PROGRESS IN ANALYSIS

Proceedings of the 8th Congress of the International Society for Analysis, its Applications, and Computation (22–27 August 2011)

Volume 2

V. I. Burenkov, M,L. Goldman,
E. B. Laneev, V. D. Stepanov
Editors

Moscow
Peoples' Friendship University of Russia
2012

УДК 517
ББК 22.1
К 64

Издание осуществлено при финансовой поддержке Российского фонда фундаментальных исследований (РФФИ) по проекту № 12-01-07116

Progress in Analysis. Proceedings of the 8th Congress of the International Society for Analysis, its Applications, and Computation. Vol. 2. — M.: Peoples' Friendship University of Russia, 2012. — 335 p.

ISBN 978-5-209-04590-8

The 8th Congress of the International Society for Analysis, its Applications, and Computation (ISAAC) was organized by Peoples' Friendship University of Russia, Division of Mathematics of the Russian Academy of Sciences, V. A. Steklov Institute of Mathematics of Russian Academy of Sciences and V. M. Lomonosov Moscow State University, and took place at Peoples' Friendship University of Russia, Moscow, through 22-27 August 2011.

The programme of the Congress included most of the topics of contemporary mathematical analysis, in particular, real, functional, complex analysis, operator theory, theory of ordinary differential equations, theory of partial differential equations, nonlinear analysis, optimization theory, variational analysis, approximation theory, applications of analysis (inverse problems, functional and difference equations, mathematics in medicine, stochastic analysis), teaching analysis at universities and schools, history of analysis.

Vol. 2 contains papers by the participants of Session III (Partial differential equations), Session IV (Applied analysis), and of Session V (Nonlinear analysis, extremal problems, approximation theory, and applications).

This volume would be of interest for mathematicians working in all main branches of contemporary mathematical analysis and its applications.

Prepared by the Organizing Committee of the 8th Congress of the ISAAC.

Contents

III. Partial Differential Equations

III.2. Dispersive equations

IV. Applied Analysis

IV.1. Inverse problems

IV.3. Medical mathematics

V. Nonlinear analysis, extremal problems, approximation theory and applications

V.1. Optimization theory, variational analysis and nonlinear analysis

V.2. Approximation theory and Fourier analysis

VIII. Other

III. Partial Differential Equations

III.2. Dispersive equations

(Sessions organizers: M. Reissig, M. Ruzhansky)

HYPERBOLIC-LIKE DISSIPATIONS FOR SECOND ORDER EQUATIONS

M. D'Abbicco, M. R. Ebert

Key words: Hyperbolic Equations, Dissipative Equations

AMS Mathematics Subject Classification: 35L15

Abstract. The main goal of this paper is to study the long time behavior of the energy for the second order hyperbolic equation with time-dependent coefficients of the form

$$u_{tt} - \lambda^2(t)u_{xx} + b(t)u_t + d(t)u_x = 0.$$

Under positivity assumptions of the coefficient $b(t)$ of the *damping term* we expect a dissipative effect, whereas the *drift term* $d(t)u_x$ should be controlled to avoid blow-up effects. We give a precise description of the behavior of the *local energy* in the *hyperbolic* zone of the extended phase space, then we look for sufficient conditions that guarantee the same energy estimate from above in all the extended phase space. We call this class of dissipations *hyperbolic-like* since the energy behavior is deeply depending on the hyperbolic structure of the equation. We refer the interested reader to [2], where it is also considered the case of a strong influence coming from the drift term $d(t)u_x$ and it is introduced a classification of mass terms $e(t)u$ which brings no contribution to the energy behavior. Moreover, some models of higher order equations are presented.

1 Introduction

In this paper, we consider the Cauchy problem for the second order linear hyperbolic equation with time-dependent coefficients of the form

$$\begin{cases} u_{tt} - \lambda^2(t)u_{xx} + b(t)u_t + d(t)u_x = 0\,, \quad t \geqslant 0, \\ u(0,x) = u_0(x), \quad u_t(0,x) = u_1(x), \end{cases} \tag{1}$$

with initial data $(u_0, u_1) \in H^1 \times L^2$ and we study the behavior of its λ-*energy*

$$E_\lambda(t) = \|u_t(t,\cdot)\|^2_{L^2} + \lambda^2(t)\|u_x(t,\cdot)\|^2_{L^2},$$

as $t \to \infty$. For the sake of simplicity we are going to consider (1) in one space dimension, but our reasoning can be easily extended to the case of space dimension $n \geqslant 2$.

Let $\lambda \equiv 1$ and $d \equiv 0$; if $b \equiv 0$ (wave equation), then the energy

$$E_1(t) = \|u_t(t,\cdot)\|_{L^2}^2 + \|u_x(t,\cdot)\|_{L^2}^2 ,$$

is constant, whereas if $b \equiv 1$ (classical damped wave equation), then the wave energy $E_1(t)$ dissipates as $t \to \infty$. In the latter case, a better decay is obtained thanks to the presence of the dissipative term u_t.

If one is interested in decay estimates for the second order homogeneous equation

$$u_{tt} - \lambda^2(t) u_{xx} = 0, \tag{2}$$

that is, the wave equation with variable speed of propagation, then one has to study properties of $\lambda = \lambda(t)$. Even if $0 < \lambda_0 \leqslant \lambda(t) \leqslant \lambda_2$, some difficulties arise for the energy estimates, since the oscillations of $\lambda = \lambda(t)$ have a deteriorating influence on the energy behavior [10]. Nevertheless if

$$|\lambda^{(k)}(t)| \leqslant C_k (1+t)^{-k}, \text{ for } k = 1, 2,$$

then the so-called *generalized energy conservation* property holds [9], that is,

$$C_0 E_1(0) \leqslant E_1(t) \leqslant C_1 E_1(0). \tag{3}$$

More recently (see [5–7, 9]), these energy estimates have been studied for (2) with speed $\lambda(t) \geqslant \lambda_0 > 0$, not bounded from above. By using an hypothesis of stabilization at $t = \infty$ and $\mathcal{C}^m$ regularity for $\lambda(t)$, it is possible to enlarge the class of admissible oscillations for $\lambda(t)$. In [7] it is derived the estimate

$$C_0(u_0, u_1) \leqslant \frac{1}{\lambda(t)} E_\lambda(t) \leqslant C_1 E(0), \tag{4}$$

where $C_0(u_0, u_1)$ is a positive constant that depends on the initial data, and the (Klein-Gordon) energy $E(0)$ is given by

$$E(0) = \|u_1\|_{L^2}^2 + \|u_0\|_{H^1}^2.$$

We remark that the L^2 norm of u_0 can not be neglected in (4) if $\lambda(t)$ is not bounded from above. We refer the interested reader to [3] and [4] too, where *generalized energy conservation* and *blow-up* effects are proved for 2 by 2 systems and some (dissipative) influence from the lower order term is considered.

If one is interested in the energy behavior for the wave equation with time-dependent dissipation

$$u_{tt} - u_{xx} + b(t)u_t = 0, \tag{5}$$

with $b(t)$ positive, then one has to explore properties of the dissipative term $b(t)$. In [11], J. Wirth derived $L^p - L^q$ estimates for the solution if among other things $\limsup_{t\to\infty} tb(t) < 1$; in particular, he proved that $E_1(t) \leqslant CE(0)\exp(-\int_0^t b(\tau)d\tau)$, and he called this a *non-effective dissipation*. Theorem 1 largely extends this energy estimate to a much more complex situation with a unified approach.

In this paper we describe new effects coming from the interaction between the wave speed $\lambda(t)$ and the term $b(t)$ in a very general setting. In particular, the function $\lambda(t)$ is not necessarily increasing and $b(t)$ is not necessarily positive; nevertheless $\lambda'(t) + b(t)\lambda(t)$ is positive. We also introduce some conditions to control the influence coming from the *drift* term $d(t)u_x$.

2 Hyperbolic-like dissipation for second order equations

We consider (1) and we assume $\lambda \in \mathcal{C}^2$, positive, $b \in \mathcal{C}^1$ real-valued, and $d \in \mathcal{C}^1$, may be complex-valued, in general. We assume that $\lambda(0) = 1$ and $\lambda(t) > 0$.

Hypothesis 1. *We define* $\Lambda(t) = 1 + \int_0^t \lambda(\tau)d\tau$, *and we assume that* $\Lambda(t)$ *remains not bounded as* $t \to \infty$, *that is,* $\lambda \notin L^1$. *We assume that the coefficients have* very slow oscillations, *that is,*

$$\frac{|\lambda^{(k)}(t)|}{\lambda(t)} + |b^{(k-1)}(t)| + |d^{(k-1)}(t)| \leqslant C\,(\eta(t))^k \text{ for } k = 1, 2, \tag{6}$$

where the function $\eta(t){:=}\lambda(t)/\Lambda(t)$ *will play a fundamental role in the following.*

In order to consider the real part of $d(t)$ as a weak perturbation which does not influence the energy behavior, we assume the following.

Hypothesis 2. *We assume that the real part of* $d(t)$ *satisfies the following:*

$$-C \leqslant \int_0^t \Re d(\tau)\, d\tau \leqslant C. \tag{7}$$

Definition 1. We put

$$\gamma(t):=\exp\Big(-\int_0^t b(\tau)d\tau\Big),\quad \Gamma(t):=1+\int_0^t \gamma(\tau)\,d\tau.$$

If $\gamma\in L^1$, that is, $\Gamma(t)$ remains bounded as $t\to\infty$, we define $\Gamma^\sharp(s):=\int_s^\infty \gamma(\tau)\,d\tau$.

Hypothesis 3. *We assume that $\lambda(t)$ and $b(t)$ satisfy the following conditions:*

$$\frac{\lambda(t)}{\gamma(t)} \quad \textit{is increasing, i.e.} \quad \lambda'(t)+b(t)\lambda(t)\geqslant 0, \tag{8}$$

$$\frac{\eta(t)\Gamma(t)}{\sqrt{\lambda(t)\gamma(t)}}\leqslant C. \tag{9}$$

In particular, if $\gamma\in L^1$, then the conditions (8)-(9) *hold if*

$$0\leqslant \lambda'(t)+b(t)\lambda(t)\leqslant 2\lambda(t)\eta(t)\,. \tag{10}$$

If the equation in (1) *satisfies* (8) *and* (9)*, we say that it is* hyperbolic-like dissipative.

Remark 1. We remark that (9) is related to the request to have an estimate of the *local energy* for small frequencies (i.e. in the *pseudo-differential* zone, see the proof of Theorem 1) which is not worst than the estimate obtained for large frequencies (i.e. in the *hyperbolic* zone). If we relax (9) we can still prove some dissipative effect for the energy, but it will be not *hyperbolic-like*.

Hypothesis 4. *We assume that there exist two functions $g(t)$ and $h(s)$ such that*

$$\int_s^t \gamma(\tau)\,d\tau\leqslant g(t)h(s),\quad \frac{\eta(t)g(t)}{\sqrt{\lambda(t)\gamma(t)}}\leqslant C, \tag{11}$$

where $g(t)\geqslant\Gamma(t)$ is increasing, $g(0)=1$ and $h(s)$ is decreasing, $g(t)h(t)\leqslant\Gamma(t)$, and

$$d(t)\leqslant C\frac{\gamma(t)}{g(t)h(t)}. \tag{12}$$

Remark 2. If $\gamma\notin L^1$, in most of the cases it is sufficient to take $g(t)=\Gamma(t)$ and $h(s)=1$ in (11), whereas if $\gamma\in L^1$, then we can take $g(t)=1$ and $h(s)=\Gamma^\sharp(s)$

in (11). Indeed (11) is satisfied thanks to (9). Nevertheless, in some cases it may be convenient to choose different $g(t)$ and $h(s)$ such that $g(t)h(t) \ll \Gamma(t)$ for large t (see Example 2), so that condition (12) is less restrictive.

Theorem 1. *We assume the Hypotheses 1 to 4. Then the following* a priori *energy estimate holds:*

$$E_\lambda(t) \leqslant C\lambda(t)\gamma(t)E(0). \tag{13}$$

In particular, from Theorem 1 it follows that $\|u_x(t,\cdot)\|_{L^2}$ is bounded from above by a decreasing function since

$$\|u_x(t,\cdot)\|_{L^2}^2 \leqslant CE(0)\gamma(t)/\lambda(t).$$

3 Examples

For the sake of brevity, we present only a few examples.

Example 1. It is easy to prove that $\Gamma(t) \leqslant C\Lambda(t)$. Therefore it makes sense to consider which dissipative effect we can expect if $\Gamma(t) = 1 + (\Lambda^\kappa(t) - 1)/\kappa$ for some $\kappa \in (0, 1]$. It follows $\gamma(t) = \Gamma'(t) = \lambda(t)/\Lambda^{1-\kappa}(t)$ and

$$b(t) = -(\log\gamma(t))' = ((1-\kappa)\log\Lambda - \log\lambda)' = (1-\kappa)\frac{\lambda(t)}{\Lambda(t)} - \frac{\lambda'(t)}{\lambda(t)}.$$

Condition (8) holds true since

$$\lambda'(t) + \lambda(t)b(t) = (1-\kappa)\lambda(t)\eta(t) \geqslant 0,$$

whereas condition (9) holds true since $\gamma/\Gamma \approx \lambda/\Lambda$. If $|d(t)| \leqslant C\eta(t)$ as in Hypothesis 1, then (12) trivially holds for $(g, h) = (\Gamma, 1)$.

Now let $\Gamma^\sharp(t) = (\kappa\Lambda^\kappa(t))^{-1}$ for some $\kappa \in [-1, 0)$. Then $\gamma(t)$, $b(t)$ and $\lambda'(t) + \lambda(t)b(t)$ are as above. Condition (10) is satisfied since $0 \leqslant 1-\kappa \leqslant 2$, and if $|d(t)| \leqslant C\eta(t)$ as in Hypothesis 1, then (12) trivially holds for $(g, h) = (1, \Gamma^\sharp)$.

Example 2. Let $b(t) = -\eta'(t)/\eta(t)$. It follows $\gamma(t) = \eta(t)$ and $\Gamma(t) = 1 + \log\Lambda(t)$. Therefore, we have

$$\frac{\lambda}{\gamma} = \Lambda, \quad \frac{\eta}{\sqrt{\lambda}}\frac{\Gamma}{\sqrt{\gamma}} = \frac{1+\log\Lambda}{\sqrt{\Lambda}},$$

hence, conditions (8) and (9) hold. Nevertheless, we remark that

$$\frac{\gamma}{\Gamma} = \frac{\lambda}{\Lambda(1+\log\Lambda)} \not\approx \frac{\lambda}{\Lambda}.$$

If we take $g = \Gamma$ and $h = 1$, then condition (12) on $d(t)$ is more restrictive than $|d(t)| \leqslant C\eta(t)$ as in Hypothesis 1. Therefore, we fix $\epsilon \in (0, 1/2)$ and we take $g(t) = \Lambda^{\epsilon}(t)$ and $h(s) = \Lambda^{-\epsilon}(s)/\epsilon$ in (11). With this choice condition (12) holds true (see Remark 2).

Examples 1 and 2 are applicable for any $\lambda(t)$ which satisfies our assumptions. For the sake of clarity we apply our reasoning to two specific choices of $\lambda(t)$.

Example 3 (Polynomial growth). Let $\lambda(t) = (1+t)^p$ for some $p > -1$; then the function $\Lambda(t)$ has a *polynomial* growth, i.e. $\Lambda(t) \approx (1+t)^{p+1}$. The case of polynomial growth has many special properties, therefore it is the easiest to manage. Let $b(t) = \mu/(1+t)$, for some $\mu \in \mathbb{R}$; we remark that the case $b \equiv 0$ is included. If $\mu = 1$ we are, asymptotically speaking, in the case considered in Example 2, i.e, $b(t) \asymp -\eta'(t)/\eta(t)$. Otherwise we can follow Example 1 for $\kappa = (1-\mu)/(p+1)$. Therefore, Hypothesis 3 is satisfied if and only if $|\kappa| \leqslant 1$, that is, $-p \leqslant \mu \leqslant p+2$. Concerning the term $d(t)$, it is sufficient to assume $|d(t)| \leqslant C/(1+t)$ as in Hypothesis 1.

Example 4 (Exponential growth). Let $\lambda(t) = e^{pt}$ for some $p > 0$; then $\Lambda(t) \approx \lambda(t)$. Condition (6) is satisfied if b, b', d, d' are bounded. Let $b \equiv \mu$ for some $\mu \neq 0$. We can follow Example 1 for $\kappa = -\mu/p$. Therefore, Hypothesis 3 is satisfied if and only if $|\mu| \leqslant p$.

