the definition of $\boldsymbol{M_n^{\alpha,a}}$, it is easy to prove the next theorem.

Theorem 1. *Let* $f\in C_B(\mathbb{R}_+^m)$. *Then*

$$\left\|\boldsymbol{M_n^{\alpha,a}}(f)\right\|_{C_B(\mathbb{R}_+^m)}\leq\|f\|_{C_B(\mathbb{R}_+^m)}$$

for all $\boldsymbol{n}\in\mathbb{N}^m$.

This paper is devoted to a study aimed at obtaining approximation results by using the modulus of continuity and the Voronovskaya asymptotic formula for the Baskakov--Durrmeyer type operators defined by (1.2) in the space of uniformly continuous and bounded functions of several variables. Approximation properties of various positive linear operators for functions of one, two and several variables have been investigated in many papers (for example [2], [5], [7], [8], [9], [10], [12], [13]).

2. Rate of convergence

In this section we shall prove two theorems on the degree of approximation of functions belonging to the class $C_B(\mathbb{R}_+^m)$ by $\boldsymbol{M_n^{\alpha,a}}$. We shall apply the method used in [6].

We denote

$$C_B^1(\mathbb{R}_+^m) := \left\{ f \in C_B(\mathbb{R}_+^m) : \ \frac{\partial f}{\partial x_k} \in C_B(\mathbb{R}_+^m), \quad 1 \le k \le m \right\}.$$

In order to prove the approximation theorem we need the following result. Let $\varphi^1(s_j) = s_j,\ s_j \in \mathbb{R}_0^+,\ j = 1,2,\ldots m,\ r \in \mathbb{N}.$

Theorem 2. *If* $g \in C_B^1(\mathbb{R}_+^m)$, *then*

$$\left| \boldsymbol{M}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{a}}(g;x) - g(x) \right|$$

$$\le \sum_{j=1}^{m} \left\| \frac{\partial g}{\partial x_j} \right\|_{C_B(\mathbb{R}_+^m)} \left(\frac{(\alpha_j+1)(\alpha_j+2)}{n_j^2} + \frac{2x_j + x_j^2}{n_j} + \frac{a_j^2 x_j^2}{n_j^2 (1+x_j)^2} + \frac{2(\alpha_j+2)a_j x_j}{n_j^2 (1+x_j)} \right)^{1/2}$$

for all $\boldsymbol{x} \in \mathbb{R}_+^m$, where $\boldsymbol{n} \in \mathbb{N}^m$, $\boldsymbol{a} \in \mathbb{R}_+^m$, $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_m)$, $\alpha_k > -1$ for $k = 1,2,\ldots,m$

Proof. Fix $\boldsymbol{x} \in \mathbb{R}_+^m$. For every $\boldsymbol{s} \in \mathbb{R}_+^m$ we have

$$g(\boldsymbol{s}) - g(\boldsymbol{x}) = \sum_{j=1}^{m} \int_{x_j}^{s_j} \frac{\partial}{\partial u_j} g(\boldsymbol{y}_j) du_j,$$

where $\boldsymbol{y}_j = (x_1,\ldots,x_{j-1},u_j,s_{j+1},\ldots,s_m)$. Observe that

$$\left| \int_{x_j}^{s_j} \frac{\partial}{\partial u_j} g(\boldsymbol{y}_j) du_j \right| \le \left\| \frac{\partial g}{\partial x_j} \right\|_{C_B(\mathbb{R}_+^m)} \left| s_j - x_j \right|$$

and

$$M_{n_j}^{\alpha_j,a_j}\left(\left| \int_{x_j}^{\varphi^1} \frac{\partial}{\partial u_j} g(\boldsymbol{y}_j) du_j \right| ; x_j \right) \le \left\| \frac{\partial g}{\partial x_j} \right\|_{C_B(\mathbb{R}_+^m)} M_{n_j}^{\alpha_j,a_j}\left(\left| \varphi^1 - x_j \right| ; x_j \right).$$

Applying the Cauchy-Schwarz inequality we obtain

$$M_{n_j}^{\alpha_j,a_j}\left(\left| \int_{x_j}^{\varphi^1} \frac{\partial}{\partial u_j} g(\boldsymbol{y}_j) du_j \right| ; x_j \right) \le \left\| \frac{\partial g}{\partial x_j} \right\|_{C_B(\mathbb{R}_+^m)} \left(M_{n_j}^{\alpha_j,a_j}\left((\varphi^1 - x_j)^2 ; x_j \right) \right)^{\frac{1}{2}}.$$

From the above, using the linearity of $\boldsymbol{M}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{a}}$ and Lemma 1, we obtain

$$\left| \boldsymbol{M}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{a}}(g;\boldsymbol{x}) - g(\boldsymbol{x}) \right|$$

$$\le \sum_{j=1}^{m} \left\| \frac{\partial g}{\partial x_j} \right\|_{C_B(\mathbb{R}_+^m)} \left(\frac{(\alpha_j+1)(\alpha_j+2)}{n_j^2} + \frac{2x_j + x_j^2}{n_j} + \frac{a_j^2 x_j^2}{n_j^2 (1+x_j)^2} + \frac{2(\alpha_j+2)a_j x_j}{n_j^2 (1+x_j)} \right)^{1/2},$$

whence the result. □

In the next theorem we will use the modulus of continuity of $f \in C_B(\mathbb{R}^m_+)$ given by

$$\omega(f;\boldsymbol{\delta}) = \sup_{\substack{0<h_1\le\delta_1 \\ \dots \\ 0<h_m\le\delta_m}} \left\|\Delta_{\boldsymbol{h}} f\right\|_{C_B(\mathbb{R}^m_+)}, \quad \boldsymbol{\delta}, \boldsymbol{h} \in \mathbb{R}^m_+ \setminus \{\overline{\mathbf{0}}\},$$

where

$$\Delta_h f(\boldsymbol{x}) = f(\boldsymbol{x}+\boldsymbol{h}) - f(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^m_+.$$

For a fixed $\boldsymbol{\beta} = (\beta_1,\dots,\beta_m)$, $0 < \beta_j \le 1$ for $j = 1,2,\dots,m$, we denote by $Lip(C_B(\mathbb{R}^m_+);\boldsymbol{\beta})$ the class of all functions $f \in C_B(\mathbb{R}^m_+)$ for which $\omega(f;\boldsymbol{\delta}) = O\left(\delta_1^{\beta_1} + \dots + \delta_m^{\beta_m}\right)$ as $\delta_j \to 0^+$ for $j = 1,2,\dots,\mathrm{m}$.

Theorem 3. *Suppose that $f \in C_B(\mathbb{R}^m_+)$. Then for all $\boldsymbol{x} \in \mathbb{R}^m_+$, it holds*

$$\left|\boldsymbol{M}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{a}}(f;\boldsymbol{x}) - f(\boldsymbol{x})\right| \le 2(m+1)\omega\left(f;\delta\right),$$

where

$$\delta_j = \left(\frac{(\alpha_j+1)(\alpha_j+2)}{n_j^2} + \frac{2x_j + x_j^2}{n_j} + \frac{a_j^2 x_j^2}{n_j^2(1+x_j)^2} + \frac{2(\alpha_j+2)a_j x_j}{n_j^2(1+x_j)}\right)^{1/2}, \quad j = 1,2,\dots,m.$$

Proof. Let $f_{\boldsymbol{\delta}}$ be the Steklov mean of $f \in C_B(\mathbb{R}^m_+)$

$$f_{\boldsymbol{\delta}}(\boldsymbol{x}) = \left(\prod_{j=1}^m \delta_j\right)^{-1} \int_{\mathbf{0}}^{\boldsymbol{\delta}} f(\boldsymbol{x}+\boldsymbol{u})\, d\boldsymbol{u}$$

for $\boldsymbol{x} \in \mathbb{R}^m_+$, $\delta_j > 0$ for $j = 1,\dots,m$. We have

$$f_{\boldsymbol{\delta}}(\boldsymbol{x}) - f(\boldsymbol{x}) = \left(\prod_{j=1}^m \delta_j\right)^{-1} \int_{\mathbf{0}}^{\boldsymbol{\delta}} (f(\boldsymbol{x}+\boldsymbol{u}) - f(\boldsymbol{x}))\, d\boldsymbol{u}$$

and

$$\frac{\partial}{\partial x_j} f_{\boldsymbol{\delta}}(\boldsymbol{x}) = \left(\prod_{j=1}^m \delta_j\right)^{-1} \int_0^{\delta_1} \dots \int_0^{\delta_{j-1}} \int_0^{\delta_{j+1}} \dots \int_0^{\delta_m} (f(\boldsymbol{x}+\boldsymbol{u}^*) - f(\boldsymbol{x}+\boldsymbol{u}_*))\, du_1 \dots du_{j-1} du_{j+1} \dots du_m$$

for $j = 1,\dots,m$, where $\boldsymbol{u}^* = (u_1,\dots,u_{j-1},\delta_j,u_{j+1},\dots,u_m)$, $\boldsymbol{u}_* = (u_1,\dots,u_{j-1},0,u_{j+1},\dots,u_m)$.

From this we obtain

$$\left\|f_{\boldsymbol{\delta}} - f\right\|_{C_B(\mathbb{R}^m_+)} \le \omega(f;\boldsymbol{\delta}), \tag{2.1}$$

$$\left\| \frac{\partial f_{\boldsymbol{\delta}}}{\partial x_j} \right\|_{C_B(\mathbb{R}_+^m)} \leq 2\delta_j^{-1}\omega(f;\boldsymbol{\delta}), \tag{2.2}$$

which implies $f_{\boldsymbol{\delta}} \in C_B^1(\mathbb{R}_+^m)$. Hence, for every $\boldsymbol{\delta} \in \mathbb{R}_+^m \setminus \{\overline{\mathbf{0}}\}$, $\boldsymbol{x} \in \mathbb{R}_+^m$ and $\boldsymbol{n} \in \mathbb{N}^m$, we can write

$$\left|M_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{a}}(f;\boldsymbol{x}) - f(\boldsymbol{x})\right| \leq \left|M_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{a}}(f - f_{\boldsymbol{\delta}};\boldsymbol{x})\right| + \left|M_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{a}}(f_{\boldsymbol{\delta}};\boldsymbol{x}) - f_{\boldsymbol{\delta}}(\boldsymbol{x})\right| + \left|f_{\boldsymbol{\delta}}(\boldsymbol{x}) - f(\boldsymbol{x})\right|.$$

Using Theorem 1 and (2.1), we get

$$\left|M_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{a}}(f - f_{\boldsymbol{\delta}};\boldsymbol{x})\right| \leq \|f_{\boldsymbol{\delta}} - f\| \leq \omega(f;\boldsymbol{\delta}).$$

By Theorem 2 and (2.2), it follows

$$\begin{aligned}
&\left|M_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{a}}(f_{\boldsymbol{\delta}};\boldsymbol{x}) - f_{\boldsymbol{\delta}}(\boldsymbol{x})\right| \\
&\quad \leq \sum_{j=1}^{m} \left\| \frac{\partial f_{\boldsymbol{\delta}}}{\partial x_j} \right\|_{C_B(\mathbb{R}_+^m)} \left(\frac{(\alpha_j+1)(\alpha_j+2)}{n_j^2} + \frac{2x_j + x_j^2}{n_j} + \frac{a_j^2 x_j^2}{n_j^2(1+x_j)^2} + \frac{2(\alpha_j+2)a_j x_j}{n_j^2(1+x_j)} \right)^{1/2} \\
&\quad \leq 2\omega(f;\boldsymbol{\delta}) \sum_{j=1}^{m} \delta_j^{-1} \left(\frac{(\alpha_j+1)(\alpha_j+2)}{n_j^2} + \frac{2x_j + x_j^2}{n_j} + \frac{a_j^2 x_j^2}{n_j^2(1+x_j)^2} + \frac{2(\alpha_j+2)a_j x_j}{n_j^2(1+x_j)} \right)^{1/2}
\end{aligned}$$

Consequently

$$\begin{aligned}
&\left|M_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{a}}(f;\boldsymbol{x}) - f(\boldsymbol{x})\right| \leq 2\omega(f;\boldsymbol{\delta}) \\
&\qquad \times \left\{ 1 + \sum_{j=1}^{m} \delta_j^{-1} \left(\frac{(\alpha_j+1)(\alpha_j+2)}{n_j^2} + \frac{2x_j + x_j^2}{n_j} + \frac{a_j^2 x_j^2}{n_j^2(1+x_j)^2} + \frac{2(\alpha_j+2)a_j x_j}{n_j^2(1+x_j)} \right)^{1/2} \right\}
\end{aligned}$$

for all $\boldsymbol{\delta} \in \mathbb{R}_+^m \setminus \{\overline{\mathbf{0}}\}$. Choosing $\boldsymbol{\delta}$ with

$$\delta_j = \left(\frac{(\alpha_j+1)(\alpha_j+2)}{n_j^2} + \frac{2x_j + x_j^2}{n_j} + \frac{a_j^2 x_j^2}{n_j^2(1+x_j)^2} + \frac{2(\alpha_j+2)a_j x_j}{n_j^2(1+x_j)} \right)^{1/2},$$

$j = 1, \ldots, m$, we obtain the assertion. □

From Theorem 3, using the properties of modulus of continuity for uniformly continous function (see [1], [4]), we can derive the following corollaries.

Corollary 1. *If* $f \in C_B(\mathbb{R}_+^m)$, *then*

$$\lim_{\boldsymbol{n} \to \overline{\infty}} M_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{a}}(f;\boldsymbol{x}) = f(\boldsymbol{x})$$

for every $\boldsymbol{x} \in \mathbb{R}_+^m$. Moreover, this convergence is uniform on every compact set $\boldsymbol{I} \subset \mathbb{R}_+^m$.

Corollary 2. *Let* $f \in Lip(C_B(\mathbb{R}_+^m);\boldsymbol{\beta})$ *with some fixed* $m \in \mathbb{N}$ *and* $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_m)$, $0 < \beta_j \le 1$ for $j = 1,2,\ldots m$. *Then for all* $\boldsymbol{a},\boldsymbol{x} \in \mathbb{R}_+^m$, $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_m)$, $\alpha_k > -1$ *for* $k = 1,2,\ldots,m$ *and* $\boldsymbol{n} \in \mathbb{N}^m$, *it holds*

$$\left|M_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{a}}(f;\boldsymbol{x}) - f(\boldsymbol{x})\right| \le \sum_{j=1}^{m}\left(\frac{(\alpha_j+1)(\alpha_j+2)}{n_j^2} + \frac{2x_j + x_j^2}{n_j} + \frac{a_j^2 x_j^2}{n_j^2(1+x_j)^2} + \frac{2(\alpha_j+2)a_j x_j}{n_j^2(1+x_j)}\right)^{\beta_j/2}.$$

3. The Voronovskaya type theorem

Let $\overline{\boldsymbol{n}} = (n,\ldots,n) \in \mathbb{N}^m$. In this part, we will consider the operator $\boldsymbol{M}_{\overline{\boldsymbol{n}}}^{\boldsymbol{\alpha},\boldsymbol{a}}$. In order to state the Voronovskaya type theorem we need the following result, which is a simple consequence of Lemma 1.

Lemma 2. *Let* $\boldsymbol{x} = (x_1,\ldots,x_m) \in \mathbb{R}_+^m$ *be a fixed point. Then*

$$\begin{aligned}
&\lim_{n\to\infty} nM_n^{\alpha_j,a_j}(\varphi^1 - x_j;x_j) = \alpha_j + 1 + \frac{a_j x_j}{1+x_j},\\
&\lim_{n\to\infty} nM_n^{\alpha_j,a_j}(\varphi^1 - x_j;x_j)M_n^{\alpha_i,a_i}(\varphi^1 - x_i;x_i) = 0, \quad i \ne j,\\
&\lim_{n\to\infty} nM_n^{\alpha_j,a_j}((\varphi^1 - x_j)^2;x_j) = 2x_j + x_j^2,\\
&\lim_{n\to\infty} n^2 M_n^{\alpha_j,a_j}((\varphi^1 - x_j)^4;x_j) = 12x_j^2 + 12x_j^3 + 3x_j^4 - 3a_j x_j^3\\
&\qquad + \frac{(3x_j - 5)a_j x_j^3}{1+x_j} - \frac{[12(\alpha_j+2)+4]a_j^2 x_j^3}{(1+x_j)^2}
\end{aligned} \tag{3.1}$$

for $j = 1,2,\ldots,m$.

Theorem 4. *Let* $f \in C_B(\mathbb{R}_+^m)$ and $\boldsymbol{x} \in \mathbb{R}_+^m$. *If* f *is of the class* $C_B^1(\mathbb{R}_+^m)$ *in a certain neighbourhood of a point* $\boldsymbol{x}$ *and* $f''(\boldsymbol{x})$ *exists (in the Fréchet sense), then for every* $\boldsymbol{x} \in \mathbb{R}_+^m$, *we have*

$$\lim_{n\to\infty} n\{\boldsymbol{M}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{a}}(f;\boldsymbol{x}) - f(\boldsymbol{x})\} = \sum_{j=1}^{m}\left(\alpha_j + 1 + \frac{a_j x_j}{1+x_j}\right)\frac{\partial f(\boldsymbol{x})}{\partial x_j} + \sum_{j=1}^{m}\left(x_j + \frac{1}{2}x_j^2\right)\frac{\partial^2 f(\boldsymbol{x})}{\partial x_j^2}.$$

Proof. Let $\boldsymbol{x}$ be a fixed point in $\mathbb{R}_+^m$. By Taylor's formula we get

$$f(\boldsymbol{s}) = f(\boldsymbol{x}) + \sum_{j=1}^{m}(s_j - x_j)\frac{\partial f(\boldsymbol{x})}{\partial x_j} + \frac{1}{2}\left\{\sum_{j=1}^{m}(s_j - x_j)^2\frac{\partial^2 f(\boldsymbol{x})}{\partial x_j^2}\right.$$
$$\left. +2\sum_{i\neq j, i,j=1}^{m}(s_i - x_i)(s_j - x_j)\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}\right\} + \psi_{\boldsymbol{x}}(\boldsymbol{s})\left(\sum_{j=1}^{m}(s_j - x_j)^4\right)^{1/2},$$

where the function ψ_x is uniformly continuous and bounded in $\mathbb{R}_+^m$ and $\lim_{\boldsymbol{s}\to\boldsymbol{x}}\psi_{\boldsymbol{x}}(\boldsymbol{s}) = 0$. From linearity of $\boldsymbol{M}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{a}}$, we obtain

$$\begin{aligned} n\{\boldsymbol{M}_{\bar{\boldsymbol{n}}}^{\boldsymbol{\alpha},\boldsymbol{a}}(f;\boldsymbol{x}) - f(\boldsymbol{x})\} &= n\sum_{j=1}^{m} M_n^{\alpha_j,a_j}(\varphi^1 - x_j; x_j)\frac{\partial f(\boldsymbol{x})}{\partial x_j} \\ &+ \frac{1}{2}n\left\{\sum_{j=1}^{m} M_n^{\alpha_j,a_j}((\varphi^1 - x_j)^2; x_j)\frac{\partial^2 f(\boldsymbol{x})}{\partial x_j^2}\right. \\ &\left. +2\sum_{i\neq j, i,j=1}^{m} M_n^{\alpha_j,a_j}(\varphi^1 - x_j; x_j)M_n^{\alpha_i,a_i}(\varphi^1 - x_i; x_i)\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}\right\} \\ &+ n\boldsymbol{M}_{\bar{\boldsymbol{n}}}^{\boldsymbol{\alpha},\boldsymbol{a}}\left(\psi_{\boldsymbol{x}}\phi_{\boldsymbol{x}}; \boldsymbol{x}\right), \end{aligned} \tag{3.2}$$

where $\phi_{\boldsymbol{x}}(\boldsymbol{s}) = \left(\sum_{j=1}^{m}(s_j - x_j)^4\right)^{1/2}$. Using the Cauchy-Schwarz inequality we obtain

$$n\left|\boldsymbol{M}_{\bar{\boldsymbol{n}}}^{\boldsymbol{\alpha},\boldsymbol{a}}(\psi_{\boldsymbol{x}}\phi_{\boldsymbol{x}}; \boldsymbol{x})\right| \leq \left|\boldsymbol{M}_{\bar{\boldsymbol{n}}}^{\boldsymbol{\alpha},\boldsymbol{a}}(\psi_{\boldsymbol{x}}^2; \boldsymbol{x})\right|^{1/2}\left|n^2\boldsymbol{M}_{\bar{\boldsymbol{n}}}^{\boldsymbol{\alpha},\boldsymbol{a}}(\phi_{\boldsymbol{x}}^2; \boldsymbol{x})\right|^{1/2}.$$

Moreover, the function $\psi_{\boldsymbol{x}}^2$ satisfies the assumption of Corollary 1. Hence

$$\lim_{n\to\infty}\boldsymbol{M}_{\bar{\boldsymbol{n}}}^{\boldsymbol{\alpha},\boldsymbol{a}}(\psi_{\boldsymbol{x}}^2; \boldsymbol{x}) = \psi_{\boldsymbol{x}}^2(\boldsymbol{x}) = 0.$$

Observe that

$$\boldsymbol{M}_{\bar{\boldsymbol{n}}}^{\boldsymbol{\alpha},\boldsymbol{a}}(\phi_{\boldsymbol{x}}^2; \boldsymbol{x}) = \boldsymbol{M}_{\bar{\boldsymbol{n}}}^{\boldsymbol{\alpha},\boldsymbol{a}}\left(\sum_{j=1}^{m}(\varphi^1 - x_j)^4; \boldsymbol{x}\right) = \sum_{j=1}^{m} M_n^{\alpha_j,a_j}((\varphi^1 - x_j)^4; x_j).$$

Using (3.1) we obtain

$$\lim_{n\to\infty} n\boldsymbol{M}_{\bar{\boldsymbol{n}}}^{\boldsymbol{\alpha},\boldsymbol{a}}(\psi_{\boldsymbol{x}}\phi_{\boldsymbol{x}}; \boldsymbol{x}) = 0. \tag{3.3}$$

From (3.2), (3.3) and Lemma 2 we get the assertion. □

Corollary 3. *Let* $\boldsymbol{x} \in \mathbb{R}_+^m$. *If* f *satisfies the assumption of Theorem 4, then*

$$\left|\boldsymbol{M}_{\bar{\boldsymbol{n}}}^{\boldsymbol{\alpha},\boldsymbol{a}}(f;\boldsymbol{x}) - f(\boldsymbol{x})\right| = O\left(\frac{1}{n}\right) \quad as \quad n \to \infty.$$

References

[1] Anastassiou G.A., Gal S.G., *Approximation theory: moduli of continuity and global smoothness preservation*, Birkhauser, Boston 2000.

[2] Atakut Ç., Büyükyazıcı İ., Serenbay S., *Approximation properties of Baskakov-Balazs type operators for functions of two variables*, "Miskolc Math. Notes" 16.2/2015, 667–678.

[3] DeVore R.A., Lorentz G.G., *Constructive Approximation*, Springer–Verlag, Berlin 1993.

[4] Ditzian Z., Totik V., *Moduli of Smoothness*, Springer–Verlag, New York 1987.

[5] Erençin A., *Durrmeyer type modification of generalized Baskakov operators*, "Appl. Math. Comput." 218/2011, 4384–4390.

[6] Firlej B., Rempulska L., *Approximation of functions of several variables by some operators of the Szasz-Mirakjan type*, "Fasc. Math." 27/1997, 15–27.

[7] Gurdek M., Rempulska L., Skorupka M., *The Baskakov operators for functions of two variables*, "Collect. Math." 50.3/1999, 289–302.

[8] İzgi A., *Order of approximation of functions of two variables by new type gamma operators*, "General Mathematics" 17.1/2009, 23–32.

[9] Kajla A., Ispir N., Agraval P.N., Goyal M., *Q-Bernstein-Schurer-Durrmeyer type operators for functions of one and two variables*, "Appl. Math. Comput." 275/2016, 372–385.

[10] Krech G., Wachnicki E., *Direct estimate for some operators of Durrmeyer type in exponential weighted space*, "Demonstratio Math." 47.2/2014, 336–349.

[11] Malejki R., Wachnicki E., *On the Baskakov-Durrmeyer type operators*, "Comment. Math." 54.1/2014, 39–49.

[12] Miheşan V., *Uniform approximation with positive linear operators generalized Baskakov method*, "Automat. Comput. Appl. Math." 7.1/1998, 34–37.

[13] Wafi, A., Khatoon S., *On the order of approximation of functions by generalized Baskakov operators*, "Indian J. Pure Appl. Math." 35.3/2004, 347–358.

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

KAMIL KULAR*

ON BASIC PROPERTIES OF δ-PRIME AND δ-SEMIPRIME RINGS

O PODSTAWOWYCH WŁASNOŚCIACH PIERŚCIENI δ-PIERWSZYCH I δ-PÓŁPIERWSZYCH

Abstract

We provide a self-contained discussion of the notions of δ-primeness and δ-semiprimeness for associative rings, possibly without identity. Some of the facts and properties presented in the article seem less known and quite difficult to find in standard reference sources.

Keywords: associative ring, derivation, δ-ideal, δ-prime ring, δ-semiprime ring, δ-prime radical, δ-nilpotent element.

AMS Mathematics Subject Classification (2010): 16W25, 16N60.

Streszczenie

Praca jest „samowystarczalnym" omówieniem pojęć δ-pierwszości i δ-półpierwszości, rozważanych w algebrze nieprzemiennej. Wszystkie udowodnione w pracy twierdzenia stosują się i do pierścieni z jedynką, i do pierścieni bez jedynki. Część faktów i własności przedstawionych w pracy wydaje się mało znana i raczej trudna do odszukania w standardowej literaturze.

Słowa kluczowe: pierścień łączny, różniczkowanie, δ-ideał, pierścień δ-pierwszy, pierścień δ-półpierwszy, radykał δ-pierwszy, element δ-nilpotentny.

DOI: 10.4467/2353737XCT.16.144.5755

This paper was prepared in LaTeX.

*Kamil Kular (kkular@pk.edu.pl), Institute of Mathematics, Cracow University of Technology.

1. Preliminaries and introduction

Throughout the article, R is an associative ring and $\delta : R \longrightarrow R$ is a derivation. We do not assume that R has an identity.

Definition 1.1. *A map $\delta : R \longrightarrow R$ is said to be a derivation, if it is additive and satisfies the Leibniz rule*

$$\forall a,b \in R:\ \delta(ab) = \delta(a)b + a\delta(b).$$

Notice that the zero map is a derivation of the ring R. We define

$$\delta^n = \begin{cases} \mathrm{id}_R, & \text{if } n = 0, \\ \underbrace{\delta \circ \ldots \circ \delta}_{n}, & \text{if } n \in \mathbb{N} \setminus \{0\}. \end{cases}$$

The center of the ring R will be denoted by $\mathrm{Z}(R)$, i.e.,

$$\mathrm{Z}(R) = \{a \in R : ab = ba \text{ for all } b \in R\}.$$

Let us remark that $\mathrm{Z}(R)$ is a subring of R. For any elements $a,b \in R$ we define $[a,b] = ab - ba$. By "ideal of the ring R" we always mean a left, right, or two-sided ideal.

Prime rings and, more generally, semiprime rings are fundamental objects of study in noncommutative algebra. For a long time the research has also been focused on various extensions of these classes of rings. Taking into account the analogues of prime and semiprime rings defined by means of ideals that are invariant with respect to either a single derivation or a family of derivations, yields important examples of such extensions. The analogues are referred to as δ-(semi)prime rings and Δ-(semi)prime rings, respectively. They still attract interest of algebraists.

The article does not bring new results. Our first purpose is to collect and systematize basic facts about δ-prime rings and δ-semiprime rings. Some of these facts seem a bit less known. The second purpose is to provide complete and self-contained proofs for all the presented theorems (the proofs are very often omitted in reference sources). The proofs we provide are mostly modifications of corresponding "nondifferential" proofs given in the classical monographs [4, 6, 7]. Two features of the article seem worth emphasizing: all the proofs are valid for rings without identity and a brief introduction to δ-nilpotent elements is included.

The article is organized as follows. In Section 2 we collect some useful facts and examples concerning δ-ideals. In Section 3 we discuss various characterizations of δ-prime rings and δ-prime ideals. Section 4 is devoted to strongly nilpotent elements and δ-nilpotent elements. Finally, in Section 5 we deal with characterizations of δ-semiprime rings.

2. δ-ideals

We begin with a few standard definitions.

Definition 2.1. *A set $S \subseteq R$ is called δ-stable, if $\delta(S) \subseteq S$. An ideal I of the ring R is said to be a δ-ideal, if it is δ-stable.*

Definition 2.2. *The two-sided ideal of R generated by the set $\{[a,b] : a,b \in R\}$ is called the commutator ideal. This ideal is denoted by* $\mathrm{C}(R)$.

Definition 2.3. *For a set $S \subseteq R$ we define*

- *the left annihilator* $\mathrm{ann}_\ell(S) = \{a \in R : ab = 0 \text{ for all } b \in S\}$,
- *the right annihilator* $\mathrm{ann}_r(S) = \{a \in R : ba = 0 \text{ for all } b \in S\}$.

Notice that if δ is the zero derivation, then every ideal of the ring R is a δ-ideal. Moreover, $\mathrm{ann}_\ell(S)$ is a left ideal of R, $\mathrm{ann}_r(S)$ is a right ideal of R, and R is commutative if and only if $\mathrm{C}(R) = \{0\}$.

Before we turn to more interesting observations, let us state an obvious but useful formula.

Lemma 2.4. *If $a,b \in R$, then $\delta([a,b]) = [\delta(a),b] + [a,\delta(b)]$.*

Take now a closer look at $\mathrm{C}(R)$, $\mathrm{Z}(R)$ and annihilators.

Proposition 2.5. *The commutator ideal $\mathrm{C}(R)$ is a δ-ideal and the center $\mathrm{Z}(R)$ is a δ-stable set. Moreover, if $S \subseteq R$ is a δ-stable set, then $\mathrm{ann}_\ell(S)$ and $\mathrm{ann}_r(S)$ are δ-ideals.*

Proof. Let us first define $A = \{[a,b] : a,b \in R\}$, $B = \{x[a,b] : a,b,x \in R\}$, $C = \{[a,b]y : a,b,y \in R\}$, and $D = \{x[a,b]y : a,b,x,y \in R\}$. Then $\mathrm{C}(R)$ coincides with the totality of finite sums of elements belonging to the set $A \cup B \cup C \cup D$. Pick arbitrary $a,b,x,y \in R$. By Lemma 2.4, we have

$$\begin{aligned}
\delta([a,b]) &= [\delta(a),b] + [a,\delta(b)] \in \mathrm{C}(R),\\
\delta(x[a,b]) &= \delta(x)[a,b] + x[\delta(a),b] + x[a,\delta(b)] \in \mathrm{C}(R),\\
\delta([a,b]y) &= [\delta(a),b]y + [a,\delta(b)]y + [a,b]\delta(y) \in \mathrm{C}(R),\\
\delta(x[a,b]y) &= \delta(x)[a,b]y + x[\delta(a),b]y + x[a,\delta(b)]y + x[a,b]\delta(y) \in \mathrm{C}(R).
\end{aligned}$$

The δ-stability of $\mathrm{C}(R)$ follows.

Now, pick an arbitrary $a \in \mathrm{Z}(R)$ and an arbitrary $b \in R$. Then $[a,b] = 0 = [a,\delta(b)]$, and hence

$$0 = \delta([a,b]) = [\delta(a),b] + [a,\delta(b)] = [\delta(a),b].$$

Consequently, $\delta(a) \in \mathrm{Z}(R)$. The δ-stability of $\mathrm{Z}(R)$ follows.

Suppose, finally, that $S \subseteq R$ is a δ-stable set. Pick an arbitrary $a \in \mathrm{ann}_\ell(S)$ and an arbitrary $b \in S$. Then $ab = 0$ and $\delta(b) \in S$. Consequently,

$$0 = \delta(ab) = \delta(a)b + a\delta(b) = \delta(a)b.$$

This yields $\delta(a) \in \mathrm{ann}_\ell(S)$. The δ-stability of $\mathrm{ann}_r(S)$ can be proved analogously. □

The intersection of any family of two-sided δ-ideals of the ring R is also a two-sided δ-ideal. Obviously, the statement remains true, if we replace the word "two-sided" by "left" or "right". We are thus enabled to consider δ-ideals generated by subsets of R.

Let us define $\langle S\rangle^\delta$, $\langle S\rangle_\ell^\delta$ and $\langle S\rangle_r^\delta$ to be the two-sided, the left and the right δ-ideal of the ring R generated by a set $S \subseteq R$ (respectively). We will write as usual $\langle a\rangle^\delta$ instead of $\langle\{a\}\rangle^\delta$, and analogously for the left and the right δ-ideal generated by the singleton $\{a\}$.

Proposition 2.6. *Let $a \in R$. Define $A = \{k\delta^n(a) : k \in \mathbb{Z}, n \in \mathbb{N}\cup\{0\}\}$, $B = \{x\delta^n(a) : x \in R, n \in \mathbb{N}\cup\{0\}\}$, $C = \{\delta^n(a)y : y \in R, n \in \mathbb{N}\cup\{0\}\}$, and $D = \{x\delta^n(a)y : x,y \in R, n \in \mathbb{N}\cup\{0\}\}$. Then*

(i) $\langle a\rangle^{\delta}$ *coincides with the totality of finite sums of elements belonging to the set $A\cup B\cup C\cup D$,*

(ii) $\langle a\rangle_{\ell}^{\delta}$ *coincides with the totality of finite sums of elements belonging to the set $A\cup B$,*

(iii) $\langle a\rangle_{r}^{\delta}$ *coincides with the totality of finite sums of elements belonging to the set $A\cup C$.*

Proof. Denote by T the totality of finite sums of elements belonging to $A\cup B\cup C\cup D$. Notice that T is an additive subgroup of the ring R. Moreover, T is a two-sided ideal and $a = \delta^0(a) \in T$. A reasoning similar to the proof of the δ-stability of $\mathrm{C}(R)$ shows that T is δ-stable. We therefore get $\langle a\rangle^{\delta} \subseteq T$. On the other hand, if $I \subseteq R$ is a two-sided δ-ideal and $a \in I$, then clearly $T \subseteq I$. The converse inclusion follows. Properties (ii) and (iii) can be proved analogously. □

It seems worth noting that in the above proposition

$$B = \bigcup_{n=0}^{\infty} R\delta^n(a),\ C = \bigcup_{n=0}^{\infty} \delta^n(a)R,\ D = \bigcup_{n=0}^{\infty} R\delta^n(a)R.$$

We conclude the section with some remarks on products and sums of δ-ideals. Let $k \in \mathbb{N}\setminus\{0\}$ and $S_1,\dots,S_k \subseteq R$. If either all the sets are two-sided ideals or all the sets are left ideals or all the sets are right ideals, then we define $S_1\cdot\ldots\cdot S_k$ to be the additive subgroup of the ring R generated by the "elementwise product" $\{a_1\cdot\ldots\cdot a_k : a_1 \in S_1,\dots,a_k \in S_k\}$ (the usual product of ideals). Otherwise, we define $S_1\cdot\ldots\cdot S_k$ to be just the elementwise product.

If all the sets $S_1,\dots,S_k$ are two-sided δ-ideals, then $S_1\cdot\ldots\cdot S_k$ is also a two-sided δ-ideal. Obviously, we can replace the word "two-sided" by "left" or "right". Hence any power of a δ-ideal is also a δ-ideal.

Notice, finally, that if $I,J \subseteq R$ are two-sided δ-ideals, then $I+J = \{a+b : a \in I, b \in J\}$ is also a two-sided δ-ideal. (We can replace "two-sided" by "left" or "right").

3. δ-prime rings and δ-prime ideals

We start with the following quite standard definition.

