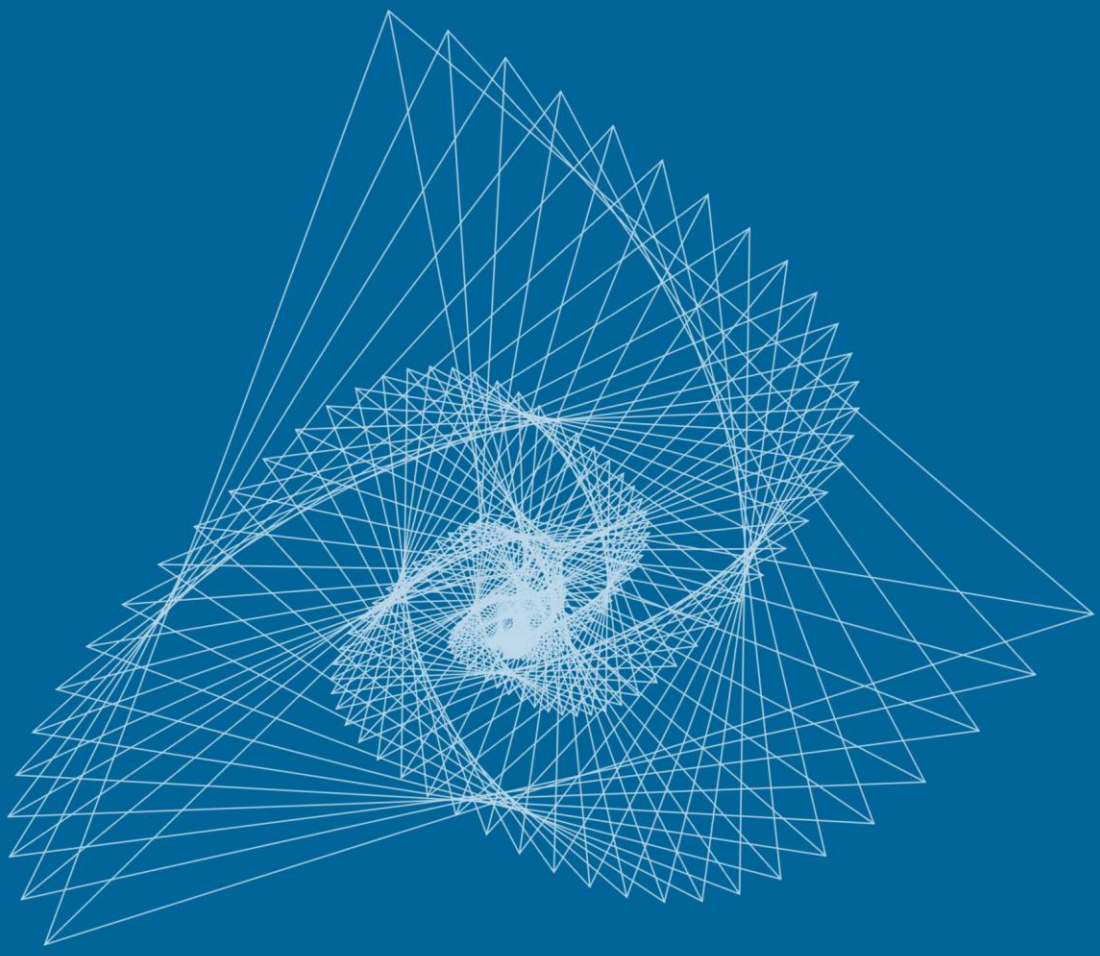
VOL. 5, NO. 2

URAL MATHEMATICAL JOURNAL

N.N. Krasovskii Institute of Mathematics and Mechanics of
the Ural Branch of Russian Academy of Sciences and
Ural Federal University named after the first President of Russia B.N.Yeltsin

ISSN: 2414-3952



UMJURAN.RU
UMJ

2019

ISSN 2414-3952
(Online)

Ural Mathematical Journal

Vol. 5, no. 2, 2019

https://umjuran.ru

Electronic Periodical Scientific Journal
Founded in 2015

The Journal is registered by the Federal Service for Supervision in the Sphere of Communication, Information Technologies and Mass Communications
Certificate of Registration of the Mass Media Эл № ФС77-61719 of 07.05.2015

Founders

N.N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of Russian Academy of Sciences

Ural Federal University named after the first President of Russia B.N. Yeltsin

Contact Information

16 S. Kovalevskaya str., Ekaterinburg, Russia, 620990
Phone: +7 (343) 375-34-73 Fax: +7 (343) 374-25-81
Email: secretary@umjuran.ru
Web-site: https://umjuran.ru

TABLE OF CONTENTS

URAL MATHEMATICAL JOURNAL, Vol. 5, No. 2, 2019, pp. 3–12

DOI: 10.15826/umj.2019.2.001

INTERPOLATING WAVELETS ON THE SPHERE[1]

Nikolai I. Chernykh

Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences,
16 S. Kovalevskaya Str., Ekaterinburg, 620990, Russia

Ural Federal University,
51 Lenin aven., Ekaterinburg, 620000, Russia

chernykh@imm.uran.ru

Abstract: There are several works where bases of wavelets on the sphere (mainly orthogonal and wavelet-like bases) were constructed. In all such constructions, the authors seek to preserve the most important properties of classical wavelets including constructions on the basis of the lifting-scheme. In the present paper, we propose one more construction of wavelets on the sphere. Although two of three systems of wavelets constructed in this paper are orthogonal, we are more interested in their interpolation properties. Our main idea consists in a special double expansion of the unit sphere in $\mathbb{R}^3$ such that any continuous function on this sphere defined in spherical coordinates is easily mapped into a 2π-periodic function on the plane. After that everything becomes simple, since the classical scheme of the tensor product of one-dimensional bases of functional spaces works to construct bases of spaces of functions of several variables.

Keywords: Wavelets, Multiresolution analysis, Scaling functions, Interpolating wavelets, Best approximation, Trigonometric polynomials.

Introduction

Different systems of wavelets on the sphere are constructed and studied in a number of works. We would like to note the constructions in the paper by Skopina [9]. They are beautiful but difficult to put in practice, as their author notes herself. In [2], the ideas of these constructions were extended to spheres in $\mathbb{R}^n$. These and some other works mentioned below contain a good analysis of the studies on the specified or close subject. In [4, 6, 7], to construct bases of wavelets on spheres in $\mathbb{S}^2$ and $\mathbb{S}^3$, the tensor product of bases of one-dimensional wavelets is used including a basis of exponential splines on a segment. In the papers [1, 5, 8], which contain much of the bibliography related or close to the subject and the analysis of the previous results, in particular, the lifting-scheme technique is used to construct biorthogonal wavelets on the sphere. This is accompanied by rejecting a number of properties of classical wavelets including, for instance, shifts with a constant step at each scaling level and with localization of the compression–stretching operation in the right places. In the present paper, we attempt to preserve the standard properties of classical wavelets on the line and on the period in the construction of wavelets on the sphere. In so doing, we give preference to interpolating wavelets. Orthogonal wavelets are only defined. The study of their approximative properties is postponed for the future. Here, for the classical schemes to construct wavelets on the sphere to work, we carry out a double expansion of the unit sphere with a special extension to it of the function originally defined on the sphere. This makes it possible to apply one-dimensional periodic interpolation and interpolation–orthogonal bases of expanding subspaces of

[1] This work was supported by the Russian Academic Excellence Project (agreement no. 02.A03.21.0006 of August 27, 2013, between the Ministry of Education and Science of the Russian Federation and Ural Federal University).

multiresolution analysis (constructed and studied in [3, 10]) to construct wavelets on the expanded sphere.

1. Construction of wavelets on the sphere

Without loss of generality, we assume that $\mathbb{S}$ is the sphere of unit radius centered at the origin of a Cartesian coordinate system, (θ, φ) are spherical coordinates of points of $\mathbb{S}$ ($\theta \uparrow_0^\pi$ denotes latitude and $\varphi \uparrow_0^{2\pi}$ denotes longitude) associated in a standard way with the Cartesian coordinates. Thus,

$$\mathbb{S} = \{M(\theta, \varphi) \in \mathbb{S} : \ 0 \le \theta \le \pi, \ 0 \le \varphi < 2\pi\}.$$

For a uniform grid with any small step $h = 2\pi/l$ ($l \in \mathbb{N}, \ l \gg 1$) in the angular coordinates θ, φ, the geometrical sizes of cells of the corresponding grids on the sphere are strongly nonuniform. We have cells with size of order $h \times h$ in $\mathbb{R}^3$ near the equator, where θ is close to $\pi/2$, and we have cells with size of order $h \times h^2$ near the poles, where θ is close to 0 in the case of the north pole N or θ is close to π in the case of the south pole S. On $\mathbb{S}$, every value $\varphi \in \mathbb{T}$ (where φ and $\varphi_{\pm 2l\pi}$ are indistinguishable) determines the φ-meridian, i.e., the great circle arc

$$M_\varphi := \{M(\theta, \varphi) : \ 0 \le \theta \le \pi\},$$

and every value $\theta \in (0, \pi)$ determines θ-latitude, i.e., the circle $M^\theta := \{M(\theta, \varphi) : \ 0 \le \varphi < 2\pi\}$ of radius $r_\theta = \sin\theta$ in the plane $z_\theta = \cos\theta$ centered at the point $(0, 0, z_\theta)$ of the Cartesian system. Despite the noted disadvantage of spherical coordinates and the specified grids on $\mathbb{S}$ uniform in θ and φ, their application is profitable and simple for both the construction of wavelets on $\mathbb{S}$ and the practical use of the wavelets in computational algorithms.

Thus, to construct basis scaling functions of the subspaces $V_j(\mathbb{S}) \subset L^2(\mathbb{S})$ ($j \in \mathbb{Z}_+$) of multiresolu-tion analysis on $\mathbb{S}$, a usual method of passage from one-dimensional to multi-dimensional wavelets can be used here by choosing as those the tensor product of the bases of the subspaces $V_j(\mathbb{T})$ of the space $L^2(\mathbb{T})$ of 2π-periodic functions and the bases of the subspaces $V_j[0, \pi] \subset L^2[0, \pi]$ as done in [4]. It is true that, in this case, one has to use a construction of wavelets on a segment, which is more complicated than that on the line or on a period, for instance, applying a "folding" operation. Instead of this, in the present paper, the sphere "doubles". Due to this, the construction of bases of the subspaces $V_j(\mathbb{S})$ reduces to the tensor product of two (possibly different) bases of the subspaces $V_j(\mathbb{T})$ in the variables φ and θ, respectively.

It is clear that any φ-meridian M_φ is connected with the opposite $(\varphi \pm \pi)$-meridian $M_{\varphi\pm\pi}$ on which as well as on M_φ, by the definition of the spherical coordinates (θ, φ), θ changes from 0 (at the pole N) to π (at the pole S). These two meridians form together the great circle C_φ on $\mathbb{S}$. Keeping the bypass direction of the φ-meridian by the points $M(\theta, \varphi)$ when θ increases on φ-meridian and changing it to the opposite on the $\varphi \pm \pi$-meridian, for every $\varphi \in [0, 2\pi]$, we define the full φ-meridian as follows:

$$C_\varphi := \{M(\theta, \varphi) \in M_\varphi \cup M_{\varphi\pm\pi} : \ 0 \le \theta \le 2\pi\},$$

where $M(0, \varphi) = M(2\pi, \varphi)$.

We note that, although the full φ-meridian C_φ, element-wisely coinciding with $M_\varphi \cup M_{\varphi\pm\pi}$, crosses the equator $\{M(\theta, \varphi) : \ \theta = \pi/2, \ 0 \le \varphi < 2\pi\}$ in the two points $(\pi/2, \varphi)$ and $(\pi/2, \varphi \pm \pi)$, this meridian is completely determined by the value of the angle φ, since its bypass direction with the increase of θ is determined by the movement direction of the point $M(\theta, \varphi)$ along the φ-meridian and is continuously extended to the $(\varphi \pm \pi)$-meridian changing its original direction from N to S to the opposite.

Any function f defined on $\mathbb{S}$ uniquely determines the function $f(\theta,\varphi)$ of the variables $\varphi\in[0,2\pi]$ and $\theta\in[0,\pi]$. In particular, this function is uniquely defined on any full φ-meridian as a function of θ, and to apply periodic wavelets in the construction of wavelets on $\mathbb{S}$, it is very important that the coordinate θ changes on C_φ over the full period from 0 to 2π, since, in C_φ, the function $f(\theta,\varphi)$ is 2π-periodic in θ because the functions $f(0,\varphi)$ and $f(\pi,\varphi)$ on $\mathbb{S}$ do not depend on φ. However, it is easy to see that C_φ and $C_{\varphi\pm\pi}$ coincide as sets of points on $\mathbb{S}$ differing only in the direction of movement of their points M with coordinate θ as $\theta\uparrow_0^{2\pi}$. As a result, every function $f(\theta,\varphi)$ single-valued on $\mathbb{S}$ generates a two-valued function $F(\theta,\varphi)$ of the variable θ on every set $C_\varphi=C_{\varphi\pm\pi}$ and, hence, on $\mathbb{S}$. Namely, for any $\varphi\in[0,2\pi)$, we have

$$F(\theta,\varphi)=\begin{cases} f(\theta,\varphi),\ \theta\uparrow_0^{\pi} & \text{on}\ \ M_\varphi,\\ f(2\pi-\theta,\varphi\pm\pi),\ \theta\uparrow_\pi^{2\pi} & \text{on}\ \ M_{\varphi\pm\pi}\end{cases} \tag{1.1}$$

on C_φ and

$$F(\theta,\varphi)=\begin{cases} f(\theta,\varphi\pm\pi),\ \theta\uparrow_0^{\pi} & \text{on}\ \ M_{\varphi\pm\pi},\\ f(2\pi-\theta,\varphi),\ \theta\uparrow_\pi^{2\pi} & \text{on}\ \ M_{\varphi}\end{cases} \tag{1.2}$$

on $C_{\varphi\pm\pi}$. Obviously, this function completely restores $f(\theta,\varphi)$ already for $0\le\varphi<\pi$. However, it is also important for us to preserve the 2π-periodicity of the function F in φ.

To avoid the two-valuedness, we use the fact that $\mathbb{S}$ is a two-sided surface and we distinguish external and internal points $M(r,\theta,\varphi)$ of $\mathbb{S}$ considering them as if for $r=1+0$ and $r=1-0$.

In what follows, we assume that the continuous passage from one side of $\mathbb{S}$ to the other is allowed only through the poles N and S of the sphere $\mathbb{S}$. In so doing, any full φ-meridian is not placed on one side of $\mathbb{S}$ but is placed in two parts on different sides of $\mathbb{S}$. We place the part M_φ of any full φ-meridian C_φ on the external side $\mathbb{S}_{1+0}$ of $\mathbb{S}$ and the part $M_{\varphi\pm\pi}$ with $\theta\uparrow_\pi^{2\pi}$ on the internal part $\mathbb{S}_{1-0}$ of $\mathbb{S}$. As a result, the function $F(\theta,\varphi)$ in (1.1) becomes a single-valued and well-defined function on $S_{1+0}\cup S_{1-0}$ coinciding with $f(\theta,\varphi)$ on $\mathbb{S}_{1+0}$. On the internal side $\mathbb{S}_{1-0}$, the function $F(\theta,\varphi)$ is defined by the part of formula (1.1) which relates to the $(\varphi\pm\pi)$-meridian. Formula (1.2) is given only to explain the reason of the two-valuedness of the function $F(\theta,\varphi)$ on $\mathbb{S}$.

Now, according the usual classical Meyer scheme, one can easily construct a multiresolution analysis on the double sphere $\widetilde{\mathbb{S}}_{(2)}=S_{1+0}\cup S_{1-0}$ with angular coordinates of points on $\widetilde{\mathbb{S}}_{(2)}$ still denoted by (θ,φ). In this case, θ changes from 0 to 2π on any full φ-meridian and values of φ can still be bounded by the interval $[0,2\pi)$. The coordinates of points $M(\theta,\varphi)$ on $\mathbb{S}_{1+0}$ are usual spherical coordinates. They are extended on $\mathbb{S}_{1-0}$ as follows: the φ-coordinate of the point $M\in\mathbb{S}_{1-0}$ coincides with its value in the original spherical coordinate system, and the value of its usual spherical latitude, say τ, is replaced by $\theta=2\pi-\tau$. It is easy to see that the point M with such coordinates (θ,φ) belongs to the part of the full $(\varphi\pm\pi)$-meridian lying on $\mathbb{S}_{1-0}$ (the sign, plus or minus, in the expression $\varphi\pm\pi$ can always be taken so that $\varphi\pm\pi\in[0,2\pi)$).

As basic functions of the subspaces $V_j(\mathbb{T})$ of multiresolution analysis on $\widetilde{\mathbb{S}}_{(2)}$ (defining $V_j(\mathbb{T})$ themselves), we take systems of 2π-periodic functions constructed on the basis of Meyer wavelets. These are the trigonometric polynomials $\Phi_s^{j,k}(x)$ $(s=1,2,3)$ generating the finite-dimensional subspaces $V_j(\mathbb{T})$. We use them because of their simplicity. Furthermore, in order not to calculate integral coefficients of function expansions in orthogonal systems, we restrict ourselves to the use only of the interpolation properties of multiresolution analysis on finite grides in θ and φ. Since the convergence of interpolation expansions for continuous (and especially smooth) functions on $\mathbb{S}$ occurs with high rate, there is no need to apply the subspaces $V_j(\widetilde{\mathbb{S}}_{(2)})$ with large indices j for practical problems. Thus, one may not be afraid of a significant concentration of grid points near the poles (especially in the case of computer implementation of algorithms of approximation of functions f on $\mathbb{S}$). The orthogonal properties of bases can be useful when approximating functions integrable only on $\mathbb{S}$.

Thus, in what follows, we use (see [3, 10]) the scaling functions of periodic multiresolution analyzes:

$$\Phi_s^{j,k}(x) = 2^{-j} \sum_{|\nu/2^j|<(1+\varepsilon)/2} \widehat{\varphi}_s\Big(\frac{\nu}{2^j}\Big) e^{i\nu(x-2\pi k/2^j)}, \quad k = \overline{0, 2^j-1}, \quad j \in \mathbb{Z}_+, \quad s = 1,2,3, \tag{1.3}$$

where

$$\widehat{\varphi}_s(\omega) = \widehat{\varphi}_\varepsilon(\omega)^2 + (1-\delta_{3,s}) i(\operatorname{sign}\omega)\widehat{\varphi}_\varepsilon(\omega)\big(\widehat{\varphi}_\varepsilon(\omega-1) + \widehat{\varphi}_\varepsilon(\omega+1)\big), \quad s = 2,3. \tag{1.4}$$

In turn, $\widehat{\varphi}_\varepsilon(\omega)$, $\varepsilon > 0$, is an even continuous real function on $\mathbb{R}$ of Meyer type supported on the interval $|\omega| < (1+\varepsilon)/2$ and such that $\widehat{\varphi}_\varepsilon(\omega) = 1$ for $|\omega| \le (1-\varepsilon)/2$ $(0 < \varepsilon \le 1/3)$, the derivative $\widehat{\varphi}'_\varepsilon(\omega)$ is a function of bounded variation, and $\widehat{\varphi}^2_\varepsilon(\omega) + \widehat{\varphi}^2_\varepsilon(\omega-1) = 1$ for $(1-\varepsilon)/2 < \omega \le (1+\varepsilon)/2$. When $s = 1$, we replace $\widehat{\varphi}_\varepsilon(\omega)$ in (1.4) by

$$\widehat{\varphi}_{1,\varepsilon}(\omega) = \frac{1}{\sqrt{2}} \sqrt{1 + \widehat{\varphi}_\varepsilon(\omega) - \widehat{\varphi}_\varepsilon(\omega-1) - \widehat{\varphi}_\varepsilon(\omega+1)}. \tag{1.5}$$

For each $s = 1,2,3$, the functions $\Phi_s^{j,k}(x)$ form the interpolation basis of the subspaces $V_s^j(\mathbb{T})$ $(j \in \mathbb{Z}_+)$ of 2π-periodic multiresolution analysis:

$$\Phi_s^{j,k}\Big(\frac{2\pi l}{2^j}\Big) = \delta_{k,l} \quad (k,l = \overline{0, 2^j-1}).$$

In addition, for $s = 1,2$ and for any $j \in \mathbb{Z}_+$, the system $\{2^{j/2}\Phi_s^{j,k}(x)\}$ is orthonormal on $\mathbb{T}$:

$$\frac{1}{2\pi} \int_0^{2\pi} 2^j \Phi_s^{j,k}(x) \overline{\Phi_s^{j,l}(x)}\, dx = \delta_{k,l} \quad (k,l = \overline{0, 2^j-1}). \tag{1.6}$$

For any j and for $k,l = 0,1,\dots,2^j-1$, we define

$$\Phi_s^{j,k,l}(\theta,\varphi) = \Phi_s^{j,k}(\theta)\Phi_s^{j,l}(\varphi) \quad \text{for} \quad (\theta,\varphi) \in \mathbb{T}\times\mathbb{T}. \tag{1.7}$$

Naturally, without any additional assumptions except for the 2π-periodicity, this is an interpolation system of functions on the grid $\{(2\pi m/2^j,\, 2\pi n/2^j) :\ m,n = \overline{0, 2^j-1}\}$:

$$\Phi_s^{j,k,l}\Big(\frac{2\pi m}{2^j}, \frac{2\pi n}{2^j}\Big) = \delta_{k,m}\cdot\delta_{l,n}.$$

This system inherits in $C(\mathbb{T}\times\mathbb{T})$ all approximative properties of system (1.3) in $C[0,2\pi]$.

2. Approximation by interpolating wavelets in $C(\mathbb{T}\times\mathbb{T})$

We denote by ${}_sV_j(\mathbb{T}^2)$ the subspace of the space $C(\mathbb{T}\times\mathbb{T})$ of 2π-periodic (in θ and φ) functions on $\mathbb{R}^2$ by setting

$${}_sV_j(\mathbb{T}^2) := \Big\{ \sum_{k=0}^{2^j-1}\sum_{l=0}^{2^j-1} C_{k,l}\Phi_s^{j,k}(\theta)\Phi_s^{j,l}(\varphi) :\ C_{k,l} \in \mathbb{R} \ \text{ for all } \ k,l = \overline{0, 2^j-1} \Big\}.$$

The interpolation projection of any function $F(\theta,\varphi) \in C(\mathbb{T}\times\mathbb{T})$ is defined as follows:

$$S_{s,2^j}F(\theta,\varphi) = P^{\mathrm{int}}_{{}_sV_j(\mathbb{T}^2)}F(\theta,\varphi) := \sum_{k=0}^{2^j-1}\sum_{l=0}^{2^j-1} F\Big(\frac{2\pi k}{2^j}, \frac{2\pi l}{2^j}\Big)\Phi_s^{j,k,l}(\theta,\varphi). \tag{2.1}$$

Obviously, $|S_{2^j}F(\theta,\varphi)| \le L_{2^j}(\theta,\varphi)\|F\|_{C(\mathbb{T}\times\mathbb{T})}$, where $L_{2^j}(\theta,\varphi)$ is the Lebesgue function of the operator $S_{2^j}: C(\mathbb{T}\times\mathbb{T}) \to {}_sV_j \subset C(\mathbb{T}\times\mathbb{T})$,

$$L_{s,2^j}(\theta,\varphi) = \sum_{k=0}^{2^j-1}\sum_{l=0}^{2^j-1} |\Phi_s^{j,k}(\theta)||\Phi_s^{j,l}(\varphi)| = L_{s,2^j}(\theta)L_{s,2^j}(\varphi), \tag{2.2}$$

and $L_{s,2^j}(x)$ is the Lebesgue function of the projection operator of continuous 2π-periodic functions on the line on the subspace $V_s^j(\mathbb{T}) \subset C(\mathbb{T})$, which was studied in Lemmas 2 and 3 of the paper [10] under the condition of smoothness of the functions $\widehat{\varphi}_s(\omega)$ on $\mathbb{R}$ for $s = 2, 3$. Using this lemmas and the remark to them on page 265 of the mentioned paper, for the $2\pi/2^j$-periodic Lebesgue function $L_{s,2^j}(x)$ with $s = 2, 3$, we obtain

$$L_{s,2^j}(x) \le \left(\bigvee_{1/2}^{(1+\varepsilon)/2} (\widehat{\varphi}_\varepsilon^2(\omega))'_\omega \Big|\frac{\sin 2^{j-1}x}{2^{j-1}x}\Big| + \delta_{s,2} \bigvee_{1/2}^{(1+\varepsilon)/2} (\widehat{\varphi}_\varepsilon(\omega)\widehat{\varphi}_\varepsilon(\omega-1))'_\omega \frac{\sin^2 \varepsilon 2^{j-2}x}{|2^{j-1}x|} \right) \frac{|\sin 2^{j-1}x|}{|2^{j-1}x|} +$$
$$+ \left[\bigvee_{1/2}^{(1+\varepsilon)/2} (\widehat{\varphi}_\varepsilon^2(\omega))'_\omega + \delta_{s,2} \bigvee_{1/2}^{(1+\varepsilon)/2} (\widehat{\varphi}_\varepsilon(\omega)\widehat{\varphi}_\varepsilon(\omega-1))'_\omega \Big(\frac{4}{\pi^2} + \varepsilon + \frac{1-4/\pi^2}{2^{2j}} \Big) |\sin 2^{j-1}x| \right.$$

for $|x| < 2\pi/2^{j+1}$. We do not write an analogous estimate for $s = 1$, noting only that this estimate is similar to the latter one with replacing $\delta_{s,2}$ by $\delta_{s,1}$ and $\widehat{\varphi}_\varepsilon(\omega)$ by $\widehat{\varphi}_{1,\varepsilon}(\omega)$ from (1.5).

For brevity, we use the following formulas from [10]:

$$\Delta^\varepsilon = \Big[\frac{1}{2}, \frac{1+\varepsilon}{2}\Big], \quad \widehat{\varphi}_3(\omega) = \widehat{\varphi}_\varepsilon^2(\omega), \quad \beta(\omega) = \widehat{\varphi}_\varepsilon(\omega)\widehat{\varphi}_\varepsilon(\omega-1).$$

Theorem 1. *Assume that, in addition to the conditions*[2] *on* $\widehat{\varphi}_s(\omega)$ *imposed in the description of formula* (1.4), *the functions* $\widehat{\varphi}_3(\omega)$ *and* $\beta(\omega)$ *are smooth in a neighbourhood of the interval* $[(1-\varepsilon)/2, (1+\varepsilon)/2]$. *Then the Lebesgue constants* $L_{s,2^j}(\theta,\varphi)$, $s = 2, 3$, *in* (2.2) *satisfy on their period* $[-2\pi/2^{j+1}, 2\pi/2^{j+1}] \times [-2\pi/2^{j+1}, 2\pi/2^{j+1}]$ *the estimates*

$$\begin{aligned} L_{s,2^j}(\theta,\varphi) \le \Big\{ \Big[\bigvee_{\Delta^\varepsilon} \widehat{\varphi}'_3(\omega) \Big|\frac{\sin 2^{j-1}\varphi}{2^{j-1}\varphi}\Big| + \delta_{s,2} \bigvee_{\Delta^\varepsilon} \beta'(\omega) \frac{\sin^2(\varepsilon 2^{j-1}\varphi/2)}{|2^{j-1}\varphi|} \Big] \frac{|\sin 2^{j-1}\varphi|}{|2^{j-1}\varphi|} + \\ + \Big[\bigvee_{\Delta^\varepsilon} \widehat{\varphi}'_3(\omega) + \delta_{s,2} \bigvee_{\Delta^\varepsilon} \beta'_3(\omega) \Big] \Big(\frac{4}{\pi^2} + \frac{1-4/\pi^2}{2^{2j}} \Big) |\sin 2^{j-1}\varphi| \Big\} \times \\ \times \Big\{ \Big[\bigvee_{\Delta^\varepsilon} \widehat{\varphi}'_3(\omega) \frac{|\sin 2^{j-1}\theta|}{2^{j-1}\theta} + \delta_{s,2} \bigvee_{\Delta^\varepsilon} \beta'(\omega) \frac{\sin^2(\varepsilon 2^{j-1}\theta/2)}{|2^{j-1}\theta|} \Big] \frac{|\sin 2^{j-1}\theta|}{|2^{j-1}\theta|} + \\ + \Big[\bigvee_{\Delta^\varepsilon} \widehat{\varphi}'_3(\omega) + \delta_{s,2} \bigvee_{\Delta^\varepsilon} \beta'(\omega) \Big] \Big(\frac{4}{\pi^2} + \frac{1-4/\pi^2}{2^{2j}} \Big) |\sin 2^{j-1}\theta| \Big\}. \end{aligned} \tag{2.3}$$

P r o o f follows from the above estimate and (2.2). □

We note that, to estimate the function (2.2) on the square $\mathbb{T}\times\mathbb{T}$, it is needed to write its estimate on every small square $[2\pi(2k-1)/2^{j+1}, 2\pi(2k+1)/2^{j+1}] \times [2\pi(2l-1)/2^{j+1}, 2\pi(2l+1)/2^{j+1}]$ contained in $\mathbb{T}\times\mathbb{T}$ by replacing on the right-hand-side of (2.3) φ by $(\varphi - 2\pi k/2^j)$ and θ by $(\theta - 2\pi l/2^j)$.

[2] Actually, this is a condition to estimate $L_{s,2^j}(x)$ in [10] allowing to drop terms outside the integrals when integrating by parts.

To estimate the Lebesgue constant, which is the norm in $C(\mathbb{T}\times\mathbb{T})$ of the function $L_{s,2^j}(\theta,\varphi)$ coinciding with the norm of the operator $\|S_{2^j}\| = \|S_{2^j}\|_{C(\mathbb{T}\times\mathbb{T})}^{C(\mathbb{T}\times\mathbb{T})}$, it is sufficient to estimate it on any period, in particular, for $|\theta| < \pi/2^j$ and $|\varphi| < \pi/2^j$. Estimating the right-hand side of (2.3) with the use of the fact that $|\sin x|/|x| \le 1$ for $|x| < \pi/2$, we obtain the following result.

Corollary 1. *Assume that the conditions of Theorem 1 are satisfied. Then the norm of the operators of the interpolation projection (2.1) from $C(\mathbb{T}\times\mathbb{T})$ to the subspace ${}_sV_j(\mathbb{T}^2) \subset C(\mathbb{T}\times\mathbb{T})$ satisfies the estimate*

$$\|S_{s,2^j}\| \le \Big(\bigvee_{\Delta^\varepsilon} \widehat{\varphi}_3'(\omega) + \delta_{s,2}\bigvee_{\Delta^\varepsilon}\beta'(\omega)\Big)^2\Big(\frac{4}{\pi^2}+\varepsilon+\frac{1-4/\pi^2}{2^{2j}}\Big)^2. \tag{2.4}$$

For one-dimensional periodic wavelets, the following well-known and easily verified remarkable fact holds: for any $\varepsilon \in (0,1/3]$, the operator of interpolation (and also orthogonal) projection on the subspaces V_j of periodic multiresolution analysis generated by any Meyer type function $\widehat{\varphi}_\varepsilon(\omega)$ is the identity operator on the subspace of trigonometric polynomials of order $N_\varepsilon = [2^{j-1}(1-\varepsilon)]$, where $[a]$ is the integer part of a for $a \ge 0$.

We verify in what form this property is preserved for the operators (2.1). Computing $S_{s,2^j}g(\theta,\varphi)$ for $g(\theta,\varphi) = e^{i\mu\theta}e^{i\eta\varphi}$ and integer μ and η, we have

$$S_{s,2^j}g(\theta,\varphi) = \sum_{k=0}^{2^j-1}\sum_{l=0}^{2^j-1} e^{2\pi i\mu k/2^j}e^{2\pi i\eta l/2^j}\Phi_s^{j,k,l}(\theta,\varphi) =$$

$$= \sum_\nu 2^{-j}\widehat{\varphi}_s\Big(\frac{\nu}{2^j}\Big)e^{i\nu\theta}\sum_{k=0}^{2^j-1}e^{2\pi i(\mu-\nu)k/2^j}\sum_{\nu'}2^{-j}\widehat{\varphi}_s\Big(\frac{\nu'}{2^j}\Big)e^{i\nu'\varphi}\sum_{l=0}^{2^j-1}e^{2\pi i(\eta-\nu')l/2^j} =$$

$$= \sum_\nu 2^{-j}\widehat{\varphi}_s\Big(\frac{\nu}{2^j}\Big)e^{i\nu\theta}\frac{e^{2\pi i(\mu-\nu)}-1}{e^{2\pi i(\mu-\nu)/2^j}-1}\sum_{\nu'}2^{-j}\widehat{\varphi}_s\Big(\frac{\nu'}{2^j}\Big)e^{i\nu'\varphi}\frac{e^{2\pi i(\eta-\nu')}-1}{e^{2\pi i(\eta-\nu')/2^j}-1} =$$

$$= \sum_\nu 2^{-j}\widehat{\varphi}_s\Big(\frac{\nu}{2^j}\Big)e^{i\nu\theta}2^j\delta_{\mu,\nu}\sum_{\nu'}^{(\mu-\nu')/2^j}2^{-j}\widehat{\varphi}_s\Big(\frac{\nu'}{2^j}\Big)e^{i\nu'\varphi}2^j\delta_{\nu',\eta} = e^{i\mu\theta}e^{i\eta\varphi}\widehat{\varphi}_s\Big(\frac{\mu}{2^j}\Big)\widehat{\varphi}_s\Big(\frac{\eta}{2^j}\Big),$$

which coincides with $g(\theta,\varphi)$ for $|\mu|/2^j \le (1-\varepsilon)/2$ and $|\eta|/2^j \le (1-\varepsilon)/2$ (where $\widehat{\varphi}_s(\omega)\equiv 1$).

Thus, we obtain the following property of interpolation projections on the subspaces ${}_sV_j(\mathbb{T}^2)$.

Assertion 1. *For the trigonometric polynomials of two variables*

$$t_{n,m}(\theta,\varphi) = \sum_{\mu=-n}^{n}\sum_{\eta=-m}^{m} a_{\mu,\nu}e^{i(\mu\theta+\nu\varphi)}$$

of order n in the variable θ and order m in the variable φ, the equalities

$$S_{s,2^j}t_{n,m}(\theta,\varphi) \equiv t_{n,m}(\theta,\varphi) \tag{2.5}$$

hold for n and m not greater than $N_{\varepsilon,j} = [2^{j-1}(1-\varepsilon)]$ and $s = 1,2,3$.

Note that the order $N_{\varepsilon,j}$ of the polynomials in (2.5) is allowed in each of the variables θ and φ, not just in the totality of variables (when the summation in the formula for $t_{n,m}(\theta,\varphi)$ is taken over μ and ν such that $|\mu|+|\nu| \le N_{\varepsilon,j}$).

According to the usual Lebesgue scheme, from inequality (2.4) and Assertion 1, we easily obtain an estimate of the error of approximation of continuous 2π-periodic functions of two variables by their interpolation projections on ${}_sV_j(\mathbb{T}^2)$. In view of the importance of this estimate for practical applications of interpolating wavelets, we state it as a theorem. We denote by $E_n(F)_{C(\mathbb{T}\times\mathbb{T})}$ the best approximation in the metric of $C(\mathbb{T}\times\mathbb{T})$ of continuous 2π-periodic functions F on the square $\mathbb{T}\times\mathbb{T}$ by trigonometric polynomials of order n in each variable.

Theorem 2. *Under the conditions of Theorem 1 on $\widehat{\varphi_s}(\omega)$, $s=2,3$, any function $F(\theta,\varphi)$ in $C(\mathbb{T}\times\mathbb{T})$ satisfies the estimates*

$$\|F(\theta,\varphi)-S_{s,2^j}F(\theta,\varphi)\|_{C(\mathbb{T}\times\mathbb{T})}\le(1+\|S_{s,2^j}\|)E_{N_{\varepsilon,j}}(F)_{C(\mathbb{T}\times\mathbb{T})}. \tag{2.6}$$

P r o o f. To justify this estimate, we note that, applying formula (2.5) to the polynomial $t_{N_{\varepsilon,j}}$ of the best approximation of the function F in $C(\mathbb{T}\times\mathbb{T})$, we obtain

$$\|F(\theta,\varphi)-S_{s,2^j}F(\theta,\varphi)\|=\|(F(\theta,\varphi)-t_{N_{\varepsilon,j}}(\theta,\varphi))+S_{s,2^j}\big(t_{N_{\varepsilon,j}}(\theta,\varphi)-F(\theta,\varphi)\big)\|,$$

From this, using the triangle inequality for norms, the definition of $\|S_{s,2^j}\|$, and Corollary 1, we get (2.6). □

An estimate of the best approximations $E_n(F)_{C(\mathbb{T}\times\mathbb{T})}$ of the Jackson type in terms of the modules of continuity or the parameters K and α of the Hölder class

$$KH^\alpha=\{f:|f(x+\Delta x)-f(x)|\le K|\Delta x|^\alpha\}$$

can be found in the known monographs on approximation theory.

The systems of functions

$$\{\Phi_s^{j+1,2k+1,2l+1}(\theta,\varphi):k=\overline{0,2^j-1}\},\ j\in\mathbb{Z}_+\ (s=1,2,3), \tag{2.7}$$

additional to (1.7) are the interpolation bases of the subspaces ${}_sW_j(\mathbb{T}^2)$ (${}_sV_{j+1}(\mathbb{T}^2)={}_s\ V_j(\mathbb{T}^2)\oplus_s W_j(\mathbb{T}^2), j\in\mathbb{Z}_+$). By their means, any function $g\in_s V_{j+1}(\mathbb{T}^2)$ is uniquely represented in the form

$$g=P^{int}_{{}_sV_j}g+P^{int}_{{}_sW_j}(g-P^{int}_{{}_sV_j}g), \tag{2.8}$$

which is easily derived from the fact that ${}_sW_j\subset{}_sV_{j+1}$. For each $s=1,2,3$, the family of systems (2.7) over all $j\in\mathbb{Z}_+$ together with $\Phi_{0,0}\equiv1$ is an interpolation basis of the whole space $C(\mathbb{T}\times\mathbb{T})$, so that any function $F(\theta,\varphi)$ is expanded in the series

$$F(\theta,\varphi)=F(0,0)+\sum_{j=0}^{\infty}\sum_{k,l=0}^{2^j-1}c_{j,k,l}\Phi_s^{j+1,2k+1,2l+1}(\theta,\varphi) \tag{2.9}$$

converging uniformly in the square $\mathbb{T}\times\mathbb{T}$ and, hence, in $\mathbb{R}^2$. According to the usual scheme, the coefficients of this series are calculated recursively in j in terms of the grid values of the function F and the partial sums

$$\Sigma_{j-1}(\theta,\varphi;F)=F(0,0)+\sum_{\lambda=0}^{j-1}\sum_{\mu,\nu=0}^{2^\lambda-1}c_{\lambda,\mu,\nu}\Phi_s^{\lambda+1,2\mu+1,2\nu+1}(\theta,\varphi) \tag{2.10}$$

of the same series, namely

$$c_{j,k,l}=F\Big(\frac{2\pi(2k+1)}{2^{j+1}},\frac{2\pi(2l+1)}{2^{j+1}}\Big)-\Sigma_{j-1}\Big(\frac{2\pi(2k+1)}{2^{j+1}},\frac{2\pi(2l+1)}{2^{j+1}}\Big).$$

It follows from (2.8) that the sum $\Sigma_{j-1}(\theta,\varphi;F)$ coincides with $P^{int}_{sV_j}F(\theta,\varphi)=S_{s,2^j}F(\theta,\varphi)$ (see (2.1)), so that we can write the values $c_{j,k,l}$ without recurrence:

$$c_{j,k,l}=(F-P^{\mathrm{int}}_{sV_j}F)\Big(\frac{2\pi(2k+1)}{2^{j+1}},\frac{2\pi(2l+1)}{2^{j+1}}\Big). \tag{2.11}$$

Hence, it is easily deduced that the series (2.9) with the coefficients (2.11) coincides with the series

$$F(\theta,\varphi)=F(0,0)+\sum_{j=0}^{\infty}(S_{s,2^{j+1}}F(\theta,\varphi)-S_{s,2^j}F(\theta,\varphi)) \tag{2.12}$$

converging uniformly in $\mathbb{T}\times\mathbb{T}$ by Theorem 2. The partial sum of order J of the latter series is

$$F(0,0)+\sum_{j=0}^{J-1}(S_{s,2^{j+1}}-S_{s,2^j})F(\theta,\varphi)=S_{s,2^J}F(\theta,\varphi)=\Sigma_{J-1}(\theta,\varphi;F).$$

3. Interpolating wavelets on the sphere and their application to the approximation of functions in $C(\mathbb{S})$

In the second section, unlike the first section, the arguments (θ,φ) of the functions F and $\Phi_s^{j,k,l}$ were treated as the Cartesian coordinates of points of the square $\mathbb{T}\times\mathbb{T}$ on the opposite sides of which the values of any function in $C(\mathbb{T}\times\mathbb{T})$ coincide in view of its 2π-periodicity. Moreover, the function $F(\theta,\varphi)$ constructed on $\widetilde{\mathbb{S}}_{(2)}$ by formula (1.1), if interpreted as a function on the square $\mathbb{T}\times\mathbb{T}$, has the additional feature that it is constant on each of the sides $\theta=0$ and $\theta=\pi$ of the square.

Let F be a function defined on the sphere $\mathbb{S}$ and continuously depending on the points of the sphere. For instance, F represented as $F(x_1,x_2,x_3)$ is a function continuous in all coordinates connected by the relation $x_1^2+x_2^2+x_3^2=1$. In particular, F is also continuous at the poles N and S of the sphere $\mathbb{S}$. Therefore, after the change $x_1=\cos\varphi\sin\theta$, $x_2=\cos\varphi\cos\theta$, $x_3=\sin\theta$, the function F becomes a function of the coordinates $\varphi\in\mathbb{T}=[0,2\pi)$ and $\theta\in[0,\pi]$ with the following specificity: $F(N)$ and $F(S)$ do not depend on θ, since $\lim_{\theta\to0}F(\theta,\varphi)=F(N)$ and $\lim_{\theta\to\pi}F(\theta,\varphi)=F(S)$ for any $\varphi\in[0,2\pi]$. Thus, the function $F(\theta,\varphi)$ defined on the double sphere $\widetilde{\mathbb{S}}_{(2)}$ by formula (1.1) and glued from the continuous functions $f(\theta,\varphi)$ on M_φ for $\theta\uparrow_0^\pi$ and $f(2\pi-\theta,\varphi\pm\pi)$ on $M_{\varphi\pm\pi}$ for $\theta\uparrow_\pi^{2\pi}$ is continuous on $\widetilde{\mathbb{S}}_{(2)}$, since the values of the function $F(0,\varphi)$ and the values of the function $F(\pi,\varphi)$ do not depend on φ at the gluing points $\theta=2\pi$ and $\theta=\pi$.

We note that the values of the function $F(\theta,\varphi)$ on $\mathbb{S}$ (i.e., for $\theta\uparrow_0^\pi$, $\varphi\uparrow_0^{2\pi}$) coincide with the values of the original function $f(\theta,\varphi)$. Therefore, approximating F on $\widetilde{\mathbb{S}}_{(2)}$, we simultaneously approximate f on $\mathbb{S}$. Of course, the latter property could be preserved for any continuous extension of f from $\mathbb{S}$ to $\widetilde{\mathbb{S}}_{(2)}\setminus\mathbb{S}$. However, if the original function f is smooth on $\mathbb{S}$, i.e., at any point $(x_1,x_2,x_3,f(x_1,x_2,x_3))$ (with $x_1^2+x_2^2+x_3^2=1$) of the graph surface of f over $\mathbb{S}$, there exists a tangent plane to the graph, then, obviously, the extension chosen by means of (1.4) preserves the smoothness of $F(\theta,\varphi)$ on any full φ-meridian C_φ and, hence, on the whole double sphere $\widetilde{\mathbb{S}}_{(2)}$, since there exists a tangent line to the graph of $F(\theta,\varphi)$ over any full φ-meridian at the points (θ,φ) $(\theta\uparrow_0^\pi)$ which is the section of the tangent plane at the point $X(\theta,\varphi)\in C_\varphi$ by the plane containing C_φ.