4 Brief sketch of the proof of Theorem 1

We perform the Fourier transform of (1) and we claim that $\mathcal{E}_\lambda(t,\xi) \leqslant C\lambda(t)\gamma(t)\mathcal{E}_0(\xi)$ uniformly with respect to $\xi \in \mathbb{R}$, where

$$\mathcal{E}_\lambda(t,\xi) = |\widehat{u}_t(t,\xi)|^2 + |\xi|^2\lambda^2(t)|\widehat{u}(t,\xi)|^2,$$
$$\mathcal{E}_0(\xi) = |\widehat{u_1}(\xi)|^2 + (1+|\xi|^2)|\widehat{u_0}(t,\xi)|^2.$$

To prove this claim we divide the extended phase space into the *pseudo-differential zone* $Z_{\mathrm{pd}}(N)$ and into the *hyperbolic zone* $Z_{\mathrm{hyp}}(N)$, where the separating curve is given by $\Lambda(t)|\xi| = N$ for a suitable $N > 0$. In $Z_{\mathrm{hyp}}(N)$ we consider the system for the micro-energy $U = (i\xi\lambda(t)b(t)\widehat{u}, \widehat{u}_t)$. We diagonalize its principal part, and then we apply a *refined diagonalization* of the lower-order term, which allow us to reduce the order of the non-diagonal part of the pseudo-differential system. This

refined diagonalizer is explicitly constructed by using Hypothesis 1, together with the condition $\Lambda(t)|\xi| \geqslant N$, which holds in $Z_{\rm hyp}(N)$. Namely, we derive:

$$\partial_t W = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \varphi(t,\xi) W + f(t) W + J(t,\xi) W, \tag{14}$$

where $\varphi(t,\xi)$ and $f(t)$ are scalar functions given by

$$\varphi(t,\xi) = i\xi\lambda(t)b(t) - d(t), \quad f(t) = \frac{1}{2}\left(\frac{\lambda'(t)}{\lambda(t)} - b(t)\right),$$

and the matrix $J(t,\xi)$ is integrable in $Z_{\rm hyp}(N)$. Since $\lambda(t)$ is real-valued and $\Re d(t)$ satisfies Hypothesis 2, the only contribution to the behavior of $|W(t,\xi)|$ is given by $\exp\int f(t)\,dt$, and we prove our claim in $Z_{\rm hyp}(N)$. In $Z_{\rm pd}(N)$ we introduce the micro-energy $V = (i\eta(t)\widehat{u}, \widehat{u}_t)$, and we set $V = \sqrt{\lambda(t)\gamma(t)}\widetilde{V}$. In this way we derive

$$\partial_t \widetilde{V} = \mathcal{A}(t,\xi)\widetilde{V} \equiv \begin{pmatrix} \dfrac{\eta'}{\eta} + \dfrac{b}{2} - \dfrac{\lambda'}{2\lambda} & i\eta \\ \dfrac{i\xi^2\lambda^2 - \xi\lambda d}{\eta} & -\dfrac{b}{2} - \dfrac{\lambda'}{2\lambda} \end{pmatrix} \tilde{V}. \tag{15}$$

To prove our claim in $Z_{\rm pd}(N)$, we have to prove that the fundamental solution $E(t,\xi)$ of (15), which satisfies $\partial_t E = \mathcal{A}(t,\xi)E$ with initial condition $E(0,\xi) = {\rm Id}$, remains bounded as $t \to \infty$. We put $E = (E_{ij})_{i,j=1,2}$ and we write down a system of integral equations. We transform this system into two Volterra-type integral equations and we derive boundedness of the solution by using integral Gronwall-like inequalities [3]. The main difficulty arises into get estimates as sharp as possible of the kernels of these equations. For instance, for $E_{1j}(t,\xi)$ we have

$$E_{1j} = \frac{\eta(t)}{\sqrt{\gamma(t)\lambda(t)}}\,\delta_{1j} + i\,\frac{\eta(t)(\Gamma(t)-1)}{\sqrt{\gamma(t)\lambda(t)}}\,\delta_{2j} + \\ + \frac{i\eta(t)}{\sqrt{\gamma(t)\lambda(t)}} \int_0^t \gamma(\tau) \int_0^\tau \sqrt{\frac{\lambda(\sigma)}{\gamma(\sigma)}}\,\frac{i\xi^2\lambda^2(\sigma) - \xi\lambda(\sigma)d(\sigma)}{\eta(\sigma)} E_{1j}(\sigma,\xi)\,d\sigma\,d\tau.$$

The boundedness of $E(t,\xi)$ then follows from Hypotheses 3 and 4. Once we proved boundedness of $E(t,\xi)$, that is, our claim in $Z_{\rm pd}(N)$, we conclude the proof by integrating over ξ and by using Plancherel's theorem.

References

1. D. Bainov and P. Simeonov : *Integral inequalities and applications. Mathematics and its Applications.* Kluver Academic Publishers, Boston, 2010. 245 pages.
2. M. D'Abbicco, M.R. Ebert : *Hyperbolic-like Dissipative Effects for Equations.* Preprint, submitted. 30 pages.
3. M. D'Abbicco, M. Reissig : *Long time asymptotics for 2 by 2 hyperbolic systems.* J. Differential Equations **250**, 2, 2011. Pp. 752–781.
4. M. D'Abbicco, M. Reissig : *Blow-up of the energy at infinity for 2 by 2 systems.* J. Differential Equations **252**, 1, 2012. Pp. 477–504.
5. M.R. Ebert, M.Reissig : *The influence of oscillations on global existence for a class of semi-linear wave equations.* Math. Meth. Appl. Sci. **34**, 2011. Pp. 1289–1307.
6. F. Hirosawa : *On the asymptotic behavior of the energy for the wave equation with time depending coefficients.* Math. Ann. **339**, 2007. Pp. 819–838.
7. F. Hirosawa, J.Wirth : *Generalised energy conservation law for wave equations with variable propagation speed.* J. Math. Anal. Appl. **358**, 2009. Pp. 56–74.
8. M. Reissig : *Optimality of the asymptotic behavior of the energy for wave models. Modern Aspects of the Theory of PDE; Operator Theory: Adv. and Appl.* **216**, Birkhäuser, Boston, 2011. Pp. 291–315.
9. M. Reissig, J. Smith : $L^p - L^q$ *estimate for wave equation with bounded time dependent coefficient.* Hokkaido Math. J. **34**, 2005. Pp. 541–586.
10. M. Reissig, K. Yagdjian : *About the influence of oscillations on Strichartz-type decay estimates.* Rend. Sem. Mat. Univ. Pol. Torino **58**, 2000. Pp. 375–388.
11. J. Wirth : *Wave equations with time-dependent dissipation I. Non-effective dissipation.* J. Differential Equations **222**, 2006. Pp. 487–514.

M. D'Abbicco
Department of Mathematics, University of Bari, Italy, 70125, Bari, Via E. Orabona 4, dabbicco@dm.uniba.it

M. R. Ebert
Departamento de Física e Matemática, Universidade de São Paulo (USP), FFCLRP, Av.Bandeirantes, 3900,Campus da USP,CEP 14040-901, Ribeirão Preto - SP - Brasil, ebert@ffclrp.usp.br

SOME BVP FOR GENERAL SEMILINEAR PARABOLIC EQUATION IN DIVERGENCE FORM

Kerime Kalli, Kamal N. Soltanov

Key words: Semilinear Parabolic Equation, Third type Boundary Value Problem, Existence and Uniqueness Theorems, Sub-linear, Linear and Super-Linear Cases.

AMS Mathematics Subject Classification: 35D30, 35K58, 35M12

Abstract. In this paper we study some third type boundary value problems for a general semilinear parabolic equation in divergence form. The existence of a generalized solution for the considered problem is proved under sufficient conditions and also the uniqueness of the solution of the considered problem is proved in a model case.

1 Introduction

Consider the problem

$$\frac{\partial u}{\partial t} + Lu + g(x,t,u) = h(x,t), \quad (x,t) \in Q_T \equiv \Omega \times (0,T), \tag{1}$$

$$u(x,0) = 0, \quad x \in \Omega \subset \mathbb{R}^n, n \geqslant 3, \tag{2}$$

$$\left(\frac{\partial u}{\partial \nu} + k(x,t)u\right)\Big|_{\Gamma_T} = \varphi(x,t), \quad \Gamma_T \equiv \partial\Omega \times [0,T], \quad T > 0. \tag{3}$$

Here Ω is a bounded domain with sufficiently smooth boundary $\partial\Omega$; L denotes a second order linear elliptic operator in divergence form:

$$Lu := -\sum_{i,j=1}^{n} D_i(a_{ij}\,(x,t)\,D_j u) + \sum_{i=1}^{n} b_i\,(x,t)\,D_i u + c(x,t)u,$$

where a_{ij}, b_i and c are given coefficient functions $(i,j = 1,\ldots,n)$;

$g : Q_T \times \mathbb{R} \longrightarrow \mathbb{R}$ and $k : \Gamma_T \longrightarrow \mathbb{R}$ are given functions; h and φ are given generalized functions.

Semilinear parabolic equations like (1) have been studied extensively in the literature by numerous scientists who used various methods (see for example [4, 5,

This research is supported by 110T558-project of TUBITAK.

7, 9, 11]). A lot of mathematical models from physics, chemistry, and mechanics correspond to this type of equations. Mostly Dirichlet or Neumann type boundary value problems are studied, different cases of L and g have been investigated [2, 9, 10, 14]. Generally differential operator L has been taken as Laplace operator [3, 4, 14] and mapping g in polynomial forms [10].

In this paper we investigate nonhomogeneous third type boundary value problem for equation (1) in divergence form, with the mapping g in general form. For the existence of the generalized solution of problem (1)–(3) we obtain sufficient conditions for L, g and k. Under these conditions we prove that problem (1)–(3) is solvable in corresponding spaces by applying a general existence theorem from [11]. Also, in a model case of the mapping g, we prove the uniqueness of the solution of the considered problem.

2 Formulation and the Main Conditions

We investigate problem (1)–(3) when $h \in L_2(0,T;(W_2^1(\Omega))^*) + L_q(Q_T)$ ([8]), where $q > 1$ and $\varphi \in L_2(0,T;(W_2^{-\frac{1}{2}}(\partial\Omega))$.

We assume the following conditions:

(1) The coefficients $a_{ij} \in L_\infty(\overline{Q_T})$, $a_{ij} = a_{ji}$ is satisfied for all $i,j = 1,\dots,n$ and there exists a constant $\vartheta > 0$ such that

$$\sum_{i,j=1}^{n} a_{ij}(x,t)\xi_i\xi_j \geqslant \vartheta\,|\xi|^2$$

holds for all $\xi \in \mathbb{R}^n$ and $a.e.(x,t) \in \overline{Q_T}$.

(2) the function g is a Caratheodory function in $Q_T \times \mathbb{R}^1$ such that there exist a number $\alpha \geqslant 0$ and some functions g_0, g_1 which satisfy the inequality

$$|g(x,t,\tau)| \leqslant g_1(x,t)\,|\tau|^\alpha + g_0(x,t).$$

(The spaces to which g_0, g_1 belong will be defined later according to α.)

(3) The function $k \in L_\infty(0,T;L_{n-1}(\partial\Omega))$.

We define the space P_0 in the following way: $\mathbf{P_0} \equiv L_2(0,T;W_2^1(\Omega)) \cap L_{\alpha+1}(Q_T) \cap W_q^1(0,T;(W_2^1(\Omega))^*) \cap \{u : u(x,0) = 0\}$ $(q = q(\alpha) > 1)$.

A solution of the considered problem is understood as following:

Definition 1. A function $u \in P_0$ is called a generalized solution of problem (1)-(3) if it satisfies the equality

$$-\int_0^T\int_\Omega u\frac{\partial v}{\partial t}dxdt+\int_\Omega u(x,T)v(x,T)dx+\int_0^T\int_\Omega\sum_{i,j=1}^n a_{ij}\,(x,t)\,D_juD_ivdxdt+$$
$$+\int_0^T\int_\Omega\sum_{i=1}^n b_i\,(x,t)\,D_iu\;vdxdt+\int_0^T\int_\Omega c(x,t)uvdxdt+\int_0^T\int_\Omega g(x,t,u)vdxdt+$$
$$+\int_0^T\int_{\partial\Omega}k(x,t)uvdxdt=\int_0^T\int_\Omega hvdxdt+\int_0^T\int_{\partial\Omega}\varphi vdxdt$$

for all $v \in L_2(0,T;W_2^1(\Omega)) \cap L_{\alpha+1}(Q_T) \cap W_2^1(0,T;L_2(\Omega))$.

The investigation of problem (1)–(3) is related to the connection between the spaces $L_2(0,T;W_2^1(\Omega))$ and $L_{\alpha+1}(Q_T)$ according to values of α in condition (2). Depending on this connection we consider the problem in the sub-linear, linear and super-linear case.

3 Solvability of Problem (1)–(3) in the Super-linear Case

In this section we investigate problem (1)–(3) while $\alpha > 1$. For this case $\mathbf{P_0} \equiv L_2(0,T;W_2^1(\Omega))\cap L_{\alpha+1}(Q_T)\cap W^1_{\frac{\alpha+1}{\alpha}}(0,T;(W_2^1(\Omega))^*)\cap\{u : u(x,0)=0\}$. We assume that the following conditions are satisfied:

(2′) Let condition (2) be fulfilled with functions $g_1 > 0$, $g_0 \geqslant 0$ such that

$$g_1 \in L_\infty(Q_T), \quad g_0 \in L_{\frac{\alpha+1}{\alpha}}(Q_T).$$

(i) Let the functions b_i and c belong to $L_\infty(0,T;L_{\frac{\alpha+1}{\alpha-1}}(\Omega))$ for all $i = 1,\dots,n)$ and $L_{\frac{\alpha+1}{\alpha-1}}(Q_T)$ respectively.

(ii) There exist some numbers $\widetilde{g_1} > 0$ and $\widetilde{g_0} \in \mathbb{R}^1$ such that

$$g(x,t,\xi)\xi \geqslant \widetilde{g_1}\,|\xi|^{\alpha+1} - \widetilde{g_0}$$

holds for all $\xi \in \mathbb{R}^n$.

(iii) Let the function k satisfy one of the following relations:

(a) There exists a number $k_0 > 0$ such that $k(x,t) \geqslant -k_0 > -\frac{\vartheta_1}{c_3}$ holds for a.e. $(x,t) \in \Gamma_T$,

(b) $\|k\|_{L_\infty(0,T;L_{n-1}(\partial\Omega))} < \frac{\vartheta_1}{c_1}$,

where $\vartheta_1 < \min\{\vartheta, \widetilde{g_1}\}$. (Here c_1, c_3 are constants of Sobolev's embedding inequalities [1][1].)

Theorem 1. Let the conditions (1),(2′),(3) and (i)–(iii) be fulfilled for $\alpha > 1$. Then the problem (1)–(3) is solvable in P_0 for any $(h, \varphi) \in [L_2(0,T;(W_2^1(\Omega))^*) + L_{\frac{\alpha+1}{\alpha}}(Q_T)] \times L_2(0,T;(W_2^{-\frac{1}{2}}(\partial\Omega))$.

Proof. To apply the existence theorem from [11] to problem (1)–(3) firstly we define the corresponding mappings

$$f := \{f_1, f_2\} : P_0 \to L_2(0,T;(W_2^1(\Omega))^*) + L_{\frac{\alpha+1}{\alpha}}(Q_T),$$

where

$$f_1 := L + g, \quad f_2 := \frac{\partial}{\partial \nu} + k(x,t) \tag{4}$$

and

$$A \equiv Id : P_0 \to P_0. \tag{5}$$

We show that the conditions of the existence theorem from [11] are satisfied by proving the following lemmas.

Lemma 1. *The mapping $f : P_0 \to L_2(0,T;(W_2^1(\Omega))^*) + L_{\frac{\alpha+1}{\alpha}}(Q_T)$ is weakly continuous under the conditions of Theorem 1.*

Lemma 2. *The mappings f and A generate a coercive pair on $L_2(0,T;W_2^1(\Omega)) \cap L_{\alpha+1}(Q_T)$ under the conditions of Theorem 1.*

Thus, problem (1)–(3) is solvable in P_0 for any $(h, \varphi) \in [L_2(0,T;(W_2^1(\Omega))^*) + L_{\frac{\alpha+1}{\alpha}}(Q_T)] \times L_2(0,T;(W_2^{-\frac{1}{2}}(\partial\Omega))$. □

4 Solvability of Problem (1)–(3) in the Linear and Sub-linear Cases

In this section we investigate problem (1)–(3) while $0 \leqslant \alpha \leqslant 1$.For this case $\mathbf{P_0} \equiv L_2(0,T;W_2^1(\Omega)) \cap W_2^1(0,T;(W_2^1(\Omega))^*) \cap \{u : u(x,0) = 0\}$.

We assume that the following conditions are satisfied:

(2″) Let the condition (2) be fulfilled with the nonnegative functions below:

$$g_1 \in \begin{cases} L_{\frac{2}{1-\alpha}}(0,T;L_{\frac{2^*}{2^*-\alpha-1}}(\Omega)), & \text{if } 0 \leqslant \alpha < 1, \\ L_\infty(0,T;L_{\frac{n}{2}}(\Omega)), & \text{if } \alpha = 1, \end{cases} \qquad g_0 \in L_2(0,T;L_{(2^*)'}(\Omega)).$$

[1] $\|u\|^2_{L_{\frac{2(n-1)}{n-2}}(\partial\Omega)} \leqslant \mathbf{c_1}\|u\|^2_{W_2^1(\Omega)}$, $\|u\|^2_{L_2(\partial\Omega)} \leqslant \mathbf{c_3}\|u\|^2_{W_2^1(\Omega)}$

(Here $2^* := \frac{2n}{n-2}$ and $(2^*)' := \frac{2^*}{2^*-1}$.)

(i′) Let $b_i \in L_\infty(0,T;L_n(\Omega))$ for all $i = 1,\dots,n$. For these functions, we define

$$\sup_i \|b_i\|_{L_\infty(0,T;L_n(\Omega))} \equiv \sigma.$$

(ii′) Let the functions c and k belong to $L_\infty(0,T;L_{\frac{n}{2}}(\Omega))$ and $L_\infty(0,T;L_{n-1}(\partial\Omega))$ respectively, and satisfy one of the following relations when $0 \leqslant \alpha < 1$ (When $\alpha = 1$, i.e. that contains the linear case, it is allowed to take the function $c - g_1$ instead of the function c.):

(a) There exists a number $\widetilde{c}$ such that $c(x,t) \geqslant \widetilde{c} > 0$ holds for a.e. $(x,t) \in Q_T$ and a number σ satisfies the inequality

$$\sigma < \frac{\vartheta_2}{c_0}, \text{ where} \vartheta_2 := \min\left\{\vartheta, \widetilde{c}\right\}.$$

In this case the function k satisfies one of the following conditions:
(a_1) $k(x,t) \geqslant -k_0$ for a.e. $(x,t) \in \Gamma_T$, where $0 < k_0 < \frac{\vartheta_2 - \sigma c_0}{c_3}$,
(a_2) $\|k\|_{L_\infty(0,T;L_{n-1}(\partial\Omega))} < \frac{\vartheta_2 - \sigma c_0}{c_1}$.