Definition 3.1. *The ring R is said to be δ-prime if it is nonzero and for any two-sided δ-ideals $I,J \subseteq R$ such that $IJ = \{0\}$, we have either $I = \{0\}$ or $J = \{0\}$.*

Notice that if δ is the zero derivation, then the δ-primeness is the same thing as the usual primeness of R (see, for instance, [6, Ch. 3]). Moreover, the ring R is prime if and only if it is d-prime for each derivation $d : R \longrightarrow R$. Let us now state and prove a fundamental characterization of δ-prime rings (cf. [1, Lemma 2]).

Theorem 3.2. *Suppose that R is a nonzero ring. The following conditions are equivalent:*

(i) R *is* δ*-prime,*

(ii) *for any elements* $a,b \in R$, *if* $\forall n \in \mathbb{N}\cup\{0\} : aR\delta^n(b) = \{0\}$, *then either* $a = 0$ *or* $b = 0$,

(iii) *for any elements* $a,b \in R$, *if* $\forall n \in \mathbb{N}\cup\{0\} : \delta^n(a)Rb = \{0\}$, *then either* $a = 0$ *or* $b = 0$,

(iv) *for any elements* $a,b \in R$, *if* $\langle a\rangle^\delta \langle b\rangle^\delta = \{0\}$, *then either* $a = 0$ *or* $b = 0$,

(v) *for any right* δ*-ideals* $I,J \subseteq R$, *if* $IJ = \{0\}$, *then either* $I = \{0\}$ *or* $J = \{0\}$,

(vi) *for any left* δ*-ideals* $I,J \subseteq R$, *if* $IJ = \{0\}$, *then either* $I = \{0\}$ *or* $J = \{0\}$,

(vii) *for an arbitrary nonzero right* δ*-ideal* $I \subseteq R$ *we have* $\operatorname{ann}_r(I) = \{0\}$,

(viii) *for an arbitrary nonzero left* δ*-ideal* $I \subseteq R$ *we have* $\operatorname{ann}_\ell(I) = \{0\}$.

Proof. Assume that R is δ-prime. Let $a,b \in R$ be such that

$$\forall n \in \mathbb{N}\cup\{0\} : aR\delta^n(b) = \{0\}. \tag{1}$$

Define I to be the totality of finite sums of elements of the set $\{c_1\delta^m(a)c_2 : c_1,c_2 \in R, m \in \mathbb{N}\cup\{0\}\}$. Furthermore, define J to be the totality of finite sums of elements of the set $\{h_1\delta^n(b)h_2 : h_1,h_2 \in R, n \in \mathbb{N}\cup\{0\}\}$. Then I and J are two-sided δ-ideals of the ring R.

Next, we will prove by induction on m that

$$\forall m,n \in \mathbb{N}\cup\{0\} : \delta^m(a)R\delta^n(b) = \{0\}.$$

If $m = 0$, then the assertion coincides with (1). Pick therefore some $k \in \mathbb{N}\cup\{0\}$ and suppose that

$$\forall n \in \mathbb{N}\cup\{0\} : \delta^k(a)R\delta^n(b) = \{0\}.$$

If $c \in R$ and $n \in \mathbb{N}\cup\{0\}$, then the induction hypothesis yields

$$0 = \delta(\delta^k(a)c\delta^n(b)) = \delta^{k+1}(a)c\delta^n(b) + \delta^k(a)\delta(c)\delta^n(b) + \delta^k(a)c\delta^{n+1}(b) =$$

$$= \delta^{k+1}(a)c\delta^n(b).$$

In this way, we have proved that $\delta^{k+1}(a)R\delta^n(b) = \{0\}$ for all $n \in \mathbb{N}\cup\{0\}$. The induction step is complete.

Pick arbitrary $m,n \in \mathbb{N}\cup\{0\}$. Since $\delta^m(a)R\delta^n(b) = \{0\}$, we get

$$(R\delta^m(a)R)(R\delta^n(b)R) \subseteq R(\delta^m(a)R\delta^n(b))R = \{0\}$$

(the products above are elementwise products of sets). Consequently, $IJ = \{0\}$. The δ-primeness therefore implies that either $I = \{0\}$ or $J = \{0\}$. It is easy to verify that $(\langle a\rangle^\delta)^3 \subseteq I$ and $(\langle b\rangle^\delta)^3 \subseteq J$ (see Proposition 2.6). Thus we have either $(\langle a\rangle^\delta)^3 = \{0\}$ or $(\langle b\rangle^\delta)^3 = \{0\}$. Since the square of a two-sided δ-ideal is also a two-sided δ-ideal, the δ-primeness yields

that either $\langle a\rangle^{\delta}=\{0\}$ or $\langle b\rangle^{\delta}=\{0\}$. This means, finally, that either $a=0$ or $b=0$. Condition (ii) follows. The implication (i) $\Longrightarrow$ (iii) can be proved analogously.

Assume that condition (ii) is satisfied. Let $a,b\in R$ be such that $\langle a\rangle^{\delta}\langle b\rangle^{\delta}=\{0\}$. Observe that for an arbitrary $n\in\mathbb{N}\cup\{0\}$, we have $aR\delta^{n}(b)\subseteq\langle a\rangle^{\delta}\langle b\rangle^{\delta}$. Hence (ii) implies that either $a=0$ or $b=0$. Condition (iv) follows. The implication (iii) $\Longrightarrow$ (iv) can be proved analogously.

Assume now that condition (iv) is satisfied. Let $I,J\subseteq R$ be right δ-ideals such that $IJ=\{0\}$. Suppose that $I\neq\{0\}$ and pick some $a\in I\setminus\{0\}$. Let $b\in J$. It is quite easy to verify that

$$\langle a\rangle^{\delta}\langle b\rangle^{\delta}\subseteq IJ+RIJ=\{0\}.$$

Condition (iv) therefore yields $b=0$. In this way, we have proved that $J=\{0\}$. Condition (v) follows. The implication (iv) $\Longrightarrow$ (vi) can be proved analogously.

It is clear that any of conditions (v) and (vi) implies the δ-primeness of the ring R. We have thus proved that conditions (i)–(vi) are pairwise equivalent.

Assume that condition (vi) is satisfied. Let $I\subseteq R$ be a nonzero left δ-ideal. Since $\operatorname{ann}_{\ell}(I)$ is a left δ-ideal and $\operatorname{ann}_{\ell}(I)I=\{0\}$, condition (vi) yields that $\operatorname{ann}_{\ell}(I)=\{0\}$. Condition (viii) follows. The implication (v) $\Longrightarrow$ (vii) can be proved analogously.

Assume, finally, that condition (viii) is satisfied. Let $I,J\subseteq R$ be two-sided δ-ideals such that $IJ=\{0\}$. Suppose that $J\neq\{0\}$. Since $I\subseteq\operatorname{ann}_{\ell}(J)$, condition (viii) implies that $I\subseteq\operatorname{ann}_{\ell}(J)=\{0\}$. The δ-primeness of the ring R follows. The implication (vii) $\Longrightarrow$ (i) can be proved analogously. The proof of the theorem is complete. □

Let us remark that if R is a ring with identity, then the totality I of finite sums of elements of the set $\{c_1\delta^{m}(a)c_2 : c_1,c_2\in R, m\in\mathbb{N}\cup\{0\}\}$, considered in the above proof, is the same thing as $\langle a\rangle^{\delta}$. But in the case where R is a ring without identity, it may happen that $a\notin I$.

Recall that if I is a two-sided δ-ideal of the ring R, then

$$\delta_I : R/I\ni a+I\longmapsto\delta(a)+I\in R/I$$

is a well-defined derivation.

Definition 3.3. *A two-sided δ-ideal $P\subseteq R$ is said to be δ-prime, if R/P is a δ_P-prime ring.*

Obviously, each δ-prime ideal is a proper ideal. It is worth noting that the ring R is δ-prime if and only if $\{0\}$ is a δ-prime ideal of R. The corollary below follows quite directly from Theorem 3.2.

Corollary 3.4. *Let P be a proper two-sided δ-ideal of the ring R. The following conditions are equivalent:*

(i) *P is δ-prime,*
(ii) *for arbitrary two-sided δ-ideals $I,J \subseteq R$, if $IJ \subseteq P$, then either $I \subseteq P$ or $J \subseteq P$,*
(iii) *for any elements $a,b \in R$, if $\forall n \in \mathbb{N}\cup\{0\} : aR\delta^n(b) \subseteq P$, then either $a \in P$ or $b \in P$,*
(iv) *for any elements $a,b \in R$, if $\forall n \in \mathbb{N}\cup\{0\} : \delta^n(a)Rb \subseteq P$, then either $a \in P$ or $b \in P$,*
(v) *for any elements $a,b \in R$, if $\langle a\rangle^{\delta}\langle b\rangle^{\delta} \subseteq P$, then either $a \in P$ or $b \in P$,*
(vi) *for arbitrary right δ-ideals $I,J \subseteq R$, if $IJ \subseteq P$, then either $I \subseteq P$ or $J \subseteq P$,*
(vii) *for arbitrary left δ-ideals $I,J \subseteq R$, if $IJ \subseteq P$, then either $I \subseteq P$ or $J \subseteq P$.*

Again, if δ is the zero derivation, then the notion of a δ-prime ideal coincides with the well-known general ("noncommutative") notion of a prime ideal. We are ready to discuss an example of a δ-prime ring which is not prime (the example is taken from [5]).

Example 3.5. *Let $\mathbb{F}$ be a field of characteristic $p \neq 0$. Consider the principal ideal P of the polynomial ring $\mathbb{F}[x]$ generated by x^p. Since $R = \mathbb{F}[x]/P$ is a commutative ring and $x+P$ is a nonzero nilpotent element of R, the ring R is not prime. (Let us recall here that a commutative ring with identity is prime if and only if it is an integral domain). Using condition (iii) of Corollary 3.4, we can prove quite easily that P is a δ-prime ideal for the natural derivation $\delta : \mathbb{F}[x] \ni f \longmapsto f' \in \mathbb{F}[x]$. Thus R is δ_P-prime.*

In the sequel we will deal with the following generalization of the prime radical. This generalization has been introduced by Burkov (see [2]).

Definition 3.6. *The intersection $\mathrm{N}_{\delta}(R)$ of the family of all δ-prime ideals of the ring R is called the δ-prime radical of R.*

Notice that $\mathrm{N}_{\delta}(R) = R$ whenever R has no δ-prime ideals.

4. δ-nilpotent elements

Consider the family

$$\mathscr{D} = \left\{ \sum_{j=0}^{n} c_j \delta^j : n \in \mathbb{N}\cup\{0\}, c_0, \ldots, c_n \in R \right\}$$

of "differential operators on the ring R".

Remark 4.1. *If $D \in \mathscr{D}$ and I is a left δ-ideal of R, then $D(I) \subseteq I$.*

The definition below is taken from [2].

Definition 4.2. *An element $a \in R$ is said to be δ-nilpotent, if for any sequence $\{D_k\}_{k=0}^{\infty}$ of elements of $\mathscr{D}$ almost all members of the sequence $\{a_k\}_{k=0}^{\infty}$ defined by*

$$\begin{cases} a_0 = a, \\ a_{k+1} = a_k D_k(a_k) \end{cases}$$

are equal to 0.

Let us also recall the well-known concept of a strongly nilpotent element.

Definition 4.3. *An element $a \in R$ is said to be strongly nilpotent, if almost all members of any sequence $\{a_k\}_{k=0}^{\infty}$ in the ring R such that $a_0 = a$ and*

$$\forall k \in \mathbb{N} \cup \{0\} : a_{k+1} \in a_k R a_k$$

are equal to 0.

Observe that an element $a \in R$ is strongly nilpotent if and only if for an arbitrary sequence $\{x_k\}_{k=0}^{\infty}$ of elements of R, almost all members of the sequence $\{a_k\}_{k=0}^{\infty}$ defined by

$$\begin{cases} a_0 = a, \\ a_{k+1} = a_k x_k a_k \end{cases}$$

are equal to 0.

It is clear that in the definitions of a δ-nilpotent element and a strongly nilpotent element (as well as in the equivalent definition of a δ-nilpotent element given in the sequel of this section), the words "almost all members of the sequence $\{a_k\}_{k=0}^{\infty}$ are equal to 0" can be replaced by "the sequence $\{a_k\}_{k=0}^{\infty}$ contains a member equal to 0". Let us now take a look at some simple but important properties.

Proposition 4.4. *For an element $a \in R$ the following hold true:*

(i) *if a is δ-nilpotent, then it is strongly nilpotent,*

(ii) *if a is strongly nilpotent, then it is nilpotent in the usual sense,*

(iii) *if a is nilpotent in the usual sense, $a \in \mathrm{Z}(R)$ and $\delta(a) = 0$, then a is δ-nilpotent,*

(iv) *if a is nilpotent in the usual sense and $a \in \mathrm{Z}(R)$, then a is strongly nilpotent,*

(v) *if δ is the zero derivation and a is strongly nilpotent, then a is δ-nilpotent.*

Proof. Assume that a is δ-nilpotent. Pick an arbitrary sequence $\{x_k\}_{k=0}^{\infty}$ in the ring R. Since

$$\forall k \in \mathbb{N} \cup \{0\} \forall b \in R : \begin{cases} bx_k b = bx_k \delta^0(b), \\ x_k \delta^0 \in \mathscr{D}, \end{cases}$$

the δ-nilpotency implies that almost all members of the sequence $\{a_k\}_{k=0}^{\infty}$ defined by

$$\begin{cases} a_0 = a, \\ a_{k+1} = a_k x_k a_k \end{cases}$$

are equal to 0. Therefore, a is strongly nilpotent.

If a is a strongly nilpotent element, then almost all members of the sequence $\{a_k\}_{k=0}^{\infty}$ of powers of a defined by

$$\begin{cases} a_0 = a, \\ a_{k+1} = a_k a a_k \end{cases}$$

are equal to 0 and hence a is nilpotent in the usual sense.

Let us turn to property (iii). It is easy to see that if $\delta(z) = 0$ for some $z \in R$, then

$$\forall t \in \mathbb{N} \setminus \{0\} \, \forall j \in \mathbb{N} \cup \{0\} \, \forall b \in R : \begin{cases} \delta(z^t) = 0, \\ \delta^j(z^t b) = z^t \delta^j(b). \end{cases} \tag{2}$$

Assume that a is nilpotent in the usual sense, $a \in \mathrm{Z}(R)$ and $\delta(a) = 0$. Let $\{D_k\}_{k=0}^{\infty}$, where

$$D_k = \sum_{j=0}^{n_k} c_{jk} \delta^j$$

for some $n_k \in \mathbb{N} \cup \{0\}$ and some $c_{0k}, \ldots, c_{n_k k} \in R$, be a sequence of elements of the family $\mathscr{D}$. Consider the sequence $\{a_k\}_{k=0}^{\infty}$ defined by

$$\begin{cases} a_0 = a, \\ a_{k+1} = a_k D_k(a_k). \end{cases}$$

We will show by induction that $a_k \in a^{2^k} R$ for an arbitrary $k \in \mathbb{N} \setminus \{0\}$. First, since $\delta(a) = 0$ and $a \in \mathrm{Z}(R)$, we have

$$a_1 = a D_0(a) = a \sum_{j=0}^{n_0} c_{j0} \delta^j(a) = a^2 c_{00}.$$

Suppose therefore that $a_\ell = a^{2^\ell} b$ for some $\ell \in \mathbb{N} \setminus \{0\}$ and some $b \in R$. In view of (2) and the fact that $a \in \mathrm{Z}(R)$, we obtain

$$a_{\ell+1} = a_\ell D_\ell(a_\ell) = a^{2^\ell} b D_\ell(a^{2^\ell} b) = a^{2^\ell} b \sum_{j=0}^{n_\ell} c_{j\ell} \delta^j(a^{2^\ell} b) =$$

$$= a^{2^\ell} b \sum_{j=0}^{n_\ell} c_{j\ell} a^{2^\ell} \delta^j(b) = a^{2^{\ell+1}} b \sum_{j=0}^{n_\ell} c_{j\ell} \delta^j(b).$$

The induction step is complete. Now, let $s \in \mathbb{N} \setminus \{0\}$ be such that $a^s = 0$ ("usual nilpotency" of a). Observe that if $k \in \mathbb{N} \setminus \{0\}$ satisfies the condition $2^k \geqslant s$, then $a_k \in a^{2^k} R \subseteq a^s R = \{0\}$. The δ-nilpotency of a follows.

Let us turn to (iv). Assume that a is nilpotent in the usual sense and $a \in \mathrm{Z}(R)$. Suppose additionally that δ is the zero derivation. Then property (iii) yields that a is δ-nilpotent. It therefore follows from (i) that the element a is strongly nilpotent.

Property (v) is an immediate consequence of the fact that if δ is the zero derivation, then $\mathscr{D} = \{c \cdot \mathrm{id}_R : c \in R\}$ (and hence the definition of a δ-nilpotent element reduces to the definition of a strongly nilpotent element). □

Notice that in the case where R is a commutative ring and δ is the zero derivation, the usual nilpotency, the strong nilpotency and the δ-nilpotency of an element are the same thing. Let us see an example of a strongly nilpotent element which is not δ-nilpotent.

Example 4.5. *With the assumptions and notations of Example 3.5, we have* $(x+P)\delta_P(x+P) = x+P$. *Hence all members of the sequence* $\{a_k\}_{k=0}^{\infty}$ *defined by*

$$\begin{cases} a_0 = x+P, \\ a_{k+1} = a_k \delta_P(a_k) \end{cases}$$

are nonzero. This yields that $x+P$ *is not a* δ_P*-nilpotent element of the ring* R. *On the other hand,* $x+P$ *is a strongly nilpotent element, because it is nilpotent in the usual sense and* R *is a commutative ring.*

The main theorem of the section is a modification of a result which has been first stated in [2].

Theorem 4.6. *Let* $a \in R$. *The following conditions are equivalent:*

(i) a *is* δ*-nilpotent,*

(ii) *for arbitrary sequences* $\{c_k\}_{k=0}^{\infty}$ *of elements of* R *and* $\{n_k\}_{k=0}^{\infty}$ *of non-negative integers, almost all members of the sequence* $\{a_k\}_{k=0}^{\infty}$ *defined by*

$$\begin{cases} a_0 = a, \\ a_{k+1} = a_k c_k \delta^{n_k}(a_k) \end{cases}$$

are equal to 0,

(iii) $a \in \mathrm{N}_{\delta}(R)$.

Proof. The implication (i) $\Longrightarrow$ (ii) is obvious (see the definition of the family $\mathscr{D}$).

Suppose that $a \in R \setminus \mathrm{N}_{\delta}(R)$. Then $a \notin P$ for some δ-prime ideal P of the ring R. Hence by condition (iii) of Corollary 3.4, there exist sequences $\{c_k\}_{k=0}^{\infty}$ of elements of R and $\{n_k\}_{k=0}^{\infty}$ of non-negative integers such that no member of the sequence $\{a_k\}_{k=0}^{\infty}$ defined by

$$\begin{cases} a_0 = a, \\ a_{k+1} = a_k c_k \delta^{n_k}(a_k) \end{cases}$$

belongs to P. It follows that $a_k \neq 0$ for all $k \in \mathbb{N} \cup \{0\}$. Therefore, condition (ii) is not satisfied. This completes the proof of the implication (ii) $\Longrightarrow$ (iii).

Now suppose that the element a is not δ-nilpotent. Then there exists a sequence $\{D_k\}_{k=0}^{\infty}$ of elements of $\mathscr{D}$ such that all members of the sequence $\{a_k\}_{k=0}^{\infty}$ defined by

$$\begin{cases} a_0 = a, \\ a_{k+1} = a_k D_k(a_k) \end{cases}$$

are different from 0. Consider the family $\mathfrak{F}$ of all two-sided δ-ideals $I \subseteq R$ with the property that $I \cap \{a_k : k \in \mathbb{N} \cup \{0\}\} = \emptyset$. Notice that $\{0\} \in \mathfrak{F}$. The family $\mathfrak{F}$ (partially) ordered by set inclusion satisfies the assumption of Zorn's lemma. Pick a maximal element $P_0 \in \mathfrak{F}$. Let us emphasize that P_0 is a proper two-sided δ-ideal of the ring R.

Let $J,K \subseteq R$ be two-sided δ-ideals such that $JK \subseteq P_0$. Assume that neither J nor K is contained in P_0. Since $P_0 \subseteq (P_0+J)\cap(P_0+K)$, $P_0 \neq P_0+J$ and $P_0 \neq P_0+K$, the maximality of P_0 implies that $P_0+J \notin \mathfrak{F}$ and $P_0+K \notin \mathfrak{F}$. But P_0+J and P_0+K are two-sided δ-ideals of the ring R. Hence there are $s,t \in \mathbb{N}\cup\{0\}$ such that $a_s \in P_0+J$ and $a_t \in P_0+K$. Let us define $u = \max\{s,t\}$. If $T \subseteq R$ is a right ideal, $x \in T$ and $D \in \mathscr{D}$, then obviously $xD(x) \in T$. It follows therefore from the definition of $\{a_k\}_{k=0}^{\infty}$ that $a_u \in (P_0+J)\cap(P_0+K)$. Next, observe that if $x \in P_0+J$, $y \in P_0+K$ and $D \in \mathscr{D}$, then by Remark 4.1 we have

$$xD(y) \in (P_0+J)(P_0+K) \subseteq P_0+JK = P_0.$$

Since $a_u \in (P_0+J)\cap(P_0+K)$, the observation yields $a_{u+1} = a_u D_u(a_u) \in P_0$. This contradicts the fact that $P_0 \in \mathfrak{F}$.

We have therefore proved that for any two-sided δ-ideals $J,K \subseteq R$, if $JK \subseteq P_0$, then either $J \subseteq P_0$ or $K \subseteq P_0$. In other words, P_0 is a δ-prime ideal of the ring R. Since $a = a_0 \notin P_0$, we get $a \notin \mathrm{N}_\delta(R)$. The proof of the implication (iii) $\Longrightarrow$ (i) is complete. □

It follows immediately from the above theorem that $\mathrm{N}_\delta(R)$ coincides with the totality of δ-nilpotent elements of the ring R. The theorem also allows us to give an equivalent definition of a δ-nilpotent element (namely, an element $a \in R$ is δ-nilpotent if and only if for arbitrary sequences $\{c_k\}_{k=0}^{\infty}$ of elements of R and $\{n_k\}_{k=0}^{\infty}$ of non-negative integers, almost all members of the sequence $\{a_k\}_{k=0}^{\infty}$ defined by

$$\begin{cases} a_0 = a, \\ a_{k+1} = a_k c_k \delta^{n_k}(a_k) \end{cases}$$

are equal to 0).

Recall that a set $S \subseteq R$ is said to be nil, if every element of S is nilpotent in the usual sense. Combining Theorem 4.6 with Proposition 4.4 yields a noteworthy corollary.

Corollary 4.7. *The δ-prime radical $\mathrm{N}_\delta(R)$ is a nil two-sided δ-ideal of the ring R.*

Let us finally notice that if δ is the zero derivation, then $\mathrm{N}_\delta(R)$ and the standard prime radical $\mathrm{rad}(R)$ are the same thing. In view of Theorem 4.6 and Proposition 4.4, we thus obtain the following classical fact.

Corollary 4.8. *The prime radical $\mathrm{rad}(R)$ coincides with the totality of strongly nilpotent elements of R.*

5. δ-semiprime rings

We will use the following definition of a δ-semiprime ring.

Definition 5.1. *The ring R is called δ-semiprime, if there exists no two-sided δ-ideal $I \subseteq R$ such that $I \neq \{0\}$ and $I^2 = \{0\}$.*

Obviously, each δ-prime ring is δ-semiprime. Recall that an ideal I of the ring R is said to be nilpotent, if $I^k = \{0\}$ for some $k \in \mathbb{N}\setminus\{0\}$. We are in a position to state and prove a fundamental characterization of δ-semiprime rings (cf. [1, Lemma 1]).

Theorem 5.2. *The following conditions are equivalent:*

(i) R *is a* δ*-semiprime ring,*

(ii) *for any element* $a \in R$, *if* $\forall\, n \in \mathbb{N} \cup \{0\} : aR\delta^n(a) = \{0\}$, *then* $a = 0$,

(iii) *for any element* $a \in R$, *if* $\forall\, n \in \mathbb{N} \cup \{0\} : \delta^n(a)Ra = \{0\}$, *then* $a = 0$,

(iv) *for any element* $a \in R$, *if* $(\langle a \rangle^{\delta})^2 = \{0\}$, *then* $a = 0$,

(v) *for an arbitrary right* δ*-ideal* $I \subseteq R$, *if* $I^2 = \{0\}$, *then* $I = \{0\}$,

(vi) *for an arbitrary left* δ*-ideal* $I \subseteq R$, *if* $I^2 = \{0\}$, *then* $I = \{0\}$,

(vii) $\{0\}$ *is the only nilpotent two-sided* δ*-ideal of the ring* R,

(viii) $\{0\}$ *is the only nilpotent right* δ*-ideal of the ring* R,

(ix) $\{0\}$ *is the only nilpotent left* δ*-ideal of the ring* R,

(x) *for any two-sided* δ*-ideals* $I, J \subseteq R$, *if* $IJ = \{0\}$, *then* $I \cap J = \{0\}$,

(xi) *for any right* δ*-ideals* $I, J \subseteq R$, *if* $IJ = \{0\}$, *then* $I \cap J = \{0\}$,

(xii) *for any left* δ*-ideals* $I, J \subseteq R$, *if* $IJ = \{0\}$, *then* $I \cap J = \{0\}$,

(xiii) R *has no nonzero* δ*-nilpotent elements,*

(xiv) $\mathrm{N}_{\delta}(R) = \{0\}$.

Proof. The equivalence of conditions (i)–(vi) can be proved analogously as in Theorem 3.2.

Assume that R is a δ-semiprime ring. Let $I \subseteq R$ be a nilpotent two-sided δ-ideal. Define $k_0 = \min\{k \in \mathbb{N} \setminus \{0\} : I^k = \{0\}\}$ (in other words, k_0 is the nilpotency index of I). Let $s \in \{0, 1\}$ be such that $k_0 + s$ is even. Then $(I^t)^2 = \{0\}$, where $t = (k_0 + s)/2$. The δ-semiprimeness now implies that $I^t = \{0\}$. Thus $k_0 \leqslant t$. The inequality is equivalent to $k_0 \leqslant s$. Therefore, $k_0 = 1$ and hence $I = \{0\}$. Condition (vii) follows. The implications (v) $\Longrightarrow$ (viii) and (vi) $\Longrightarrow$ (ix) can be proved analogously.

The implications (vii) $\Longrightarrow$ (i), (viii) $\Longrightarrow$ (v) and (ix) $\Longrightarrow$ (vi) are obvious.

Assume again that R is a δ-semiprime ring. Let $I, J \subseteq R$ be two-sided δ-ideals such that $IJ = \{0\}$. Since $I \cap J$ is also a two-sided δ-ideal and $(I \cap J)^2 \subseteq IJ$, the δ-semiprimeness yields that $I \cap J = \{0\}$. Condition (x) follows. The implications (v) $\Longrightarrow$ (xi) and (vi) $\Longrightarrow$ (xii) can be proved analogously.

Assume that condition (x) is satisfied. Let $I \subseteq R$ be a two-sided δ-ideal such that $I^2 = \{0\}$. Then $I = I \cap I = \{0\}$. The δ-semiprimeness of the ring R follows. The implications (xi) $\Longrightarrow$ (v) and (xii) $\Longrightarrow$ (vi) can be proved analogously. Hence, we have proved that conditions (i)–(xii) are pairwise equivalent.

Now assume that condition (ii) is satisfied. Let $a \in R \setminus \{0\}$. Then there exist sequences $\{c_k\}_{k=0}^{\infty}$ of elements of the ring R and $\{n_k\}_{k=0}^{\infty}$ of non-negative integers such that every member of the sequence $\{a_k\}_{k=0}^{\infty}$ defined by

$$\begin{cases} a_0 = a, \\ a_{k+1} = a_k c_k \delta^{n_k}(a_k) \end{cases}$$

is different from 0. Consequently, the element a is not δ-nilpotent (cf. the proof of Theorem 4.6). Condition (xiii) follows.

The equivalence (xiii) $\Longleftrightarrow$ (xiv) follows immediately from Theorem 4.6.

Assume, finally, that $\mathrm{N}_{\delta}(R) = \{0\}$. Let $I, J \subseteq R$ be two-sided δ-ideals such that $IJ = \{0\}$. Moreover, let P be a δ-prime ideal of the ring R. Since $IJ \subseteq P$, we get either $I \subseteq P$ or $J \subseteq P$.

Hence $I \cap J \subseteq P$. We have therefore proved that $I \cap J$ is contained in any δ-prime ideal of R. This means exactly that $I \cap J \subseteq \mathrm{N}_\delta(R)$. Condition (x) follows. The proof is complete. □

As an obvious consequence of the above theorem, we obtain a quite important fact.

Corollary 5.3. *Suppose that the ring R is δ-semiprime. Let $I \subseteq R$ be a δ-ideal. Then*
(i) $I \cap \mathrm{ann}_r(I) = \{0\}$ *whenever I is a right ideal,*
(ii) $I \cap \mathrm{ann}_\ell(I) = \{0\}$ *whenever I is a left ideal.*

In the case where δ is the zero derivation, the definition of a δ-semiprime ring is just the well-known definition of a semiprime ring. Clearly, the ring R is semiprime if and only if it is d-semiprime for all derivations $d : R \longrightarrow R$. Notice that in fact, the ring R considered in Examples 3.5 and 4.5 is not semiprime (a commutative ring with identity is semiprime if and only if it has no nilpotent elements different from 0).

Though a δ-semiprime ring has no δ-nilpotent elements different from 0 and no nonzero nilpotent δ-ideals, it can have a nonzero nil δ-ideal. For an example we refer to [3, p. 332].

Finally, let us see how another important fact about semiprime rings generalizes to δ-semiprime rings (cf. [1, Lemma 5]).

Proposition 5.4. *Suppose that R is a δ-semiprime ring. Let $I \subseteq R$ be a two-sided δ-ideal. Then* $\mathrm{ann}_\ell(I) = \mathrm{ann}_r(I)$.

Proof. Define $K = \mathrm{ann}_r(I)I$ (product of right δ-ideals). We have

$$K^2 = (\mathrm{ann}_r(I)I)(\mathrm{ann}_r(I)I) \subseteq \mathrm{ann}_r(I)(I\mathrm{ann}_r(I))I = \{0\}$$

and hence, by the δ-semiprimeness, $\mathrm{ann}_r(I)I = K = \{0\}$. Therefore, $\mathrm{ann}_r(I) \subseteq \mathrm{ann}_\ell(I)$. The converse inclusion can be proved analogously. □

*T*he author would like to thank the anonymous referees for carefully reading the manuscript and giving a number of constructive comments which helped him to substantially improve the text.

References

[1] Artemovych O. D. & Lukashenko M. P., Lie and Jordan structures of differentially semiprime rings, *Algebra Discrete Math.* **20**, No. 1, 2015, 13–31.

[2] Burkov V. D., On derivationally prime rings, *Russ. Math. Surv.* **35**, No. 5, 1980, 253–254.

[3] Cohn P. M., *Further Algebra and Applications*, Springer, London, 2003.

[4] Herstein I. N., *Noncommutative Rings*, The Carus Mathematical Monographs, Mathematical Association of America, Washington, DC, 1996.

[5] Jordan C. R. & Jordan D. A., Lie rings of derivations of associative rings, *J. Lond. Math. Soc., II. Ser.* **17**, 1978, 33–41.

[6] Lambek J., *Lectures on Rings and Modules*, Blaisdell Publishing Company, 1966.

[7] McCoy N. H., *The Theory of Rings*, Chelsea Publishing Company, Bronx, New York, 1973.

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

ANNA MILIAN*

ON SOME VOLATILITY REDUCTION OF RETURNS ON SHARES

O REDUKCJI ZMIENNOŚCI STOPY ZWROTU Z AKCJI

A b s t r a c t

In this paper we consider derivatives which are binary options of asset-or-nothing type with a payoff function depending on a parameter. The payoff is modelled on the payoff of catastrophe bonds. We examine the influence of the derivative on returns on shares. For this purpose two portfolios are compared: one consisting of stocks and a second additionally containing the derivative. Using the Black-Scholes model we derive an explicit formula for the standard deviation of the returns on the investment portfolios. Numerical examples show that the derivative reduces the volatility of returns on shares. For typical values of stock price volatility we indicate the value of the parameter appearing in the payoff for which the volatility of returns on shares reaches a minimum. All numerical calculations were made with MAPLE.

Keywords: Black-Scholes model, *risk-reducing derivatives*, *MAPLE*

S t r e s z c z e n i e

W artykule rozważa się pewien pochodny instrument finansowy którego funkcja wypłaty jest wzorowana na funkcji wypłaty z obligacji katastroficznych. Analizuje się wpływ tego instrumentu na stopę zwrotu z akcji porównując portfel akcji z portfelem zawierającym dodatkowo rozważany instrument pochodny. Stosując model Blacka-Scholesa wyprowadza się dokładny wzór na odchylenie standardowe stóp zwrotu z każdego z tych portfeli. Analizowane przykłady pokazują, że rozważany instrument pochodny redukuje zmienność stóp zwrotu z akcji. Obliczenia do podanych przykładów zostały wykonane przy pomocy programu MAPLE.

Słowa kluczowe: model Blacka-Scholesa, *instrument pochodny redukujący ryzyko*, *MAPLE*

DOI: 10.4467/2353737XCT.16.145.5756

* Anna Milian (amilian@pk.edu.pl), Institute of Mathematics, Faculty of Physic, Mathematics and Computer Sciences, Cracow University of Technology.

1. Introduction

The subject of this paper is a derivative, considered in [4] as a risk-reducing derivative. The payment of the derivative depends on a parameter. Using Monte Carlo simulations, for each of the typical value of the volatility of stocks a variant of the derivative (a proper parameter in a payoff function) reducing the risk of a large loss by more than 10% on a confidence level of 95% was indicated.

In this paper we examine volatility of rate of return from stocks, when portfolio apart from stocks additionally includes a derivative. We obtain an analytical closed form formula for the volatility expressed as standard deviation of related, discounted percentage of profit from a portfolio. We show that the derivative reduces volatility of rate of return on stocks.

In this paper we use the Black-Scholes model with one risk-free asset and one risky instrument – a stock – regarded as the underlying. We consider the simplest case of the model which is based on the following assumptions: security trading is continuous, there are no riskless arbitrage opportunities, there are no transaction costs and no dividends during the life of a derivative, the risk-free rate of interest and the volatility of an underlying asset are constant. The annualized volatility of the stock, from now on called briefly volatility, is typically between 15% and 60% [6].

2. Model description

Let $\sigma > 0$ be a stock price volatility and r be the risk-free interest rate. We assume the price of the stock follows a geometric Brownian motion

$$S_t = S\exp\left(\left(r-\frac{1}{2}\sigma^2\right)t+\sigma W_t\right), \quad t\in[0,T] \tag{1}$$

where S is the stock price at time 0, $W=\{W_t, t\in[0,T]\}$ is a standard Brownian motion under the risk-neutral probability P and T is the expiry date. Let E^P denote the expectation operator under the P measure and let $\{\mathcal{F}_t\}$ be a filtration for Brownian motion W. Let us consider a financial derivative instrument dependent on parameter $a > 0$, with the following payoff function

$$f(S_T)=\begin{cases} S_T & if \quad S_T \le aS, \\ 0 & if \quad S_T > aS. \end{cases} \tag{2}$$

The instrument provides some protection against a decline in the stock price i.e. against the event $S_T \le aS$ and can be considered as an obligation transferring the risk from the holder of the derivative to the issuer [4]. We will analyse a portfolio composed of one stock and one derivative with payoff function (2). We will calculate the variance of the discounted profit from the portfolio. According to the volatility of the stock we will indicate value of a in the interval $[0,2]$ which minimizes the variance.