The basis functions $\Phi_s^{j,k,l}(\theta,\varphi)$ are defined on the whole $\widetilde{\mathbb{S}}_{(2)}$ as 2π-periodic in θ and φ, since the parameter θ changes from 0 to 2π on any full φ-meridian and the parameter φ defining C_φ changes similarly. Of course, not each of these functions is constant for $\theta=0$ $(\theta=2\pi)$ or $\theta=\pi$ like $F(\theta,\varphi)$ (these are the functions $\Phi_s^{j,0,l}(\theta,\varphi)$ and $\Phi_s^{j,2^{j-1},l}(\theta,\varphi)$). However, only the continuity of F is important to apply formula (2.1), estimates (2.3), (2.4), and (2.6), and formulas (2.9)–(2.12) to

the functions $F(\theta,\varphi)$ defined by (1.1). Thus, the functions $\Phi_s^{j,k,l}(\theta,\varphi)$ defined by (1.7) determine multiresolution analysis on $\widetilde{\mathbb{S}}_{(2)}$, i.e. the subspaces ${}_sV_j(\widetilde{\mathbb{S}}_{(2)})$ and ${}_sW_j(\widetilde{\mathbb{S}}_{(2)})$. In so doing, a pair (θ,φ) should be treated everywhere as parameters defining the points $M(\theta,\varphi)$ on $\widetilde{\mathbb{S}}_{(2)}$. The only useful thing remaining is to rewrite formulas (2.1) in terms of the function $f(\theta,\varphi)$ on $\mathbb{S}$ defining $F(\theta,\varphi)$ on $\widetilde{\mathbb{S}}_{(2)}$. Using (2.1) and (1.1), we set

$$P_{{}_sV_j}^{int}F(\theta,\varphi)=\sum_{k=0}^{2^{j-1}-1}\sum_{l=0}^{2^j-1}f\Big(\frac{2\pi k}{2^j},\frac{2\pi l}{2^j}\Big)\Phi_s^{j,k,l}(\theta,\varphi)+$$

$$+\sum_{k=2^{j-1}}^{2^j-1}\Bigg(\sum_{l=0}^{2^{j-1}-1}f\Big(\frac{2\pi(2^j-k)}{2^j},\frac{2\pi(l+2^{j-1})}{2^j}\Big)\Phi_s^{j,k,l}(\theta,\varphi)+\sum_{l=2^{j-1}}^{2^j-1}f\Big(\frac{2\pi(2^j-k)}{2^j},\frac{2\pi l}{2^j}\Big)\Phi_s^{j,k,l}(\theta,\varphi)\Bigg).$$

By Theorem 2, one can estimate the error of approximation of the function $F(\theta,\varphi)$ by means of $P_{{}_sV_j}^{int}F(\theta,\varphi)$ in terms of the best approximation $E_{N_{\varepsilon,l}}(F)_{C(\widetilde{\mathbb{S}}_{(2)})}$. In real applied problems, it is unlikely to be required to approximate functions defined on both inner and outer sides of the sphere $\mathbb{S}$. Therefore, to approximate the original function $f(\theta,\varphi)$, it is sufficient to estimate the deviation $|f(\theta,\varphi)-P_{{}_sV_j}^{int}F(\theta,\varphi)|_{C([0,\pi]\times\mathbb{T})}$ that does not exceed the approximation error (2.6).

There are studies of the problem of approximation by trigonometric polynomials on an interval less than the period. Here, one can expect an essential improvement of the estimate (2.6) by learning to use the specificity of the function f on $\mathbb{S}$, in particular, its singularities on $\mathbb{S}$ which make it hardly changing in a neighborhood of the sides of the rectangle $[0,\pi]\times\mathbb{T}$ with $\theta=0$ and $\theta=\pi$ on which the function $f(\theta,\varphi)$ is naturally transferred from $\mathbb{S}$.

Until now, we have discussed the use of interpolation properties of the wavelets $\Phi_s^{j,k,l}(\theta,\varphi)$. As noted, in view of (1.6) the systems $\{2^{j/2}\Phi_s^{j,k}(x):k=\overline{0,2^j-1}\}$ are orthonormal for $s=1$ or $s=2$ and for every $j\in\mathbb{N}$. This implies that, for any $j\in\mathbb{N}$, the systems $\{2^{j/2}\Phi_2^{j,k,l}(\theta,\varphi):k,l=\overline{0,2^j-1}\}$ are also orthonormal:

$$\Big(\frac{1}{2\pi}\Big)^2\int_0^{2\pi}\int_0^{2\pi}2^{j/2}\Phi_2^{j,k,l}(\theta,\varphi)2^{j/2}\overline{\Phi_2^{j,m,n}(\theta,\varphi)}d\theta d\varphi=\frac{1}{2\pi}\int_0^{2\pi}2^{j/2}\Phi_2^{j,k}(\theta)\overline{\Phi_2^{j,m}(\theta)}d\theta\times$$

$$\times\frac{1}{2\pi}\int_0^{2\pi}2^{j/2}\Phi_2^{j,l}(\varphi)\overline{\Phi_2^{j,n}(\varphi)}d\varphi=\delta_{k,m}\cdot\delta_{l,n}=\begin{cases}1, & (k,l)=(m,n),\\ 0, & (k,l)\neq(m,n).\end{cases}$$

However, the question on the application of the orthonormality properties of these systems to the approximation of functions on the sphere in $L^2(\mathbb{S})$-norm requires separate consideration.

4. Conclusion

In this paper, we have considered the question of approximation of continuous functions on the sphere $\mathbb{S}\subset\mathbb{R}^3$ and have proposed once more approach to the construction of corresponding interpolating wavelets. Due to a special double expansion of the sphere, this approach reduces to the simple and well-studied problem on the construction of interpolating periodic wavelets on the plane $\mathbb{R}^2$. Two of the constructed wavelet systems are orthogonal on the expanded sphere $\mathbb{S}$. This property can be useful when the approximated function is inaccurately defined (for instance, is obtained experimentally). The problem of accuracy of approximation of functions on the sphere in L^2 was not studied in this paper.

REFERENCES

1. Arfaoui S., Rezgui I., Mabrouk A. B. *Wavelet Analysis on the Sphere: Spheroidal Wavelets.* Berlin: Walter de Gruyter GmbH & Co KG, 2017. 144 p.

2. Askari-Hemmat A., Dehghan M. A., Skopina M. Polynomial Wavelet-Type Expansions on the Sphere. *Math. Notes*, 2003. Vol. 74, No. 2. P. 278–285. DOI: 10.1023/A:1025016510773
3. Chernykh N. I., Subbotin Yu. N. Interpolating-orthogonal wavelet systems. *Proc. Steklov Inst. Math.*, 2009. Vol. 264, Suppl. 1. P. 107–115. DOI: 10.1134/S0081543809050083
4. Dahlke S., Dahmen W., Weinreich I., Schmitt E. Multiresolution analysis and wavelets on $\mathbb{S}^2$ and $\mathbb{S}^3$. *Numer. Funct. Anal. Optim.*, 1995. Vol. 16, No. 1–2. P. 19–41. DOI: 10.1080/01630569508816605
5. Dai F. Characterizations of function spaces on the sphere using frames. *Trans. Amer. Math. Soc.*, 2007. Vol. 359, No. 2. P. 567–589. DOI: 10.1090/S0002-9947-06-04030-X
6. Farkov Yu. B-spline wavelets on the sphere. In: *Proc. of the Intern. Workshop "Self-Similar Systems"*, 1999. Vol. 30. P. 79–82.
7. Freeden W., Schreiner M. Orthogonal and nonorthogonal multiresolution analysis, scale discrete and exact fully discrete wavelet transform on the sphere. *Constr. Approx.*, 1998. Vol. 14, No. 4. P. 493–515. DOI: 10.1007/s003659900087
8. Schröder P., Sweldens W. Spherical wavelets: Efficiently representing functions on the sphere. In: *Wavelets in the Geosciences. Lect. Notes in Earth Sci.*, vol. 90. 1995. P. 158–188. DOI: 10.1007/BFb0011096
9. Skopina M. *Polynomial Expansions of Continuous Functions on the Sphere and on the Disk.* IMI Research Reports, Department of Mathematics, University of South Carolina, 2001. Preprint, Vol. 5. 13 p. http://imi.cas.sc.edu/django/site_media/media/papers/2001/2001_05.pdf
10. Subbotin Yu. N., Chernykh N. I. Interpolation wavelets in boundary value problems. *Proc. Steklov Inst. Math.*, 2018. Vol. 300, Suppl. 1. P. 172–183. DOI: 10.1134/S0081543818020177

URAL MATHEMATICAL JOURNAL, Vol. 5, No. 2, 2019, pp. 13–20

DOI: 10.15826/umj.2019.2.002

IMPULSE CONTROL OF THE MANIPULATION ROBOT

Yurii F. Dolgii[1,2,†], **Alexander N. Sesekin**[1,2,††], **Ilya A. Chupin**[1,†††]

[1]Ural Federal University,
19 Mira str., Ekaterinburg, 620002, Russia

[2]Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences,
16 S. Kovalevskaya Str., Ekaterinburg, 620990, Russia

[†]juri.dolgy@urfu.ru, [††]a.n.sesekin@urfu.ru, [†††]mr.tchupin@yandex.ru

Abstract: A nonlinear control problem for a manipulation robot is considered. The solvability conditions for the problem are obtained in the class of special impulse controls. To achieve the control goal, the kinetic energy of the manipulation robot is used. When finding analytical formulas for controls, the classical first integrals of Lagrangian mechanics were used. The effectiveness of the proposed algorithm is illustrated by computer simulation.

Keywords: Manipulation robot, Impulse controls, First integrals.

Introduction

The purpose of controlling the manipulation robot is to transfer it from the initial position to the final. The significant nonlinearity of the mathematical model describing the movements of the manipulation robot does not allow the use of the methods of the mathematical control theory directly for the original model. Decomposition methods make it possible to reduce the dimension of the control problem, passing to approximate linear or integrable mathematical control models [1, 2]. The work considers a manipulation robot with three degrees of freedom, imitating the movement of a human hand, described in the monograph [3]. In [4], the problem of controlling the horizontal two-dimensional motion of this robot was studied. The original nonlinear control model was replaced with an integrable controlled model. The Pontryagin maximum principle was used to find controls in the performance problem. The use of special impulse controls in this work allows the kinetic energy of the manipulation robot to be used to reach the final position. This approach preserves the integrability of the original controlled mathematical model and, when solving the problem, uses the classical first integrals of Lagrangian mechanics. The article concludes with the results of numerical simulation of the algorithm.

1. Mathematical statement of the problem

A manipulation robot with three degrees of freedom, imitating the movement of a human hand, described in the monograph [3, p. 263]. Figure 1 shows a robot manipulator. The number 1 indicates the base of the robotic arm, 2 is a rack of vertically oriented shaft. This shaft is rigidly connected to the guide beam 4 and hand 5.

Kinetic energy of the manipulation robot is determined by the formula

$$T = \frac{1}{2}\left(m_1 z'^2 + (J_1 + J_2)\varphi'^2 + m_2(z'^2 + x'^2 + x^2\varphi'^2)\right),$$

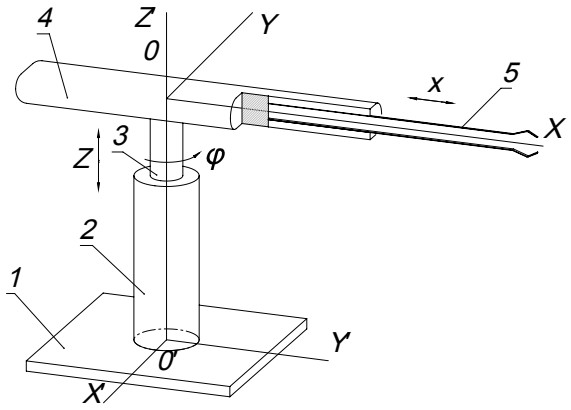


Figure 1

where x is the coordinate of the center of mass of the arm, $x > 0$; z is the height of the arm, $z > 0$; φ is the angle of rotation of the hand; m_1 is the total mass of the shaft that rotates the arm, and the guides that implement the horizontal displacement of the arm; m_2 is the mass of the arm; J_1 is the moment of inertia of the shaft and guides relative to the vertical axis; J_2 is the central moment of inertia of the arm relative to the vertical axis. The potential energy of gravity is $V = (m_1 + m_2)g$, where g is the gravitational acceleration.

The second-order Lagrange equations for the mechanical system under consideration are of the form [3, p. 263]

$$z'' + g = u_1, \tag{1.1}$$

$$x'' - x\varphi'^2 = u_2, \tag{1.2}$$

$$\varphi'' + \alpha^2 \left(x^2\varphi'\right)' = u_3. \tag{1.3}$$

Here $u_1 = P/(m_1 + m_2)$, $u_2 = F/m_2$, $u_3 = M/(J_1 + J_2)$ are the control force actions, where P is the magnitude of the longitudinal force acting along the vertical axis, F is the magnitude of the longitudinal force acting along the horizontal guides, M is the magnitude of the moment of force directed along the vertical axis, $\alpha = \sqrt{m_2/(J_1 + J_2)}$.

It is required to find the controls u_1, u_2, u_3 that translate the system (1.1)–(1.3) from the initial equilibrium position $(z_0, x_0, \varphi_0)^\top$, $z_0 > 0$, $x_0 > 0$, to the given end position $(z_*, x_*, \varphi_*)^\top$, $z_* > 0$, $x_* > 0$, $z_* \neq z_0$, $x_* \neq x_0$, $\varphi_* \neq \varphi_0$.

To solve this problem, we use a special set $\mathbb{U}$ of impulse controls defined by the formulas

$$u_1(t) = \dot{z}_0\delta(t), \quad u_2(t) = \dot{x}_0\delta(t), \quad u_3(t) = \dot{\varphi}_0\delta(t), \quad t \in \mathbb{R},$$

where $\delta(\cdot)$ is the Dirac impulse function. Impulse controls at the initial moment of time $t = 0$ to the dynamic system (1.1)–(1.3), which are in equilibrium, report the initial velocities $z'(0) = \dot{z}_0$, $x'(0) = \dot{x}_0$, $\varphi'(0) = \dot{\varphi}_0$. The initial speeds are control parameters, the choice of which should ensure that the dynamic system (1.1)–(1.3) falls into the final position.

2. Impulse control parameters

The problem is divided into two problems. In the first problem, for the equation (1.1) that determines the vertical movement of the manipulation robot, we use the set $\mathbb{U}_1$ of impulse controls defined by the formula $u_1(t) = \dot{z}_0\delta(t)$, $t \in \mathbb{R}$. It is required to find the control u_1, which transfers the dynamical system (1.1) from the initial equilibrium position z_0, $z_0 > 0$, to the given final

position z_*, $z_* > 0$, $z_* \neq z_0$. In the second problem, for the system of equations (1.2), (1.3), which determines the horizontal movement of the manipulation robot, we use the set $\mathbb{U}_2$ of impulse controls defined by the formulas $u_2(t) = \dot{x}_0\delta(t)$, $u_3(t) = \dot{\varphi}_0\delta(t)$, $t \in \mathbb{R}$. It is required to find controls u_2, u_3 that translate the dynamical system (1.2), (1.3) from the initial equilibrium position $(x_0, \varphi_0)^\top$, $x_0 > 0$, to the given end position $(x_*, \varphi_*)^\top$, $x_* > 0$, $x_* \neq x_0$, $\varphi_* \neq \varphi_0$.

Lemma 1. *In the set $\mathbb{U}_1$ there are impulse controls u_1 that transfer the dynamical system (1.1) from an arbitrary initial position z_0, $z_0 > 0$, to an arbitrary end position z_*, $z_* > 0$, $z_* \neq z_0$.*

P r o o f. The impulse control $u_1 \in \mathbb{U}_1$ provides the dynamic system (1.1) at the initial moment $t = 0$ the initial velocity $z'(0) = \dot{z}_0$. For $t > 0$, the motion of a free dynamic system is determined by the differential equation $z'' + g = 0$ with the initial conditions $z(0) = z_0$, $z'(0) = \dot{z}_0$. The vertical movement is described by the formula $z(t) = -gt^2/2 + \dot{z}_0 t + z_0$, $t > 0$. For any $\tau_1 > 0$ there is a unique value of the control parameter z_0 for which the equality $z(\tau_1) = z_*$, $\dot{z}_0 = \tau_1^{-1}(z_* - z_0) + g\tau_1/2$ is true.

We show that the velocity at the finite moment of time is minimal in absolute value if $\tau_1 = \sqrt{2|z_* - z_0|/g}$. The velocity at the finite moment of time is determined by the formula

$$z'(\tau_1) = -g\tau_1 + \dot{z}_0 = -g\tau_1/2 + \tau_1^{-1}(z_* - z_0).$$

If $z_* > z_0 > 0$, then we have $z'(\tau_1) = 0$ under the condition $\tau_1 = \sqrt{2(z_* - z_0)/g}$. The function $f(\tau) = g_1/2 + \tau^{-1}(z_0 - z_*)$ has a minimum at $\tau = \tau_1 = \sqrt{2(z_0 - z_*)/g}$ which is equal to $\sqrt{2g(z_0 - z_*)}$ if $z_0 > z_* > 0$.

The impulse controls $u_2, u_3 \in \mathbb{U}_2$ prescribe the dynamic system (1.2), (1.3) at the initial moment $t = 0$ the initial speeds $x'(0) = \dot{x}_0$, $\varphi'(0) = \dot{\varphi}_0$. For $t > 0$ the motion of a free dynamic system is determined by differential equations $x'' - x\varphi'^2 = 0$, $\varphi'' + \alpha^2\left(x^2\varphi'\right)' = 0$ with initial conditions $x(0) = x_0$, $\varphi(0) = \varphi_0$, $x'(0) = \dot{x}_0$, $\varphi'(0) = \dot{\varphi}_0$. For the horizontal free movement of the manipulation robot, kinetic energy and momentum are kept the same [5]

$$T_2 = \frac{1}{2}\left((J_1 + J_2)\varphi'^2 + m_2(x'^2 + x^2\varphi'^2)\right) = \text{const},$$

$$p_2 = \frac{\partial T_2}{\partial \varphi'} = (J_1 + J_2 + m_2x^2)\varphi' = \text{const}.$$

When describing horizontal motion for $t > 0$, we replace the system (1.1), (1.2) with the system of differential equations

$$\alpha^2x'^2 + (1 + \alpha^2x^2)\varphi'^2 = c_1, \tag{2.1}$$

$$(1 + \alpha^2x^2)\varphi' = c_2, \tag{2.2}$$

where

$$c_1 = \alpha^2\dot{x}_0^2 + (1 + \alpha^2x_0^2)\dot{\varphi}_0^2, \quad c_2 = (1 + \alpha^2x_0^2)\dot{\varphi}_0.$$

□

Lemma 2. *Let the conditions*

$$|\varphi_* - \varphi_0| \leq \int_{x_0}^{x_*} \frac{\sqrt{1 + \alpha^2x_0^2}ds}{\sqrt{(1 + \alpha^2s^2)(s^2 - x_0^2)}}, \quad 0 < x_0 < x_*, \tag{2.3}$$

$$|\varphi_* - \varphi_0| \leq \int_{x_*}^{x_0} \frac{\sqrt{1 + \alpha^2x_*^2}ds}{\sqrt{(1 + \alpha^2s^2)(s^2 - x_*^2)}}, \quad 0 < x_* < x_0, \tag{2.4}$$

hold. Then in the set $\mathbb{U}_2$ *there are impulse controls* u_1, u_2 *that move the dynamical system* (1.2), (1.3) *from the starting position* $(x_0, \varphi_0)^\top$, $x_0 > 0$, *to the ending position* $(x_*, \varphi_*)^\top$, $x_* > 0$, $x_* \neq x_0$, $\varphi_* \neq \varphi_0$.

P r o o f. Under the condition $\dot{x}_0 \neq 0$, the system of differential equations (2.1), (2.2) is transformed to the following form

$$x' = \operatorname{sgn} \dot{x}_0 \sqrt{\dot{x}_0^2 + (1 + \alpha^2 x_0^2)\, \dot{\varphi}_0^2 \, \frac{x^2 - x_0^2}{1 + \alpha^2 x^2}}, \quad x \in \mathbb{X}, \tag{2.5}$$

$$\varphi' = \frac{(1 + \alpha^2 x_0^2)\dot{\varphi}_0}{1 + \alpha^2 x^2}, \quad x \in \mathbb{X}, \tag{2.6}$$

where

$$\mathbb{X} = \Big\{ x \in \mathbb{R}^+ : \ \dot{x}_0^2 + (1 + \alpha^2 x_0^2)\dot{\varphi}_0^2 \, \frac{x^2 - x_0^2}{1 + \alpha^2 x^2} \geq 0 \Big\}.$$

To move the motion of the dynamical system (2.5), (2.6) from the initial to the final position, the control parameters $\dot{x}_0$, $\dot{\varphi}_0$ must satisfy the conditions

$$\dot{\varphi}_0 \neq 0, \quad \operatorname{sgn} \dot{x}_0 = \operatorname{sgn} (x_* - x_0), \quad \operatorname{sgn} \dot{\varphi}_0 = \operatorname{sgn} (\varphi_* - \varphi_0).$$

We introduce the parameter $p = |\dot{x}_0|/|\dot{\varphi}_0|$. Now the description of the set $\mathbb{X} = \mathbb{X}(p)$ is simplified. As a result, we have

$$\mathbb{X}(p) = \mathbb{R}^+ \text{ under } \ p \geq x_0 \sqrt{1 + \alpha^2 x_0^2},$$

$$\mathbb{X}(p) = \mathbb{R}^+ / (0, x_1(p)) \text{ under } \ p < x_0 \sqrt{1 + \alpha^2 x_0^2},$$

where

$$x_1(p) = \sqrt{\frac{x_0^2(1 + \alpha^2 x_0^2) - p^2}{1 + \alpha^2 x_0^2} + \alpha^2 p^2}.$$

The equation (2.5) is converted to

$$x' = \operatorname{sgn} (x_* - x_0) |\dot{\varphi}_0| \sqrt{p^2 + (1 + \alpha^2 x_0^2) \frac{x^2 - x_0^2}{1 + \alpha^2 x^2}}, \quad x \in \mathbb{X}. \tag{2.7}$$

It is also valid for $p = 0$. Using (2.7) and (2.6), we obtain a differential equation for finding the trajectory of a horizontal movement

$$\frac{dx}{d\varphi} = \frac{\operatorname{sgn} (x - x_0)(1 + \alpha^2 x^2)}{\operatorname{sgn} (\varphi_* - \varphi_0)(1 + \alpha^2 x_0^2)} \sqrt{p^2 + (1 + \alpha^2 x_0^2) \frac{x^2 - x_0^2}{1 + \alpha^2 x^2}}, \quad x \in \mathbb{X}(p).$$

Integrating the differential equation, we find the equation of the trajectory of a horizontal movement

$$\operatorname{sgn} (x_* - x_0) \int\limits_{x_0}^{x} \frac{(1 + \alpha^2 x_0^2) ds}{(1 + \alpha^2 s^2) \sqrt{p^2 + (1 + \alpha^2 x_0^2) \dfrac{s^2 - x_0^2}{1 + \alpha^2 s^2}}} = \operatorname{sgn} (\varphi_* - \varphi_0)(\varphi - \varphi_0), \quad x \in \mathbb{X}(p).$$

The trajectory passes through the end point if the condition is true

$$\left| \int\limits_{x_0}^{x_*} \frac{(1 + \alpha^2 x_0^2) ds}{(1 + \alpha^2 s^2) \sqrt{p^2 + (1 + \alpha^2 x_0^2) \dfrac{s^2 - x_0^2}{1 + \alpha^2 s^2}}} \right| = |\varphi_* - \varphi_0|, \quad x_* \in \mathbb{X}(p).$$

We select the value of the parameter p to satisfy the condition obtained. For $0 < x_0 < x_*$, the required value of the parameter p is determined by the equation

$$\int_{x_0}^{x_*} \frac{ds}{(1+\alpha^2 s^2)\sqrt{p^2 + (1+\alpha^2 x_0^2)\dfrac{s^2 - x_0^2}{1+\alpha^2 s^2}}} = \frac{|\varphi_* - \varphi_0|}{1+\alpha^2 x_0^2}, \quad 0 \le p < +\infty. \tag{2.8}$$

For $0 < x_* < x_0$, the required value of the parameter p is determined by the equation

$$\int_{x_*}^{x_0} \frac{ds}{(1+\alpha^2 s^2)\sqrt{p^2 + (1+\alpha^2 x_0^2)\dfrac{s^2 - x_0^2}{1+\alpha^2 s^2}}} = \frac{|\varphi_* - \varphi_0|}{1+\alpha^2 x_0^2}, \quad p_{kp} \le p < +\infty, \tag{2.9}$$

where

$$p_{kp} = \sqrt{\frac{(1+\alpha^2 x_0^2)(x_0^2 - x_*^2)}{1+\alpha^2 x_*^2}}.$$

The equation (2.8) has a unique solution $p = p_*$ under the condition (2.3) and equation (2.9) also has a unique solution $p = p_*$ if the conditions (2.4) hold.

Suppose that the condition (2.3) holds for $0 < x_0 < x_*$ and for $0 < x_* < x_0$ the condition (2.4) is satisfied. Then impulse control $u_2(t) = \dot{x}_0\delta(t)$, $u_3(t) = \dot{\varphi}_0\delta(t)$ $(t \in \mathbb{R})$ bring the dynamic system (1.2), (1.3) to the given final position, if the control parameters are determined by the formulas

$$\dot{x}_0 = |\dot{x}_0|\,\mathrm{sgn}\,(x_* - x_0), \quad \dot{\varphi}_0 = |\dot{\varphi}_0|\,\mathrm{sgn}\,(\varphi_* - \varphi_0), \quad |\dot{x}_0| = p_*|\dot{\varphi}_0|,$$

where p_* is the root of the equation (2.8) for $0 < x_0 < x_*$, and for $0 < x_* < x_0$ there is a root of the equation (2.9).

Integrating the differential equation (2.7), we find the arrival time of the motion of the dynamical system (1.2), (1.3) at the end point

$$\tau_2 = \frac{1}{|\dot{\varphi}_0|}\left|\int_{x_0}^{x_*} \frac{ds}{\sqrt{p_*^2 + (1+\alpha^2 x_0^2)\dfrac{s^2 - x_0^2}{1+\alpha^2 s^2}}}\right|.$$

We synchronize the arrival times of the movements of the dynamical systems (1.1)–(1.3) to the end points. □

Theorem 1. *Let the conditions of Lemma 2 be satisfied. Then the values of the parameters of the impulse controls that move the dynamical system (1.1)–(1.3) from the initial position* $(z_0, x_0, \varphi_0)^\top$, $z_0, x_0 > 0$, *to end position* $(z_*, x_*, \varphi_*)^\top$, $z_* > 0$, $x_* > 0$, $z_* \ne z_0$, $x_* \ne x_0$, $\varphi_* \ne \varphi_0$ *are defined by formulas*

$$\dot{z}_0 = \sqrt{2g(z_* - z_0)} \quad \textit{for} \quad 0 < z_0 < z_*,$$
$$\dot{z}_0 = 0 \quad \textit{for} \quad 0 < z_* < z_0,$$
$$\dot{\varphi}_0 = \frac{\sqrt{g}\,\mathrm{sgn}\,(\varphi_* - \varphi_0)}{\sqrt{2|z_* - z_0|}}\left|\int_{x_0}^{x_*} \frac{ds}{\sqrt{p_*^2 + (1+\alpha^2 x_0^2)\dfrac{s^2 - x_0^2}{1+\alpha^2 s^2}}}\right|,$$
$$\dot{x}_0 = \frac{\sqrt{g}p_*}{\sqrt{2|z_* - z_0|}}\int_{x_0}^{x_*} \frac{ds}{\sqrt{p_*^2 + (1+\alpha^2 x_0^2)\dfrac{s^2 - x_0^2}{1+\alpha^2 s^2}}}.$$

P r o o f. Using Lemma 1 and the arrival time τ_1 of the motion of the dynamical system (1.1) to the end point, we find the value of the control parameter $\dot{z}_0$. The synchronization condition $\tau_1 = \tau_2$ of the arrival times of the motions of dynamical systems (1.1)–(1.3) at end points determines the arrival time $\tau = \sqrt{2|z_* - z_0|/g}$ of dynamic system movements (1.1)–(1.3) to the end point $(z_*, x_*, \varphi_*)^\top$ and the equation for the control parameter $\dot{\varphi}_0$. From this equation we find

$$|\dot{\varphi}_0| = \frac{\sqrt{g}}{\sqrt{2|z_* - z_0|}} \left| \int_{x_0}^{x_*} \frac{ds}{\sqrt{p_*^2 + (1+\alpha^2 x_0^2)\dfrac{s^2 - x_0^2}{1+\alpha^2 s^2}}} \right|.$$

Using Lemma 2, we find the control parameters $\dot{\varphi}_0$, $\dot{x}_0$. □

3. Stabilization of manipulation robot in a final position

When stabilizing the manipulation robot in a small neighborhood of the final position, we use special positional controls, the choice of which turns the final position into a stable equilibrium position of the controlled system. For this purpose we use substitutions for coordinates

$$z = z_* + \hat{z}, \quad x = x_* + \hat{x}, \quad \varphi = \varphi_* + \hat{\varphi}$$

and controls

$$u_1 = \hat{u}_1 + g, \quad u_2 = \hat{u}_2, \quad u_3 = \hat{u}_3(1 + \alpha^2 x_*^2).$$

In a small neighborhood of the final equilibrium, the controlled system (1.1)–(1.3) is replaced by the following controlled system

$$\hat{z}'' = \hat{u}_1, \quad \hat{x}'' = \hat{u}_2, \quad \hat{\varphi}'' = \hat{u}_3. \tag{3.1}$$

We find the stabilizing control using the theory of optimal stabilization for linear systems with quadratic quality criteria. Choosing the quality criterion

$$J_1 = \int_0^{+\infty} \left(\hat{z}^2(t) + k_1^2 \hat{z}'^2(t) + \hat{u}_1^2(t)\right) dt, \quad k_1 > 0, \tag{3.2}$$

for the first control of the system (3.1), we find the stabilizing control

$$\hat{u}_1 = -\hat{z} - \sqrt{k_1^2 + 2}\hat{z}'.$$

We also can find stabilizing controls for the second and third equations in (3.1)

$$\hat{u}_2 = -\hat{x} - \sqrt{k_2^2 + 2}\hat{x}', \quad \hat{u}_3 = -\hat{\varphi} - \sqrt{k_3^2 + 2}\hat{\varphi}',$$

using a quality criteria similar to (3.2) with constants k_2 and k_2, respectively.

4. Numerical modeling

In the numerical simulation of the system motions (1.1)–(1.3), the following values of the parameters of the mechanical system were used

$$m_1 = 20, \quad m_2 = 8, \quad J_1 = 12, \quad J_2 = 6, \quad g = 9.8.$$

The start and the end positions are $z_0 = 0$, $x_0 = 0$, $\varphi_0 = 0$ and $z_* = 1.4$, $x_* = 0.5$, $\varphi_* = 1.2$.

We take the controls

$$u_1(t,z) = u_1^{pr}(t) + u_1^{ps}(z), \quad u_2(t,x) = u_2^{pr}(t) + u_2^{ps}(x), \quad u_3(t,\varphi) = u_3^{pr}(t) + u_3^{ps}(\varphi).$$

The program control is defined by formulas

$$u_1^{pr}(t) = \dot z_0 \delta(t), \quad u_2^{pr}(t) = \dot x_0 \delta(t), \quad u_3^{pr}(t) = \dot\varphi_0 \delta(t), \quad t \in \mathbb{R},$$

where the parameters are given by formulas

$$\dot z_0 = \sqrt{2gz_*}, \quad \dot x_0 = p_* \dot\varphi_0, \quad \dot\varphi_0 = \sqrt{\frac{g}{2z_*}} \int\limits_0^{x_*} \sqrt{\frac{1+\alpha^2 s^2}{p_*^2(1+\alpha^2 s^2)+s^2}}\, ds.$$

Here $p = p_*$ is the positive root of the equation

$$\int\limits_0^{x_*} \frac{ds}{\sqrt{(1+\alpha^2 s^2)((1+\alpha^2 s^2)p^2 + s^2)}} = \varphi_*.$$

Impulse controls moves the mechanical system into equilibrium, the initial speeds are

$$z'(+0) = \dot z_0, \quad x'(+0) = \dot x_0, \quad \dot\varphi(+0) = \dot\varphi_0.$$

We also consider software controls in the form of rectangular impulses, which are approximations of ideal impulses

$$u_1^{pr}(t) = \dot z_0 \delta_\Delta(t), \quad u_2^{pr}(t) = \dot x_0 \delta_\Delta(t), \quad u_3^{pr}(t) = \dot\varphi_0 \delta_\Delta(t), \quad t \in \mathbb{R},$$

where

$$\delta_\Delta(t) = 1/\Delta, \quad t \in (0,\Delta), \quad \delta_\Delta(t) = 0, \quad t \in \mathbb{R}/(0,\Delta), \quad \Delta = 0.1.$$

For these controls, the initial velocities of the equilibrium mechanical system are determined by the formulas $z'(0) = 0$, $x'(0) = 0$, $\dot\varphi(0) = 0$.

Positional controls are determined by the following formulas

$$\begin{aligned}
&u_1^{ps}(z) = 0, && 0 < z \le z_* - \epsilon_1,\\
&u_1^{ps}(z) = g - (z - z_*) - \sqrt{k_1^2 + 2}\, z', && z > z_* - \epsilon_1,\\
&u_2^{ps}(x) = 0, && 0 < x \le x_* - \epsilon_2,\\
&u_2^{ps}(x) = -(x - x_*) - \sqrt{k_2^2 + 2}\, x', && x > x_* - \epsilon_2,\\
&u_3^{ps}(\varphi) = 0, && 0 < \varphi \le \varphi_* - \epsilon_3,\\
&u_3^{ps}(\varphi) = -(1+\alpha^2 x_*^2)\big((\varphi - \varphi_*) + \sqrt{k_3^2 + 2}\,\varphi'\big), && \varphi > \varphi_* - \epsilon_3,
\end{aligned}$$

where

$$k_1 = 1, \quad k_2 = 1, \quad k_3 = 1, \quad \epsilon_1 = 0.1, \quad \epsilon_2 = 0.1, \quad \epsilon_3 = 0.1.$$

In the final position, the following conditions must be met

$$|z(\tau_f)| < \epsilon, \quad |x(\tau_f)| < \epsilon, \quad |\varphi(\tau_f)| < \epsilon, \quad |z'(\tau_f)| < \epsilon, \quad |x'(\tau_f)| < \epsilon, \quad |\varphi'(\tau_f)| < \epsilon.$$

In the computational experiment, we assumed that $\epsilon = 0.01$.

The time movement to the final position is $\tau_f = 7.5\, sec.$ for impulse controls and $\tau_f = 7.51\, sec.$ for rectangular impulses approximating ideal impulse actions. Projections of phase trajectories on state planes (x, x') and (φ, φ') are shown in Fig. 2 and Fig. 3.

The trajectories corresponding to impulse controls are shown in blue, the trajectories corresponding to the approximations of ideal impulse controls are shown in brown.

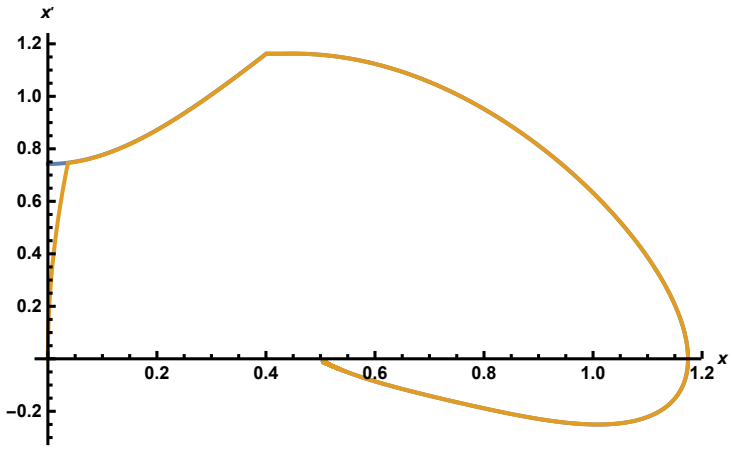


Figure 2. State plane (x, x').

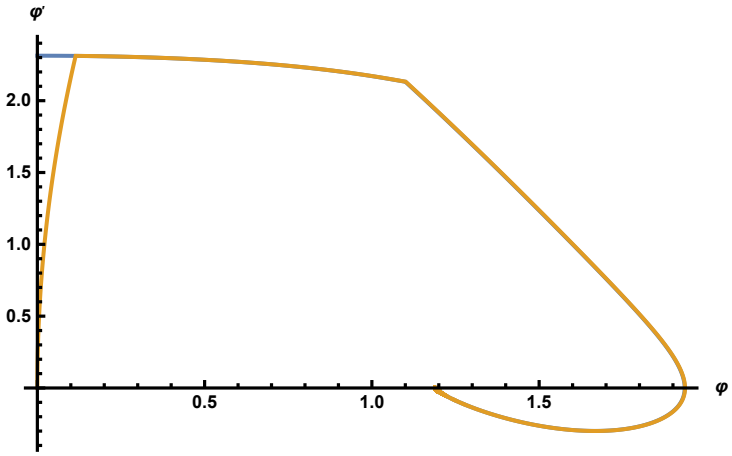


Figure 3. State plane (φ, φ').

5. Conclusion

The impulse control is constructed in the work that transfers the manipulator from a given position to its final position. A computational experiment showing the efficiency of the proposed algorithm is presented. The proposed algorithm is simulated in the case when the ideal impulse is approximated by the usual bounded control.

REFERENCES

1. Kozowski K. *Modelling and Identification in Robotics.* Ser. Adv. Ind. Control. London: Springer–Verlag, 1998. 261 p. DOI: 10.1007/978-1-4471-0429-2
2. Chernous'ko F. L., Ananievski I. M., Reshmin S. A. *Control of Nonlinear Dynamical Systems. Methods and Applications.* Comm. Control Engrg. Ser. Berlin, Heidelberg: Springer–Verlag, 2008. 396 p. DOI: 10.1007/978-3-540-70784-4
3. Chernousko F. L., Bolotnik N. N., Gradetsky V. G. *Manipulation Robots: Dynamics, Control and Optimization.* Boca Raton: CRC Press, 1994. 268 p.
4. Akulenko D. D., Bolotnik N. N., Kaplunov A. A. Some control modes of industrial manipulators. *Izv. AN SSSR. Tekhnicheskaia Kibernetika,* 1985. No. 6. P. 44–50.
5. Whittaker E. T. *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies.* Cambridge: Cambridge University Press. 1988. 456 p. DOI: 10.1017/CBO9780511608797

URAL MATHEMATICAL JOURNAL, Vol. 5, No. 2, 2019, pp. 21–30

DOI: 10.15826/umj.2019.2.003

CONTROL AND ESTIMATION FOR A CLASS OF IMPULSIVE DYNAMICAL SYSTEMS

Tatiana F. Filippova

Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences,
16 S. Kovalevskaya Str., Ekaterinburg, 620990, Russia

Ural Federal University,
19 Mira str., Ekaterinburg, 620002, Russia

ftf@imm.uran.ru

Abstract: The nonlinear dynamical control system with uncertainty in initial states and parameters is studied. It is assumed that the dynamic system has a special structure in which the system nonlinearity is due to the presence of quadratic forms in system velocities. The case of combined controls is studied here when both classical measurable control functions and the controls generated by vector measures are allowed. We present several theoretical schemes and the estimating algorithms allowing to find the upper bounds for reachable sets of the studied control system. The research develops the techniques of the ellipsoidal calculus and of the theory of evolution equations for set-valued states of dynamical systems having in their description the uncertainty of set-membership kind. Numerical results of system modeling based on the proposed methods are included.

Keywords: Control systems, Nonlinearity of quadratic type, Uncertainty, Impulse control, Ellipsoidal calculus, Tube of trajectories

Introduction

The paper is devoted to the state estimation problems for nonlinear control systems with uncertainty in description of their models. One of the central places in the theory of optimal control of dynamical systems is occupied by questions of constructing the corresponding reachable sets of the studied controlled systems, that is, the sets of all system positions obtained at a given time from a fixed initial state (or a set of such states) when all admissible controls are applied. Analysis of reachable sets and the construction of their different estimates may greatly facilitate the solution of many theoretical and applied problems of mathematical control theory. For linear controlled systems, the problem of describing and finding reachable sets has been considered in many papers and numerous ideas were involved to obtain external and internal estimates of reachable sets, basing on the corresponding versions of the ellipsoidal and polyhedral calculus [7, 8, 24, 26, 28, 35]. Note that even for linear systems studied at that time, the assumption that there are different kinds of uncertainties in describing the dynamics of systems significantly complicated the problem and transferred it to the class of nonlinear optimization problems.

A new stage in the development of approaches to solving nonlinear problems of estimating the states of control systems with uncertainty was carried out in connection with important researches in the field of set-valued analysis and in the theory of differential inclusions, including studies of sets of trajectories of control systems or differential inclusions with additional state constraints (the viability theory) [2, 27, 29, 32, 36, 37].

In this paper we study the case of a set-membership uncertainty [26–29, 32, 35] when only upper bounds on uncertain items are known and any additional probability characteristics for uncertainties are not done. Under such informational assumptions it is not possible to construct

precisely related reachable sets of the dynamical control system but instead we may find external and (or) internal estimating sets for them using simple canonical structures (for example, ellipsoids or polyhedra). The proposed approaches are motivated by the development of the theory of uncertain control systems and can be used in further researches related to filtering, forecasting and smoothing problems for mechanical systems described by stochastic differential equations, multistep equations and inclusions, these results may help in solving a range of optimization problems for nonlinear controlled systems with impulse control, state constraints and uncertainty, they may be used also in the study of irregular problems of optimal control and in studies of resistance movements systems with generalized controls and with a delay uncertainty.

The approaches presented in this paper are based on main ideas of early research [2, 9, 27] and are further developed for a different and more complicated classes of uncertain systems, the research continues and develops the results of the most recent studies [8, 11, 13, 15–19] for a wider class of control systems. Here we study the problems of constructing and estimating reachable sets of dynamical systems with impulse control [10, 12] and with uncertainty in the parameters of the systems dynamics and in the specification of its initial state. We further develop here the approaches related to consideration of bilinear uncertainties using the Minkowski gauge functionals [20].

Here we consider a more complicated case of a dynamic system than in papers [10, 12, 21], and we assume here that the impulse controls in the system are vectorial, which somewhat complicates both the previous analysis of the system dynamics and the corresponding proposed constructions, as well as the basic algorithm for constructing external estimates of reachable sets. Note that the issues of constructing internal ellipsoidal estimates of reachable sets of control systems with generalized (impulse) controls in both scalar and vector cases are much more complicated and are under development.

The results given here may be used in model-based advanced control of complex systems, such as adaptive control, robust control, sliding-mode control, H-infinite control, etc. [1, 3–6, 23, 25, 30]. Methods and schemes proposed in the paper possess such features as reliability, sufficient simplicity of computational algorithms and relatively high speed of their processing, so these schemes allow using them in real time e.g. in problems of robust control, stability, problems of control synthesis for dynamic systems of various types including problems of forecasting financial results in economic planning and other fields.

The paper is organized as follows. We introduce first some notations and definitions and formulate the main problem in Section 2. The approach related to upper estimates of reachable sets in nonlinear case under study is described in Section 3. Example illustrating the results is given in Section 4. Finally, some concluding remarks are given.

1. Problem formulation

In this section we introduce some basic notations and constructions and formulate further the main problem of state estimation for nonlinear control system with uncertainty and with impulsive controls of vector type.

1.1. Main notations

Let $\mathbb{R}^n$ denote the n–dimensional Euclidean space and $x'y$ is the usual inner product of vectors $x, y \in \mathbb{R}^n$ (the prime corresponds to a transpose), $\|x\| = (x'x)^{1/2}$. We will use also other norms of $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, namely $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ for $1 \le p < \infty$. The symbol comp $\mathbb{R}^n$ stands for the variety of all compact subsets $A \subset \mathbb{R}^n$ and conv $\mathbb{R}^n$ corresponds to a variety of all compact convex subsets $A \subset \mathbb{R}^n$.

Denote by clconv $\mathbb{R}^n$ the set of all closed convex subsets $A \subseteq \mathbb{R}^n$. Let $\mathbb{R}^{n\times m}$ be the set of all $n \times m$-matrices, $\mathrm{diag}\,\{v\}$ denotes a diagonal matrix with elements of a vector v standing at the main diagonal (and with zeros at other places). Let $I \in \mathbb{R}^{n\times n}$ be the identity matrix and $\mathrm{Tr}\,(A)$ be the trace of $n \times n$-matrix $A = \{a_{ij}\}$ (the sum of its diagonal elements, $\mathrm{Tr}\,(A) = \sum_{i=1}^{n} a_{ii}$).

We denote also as $B(a, r) = \{x \in \mathbb{R}^n : \|x - a\| \le r\}$ the ball in $\mathbb{R}^n$ with a center $a \in \mathbb{R}^n$ and a radius $r > 0$ and denote as

$$E(a, Q) = \{x \in \mathbb{R}^n : \ (Q^{-1}(x - a), (x - a)) \le 1\}$$

the ellipsoid in $\mathbb{R}^n$ with a center $a \in \mathbb{R}^n$ and with a symmetric positive definite $n \times n$-matrix Q.

1.2. Problem description

Consider the following impulsive control system ($t_0 \le t \le T$, $x \in \mathbb{R}^n$)

$$\begin{gathered} dx(t) = \big(A(t)x(t) + x'Bx \cdot d + u(t)\big)dt + Cdv(t), \\ x(t_0 - 0) = x_0 \in X_0 = E(a_0, Q_0). \end{gathered} \tag{1.1}$$

Here a matrix $A(t)$ is unknown but satisfies the constraint

$$A(t) \in \mathcal{A} = A^0 + \mathcal{A}^1, \quad t_0 \le t \le T,$$

where A^0 is a given matrix and

$$\begin{gathered} \mathcal{A}^1 = \big\{A{=}\{a_{ij}\}{\in}\mathbb{R}^{n\times n} : a_{ij} = 0 \ \text{ for } \ i \ne j, \ \text{ and} \\ a_{ii} = a_i, \ \ i = 1, \dots, n, \ \ a = (a_1, \dots, a_n), \ \ a'Da \le 1\big\}, \end{gathered} \tag{1.2}$$

with $D \in \mathbb{R}^{n\times n}$ being a given symmetric and positive definite matrix.