(b) There exists a number k_0 such that $k(x,t) \geqslant k_0 > 0$ holds for a.e. $(x,t) \in \Gamma_T$ and a number σ satisfies the inequality

$$\sigma < \frac{\vartheta_2 c_2}{c_0} \text{ where} \vartheta_2 := \min\{\vartheta, k_0\}.$$

In this case the function c satisfies one of the following conditions:
(b_1) $c(x,t) \geqslant -\widetilde{c}$ for a.e. $(x,t) \in Q_T$, where $0 < \widetilde{c} < \vartheta_2 c_2 - \sigma c_0$,
(b_2) $\|c\|_{L_\infty(0,T;L_{\frac{n}{2}}(\Omega))} < \frac{\vartheta_2 c_2 - \sigma c_0}{c_0^2}$.

(Here, c_0, c_1, c_3 are constants of Sobolev's embedding inequalities [1] [1] and c_2 comes from the inequality[2] [8, 13].)

Theorem 2. Let the conditions (1), (2″), (3), (i′) and (ii′) be fulfilled for $0 \leqslant \alpha \leqslant 1$. Then the problem (1)–(3) is solvable in P_0 for any $(h,\varphi) \in L_2(0,T;(W_2^1(\Omega))^*) \times L_2(0,T;(W_2^{-\frac{1}{2}}(\partial\Omega))$.

Proof. For the proof of Theorem 2. we again apply the existence theorem from [11] to problem (1)–(3). Here we take the corresponding mappings as in (4)–(5).

[1] $\|u\|_{L_{\frac{2n}{n-2}}(\Omega)} \leqslant \mathbf{c_0}\|u\|_{W_2^1(\Omega)}$, $\|u\|^2_{L_{\frac{2(n-1)}{n-2}}(\partial\Omega)} \leqslant \mathbf{c_1}\|u\|^2_{W_2^1(\Omega)}$, $\|u\|^2_{L_2(\partial\Omega)} \leqslant \mathbf{c_3}\|u\|^2_{W_2^1(\Omega)}$

[2] $\mathbf{c_2}\|u\|^2_{W_2^1(\Omega)} leq \|Du\|^2_{L_2(\Omega)} + \|u\|^2_{L_2(\partial\Omega)}$

We obtain again that the conditions of the existence theorem in [11] are satisfied from the following lemmas:

Lemma 3. *The mapping $f : P_0 \to L_2(0,T;(W_2^1(\Omega))^*)$ is weakly continuous under the conditions of Theorem 2.*

Lemma 4. *The mappings f and A generate a coercive pair on $L_2(0,T;W_2^1(\Omega))$ under the conditions of Theorem 2.*

Thus, problem (1)–(3) is solvable in P_0 for any $(h,\varphi) \in L_2(0,T;(W_2^1(\Omega))^*) \times L_2(0,T;(W_2^{-\frac{1}{2}}(\partial\Omega))$. □

5 Uniqueness of the Solution of Problem (1)–(3) in a Model Case

We investigate the uniqueness of solutions of problem (1)–(3) in a model case of the mapping g in $\mathbf{P_0} \equiv L_2(0,T;W_2^1(\Omega)) \cap L_{\rho+1}(Q_T) \cap W^1_{\frac{\rho+1}{\rho}}(0,T;(W_2^1(\Omega))^*) \cap \{u : u(x,0) = 0\}$. We assume that the following conditions are satisfied:

(i_0) Let $b_i \in L_\infty(0,T;L_n(\Omega))$ for all $i = 1,\dots,n$. For these functions we define

$$\sup_i \|b_i\|_{L_\infty(0,T;L_n(\Omega))} \equiv \sigma.$$

(ii_0) Let the functions c and k belong to $L_\infty(0,T;L_{\frac{n}{2}}(\Omega))$ and $L_\infty(0,T;L_{n-1}(\partial\Omega))$ respectively, and satisfy one of the following conditions:

(a) There exists a number $\widetilde{c}$ such that $c(x,t) \geqslant \widetilde{c} > 0$ holds for a.e. $(x,t) \in Q_T$ and a number σ satisfies the inequality

$$\sigma < \frac{\vartheta_2}{c_0}, \text{ where} \vartheta_2 := \min\left\{\vartheta, \widetilde{c}\right\}.$$

In this case the function k satisfies one of the following conditions:
(a_1) $k(x,t) \geqslant -k_0$ for a.e. $(x,t) \in \Gamma_T$, where $0 < k_0 < \frac{\vartheta_2 - \sigma c_0}{c_3}$,
(a_2) $\|k\|_{L_\infty(0,T;L_{n-1}(\partial\Omega))} < \frac{\vartheta_2 - \sigma c_0}{c_1}$.

(b) There exists a number k_0 such that $k(x,t) \geqslant k_0 > 0$ holds for a.e. $(x,t) \in \Gamma_T$ and a number σ satisfies the inequality

$$\sigma < \frac{\vartheta_2 c_2}{c_0}, \quad \text{where} \vartheta_2 := \min\{\vartheta, k_0\}.$$

In this case the function c satisfies one of the following conditions:
(b_1) $c(x,t) \geqslant -\widetilde{c}$ for a.e. $(x,t) \in Q_T$, where $0 < \widetilde{c} < \vartheta_2 c_2 - \sigma c_0$,
(b_2) $\|c\|_{L_\infty(0,T;L_{\frac{n}{2}}(\Omega))} < \frac{\vartheta_2 c_2 - \sigma c_0}{c_0^2}$.

Theorem 3. Let the mapping g in problem (1)–(3) be defined as

$$g(x,t,\tau) = d(x,t)\,|\tau|^{\rho-1}\,\tau,$$

where $\rho > 0$, $d(x,t) \geqslant 0$ for a.e. $(x,t) \in Q_T$ and

$$d \in \begin{cases} L_{\frac{2}{1-\rho}}(0,T; L_{\frac{2^*}{2^*-\rho-1}}(\Omega)), & \text{if } 0 \leqslant \rho < 1, \\ L_\infty(0,T; L_{\frac{n}{2}}(\Omega)), & \text{if } \rho = 1, \\ L_\infty(Q_T), & \text{if } \rho > 1. \end{cases}$$

Let additionally the conditions (1), (3), (i_0) and (ii_0) be fulfilled. Then the solution of problem (1)–(3) in P_0 is unique if it exists.

For the proof of Theorem 3 we use a well-known method. We assume that problem (1)–(3) has two different solutions and after making necessary operations we obtain a contradiction.

References

1. R. A. Adams; *Sobolev Spaces*, Academic Press, New York, 1975
2. C. Bandle, H. A. Levine, Qi S. Zhang; *Critical exponents of Fujita type for Inhomogeneous Parabolic Equations and systems*, Journal of Mathematical Analysis and Applications, 251 (2000) 624-648
3. T. Cheng, G-F. Zheng; *Some Blow-up Problems for a Semilinear Parabolic Equation with a Potential*, J. Differential Equations, 244 (2008) 766-802
4. H. Fujita; *On the blow up of solutions of the Cauchy problem for* $u_t = \Delta u + u^{\sigma+1}$, J. Fac. Sci. Univ., Tokyo, Sect. I 13 (1966) 109-124
5. V. A. Galaktionov, J. L. Vazquez; *Continuation of blow up solutions of nonlinear heat equations in several space dimensions*, Comm. Pure Appl. Math., 50 1 (1997) 1-67
6. K. Kalli, K. N. Soltanov; *On Some Semilinear Elliptic Equations*, AIP Conference Proceedings, 1168 (2009) 298-301
7. J. L. Lions; *Quelques methodes de resolution des problemes aux Limities non lineaires*, Dunod, Gauthier-Villars, Paris 1969
8. J. L. Lions, E. Magenes; *Non-Homogeneous Boundary Value Problems and Applications*, Springer-Verlag, Berlin, Heidelberg, New York, Volume 1 1972
9. L. Ma; *Boundary Value Problem for a Classical Semilinear Parabolic Equation*, arXiv:1012.5861v1 [math.AP] (2010)
10. N. Mizoguchi; *Grow up Solutions for a Semilinear Heat Equation with Supercritical Nonlinearity*, J. Differential Equations, 227 (2006) 652-669

11. K. N. Soltanov; *On some modification on Navier-Stokes equations*, Nonlinear Analysis- Theory Methods and Applications 52 3 (2003) 769-793
12. K. N. Soltanov; *Some Boundary Problem for Emden-Fowler Type Equations, Function Spaces*, Differential Operators and Nonlinear Analysis, (FSDONA, 2004) May 27-June 2, 2004, Svratka, Czech Republic, 2005 311-318
13. M. Struwe; *Variational Methods Applications to Nonlinear Partial Differential Equations and Hamiltonian Systems*, Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong, Barcelona, 1990
14. G-F. Zheng; *a Quasi-monotonicity Formula and Partial Regularity for Borderline Solutions to a Parabolic Equation*, Ann. I. H. Poincare- AN 27 (2010) 1333-1360

Kerime Kalli
Department of Mathematics, Hacettepe University, 06800 Beytepe, Ankara, Turkey,
kerime@hacettepe.edu.tr

Kamal N. Soltanov
Department of Mathematics, Hacettepe University, 06800 Beytepe, Ankara, Turkey,
soltanov@hacettepe.edu.tr

OPTIMIZATION OF BOUNDARY CONTROL PRODUCED BY THE THIRD BOUNDARY CONDITION

A. A. Nikitin

Key words: PDE, Optimization

AMS Mathematics Subject Classification:

Abstract. In this paper we study the boundary control, produced the third boundary condition at the left end of a string with fixed right. Optimality criterion is established based on the minimization of the integral of a linear combination of the control and its primitive, built in an arbitrary integer power $p > 1$. In this paper, optimal control is presented in the form of a well-established solutions Volterra integral equation of convolution type.

1 Introduction

Consider the vibrations of a string governed by the wave equation

$$u_{xx}(x,t) - u_{tt}(x,t) = 0. \tag{1}$$

over the time interval $0 < t \leqslant T$. It is assumed that the end $x = l$ is fixed, while the other end $x = 0$ is controlled by applying the third boundary condition $u_x(0,t) - hu(0,t) = \mu(t)$. In an arbitrary time T that is a multiple of $4l$, the vibration process transfers the string from the arbitrarily given initial state

$$\{u(x,0) = \varphi(x);\ u_t(x,0) = \psi(x)\} \tag{2}$$

to the arbitrarily given terminal state

$$\left\{u(x,T) = \widehat{\varphi}(x);\ u_t(x,T) = \widehat{\psi}(x)\right\}. \tag{3}$$

The consideration is performed in terms of a weak solution to wave equation (1) from the class $\widehat{W}_p^1(Q_T)$, where Q_T is the rectangle $[0 \leqslant x \leqslant l] \times [0 \leqslant t \leqslant T]$. This class was first introduced by V.Il'in in [1]. It is defined as the set of continuous functions $u(x,t)$ in Q_T with generalized partial derivatives $u_x(x,t), u_t(x,t)$, that belong not only to the class $L_p(Q_T)$, but also to $L_p[0,l]$ for all $t \in [0,T]$ and to $L_p[0,T]$ for all $x \in [0,l]$.

This work was supported by the program MK-6025.2010.1.

The definition of $\widehat{W}_p^1(Q_T)$ implies the membership conditions

$$\begin{gathered} u(x,0)=\varphi(x)\in W_p^1[0,l], \quad u_t(x,0)=\psi(x)\in L_p[0,l]; \\ u(x,T)=\widehat{\varphi}(x)\in W_p^1[0,l], \quad u_t(x,T)=\widehat{\psi}(x)\in L_p[0,l]; \\ \mu(t)\in L_p[0,T]. \end{gathered} \tag{4}$$

The fixed end conditions at $x=l$ are

$$\varphi(l)=0, \quad \widehat{\varphi}(l)=0. \tag{5}$$

2 Mixed problem

Before stating further results, we consider the following mixed problem for the wave equation with initial and boundary conditions:

$$u_{xx}(x,t)-u_{tt}(x,t)=0, \tag{6}$$

$$u(x,0)=\varphi(x), \quad u_t(x,0)=\psi(x), \tag{7}$$

$$u_x(0,t)-h\cdot u(0,t)=\mu(t), \quad u(l,t)=0, \tag{8}$$

where $\varphi(x), \psi(x)$ and $\mu(t)$ belong to classes (4) and obey conditions (5).

Definition 1. The weak solution to the mixed problem in the class $\widehat{W}_p^1(Q_T)$ is a function $u(x,t)\in\widehat{W}_p^1(Q_T)$, satisfying the integral identity

$$\int_0^l\int_0^T u(x,t)[\Phi_{tt}(x,t)-\Phi_{xx}(x,t)]dxdt+\int_0^T\mu(t)\Phi(0,t)dt+ \\ +\int_0^l[\varphi(x)\Phi_t(x,0)-\psi(x)\Phi(x,0)]\,dx=0, \tag{9}$$

where $\Phi(x,t)$ is an arbitrary function from $C^2(Q_T)$, satisfying the conditions: $\Phi_x(0,t)-h\Phi(0,t)\equiv 0$, $\Phi(l,t)\equiv 0$ for $0\leqslant t\leqslant T$ and $\Phi(x,T)\equiv 0, \Phi_t(x,T)\equiv 0$ for $0\leqslant x\leqslant l$.

The assertion below follows from [2] **Proposition 1.** For every $T>0$, the mixed problem has a unique weak solution from the class $\widehat{W}_p^1(Q_T)$.

Definition 2. The solution to the corresponding boundary control problem is a function $\mu(t)\in L_p[0,T]$ such that the weak solution $u(x,t)\in\widehat{W}_p^1(Q_T)$ to mixed problem (6) - (8) satisfies terminal conditions (3).

Consider this problem when $T = 4l(n+1)$, $n = 0, 1, 2, \ldots$. For $T > 2l$, the boundary control problem has infinitely many solutions. Therefore, the problem arises of finding an *optimal* solution among them. The optimality of this boundary control is achieved by minimizing the pth power of the integral of a linear combination of the control itself and its primitive, where $p > 1$ is an arbitrary integer. The result is a generalization of the control problem for a force at the string's left end with its right end being fixed (i.e., at $h = 0, p = 2$), which was investigated in detail in [3].

3 Optimization

To formulate the optimization problem, we define the function

$$\mathbf{H}_m(\tau) = \left\{ e^{-h\tau} \cdot \left[\mathbf{L}^1_{2n-m+1}(2h\tau) + \mathbf{L}^1_{2n-m}(2h\tau) \right], \quad m = \overline{0, 2n+1} \right\}, \tag{10}$$

where $\mathbf{L}^1_k(2h\tau)$ is a Laguerre polynomial, see [4].

Now define function $\mathbf{H}(t, \tau)$ as

$$\mathbf{H}(t, \tau) = \left\{ \mathbf{H}_m(\tau), \text{ for } 2lm < t \leqslant 2l(m+1), \ m = \overline{0, 2n+1} \right\}. \tag{11}$$

The task is, among all the boundary controls $\mu(t) \in L_p[0, T]$, to find a function that minimizes the integral

$$\int_0^T \left| \mu(t) - h \cdot \int_0^t \mathbf{H}(t, t-\xi)\mu(\xi)\, d\xi \right|^p dt \tag{12}$$

subject to the coupling conditions derived from the arbitrarily given initial and terminal conditions.

The functions $\varphi(x)$ and $\psi(x)$ are extended from initial conditions (2) and functions $\widehat{\varphi}(x)$ and $\widehat{\psi}(x)$ are extended from terminal conditions (3) in an odd manner with respect to the point $x = l$ from the interval $[0, l]$ to $[l, 2l]$. The fixed point conditions (5) guarantee that the functions thus extended to $[0, 2l]$ belong to the classes

$$\varphi(x),\ \widehat{\varphi}(x) \in W^1_p[0, 2l], \quad \psi(x),\ \widehat{\psi}(x) \in L_p[0, 2l]; \tag{4*}$$

The assertion below follows [5]

Proposition 2. The weak solution $\widehat{u}(x, t)$ to mixed problem (6)–(8) with zero initial conditions (7) for $T \leqslant 4l(n+1)$ is given by the identity

$$\widehat{u}(x,t) = -\sum_{k=0}^{2n+1}(-1)^k \int_0^{t-x-2kl} \underline{\mu}(\xi)d\xi - \sum_{k=1}^{2n+2}(-1)^k \int_0^{t+x-2kl} \underline{\mu}(\xi)d\xi + h\sum_{k=0}^{2n+1}(-1)^k \times$$
$$\times \int_0^t e^{-h\tau}\mathbf{L}_k^1(2h\tau)\Big[\int_0^{t-2kl-x-\tau} \underline{\mu}(\xi)d\xi - \int_0^{t-2l(k+1)+x-\tau} \underline{\mu}(\xi)d\xi \int_0^{t-2l(k+1)-x-\tau} -$$
$$- \underline{\mu}(\xi)d\xi \int_0^{t-2l(k+2)+x-\tau} + \underline{\mu}(\xi)d\xi\Big]d\tau, \quad (13)$$

where $\underline{\mu}(t)$ denotes the function that equals $\mu(t)$ for $t \geqslant 0$ and vanishes for $t < 0$. For brevity, we introduce the notations

$$\mu_m(x) = \mu(x + 2lm), \quad m = 0, 1, 2, \ldots; \quad \boldsymbol{l}\,_k^1(2h\tau) = e^{-h\tau} \cdot \mathbf{L}_k^1(2h\tau).$$

Applying the technique proposed in [6] and identity (13), we find the coupling condition to be considered together with integral (12):

$$-\sum_{m=0}^{2n+1}(-1)^m \cdot \mu_m(x) +$$
$$+ h\sum_{m=0}^{2n+1}(-1)^m \int_0^{2lm+x} \left[\boldsymbol{l}\,_{2n-m+1}^1(2h\tau) + \boldsymbol{l}\,_{2n-m}^1(2h\tau)\right] \cdot \mu_m(x-\tau)d\tau = \mathbf{D}(x),$$

where

$$\mathbf{D}(x) = \frac{1}{2}\left[\widehat{\varphi}'(x) + \widehat{\psi}(x)\right] - \widetilde{\mathbf{A}}(x) + h \cdot \int_0^x \left[\boldsymbol{l}\,_{2n+1}^1(2h\tau) + \boldsymbol{l}\,_{2n}^1(2h\tau)\right] \cdot \widetilde{\mathbf{A}}(x-\tau)d\tau,$$
$$\widetilde{\mathbf{A}}(x) = \frac{1}{2} \cdot \left\{\varphi'(x) + \psi(x) - h \cdot \left[\varphi(x) + \varphi(0) + \int_0^x \psi(\xi)\,d\xi\right]\right\} \quad (14)$$

Note that this condition is necessary and sufficient for function μ to be a solution of the considered boundary control problem.