3. Volatilities of portfolios

In Black-Scholes model, today's arbitrage price of the derivative instrument expresses as the expected value of its discounted payoff function, taken with respect to the risk-neutral measure P [2]:

$$c = E^P(\exp(-rT)f(S_T)) \tag{3}$$

In [4] the following closed form formula for pricing the derivative was derived

$$c = SN\left(\frac{lna - \left(r + \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right) \tag{4}$$

where N denotes the cumulative probability distribution function for a standardized normal distribution. The formula can also be found in [2] and [5]. The today's price of considered stock equals S so the discounted gain from a portfolio is

$$(S_T + f(S_T))\exp(-rT) - (S + c)$$

and the related, discounted percentage of profit from the portfolio equals

$$R = \frac{(S_T + f(S_T))\exp(-rT) - (S + c)}{S + c} \cdot 100\%. \tag{5}$$

To calculate standard deviation of R let us first denote:

- Φ – cumulative probability distribution function of σW_T,
- ϕ – probability density function of σW_T,
- F – cumulative probability distribution function of S_T,
- f – probability density function of S_T,

$$k = S\exp\left[\left(r - \frac{1}{2}\sigma^2\right)T\right]. \tag{6}$$

By it follows that $S_T = k\exp(\sigma W_T)$ and consequently

$$f(x) = \frac{1}{x}\varphi\left(\ln\frac{x}{k}\right) \quad \text{for} \quad x > 0$$

and

$$f(x) = 0 \quad \text{for} \quad x \le 0 \tag{7}$$

where

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi T}}\exp\left(-\frac{x^2}{2\sigma^2 T}\right).$$

Hence

$$D^2(R) = \left(\frac{100}{S+c}\right)^2 D^2[(S_T + f(S_T))\exp(-rT) - (S+c)] =$$

$$= \left(\frac{100}{S+c}\right)^2 \exp(-2rT) D^2(S_T + f(S_T)).$$

Since

$$f(S_T) = S_T 1_{\{S_T \le aS\}}$$

it follows that

$$(S_T + f(S_T))^2 = S_T^2 + 3S_T^2 1_{\{S_T \le aS\}}.$$

By (3) we have $E^P(f(S_T)) = e^{rT}c$. The process $\exp(-rT)S_t$, $t \ge 0$ is a martingale which implies

$$E^P(S_T) = \exp(rT)S. \tag{8}$$

Hence

$$E^P(S_T + f_T) = \exp(rT)(S+c) \tag{9}$$

and variance of R expresses as follows:

$$D^2(R) = \left(\frac{100}{S+c}\right)^2 \left[e^{-2rT}(E^P(S_T^2 + 3S_T^2 1_{\{S_T \le aS\}}) - (S+c)^2)\right] \tag{10}$$

Using (7) we calculate $E^P(S_T^2)$ as $\int_0^\infty x\varphi\left(\ln\left(\frac{x}{k}\right)\right)dx$.

Substituting $\ln\left(\frac{x}{k}\right) = t$ we have

$$E^P(S_T^2) = \int_{-\infty}^{\infty} k^2 \exp(2t)\varphi(t)dt = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi T}} \exp\left(2t - \frac{t^2}{2\sigma^2 T}\right)dt.$$

But

$$2t - \frac{t^2}{2\sigma^2 T} = 2T\sigma^2 - \left(\frac{t}{\sigma\sqrt{2T}} - \sigma\sqrt{2T}\right)^2$$

and consequently

$$E^P(S_T^2) = \frac{k^2 \exp(2T\sigma^2)}{\sigma\sqrt{2\pi T}} \int_{-\infty}^{\infty} \exp\left(-\left(\frac{t}{\sigma\sqrt{2T}} - \sigma\sqrt{2T}\right)^2\right)dt.$$

Taking into account (6) and substituting $\frac{t}{\sigma\sqrt{2T}} - \sigma\sqrt{2T} = \frac{u}{\sqrt{2}}$ we have

$$E^P(S_T^2) = S^2 \frac{\exp(2r+\sigma^2)T}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}u^2\right) du = S^2 \exp[(2r+\sigma^2)T]. \tag{11}$$

Similarly, we calculate $E^P(S_T^2 1_{(S_T \le aS)})$. Namely, with the change of variables we have

$$E^P(S_T^2 1_{(S_T \le aS)}) = \int_{-\infty}^{aS} x^2 f(x)dx.$$

Using substitutions as above, i.e. $\ln\left(\frac{x}{k}\right) = t$ and $\frac{t}{\sigma\sqrt{2T}} - \sigma\sqrt{2T} = \frac{u}{\sqrt{2}}$ and we obtain

$$\begin{aligned}
E^P(S_T^2 1_{(S_T \le aS)}) &= \int_0^{aS} x\varphi\left(\ln\left(\frac{x}{k}\right)\right)dx \\
&= \int_{-\infty}^{A} k^2 \frac{1}{\sigma\sqrt{2\pi T}} \exp\left(2t - \frac{t^2}{2\sigma^2 T}\right)dt \\
&= S^2 \frac{\exp((2r+\sigma^2)T)}{\sigma\sqrt{2\pi T}} \int_{-\infty}^{A} \exp\left(-\left(\frac{t}{\sigma\sqrt{2T}} - \sigma\sqrt{2T}\right)^2\right)dt \\
&= S^2 \frac{\exp((2r+\sigma^2)T)}{\sqrt{2\pi}} \int_{-\infty}^{B} \exp\left(-\frac{1}{2}u^2\right)du \\
&= S^2 \exp((2r+\sigma^2)T)N(B)
\end{aligned}$$

where

$$A = \ln\left(\frac{aS}{k}\right) \quad \text{and} \quad B = \sqrt{2}\left(\frac{\mathrm{A}}{\sigma\sqrt{2T}} - \sigma\sqrt{2T}\right).$$

Using (6) we obtain

$$E^P(S_T^2 1_{(S_T \le aS)}) = S^2 \exp((2r+\sigma^2)T)N\left(\frac{\ln a - \left(r + \frac{3}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right) \tag{12}$$

Finally, substituting (4), (11) and (12) into (10) we obtain

$$D^2(R)=10^4\left[\frac{\exp(\sigma^2T)\left[1+3N\left(\frac{\ln a-\left(r+\frac{3}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right)\right]}{\left(1+N\left(\frac{lna-\left(r+\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right)\right)^2}-1\right] \tag{13}$$

To examine the impact of the derivative, defined by (2), on the rate of return on investment in shares, we are going to compare the above variance with variance of analogous rate of return from a portfolio composed of a stock only. Namely, let S be the today's price of considered stock. Then, the discounted gain from a portfolio is

$$S_T\exp(-rT)-S$$

and the related, discounted percentage of profit from the portfolio equals

$$Z=\frac{S_T\exp(-rT)-S}{S}\cdot 100\%. \tag{14}$$

Using (8) and (11) we obtain

$$D^2Z=10^4(\exp(\sigma^2T)-1). \tag{15}$$

4. Comparison of volatility of *R* and *Z* with MAPLE

In this section we compare two portfolios, one composed of one stock with value $S = 1$ at time 0 and with one derivative with payoff (2), at price c, given by (4).

The second portfolio is composed of one stock with value $S = 0$ at time 0 only. As in the previous section, R and Z denote the related, discounted percentages of profit from the portfolios, respectively. We calculate standard deviations of R and Z using MAPLE. Let us consider standard deviation of R as a function of parameter a.

In the screenshot presented below, due to the requirements of MAPLE, standard deviation of R is denoted as σR and F denotes the cumulative probability distribution function for a standardized normal distribution:

$$\sigma R := a \to 100 \cdot \mathrm{sqrt}\left(\exp(\sigma^2 T) \cdot \frac{1 + 3 \cdot F\left(\frac{\ln(a) - \left(r + \frac{3}{2}\sigma^2 \right) \cdot T}{(\sigma \cdot \mathrm{sqrt}(T))} \right)}{1 + F\left(\frac{\left(\ln(a) - \left(r + \frac{1}{2}\sigma^2 \right) \cdot T \right)}{(\sigma \cdot \mathrm{sqrt}(T))} \right)^2} - 1 \right).$$

Let $\sigma_{\min}(R)$ denote the minimum of σR, considered as a function of parameter $a \in \left[\frac{1}{10}, 2 \right]$ and let a^* be the value of the parameter for which the function takes the minimum value.

We obtain $\sigma_{\min}(R)$ and a^* using command of MAPLE:

$$NLPSolve(\sigma R(a),\ a = 1/10..2).$$

In Table 1 one can see dependence of a^* and $\sigma_{\min}(R)$ from σ.

For every σ appearing in the table, a^* and $\sigma_{\min}(R)$ take the same values, independently of $r \in \{1\%, 2\%, 3\%, 4\%, 5\%, 6\%\}$.

Table 1

σ [%]	10	20	30	40	50	60	70	80	90
a^*	1.36	1.31	1.03	1.02	1.07	1.14	1.25	1.39	1.59
$\sigma_{\min}(R)$ [%]	10	17.99	20.9	24.43	31.25	40.03	50.18	61.53	74.16

Standard deviation $\sigma(Z)$ of does not depend on the risk-free interest rate r but it does on stock price volatility σ.

In Table 2, we present values of $\sigma(Z)$ depending on stock price volatility σ. As we can see, the stock price volatility σ and standard deviation $\sigma(Z)$ of Z are approximately equal (both are expressed in percentage):

Table 2

σ [%]	10	20	30	40	50	60	70	80	90
$\sigma(Z)$ [%]	10.025	20.202	30.688	41.655	53.294	65.828	79.518	94.68	111.71

Example

We present a sample screenshot with the calculations for the following parameters:

> X:= *Random Variable* (*Normal*(0, 1)) :
> $F(x)$:= *CDF* (X,x) :
> σ := 0.3 :
> T := 1 :
> σZ := 100·sqrt(exp ($\sigma^2 T$) – 1);

30.6878288

still assuming that $\sigma = 0.3$, $T = 1$. As you can see in the screenshot below, standard deviation of R, considered as the function σR on interval $\left[\frac{1}{10}, 2\right]$, achieves minimum equal to 20.9011… for the argument $a = 1.05528…$:

with(*Optimization*):

$$NLPSolve\left(\sigma R(a),\ a = \frac{1}{10}..2\right);$$

[20.9011145934436868, $a = 1.05528207323025392$]

The same can be seen in a graph of function σR:

$$> plot\left(\sigma R(a),\ a = \frac{1}{10}..2\right)$$

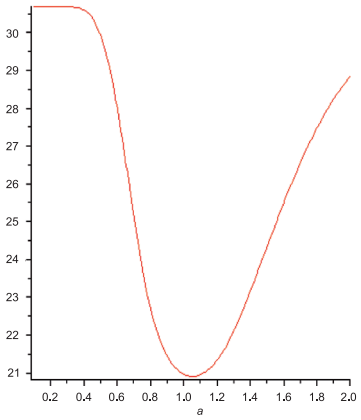

Fig. 1. Dependence of standard deviation of R from parameter a

The following graph allows us to compare the volatilities of return of the considered portfolios:

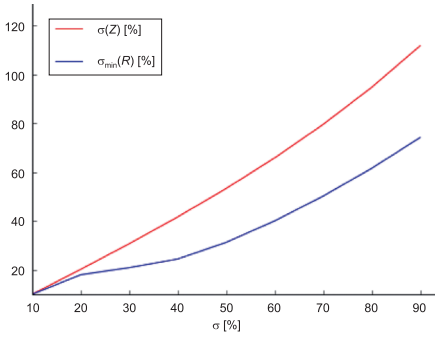

Fig. 2. Volatilities of considered portfolios

5. Conclusions

As shown, σ and $\sigma(Z)$ are approximately equal when $\sigma \leq 30\%$. If $\sigma \leq 30\%$ then $\sigma(Z) > \sigma$ and their difference increases with increasing σ.

Tables 1 and 2 allow us to compare volatilities of Z and R, expressed as standard deviations of Z and R. We see that for every stock price volatility observed in the financial market we can point to such version of considered derivative (with such a parameter a^*) with payoff function (2) that most reduces the volatility of return on the portfolio, thus reducing the risk of investing in stocks.

References

[1] Cuthbertson K., Nitzsche D., *Financial Engineering Derivatives and Risk Management*, John Wiley and Sons, LTD, 2003.

[2] Hull J.C., *Options, Futures and Other Derivatives*, Eighth edition, Prentice Hall, 2012.

[3] Jakubowski J., Palczewski A., Rutkowski M., Stettner Ł., *Matematyka finansowa – instrumenty pochodne*, WNT, Warszawa 2003.

[4] Milian A., *On some risk-reducing derivatives*, OPTIMUM, Studia Ekonomiczne, Nr 5(71), 2014, 198-207.

[5] Musiela M., Rutkowski M., *Martingale Methods in Financial Modelling*, Second Edition, Springer, 2007.

[6] Romaniuk M., Ermolieva T., *Wycena obligacji katastroficznych metodami symulacyjnymi. Badania Operacyjne Systemowe. Zastosowania*, 2004, 109-120.

[7] Weron A., Weron R., *Inżynieria finansowa. Wycena instrumentów pochodnych. Symulacje komputerowe. Statystyka rynku*, WNT, Warszawa 1999.

[8] Wiklund E., 2012, *Asian Option Pricing and Volatility*, Thesis, Royal Institute of Technology in Stockholm, Electronic document: http://www.math.kth.se/matstat/seminarier/reports/M-exjobb12/120412a.pdf (downloaded: 03.10.2014).

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

KATARZYNA PAŁASIŃSKA*

EXPANSION BY A NEW CONSTANT MAY CHANGE THE FINITE AXIOMATIZATION PROPERTY OF A MATRIX

ROZSZERZENIE SYGNATURY MATRYCY LOGICZNEJ O STAŁE WPŁYWA NA WŁASNOŚĆ SKOŃCZONEJ AKSJOMATYZACJI

A b s t r a c t

We give an example of a finite matrix with the property that expanding its language with a constant changes its finite axiomatization property: in the language with one binary operation the tautologies of the matrix are finitely axiomatizable while in the expanded language they are not. The constant we add is not definable in the original language.The deductive system generated by this matrix is not algebraizable.

Keywords: logical matrix, *finite axiomatization*

S t r e s z c z e n i e

Podajemy przykład skończonej matrycy logicznej, która jest skończenie aksjomatyzowalna, ale po dodaniu stałej do sygnatury tej matrycy, własność ta się psuje. Dodawana stała nie jest definiowalna w języku matrycy, a operator konsekwencji wyznaczony przez tę matrycę nie jest algebraizowalny

Słowa kluczowe: matryca logiczna, *skończona aksjomatyzowalność*

DOI: 10.4467/2353737XCT.16.146.5757

* Katarzyna Pałasińska (kpalasin@pk.edu.pl), Institute of Mathematics, Faculty of Physics, Mathematics and Computer Science, Cracow University of Technology.

1. Introduction

By a (logical) *matrix* we mean an algebra with a designated subset. *Tautologies* of the matrix are the terms that under every valuation take a designated value. By a *valid rule*, or simply a *rule,* we mean a pair $\langle X, \alpha \rangle$, where $X \cup \{\alpha\}$ is a finite set of terms such that for every valuation assigning designated values to all members of X, the term α also takes a designated value. For $X = \varnothing$ the rule $\langle X, \alpha \rangle$ is called *axiomatic* and is identified with α. The set of all tautologies of a matrix $\mathfrak{N}$ is denoted by $E(\mathfrak{N})$; the set of all valid rules of $\mathfrak{N}$ is denoted by $R(\mathfrak{N})$. We say that a matrix is *finitely axiomatizable* if there exists a finite set of rules valid in this matrix from which all its tautologies can be derived. This differs from the *finite basis property*, which is the property that there exists a finite set of rules from which all valid rules can be derived.

Consider the 5-element matrix

$$\mathfrak{M} = \langle \{0,1,2,3,4\}, \{\cdot\}, \{2,3\} \rangle$$

with $\cdot$ given by the following table.

$\cdot$	0	1	2	3	4
0	1	2	2	2	2
1	1	2	2	2	2
2	1	2	2	2	2
3	4	3	3	3	3
4	3	3	3	3	3

Although this matrix is finitely axiomatizable (Proposition 1), we will show that the matrix

$$\mathfrak{M}_1 = \langle \{0,1,2,3,4\}, \{\cdot, 3\}, \{2,3\} \rangle$$

does not have a finite axiomatization for the set of its tautologies (Theorem 2). The constant 3 is not a definable constant of $\mathfrak{M}$, so $\mathfrak{M}$ and $\mathfrak{M}_1$ are not term equivalent. Let us observe that the deductive systems determined by these matrices are not algebraizable.

Proposition 1. *The consequence operation of neither* $\mathfrak{M}$ *nor* $\mathfrak{M}_1$ *is algebraizable.*

Proof. Let $\mathfrak{N}$ be either $\mathfrak{M}$ or $\mathfrak{M}_1$. We will show that $\mathfrak{N}$ is not even protoalgebraic, a weaker condition than algebraizable. Suppose that there is a finite set of binary terms $\Delta(x,y)$ such that all terms in $\Delta(x,x)$ are tautologies and such that y is a consequence of $\Delta(x,y)$ and x. Such a set $\Delta(x,y)$ must exist for a protoalgebraic deductive system, [1]. As no variable is a tautology of $\mathfrak{N}$, it follows that no term in $\Delta(x,x)$ is a variable, so neither is any term in $\Delta(x,y)$. Evaluating x as 3 and y as 4, we get that x and $\Delta(x,y)$ evaluate to 3, while y is 4. This contradicts the condition that y is a consequence of $\Delta(x,y)$ and x. ■

In [4] Katarzyna Idziak has shown a finite equivalential algebra with a similar property: the quasi-equational theory of this algebra is finitely based but adding a nondefinable constant to the language of the algebra results in a nonfinitely based quasi-equational theory. Her example and ours differ in two aspects. First, the deductive system generated by the

matrix $\mathfrak{M}$ is not algebraizable, while the deductive system equivalent to any equivalential algebra is obviously algebraizable. The role of rules in the deductive system associated with this algebra is played by the quasi-identities, while the role of tautologies – by identities valid in it. Therefore the second difference lies in the difference between finite axiomatization and finite basis: the example of [4] is an example that the finite basis property is fragile under adding new constants, while ours shows the same for the finite axiomatization property. Every finitely based deductive system is finitely axiomatizable, but a system that is not finitely based may still be finitely axiomatizable. The example given in [4] has 6 elements, so ours is smaller by one element.

2. Main result

Let $V = \{x_1, x_2, \ldots\}$ be a countable set of pairwise distinct variables. Let Te denote the set of all terms written by means of these variables in the language $\{\cdot, 3\}$. When writing terms we omit the symbol of the binary operation $\cdot$ and assume the association to the left. The length of a term t is denoted by $|t|$. By θ we mean the valuation in the algebra $\langle\{0,1,2,3,4\}, \{\cdot, 3\}\rangle$ assigning 0 to every variable.

Observe that every term $t \in$ Te is of the form

$$t = t_0 t_1 \cdots t_n, \tag{1}$$

where n is a nonnegative integer, all $t_i's$ are terms and t_0 is either a variable or the constant 3. Immediately from the table we see that for a term of the form (1):

$$\textit{if } t_0 \textit{ is variable, then } \theta(t) \in \{0,1,2\}. \tag{2}$$

By our next proposition the set $E(\mathfrak{M})$ is a consequence of one single axiom, so $\mathfrak{M}$ is finitely axiomatizable.

Proposition 2. *The tautologies of the matrix* $\mathfrak{M}$ *all follow from the axiom* $x(yz)$.

Proof. Clearly, $x(yz)$ is a tautology of $\mathfrak{M}$ and no variable is. If a term of the form $t = rs$ is in $E(\mathfrak{M})$ then s cannot be a variable; for otherwise, by (2), $\theta(r) \in \{0,1,2\}$ and $\theta(rs) = \theta(r)\theta(s) = \theta(r)0 = 1$. Therefore $E(\mathfrak{M})$ contains only terms of the form $r(su)$. ■

Theorem 3. *The matrix* $\mathfrak{M}_1$ *is not finitely axiomatizable.*

Proof. Let E be the set of all tautologies of $\mathfrak{M}_1$ in Te. We will call a term *t left associated* if t is a variable or is of the form t_1x, where t_1 is left associated and $x \in V$. For the proof by contradiction let R be a finite subset of $R(\mathfrak{M}_1)$ and assume that all tautologies of $\mathfrak{M}_1$ are derivable from R. Then there is a number n such that the length of the conclusion of any rule in R is no longer than $2n$. Let

$$\alpha_0 := 3x_1x_2 \ldots x_{2n}$$

and consider the set F consisting of all left associated tautologies of $\mathfrak{M}_1$ having α_0 as a subterm. Notice that the term

$$\alpha_0 x_1 x_1 x_3 x_3 \ldots x_{2i-1} x_{2i-1} \ldots x_{2n-1} x_{2n-1}$$

belongs to F. So

$$F \neq \emptyset. \tag{3}$$

Lemma 4

Assume that $\alpha = \alpha_0 y_l y_{l-1} \ldots y_2 y_1 \in F$, where l is a nonnegative integer and $y_1, y_2, \ldots, y_l \in V$.

Then

$$l \text{ is even} \tag{4}$$

$$\forall_{i \leq l} [i \text{ is even} \Rightarrow \exists_{j<i}\, y_i = y_j] \tag{5}$$

$$\forall_{i \leq 2k} [i \text{ is odd} \Rightarrow \exists_{j<i}\, x_i = y_j] \tag{6}$$

$$l \leq 2n. \tag{7}$$

Proof of the Lemma. Use the valuation θ to see (4). For (5) assume that for some even i there is no $j < i$ such that $y_i = y_j$. Let i be the smallest such. Assign 3 to y_i and 0 to every variable other than y_i. Then the value of α is $3 \cdot 0 \cdots 0$, with an odd number of 0′s in this expression. Hence α takes 4 under this valuation, a contradiction. Condition (6) is proved similarly and (7) follows from (6) and (5). ■

By (3) there exists a proof using the rules from the set R that proves some term $\alpha \in F$. Consider a shortest such proof π and let $\alpha \in F$ be the term proved by this proof. Consider the last rule $\langle X, \beta \rangle \in R$ used in π. So

$$|\beta| \leq 2n \tag{8}$$

and there is a substitution σ such that:

$$\sigma(\beta) = \alpha \tag{9}$$

and all terms $\sigma(\gamma)$ for $\gamma \in X$ occur in the proof π earlier than α. Since π is a shortest proof proving a formula in F, it follows that for every $\gamma \in X$

$$\sigma(\gamma) \in E \setminus F. \tag{10}$$

Since α satisfies the assumptions of Lemma 4, by (7), (8) and (9) we get that

$$\beta = u v_m \ldots v_1,$$

where $u, v_1 \ldots, v_m \in V$, $m < l$, $\sigma(u) = \alpha_0 y_1 \ldots y_{m+1}$ and $\sigma(v_i) = y_i$ for each $i = 1, \ldots, m$. Obviously, $u \neq v_i$, for any $i \in \{1, \ldots, m\}$.

Let us define the valuation φ such that $\varphi(u) = 3$ if m is odd, $\varphi(u) = 4$ if m is even and $\varphi(x) = \theta(\sigma(x))$ for every $x \in V \setminus \{u\}$. Notice that then for $i \in \{1, \ldots, m\}$, $\varphi(v_i) = 0$, so $\varphi(\beta) = 4$. Since the rule $\langle X, \beta \rangle$ is valid in $\mathfrak{M}_1$, there must be a term $\gamma \in X$ such that

$$\varphi(\gamma) \in \{0, 1, 4\}. \tag{11}$$

By (10), $\theta(\sigma(\gamma)) \in \{2,3\}$. So

$$\varphi(\gamma) \neq \theta(\sigma(\gamma)). \tag{12}$$

By the definition of φ and by (12), the term γ contains u. Moreover, by (1), the term γ takes one of the following three forms: $\gamma = ut_1 \cdots t_k$, $\gamma = xt_1 \cdots t_k$ with $x \in V$ and $x \neq u$, or $\gamma = 3t_1 \cdots t_k$, for some k and some sequence $t_1, \ldots, t_k$ of terms. In the last two cases, $\varphi(\gamma) = \theta(\sigma(\gamma))$, because on positions other than the initial one, the value 3 behaves the same as the value 4. Similarly, if any of the terms t_i would be composed, then we would have $\varphi(\gamma) = \theta(\sigma(\gamma))$. So it follows by (12) that the only form γ may take is

$$\gamma = uz_1 \ldots z_k,$$

where $z_1, \cdots, z_k$ are variables. But then σ(γ) is a left associated tautology of $\mathfrak{M}_1$ with a subterm α_0 which contradicts (10). ■

The technique of the proof is similar to the one used in [2, 6, 7]. The idea of the example is similar to that of [5].

3. Questions

One may ask if there is a matrix of a smaller size or a matrix with a smaller number of designated values that has the same property as presented here.

Question 5. *Is there a non-algebraizable matrix with less than 5 elements with the property that its tautologies are finitely axiomatizable while the tautologies of the same matrix in the language expanded by a constant are not finitely axiomatizable?*

Question 6. *Find such a matrix with only one designated value.*

The finite basis property mentioned in the Introduction is related but different from the finite axiomatization property. Our example does not answer the following

Question 7. *Find a non-algebraizable finitely based matrix that expanded by a constant becomes non-finitely based.*

An open problem, due to W. Rautenberg is whether the finite basis property of finite matrices is independent of the language. More precisely, given a finite finitely based matrix, is every matrix term-equivalent to it also finitely based? See [3]. The constant 3 added to the language of our matrix $\mathfrak{M}$ is clearly not definable.

Question 8. *Is there a finitely based (resp. finitely axiomatizable) matrix* $\mathfrak{M} = \langle M, F, D \rangle$ *and a constant c definable in its language such that the consequence operation of the matrix* $\mathfrak{M}_1 = \langle M, F \cup \{c\}, D \rangle$ *is not finitely based (not finitely axiomatizable, resp.)?*

If such a matrix $\mathfrak{M}$ exists its consequence operation is necessarily non-algebraizable.

References

[1] Blok W., Pigozzi D., *Protoalgebraic Logics*, "Studia Logica", 45 /1986, 337-369.
[2] Dziobiak W., *A finite matrix whose set of tautologies is not finitely axiomatizable*, "Reports on Mathematical Logic", 25/1991/, 113-117.
[3] Herrmann B., Rautenberg W., *Finite replacement and Finite Hilbert-Style Axiomatizability*, "Mathematical Logic Quarterly", 38/1992, 327-344.
[4] Idziak K., *Equivalential Logics with Constants*, Doctoral dissertation, Jagiellonian University, Kraków 1998.
[5] Pałasińska K., *Three-element nonfinitely axiomatizable matrices*, "Studia Logica", 53/1994, 361-372.
[6] Wojtylak P., *Strongly Finite Logics: Finite Axiomatizability and the Problem of Supremum*, "Bulletin of the Section of Logic", PAN 8/1979, 99-111.
[7] Wojtylak P., *An example of a finite though finitely non-axiomatizable matrix*, "Reports on Mathematical Logic", 17/1984, 39-46.

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

ELIZA WAJCH*

CONVERGENCE IN MEASURE THROUGH COMPACTIFICATIONS

ZBIEŻNOŚĆ WEDŁUG MIARY POPRZEZ UZWARCENIA

Abstract

For a metrizable space X, concepts of metric convergence in measure and of functional convergence in measure of sequences of measurable mappings taking their values in X are introduced and applied to a comparison of compactifications of X.

Keywords: *Metrizable space*, *Hausdorff compactification, mimimum uniform compactification*, *infinite σ-finite measure*, *metric convergence in measure*, *functional convergence in measure*

Streszczenie

Dla przestrzeni metryzowalnej X, wprowadza się pojęcia metrycznej zbieżności według miary i funkcyjnej zbieżności według miary ciągów odwzorowań mierzalnych, przyjmujących swe wartości w X oraz stosuje się te pojęcia do porównywania uzwarceń przestrzeni X.

Słowa kluczowe: *przestrzeń metryzowalna*, *uzwarcenie Hausdorffa*, *minimalne uzwarcenie jednostajne*, *nieskończona miara σ-skończona*, *metryczna zbieżność według miary*, *funkcyjna zbieżność według miary*

DOI: 10.4467/2353737XCT.16.147.5758

* Eliza Wajch (eliza.wajch@wp.pl), Institute Mathematics and Physics, University of Natural Sciences and Humanities in Siedlce.

In memory of Stanisław Łojasiewicz [*1926–2002*]

1. Introduction

A convenient interpretation of ZFC which agrees with that of [8] is our basic set-theoretic assumption. An evident frequent use of the axiom of countable choice (CC) makes it impossible to rewrite in ZF most of the results of this work (cf. [4] and [6]–[8]); however, some of the theorems presented below are also theorems of, for instance, ZF+UFT+CC (cf [6]).

In what follows, $X = (X,\tau)$ is a non-void metrizable space, $\mathfrak{B}(X)$ is the σ-field of all Borel sets in X, and $\mathfrak{B}_s(X)$ is the collection of all separable Borel sets in X. Moreover, $\mathfrak{M}$ is a σ-field of subsets of a set E, while μ is an infinite σ-finite measure on $\mathfrak{M}$. Let $\mathcal{M}(E,X)$ be the family of all $(\mathfrak{M},\mathfrak{B}(X))$-measurable functions $f: E \to X$ such that $\mu[f^{-1}(X \setminus B_f)] = 0$ for some $B_f \in \mathfrak{B}_s(X)$. Of course, a function $f: E \to X$ is $(\mathfrak{M},\mathfrak{B}(X))$-measurable if and only if $f^{-1}(V) \in \mathfrak{M}$ whenever $V \in \tau$. If one wants to try to work without CC, since second countability and separability are not equivalents in ZF+¬CC, it might be more preferable to define $\mathfrak{B}_s(X)$ as the collection of all these Borel sets of X that are second-countable as topological subspaces of X. Clearly, the second definition of $\mathfrak{B}_s(X)$ is equivalent in ZFC to our previous definition of $\mathfrak{B}_s(X)$ but not equivalent to it in ZF.

Every compactification of X is assumed to be Hausdorff. For a compactification αX of X, the collection of all bounded continuous real functions on X that are continuously extendable over αX is denoted by $C_\alpha(X)$. As usual, βX stands for the Čech-Stone compactification of X. The collection of all bounded continuous real functions on X is $C_\beta(X)$. A great role in the theory of compactifications is played by the collection $\mathcal{E}(X)$ of all sets $F \subseteq C_\beta(X)$ such that the evaluation mapping $e_F : X \to \mathbb{R}^F$ is a homeomorphic embedding where $[e_F(x)](f) = f(x)$ for all $f \in F$ and $x \in X$ (cf. e.g.[1], [2] and [11]–[13]). If $F \in \mathcal{E}(X)$, then the closure in $\mathbb{R}^F$ of the set $e_F(X)$ is a compactification of X called generated by F and denoted by $e_F X$. In particular, every compactification αX of X is generated by $C_\alpha(X)$. Since, in ZF, the sentence that Tikhonov cubes (called Hilbert cubes in [6]) are compact is equivalent with the ultrafilter theorem UFT (cf. Theorem 4.70 of [6]), it is true in ZF+UFT that, for every $F \in \mathcal{E}(X)$, the compactification $e_F X$ of X exists. This is why some theorems on compactifications in ZFC are also theorems of ZF+UFT. It is still an open problem to investigate all significant details on compactifications in ZF+UFT and show possible differences between the theories of compactifications in ZFC and in ZF+UFT. Let us leave this problem for future considerations not described in this article and, for simplicity, let us work in ZFC to avoid troublesome disasters without AC. All topological and set-theoretic concepts that we use are standard and they can be found in [2], [3], [6]–[8] and [10]. Useful facts of measure theory are taken from [5] and [9].

The paper is mainly about the following concepts of metric and functional convergence in measure:

Definition 1. Let d be a compatible metric on X and let f_n, f be functions from $\mathcal{M}(E,X)$ where $n \in \omega$. We say that the sequence $\langle f_n \rangle$ is d-convergent in measure μ to f if, for each positive real number ε, the sequence

$$\langle \mu(\{t \in E : d(f_n(t), f(t)) \geq \varepsilon\}) \rangle$$

converges to zero in $\mathbb{R}$ with the usual topology. For every compatible metric ρ on X, the ρ-convergence in μ will be called a metric convergence in μ.

Definition 2. Suppose that $\varnothing \neq F \subseteq C_\beta(X)$ and let f_n, f be functions from $\mathcal{M}(E, X)$ where $n \in \omega$. We say that the sequence $\langle f_n \rangle$ is F-convergent in measure μ to f if, for each positive real number ε and for each $\phi \in F$, the sequence

$$\langle \mu(\{t \in E : |\phi(f_n(t)) - \phi(f(t))| \geq \varepsilon\}) \rangle$$

converges to zero, i.e. if for each $\phi \in F$, the sequence $\langle \phi \circ f_n \rangle$ converges in μ to $\phi \circ f$. For each set $H \in \mathcal{E}(X)$, the H-convergence in μ will be called a functional convergence in μ.

Definition 3. Let d, ρ be compatible metrics on X and let F, H be non-void subsets of $C_\beta(X)$. For $i, j \in \{d, \rho, F, H\}$, we say that:

1. i-convergence in μ implies j-convergence in μ if every sequence of functions from $\mathcal{M}(E, X)$ which is i-convergent in μ to a function $f \in \mathcal{M}(E, X)$ is also j-convergent in μ to f;
2. i-convergence in μ is equivalent with j-convergence in μ if i-convergence in μ implies j-convergence in μ and j-convergence in μ implies i-convergence in μ.

In the sequel, the notions of d-convergence and F-convergence in μ are applied to a comparison of compactifications of X. Recall that, for compactifications αX and γX of X, the inequality $\alpha X \leq \gamma X$ holds if and only if $C_\alpha(X) \subseteq C_\gamma(X)$; moreover, αX and γX are equivalent if and only if $C_\alpha(X) = C_\gamma(X)$. We write $\alpha X = \gamma X$ to say that αX is identified with γX, i.e. to denote that αX and γX are equivalent. One of the most interesting theorems of this paper asserts that if there exists a metrizable compactification αX of X such that $C_\alpha(X)$-convergence in μ implies $C_\beta(X)$-convergence in μ, then the space X is compact. Moreover, among other results, it is shown that if αX and γX are metrizable compactifications of X, then $\alpha X \leq \gamma X$ if and only if $C_\gamma(X)$-convergence in μ implies $C_\alpha(X)$-convergence in μ. Ideas of simple examples relevant to convergence in μ are described.

2. Metric convergence in measure and minimum uniform compactifications

For a compatible metric d on X, R. Grant Woods investigated in [14] the compactification $u_d X$ generated by the collection $\mathcal{U}_d^*(X)$ of all these bounded real functions on X that are uniformly continuous with respect to d and the standard metric induced by the absolute value on $\mathbb{R}$. If the metric d is not totally bounded, $u_d X$ is not metrizable (cf. Theorem 3.3 (b) of [14]). If the metric d is totally bounded, then $u_d X$ is the Hausdorff metric completion of the metric space (X, d) (cf. Theorem 3.3 (a) of [14] and Problem 4.5.6 of [3]). If one wants to consider minimum uniform compactifications in ZF, one should be warned that models of ZF in which there are infinite Dedekind-finite dense subsets of $\mathbb{R}$ (cf. [6]–[8]) can be used to deduce the following:

Proposition 1. *If X is an infinite Dedekind-finite dense subset of the unit interval* $[0;1]$ *and* $d(x,y)=|x-y|$ *for* $x,y\in X$, *then d is a totally bounded complete metric on X such that* $u_dX=[0;1]$, *while the Hausdorff metric completion of* (X,d) *is (up to an obvious isometry)* (X,d). *Therefore, it is unprovable in ZF that, for every totally bounded metric space* (X,d), *the minimum uniform compactification u_dX is the Hausdorff metric completion of* (X,d).