We assume that the impulsive part $v : [t_0, T] \to \mathbb{R}^m$ of the control pair $\{u(\cdot), v(\cdot)\}$ in (1.1) is of bounded variation on $[t_0, T]$, with

$$\operatorname*{Var}_{t\in[t_0,T]} v(t) = \sup_{\{t_i\}} \Big\{ \sum_{i=1}^{k} ||v(t_i) - v(t_{i-1})||_1 : \ \ \forall t_i : t_0 \le t_1 \le \ldots \le t_k = T \Big\} \le \mu, \tag{1.3}$$

where $\mu > 0$ is given. Denote the above class of functions $v(\cdot)$ as $\mathcal{V}$.

We assume also that $u(t) \in \mathcal{U} = E(\hat{a}, \hat{Q})$ where the center $\hat{a}$ and the matrix $\hat{Q}$ of the ellipsoid $\mathcal{U}$ are known.

The guaranteed estimation problem consists in describing the set

$$\begin{gathered} \mathcal{X}(t) = \mathcal{X}(t;\ t_0, X_0) = \big\{x \in \mathbb{R}^n : \ \exists\ x_0 \in X_0,\ \exists\ u(\cdot) \in \mathcal{U},\ \exists\ v(\cdot) \in \mathcal{V}, \ \ \exists\ A(\cdot) \in \mathcal{A}^1 \\ \text{such that } x = x(t) = x(t;\ u(\cdot),\ v(\cdot),\ x_0,\ A(\cdot))\big\}. \end{gathered}$$

of solutions to the system (1.1)–(1.2).

The problem studied here is to construct external ellipsoidal estimates for reachable sets $\mathcal{X}(t)$ ($t_0 < t \le T$) basing on recent results and on related techniques of the estimation theory for control systems with uncertainty and nonlinearity. We investigate a more complicated case than in [15, 17] and use here the technique recently developed in [19]. The main ideas used to solve the estimation problem go back to the results and reparametrization procedure of the papers [10, 31], with corresponding changes and improvements caused by the presence of vector measures (generalized controls).

2. Problem solution

The main result of the paper is connected with a special scheme of transition from a system of impulse type to a control system (or the corresponding differential inclusion) that does not contain impulse control components. Note that the proposed construction differs from the schemes of [10, 12, 21] where the case of scalar impulse components of control components was investigated.

2.1. Auxiliary constructions: impulsive differential inclusions

Consider a differential inclusion of the following type

$$dx(t) \in F(t, x(t))dt + C(t)dv(t), \tag{2.1}$$

with the initial condition

$$x(t_0 - 0) = x_0, \quad x_0 \in X_0.$$

Here we use the notation

$$F(t, x) = f(t, x, U) = \bigcup\{f(t, x, u) : \ u \in U\}.$$

Definition 1 [33]. *A function* $x[t] = x(t, t_0, x_0)$ $(x_0 \in X_0,\ t \in [t_0, T])$ *will be called a solution (a trajectory) of the differential inclusion* (2.1) *if for all* $t \in [t_0, T]$ *the following equality holds true*

$$x[t] = x_0 + \int_{t_0}^{t} \psi(t)dt + \int_{t_0}^{t} C(t)dv(t), \tag{2.2}$$

where $\psi(\cdot) \in L_1^n[t_0, T]$ *is a selector of* F*, that is* $\psi(t) \in F(t, x[t])$ *a.e. (the last integral in* (2.2) *is taken as the Riemann–Stieltjes integral).*

Following the scheme of the proof of the well-known Caratheodory theorem we can prove the existence of solutions $x(\cdot) = x(\cdot, t_0, x_0) \in BV^n[t_0, T]$ for all $x_0 \in X_0$ where $BV^n[t_0, T]$ is the space of n-vector functions with bounded variation at $[t_0, T]$.

2.2. Discontinuous time replacement

Let us introduce a new time variable [10, 31, 34],

$$\eta(t) = t + \int_{t_0}^{t} ||dv(t)||_1,$$

and a new state coordinate $\tau(\eta) = \inf\ \{t : \ \eta(t) \geq \eta\}$. Consider the following auxiliary differential inclusion

$$\frac{d}{d\eta} \begin{pmatrix} z \\ \tau \end{pmatrix} \in H(\tau, z) \tag{2.3}$$

with the initial condition

$$z(t_0) = x^0, \quad \tau(t_0) = t_0, \quad t_0 \leq \eta \leq T + \mu.$$

Here we denote

$$H(\tau, z) = \bigcup_{0 \leq \nu \leq 1} \left\{ \nu \begin{pmatrix} C^* \\ 0 \end{pmatrix} + (1 - \nu) \begin{pmatrix} Az + z'Bz \cdot d + E(\hat{a}, \hat{Q}) \\ 1 \end{pmatrix} \right\}, \tag{2.4}$$

where $C^* = \mathrm{co}\,\{c^{(1)}, \dots, c^{(m)}\}$ and $c^{(i)} \in \mathbb{R}^n$ $(i = 1, \dots, m)$ are columns of the matrix $C \in \mathbb{R}^{n\times m}$.

Under the above assumptions on the impulsive system we have two lemmas which will be used in further analysis.

Lemma 1. *The map $H(\tau, z)$ is convex and compact valued*

$$H : [t_0, T+\mu] \times \mathbb{R}^n \to \mathrm{conv}\,\mathbb{R}^{n+1}$$

and $H(\tau, z)$ is Lipschitz continuous in both variables τ, z.

P r o o f. The required properties can be easily derived from the specific type of set-valued map $H(\tau, z)$ defined above. □

Remark 1. Note that the design of the auxiliary differential inclusion (2.3) is different from the scheme used in [14]. The reason is the assumption of a vector type for impulse controls in (2.3)–(2.4). We also indicate that in the paper [14] a different type of constraints on undefined elements of the matrix $\mathcal{A}^1$ (in (1.2)) was investigated.

Denote $w = \{z, \tau\}$ the extended state vector of the system (2.3) and consider trajectory tube of this differential inclusion (which has no measure or impulse components):

$$W(\eta) = \bigcup_{w^0 \in X^0 \times \{t_0\}} w(\eta, t_0, w^0), \quad t_0 \le \eta \ \le T + \mu.$$

The next lemma explains the construction of the auxiliary differential inclusion (2.3)–(2.4).

Lemma 2. *The set $\mathcal{X}(T)$ is the projection of $W(T+\mu)$ at the subspace of state variables z:*

$$\mathcal{X}(T) = \pi_z W(T+\mu).$$

P r o o f. The proof of this result can be carried out according to the scheme of the paper [10], with a slight modification due to a more complicated case of the vector measure $dv(t)$ in (1.1) considered here. □

Denote as $h_M(z)$ the Minkowski (gauge) functional for a set $M \subset \mathbb{R}^n$ [9, 20],

$$h_M(z) = \inf\{t > 0 : z \in tM, \ x \in \mathbb{R}^n\},$$

and let $W(t; t_0, X_0 \times \{t_0\})$ be a trajectory tube of the inclusion (2.3)–(2.4).

Denote as $E(\tilde{c}, \tilde{Q})$ the ellipsoid with minimal volume and such that

$$C^* \subseteq E(\tilde{c}, \tilde{Q}). \tag{2.5}$$

Theorem 1. *For any $\sigma > 0$ the following inclusion is true*

$$W(t_0 + \sigma) \subseteq \mathcal{W}(t_0, \sigma, \nu) + o(\sigma)B_*(0,1), \quad \lim_{\sigma \to +0} \sigma^{-1} o(\sigma) = 0,$$

where

$$\mathcal{W}(t_0, \sigma, \nu) = \begin{pmatrix} E\big(a^*(\sigma,\nu), Q^*(\sigma,\nu)\big) \\ \\ t_0 + \sigma(1-\nu) \end{pmatrix},$$

$$a^*(\sigma,\nu) = a_0 + \sigma\big((1-\nu)(a_0{}'Ba_0 \cdot d + k^2 d + \hat{a}) + \nu\tilde{c}\big), \tag{2.6}$$

$$Q^*(\sigma,\nu) = (p^{-1}+1)\tilde{Q}(\sigma,\nu) + (p+1)\sigma^2 \hat{Q}^*_\nu,$$

with $E(\hat{a}_\nu, \hat{Q}^*_\nu)$ *being the ellipsoid with minimal volume such that*

$$\nu E(\tilde{c}, \tilde{Q}) + (1-\nu)E(\hat{a}, \hat{Q}) + 2(1-\nu)d \cdot a_0' B \cdot E(0, k^2 B^{-1}) \subseteq E(\hat{a}_\nu, \hat{Q}^*_\nu),$$

$$\hat{a}_\nu = \nu\tilde{c} + (1-\nu)\hat{a},$$

and where the function $\tilde{Q}(\sigma, \nu)$ *in* (2.6) *is defined as follows,*

$$\tilde{Q}(\sigma, \nu) = \operatorname{diag}\{(p^{-1}+1)\sigma^2 a_{0i}^2 + (p+1)r^2(\sigma) : \ i = 1, \dots, n\}, \tag{2.7}$$

with

$$r(\sigma) = \max_z ||z|| \cdot (h_{(I+\sigma\mathcal{A})*\mathcal{X}_0}(z, \sigma))^{-1}, \tag{2.8}$$

and $p = p(\sigma, \nu)$ *is the unique positive root of the equation*

$$\sum_{i=1}^{n} \frac{1}{p+\lambda_i} = \frac{n}{p(p+1)},$$

with numbers $\lambda_i = \lambda_i(\sigma, \nu) \geq 0$ $(i = 1, \dots, n)$ *satisfying the equation* $|\tilde{Q}(\sigma, \nu) - \lambda\sigma^2 \hat{Q}^*_\nu| = 0$.

P r o o f. In order to calculate the upper estimate for $W[t_0 + \sigma]$ we use first the inclusion (2.5) and therefore we may weaken the estimate (2.3)–(2.4) in the following way, considering the modified differential inclusion

$$\frac{d}{d\eta}\begin{pmatrix} z \\ \tau \end{pmatrix} \in H^*(\tau, z)$$

with the initial condition

$$z(t_0) = x^0, \quad \tau(t_0) = t_0, \quad t_0 \leq \eta \leq T + \mu,$$

where the set-valued map $H^*(\tau, z)$ is defined as

$$H^*(\tau, z) = \bigcup_{0 \leq \nu \leq 1} \left\{ \nu \begin{pmatrix} E(\tilde{c}, \tilde{C}) \\ 0 \end{pmatrix} + (1-\nu) \begin{pmatrix} Az + z'Bz \cdot d + E(\hat{a}, \hat{Q}) \\ 1 \end{pmatrix} \right\}.$$

Estimating the sum of two ellipsoids $\nu E(\tilde{c}, \tilde{C})$ and $(1-\nu)E(\hat{a}, \hat{Q})$ in the above formula (see, e.g., related procedures in [7, 28]) and using the results of Theorem 3 in [22] we come to the relations (2.6)–(2.8). □

Remark 2. To determinate a better estimate of the reachable set $\mathcal{W}(t_0 + \sigma)$ we may introduce a small parameter $\varepsilon > 0$ and embed the set $\mathcal{W}(t_0, \sigma, \nu)$ into a nondegenerate ellipsoid $E_\varepsilon\big(w(t_0, \sigma, \nu), O_\varepsilon(t_0, \sigma, \nu)\big)$:

$$\mathcal{W}(t_0, \sigma, \nu) \subseteq E_\varepsilon\big(w(t_0, \sigma, \nu), O_\varepsilon(t_0, \sigma, \nu)\big),$$

$$w(t_0, \sigma, \nu) = \begin{pmatrix} a^*(\sigma, \nu) \\ t_0 + \sigma(1-\nu) \end{pmatrix}, \quad O_\varepsilon(t_0, \sigma, \nu) = \begin{pmatrix} Q^*(\sigma, \nu) & 0 \\ 0 & \varepsilon^2 \end{pmatrix}.$$

For small $\varepsilon > 0$ we will have

$$\mathcal{W}(t_0, \sigma) \subset \mathcal{W}_\varepsilon(t_0, \sigma) = \bigcup_{0 \leq \nu \leq 1} E_\varepsilon\big(w(t_0, \sigma, \nu), O_\varepsilon(t_0, \sigma, \nu)\big) \subset E_\varepsilon(w^+(\sigma), O^+(\sigma)),$$

$$\lim_{\varepsilon \to +0} h(\mathcal{W}(t_0, \sigma), \mathcal{W}_\varepsilon(t_0, \sigma)) = 0,$$

here $h(A, B)$ is the Hausdorff distance between compact sets $A, B \subset \mathbb{R}^n$.

Further as the next step of the describing estimation procedure we may use the algorithms developed in [21] and applying them we construct the upper estimate $E_\varepsilon(w^+(\sigma), O^+(\sigma))$ for the union of ellipsoids $\mathcal{W}_\varepsilon(t_0, \sigma)$. Thus we get the ellipsoidal estimate of the reachable set $\mathcal{W}(t_0 + \sigma)$

$$\mathcal{W}(t_0 + \sigma) \subset E_\varepsilon(w^+(\sigma), O^+(\sigma)) + o(\sigma)B(0, 1).$$

Now we can formulate a new computational algorithm for the numerical construction of external ellipsoidal estimates for reachable sets of the system (1.1), this algorithm essentially uses the Theorem 1.

Algorithm (External Estimation of Reachable Sets).

Subdivide the time segment $[t_0, T + \mu]$ into subsegments $\{[t_i, t_{i+1}]\}$, where $t_i = t_0 + ih$ $(i = 1, \dots, m)$, $h = (T + \mu - t_0)/m$, $t_m = T + \mu$. Subdivide also the segment $[0, 1]$ into subsegments $[\nu_j, \nu_{j+1}]$, where $\nu_i = ih_*$, $h_* = 1/m$, $\nu_0 = 0$, $\nu_m = 1$.

1. Repeated steps begin with Step 1:

   - Take $\sigma = h$ and for given $X_0 = E(a_0, k^2 B^{-1})$ define by Theorem 1 the sets $\mathcal{W}(t_0, \sigma, \nu_i)$ $(i = 0, \dots, m)$.
   - Find ellipsoid $E_\varepsilon(w_1(\sigma), O_1(\sigma))$ in $\mathbb{R}^{n+1}$ such that

     $$\mathcal{W}(t_0, \sigma, \nu_i) \subseteq E_\varepsilon(w_1(\sigma), O_1(\sigma)) \quad (i = 0, \dots, m).$$

     At this step we find the ellipsoidal estimate for the union of a finite family of ellipsoids [21].
   - Find the projection $E(a_1, Q_1) = \pi_z E_\varepsilon(w_1(\sigma), O_1(\sigma))$ by Lemma 2.
   - Find the smallest $k_1 > 0$ such that $E(a_1, Q_1) \subseteq E(a_1, k_1^2 B^{-1})$ (k_1^2 is the maximal eigenvalue of the matrix $B^{1/2} Q_1 B^{1/2}$).
   - Consider the system on the next subsegment $[t_1, t_2]$ with $E(a_1, k_1^2 B^{-1})$ as the initial ellipsoid at instant t_1.

2. The next step repeats the previous iteration beginning with new initial data. At the end of the process we will get the external estimate $E(a^+(T), Q^+(T))$ of the reachable set of the system (1.1)–(1.3).

Remark 3. One of the subsequent steps of the above algorithm contains the projection of an ellipsoid on the subspace of the part of state variables, it complicates a bit the whole estimation procedure. But it is not possible to avoid this difficult step of the whole estimation process because of the presence of impulsive components in the control functions. One of the main goals of this paper is to overcome this complication.

3. Example

In this section we illustrate the main ideas and results obtained above by an example of an impulsive control system with uncertain initial set and with nonlinearity in dynamics.

Example. Consider the following control system

$$\begin{cases} dx_1 &= a_1 x_1 dt + u_1(t)dt + dv_1, \\ dx_2 &= a_2 x_2 dt + x_1^2 dt + x_2^2 dt + u_2(t)dt + 0.01 dv_2, \end{cases} \quad 0 \le t \le T,$$

with unknown initial state which belongs to a unit ball

$$x_0 \in X_0 = B(0,1).$$

Here we take $t_0 = 0$, $T = 0.4$, $\mathcal{U} = B(0,r)$, with $r = 0.01$. We have also $A = 2I$, $B = I$, $d_1 = 0$, $d_2 = 1$.

External ellipsoidal tube $E^+(t) = E(a^+(t), Q^+(t))$ is shown at Fig. 1, it is found using the main result of Theorem 1 and is constructed according to the the main Algorithm. The first estimating ellipsoid $E(0, k_0^2 B^{-1})$ is shown in red color and it contains X_0 (it is shown in blue color). It is worth recalling that the construction of the set $E(0, k_0^2 B^{-1})$ begins the whole iterative estimation process described by the Algorithm.

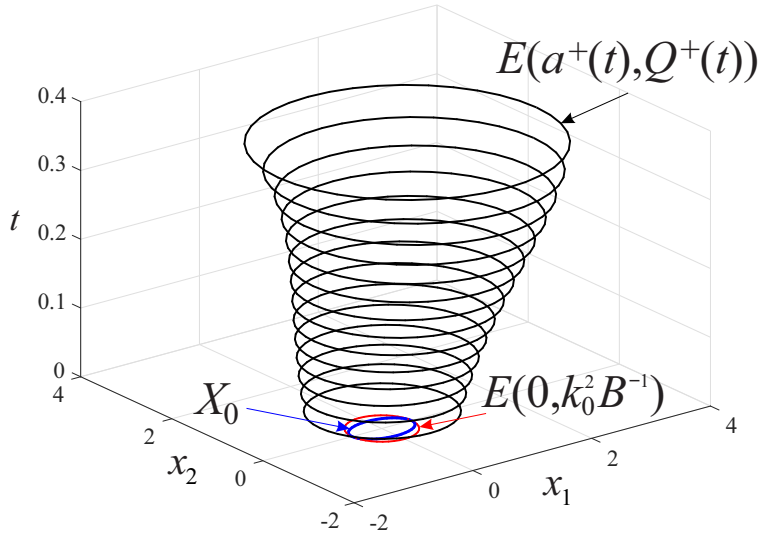


Figure 1. External ellipsoidal tube $E^+(t) = E(a^+(t), Q^+(t)), t \in [0, 0.4]$.

Remark 4. The example shows that the estimation errors can increase with time (accumulation effect). However, this is due to two factors of the model, the presence of nonlinear terms in the equations of dynamics and the presence of impulse controls.

4. Conclusions

The problems of state estimation for nonlinear impulsive control systems with unknown but bounded initial states were studied here. The solution was implemented based on the techniques of trajectory tubes of differential inclusions theory and also based on results of ellipsoidal calculus developed recently for these class of problems.

We study here the case when the system nonlinearity is generated by the combination of two types of functions in related differential equations, one of which is bilinear and the other one is quadratic. Additional difficulties in solving the considered problems were caused by the presence in the dynamic system of impulsive actions of a vector type.

The applications of the problems studied in this paper are in guaranteed state estimation for nonlinear systems with unknown but bounded errors and in related applied fields (e.g., in robotics, in problems of motor actuation, hydraulic actuation and others fields), the approaches developed here may be used in the model-based advanced control of complex systems, such as adaptive control, robust control, sliding-mode control, H-infinite control, etc.

Directions for further investigation continuing the paper research may be motivated by the studies in the theory of dynamic systems with uncertainty and with vector impulse controls under more complicated assumptions e.g. when the right hand sides of differential equations describing the system dynamics contain the product of state coordinates and the generalized (impulse) controls.

REFERENCES

1. Asselborn L., Groß D., Stursberg O. Control of uncertain nonlinear systems using ellipsoidal reachability calculus. *IFAC Proc. Volumes*, 2013. Vol. 46, no. 23. P. 50–55. DOI: 10.3182/20130904-3-FR-2041.00204
2. Aubin J.-P., Frankowska H. *Set-Valued Analysis.* Basel: Birkhäuser, 1990. 461 p.
3. August E., Lu J., Koeppl H. Trajectory enclosures for nonlinear systems with uncertain initial conditions and parameters. In: *Proc. of the 2012 American Control Conf., June 27–29, 2012, Montréal, Canada.* QC. IEEE Computer Soc., 2012. P. 1488–1493.
4. Blanchini F., Miani S. *Set-Theoretic Methods in Control.* Ser. Syst. Control: Foundations & Applications. Birkhäuser, Basel, 2015. XV+487 p. DOI: 10.1007/978-0-8176-4606-6
5. Boscain U., Chambrion T., Sigalotti M. On some open questions in bilinear quantum control. In: *Proc. of the European Control Conf. (ECC), July 17–19, 2013, Zurich, Switzerland.* IEEE Xplore, 2013. P. 2080–2085. DOI: 10.23919/ECC.2013.6669238
6. Ceccarelli N., Di Marco M., Garulli A., Giannitrapani A. A set theoretic approach to path planning for mobile robots. In: *Proc. 43rd IEEE Conf. on Decision and Control (CDC) Dec. 14–17, 2004, Nassau, Bahamas.* IEEE Xplore, 2004. P. 147–152. DOI: 10.1109/CDC.2004.1428621
7. Chernousko F. L. *State Estimation for Dynamic Systems.* CRC Press: Boca Raton, 1994. 320 p.
8. Chernousko F. L., Rokityanskii D. Ya. Ellipsoidal bounds on reachable sets of dynamical systems with matrices subjected to uncertain perturbation. *J. Optim. Theory Appl.*, 2000. Vol. 104, No. 1. P. 1–19. DOI: 10.1023/A:1004687620019
9. Demyanov V. F., Rubinov A. M. *Quasidifferential Calculus.* New York: Optimization Software Inc., 1986.
10. Filippova T. F. Set-valued solutions to impulsive differential inclusions. *Math. Comput. Model. Dyn. Syst.*, 2005. No. 11. P. 149–158.
11. Filippova T. F. Differential equations of ellipsoidal state estimates in nonlinear control problems under uncertainty. *Discrete Contin. Dyn. Syst.*, Vol. Suppl.-2011. 2011 P. 410–419. DOI: 10.3934/proc.2011.2011.410
12. Filippova T. F. Approximation techniques in impulsive control problems for the tubes of solutions of uncertain differential systems. *Advances in Applied Mathematics and Approximation Theory.* Springer Proc. Math. Stat., vol. 41. New York: Springer, 2013. P. 385-396. DOI: 10.1007/978-1-4614-6393-1_25
13. Filippova T. F. State estimation for a class of nonlinear dynamic systems with uncertainty through dynamic programming technique. In: *Proc. of the 6th Int. Conf. PhysCon2013, August 26–29, 2013, San Lois Potosi, Mexico*, 2013. P. 1–6.
14. Filippova T. F. Estimating reachable sets of control systems with uncertainty on initial data and with nonlinearity of a special kind. In: *Proc. of the Int. Conf. Stability and Oscillations of Nonlinear Control Systems (Pyatnitskiy's Conference), June 1–3, 2016, Moscow, Russia.* IEEE Xplore, 2016. P. 1–4. DOI: 10.1109/STAB.2016.7541183
15. Filippova T. F. Ellipsoidal estimates of reachable sets for control systems with nonlinear terms. *IFAC-PapersOnLine*, 2017. Vol. 50, No. 1. P. 15355–15360. DOI: 10.1016/j.ifacol.2017.08.2460
16. Filippova T. F. Dynamics and estimates of star-shaped reachable sets of nonlinear control systems. *J. Chaotic Modeling and Simulation (CMSIM)*, 2017. No. 4. P. 469–478.
17. Filippova T. F. Estimation of star-shaped reachable sets of nonlinear control system. In: *Lecture Notes in Comput. Sci., vol. 10665: Proc. Large-Scale Scientific Computing. LSSC 2017.* Cham: Springer, 2018. P. 210–218. DOI: 10.1007/978-3-319-73441-5_22

18. Filippova T. F. Differential equations for ellipsoidal estimates of reachable sets for a class of control systems with nonlinearity and uncertainty. *IFAC-PapersOnLine*, 2018. Vol. 51, No. 32. P. 770–775. DOI: 10.1016/j.ifacol.2018.11.452
19. Filippova T. F. Description of dynamics of ellipsoidal estimates of reachable sets of nonlinear control systems with bilinear uncertainty. In: *Lecture Notes in Comput. Sci., vol. 11189: Numerical Methods and Applications. NMA 2018.* Cham: Springer, 2019. P. 97–105. DOI: 10.1007/978-3-030-10692-8_11
20. Filippova T. F., Lisin D. V. On the estimation of trajectory tubes of differential inclusions. *Proc. Steklov Inst. Math.*, 2000. Vol. Suppl. 2. P. S28–S37.
21. Filippova T. F., Matviychuk O. G. Algorithms to estimate the reachability sets of the pulse controlled systems with ellipsoidal phase constraints. *Autom. Remote Control*, 2011. Vol. 72, No. 9. P. 1911–1924. DOI: 10.1134/S000511791109013X
22. Filippova T. F., Matviychuk O. G. Algorithms of estimating reachable sets of nonlinear control systems with uncertainty. *J. Chaotic Modeling and Simulation*, 2015. No. 3. P. 205–214.
23. Kishida M., Braatz R. D. Ellipsoidal bounds on state trajectories for discrete-time systems with linear fractional uncertainties. *Optim. Eng.*, 2015. Vol. 16. P. 695–711. DOI: 10.1007/s11081-014-9255-9
24. Kostousova E. K., Kurzhanski A. B. Theoretical framework and approximation techniques for parallel computation in set-membership state estimation. In: *Proc. of the Symposium on Modelling Analysis and Simulation, July 9–12, 1996, Lille, France*, 1996. No. 2. P. 849–854.
25. Kuntsevich V. M., Volosov V. V. Ellipsoidal and interval estimation of state vectors for families of linear and nonlinear discrete-time dynamic systems. *Cybernet. Systems Anal.*, 2015. Vol. 51. P. 64–73. DOI: 10.1007/s10559-015-9698-9
26. Kurzhanski A. B. *Upravlenie i nablyudenie v usloviyah neopredelennosti* [Control and Observation under Conditions of Uncertainty]. Moscow: Nauka, 1977. 392 p. (in Russian)
27. Kurzhanski A. B., Filippova T. F. On the theory of trajectory tubes — a mathematical formalism for uncertain dynamics, viability and control. In: *Advances in Nonlinear Dynamics and Control: a Report from Russia,* ed. A. B. Kurzhanski. Progress in Systems and Control Theory, vol. 17. Boston: Birkhäuser, 1993. P. 122–188. DOI: 10.1007/978-1-4612-0349-0_4
28. Kurzhanski A. B., Valyi I. *Ellipsoidal Calculus for Estimation and Control.* Systems Control Found. Appl. Basel: Birkhäuser, 1997. 321 p.
29. Kurzhanski A. B., Varaiya P. *Dynamics and Control of Trajectory Tubes: Theory and Computation.* Systems Control Found. Appl., vol. 85. Basel: Birkhäuser, 2014. 445 p. DOI: 10.1007/978-3-319-10277-1
30. Malyshev V. V., Tychinskii Yu. D. Construction of sets of attainability and maneuver optimization for low-thrust artificial satellites of the earth in a strong gravitational field. *J. Comput. Syst. Sci. Int.*, 2005. Vol. 44, No. 4. P. 622–630.
31. Miller B. M. Method of discontinuous time change in problems of control for impulse and discrete-continuous systems. *Autom. Remote Control*, 1993. Vol. 54, No. 12. P. 1727–1750.
32. Panasyuk A. I. Equations of attainable set dynamics. Part 1: Integral funnel equations. *J. Optimiz. Theory Appl.*, 1990. Vol. 64. P. 349–366. DOI: 10.1007/BF00939453
33. Pereira F. L., Filippova T. F. On a solution concept to impulsive differential systems. In: *Proc. of 4th Int. Conf. Tools for Mathematical Modelling (MathTools'03), June 23–28, 2003, St.Petersburg, Russia.* 2003. P. 350–355.
34. Rishel R. An extended Pontryagin principle for control system whose control laws contain measures. *SIAM J. Control*, 1965. Vol. 3. P. 191–205.
35. Schweppe F. C. *Uncertain Dynamic Systems.* New Jersey: Prentice-Hall, Englewood Cliffs, 1973. 563 p.
36. Veliov V. M. Second order discrete approximations to strongly convex differential inclusions. *Systems Control Lett.*, 1989. Vol. 13. P. 263–269.
37. Veliov V. M. Second-order discrete approximation to linear differential inclusions. *SIAM J. Numer. Anal.*, 1992. Vol. 29, no. 2. P. 439–451.

URAL MATHEMATICAL JOURNAL, Vol. 5, No. 2, 2019, pp. 31–54

DOI: 10.15826/umj.2019.2.004

LOCAL EXTENSIONS WITH IMPERFECT RESIDUE FIELD

Akram Lbekkouri

10507 Casa-Bandoeng, 20002 Casablanca, Morocco

lbeka11@gmail.com

Abstract: The paper deals with some aspects of general local fields and tries to elucidate some obscure facts. Indeed, several questions remain open, in this domain of research, and literature is getting scarce. Broadly speaking, we present a full description of the absolute Galois group in all cases with answers on the solvability, prosolvability and procyclicity. Furthermore, we give a result that makes "some" generalization to Abhyankar's Lemma in local case. Half-way a short section, containing a view of some future research loosely discussed, presents an attempt in the development of the theory. An Annexe elucidate several important points, concerning Hilbert's theory.

Keywords: Inertia group, Abhyankar's Lemma, Imperfect residue field, Weakly unramified, Solvability, Monogenity.

Introduction

Local fields with perfect residue field (or more generally when the residue extension is assumed to be separable) were deeply studied. The general case, when dropping off the separability of the residue extension, considered for the first time in [25] still needs more work.

This condition of separability implies that the extension of the valuation rings is monogenic and plays an imminent role in the proofs of some standard results for example Hilbert formula, Herbrand property and Hasse-Arf Theorem which remain true under the less strong condition of monogeneity. Meanwhile the property of the congruence of the ramification breaks modulo the residual characteristic, (necessarily $p > 0$) does not hold if the residue extension is not separable, even by assuming the monogeneity of the respective valuation rings extension.

The residue field is only a "fair" field, and does not have to be CDV. When assuming it as local, we can characterize a large "family" of general local fields, more precisely "the higher dimensional local fields" (such fields need not be necessarily monogenic). Paršin introduced the "2-dimensional" local fields and constructed a class field theory of them, then Hyodo defined "upper" ramification breaks, as m-tuples, for a Galois extension of "m-dimensional" local fields (with finite last residue field).

The perfectness of the residue field ($\mathrm{char}\,(\overline{K}) = p > 0$), implies necessarily the separability of the residue extension. So, by assuming the less strong condition $[\overline{K} : \overline{K}^p] = p^c < \infty$, we make a step ahead to the generalization ("c" is called the degree of imperfectness). By taking $c = 1$, I. Zhukov in [26, §1] defines a good ramification theory under the hypothesis $[\overline{K} : \overline{K}^p] = p$ (i.e. $\overline{K}$ has a p basis of length 1). Especially for such fields, he proved that all weakly unramified extensions are well ramified and then monogenic. Zhukov's theory was for "2-dimensional" local fields only, then later it was generalized to "n-dimensional" local fields by V. A. Abrashkin [2].

It depends on the choice of a subfield of "1-dimensional constants" $\mathcal{K}$ in K (a field is "1-dimensional" if it is complete with respect to its discrete valuation and has a finite residue field, it is said to be "2-dimensional" if it is complete with respect to its discrete valuation and has a residue field which is itself "1-dimensional", and so on and so forth, we can define an "n-dimensional" local field).

The theory is presented by a ramification filtration on $\mathrm{gal}\,(K^{sep}/K)$, the absolute Galois group of K, by steps beginning with $\mathrm{gal}\,(\mathcal{K}^{sep}/\mathcal{K})$, the absolute Galois group of $\mathcal{K}$.

In fact, in characteristic zero he defined $\mathcal{K}$ as the set of all $x \in K$ which are algebraic over the fraction field $\mathcal{K}_0$ of $W(F)$ where $F = \cap \overline{K}^{p^i}$ and $W(F)$ is the Witt ring of F. Such $\mathcal{K}$ is the maximal for this property and is complete with a perfect residue field. Meanwhile, in characteristic $p > 0$ it is possible to fix a "base" subfield $\mathcal{B}$ in K, complete with respect to the valuation of K and having $\mathbb{F}_p$ as a residue field. That are the $\mathbb{F}_p((\tau))$ with $v_K(\tau) > 0$, if $\mathcal{K}$ is the algebraic closure in K of the completion of $\mathcal{B}(\mathcal{R}_K)$. Here $\mathcal{R}_K$ consists of Teichmüller representatives of elements of the maximal perfect subfield in $\overline{K}$.

Defining first a ramification filtration in classical way on $\mathrm{gal}\,(\mathcal{K}^{sep}/\mathcal{K})$ he introduces then a new lower filtration on $\mathrm{gal}\,(L/K)$ indexed by a special linear ordered set $\mathbb{I} \subset \mathbb{Q}^2$ (lexicographic order). Then a new Hasse-Herbrand function $\Phi : \mathbb{I} \mapsto \mathbb{I}$ is defined with all the usual properties. Therefore, a theory of upper ramification groups, in this case, is stated. He uses the method of "eliminating wild ramification" due to Epp [6] to reduce, in a canonical way, the study of completely ramified extensions to the last one of ferociously ramified extensions. For such extension the hypothesis on $[\overline{K} : \overline{K}^p]$ implies that the extension to consider is in fact ferociously ramified with $\overline{L}/\overline{K}$ generated by only one element i.e. it is monogenenic (Section 5), for which L. Spriano defines a more general ramification theory what he calls "case II", see [20, §5], [21]. Particularly in this case, the question of the "passage of the ramification to the quotient" is affirmatively solved.

Lastly, Abbes and Saito, using techniques of rigid geometry, define an upper ramification filtration in the general case successfully. Till now they cannot make the two filtrations (namely the lower of Hilbert–Zariski–Samuel and their upper) corresponding in a satisfactory way.

To sum up, the assumption of the monogeneity remains the first important step to generalization without losing the trueness of large number of important results.

Section Progression:

Here are three main sections, then a section of limelight questions and a last as annexe.

In Section 1 we prove the solvability of the inertia group of any finite extension regardless of the residual extension, then we give a discussion on the solvability of the Galois group.

In Section 2 we give a full description of the absolute Galois group in all cases.

In Section 3 Theorem 9 makes some generalization of Abhyankar's Lemma in local case.

Section 4 contains a view of some future research, an attempt to develop of the theory.

Section 5 is an Annexe section destined to briefly elucidate several important points, necessary for the study, concerning Hilbert's theory.

The main results are Theorems 1, 2, 3, 4, 5 and 9, Propositions 1 and 4, Lemma 1.

Nowhere else in the realm of abstract algebra does one see such an elegant interaction of topics as in the subject of General Theory of Local Fields.

*By **local field** we mean a complete discrete-valued field (CDVF), the residue field being not necessarily perfect. We say **classical case** when the residue field is perfect or at least when the residue extension is separable, otherwise we name it as **general case**.*

1. On the solvability in finite local extensions

Here, we study the solvability of the Galois group of some local extensions with possibly imperfect residue field. Theorem 1 is a direct proof of the solvability of the inertia group in general case, then results on solvability of n-dimensional local fields are given.

1.1. On the solvability of the inertia group

Let L/K be a finite Galois extension of local fields. The residue extension $\overline{L}/\overline{K}$ is normal, see [18, Proposition I.7.20], but need not be separable. Consider D the set of all automorphisms of $\overline{L}$ unvarying all elements of $\overline{K}$, there is a natural surjective homomorphism $\varphi : G \to D$. Indeed, let $g \in G$, g preserves $\mathcal{O}_L$ as well as $\mathcal{M}_L$, therefore g induces an automorphism of $\overline{L} = \mathcal{O}_L/\mathcal{M}_L$. Since g fixes each element of K it fixes each element of $\overline{K}$ as well, for the surjectivity of φ see the same reference. So, the inertia group of L/K is $G_0 = ker(\varphi)$, also G is solvable if and only if D and G_0 are too.

Theorem 1. *Let L/K be a finite Galois extension of any local fields without any assumption on the residual extension. Then the inertia group G_0 of L/K is solvable, furthermore it is cyclic when the residual characteristic is zero.*

It is a generalization of Serre's results in [18, Proposition IV.2.7] *and its corollaries. Published in* [11]*, the proof needs some necessary retouches that can be found here.*

P r o o f. An uniformizer π of L being fixed, let us fix a set of generators of the residue field extension and their lifts $u_1, ..., u_n$ to $\mathcal{O}_L$. Put it in another way, $\mathcal{O}_L$ is generated by $\pi, u_1, ..., u_n$ as an $\mathcal{O}_K$-algebra, with $v_(\pi) = 1$, u_i being units. Consider the map:

$$\begin{array}{rccc} \varphi_1 : & G_0 & \to & \overline{K}^\star, \\ & g & \to & \overline{g(\pi)/\pi}. \end{array}$$

It is clear that this is a homomorphism, write $J_1 = ker(\varphi_1)$ for the kernel of this map, so $J_1 = H_1$; we use Zariski–Samuel notation [25, ch. V, § 10]. Then again consider the homomorphism:

$$\begin{array}{rccc} \varphi_2 : & J_1 & \to & \overline{K} \oplus ... \oplus \overline{K};\ \text{(n+1) of them}, \\ & g & \to & (\overline{(g(\pi) - \pi)/\pi^2}, \overline{(g(u_1) - u_1)/\pi}, \quad ..., \overline{(g(u_n) - u_n)/\pi})), \end{array}$$

where $\overline{(g(\alpha) - \alpha)/\pi^i}$ is the class of $(g(\alpha) - \alpha)/\pi^i$ mod π. Set $J_2 = ker(\varphi_2)$. Again by considering,

$$\begin{array}{rccc} \varphi_3 : & J_2 & \to & \overline{K} \oplus ... \oplus \overline{K};\ \text{(n+1) of them}, \\ & g & \to & (\overline{(g(\pi) - \pi)/\pi^3}, \overline{(g(u_1) - u_1)/\pi^2}, \quad ..., \overline{(g(u_n) - u_n)/\pi^2})) \end{array}$$

and so on and so forth, until the filtration stabilizes (of course, since $\mathcal{O}_L \simeq \varprojlim \mathcal{O}_L/\mathcal{M}_L^i$) and we get a trivial J_r. From this, we conclude

1. **If the residual characteristic is** $p > 0$: it is clear that J_1 has a filtration by normal subgroups J_i, where the subquotients J_i/J_{i+1} are p-elementary abelian groups as J_i/J_{i+1} injectively maps to $(1+\mathcal{M}_L^i)/(1+\mathcal{M}_L^{i+1})$ which is canonically isomorphic to $(\overline{L}, +)$ for $i \geq 1$. Furthermore, G_0/J_1 is cyclic as it injectively maps to $\mathcal{R}_L^\star/(1 + \mathcal{M}_L) \simeq (\overline{L}^\star, \times)$, and to $Aut_{\overline{L}}(\mathcal{M}_L/\mathcal{M}_L^2) \simeq (\overline{L}^\star, \times)$ as well, and the field $\overline{L}$ is of characteristic p. (Remark the order of G_0/J_1 is prime to p if $p \geq 3$). Furthermore, worthy to note that the maximal tamely ramified subfield T of L corresponds to the subgroup J_1. Finally, J_1 is a p-group (the unique Sylow p-subgroup of G_0) it is of order $e_{wild}f_{insep}$, which then implies the solvability of G_0.

2. **When the residual characteristic is zero**: for $i \geq 1$ the subquotients J_i/J_{i+1} being isomorphic to a subgroup of $(\overline{L}, +)$ (additive), which has no finite subgroup except $\{0\}$, J_i are trivial for all $i \geq 1$ and G_0 is cyclic. □

Remark 1. J.P. Serre in [18, Corollary IV.2.5 of Proposition 7], inspired by Zariski–Samuel in [25], gives a proof of this theorem in the classical case. Unhappily his proof breaks down in the general case because he uses the $(G_n)_n$ (lower ramification subgroups Hilbert–Zariski–Samuel's filtration of G_0). Of course, in general G_0/G_1 need not be abelian, see [25, page 297, last line], the purely inseparable part of the residue extension playing main role. Indeed, Theorem 1 in the same reference claims that the group G_0/G_1 contains a normal subgroup G_1' which is reduced to the identity in separable case (see § 5.1).

1.2. Consequences

From Theorem 1 we straightforwardly deduce the corollaries:

Corollary 1. *Let K be a local field, and let L/K be a finite Galois extension. Then L/K is solvable if and only if the maximal separable subextension of $\overline{L}/\overline{K}$ is solvable.*

Corollary 2. *Consequently, in the classical case the Galois group of L/K is solvable if and only if the Galois group of $\overline{L}/\overline{K}$ is solvable.*

1.3. On n-dimensional local fields

A complete discrete-valued field K is said to have the structure of an n-dimensional local field if there is a chain of fields $K = K_n, K_{n-1}, ..., K_1, K_0$ where K_{i+1} is a complete discrete valuation field with residue field K_i and K_0 is a finite field. The field $\overline{K} = \overline{K_n} = K_{n-1}$ is said to be the first residue field of K, respectively K_0 is the last.
Recall some facts about n-dimensional local fields:

- When assuming the last residue K_0 is perfect rather than finite, we preserve most of the properties of n-dimensional local fields.
- Some authors referred to as an n-dimensional local field over a perfect field, rather than a finite field. But we consider an n-dimensional local field over an arbitrary field K_0 as well.
- Let L/K be a finite extension. If K is an n-dimensional local field, then so is L.

Since finite extensions of a finite field are cyclic, by induction (use Corollary 1) we get:

Corollary 3. *Every finite Galois extension of a "n-dimensional" local field with the residue field of the corresponding "1-dimensional" field is finite, has a solvable Galois group.*

In "Serre's sense" a field is said to be **quasi-finite** if it is perfect and $\mathrm{gal}\,(K^{sep}/K) \simeq \widehat{\mathbb{Z}}$ (K^{sep} being a separable closure of K and $\widehat{\mathbb{Z}}$ the profinite completion of $\mathbb{Z}$). Every finite quotient of $\widehat{\mathbb{Z}}$ is cyclic ($\widehat{\mathbb{Z}}$ is a profinite group as the projective limit of the finite cyclic groups $\mathbb{Z}/n\mathbb{Z}$) and thus is abelian and procyclic). Some authors allow themselves to say $\widehat{\mathbb{Z}}$ is cyclic as a topological group, even if it is not countable since the natural homomorphism $\mathbb{Z} \to \widehat{\mathbb{Z}}$ has a dense image.

So, Corollary 3 can be immediately generalized (in some sense) to the case when the residue field of the "1-dimensional" field is assumed to be quasi-finite only, if we allow ourselves to generalize the notion of "high-dimensional" local fields such way (replacing the finiteness of the residue field of the "1-dimensional" field by its perfectness). Even the perfectness of the residue field is not necessary. We can only assume that $\mathrm{gal}\,(\overline{K}^{sep}/\overline{K}) \simeq \widehat{\mathbb{Z}}$, or more generally prosolvable ($\overline{K}$ being the residue field of the "1-dimensional" field K).

But we cannot say that the result remains true when $\mathrm{gal}\,(\overline{K}^{sep}/\overline{K})$ is any profinite group. Indeed, a finite quotient of a profinite group need not be solvable. For this it is easy to construct a counter-example of course, $PSL(2, \mathbb{F}_q)$ is very often simple.

Corollary 4. *Every finite Galois extension of a "n-dimensional" local field, has necessarily a solvable Galois group if the residue field $\overline{K}$ of the corresponding "1-dimensional" form $\overline{K} = k((T))$ with k being an algebraically closed field of characteristic zero.*

P r o o f. It suffices to use the Corollary of the Proposition IV.2.8 in [18]. Indeed, we have the Galois group of the algebraic closure of $\overline{K}$ which is isomorphic to $\widehat{\mathbb{Z}}$. □

Notice that if $\mathrm{gal}\,(K^{sep}/K) \simeq \widehat{\mathbb{Z}}$ (e.g. if K is quasi-finite) then for every supernatural number n, K has only one Galois extension of degree n. Since $\widehat{\mathbb{Z}}$ has a unique closed subgroup of a given index n, see Theorem 2.7.2 in [14].

2. Absolute groups

Here, we give a whole description of the absolute groups and their classification by residual characteristic in the general case. More precise facts are found in § 2.4. Then we answer questions concerning: the nature of these groups.

By absolute groups of K a CDVF, we mean the Galois group G, the inertia group G_0 and G_W the wild ramification subgroup of a separable closure K^{sep}/K.

2.1. Hilbert decomposition of the separable closure

2.1.1. Presentation

For K being any field, consider K^{sep}/K a separable closure (that is the union of all finite Galois extensions of K), it is necessarily normal and then Galois. In general $K^{sep} \subseteq K^{alg}$, nevertheless $K^{sep} = K^{alg}$ if and only if K is perfect. Now, if K is a complete discrete-valued field then its valuation extends uniquely to K^{sep} but it is no more discrete on it, actually $v((K^{sep})^{\star}) = \mathbb{Q}$; furthermore, K^{sep} is not complete for the discussed valuation.