By using relation (10), this condition can be rewritten as

$$\sum_{m=0}^{2n+1}(-1)^m\cdot\left\{\mu_m(x)-h\cdot\int\limits_0^{2lm+x}\mathbf{H}_m(\tau)\cdot\mu_m(x-\tau)d\tau\right\}=-\mathbf{D}(x). \tag{15}$$

The substitution $\{\tau=t-\xi\}$ reduces integral (12) to the form

$$\int\limits_0^T\left|\mu(t)-h\cdot\int\limits_0^t\mathbf{H}(t,\tau)\mu(t-\tau)\,d\tau\right|^p dt. \tag{16}$$

For the time interval $T=4l(n+1)$, this integral can be written as

$$\int\limits_0^{2l}\sum_{m=0}^{2n+1}\left|(-1)^m\cdot\left\{\mu_m(x)-h\cdot\int\limits_0^{2lm+x}\mathbf{H}_m(\tau)\mu_m(x-\tau)\,d\tau\right\}\right|^p dx, \tag{17}$$

where the variable t was replaced by x.

Thus, the optimization problem is reduced to finding the minimum of integral (12) with the coupling conditions given by (14) and (15). The lemma proved in [6] reduces the minimization of the integral in sum (17) to finding the pointwise minimum of the sum

$$\sum_{m=0}^{2n+1}\left|(-1)^m\cdot\left\{\mu_m(x)-h\int\limits_0^{2lm+x}\mathbf{H}_m(\tau)\mu_m(x-\tau)\,d\tau\right\}\right|^p. \tag{18}$$

This minimum can be found by the Lagrange method. Without going into technical details, the result can be described as follows. For all $p>1$, the desired solution $\mu(x)$ to problem (14), (15), (18) on each interval is found by solving the integral equation

$$\mu_m(x)-h\int\limits_0^{2lm+x}\mathbf{H}_m(\tau)\mu_m(x-\tau)\,d\tau=\frac{(-1)^{m+1}\,\mathbf{D}(x)}{2n+2},\ 0\leqslant x\leqslant 2l;\ m=\overline{0,2n+1}.$$

For all $m=0,1,\ldots,2n+1$, the number $2lm+x$ is redenoted by y and the substitution $y-\tau=\xi$ is made in the integrand. Moreover, the result is reduced to the solution to the *convolution-type Volterra* integral equation the of the second

kind

$$\mu(y) - h\int\limits_0^y \mathbf{H}_m(y-\xi)\mu(\xi)d\xi = \frac{(-1)^{m+1}\mathbf{D}(y-2lm)}{2n+2},$$

$$\text{for } 2lm \leqslant y \leqslant 2l(m+1);\ m = \overline{0, 2n+1}, \quad (19)$$

where the kernel $\mathbf{H}_m(y-\xi)$ is given by (10).

Such equations are studied using the Laplace transform $\widetilde{\mathbf{F}}(z) = \int\limits_0^\infty e^{-zt}\mathbf{F}(t)\,dt$, since, under certain constraints associated with its applicability, this operator brings the convolution to an ordinary product. Applying direct and inverse Laplace transform to the equation (19) we can write the solution of this equation, the function $\mu(y)$ in the form of

$$\mu(y) = \frac{(-1)^{m+1}\,\mathbf{D}(y-2lm)}{2n+2} + h\int\limits_0^y \mathbf{R}_m(y-t)\,\frac{(-1)^{m+1}\,\mathbf{D}(t-2lm)}{2n+2}\,dt, \quad (20)$$

where

$$\mathbf{R}_m(y-t) = \sum_{i=0}^{2n-m+2} h^{i-1}(y-t)^{i-1}\begin{pmatrix} 2n-m+2 \\ i \end{pmatrix}_1 \tilde{\mathbf{F}}_1\big(2n-m+1;\, i;\, h(y-t)\big),$$

${}_1\tilde{\mathbf{F}}_1(a;c;z)$ — *Kummer confluent hypergeometric function*, see. [7].

Remark: Since $\mu(y)$ is defined on the finite interval $[2lm, 2l(m+1)]$, it can be extended by zero outside this interval. Thus, these conditions can be assumed to hold. Expressions (14) and (19) imply the following estimate for integral (12):

$$\int\limits_0^T \left|\mu(t) - h\cdot\int\limits_0^t \mathbf{H}(t, t-\xi)\mu(\xi)\,d\xi\right|^p dt = \underline{O}\left(\frac{1}{T^{p-1}}\right), \quad (21)$$

where the constant limiting the growth of the $\underline{O}$ - terms depends only on the norms of $\varphi(x)$, $\psi(x)$, $\widehat{\varphi}(x)$ and $\widehat{\psi}(x)$, in classes (4). Estimate (21) shows that integral (12) tends to zero as $T \to \infty$ for any $p > 1$. Therefore, choosing a sufficiently long time interval T, we can avoid the resonance of the process.

To conclude, we analyze the uniqueness of the optimal solution to the boundary control problem. Since the integral is linear, using the Minkowski inequality and an argument similar to that used in [6], we can prove that integral (12) for $p > 1$

has its minimum at a unique function $\mu(t)$. And for $p = 1$ we get infinitely many solutions.

References

1. V. A. Il′in, *Differ. Equations. 2000. 36, N11. p. 1513–1528.*
2. V. A. Il′in, *Usp. Mat. Nauk 15 (2), 97–154 (1960).*
3. V. A. Il′in, *Dokl. Math.* 71, 10–14 (2005) [Dokl. Akad.Nauk 400, 731–735 (2005)].
4. *Higher Transcendental Functions* (Bateman Manuscript Project), Ed. by A. Erdelyi (McGraw-Hill, New York, 1953; Nauka, Moscow, 1966), Vol. 2.
5. A. A. Nikitin, *Differ. Equations. 2007. 43, №12, p. 1692–1700.*
6. V. A. Il′in and E. I. Moiseev, *Differ. Equations 42, 1633–1644 (2006)* [Differ. Uravn. 42, 1558–1570 (2006)].
7. *Higher Transcendental Functions* (Bateman Manuscript Project), Ed. by A. Erdelyi (McGraw-Hill, New York, 1953; Nauka, Moscow, 1965), Vol. 1.

A. A. Nikitin
Contacts for the first author: Russia, Moscow State University, 117296, Moscow,
8-(905)-762-0653, nikitin@cs.msu.su

ROBIN AND INITIAL VALUE PROBLEM FOR A YAMABE TYPE PARABOLIC EQUATION

Eylem Öztürk, Kamal N. Soltanov

Key words: Yamabe type Parabolic Equation, Third type Boundary Value Problem, Existence and Uniqueness Theorems, Sublinear and Super Linear Cases.

AMS Mathematics Subject Classification: 35D30, 35K58, 35M12

Abstract. In this article we investigate mixed problem with Robin boundary condition for a Yamabe type Parabolic Equation. We show that there exists a generalized solution of the considered problem under more general conditions. We also show the uniqueness of the solution of the considered problem in a special case.

1 Introduction

We consider the problem;

$$\frac{\partial u}{\partial t} - \Delta u + a(x,t)\,|u|^{\rho}\,u - b(x,t)\,|u|^{\nu}\,u = h(x,t),\ (x,t) \in Q_T \tag{1}$$

$$\left(\frac{\partial u}{\partial \eta} + k(x',t)u\right)\Big|_{\Sigma_T} = \varphi(x',t), \quad (x',t) \in \Sigma_T \tag{2}$$

$$u(x,0) = 0, \quad x \in \Omega \tag{3}$$

Where $\Omega \subset \mathbb{R}^n$, $n \geqslant 3$, is a bounded domain with sufficiently smooth boundary $\partial\Omega$; $\rho, \nu > -1$ are given some numbers; $Q_T = \Omega \times (0,T)$, $\Sigma_T = \partial\Omega \times [0,T]$; Δ is the n dimensional Laplace operator; $a : Q_T \to \mathbb{R}^1$, $b : Q_T \to \mathbb{R}^1$ and $k : \Sigma_T \to \mathbb{R}^1$ are given functions; In general h and φ are given generalized functions.

The elliptic part of equation (1) is so-called "Yamabe type equation" when it's homogenous and $a(x,t) = 1$, $b(x,t) = 1$, $\rho = 0$, $\nu = \frac{4}{n-2}$ (see [2, 10]). Hence, equation (1) is the so-called "Yamabe type parabolic equation". This semilinear parabolic equation has been studied extensively in homogeneous form (see [3, 5-8, 12, 15]) for $a(x,t) = 0$, $b(x,t) = 1$ and various restriction on ν. In [4], the existence of global solutions of homogeneous form of equation (1) when $a(x,t) = 0$, $b(x,t) = 1$, $\nu > \frac{2}{n-2}$ with nonhomogeneous Dirichlet boundary condition (also for Neumann boundary condition) was proved in a domain Ω which has special conditions on

This research is supported by 110T558-project of TUBITAK..

boundary $\partial\Omega$. In [11], global nonnegative weak solutions of Cauchy problem for homogeneous form of equation (1) when $a(x,t)=0$, $b(x,t)=1$ was researched. In [9], global positive solutions of homogenous form of (1) when $a(x,t)=0$, $b(x,t)=1$, $\nu=p-1$ for special cases of p with homogenous Dirichlet condition was investigated.

In this paper, we investigate mixed problem with Robin boundary condition for a Yamabe type parabolic equation. We investigate the problem (1)–(3) in sublinear, linear and super linear cases, by depending on nonlinear part. For the existence of generalized solution of problem (1)–(3) and for the uniqueness of generalized solution of the problem in a special case, we obtained sufficient conditions for functions a, b and k and relations between ρ and ν. And under these conditions we showed the existence of generalized solution of problem (1)–(3) and the uniqueness of the solution in corresponding spaces, by applying a general existence theorem from [13].

2 Formulation of Problem (1)–(3) and Main Conditions

We shall assume $h \in L_2(0,T;(W_2^1(\Omega))^*) + L_{\frac{\rho+2}{\rho+1}}(Q_T)$ and $\varphi \in L_2(0,T;W_2^{-\frac{1}{2}}(\partial\Omega))$. We assume that the following conditions are satisfied;

(i) *a and b are nonnegative functions such that $a \in L_{p_1}(0,T;L_{p_2}(\Omega))$, $b \in L_{r_1}(0,T;L_{r_2}(\Omega))$ for some numbers p_1, r_1, p_2, $r_2 > 1$ which will be defined later.*

(ii) $k \in L_\infty(0,T;L_{n-1}(\partial\Omega))$

We introduce following space; $P_0 := L_2(0,T;W_2^1(\Omega)) \cap L_{\rho+2}(Q_T) \cap W_\alpha^1(0,T;(W_2^1(\Omega))^*) \cap \{u : u(x,0)=0\}$, $(\alpha=\alpha(\rho)>1)$

We will understand the solution of the considered problem as the following sense:

Definition 1. A function $u \in P_0$ is called the generalized solution of problem (1)–(3) if it satisfies the equality;

$$
\begin{aligned}
&-\int_0^T\int_\Omega u\frac{\partial v}{\partial t}dxdt+\int_\Omega u(x,T)v(x,T)dxdt+\\
&\quad+\int_0^T\int_\Omega Du.Dvdxdt+\int_0^T\int_\Omega\left(a(x,t)\left|u\right|^\rho u-b(x,t)\left|u\right|^\nu u\right)vdxdt+\\
&\quad+\int_0^T\int_{\partial\Omega}k(x',t)uvdx'dt=\int_0^T\int_\Omega hvdxdt+\int_0^T\int_{\partial\Omega}\varphi vdx'dt
\end{aligned}
$$

for all $v \in W_2^1(0,T;L_2(\Omega)) \cap L_2(0,T;W_2^1(\Omega)) \cap L_{\rho+2}(Q_T)$.

We'll consider the problem in three different sections: sublinear, linear, super linear cases. This division will be done according to nonlinear part of the problem.

3 Existence of the Solution of Problem (1)–(3) in Subliner Case

Let $-1 < \rho < 0$, $-1 < \nu \leqslant 0$ or $-1 < \rho \leqslant 0$, $-1 < \nu < 0$. Then, $P_0 \equiv L_2(0,T;W_2^1(\Omega)) \cap W_2^1(0,T;(W_2^1(\Omega))^*) \cap \{u : u(x,0) = 0\}$ since $L_2(0,T;W_2^1(\Omega)) \subset L_{\rho+2}(Q_T)$.

In this case, condition ***(i)*** is denoted as (***i'***) with following parameters p_1, r_1, p_2, r_2

$$p_1 := \begin{cases} \dfrac{2}{|\rho|}, \text{ if } -1 < \rho < 0 \\ \infty, \text{ if } \rho = 0 \end{cases} \quad p_2 := \begin{cases} \dfrac{2n}{\rho(2-n)+4}, \text{ if } -1 < \rho < 0 \\ \dfrac{n}{2}, \text{ if } \rho = 0 \end{cases}$$

$$r_1 := \begin{cases} \dfrac{2}{|\nu|}, \text{ if } -1 < \nu < 0 \\ \infty, \text{ if } \nu = 0 \end{cases} \quad r_2 := \begin{cases} \dfrac{2n}{\nu(2-n)+4}, \text{ if } -1 < \nu < 0 \\ \dfrac{n}{2}, \text{ if } \nu = 0 \end{cases}$$

Theorem 1. *Let $-1 < \rho < 0$, $-1 < \nu \leqslant 0$ or $-1 < \rho \leqslant 0$, $-1 < \nu < 0$ and conditions (**i'**), **(ii)** be fulfilled. Additionally the following conditions are satisfied;*

(iii) *There exists a number $k_0 > 0$ such that $k(x',t) \geqslant k_0$ for almost every $(x',t) \in \Sigma_T$*

(iv) *If $\nu = 0$ then $\|b\|_{L_{r_1}(0,T;L_{r_2}(\Omega))} < \frac{1}{\widetilde{c}\,c_1^2} \min\{1, k_0\}$*[1]

Then problem (1)–(3) is solvable in P_0 for any $(h,\varphi) \in [L_2(0,T;(W_2^1(\Omega))^*)] \times L_2(0,T;W_2^{-\frac{1}{2}}(\partial\Omega))$.

Proof. We want to apply the existence Theorem from [13] to problem (1)–(3), firstly we define corresponding mappings and spaces for the problem: $f = \{f_1, f_2\}$ such that

$$f_1(u) := -\Delta u + a(x,t)|u|^\rho u - b(x,t)|u|^\nu u$$

[1]Here c_1 is constant of Sobolev's Imbedding inequality [1]: $\|u\|_{L_{\frac{2n}{n-2}}(\Omega)} \leqslant c_1 \|u\|_{W_2^1(\Omega)}$ $\widetilde{c}$ comes from this inequality [14]: $\|u\|^2_{W_2^1(\Omega)} \leqslant \widetilde{c}(\|Du\|^2_{L_2(\Omega)} + \|u\|^2_{L_2(\partial\Omega)})$

$$f_2(u) := \frac{\partial u}{\partial \eta} + k(x', t)u \tag{4}$$

$$A := Id$$

Here, $f : P_0 \to L_2(0, T; (W_2^1(\Omega))^*) \times L_2(0, T; W_2^{-\frac{1}{2}}(\partial\Omega))$; $A : P_0 \to P_0$.