That $u_dX=[0;1]$ for each dense in $[0;1]$ infinite Dedekind-finite set X in the proposition above can be shown in ZF by using Lemma 4.3.16 of [3]. Interesting problems on Hausdorff metric completions of metric spaces in ZF are described in [4]. To avoid misunderstanding, let us recall that ZFC is our basic assumption throughout this paper.

For every metrizable compactification αX of X, there exists a totally bounded metric ρ on X such that $\alpha X=u_\rho X$. To apply metric convergence in measure to minimum uniform compactifications, the following notion is useful:

Definition 4. Let d and ρ be compatible metrics on X. We say that d is uniformly smaller than ρ if the following condition holds:

$$\forall_{\varepsilon\in(0;+\infty)}\exists_{\delta\in(0;+\infty)}\forall_{x,y\in X}[\rho(x,y)<\delta\Rightarrow d(x,y)<\varepsilon].$$

Proposition 2. *Let d and ρ be compatible metrics on X such that d is not uniformly smaller than* ρ. *Then there exist functions* $f_n,f\in\mathcal{M}(E,X)$ *where* $n\in\omega$, *such that the sequence* $\langle f_n\rangle$ *is* ρ*-convergent in* μ *to f but* $\langle f_n\rangle$ *is not d-convergent in* μ *to f.*

Proof. Let us take $\varepsilon\in(0,+\infty)$ such that, for each $\delta\in(0,+\infty)$, there are $x,y\in X$ such that $\rho(x,y)<\delta$, while $d(x,y)\geq\varepsilon$. Using CC, we find sequences $\langle x_n\rangle$ and $\langle y_n\rangle$ of points of X such that $\lim_{n\to+\infty}\rho(x_n,y_n)=0$, while $d(x_n,y_n)\geq\varepsilon$ for each $n\in\omega$. Let $\langle E_n\rangle$ be a sequence of sets from $\mathfrak{M}$ such that $\bigcap_{n\in\omega}E_n=\varnothing$, $\mu(E\setminus E_n)<+\infty$, $\mu(E_n)=+\infty$ and $E_{n+1}\subsetneq E_n$ for all $n\in\omega$. Such a sequence $\langle E_n\rangle$ exists because the measure μ is infinite and σ-finite. Define $f_n(t)=y_0$ for $t\in E\setminus E_0$ and, for each $t\in E_i\setminus E_{i+1}$, let $f_n(t)=y_i$ if $i\leq n$, while $f_n(t)=x_i$ if $i>n$. Moreover, put $f(t)=y_0$ for $t\in E\setminus E_1$ and, for each $i\in\omega$, let $f(t)=y_i$ when $t\in E_i\setminus E_{i+1}$. The sequence $\langle f_n\rangle$ ρ-converges in μ to f but it does not d-converge in μ to f. □

The proof to Proposition 2 can serve as a scheme of examples of sequences ρ-convergent in μ that are not d-convergent in μ.

In much the same way as for the classical convergence in measure, one can prove Propositions 3–5.

Proposition 3. *Let d be a compatible metric on X and let $f,g\in\mathcal{M}(E,X)$. If a sequence of functions from $\mathcal{M}(E,X)$ is d-convergent in* μ *to f and to g, then* $\mu(\{t\in E:f(t)\neq g(t)\})=0$.

Definition 5. When d is a compatible metric on X, then we say that a sequence $\langle f_n\rangle$ of functions from $\mathcal{M}(E,X)$ converges (d,μ)-uniformly on E to a function $f\in\mathcal{M}(E,X)$ if, for each $\varepsilon\in(0,+\infty)$, there exists a set $A\in\mathfrak{M}$ such that $\mu(E\setminus A)<\varepsilon$ and the convergence of $\langle f_n\rangle$ to f is uniform with respect to d on A.

Proposition 4. *When d is a compatible metric on X, then a sequence* $\langle f_n \rangle$ *of functions from* $\mathcal{M}(E,X)$ *is d-convergent in* μ *to a function* $f \in \mathcal{M}(E,X)$ *if and only if each subsequence of* $\langle f_n \rangle$ *contains a subsequence which converges* (d,μ)*-uniformly on E to f.*

Proposition 5. *If d is a compatible metric on X, then every sequence of functions from* $\mathcal{M}(E,X)$ *which is d-convergent in* μ *to a function* $f \in \mathcal{M}(E,X)$ *contains a subsequence which pointwise converges* μ*-almost everywhere on E to f.*

In the light of Proposition 5, for every pair d, ρ of compatible metrics on X and for every pair f, g of functions from $\mathcal{M}(E,X)$, it is true that if there exists a sequence $\langle f_n \rangle$ of functions from $\mathcal{M}(E,X)$ such that $\langle f_n \rangle$ is both d-convergent in μ to f and ρ-convergent in μ to g, then $f = g$ μ-almost everywhere on E, i.e. $\mu(\{t \in E : f(t) \neq g(t)\}) = 0$. Therefore, if for compatible metrics d and ρ on X, a sequence $\langle h_n \rangle$ of functions from $\mathcal{M}(E,X)$ is d-convergent in μ to a function $h \in \mathcal{M}(E,X)$ and the same sequence $\langle h_n \rangle$ is not ρ-convergent in μ to h, then there does not exist a function in $\mathcal{M}(E,X)$ such that $\langle h_n \rangle$ is ρ-convergent in μ to it.

Theorem 1. *For every pair* d, ρ *of compatible metrics on X, the following conditions are equivalent:*

1. *d is uniformly smaller than* ρ;
2. $\mathcal{U}_d^*(X) \subseteq \mathcal{U}_\rho^*(X)$;
3. $u_d X \leq u_\rho X$;
4. *for every pair A, B of subsets of X such that* $d(A,B) > 0$, *the inequality* $\rho(A,B) > 0$ *holds*;
5. ρ*-convergence in* μ *implies d-convergence in* μ.

Proof. It is obvious that implications (i)⇒(ii)⇒(iii) and (i)⇒(v) are true. Suppose that (iii) holds and consider an arbitrary pair A, B of subsets of X such that $d(A,B) \neq 0$. Then $\mathrm{cl}_{u_d X} A \cap \mathrm{cl}_{u_d X} B = \varnothing$ by Theorem 2.5 of [14]. Since $u_d X \leq u_\rho X$, in the light of 4.2(h) of [10], we have $\mathrm{cl}_{u_\rho X} A \cap \mathrm{cl}_{u_\rho X} B = \varnothing$. This, together with Theorem 2.5 of [14], gives that $\rho(A,B) \neq 0$. Hence (iv) follows from (iii). Now, assume that (i) is not fulfilled. Then, with CC in hand, we deduce that, for some $\varepsilon \in (0,+\infty)$, there are sequences $\langle x_n \rangle$ and $\langle y_n \rangle$ of X such that $\lim_{n\to+\infty} \rho(x_n, y_n) = 0$ and $d(x_n, y_n) \geq \varepsilon$ for all $n, m \in \omega$ (cf. hint to 8.5.19 of [3]). If $A = \{x_n : n \in \omega\}$ and $B = \{y_n : n \in \omega\}$, then $\rho(A,B) = 0$, while $d(A,B) \neq 0$. Therefore, (i) is a consequence of (iv). That (v) implies (i) follows from Proposition 2. □

Corollary 1. *Let d and* ρ *be compatible metrics on X. If* ρ *is totally bounded and* ρ*-convergence in* μ *implies d-convergence in* μ, *then d is totally bounded.*

Proof. It is clear that if d is uniformly smaller than ρ and the metric ρ is totally bounded, then d is also totally bounded. To complete the proof, it suffices to use the equivalence of (i) and (v) of Theorem 1. □

Theorem 2. *Assume that d is a totally bounded compatible metric on X. Then d-convergence in* μ *is equivalent with* $\mathcal{U}_d^*(X)$ *-convergence in* μ.

Proof. It is obvious that *d*-convergence in μ implies $\mathcal{U}_d^*(X)$ -convergence in μ. Since *d* is totally bounded, u_dX is a metrizable compactification of *X*. By, for example, Propositions 3.4 and 3.5 of [11] or by Theorem 7 of [12], there is a countable collection $F \subseteq \mathcal{U}_d^*(X)$ such that $e_FX = u_dX$ and, moreover, $\phi(X) \subseteq [0;1]$ for each $\phi \in F$. Let $F = \{\phi_i : i \in \omega\}$ and define $\rho(x,y) = \sum_{i\in\omega} \frac{1}{2^{i+1}} |\phi_i(x) - \phi_i(y)|$ for all $x,y \in X$. Then ρ is a totally bounded metric on *X* such that $u_dX = u_\rho X$. Hence, in view of Theorem 1, *d*-convergence in μ is equivalent with ρ-convergence in μ. However, one can easily check that *F*-convergence in μ implies ρ-convergence in μ. In consequence, *F*-convergence in μ implies *d*-convergence in μ, which concludes the proof. □

Question 1. If *d* is a compatible but not totally bounded metric on *X*, must $\mathcal{U}_d^*(X)$ -convergence in μ imply *d*-convergence in μ?

A familiar theorem of ZFC states that a metrizable space *X* is compact if and only if every compatible metric on *X* is totally bounded. The standard proof to this theorem involves CC. However, one can easily prove in ZF that if *X* is a metrizable space such that every compatible metric on *X* is totally bounded, then *X* is closed in every metrizable space that contains *X* as a subspace. Indeed, let (Y,d) be a metric space and let $X \subseteq Y$ be not closed in (Y,d). Choose a point $x_0 \in (\mathrm{cl}_Y X) \setminus X$ and, for $x,y \in X$, define

$$\rho(x,y) = d(x,y) + \left| \frac{1}{d(x,x_0)} - \frac{1}{d(y,x_0)} \right|$$

to get a compatible but not totally bounded metric ρ on *X* in ZF (cf. 4.3.E.(c) of [3]). Unfortunately, this does not give a satisfactory answer to the following interesting question:

Question 2. Is it consistent with ZF+¬CC that there exists a non-compact metrizable space *X* such that every compatible metric on *X* is totally bounded?

3. Functional convergence in measure

It has not been explained so far why it is assumed here that, for each function $f \in \mathcal{M}(E,X)$, there exists $B_f \in \mathfrak{B}_s(X)$ such that $\mu[f^{-1}(X \setminus B_f)] = 0$. In fact, this assumption was needless in the previous section; however, it is helpful to get the following theorem:

Theorem 3. *Let us suppose that* $F \in \mathcal{E}(X)$, *while* $\langle f_n \rangle$ *is a sequence of functions from* $\mathcal{M}(E,X)$ *such that* $\langle f_n \rangle$ *is F-convergent in* μ *to functions* $f,g \in \mathcal{M}(E,X)$. *Then the following conditions hold:*

1. $\mu(\{t \in E : f(t) \neq g(t)\}) = 0$;
2. *each subsequence of* $\langle f_n \rangle$ *contains a subsequence that pointwise converges* μ*-almost everywhere on E to f*;
3. *if* $G \in \mathcal{E}(X)$ *is such that the sequence* $\langle f_n \rangle$ *is not G-convergent in* μ *to f, then there does not exist a function* $h \in \mathcal{M}(E,X)$ *such that* $\langle f_n \rangle$ *is G-convergent in* μ *to h.*

Proof. Using CC, we deduce that there is a sequence $\langle B_n \rangle$ of separable Borel subsets of X and there are sets $B_f, B_g \in \mathfrak{B}_s(X)$, such that the sets $X_0 = B_f \cup B_g \cup \bigcup_{n\in\omega} B_n$ and $E_0 = E \setminus [f^{-1}(X \setminus X_0) \cup g^{-1}(X \setminus X_0) \cup \bigcup_{n\in\omega} f_n^{-1}(X \setminus X_0)]$ have the properties that $\mu(E \setminus E_0) = 0$ and all functions f_n, f, g restricted to E_0 transform E_0 into the separable Borel in X set X_0. It follows from Theorem 6 of [12] that there exists a countable collection $H \subseteq F$ such that the restriction to X_0 of the evaluation map e_H is a homeomorphic embedding of X_0 into $\mathbb{R}^H$. Let $H = \{\phi_i : i \in \omega\}$. For each $i \in \omega$, choose a positive real number a_i such that $|\phi_i| \leq a_i$ and, for $x, y \in X$, define $\rho(x,y) = \sum_{i\in\omega} \frac{|\phi_i(x) - \phi_i(y)|}{a_i 2^{i+1}}$. Then ρ is a compatible metric on X_0. It is not difficult to check that the sequence $\langle f_n|_{E_0} \rangle$ of the restrictions $f_n|_{E_0}$ of functions f_n to E_0 is ρ-convergent in μ to $f_n|_{E_0}$ and $g|_{E_0}$. Hence, in view of Proposition 3, $\mu(\{t \in E : f(t) \neq g(t)\}) = 0$. Now, to conclude that (ii) holds, it suffices to use Proposition 5. Condition (iii) follows from (ii). □

Theorem 4. *Let* αX *be a compactification of X and let* $F \in \mathcal{E}(X)$ *generate* αX*, i.e.* $\alpha X = e_F X$*. Then F-convergence in* μ *and* $C_\alpha(X)$*-convergence in* μ *are equivalent.*

Proof. Since $F \subseteq C_\alpha(X)$, it is obvious that $C_\alpha(X)$-convergence in μ implies F-convergence in μ. Now, assume that a sequence $\langle f_n \rangle$ of functions from $\mathcal{M}(E,X)$ is F-convergent in μ to a function $f \in \mathcal{M}(E,X)$. Consider an arbitrary function $\phi \in C_\alpha(X)$ and a positive real number ε. By Theorem 4 of [13], there exist a non-void finite set $H \subseteq F$ and a positive real number δ, such that if

$$d_H(x,y) = \max\{|\psi(x) - \psi(y)| : \psi \in H\}$$

for $x, y \in X$, then $|\phi(x) - \phi(y)| < \varepsilon$ whenever $d_H(x,y) < \delta$. It follows from the F-convergence in μ of $\langle f_n \rangle$ to f that

$$\lim_{n\to+\infty} \mu(\{t \in E : d_H(f_n(t), f(t)) \geq \delta\}) = 0.$$

In addition,

$$\{t \in E : |\phi[f_n(t)] - \phi[f(t)]| \geq \varepsilon\} \subseteq \{t \in E : d_H(f_n(t), f(t)) \geq \delta\}$$

for all $n \in \omega$. In consequence,

$$\lim_{n\to+\infty} \mu(\{t \in E : |\phi[f_n(t)] - \phi[f(t)]| \geq \varepsilon\}) = 0.$$

This means that $\langle f_n \rangle$ is $C_\alpha(X)$-convergent in μ to f. □

We consider the set $C_\beta(X)$ as the metric space $(C_\beta(X), \sigma)$ where the metric σ on $C_\beta(X)$ is defined as follows: $\sigma(f,g) = \sup\{|f(x) - g(x)|; x \in X\}$ for $f, g \in C_\beta(X)$. In view of, for example, Theorem 7 of [12], when $F \in \mathcal{E}(X)$, then the compactification $e_F X$ of X is metrizable if and only if F is second-countable in $(C_\beta(X), \sigma)$. In what follows, every subset of $C_\beta(X)$ is equipped with the topology inherited from the topology on $C_\beta(X)$ induced by the metric σ.

Theorem 5. *Let αX and γX be compactifications of X such that αX is metrizable and $C_\alpha(X)$-convergence in μ implies $C_\gamma(X)$-convergence in μ. Then γX is also metrizable and $\gamma X \leq \alpha X$.*

Proof. Since αX is metrizable, there exists a totally bounded compatible metric ρ on X such that $u_\rho X = \alpha X$. Consider any function $\phi \in C_\gamma(X)$ and let $F = C_\alpha(X) \cup \{\phi\}$. Of course, $F \in \mathcal{E}(X)$. The compactification $e_F X$ is metrizable because F is second-countable. There is a totally bounded metric d on X such that $e_F X = u_d X$. It follows from Theorem 2 that ρ-convergence in μ implies d-convergence in μ. Therefore, $u_d X \leq u_\rho X$ by Theorem 1. This implies that $F \subseteq C_\alpha(X)$ and, in consequence, $C_\gamma(X) \subseteq C_\alpha(X)$. Then $\gamma X \leq \alpha X$ and $C_\gamma(X)$ is second-countable. Hence γX is metrizable. □

Corollary 2. *Let αX and γX be metrizable compactifications of X. Then $\alpha X \leq \gamma X$ if and only if $C_\gamma(X)$-convergence in μ implies $C_\alpha(X)$-convergence in μ.*

From Theorems 4 and 5 we immediately deduce the following:

Corollary 3. *Suppose that $F, G \in \mathcal{E}(X)$ are such that F-convergence in μ implies G-convergence in μ. If F is second-countable, then G is second-countable and $e_G X \leq e_F X$.*

Our final theorem is a nice conclusion from Theorem 5.

Theorem 6. *If there exists a metrizable compactification αX of X such that $C_\alpha(X)$--convergence in μ implies $C_\beta(X)$-convergence in μ, then X is compact.*

Proof. Let us assume that αX is a metrizable compactification of X such that $C_\alpha(X)$--convergence in μ implies $C_\beta(X)$-convergence in μ. Since $\alpha X \leq \beta X$, it follows from Theorem 5 that βX is metrizable and $\beta X = \alpha X$. If X were non-compact, βX would be non--metrizable (cf. 3.6. G of [3]). □

Corollary 4. *A metrizable space X is compact if and only if there exists a totally bounded metric d on X such that d-convergence in μ implies $C_\beta(X)$-convergence in μ.*

References

[1] Ball B.J., Shoji Yokura, *Compactifications determined by subsets of C*(X)*, Topology Appl., 13 (1982), 1–13.

[2] Chandler R., *Hausdorff Compactifications*, Marcel Dekker, New York 1976.

[3] Engelking R., *General Topology*, PWN, Warsaw 1989.

[4] Gutierres G., *The Axiom of Countable Choice in Topology*, thesis, Department of Mathematics, University of Coimbra, 2004.

[5] Halmos P., *Measure Theory*, New York 1950.

[6] Herrlich H., *Axiom of Choice*, Springer–Verlag 2006.

[7] Jech T.J., *The Axiom of Choice*, North-Holland, Amsterdam 1973.

[8] Kunen K., *The Foundations of Mathematics*, College Publications, London 2009.

[9] Łojasiewicz S., *An Introduction to the Theory of Real Functions*, PWN, Warsaw 1976 (in Polish), Wiley 1988 (translation into English).

[10] Porter J.R., Woods R.G., *Extensions and Absolutes of Hausdorff Spaces*, Springer–Verlag 1987.

[11] Wajch E., *Subsets of C*(X) generating compactifications*, Topology Appl., 29, 1988, 29–39.

[12] Wajch E., *Compactifications and L-separation*, Comment. Math. Univ. Carolinae, 29 (3), 1988, 477–484.

[13] Wajch E., *Sets of functions generating compactifications via uniformities. Connectedness*, Demonstratio Math., 26 (2), 1993, 501–511.

[14] Woods R.G., *The minimum uniform compactification of a metric space*, Fund. Math., 147, 1995), 39–59.

COMPUTER SCIENCES

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

JAN KUCWAJ*

COMPUTATIONAL EXPERIMENTS OF A REMESHING ALGORITHM BASED ON MESH GENERATOR

NUMERYCZNA EFEKTYWNOŚĆ ALGORYTMU OPARTEGO NA GENERATORZE SIATEK

A b s t r a c t

The main goal of the presented paper are numerical experiments of the convergence of the adaptation algorithm [4] developed by the author based on remeshing, which form the proof of the concept for the presented algorithm. The main feature of the considered algorithm is an application of the mesh generator to the adaptation with a mesh size function [5]. The proposed method uses a sequence of meshes obtained by successive modification of the mesh size function. The rate of the convergence is obtained numerically considering a known solution. The analysis of the unknown solution was restricted to the assessment of some properties of the strict solution.

Keywords: adaptivity, *mesh generation*, *error estimation*, *finite element method*, *finite difference method*, *nolinearity*

S t r e s z c z e n i e

Głównym celem artykułu są numeryczne eksperymenty zbieżności algorytmu adaptacji [4] rozwijanego przez autora opartego na „remeshingu", które potwierdzają koncepcję prezentowanego algorytmu. Główną cechą analizowanego algorytmu jest zastosowanie generatora siatek [5] z zadaną funkcją rozmiaru siatki. Rozwijana metoda wykorzystuje ciąg siatek pokrywających obszar otrzymanych z odpowiednio modyfikowaną funkcją rozmiaru siatki. Otrzymane tempo numerycznej zbieżności zostało uzyskane na znanym rozwiązaniu. Analiza nieznanego rozwiązania została sprowadzona do oceny znanych własności danego rozwiązania, które mogą być obserwowane na rozwiązaniach przybliżonych

Słowa kluczowe: adaptacja, *generowanie siatek*, *wskaźnik błędu*, *metoda elementów skończonych*, *metoda różnic skończonych*, *zagadnienia nieliniowe*

DOI: 10.4467/2353737XCT.16.148.5759

* Jan Kucwaj (jkucwaj@pk.edu.pl), Institute of Computer Science, Cracow University of Technology.

1. Introduction

The paper concerns numerical speed of the convergence of the adaptive algorithm based on a grid generator with a mesh size function [6, 7]. The rate of convergence will be calculated by a description of the dependence between the number of degrees of freedom and norm of error defined as the difference between a strict solution and an approximate solution for a given mesh, provided that the strict solution is known. In case of an unknown solution some properties of the solution are known and their fulfilment can be assessed.

For the sake of the numerical solution the infinite space is approximated by a finite dimensional space spanned by a given set of basis functions [7, 11] of the finite element method [10] generated by linear shape functions [10], the approximated solution to the problem is equal to a linear combination of the basis functions. The coefficients of the linear combination are found from the nonlinear algebraic system of equations. The system is led out from stationarity conditions. The system of nonlinear algebraic equations is solved by using the Newton-Raphson method. In consecutive remeshing (this means separate finite element problems) steps of the adaptation algorithm the values of the mesh size function taken at the nodes are so modified that at the points with greatest values of an error indicator [2, 5] the values of mesh size function are the most diminished. Having the values of the mesh size function at nodes the new mesh size function is defined by the linear interpolation. The process is performed till the error indicator attains the assumed value. The error indicator is found at every node as an approximated residual by the finite difference method for the appropriate local formulation.

The presented numerical analysis of the convergence suggests better than linear dependence between number of degrees of freedom and error norm for derivatives. In further development it is planned to generalize the method to apply anisotropic meshes. The proposed method was applied to both problems, in which the solution is known and unknown. The obtained results were consistent with physical interpretations [4].

The adapted mesh for an example problem, where the strict solution is known, is presented. It can be observed that the rapid change of the size function corresponds to the great gradient of the solution. Additionally, it can be said that the final t mesh depends on both the solution and the assumed error indicator. As an example problem the Poisson equation was taken with known solution and elastic-plastic problem of twisting of bars with hardening, where some physical properties of the solution to the problem are known.

2. Example problem

2.1. The Poisson equation

The boundary value problem for the Poisson equation is formulated as follows:

$$\Delta u = f(x, y), \quad \text{in} \quad \Omega, \tag{1}$$

$$u = 0, \quad \text{in} \quad \partial\Omega. \tag{2}$$

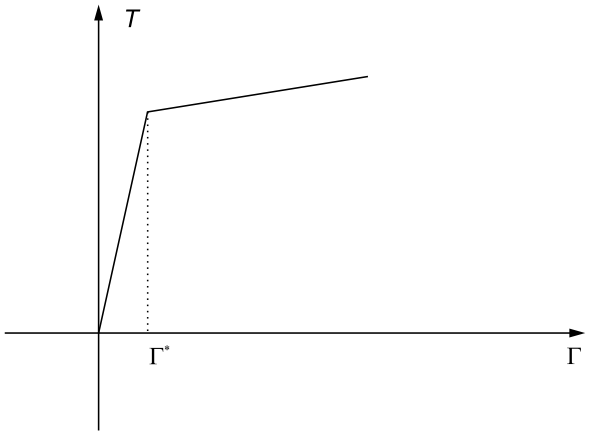

Fig. 1. The dependence between the strain and stress intensity

This equation is used for error indicator calculation:

$$e_i = \tilde{\Delta} u_h(P_i) - f(x,y) \quad \text{at } i\text{-th node} \tag{3}$$

where $\tilde{\Delta}$ is the finite difference approximation of Δ. In this case of the Poisson equation. This problem is equivalent to search for the stationary point of the following functional:

$$I(u) = \int_{\Omega} u_x^2(x,y) + u_y^2(x,y))\, d\Omega. \tag{4}$$

2.2. The elastic-plastic twisting of bars with hardening

In this section the elastic-plastic problem of twisting of bars with hardening is formulated. According to [3] the problem can be led to search for the extremum of the following functional:

$$I(u) = \iint_{\Omega} \left[\int_0^T sg(s)ds - 2\omega u \right] d\Omega, \tag{5}$$

where T is the stress intensity:

$$T = \sqrt{\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2}, \quad \tau_{13} = \frac{\partial u}{\partial x}, \quad \tau_{13} = -\frac{\partial u}{\partial y}.$$

The function g defines the dependence between the effective stress and the effective strain: $T = g(\Gamma)\Gamma$ (Fig. 1), where $\Gamma = \sqrt{\varepsilon_{ij}\varepsilon_{ij}}$, ε_{ij} is the strain tensor and ω is the angle of the torsion.

After the substitution $s = \sqrt{r}$, it is obtained:

$$I(u) = \iint_{\Omega} \left[\int_0^{T^2} sg(\sqrt{s}) \frac{1}{2} ds - 2\omega u \right] d\Omega. \tag{6}$$

In both problems the current function u varies in the Sobolev space

$$H^1(\Omega) = \left\{ v \in L^2(\Omega), \frac{\partial v}{\partial x_i} \in L^2(\Omega), i = 1,2 \right\}.$$

In both problems the current function u varies in the Sobolev [3] space.

For the sake of the approximation the finite dimensional space of functions is defined:

$$V^0 = \left\{ v : \bigcup_{i=0}^{N_T} \bar{T}_i \to \Re,\ v \ is\ continuos,\ v|_{\bar{T}_i} \in P_1 \right\},$$

where $T = \{T_i : i = 1,\ldots,N_T\}$ is a set of non-intersecting triangles covering the domain. For the finite element approximation the approximate solution is defined as finite linear combination of basis functions [10] of the space V^0. The unknown coefficients of the linear combination are found by solving the nonlinear system of algebraic equations, obtained from the stationarity condition [6].

3. The unstructured grid generation with mesh size function in arbitrary domains

Grid generation with arbitrary mesh size function is performed using a 2-D generator [5, 6]. The main idea of grid generation is based upon the algorithm of the advancing front technique and a generalization of the Delaunay triangulation [5, 8] for wide class of $2 - D$ domains. It is assumed that the domain is multiconnected with an arbitrary number of internal loops. The boundary of the domain may be composed of the following curves:

– A straight line segment,
– An arc of circle,
– A B-spline curve.

In case of the advancing front technique combined with the Delaunay triangulation the point insertion and triangulation can be divided into the following steps:

1. Point generation on the boundary,
2. Internal point generation by the advancing front technique,
3. Delaunay triangulation of the previously obtained set of points,
4. The Laplacian smoothing of the obtained mesh.

The algorithm for boundary point generation depends upon the type of boundary segment: [5].

4. Algorithm of remeshing

The whole algorithm of the adaptation is realized in the successive generation of a sequence of meshes $\{T_\nu\}$, where $\nu = 0,1,2,\ldots$ with a modified mesh size function. By using every mesh of the sequence the approximate solution to the problem is obtained and then appropriate error indicators at each node are calculated. Having values of errors at nodes a continuous error function in the whole domain is constructed by using piecewise

linear interpolation at all elements. The error function is appropriately transformed to obtain a multiplier for mesh size function.

The proposed approach gives the possibility to solve the considered problem on well--conditioned meshes and to obtain optimal graded meshes.

4.1. Remeshing scheme

The algorithm of remeshing [4, 13] can be divided into the following steps:

1. Preparation of the information about the geometry and boundary conditions of the problem to be solved,
2. Arrangement of an initial mesh size function,
3. Mesh generation with mesh size function,
4. Solution to the considered on the generated mesh,
5. Evaluation of error indicator at each node,
6. Definition of the new mesh size function by using the errors found at every point of the computational grid,
7. If the error not small enough go to the point 3,
8. End of computations.

In the examples solved by the author it was sufficient to make from 5 to 9 steps of adaptation.

4.2. Error indicators

The applied error indicators are calculated directly for every node, not in elements like in [6, 9].

Let e_i for $i = 1,\ldots,n_P$ be an error indicator at i-the apex of the mesh $\tilde{T}_\nu$, and $\tilde{P}_\nu = \{P_i,\ i = 1,\ldots,n_P\}$ set of nodes. We define a patch of elements for every node P_i as:

$$L_i = \{k,\ P_i \in \bar{T}\} \quad \text{for} \quad i = 1,\ldots,n_P \tag{7}$$

where T_P is the k-th element of the mesh.

1. The first proposed error indicator is biased on the discretized form of the equation (1). At every node partial derivatives are found according to the following recipe:

Having found

$$\frac{\partial u_h}{\partial x}(P_i) = \frac{\sum_{k\in L_i} \frac{\partial u}{\partial x}(P_i)\, area(T_k)}{\sum_{k\in L_i} area(T_k)}. \tag{8}$$

where u_h^k is the restriction of the solution u_h to the k-th element is a linear combination of shape functions of k-th element, then:

$$u_h^k = \sum_{j=0}^{n_e} \lambda_j N_j^k, \quad \text{which gives} \quad \frac{\partial u_h}{\partial x} = \sum_{j=0}^{n_e} \lambda_j \frac{\partial (N_j^k)}{\partial x}, \tag{9}$$

where $u_h = N_j^k$ is a shape function of the *k*-th element. Formula 9 is applied at nodal points. The derivatives $\frac{\partial u_h}{\partial x}(P_i)$, $\frac{\partial u_h}{\partial y}(P_i)$, $i = 1,\ldots,N_P$ found in that way are used for calculation of second order derivatives at nodes in the similar way by using the recurrent formula:

$$\frac{\partial^2 u_h}{\partial x^2}(P_i) = \frac{\partial}{\partial x}\left(\frac{\partial (u_h)}{\partial x}\right)(P_i). \tag{10}$$

In the similar way it is possible to calculate the derivatives of arbitrary order and put them into formula 2 to obtain the value of the error indicator at *i*-th node.

2. In this case it is suggested to evaluate directly derivatives values of error indicator at every node in the following way:

$$T = \sqrt{\sum_{k\in L_i, l\in L_i, l\neq k}\left(\frac{\partial u_i}{\partial x} - \frac{\partial u_k}{\partial x}\right)^2 + \left(\frac{\partial u_i}{\partial y} - \frac{\partial u_k}{\partial y}\right)^2}. \tag{11}$$

where L_i is the set of elements meeting at *i*-th node.

4.3. Error indicators

The modification of the mesh size function is performed at every adaptation step for the realization of the next one. The main idea of this part of the algorithm relies on a multiplication of the values of the mesh size function by an appropriately chosen function. The chosen function should be continuous, linear and should have the smallest value at the node where the value of the error indicator is maximal and the greatest where the value of the error is maximal. It should increase when the error decreases.

The error indicators $\tilde{e}_k$ are calculated at each node of the current mesh, then the minimal and maximal values of the error are found:

$$\alpha = \min_{k=1,2,..,N_{NOD}} \tilde{e}_k, \quad \alpha = \max_{k=1,2,..,N_{NOD}} \tilde{e}_k, \tag{12}$$

where N_{NOD} is the number of nodes. Certainly, $\alpha \le \tilde{e}_k \le \beta$ for $k = 1,2,\ldots,N_{NOD}$.

The following values are introduced:

λ – a value indicating the greatest mesh size reduction.

μ – a value indicating the greatest mesh size reduction.

The values of λ and μ usually should be greater than 0.5, which means that the mesh size does not change too rapidly, which would have an influence on mesh quality in the vicinity, where there are big errors. Usually it is assumed that λ varies from 0.5 to 0.6 and μ from 0.8 to 1.0.

The following affine transformation is defined:

$$l : [\alpha,\beta] \to [\mu,\lambda], \tag{13}$$

which satisfies the conditions $l(\alpha) = \lambda$ and $l(\alpha) = \mu$. By these assumptions it can be observed that $\mu \le l(x) \le \lambda, \ \forall x \in [\alpha, \beta]$.

Provided that

$$Q_i = l(\tilde{e}_i) \le \lambda, \ \forall i = 1, \ldots, N_{NOD}, \tag{14}$$

then we have

$$\min_{i=1,2,\ldots,N_{NOD}} Q_i = \mu, \qquad \max_{i=1,2,\ldots,N_{NOD}} Q_i = \lambda.$$

Introducing the function $r : \bar{D} \to \Re$, as follows: $r(x) = \Pi(x)$, if $x \in \bar{T}_s$, where Π is an affine mapping of two variables satisfying the following equalities:

$$\Pi(P_i) = Q_i, \quad \text{for} \quad i = 1, 2, 3, \tag{15}$$

where P_1, P_2, P_3 are the vertices of the triangle T_s of the triangulation of Ω, and appropriately Q_1, Q_2, Q_3 are the values defined by the formula (14). The function $r(x)$ is defined in the whole domain because the triangles $\{\bar{T}_s\}_{s=1}^{N_T}$ cover it. The new mesh size function is defined as follows:

$$\gamma_{i+1}(x) = \gamma_i(x) r(x). \tag{16}$$

As $\mu \le r(x) \le \lambda$ then $\mu\gamma_i(x) \le \gamma_{i+1}(x) \le \lambda\gamma_i(x)$.

It can be checked that: $\exists x, \ y \in \bar{\Omega}$ such that:

$$\mu\gamma_i(x) = \gamma_{i+1}(x) \quad \text{and} \quad \gamma_{i+1}(x) = \lambda\gamma_i(x). \tag{17}$$

It can be shown, that $\|\gamma_{i+1} - \gamma_i\|_{\bar{\Omega},\max} \le \|\gamma_i\|_{\bar{\Omega},\max} \max\{|1-\lambda|, |1-\mu|\}$, where

$$\|\gamma\|_{\bar{\Omega},\max} = \max_{x \in \bar{\Omega}} \{|\gamma(x)|\}. \tag{18}$$

5. Numerical examples

The manner of size function modification depends on the error indicator and on the coefficients λ, μ, which determine the details of the mesh size function modification. If the values of the coefficients λ, μ, are small then fewer adaptation steps is necessary. How quickly an adapted grid will be close enough to an optimal mesh, besides of error indicator function, depends on the initial mesh too. For the solved problems it was assumed that $\lambda = 0.6$ and $\lambda = 0.8$, which caused performing greater number of adaptation steps, what may lead to a better solution. In the plasticity theory problems it can observed (figures 5, 6), that the adapted mesh densities at the border between elastic and plastic zones and the adapted mesh (Fig. 5) coincide with the sand heap analogy [3]. It would be rather impossible to obtain the effect by the methods based on mesh enrichment [1, 9].