The Galois group $\mathrm{gal}\,(K^{sep}/K) = \mathcal{A}ut_K(K^{sep}/K)$, called absolute Galois group of K, is a compact topological group with respect to the profinite topology. Indeed, going over all finite extensions L/K, denote by $\mathfrak{L}$ the set of all finite Galois extensions L of K contained in K^{sep}/K, then we can write,

$$K^{sep} = \bigcup_{L\in\mathfrak{L}} L; \quad \text{and } \mathrm{gal}\,(K^{sep}/K) = \varprojlim_{L\in\mathfrak{L}} \mathrm{gal}\,(L/K).$$

Now, the maximal unramified extension K^{unr} of K in K^{sep} is the union of all fields L_0 (L_0 being the maximal unramified extension of K in L and is Galois over K), we too find that K_W, the union of all fields L_w (where L_w is a tamely ramified Galois extension in L that contains every tamely ramified extension of K in L), is a tamely ramified extension of K in K^{sep}. That is we have the tower:

$$K \text{ —— } K^{unr} \text{ —— } K_W \text{ —— } K^{sep}\ .$$

K^{unr}/K and K_W/K both are Galois and $G_W = \mathrm{gal}\,(K^{sep}/K_W)$ os the absolute wild ramification group (maybe trivial), which can be considered as the projective limit of a sequence of corresponding finite wild ramification p-subgroups (in all cases the ramification filtration always exists). So, G_W is prosolvable even more pronilpotent, but in general not solvable. That is the p-Sylow subgroup of $G_0 = \mathrm{gal}\,(K^{sep}/K^{unr})$ (the absolute inertia group), and a closed normal pro-p-subgroup of $G = \mathrm{gal}\,(K^{sep}/K)$. Furthermore, write $\overline{K}^{sep}$ as a separable closure of $\overline{K}$, $\overline{K}^{sep} = \mathcal{O}_{K^{unr}}/\mathcal{M}_{K^{unr}} =$

$\overline{K^{unr}}$. Indeed, the residue field of the maximal unramified extension of K is a separable closure of $\overline{K}$. Furthermore, $\mathrm{gal}\,(K^{unr}/K) = \mathrm{gal}\,(\overline{K}^{sep}/\overline{K})$, see [13, ch. II, Proposition 7.5].

Remark 2. It is well known that G_0/G_W is a torsion free abelian group, the q-Sylow subgroups of which are free $\mathbb{Z}_q$-modules of rank $dim_{\mathbb{F}_q}\Gamma/q\Gamma$ where Γ is the additive value group. The prime numbers q are necessarily different from the residue characteristic.

2.2. General description

First, let us notice some relatiohship between the unit group and the Galois group. Recall that the unit group is abelian and the absolute Galois group is not. However we know that there is some correspondence between the unit group and the Galois group of certain subextension of K^{sep}. Indeed, when K is a local field with finite residue field its unit group is isomorphic to the Galois group of a certain totally ramified abelian extension of K. For example, the extension of $\mathbb{Q}_p$ obtained by adjoining all p-power roots of unity has Galois group $\mathbb{Z}_p^\star$ over $\mathbb{Q}_p$. This generalizes to other base fields using Lubin–Tate formal groups.

If K is any complete discrete-valued field then the unit group $\mathcal{O}_K^\star$ (as well as $\mathcal{O}_K$) is compact if and only if the residue field of K is finite, so in the case of an infinite residue field the topological group $\mathcal{O}_K^\star$ could not be a Galois group since profinite groups are compact.

2.2.1. Classification by residual characteristic

Let us proceed by cases:

1. If $\mathrm{char}\,(\overline{K}) = 0$, all Galois extensions are tamely ramified, the inertia group of every finite extension is cyclic and its wild ramification subgroup is trivial, see the proof of Theorem 1, hence the absolute inertia group G_0 of the absolute Galois group is the profinite completion of $\mathbb{Z}$ i.e. is isomorphic to $\widehat{\mathbb{Z}}$ so it is procyclic (by the way abelian), meanwhile G_W is trivial. In consequence the absolute Galois group is a semi-direct product of the absolute inertia group by the absolute Galois group of the residue field i.e. $G \simeq \widehat{\mathbb{Z}} \rtimes \mathrm{gal}\,(\overline{K^{sep}}/\overline{K})$.
   Now, when the residue field $\overline{K}$ is algebraically closed $\overline{K}^{sep} = \overline{K}$, the maximal unramified extension is trivial, in consequence the absolute inertia group equals the absolute Galois group $G_0 = G$. So, we find the main result of Theorem 4 that comes.

2. If $\mathrm{char}\,(\overline{K}) = p > 0$, the absolute inertia subgroup G_0 of G is isomorphic to the extension of G_W by $\prod_{q\neg p}\mathbb{Z}_q$, where $\mathbb{Z}_q$ is the ring of q-adic integers with $q \neq p$. With K^{unr} being the field fixed by G_0 in K^{sep}, K^{unr}/K is a Galois extension such that $\mathrm{gal}\,(K^{unr}/K)$ is isomorphic to $G_{\overline{K}}$ where $G_{\overline{K}}$ is the absolute Galois group of the residue field $\overline{K}$. That is $G_0 = G/G_W \simeq \prod_{q\neq p} Z_q \rtimes G_{\overline{K}}$ with its Galois action. Indeed, for each integer q prime to p, the group of q-th roots of unity $\mu_q(\overline{K}^{sep})$ is cyclic of order q. Consider $\mathcal{Q}$ the set of all integers q prime to p ordered by divisibility, if $q' = q.m$ by means of the transition map (rising to power m) $\mu_{q'}(\overline{K}^{sep}) \to \mu_q(\overline{K}^{sep})$ we have a canonical isomorphism $G_0/G_W \simeq \varprojlim_{q\in\mathcal{Q}} \mu_q(\overline{K}^{sep})$.

The Tate twist of $\mathbb{Z}_q$ being defined by $\mathbb{Z}_q(1) = \mu_{q^\infty}(\overline{K}^{sep})$, write, $\widehat{\mathbb{Z}}' = \prod_{r\neq p}\mathbb{Z}_r$, and $\widehat{\mathbb{Z}}'(1) = \prod_{r\neq p}\mathbb{Z}_r(1)$, we have that $\widehat{\mathbb{Z}}'(1) \simeq \widehat{\mathbb{Z}}'$ the isomorphism being not canonic. Then we get $G_0/G_W \simeq \prod_{r\neq p}\mathbb{Z}_r(1)$. Since, $G/G_0 \simeq G_{\overline{K}}$ the action by conjugation of $G_{\overline{K}}$ on G/G_0 gives the natural action on $\widehat{\mathbb{Z}}'(1)$.

Furthermore, $G/G_0 \simeq G_{\overline{K}}$ and T the absolute "tame-inertia" subgroup $T \simeq \prod_{r\neq p}\mathbb{Z}_r$ is a normal subgroup of G/G_W. In other words, we have: $(G/G_W)/T \simeq G_{\overline{K}}$.

Remark 3. For q prime, $q \neq p$ and $n \in \mathbb{N}^\star$, any cyclic finite extension of K of degree q^n, if it exists, corresponds to a quotient of $\mathrm{gal}\,(K^{sep}/K)/G_W$ that looks like $\mathbb{Z}/q^n\mathbb{Z}$. Indeed, if L/K is cyclic of degree q^n, then $[K_W L : K_W]$ has degree q^m with $m \neq n$. G_W being pro-p-group, $m = 0$, so $L \subset K_W$ hence $\mathrm{gal}\,(L/K)$ is a quotient of $\mathrm{gal}\,(K^{sep}/K)/G_W$.

2.3. Pro-solvability, pro-cyclicity and solvability

2.3.1. When is the absolute Galois group prosolvable?

Remark 4. The absolute Galois group of any Henselian discrete-valued field need not be prosolvable in general. Indeed, it admits a canonical surjection onto the absolute Galois group of the residue field given by the action on the maximal unramified extension, so if the latter is not prosolvable, the former cannot be either. See the following example.

Example 1. The absolute Galois group of $\mathbb{Q}$ is a quotient of the absolute Galois group of $\mathbb{Q}((X))$. The first is not prosolvable so, neither is the last. More generally, if K is any Henselian discrete-valued field, then the maximal unramified extension K^{unr} of K has a Galois group $\mathrm{gal}\,(K^{unr}/K)$ isomorphic to the absolute Galois group of $\overline{K}$ (i.e. $\mathrm{gal}\,(K^{unr}/K) \simeq \mathrm{gal}\,(\overline{K}^{sep}/\overline{K})$), since $\mathrm{gal}\,(K^{unr}/K)$ is a quotient of $\mathrm{gal}\,(K^{sep}/K)$ then so is $\mathrm{gal}\,(\overline{K}^{sep}/\overline{K})$.

More precisely, we have the following result.

Theorem 2. *For any Henselian discrete-valued field,*

- *the absolute wild ramification group and all wild ramification subgroups are always pronilpotent. Meanwhile,*
- *its absolute Galois group is prosolvable if and only if this is true for the absolute Galois group of the residue field.*

P r o o f. Indeed, in all cases G_W (maybe trivial) is a closed normal pro-p-subgroup of the absolute Galois group $G = \mathrm{gal}\,(K^{sep}/K)$ and is then pronilpotent. See § 2.1.1.

Consider first the case of a positive residual characteristic $p > 0$.

Denote by $(v((K^{sep})^\star)^p/v((K)^\star))$ the p-free part of the abelian torsion group $(v((K^{sep})^\star)/v((K)^\star))$ (a quotient group of $\mathbb{Q}$), then we have the exact sequence see [18]:

$$1 \to (v((K^{sep})^\star)^p/v((K)^\star))^\vee \to \mathrm{gal}\,(K_W/K) \to \mathrm{gal}\,(\overline{K}^{sep}/\overline{K}) \to 1,$$

where, $(v((K^{sep})^\star)^p/v((K)^\star))^\vee$ is the dual of $(v((K^{sep})^\star)^p/v((K)^\star))$ in the sense that is the full character group of

$$(v((K^{sep})^\star)^p/v((K)^\star)) \text{ i.e.}$$
$$(v((K^{sep})^\star)^p/v((K)^\star))^\vee = Hom((v((K^{sep})^\star)^p/v((K)^\star)), \overline{K}^{sep}).$$

In consequence we have that,

$$\mathrm{gal}\,(K_W/K) \text{ is an extension of } \mathrm{gal}\,(\overline{K}^{sep}/\overline{K}) \text{ by}$$
$$(v((K^{sep})^\star)^p/v((K)^\star))^\vee \simeq \mathrm{gal}\,(K_W/K^{unr}) = G_0/G_W \simeq \textstyle\prod_{r\neq p} \mathbb{Z}_r(1).$$

It follows that all its Sylow subgroups are normal. Then the results follow.

Furthermore, $G/G_0 \simeq G_{\overline{K}}$ and $(G/G_W)/T \simeq G_{\overline{K}}$; T the absolute "tame-inertia" subgroup. See § 2.2.1. Finally, we get that $\mathrm{gal}\,(K^{sep}/K)$ is prosolvable if and only if this is true for $G_{\overline{K}}$ the absolute Galois group of the residue field.

Consider now the case when the characteristic is zero. It still holds, indeed, G_W is trivial, so $G_0 \simeq \prod_{r\neq p} \mathbb{Z}_r(1)$. □

Remark 5. It is worthy to notice that

- G_0 need not be pronilpotent. Indeed, the tame quotient can act by a non trivial outer automorphisms on the wild subgroup.
- The equivalence in Theorem 2 concerns the prosolvability only but not the solvability. Of course take for example p-adic field $\mathbb{Q}_p$ its absolute Galois group is prosolvable (but not solvable) since every finite Galois extension of it is solvable see Proposition 1, meanwhile the absolute Galois group of its residue field, $\mathbb{F}_p$, is procyclic.
- Since, any finite quotient of a pronilpotent profinite group is nilpotent. In general, the absolute Galois group of any Henselian discrete-valued field need not be pronilpotent. See Theorem 3.

Theorem 3. *Every finite normal totally ramified extension of $\mathbb{Q}_p$ for p being an odd prime number is either cyclic or nonnilpotent. Moreover if the extension is wildly ramified, then it is cyclic.*

P r o o f. Consider such extension $K/\mathbb{Q}_p$ with the Galois group G. Suppose first that G is a p-group and let $\Phi(G)$ be its Frattini subgroup. Since $G/\Phi(G)$ is an elementary abelian p-group thus the group $G/\Phi(G)$ is cyclic and therefore from a property of Frattini subgroups G is itself cyclic. Now let the group G be nilpotent, then it is the direct sum of its Sylow subgroups. Consequently $G = G_1 \times R$ where G_1 is a p-group and the order of R is prime to p. Remark that G_1 is the ramification group of $K/\mathbb{Q}_p$. Since $K^R/\mathbb{Q}_p$ is a normal totally ramified extension (K^R the fixed field by the elements of R) and its Galois group $G_{K^R/\mathbb{Q}_p} = G_{K/Q_p}/G_{K/K^R} = G/R = G_1$ is a p-group it follows from above that G_1 is cyclic. Since for $p \neq 2$ every normal totally and tamely ramified extension K/Q_p is cyclic of degree dividing $p-1$, furthermore with $M = K^{G_1}$ we get that the group $G_{M/\mathbb{Q}_p} = G_{K^{G_1}/\mathbb{Q}_p} = G/G_1 = R$ is cyclic of order prime to p. Consequently the group $G = G_1 \times R$ is cyclic. □

2.3.2. When is the absolute Galois group procyclic?

Here we prove the converse of Proposition IV.2.8 in [18].

Theorem 4. *For a complete discrete-valued field K, the absolute Galois group is isomorphic to $\widehat{\mathbb{Z}}$ if and only if the residue field $\overline{K}$ of K is algebraically closed and is of characteristic* 0.

P r o o f. If $\mathrm{char}\,(\overline{K}) = p > 0$ then the structure of the inertia group is not commutative since it has non-Galois separable finite extensions (discreteness of the valuation bounds the amount of p-power roots of unity in the maximal unramified extension when $\mathrm{char}\,(K) = 0$, so p^n-th root extractions of a uniformizer will be non-Galois for large n; in characteristic p one can use Artin-Schreier extensions of some tamely ramified extensions to make non-Galois extensions). So if the Galois group is commutative then $\mathrm{char}\,, (\overline{K}) = 0$, so by completeness the field must be $K = \overline{K}((T))$ for a field $\overline{K}$ of characteristic 0, and then the Galois group is an extension of $\mathrm{gal}\,(\overline{K}^{sep}/\overline{K})$ by $\widehat{\mathbb{Z}}$, but this $\widehat{\mathbb{Z}}$ being a closed subgroup of $\widehat{\mathbb{Z}}$ can only happen in case of equality, so can only happen when $\overline{K}^{sep} = \overline{K}$, which is to say $\overline{K}$ is algebraically closed. Nothe that the necessary condition is proved in Proposition IV.2.8 in [18]. In such case the absolute inertia subgroup equals the absolute Galois group, $G_0 = G$. □

2.3.3. When is the absolute Galois group solvable?

If a profinite group G is solvable then it is prosolvable, the converse is not true. Of course prosolvable does not mean that $G^{(n)} = \{1\}$ for some finite n (i.e. the derived length of G is finite, $G^{(n)}$ being the n-th commutator subgroup of G), but it only means that the series $G^{(n)}$ of higher commutator groups converges to $\{1\}$, i.e. every neighbourhood of $\{1\}$ contains almost all higher commutator subgroups.

For K being any CDVF, with the current notations, $G_W = \mathrm{gal}\,(K^{sep}/K_W)$ is the absolute wild ramification group, maybe trivial, otherwise, it is a free pro-p-group of infinite rank, where p is the residual characteristic. It, is prosolvable, pronilpotent, but in general not solvable. By Corollary 5, we have G/G_W is metabelian if and only if the absolute Galois group of the residue field of K, $\overline{G} = \mathrm{gal}\,(\overline{K^{sep}}/\overline{K})$ is too.

We have also the following properties.

1. If $\mathrm{char}\,(\overline{K}) = p$ and $\mathrm{char}\,(K) = p > 0$.
   Since a free pro-p-group is either isomorphic to $\widehat{\mathbb{Z}}$ when it is of rank 1 otherwise it is non-solvable, $G(p)$ being the biggest quotient of G which is a pro-p-group. For more details see § 2.5 is then non-solvable, neither is G; (indeed, $G(p)$ is a factor group of it). So, we get the following result.

   Let K be any CDVF of characteristic $p > 0$ the residue field being not algebraically closed i.e. $\overline{G}$ is not trivial, (with no further assumption on the residue field). Then the absolute Galois group of K is not solvable.

2. If $\mathrm{char}\,(\overline{K}) = p$ and $\mathrm{char}\,(K) = 0$.
   Then G_W the absolute wild ramification group is not trivial and a free pro-p-group of infinite rank. Therefore, G_W is not solvable. And consequently G is too.

So, we have the recapitulative result:

Proposition 1. *With the current notations, for K any CDVF regardless of its characteristic, if* $\mathrm{char}\,(\overline{K}) = p$, *then if the absolute Galois group is not trivial it is then not solvable as an abstract group.*

Now, let us prove a nice and necessary result on profinite groups.

Proposition 2. *Let N be an abelian profinite group whose automorphisms group $\mathcal{A}ut(N)$ being abelian profinite too. Consider the profinite group (semi-direct product) $G = N \rtimes H$. Then we have:*
1. If H is metabelian then G' (derived group) is abelian (i.e. G is metabelian too). Consequently:
2. G is metabelian if and only if H is too.

P r o o f. 1. Let $K = C_{H(N)}$, the centralizer of N within H (the set of elements in H that commute with every element of N, in the semi-direct product). As the action of H on N by automorphisms is given by a homomorphism $H \to \mathcal{A}ut(N)$ the kernel of which is K so H/K embeds in $\mathcal{A}ut(N)$, and as $\mathcal{A}ut(N)$ is abelian, H/K is abelian as well. In other words, K contains H' the group generated by the commutators of H, so H' centralizes N. Furthermore, since H is metabelian then H' is commutative, knowing that, $G' = N \rtimes H'$, we get $N \rtimes H' = N \times H'$ is commutative. Hence, G' is abelian.
2. Consequently G/G' is commutative. Conversely, if G is metabelian then H is too. □

Remark 6. Two important remarks are worthy to be noticed:

1. For N a profinite group $\mathcal{A}ut(N)$ need not be profinite, see Example 4.4.6 in [14].
2. If $\mathcal{A}ut(N)$ is abelian, N need not be abelian too, even when N is a finite group. (There are nonabelian finite p-groups for each prime p such that the automorphism groups are abelian see [8].)

Now, from Proposition 2, and since $\mathcal{A}ut(\prod_{q\neq p} Z_q)$ is abelian, we get the following result.

Corollary 5. *Let K be any CDVF of characteristic $p > 0$ with no assumption on $\overline{K}$, G being the absolute Galois group and G_W the absolute (wild) ramification subgroup of G. Then, G/G_W is metabelian if and only if the absolute Galois group of $\overline{K}$, $\overline{G}$ is too.*

2.4. Recapitulation

Let K being any CDVF with no special assumption on $\overline{K}$.

1. Let $\mathrm{char}\,(K) = p > 0$. The absolute Galois group of K is not solvable, see Remark 7.
2. Let $\mathrm{char}\,(K) = 0$ and $\mathrm{char}\,(\overline{K}) = p$. We have wild ramification, so a non trivial G_W which is not solvable, neither is the absolute Galois group is not solvable as an abstract group, see Proposition 1.
3. Let $\mathrm{char}\,(K) = 0$ and $\mathrm{char}\,(\overline{K}) = 0$. There is no wild ramification, so the subgroup G_W is trivial, and the absolute inertia group $G_0 \simeq \widehat{\mathbb{Z}}$. Now since the absolute Galois group is isomorphic to a semi-direct product of G_0 by $\mathrm{gal}\,(\overline{K}^{sep}/\overline{K})$ i.e. $\mathrm{gal}\,(K^{sep}/K) \simeq \widehat{\mathbb{Z}} \rtimes \mathrm{gal}\,(\overline{K}^{sep}/\overline{K})$. We may have the three following cases:
   (a) If $\overline{K}$ is algebraically closed then, $\mathrm{gal}\,(\overline{K}^{sep}/\overline{K})$ is trivial. So, $\mathrm{gal}\,((\mathbb{K})^{sep}/\mathbb{K}) \simeq \widehat{\mathbb{Z}}$ it procyclic hence abelian, see Theorem 4.
   (b) If $\overline{K}$ is not algebraically closed but can be endowed with a structure of C.D.V.F with residual characteristic $p > 0$. we still have the non solvability straightforwardly with respect to Proposition 1. (Particularly if the field K is a High dimensional local field).
   (c) The only case that remains to study, is that when $\overline{K}$ is not algebraically closed and cannot be endowed with a structure of C.D.V.F with residual characteristic $p > 0$. Note that, for example, if the residue field is $\mathbb{Q}$ it is clearly not solvable, whereas if the residue field is the fixed field of a single element from the absolute Galois group of $\mathbb{Q}$ then it is solvable. (For more details on the solvable profinite groups occurring as absolute Galois groups see [9].)

Question 1. A question that is staring immediately in the face is: "*Is an absolute Galois group either procyclic or else nonabelian?*"

But the answer is surprisingly simple, it is negative! See the following Example 2:

Example 2. Take the field $K = \mathbb{C}((X))((Y))$ with $\mathbb{C}$ the field of complex numbers. It is Henselian according to the discrete Y-adic valuation, (the residue field being $\mathbb{C}((X))$).But the absolute Galois group G of K is the direct product of two copies of $\widehat{\mathbb{Z}}$, $G \simeq \widehat{\mathbb{Z}} \times \widehat{\mathbb{Z}}$, hence abelian but non procyclic.

Note 1. With respect to our study, the result in [9]: "*For any commutative field, if the absolute Galois group is solvable then it is metabelian,*" turns out to be more relevant in global case then for CDVF, except in the single case when $\mathrm{char}\,(K) = 0$ and $\mathrm{char}\,(\overline{K}) = 0$ and no structure of CDVF with residual characteristic $p > 0$, can be defined on $\overline{K}$.

2.5. On the p-maximal extension

For details see § 2.2.1.

First case same characteristic. *Here let us assume that* $\operatorname{char}\overline{K} = p > 0$.

Let K be any CDVF of characteristic $p > 0$ with no special assumption on $\overline{K}$, the residue field of K. Write $G(p)$ for the biggest quotient of G which is a pro-p-group. $G(p)$ is the Galois group of the maximal p-extension $K(p)/K$ i.e. the compositum of all Galois extensions of p-power order. It is a free pro-p-group of rank > 1, see [19, Chap. II., § 2.2, Corollary 1, p. 75], (i.e. $G(p)$ is the profinite completion of a free group with respect to a system of normal subgroups the quotients of which are finite p-groups) such that $H^1(G(p))$ can be identified with $K/\wp(K)$ (where $\wp : x \mapsto x^p - x$) which is a vector space of infinite dimension over $\mathbb{F}_p$ (the field of p elements), since the powers T^n (with n ranging over $\mathbb{N}$ and prime to p, T being a prime element in the DVF) are linearly independent over $\mathbb{F}_p$.
First let us recall the following well known results:

Proposition 3. *Let* $L = K((t))$ *(Laurent Series field) with* $\operatorname{char}(K) = p > 0$, $K(p)/K$ *being the maximal* p*-extension (compositum of all Galois extensions of* p*-power order), then:*

- *If* K *is finite or countable then* $G(p) = \operatorname{gal}(K(p)/K)$ *is a free pro-*p*-group of countably infinite rank,*
- *If* K *is uncountable then* $G(p) = \operatorname{gal}(K(p)/K)$ *is a free pro-*p*-groups of uncountable rank* (*see* [12, Proposition 6.1.7]).

In other words and in classical case more precisely for any local field K with finite residue field we have:

Theorem 5. *Let* K *be any local field with finite residue field* $\overline{K}$*, let* $\operatorname{char}(\overline{K}) = p$*, then* $G(p)$ *as well as* G_W *(the wild ramification group) are free pro-*p*-groups of countably infinite rank.*

P r o o f. See Proposition 7.5.1 and [12, Theorem 7.5.10]. □

Remark 7. Since a free pro-p-group is either isomorphic to $\widehat{\mathbb{Z}}$ when it is of rank 1 otherwise it is non-solvable. Then $G(p)$, being a free pro-p-group of rank > 1, is non-solvable, neither is G as $G(p)$ is a factor group of it.

So, we get the following result:

Theorem 6. *Let* K *be any complete discrete-valued field of characteristic* $p > 0$ *with no assumption on the residue field. Then the absolute Galois group of* K *is not solvable.*

Remarks on the q-maximal extension with $q \neq \operatorname{char}(K)$: K being some field containing a q-th root of unity, q being an odd prime number and different from the characteristic of K. Write $G(q)$ for the Galois group of the q-maximal extension of K, and assume that $G(q)$ has a finite normal series with abelian factor groups (i.e. solvable). Then the derived subgroup $G(q)'$ of $G(q)$ is abelian, moreover, $G(q)$ has a normal abelian subgroup with a pro-cyclic factor group. Furthermore, we have the following result:

Theorem 7 [22]. *Under the current hypotheses and notations the following statements are equivalent:*

- $G(q)$ *is solvable.*

- *$G(q)$ is metabelian.*
- *$G(q)$ does not contain a free non-abelian subgroup.*

Now, G_W is a pro-p-group therefore, the absolute Galois group of K is prosolvabe if and only if $\mathrm{gal}\,(K_W/K)$ is too (a pro-p-group is pro-nilpotent but need not be solvable). See [5].

Second case: mixed characteristic. In this case, Safarevič in [15] showed that for $K/\mathbb{Q}_p$ an extension of degree n not containing the p-th roots of unity and if $K/\mathbb{Q}_p$ is finite of degree $n < +\infty$ then $G(p)$ is a free pro-p-group of rank $n+1$. Now if K contains μ_p (the group of the p-th roots of unity) then $G(p)$ is a Poincare group of dimension 2 that is a Demuskin group of rank $n+2$. See Theorem 7.5.11 in [12]. So, in both cases if $K/\mathbb{Q}_p$ is finite of degree $n < +\infty$ then the absolute Galois group G of K can be generated by $n+2$ elements. See Theorem 7.4.1 in [12].

Furthermore, we have the following:

- By local class field theory, the abelianized group $G(p)/G(p)'$ is isomorphic to the pro-p-completion of $K^\star$ hence it is isomorphic to $U_K^1 \times \mathbb{Z}_p$ which is not procyclic, of course U_K^1, the subgroup of 1-units in $K^\star$ is not procyclic, but it is free abelian for $p > 2$ and K not containing the p-th roots of unity.
- Any complete discrete-valued field of residue characteristic $p > 0$ has an (unramified) procyclic extension $K(p')$ generated over K by all the ℓ-th roots of unity for ℓ describing all natural integers not divisible by p, thanks to Hensel's Lemma. Of course, from Galois theory of finite fields, by adjoining such roots of unity at residual level is obtained from doing so over the prime subfield $\mathbb{F}_p$ of the residue field $\overline{K}$. For more details see § 3.4.

Note that the unramified extension $K(p')/K$ maybe trivial. For example if k is algebraically closed of characteristic p, then $k((t))$ has no unramified extension.

It is worthy to notice the following result:

Lemma 1. *In case if G has $G(p)$ as free pro-p-group of "$1 < rank \leq +\infty$" with $(\mathrm{char}\,(\overline{K}) = p)$, we can add that G is a semi-direct product of $\mathrm{gal}\,(K^{sep}/K(p))$ by a subgroup isomorphic to $G(p)$.*

P r o o f. Indeed, according to Theorem 7.7.4 in [14] "$G(p)$ is a free pro-p-group if and only if $G(p)$ is projective group (in the category of profinite groups)", that is it has the lifting property for every extension, which is equivalent to say that for every surjective morphism from any profinite group $H \to G(p)$ there is a section (a right inverse of the morphism in question) $G(p) \to H$. So, if f is an epimorphism from G onto $G(p)$ by the projectivity of $G(p)$ there exists a homomorphism h from $G(p)$ to G such that fh is the identity map on $G(p)$. Hence, G is a semi-direct product $\ker(f)$ and $h(G(p))$ (which is isomorphic to $G(p)$). □

2.6. On the maximal unramified extension

Let K be any complete discrete-valued field of residue characteristic $p > 0$ with $\overline{K}$ being the residue field of K, write $\overline{K}^{sep} = \mathcal{O}_{K^{unr}}/\mathcal{M}_{K^{unr}}$, it is a separable closure of $\overline{K}$; (K^{unr} being the maximal unramified extension of K that is the composite of all unramified extensions inside an algebraic closure of K).

From [13, ch. II, § 7] in the general case that is when K is assumed to be Henselian only K^{unr} contains all roots of unity of order m not divisible by the residue characteristic, because the separable polynomial $X^m - 1$ splits over the separable closure of the residue field of K, and hence also over the maximal unramified extension K^{unr} of K, by Hensel's Lemma. Now write $K(p')$ for

the (unramified) pro-cyclic extension of K generated by all the ℓ-th roots of unity for ℓ describing all natural integers not divisible by p, it contains a subextension $K(p'')$ that is generated over K by all the q-th roots of unity for q describing all the primes different from p.

The question remains to prove that $K(p') = K(p'')$.

First, notice that the question is certainly a question of residue fields, par excellence.

Consider, the largest finite field contained in $\overline{K}$, $\mathbb{F}_\ell$ (the finite field of ℓ elements) where ℓ is power of p. Since the finite field $\mathbb{F}_\ell$ consists of the $(\ell-1)$-th roots of unity and 0, the said roots of unity are contained in $\overline{K}$. Now, if a complete discrete valuation field has residue field containing $\mathbb{F}_\ell$, then K contains the $(\ell-1)$–th roots of unity (Hensel's Lemma). This is an if and only if statement.

Now, we can say that $\overline{K}(p')$ (defined as above) is included in the residue field of $K(p')/K$ and then $\mathbb{F}_\ell(p')$ is included in the residue field of $K(p')/K$ too, as well as $\overline{K}(p'')$.

(Note that $\mathbb{F}_\ell(p')$ and $\mathbb{F}_\ell(p'')$ are no more finite, but infinite fields of characteristic $p > 0$.)

In other words, we can replace K by $\mathbb{F}_\ell$.

Hence, if we prove that: $\mathbb{F}_\ell(p'')$ is an algebraic closure of $\mathbb{F}_\ell$. Then we get that $\mathbb{F}_\ell(p') = \mathbb{F}_\ell(p'')$ and consequently that $K(p') = K(p'')$.

Since to get a primitive f-th root of unity in a field is equivalent to getting primitive roots of unity of order equal to each prime-power factor of f, our question amounts to asking if for a given prime p and prime power ℓ^r (allowing $\ell = p$), does there exist a square-free n not divisible by p such that p mod n has order divisible by ℓ^r (so then adjoining a primitive n-th root of unity to $\mathbb{F}_p$ would give an extension of degree divisible by ℓ^r, and then do this for several such prime powers to get an extension of $\mathbb{F}_p$ generated by prime-order roots of unity such that its degree is divisible by whatever we want).

But if $(\mathbb{Z}/n\mathbb{Z})^\star$ is going to contain a cyclic subgroup of order ℓ^r then under the decomposition $(\mathbb{Z}/n\mathbb{Z})^\star = \prod(\mathbb{Z}/q_i\mathbb{Z})^\star$ for the prime factors q_i of the square-free n we see that one of the projections $(\mathbb{Z}/n\mathbb{Z})^\star \to (\mathbb{Z}/q_i\mathbb{Z})^\star$ is injective on that cyclic subgroup of order ℓ^r. Hence, if some such n is going to exist then even a prime n will have to exist which does the job. In other words, the question is exactly asking this:

Given a prime p and a prime power ℓ^r (allowing $\ell = p$), does there exist a prime q distinct from p such that p mod q has order divisible by ℓ^r?

Since $(\mathbb{Z}/q\mathbb{Z})^\star$ is cyclic, the only way it contains an element with order divisible by ℓ^r is if the size of this cyclic group is divisible by ℓ^r, which is to say $q = 1 \bmod \ell^r$.

Lemma 2. *Let p be prime. To generate an algebraic closure of $\mathbb{F}_p$ it is enough to adjoin the q-th roots of unity for all prime q different from p.*

Here we must use Čebotarev's Theorem (see, [13, ch. VII, § 13, Theorem 13.4]). Indeed, Čebotarev density Theorem reduces the problem of classifying Galois extensions to that of describing the splitting of primes in extensions. Specifically, it implies that as a Galois extension of K, L is uniquely determined by the set of primes of K that split completely in it. A related corollary is that if almost all prime ideals of K split completely in L, then in fact $L = K$.

P r o o f of Lemma 2. By a simple application of non-abelian Čebotarev result, it is enough to settle that "For (possibly equal) primes p and ℓ and any integer $r > 0$ that there are lots of primes $q = 1 \bmod \ell^r$ such that p mod q has order divisible by ℓ^r", (e.g., lots of $q = 1 \bmod 9$ such that 5 mod q has order divisible by 9). Since $(\mathbb{Z}/q\mathbb{Z})^\star$ is cyclic with size divisible by ℓ^r, a sufficient condition for an element to have order divisible by ℓ^r is that it "not" be an ℓ-th power. So one way to ensure that p mod q has order divisible by ℓ^r is to make sure that p mod q is not an ℓ-th power.

So consider the non-abelian Galois extension $K = \mathbb{Q}(\zeta_\ell^r, p^{1/\ell})$ of $\mathbb{Q}$. We have $\mathrm{gal}\,(K/\mathbb{Q}) \to \mathrm{gal}\,(\mathbb{Q}(\zeta_{\ell^r})/\mathbb{Q}) = (\mathbb{Z}/\ell^r\mathbb{Z})^\star$ carrying a Frobenius element $Frob_q$ onto q mod ℓ^r, hence Čhebotarev

provides many q such that $Frob_q$ is nontrivial but $q = 1 \bmod \ell^r$. For any such q, not only is q totally split in $\mathbb{Q}(\zeta_{\ell^r})$ but the extension give by adjoining an ℓ-th root of p is "non-trivial " over $\mathbb{F}_q = (\mathbb{Z}/q\mathbb{Z})^\star$. Hence, $X^\ell - p$ has no root in $\mathbb{F}_q$ (since if it has one root then it completely splits, as $\mathbb{F}_q$ contains a primitive ℓ -th root of 1 by design).

Applying this with a fixed p but several ℓ^rs (for different ℓs) and considering pairwise distinct q s thereby obtained, it follows that every finite extension of $\mathbb{F}_p$ is contained in an extension generated by prime-order roots of unity, that is exactly what we wish. □

Also, we have the following result:

Proposition 4. *Let K be any complete discrete-valued field of residue characteristic $p > 0$ with no more assumption on the residue field, then $K(p') = K(p'')$, namely the (unramified) procyclic extension of K generated by all the ℓ-th roots of unity for ℓ describing all natural integers not divisible by p equals the last one generated by all the q-th roots of unity for q describing all the primes different from p .*

P r o o f. The proposition follows from Lemma 2 immediately. □

Remark that if $\overline{K}$ is finite then $K^{unr} = K(p')$ (see [13, ch. II, § 7]). So, we have:

Corollary 6. *Let K be any complete discrete-valued field with a finite residue field of characteristic $p > 0$, then the maximal unramified extension of K is the extension generated over K by all the q-th roots of unity for all prime q different from p.*

Notice that Corollary 6 above is no more true if $\overline{K}$ is not finite, indeed:

Example 3. If $k = \mathbb{F}_p(u)$ with u transcendent on $\mathbb{F}_p$ (the field of p elements) and $K = k((x))$, then $K(v)$ with $v^n = u$ is an unramified extension of K (in the sense that $e = 1$, and in the strict sense if p does not divide n). Obviously, $K(v)$ cannot be generated by a root of unity.

3. On Abhyankar's Lemma

The aim of this section, is the proof of Theorem 9 that is "some" generalization of Abhyankar's Lemma in local case, by use of the following EPP's Theorem 8 (see [6]).

First, let us recall both the Abhyankar Lemma [7] and EPP Theorem[1].

Lemma 3 (Abhyankar, [7]). *Let $L = L_1L_2$ be the compositum of two finite algebraic extension fields of K, let $\mathcal{P}$ be prime divisor of L, which is ramified in L_i/K of order e_i $(i = 1, 2)$; then if $e_2|e_1$ and $\mathcal{P}$ is tame in L_2/K, then $\mathcal{P}$ is unramified in L/L_1.*

Theorem 8. (EPP[2]) *Let L/K be any non-trivial finite extension of discretely valued fields, it is possible to eliminate wild ramification, that is to ensure that $e_{k'L/k'K} = 1$ for some finite extension k'/k, where k is a "constant[3] subfield".*

Now, our generalization of Abhyankar's Lemma in local case can be announced as follows.

Theorem 9. *Given any finite Galois extension L/K of complete discrete-valued fields with a non necessarily perfect residue field of characteristic $p > 0$. Then there exist two separable overextensions K' and M of K such that:*

[1] EPP Theorem is an existence theorem of a reduced extension but non-constructive.

[2] Worthy to note that in [10] F.V. Kuhlmann has corrected an error in the proof of Theorem 8 of EPP's article [6]. Happily the error does not hurt any of the wording of all results in the said article.

[3] A subfield k of K is said to be constant, if it is a maximal subfield of K having a perfect residue field. Note that, such k is canonical in the mixed characteristic case).

- $K \subset K' \subsetneq M \subseteq LK'$,
- *LK'/K' is weakly unramified, so a uniformizer in K' remains uniformizer in LK'.*
- *M/K' is unramified.*
- *LK'/M is ferociously ramified, then the Galois group* $\mathrm{gal}\,(LK'/M)$ *is a p-group.*

P r o o f. First, according to Theorem 8, there exists a finite extension K'/K such that LK'/K' is weakly unramified, therefore $[LK' : K'] = [\overline{LK'} : \overline{K'}]$, i.e. $e = 1$ and $f = [LK' : K']$ (where $\overline{K'}$ is the residue field of K'). This condition implies that a uniformizer of K' remains uniformizer in LK', but the residue extension can be inseparable, furthermore it is not evident that Epp's extension K'/K is separable.

Let M be the maximal unramified (i.e. etale) extension of K' that is contained in LK'.
Characterization of M: The ramification index $e_{(M/K')} = 1$, the residue extension of M/K' is separable so that M/K' is unramified, but the residue extension of LK'/M is purely inseparable (if LK'/K' is not unramified, see Remark 8).

Note that, if K has characteristic zero, then we can certainly take K'/K Galois, because if K'/K is not Galois, then we can always use its Galois closure instead.

Let T be the maximal tamely ramified subextension of LK'/K'. Characterization of T (see § 5.3): $e_{(T/K')}$ is prime to p, $e_{(LK'/T)}$ is a power of p, the residue extension of T/K' is separable, and the residue extension of LK'/T is purely inseparable. Hence if $e_{(LK'/K')} = 1$, we have $T = M$ and $[LK' : M]$ is a power of p.

Indeed, more precisely, in case of K'/K is separable LK'/M is then weakly unramified and the residue field extension is purely inseparable i.e. LK'/M is ferociously ramified (if it is not trivial). $[LK' : M] = [\overline{LK'} : \overline{M}]_{insep}$. In such case the inertia group of LK'/M is the full Galois group of LK'/M, and this group is a p-group.
In case of K'/K is purely inseparable, LK'/K' being weakly unramified, then it cannot be the case that the inertia group of L/K has a prime-to-p part, as tame ramification cannot be eliminated by an inseparable extension, in other words if the tame ramification index $e_{tame} > 1$, and if K'/K is a purely inseparable extension, then LK'/K has the same tame ramification index, so it cannot be weakly unramified, this follows from the multiplicativity of the tame ramification index. So, assuming LK'/K' is weakly unramified, then it is true that LK'/M is ferociously ramified. The proposition follows. □

Remark 8. When considering the particular case of perfect residue fields with L/K tamely ramified we get $M = LK'$, that is the Abhyankar's Lemma.

Remark 9. [26, § 1] If furthermore, we assume the hypothesis $[\overline{K} : \overline{K}^p] = p$ (i.e. $\overline{K}$ has a p basis of length 1), we get that LK'/K' is well ramified and then monogenic.

Remark 10. The usefulness of Theorem 9 is alluded to in the construction of a translated weakly unramified extension that is decomposable in an unramified and a ferociously ramified extensions. Worthy to note that such extensions arise in some situations in algebraic geometry. They are almost as important as selected in algebraic setting. For example, the book [4] which considers local extensions of discrete-valued rings having $e = 1$ in the more general case, such situations are called there as with "ramification index 1".

In a similar question of ours, Abbes and Saito proved the following different Corollary, see [1, Corollary A.2, p. 31]. *However, in their result they eliminate the fierce extension and allow to get an unfiercely ramified extension. They use the term* ***unfiercely*** *ramified for the case of finite separable extensions with separable residue extensions.*

Corollary 7 (A.2)(Abbes–Saito). *Given any finite separable extension of complete discrete-valued fields L/K, there exists a tower $K \subseteq K' \subseteq LK'$, K'/K finite separable such that*
- *a uniformizing element of K remains uniformizing element in K';*
- *LK'/K' is unfiercely ramified.*

4. Questions in the limelight in the general case

In this section some important and still open questions, that can make a fruitful subject of research, are given:

- How to completely specify the extensions having $e_{wild} > 1$ and $f_{insep} > 1$ for which there exists a normal subgroup that can "separate" ferocious from wild ramification? Note that some steps have been already done by L. Spriano, but the question is still very far from being entirely solved.

- In his paper [17, Corollary 1.3.4, p. 790] (in the equal characteristic case) and also in [16, Theorem 2, p. 568] (in the mixed characteristic case), Saito considered the following natural injective map the refined Swan conductor homomorphism ("rsw" initially defined by Kato) from the graded quotients piece of the Abbes–Saito filtration into the differentials. To be more precise, in the general case we have

$$rsw : Hom(G^r_{K,log}/G^{r+}_{K,log}, \mathbb{F}_p) \to \Omega^1_{\mathcal{O}_K}(log) \otimes_{\mathcal{O}_K} \pi_K^{-r}\overline{K}^{sep}, \tag{4.1}$$

  where K is a complete field with respect to a discrete-valuation, the residue field $\overline{K}$ being not necessarily perfect, $\overline{K}^{sep}$ a separable closure of it, G_K is the absolute Galois group, $r \in \mathbb{Q}_{>0}$, $\mathcal{O}_K$ is the ring of integers of K, π_K is uniformizer and $\Omega^1_K(log)$ is the logarithmic differential.

  It is likely that the said map is also surjective, if the residue field is perfect. Of course, when the residue field $\overline{K}$ is perfect, the right hand side (target) is just a one-dimensional vector space over the separable closure $\overline{K}^{sep}$. But there is no canonical basis. So, (4.1) reduces to

$$rsw : Hom(G^r_{K,log}/G^{r+}_{K,log}, \mathbb{F}_p) \to \pi_K^{-r} \otimes \overline{K}^{sep}.$$

  We cannot say that, the right hand side $\pi_K^{-r} \otimes \overline{K}^{sep}$ is exactly the residual ring $\mathcal{M}^r_{K^{sep}}/\mathcal{M}^{r+1}_{K^{sep}}$ where $\mathcal{M}^r_{K^{sep}} = \{x \in K^{sep}, v_{K^{sep}}(x) \geq r\}$, and $\mathcal{M}^{(r+1)}_{K^{sep}} = \{x \in K^{sep}, v_{K^{sep}}(x) \geq r+1\}$ with $v_{K^{sep}}$ the extension of the normalized valuation v_K to K^{sep}) since K^{sep} is not discretely valued. It is more correct to write $\pi_K^{-r} \otimes \overline{K}^{sep}$ differently as $\mathcal{M}^r_{K^{sep}}/\bigcup_{\epsilon>0} \mathcal{M}^{r+\epsilon}_{K^{sep}}$. For a proof of this result in perfect residue field case and for $r \in \mathbb{Z}_{>0}$, it is used to make working some means of local class field theory, then the case $r \in \mathbb{Q}_{>0}$ follows from certain base change result. The p-adic differential modules being out of the frame of this study, this question will appear in a next work.

  I think we can conjecture that this map remains surjective, even when dropping the hypothesis of the perfectness of the residue field. I have been told that some experts have pinned down the exact image of the abelian part. I think if we can run a base change argument to reach the rest of differential forms on the target, as in the case of perfect residue field, the problem will be solved. Probably, one needs to avoid the case when p is absolutely unramified in a mixed characteristic field.

5. Annexe on Hilbert's theory in the general case

The transition from the classical to the general case, requires a recall of special notions. So, let us consult our notes on Zariski–Samuel filtration as well as on Abbes–Saito ramification filtration, where some subtle and essential differences between the general and the classical cases appear. Furthermore, some important remarks and some original examples and counterexamples are given.

5.1. Hilbert–Zariski–Samuel filtration

Let L/K be any finite Galois extension of local fields with no special assumption on the residual extension and G is its Galois group.

Indeed, following Hilbert's way, in [25, ch. V] Zariski and Samuel define their lower ramification subgroups filtration as follows.

Then for any positive integer $n \geq 1$, they define the n-th ramification group G_n as the subset of G consisting of all automorphisms $\sigma \in G$ such that $\sigma(x) \equiv x \mod \mathcal{M}_L^{n+1}$ for every $x \in \mathcal{O}_L$. G_n is the kernel of the action on $\mathcal{O}_L/\mathcal{M}_L^n$. They establish that G_n are invariant subgroups of G, and the quotients G_n/G_{n+1} are abelian for $n \geq 1$ [25, Lemma 1, p. 295]. Meanwhile, $G_0/G_1 (= G_T/G_{V_2}$ in Zariski–Samuel notation) need not be abelian in general case [25, ch. V, § 10, p. 297]. Indeed, there are extensions where $f_{insep} > 1$ and $e_{wild} > 1$, for which there does not exist a normal subgroup which can "separate" ferocious from wild ramification [20, § 1,page 1273]. So a second filtration H_n was necessary. By use of the homomorphism, we have

$$\begin{aligned} \lambda : \quad & G_0 \to \overline{L}^\star, \\ & \sigma \mapsto \lambda(\sigma) = \overline{(\sigma(\pi)/\pi)} = u_\sigma, \end{aligned}$$

H_1 is defined as the kernel of λ, that is the subgroup of all automorphisms σ in G_0 such that $u_\sigma \equiv 1 \mod \mathcal{M}_L$; that is such that $\sigma(\pi) - \pi \in \mathcal{M}_L^2$.