Now, we shall give the following lemmas to see that mappings f and A satisfy the conditions of existence theorem from [13]:

Lemma 1. *f is bounded from P_0 to $L_2(0, T; (W_2^1(\Omega))^*)$, under the conditions of Theorem 1.*

Lemma 2. *f is weakly continuous from P_0 to $L_2(0, T; (W_2^1(\Omega))^*)$, under the conditions of Theorem 1.*

Lemma 3. *f and A generate a coercive pair on $L_2(0, T; W_2^1(\Omega))$, under the conditions of Theorem 1.*

From these Lemmas, we obtain that all conditions of existence theorem from [13] are satisfied for the mappings f and A. So we apply existence theorem to problem (1)–(3) then we see that problem (1)–(3) is solvable in P_0 for any $(h, \varphi) \in L_2(0, T; (W_2^1(\Omega))^*) \times L_2(0, T; W_2^{-\frac{1}{2}}(\partial\Omega))$. □

4 Existence and Uniqueness of the Solution of problem (1)–(3) in Linear Case

Let $\rho = \nu = 0$. Then, $P_0 \equiv L_2(0, T; W_2^1(\Omega)) \cap W_2^1(0, T; (W_2^1(\Omega))^*) \cap \{u : u(x, 0) = 0\}$. In this case, condition ***(i)*** is denoted as (***i′′***) with following parameters p_1, r_1, p_2, r_2 $p_1 = r_1 = \infty$ and $p_2 = r_2 = \frac{n}{2}$

Theorem 2. *Let $\rho = \nu = 0$ and conditions (**i′′**), **(ii)** be fulfilled. Additionally one of the following conditions is satisfied:*

(iii) *There exists a number $k_0 > 0$ such that $k(x', t) \geqslant k_0$ for almost every $(x', t) \in \Sigma_T$. In this case, one of the following conditions is satisfied*[1]

(a$_1$) *There exists a number $b_0 > 0$ such that $(a(x, t) - b(x, t)) \geqslant -b_0$ for almost every $(x, t) \in Q_T$, and $b_0 < \dfrac{\min\{k_0, 1\}}{\widetilde{c}c_2^2}$*

[1]Here $\widetilde{c}$ is the constant coming from the inequality, [14]: $\|u\|^2_{W_2^1(\Omega)} \leqslant \widetilde{c}(\|Du\|^2_{L_2(\Omega)} + \|u\|^2_{L_2(\partial\Omega)})$ c_1, c_2, c_3, c_4 are constants of Sobolev's Imbedding inequalities, [1]: $\|u\|_{L_{\frac{2n}{n-2}}(\Omega)} \leqslant c_1 \|u\|_{W_2^1(\Omega)}, \|u\|_{L_2(\Omega)} \leqslant c_2\|u\|_{W_2^1(\Omega)}, \|u\|_{L_2(\partial\Omega)} \leqslant c_3\|u\|_{W_2^1(\Omega)}, \|u\|_{L_{\frac{2(n-1)}{n-2}}(\partial\Omega)} \leqslant c_4\|u\|_{W_2^1(\Omega)}$

(b_1) $\|a-b\|_{L_\infty(0,T;L_{\frac{n}{2}}(\Omega))} < \dfrac{\min\{k_0,1\}}{\widetilde{c}c_1^2}$

(c_1) $\|b\|_{L_\infty(0,T;L_{\frac{n}{2}}(\Omega))} < \dfrac{\min\{1,k_0\}}{\widetilde{c}c_1^2}$

(iv) *There exists a number $d_0 > 0$ such that $(a(x,t)-b(x,t)) \geqslant d_0$ for almost every $(x,t) \in Q_T$. In this case, the function k satisfies one of the followings:*

(a_2) *There exists a number $k_1 > 0$ such that $k(x',t) \geqslant -k_1$ for almost every $(x',t) \in \Sigma_T$, and $k_1 < \dfrac{\min\{d_0,1\}}{c_3^2}$*

(b_2) $\|k\|_{L_\infty(0,T;L_{n-1}(\partial\Omega))} < \dfrac{\min\{d_0,1\}}{c_4^2}$

Then the solution of problem (1)–(3) uniquely exists in P_0 for any $(h,\varphi) \in [L_2(0,T;(W_2^1(\Omega))^*)] \times L_2(0,T;W_2^{-\frac{1}{2}}(\partial\Omega))$.

Proof. To prove this theorem we again make use of the existence theorem from [13]. We define corresponding mappings as (4), (5), (6) such that,

$$f = \{f_1, f_2\} : P_0 \to L_2(0,T;(W_2^1(\Omega))^*) \times L_2(0,T;W_2^{-\frac{1}{2}}(\partial\Omega)); A : P_0 \to P_0$$

We obtain that the conditions of the existence theorem [13] are satisfied from the following lemmas:

Lemma 4. *f is bounded from P_0 to $L_2(0,T;(W_2^1(\Omega))^*)$, under the conditions of Theorem 2.*

Lemma 5. *f is weakly continuous from P_0 to $L_2(0,T;(W_2^1(\Omega))^*)$, under the conditions of Theorem 2.*

Lemma 6. *f and A generate a coercive pair on $L_2(0,T;W_2^1(\Omega))$, under the conditions of Theorem 2.*

Thus problem (1)–(3) is solvable in P_0 for any $(h,\varphi) \in L_2(0,T;(W_2^1(\Omega))^*) \times L_2(0,T;W_2^{-\frac{1}{2}}(\partial\Omega))$. For the uniqueness, we assume that the problem has two different solutions: $u(x,t)$ and $v(x,t)$ from P_0. If we denote by $w := u - v$ then we have the following:

$$0 = \int_0^T \int_\Omega \frac{\partial w}{\partial t} w dx dt + \int_0^T \int_\Omega (Dw)^2 dx dt +$$

$$+\int_0^T\int_\Omega (a(x,t)-b(x,t))w^2dxdt+\int_0^T\int_{\partial\Omega} k(x',t)w^2dx'dt$$

Here from the conditions of Theorem 2, we obtain contradiction of $0>0$. Hence, the solution of problem (1)–(3) uniquely exists. □

5 Existence of the solution of Problem (1)–(3) in Super Linear Case

Let $\rho>0, -1<\nu\leqslant\rho$. In this section, $P_0=L_2(0,T;W_2^1(\Omega))\cap L_{\rho+2}(Q_T)\cap W^1_{\frac{\rho+2}{\rho+1}}(0,T;(W_2^1(\Omega))^*)\cap\{u:u(x,0)=0\}$ In this case, condition ***(i)*** is denoted as (***i'''***) with the following parameters p_1, r_1, p_2, r_2:

$$p_1=p_2:=\infty\quad r_1=r_2:=\begin{cases}\dfrac{\rho+2}{\rho-\nu}\text{ if }\nu<\rho,\\ \infty\text{ if }\nu=\rho.\end{cases}$$

Theorem 3. *Let $\rho>0$, $-1<\nu\leqslant\rho$ and let the conditions (**i'''**), **(ii)** be fulfilled. Additionally the following conditions are satisfied:*

(iii) *If $-1<\nu<\rho$, then there exists a number $a_0>0$ such that $a(x,t)\geqslant a_0$ for almost every $(x,t)\in Q_T$. If $\nu=\rho$, then there exists a number $b_0>0$ such that $a(x,t)-b(x,t)\geqslant b_0$ for almost every $(x,t)\in Q_T$.*

(iv) *The function k satisfies one of the following conditions:*

(a) *There exists a number k_0 such that $k(x',t)\geqslant -k_0$ for almost every $(x',t)\in\Sigma_T$, and*[1]

$$k_0<\begin{cases}\dfrac{\min\{a',1\}}{c_3^2}\ \textit{if}\ -1<\nu<\rho,\\ \dfrac{\min\{b_0,1\}}{c_3^2}\ \textit{if}\ \nu=\rho.\end{cases}$$

[1] Here a' is a positive number such that $a'<a_0$ and, c_3, c_4 are constants of Sobolev's imbedding inequalities: $\|u\|_{L_2(\partial\Omega)}\leqslant c_3\|u\|_{W_2^1(\Omega)}$, $\|u\|_{L_{\frac{2(n-1)}{n-2}}(\partial\Omega)}\leqslant c_4\|u\|_{W_2^1(\Omega)}$.

(b) *It holds*

$$\|k\|_{L_\infty(0,T;L_{n-1}(\partial\Omega))} < \begin{cases} \dfrac{\min\{a',1\}}{c_4^2} \ \textit{if} \ -1 < \nu < \rho, \\ \dfrac{\min\{b_0,1\}}{c_4^2} \ \textit{if} \ \nu = \rho. \end{cases}$$

Then the problem (1)–(3) is solvable in P_0 for any $(h,\varphi) \in [L_2(0,T;(W_2^1(\Omega))^*) + L_{\frac{\rho+2}{\rho+1}}(Q_T)] \times L_2(0,T;W_2^{-\frac{1}{2}}(\partial\Omega))$.

Proof. To prove this theorem we again make use of the existence theorem from [13]. We define the corresponding mappings as (4), (5), (6) such that,

$$f = \{f_1, f_2\} : P_0 \to [L_2(0,T;(W_2^1(\Omega))^*) + L_{\frac{\rho+2}{\rho+1}}(Q_T)] \times \\ \times L_2(0,T;W_2^{-\frac{1}{2}}(\partial\Omega)); A : P_0 \to P_0.$$

Lemma 7. *The mapping f is bounded from P_0 to $L_2(0,T;(W_2^1(\Omega))^*) + L_{\frac{\rho+2}{\rho+1}}(Q_T)$ under the conditions of Theorem 3.*

Lemma 8. *The mapping f is weakly continuous from P_0 to $L_2(0,T;(W_2^1(\Omega))^*) + L_{\frac{\rho+2}{\rho+1}}(Q_T)$ under the conditions of Theorem 3.*

Lemma 9. *The mappings f and A generate a coercive pair on $L_2(0,T;W_2^1(\Omega)) \cap L_{\rho+2}(Q_T)$ under the conditions of Theorem 3.*

From these lemmas we obtain that all conditions of the existence theorem are satisfied for the mappings f and A. So, if we apply this theorem to problem (1)–(3), then we see that problem (1)–(3) is solvable in P_0 for any $(h,\varphi) \in [L_2(0,T;(W_2^1(\Omega))^*) + L_{\frac{\rho+2}{\rho+1}}(Q_T)] \times L_2(0,T;W_2^{-\frac{1}{2}}(\partial\Omega))$. □

6 On Uniqueness of the Solution to Problem (1)–(3)

In this section we investigate the uniqueness of the solution of the problem (1)–(3) for a special case (i.e. $\nu = \rho > -1$, $n \geqslant 3$).

Theorem 4. *Let the following conditions be satisfied for the problem (1)–(3):*
(1) *Let $\nu = \rho > -1$.*

(2) *The functions* $a, b \in \begin{cases} L_\infty(Q_T), & \rho > 0 \\ L_\infty(0,T; L_{\frac{n}{2}}(\Omega)), & \rho = 0 \\ L_{\frac{2}{|\rho|}}(0,T; L_{\frac{2n}{\rho(2-n)+4}}(\Omega)), & \rho < 0 \end{cases}$ *such that* $a(x,t) \geqslant 0$, $b(x,t) \geqslant 0$ *and* $a(x,t) \geqslant b(x,t)$ *for almost every* $(x,t) \in Q_T$.

(3) *The function* $k \in L_\infty(0,T; L_{n-1}(\partial\Omega))$, *there exists a number* $k_0 > 0$ *such that* $k(x',t) \geqslant k_0$ *for almost every* $(x',t) \in \Sigma_T$.

Then the solution of problem (1)–(3) is unique if it exists in P_0.

References

1. R. A. Adams *Sobolev Spaces*, Academic Press, New York, 1975. 268 pages.
2. B. Amman, E. Humbert *The second Yamabe Invariant*, Journal of Functional Analysis 235, 2006. Pp. 377–412.
3. C. Bandle and H. A. Levine *On the Existence and Nonexistence of Global Solutions of Reaction- Diffusion Equations in Sectorial Domains*, Transactions of the American Mathematical society Vol. 316, Num. 2, 1989. Pp. 595–622.
4. C. Bandle, H. A. Levine, Qi S. Zhang *Critical exponents of Fujita type for Inhomogeneous Parabolic equations and systems*, Journal of Mathematical Analysis and Applications 251, 2000. Pp. 624–648.
5. H. Fujita *On the blow up of solutions of the Cauchy problem for* $u_t = \Delta u + u^{\sigma+1}$ J. Fac. Sci. Univ., Tokyo, Sect. I 13, 1966. Pp. 109–124.
6. J. L. Lions *Quelques methodes de resolution des problemes aux Limities non lineaires*, Dunod, Gauthier-Villars, Paris, 1969. 554 pages.
7. J. L. Lions, E. Magenes *Non-Homogeneous Boundary Value Problems and Applications*, Volume 1, Springer-Verlag, Berlin, Heidelberg, New York, 1972. 357 pages.
8. Tor A. Kwembe *A remark on the existence anduniqueness of solutions of a semilinear parabolic equation*, Nonlinear Analysis 50, 2002. Pp. 425–432.
9. L. Ma *Boundary Value problem for a Classical Semilinear Parabolic Equation,* arXiv:1012.5861v1 [math.AP], 2010.
10. E. Öztürk , K. N. Soltanov *On Some Yamabe Type Equations*, AIP Conference Proc., Volume 1168, 2009. Pp. 377–380.
11. A. Pisante *Hardy inequalities and dynamic instability of singular Yamabe metrics*, Ann. I. H. Poincare-AN 23, 2006. Pp. 591–628.
12. S. Sato *Singular Backward Self-Similar Solutions of a Semilinear Parabolic Equation*, Discrete and Continous Dynamical Systems Series Volume 4, Number 4, 2011. Pp. 897–906.

13. K. N. Soltanov *On some modification on Navier-Stokes equations*, Nonlinear Analysis-Theory Methods and Applications, Vol. 52, Iss. 3, 2003. Pp. 769–793.
14. M. Struwe *Variational Methods Applications to Nonlinear Partial Differential Equations and Hamiltonian Systems*, Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong, Barcelona, 1990. 244 pages.
15. Qi S. Zhang *A critical behavior for some semilinear parabolic equations involving sign changing solutions*, Nonlinear Analysis 50, 2002. Pp. 967–980.

Eylem Öztürk
Department of Mathematics, Hacettepe University, 06800 Beytepe, Ankara, Turkey,
eozturk@hacettepe.edu.tr

Kamal N. Soltanov
Department of Mathematics, Hacettepe University, 06800 Beytepe, Ankara, Turkey,
soltanov@hacettepe.edu.tr

ε-ENERGIES FOR WEAKLY HYPERBOLIC OPERATORS

Giovanni Taglialatela

Key words: weakly hyperbolic equations, energy estimates

AMS Mathematics Subject Classification: 35L25

Abstract. Energy estimates are a fundamental tool to obtain many results for linear and nonlinear hyperbolic equations: well-posedness, dispersive estimates, regularity of the solutions, ... Approximated energies (or ε-energies) were introduced in [1] and [2] in order to treat non regular or degenerate hyperbolic operators of second order. There are (at least) three methods to extend the notion of ε-energies to higher order equations. We prove here that all these methods are essentially equivalent.

1 Introduction

It is well-known that for the solutions to the wave equation

$$\Box u \equiv \partial_t^2 u - \partial_x^2 u = 0$$

the *energy*

$$E(u;t) := \big\|\partial_t u(t,\cdot)\big\|_{L^2}^2 + \big\|\partial_x u(t,\cdot)\big\|_{L^2}^2$$

is constant, i.e. $E(u;t) = E(u;0)$. Similarly the solutions to the strictly hyperbolic equation

$$Pu \equiv \partial_t^2 u - a(t)\partial_x^2 u = 0$$

with $a \in \mathcal{C}^1\big([0,T]\big)$ such that $a(t) \geqslant \delta > 0$ for all $t \in [0,T]$, the energy

$$E_P(u;t) := \big\|\partial_t u(t,\cdot)\big\|_{L^2}^2 + a(t)\big\|\partial_x u(t,\cdot)\big\|_{L^2}^2 \tag{1}$$

verifies the estimate

$$E_P'(u;t) \leqslant C\, E_P(u;t),$$

for some $C \in \mathbb{R}$ depending on $\|a\|_{C^1}$ and δ.

Energy estimates are used to get many results for linear and nonlinear hyperbolic equations: well-posedness, dispersive estimates, stability of the solutions, ... Hence it is of great importance to extend the notion of energy to general higher order equations and systems of hyperbolic type. In this note we restrict ourselves

to scalar equations with coefficients depending only on the time variable. Also, for the sake of brevity, we consider only operators in one space variable.

Let

$$P(t;\partial_t,\partial_x)=\partial_t^m-\sum_{j=1}^{m}a_j(t)\partial_t^{m-j}\partial_x^j \tag{2}$$

be a hyperbolic operator of order m, that is

$$P(t;\tau,\xi)=\prod_{j=1}^{m}\big(\tau-\tau_j(t)\xi\big) \tag{3}$$

and the characteristic roots $\tau_j(t)$ are real functions for $j=1,\dots,m$. Taking the Fourier transform in the space variable we transform the Cauchy problem for P to an initial value problem for an ordinary differential equation depending on ξ:

$$\begin{cases}P(t;\partial_t,\partial_x)u=0\\ u|_{t=0}=u_0\end{cases}\quad\Longleftrightarrow\quad\begin{cases}P(t;\partial_t,i\xi)v=0\\ v|_{t=0}=v_0\end{cases},$$

where $v(t,\xi):=\mathcal{F}_{x\to\xi}\big(u(t,\cdot)\big)$ and $v_0(\xi):=\mathcal{F}_{x\to\xi}\big(u_0(\cdot)\big)$.