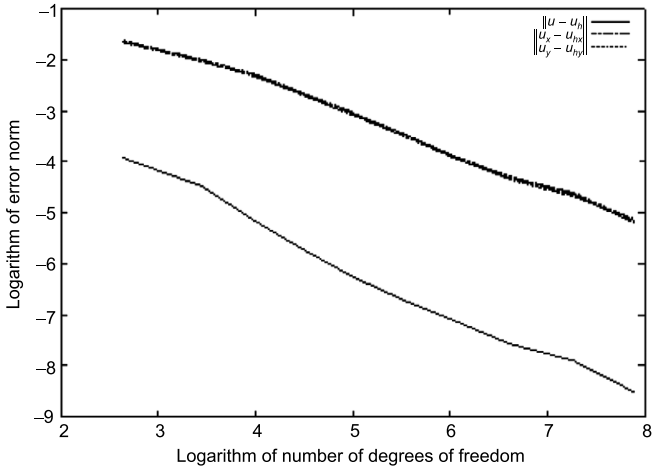

Fig. 2. The convergence curves for u, u_x, u_y for problem 1 with respect to the norms

For the sake of numerical rate of the convergence of the proposed method for the problem defined in 4 the function f was defined in the way that the solution to the problem is the function [13]: $u(x,y) = x(1-x)y(1-y)\arctan\left(a\left(\frac{x+y}{\sqrt{2}} - \xi\right)\right)$, where a = 20 and ξ = 0.8. The figure 2 presents the dependence between number of nodes and norms $\|u-u_h\|$, $\left\|\frac{\partial u}{\partial x} - \frac{\partial u_h}{\partial x}\right\|$ and $\left\|\frac{\partial u}{\partial y} - \frac{\partial u_h}{\partial y}\right\|$. The graphs of the norms $\left\|\frac{\partial u}{\partial x} - \frac{\partial u_h}{\partial x}\right\|$ and $\left\|\frac{\partial u}{\partial y} - \frac{\partial u_h}{\partial y}\right\|$ almost cover each other.

The figure 3 presents the adaptive mesh.

It can seen that the mesh for the example problem 1 and its strict solution 5 coincide.

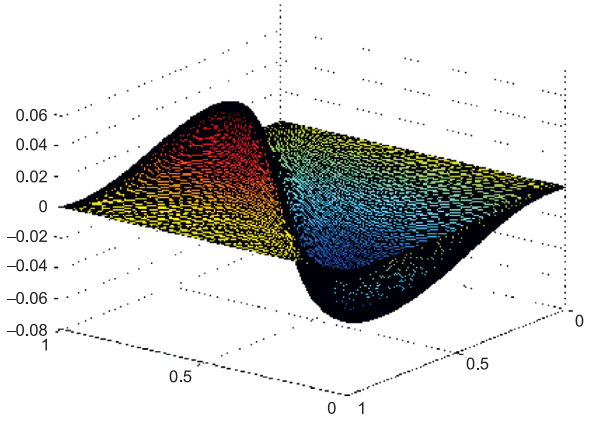

Fig. 3. Strict solution for the problem 1

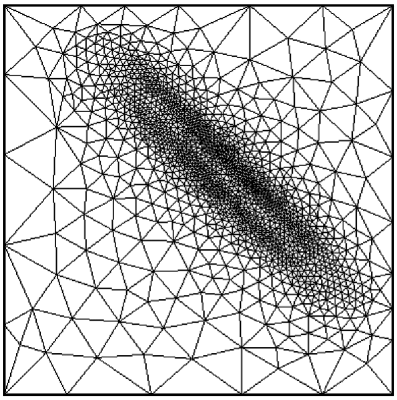

Fig. 4. Adapted mesh for the problem 1

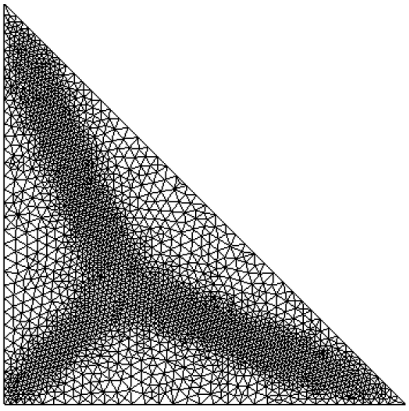

Fig. 5. Final mesh after 7 adaptation steps for problems 6

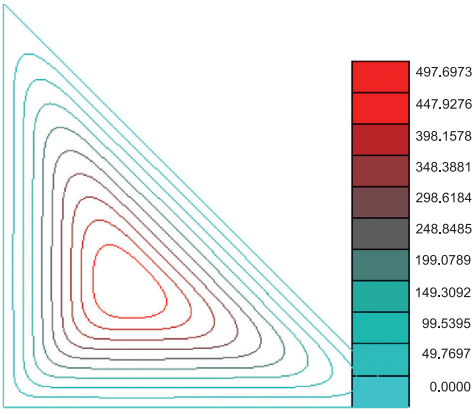

Fig. 6. Adapted mesh for the problem 6

6. Summary

- New error indicators based on generalized finite difference method were introduced applied to the proposed adaptive remeshing.
- The numerical rate of the convergence was calculated by using the known strict solution.
- The optimal mesh size function is obtained iteratively and depends on values of error indicators at nodes.
- The generator based on Delaunay condition and advancing front technique seems very suitable to the class of problems where different zones of the domain are to be appointed.
- For further investigations the anisotropic mesh generation algorithm will be developed an appropriate anisotropic adaptation algorithms as well too.

References

[1] Bank R., Sherman A., Wieser A., *Some Refonement Algorithms and Data Structures for regular local mesh refinement*, Scientific Computing, IMACS, 1983.

[2] Huerta A., Diez P., *Error estimation including pollution assessement for nonlinear finite element analysis*, Comp. Meth. Appl. Mech. Engn., vol. 181, 2000, 21-24.

[3] Kachanov L.M., *Fundamentals of plasticity theory, Dover Publications Inc*., 2004, 479 p. (ISBN 0-486-43483-0), Mineola, NY, USA, Moscow, 1968.

[4] Kucwaj J., *The Algorithm of Adaptation by using graded meshes generator*, Computer Assisted Mechanics and Engineering Sciences, vol. 7, 2000, 615-624.

[5] Thompson J.F., Soni B.K., Weatherwill N.P., *Handbook of grid generation*, CRC Press, Boca Raton, 1999.

[6] Kucwaj J., *Numerical Investigation of the convergence of remeshing algorithm na an axample of subsonic flow*, Computer Assisted Mechanics and Engineering Sciences, vol. 17, 2010, 147-160.

[7] Kucwaj J., *Adaptive unstructured solution to the problem of elastic-plastic hardening of prismatic bars*, Technical Transactions, vol. 111, 2014, 63-79.

[8] Lo S.H., *Finite element mesh generation and adaptive meshing*, Progress in Structural Engineering and Materials, vol. 4, 2002, 381-399.

[9] Oden J.T., Demkowicz L., Rachowicz W., Westermann T.A., *Towards a universal h-p adaptive finite element strategy, part 2, aposteriori error estimation*, Comp. Meth. Appl. Mech. Engn., vol. 77, 1989, 113-180.

[10] Zienkiewicz O.C., Taylor R.L., *The finite element method*, 4-th edition, vol. 1, basic formulation and linear problems, McGraw-Hill Book Company, London, Washington, 1989.

[11] Zienkiewicz O.C., *Achievements and some unsolved problems of the finite element method*, Int. J. Num. Meth. Engn., vol. 47, 2000, 9-28.

[12] Zienkiewicz O.C., Zhu J.Z., *Adaptavity and mesh generation*, Int. J. Num. Meth. Engn., vol. 32, 1991, 783-810.

[13] Madlib: An open source mesh adaptation library, http://sites.uclouvain.be/madlib/, 2010.

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

ARTUR NIEWIAROWSKI*

SHORT TEXT SIMILARITY ALGORITHM BASED ON THE EDIT DISTANCE AND THESAURUS

ALGORYTM PODOBIEŃSTWA KRÓTKICH FRAGMENTÓW TEKSTÓW OPARTY NA ODLEGŁOŚCI EDYCYJNEJ I SŁOWNIKU WYRAZÓW BLISKOZNACZNYCH

Abstract

This paper proposes a method of comparing the short texts using the Levenshtein distance algorithm and thesaurus for analysing terms enclosed in texts instead of popular methods exploiting the grammatical variations glossary. The tested texts contain a variety of nouns and verbs together with grammatical or orthographical mistakes. Based on the proposed new algorithm the similarity of such texts will be estimated. The described technique is compared with methods: Cosine distances, distance Dice and Jaccard distance constructed on the term frequency method. The proposition is competitive against well-known algorithms of stemming and lemmatization.

Keywords: *Levenshtein distance algorithm*, *the edit distance*, *thesaurus*, *the measure of texts similarity*, *plagiarism detection, text mining*, *Natural Language Processing*, *Natural Language Understanding*, *stemming*, *lemmatization*

Streszczenie

Artykuł przedstawia propozycję metody porównywania krótkich fragmentów tekstów bazującą na algorytmie odległości Levenshteina i słowniku wyrazów bliskoznacznych. Porównywane teksty zawierają odmienione terminy oraz celowe błędy ortograficzne i gramatyczne. Opisany mechanizm zestawiony został z popularnymi metodami porównywania tekstów, takimi jak: odległości Kosinusowa, Dice'a i Jaccard'a, dla których wartości wektorów obliczane są metodą częstości terminów. Zastosowanie w mechanizmie słownika wyrazów bliskoznacznych jest alternatywą wobec znanych algorytmów określania rdzenia terminu i lematyzacji w analizie danych tekstowych.

Słowa kluczowe: *odległość Levenshteina*, *odległość edycyjna*, *słownik wyrazów bliskoznacznych*, *miara podobieństwa tekstów*, *detekcja plagiatu*, *analiza danych tekstowych*, *przetwarzanie języka naturalnego*, *stemming*, *lematyzacja*

DOI: 10.4467/2353737XCT.16.149.5760

* Artur Niewiarowski (aniewiarowski@pk.edu.pl), Institute of Computer Science, Faculty of Physics, Mathematics and Computer Science of Cracow University of Technology.

1. Introduction

Most of the mechanisms estimating measures of similarity between text documents are based on vector space models and weight methods [1, 2, 3] with compilation of other methods, such as probability methods [4], semantic networks (e.g. WordNet) [5, 6] or genetic algorithms [7], etc. The traditional text identification mechanisms usually use glossary of grammatical variations which in some cases are difficult to implement. Most of the mechanisms are composed of stemming algorithms [8], such as popular: Lovins stemming algorithm [9], Porter stemming algorithm [10], Dawson stemming algorithm, Krovetz stemming algorithm, etc. [11], causing increasing of time consumed by identification algorithm during the data analysing process.

In our research we propose a model of effective mechanism for the calculation of similarity between two short texts (e.g. sentences) mainly based on Levenshtein distance algorithm (Lda) combined with the word coding technique. Our research indicates that the terms coding technique with its implementation for measure of text similarity improves the results of text identification significantly. The proposed technique seems to be easy for implementation in most programming technologies and open to most European languages.

2. Description of the problem

The main idea of the mechanism of measuring the similarity of texts consists of:

– implementation of function of terms coding based on Levenshtein distance [16] (*presented in subsection 2.2*) and thesaurus (*described in subsection 3.2*),
– calculation of similarity measure between two sentences based on Levenshtein distance between encoded terms.

2.1. Levenshtein distance algorithm

The concept of the Levenshtein distance algorithm (*Levenshtein Distance function*) may be depicted by the following pseudo-code:

Pseudo-code 1

```
input variables: char Text1[0..M-1], char Text2[0..N-1]
     declare: int d[0..M, 0..N]
     for i from 0 to M
      d[i, 0] := i
     for j from 0 to N
      d[0, j] := j

     for i from 1 to M
      for j from 1 to N

       if char of Text1 at (i - 1) = char of Text2 at (j - 1) then
                    cost := 0 else cost := 1
       end if
                    d[i, j] :=
                        Minimum(d[i - 1, j] + 1,
```

```
                    d[i, j - 1] + 1,
                    d[i - 1, j - 1] + cost)
      end for (variable j)
    end for (variable i)

return d[M, N];
```

where:

- d – Levenshtein matrix of the size N+1, M+1, formed for two terms: Text1 and Text2,
- M, N – lengths of two terms respectively,
- d [i, j] – (i, j) – element of Levenshtein matrix d,
- min – a function to calculate minimum of three variables,
- cost – variable that gets values either 0 or 1

The Levenshtein distance K is the minimum number of operations (insertion, deletion, substitution) required to change one term into the other

$$K = d(M, N)$$

2.2. The measure of similarity between terms

Measure of similarity P is the quotient of number of Levenshtein operations (*after calculation of Lda*) by the number of all Levenshtein operations in pessimistic case. This means, before the calculations of Lda will be completed but with the maximum possible number of Levenshtein operations well known.

The similarity measures P is calculated by the formula:

$$P = 1 - \left(\frac{K}{K_{\max}} \right); \quad K_{\max} = \max(N, M), \quad \begin{matrix} K \geq 0,\ M > 0,\ N > 0 \\ P \in \langle 0,1 \rangle \end{matrix} \tag{1}$$

where:

- $K_{\max}$ – the length of the longest of analysed two terms/text strings (i.e. pessimistic case where K is equal to the length of the longest term).

Table 1

Examples of the Levenshtein distance and the measure of similarity between two short texts

No.	Text1	Text 2	*K*	*P*
1.	Car	Cars	1	0.75
2.	University	Universities	3	0.75
3.	Tom is writing a letter	Tom is writin letters	4	0.82

2.3. The algorithm to measure similarity between two sentences

The algorithm for measuring of similarity between two sentences, based on Lda, is described by the formula (2) presented below:

$$\mathop{\mathrm{I}}_{i_S=1}^{N_S}\mathop{\mathrm{I}}_{j_S=1}^{M_S} d_S(i_S, j_S) = \min\left(d_S(i_S-1, j_S)+1, d_S(i_S, j_S-1)+1, d_S(i_S-1, j_S-1)+\beta_S\right)$$

$$\begin{cases} \beta_S = 0 : \Lambda(a_S(i_S), b_S(j_S)) >= q \\ \beta_S = 1 : \Lambda(a_S(i_S), b_S(j_S)) < q \\ d_S(i_S, 0) = i_S \\ d_S(0, j_S) = j_S \\ d_S(0,0) = 0 \end{cases} \tag{2}$$

where:

I	–	symbol for iteration (for loop presented in the pseudo-code 1)
N_S+1, M_S+1	–	matrix sizes made from two sentences,
d_S	–	matrix made from two sentences,
$d_S(i_S,j_S) - (i_S,j_S)$	–	element of matrix d_S,
Λ	–	function which returns the measure of similarity between two terms P, calculated based on Levenshtein distance algorithm (pseudo-code 1), few terms creates sentence,
β_S	–	variable: 0 or 1,
$a_S(i_S)$– i_S	–	term of sentence a_S,
$b_S(j_S) - j_S$	–	term of sentence b_S
q	–	acceptable boundary of value of similarity measure for two texts (terms in this case). This value is set by the user and it depends on data (e.g. texts from old books, texts from newspapers).

The asymptotic computational complexity of the algorithm is $O(n^4)$. This derives from the construction of the algorithm which consists of four nested loops[1].

The similarity measures P_S between two sentences may be estimated in the rule:

$$P_S = 1 - \left(\frac{K_S}{K_{S\max}}\right); \quad K_{S\max} = \max(N_S, M_S), \quad \begin{matrix} K_S \geq 0,\ M_S > 0,\ N_S > 0 \\ P_S \in \langle 0,1 \rangle \end{matrix} \tag{3}$$

Table 2

Examples of the Levenshtein distance and the measure of similarity between two sentences in whose terms are treated as chars

No.	Sentence 1	Sentence 2	K_S	P_S	q
1.	My car isn't working	My bicycle isn't working	1	0,75	1; 0,75; 0,3
2.	What did you do yesterday?	What have you done?	3	0,40	1; 0,6
3.	What did you do yesterday?	What have you done?	3	0,60	0,5; 0,1
4.	Tom is writing a letter	Tom is writin letters	3	0,40	1; 0,9
5.	Tom is writing a letter	Tom is writin letters	3	0,80	0,85; 0,05

[1] Interesting research about parallelization of the Levenshtein distance algorithm (and Levenshtein--Damerau distance algorithm) for accelerate the calculations is described in [12].

3. Procedure of Terms Coding based on thesaurus

On the figures and tables below concept of the coding process is described. Tables 3–5 include sample data to be encoded. Formulas 4–7 describe all steps of the coding process.

3.1. The data model

The database model of thesaurus (τ) with tables of unique terms, groups of terms and table of terms associations is presented below.

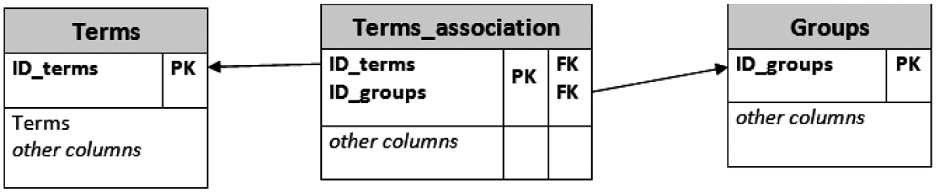

Fig. 1. The database model of thesaurus (τ)

Example of use of the model above:

Table 3

Terms

ID_terms	Terms
1	Tom
2	Mary
3	John
4	car
5	auto
6	vehicle

Table 4

Terms_association

ID_terms	ID_groups
1	1
2	1
3	1
1	2
2	2
4	4
5	5
6	6

Table 5

Groups

ID_groups	Describe
1	names
2	my best friends
3	vehicles

It is easy to see that a term can belong to a few groups. The example shows that *Tom* and *Mary* can belong to the groups: *names* and *my best friends*. It means that *Tom*, *Mary* and *John* are the same terms (names) because they have the same meaning. Following terms: *car, auto, vehicle* are the same terms – *vehicle*.

3.2. The coding process

The process of common group analysis proceeds with the following steps:

1. Get all codes of each terms of both texts (a_S, b_S) from thesaurus (e.g. array of codes for cases where one term can has a multiple meanings).

$$\Psi(a_s, \tau, q) \rightarrow \mathbf{C}_{\mathbf{a_S}}(i_S, t_{i_S}) \quad \text{and} \quad \Psi(b_s, \tau) \rightarrow \mathbf{C}_{\mathbf{b_S}}(j_S, t_{j_S}) \tag{4}$$

where:

Ψ – function to get codes of terms,

τ – thesaurus,

t_{i_S} – the number of term's codes variants,

$\mathbf{C}_{\mathbf{a_S}}$ – array of codes of terms of sentence a_S.

2. Calculate number of occurrences of codes terms based on their frequency in texts (i.e. matching process of common meanings).

$$\Gamma(\mathbf{C}_{\mathbf{a_S}}, \mathbf{C}_{\mathbf{b_S}}) \rightarrow \mathbf{C}_{\mathbf{ab_S}}(h) \tag{5}$$

where:

Γ – function which calculates frequency of occurrences of codes,

$\mathbf{C}_{\mathbf{ab_S}}$ – array of the best codes of terms,

h – ID of the term.

3. Replace each term in both texts with the most frequent code

$$\Phi(\mathbf{C}_{\mathbf{ab_S}}, \mathbf{C}_{\mathbf{a_S}}) \rightarrow \mathbf{NC}_{\mathbf{a_S}}(i_S) \quad \text{and} \quad \Phi(\mathbf{C}_{\mathbf{ab_S}}, \mathbf{C}_{\mathbf{b_S}}) \rightarrow \mathbf{NC}_{\mathbf{b_S}}(j_S) \tag{6}$$

where:

Φ – function which replaces the most frequent code,

$\mathbf{NC}_{\mathbf{b_S}}$ – new sentence with encoded terms.

4. Function which calculates similarity K_S between two sentences (short texts).

$$\Omega(\mathbf{NC}_{\mathbf{a_S}}, \mathbf{NC}_{\mathbf{b_S}}, q, q_\tau) \rightarrow P_S \tag{7}$$

where:

Ω – function which calculates similarity K_S measures between sentences,

q_τ – acceptable border value of similarity measure for two terms – between term includes in text and term derives from thesaurus.

4. Verification of proposed similarity measures mechanism

The following tests show how term coding methods improve the mechanism of similarity between two short texts. As a test of the proposed algorithms, 170 pars of correct and incorrect sentences written in various tenses were checked. 10% of interesting sentences were chosen and described below. Some examples based on the three popular languages of Eastern Europe are provided in Appendix 1.

The terms and sentences used for the tests were presented in the tables below:

1. synonyms (thesaurus) (Table 6)
2. grammatically and spelling correct sentences written in various tenses (Table 7)
3. grammatically and spelling incorrect sentences written based on the correct sentences (Table 7)
4. encoded texts based on the thesaurus (Table 8)
5. results of tests (Table 9)

For all tests and in all cases the acceptable boundaries of similarity P are: $q_\tau = 0.80$ for **thesaurus** and $q = 0.75$ for similarity between **terms** in sentences (formula 2).

Table 6

Example of the thesaurus schema with terms groups by common ID (code in text). Similar term from column *Describe* in analysing text will be replaced by ID

ID of groups	Describe (e.g. name of the group)	Terms
#1	names	Tom, Mary, John, Jimmy, Jane, Derek, Gina
#2	cars	car, auto, automobile, taxi, vehicle
#3	numbers	one, two three, four, five, six, seven, eight, nine, ten
#4	seasons	spring, summer, autumn, winter
#5	fruits	apple, pear, cherry, mango, kiwi, watermelon
#6	cities	Warsaw, Berlin, London
#7	phones	phone, telephone, iPhone, mobile phone
#8	very	very, extremely
#9	shortcuts1	is not, isn't
#10	shortcuts2	are not, aren't
#11	shortcuts3	don't, do not
#12	fluid	milk, water
#13	my_friends	Tom, Jack, Ella, Olivia

Table 7

Correct and incorrect sentences for the tests

No.	Correct sentences	Incorrect sentences with synonyms
s1	Tom is writing a letter	Dere is writin a letters
s2	We are waiting for a taxi	We are waitin for car
s3	Is Mary having breakfast?	Is Jane hasing brekfest?
s4	Tom is not writing a letter	Jimm isn't writin leter
s5	He isn't looking at the stars	He is not look at the start
s6	He drinks milk twice a day	He is drinks water twice a day
s7	We go to work six times a week	We goes to works seven times a week
s8	I always feel great in spring	I alway feel great in summer
s9	Do you like apples?	Does you likes pear?
s10	I don't like milk	I do not likes water
s11	Tom was writing the letter all day yesterday	Jimmy writting the leter all day yestaredy
s12	They met when they were studying in Berlin	They met when they were studying in Warsaw
s13	I was working in London this time last year	I was work in Berlin this times last years

s14	I have found his telephone number	I have found his phone number
s15	I was shocked when I found out that Derek and Gina had got divorced	I was shock when I found out that John and Mary has gotten divorced
s16	I have been working for five hours	I has been working for six hour
s17	It had been raining for days so when they finally left, the roads were very muddy	It has been raining for days so when they finaly left, the roads were extremly muddy

Table 7 shows the correct sentences with synonyms in the left column and incorrect sentences with synonyms in the right column. Synonyms (not all) include mistakes, like for example Dere instead of Derek in the first row.

Table 8

Correct and incorrect sentences after terms coding (based on the thesaurus)

No.	Correct sentences after terms coding	Incorrect sentences with synonyms after terms coding method
s1	#1 is writing a letter	#1 is writin a letters
s2	we are waiting for a #2	we are waitin for #2
s3	is #1 having breakfast?	is #1 hasing brekfest?
s4	#1 #9 writing a letter	#1 #9 writin leter
s5	he #9 looking at the stars	he #9 look at the start
s6	he drinks #12 twice a day	he is drinks #12 twice a day
s7	we go to work #3 times a week	we goes to works #3 times a week
s8	i always feel great in #4	i alway feel great in #4
s9	do you like #5?	does you likes #5?
s10	i #11 like #12	i #11 likes #12
s11	#1 was writing the letter all day yesterday	#1 writting the leter all day yestaredy
s12	they met when they were studying in #6	they met when they were studying in #6
s13	i was working in #6 this time last year	i was work in #6 this times last years
s14	i have found his #7 number	i have found his #7 number
s15	i was shocked when i found out that #1 and #1 had got divorced	i was shock when i found out that #1 and #1 has gotten divorced
s16	i have been working for #3 hours	i has been working for #3 hour
s17	it had been raining for days so when they finally left, the roads were #8 muddy	it has been raining for days so when they finaly left, the roads were #8 muddy

Table 8 includes sentences from table 7 after coding by the proposed algorithm. The similarity of these sentences was calculated with (1)(2) and presented below.

Table 9

Values of similarity of the sentences with and without the terms coding method based on Levenshtein distance algorithm. Description of columns: Col. 1 – Similarity between correct and incorrect sentences without methods of the: similarity measure between terms; coding terms (based on Table 7 data); Col. 2 – Similarity between correct and incorrect sentences (without using method of terms coding) based on Table 7 data; Col. 3 – Similarity between correct and incorrect sentences (with using method of terms coding) based on Table 8 data; No – number of sentence

No.	Col. 1	Col. 2	Col. 3
s1	0.40	0.80	1.00
s2	0.50	0.67	0.83
s3	0.25	0.75	1.00
s4	0.00	0.33	0.80
s5	0.43	0.57	0.83
s6	0.71	0.71	0.86
s7	0.62	0.75	0.88
s8	0.67	0.83	1.00
s9	0.25	0.50	0.75
s10	0.20	0.40	1.00
s11	0.38	0.62	0.75
s12	0.88	0.88	1.00
s13	0.56	0.78	0.89
s14	0.83	0.83	1.00
s15	0.64	0.71	0.86
s16	0.57	0.71	0.86
s17	0.81	0.88	0.94

The obtained results show that the method of coding terms (column no. 3) increases the precision of similarity estimation in some cases from 0–20% even up to 75–100%.

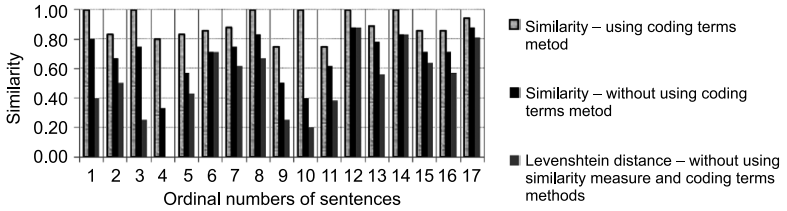

Fig. 2. Graphical results of quality test of English sentences. For all tests in this case acceptable boundaries of similarity P are: $q = 0.75$, $q_\tau = 0.80$

5. Comparison quality of described method with popular methods

Results in table 10 (and in tables 11–16 in Appendix 1) show that the similarity methods based on Levenshtein distance algorithm (i.e. Lda without coding terms method and Lda with coding terms method – Table 9/Col. 3) are more precise than popular methods like: Dice distance, Jaccard distance or Cosine distance [17]. Formulas (8–12) refer to these distances.

Dice distance is described by the formula below:

$$\text{Dice\_dist}(a_i, b_k) = 2\frac{\sum_{j=1}^{n} a_{ij} b_{kj}}{\sum_{j=1}^{n} a_{ij}^2 \sum_{j=1}^{n} b_{kj}^2} \tag{8}$$

where:

a_i – weight of the term in i position of the vector of text document a,

n – length of the two vectors created from two text documents a and b.

Jaccard distance is described by the formula below:

$$\text{Jaccard\_dist}(a_i, b_k) = \frac{\sum_{j=1}^{n} a_{ij} b_{kj}}{\sum_{j=1}^{n} a_{ij}^2 + \sum_{j=1}^{n} b_{kj}^2 - \sum_{j=1}^{n} a_{ij} b_{kj}} \tag{9}$$

Cosine distance is described by the formula below:

$$\text{Cosine\_dist}(a_i, b_k) = \frac{\sum_{j=1}^{n} a_{ij} b_{kj}}{\sqrt{\sum_{j=1}^{n} a_{ij}^2} \sqrt{\sum_{j=1}^{n} b_{kj}^2}} \tag{10}$$

Weights are calculated by special formulas, i.e. *term frequency method* (*tf*) or *term frequency – inversed document frequency method* (*tf – idf*) [18].

The *idf* method is described by the formula below:

$$idf_t = \log \frac{N}{df_t} \tag{11}$$

where:

N – the number of analysed documents,

df_t – the number of documents where term t occurs.

The *tf – idf* method is described by the formula below:

$$tf - idf_{t,d} = \log \frac{N}{df_t} \times tf_{t,d} \tag{12}$$

where:

$tf_{t,d}$ – the number of times that term t occurs in document d.

Table 10

Values of similarity of the sentences using popular methods. Describe of columns (experiments): Col. 1 – Similarity method based on Lda without coding terms method; Col. 2 – Cosine distance based on term frequency weight method (tf); Col. 3 – Dice distance based on tf weight method; Col. 4 – Jaccard distance based on tf weight method; No. – number of sentence. Experiments 2–4 based on Table 8 data (i.e. method of coding synonym terms was used)

No.	Col. 1	Col. 2	Col. 3	Col. 4
s1	0.80	0.40	0.08	0.25
s2	0.67	0.55	0.10	0.37
s3	0.75	0.25	0.06	0.14
s4	0.33	0.00	0.00	0.00
s5	0.57	0.46	0.07	0.30
s6	0.71	0.77	0.12	0.62
s7	0.75	0.63	0.07	0.45
s8	0.83	0.66	0.11	0.50
s9	0.50	0.25	0.62	0.14
s10	0.40	0.22	0.50	0.12
s11	0.62	0.40	0.53	0.25
s12	0.88	0.90	0.00	0.81
s13	0.78	0.56	0.06	0.38
s14	0.83	0.83	0.13	0.71
s15	0.71	0.69	0.04	0.52
s16	0.71	0.57	0.08	0.40
s17	0.88	0.81	0.05	0.68

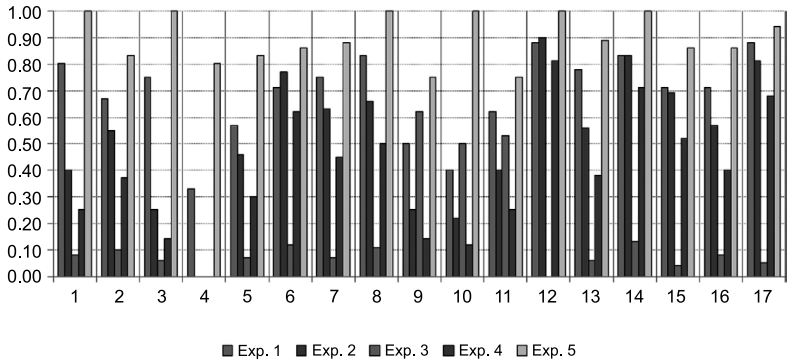

Fig. 3. Graphical results of similarity of the incorrect sentences using popular methods. Description of experiments: Exp. 1 – Similarity method based on Lda without coding terms method; Exp. 2 – Cosine distance based on term frequency weight method (*tf*); Exp. 3 – Dice distance based on *tf* weight method; Exp. 4 – Jaccard distance based on *tf* weight method; Exp. 5 – Similarity between correct and incorrect sentences (with using method of terms coding) based on Table 8 data

As can be seen, the described similarity methods (based on Lda) include a fully different algorithm compared with the popular methods (based on terms weights and vector space models) because of the identification of the distribution of terms. Described popular methods based on the number of occurrences of terms in documents only. Methods based on Lda do not need other documents instead of e.g. term frequency – inversed document frequency method to estimate weight of terms (in case of two sentences impossible to apply). Graphical interpretation includes all described methods is shown below (Figure 3). The best results includes exp. no. 5.

6. Summary

The method of coding terms described in this paper increases the precision of calculation of the similarity measures based on Levenshtein distance significantly. This method is characterized by the speed of data analysis and simplicity of implementation. The coding method of terms in combination with the Levenshtein distance and the similarity measures can be used in: detecting plagiarism (resignation of variety of nouns and verbs based on standard thesaurus and stemming algorithms), finding phrases in text documents [8] (or web documents [13], etc.), algorithms for correcting mistakes, mechanism of identification and classification of content based on term weighted methods [1, 14, 15], etc.

The proposed solutions are applied and have been tested in the mechanism of topic analysis and descriptions of selected written works (diploma thesis) to automatic selection of supervisors and reviewers at the Faculty of Physics, Mathematics and Computer Science of the Cracow University of Technology[2]. The solution also was included in Anti-plagiarism System of Faculty of Physics, Mathematics and Computer Science[3]. Tests results show a high quality of the text mining analysis.

References

[1] Niewiarowski A., *Term frequency optimization for the vector space model*, "Technical Transactions", 9-M/2012, 155-165.

[2] Yih W., Meek Ch., *Improving Similarity Measures for Short Segments of Text*, Microsoft Research, USA 2007.

[3] Long-Scheng Cz., Chia-Wei Ch., *A New Term Weighting Method by Introducing Class Information for Sentiment Classification of Textual Data*, "Proceedings of the International MultiConference of Engineers and Computer Scientists", IMECS 2011, 394-397.

[4] Metzler D., Dumais S., Meek Ch., *Similarity Measures for Short Segments of Text*, Microsoft Research, USA 2007.

[5] Piasecki M., Broda B., *Semantic similarity measure of Polish nouns based on linguistic features*, "Business Information Systems 10th International Conference, Lecture Notes in Computer Science, Springer", BIS 2007, 381-390.

[2] Manager of Diploma Thesis web site: https://dyplomy.fmi.pk.edu.pl

[3] Screenshot of our .NET application (for MS Windows) for analyze plagiarism: www.pk.edu.pl/~aniewiarowski/programy/antyplagius.png

[6] Novay L.G., Novay Ch. W., Brussee R., *Thesaurus Based Term Ranking for Keyword Extraction*, "DEXA'10 Proceedings of the 2010 Workshops on Database and Expert Systems Applications, Computer Society", 2010.
[7] Castillo Sequera J.L., Fernandez del Castillo Diez J.R., Gonzales Sotos L., *A clustering algorithm based on a recursive function of distance and similarity*, "IADIS European Conference Data Mining" 2011, 43-50.
[8] Szwed P., *Conecpts extraction from unstructered Polish texts: a rule based approach*, "Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, Springer", 355-364.
[9] Lovins J.B., *Development of a Stemming Algorithm*, "Mechanical Translation and Computational Linguistics" vol. 11, nos. 1 and 2 1968.
[10] Willett P., *The Porter stemming algorithm: then and now*, "Program", Vol. 40, 219-223.
[11] Abramowicz W., Filipowska A., Małyszko J., Wagner T., *Lemmatization of Multi-Word Entity Names for Polish Language Using Rules Automatically Generated Based on the Corpus Analysis*, "Language and Technology Conference", 2009, 540-544.
[12] Niewiarowski A., Stanuszek M., *Parallelization of the Levenshtein distance algorithm*, "Technical Transactions", 3-NP/2014, 109-122.
[13] Niewiarowski A., *Działanie parsera Part-of-Speech Tagging w ujęciu mechanizmu Web Content Mining*, 6'th National Conference „Science and Industry", 2011, 93-100.
[14] Niewiarowski A., Stanuszek M., *The mechanism of identification and classification of content*, "Studia Informatica", Volume 34, Number 2B (112), 2013, 205-222.
[15] Niewiarowski A., *Mechanism of plagiarism detection based on the variation of the Levenshtein distance algorithm*, 5'th National Conference „Science and Industry", 2010, 86-103.
[16] Левенштейн В.И., *Двоичные коды с исправлением выпадений, вставок и замещений символов*, „Доклады Академий Наук СССР", 163 (4), 1965, 845-848.
[17] Singhal, Amit., *Modern Information Retrieval: A Brief Overview*, "Bulletin of the IEEE Computer Society Technical Committee on Data Engineering", 24 (4), 2001, 35-43.
[18] Rajaraman, A., Ullman, J.D., *Data Mining*, "Mining of Massive Datasets", Cambridge University Press, 2014, 1-17.