Likewise, H_i (for $i > 1$) is defined to be the kernel of the homomorphism

$$\begin{aligned} \lambda_i : \quad & G_i \to (\overline{L}, +), \\ & \sigma \mapsto \lambda_i(\sigma) = y_\sigma, \end{aligned}$$

that is the subgroup of all automorphisms σ in G_i such that $y_\sigma \equiv 0 \mod \mathcal{M}_L$, where y_σ is the integer $y_\sigma \in \mathcal{O}_L$ satisfying $\sigma(\pi) - \pi = y_\sigma \pi^i$ (i.e. $\sigma(\pi) - \pi \in \mathcal{M}_L^i$).

We have $G_i \supseteq H_i$ for every $i \geq 1$ (the equality occurs when the residue fields extension is separable, see [25, ch. V, § 10, p. 296]). So, $\sigma \in H_i$ implies that $\sigma(x) \equiv x \mod \mathcal{M}_L^{i+1}$ for every $x \in \mathcal{O}_L$. H_i is then the kernel of the action on $\mathcal{M}_L/\mathcal{M}_L^i$ for $i \geq 1$.

Intertwining both two filtrations of the Galois group with ramification groups, they used to define a unique filtration $G_{(n,i)}$ such that $G_n = G_{(n+1,0)}$ and $H_n = G_{(n,1)}$, as follows: for $n, i \in \mathbb{N}$ the (n,i)-ramification group $G_{(n,i)}$ of $G = \mathrm{gal}\,(L/K)$ is the subgroup of those K-automorphisms of L that induce the identity on $\mathcal{M}_L^i/\mathcal{M}_L^{n+i}$, i.e.

$$G_{(n,i)} = \{\sigma \in G; v_L(\sigma(x) - x) \geq i + n \forall x \in \mathcal{M}_L^i\} = \{\sigma \in G; \forall x \in \mathcal{M}_L^i; x - \sigma(x) \in \mathcal{M}_L^{n+i}\}.$$

Since $G_{(n,i)}$ is the kernel of the homomorphism $G \longrightarrow \mathcal{A}ut(\mathcal{M}_L^i/\mathcal{M}_L^{n+i})$ it is then a normal subgroup of G. Then we get, in the Zariski-Samuel filtration,

$$G_n = G_{(n+1,0)} \text{ and } H_n = G_{(n,1)}.$$

The $G_{(n,i)}$ with $i > 0$ makes sense, in the non-classical case only. Now, in the classical sense, *the G_n meet the usual ramification groups, see* [18]. *Explicitly, for $n \geq -1$ the n-th ("lower") ramification subgroup is defined as* $G_n = \{\sigma \in G; i_G(\sigma) \geq n + 1\}$.

Consequences for the classical case:

The usual ramification subgroups in the classical case, are $(H_n = G_n)_{n\geq 1}$, and G_1 is called ramification group. From this Serre in [18] obtained the upper filtration by use of the Hasse-Herbrand functions ϕ defined by:

$$\phi_{L/K}(x) = \int_0^x \frac{dt}{|G_0 : G_t|},$$

and its inverse ψ (remember that φ and ϕ are only defined in case when the residue extension is separable). The upper $(G^n)_{n\geq 1}$ is related to the lower filtration by the formula $(G_n = G^{\phi(n)})_{n\geq 1}$ and $(G^n = G_{\psi(n)})_{n\geq 1}$. Note that the upper one behaves well under quotient subgroups; meanwhile, the lower one behaves well when taking subgroups.

The i such that $G_i \neq G_{i+1}$ (resp. $G^i \neq G^{i+1}$) are called lower (resp. upper) breaks.

5.2. Outline of Abbes–Saito ramification filtration

Let L/K be a finite Galois extension of local fields, then respectively we will write G_K and G_L for the absolute Galois groups of K and L. It is worthy to note that a separable closure is not complete as valued field in general.Nevertheless, a filtration on the absolute Galois group can be defined by taking inverse limit, as well as breaks.

Indeed, using techniques of rigid geometry, A. Abbes and T. Saito in [1] defined two decreasing filtrations, the first $(G_K^a)_{a\in\mathbb{Q}_{\geq 0}}$ and the second by logarithmic ramification groups $(G^a_{log,K})_{a\in\mathbb{Q}_{>0}}$ (closed normal subgroups of G_K). The filtration coincides with the classical upper numbering ramification filtration shifted by one, if the residue field of K is perfect in the sense that $G_K^{a-1} = G^a_{log,K}$ agrees with the upper numbered ramification filtration labeled by a. It is noteworthy that the filtration is left continuous and their jumps are rational.

For a real number $a > 0$, they define G^{a+} to be the topological closure of $G_K^{a+} = \overline{\cup_{b>a} G_K^b}$ and $G_K^{a-} = \cap_{b<a} G_K^b$, where b denotes a rational number. Then the following holds,

- $G_K^{a-} = G_K^a$ if $a \in \mathbb{Q}$, and $G_K^{a-} = G_K^{a+}$ if a not in $\mathbb{Q}$. It holds for the logarithmic too.
- The two filtrations by ramification groups are related as follows:

Let $j > 0$ be a rational number, then we have the following inclusions $G_K^j \supset G^j_{K,log} \supset G_K^{j+1}$, see [1, Proposition 3.15]

- G_K^1 is the absolute inertia subgroup of G_K; and G_K^{1+} the absolute wild inertia group of G_K.
- From the filtration above they define for any Galois extension L over K, the ramification filtration of the Galois group $\mathrm{gal}\,(L/K)$ by $G_K^a/(G_K^a \cap G_L)$. As a consequence, in the more general case, we have:
- $G_K^1/(G_K^1 \cap G_L)$ is the inertia subgroup of $\mathrm{gal}\,(L/K)$.
- $G_K^{1+}/(G_K^{1+} \cap G_L)$ is the wild inertia subgroup of $\mathrm{gal}\,(L/K)$.
- $\#(G_K^{1+}/(G_K^{1+} \cap G_L)) = e_{wild} f_{insep}$.
- If L/K finite unramified extension then $G_K^a = G_L^a$.
- If L/K finite tamely ramified extension with ramification index m then $G^{ma}_{log,L} = G^a_{log,K}$.

Furthermore, the logarithmic ramification filtration groups satisfy the following theorem [24, Theorem 3.7.3].

Theorem 10 [24]. *Assume that the residue field is of characteristic $p > 0$. Then the subquotients groups of the logarithmic ramification filtration $G^a_{log,K}/G^{a+}_{log,K}$ are abelian and annulated by p if $a \in \mathbb{Q}_{>0}$ and are trivial if a is irrational.*

Remark 11. It is worthy to note that we cannot make the filtration of Hilbert–Zariski–Samuel type and the last one of Abbes–Saito corresponding to each other, in a satisfactory way for example by use of some means like the well-known Hasse–Herbrand φ, ψ functions. Furthermore, the basic ramification degrees do not seem to work well as when the residue field fails to be perfect. Of course, the unramified part and the tame part are still okay, but it is not practical to separate the wild part from the residually inseparable part. Some attempts have been done, trying to describe ramification using more complex objects as ramification invariants. E.g. I.B. Zhukov used the "cutting-by-curves" method by considering the Abbes–Saito Swan conductor which is defined by looking at the generic points of the divisors. For details see [27] and [28], especially the results Theorems 2.2 and Theorems 2.4 in [27], and Remark 2.5.3 in [28]. But these notions are very far from our study.

5.3. Ramification cases

Consider a finite Galois extension L/K of local fields with Galois group $G = \mathrm{gal}\,(L/K)$, the residue extension $\overline{L}/\overline{K}$ being of characteristic $p > 0$ and not necessarily separable.

Write $K_{unr,L} = L \cap K^{unr}$ (for the maximal unramified extension of K in L i.e. the inertia field of L/K), and $G_0 = \mathrm{gal}\,(L/K_{unr,L})$ for the inertia group of L/K; so

$$G/G_0 = \mathrm{gal}\,(L/K)/\mathrm{gal}\,(L/K_{unr,L}) \simeq \mathrm{gal}\,(\overline{K}_{sep,\overline{L}}/\overline{K}),$$

where $\overline{K}_{sep,\overline{L}} = \overline{L} \cap \overline{K}^{sep}$, $\overline{K}^{sep}$ being a separable closure of the residue field $\overline{K}$.

Consider the ramification index e of the extension L/K, and f as its residue degree. Then we can write $e = e_{tame}.e_{wild}$ and $f = f_{sep}.f_{insep}$. So, we have

$$f_{sep} = \#(G/G_0) = [\overline{K}_{sep,\overline{L}} : \overline{K}] = [\overline{L} : \overline{K}]_{sep}; f_{insep} = [\overline{L} : \overline{K}_{sep,\overline{L}}] = [\overline{L} : \overline{K}]_{insep}.$$

L/K is *unramified* if f_{sep} is arbitrary and $f_{insep} = e = 1$.
L/K is *tamely ramified* if f_{sep} is arbitrary, e prime to p and $f_{insep} = 1$.
L/K is *completely ramified* if $f_{sep} = 1$, f_{insep} is arbitrary and e is a power of p.
L/K is *totally ramified* if $f_{sep} = f_{insep} = 1$ and e is arbitrary; in such case $\overline{L} = \overline{K}$.
L/K is *totally and wildly ramified* if $f_{sep} = f_{insep} = 1$ and e is a power of p.
L/K is *weakly unramified* if f_{sep}, f_{insep} are arbitrary and $e = 1$.
L/K is *ferociously ramified* or *fierce extension* if $f_{insep} > 1$ is arbitrary and $e = f_{sep} = 1$.

Note 2. "If L/K is fierce extension then it is weakly unramified, so that K contains a prime element of L."

5.4. Some well-known formulas and theorems (classical case)

$L = K(\alpha)/K$ being a finite Galois extension with Galois group G, the residue extension $\overline{L}/\overline{K}$ being separable of characteristic $p > 0$, we write f for the minimal polynomial of α.

Then we have the following useful summary of formulas and theorems, see for example [18, Ch. IV]. Meanwhile, for the general case, in [20, Examples 3.3, 3.4 and 8.1] beautiful counterexamples are given.

1. Hilbert's formula

$$v_L(\mathcal{D}_{L/K}) = \sum_{\sigma \neq I} i_G(\sigma) = \sum_{i \geq 0}(|G_i| - 1) = v_L(f'(\alpha)),$$

where $\mathcal{D}_{L/K}$ is the different, $|G_i|$ the order of the i-th lower ramification group, and i_G the function:

$$\begin{aligned} i_G : G &\to \mathbb{Z} \cup \{\infty\}, \\ \sigma &\mapsto i_G(\sigma) = \inf_{x \in \mathcal{O}_L^*} v_L(\sigma(x) - x) \quad \text{for} \quad \sigma \neq 1. \end{aligned}$$

2. Herbrand's theorem

Let L/K be a finite Galois extension and L'/K a Galois subextension. Write $G = \mathrm{gal}\,(L/K)$ and $H = \mathrm{gal}\,(L'/K)$, H is then a normal subgroup of G naturally.

Theorem 11 (Herbrand). *For any $i \geq -1$ we have,*

$$(G/H)^i = G^i H/H, \quad \textit{i.e. } (G/H)_i = G_{\psi_{L/L'}(i)} H/H,$$

see [18, Proposition IV.3.14 and Lemma IV.3.5]. Then we straightforwardly can deduce the following result

Corollary 8. *If H is itself a ramification subgroup of G, i.e. $H = G_j$ for some j. Then*

$$(G/H)_i = \begin{cases} G_i/H, & \text{if} \quad i \leq j, \\ \{1\}, & \text{if} \quad i \geq j. \end{cases}$$

An important consequence of Herbrand's Theorem is that we can define upper ramification filtration $\{\mathrm{gal}\,(L/K)^i\}_i$ for an infinite Galois extension L/K as inverse limit as follows,

$$\mathrm{gal}\,(L/K)^i = \varprojlim_{L'/K \text{ finite} L' \subset L} \mathrm{gal}\,(L'/K)^i.$$

In particular, we can define an upper ramification filtration on the whole absolute Galois group as it is done in § 1.2.

3. Congruence formula The integers i such that

$$G_i \neq G_{i+1}, \quad \text{i.e. the breaks} = \text{lower ramification numbers}, \tag{5.1}$$

are congruent modulo p, see [18, Proposition IV.2.11]. This formula is no more true in every well-ramified extension, see § 1.5 that comes.

4. Hasse–Arf theorem

Theorem 12 (Hasse–Arf). *Let L/K be a finite abelian residually separable extension of any local fields. If i is such that $G_i \neq G_{i+1}$ then $\phi(i)$ is an integer.*

5.5. On the monogenic case (a step in the generalization)

L/K is said to be *monogenic* if $\mathcal{O}_L$ is generated by only one element as $\mathcal{O}_K$-algebra, the generator being not necessarily uniformizer, in general.

The Hasse–Arf theorem, Herbrand's Theorem, more generally Sen's theorem, and Hilbert's formula which are true under the strong hypothesis "$\overline{L}/\overline{K}$ separable" (see for example [18]); however remain true in the more general case when "L/K is assumed to be monogenic" see [3, 20, 23, 24] for Hasse–Arf theorem.

Except the Congruence formula (5.1) that requires necessarily the separability of $\overline{L}/\overline{K}$ see § 5 in [20].

Furthermore, from [20, Theorem 5.1] we have:

Definition 1. *A well-ramified extension L/K is defined as a finite Galois and completely ramified extension satisfying one of the three equivalent conditions:*

- *L/K is monogenic,*
- *Hilbert's formula holds,*
- *Herbrand's theorem holds for any normal subgroup.*

Remark 12. Note here the following important facts due to the monogenity.

- If L/K is monogenic then $\overline{L}/\overline{K}$ is too, but the converse is not true, see Counter-example 1.
- If $\overline{L}/\overline{K}$ is separable then L/K is monogenic, the converse is not true for a counter-example take a Galois extension of degree p such that the residue fields extension is purely inseparable, see Counter-example 1.
- In monogenic case, even by assuming that the residue extension is separable, the generator of the respective DVR need not be a uniformizer unless we are the setting of a totally ramified extension. If L/K is not totally ramified, it's very easy to give counter-examples.
- If L is the compositum of two linearly disjoint extensions L_1 and L_2 such that the residue extensions $\overline{L_1}/\overline{K}$ and $\overline{L_2}/\overline{K}$ are separable the compositum $\overline{L}/\overline{K}$ need neither be separable nor monogenic. A main example arises as follows, see Counterexample 1.

Counterexample 1. Let K be any complete discretely valued field of characteristic 0, containing a primitive p-th root of unity with residue field $\overline{K}$ of characteristic $p > 0$. Write π for a uniformizer of K and consider $L_1 = K(\sqrt[p]{\pi u})$, where u has a valuation zero, u is not a p-th power in K and does not reduce to a p-th power in $\overline{K}$, and $L_2 = K(\sqrt[p]{\pi})$. Then each is totally ramified of degree p; $\overline{L}_1/\overline{K}$ and $\overline{L}_2/\overline{K}$ are both trivial (so separable). Also $\sqrt[p]{\pi u}$ and $\sqrt[p]{\pi}$ are both roots of $f(X) = X^{2p} - \pi(1+u)X^p + \pi^2 u$ which is irreducible over K according to Schönmann criterion, so $\sqrt[p]{\pi u}$ and $\sqrt[p]{\pi}$ are linearly independent over $\mathcal{O}_K$. The compositum $L = L_1.L_2$, is an elementary abelian extension of degree p^2 since its ramification index is p and the residue field is $\overline{L} = \overline{K}(\sqrt[p]{u})$, which is inseparable of degree p over $\overline{K}$. Then we get $\overline{L}/\overline{K}$ monogenic since it is of prime degree. Prove that L/K is not monogenic.

We know that if Herbrand Theorem does not hold then the extension is not well ramified and then it is not monogenic, see [20, Lemma 5.2]. That is if there exists a normal subgroup H of G, such that $i_{G/H}(\tau) \neq 1/e_{(L/L^H)} \sum_{\sigma>\tau} i_G(\sigma)$, where $i_{G(.)}$ is the Artin ramification number.

Let $H =< \sigma >$ be the cyclic group of order p such that $\sigma(u^{1/p}) = \zeta.u^{1/p}$, where ζ is a primitive p-th root of unity. So $L^H = K(\pi^{1/p})$ with L/L^H is ferociously ramified meanwhile L^H/K is wildly ramified both of prime degree and has each a single Artin ramification number. Also $\mathrm{char}\,(K) = 0$, L/L^H is ferociously ramified and $v(u) = 0$ implies that $i_G(\sigma) = s_G(\sigma)$ where $s_{G(.)}$ is the Swan ramification number.

So $i_G(\sigma) = s_G(\sigma) = v_L(\zeta - 1) = e_L/(p-1)$ for every σ. Since L^H/K is wildly ramified and $v(a) = 1$ hence $i_G(\tau) = s_G(\tau) + 1$; if τ is not the identity. That is

$$i_G(\tau) = e_{L^H}/(p-1) + 1, \quad e_{L^H} = pe, \quad e = v_K(p)$$

the absolute ramification index and $e_{L/L^H} = 1$. In this case we have

$$1/e_{(L/L^H)} \sum_{\sigma>\tau} i_G(\sigma) = 1/e_{(L/L^H)} \sum_{\sigma\in H} i_G(\sigma),$$

so, Herbrand does not hold. Then $\mathcal{O}_L$ is not monogenic over $\mathcal{O}_K$.

- The separability of a finite extension does not imply the separability of the residue extension. Indeed, it easy to construct a Counterexample 2.

Counterexample 2. Let K be CDVF with $\overline{K}$ imperfect. Regardless of the characteristic of K, consider $a \in \overline{K} \setminus \overline{K}^p$, thus $X^p - a$ is irreducible in $\overline{K}[X]$. Take

$$f(X)=X^p-bX-a$$

with $b \in \mathcal{M}_K$ ($\mathcal{M}_K$ is the maximal ideal of $\mathcal{O}_K$) requiring that

$$b \neq p^p(a/(1-p))^{(p-1)}.$$

Here f is separable and has reduction $X^p - a \in k[X]$. Here $L = K[X]/(f)$ is a degree p separable extension of K and its subring $\mathcal{O}_0 = \mathcal{O}_K[X]/(f)$ is a domain that is $\mathcal{O}_K$-finite and $\mathcal{O}_0/\pi\mathcal{O}_0 = k[X]/(X^p - a)$, is a field where an uniformiser $\pi \in \mathcal{O}_K$. To prove that $\mathcal{O}_0$ is a DVR see the proof of Lemma 4. We get $\mathcal{O}_0$=$\mathcal{O}_L$ and $\mathcal{O}_0/\mathcal{M}_0$=$l$, is the residue field of L and $e_{L/K}$=1 because the chosen uniformiser of $\mathcal{O}_K$ is an uniformiser of $\mathcal{O}_L$ too. L/K is a degree p separable extension with ramification index $e_{L/K}$=1 and the residual extension l/k is purely inseparable of degree p so L/K is not unramified. □

Assume K being CDVF with imperfect residue field $\overline{K}$. Regardless of the characteristic of K, consider any irreducible and separable polynomial f of $K[X]$ lying above $X^p - a$ with $a \in \overline{K} \setminus \overline{K}^p$. Then we have,

Lemma 4. *For θ a root of f, $L = K(\theta)/K$ is separable extension, meanwhile, its residue extension $\overline{K}(\sqrt[p]{a})/\overline{K}$ is inseparable.*

P r o o f. $X^p - a$ is separable, since it is irreducible, adjoining a root of f (which is separable since irreducible) to get $L = K(\theta) = K[X]/(f)$ gives a degree p separable extension with $\overline{K}(\sqrt[p]{a})/\overline{K}$ inside the residue field. Now, its subring $\mathcal{O}_0 = \mathcal{O}_K[X]/(f)$, is a domain that is $\mathcal{O}_K$-finite and $\mathcal{O}_0/\pi\mathcal{O}_0 = \overline{K}[X]/(X^p - a)$, is a field where $\pi \in \mathcal{O}_K$ denotes an uniformiser. Now prove that $\mathcal{O}_0$ is a DVR or equivalently that $\mathcal{O}_0$ is the integral closure of $\mathcal{O}_K$ in L. (That is true, indeed, if $\mathcal{O}$ is a DVR and f in $\mathcal{O}[X]$ has an irreducible reduction, then $\mathcal{O}[X]/f$ is again a DVR). More precisely, $\mathcal{M}_0 = \pi\mathcal{O}_0$, is a principal maximal ideal in $\mathcal{O}_0$. This is the only maximal ideal of $\mathcal{O}_0$ because any nonzero prime ideal of $\mathcal{O}_0$ intersects $\mathcal{O}_K$ in its unique nonzero prime ideal $\pi\mathcal{O}_K$ and so contains $\pi\mathcal{O}_0$. It follows that $\mathcal{O}_0$ must be DVR. Then the fundamental inequality implies the residue field is exactly $\overline{K}(\sqrt[p]{a})/\overline{K}$ and the ramification index is 1. So, you have a separable L/K with purely inseparable residue extension. □

Note 3. The hypothesis "f irreducible and separable polynomial" is necessary if $\mathrm{char}\,(K) > p$. Of course in such case irreducible doesn't mean separable.

Remark 13. Much more, the solvability of a finite extension does not imply the separability of the residue extension. Indeed, see the following example.

Example 4. Consider $k = \mathbb{F}_p((T_1))$, and $K = k((T_2))$, and α to be a root of the Artin–Schreier equation $f(X) = X^p - T_2^{p-1}X - T_1$ (f is obviously separable since $f' \neq 0$) and write $L = K(\alpha)$. The roots of f are $\alpha + nT_2$, with $0 \leq n \leq p-1$, thus the Galois group of L/K is solvable. Therefore, $\alpha \in \mathcal{O}_L$ (the ring of integers of L) hence is integer over $k[[T_2]]$ (the ring of integers of K), so modulo the maximal ideal we have $\alpha^p = T_1$, the residue extension is then $k(\sqrt[p]{T_1})/k$, which is purely inseparable. □

REFERENCES

1. Abbes A., Saito T. Ramification of local fields with imperfect residue fields. *Amer. J. Math.*, 2002. Vol. 124, No. 5, P. 879–920.
2. Abrashkin V. A. *Towards Explicit Description of Ramification Filtration in the 2-dimensional Case.* Prepint of Nottingham Univ., 2000. No. 00-01.
3. Borger J. A monogenic Hasse–Arf theorem. In: *Proc. of the Conf. on Ramification Theory for Arithmetic Schemes, Luminy, 1999.*
4. Bosch S., Lütkebohmert W., Raynaud M. *Néron Models.* Ergeb. Math. Grenzgeb. (3), vol. 21. Berlin, Heidelberg: Springer–Verlag, 1990. 328 p. DOI: 10.1007/978-3-642-51438-8
5. Engler A.J., Prestel A. *Valued Fields.* Springer Monogr. Math. Berlin, Heidelberg: Springer–Verlag, 2005. 208 p. DOI: 10.1007/3-540-30035-X
6. Epp H. P. Eliminating wild ramification. *Invent Math.*, 1973. Vol. 19. P. 235–249. DOI: 10.1007/BF01390208
7. Gold R., Madan M.L. Some applications of Abhyankar's Lemma. *Math. Nachr.*, 1978. Vol. 82, No. 1. P. 115–119. DOI: 10.1002/mana.19780820112
8. Jonah D., Konvisser M. Some nonabelian p-groups with abelian automorphism groups. *Arch. Math.*, 1975. Vol. 2, No. 1. P. 131—133. DOI: 10.1007/BF01229715
9. Koenigsmann J. Solvable absolute Galois groups are metabelian. *Invent. Math.*, 2001. Vol. 144. P. 1–22. DOI: 10.1007/s002220000117
10. Kuhlmann F. V. *A Correction to Epp's paper "Elimination of Wild Ramification"*, 2010. arXiv:1003.5687v1 [math.AC]
11. Lbekkouri A. On the solvability in local extensions. *An. Şt. Univ. Ovidius Constanţa*, 2014. Vol. 22. No. 2. P. 121–127. DOI: 10.2478/auom-2014-0037
12. Neukirch J., Shmidt A., Wingberg K. *Cohomology of Number Fields.* Grundlehren Math. Wiss., vol. 323. Berlin: Springer–Verlag, 2000. 720 p.
13. Neukirch J. *Algebraic Number Theory.* Berlin, Heidelberg: Springer–Verlag, 1999. 322 p. DOI: 10.1007/978-3-662-03983-0
14. Ribes L., Zalesskii P. *Profinite Groups.* Ergeb. Math. Grenzgeb. (3), vol. 40. Berlin, Heidelberg: Springer–Verlag, 2000. 483 p. DOI: 10.1007/978-3-642-01642-4
15. Safarevič I.R. On p-extensions. *Amer. Math. Soc. Transl. Ser. 2*, 1954. Vol. 4. P. 59–72.
16. Saito T. Ramification of local fields with imperfect residue fields III. *Math. Ann.*, 2012. Vol. 352. P. 567–580. DOI: 10.1007/s00208-011-0652-5
17. Saito T. Wild ramification and the characteristic cycle of an l-adic sheaf. *J. Inst. Math. Jussieu*, 2008. Vol. 8, No. 4. P. 769–829. DOI: 10.1017/S1474748008000364
18. Serre J.-P. *Local Fields.* Grad. Texts in Math., vol. 67. New York: Springer–Verlag, 1979. 241 p. DOI: 10.1007/978-1-4757-5673-9
19. Serre J.-P. *Cohomologie Galoisienne.* Lecture Notes in Math., vol. 5. Berlin Heidelberg: Springer–Verlag, 1997. 181 p. DOI: 10.1007/BFb0108758
20. Spriano L. Well ramified extensions of complete discrete valuation fields with application to the Kato Conductor. *Canad. J. Math.*, 2000. Vol. 52, No. 6. P. 1269–1309. DOI: 10.4153/CJM-2000-053-1
21. Spriano L. On ramification theory of monogenic extensions. In: *Geom. Topol. Monogr. Vol. 3: Invitation to Higher Local Fields,* eds. I. Fesenko and M. Kurihara, 2000. Part I, Sect. 18. P. 151–164.
22. Ware R. On Galois groups of maximal p-extension. *Trans. Amer. Math. Soc.*, 1992. Vol. 333, No. 2. P. 721–728. DOI: 10.2307/2154057
23. Xiao L. On ramification filtrations and p-adic differential modules, I: the equal characteristic case. *Algebra Number Theory*, 2010. Vol. 4, No. 8. P. 969–1027. DOI: 10.2140/ant.2010.4.969
24. Xiao L. On ramification filtrations and p-adic differential equations, II: mixed characteristic case. *Compos. Math.*, 2012. Vol. 148, No. 2. P. 415–463. DOI: 10.1112/S0010437X1100707X
25. Zariski O., Samuel P. *Commutative Algebra I.* Grad. Texts in Math., vol. 28. New York: Springer–Verlag, 1975. 334 p.
26. Zhukov I.B. *On Ramification Theory in the Imperfect Residue Field Case.* Prepint No. 98-02, Nottingham Univ., 1998. Accessible on arXiv:math/0201238 [math.NT]
27. Zhukov I.B. *Ramification of Surfaces Artin-Schreier Extensions*, 2002. arXiv:math/0209183 [math.AG]

28. Zhukov I.B. *Ramification of Surfaces: Sufficient Jet Order for Wild Jumps*, 2002. arXiv:math/0201071 [math.AG]

URAL MATHEMATICAL JOURNAL, Vol. 5, No. 2, 2019, pp. 55–63

DOI: 10.15826/umj.2019.2.005

RESTRAINED DOUBLE MONOPHONIC NUMBER OF A GRAPH[1]

A.P. Santhakumaran[2]

Department of Mathematics, Hindustan Institute of Technology and Science,
Chennai – 603 103, India
apskumar1953@gmail.com

K. Ganesamoorthy

Department of Mathematics, Coimbatore Institute of Technology,
Coimbatore – 641 014, India
kvgm_2005@yahoo.co.in

Abstract: For a connected graph G of order at least two, a double monophonic set S of a graph G is a *restrained double monophonic set* if either $S = V$ or the subgraph induced by $V - S$ has no isolated vertices. The minimum cardinality of a restrained double monophonic set of G is the *restrained double monophonic number* of G and is denoted by $dm_r(G)$. The restrained double monophonic number of certain classes graphs are determined. It is shown that for any integers a, b, c with $3 \leq a \leq b \leq c$, there is a connected graph G with $m(G) = a$, $m_r(G) = b$ and $dm_r(G) = c$, where $m(G)$ is the monophonic number and $m_r(G)$ is the restrained monophonic number of a graph G.

Keywords: Monophonic set, Restrained monophonic set, Restrained monophonic number, Restrained double monophonic set, Restrained double monophonic number.

AMS Subject Classification: 05C12.

Introduction

By a graph $G = (V, E)$ we mean a finite undirected connected graph without loops or multiple edges. The order and size of G are denoted by p and q respectively. For basic graph theoretic terminology we refer to Harary [5]. For vertices u and v in a connected graph G, the *distance* $d(u, v)$ is the length of a shortest $u - v$ path in G. An $u - v$ path of length $d(u, v)$ is called an $u - v$ *geodesic*. It is known that d is a metric on the vertex set V of G [2]. The *neighborhood* of a vertex v is the set $N(v)$ consisting of all vertices u which are adjacent with v. A vertex v is an *extreme vertex* if the subgraph induced by its neighbors is complete.

The *closed interval* $I[x, y]$ consists of all vertices lying on some $x - y$ geodesic of G, while for $S \subseteq V$, $I[S] = \bigcup_{x,y \in S} I[x, y]$. A set S of vertices of G is a *geodetic set* if $I[S] = V$, and the minimum cardinality of a geodetic set is the *geodetic number* $g(G)$. The geodetic number of a graph was introduced in [2, 6] and further studied in [3, 4]. It was shown in [6] that determining the geodetic number of a graph is an NP-hard problem. A geodetic set S of a graph G is a *restrained geodetic set* if the subgraph $G[V - S]$ induced by $V - S$ has no isolated vertex. The minimum cardinality of a restrained geodetic set of G is the *restrained geodetic number* of G. The restrained geodetic

[1]The second author research work was supported by National Board for Higher Mathematics, INDIA (Project No. NBHM/R.P.29/2015/Fresh/157).

[2]Former Professor

number of a graph was introduced and studied in [1]. Let 2^V denote the set of all subsets of V. The mapping $I : V \times V \to 2^V$ defined by $I[u, v] = \{w \in V : w$ lies on a $u - v$ geodesic in $G\}$ is the *interval function* of G. One of the basic properties of I is that $u, v \in I[u, v]$ for any pair $u, v \in V$. Hence the interval function captures every pair of vertices and so the problem of double geodetic sets is trivially well-defined while it is clear that this fails in many graphs already for triplets (for example, complete graphs). This motivated us to introduce and study double geodetic sets in [7] and further double monophonic sets in [9]. Also, double monophonic parameters like the upper double monophonic number of a graph and the connected double monophonic number of a graph were studied in [10, 11]. This is the basis behind the introduction and study of the restrained double monophonic number of a graph. A set S of vertices is called a *double geodetic set* of G if for each pair of vertices x, y in G there exist vertices $u, v \in S$ such that $x, y \in I[u, v]$. The *double geodetic number* $dg(G)$ is the minimum cardinality of a double geodetic set. The double geodetic number of a graph was introduced and studied in [7].

A *chord* of a path P is an edge joining two non-adjacent vertices of P. A path P is called a *monophonic path* if it is a chordless path. A set S of vertices of G is a *monophonic set* of G if each vertex v of G lies on an $x - y$ monophonic path for some $x, y \in S$. The minimum cardinality of a monophonic set of G is the *monophonic number* of G and is denoted by $m(G)$. The monophonic number of a graph was studied and discussed in [8]. A set S of vertices of G is called a *double monophonic set* of G if for each pair of vertices x, y in G there exist vertices u, v in S such that x, y lie on a $u - v$ monophonic path. The *double monophonic number* $dm(G)$ of G is the minimum cardinality of a double monophonic set of G. The concept of double monophonic number of a graph was introduced and studied in [9].

The concept of distance in graphs is a major component in graph theory with its centrality and convexity concepts having numerous applications to real life problems. There are several interesting applications of these concepts to facility location in real life situations, routing of transport problems and communication network designs. As the paths involved in the discussion of this paper are monophonic, no intervention by hackers or enemies is possible to the respective facilities provided. Further, as monophonic paths are secured and longer than geodesic paths, it is advantageous to more customers in getting the service with protection.

The following theorems will be used in the sequel.

Theorem 1 [8]. *Each extreme vertex of a connected graph G belongs to every monophonic set of G.*

Theorem 2 [8]. *For the complete graph $K_p (p \geq 2)$, $m(K_p) = p$.*

Theorem 3 [7]. *Each extreme vertex of a connected graph G belongs to every double geodetic set of G.*

Throughout this paper G denotes a connected graph with at least two vertices.

1. Restrained double monophonic number

To study the main concepts of this paper, we introduce first the restrained monophonic number of a graph and the restrained double geodetic number of a graph, and then prove some basic results and proceed.

Definition 1. *A **restrained monophonic set** S of a graph G is a **monophonic set** such that either $S = V$ or the subgraph induced by $V - S$ has no isolated vertices. The minimum cardinality of a restrained monophonic set of G is the **restrained monophonic number** of G and is denoted by $m_r(G)$.*

Example 1. For the cycle $C_5 : u, v, w, x, y, u$ of order 5, it is easily verified that $S = \{u, w\}$ is a minimum monophonic set of C_5 and so $m(C_5) = 2$. Since the subgraph induced by $V - S$ has an isolated vertex v, S is not a restrained monophonic set of C_5. It is clear that, $S \cup \{v\}$ is a minimum restrained monophonic set of C_5 so that $m_r(C_5) = 3$. Thus the monophonic number and the restrained monophonic number of a graph are different.

It is clear that every restrained monophonic set of G is a monophonic set of G and so Theorem 1 gives the next result.

Theorem 4. *Each extreme vertex of a connected graph G belongs to every restrained monophonic set of G.*

Corollary 1. *For the complete graph $K_p (p \geq 2)$, $m_r(K_p) = p$.*

Definition 2. *A double geodetic set S of a graph G is a **restrained double geodetic set** if either $S = V$ or the subgraph induced by $V - S$ has no isolated vertices. The minimum cardinality of a restrained double geodetic set of G is the **restrained double geodetic number** of G and is denoted by $dg_r(G)$.*

Example 2. For the cycle C_4 of order 4, it is clear that any set S of two non-adjacent vertices of C_4 is a minimum double geodetic set of C_4 and so $dg(C_4) = 2$. Since the subgrph induced by $V - S$ has an isolated vertices, S is not a restrained double geodetic set of C_4. Also, no 3-element subset of $V(C_4)$ is a restrained double geodetic set of C_4. Thus $V(C_4)$ is the unique minimum restrained double geodetic set of C_4 and so $dg_r(C_4) = 4$. Hence the double geodetic number and the restrained double geodetic number of a graph are different.

It is clear that every restrained double geodetic set of G is a double geodetic set of G and so Theorem 3 gives the next result.

Theorem 5. *Each extreme vertex of a connected graph G belongs to every restrained double geodetic set of G.*

Now, we introduce the restrained double monophonic number of a graph and investigate.

Definition 3. *A double monophonic set S of a graph G is a **restrained double monophonic set** if either $S = V$ or the subgraph induced by $V - S$ has no isolated vertices. The minimum cardinality of a restrained double monophonic set of G is the **restrained double monophonic number** of G and is denoted by $dm_r(G)$.*

Example 3. For the graph G given in Fig. 1, it is clear that no 2-element subset of vertices of G is a monophonic set of G. Since $S = \{v, x, z\}$ is a monophonic set of G, $m(G) = 3$. It is the only minimum monophonic set of G. Since the subgraph induced by $V - S$ has the isolated vertex v_4, it is not a restrained monophonic set of G. Now, the set $S_1 = \{v, x, z, v_4\}$ is a restrained monophonic set of G so that $m_r(G) = 4$.

It is easily seen that no 4-element subset of vertices of G containing the vertex v is a double monophonic set of G. Also, it is clear that the set $S_2 = \{v, v_1, v_2, v_3, v_4\}$ is the unique minimum double monophonic set of G. Since the subgraph induced by $V - S_2$ has no isolated vertices, S_2 is a minimum restrained double monophonic set of G so that $dm_r(G) = 5$. Thus the monophonic number, the restrained monophonic number and the restrained double monophonic number of a graph are different.

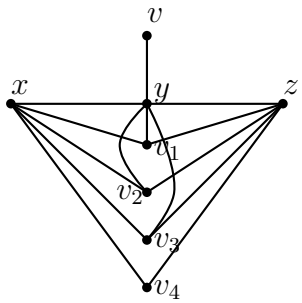


Figure 1. Graph G.

Theorem 6. *Every extreme vertex of a connected graph G belongs to every restrained double monophonic set of G. In particular, if the set of all extreme vertices of G is a restrained double monophonic set, then it is the unique minimum restrained double monophonic set of G.*

P r o o f. Since every restrained double monophonic set is a monophonic set, the result follows from Theorem 1. □

The following results are easy consequences of Theorem 6.

Result 1. *For the complete graph $K_p (p \geq 2)$, $dm_r(G) = p$.*

Result 2. *For a graph G of order p with k extreme vertices,* $\max\{2, k\} \leq dm_r(G) \leq p$.

Result 3. *If T is a tree of order p with k end-vertices and $p - k \geq 2$, then $dm_r(T) = k$.*

Theorem 7. *For any graph G of order p,*

$$2 \leq m(G) \leq m_r(G) \leq dm_r(G) \leq p, \quad m_r(G) \neq p - 1 \neq dm_r(G).$$

P r o o f. Any monophonic set needs at least two vertices and hence $m(G) \geq 2$. Since every restrained monophonic set is also a monophonic set of G, it follows that $m(G) \leq m_r(G)$. It is clear that every restrained double monophonic set of G is also a restrained monophonic set and so $m_r(G) \leq dm_r(G)$. Since the set of all vertices of G is a restrained double monophonic set of G, $dm_r(G) \leq p$. From the definitions of restrained monophonic number and the restrained double monophonic number, it is clear that $m_r(G) \neq p - 1 \neq dm_r(G)$. □

Remark 1. The bounds in Theorem 7 are sharp. The two end-vertices of a nontrivial path P_n on n vertices is its unique minimum monophonic set so that $m(P_n) = 2$ and for the complete graph $K_p (p \geq 2)$, we have $dm_r(K_p) = p$. Also, all the inequalities in Theorem 7 can be strict, for the graph G of order 8 given in Fig. 1, $m(G) = 3$, $m_r(G) = 4$ and $dm_r(G) = 5$. Thus we have $2 < m(G) < m_r(G) < dm_r(G) < p$.

Theorem 8. *For any graph G of order p, $2 \leq dm_r(G) \leq dg_r(G) \leq p$, $dm_r(G) \neq p - 1 \neq dg_r(G)$.*

P r o o f. Any restrained double monophonic set needs at least two vertices and so $dm_r(G) \geq 2$. It is clear that every restrained double geodetic set of G is also a restrained double monophonic set and so $dm_r(G) \leq dg_r(G)$. Since the set of all vertices of G is a restrained double geodetic set of G, $dg_r(G) \leq p$. From the definitions of restrained double monophonic number and the restrained

double geodetic number, it is clear that $dm_r(G) \neq p-1 \neq dg_r(G)$. □

Remark 2. The bounds in Theorem 8 are sharp. For the path $P_n(n \geq 4)$, $dm_r(G) = dg_r(G) = 2$ and for the complete graph $K_p(p \geq 3)$, $dm_r(K_p) = dg_r(K_p) = p$. All the inequalities in Theorem 8 can be strict. For the graph G of order 7 given in Fig. 2, no 2-element subset of $V(G)$ forms a minimum restrained double monophonic set of G. The minimum restrained double monophonic sets of G are $S_1 = \{v_1, v_2, v_5\}$ and $S_2 = \{v_1, v_2, v_6\}$ so that $dm_r(G) = 3$. Also, there is no 3-element or 4-element subset of $V(G)$ forms a minimum restrained double geodetic set of G. It is easy to verify that $S_3 = \{v_1, v_2, v_4, v_5, v_6\}$ is a minimum restrained double geodetic set of G and so $dg_r(G) = 5$. Thus we have $2 < dm_r(G) < dg_r(G) < p$.

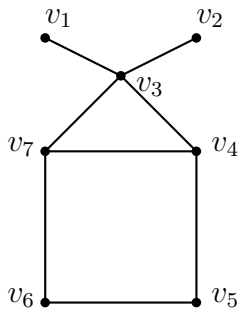


Figure 2. Graph G.

The following results are easy to prove.

Result 4. *For any cycle* C_p, $dm_r(C_p) = \begin{cases} p, & \text{if } p = 3, 4, \\ 3, & \text{if } p = 5, \\ 2, & \text{if } p \geq 6. \end{cases}$

Result 5. *For any wheel* $W_p = K_1 + C_{p-1}$, $(p \geq 4)$, $dm_r(W_p) = \begin{cases} 4, & \text{if } p = 4, \\ 2, & \text{if } p \geq 5. \end{cases}$

Result 6. *For the complete bipartite graph* $G = K_{m,n}(2 \leq m \leq n)$,

$$dm_r(G) = \begin{cases} n+2, & \text{if } 2 = m \leq n, \\ 4, & \text{if } 3 \leq m \leq n. \end{cases}$$

In view of Theorem 7, we have the following realization theorem.

Theorem 9. *For any three integers* a, b, c *with* $3 \leq a \leq b \leq c$, *there is a connected graph* G *with* $m(G) = a$, $m_r(G) = b$ *and* $dm_r(G) = c$.

P r o o f. This theorem is proved by considering four cases.

Case 1. $a = b = c$. Then, for the complete graph $G = K_a$, by Theorem 2, Corollary 1 and Result 1, $m(G) = m_r(G) = dm_r(G) = a$.

Case 2. $a = b < c$. Let G be the graph in Fig. 3 is got by adding $a-1$ new vertices $w_1, w_2, \cdots, w_{a-2}, x$ to the complete bipartite graph $K_{2,c-a+1}$ with the partite sets $U = \{u_1, u_2\}$ and $W = \{v_1, v_2, \cdots, v_{c-a+1}\}$, joining each vertex $w_i(1 \leq i \leq a-2)$ to the vertex u_1 and joining the vertex x to the vertex v_1. By Theorems 1, 4 and 6, every monophonic set, every restrained monophonic

set and every restrained double monophonic set of G contain the set $S = \{w_1, w_2, \cdots, w_{a-2}, x\}$ of all extreme vertices of G. Clearly, S is not a monophonic set of G. It is easy to verify that $S_1 = S \cup \{u_2\}$ is the unique minimum monophonic set of G and so $m(G) = a$. Since the subgraph induced by $V - S_1$ has no isolated vertices, S_1 is the unique minimum restrained monophonic set of G so that $m_r(G) = a = m(G)$. It is clear that the pair of vertices $x, v_i\,(i = 2, 3, \cdots, c-a+1)$ do not lie on any $u-v$ monophonic path, for any $u, v \in S_1$ and so S_1 is not a restrained double monophonic set of G. It is easy to verify that $S_2 = S \cup \{u_2, v_2, \cdots, v_{c-a+1}\}$ is a minimum double monophonic set of G and the subgraph induced by $V - S_2$ has no isolated vertices so that $dm_r(G) = c$.

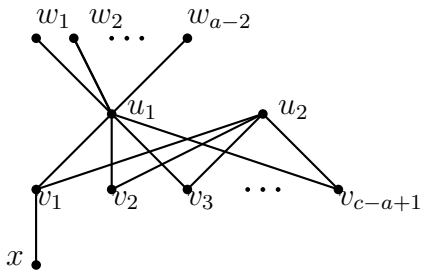


Figure 3. Graph G.

Case 3. $a < b = c$. Let G be the graph in Fig. 4 formed from the path $P_3 : v_1, v_2, v_3$ of order 3, by adding b new vertices $u_1, u_2, \cdots, u_{a-2}, w_1, w_2, ..., w_{b-a+2}$ to P_3 and joining each vertex $u_i (1 \leq i \leq a-2)$ to v_2; and joining each vertex $w_j (1 \leq j \leq b-a+2)$ to v_1 and v_3. By Theorems 1, 4 and 6, every monophonic set, every restrained monophonic set and every restrained double monophonic set of G contain the set $S = \{u_1, u_2, \cdots, u_{a-2}\}$ of all extreme vertices of G. Clearly, S is not a monophonic set of G. Also, for any $x \in V - S$, $S \cup \{x\}$ is not a monophonic set of G. It is easy to verify that $S_1 = S \cup \{v_1, v_3\}$ is a minimum monophonic set of G and so $m(G) = a$. Since the subgraph induced by $V - S_1$ has the isolated vertices $w_1, w_2, \cdots, w_{b-a+2}, v_2$, S_1 is not a restrained monophonic set of G. It is clear that, every restrained monophonic set and every restrained double monophonic set of G contains $\{w_1, w_2, \cdots, w_{b-a+2}\}$ and it follows that $S_2 = S \cup \{w_1, w_2, \cdots, w_{b-a+2}\}$ is a minimum restrained monophonic set and a minimum restrained double monophonic set of G so that $m_r(G) = dm_r(G) = b$.