Assume at first that P is *strictly hyperbolic*, i.e. the functions $\tau_j(t)$ are real and distinct

$$\tau_j(t)\neq\tau_k(t)\quad\text{for any } j\neq k \text{ and } t\in[0,T].$$

We define the *energy* (*density*) of v (cf. [6, 9]) by:

$$E_P(v;t;\xi):=\sum_{j=1}^{m}\big|P_{\hat{\jmath}}(t;\partial_t,i\xi)v(t;\xi)\big|^2,\quad\text{where}\quad P_{\hat{\jmath}}(t;\tau,\xi):=\frac{P(t;\tau,\xi)}{\tau-\tau_j(t)\xi}.$$

The *energy* of u is obtained by

$$E_P(u;t)=\int_{\mathbb{R}}E_P(v;t;\xi)\,d\xi.$$

Example 1. Let $Pu\equiv\partial_t^2u-a(t)\partial_x^2u$ with $a(t)\geqslant\delta>0$ for all $t\in[0,T]$, and let

$$P_{\{1\}}(t;\tau,\xi):=\tau-\sqrt{a(t)}\,\xi\quad P_{\{2\}}(t;\tau,\xi):=\tau+\sqrt{a(t)}\,\xi,$$

hence,

$$E_P(v;t;\xi):=|P_{\{1\}}(t;\partial_t,i\xi)v|^2+|P_{\{2\}}(t;\partial_t,i\xi)v|^2=2\Big[|v_t|^2+a(t)\xi^2|v|^2\Big]. \tag{4}$$

Example 2 ($m = 3$). Let P be as in (2)-(3) with $m = 3$. Then

$$E_P(v;t;\xi) := |P_{\{1,2\}}(t;\partial_t, i\xi)v|^2 + |P_{\{2,3\}}(t;\partial_t, i\xi)v|^2 + |P_{\{1,3\}}(t;\partial_t, i\xi)v|^2,$$

where

$$P_{\{j,k\}}(t;\tau,\xi) := \tau^2 - (\tau_j + \tau_k)\tau\xi + \tau_j\tau_k\,\xi^2.$$

It is not difficult to prove that if P is strictly hyperbolic, then we have the estimates

$$C_1\Big(\sum_{j=0}^{m-1} |\xi|^{2(m-j-1)}|\partial_t^j v|^2\Big) \leqslant E_P(v;t;\xi) \leqslant C_2\Big(\sum_{j=0}^{m-1} |\xi|^{2(m-j-1)}|\partial_t^j v|^2\Big) \tag{5}$$

and

$$E_P'(v;t;\xi) \leqslant C_3\, E_P(v;t;\xi), \tag{6}$$

where the constants C_1, C_2, C_3 depend only on the coefficients of P and do not depend on $t \in [0,T]$, $\xi \in \mathbb{R}$ and v. From (5) and (6), via the Paley-Wiener Theorem, we can deduce well-posedness results in a wide class of spaces ($\mathcal{C}^\infty$, Gevrey, ...).

In [1] the notion of ε-approximate energy is introduced in order to treat strictly hyperbolic operators with non regular coefficients. In [2] a similar notion is employed to treat also weakly (i.e. non strictly) hyperbolic operators. Both papers considered only second order equations. The basic idea is to replace in the definition of energy (4) a suitable approximation a_ε of a and to choose ε dependent on ξ, so that the estimates (5) and (6) can be used again to obtain well-posedness results.

In this note we present three methods to define ε-energies for higher order operators. These methods are not entirely new, but our presentation differs slightly from the original ones. However, our principal contribution is in showing that all these methods are equivalent.

2 The method of Jannelli [7]

Let $P(t;\tau,\xi)$ be as in (2), and let

$$P_\varepsilon(t;\tau,\xi) = \prod_{j=1}^{m}\big(\tau - \tau_{j,\varepsilon}(t)\,\xi\big), \quad \varepsilon \in]0,1], \tag{7}$$

be a sequence of polynomials approximating P such that

$$\big|\tau_{j,\varepsilon}(t) - \tau_{k,\varepsilon}(t)\big| \geqslant \varepsilon, \quad (j \neq k), \quad \big|\tau_j(t) - \tau_{j,\varepsilon}(t)\big| \leqslant C^*\,\varepsilon. \tag{8}$$

Then let us consider the energies

$$\mathcal{E}^J_\varepsilon(v;t;\xi) := E_{P_\varepsilon}(v;t;\xi).$$

There are several methods to construct the approximated polynomials P_ε.

Example 3 (see [10]). Let $T_\varepsilon P := P + \varepsilon\xi\,\partial_\tau P$. If P is a hyperbolic polynomial with multiplicity $\leqslant r$, then $T_\varepsilon P$ is a hyperbolic polynomial with multiplicity $\leqslant r-1$. Then define $P_\varepsilon := T_\varepsilon \cdots T_\varepsilon P$ (iterated r times).

Example 4 (see [7]). Let $T_\varepsilon P := P - (\varepsilon\xi)^2\partial_\tau^2 P$. If P is a hyperbolic polynomial with multiplicity $\leqslant r$, then $T_\varepsilon P$ is a hyperbolic polynomial with multiplicity $\leqslant r-2$. Then define $P_\varepsilon := T_\varepsilon \cdots T_\varepsilon P$ (iterated $r/2$ times).

Example 5 (see [4]). Let $\tau_{j,\varepsilon} := \tau_j + \sqrt{-1}\,\varepsilon\, j$, then define P_ε according to (7).

Note that in this case the P_ε are not hyperbolic, nevertheless the corresponding energy is positive.

3 The method of D'Ancona-Spagnolo [5]

Let

$$\mathcal{E}^{DS}_\varepsilon(v;t;\xi) := E_P(v;t;\xi) + \sum_{r=0}^{m-1} (\varepsilon\xi)^{2(m-r)} \sum E_R(v;t;\xi),$$

where the second sum is extended over all the monic polynomials R of degree r which divide P.

Example 6 ($m = 3$). Let P be as in (2)-(3) with $m = 3$. Then

$$\begin{aligned}\mathcal{E}^{DS}_\varepsilon := {} & |P_{\{1,2\}}v|^2 + |P_{\{2,3\}}v|^2 + |P_{\{1,3\}}v|^2 \\ & + (\varepsilon\xi)^2\Big(|P_{\{1\}}v|^2 + |P_{\{2\}}v|^2 + |P_{\{3\}}v|^2\Big) + (\varepsilon\xi)^4|v|^2, \quad (9)\end{aligned}$$

where

$$P_{\{j,k\}}(t;\tau,\xi) := \tau^2 - (\tau_j+\tau_k)\tau\xi + \tau_j\tau_k\,\xi^2\,, \qquad j,k\in\{1,2,3\}\,,\ j\neq k\,, \quad (10)$$

$$P_{\{j\}}(t;\tau,\xi) = \tau - \tau_j\,\xi\,, \qquad j\in\{1,2,3\}. \quad (11)$$

4 The method of Peyser [11]

Let $P^{(j)}(t;\tau,\xi) = \dfrac{(m-j)!}{m!}\partial_\tau^j P(t;\tau,\xi)$. We set

$$\mathcal{E}_\varepsilon^P := E_P + (\varepsilon\xi)^2\, E_{P^{(1)}} + (\varepsilon\xi)^4\, E_{P^{(2)}} + \cdots + (\varepsilon\xi)^{2(m-1)}\, E_{P^{(m-1)}}.$$

Example 7 ($m = 3$). Let P be as in (2)-(3) with $m = 3$. Then

$$\mathcal{E}_\varepsilon^P := |P_{\{1,2\}}v|^2 + |P_{\{2,3\}}v|^2 + |P_{\{1,3\}}v|^2 + (\varepsilon\xi)^2\Big(|R_{\{1\}}v|^2 + |R_{\{2\}}v|^2\Big) + (\varepsilon\xi)^4|v|^2, \quad (12)$$

where $P_{\{j,k\}}(t;\tau,\xi)$ are defined in (10), whereas

$$R(t;\tau,\xi) := \frac{1}{3}\partial_\tau P(t;\tau,\xi) = \big(\tau - \lambda_1(t)\,\xi\big)\big(\tau - \lambda_2(t)\,\xi\big),$$

$$R_{\{1\}}(t;\tau,\xi) = \tau - \lambda_1(t)\,\xi, \quad R_{\{2\}}(t;\tau,\xi) = \tau - \lambda_2(t)\,\xi.$$

5 Equivalence of the ε-energies

The ε-energies $\mathcal{E}_\varepsilon^J$, $\mathcal{E}_\varepsilon^{DS}$ and $\mathcal{E}_\varepsilon^P$ are equivalent. More precisely, the following statement holds:

Theorem 1. *There exist constants C_1, C_2, C_3 independent of v such that*

$$\mathcal{E}_\varepsilon^P \leqslant C_1\,\mathcal{E}_\varepsilon^J, \quad \mathcal{E}_\varepsilon^J \leqslant C_2\,\mathcal{E}_\varepsilon^{DS}, \quad \mathcal{E}_\varepsilon^{DS} \leqslant C_3\,\mathcal{E}_\varepsilon^P.$$

Here we only sketch the proof in the case $m = 3$, but the general case can be treated in a similar way.

To show that $\mathcal{E}_\varepsilon^P \leqslant C_1\,\mathcal{E}_\varepsilon^J$ we use the Newton-Lagrange interpolation formula: If $P(\tau)$ is a polynomial with simple roots and $\deg R(\tau) < \deg P(\tau)$, then

$$R(\tau) = \sum_{j=1}^{m} \frac{R(\tau_j)}{P_{\widehat{j}}(\tau_j)}\, P_{\widehat{j}}(\tau), \quad \text{where} \quad P_{\widehat{j}}(\tau) := \frac{P(\tau)}{\tau - \tau_j}. \quad (13)$$

Thanks to (13) and (8) we can show that the $P_{\{j,k\}}$ are linear combinations with bounded coefficients of the $P_{\varepsilon,\{j,k\}}$, since we have, e.g.,

$$P_{\{1,2\}} = \frac{(\tau_3^\varepsilon - \tau_1)(\tau_3^\varepsilon - \tau_2)}{(\tau_3^\varepsilon - \tau_1^\varepsilon)(\tau_3^\varepsilon - \tau_2^\varepsilon)}\, P_{\varepsilon,\{1,2\}} + \cdots \quad \text{(similar terms)}.$$

Thus,

$$|P_{\{1,2\}}v|^2 \leqslant C\Big(|P_{\varepsilon,\{1,2\}}v|^2 + |P_{\varepsilon,\{2,3\}}v|^2 + |P_{\varepsilon,\{1,3\}}v|^2\Big)$$

Similarly,

$$\varepsilon\xi R_{\{1\}} = \frac{\varepsilon\,(\tau_3^\varepsilon - \lambda_1)}{(\tau_3^\varepsilon - \tau_1^\varepsilon)(\tau_3^\varepsilon - \tau_2^\varepsilon)}\,P_{\varepsilon,\{1,2\}} + \cdots \quad \text{(similar terms)}.$$

Since $\tau_1 \leqslant \lambda_1 \leqslant \tau_2$, by using again (8) we can show that the $\varepsilon\xi R_{\{j\}}$ are linear combinations with bounded coefficients of the $P_{\varepsilon,\{j,k\}}$. A similar estimate holds for $(\varepsilon\xi)^4|v|^2$.

Now we show that $\mathcal{E}_\varepsilon^J \leqslant C_2\,\mathcal{E}_\varepsilon^{DS}$. Let $\varepsilon_j := \tau_{\varepsilon,j} - \tau_j$, $j = 1,2,3$, so that $|\varepsilon_j| \leqslant \varepsilon$. We have the identity

$$P_{\varepsilon,\{j,k\}} = P_{\{j,k\}} - \varepsilon_j\,\xi\,P_{\{k\}} - \varepsilon_k\,\xi\,P_{\{j\}} + \varepsilon_j\varepsilon_j\xi^2$$

which gives

$$|P_{\varepsilon,\{j,k\}}v|^2 \leqslant 4\,\Big(|P_{\{j,k\}}v|^2 + (\varepsilon\xi)^2|P_{\{k\}}v|^2 + (\varepsilon\xi)^2|P_{\{j\}}v|^2 + (\varepsilon\xi)^4|v|^2\Big).$$

Finally, we show that $\mathcal{E}_\varepsilon^{DS} \leqslant C_3\,\mathcal{E}_\varepsilon^P$. Comparing (9) and (12) it is sufficient to show that the $P_{\{j\}}$, $j = 1,2,3$, are linear combinations with bounded coefficients of the $R_{\{j\}}$, $j = 1,2$. We use again Newton-Lagrange interpolation formula (13) to get

$$P_{\{1\}} = \frac{\tau_1 - \lambda_2}{\lambda_1 - \lambda_2}\,R_{\{1\}} + \frac{\tau_1 - \lambda_1}{\lambda_2 - \lambda_1}\,R_{\{2\}},$$
$$P_{\{2\}} = \frac{\tau_2 - \lambda_2}{\lambda_1 - \lambda_2}\,R_{\{1\}} + \frac{\tau_2 - \lambda_1}{\lambda_2 - \lambda_1}\,R_{\{2\}}, \quad P_{\{3\}} = \ldots.$$

Note that since $\tau_1 \leqslant \lambda_1 \leqslant \tau_2 \leqslant \lambda_2 \leqslant \tau_3$ we immediately get the boundness of $\dfrac{\tau_2 - \lambda_2}{\lambda_1 - \lambda_2}$ and $\dfrac{\tau_2 - \lambda_1}{\lambda_2 - \lambda_1}$. To estimate the other coefficients we need the following lemma on the location of the roots of the derivative of a hyperbolic polynomial.

Lemma 1 (Peyser [12]). *Let $P(\tau)$ be a hyperbolic polynomial of degree m, and let $R(\tau) := \dfrac{1}{m}\partial_\tau^j P(t;\tau,\xi)$. Denote by τ_j, $j = 1,\ldots,m$, the roots of $P(\tau)$ and by λ_j, $j = 1,\ldots,m-1$, the roots of $R(\tau)$. Then we have:*

$$\tau_j + \frac{1}{m}(\tau_{j+1} - \tau_j) \leqslant \lambda_j \leqslant \tau_{j+1} - \frac{1}{m}(\tau_{j+1} - \tau_j).$$

In particular, for $m = 3$ we have

$$\tau_1 + \frac{1}{3}(\tau_2 - \tau_1) \leqslant \lambda_1 \leqslant \tau_2 - \frac{1}{3}(\tau_2 - \tau_1),$$
$$\tau_2 + \frac{1}{3}(\tau_3 - \tau_2) \leqslant \lambda_2 \leqslant \tau_3 - \frac{1}{3}(\tau_3 - \tau_2),$$

hence,

$$\frac{1}{3}(\tau_3 - \tau_1) \leqslant \lambda_2 - \lambda_1 \leqslant \frac{2}{3}(\tau_3 - \tau_1).$$

Using this inequality we get the boundness of the coefficients $\dfrac{\tau_1 - \lambda_2}{\lambda_1 - \lambda_2}$, $\dfrac{\tau_1 - \lambda_1}{\lambda_2 - \lambda_1}$, ...

6 Applications

In this section we present some results that can been obtained by using ε-energies. Here γ^s is the space of the (uniform) Gevrey functions of order s on $\mathbb{R}$, i.e. the space of $f \in \mathcal{C}^\infty(\mathbb{R})$ for which there exists constants C, R such that

$$\sup_{x\in\mathbb{R}} \big|f^{(j)}(x)\big| \leqslant C\, R^{-j}\, j!^s \quad \text{for any } j \in \mathbb{N}.$$

Theorem 2 (see [5]). *Assume that the coefficients of P are $\mathcal{C}^2\big([0,T]\big)$, and let $B(t,x,\partial_t,\partial_x)$ be any differential operator of order $< m$. Then the Cauchy problem for $P + B$ is well-posed in γ^s if $s < \dfrac{m}{m-1}$.*

Theorem 3 (see [5], see also [13]). *Assume that the coefficients of P are $\mathcal{C}^m\big([0,T]\big)$, and let f be an entire analytic function. Let $u \in \gamma^s$, $s < \dfrac{m}{m-1}$, be a Gevrey solution of the equation*

$$Pu = f(u).$$

If $u|_{t=0}$ is analytic, then u is analytic.

Theorem 4 (see [3]). *Assume that the coefficients of P are $\mathcal{C}^m\big([0,T]\big)$, the discriminant of P, i.e. $\prod\limits_{j<k} \big(\tau_j(t) - \tau_k(t)\big)^2$, vanishes at finite order in $t = 0$ and the $\tau_j(t)$ satisfy the estimate*

$$|t|\big|\tau_j'(t)\big| \leqslant C\big|\tau_j(t) - \tau_k(t)\big| \quad \textit{for any } j \neq k.$$

Let

$$B(t;\tau,\xi)=\sum_{j=0}^{m-1}B_j(t;\tau,\xi)\quad \textit{with } \deg(B_j)=j,$$

and let us assume

$$\left|t^{m-j}\,B_j\big(t;\tau_k(\xi),\xi\big)\right|\leqslant C\left|P^{(m-j)}\big(t;\tau_k(\xi),\xi\big)\right|\quad \textit{for } k=1,\dots,j+1.$$

Then the Cauchy problem for $P+B$ is well-posed in $\mathcal{C}^\infty$ in a neighborhood of 0.

References

1. F. Colombini, E. De Giorgi, and S. Spagnolo, *Sur les équations hyperboliques avec des coefficients qui ne dépendent que du temps*, Ann. S.N.S. Pisa Cl. Sci. (4) **6** (1979), Pp. 511–559.
2. F. Colombini, E. Jannelli, and S. Spagnolo, *Well-posedness in the Gevrey classes of the Cauchy problem for a nonstrictly hyperbolic equation with coefficients depending on time*, Ann. S.N.S. Pisa Cl. Sci. (4) **10** (1983), Pp. 291–312.
3. F. Colombini and G. Taglialatela, *Well-posedness for hyperbolic higher order operators with finite degeneracy*, J. Math. Kyoto Univ. **46** (2006), Pp. 833–877.
4. P. D'Ancona and T. Kinoshita, *On the wellposedness of the Cauchy problem for weakly hyperbolic equations of higher order*, Math. Nachr. **278** (2005), Pp. 1147–1162.
5. P. D'Ancona and S. Spagnolo, *Quasi-symmetrization of hyperbolic systems and propagation of the analytic regularity*, Boll. U.M.I. (8) **1** (1998), Pp. 169–185.
6. L. Gårding, *Solution directe du problème de Cauchy pour les équations hyperboliques*, Colloques Internat. C.N.R.S. 71, Pp. 71–90 (1956).
7. E. Jannelli, *On the symmetrization of the principal symbol of hyperbolic equations*, Comm. P.D.E. **14** (1989), 1617–1634.
8. E. Jannelli, *The hyperbolic symmetrizer: theory and applications*, (A. Bove and al., eds.), Birkhäuser, 2009, Workshop in Honor of F. Colombini's 60th birthday, Siena, Italy, October 2007, pp. 113–139.
9. J. Leray, Hyperbolic Differential Equations, Princeton Univ. Press, Princeton, NJ, 1953.
10. W. Nuij, *A note on hyperbolic polynomials*, Math. Scand. **23** (1968), Pp. 69–72 (1969).
11. G. Peyser, *Energy inequalities for hyperbolic equations in several variables with multiple characteristics and constant coefficients*, Trans. A.M.S. **108** (1963), Pp. 478–490.