Appendix 1

Table 11

Example of Polish sentences. Stars define the same origin (i.e. method of coding synonyms was used)

No.	Correct sentence	Incorrect sentence
s1	Jutro będzie nowy* dzień	Jutro bedzie nowiutki* dzien
s2	Gdy** chcesz opisać prawdę, elegancję pozostaw*** krawcom.	Kiedy** chcesz opisac prawde, elegancje zostaw*** krawcom.
s3	Kto się lęka**** już przegrał	Kto sie boi**** juz przegral

Table 12

Values of similarity of the sentences using popular methods. Description of columns (experiments): Col. 1 – Similarity method based on Lda without coding terms method; Col. 2 – Similarity method based on Lda with coding terms method; Col. 3 – Cosine distance based on term frequency weight method (tf); Col. 4 – Dice distance based on tf; Col. 5 – Jaccard distance based on tf weight method; No. – number of sentence

No.	Col. 1	Col. 2	Col. 3	Col. 4	Col. 5
s1	0.75	1.00	0.25	0.06	0.14
s2	0.86	1.00	0.57	0.08	0.40
s3	0.40	0.60	0.40	0.08	0.25

Table 11 includes correct and incorrect Polish sentences. Sentences have similar meaning but include different terms. Column no. 2 in table 12 includes the best values of the tests.

Table 13

Example of Russian sentences. Stars define the same origin (i.e. method of coding synonyms was used)

No.	Correct sentence	Incorrect sentence
s1	Завтра будет новый* день	Завтра будетъ новенкий* день
s2	Если** вы хотите сказать*** правду, оставьте элегантность портным	Когда** вы хатите рассказать*** правду,оставте элегантность портным
s3	Кто боится уже проиграл	Кто баиться уже праиграл

Table 14

Values of similarity of the sentences using popular methods. Description of columns (experiments): Col. 1 – Similarity method based on Lda without coding terms method; Col. 2 – Similarity method based on Lda with coding terms method; Col. 3 – Cosine distance based on term frequency weight method (tf); Col. 4 – Dice distance based on tf; Col. 5 – Jaccard distance based on tf weight method; No. – number of sentence

No.	Col. 1	Col. 2	Col. 3	Col. 4	Col. 5
s1	0.75	1.00	0.75	0.18	0.60
s2	0.75	1.00	0.75	0.09	0.60
s3	0.75	0.75	0.50	0.12	0.33

Table 13 includes correct and incorrect Russian sentences. Sentences have similar meaning but include different terms. Column no. 2 in table 14 includes the best values of the tests.

Table 15

Example of Belarusian sentences. Stars define the same origin (i.e. method of coding synonyms was used)

No.	Correct sentence	Incorrect sentence
s1	Заўтра* будзе новы дзень	Заутра* будзіць новый дзень
s2	Калі вы хочаце сказаць праўду, пакіньце** элегантнасць для краўцоў	Калі вы хочаце сказать правду, пазастаўце** элегантнасць для краўцоў
s3	Хто баіцца ўжо прайграў	Кто баіца ужо праіграў

Table 16

Values of similarity of the sentences using popular methods. Description of columns (experiments): Col. 1 – Similarity method based on Lda without coding terms method; Col. 2 – Similarity method based on Lda with coding terms method; Col. 3 – Cosine distance based on term frequency weight method (tf); Col. 4 – Dice distance based on tf; Col. 5 – Jaccard distance based on tf weight method; No. – number of sentence

No.	Col. 1	Col. 2	Col. 3	Col. 4	Col. 5
s1	0.75	0.75	0.50	0.12	0.33
s2	0.89	1.00	0.77	0.08	0.63
s3	0.50	0.50	0	0	0

Table 15 includes correct and incorrect Belarusian sentences. Sentences have similar meaning but include different terms. Column no. 2 in table 16 includes the best values of the tests.

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

JOANNA PŁAŻEK*

THREE-DIMENSIONAL PATTERN RECOGNITION FOR LINEAR SECTIONS OF FORWARD TRACKER IN PANDA EXPERIMENT

ROZPOZNAWANIE ŚLADU W TRZECH WYMIARACH DLA SEKCJI LINIOWYCH TRACKERA W EKSPERYMENCIE PANDA

Abstract

Straw tube detectors operating as drift detectors are commonly used in nuclear and particle physics experiments. The straw tubes are arranged in layers and placed in areas both with and without magnetic field. In order to resolve the 3D coordinates of a track, the straws are placed in multiple overlapping layers at varying angles of inclination. By measuring the drift time and then converting it to the drift distance, a position resolution per straw can be achieved. Knowing the geometry of the detector, the position of hit straws and the drift distances, it is possible to determine the tracks of the moving particles. In the paper we present the algorithm for the 3D track reconstruction in areas without magnetic field by using vertical and skewed straws. The algorithm has been prepared for the Forward Tracker in the PANDA experiment.

Keywords: straw tube detector, *track recognition*, *analytical method*

Streszczenie

Detektory słomkowe działające jako detektory dryfowe są powszechnie stosowane w eksperymentach fizyki jądrowej i cząstek. Słomki są układane w warstwy i umieszczane zarówno w obszarach z polem magnetycznym jak i bez niego w wielu nakładających się warstwach pod różnymi kątami nachylenia. W celu wyznaczenia współrzędnych toru w trzech wymiarach, dokonuje się pomiaru czasu dryfu, a następnie przekształca się go na odległość punktu interakcji cząstki z materiałem detektora od drutu sygnałowego. Znając geometrię detektora, położenia zapalonych słomek i skojarzone z nimi odległości dryfu, jest możliwe wyznaczenie torów, po których poruszają się cząstki. W pracy przedstawiono algorytm rekonstrukcji torów 3D na obszarach bez pola magnetycznego za pomocą pionowych i ukośnych słomek. Algorytm został przygotowany dla Forward Trackera w eksperymencie PANDA.

Słowa kluczowe: detektor słomkowy, *rozpoznawanie śladu*, *metoda analityczna*

DOI: 10.4467/2353737XCT.16.150.5761

* Joanna Płażek (joannaplazek@gmail.com), Institute of Teleinformatics, Faculty of Physics, Mathematics and Computer Science, Cracow University of Technology.

1. Introduction

In experimental and applied particle physics, nuclear physics and nuclear engineering, a particle detector is a device used to detect, track, or identify high-energy particles. It may also deliver information on other attributes such as its momentum or charge.

Drift chambers are used to measure the space coordinates of the charged particle trajectory. This is achieved by measuring the drift time of the ionization electrons to the sensitive electrodes [1]. This technology is applied also in the straw drift tube chambers [2]. The straw type detectors differ in the number of the straws and also their position or orientation. The path is determined by the best fit to coordinates calculated using information coming from hit straws. Additionally, the measured drift time, which is proportional to the distance of the particle's closest approach to that chamber's sense wire, allows the coordinate to be determined with precision better than the straw radius.

The track pattern recognition in detectors has been developed since the first detector was built. A review can be found in [3]. The author after a brief introduction discusses different approaches in global and local methods of track pattern recognition including their strengths and shortcomings. In [4] a novel track finding algorithm, named the Drift Tube Hough Transform (DTHT) algorithm, is presented. The DTHT algorithm uses the possible explanations for a lack of particle hits as additional information, and takes into account all possible scenarios that may occur in the tubes.

It is quite clear that not only the accuracy of the determination of the particle track properties should be taken into account. One should also stress the importance of the analysis time especially in case of on-line processing. For this reason a unique algorithm for each detector is needed.

In this paper the algorithm for the 3D track recognition for a linear forward tracker segment is presented. It is designed for the PANDA experiment [5]. This experiment is one of the key experiments at the Facility for Antiproton and Ion Research (FAIR) in Darmstadt, Germany. It is foreseen to study the collisions of an anti-proton beam with different fixed targets.

2. Construction of Forward Tracking Stations in PANDA experiment

The Forward Tracker (FT) in the PANDA experiment consists of three pairs of planar tracking stations (Fig. 1). One pair (FT1, FT2) is placed in front of the magnet gap, the second (FT3, FT4) is placed inside the magnet gap (dipole field) and the third pair (FT5, FT6) is placed behind the magnet gap, in order to track the low transverse momentum particles exiting the magnet yoke [6].

Each tracking station consists of four double layers of straw tubes oriented respectively at 0°, +5°, –5°, 0° (Fig. 2) with respect to the vertical direction.

Each double layer contains a different numbers of straws and has the beam pipe openings of different dimensions. The details of the geometry of active areas, positions along the beam direction and the number of straw tubes in individual tracking stations can be found in [7].

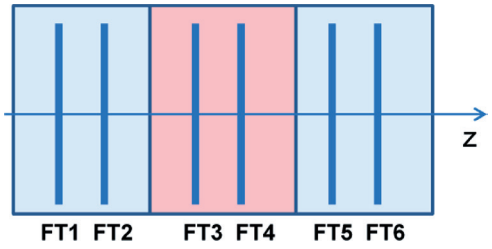

Fig. 1. Forward Tracking Stations in PANDA experiment; FT1, FT2 – in front of the magnet gap, FT3, FT4 – inside the magnet gap, FT5, FT6 – behind the magnet gap

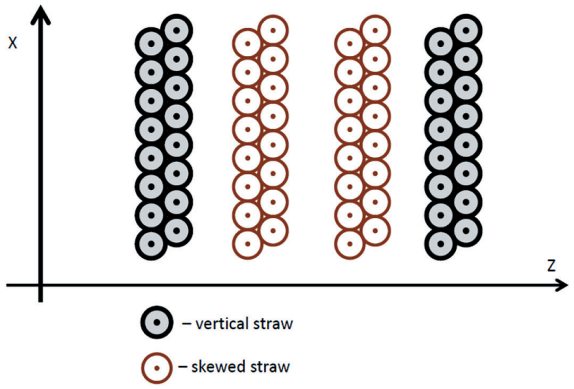

Fig. 2. One Forward Tracking Station

The most important properties of the straws for the track pattern recognition are:

– the straw diameter – 10.1 mm,
– the Mylar straw tube wall thickness – 0.03 mm,
– the tungsten, sense wire diameter – 0.02 mm,
– the gas mixture: 90%Ar + 10% CO at 2 bar.

The positions of individual sense wires in the FT straw tubes are described by straight line equations. The equations are given in a right-handed coordinate system with origin located in the nominal PANDA interaction point, Z-axis is parallel to the beam direction and Y-axis is oriented in the vertical direction.

Each straw has its unique ID number which can be used to access the layer and the tracking station numbers as well as the set of parameters describing the position of the straw sense wire.

Since the outer stations (FT1, FT2 marked as FT12 and FT5, FT6 marked as FT56) are situated outside the magnetic field it can be roughly assumed, neglecting the multiple scattering effects in light material, that in these areas particles will move along a straight line. In contrast, the charged particle trajectory in the central stations (FT3, FT4 marked as FT34) will be close to a helix: circle in X-Z and line in Y-Z projection.

To determine the particle track in three-dimensional space it is enough to calculate the track parameters in two independent two-dimensional spaces:

– horizontal plane ZOX using the vertical straws,
– vertical plane ZOY using the inclined straws.

This paper presents the three-dimensional track recognition using the FT12 and FT56 stations situated outside the magnetic field.

3. Description of the method

It is clear that the same algorithm the particle track finding can be used in the FT12 or FT56 stations since both of them are located in the regions free of the magnetic field.

In the FT12 tracking station there are four double layers of straw tubes. Half of them are vertical and the others are skewed. A straw located above the beam pipe opening matches the direction of the corresponding straw located below the opening i.e. both are described by the same equation.

The first step of the algorithm is to read in the forward detector geometry data which describe all the straw tubes arranged in 48 layers. In turn, each straw is attributed a set of five numbers, {ID, *l*, *x*, *y*, *z*}, where ID is the unique straw ordinal number, *l* is the layer number, *x*, *y*, *z* are the three coordinates which enable the determination of the equation describing a wire in a give area of the detector. The next step is to load input data generated by the PANDAROOT software. The simulator delivers events, containing for example particles of selected energies and selected angles with respect to the beam, which are passed through the detector simulation. During this operation the event number, ID-s of all hit straws (hits) are stored, and the drift radius *r* for each hit is calculated. Also, real (true) coordinates of the track are stored for each hit. This information is necessary at the later stage of the track pattern recognition to verify the correctness of the obtained results. Input data are loaded into two arrays. One stores information considering the vertical straws, the other one the skewed straws.

3.1. Track recognition in the ZOX plane using the vertical straws

To determine the particle track in the ZOX plane the layers with vertical straws are required. In the case of the tracking stations FT12 the processed layers are: 1, 2, 7, 8, 9, 10, 15 and 16.

At the beginning the track candidates are searched for. Based on the list of vertical hit straws we choose all pairs of the straws ($S1$, $S2$), where the straw $S1(z1, x1)$ with the drift radius $r1$ belongs to the layer 1 or 2, and $S2(z2, x2)$ with the drift radius $r2$ belongs to the layer 15 or 16. Points $(z1, x1)$ and $(z2, x2)$ define the place of the intersection wires with the ZOX plane. Next, the straight line $L: x=A*z + B$ passing through these points is constructed. In consequence, the algorithm then looks for all hit vertical straws $S_i(z_i, x_i)$, whose distance from this line is smaller than a predetermined value d (Fig. 3):

$$\frac{\left|A * z_i + B - x_i\right|}{\sqrt{1 + A^2}} < d, \tag{1}$$

with d defined as:

$$d = \max(r1, r2) + 0.5 \text{ cm} \tag{2}$$

where d is measured in centimetres and the constant 0.5 cm is the inner radius of a straw tube.

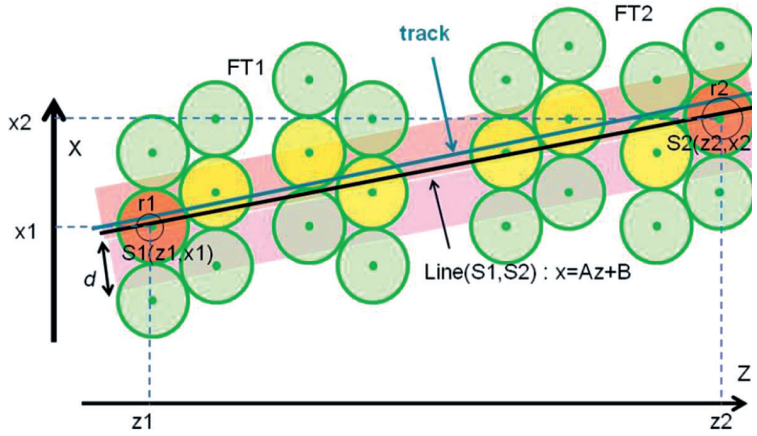

Fig. 3. Candidates to the track on ZOX plane

If the number of selected straws (with $S1$ and $S2$) is greater than 6 the case is accepted and two circles $c(S1, r1)$ and $c(S2, r2)$ are used to construct four tangent lines, see Fig. 4. Later, for each tangent line a new straw search is performed. Again the distance between the selected tangent line and the straw centre is calculated.

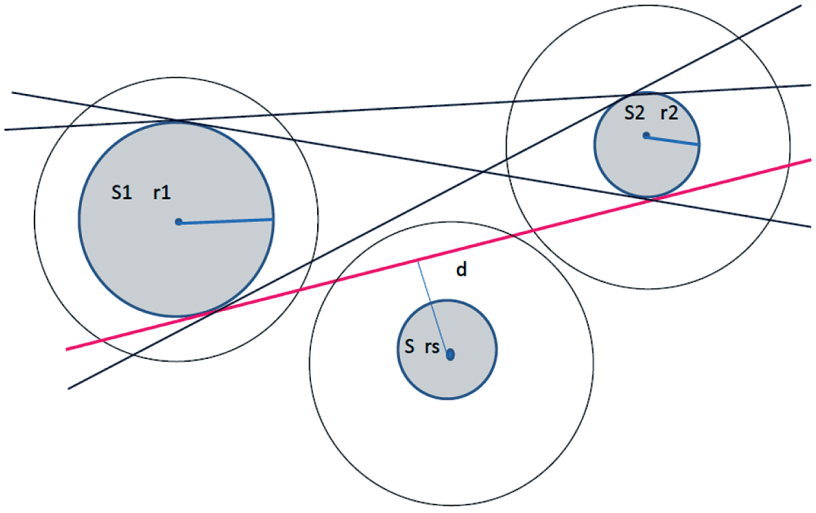

Fig. 4. Selection of the optimal tangent

If this distance diminished by the size of the drift radius *rs* fulfills the condition $|d - rs| < \Delta 1$, assuming $\Delta 1 = 0.5$ cm (size of the inner radius of a straw tube), then the straw is accepted and added to the straw list associated with a hypothetical track and the sum

$$dd = \min \sum |d(S, \text{tangent}) - rs| \tag{3}$$

is calculated. If more than 6 straws meet this criterion then it is assumed that the tangent is a candidate for the track. From all selected tangent lines the one characterised by the smallest value of dd is accepted as the track candidate.

Initially, the algorithm was initialised only for the pairs of the straws belonging to layers 1 and 16 which prevented the track determination if there was no hit in one of these layers. To improve the algorithm performance it was assumed that the signal due to the particle passage was generated in at least one layer of straws belonging to the double layer structure. Therefore, the initial straw can belong either to layer 1 or 2, and respectively 15 or 16. For this reason, the algorithm considers many more pairs and in cases in which all the layers have a hit some tracks are duplicated. This implies that an elimination procedure has to be carried out. Two track candidates are considered to be an "repetition event" if they contain the same hit straws in at least 7 out of 8 layers or also if out of 7 hit straws no more than 4 straws are not exactly the same but have neighboring numbers. Eventually, out of two such candidates the one with more hits or with smaller value of *dd* is accepted.

The pseudo-code of the algorithm described above is presented in Fig. 5.

```
for each straw S1 from layer 1 or 2
  for each straw S2 from layer 15 or 16
    line(S1,S2);
    find all straws S for which |d(S,line)-rs| < Δ1;
    if(number of found straws < 6) take next pair;
    else construct four lines tangent to c(S1,r1) and
c(S2,r2) and compute
          dd= ∑|d(S,tangent)- rs|;
          select the tangent line having min(dd);
    compare found tracks and eliminate duplicates
```

Fig. 5. Reconstruction of traces in the plane XOZ in FT12

The result of track recognition on ZOX plane is a set of hits belonging to the track and two parameters α and β of $x = \alpha * z + \beta$ forming the track in this plane.

3.2. Track recognition in ZOY plane using the skewed straws

To determine the particle trace in the plane ZOY the layers with skewed straws are used. In tracking stations FT12 the processed layers are: 3, 4, 5, 6, 11, 12, 13 and 14.

For each track found in the ZOX plane the plane Z′OY vertical to ZOX and containing the found track is constructed (see Fig. 6).

Then for each processed straw (described by equation $y = a*(x - x_0) + y_0$) the point $P(z, y)$ of intersection of the straw with the plane is calculated. Given the drift radius the coordinates of points $P1(z, y1)$ and $P2(z, y2)$ belonging to the track (4) are determined.

$$y1 = y + cc; \quad y2 = y - cc; \quad cc = r\sqrt{1 + a^2} \tag{4}$$

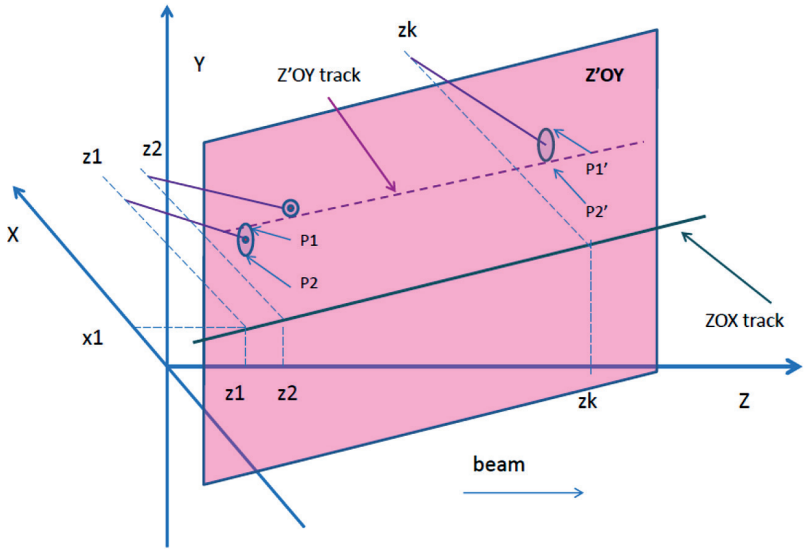

Fig. 6. Points – candidates to the track on Z'OY plane

From the list of all points $P1$ and $P2$ (see Fig. 7) only those pairs of points (P, P') are selected for which point P belongs to the layer 3 or 4, and P' to the layer 13 or 14. Next for each accepted pair a line passing through it points is determined.

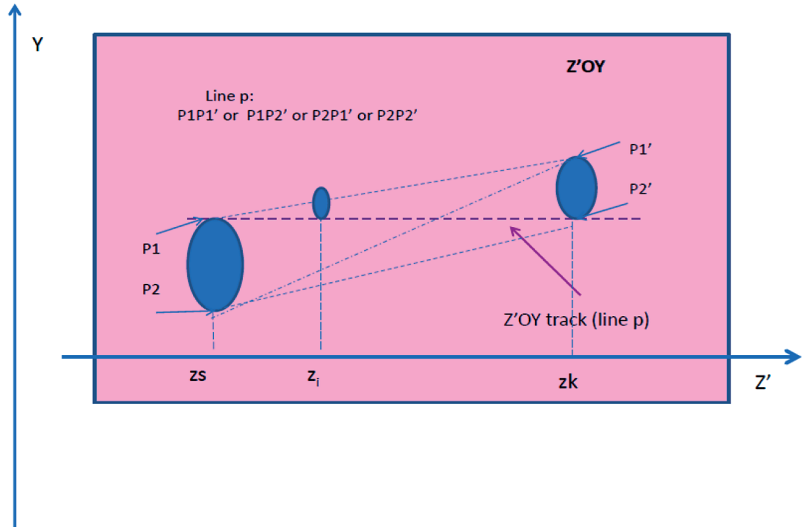

Fig. 7. The track candidates in the Z′OY plane

At the next step, all points, one in each layer, whose distance to this line is the shortest, but no greater than value of Δ2 (Δ2 = 0.5 cm; size of the inner radius of a straw tube), are considered and out of all constructed lines the one with the smallest value of the sum:

$$dd = \sum_{\text{layer}} d\min(P, \text{line}) \tag{5}$$

is selected. If there are more than 6 points in the sum then the line is the track candidate in the Z′OY plane. It is quite clear that the transformation of this line from the Z′OY to the ZOY plane is required (Fig. 8). The results of the track recognition in the ZOY plane is a set of hit straws belonging to the track and two parameters α and β defining the track line $x = \alpha * z \quad \beta$ in this plane.

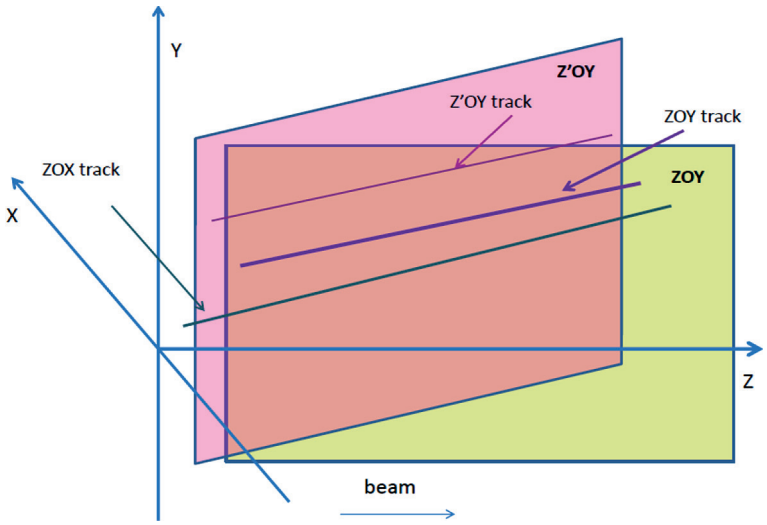

Fig. 8. Transformation of the track from the Z′OY to the ZOY plane

The pseudo-code of the discussed algorithm is presented below in Fig. 9.

```
for each track in ZOX plain
    for each straw Si
        compute points P1(zi,y1) and P2(zi,y2)
    for each point K belonging to layer 3 or 4
     for each point M  belonging to layer 13 or 14
           line(K, M);
           dd= ∑layer dmin(P,line);
```

Fig. 9. Reconstruction of traces in the plane XOY in FT12

4. Results

The algorithm was tested for input data generated by the Pandaroot. The data extracted from the simulations were ordered in the form of rows with a fixed number of columns defining the order:

– the event number,
– the track number,
– whether it is a part of the primary particle (equal –1),
– the layer number,
– the global number of the hit straw,
– the radius,
– x, y, z coordinates.

The last three numbers are the coordinates of the point crossed by the particle allowing to verify the obtained results with the data from the simulator.

Calculations were made for muons with energies of: 0.5 GeV, 2.55 GeV and 5.55 GeV. The angle of incidence of a particle was within the range (2.5°; 5.0°). The generated events contained one, three or five tracks. Only the tracks with at least one hit in each of the double layer were considered in the present analysis.

Fig. 10 shows the distribution of the simulated track position at the first layer for muons of 5.55 GeV energy.

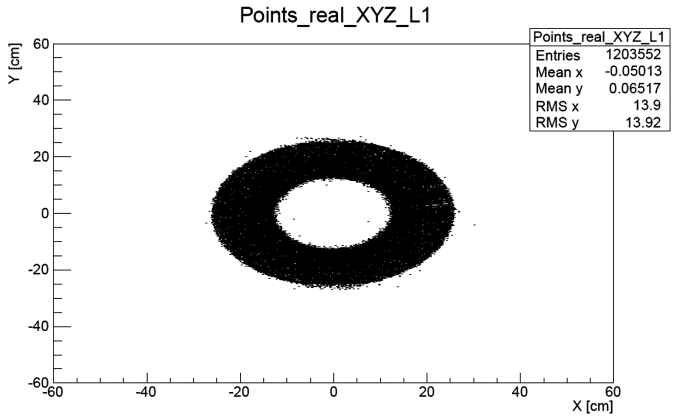

Fig. 10. Distribution of the simulated track position at the first layer for muons of 5.55 GeV

In Fig. 11 and Fig. 12 the difference in X and Y coordinates at each layer of F12 between simulated and reconstructed tracks are presented. The difference in X coordinate was computed using the vertical straws in the ZOX plane, the difference in Y coordinate on skewed straws in the ZOY plane. The difference in X coordinate is about ten times smaller than in Y coordinate.

The efficiency of the track recognition algorithm in the forward tracker F12 is illustrated in Fig. 13. It is a function of the number of found tracks in generated events. The found track is a track with minimum 14 hits in 16 layers.

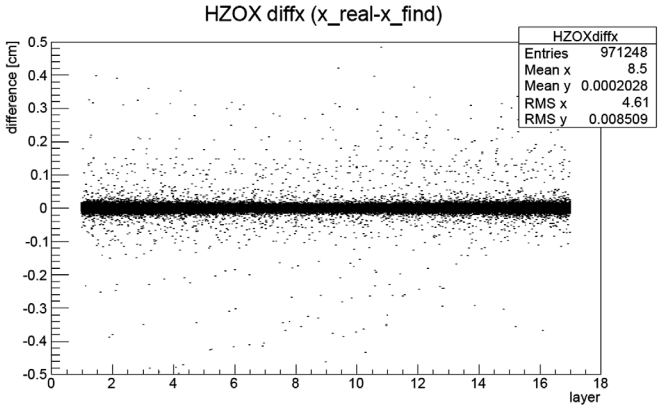

Fig. 11. The difference in X coordinate between the simulated and reconstructed track position for muons of 5.55 GeV

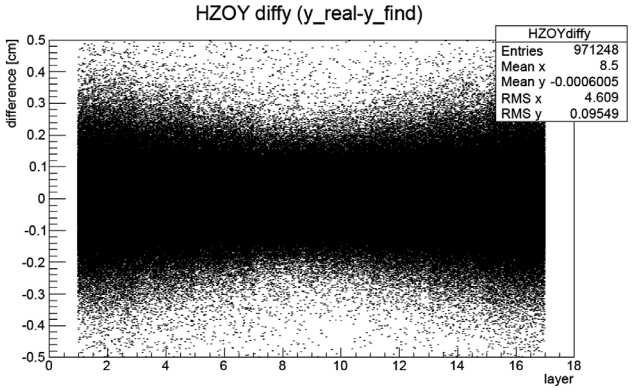

Fig. 12. The difference in Y coordinate between the simulated and reconstructed track position for muons of 5.55 GeV

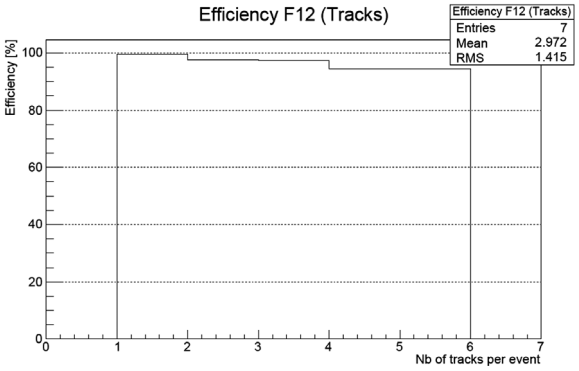

Fig. 13. Efficiency of tracks recognition in FT12 for muons of 5.55 GeV

5. Conclusions

In this paper, the three-dimensional track pattern recognition algorithm outside the magnetic field for PANDA experiment in FAIR was presented. It is based on an analytical solution. The algorithm uses the detector geometry, information about the particle hits and the drift time/radius values.

Results indicate that the obtained accuracy of the particle path determination using the vertical straws is much better than that which can be obtained using the skewed straws. The efficiency of the track finding is still above 95% for events with five tracks.

The algorithm returns a list of straws associated with a track as well as the parameters of the straight line allowing the three-dimensional determination of the particle path before the magnetic field area (in FT12), and after leaving it (in FT56). This information is necessary input for the determination of the particle trajectory of a particle in the magnetic dipole field.

I am indebted to Cracow PANDA Group for stimulating discussions and help and to ACK CYFRONET – Kraków for possibility of using the computing resources.

References

[1] Blum W., Riegler W., Rolandi L., "Particle Detection with Drift Chambers", Springer–Verlag Berlin Heidelberg, 2008.
[2] http://cms.web.cern.ch/
[3] Mankel R., "Pattern Recognition and Event Reconstruction in Particle Physics Experiment", Reports on Progress in Physics – IOPscience, 2004.
[4] Kortner O., Messer H., Mikenberg G., Primor D., "A novel approach to track finding in a drift tube chamber", Published by Institute of Physics Publishing and SISSA, 2007, http://iopscience.iop.org/article/10.1088/1748-0221/2/01/P01009/pdf.
[5] https://panda.gsi.de/
[6] Smyrski J., "Overview of the PANDA Experiment", Proceedings of the 2nd International Conference on Technology and Instrumentation in Particle Physics (TIPP 2011), Physics Procedia, vol. 37, 2012, pages 85-95.
[7] Smyrski J., "Geometry of the Forward Tracking Stations", Internal document, 2013.

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

JERZY RASZKA*, LECH JAMROŻ*

ANALYSIS AND DESIGN OF CONTROL DATA PROCESSING AS DISCRETE EVENT SYSTEMS

ANALIZA I PROJEKTOWANIE STEROWANIA PRZETWARZANIEM DANYCH JAKO SYSTEMU ZDARZEŃ DYSKRETNYCH

A b s t r a c t

The following article presents the application of the Max-Plus Linear System (MPLS) method in the synthesis of different IT process control structures, taking place in the systems belonging to the class of Discrete Event Systems (DES). Modelling and analysis of these systems are based on the state equations, expressed in MPLS categories and classical principles of control theory, adapted for them. MPLS is based on max-plus algebra formalism and is supported with graphical representation in the form of Timed Event Graph (TEG), which is the special case of Timed Petri Nets (TPN). The article contains an overview of the theoretical work on discrete control processes, with a particular focus on the synthesis of control signals in the open control to obtain the desired output system. A practical example has been used for distributed computational process and data transmission for the system controlling selected technological process. Numerical results are presented.

Keywords: max-plus linear system, *timed event graph*, *event system*, *control*

S t r e s z c z e n i e

W artykule przedstawiono zastosowanie metod opartych na max-plus liniowym systemie (MPLS) w syntezie struktur sterujących różnych procesów IT, zachodzących w systemach należących do klasy systemów zdarzeń dyskretnych. Modelowanie i analiza tych systemów bazuje na równaniach stanu, wyrażonych w kategoriach MPLS i przystosowanych do nich, klasycznych zasadach teorii sterowania. MPLS opiera się na formalizmie max-plus algebry i posiada reprezentację graficzną w postaci czasowego grafu zdarzeń, który jest szczególnym przypadkiem czasowych sieci Petriego. Artykuł zawiera przegląd prac teoretycznych dotyczących sterowania procesami zdarzeń dyskretnych ze szczególnym przedstawieniem syntezy sygnałów w otwartym układzie sterowania mających na celu wymuszenie zadanej odpowiedzi. Praktyczny przykład zastosowano dla rozproszonego procesu obliczeniowego z przesyłaniem danych do systemu sterowania wybranego procesu technologicznego oraz przedstawiono wyniki obliczeń numerycznych.

Słowa kluczowe: max-plus liniowy system, *czasowy graf zdarzeń*, *system zdarzeń*, *sterowanie*

DOI: 10.4467/2353737XCT.16.151.5762

* Jerzy Raszka (jraszka@riad.pk.edu.pl), dr inż. Lech Jamroż, Institute of Computer Science, Cracow University of Technology.

1. Introduction

The increasing complexity of information processes in distributed computer systems and microprocessor systems increases the probability of faults and disturbances. Hence, there is the need to include them in the design process. These processes were treated as occurring in class of discrete event systems (DES).

The DES class is very wide, covering manufacturing systems (including flexible and assembly lines) [5, 6], road, railway and air transport systems [3, 19], as well as computer networks [15]. In addition, one can also qualify processes in the field of Human System Interaction, e.g. resource and task management, as well as process control technologies. The variety of DES systems leads to different models. The MPLS model has been adopted in this article, based on the algebra theory (max, +). One of the first scientists, who described this theory was R.A. Cuninghame-Green [12] and then it was developed by the INRIA team [7]. The authors of the articles showed that the behaviour of certain DES systems can be described using linear equations. They have also pointed at many cases of analogies to the problems in the systems theory, automation and control. Joint researches led to the publication of collective work [2]. In the following article, based on the bibliographic data, the theoretical basis of modelling and control in the DES systems are discussed, supplemented with the authors' results from the scope of this problem.

In the area of research, in particular should be mentioned:

- Optimal control in the open loop. Structure described by the Cohen [9], in which a well-known system model and time sequence of output signals are assumed, while the optimal trajectory of the input signals is calculated,
- Preliminary compensator [14, 18]. The time sequence of output signals is not known, but the reference model, which imposes the behaviour of the output relative to the input is assumed.
- Corrector with the feedback. In the control structure the system output is modified by the correctors in the feedback [10]. They converge the behaviour of the whole system to the behaviour set in the reference model.
- R, S, T type correction. Strategy based on the introduction of three correctors into the structure, has been inspired by the Åström [1]. This leads to the better results than those obtained with the single corrector [20].
- Control in the presence of disturbances [16]. System is exposed to the acting of uncontrolled inputs. To reduce their impact, the control in the closed-loop is taken into account.
- Robust control [17]. It is assumed that the system parameters are random, but are in the specific range of values. Synthesis of control is based on the feedback.

The main part of the following article concerns the design of the open control structures, including control under conditions of uncertainty in the data transmission aspect.