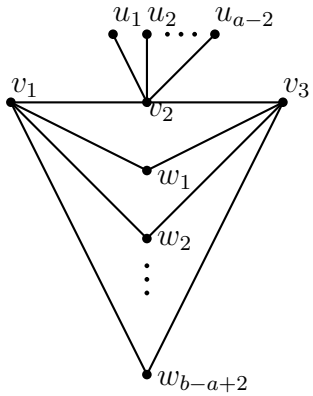


Figure 4. Graph G.

Case 4. $a < b < c$. Let G be the graph in Fig. 5 formed from the path $P_3 : x, y, z$ of order 3 by adding c new vertices $u_1, u_2, \ldots, u_{a-2}, w_1, w_2, \ldots, w_{b-a}, v_1, v_2, \ldots, v_{c-b+2}$ to P_3 and joining each vertex $v_i (1 \leq i \leq c-b+2)$ to the vertices x, y and z; joining each vertex $w_j (1 \leq j \leq b-a)$ to the

vertices x and z; joining each vertex $u_i (1 \leq k \leq a-2)$ to the vertex y. By Theorems 1, 4 and 6, every monophonic set, every restrained monophonic set and every restrained double monophonic set of G contain the set $S = \{u_1, u_2, \cdots, u_{a-2}\}$ of all extreme vertices of G. Clearly, S is not a monophonic set of G and also for any $u \in V(G) - S$, $S \cup \{u\}$ is not a monophonic set of G. It is easily verified that $S_1 = S \cup \{x, z\}$ is a minimum monophonic set of G and so $m(G) = a$. Since the subgraph induced by $V - S_1$ has the isolated vertices $w_2, w_3, \cdots, w_{b-a}$, S_1 is not a restrained monophonic set of G. It is clear that every restrained monophonic set of G contains $\{w_2, w_3, \cdots, w_{b-a}\}$. Then $S_2 = S_1 \cup \{w_1, w_2, \ldots, w_{b-a}\}$ is a minimum restrained monophonic set of G and so $m_r(G) = b$.

Now, any double monophonic set of G should contain the set S. It is easily verified that the set $S' = \{u_1, u_2, \cdots, u_{a-2}, w_1, w_2, \ldots, w_{b-a}, v_1, v_2, \ldots, v_{c-b+2}\}$ is the unique minimum double monophonic set of G. Since the subgraph induced by $V - S'$ has no isolated vertices, it follows that S' is a minimum restrained double monophonic set of G so that $dm_r(G) = c$. □

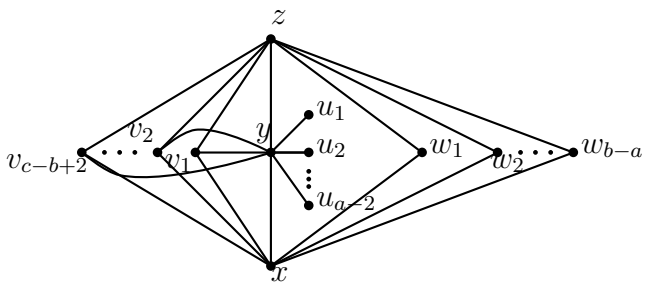


Figure 5. Graph G.

Theorem 10. *For any integer $p \geq 4$ with $2 \leq k \leq p$, $k \neq p-1$, there is a connected graph G of order p such that $dm_r(G) = k$.*

P r o o f. Let $G \neq K_{1,p-1}$ be any tree of order p with k end-vertices. Then clearly, $dm_r(G) = k$. □

In view of Theorem 8, we have the following realization theorem.

Theorem 11. *For every pair a, b of integers with $3 \leq a \leq b$, there is a connected graph G with $dm_r(G) = a$ and $dg_r(G) = b$.*

P r o o f. For $3 \leq a = b$, by Theorem 5 and Result 1, the complete graph K_a of order a has the desired properties. So, assume that $3 \leq a < b$. Let H be the graph obtained from the complete graph K_{a-1} and the path $P_4 : u, v, w, x$ of order 4 by joining all the vertices of K_{a-1} to the vertex u of P_4. Let G be the graph in Fig. 6 obtained from H by taking '$b-a-1$' copies of the path $P_i : x_i, y_i, z_i (1 \leq i \leq b-a-1)$ of order 3 and joining each vertex $x_i (1 \leq i \leq b-a-1)$ in P_i and u in H; and also joining each vertex $z_i (1 \leq i \leq b-a-1)$ in P_i and w in H. Let $S = V(K_{a-1}) \cup \{x\}$ be the set of all extreme vertices of G. By Theorems 5 and 6, every restrained double geodetic set and every restrained double monophonic set of G contain S. Clearly, S is the unique minimum restrained double monophonic set of G and so $dm_r(G) = a$. Since the pair of vertices v, y_i $(i = 1, 2, 3, \cdots, b-a-1)$, do not lie on any geodesic joining a pair of vertices from S, S is not a restrained double geodetic set of G. Let $S' = S \cup \{v, y_1, y_2, \cdots, y_{b-a-1}\}$. It is easy to verify that S' is a minimum double geodetic set of G and the subgraph induced by $V - S'$ has no isolated vertices. Thus S' is a minimum restrained double geodetic set of G and so $dg_r(G)$=b. □

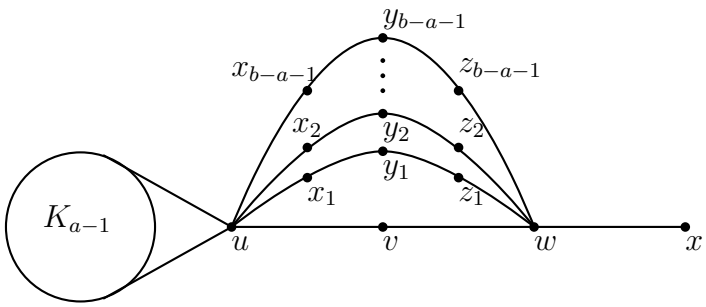


Figure 6. Graph G.

Theorem 12. If $G^{'}$ is a graph obtained by adding k pendant vertices to a connected graph G, then $dm_r(G) \leq dm_r(G^{'}) \leq dm_r(G) + k$.

P r o o f. Let $G^{'}$ be the connected graph obtained from G by adding k pendant vertices $v_i(1 \leq i \leq k)$ to the vertices $u_l(1 \leq l \leq k)$ where each u_l is a vertex of G and each $v_i(1 \leq i \leq k)$ is not a vertex of G. Note that $u_1, u_2, \cdots, u_k$ need not be distinct. Let S be a minimum restrained double monophonic set of G. Then $S \cup \{v_1, v_2, ..., v_k\}$ is a restrained double monophonic set of $G^{'}$ and so $dm_r(G^{'}) \leq dm_r(G) + k$.

Now, we claim that $dm_r(G) \leq dm_r(G^{'})$. Suppose that $dm_r(G) > dm_r(G^{'})$. Let $S^{'}$ be a restrained double monophonic set of $G^{'}$ with $|S^{'}| < dm_r(G)$. Since each v_i $(1 \leq i \leq k)$ is an extreme vertex of $G^{'}$, it follows from Theorem 6 that $\{v_1, v_2, ..., v_k\} \subseteq S^{'}$. Let $S = (S^{'} - \{v_1, v_2, ..., v_k\}) \bigcup \{u_1, u_2, ..., u_l\}$ $(1 \leq l \leq k)$. Then S is a subset of $V(G)$ and $|S| \leq |S'| - k + l = |S'| - (k - l) \leq |S^{'}| < dm_r(G)$. Now, we show that S is a restrained double monophonic set of G. Let $u, v \in V(G) - S$. Then $u, v \in V(G')$ also. Since $S^{'}$ is a restrained double monophonic set of $G^{'}$, u and v lie on a $x - y$ monophonic path P in G' for some vertices $x, y \in S^{'}$. If neither x nor y is $v_i(1 \leq i \leq k)$, then $x, y \in S$. If exactly one of x, y is $v_i(1 \leq i \leq k)$, say $x = v_i$, then u and v lie on a $x - y$ monophonic path in G, where $y = u_j$ and u_j is adjacent to v_j in G' where $i \neq j$. If both $x, y \in \{v_1, v_2, ..., v_k\}$, then let $x = v_i$ and $y = v_j$ where $i \neq j$. Hence u and v lie on the $u_s - u_t$ monophonic path in G, where u_s is adjacent to v_i and u_t is adjacent to v_j in G'. Thus S is a restrained double monophonic set of G. Hence $dm_r(G) \leq |S| < dm_r(G)$, which is a contradiction. □

Remark 3. The bounds for $dm_r(G^{'})$ in Theorem 12 are sharp. Consider a tree T with number of end-vertices $l \geq 3$ and at least two internal vertices. Let $S = \{v_1, v_2, ..., v_l\}$ be the set of all end-vertices of T. Then by Result 3, $dm_r(T) = l$. If we add a pendant vertex to an end-vertex of T, then we obtain another tree $T^{'}$ with l end-vertices. Hence $dm_r(T) = dm_r(T^{'})$. On the other hand, if we add k pendant vertices to a cut-vertex of T, then we obtain a tree $T^{'}$ with $k + l$ end-vertices. Then by Result 3, $dm_r(T^{'}) = dm_r(T) + k$.

2. Conclusions

In this paper, the concept of restrained double monophonic number of a graph is introduced and certain general properties satisfied by this parameter are studied. This parameter is determined for several standard graphs. Also, certain realisation results of this parameter are proved with regard to certain other parameters like monophonic number, restrained monophonic number and restrained double geodetic number of a graph. As a future work of this paper, new parameters like

upper restrained double monophonic number of a graph, forcing restrained double monophonic number of a graph can be developed and investigated.

Acknowledgement

The authors are thankful to the reviewers for their useful comments for the improvement of this paper.

REFERENCES

1. Abdollahzadeh Ahangar H., Samodivkin V., Sheikholeslami S. M. and Abdollah Khodkar. The Restrained Geodetic Number of a Graph. *Bull. Malays. Math. Sci. Soc.,* 2015. Vol. 38. P. 1143–1155. DOI: 10.1007/s40840-014-0068-y
2. Buckley F., Harary F. *Distance in Graphs*. Addison-Wesley, Redwood City, CA, 1990. 335 p.
3. Chartrand G., Harary F. and Zhang P. On the geodetic number of a graph. *Networks,* 2002. Vol. 39, No. 1. P. 1–6. DOI: 10.1002/net.10007
4. Chartrand G., Johns G. L., and Zhang P. On the detour number and geodetic number of a graph. *Ars Combin.*, 2004. Vol. 72. P. 3–15.
5. Harary F. *Graph Theory*, Addison-Wesley, 1969.
6. Harary F., Loukakis E. and Tsouros C. The geodetic number of a graph. *Math. Comput. Modelling*, 1993. Vol. 17, No. 11. P. 89–95. DOI: 10.1016/0895-7177(93)90259-2
7. Santhakumaran A. P. and Jebaraj T. Double geodetic number of a graph. *Discuss. Math. Graph Theory*, 2012, Vol. 32, No. 1. P. 109–119. DOI: 10.7151/dmgt.1589
8. Santhakumaran A. P., Titus P. and Ganesamoorthy K. On the monophonic number of a graph. *J. Appl. Math. Inform.*, 2014. Vol. 32, No. 1–2. P. 255–266. DOI: 10.14317/jami.2014.255
9. Santhakumaran A. P. and Venakata Raghu T. Double monophonic number of a graph. *Int. J. Comput. Appl. Math.*, 2016. Vol. 11, No. 1. P. 21–26. https://www.ripublication.com/ijcam16/ijcamv11n1_03.pdf
10. Santhakumaran A. P. and Venakata Raghu T. Upper double monophonic number of a graph. *Proyecciones*, 2018, Vol. 37, No. 2. P. 295–304. DOI: 10.4067/S0716-09172018000200295
11. Santhakumaran A. P. and Venakata Raghu T. Connected double monophonic number of a graph. *Int. J. Math. Comb.,* 2018, Special Issue 1. P. 54–60.

URAL MATHEMATICAL JOURNAL, Vol. 5, No. 2, 2019, pp. 64–71

DOI: 10.15826/umj.2019.2.006

ORDER OF THE RUNGE–KUTTA METHOD AND EVOLUTION OF THE STABILITY REGION[1]

Hippolyte Séka[†], **Kouassi Richard Assui**[††]

Institut National Polytechnique Houphouët–Boigny,
BP 1093 Yamoussoukro, Côte d'Ivoire

[†]hippolyte.seka@inphb.ci, [††]r.assui@yahoo.fr

Abstract: In this article, we demonstrate through specific examples that the evolution of the size of the absolute stability regions of Runge–Kutta methods for ordinary differential equation does not depend on the order of methods.

Keywords: Stability region, Runge–Kutta methods, Ordinary differential equations, Order of methods.

Introduction

Representations of the stability regions of Runge–Kutta methods are presented in several literatures [1–8, 11, 13]. It has been found that the stability region varies according to the order of the method. However, it is not proven in the literature whether or not there is a relation between the evolution of the size of the region of stability and the order of the method. In this article, we demonstrate that the evolution of the size of the stability region does not depend on the order of the methods. For that we exhibit methods whose regions of stability grow according to the order. Subsequently, we give a counter-example where we introduce a new 8 order method [12]. We compare the stability region of this new 8 order method with those of certain lower order methods. We show that the stability regions of lower order methods are larger than that of the new 8 order method. The study will be done in accordance with the following plan: in Section 2 we describe some generalities on the stability regions, in Section 3 we present some stability functions, in Section 4 we present the new 8 order method and its stability regions, Section 5 we give a conclusion.

1. Generalities on the stability regions

Consider a general form of the first-order ODE given below:

$$y' = f(x, y(x)), \tag{1.1}$$

with the initial condition $y(x_0) = y_0$ for the interval $x_0 \leq x \leq x_n$. Here, x is the independent variable, y is the dependent variable, n is the number of point values, and f is the function of the derivation. The goal is to determine the unknown function $y(x)$ whose derivative satisfies (1.1) and the corresponding initial values. In doing so, let us discretize the interval $x_0 \leq x \leq x_n$ to be

$$x_0, \quad x_1 = x_0 + h, \quad x_2 = x_0 + 2h, ..., \quad x_n = x_0 + nh,$$

[1] We would like to express our deepest appreciation and gratitude to Professor Sergey Khashin of Ivanovo State University who provided us the possibility to coordinate and complete this article.

where h is the fixed step size. With the initial condition $y(x_0) = y_0$, the unknown grid function $y_1, y_2, y_3, \cdots, y_n$ can be calculated by using the Runge–Kutta method of the order 8 (RK8 method).

The 8-th order method is thus obtained by the resolution of the 200 equations with 11 stages [12] on Maple.

Lets consider the Butcher tableau of 8 order and 11 steps RK method (see Fig. 1):

0											
c_2	$a_{2,1}$										
c_3	$a_{3,1}$	$a_{3,2}$									
c_4	$a_{4,1}$	$a_{4,2}$	$a_{4,3}$								
c_5	$a_{5,1}$	$a_{5,2}$	$a_{5,3}$	$a_{5,4}$							
c_6	$a_{6,1}$	$a_{6,2}$	$a_{6,3}$	$a_{6,4}$	$a_{6,5}$						
c_7	$a_{7,1}$	$a_{7,2}$	$a_{7,3}$	$a_{7,4}$	$a_{7,5}$	$a_{7,6}$					
c_8	$a_{8,1}$	$a_{8,2}$	$a_{8,3}$	$a_{8,4}$	$a_{8,5}$	$a_{8,6}$	$a_{8,7}$				
c_9	$a_{9,1}$	$a_{9,2}$	$a_{9,3}$	$a_{9,4}$	$a_{9,5}$	$a_{9,6}$	$a_{9,7}$	$a_{9,8}$			
c_{10}	$a_{10,1}$	$a_{10,2}$	$a_{10,3}$	$a_{10,4}$	$a_{10,5}$	$a_{10,6}$	$a_{10,7}$	$a_{10,8}$	$a_{10,9}$		
c_{11}	$a_{11,1}$	$a_{11,2}$	$a_{11,3}$	$a_{11,4}$	$a_{11,5}$	$a_{11,6}$	$a_{11,7}$	$a_{11,8}$	$a_{11,9}$	$a_{11,10}$	
	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}

Figure 1. Butcher tableau of RK8 method.

The numerical solution is given by the formula

$$y_{i+1} = y_i + h\Big(\sum_{s=1}^{11} b_s k_s\Big), \tag{1.2}$$

with

$$k_s = f\Big(x_i + c_s h, y_i + h\sum_{j=1}^{s-1} a_{s,j} k_j\Big), \quad x_{i+1} = x_i + h. \tag{1.3}$$

The concept of absolute stability, in its simplest form, is based on the analysis of the behavior, according to the values of the step h, of the numerical solutions of the model equation [9–12]:

$$u'(t) = \lambda u(t). \tag{1.4}$$

Using (1.3) and (1.4), we obtain:

$$\text{for} \quad s \geq 1, \quad k_s = \lambda\Big(y_i + h\sum_{j=1}^{s-1} a_{s,j} k_j\Big);$$

which gives:

$$y_{i+1} = \zeta\,(h\lambda)\, y_i.$$

If $z = h\lambda$, then the absolute stability region is the set

$$\{z \in \mathbb{C}| \ \ |\zeta(z)| \leq 1\}\,.$$

2. Presentation of some stability functions

Consider the standard Runge-Kutta methods of orders 1 to 4. When (1.2) and (1.3) are applied to the model problem (1.4), the resulting equations are

$$\begin{aligned}
&\text{RK1:} && y_{i+1} = (1+z)\, y_i;\\
&\text{RK2:} && y_{i+1} = \left(1+z+\frac{z^2}{2}\right) y_i;\\
&\text{RK3:} && y_{i+1} = \left(1+z+\frac{z^2}{2}+\frac{z^3}{6}\right) y_i;\\
&\text{RK4:} && y_{i+1} = \left(1+z+\frac{z^2}{2}+\frac{z^3}{6}+\frac{z^4}{24}\right) y_i.
\end{aligned}$$

The stability regions are shown at the next figure:

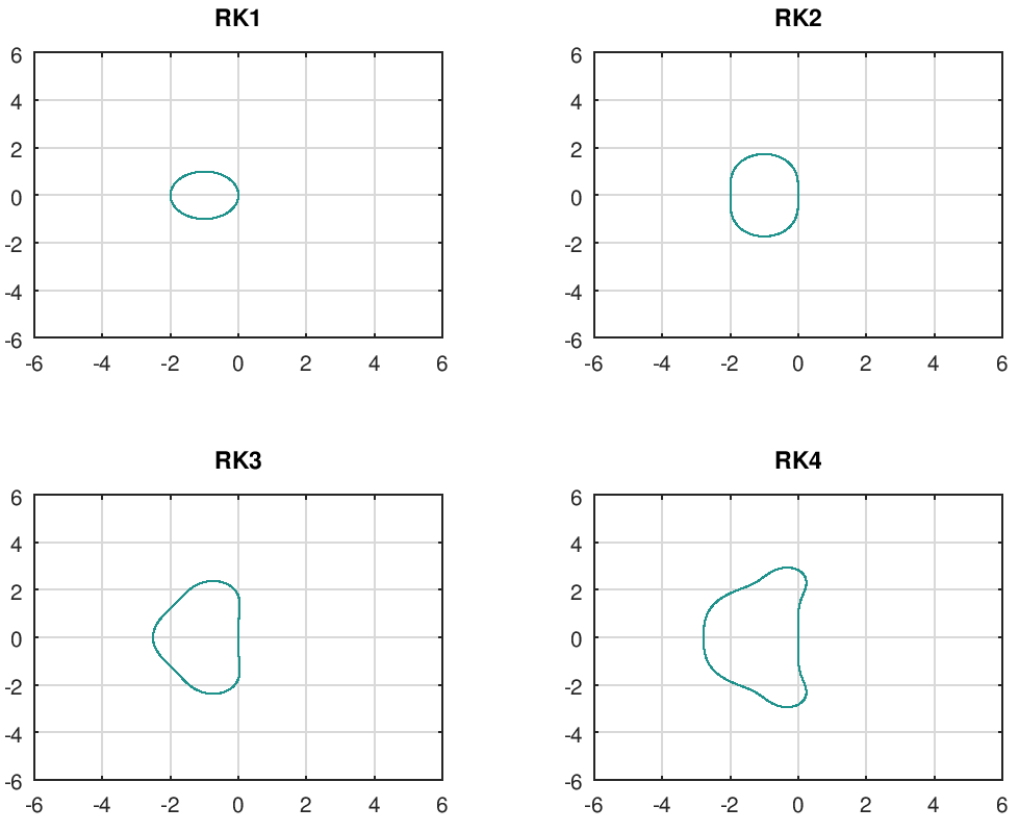


Figure 2. Evolution of the stability region according to the order.

We can see the evolution of the size of the region of stability as the order of the method increases. Let's now give a counterexample for which the stability region is very small.

3. Presentation of the new 8 order method and its stability regions

The family of the 8^{th} order method is thus obtained by the resolution of the 200 equations with 11 stages [12] on Maple. This method depends on free parameters b_8 and $a_{10,5}$ [12].

Some of related coefficients have fixed values, not depending on b_8 and $a_{10,5}$, these coefficients are:

$$b_1 = \frac{1}{20}; \quad b_2 = 0; \quad b_3 = 0; \quad b_4 = 0; \quad b_5 = 0; \quad b_6 = 0; \quad b_9 = \frac{16}{45}; \quad b_{10} = \frac{49}{180}; \quad b_{11} = \frac{1}{20};$$

$$c_2 = \frac{1}{2}; \quad c_3 = \frac{1}{2}; \quad c_4 = \frac{7+\sqrt{21}}{14}; \quad c_5 = \frac{7+\sqrt{21}}{14}; \quad c_6 = \frac{1}{2};$$

$$c_7 = \frac{7-\sqrt{21}}{14}; \quad c_8 = \frac{7-\sqrt{21}}{14}; \quad c_9 = \frac{1}{2}; \quad c_{10} = \frac{7+\sqrt{21}}{14}; \quad c_{11} = 1;$$

$$a_{2,1} = \frac{1}{2};$$

$$a_{3,1} = \frac{1}{4}; \quad a_{3,2} = \frac{1}{4};$$

$$a_{4,1} = \frac{1}{7}; \quad a_{4,2} = \frac{-7-3\sqrt{21}}{98}; \quad a_{4,3} = \frac{21+5\sqrt{21}}{49};$$

$$a_{5,1} = \frac{11+\sqrt{21}}{84}; \quad a_{5,2} = 0; \quad a_{5,3} = \frac{4\sqrt{21}}{63} + \frac{2}{7}; \quad a_{5,4} = \frac{21-\sqrt{21}}{252};$$

$$a_{6,1} = \frac{5+\sqrt{21}}{48}; \quad a_{6,2} = 0; \quad a_{6,3} = \frac{9+\sqrt{21}}{36}; \quad a_{6,4} = \frac{-231+14\sqrt{21}}{360}; \quad a_{6,5} = \frac{63-7\sqrt{21}}{80};$$

$$a_{7,1} = \frac{10-\sqrt{21}}{42}; \quad a_{7,2} = 0;$$

$$a_{9,1} = \frac{1}{32}; \quad a_{9,2} = 0;$$

$$a_{10,1} = \frac{1}{14}; \quad a_{10,2} = 0; \quad a_{10,9} = \frac{4\sqrt{21}}{35} + \frac{132}{245};$$

$$a_{11,1} = 0; \quad a_{11,2} = 0; \quad a_{11,9} = \frac{28-28\sqrt{21}}{45}; \quad a_{11,10} = \frac{49-7\sqrt{21}}{18}.$$

And the others are expressed in terms of b_8 and $a_{10,5}$:

$$b_7 = -b_8 + \frac{49}{180};$$

$$a_{7,3} = -(24/35)a_{10,5} - 136/105 - (12/245)a_{10,5}\sqrt{21} + (656/2205)\sqrt{21};$$

$$a_{7,4} = 7 - (3/10)a_{10,5}\sqrt{21} - (71/45)\sqrt{21} + (3/10)a_{10,5};$$

$$a_{7,5} = -(3/10)a_{10,5} + (3/10)a_{10,5}\sqrt{21} - 43/6 + (169/105)\sqrt{21};$$

$$a_{7,6} = -(277/735)\sqrt{21} + 181/105 + (12/245)a_{10,5}\sqrt{21} + (24/35)a_{10,5};$$

$$a_{8,1} = -\frac{180b_8\sqrt{21} - 49\sqrt{21} - 1800b_8 + 343}{7560b_8}; \quad a_{8,2} = 0;$$

$$a_{8,5} = -\frac{441a_{10,5}\sqrt{21} - 3240a_{7,5}b_8 - 28\sqrt{21} + 882a_{7,5} - 2205a_{10,5} + 147}{3240b_8};$$

$$a_{8,6} = \frac{72a_{10,5}\sqrt{21} + 1620a_{7,6}b_8 - 29\sqrt{21} - 441a_{7,6} - 252a_{10,5} + 119}{1620b_8};$$

And also:

$$a_{8,3} = -\frac{900b_8\sqrt{21} + 11340a_{7,2}b_8 + 11340a_{8,6}b_8 - 98\sqrt{21} - 3087a_{7,2} - 4860b_8 + 686}{11340b_8};$$

$$a_{8,7} = \frac{49}{1620b_8};$$

$$a_{8,4} = \frac{(c_8^2/2) - a_{8,2}c_2 - a_{8,3}c_3 - a_{8,5}c_5 - a_{8,6}c_6 - a_{8,7}c_7}{c_4};$$

$$a_{9,3} = (1/8)a_{10,5}\sqrt{21} - (1/8)a_{10,5} - (1/72)\sqrt{21} + 1/72;$$

$$a_{9,4} = -49/288 - (7/32)a_{10,5}\sqrt{21} + (7/288)\sqrt{21} + (49/32)a_{10,5};$$

$$a_{9,5} = (7/32)a_{10,5}\sqrt{21} - (35/576)\sqrt{21} - (49/32)a_{10,5} + 21/64;$$

$$a_{9,6} = -(1/8)a_{10,5}\sqrt{21} + (1/8)a_{10,5} + (1/72)\sqrt{21} + 5/36;$$

$$a_{9,7} = 91/576 + (7/192)\sqrt{21} - (585/1568)b_8\sqrt{21} - (405/224)b_8;$$

$$a_{9,8} = (585/1568)b_8\sqrt{21} + (405/224)b_8;$$

$$a_{10,3} = -(6/49)a_{10,5}\sqrt{21} - (2/7)a_{10,5} + (2/147)\sqrt{21} + 2/63;$$

$$a_{10,4} = 1/9 - a_{10,5};$$

$$a_{10,6} = (2/7)a_{10,5} - 803/2205 + (6/49)a_{10,5}\sqrt{21} - (59/735)\sqrt{21};$$

$$a_{10,7} = 1/9 + (1/42)\sqrt{21} + (2295/686)b_8 + (495/686)b_8\sqrt{21};$$

$$a_{10,8} = -(2295/686)b_8 - (495/686)b_8\sqrt{21};$$

$$a_{11,3} = (2/3)a_{10,5}\sqrt{21} - (2/3)a_{10,5} - (2/27)\sqrt{21} + 2/27;$$

$$a_{11,4} = -(7/6)a_{10,5}\sqrt{21} + (7/54)\sqrt{21} + (49/6)a_{10,5} - 49/54;$$

$$a_{11,5} = (7/27)\sqrt{21} - 77/54 - (49/6)a_{10,5} + (7/6)a_{10,5}\sqrt{21};$$

$$a_{11,6} = (2/3)a_{10,5} - 64/135 - (2/3)a_{10,5}\sqrt{21} + (94/135)\sqrt{21};$$

$$a_{11,7} = 7/18 - (265/98)b_8\sqrt{21} - (215/14)b_8;$$

$$a_{11,8} = (265/98)b_8\sqrt{21} + (215/14)b_8.$$

The numerical solution is given by the formula (1.2). The values of k_s are given by the formula (1.3). We can notice that if $b_8 = 49/180$ and $a_{10,5} = 1/9$, then we find the method of Cooper–Verner [1, 12].

With the help of Maple, the stability function depends on $a_{10,5}$ and is given by [12]:

$$\begin{aligned}\zeta(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4 + \frac{1}{120}z^5 + \frac{1}{720}z^6 + \frac{1}{5040}z^7 + \frac{1}{40320}z^8 \\ + \left(- \frac{797}{50803200} + \frac{1}{25200}a_{10,5} + \frac{37}{4233600}\sqrt{21}a_{10,5} - \frac{499}{152409600}\sqrt{21}\right)z^9 \\ + \left(\frac{1}{470400} + \frac{1}{2083725}\sqrt{21} - \frac{31}{940800}a_{10,5} - \frac{61}{8467200}\sqrt{21}a_{10,5}\right)z^{10} \\ + \left(- \frac{1}{29030400} - \frac{13}{4267468800}\sqrt{21} + \frac{11}{1612800}a_{10,5} + \frac{353}{237081600}\sqrt{21}a_{10,5}\right)z^{11}.\end{aligned}$$

For $a_{10,5} = 10^6$ we find

$$\zeta(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4 + \frac{1}{120}z^5 + \frac{1}{720}z^6 + \frac{1}{5040}z^7$$
$$+\frac{1}{40320}z^8 + \frac{2015999203}{50803200}z^9 - \frac{15499999}{470400}z^{10} + \frac{197999999}{29030400}z^{11}$$
$$+\frac{190285643}{21772800}\sqrt{21}z^9 - \frac{60046871}{8334900}\sqrt{21}z^{10} + \frac{6353999987}{4267468800}\sqrt{21}z^{11}.$$

The stability region of the new RK8 method for $a_{10,5} = 10^6$ is given by Fig. 3.

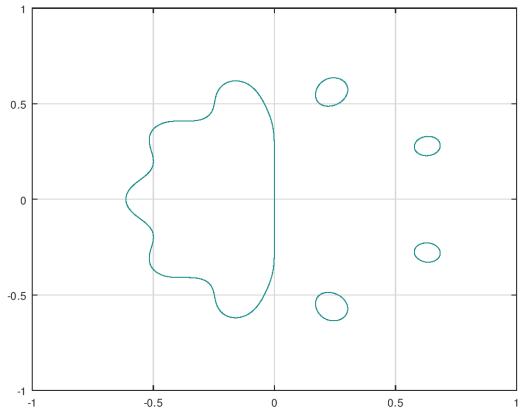


Figure 3. Stability region of the new RK8 method for $a_{10,5} = 10^6$.

We can see that the stability region of the new method of order 8 is smaller than $2, 3, 4$. There is a decrease in the values of x and y.

For $a_{10,5} = 10^{12}$ the stability region is the following (see Fig. 4):

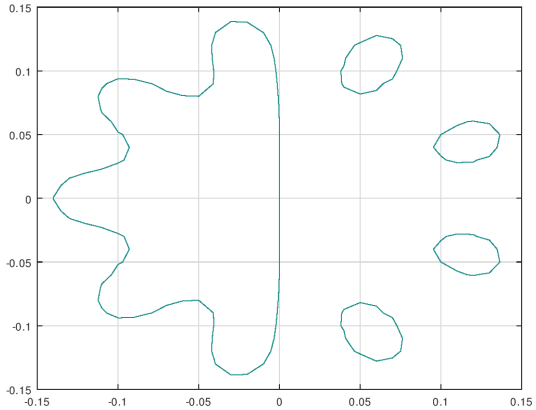


Figure 4. Stability region of the new RK8 method for $a_{10,5} = 10^{12}$.

We can see that the stability region of the new method of order 8 is smaller than those of ordering regions $1, 2, 3, 4$. There is a decrease in the values of x and y.

For $a_{10,5} = \underbrace{9\ldots9}_{37\ times}$ the stability region is is shown at the next figure:

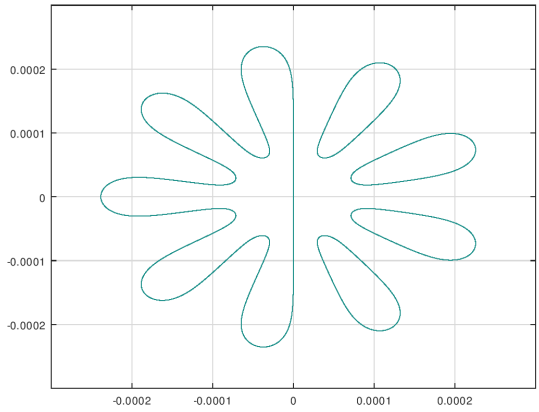


Figure 5. Stability region of the new RK8 method.

We find that the values of x and y have very strongly diminished and the region of stability is very small.

4. Conclusion

Presumably, by representing the domains of stability of methods of the order of 1, 2, 3, 4, one could assume that the higher the order, the greater the area of stability. However, a new 8 order method is discovered. The stability region of this 8 order method is smaller than that of the regions of orders $2, 3, 4$. We can therefore conclude that the evolution of the size of the stability regions of Runge-Kutta methods does not depend on the order of the method.

REFERENCES

1. Butcher J.-C. *Numerical Methods for Ordinary Differential Equations.* 2nd ed. John Wiley & Sons Ltd., 2008. 175 p. DOI: 10.1002/9780470753767
2. Calvo M., Montijano J. I., Randez L. A new embedded pair of Runge–Kutta formulas of orders 5 and 6. *Comput. Math. Appl.*, 1990. Vol. 20, No. 1. P. 15–24. DOI: 10.1016/0898-1221(90)90064-Q
3. Cassity C. R. The complete solution of the fifth order Runge–Kutta equations. *SIAM J. Numer. Anal.*, 1969. Vol. 6, No. 3. P. 432–436. DOI: 10.1137/0706038
4. Feagin T. A tenth-order Runge–Kutta method with error estimate. In: *Proc. of the IAENG Conf. on Scientific Computing.* Hong Kong, 2007. Accessible at https://sce.uhcl.edu/feagin/courses/rk10.pdf
5. Feagin T. *High-Order Explicit Runge-Kutta Methods.* 2013. Accessible at http://sce.uhcl.edu/rungekutta
6. Hairer E., Nørsett S. P., Wanner G. *Solving Ordinary Differential Equations I. Nonstiff Problems.* Springer Ser. Comput. Math., vol. 8. Berlin, Heidelberg: Springer–Verlag, 1993. 528 p. DOI: 10.1007/978-3-540-78862-1
7. Houben S. *Stability Regions of Runge–Kutta Methods.* Eindhoven University of Technology, 2002. Accessible at https://www.win.tue.nl/casa/meetings/seminar/previous/_abstract020220_files/talk.pd
8. Jackiewicz Z. *General Linear Methods for Ordinary Differential Equations.* John Wiley & Sons, Inc., 2009. 482 p. DOI: 10.1002/9780470522165
9. Khashin S. I. *List of Some Known Runge-Kutta Methods Family.* Preliminary version. 2013. Accessible at http://math.ivanovo.ac.ru/dalgebra/Khashin/rk/sh_rk.html
10. Kashin S. I. Estimating the error in classical Runge–Kutta methods. *Comput. Math. Math. Phys.*, 2014. Vol. 54, No. 5. P. 767–774. DOI: 10.1134/S0965542514050145

11. Liu M. Z., Song M. H., Yang Z. W. Stability of Runge–Kutta methods in the numerical solution of equation $u'(t) = au(t) + a_0u([t])$. *J. Comput. Appl. Math.*, 2004. Vol. 166, No. 2. P. 361–370. DOI: 10.1016/j.cam.2003.04.002
12. Seka H., Assui K. R. A New Eighth Order Runge-Kutta Family Method. *J. Math. Res.*, 2019. Vol. 11, No. 2. P. 190–199. DOI: 10.5539/jmr.v11n2p190
13. Velagala S. R. *Stability Analysis of the 4th order Runge–Kutta Method in Application to Colloidal Particle Interactions.* Master's thesis. University of Illinois, Urbana-Champaign, USA, 2014. Accessible at http://hdl.handle.net/2142/72750

URAL MATHEMATICAL JOURNAL, Vol. 5, No. 2, 2019, pp. 72–79

DOI: 10.15826/umj.2019.2.007

ON A DYNAMIC GAME PROBLEM WITH AN INDECOMPOSABLE SET OF DISTURBANCES[1]

Dmitriy A. Serkov

Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences,
16 S. Kovalevskaya Str., Ekaterinburg, 620990, Russia

serkov@imm.uran.ru

Abstract: For an abstract dynamic system, a game problem of retention of the motions in a given set of the motion histories is considered. The case of an indecomposable set of disturbances is studied. The set of successful solvability and a construction of a resolving quasistrategy based on the method of programmed iterations is proposed.

Keywords: Indecomposable disturbances, Quasistrategy, Retention problem.

Introduction

The theory and methods of solving control problems are widely developed in the case when the instantaneous (geometric) restrictions in couple with measurability are the only claim describing the set of admissible disturbances. For these important problems, such a fundamental results as the theorem on the alternative and the extreme aiming method are established (see [5] and the references therein). This kind of restrictions imply the decomposability property [7] of the set of admissible disturbances. In control problems, this property is known as the possibility to "glue" any admissible disturbances at any time for composing a new admissible disturbance. On the other hand, a notable family of control problems is characterized by the absence of this property: typical cases are given by the systems with continuous or constant disturbances.

In the present paper we consider a dynamic control problem with an indecomposable set of disturbances. The consideration is carried out on the example of a retention problem — a simple case of the positional differential game. We search a solution of the problem in a set of quasistrategies. The description of controlled system is given in an abstract form and, in general, follows the scheme [9]. The proposed solution is based on the method of programmed iterations (see [1] and the references therein; see also [9]). The formalization studied lies in direction of the problems that use an additional information on the functional properties of the set of disturbances (see, e.g., [4, 6]). The absence of topological requirements on the elements of the retention problem is compensated by an increasing of the iterations "number" [8]. As usual in the abstract setting, the control time "interval" is not assumed to be bounded or connected.

[1]The reported study was funded by RFBR, project number 19-01-00573.

1. Problem statement

1.1. Dynamic system

Denote by $\mathcal{P}(T)$ (by $\mathcal{P}'(T)$) all (all non-empty) subsets of the set T. For non-empty sets A and B, let B^A be the set of all mappings defined on the set A with values in the set B. If, in addition, $f \in B^A$ and $C \in \mathcal{P}'(A)$, then $(f \,|\, C) \in B^C$ denotes the restriction of the mapping f to the set C: $(f \,|\, C)(x) \triangleq f(x)\ \forall x \in C$. In case when $F \in \mathcal{P}'(B^A)$, we denote $\big(F \,|\, C\big) \triangleq \{(f \,|\, C) : f \in F\}$.

Choose and fix a non-empty subset $\mathbf{I}$ of real numbers $\mathbb{R}$ as an analogue of a time interval. Non-empty sets $\mathbf{X}$ and $\mathbf{Y}$ specify the ranges of the spatial variables and the disturbance values respectively. If $t \in \mathbf{I}$, then we denote $\mathbf{I}^t \triangleq \{\xi \in \mathbf{I} \mid \xi \le t\}$ and $\mathbf{I}_t \triangleq \{\xi \in \mathbf{I} \mid \xi \ge t\}$. Let $\mathbf{C} \in \mathcal{P}'(\mathbf{X}^{\mathbf{I}})$ and $\Omega \in \mathcal{P}'(\mathbf{Y}^{\mathbf{I}})$ be the sets of admissible trajectories and disturbances respectively. Denote by $\mathcal{D} \triangleq \mathbf{I} \times \mathbf{C} \times \Omega$ the state space of the controlled process. For any $t \in \mathbf{I}$, $x \in \mathbf{C}$, we denote $\mathrm{Z}_0(x \,|\, \mathbf{I}^t) \triangleq \{x' \in \mathbf{C} \mid (x' \,|\, \mathbf{I}^t) = (x \,|\, \mathbf{I}^t)\}$.

As an analogue of the dynamic system, we fix a mapping $\mathcal{S} : \mathcal{D} \mapsto \mathcal{P}'(\mathbf{C})$ such that $\forall t \in \mathbf{I}$, $\forall \tau \in \mathbf{I}_t\ \forall x, x' \in \mathbf{C}$ and $\forall \omega, \omega' \in \Omega$

$$\mathcal{S}(t, x, \omega) \in \mathcal{P}'(\mathrm{Z}_0(x \,|\, \mathbf{I}^t)), \tag{1.1}$$

$$((x \,|\, \mathbf{I}^t) = (x' \,|\, \mathbf{I}^t)) \Rightarrow \big(\mathcal{S}(t, x, \omega) = \mathcal{S}(t, x', \omega)\big)\,, \tag{1.2}$$

$$(h \in \mathcal{S}(t, x, \omega)) \Rightarrow (h \in \mathcal{S}(\tau, h, \omega)), \tag{1.3}$$

$$\Big(\big(\mathcal{S}(t, x, \omega) \,|\, \mathbf{I}^\tau\big) = \big(\mathcal{S}(t, x, \omega') \,|\, \mathbf{I}^\tau\big) \& (h \in \mathcal{S}(t, x, \omega)) \& (h' \in \mathcal{S}(\tau, h, \omega'))\Big) \Rightarrow (h' \in \mathcal{S}(t, x, \omega')). \tag{1.4}$$

Thus, for every $(t, x, \omega) \in \mathcal{D}$, $\mathcal{S}(t, x, \omega)$ denotes the set of all trajectories of the system (1.1)–(1.4) corresponding to the history x up to the "moment" t and to the disturbance realization ω after t.

Choose and fix some initial history $(t_0, x_0) \in \mathbf{I} \times \mathbf{C}$. All further constructions are carried out in order to formulate and solve the retention problem for this initial history. Let us define the set $\mathbf{SP}_{(t_0,x_0)} \in \mathcal{P}'(\mathcal{D})$ of all the states of the controlled process arising in the system from the initial history (t_0, x_0) when all admissible disturbances are implemented:

$$\mathbf{SP}_{(t_0,x_0)} \triangleq \Big\{(t, x, \omega) \in \mathcal{D} \mid t \in \mathbf{I}_{t_0}\ (x \,|\, \mathbf{I}^t) \in \big(\mathcal{S}(t_0, x_0, \omega) \,|\, \mathbf{I}^t\big)\Big\}. \tag{1.5}$$

For a state $(t, x, \omega) \in \mathbf{SP}_{(t_0,x_0)}$, we determine the set $\Omega(t, x, \omega)$ of all disturbances that are compatible with this state:

$$\Omega(t, x, \omega) \triangleq \Big\{\omega' \in \Omega \mid \big(\mathcal{S}(t_0, x_0, \omega) \,|\, \mathbf{I}^t\big) = \big(\mathcal{S}(t_0, x_0, \omega') \,|\, \mathbf{I}^t\big)\Big\}. \tag{1.6}$$

So (see (1.2)), we have $\Omega(t_0, x_0, \omega) = \Omega$ for all $\omega \in \Omega$.

1.2. Control procedures and the retention problem

We assume that for the formation of the trajectories the controlling side uses non-empty-valued and non-anticipatory multifunctions from $\mathcal{P}(\mathbf{C})^\Omega$. So, for a state $(t, x, \omega) \in \mathbf{SP}_{(t_0,x_0)}$, we determine the set $\mathbb{M}_{(t,x,\omega)}$ of all admissible control procedures — of quasistrategies — as follows:

$$\begin{aligned}\mathbb{M}_{(t,x,\omega)} \triangleq \Big\{\alpha \in \prod_{\omega' \in \Omega(t,x,\omega)} \mathcal{P}'(\mathcal{S}(t, x, \omega'))\ \mid\ \forall \omega_1, \omega_2 \in \Omega(t, x, \omega)\ \forall \tau \in \mathbf{I}\\ \Big(\big(\mathcal{S}(t_0, x_0, \omega_1) \,|\, \mathbf{I}^\tau\big) = \big(\mathcal{S}(t_0, x_0, \omega_2) \,|\, \mathbf{I}^\tau\big)\Big) \Rightarrow \Big(\big(\alpha(\omega_1) \,|\, \mathbf{I}^\tau\big) = \big(\alpha(\omega_2) \,|\, \mathbf{I}^\tau\big)\Big)\Big\}.\end{aligned} \tag{1.7}$$

Let the set $\mathbf{D} \in \mathcal{P}'(\mathbf{I} \times \mathbf{C})$, which describes the phase constraints, satisfies the conditions

$$(t_0, x_0) \in \mathbf{D}, \quad ((t, x) \in \mathbf{D}) \Rightarrow (\{t\} \times \mathrm{Z}_0(x \,|\, \mathbf{I}^t) \subset \mathbf{D}). \tag{1.8}$$

On the basis of $\mathbf{D}$, we define the set $\mathcal{N}$ in the following way:

$$\mathcal{N} \triangleq (\mathbf{D} \times \Omega) \cap \mathbf{SP}_{(t_0,x_0)} \tag{1.9}$$

and consider the retention of states of the control process in the set $\mathcal{N}$ as the aim of control. Namely, we say that the aim is attainable for the initial state (t, x, ω) if the inclusions

$$(\tau, h, \nu) \in \mathcal{N}, \quad \forall \tau \in \mathbf{I}_t, \ \forall h \in \alpha(\nu), \ \forall \nu \in \Omega(t, x, \omega)$$

hold for some quasistrategy $\alpha \in \mathbb{M}_{(t,x,\omega)}$. For the initial history (t_0, x_0), this definition means that projections of current states of the control process on the set $\mathbf{I} \times \mathbf{C}$ remain in $\mathbf{D}$ for any disturbance $\nu \in \Omega$.