12. G. Peyser, *On the roots of the derivative of a polynomial with real roots*, Amer. Math. Monthly **74** (1967), Pp. 1102–1104.
13. S. Spagnolo and G. Taglialatela, *Analytic propagation for nonlinear weakly hyperbolic systems*, Commun. P.D.E. **35** (2010), Pp. 2123–2163.

Giovanni Taglialatela
Universitá di Bari, Italy, 70124 Bari, Via C. Rosalba 53, +390805049217,
taglia@dse.uniba.it

IV. Applied Analysis

IV.1. Inverse problems

(Sessions organizers: A. Hasanov, S. Kabanikhin, Y. Kurylev, A. Yagola, M. Yamamoto)

USING PARALLEL COMPUTING FOR SOLVING MULTIDIMENSIONAL ILL-POSED PROBLEMS

D. V. Lukyanenko, A. G. Yagola

Key words: Inverse problem, ill-posed problem, Tikhonov regularization, parallel computing.

AMS Mathematics Subject Classification: 45B05, 45Q05, 65F22, 68W10

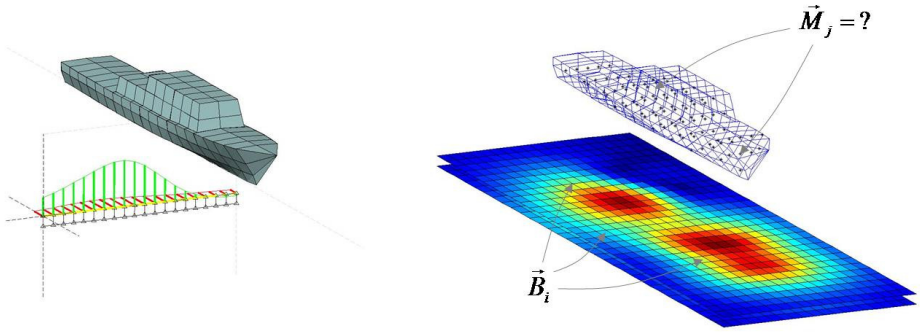
Abstract. Solving multidimensional ill-posed problems has attracted wide interests and found many practical applications. However, the most modern applications require processing a large amount of data that often very difficult to perform on personal computers. In these cases usual different methods are applied for simplification of the problem statements but these simplifications degrade the accuracy of the inverted parameters. It is supposed to solve computationally difficult applications in general form (without any simplifications) by using parallel computing that gives us an advantage the time and the accuracy. The proposed method can be efficiently applied for solving multidimensional Fredholm integral equations of the 1st kind in many areas of physics. In this paper we consider the main propositions using as basis a practical problem of restoring the magnetization parameters of a magnetic object.

1 Introduction

Key propositions of this paper we consider on the example of a practical problem of restoring magnetization parameters over a ship body. Formulation of the problem is as follows: a ship passes over a system of triaxial sensors (fig. 1a) which measure the value of the induced magnetic field. According to these values of induced magnetic fields it is necessary to restore the magnetization parameters over the hull of the ship [1]. This formulation of the problem is equal to a problem where the ship stand over the system of sensor arrays (fig. 1b).

The main difficulty of this problem solving (as in most of the other multidimensional problems) is that general solution of the problem is extremely time-taking. As a result, various simplifications are typically used to simplify the problem that reduce the dimension of the problem that allows us to solve it by using simpler methods [2].

The work was partially supported by RFBR grand 08-01-00160 and 09-01-00586-a.



(a) (b)

Figure 1. a) The ship passes over the system of triaxial sensors. b) The ship stands over the system of sensor arrays.

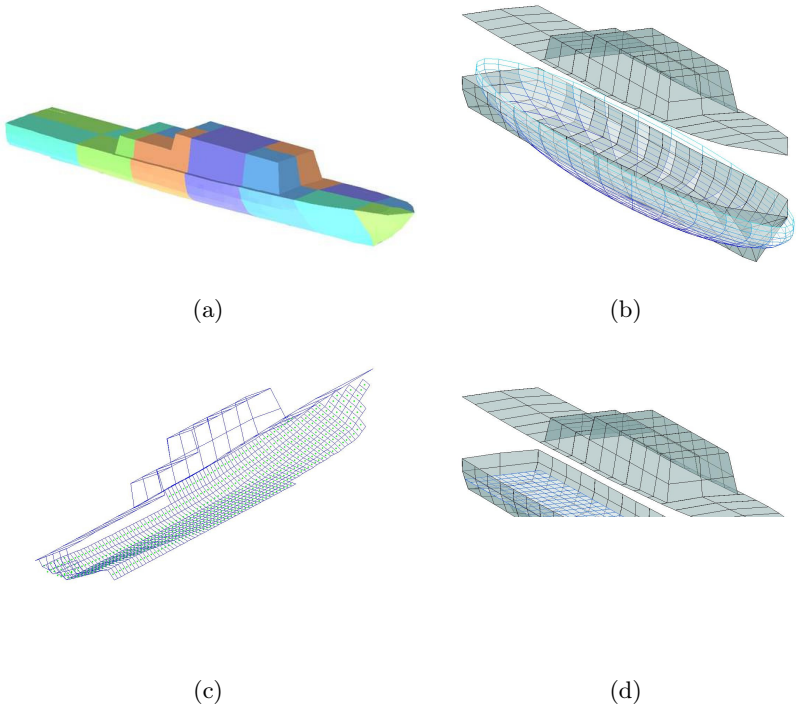
For example, it is used a partition of the vessel at sufficiently large subvolumes (fig. 2a) and restoring of the magnetization parameters of these particular elements of the partition. It is obvious that this approach can reduce the dimension of the problem but gives us only a qualitative description of the object.

Second approach is to approximate the hull of the ship by an ellipsoid of revolution (fig. 2b) for which well-known analytical transformations can be applied that can significantly reduce the dimension of the problem. But in this case we apply an assumption that only a hull is the magnetized part of the ship and inner magnetized parts are not counted. Along with the often rough approximation of the hull of the ship by ellipsoid of revolution this type of simplification also gives us only a partial picture of the object under study.

The third approach is to approximate the hull of the ship by a plane (which is applicable for very large ships) (fig. 2d) that leads to the need to solve two-dimensional integral equation of convolution type for vector functions.

But in all of these cases (similar to other multidimensional inverse problems) the simplifications decrease dimension of the problem but give us results which are useful only in specific situations [2].

And what to do if it is need to solve the problem in general? For example, what to do if the dimension of grids too huge (fig. 3)? In this case, there is only one way for solving of these problems: using parallel computing [3].



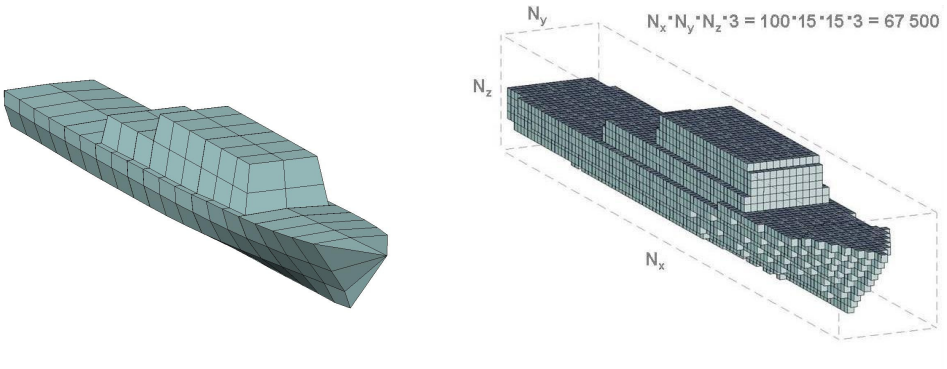
(a) (b)

(c) (d)

Figure 2. Different types of simplifications.

2 Using parallel computing

Parallel computing is a form of computation in which many calculations are carried out simultaneously, operating on the principle that large problems can often be divided into smaller ones, which are then solved concurrently ("in parallel") [4], [5]. Parallel computation can be performs on multi-processor clusters or on multi-core computers having multiple processing elements within a single machine. But not every problem can be parallelized efficiently. The speed-up of a program as a result of parallelization is observed as Amdahl's law. It states that a small portion of the program which cannot be parallelized will limit the overall speed-up available from parallelization. Any large mathematical or engineering problem will typically consist of several parallelizable parts and several non-parallelizable (sequential) parts. This relationship is given by equation $S = \frac{1}{(1-P)+\frac{P}{N}}$, where S is the speed-up of the program (as a factor of its original sequential runtime), N is number of processors and P is the fraction that is parallelizable. This puts an upper limit on



(a) (b)

Figure 3. Example of a large segmentation of the body of the ship.

the usefulness of adding more parallel execution units. It is possible to prove that parallelizable fraction for multidimensional Fredholm integral equation of the 1st kind is $\sim 0.(9)$ that gives us very high effectiveness of parallelization [3].

3 Parallelization of multidimensional ill-posed problem

The equation describing the magnetic field $\boldsymbol{B}$ of dipole sources in term of the field point position relative to the source r and equivalent magnetic moment $\boldsymbol{M}$ is defined as

$$\boldsymbol{B}_i(x_s, y_s, z_s) = \sum_{j=1}^{N} \frac{\mu_0}{4\pi} \left[\frac{3(\boldsymbol{M}_j, \boldsymbol{r}_{ij})\boldsymbol{r}_{ij}}{|\boldsymbol{r}_{ij}|^5} - \frac{\boldsymbol{M}_j}{|\boldsymbol{r}_{ij}|^3} \right] \tag{1}$$

where x_s, y_s, z_s are coordinates of a point located on the sensor planes in the Cartesian system of coordinates (x, y, z) and correspond to coordinates of the sensors i, r_{ij} is a distance between the point (x_s, y_s, z_s) and the point of the dipole source j, μ_0 is a permeability in vacuum, N is number of the dipole sources. The equation (1) can be expressed by equivalent mathematical model of the three-dimensional Fredholm integral equation of the 1st kind for vector-function [3]

$$\mathbf{A}\boldsymbol{M} = \int_{L_x}^{R_x} \int_{L_y}^{R_y} \int_{L_z}^{R_z} \mathbf{K}(s,t,r,x,y,z)\boldsymbol{M}(x,y,z)dxdydz = \boldsymbol{B}(s,t,r). \tag{2}$$

We assume that $\boldsymbol{M} \in W_2^2$, $\boldsymbol{B} \in L_2$, and operator $\mathbf{A}$ with kernel $\mathbf{K}$ is continuous and unique. Suppose that instead of accurately known $\bar{\boldsymbol{B}}$ and operator $\mathbf{A}$ their approximate values $\boldsymbol{B}_\delta$ and $\mathbf{A}_h$ are known, such that $\|\boldsymbol{B}_\delta - \bar{\boldsymbol{B}}\|_{L_2} \leqslant \delta$, $\|\mathbf{A} - \mathbf{A}_h\|_{W_2^2 \to L_2} \leqslant h$. The problem (2) is ill-posed and it is necessary to build the regularizing algorithm based on the minimization of the Tikhonov functional [6]. When we solve minimization problem by conjugate gradient method it is necessary to calculate values of the Tikhonov functional $F^\alpha[\boldsymbol{M}]$ and its gradient $\operatorname{grad} F^\alpha[\boldsymbol{M}]$.

The finite-difference approximation of the Tikhonov functional is

$$F^\alpha[\boldsymbol{M}] = \Phi[\boldsymbol{M}] + \alpha\Omega[\boldsymbol{M}], \tag{3}$$

$$\Phi[\boldsymbol{M}] = \sum_{j_1=1}^{N_s}\sum_{j_2=1}^{N_t}\sum_{j_3=1}^{N_r}\sum_{n=1}^{3} h_s h_t h_r \times \times \left[\sum_{i_1=1}^{N_x}\sum_{i_2=1}^{N_y}\sum_{i_3=1}^{N_z}\sum_{m=1}^{3} h_x h_y h_z K^{nm}_{j_1 j_2 j_3 i_1 i_2 i_3} M^m_{i_1 i_2 i_3} - B^n_{j_1 j_2 j_3}\right]^2, \tag{4}$$

$$\begin{aligned}\Omega[\boldsymbol{M}] = {} & h_x h_y h_z \sum_{i_1=1}^{N_x}\sum_{i_2=1}^{N_y}\sum_{i_3=1}^{N_z}\sum_{m=1}^{3} \left(M^m_{i_1 i_2 i_3}\right)^2 + \ldots \\ & + \frac{h_y h_z}{h_x^3} \sum_{i_1=2}^{N_x-1}\sum_{i_2=1}^{N_y}\sum_{i_3=1}^{N_z}\sum_{m=1}^{3} \left(M^m_{i_1+1 i_2 i_3} - 2M^m_{i_1 i_2 i_3} + M^m_{i_1-1 i_2 i_3}\right)^2 + \ldots \\ & + \frac{h_x h_z}{h_y^3} \sum_{i_1=1}^{N_x}\sum_{i_2=2}^{N_y-1}\sum_{i_3=1}^{N_z}\sum_{m=1}^{3} \left(M^m_{i_1 i_2+1 i_3} - 2M^m_{i_1 i_2 i_3} + M^m_{i_1 i_2-1 i_3}\right)^2 + \ldots \\ & + \frac{h_x h_y}{h_z^3} \sum_{i_1=1}^{N_x}\sum_{i_2=1}^{N_y}\sum_{i_3=2}^{N_z-1}\sum_{m=1}^{3} \left(M^m_{i_1 i_2 i_3+1} - 2M^m_{i_1 i_2 i_3} + M^m_{i_1 i_2 i_3-1}\right)^2, \end{aligned} \tag{5}$$

approximation of its gradient is

$$\begin{aligned}\left(\operatorname{grad} F^\alpha[\boldsymbol{M}]\right)^m_{i_1 i_2 i_3} &= \frac{\partial F^\alpha[\boldsymbol{M}]}{\partial M^m_{i_1 i_2 i_3}} = \ldots \\ &= 2h_x h_y h_z \sum_{j_1=1}^{N_s}\sum_{j_2=1}^{N_t}\sum_{j_3=1}^{N_r}\sum_{n=1}^{3} h_s h_t h_r K^{nm}_{j_1 j_2 j_3 i_1 i_2 i_3} \times\end{aligned}$$

$$\times \left[\sum_{l_1=1}^{N_x} \sum_{l_2=1}^{N_y} \sum_{l_3=1}^{N_z} \sum_{p=1}^{3} h_x h_y h_z K^{np}_{j_1 j_2 j_3 l_1 l_2 l_3} M^p_{l_1 l_2 l_3} - B^n_{j_1 j_2 j_3} \right] + \alpha \frac{\partial \Omega[\boldsymbol{M}]}{\partial M^m_{i_1 i_2 i_3}}, \quad (6)$$

These formulas (4), (6) contain large groups of independent summands. It allows to divide large problem of calculating the functional and its gradient into smaller ones which are then solved "in parallel" [3–5].

We propose schemes of calculating value of the Tikhonov functional (residual) (fig. 4) and its gradient (fig. 5). The main idea of the schemes is that each process calculates its own "big summands" (for examle, "square" of functional (4) at the first schem). It is clear that all other calculations are carried out sequentially. But since the time of the serial code is negligible for large dimensions of the grid, the time spent on all these calculations can be considered equal to zero. It remains a question whether the impact on the effectiveness of parallelizing sequential computations of the smoothing functional $\Omega[\boldsymbol{M}]$ in the calculation of the Tikhonov functional is or not.

From formula (4) for the finite-difference approximation of the Tikhonov functional it can be seen that the residual consist of $3 \times N_s \times N_t \times N_r$ independent summands, and the smoothing functional (5) consists of $3 \times (\leqslant 4)$ equivalent (in terms of time spent on computing) groups of summands. So for the calculation of the Tikhonov functional part of parallelizable actions is $P \geqslant \frac{3N_sN_tN_r}{3N_sN_tN_r+12} = \frac{N_sN_tN_r}{N_sN_tN_r+4}$, that shows that even with a relatively small number of input data, for example $N_s = N_t = N_r = 10$ (the grids correspond to a domain of a known vector function $\boldsymbol{B}$), parallelized part of the computation is more than 0.97. Modern applied problems require the handling of a much larger number of input data, and therefore the proportion of parallelizable computations tends to 1.0, which proves a very high effectiveness of algorithms parallelization.

In our case the amount of sequential code is closer to 0.0 with increasing N_s, N_t and N_r that provides strong efficiently parallelizing of proposed algorithm that allows to process large problems.

4 Some examples of calculations

As a result of implementation of the describing method distribution of the magnetization parameters over the volume of the ship was obtained. Some results of calculations are represented (fig. 6). Typical dimensions that correspond to real applications are $N_x = 100$, $N_y = 15$, $N_z = 15$. Input data simulated a real experiment and correspond to grids $N_s = 4000$, $N_t = 3$, $N_r = 2$ that relevant to 67500 unknowns and 72000 equations (error of input data is equal to 1,5%).