This article is organized as follows. Section 2 introduces the maxplus algebra theory and MPLS modelling. In section 3, based on the literature review, the authors' results related to the selected problems of processes control in the DES systems have been presented. Also some problems in the disturbances conditions, damages and uncertainty have been discussed. In section 4 there is the general theory related with the open control. In section 5

control system synthesis has been described and the practical results for the computational processes and data transmission in model of IT systems have been presented.

2. Mathematical fundamentals

This section contains selected basic concepts of max-plus algebra providing the basis to formulate a model MPLS. They are widely discussed in the Chapters 3 and 4 of publication [2]. Max-plus algebra formalism is based largely on the lattice theory and partially ordered sets, and residuation theory. In turn, the theoretical basis, allowing representation of MPLS in the categories of the state equations system is presented in Chapters 5 and 6 of [2].

2.1. Max plus algebra [25]

In recent years, the concept of a max-plus-linear system (MPLS) has been increasingly frequently used in the literature. It is based on a mathematical formalism, namely max-plus algebra. The basic operations of max-plus algebra are maximization and addition, which will be represented, respectively, by $\oplus$ and $\otimes: x \oplus y = \max(x, y)$ and $x \otimes y = x + y$ for $x, y \in R_\varepsilon$, $R_\varepsilon =^{def} R \cup \{-\infty\}$

The reason for using these symbols is that there is a remarkable analogy between $\oplus$ and conventional addition, and between $\otimes$ and conventional multiplication: many concepts and properties from linear algebra (such as the Cayley-Hamilton theorem, eigenvectors and eigenvalues, Cramer's rule) can be translated to max-plus algebra by replacing + with $\oplus$ and $\times$ with $\otimes$. Hence we also call $\oplus$ the max-plus-algebraic addition, and $\otimes$ the max--plus-algebraic multiplication. Note, however, that a major difference between conventional algebra and max-plus algebra is that, in general, there are no inverse elements with respect to $\oplus$ in R_ε. The zero element for $\oplus$ is $\varepsilon =^{def} -\infty$ and we have $a \oplus \varepsilon = a = \varepsilon + a$ for all $a \in R_\varepsilon$. The structure $(R_\varepsilon, \oplus, \otimes)$ is referred to as max-plus algebra. Let $r \in R$. The r^{th} max-plus-algebraic power of $x \in R$ is denoted by $x^{\otimes r}$ and corresponds to rx in conventional algebra. If $x \in R_\varepsilon$, then $x^{\otimes 0} = 0$ and the inverse element of x w.r.t. $\otimes$ is $x^{\otimes -1} = -x$. There is no inverse element for ε since ε is absorbing for $\otimes$. If $r > 0$, then $\varepsilon^{\otimes r} = \varepsilon$, and if $r < 0$, then $\varepsilon^{\otimes r}$ is not defined. In this paper, we have $\varepsilon^{\otimes 0} = 0$ by definition.

The implicit equation $x = a \otimes x \oplus b$ determines $a = a^* \otimes b$ where the Kleene star operator:

$$a^* = \bigoplus_{i=0}^{\infty} a^i$$

The rules for the order of evaluation of max-plus algebraic operators correspond to those of conventional algebra. So the max-plus-algebraic power has the highest priority, and max--plus-algebraic multiplication has a higher priority than max-plus-algebraic addition.

The basic max-plus-algebraic operations are extended to matrices as follows.

If $\mathbf{A}, \mathbf{B} \in R_\varepsilon^{m \times n}$ and $\mathbf{C} \in R_\varepsilon^{m \times p}$, then:

$$(\mathbf{A} \oplus \mathbf{B})_{ij} = a_{ij} \oplus b_{ij} = \max(a_{ij}, b_{ij})$$

$$(\mathbf{A} \otimes \mathbf{C})_{ij} = \bigoplus_{k=1}^{n} a_{ik} \otimes c_{kj} = \max_{k=1\ldots n} (a_{ik} + c_{ki})$$

for all *i, j*. Note the analogy with the definitions of the matrix sum and the product in conventional linear algebra.

The matrix $\boldsymbol{\varepsilon}_{m\times n}$ is the $m\times n$ max-plus-algebraic zero matrix: $(\boldsymbol{\varepsilon}_{m\times n})_{i,j} = \varepsilon$ for all i, j; and the matrix $\mathbf{E}_n$ is the $n \times n$ max-plus-algebraic identity matrix: $(\mathbf{E}_n)_{ii} = 0$ for all i and $\mathbf{E}(\mathbf{E}_n)_{ij} = \varepsilon$ i, j with $i \neq j$. If the size of the max-plus-algebraic identity matrix or the max--plus-algebraic zero matrix is not specified, it should be clear from the context. The max-plus--algebraic matrix power of $\mathbf{A} \in R_{\varepsilon}^{n\times n}$ is defined as follows: $\mathbf{A}^{\otimes 0} = \mathbf{E}_n$ and $\mathbf{A}^{\otimes k} = \mathbf{A} \otimes \mathbf{A}^{\otimes(k-1)}$ for $k = 1,2, \ldots$

The Kleene star operator can also be applied to matrices:

$$\mathbf{A}^{*} = \bigoplus_{i=0}^{\infty} \mathbf{A}^{i} \quad \text{with} \quad \mathbf{A}^{i+1} = \mathbf{A} \otimes \mathbf{A}^{i} \text{ and } \mathbf{A}^{0} = \mathbf{E} \tag{1}$$

where:

$\mathbf{E}$ – the identity matrix.

Equation (1), which has nilpotent matrix, achieves convergence (all coefficients equal ε).

2.2. Model of the system

In article [5] Cohen showed that the nonlinear dynamic systems, whose structure and behaviour is based on the timed event graph (TEG) may be described using the linear equation system The example of a TEG with the determined holding time of 2 units in place P1 is given in Fig. 1 [4].

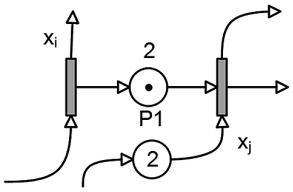

Fig. 1. Graphical representation of a TEG

State-space descriptions in the max-plus algebra for a certain class of discrete-event-systems become linear representations which are similar to state-space equations in the traditional model control theory. Generally speaking, for any TEG system, one obtains the following kind of equations as an MPLS [8]:

$$\mathbf{x}(k) = \bigoplus_{i=0}^{M} \mathbf{A}_i \ \mathbf{x}(k-i) \oplus \mathbf{B}_i \ \mathbf{u}(k-i) \tag{2.1}$$

$$\mathbf{y}(k) = \bigoplus_{i=0}^{M} \mathbf{C}_i \ \mathbf{x}(k-i) \tag{2.2}$$

where **x**, **u**, and **y** are vectors of dimensions equal to the numbers of internal, input and output transitions, respectively. $\mathbf{A}_i$, $\mathbf{B}_i$, and $\mathbf{C}_i$ are matrices of the appropriate dimensions with entries in the max-plus algebra, and M is the maximal number of tokens in the initial marking. The variables of (2) are time instances and the represented events occur at k-times. The coefficients of matrices **A**, **B**, **C** represent parameters associated with the places located between these transitions. The classical theory of the continuous and discrete systems in the time domain, revolutionized the integral transforms (e.g. Laplace, Fourier, Z-transform). Similar transformations have become useful in the theory of discrete processes. Each transition in the TEG model can be assigned to the appropriate of both, input or output vector's components, as well as to internal state.

In the article [9] is derived model, of the system is represented by 2-dimentional (γ, δ) – transform noted as $\mathbf{M}_{\min}^{\max}[[\gamma,\delta]]$ a set of formal power series for two variables γ and δ. A finite series of $\mathbf{M}_{\min}^{\max}[[\gamma,\delta]]$ is a polynomial and is used to code a set of information concerning the transition of a TEG. The monomial $\gamma^k\delta^t$ may be interpreted as the k-th event occurring at least at time t.

Using transform by $\mathbf{M}_{\min}^{\max}[[\gamma,\delta]]$ the TEG system (2) has implicit form as

$$\mathbf{x} = \mathbf{A}\mathbf{x} \oplus \mathbf{B}\mathbf{u} \tag{3.1}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} \tag{3.2}$$

where $\mathbf{A} \in \mathbf{M}_{\min}^{\max}[[\gamma,\delta]]_{n\times n}$, $\mathbf{B} \in \mathbf{M}_{\min}^{\max}[[\gamma,\delta]]_{n\times p}$, $\mathbf{C} \in \mathbf{M}_{\min}^{\max}[[\gamma,\delta]]_{m\times \boldsymbol{n}}$,

System equation (3) by Kleene star (1) transform, gives the explicit form as

$$\mathbf{x} = \mathbf{A}^*\mathbf{B}\mathbf{u} \tag{4.1}$$

$$\mathbf{y} = \mathbf{H}\mathbf{u} \tag{4.2}$$

where $\mathbf{H} = \mathbf{C}\mathbf{A}^*\mathbf{B} \in \mathbf{M}_{\min}^{\max}[[\gamma,\delta]]_{p\times m}$ is input/output transfer matrix relation

System equation (3) and matrix **H** is modelled as block schema (Fig. 2).

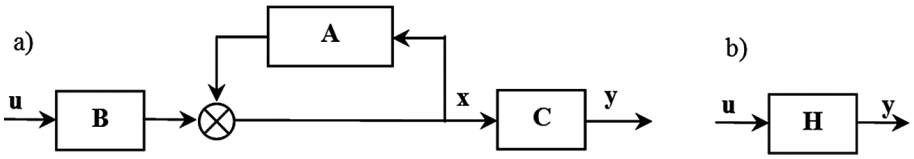

Fig. 2. Block schema of the system (a), and its substitute (b)

3. Control systems

3.1. Open control

First results concerning the open control, obtained using the (max,+) algebra are included in the Cohen [9] and Menguy [21] articles. Control was proposed for developing set *a priori* output trajectory, specified as open control. This control plays a key role in event

planning [21] and scheduling of tasks. For example these tasks can be executed by the some processes in a distributed computing micro-processors system. This issue will be presented in detail with example in sections 5 and 6.

3.2. Feedback control from the output

Controlled structure consists with the corrector between the system's output **y** and its input **v**. Output signals of events are modified by the corrector and put together with the input events. This problem is described in details in the work of B. Cottenceau [10] and in the articles [15]. Structure of control with output feedback as the block diagram is explained in the Fig. 3.

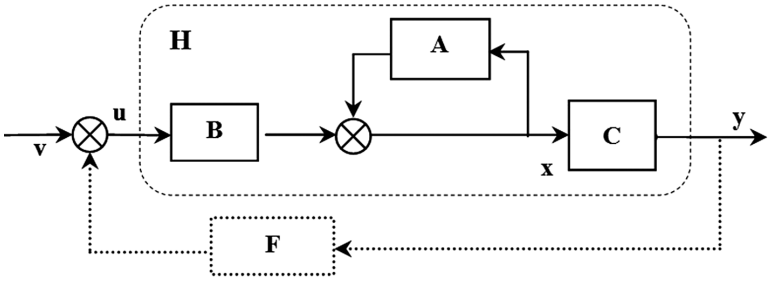

Fig. 3. Control with of the feedback from the output

For this system $\mathbf{u} = \mathbf{Fy} \oplus \mathbf{v}$

$$\mathbf{y} = \mathbf{H}(\mathbf{Fy} \oplus \mathbf{v}) = \mathbf{HFy} \oplus \mathbf{Hv}$$

and using (1) $\mathbf{y} = (\mathbf{HF})^{*}\mathbf{Hv} = \mathbf{G_F v}$

where $$\mathbf{G_F} = (\mathbf{HF})^{*}\mathbf{H} \leqslant \mathbf{G_z} \tag{5}$$

Expression (5) may be used to find **F** as the best control for applied desired characteristics and may be at least as fast as the reference $\mathbf{G_z}$.

3.3. Feedback control from the state

Structure of control with state feedback is explained in Fig. 4. In this case change of control structure consists of the corrector between the system's output and input. System is being controlled using by signal of state system's events, and changing by the corrector **F** analogically as in previous subsection 3.2 modified input the system. This problem is described in details in the work of B. Cottenceau [10] and in the article [15].

For this system $\mathbf{u} = \mathbf{Fx} \oplus \mathbf{v}$

and $$\mathbf{x} = \mathbf{Ax} \oplus \mathbf{Bu} = \mathbf{Ax} \oplus \mathbf{BKx} \oplus \mathbf{Bv}$$

$$\mathbf{x} = (\mathbf{A} \oplus \mathbf{BF})\mathbf{x} \oplus \mathbf{Bv}$$

and solve using (1) $\mathbf{x} = (\mathbf{A} \oplus \mathbf{BF})^{*}\mathbf{Bv}$

Output $$\mathbf{y} = \mathbf{C}(\mathbf{A} \oplus \mathbf{BF})^{*}\mathbf{Bv} = \mathbf{G}_{\mathbf{F}}\mathbf{v} \quad (6)$$

where $$\mathbf{G}_{\mathbf{F}} = \mathbf{CA}^{*}(\mathbf{A}^{*}\mathbf{BF})^{*}\mathbf{A}^{*}\mathbf{B} = \mathbf{CA}^{*}\mathbf{B}(\mathbf{FA}^{*}\mathbf{B})^{*} \quad (7)$$

$$\mathbf{G}_{\mathbf{y}} \leqslant \mathbf{G}_{\mathbf{z}} \quad (8)$$

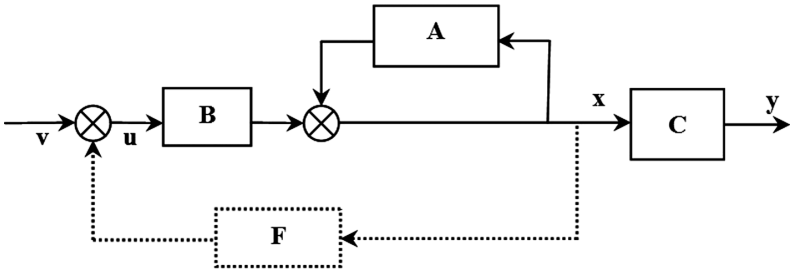

Fig. 4. Feedback control from the state

Expressions (7) may be used to find **F** as the best control for applied desired characteristics and may be at least as fast as the reference $\mathbf{G}_{\mathbf{z}}$ (8).

3.4. Control with the observer

The availability of state of the system in the previous point, is an important condition but not always possible to fulfil. There is, however, based on a known model of the system can calculate the analytical condition which is reconstructed state. On the basis of the analogue, a conventional approach Fig. 5 shows the structure of an observer [26].

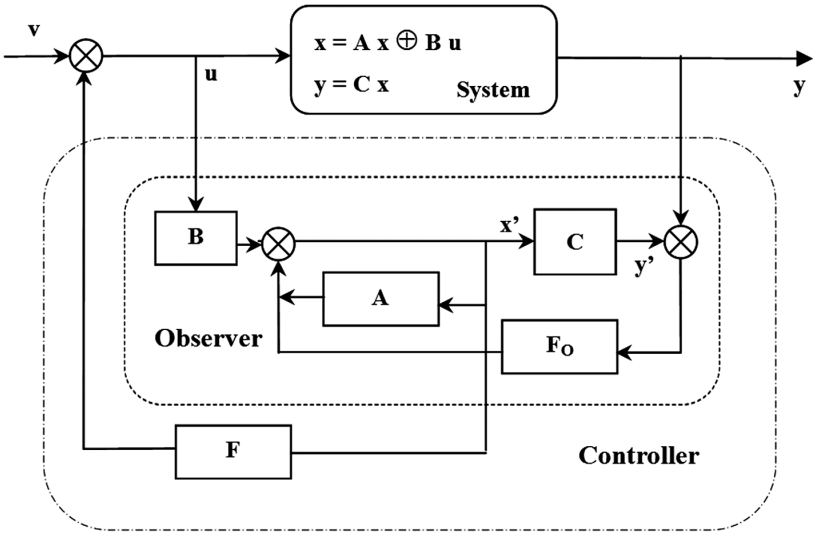

Fig. 5. Structure of control with the observer

Assuming the existence of matrices **A**, **B** and **C** the real system equation (3) in explicit form is

$$\mathbf{x} = \mathbf{A}^{*}\mathbf{B}\mathbf{u}$$

and the equation of the observer is

$$\mathbf{x}' = \mathbf{A}^{*}\mathbf{B}\mathbf{u} \oplus \mathbf{F}\mathbf{o}(\mathbf{y} \otimes \mathbf{y}') \tag{9}$$

Our goal is to calculate the matrix $\mathbf{F}_{\mathbf{O}}$ to ensure that the estimated output $\mathbf{y}'$ is less than or equal to the measured output **y**.

Typically, an observer is used to estimate the conditions necessary for feedback from the state and it is possible now to use control as in subsection 3.3 with equation (6)

$$\mathbf{y} = \mathbf{C}(\mathbf{A} \oplus \mathbf{B}\mathbf{F})^{*}\mathbf{B}\mathbf{v} \tag{10}$$

Now expressions (9) and (10) may be used to find **F** and $\mathbf{F}_{\mathbf{O}}$ as the best control for applied desired characteristics.

4. Data processing with control

Let us consider a data process that allows event-driven applications to take advantage of multiprocessors by running the code for event handlers in parallel. To achieve high performance, servers must overlap computation with the I/O. Programs typically achieve this overlap by using threads or events. Threaded programs usually process every request in a separate thread; while one thread block is waiting for the I/O, another thread can run. Event-based programs are structured as a collection of call-back functions which are called by the main loop when I/O events occur. Threads provide an intuitive programming model, but require coordinating the access of different threads to the shared state, even on a uniprocessor. Event-based programs execute call-backs sequentially so the programmer need not worry about concurrency control; however, event-based programs have so far been unable to make good use of multiprocessors. Much of the effort required to make existing event-driven programs take advantage of multiprocessors is in specifying which events can be handled in parallel.

This article presents a simple problem of designing the control of a system in which the cost is chosen so that it provides a trade-off between minimizing the delays of the end time of computational process operations (the real time to complete all the tasks in a cyclic computational process, times of final results of one cycle) and the periodicity of the desired output (the time desired or needed) to complete the process.

This problem was presented with no disturbances [22] and it was solved in max-plus algebraic functions as dater equation. Now we introduce disturbances and this problem is modelled in the 2-dimensional $\mathbf{M}_{\min}^{\max}[[\gamma, \delta]]$ domain.

Simple data processing consists of several tasks linked by the wait for I/O data (Fig. 6). To illustrate our approach, let us consider a process that consists of some tasks: T_{0i} which runs on microprocessors: μP_{0i}, for i =1, ..., n. Each of these tasks is executed on a dedicated microprocessor. In this process, the digital information flows as input/output processing data

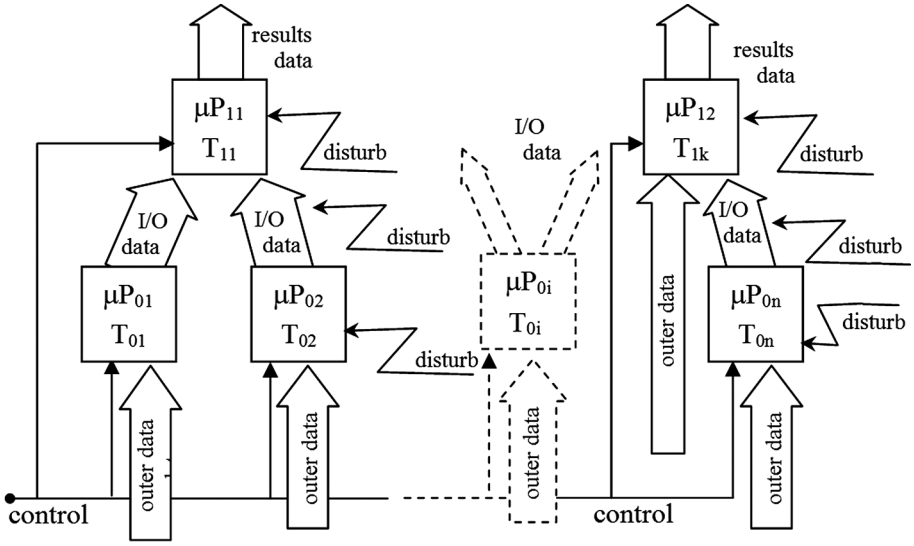

Fig. 6. The structure of the disturbing processes

and a control signal. Input data (i.e. from outer system) is processed as the first task on the μP_{01} and its output data has to be saved to memory while it waits to be processed. The other microprocessors operate in the same way, but their input data simultaneously constitutes the output result from the previous microprocessor and /or may need extra outer data too.

5. Control in open structure

In this section, the specified project is considered to obtain open control problem. In the context of data processing, this problem is to give final results and at the same time minimizing the size of the memory. Specifically, the control problem in open loop resolved as follows. It is system (a TEG with p inputs and q outputs) whose transfer matrix is known to transfer $\mathbf{H} = \mathbf{CA}^{*}\mathbf{B} \in \mathbf{M}_{\min}^{\max}[[\gamma, \delta]]_{p \times m}$. It is desired, using inputs $\mathbf{u} \in \mathbf{M}_{\min}^{\max}[[\gamma, \delta]]_{p}$ to ensure that the system outputs follow the best trajectory determined by $\mathbf{z} \in \mathbf{M}_{\min}^{\max}[[\gamma, \delta]]_{p}$.

In [8], it is shown that this problem has an optimal solution, that there is a greater input control $\mathbf{u_{opt}} \in \mathbf{M}_{\min}^{\max}[[\gamma, \delta]]_{p}$ such that the output resulting from that input ($\mathbf{y_{opt}} = \mathbf{Hu_{opt}}$) is less than or equal to the desired output $\mathbf{z}$. The $\mathbf{u_{opt}}$ order is optimal from the point of view the just-in-time criteria ($\mathbf{y_{opt}}$ the output is just-in-time). Here we implement restrictions.

– Input reference can be updated. For example, in the context of data processing the final results may lead to modifications of outer processes.
– Deadlines for the firing of some of the input transition can't be modified, which may provide input data to the actual processes.

Formally, transformation of $L_H : \mathbf{M}_{\min}^{\max}[[\gamma,\delta]]_p \Rightarrow \mathbf{M}_{\min}^{\max}[[\gamma,\delta]]_p$, $u \Rightarrow H \otimes u$, defines optimal control.

$$\{\mathbf{u} \in \mathbf{M}_{\min}^{\max}[[\gamma,\delta]]_p \,|\, L_H(u_{\text{opt}}) \leqslant z\}.$$

More specifically this is the upper limit (marked u_{opt}), which gives you the greatest control satisfying the condition of $L_H(u_{\text{opt}}) \leqslant z$. We can already see that this set is not empty since $u = \varepsilon$ is the solution, that is $L_H(\varepsilon) \leqslant z$ and it is inversion problem which the theory residuation solves this problem directly.

The optimal command u_{opt} exists and is given by:

$$\mathbf{u_{opt}} = \{\mathbf{u} \in \mathbf{M}_{\min}^{\max}[[\gamma,\delta]]_p \,|\, L_H(u) \leqslant z\} L_H(z) = H \setminus z \tag{11}$$

The optimal control for TEG corresponds to the order by entering the markers to the system as late as possible.

6. Example

In order to accomplish achieve the results, we'll look at an example of the system processes. Consider the TEG model in Fig. 7. As mentioned in section 4 this model can represent i.e. a tasks of a process in a distributed computing system constructed of some micro-processors P and memory units T. In this example data results from P1, P2 and P5 is buffered in T5, T6, T7 and then there are processed by P3 and P4. Note that processors P1, … P5 have different cycle times: i.e P1 can handle a task every 2 units while P2 every 4 units of time etc. For this system, according to $\mathbf{M}_{\min}^{\max}[[\gamma,\delta]]$ representation (3,4) we have

$$\mathbf{A} = \begin{bmatrix}
\varepsilon & \gamma & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\
\delta^2 & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\
\varepsilon & \varepsilon & \varepsilon & \gamma & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\
\varepsilon & \varepsilon & \delta^4 & \varepsilon & \varepsilon & \gamma & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\
\varepsilon & \delta^5 & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\
\varepsilon & \varepsilon & \varepsilon & \boldsymbol{\varepsilon} & \delta^2 & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\
\varepsilon & \varepsilon & \varepsilon & \delta^6 & \varepsilon & \varepsilon & \varepsilon & \gamma & \varepsilon & \varepsilon \\
\varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \delta & \varepsilon & \varepsilon & \varepsilon \\
\varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \gamma \\
\varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \delta^3 & \varepsilon
\end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix}
\delta^1 & \varepsilon & \varepsilon \\
\varepsilon & \varepsilon & \varepsilon \\
\varepsilon & \delta^1 & \varepsilon \\
\varepsilon & \varepsilon & \varepsilon \\
\varepsilon & \varepsilon & \varepsilon \\
\varepsilon & \varepsilon & \varepsilon \\
\varepsilon & \varepsilon & \varepsilon \\
\varepsilon & \varepsilon & \varepsilon \\
\varepsilon & \boldsymbol{\varepsilon} & \delta^2 \\
\varepsilon & \varepsilon & \varepsilon
\end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix}
\varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & e & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\
\varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & e & \varepsilon & \varepsilon
\end{bmatrix}$$

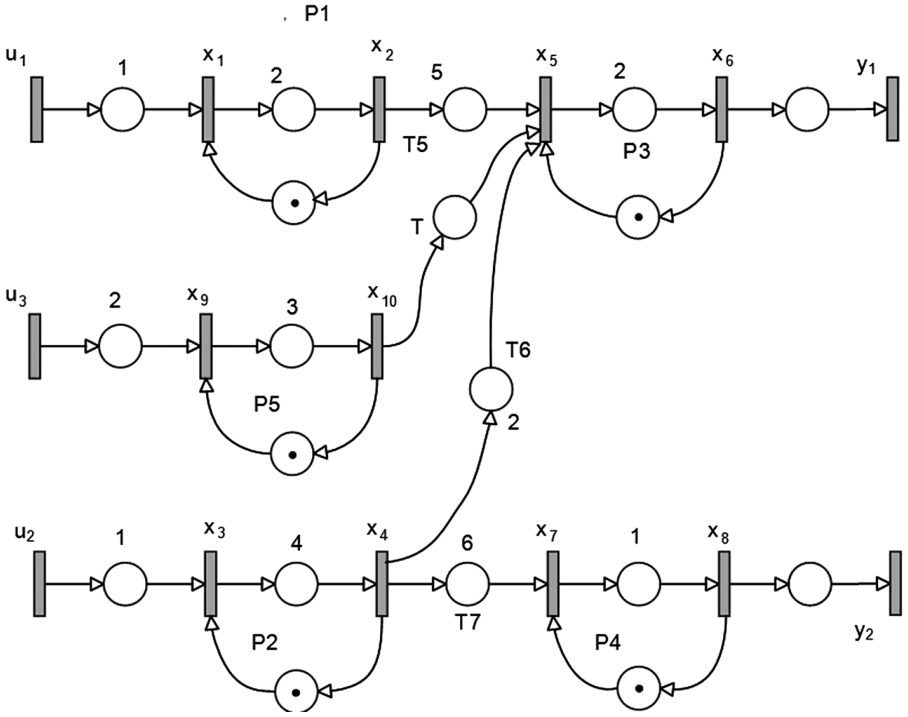

Fig. 7. TEG of the system processes

According to (4) and, we can rewrite system transfer

$$\mathbf{H} = \mathbf{C}\mathbf{A}^{*}\mathbf{B} = \begin{bmatrix} \delta^{10}(\gamma\delta^{2})^{*} & \varepsilon & \varepsilon \\ \varepsilon & \delta^{12}(\gamma\delta^{4})^{*} & \varepsilon \end{bmatrix}$$

We may to determine a desired output i.e.

$$\mathbf{z} = \begin{bmatrix} \delta^{10} \oplus \gamma\delta^{22} \oplus \gamma^{4}\delta^{30}(\gamma\delta^{6})^{*} \oplus \gamma^{10}\delta^{+\infty} \\ \varepsilon \end{bmatrix}$$

By convention, the first event is the number 0 and the trajectory of this should be interpreted as follows: 0 task should be done no later than 10 time, and the task 1, 2 and 3 at the latest during the 22. Then there is to be executed task 4, at 32 and then each one next task every 6 units of time. The final monomial $\gamma^{9}\delta^{+\infty}$ means that the task 8 is the last for this process. It also means that the task 9 and the next is not implemented (the term is infinite).

Calculation of optimal control is determined by (11)

$$\mathbf{u} = \begin{bmatrix} e \oplus \gamma\delta^{8} \oplus \gamma^{2}\delta^{10} \oplus \gamma^{3}\delta^{12} \oplus \gamma^{4}\delta^{20} \oplus \gamma^{5}\delta^{26} \oplus \gamma^{6}\delta^{32} \oplus \gamma^{7}\delta^{38} \oplus \gamma^{8}\delta^{44} \oplus \gamma^{9}\delta^{+\infty} \\ \varepsilon \\ \varepsilon \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} \delta^{10} \oplus \gamma\delta^{18} \oplus \gamma^2\delta^{20} \oplus \gamma^3\delta^{22} \oplus \gamma^4\delta^{30} \oplus \gamma^5\delta^{36} \oplus \gamma^6\delta^{42} \oplus \gamma^7\delta^{48} \oplus \gamma^8\delta^{54} \oplus \gamma^9\delta^{+\infty} \\ \varepsilon \end{bmatrix}$$

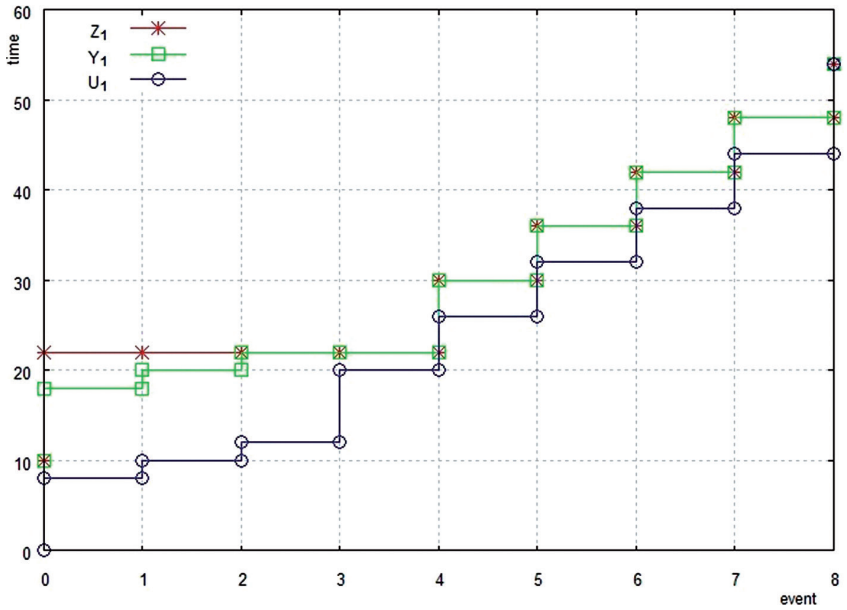

Fig. 8. Graphical representation of z_1, y_1 and u_1

Results as trajectories z_1, y_1 and u_1 are shown in Fig. 8. We can check that the optimal control **u** well meets the specification, i.e. that the output y is less than or equal to the **z**

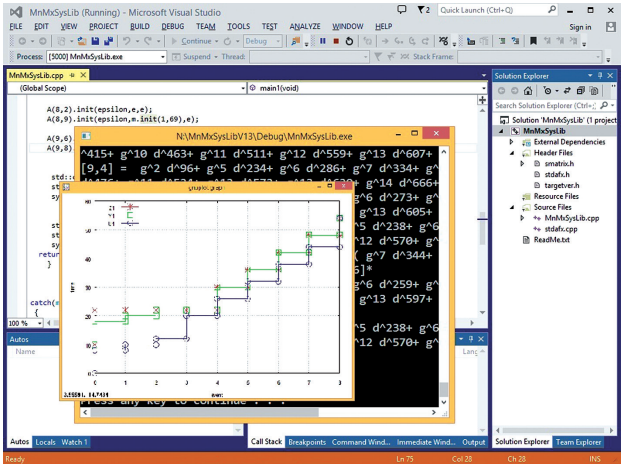

Fig. 9. VC++ platform of software tools for MPLS (in the implementation)

Calculations were performed and graphically presented using own software package currently developed on the Visual Studio 2013 platform (Fig. 9) with software library [28] and plot application Gnuplot [27].

7. Conclusions

In this article an overview of selected control structures and particular consideration of the control of the open MPLS has been presented. The main purpose is the synthesis of the control input, when we know global transfer **H** and reference (desired) input is updated. In the next step the problem of permissible deviations of real **H** should be elaborated, and the presence of uncontrollable input transitions. The problem formulated in the article has the close analogy with the problems encountered in classical control theory. There is not only feedback control but also predictive and robust control. There may be a need for effective control to use decoupling in multidimensional systems with cross-coupling interaction (like in computer control system [24]). Other problems concern different failures – events of data loss and damage while transmission. The solutions obtained do not completely eliminate the consequences of failures (i.e. delay), but are used for maintenance of the stability and elimination of memory overflow [23].

It is important to follow the new solutions and development of theoretical researches, concerning the classical theory of the system. It is planned to evolve practical applications and to create new or expand existing informatic tools. Further development of this software is planned.

References

[1] Ästrom K., *Robustness of a design method based on assignment of poles and zeros*, IEEE Trans. on Automatic Control, 25, 588-591.

[2] Baccelli F., Cohen G., Olsder, G. Quadrat, J., *Synchronization and Linearity, An Algebra for Discrete Event Systems*, Wiley and Sons, 1992.

[3] Braker H., *Algorithms and Applications in Timed Discrete Event Systems*, PhD thesis, Delft University of Technology, 1993.

[4] Brunsch T., Hardouin L., Raisch J., *Modelling Manufacturing Systems in a Dioid Framework*, [in:] *Formal Methods in Manufacturing*, ed. Campos J., 29-94.

[5] Cohen G., Dubois D., Quadrat J., Viot M., *Analyse du comportement périodique des systèmes de production par la théorie des dioïdes*, Rapport de recherche 191, INRIA, Le Chesnay, France, 1983.

[6] Cohen G., Dubois, D. Quadrat J., Viot M., *A linear system theoretic view of discrete event processes and its use for performance evaluation in manufacturing*, IEEE Trans. on Automatic Control, AC–30, 1985, 210-220.

[7] Cohen G., Gaubert S., Quadrat J.P., *Linear Projectors in the max-plus Algebra*, Proceedings of the IEEE Med. Conf, Cyprus, Jul, 1997.

[8] Cohen G., Gaubert S. Quadrat J., *Max-plus algebra and system theory: Where we are and where to go now*, Annual Rev. Elsevier-IFAC, Vol. 23, No. 1, 1999, 207-219.