2. The main results

2.1. The programmed absorption operator and its iterations

For $H \in \mathcal{P}(\mathbf{SP}_{(t_0,x_0)})$, $(t, x, \omega) \in \mathbf{SP}_{(t_0,x_0)}$, and $\nu \in \Omega(t, x, \omega)$, define

$$\Pi(\nu \,|\, (t, x, \omega), H) \triangleq \{h \in \mathcal{S}(t, x, \nu) \mid (\tau, h, \nu) \in H \ \ \forall \tau \in \mathbf{I}_t\}. \tag{2.1}$$

In terms of (2.1), we introduce the operator $\mathbf{A}$, $\mathbf{A} \in \mathcal{P}(\mathbf{SP}_{(t_0,x_0)})^{\mathcal{P}(\mathbf{SP}_{(t_0,x_0)})}$, (the programmed absorption operator) by setting

$$\mathbf{A}(H) \triangleq \{(t, x, \omega) \in H \mid \Pi(\nu \,|\, (t, x, \omega), H) \neq \varnothing \ \ \forall \nu \in \Omega(t, x, \omega)\} \tag{2.2}$$

for any $H \in \mathcal{P}(\mathbf{SP}_{(t_0,x_0)})$. The definition of $\mathbf{A}$ immediately implies that

$$\mathbf{A}(H) \subset H. \tag{2.3}$$

Then, following the transfinite induction method (see, for example, [3, sec. I.3]), let us introduce α-iteration $\mathbf{A}^\alpha$, $\mathbf{A}^\alpha \in \mathcal{P}(\mathbf{SP}_{(t_0,x_0)})^{\mathcal{P}(\mathbf{SP}_{(t_0,x_0)})}$, of the operator $\mathbf{A}$ for every ordinal α. For $\alpha = 0$, we assume $\mathbf{A}^0(H) \triangleq H$, $\forall H \in \mathcal{P}(\mathbf{SP}_{(t_0,x_0)})$; if α has a predecessor (let it be an ordinal γ), we write $\mathbf{A}^\alpha \triangleq \mathbf{A} \circ \mathbf{A}^\gamma$; and, if α is a limit ordinal, let $\mathbf{A}^\alpha(H) \triangleq \cap_{\beta \prec \alpha} \mathbf{A}^\beta(H)$, $\forall H \in \mathcal{P}(\mathbf{SP}_{(t_0,x_0)})$. Here, by $\prec$, the strict order relation on the class of ordinals is denoted. Then, according to the transfinite induction principle, α-iteration $\mathbf{A}^\alpha$ of the operator $\mathbf{A}$ is correctly defined for every ordinal α. As a consequence of the definitions and (2.3) (see [8, (4.4)]), we get the following embedding for any ordinal α:

$$\mathbf{A}^\alpha(H) \subset H. \tag{2.4}$$

2.2. Quasistrategies solving the retention problem

Let us study the issue of solvability of the retention problem in the chosen class of quasistrategies. In the following, let the ordinal σ be strictly greater than the cardinality of the set $\mathcal{N}$: $|\mathcal{N}| \prec \sigma$.

Lemma 1. *The inclusions below hold:*

$$\Pi(\cdot \,|\, (t, x, \omega), \mathbf{A}^\sigma(\mathcal{N})) \in \mathbb{M}_{(t,x,\omega)} \quad \forall (t, x, \omega) \in \mathbf{A}^\sigma(\mathcal{N}). \tag{2.5}$$

It follows from lemma 1, (2.4), and definition (2.1), that for any $(t,x,\omega)\in\mathbf{A}^{\sigma}(\mathcal{N})$,

$$(\tau,h,\nu)\in\mathcal{N}\quad\forall\tau\in\mathbf{I}_t\ \forall h\in\Pi(\nu\,|\,(t,x,\omega),\mathbf{A}^{\sigma}(\mathcal{N}))\ \forall\nu\in\Omega(t,x,\omega).$$

Thus, for any $(t,x,\omega)\in\mathbf{A}^{\sigma}(\mathcal{N})$, we have obtained an explicit form of a quasistrategy solving the retention problem in $\mathcal{N}$.

Theorem 1. *The following equality holds:*

$$\mathbf{A}^{\sigma}(\mathcal{N})=\{(t,x,\omega)\in\mathcal{N}\mid\exists\alpha\in\mathbb{M}_{(t,x,\omega)}:\ (\tau,h,\nu)\in\mathcal{N}\ \forall\tau\in\mathbf{I}_t\ \forall h\in\alpha(\nu)\ \forall\nu\in\Omega(t,x,\omega)\}.\quad(2.6)$$

Theorem 1 states that the set $\mathbf{A}^{\sigma}(\mathcal{N})$ is the greatest of the subsets of initial positions from $\mathcal{N}$ that admit a solution of the retention problem in $\mathcal{N}$ in the class of quasistrategies. By Theorem 1, the original retention problem is solvable if and only if $(t_0,x_0,\omega_0)\in\mathbf{A}^{\sigma}(\mathcal{N})$ for some $\omega_0\in\Omega$; as already mentioned, when it is solvable, the quasi-strategy $\Pi(\cdot\,|\,(t_0,x_0,\omega_0),\mathbf{A}^{\sigma}(\mathcal{N}))$ implements this solution (see Lemma 1).

3. Proofs of the results

3.1. Preliminaries

We begin with some auxiliary results. Lemma 2 is based on the properties (1.2)–(1.4).

Lemma 2. *For any $(t,x,\omega)\in\mathcal{D}$, $\tau\in\mathbf{I}^t$, and $h\in\mathcal{S}(t,x,\omega)$, the equality below is true*:

$$\mathcal{S}(\tau,h,\omega)=\mathcal{S}(t,x,\omega)\cap\mathrm{Z}_0(h\,|\,\mathbf{I}^{\tau}).\quad(3.1)$$

Lemma 3 follows from the definitions (1.5) and (1.6).

Lemma 3. *For any $(t,x,\omega)\in\mathbf{SP}_{(t_0,x_0)}$, $\nu\in\Omega(t,x,\omega)$, and $y\in\mathrm{Z}_0(x\,|\,\mathbf{I}^t)$, the relations $(t,x,\nu),(t,y,\nu),(t,y,\omega)\in\mathbf{SP}_{(t_0,x_0)}$, $\Omega(t,x,\nu)=\Omega(t,y,\nu)=\Omega(t,y,\omega)=\Omega(t,x,\omega)$ are fulfilled.*

Lemma 4 is some generalization of the results [1, 2, 9] on the fixed points of the programmed absorption operator. Here, a lack of topological properties of the operator $\mathbf{A}$ and the set $\mathcal{N}$ is compensated by increasing the countable "number" of iterations up to the ordinal σ. The proof of Lemma 4 is based on property (2.3) and statement [8, Prop. 2].

Lemma 4. *For any $H\in\mathcal{P}(\mathbf{SP}_{(t_0,x_0)})$ and an ordinal α, $|H|\prec\alpha$, the following equality holds*:

$$\mathbf{A}^{\alpha}(H)=\mathbf{A}(\mathbf{A}^{\alpha}(H)).\quad(3.2)$$

Lemma 5. *Let $(\tau,h,\nu),(\tau,h',\nu')\in\mathbf{SP}_{(t_0,x_0)}$ and*

$$(h\,|\,\mathbf{I}^{\tau})=(h'\,|\,\mathbf{I}^{\tau}),\quad(3.3)$$

$$\Omega(\tau,h,\nu)=\Omega(\tau,h',\nu').\quad(3.4)$$

Then, $((\tau,h,\nu)\in\mathbf{A}^{\eta}(\mathcal{N}))\Leftrightarrow((\tau,h',\nu')\in\mathbf{A}^{\eta}(\mathcal{N}))$ holds for any ordinal η.

P r o o f. We show the implication

$$((\tau,h,\nu)\in\mathbf{A}^{\eta}(\mathcal{N}))\Rightarrow((\tau,h',\nu')\in\mathbf{A}^{\eta}(\mathcal{N})). \tag{3.5}$$

1. Let $(\tau,h,\nu)\in\mathbf{A}^{0}(\mathcal{N})$ satisfy the conditions of Lemma 5. By definition, we have (see (1.9)) $\mathbf{A}^{0}(\mathcal{N})=\mathcal{N}=\mathbf{SP}_{(t_0,x_0)}\cap(\mathbf{D}\times\Omega)$. Then, $(\tau,h)\in\mathbf{D}$. Using (3.3) and (1.8), we derive $(\tau,h')\in\mathbf{D}$. Since $(\tau,h',\nu')\in\mathbf{SP}_{(t_0,x_0)}$, this implies (see (1.9)) $(\tau,h',\nu')\in\mathcal{N}=\mathbf{A}^{0}(\mathcal{N})$. Suppose in general that the ordinal α is such that for all ordinals β, $\beta\prec\alpha$, implication (3.5) holds.

2. If α is a limit ordinal, then by definition, $\mathbf{A}^{\alpha}(\mathcal{N})=\bigcap_{\beta\prec\alpha}\mathbf{A}^{\beta}(\mathcal{N})$. If, in addition, $(\tau,h,\nu)\in\mathbf{A}^{\alpha}(\mathcal{N})$, then $(\tau,h,\nu)\in\mathbf{A}^{\beta}(\mathcal{N})$ for all $\beta\prec\alpha$. Hence, by the induction hypothesis from (3.5), we obtain the inclusions $(\tau,h',\nu')\in\mathbf{A}^{\beta}(\mathcal{N})$ for all $\beta\prec\alpha$. And, therefore, by the definition of $\mathbf{A}^{\alpha}$, we have the inclusion $(\tau,h',\nu')\in\mathbf{A}^{\alpha}(\mathcal{N})$.

3. If α has a predecessor (let it be an ordinal γ), then, by definition, $\mathbf{A}^{\alpha}(\mathcal{N})=\mathbf{A}(\mathbf{A}^{\gamma}(\mathcal{N}))$. It follows from the definition of $\mathbf{A}$ (see (2.2)), the inclusion $(\tau,h,\nu)\in\mathbf{A}^{\alpha}(\mathcal{N})$, and induction hypothesis (3.5) that

$$(\tau,h',\nu')\in\mathbf{A}^{\gamma}(\mathcal{N}), \tag{3.6}$$

$$\forall\eta\in\Omega(\tau,h,\nu)\ \exists g\in\mathcal{S}(\tau,h,\eta):\ (\xi,g,\eta)\in\mathbf{A}^{\gamma}(\mathcal{N})\ \forall\xi\in\mathbf{I}_{\tau}. \tag{3.7}$$

From (3.3) (see (1.2)), we obtain $\mathcal{S}(\tau,h,\eta)=\mathcal{S}(\tau,h',\eta)$ for all $\eta\in\Omega$. Then, in view of (3.4), we can rewrite (3.7) in the following form:

$$\forall\eta\in\Omega(\tau,h',\nu')\ \exists g\in\mathcal{S}(\tau,h',\eta):\ (\xi,g,\eta)\in\mathbf{A}^{\gamma}(\mathcal{N})\ \forall\xi\in\mathbf{I}_{\tau}. \tag{3.8}$$

By the definition of $\mathbf{A}$ (see (2.2)), relations (3.6), (3.8) mean that $(\tau,h',\nu')\in\mathbf{A}(\mathbf{A}^{\gamma}(\mathcal{N}))$. Hence, we have again $(\tau,h',\nu')\in\mathbf{A}^{\alpha}(\mathcal{N})$.

4. Thus, by virtue of the principle of transfinite induction, implication (3.5) holds for any ordinal η. Since the triples (τ,h,ν) and (τ,h',ν') are included in the conditions of Lemma 5 symmetrically, the state of the lemma follows from (3.5). □

3.2. Proof of Lemma 1

1. Let $(t,x,\omega)\in\mathbf{A}^{\sigma}(\mathcal{N})$. Then, $(t,x,\omega)\in\mathbf{SP}_{(t_0,x_0)}$. Denote $\alpha\triangleq\Pi(\cdot\,|\,(t,x,\omega),\mathbf{A}^{\sigma}(\mathcal{N}))$. By the definition (see (2.1)), we have

$$\alpha\in\prod_{\nu\in\Omega(t,x,\omega)}\mathcal{P}(\mathcal{S}(t,x,\nu)), \tag{3.9}$$

$$(\tau,h,\eta)\in\mathbf{A}^{\sigma}(\mathcal{N})\quad\forall\tau\in\mathbf{I}\ \forall h\in\alpha(\eta)\ \forall\eta\in\Omega(t,x,\omega). \tag{3.10}$$

In view of (3.2), we have $(t,x,\omega)\in\mathbf{A}(\mathbf{A}^{\sigma}(\mathcal{N}))$. Then, $\Pi(\nu\,|\,(t,x,\omega),\mathbf{A}^{\sigma}(\mathcal{N}))\neq\varnothing\ \forall\nu\in\Omega(t,x,\omega)$. So (see (3.9)), we get

$$\alpha\in\prod_{\nu\in\Omega(t,x,\omega)}\mathcal{P}'(\mathcal{S}(t,x,\nu)). \tag{3.11}$$

2. Suppouse $\nu,\nu'\in\Omega(t,x,\omega)$ and $\theta\in\mathbf{I}_t$ satisfy the equality

$$\big(\mathcal{S}(t_0,x_0,\nu)\,|\,\mathbf{I}^{\theta}\big)=\big(\mathcal{S}(t_0,x_0,\nu')\,|\,\mathbf{I}^{\theta}\big). \tag{3.12}$$

We show the embedding $\Gamma\subset\Gamma'$ for the sets $\Gamma\triangleq\big(\alpha(\nu)\,|\,\mathbf{I}^{\theta}\big)$, $\Gamma'\triangleq\big(\alpha(\nu')\,|\,\mathbf{I}^{\theta}\big)$.

Let $\gamma \in \Gamma$, then, the equality $\gamma = (h \,|\, \mathbf{I}^\theta)$ is true for some $h \in \alpha(\nu)$. Let us verify that

$$(\theta, h, \nu) \in \mathbf{SP}_{(t_0,x_0)}, \quad \nu' \in \Omega(\theta, h, \nu). \tag{3.13}$$

By the choice of h, we have (see (3.11)) $h \in \mathcal{S}(t, x, \nu)$. From this inclusion, taking into account the relations $x \in \mathcal{S}(t_0, x_0, \omega)$, $\nu \in \Omega(t, x, \omega)$ and property (1.4) of the system, we get $h \in \mathcal{S}(t_0, x_0, \nu)$. In particular (see (1.5)), we get the first inclusion in (3.13). Then, the second inclusion in (3.13) is well defined and is fulfilled in virtue of (1.6) and (3.12). According to inclusions (3.10), we have $(\theta, h, \nu) \in \mathbf{A}^\sigma(\mathcal{N})$. From the inclusion and the equality $\mathbf{A}^\sigma(\mathcal{N}) = \mathbf{A}(\mathbf{A}^\sigma(\mathcal{N}))$ (Lemma 4), we obtain $(\theta, h, \nu) \in \mathbf{A}(\mathbf{A}^\sigma(\mathcal{N}))$. Hence (see (2.2)), the relation $\Pi(\eta \,|\, (\theta, h, \nu), \mathbf{A}^\sigma(\mathcal{N})) \neq \varnothing$ is true for all $\eta \in \Omega(\theta, h, \nu)$. In particular, taking into account the second inclusion in (3.13), we get the relation $\Pi(\nu' \,|\, (\theta, h, \nu), \mathbf{A}^\sigma(\mathcal{N})) \neq \varnothing$. Let us choose some element $h' \in \Pi(\nu' \,|\, (\theta, h, \nu), \mathbf{A}^\sigma(\mathcal{N}))$. Then (see (2.1)),

$$(\tau, h', \nu') \in \mathbf{A}^\sigma(\mathcal{N}) \quad \forall \tau \in \mathbf{I}_\theta \tag{3.14}$$

and

$$h' \in \mathcal{S}(\theta, h, \nu'). \tag{3.15}$$

It follows from (1.1) and (3.15) that

$$h' \in \mathrm{Z}_0(h \,|\, \mathbf{I}^\theta). \tag{3.16}$$

Let us verify the relations

$$(\tau, h', \nu') \in \mathbf{A}^\sigma(\mathcal{N}) \quad \forall \tau \in \mathbf{I}^\theta_t. \tag{3.17}$$

Fix any $\tau \in \mathbf{I}^\theta_t$. Then, from (3.16), (3.12) and the choice of h, we get $(\tau, h, \nu) \in \mathbf{A}^\sigma(\mathcal{N})$, $(h \,|\, \mathbf{I}^\tau) = (h' \,|\, \mathbf{I}^\tau)$, $(\tau, h, \nu), (\tau, h', \nu') \in \mathbf{SP}_{(t_0,x_0)}$, $\Omega(\tau, h, \nu) = \Omega(\tau, h', \nu')$ by Lemma 3. From these relations, by Lemma 5, we get $(\tau, h', \nu') \in \mathbf{A}^\sigma(\mathcal{N})$. Since the choice of τ was arbitrary, (3.17) holds. Combining (3.14) and (3.17), we obtain

$$(\tau, h', \nu') \in \mathbf{A}^\sigma(\mathcal{N}) \quad \forall \tau \in \mathbf{I}_t. \tag{3.18}$$

Using $(t, x, \omega) \in \mathbf{SP}_{(t_0,x_0)}$ and $\nu, \nu' \in \Omega(t, x, \omega)$, by Lemma 3, we get inclusions $(t, x, \nu), (t, x, \nu') \in \mathbf{SP}_{(t_0,x_0)}$. Then, there exist $y, y' \in \mathrm{Z}_0(x \,|\, \mathbf{I}^t)$ such that $y \in \mathcal{S}(t_0, x_0, \nu)$ and $y' \in \mathcal{S}(t_0, x_0, \nu')$. From (3.12), using (1.2) and (3.1), we obtain

$$\begin{gathered}\big(\mathcal{S}(t, x, \nu) \,|\, \mathbf{I}^\theta\big) = \big(\mathcal{S}(t, y, \nu) \,|\, \mathbf{I}^\theta\big) = \big(\mathcal{S}(t_0, x_0, \nu) \cap \mathrm{Z}_0(y \,|\, \mathbf{I}^t) \,|\, \mathbf{I}^\theta\big) \\ = \big(\mathcal{S}(t_0, x_0, \nu') \cap \mathrm{Z}_0(y' \,|\, \mathbf{I}^t) \,|\, \mathbf{I}^\theta\big) = \big(\mathcal{S}(t, y', \nu') \,|\, \mathbf{I}^\theta\big) = \big(\mathcal{S}(t, x, \nu') \,|\, \mathbf{I}^\theta\big).\end{gathered} \tag{3.19}$$

From (3.15), (3.19), and $h \in \mathcal{S}(t, x, \nu)$, using (1.4), we get $h' \in \mathcal{S}(t, x, \nu')$. The last inclusion and relations (2.1), (3.18) imply the inclusion $h' \in \alpha(\nu')$. Combining it with (3.16), we get the relations $\gamma \triangleq (h \,|\, \mathbf{I}^\theta) = (h' \,|\, \mathbf{I}^\theta) \in \big(\alpha(\nu') \,|\, \mathbf{I}^\theta\big)$. In other words, $\gamma \in \Gamma'$. Because of arbitrary choice of γ, the embedding $\Gamma \subset \Gamma'$ holds. Due to symmetry considerations, we have $\Gamma = \Gamma'$. Since the choice of θ, ν, ν' was arbitrary, we finally get that $\forall \nu, \nu' \in \Omega(t, x, \omega)$ $\forall \tau \in \mathbf{I}$

$$\Big(\big(\mathcal{S}(t_0, x_0, \nu) \,|\, \mathbf{I}^\tau\big) = \big(\mathcal{S}(t_0, x_0, \nu') \,|\, \mathbf{I}^\tau\big)\Big) \Rightarrow \Big(\big(\alpha(\nu) \,|\, \mathbf{I}^\tau\big) = \big(\alpha(\nu') \,|\, \mathbf{I}^\tau\big)\Big). \tag{3.20}$$

From (1.7), (3.11), and (3.20), we obtain $\alpha \in \mathbb{M}_{(t,x,\omega)}$ and, hence, (2.5).

3.3. Proof of Theorem 1

From the relation $\mathbf{A}^{\sigma}(\mathcal{N}) \subset \mathcal{N}$ (see (2.4)), in view of Lemma 1, we find that, under the inequality $|\mathcal{N}| \prec \sigma$, the embedding below holds:

$$\mathbf{A}^{\sigma}(\mathcal{N}) \subset \{(t,x,\omega) \in \mathcal{N} \mid \exists \alpha \in \mathbb{M}_{(t,x,\omega)} : (\tau,h,\nu) \in \mathcal{N}\ \forall \tau \in \mathbf{I}_t\ \forall h \in \alpha(\nu)\ \forall \nu \in \Omega(t,x,\omega)\}. \tag{3.21}$$

1. Denote the set from the right-hand side of (2.6) by Λ. In view of (3.21), to prove the theorem, it is sufficient to establish the relation

$$\Lambda \subset \mathbf{A}^{\sigma}(\mathcal{N}). \tag{3.22}$$

Since $\Lambda \subset \mathcal{N}$, we have

$$\Lambda \subset \mathbf{A}^{0}(\mathcal{N}). \tag{3.23}$$

Let the ordinal ζ be such that, for all $\xi \prec \zeta$,

$$\Lambda \subset \mathbf{A}^{\xi}(\mathcal{N}). \tag{3.24}$$

We show that $\Lambda \subset \mathbf{A}^{\zeta}(\mathcal{N})$. If ζ is a limit ordinal, then we obtain

$$\Lambda \subset \bigcap_{\xi \prec \zeta} \mathbf{A}^{\xi}(\mathcal{N}) = \mathbf{A}^{\zeta}(\mathcal{N}). \tag{3.25}$$

2. If ζ has a predecessor η, we shall verify that $\Lambda \subset \mathbf{A}(\mathbf{A}^{\eta}(\mathcal{N})) = \mathbf{A}^{\zeta}(\mathcal{N})$. Assume the contrary: there is a state $(t_*,x_*,\nu_*) \in \mathcal{N} \subset \mathbf{SP}_{(t_0,x_0)}$ such that

$$(t_*,x_*,\nu_*) \in \Lambda \setminus \mathbf{A}^{\eta+1}(\mathcal{N}). \tag{3.26}$$

Then, it follows from (3.24) and (3.26) that $(t_*,x_*,\nu_*) \in \mathbf{A}^{\eta}(\mathcal{N}) \setminus \mathbf{A}(\mathbf{A}^{\eta}(\mathcal{N}))$. Hence, there exists $\omega^* \in \Omega(t_*,x_*,\nu_*)$ such that $\Pi(\omega^* \,|\, (t_*,x_*,\nu_*), \mathbf{A}^{\eta}(\mathcal{N})) = \varnothing$. From this equality, in view of the inclusion $(t_*,x_*,\nu_*) \in \mathbf{A}^{\eta}(\mathcal{N})$ and definition (2.1), we obtain

$$\forall s \in \mathcal{S}(t_*,x_*,\omega^*) \quad \exists t \in \mathbf{I}_{t_*} : \quad (t,s,\omega^*) \notin \mathbf{A}^{\eta}(\mathcal{N}). \tag{3.27}$$

At the same time (see (3.26)), $(t_*,x_*,\nu_*) \in \Lambda$, and, hence, there exists a quasistrategy $\alpha_* \in \mathbb{M}_{(t_*,x_*,\nu_*)}$ for which, by the definition of Λ,

$$(t,s,\omega) \in \mathcal{N} \quad \forall t \in \mathbf{I}_{t_*} \quad \forall s \in \alpha_*(\omega) \quad \forall \omega \in \Omega(t_*,x_*,\nu_*). \tag{3.28}$$

In particular, as $\omega^* \in \Omega(t_*,x_*,\nu_*)$, we have

$$(t,s,\omega^*) \in \mathcal{N} \quad \forall t \in \mathbf{I}\ \forall s \in \alpha_*(\omega^*). \tag{3.29}$$

Choose arbitrary $s^* \in \alpha_*(\omega^*)$. Then (see (1.7)), we have $s^* \in \mathcal{S}(t_*,x_*,\omega^*)$. According to (3.27), we have $(t^*,s^*,\omega^*) \notin \mathbf{A}^{\eta}(\mathcal{N})$ for some moment $t^* \in \mathbf{I}_{t_*}$. Hence, by (3.24), we have $(t^*,s^*,\omega^*) \notin \Lambda$. In addition (see (3.29)), $(t^*,s^*,\omega^*) \in \mathcal{N}$ (and, therefore, $(t^*,s^*,\omega^*) \in \mathbf{SP}_{(t_0,x_0)}$). By the definition of Λ,

$$\forall \alpha \in \mathbb{M}_{(t^*,s^*,\omega^*)} \quad \exists \omega \in \Omega(t^*,s^*,\omega^*) \quad \exists s \in \alpha(\omega) \quad \exists t \in \mathbf{I}_{t^*} : \ (t,s,\omega) \notin \mathcal{N}. \tag{3.30}$$

3. We define a multi-valued mapping $\beta \in \mathrm{Z}_0(s^* \,|\, \mathbf{I}^{t^*})^{\Omega(t^*,s^*,\omega^*)}$ by the rule

$$\beta(\omega) \triangleq \alpha_*(\omega) \cap \mathrm{Z}_0(s^* \,|\, \mathbf{I}^{t^*}) \quad \forall \omega \in \Omega(t^*,s^*,\omega^*). \tag{3.31}$$

We verify that β is well defined: for any $\omega' \in \Omega(t^*, s^*, \omega^*)$, $\big(\mathcal{S}((t_0, x_0), \omega') \,|\, \mathbf{I}^{t^*}\big) = \big(\mathcal{S}((t_0, x_0), \omega^*) \,|\, \mathbf{I}^{t^*}\big)$ holds (see (1.6)). In view of $\omega^* \in \Omega(t_*, x_*, \nu_*)$ and (1.6), we have

$$\big(\mathcal{S}((t_0, x_0), \omega') \,|\, \mathbf{I}^{t_*}\big) = \big(\mathcal{S}((t_0, x_0), \omega^*) \,|\, \mathbf{I}^{t_*}\big) = \big(\mathcal{S}((t_0, x_0), \nu_*) \,|\, \mathbf{I}^{t_*}\big).$$

Wherefrom, we derive $\omega' \in \Omega(t_*, x_*, \nu_*)$. As ω' was chosen arbitrary, the embedding $\Omega(t^*, s^*, \omega^*) \subset \Omega(t_*, x_*, \nu_*)$ holds. So, the mapping β is well defined for all $\omega \in \Omega(t^*, s^*, \omega^*)$. Let us show the inclusion $\beta \in \mathbb{M}_{(t^*, s^*, \omega^*)}$.

For any $\nu \in \Omega(t^*, s^*, \omega^*)$ and $h \in \beta(\nu)$, by definition (3.31), we have $(h \,|\, \mathbf{I}^{t^*}) = (s^* \,|\, \mathbf{I}^{t^*})$, $h \in \mathcal{S}(t_*, x_*, \nu)$. These relations by using of (1.3) and (1.2) imply $h \in \mathcal{S}(t^*, s^*, \nu)$. As h and ν were chosen arbitrary, the inclusion $\beta \in \prod_{\chi \in \Omega(t^*, s^*, \omega^*)} \mathcal{P}(\mathcal{S}(t^*, s^*, \chi))$ holds.

Let $\omega \in \Omega(t^*, s^*, \omega^*)$. We have (see (1.6)) $\big(\mathcal{S}(t_0, x_0, \omega) \,|\, \mathbf{I}^{t^*}\big) = \big(\mathcal{S}(t_0, x_0, \omega^*) \,|\, \mathbf{I}^{t^*}\big)$. Then, taking into account that $\alpha_* \in \mathbb{M}_{(t_*, x_*, \nu_*)}$, $\omega, \omega^* \in \Omega(t_*, x_*, \nu_*)$ and (1.7), we obtain $\big(\alpha_*(\omega) \,|\, \mathbf{I}^{t^*}\big) = \big(\alpha_*(\omega^*) \,|\, \mathbf{I}^{t^*}\big)$. By using the equality and the inclusion $s^* \in \alpha_*(\omega^*) \cap \mathrm{Z}_0(s^* \,|\, \mathbf{I}^{t^*})$, we derive the relations

$$(s^* \,|\, \mathbf{I}^{t^*}) \in \big(\alpha_*(\omega^*) \cap \mathrm{Z}_0(s^* \,|\, \mathbf{I}^{t^*}) \,|\, \mathbf{I}^{t^*}\big) = \big(\alpha_*(\omega) \cap \mathrm{Z}_0(s^* \,|\, \mathbf{I}^{t^*}) \,|\, \mathbf{I}^{t^*}\big) = \big(\beta(\omega) \,|\, \mathbf{I}^{t^*}\big).$$

So, the inequality $\big(\beta(\omega) \,|\, \mathbf{I}^{t^*}\big) \neq \varnothing$ holds. Hence, $\beta(\omega) \neq \varnothing$. In virtue of arbitrary choice of ω, we have $\beta \in \prod_{\chi \in \Omega(t^*, s^*, \omega^*)} \mathcal{P}'(\mathcal{S}(t^*, s^*, \chi))$.

Concerning the second property of quasistrategies in (1.7) (non-anticipatory), the mapping β inherits it from the quasistrategy α_*. Indeed, by definition (see (3.31)), the mapping β is an intersection of two non-anticipating mappings. So, $\beta \in \mathbb{M}_{(t^*, s^*, \omega^*)}$.

4. Then (see (3.30)),

$$\exists \bar{\omega} \in \Omega(t^*, s^*, \omega^*), \exists \bar{s} \in \beta(\bar{\omega}), \exists \bar{t} \in \mathbf{I}_{t^*} : \quad (\bar{t}, \bar{s}, \bar{\omega}) \notin \mathcal{N}. \tag{3.32}$$

But, by definition, $\bar{t} \in \mathbf{I}_{t_*}$, $\bar{\omega} \in \Omega(t_*, x_*, \nu_*)$ and $\bar{s} \in \alpha_*(\bar{\omega})$. In other words, (3.32) contradicts to (3.28). Hence, assumption (3.26) was wrong, and $\Lambda \subset \mathbf{A}(\mathbf{A}^{\eta}(\mathcal{N}))$. From the embedding, relations (3.23) and (3.25), on the basis of transfinite induction principle we get $\Lambda \subset \mathbf{A}^{\delta}(\mathcal{N})$ for any ordinal δ. When $\delta = \sigma$, the embedding turns into desired relation (3.22).

REFERENCES

1. Chentsov A. G. On a game problem of converging at a given instant of time. *Math. USSR-Sb.*, 1976. Vol. 28, No. 3. P. 353–376. DOI: 10.1070/sm1976v028n03abeh001657
2. Chentsov A. G. On a game problem of guidance with information memory. *Dokl. Akad. Nauk SSSR*, 1976. Vol. 227, No. 2. P. 306–309. (in Russian) http://mi.mathnet.ru/eng/dan40223
3. Engelking R. *General Topology.* Sigma Ser. Pure Math., vol. 6. Warszawa: Panstwowe Wydawnictwo Naukowe, 1985. 540 p.
4. Gomoyunov M. I., Serkov D. Control with a guide in the guarantee optimization problem under functional constraints on the disturbance. *Proc. Steklov Inst. Math.*, 2017. Vol. 299, Suppl. 1. P. S49–S60. DOI: 10.1134/S0081543817090073
5. Krasovskii N. N., Subbotin A. I. *Game-Theoretical Control Problems.* New York: Springer-Verlag, 1988. 517 p.
6. Ledyaev Y. S. Program-predictive feedback control for systems with evolving dynamics. *IFAC-PapersOnLine*, 2018. Vol. 51, No. 32. P. 723–726. DOI: 10.1016/j.ifacol.2018.11.461
7. Rockafellar R. T. Integrals which are convex functionals. *Pacific J. Math.*, 1968. Vol. 24, No. 3. P. 525–539. DOI: 10.2140/pjm.1968.24.525
8. Serkov D. A. Transfinite sequences in the programmed iteration method. *Proc. Steklov Inst. Math.*, 2018. Vol. 300, Suppl. 1. P. S153–S164. DOI: 10.1134/S0081543818020153
9. Serkov D. A., Chentsov A. G. The elements of the operator convexity in the construction of the programmed iteration method. *Bull. South Ural State Univ. Ser. Math. Modell. Program. Comp. Software*, 2016. Vol. 9, No. 3. P. 82–93. DOI: 10.14529/mmp160307

URAL MATHEMATICAL JOURNAL, Vol. 5, No. 2, 2019, pp. 80–93

DOI: 10.15826/umj.2019.2.008

IDENTITIES IN BRANDT SEMIGROUPS, REVISITED[1]

Mikhail V. Volkov

Ural Federal University,
51 Lenin aven., Ekaterinburg, 620000, Russia

m.v.volkov@urfu.ru

Abstract: We present a new proof for the main claim made in the author's paper "On the identity bases of Brandt semigroups" (Ural. Gos. Univ. Mat. Zap., **14**, no.1 (1985), 38–42); this claim provides an identity basis for an arbitrary Brandt semigroup over a group of finite exponent. We also show how to fill a gap in the original proof of the claim in loc. cit.

Key words: Brandt semigroup, Semigroup identity, Identity basis, Finite basis problem.

1. Introduction

We assume the reader's acquaintance with the concepts of an identity and an identity basis as well as other rudiments of the theory of varieties; they all may be found, e.g., in [3, Chapter II]. Our paper deals with identity bases of a certain species of semigroups which we introduce now.

Let G be a group, I a set with at least 2 elements, and 0 a "fresh" symbol that does not belong to $G \cup I$. We define a multiplication on the set $B(G,I) = I \times G \times I \cup \{0\}$ as follows:

$$(i,g,j)(k,h,\ell) = \begin{cases} (i,gh,\ell) & \text{if } j = k, \\ 0 & \text{otherwise,} \end{cases} \quad \text{for all } i,j,k,\ell \in I \text{ and all } g,h \in G, \tag{1.1}$$
$$0x = 0,\ x0 = 0 \quad \text{for all } x \in B(G,I).$$

It is easy to verify that the multiplication (1.1) is associative so that $B(G,I)$ becomes a semigroup. The semigroup is called the *Brandt semigroup over the group* G, and the group G in this context is referred to as the *structure group* of $B(G,I)$ while I is called the *index set*.

Recall that an element a of a semigroup S is said to be *regular* if there exists an element $b \in S$ satisfying $aba = a$ and $bab = b$; it is common to say that b is an *inverse of* a. A semigroup is called *regular* [respectively, *inverse*] if every its element has an inverse [respectively, a unique inverse]. The semigroup $B(G,I)$ is inverse: one can easily check that for all $i,j \in I$ and all $g \in G$, the unique inverse of (i,g,j) is (j,g^{-1},i) and the unique inverse of 0 is 0.

Brandt semigroups arose from a concept invented by Brandt [2] in his studies on composition of quaternary quadratic forms; a distinguished role played by Brandt semigroups in the structure theory of inverse semigroups was revealed by Clifford [4] and Munn [19]. From the varietal viewpoint, Brandt semigroups are of importance as well (see, e.g., [26, Section 7]), and this justifies the study of their identities. Since Brandt semigroups happen to be inverse, there is a bifurcation in this study: along with plain identities $u = v$, in which the terms u and v are plain semigroup words, that is, products of variables, one can consider also inverse identities whose terms involve

[1]This work was supported by the Russian Foundation for Basic Research, project no. 17-01-00551, the Ministry of Science and Higher Education of the Russian Federation, project no. 1.580.2016, and the Competitiveness Program of Ural Federal University.

both multiplication and the unary operation of taking the inverse. We notice that even though plain identities form a special instance of inverse ones, this does not imply that the study of the former fully reduces to the study of the latter; see Section 4 for a more detailed discussion.

Kleiman [13] comprehensively analyzed inverse identities of Brandt semigroups. In particular, he showed how to derive a basis for such identities of $B(G, I)$ from any given identity basis of the group G. Mashevitzky [17] gave a characterization of the set of all plain identities holding in a given Brandt semigroup modulo the plain identities of its structure group. Trahtman [27] found a basis for plain identities of the 5-element Brandt semigroup B_2 in which the construction $B(G, I)$ results provided that G is the trivial group E and $|I| = 2$; this basis consists of the following identities:

$$x^2 = x^3, \quad xyx = xyxyx, \quad x^2y^2 = y^2x^2. \tag{1.2}$$

This fact was frequently cited and used in many applications, including quite important ones such as the positive solution to the finite basis problem for 5-element semigroups [15, 28, 29].

In [30], the present author applied Kleiman's result from [13] along with a generalization of Trahtman's argument from [27] in order to obtain a basis of plain identities for an arbitrary Brandt semigroup over a group of finite exponent. Recall that a group G is said to be of *finite exponent* if there exists a positive integer n such that $g^n = 1$ for all $g \in G$. The least number n with this property is called the *exponent* of G. Clearly, if G is a group of exponent $n > 1$, then $g^{-1} = g^{n-1}$ for all $g \in G$, whence every terms, which is built from certain variables with the help of the unary operation of taking the inverse along with the multiplication, is equal in G to a semigroup word over the same variables. In particular, identities of G (both inverse and plain) admit a basis $\{w_\lambda = 1\}_{\lambda\in\Lambda}$ such that each w_λ is a plain semigroup word; we refer to such a basis as a *positive identity basis* of G. The following is the main result of [30]:

Theorem 1. *Let G be a group of exponent $n > 1$, $\{w_\lambda = 1\}_{\lambda\in\Lambda}$ a positive identity basis of G, and I a set with at least 2 elements. The identities*

$$w_\lambda^2 = w_\lambda \quad (\lambda \in \Lambda), \tag{1.3}$$
$$x^2 = x^{n+2}, \tag{1.4}$$
$$xyx = (xy)^{n+1}x, \tag{1.5}$$
$$x^n y^n = y^n x^n \tag{1.6}$$

constitute a basis for plain identities of the Brandt semigroup $B(G, I)$.

This result also has some important consequences, e.g., it implies a classification of finite inverse semigroups whose plain identities admit a finite basis ([30, Corollary 3], see also Section 4).

For more than 25 years there was no doubt in the validity of Trahtman's argument in [27] until Reilly [24] observed that the argument in fact contained a lacuna. Nevertheless, the claim made in [27] turned out to persist since Reilly managed to prove that the identities (1.2) do form a basis for plain identities of the semigroup B_2, see [24, Theorem 5.4]. A crucial step in Reilly's proof employs a solution to the word problem in the free objects of the variety generated by B_2; this solution (first provided by Mashevitsky in [17]) has quite a complicated formulation and a somewhat bulky justification. Independently and simultaneously, Lee and the present author [16] invented an alternative way to save Trahtman's claim; their approach bypassed the word problem and resulted in a proof which was short and rather straightforward modulo an elementary yet powerful argument known as Kublanovskii's Lemma, see [7, Lemma 3.2]. This technique stems from the present author's paper [32].

Since the proof of Theorem 1 in [30] uses a version of Trahtman's argument, it suffers from the same problem as the proof in [27], and therefore, cannot be considered as truly complete. In fact,

the gap in the proof in [30] can be filled, and we show below how to rescue that proof. However, the main aim of the present paper is to present a new proof of Theorem 1; this new proof follows the approach in [16, 32] and relies on a suitable version of Kublanovskii's Lemma. We have made a fair effort to make our proof self-contained so that, in particular, it should be understandable without any acquaintance with [30] as a whole nor with specific results therein.

2. Preliminaries

Here we collect a few auxiliary results that we need; they all either are known or constitute minor variations of known facts. Some of these results and/or their proofs involve certain concepts of semigroup theory, which all can be found in the early chapters of any general semigroup theory text such as, e.g., [5, 8].

Lemma 1. *Let G be an arbitrary group, I a set with at least 2 elements. An identity $u = v$ holds in the Brandt semigroup $B(G, I)$ if and only if $u = v$ holds in both G and the 5-element Brandt semigroup B_2.*

P r o o f. This was established in [13, Lemma 5] for inverse identities. As plain identities are special instances of inverse ones, the claim holds for plain identities as well. □

Lemma 2. *Let G be a group and I a set such that $|G|, |I| \geq 2$. If G satisfies the identity $w = 1$ where w is a semigroup word, then the Brandt semigroup $B(G, I)$ satisfies the identity $w^2 = w$.*

P r o o f. This fact was also mentioned in [13, p. 214] for inverse identities, and we could have specialized it to plain identities as we did in the proof of Lemma 1. However, the proof in [13] is only briefly outlined, and the outline involves several advanced notions and results from the theory of inverse semigroups. For the sake of completeness, we provide here a direct and elementary proof.

Clearly, G satisfies the identity $w^2 = w$. In view of Lemma 1 it remains to verify that the identity holds in the semigroup B_2. Let $\mathcal{P}(G)$ stand for the set of all non-empty subsets of G. We define a multiplication $\cdot$ on the set $\mathcal{P}(G) \times G$ by the following rule: for $A, B \subseteq G$, $g, h \in G$,

$$(A, g) \cdot (B, h) = (A \cup gB, gh) \quad \text{where} \quad gB = \{gb : b \in B\}. \tag{2.1}$$

It is routine to verify that $\cdot$ is associative so that $(\mathcal{P}(G) \times G, \cdot)$ becomes a semigroup which, for brevity, we denote by S.

Let $\mathrm{alph}(w)$ denote the set of variables that occur in the word w. If we evaluate the variables $x_1, x_2, \dots \in \mathrm{alph}(w)$ at some elements $(A_1, g_1), (A_2, g_2), \dots$ of S and calculate the corresponding value of w, then, according to (2.1), we get an element of the form $(A, w(g_1, g_2, \dots))$ for a certain set $A \in \mathcal{P}(G)$. Since the identity $w = 1$ holds in G, we have $w(g_1, g_2, \dots) = 1$, so that the value is actually of the form $(A, 1)$. Clearly, $(A, 1) \cdot (A, 1) = (A \cup A, 1) = (A, 1)$ for every $A \in \mathcal{P}(G)$, whence S satisfies the identity $w^2 = w$.

Consider the Brandt semigroup $B(E, G)$ over the trivial group $E = \{1\}$; observe that here we make the set G play the role of the index set! Let $J = \{(A, g) \in S : |A| \geq 2\}$ and define a map $\varphi\colon S \to B(E, G)$, letting $s\varphi = 0$ for all $s \in J$ and $(\{a\}, g)\varphi = (a, 1, g^{-1}a)$ for all $(\{a\}, g) \in S \setminus J$. It is easy to see that φ is onto: indeed, an arbitrary triple $(k, 1, \ell) \in B(E, G) \setminus \{0\}$, where $k, \ell \in G$, has a unique preimage in $S \setminus J$, namely, the pair $(\{k\}, k\ell^{-1})$, and for 0, every element of J is a preimage. Let us verify that φ is a semigroup homomorphism. Clearly, $(s \cdot t)\varphi = 0 = s\varphi\, t\varphi$ whenever at least

one of the elements s and t lies in J. For $(\{a\},g),(\{b\},h)\in S\setminus J$, we have

$$\big((\{a\},g)\cdot(\{b\},h)\big)\varphi=\big((\{a,gb\},gh)\big)\varphi=\left\{\begin{array}{ll}[\text{if } a=gb] & (a,1,(gh)^{-1}a)\\ [\text{if } a\ne gb] & 0\end{array}\right.=$$
$$\left.\begin{array}{ll}(a,1,h^{-1}b) & [\text{if } g^{-1}a=b]\\ 0 & [\text{if } g^{-1}a\ne b]\end{array}\right\}=(a,1,g^{-1}a)(b,1,h^{-1}b)=(\{a\},g)\varphi\,(\{b\},h)\varphi.$$

Summing up the established properties of φ, we conclude that the Brandt semigroup $B(E,G)$ is a homomorphic image of the semigroup S, and therefore, $B(E,G)$ also satisfies the identity $w^2=w$.

Since $|G|\ge 2$, we can fix any 2-element subset K in G and "restrict" $B(E,G)$ to K, that is, consider the subsemigroup $\{(k,1,\ell)\in B(E,G):k,\ell\in K\}\cup\{0\}$ of $B(E,G)$. Then the identity $w^2=w$ holds in this subsemigroup, which clearly is isomorphic to B_2. □

Remark 1. The reader may wonder why Lemma 2 could not have been proved by a direct evaluation of the word w in the Brandt semigroup $B(G,I)$. The difficulty is that on this way one should have verified that w and w^2 take value 0 under the same evaluations of the variables from $\mathrm{alph}(w)$ in $B(G,I)$. Of course, not every word w enjoys this property so that one should have analyzed the structure of w, relying entirely on the fact that the identity $w=1$ holds in some non-trivial group. Such an analysis is possible but is rather cumbersome (it amounts to characterizing words w such that the normal closure of w in the free group on the set $\mathrm{alph}(w)$ coincides with the whole group).

Lemma 3. *Let G be a group and I a set with at least 2 elements. If the Brandt semigroup $B(G,I)$ satisfies an identity $u=v$ such that $u=u'yu''$ where y is a variable with $y\notin\mathrm{alph}(u'u'')$ and $\mathrm{alph}(u')\cap\mathrm{alph}(u'')=\varnothing$, then v can be decomposed as $v=v'yv''$ with $\mathrm{alph}(v')=\mathrm{alph}(u')$, $\mathrm{alph}(v'')=\mathrm{alph}(u'')$, and the identities $u'=v'$ and $u''=v''$ hold in $B(G,I)$.*

P r o o f. One could have deduced Lemma 3 by combining Proposition 3.2(ii) of [16] with its left-right dual. However, since the proof of Proposition 3.2(ii) is omitted in [16], we prefer to prove the lemma from scratch by a straightforward argument.

Fix two elements $k,\ell\in I$. Suppose that there exists a variable that occurs in only one of the words u and v. Evaluating this variable at 0 and other variables at $(k,1,k)$, we get that one of the words u and v takes value 0 while the value of the other is $(k,1,k)$, a contradiction. Hence, $\mathrm{alph}(u)=\mathrm{alph}(v)$. Define an evaluation $\zeta\colon\mathrm{alph}(u)\to B(G,I)$ as follows:

$$x\zeta=\begin{cases}(k,1,k) & \text{if } x\in\mathrm{alph}(u'),\\ (k,1,\ell) & \text{if } x=y,\\ (\ell,1,\ell) & \text{if } x\in\mathrm{alph}(u'').\end{cases}$$

Using the multiplication rules (1.1), one readily calculates that the value of the word u under ζ is $(k,1,\ell)$. Since $B(G,I)$ satisfies the identity $u=v$, the value of v under ζ is $(k,1,\ell)$ as well. This value is a product of the triples $(k,1,k)$, $(k,1,\ell)$, and $(\ell,1,\ell)$ in the same order in which the variables from $\mathrm{alph}(u')$, the variable y, and the variables from $\mathrm{alph}(u'')$, respectively, occur in the word v. Fix an occurrence of y in v and let $v'y$ be the prefix of v ending with this occurrence and yv'' the suffix of v starting with this occurrence. Then $v=v'yv''$. Since

$$(k,1,\ell)(k,1,\ell)=(k,1,\ell)(k,1,k)=(k,1,k)(\ell,1,\ell)=(\ell,1,\ell)(k,1,\ell)=(\ell,1,\ell)(k,1,k)=0,$$

none of the factors y^2,yx,xz,zy,zx with $x\in\mathrm{alph}(u')$ and $z\in\mathrm{alph}(u'')$ may occur in v. Therefore, every variable that appears in v' must come from $\mathrm{alph}(u')$ while every variable that appears in

v'' must belong to $\mathrm{alph}(u'')$. We see that $\mathrm{alph}(v') \subseteq \mathrm{alph}(u')$, $\mathrm{alph}(v'') \subseteq \mathrm{alph}(u'')$, and from the equality $\mathrm{alph}(u) = \mathrm{alph}(v)$ shown above, we conclude that $\mathrm{alph}(v') = \mathrm{alph}(u')$, $\mathrm{alph}(v'') = \mathrm{alph}(u'')$.

It remains to verify that the identities $u' = v'$ and $u'' = v''$ hold in $B(G, I)$. The semigroup $B(G, I)$ is inverse, and every inverse semigroup is isomorphic to its left-right dual via the bijection that maps each element to its unique inverse. Therefore $B(G, I)$ satisfies an identity $p = q$ if and only if it satisfies its mirror image $\overleftarrow{p} = \overleftarrow{q}$, where $\overleftarrow{w}$ denotes the word w read backwards. In view of this symmetry, it suffices to verify that $u' = v'$ holds in $B(G, I)$. Arguing by contradiction, consider an evaluation $\varphi\colon \mathrm{alph}(u') \to B(G, I)$ such that the values of u' and v' under φ are different. Then one of these values is not equal to 0; assume, for certainty, that the value of u' is some triple $(i, g, j) \in B(G, I) \setminus \{0\}$. We extend φ to an evaluation $\psi\colon \mathrm{alph}(u) \to B(G, I)$, letting $x\psi = x\varphi$ for all $x \in \mathrm{alph}(u')$ and $y\psi = z\psi = (j, 1, j)$ for all $z \in \mathrm{alph}(u'')$. The value of u under ψ is $(i, g, j)(j, 1, j) = (i, g, j)$; we aim to show that the value of v under ψ is different from (i, g, j). Indeed, if the value of v' under φ is 0, so is the value of v under ψ. If the value of v' under φ is a triple $(i', g', j') \neq (i, g, j)$, then the value of v under ψ is

$$(i', g', j')(j, 1, j) = \begin{cases} (i', g', j) & \text{if } j' = j, \\ 0 & \text{if } j' \neq j, \end{cases} \quad \neq \quad (i, g, j).$$

This contradicts the premise of $u = v$ holding in $B(G, I)$. □

A [0]-*minimal ideal* of a semigroup S is its minimal (with respect to the set inclusion) non-zero ideal if S has a zero and its least ideal otherwise. A non-trivial semigroup S is [0]-*simple* if $S = S^2$ and S is a [0]-minimal ideal of itself. A [0]-simple semigroup is *completely* [0]-*simple* if it contains an idempotent e such that every idempotent f satisfying $ef = fe = f$ is equal to either e or 0.

Lemma 4. *If a semigroup satisfies the identities* (1.5) *and* (1.6) *for some* $n \geq 1$*, then every its* [0]*-minimal ideal that contains a regular element is an inverse completely* [0]*-simple semigroup.*

P r o o f. It suffices to combine a few standard facts of semigroup theory. First, in any semigroup, a [0]-minimal ideal with a regular element is a [0]-simple semigroup, see [5, Theorem 2.29] or [8, Proposition 3.1.3]. Second, every [0]-simple semigroup that satisfies (1.5) is completely [0]-simple; this is a special case of Munn's theorem, see [5, Theorem 2.55] or [8, Theorem 3.2.11]. Each completely [0]-simple semigroup is regular, and a regular semigroup with commuting idempotents is inverse, see [5, Theorem 1.17] or [8, Theorem 5.1.1]. It remains to observe that idempotents commute in every semigroup satisfying (1.6). □

We say that a map $\varphi\colon S \to T$ *separates elements* $a, b \in S$ if $a\varphi \neq b\varphi$.

Lemma 5. *If a semigroup* S *satisfies the identities* (1.5) *and* (1.6) *for some* $n \geq 1$*, then any distinct regular elements* $a, b \in S$ *are separated by a homomorphism of* S *onto an inverse completely* [0]*-simple semigroup.*

P r o o f. This is a version of Kublanovskii's Lemma [7, Lemma 3.2] adapted for the purposes of the present paper. For the reader's convenience, we provide a complete proof, even though it quite closely follows the proof of Kublanovskii's Lemma in [7].

For each regular element $z \in S$, we let $I_z = \{u \in S : z \notin SuS\}$. Observe that $z \notin I_z$: indeed, if z' is an inverse of z, we have $z = zz'zz'z \in SzS$. The set I_z may be empty but if it is not empty, it forms an ideal of S. Indeed, $SutS \subseteq SuS$ and $StuS \subseteq SuS$ for any $u, t \in S$, and hence, if u lies in I_z, so do ut and tu for every $t \in S$. Define the following equivalence relation on S:

$$x \equiv y \pmod{I_z} \text{ if and only if either } x = y \text{ or } x, y \in I_z.$$

Clearly, it is just the equality relation if I_z is empty; otherwise it is nothing but the Rees congruence ι_z corresponding to the ideal I_z. Now define a further equivalence relation ρ_z on S as follows:

$$\rho_z = \{(x, y) \in S \times S : xt \equiv yt \pmod{I_z} \text{ for every } t \in SzS\}.$$

It can be easily verified that ρ_z is a congruence on S; in fact, as observed in [7], ρ_z is the kernel of the so-called Schützenberger representation for S, see [5, Section 3.5].

Clearly, $\rho_z = S \times S$ if $z = 0$. Now we aim to prove the following claim: *if $z \neq 0$, then the quotient semigroup S/ρ_z is an inverse completely $[0]$-simple semigroup.*

If $I_z \neq \varnothing$, the congruence ρ_z contains the Rees congruence ι_z. Then we may substitute S by its quotient S/ι_z as the quotient also satisfies the identities (1.5) and (1.6); in other words, we may (and will) assume that either $I_z = \varnothing$ or $I_z = \{0\}$. Then by the definition of the set I_z, every non-zero element $u \in SzS$ must fulfil $z \in SuS$ whence $SuS = SzS$. We see that SzS is a $[0]$-minimal ideal of S; as SzS contains z which is a regular element, Lemma 4 applies showing that SzS is an inverse completely $[0]$-simple semigroup. So is any homomorphic image of SzS; in particular, so is the image of SzS in the quotient semigroup S/ρ_z. Therefore, it remains to show that the image of S in S/ρ_z coincides with that of SzS, which means that for each $x \in S$, there exists $y \in SzS$ such that $(x, y) \in \rho_z$.

If $x \in SzS$, there is nothing to prove. If $x \notin SzS$, then in particular, $x \notin I_z$ whence $z = pxq$ for some $p, q \in S$. We have $z = pxqz'pxq$, where, as above, z' stands for an inverse of z. Put $w = qz'p$; then $w \in SzS$ because $z' = z'zz' \in SzS$ and $xwx \neq 0$ because $z = pxwxq \neq 0$. Now take an arbitrary element $t \in SzS$. We have already noticed (in the preceding paragraph) that $SuS = SzS$ for every non-zero element $u \in SzS$. Applying this to $u = xwx$, we conclude that $t = rxwxs$ for some $r, s \in S$. Now we have the following chain of equalities:

$$\begin{aligned}
xt = xrxwxs &= (xr)^{n+1}(xw)^{n+1}xs && \text{by applying (1.5) to } xrx \text{ and } xwx \\
&= xr(xr)^n(xw)^n xwxs \\
&= xr(xw)^n(xr)^n xwxs && \text{by applying (1.6)} \\
&= xr(xw)^n(xr)^{n-1}xrxwxs \\
&= xr(xw)^n(xr)^{n-1}xt.
\end{aligned}$$

We see that $\big(x, xr(xw)^n(xr)^{n-1}x\big) \in \rho_z$, and the element $xr(xw)^n(xr)^{n-1}x$ lies in the ideal SzS because so does w. Thus, $xr(xw)^n(xr)^{n-1}x$ can play the role of y, and our claim is proved.

Now we are ready to complete the proof of the lemma. Given an arbitrary pair (a, b) of distinct regular elements is S, we will show that at least one of the congruences ρ_a and ρ_b excludes (a, b). Then the natural homomorphism of S onto the quotient over this congruence separates a and b, and the quotient is an inverse completely $[0]$-simple semigroup by the claim just proved. (One has to take into account that if a congruence of the form ρ_z excludes some pair, then $z \neq 0$ and the claim applies.)

If $a \notin SbS$, then $b \in I_a$. Let a' be an inverse of a. We have then $a'a \in SaS$ and $a(a'a) = a \notin I_a$ while $b(a'a) \in I_a$ since I_a is an ideal. Hence $(a, b) \notin \rho_a$. Similarly, if $b \notin SaS$, we have $(a, b) \notin \rho_b$. Now suppose that $a \in SbS$ and $b \in SaS$. In this case, $SaS = SbS$ and $a, b \notin I_a = I_b$. If we assume that $(a, b) \in \rho_a$, then for every element $t \in SaS$ such that either $at \notin I_a$ or $bt \notin I_a$, we must have $at = bt$. In particular, the latter equality must hold for $t = a'a$ since $a(a'a) = a \notin I_a$ and for $t = b'b$, where b' is an inverse of b, since $b(b'b) = b \notin I_a$. Taking into account that both $a'a$ and $b'b$ are idempotents and that idempotents commute in every semigroup satisfying the identity (1.6), we have

$$a = a(a'a) = b(a'a) = b(b'b)(a'a) = a(b'b)(a'a) = a(a'a)(b'b) = a(b'b) = b(b'b) = b,$$

a contradiction. □

Remark 2. One can call our Lemma 5 "Kublanovskii's Lemma with commuting idempotents". The presence of the identity (1.6) ensures that idempotents commute, and this streamlines the proof. The most important simplification in comparison with the proof of Kublanovskii's Lemma in [7] is that we manage to avoid invoking, along with the congruences ρ_a and ρ_b, their dual versions, that is, the kernels of the corresponding Schützenberger anti-representations.

If S is an arbitrary semigroup and 0 is a "fresh" symbol that does not belong to S, we let S^0 stand for the semigroup on the set $S \cup \{0\}$ with multiplication that extends the multiplication of S and makes all products involving 0 be equal to 0. If G is a group, G^0 is known under the (standard though somewhat oxymoronic) name "*group with zero*". The following fact is a classical result of semigroup theory, see [5, Theorem 3.9] or [8, Theorem 5.1.8].

Lemma 6. *An inverse completely* $[0]$*-simple semigroup is either a group, or a group with zero, or a Brandt semigroup.*

3. Proof of Theorem 1

Recall that we aim to prove that for every group G of exponent $n > 1$ and every set I with at least 2 elements, the identities (1.3)–(1.6) constitute a basis of the plain identities of the Brandt semigroup $B(G, I)$, provided that the set $\{w_\lambda = 1\}_{\lambda\in\Lambda}$ is a positive identity basis of G.

To start with, observe that the identities (1.3)–(1.6) hold in $B(G, I)$. For (1.3) this follows from Lemma 2. As for the identities (1.4)–(1.6), it is obvious that they hold in each group of exponent n. On the other hand, comparing these identities with the identity basis (1.2) of the semigroup B_2, one readily sees that they hold in B_2 as well. Now the "if" part of Lemma 1 ensures that (1.4)–(1.6) hold in $B(G, I)$.

Let $\mathbf{A}$ be the semigroup variety defined by the identities (1.3)–(1.6) and $\mathbf{B}$ the variety generated by the Brandt semigroup $B(G, I)$. The fact established in the preceding paragraph is equivalent to the inclusion $\mathbf{B} \subseteq \mathbf{A}$ and the theorem being proved means the equality $\mathbf{B} = \mathbf{A}$. Arguing by contradiction, assume that the inclusion is strict. Then there exists an identity that holds in the semigroup $B(G, I)$ but fails in the variety $\mathbf{A}$. We choose an identity $u = v$ with this property and with the least value of $|\operatorname{alph}(u)|$. We first check that the words u and v are repeated, where a word w is called *repeated* if each variable from $\operatorname{alph}(w)$ occurs in a factor of w of the form ypy where y is a variable and p is a (possibly empty) word[2]. It is convenient to have a short name for such factors; let us refer to them as to *cells*.

Assume for a moment that, say, u is not repeated. This means that there exists a variable y that occurs in u but does not occur in any cell of u. In particular, y occurs in u exactly once, and moreover, $u = u'yu''$ with $\operatorname{alph}(u') \cap \operatorname{alph}(u'') = \varnothing$. We are in a position to employ Lemma 3 to conclude that v decomposes as $v = v'yv''$ where $\operatorname{alph}(v') = \operatorname{alph}(u')$, $\operatorname{alph}(v'') = \operatorname{alph}(u'')$ and both the identities $u' = v'$ and $u'' = v''$ hold in $B(G, I)$. Since $|\operatorname{alph}(u')|, |\operatorname{alph}(u'')| < |\operatorname{alph}(u)|$, our choice of the identity $u = v$ ensures that the identities $u' = v'$ and $u'' = v''$ hold in the variety $\mathbf{A}$. However, together they imply the identity $u = v$ that cannot hold in $\mathbf{A}$, a contradiction.

Let F stand for the free semigroup of countable rank and let α denote the fully invariant congruence on F that corresponds to the variety $\mathbf{A}$. Then the quotient semigroup F/α satisfies the identities (1.3)–(1.6) and the α-classes $u^\alpha = \{w : (w, u) \in \alpha\}$ and $v^\alpha = \{w : (w, v) \in \alpha\}$ are different in F/α. For the next step of our proof we need the following fact:

[2]The term "repeated" comes from [27, 30]; in [16] words with this property were called "semiconnected".

Lemma 7. *Every α-class that contains a repeated word is a regular element of F/α.*

We proceed with proving Theorem 1 modulo Lemma 7 and prove the lemma afterwards.

By Lemma 7, the α-classes u^α and v^α are regular elements of F/α. Applying Lemma 5, we conclude that u^α and v^α are separated by an onto homomorphism $\chi\colon F/\alpha \to T$, where T is an inverse completely [0]-simple semigroup. Lemma 6 implies the existence of a group Q such that either 1) $T = Q$, or 2) $T = Q^0$, or 3) $T = B(Q, J)$ for some set J with $|J| \geq 2$. In any case, Q is a subgroup of a homomorphic image of F/α, whence the identities (1.3) hold in Q. Clearly, if for some word w, a group satisfies the identity $w^2 = w$, then the group satisfies the identity $w = 1$ as well. Therefore the group Q satisfies the identities $w_\lambda = 1$ for all $\lambda \in \Lambda$. Since these identities form a basis for the identities of the structure group G of our semigroup $B(G, I)$, the group Q belongs to the semigroup variety generated by G, and hence, to the variety $\mathbf{B}$ generated by $B(G, I)$. The 5-element Brandt semigroup B_2 also belongs to $\mathbf{B}$; this follows, for instance from the "only if" part of Lemma 1. Applying the "if" part of Lemma 1, we conclude that the Brandt semigroup $B(Q, J)$ lies in $\mathbf{B}$. From this, we have $T \in \mathbf{B}$ as T is isomorphic to a subsemigroup in $B(Q, J)$ in the cases 1) or 2) and $T = B(Q, J)$ in the case 3). In particular, T satisfies the identity $u = v$. However, the composition of the natural homomorphism $F \to F/\alpha$ with the homomorphism $\chi\colon F/\alpha \to T$ gives rise to an evaluation under which the values of the words u and v are different. This contradiction completes the proof of Theorem 1 modulo Lemma 7.

P r o o f of Lemma 7. Take any α-class h that contains a repeated word, say, w. If some variable y occurs in w only once, then by the definition of a repeated word, y occurs in some cell zpz of w, where p is non-empty. Using the identity (1.5), we substitute the factor zpz by the factor $(zp)^{n+1}z$ and get a new word in the same α-class h in which y occurs at least twice. If this new word still contains some variable x with a single occurrence, we apply the same transformation again, etc. Thus, we may assume that h contains a word q in which every variable occurs at least twice. Now we prove that h contains also a word which is a product of cells, that is, has the form

$$y_1p_1y_1 \cdot y_2p_2y_2 \cdot \ldots \cdot y_kp_ky_k, \tag{3.1}$$

where $y_1, y_2, \ldots, y_k$ are variables and $p_1, p_2, \ldots, p_k$ are (possibly empty) words. For this, we employ a sort of greedy algorithm. Let y_1 be the leftmost variable of the word q. If q ends with y_1, the word q itself is a cell. Otherwise we find the rightmost occurrence of y_1 in q so that $q = y_1p_1y_1 \cdot q_1$ where q_1 is a non-empty word in which y_1 does not occur, and so $|\operatorname{alph}(q_1)| < |\operatorname{alph}(q)|$. Let y_2 be the leftmost variable of q_1. There are two cases to consider, depending on whether y_2 occurs in q_1 at least twice or only once. In the former case, we find the rightmost occurrence of y_2 in q_1 and represent q as $q = y_1p_1y_1 \cdot y_2p_2y_2 \cdot q_2$, where y_1, y_2 do not occur in q_2, and so $|\operatorname{alph}(q_2)| < |\operatorname{alph}(q_1)|$. Let us show that h contains a word with a similar structure also in the latter case. Indeed, the variable y_2 occurs in q at least twice and if it occurs in q_1 only once, then it must occur in p_1. Hence, $p_1 = ry_2s$ for some (possibly empty) words r and s. Then q contains the word $y_2sy_1y_2$ as a factor. Using the identity (1.5), we substitute this factor by $(y_2sy_1)^{n+1}y_2$ and transform q into a new word q' in the same α-class h; this new word can be represented as $q' = y_1p_1'y_1 \cdot y_2p_2'y_2 \cdot q_2'$, where $p_1' = r(y_2sy_1)^{n-1}y_2s$, $p_2' = sy_1$, and q_2' is obtained from q_1 by removing its leftmost variable. Then y_1, y_2 do not occur in q_2', whence $|\operatorname{alph}(q_2')| < |\operatorname{alph}(q_1)|$. Now we can apply the same procedure to the leftmost variable of q_2 or q_2', and so on. On the i-th step of the procedure we create a new cell $y_ip_iy_i$ while the yet unprocessed "remainder" omits the variables $y_1, \ldots, y_i$. Clearly, the procedure terminates after a finite number of steps and yields a word of the form (3.1) in the α-class h.

Now let h^* be the α-class that contains the word

$$(p_ky_k)^{2n-2}p_k \cdot (p_{k-1}y_{k-1})^{2n-2}p_{k-1} \cdot \ldots \cdot (p_1y_1)^{2n-2}p_1.$$

We show that h^* is an inverse of h by induction on k. If $k = 1$, that is, $h = (y_1p_1y_1)^\alpha$, the α-class hh^*h contains the word

$$y_1p_1y_1 \cdot (p_1y_1)^{2n-2}p_1 \cdot y_1p_1y_1 = (y_1p_1)^{2n+1}y_1.$$

Applying the identity (1.4) if the word p_1 is empty and the identity (1.5) otherwise, we can transform this word to the word $y_1p_1y_1 \in h$. Thus, $hh^*h = h$. Similarly, the α-class h^*hh^* contains the word

$$(p_1y_1)^{2n-2}p_1 \cdot y_1p_1y_1 \cdot (p_1y_1)^{2n-2}p_1 = (p_1y_1)^{4n-2}p_1$$

that can be transformed to $(p_1y_1)^{2n-2}p_1 \in h^*$. Hence, $h^*hh^* = h^*$ and thus, h^* is an inverse of h.

For the induction step, suppose that $k > 1$ and let f and g be the α-classes containing the words $y_1p_1y_1$ and $y_2p_2y_2 \cdot \ldots \cdot y_kp_ky_k$ respectively. Then $h = fg$, $h^* = g^*f^*$ and, by the induction assumption, f^* and g^* are inverses of f and g, respectively. The equalities $ff^*f = f$ and $gg^*g = g$ imply that the α-classes f^*f and gg^* are idempotents. Taking into account that the idempotents of F/α commute due to the identity (1.6), we obtain

$$\begin{aligned} hh^*h &= fg \cdot g^*f^* \cdot fg \\ &= f(gg^*)(f^*f)g \\ &= f(f^*f)(gg^*) \\ &= ff^*f \cdot gg^*g \\ &= fg = h, \end{aligned} \qquad \begin{aligned} h^*hh^* &= g^*f^* \cdot fg \cdot g^*f^* \\ &= g^*(f^*f)(gg^*)f^* \\ &= g^*(gg^*)(f^*f)f^* \\ &= g^*gg^* \cdot f^*ff^* \\ &= g^*f^* = h^*. \end{aligned}$$

We see that h^* is an inverse of h, and the lemma is proved. □

Now we are in a position to discuss a gap in the original proof of Theorem 1 in [30] and to explain how the gap can be filled.

The proof of Theorem 1 in [30] develops as follows. As above, it works with F, the free semigroup of countable rank, and α, the fully invariant congruence on F that corresponds to the variety $\mathbf{A}$ defined by the identities (1.3)–(1.6). In the quotient semigroup F/α, one considers the set H of all α-classes containing a repeated word. Obviously, the product of two repeated words is a repeated word whence H is a subsemigroup of F/α. The idempotents of H commute because H, being a subsemigroup of F/α, satisfies the identity(1.6). By Lemma 7 (which appears in [30] as a part of the proof of Theorem 1), H is regular. Now one can apply the textbook fact that a regular semigroup with commuting idempotents is inverse, see [5, Theorem 1.17] or [8, Theorem 5.1.1]. Thus, H is an inverse subsemigroup of F/α. At this point, the proof under discussion invokes the main result from Kleiman's paper [13], which implies that the identities (1.3)–(1.6) form a basis for the inverse identities of the Brandt semigroup $B(G, I)$. In particular, these identities hold in $B(G, I)$ whence $\mathbf{A} \supseteq \mathbf{B}$, where as above, $\mathbf{B}$ stands for the variety generated by $B(G, I)$. In the language of fully invariant congruences this means that $\alpha \subseteq \beta$, where β denotes the fully invariant congruence on F that corresponds to the variety $\mathbf{B}$. Let β/α be the induced congruence on F/α so that $(F/\alpha)\,/\,(\beta/\alpha) \cong F/\beta$. The rest of the proof relies on the following claim: *the congruence β/α separates the elements of the subsemigroup H*, that is, β/α restricted to H is the equality relation. In [30] this claim is justified by observing that H lies in the variety $\mathbf{B}$—this follows from the fact that H is inverse and satisfies the identities (1.3)–(1.6) which, according to the quoted result from [13], define the variety of inverse semigroups generated by $B(G, I)$. However, the justification is not sufficient. The membership $H \in \mathbf{B}$ only guarantees that the **least** element in the set Γ

of all congruences γ on H with $H/\gamma \in \mathbf{B}$ is the equality relation; while β/α restricted to H is a congruence in Γ, it is not immediately clear that the restriction is indeed the least element in Γ.

Let us show that the italicized claim does hold. Arguing by contradiction, assume that some distinct elements $a, b \in H$ satisfy $(a, b) \in \beta/\alpha$. Since a and b are distinct regular elements of the semigroup F/α, which satisfies the identities (1.5) and (1.6), Lemma 5 applies. Thus, a and b are separated by an onto homomorphism $\chi\colon F/\alpha \to T$, where T is an inverse completely [0]-simple semigroup. Arguing as in the last paragraph of the above proof of Theorem 1 modulo Lemma 7, one can show that T lies in the variety $\mathbf{B}$. Then the homomorphism χ must factor through the natural homomorphism $\eta\colon F/\alpha \to F/\beta$ because F/β is the $\mathbf{B}$-free semigroup of countable rank. However, $a\eta = b\eta$ since $(a, b) \in \beta/\alpha$ while $a\chi \neq b\chi$, a contradiction.

4. Corollaries and discussions

For the reader's convenience, we reproduce the main corollaries of Theorem 1, following [30]. The first of them specializes Theorem 1, providing an explicit identity basis for Brandt semigroups over abelian groups of finite exponent.

Corollary 1 [30, Corollary 1]. *Let G be an abelian group of exponent $n > 1$ and I a set with at least 2 elements. The identities (1.4), (1.5), and*

$$x^2y^2 = y^2x^2, \tag{4.1}$$

$$xyxzx = xzxyx \tag{4.2}$$

constitute a basis for plain identities of the Brandt semigroup $B(G, I)$.

This is in fact a consequence of the proof of Theorem 1 rather than the theorem itself. The corresponding arguments were omitted in [30]; therefore, we provide a proof outline here.

P r o o f (outline). First, we show that the identities (1.4), (1.5), (4.1), (4.2) hold in $B(G, I)$. By the "if" part of Lemma 1, it suffices to verify that they hold in both G and the 5-element Brandt semigroup B_2. Obviously, the identities (1.4) and (1.5) hold in every group of exponent n while the identities (4.1) and (4.2) hold in every abelian group. Thus, (1.4), (1.5), (4.1), (4.2) hold in G. Inspecting the identity basis (1.2), one readily sees that (1.4), (1.5), (4.1) hold in B_2. The identity (4.2) also holds in B_2 as the following calculation shows:

$$\begin{aligned} xyxzx &= (xy)^2(xz)^2x && \text{in view of } xyx = xyxyx \\ &= (xz)^2(xy)^2x && \text{in view of } x^2y^2 = y^2x^2 \\ &= xzxyx && \text{in view of } xyx = xyxyx. \end{aligned}$$

Now we proceed exactly as in the proof of Theorem 1. Denote by $\mathbf{A}$ the semigroup variety defined by the identities (1.4), (1.5), (4.1), (4.2) and let $\mathbf{B}$ be the variety generated by the semigroup $B(G, I)$. The fact that $B(G, I)$ satisfies (1.4), (1.5), (4.1), (4.2) implies that $\mathbf{B} \subseteq \mathbf{A}$. Assuming that the inclusion is strict, choose an identity $u = v$ with the least value of $|\operatorname{alph}(u)|$ such that $u = v$ holds in $B(G, I)$ but fails in $\mathbf{A}$. Then the words u and v are repeated due to the argument in the 4th paragraph of Section 3.

Let F be the free semigroup of countable rank and α its fully invariant congruence corresponding to the variety $\mathbf{A}$. The α-classes u^α and v^α are distinct elements of F/α and, by Lemma 7, they are regular. Then Lemmas 5 and 6 imply that u^α and v^α are separated by an onto homomorphism $\chi\colon F/\alpha \to T$, where T is either a group, or a group with zero, or a Brandt semigroup. Let Q stand for the structure group of T in the latter case and for T or $T \setminus \{0\}$ in the two former cases. Then Q is a subgroup of a homomorphic image of F/α, whence the identities (1.4) and (4.2) hold in Q.

Clearly, the exponent of every group satisfying (1.4) divides n and every group satisfying (4.2) is abelian. Thus, Q is an abelian group of exponent dividing n. A well known classification of abelian group varieties (cf. [20, Theorem 19.5] or [21, Item 13.51]) ensures that the variety of all abelian groups of exponent dividing n is generated by any abelian group of exponent n, in particular, by the structure group G of $B(G, I)$. Thus, Q belongs to the variety generated by G, and hence, to the variety $\mathbf{B}$. As the 5-element Brandt semigroup B_2 also belongs to $\mathbf{B}$, the "if" part of Lemma 1 implies that every Brandt semigroup over Q lies in $\mathbf{B}$. From this, we have $T \in \mathbf{B}$ whence T must satisfy $u = v$. On the other hand, the composition of the natural homomorphism $F \to F/\alpha$ with the homomorphism $\chi\colon F/\alpha \to T$ separates u and v in T, a contradiction. □

Remark 3. We do not know any basis for plain identities of the Brandt semigroup over the infinite cyclic group $\mathbb{Z}$ (or any other abelian group of infinite exponent); moreover, it is not known whether or not the plain identities of this semigroup admit a finite basis. A finite basis for inverse identities of the Brandt semigroup over $\mathbb{Z}$ can be found in [13, Corollary 6] or [23, Theorem XII.5.4(iii)].

In connection with Remark 3, it appears appropriate to discuss in more detail how the *finite basis property*, i.e., the property of a Brandt semigroup $B(G, I)$ to have a finite identity basis, may depend on the type of identities—inverse or plain—under consideration. It turns out that the picture is rather non-trivial here. On the one hand, the additional operation increases the expressivity of the equational language so that the inverse identities of $B(G, I)$ are "richer" than the plain ones. This indicates that $B(G, I)$ may have more chances to possess no finite basis for its inverse identities. On the other hand, the inference power of the language increases too. Hence one can encounter the situation when some identity of $B(G, I)$ does not follow from an identity system Σ as a "plain" identity but follows from Σ as an "inverse" identity. This indicates that the inverse identities of $B(G, I)$ may admit a finite basis even if its plain identities do not. The cumulative effect of the trade-off between increased expressivity and increased inference power is hard to predict in general, as the following examples demonstrate[3].

Example 1. Let G be the wreath product of the countably generated free group of exponent 4 with the countably generated free abelian group and I a set with at least 2 elements. The Brandt semigroup $B(G, I)$ satisfies only trivial plain identities but its inverse identities have no finite basis.

P r o o f. The fact that $B(G, I)$ satisfies only trivial plain identities follows from the observation that G contains the countably generated free semigroup as a subsemigroup, see, e.g., [1]. If we assume that the inverse identities of $B(G, I)$ admit a finite basis, then appending the identity $xx^{-1} = yy^{-1}$ to the basis would yield a finite basis of group identities of the group G. However, by [20, Corollary 22.22] G generates the varietal product of the variety of all groups of exponent dividing 4 with the variety of all abelian groups, and by [14, Remark 2] this product possesses no finite identity basis, a contradiction. □

In Example 1, an increase in the expressivity of the equational language dominates; now we exhibit an "opposite" example in which one sees the effect of an increase in the inference power.

Example 2. Let G be the direct product of the infinite cyclic group $\mathbb{Z}$ with the group $\mathbb{S}_3$ of all permutations of a 3-element set and I a set with at least 2 elements. The Brandt semigroup $B(G, I)$ admits a finite basis of inverse identities but its plain identities have no finite basis.

[3] Our examples are adaptations of known ones (see, e.g., [31, Section 2]) to the case of Brandt semigroups.

P r o o f. Since the group $\mathbb{S}_3$ is metabelian, so is $G = \mathbb{Z} \times \mathbb{S}_3$. It is known [6] that the group identities of any metabelian group possess a finite basis. By [13, Corollary 2], the inverse identities of a Brandt semigroup admit a finite basis whenever so do the group identities of its structure group. Thus, we may conclude that $B(G, I)$ has a finite basis of inverse identities.

Now consider the following series of identities:

$$L_n: \ x^2 y_1 \cdots y_n y_n \cdots y_1 = y_1 \cdots y_n y_n \cdots y_1 x^2, \quad n = 1, 2, \ldots .$$

We aim to show that all identities L_n hold in $B(G, I)$. Due to the "if" part of Lemma 1, it amounts to verifying that they hold in both G and the 5-element Brandt semigroup B_2. Since the group $\mathbb{S}_3$ satisfies the identity (4.1), this identity, which is equivalent to L_1, holds in $G = \mathbb{Z} \times \mathbb{S}_3$. Now it easy to verify that G satisfies the identity L_n by induction on n. Indeed, for $n > 1$ we have

$$\begin{aligned}
x^2 y_1 y_2 \cdots y_n y_n \cdots y_2 y_1 &= y_1 (y_1^{-1} x y_1)^2 y_2 \cdots y_n y_n \cdots y_2 y_1 \\
&= y_1 y_2 \cdots y_n y_n \cdots y_2 (y_1^{-1} x y_1)^2 y_1 && \text{by the inductive assumption} \\
&= y_1 y_2 \cdots y_n y_n \cdots y_2 y_1^{-1} x^2 y_1^2 \\
&= y_1 y_2 \cdots y_n y_n \cdots y_2 y_1^{-1} y_1^2 x^2 && \text{by using (4.1)} \\
&= y_1 y_2 \cdots y_n y_n \cdots y_2 y_1 x^2.
\end{aligned}$$

In order to show that each of the identities L_n holds in $B_2 = B(E, \{1, 2\})$, it suffices to observe that the values of the words $x^2 y_1 \cdots y_n y_n \cdots y_1$ and $y_1 \cdots y_n y_n \cdots y_1 x^2$ under every evaluation $\varphi\colon \{x, y_1, \ldots, y_n\} \to B_2$ are equal to 0 unless $x\varphi = y_k\varphi = (1, 1, 1)$ or $x\varphi = y_k\varphi = (2, 1, 2)$ for all $k = 1, \ldots, n$, in which case the values of these words are equal to $(1, 1, 1)$ or $(2, 1, 2)$ respectively.

Isbell [9] proved that no finite set of plain semigroup identities true in the groups $\mathbb{Z}$ and $\mathbb{S}_3$ implies all identities L_n. Hence, the plain identities of $B(G, I)$ admit no finite basis. □

Our next result also deals with the finite basis property. It immediately follows from Theorem 1.

Corollary 2 [30, Corollary 2]. *If a group G of finite exponent admits a finite identity basis, then so does every Brandt semigroup over G.*

In particular, since every finite group possesses a finite identity basis ([22], see also [21, Section 5.2]), we conclude that the plain identities of each finite Brandt semigroup have a finite basis.

Two algebraic structures of the same type are said to be *equationally equivalent* if they satisfy the same identities. Results in [13], see also [23, Proposition XII.4.13], imply that the following dichotomy holds for an arbitrary inverse semigroup S: either

(1) S is equationally equivalent to an inverse semigroup that is either a group, or a group with zero, or a Brandt semigroup and that can be chosen to be finite whenever S is finite, or

(2) the inverse semigroup variety generated by S contains the 6-*element Brandt monoid* B_2^1 obtained by adjoining a "fresh" symbol 1 to the 5-element Brandt semigroup B_2 and extending the multiplication of B_2 so that 1 becomes the identity element.

If S and T are inverse semigroups and S satisfies all inverse identities of T, then the same holds for the plain identities of T since the latter are special instances of the former. (In the language of varieties, this means that S lies in the semigroup variety generated by T whenever it belongs to the inverse semigroup variety generated by T.) In particular, if S and T are equationally equivalent as inverse semigroups, they are equationally equivalent as plain semigroups as well. In view of these observations, we see that the above dichotomy persists if one considers plain semigroup identities and varieties. Thus, if S is an arbitrary inverse semigroup, then either

(1') S is equationally equivalent as a plain semigroup to either a group, or a group with zero, or a Brandt semigroup, each of which can be chosen to be finite whenever S is finite, or

(2') the plain semigroup variety generated by S contains the 6-element Brandt monoid B_2^1.

This dichotomy, combined with a powerful result by Sapir [25], allows us to give the following classification of finite inverse semigroups with respect to the finite basis property.

Corollary 3 [30, Corollary 3]. *A finite inverse semigroup S admits a finite basis of plain identities if and only if the plain semigroup variety generated by S excludes the monoid B_2^1.*

P r o o f. The "only if" part follows from [25, Corollary 6.1], according to which every (not necessarily inverse) finite semigroup that generates a variety containing B_2^1 has no finite identity basis. For the proof of the "if" part, we invoke the above dichotomy that allows us to assume that S is either a finite group, or a finite group with zero, or a finite Brandt semigroup. We have already mentioned that every finite group possesses a finite identity basis, and so does every finite Brandt semigroup by Corollary 2. The remaining case of finite groups with zero easily follows from a general result by Melnik [18, Theorem 4] ensuring that if a (not necessarily finite) semigroup T has a finite identity basis, then so does the semigroup T^0. (See [31, Section 3] for a detailed explanation of how [18, Theorem 4] implies this claim.) □

Remark 4. As it has been observed by Kalicki [12], there exists an algorithm to decide, given two finite algebraic structures of the same type, whether one of them belongs to the variety generated by the other. Hence, Corollary 3 provides an algorithm to decide whether or not a given finite **inverse** semigroup admits a finite basis of **plain** identities. Recall that the existence of such an algorithm remains open for each of the following two situations: when one wants to decide whether or not a given finite **plain** semigroup admits a finite basis of **plain** identities (see [31, Section 2] for a discussion) and when one wants to decide whether or not a given finite **inverse** semigroup admits a finite basis of **inverse** identities. In particular, it is not known if for a finite inverse semigroup S, the plain and the inverse versions of the finite basis property are equivalent. Kaďourek [10] has proved that they are equivalent provided that all subgroups of S are solvable.

Acknowledgements

The author thanks Dr. Jiří Kaďourek who carefully examined a number of publications of the 1980s, a notable "Sturm und Drang" period in the theory of semigroup varieties (and corrected inaccuracies in some of these publications, see, e.g., [11]). In the course of his critical studies, Dr. Kaďourek observed a gap in [30] and drew the author's attention to the fact that this gap had not been properly discussed in the literature. The present paper is a response to this fair remark.

REFERENCES

1. Belyaev V. V., Sesekin N. F. Free subsemigroups in soluble groups. *Ural. Gos. Univ. Mat. Zap.*, 1981. Vol. 12, No. 3. P. 13–18. (In Russian)
2. Brandt H. Über eine Verallgemeinerung des Gruppenbegriffes. *Math. Ann.*, 1927. Vol. 96, No. 1. P. 360–366. DOI: 10.1007/BF01209171
3. Burris S., Sankappanavar H. P. *A Course in Universal Algebra*. Berlin–Heidelberg–New York: Springer-Verlag, 1981. xvi+276 p.
4. Clifford, A. H. Matrix representations of completely simple semigroups. *Amer. J. Math.*, 1942. Vol. 64, No. 1. P. 327–342. DOI: 10.2307/2371687
5. Clifford A. H., Preston G. B. *The Algebraic Theory of Semigroups*. Vol. I. 2nd ed. Providence, RI: Amer. Math. Soc., 1964. xvi+224 p.

6. Cohen D. E. On the laws of a metabelian variety. *J. Algebra*, 1967. Vol. 5, No. 3. P. 267–273. DOI: 10.1016/0021-8693(67)90039-7
7. Hall T. E., Kublanovskii S. I., Margolis S., Sapir M. V., Trotter P. G. Algorithmic problems for finite groups and finite 0-simple semigroups. *J. Pure Appl. Algebra*, 1997. Vol. 119, No. 1. P. 75–96. DOI: 10.1016/S0022-4049(96)00050-3
8. Howie J. M. *Fundamentals of Semigroup Theory*. 2nd ed. Oxford: Clarendon Press, 1995. xvi+352 p.
9. Isbell J. R. Two examples in varieties of monoids. *Proc. Cambridge Philos. Soc.*, 1970. Vol. 68, No. 2. P. 265–266. DOI: 10.1017/S0305004100046065
10. Kaďourek J. On bases of identities of finite inverse semigroups with solvable subgroups. *Semigroup Forum*, 2003. Vol. 67, No. 3. P. 317–343. DOI: 10.1007/s00233-001-0005-x
11. Kaďourek J. On finite completely simple semigroups having no finite basis of identities. *Semigroup Forum*, 2018. Vol. 97, No. 1. P. 154–161. DOI: 10.1007/s00233-017-9907-0
12. Kalicki J. On comparison of finite algebras. *Proc. Amer. Math. Soc.*, 1952. Vol. 3, No. 1. P. 36–40. DOI: 10.2307/2032452
13. Kleiman E. I. On bases of identities of Brandt semigroups. *Semigroup Forum*, 1977. Vol. 13, No. 3. P. 209–218. DOI: 10.1007/BF02194938
14. Kleiman Ju. G. On a basis of the product of varieties of groups. *Math. USSR. Izv.*, 1973. Vol. 7, No. 1. P. 91–94. DOI: 10.1070/IM1973v007n01ABEH001927
15. Lee E. W. H. Finite basis problem for semigroups of order five or less: generalization and revisitation. *Studia Logica*, 2013. Vol. 101, No. 1. P. 95–115. DOI: 10.1007/s11225-012-9369-z
16. Lee E. W. H., Volkov M. V. On the structure of the lattice of combinatorial Rees–Sushkevich varieties. *Semigroups and Formal Languages*. Hackensack, NJ: World Sci. Publ., 2007. P. 164–187. DOI: 10.1142/9789812708700_0012
17. Mashevitzky G. I. Identities in Brandt semigroups. *Polugruppovye mnogoobrazija i polugruppy endomorfizmov* [Semigroup varieties and semigroups of endomorphisms]. Leningrad: Leningrad State Pedagogical Institute, 1979. P. 126–137. (In Russian)
18. Mel'nik I. I. On varieties and lattices of varieties of semigroups. *Issledovaniya po algebre* [Investigations in algebra]. Saratov: Saratov State Univ., 1970. Vol. 2. P. 47–57. (In Russian)
19. Munn W. D. Matrix representations of semigroups. *Proc. Cambrdige Philos. Soc.*, 1957. Vol. 53, No. 1. P. 5–12. DOI: 10.1017/S0305004100031935
20. Neumann B. H. Identical relations in groups. I. *Math. Ann.*, 1937. Vol. 114, No. 1. P. 506–525. DOI: 10.1007/BF01594191
21. Neumann H. *Varieties of groups*. Berlin–Heidelberg–New York: Springer–Verlag, 1967. xii+192 p.
22. Oates S., Powell M. B. Identical relations in finite groups. *J. Algebra*, 1964. Vol. 1, No. 1. P. 11–39. DOI: 10.1016/0021-8693(64)90004-3
23. Petrich M. *Inverse semigroups*. New York: John Wiley & Sons, 1984. xii+674 p.
24. Reilly N. R. The interval $[\mathbf{B}_2, \mathbf{NB}_2]$ in the lattice of Rees–Sushkevich varieties. *Algebra Universalis*, 2008. Vol. 59, No. 3-4. P. 345–363. DOI: 10.1007/s00012-008-2091-z
25. Sapir M. V. Problems of Burnside type and the finite basis property in varieties of semigroups. *Math. USSR. Izv.*, 1988. Vol. 30, No. 2. P. 295–314. DOI: 10.1070/IM1988v030n02ABEH001012
26. Shevrin L. N., Sukhanov E. V. Structural aspects of the theory of varieties of semigroups. *Soviet Math. (Iz. VUZ)*, 1989. Vol. 33, No. 6. P. 1–34.
27. Trahtman A. N. An identity basis of the five-element Brandt semigroup. *Ural. Gos. Univ. Mat. Zap.*, 1981. Vol. 12, No. 3. P. 147–149. (In Russian)
28. Trahtman A. N. The finite basis problem for semigroups of order less than six. *Semigroup Forum*, 1983. Vol. 27. P. 387–389. DOI: 10.1007/BF02572749
29. Trahtman A. N. Finiteness of identity bases of 5-element semigroups. *Polugruppy i ikh gomomorfizmy* [Semigroups and their Homomorphisms]. Leningrad: Leningrad State Pedagogical Institute, 1991. P. 76–97. (In Russian)
30. Volkov M. V. On the identity bases of Brandt semigroups. *Ural. Gos. Univ. Mat. Zap.*, 1985. Vol. 14, No. 1. P. 38–42. (In Russian)
31. Volkov M. V. The finite basis problem for finite semigroups. *Sci. Math. Japon.*, 2001, Vol. 53, No. 1. P. 171–199.
32. Volkov M. V. On a question by Edmond W. H. Lee. *Proc. Ural State Univ.*, 2005. No. 36 (*Mathematics and Mechanics*, No. 7). P. 167–178.

Amendment to my article

HARMONIC INTERPOLATING WAVELETS IN NEUMANN BOUNDARY VALUE PROBLEM IN A CIRCLE

Published in: Ural Mathematical Journal, 2019. Vol. 5, No. 1, pp. 91–100

DOI: 10.15826/umj.2019.1.009

In the article indicated, the following omission need to be amended. Namely, on the first page, a footnote to the article title with reference to grant "This work was supported by Russian Science Foundation (project no. 14-11-00702)" should be added.

Dmitry A. Yamkovoi
dmitriiyamkovoi@bk.ru

Editor: Tatiana F. Filippova

Managing Editor: Oxana G. Matviychuk

Design: Alexander R. Matviychuk

Contact Information

16 S. Kovalevskaya str., Ekaterinburg, Russia, 620990

Phone: +7 (343) 375-34-73

Fax: +7 (343) 374-25-81

Email: secretary@umjuran.ru

Web-site: https://umjuran.ru

N.N. Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences

Ural Federal University named after the first President of Russia B.N. Yeltsin

Distributed for free