[9] Cohen G., Moller, P., Quadrat, J., Viot, M., *Algebraic Tools for the Performance Evaluation of Discrete Event Systems*, IEEE Proceedings, Special issue on Discrete Event Systems, 77(1), 1989, 39-58.
[10] Cottenceau B., *Contribution à la commande de systèmes à événements discrets, synthèse de correcteurs pour les graphes d'événements temporisés dans les dioïdes*, Thèse, LISA – Université d'Angers, 1999.
[11] Cottenceau B., Hardouin L., Boimond J.-L., Ferrier J.-L., *Model reference control for timed event graphs in dioids*, Automatica, vol. 37, 2001, 1451-1458.
[12] Cuninghame-Green R., *Minimax Algebra*, "Economics and Mathematical Systems", Springer, 1979.
[13] Davey B., Priestley H., *Introduction to Lattices and Order*, Cambridge University Press, 1990.
[14] Gruet B., *Structure de commande en boucle fermée des systèmes à événements discrets*, DEA, LISA – Université d'Angers – France, 1995.
[15] LeBoudec J.-Y., Thiran P., *Network Calculus*, Springer Verlag, 2002.
[16] Lhommeau M., *Etude de systèmes à événements discrets dans l'algèbre (max,+), 1. Synthèse de correcteurs robustes dans un dioïde d'intervalles. 2. Synthèse de correcteurs en présence de perturbations*, Thèse, LISA – Université d'Angers, 2003.
[17] Lhommeau M., Hardouin L., Cottenceau B., Jaulin L., *Interval analysis and dioid: application to robust controller design for timed event graphs*, "Automatica", vol 40(11), 2004, 1923-1930.
[18] Libeaut L., Loiseau J.-J., *On the control of timed event graphs*, [in:] Proceedings of WODES'96, Edinburgh, 1996.
[19] Lotito P., Mancinelli E., Quadrat J.-P., *A minplus derivation of the fundamental car-traffic law*, Report 324, 2001, INRIA.
[20] Maia C., Santos-Mendes R., Hardouin L., *Some Results on Identification of Timed Event Graphs in Dioid*, [in:] 11th IEEE Mediterranean Conference on Control and Automation, MED'2003, Rhodes, Grèce.
[21] Menguy E., Boimond J.-L., Hardouin L., Ferrier J.-L., *Just in time control of timed event graphs, update of reference input, presence of uncontrollable input*, "IEEE Transactions on Automatic Control", vol. 45, no. 11, 2000, 2155-2159.
[22] Raszka J., Jamroż L., *Max-Plus linear system in control of data processing*, "Technical Transactions. Fundamental Sciences", Y. 111, iss. 16, 2-NP, 2014.
[23] Raszka J., Jamroż L., *Reducing human resources in management of information technology (IT) projects*, HSI, IEEE, 2015.
[24] Raszka J., *Wspomagana komputerowo analiza stanów dynamicznych walcowni ciągłej blach walcowanych na zimno oparta na kompleksowym modelu matematycznym*, Biblioteka AGH, Kraków 1987.
[25] Schutter B.D., van den Boom T., *Max-plus algebra and max-plus linear discrete event systems: An introduction*, Proceedings of the 9th International Workshop on Discrete Event Systems (WODES'08), Göteborg, Sweden, pp. 36-42, May 2008.
[26] Hardouin L., *Sur la commande lineaire de systemes a evenements discrets dans l'algebre (max,+)*, Automatic. Universite d'Angers, 2004.
[27] Gnuplot homepage [on line], http://www.gnuplot.info (access: 01.07.2016).
[28] Software tools to handle periodic series in dioid http://perso-laris.univ-angers.fr/~hardouin/outils.html (access: 01.07.2015).

TECHNICAL TRANSACTIONS | **CZASOPISMO TECHNICZNE**
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
1-NP/2016

DARIUSZ ŻELASKO*, KRZYSZTOF CETNAROWICZ**, KRZYSZTOF WAJDA***, JAROSŁAW KOŹLAK**

PAY&REQUIRE AS CONCEPT OF VARIABLE COST ROUTING IN DYNAMICALLY RECONFIGURED NETWORKS

PAY&REQUIRE JAKO KONCEPCJA TRASOWANIA O ZMIENNYM KOSZCIE DLA DYNAMICZNIE REKONFIGUROWANYCH SIECI

Abstract

This article presents a new concept of providing service quality with a level required by customers as outcome of an agent system decision making. Decisions are based on network state information and customer requirements concerning transmission quality with accepted costs of transmission. The proposed mechanism, which is called Pay&Require, combines features of decentralized routing systems in computer networks and the newly implemented concept of centralized control determined as SDN (Software-Defined Networking).

Keywords: routing, *QoS*, *agent systems*, *SDN*, *PBR*, *Pay&Require*

Streszczenie

W artykule zaprezentowano koncepcję zapewniania jakości usługi na poziomie wymaganym przez klienta będącą wynikiem procesu decyzyjnego system agentowego. Decyzje oparte są na rzeczywistym stanie sieci oraz wymaganiach klienta dotyczących jakości transmisji z zaakceptowanym kosztem. Zaproponowany mechanizm, nazwany Pay&Require, jest kombinacją zdecentralizowanego trasowania w sieci i nowej koncepcji scentralizowanego zarządzania siecią nazwanej SDN (Software-Defined Networking).

Słowa kluczowe: ruting, *QoS*, *systemy agentowe*, *SDN*, *PBR*, *Pay&Require*

DOI: 10.4467/2353737XCT.16.152.5763

* Dariusz Żelasko (dzelasko@pk.edu.pl), ICT Institute, Faculty of Physics, Mathematics and Computer Science, Cracow University of Technology.

** Krzysztof Cetnarowicz, Jarosław Koźlak, Department of Computer Science, Faculty of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology.

*** Krzysztof Wajda, Department of Telecommunications, Faculty of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology.

1. Introduction

A modern network ensures the efficient and effective information transfer between different types of users: individuals or corporate. For many years research was carried out in order to define and implement mechanisms and architectures, which would provide the diversifying qualities of carried out data transmission, and additionally possibility to define and implement transmission according to determined QoS (Quality of Service) parameters.

Service providers usually charge a fee based on the maximum bandwidth possible to obtain. Unfortunately, this bandwidth is rarely available mainly due to the temporary load of individual nodes and links in the network. Temporal throughput is determined by the least efficient bond in transmission between sender and recipient. Different types of methods and architectures were defined, which enable the management of network traffic using different approaches to the concept of service type, priority and category of transmitted information. For classical IP networks models of IntServ and DiffServ services were proposed, for multiservice network a complete ATM technology was defined, displaced by MPLS. Nowadays, there are novel concepts, such as Software-Defined Networking (SDN), in which the assumed control of the transport layer separation causes a greater network performance. In the case of SDN it was assumed that management is carried out centrally i.e. there is a central repository storing essential rules for network management, created based on information collected from all over the network. Centralization does not always seem to be a good solution due to performance issues (scalability of solution), safety of collected and stored information (one node collecting information).

For the implementation of decentralized network management it is possible to propose the use of agent system concept. The agent approach for the implementation of routing in the graph was proposed in the work [1]. The recalled agent approach in the present article was enriched with the Pay&Require (P&R) concept, in which the separation of the transport layer was assumed (devices physically responsible for transport) from the control (the logic of the system), and additionally decentralization of the control carried out in the network. For that purpose an agent technology was used, which enables the avoidance of the application of a central repository. One of the key aspects of this concept is the fact that the user pays for a particular required connection quality. An important aspect of the article is a reference to Software-Defined Networking, which is still a novel concept, but it seems that it may provide a solution widely accepted in defining the future of computer networks.

In order to analyse the proposed concepts as well as to determine whether the use of P&R is reasonable, an emulating environment of the proposed mechanism was carried out, and the study results conducted on the network model (built for the project's purposes) were presented in this article.

2. The concept of SDN decentralization

SDN (Software-Defined Networking) [2, 3, 9, 10, 11, 12, 14] is a network architecture, in which control layer and data transmission are separated. The separation of layers allows for the introduction of a certain level of abstraction that facilitates the configuration

process – the administrator does not need to have a specialized knowledge concerning the configuration of the transport layer, just knowledge of the control layer management, which takes care of the correctness of information provided to the transport layer. The management process is centralized, and the architecture is independent of the topology or the applied network technique. Centralized management is supposed to be among others able to obtain information about full network topology. Important is the assumption that the interface of information exchange between layers is supposed to be available on the principles of open standards and protocols. As a result, modification of the network by external applications will be possible. It consists in the fact that there are established rules concerning (e.g. packet management) the operation of the transport layer.

Generally, the effect of the implemented SDN concept is supposed to be: increased flexibility of solution, centralization of control, simplification of the construction of network equipment and independence from individual producer solutions.

The solution applying SDN concept significantly facilitates network management, but it also has some imperfections. The primary one is a large amount of data that must be stored in a central repository. This results in a situation that when you want to download some rules, it is necessary to search through an entire database of substantial size. Also, the use of storage mechanisms in the cache may not resolve the problem. The downloading and uploading process causes an additional load on the network, which can lead to slow transmission and this in turn can directly cause delays. The suggested solution seems to be also susceptible to failures mostly caused by the centralized repository of the rules. A breakdown of a connection to the repository will cause the entire network to fail.

Additionally, there is a problem of transfer security. Assuming that the control information is sent with the same connections as customer data, it is necessary to think about ensuring the security of the system operation. Let us consider the case where a customer eavesdrops on the control data and then their modified version is sent to the network in order to prevent proper operation of the network or for other reasons, causing disruption to the entire network. In this case, it is necessary to apply appropriate security – e.g. transmission encryption or other mechanisms. These types of mechanisms will cause additional delays in the transmission (exchange) of rules. Another problem constitutes the fact that when decisions are supposed to be taken with regard to individual packets the process of searching for relevant rules

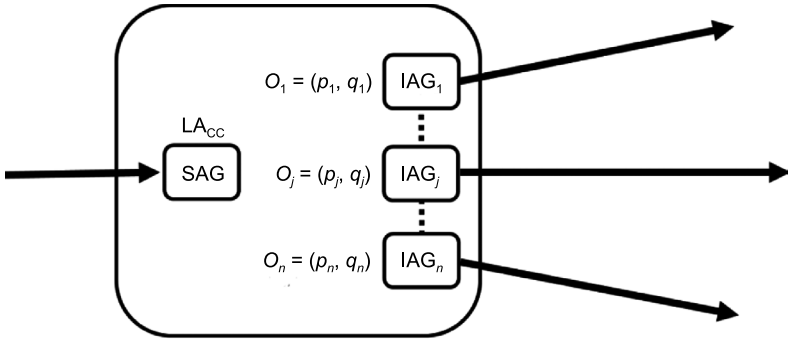

Fig. 1. Scheme of decision making process for packets routing

in a central repository will last long enough so that it will not be possible to call it real-time processing.

All these imperfections result from applying centralized system, created according to the SDN concept. It seems that a decentralization of such a system would solve these problems. Since an agent approach is one of methods of creating decentralized systems, it is necessary to consider exactly this approach for the implementation of a decentralized quality support system for the customer (user).

3. The concept of an agent system as SDN decentralized

To implement the concept of a decentralized SDN it is possible to use the agent approach, which is a known method for implementing decentralized systems. The adoption of the agent system concept requires defining agents appearing in the system. In the proposed solution the following types of agents were introduced:

- NAG – agent in one copy on each node (router). Its task is to manage a given node particularly in directing packets in the right direction.
- SAG – agent that represents the stream of information. Depending on a specific solution, this agent can be a packet carrying information or a packet setting the way (routing).
- IAG – agent that represents the output interface.

The SAG agent further way choice algorithm can be carried out as follows (Fig. 1):

- SAG agent comes to a given node (router).
- IAG agent presents the offer for SAG agent in relation to a further way commencing on a linked interface with this IAG agent.
- SAG agent based on its needs chooses the most convenient offer from those offered by IAG agents. After choosing the recalled offer the SAG agent continues its way through the chosen interface. The NAG agent can help the SAG agent in decision-making regarding the choice of offer.

It remains to define the algorithm by which a decision is taken by the *SAG* agent of the proposed offers by the IAG agent. Of course, it depends on the form of offer and the way of its determination by the IAG agent.

In the presented solution the market approach to determine choice of offer was proposed. For this purpose the Pay&Require conception was developed.

4. The Pay&Require concept

The main aim of the Pay&Require approach is to provide the quality of service (transmission) that meets client requirements. It is worth applying a market approach as a form of customer-supplier negotiations of services/resources.

Requests to provide quality services required by a customer can be implemented in such a way that the user pays for the guaranteed transmission quality i.e. the actual transmission parameters for a packet from the source node to the destination node. This quality is guaranteed by the application of an appropriate routing protocol version, which combines static and

dynamic protocol, as well as agent technique. In this case, the concept of quality refers to the most often used parameters of the QoS network such as delay, delay fluctuation, transmission time or the level of packet losses. Introducing such a concept of quality makes it possible to determine certain rates for guaranteed quality (and not, as it is in computer networks for maximum possible to obtain bandwidth, which in many cases is never achieved). It can be assumed that all packets from a given user will be transmitted at a fixed path defined statically in the network layer. To obtain a situation in which there will be a separate route for each user is in practice usually impossible and users must share the same path. When there are many users in the network who require a high level of quality and packets from these users are transmitted in the same route, it may happen that some of the connections will be overloaded. A deterioration of transmission, below the level for which individual users paid, will be a consequence of this occurrence. Then a change of routes should be carried out, i.e. establishing new routes for recalled users. For that purpose a control layer was defined (system logic), which monitors the state of individual connections and in the case of deterioration in quality, it reconfigures the network by defining new routing tables. [1, 4]

The proposed solution agents (NAG) residing on routers, are concerned with monitoring the state of the connections and the configuration of the network layer. One router reports to each agent (or group of routers). Agents exchange information necessary for defining the process of the current form of the routing tables in order to obtain their mutual cohesion and send recalled updated (reconfigured) routing tables to routers which after receiving the tables start to route packets according to the adopted P&R algorithm of the defining paths. In the end packet which do not require the quality on a high level can be directed by a completely different path on which e.g. long delays appear, and the customer will pay less for such a path agreeing to the lower quality of the transmission. As a result, the quality of service will be adapted to requirements, which are expressed by the amount of fees a customer (packet) is ready to incur. In consequence it is possible to state that such an approach also enables pricing (by SAG agents) based on market methods. At fixed prices the users can systematically bid individual levels of quality, and hence the path depending on the demand and availability of paths at the given moment. The proposed approach enables the application of different market methods to buy quality, which will affect the dynamic pricing of individual paths – it will enable development and make use of the supply and demand model of price determination.

From an agent point of view the control layer system constitutes agents, which will reside on routers as well as monitor and change corresponding parameters. As a result of the application of the agent approach decentralization was obtained, which causes reduced susceptibility to network failures (there is no central database containing rules). It is possible to consider two solution levels to the decentralization problem:

– Level 1 – each agent has full knowledge of network topology. This type of approach means that each agent provides to all agents across the network information about the networks connected to the router, on which it resides, in effect each agent has a full information about all routers. In order to achieve consistency of such information, it is necessary to exchange large amounts of data at every change in the network. However, in the case of failure there is no need to exchange additional information the agent can immediately and independently update relevant data or reconfigure the network.

– Level 2 – every agent has only the necessary information i.e. about networks connected to the given router. This approach assumes that the agent stores in its local base only this information that is necessary from the point of view of its function. As a result, when you start an agent it can start working in a relatively short time. In the case of breakdown or a change of network topology there may be a need to obtain additional information. An essential extension can be such a solution in which each agent residing on the router has partial information about the entire network and full information about the nearest neighbourhood.

5. Implementation of a model solution

A study of the proposed agent routing concept, using the Pay&Require approach, was carried out by creating a solution for level 1, in which every agent has exactly the same set of information and knowledge about the entire network topology. It seems that this case will allow us to state whether an application of the proposed concept will affect the quality of data transmission. In individual nodes (routers) an SAG agent selects appropriate further routes (paths) by comparing the conditions offered by the IAG agents associated with individual output routers. In the studied solution each SAG agent that represents packets sent by an established sender has a determined level of price lpacc approval – an established level of maximum price, and quality which is determined by $\mathrm{Par}_{\min}$ and $\mathrm{Par}_{\max}$ parameter defined as percentage deviation from quality level.

T a b l e 1

Exemplary features of links for various service levels

Rate	Bandwidth [Mbit/s]
4	100
3	50
2	10
1	5
0	1
–1	link inactive

On every router there are IAG agents – each of them is connected to one interface it represents. Every IAG_j agent presents to SAG agent the connection offer $O_j = (p_j, q_j)$ where p_j – price of the connection offered by IAG_j agent, and q_j – quality of the offered connection. Next the SAG agent (if necessary with the help of the NAG agent) determines a set of acceptable Of_{acc} offers:

$$Of_{\mathrm{acc}} = \{Of : Of = (p,q),\ p \le l_{p_{\mathrm{acc}}} \wedge q \in [Q_{\min}, Q_{\max}]\} \tag{1}$$

where

$$Q_{\min} = q - (\mathrm{Par}_{\min} * q), \quad Q_{\max} = q + (\mathrm{Par}_{\max} * q)$$

Next the SAG agent selects the best Of_k from acceptable offers, according to the rule:

$$Of_k = \min_{Of_j \in Of_{\text{acc}}} \{p_j : Of_j = (p_j, q_j)\} \tag{2}$$

Summing up: the presented offers by the IAG agents determine a further connection (related to quality and price) and the SAG agent chooses the best quality offer out of those which price is within acceptable limits.

6. Implementation of the emulation studied examples

For the purpose of the project an emulator was prepared, consisting of software providing functions of control layer and routers in the transport layer. The Vyatta system (VyOS) was used for routing. The first activity which the control layer carries out is to download the initial router configuration concerning the interfaces. The downloaded configuration is carried out for each router individually. When the software has downloaded configuration of all routers an analysis of the information takes place. The analysis consists of searching for active connections between individual routers.

It was assumed that between routers there are point-to-point connections. Software stores information about all connections between the routers. Next, a bandwidth of individual connections is established in order to carry this out, through every single connection a file of fixed size is sent (e.g. 10 MB). The size of the file can be chosen arbitrarily, however, in the study it was stated that 10 MB was an appropriate size. Information concerning transmission time and average bandwidth in bit/s was obtained in this way. Information of this type is stored for each link. Then based on measured bandwidth an evaluation of the route is appointed. A rating scale resulting from measured parameters was defined. This scale can be freely modified by the administrator – it is possible to define any rating scale.

For the presented emulation purposes a quality scale expressed by the allotted bandwidth was used, presented in Table 1. A maximum bandwidth of 100 Mbit/s is caused by limitations of virtualizing software used for emulation. Evaluation is assigned to individual links.

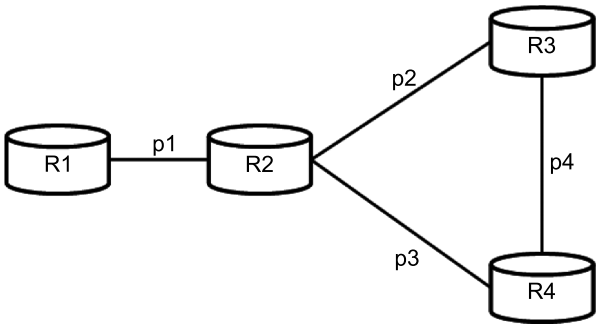

Fig. 2. Network example – characteristics of routing algorithm

The next step is to define paths. Paths lead from the source to the destination router through other routers. In order to outline all possible paths the following algorithm is applied:

1. Take the first free path of length 1 (linking only two routers). Proceed to point 2.
2. Outline all possible non-repeating paths (no loops - point-to-point connection can be used only once) for length +1. Follow step 2 for so long until there is no longer path to choose from – you cannot proceed further because all available point-to-point links were used.
3. Repeat the step from 1–2 for all point-to-point connections.

In Figure 2 a network used for the emulation was presented. Let R be a set of routers, R = {R1, R2, R3, R4}, let P be a set of connections between routers, P = {p1, p2, p3, p4} where p1 = {R1, R2}, p2 = {R2, R3}, p3={R2, R4}, p4={R3, R4}.

PR1={{p1}}
PR1={{p1},{p1,p2},{p1,p3}}
PR1={{p1},{p1,p2},{p1,p3}}
PR1={{p1},{p1,p2},{p1,p3},{p1,p2,p4},{p1,p3,p4}}
PR1={{p1},{p1,p2},{p1,p3},{p1,p2,p4},{p1,p3,p4},{p1,p2,p4,p3},{p1,p3,p4,p2}}

Evaluations are assigned to individual paths. It is possible to consider two possible approaches:

– appoint an average evaluation of the path based on evaluations of individual point-to-point connections on a way from the source (the first router in a given path) to the destination (the last router).
– as the evaluation of entire path taking the lowest evaluation of a point-to-point connection on the way from the source to the target.

Use of the second approach seems to be preferable because a connection with the lowest bandwidth will reduce the bit rate on the entire route. Evaluations of paths are stored along with information about the number of routers which a packet must go through in order to reach transmission target in a given path. This information will be used later to take a decision about the choice of paths in a situation where several paths have the same evaluation. Another step is to select a path for the customer. In the application information about a client ID and expected service level are stored (in accordance with a rating scale). At first, there is a verification to which routers the customers are connected (network configuration enables transmission between individual customers). Next all possible paths are outlined from one customer to all remaining customers (this process is carried out for all customers). For accepting a given route its evaluation decides – if this is what a client expects or higher (when there is no expected), then this route will be selected.

In the case that there is no route with the quality level for which the customer paid, or higher, a message will appear about the lack of routes. In the future it is necessary to consider how to solve this problem (e.g. a refund, negotiating with the customer). From acceptable paths one is chosen – the one which has a low hop count. If there is more than one route with the same rating then the first from the list will be chosen. When all paths will be chosen for all customers, it will be followed by a configuration of routing layer (routers).

Routing and forwarding in computer networks is based on information brought in the header of packet, i.e. for transmission purposes. In case of classical routing the device compares the address of the network to which the packet is supposed (longest prefix matching) to be sent with addresses included in its own routing table and based on this information router redirects the packet to the next router or target device. In Figure 3 a scheme of computer network is presented, which will be used to describe the operation

method. In the case of the described network between PC1 and PC2 computers and a PC3 computer there are two paths. The first path leads through R1-R2-R3, and the second leads through R1-R4-R3. Metrics were assigned to individual connections. It was assumed that if all the metrics have a lower value, the connection is better. Therefore, according to the rules of classical routing the transmission between PC1-PC3 and PC2-PC3 will proceed exactly along the same route (R1-R4-R3).

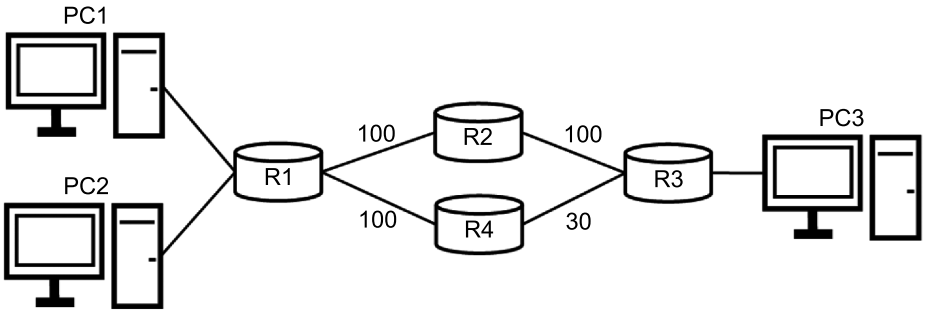

Fig. 3. Network example – classic routing and P&R comparison

In computer networks the PBR concept was defined (Policy Based Routing). [5–8, 13]. This technique allows the administrator to define complex routing rules, i.e. a decision about the next change can be taken not only based on destination address, but also e.g. on source of transmission, port number. PBR enables the definition of more than one routing table – for each user (groups) the administrator can independently define which transmission path will be stored. Of course, a substantial matter is that there should be more than one path between the source and target – then PBR makes sense. By analysing the case presented in Figure 3 the administrator has the possibility to configure routers, so that packets from a PC1 computer sent to a PC3 can travel a different route (e.g. R1-R2-R3) than packets from PC2 sent to PC3 (e.g. R1-R4-R3). Therefore, with the use of PBR it is possible to diversify the routes depending on the transmission source. PBR technique was used in the emulator for the routing purpose. After conducting the configuration of the transport layer (the routing tables) the network begins to operate in accordance with customer expectations.

7. Research results

Emulation was started by setting the network of 4 routers (Figure 2). To router R1 and R3 – 4 users were connected. Users were simulated with the following requirements concerning quality: two users needed the best quality (=4) – to router R1 and R3 one such user per router was connected, two users needed low quality (=2) – to router R1 and R3 one such user per router was connected.

An aim of the emulation was to present changes in the choice of paths depending on the bandwidth of the individual connections. In the case of the used network a total separation of transfer does not appear, since between router R1 and R2 an alternative connection does

not appear. In the emulation paths outlined between users connected to R1 (high and low quality) and R3 were taken into account. It was started by checking selected paths in a time when there is no fault in the network.

Figure 4 presents the result for the network with maximum operating parameters. A chosen route was marked with a thickened line in this case for both customers the same path was chosen. This path has the smallest number hops, and parameters concerning quality are comparable with the available paths. Another stage of the emulation assumes a degeneration of the connection parameters between R1 and R2. This fault did not cause a change in the chosen path since an alternative connection does not exist in the above network for R1-R2. It means that still the chosen path corresponds to the one presented in Figure 4. The next emulation was carried out when a bandwidth decreased between R2 and R4. In the case of a customer requiring the best quality the path did not change – this path has the maximum quality (Fig. 4). In turn, the path for a customer requiring low quality changed. A chosen path was presented in Fig. 5. It is possible to observe that a chosen path is longer than the one which was chosen in a previous emulation. It is due to the fact that the user paid for low quality and such quality he received. In the previous emulation there were no possibilities to provide quality at the expected level – only the best quality

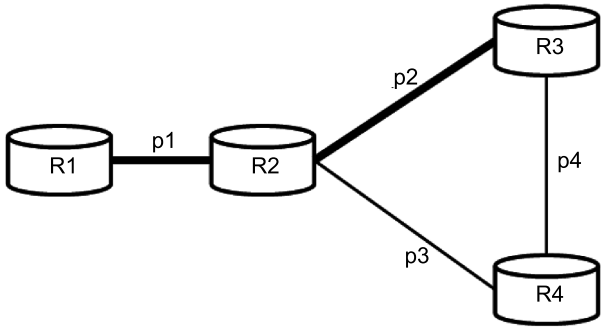

Fig. 4. The selected path for the network operating with maximum throughput

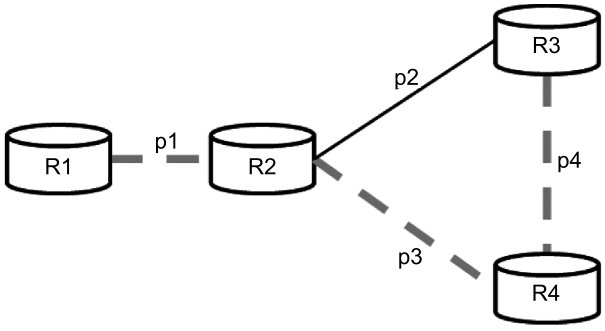

Fig. 5. The selected path for the customer who accepts low quality in case of network with worse R2-R4 link parameters

appeared and so the user also received it. This type of decision results from the characteristics of the algorithm of path choice.

The next emulation was a reduction in bandwidth R3-R4. In this case, this fault did not cause the path change of the user requiring the best quality – the earlier path is still the best choice (Fig. 4). However, in case of the user expecting low quality the same path was chosen as in case of the deterioration connection parameters R2-R3 (Fig. 5). Choice of this type results from the fact that this path still provides the expected quality i.e. low level.

The last emulation case was a reduction in bandwidth between R2 and R3. In this case, for the user requiring best quality a chosen path was presented in Figure 6. Next for the user requiring low quality a chosen path was presented in Figure 7. Such a choice of path results from the fact that route R1-R2-R4-R3 guarantees quality at a high level, despite the fact that number of routers through which the packet must pass is higher. In turn, route R1-R2-R3 in this case has low quality what is a result of the degeneration connection parameters between R2 and R3.

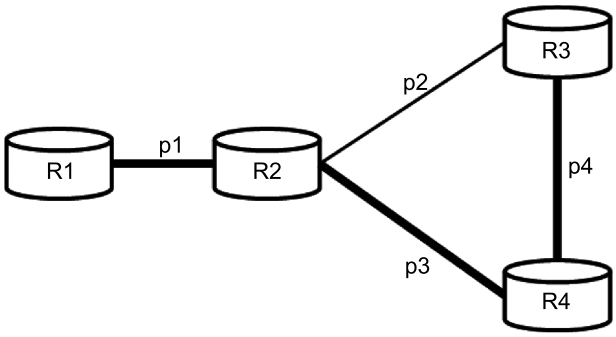

Fig. 6. The selected path for the customer who accepts high quality in case of network with worse R2-R3 link parameters

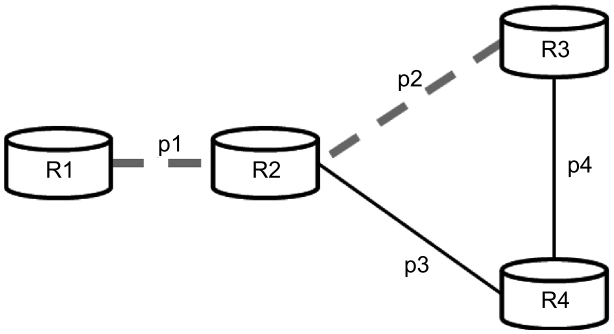

Fig. 7. The selected path for the customer who accepts low quality in case of network with worse R2-R3 link parameters

Figure 8 presents comparison of transmission times in the case of classical routing and Pay&Require. In order to determine times two measuring tools were used: ping and

transmission of a file with a fixed size of 100 MB. Ping was carried out 1000 times for each case, and file transfer was carried out 30 times. In order to determine the reference time measured when the network operated with maximum parameters and the same path was chosen for both customers. Results for both qualities are similar and the differences are slight.

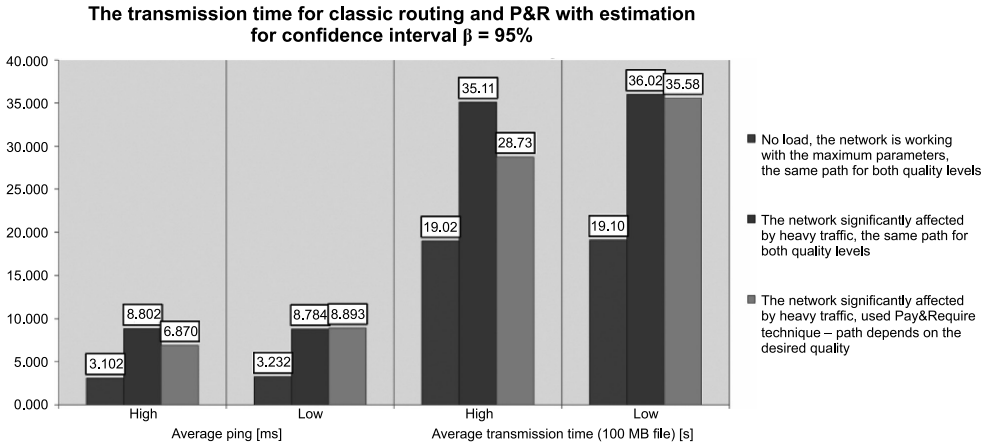

Fig. 8. Comparison of the transmission time for classic routing and P&R

In a further step it is necessary to state how transmission time changes in the case of overload network. The network was overloaded by the initiation of many simultaneous transmissions of large data between devices connected to R1 and R3. At the same time measurements of transmission time were carried out in the case of requiring high and low quality. It is possible to observe a big increase in transmission time – in the case of ping the time increased almost threefold.

In turn, the average transmission time of a file with sizes of 100 MB increased almost twofold. Because of the above two measurements it is a point of reference for the last research stage.

In the last stage a measurement of times for the same overload network was carried out using Pay&Require. The customer requiring high transmission quality received a different path than the one requiring low quality. The path for low quality is the same as the one which packets that overload the network are sent. It is possible to observe that transmission time for low quality in comparison to the previous case practically did not change, however, in the case of the best quality a significant improvement was obtained. Transmission times significantly decreased. Unfortunately, the connection between R1 and R2 constitutes a section that affects transmission quality, since every transmission between R1-R3 must go through it. Therefore, it is possible to suspect that if there was an alternative for connection R1-R2, then transmission time for the best quality would be reduced and similarly for transmission in the case of the unloaded network.

The conducted research suggests that the Pay&Require concept has merit and constitutes an alternative to methods of providing quality and pricing in computer networks.

8. Conclusions

The Pay&Require concept presented in this article may constitute an alternative to methods providing quality and pricing in computer networks. The user pays for guaranteed transmission parameters, which are practically implemented as a result of the choice of appropriate path for transmission. Quality parameters of individual paths are systematically monitored and, if such a need occurs, paths are modified, because of decentralized function of the agent system.

The presented concept refers to Software-Defined Networking technology, which constitutes a good starting point for a definition of a new mechanism for the separation of the control layer from the transport layer. There was an attempt to remove SDN imperfections specified in the article, such as centralization of the solution. In the case of the P&R mechanism a decentralization of control and agent technique was used. It was necessary to carry out a study aimed at the state of the legitimacy of the application of the P&R concept.

Research results show that the use of the P&R mechanism to provide specific quality parameters caused the desired effects, i.e. a significant improvement was obtained in relation to classical routing. Thus, it is possible to state that quality was provided at a level expected by the customer. The conclusion from the conducted study is clear, i.e. the established effects of the P&R mechanism application were achieved, therefore this approach is promising and should be developed. It is necessary to carry out further research for more complex networks in order to verify the performance of the algorithm and optimization of its operation in different conditions.

References

[1] Maj A., Jurowicz J., Kozlak J., Cetnarowicz K., *A Multi-agent System for Dynamic Network Reconfiguration*, "Proceedings of the 3rd international Central and Eastern European Conference on Multi-Agent Systems, CEEMAS 2003, Prague, Czech Republic", 2003.

[2] Gelberger A., Yemini N., Giladi R., *Performance Analysis of Software-Defined Networking (SDN)*, "IEEE 21st International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems", 2013, 389-393.

[3] Santos M.A.S., Nunes B.A.A., Obraczka K., Turletti T., De Oliveira B.T., Margi C.B., *Decentralizing SDN's control plane*, "IEEE 39th Conference on Local Computer Networks", 2014, 402-405.

[4] Subramanian Neelakantan, Praveen Ampatt, Muraleedharan N., Subhash Chandra Bose Korimilli, Arun Parmar, Manish Kumar, Anupam Roy, *An Architecture for self-configuration of network for QoS and Security*, "Communication Systems and Networks and Workshops", 2009, 1-5.

[5] Nanda P., *Supporting QoS Guarantees Using Traffic Engineering and Policy Based Routing*, "International Conference on Computer Science and Software Engineering", 3/2008, 137-142.

[6] Boschi E., Carle G., *Active control architecture implementing policy-based routing*, "10th International Conference on Telecommunications", 1/2003, 53-57.

[7] Chi-Kin Chau, Gibbens R., Griffin T.G., *Towards a Unified Theory of Policy-Based Routing*, "25th IEEE International Conference on Computer Communications, Proceedings", 2006, 1-12.

[8] Żelasko D., *Policy-Based Routing na przykładzie systemu Vyatta*, „Przegląd Elektrotechniczny", 8/2014, 46-49.

[9] Sezer S., Scott-Hayward S., Chouhan P.K., Fraser B., Lake D., Finnegan J., Viljoen N., Miller M., Rao N., *Are we ready for SDN? Implementation challenges for software-defined networks*, "Communications Magazine", 51-7/2013, 36-43.

[10] Scott-Hayward S., O'Callaghan G., Sezer S., *SDN Security: A Survey*, "IEEE SDN for Future Networks and Services", 2013, 1-7.

[11] Jie H., Chuang L., Xiangyang L., Jiwei H., *Scalability of control planes for Software defined networks: Modeling and evaluation*, "IEEE 22nd International Symposium of Quality of Service", 2014, 147-152.

[12] Alvizu R., Maier G., *Can open flow make transport networks smarter and dynamic? An overview on transport SDN*, "International Conference on Smart Communications in Network Technologies", 2014, 1-6.

[13] Smith B.R., Garcia-Luna-Aceves J.J., *Efficient policy-based routing without virtual circuits*, "Quality of Service in Heterogeneous Wired/Wireless Networks", 2004, 242-251.

[14] Seongbok B., Chankyou H., Youngwoo L., *SDN-based architecture for end-to-end path provisioning in the mixed circuit and packet network environment*, "16th Asia-Pacific Network Operations and Management Symposium", 2014, 1-4.

CONTENTS

TREŚĆ