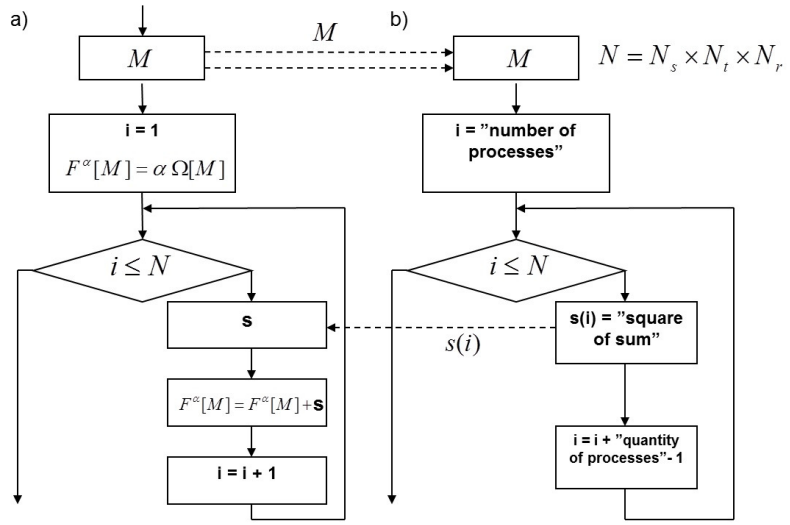


Figure 4. The scheme of calculating value of the Tikhonov functional for a) zero process, b) non-zero processes.

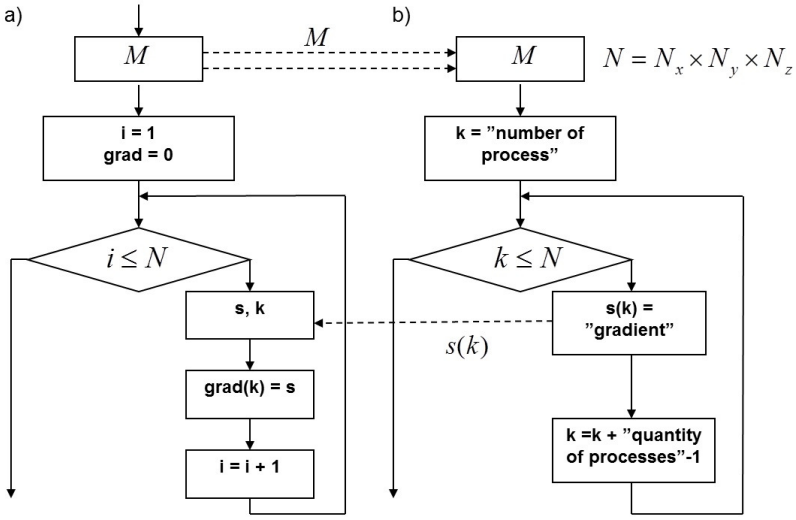


Figure 5. The scheme of calculating value of the gradient of the Tikhonov functional for: a) zero process; b) non-zero processes.

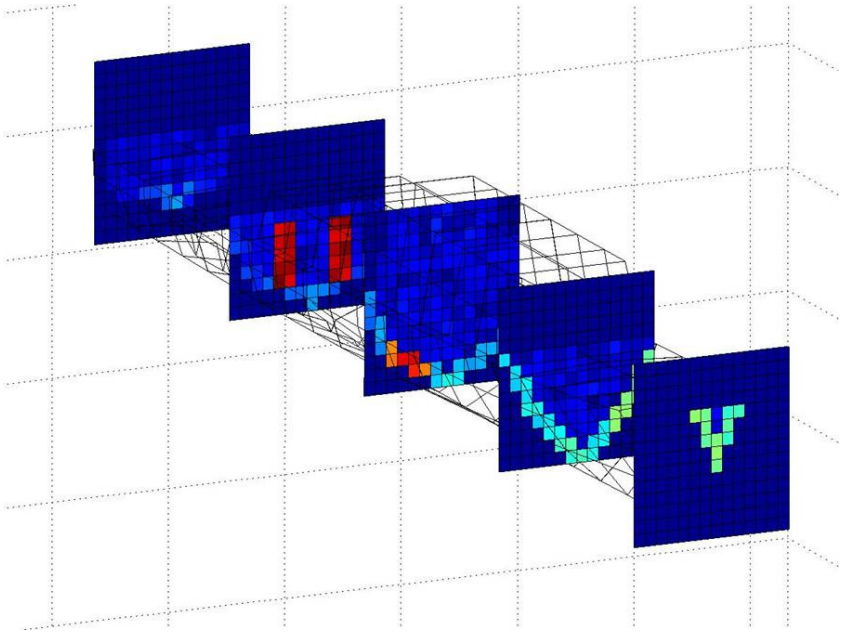


Figure 6. The results of the inversion of the magnetization parameters over the volume of the ship (it represented 5 slices of the module inverted vector function $\boldsymbol{M}$).

The computation time was approximately 29 hours with using 200 processors (Intel Xeon E5472 3.0 GHz). So long computations associated with the using of the regularizing algorithm, which requires repeated finding the minimum of the functional to be minimized for each value of the regularization parameter α.

For calculations were used Computing Cluster of the Moscow State University.

References

1. Y.H. Pei, A.G. Yagola *Constraint Magnetization Parameter Inversion by Iterative Tikhonov Regularization* In "International Conference "Inverse and Ill-Posed Problems of Mathematical Physics", dedicated to Professor M. M. Lavrentiev on the occasion of his 75-th birthday August 20-25, 2007, Novosibirsk, Russia. Abstracts of Section № 3.", 2007, p. 1-2, http://www.math.nsc.ru/conference/ipmp07/section3.htm.
2. D. V. Lukyanenko, Y. H. Pei, A. G. Yagola, Liu Gui-Rong, N. A. Evdokimova *Numerical methods for solving ill-posed problems with constraints and applications to inversion of the magnetic field* In "International Conference "Inverse and Ill-Posed Problems of Mathematical Physics", dedicated

to Professor M.M. Lavrentiev on the occasion of his 75-th birthday August 20-25, 2007, Novosibirsk, Russia. Abstracts of Section № 3.", 2007, p. 1-2, http://www.math.nsc.ru/conference/ipmp07/section3.htm.

3. D. V. Lukyanenko, A. G. Yagola *Application of multiprocessor systems for solving three-dimensional Fredholm integral equations of the first kind for vector functions* Numerical Methods and Programming, 2010, v. 11, pp. 336-343 (in Russian).
4. V. V. Voevodin *Mathematical foundations of parallel computing* World Scientific Publ, Co, Singapore, 1992.
5. Vl. V. Voevodin *Theory and Practice of Parallelism Detection in Sequential Programs* Programming and computer software, 1992.
6. A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov, A. G. Yagola *Numerical methods for the solution of ill-posed problems* Kluwer Academic Publishers, Dordrecht, 1995.

D. V. Lukyanenko
Contacts for the first author: Lomonosov Moscow State University, Faculty of Physics, Department of Mathematics, Russia, 119991, Moscow, lukyanenko@physics.msu.ru

A. G. Yagola
Contacts for the second author: Lomonosov Moscow State University, Faculty of Physics, Department of Mathematics, Russia, 119991, Moscow, yagola@physics.msu.ru

ERROR ESTIMATIONS IN LINEAR INVERSE PROBLEMS IN ORDERED SPACES

A. G. Yagola, Yu. M. Korolev

Key words: ill-posed problems, error estimation

AMS Mathematics Subject Classification: 47A52, 65F22, 65N20

Abstract. We consider an inverse problem for an operator equation $Az = u$. The exact operator A and the exact right-hand side u are unknown. Only their upper and lower estimations are available. We provide techniques of calculating upper and lower estimations for the exact solution belonging to a compact set in this case, as well as a posteriori error estimations. We obtain approximate solutions with an optimal a posteriori error estimate. We also make use of a priori information about the exact solution, e.g. its monotonicity and convexity. The developed software package was applied to solving practical ill-posed problems.

1 Introduction

In this paper we consider operator equations

$$Az = u, \tag{1}$$

where $z \in Z$, $u \in U$, $A\colon Z \to U$ is a linear injective operator, Z and U are normed spaces. According to J. Hadamard, the problem (1) is called well-posed, if

1) the solution exists $\forall u \in U$;

2) it is unique;

3) it depends continuously on the problem parameters, e.g. small perturbations in the right hand side u and the operator A lead to small changes in the solution.

Unfortunately, there are many applications, for which the condition 3) does not hold. In many cases, a stable solution of this problem can be found using the Tikhonov regularization method [4]. This solution tends to the exact solution as the error in the initial data tends to zero.

It is well known that it is impossible to estimate the error of an approximate solution of an ill-posed problem without a priori information about the exact solution. In this paper we consider the case when the exact solution belongs to a compact set $M \subset Z$. In this case it is possible to estimate the error of the approximate solution.

This work was supported by the RFBR grants 11-01-00040a and 09-01-00586-a.

Other a priori information can be also used (for example, sourcewise representation of the exact solution, see [1]).

2 A posteriori error estimates

In practice the exact operator A and the exact right-hand side u are unknown and only their estimations A_h and u_δ are available, such that

$$\|u - u_\delta\| \leqslant \delta, \quad \|Az - A_h z\| \leqslant h\|z\|. \tag{2}$$

The pair $\eta = (h, \delta)$ describes the error in the initial data.

If we know a priori that the exact solution belongs to a convex compact set $M \subset Z$, then the set of approximate solutions

$$Z_\eta = \{z \in M \colon \|A_h z - u_\delta\| \leqslant \delta + h\|z\|\} \tag{3}$$

has a finite diameter ([8]). The set Z_η is not convex. Hence, we consider another set of approximate solutions

$$Z_\eta^C = \{z \in M \colon \|A_h z - u_\delta\| \leqslant \delta + hC\}, \tag{4}$$

where $C = \max_{z \in M} \|z\|$. Obviously, $Z_\eta \subseteq Z_\eta^C$.

Consider the following functional on Z_η^C:

$$\varphi(z) = \max_{\zeta \in Z_\eta^C} \|\zeta - z\|. \tag{5}$$

It is interpreted as the a posteriori error estimate of the approximate solution z. We will come back to the properties of $\varphi(z)$ later on in this paper. Note that the value $\max_{z \in Z_\eta^C} \varphi(z)$ is the a priori error estimate for the solution of Eqn. (1).

Alternatively, if Z and U are ordered spaces, the approximate initial data may by given by the following inequalities.

$$u^l \leqslant u \leqslant u^u, \quad \forall z \geqslant 0 \quad A^l z \leqslant Az \leqslant A^u z, \tag{6}$$

where A^l and A^u are linear operators. In $\mathbb{R}^n$, for example, partial order may be introduced as following:

$$x \leqslant y \quad \Leftrightarrow \quad x_i \leqslant y_i, \quad i = 1, \ldots, n. \tag{7}$$

Another example of an ordered set is the set of integrable functions on D. We say of two integrable functions f and g that $f \leqslant g$ if $f(\omega) \leqslant g(\omega)$ for any subset $\omega \subset D$, which has non-zero measure.

Consider an integral equation

$$\int\limits_D K(\xi, x)z(x)dx = u(\xi), \quad x \in D \subset \mathbb{R}^n, \quad \xi \in \mathbb{T} \subset R^m. \tag{8}$$

Here $K(\xi, x)$ is a continuous positive bounded function on $T \times D$, $z \in L_1(D)$, $u \in C(T)$. If we know upper and lower estimations of $K(\xi, x)$

$$0 \leqslant K^l(\xi, x) \leqslant K(\xi, x) \leqslant K^u(\xi, x) \leqslant C, \quad (\xi, x) \in T \times D, \tag{9}$$

where $C < +\infty$, then

$$A^l z = \int\limits_D K^l(\xi, x)z(x)dx, \quad A^u z = \int\limits_D K^u(\xi, x)z(x)dx. \tag{10}$$

Since $u^l \leqslant Az = u \leqslant u^u$ and $A^l z \leqslant Az \leqslant A^u z$, we come up with following inequalities

$$A^l z \leqslant u^u, \quad A^u z \geqslant u^l, \tag{11}$$

which hold for the exact solution.

In this case the set of approximate solutions is

$$Z_{app} = \{z \in M \colon A^l z \leqslant u^u, \quad A^u z \geqslant u^l\}. \tag{12}$$

The set Z_{app} is convex since the operators A^l and A^u are linear.

2.1 Solutions with an optimal a posteriori error estimate

Consider the functional $\varphi(z)$, which is defined on a convex set of approximate solutions $\tilde{Z}$ ($\tilde{Z} = Z_\eta^C$ or $\tilde{Z} = Z_{app}$). The following proposition holds true.

Proposition 1. The functional $\varphi(z)$ is convex.

Proof. Consider $z = \lambda z_1 + (1 - \lambda)z_2$, $z_1, z_2 \in \tilde{Z}$, $\lambda \in [0, 1]$. Since $\tilde{Z}$ is convex, $z \in \tilde{Z}$.

$$\begin{aligned} \|\zeta - z\| &= \|\zeta - \lambda z_1 - (1 - \lambda)z_2\| = \\ &= \|\lambda(\zeta - z_1) + (1 - \lambda)(\zeta - z_2)\| \leqslant\leqslant \lambda\|\zeta - z_1\| + (1 - \lambda)\|\zeta - z_2\|. \end{aligned} \tag{13}$$

Hence,

$$\varphi(z) = \max_{\zeta\in\tilde{Z}} \|\zeta - z\| \leqslant \max_{\zeta\in\tilde{Z}} (\lambda\|\zeta - z_1\| + (1-\lambda)\|\zeta - z_2\|) \leqslant$$
$$\leqslant \lambda \max_{\zeta\in\tilde{Z}} \|\zeta - z_1\| + (1-\lambda) \max_{\zeta\in\tilde{Z}} \|\zeta - z_2\| = \lambda\varphi(z_1) + (1-\lambda)\varphi(z_2), \quad (14)$$

which proves the proposition. □

Since a convex function achieves its minimum on a convex set, we can introduce a solution

$$z^* = \arg\min_{z\in\tilde{Z}} \varphi(z) = \arg\min_{z\in\tilde{Z}} \max_{\zeta\in\tilde{Z}} \|\zeta - z\|, \quad (15)$$

which has an optimal a posteriori error estimate.

3 Finite dimensional approximation

Let us return to Eqn. (8). Consider a finite family of subsets $\{\omega_i\}_{i=1}^n$, such that $\cup_{i=1}^n \omega_i = D$, $\omega_i \cap \omega_j = 0$, $i \neq j$. Approximate $z(x)$ by a piecewise-constant function $\tilde{z}(x)$, so that $\tilde{z}(x) = \tilde{z}_i = \int\limits_{\omega_i} z(x)dx$, $x \in \omega_i$. Denote by $\tilde{z}$ the vector of approximation coefficients. Introduce a grid $\{\xi_j\}_{j=1}^m$ on T. Then we can rewrite (11) as follows.

$$\begin{aligned} \sum_{i=1}^n S_{i,j}^l \tilde{z}_i &\leqslant u_j^u, \quad j = 1, \ldots, m, \\ \sum_{i=1}^n S_{i,j}^u \tilde{z}_i &\geqslant u_j^l, \quad j = 1, \ldots, m, \end{aligned} \quad (16)$$

where

$$u_j^{l,u} = u^{l,u}(\xi_j), \quad S_{i,j}^{l,u} = \int\limits_{\omega_i} K^{l,u}(\xi_j, x)dx. \quad (17)$$

Other approximations for $z(x)$ could be used (for example, piecewise-linear approximation). Then we would get other expressions for $S_{i,j}^{l,u}$. Introduce vectors $u^{l,u} = \{u_j^{l,u}\}$ and matrices $S^{l,u} = \{S_{i,j}^{l,u}\}$. Then we can rewrite (4.1) in matrix form.

$$S^l\tilde{z} \leqslant u^u, \quad S^u\tilde{z} \geqslant u^l. \quad (18)$$

Approximate the convex compact set of a priori restrictions $M \subset Z$ by a convex polyhedron $\hat{M} \subset \mathbb{R}^n$. Then the set of approximate solutions is approximated by a

convex polyhedron

$$\hat{Z}_{app} = \{z \in \hat{M} \colon S^l z \leqslant u^u, \quad S^u z \geqslant u^l\}. \tag{19}$$

Let us find lower and upper estimations for the exact solution $\tilde{z}^l$ and $\tilde{z}^u$. Since $\tilde{z} \in \hat{Z}_{app}$, we can take any lower and upper bound of $\hat{Z}_{app}$ as $\tilde{z}^l$ and $\tilde{z}^u$, respectively. They solve following problems

$$\begin{aligned} \tilde{z}_i^l &= \arg \min_{z \in \hat{Z}_{app}} z_i, \quad i = 1, \dots n, \\ \tilde{z}_i^u &= \arg \max_{z \in \hat{Z}_{app}} z_i, \quad i = 1, \dots n. \end{aligned} \tag{20}$$

These problems can be efficiently solved using standard linear programming algorithms [3]. We use standard MATLAB® algorithms in our computations. Note that in general $\tilde{z}^l$ and $\tilde{z}^u$ do not belong to $\hat{Z}_{app}$ or even to $\hat{M}$.

4 Computation of the error estimate

To calculate the a posteriori error estimate (5), we need to maximize a convex function φ on a convex polyhedron $\hat{Z}_{app}$, which is defined by

$$\hat{Z}_{app} = \{z \colon Gz \leqslant q\}, \tag{21}$$

where the matrix inequality $Gz \leqslant q$ consists of two inequalities $S^l z \leqslant u^u$ and $-S^u z \leqslant -u^l$ and the inequalities which define the set of a priori restrictions $\hat{M}$. In fact, it is a nonstandard problem of quadratic programming.

A convex function, which is defined on a polyhedron, achieves its maximum at a vertex of this polyhedron ([6]). So, our aim is to find all vertices of $\hat{Z}_{app}$.

We use the following algorithm to find them [7]. Start with any convex polyhedron $W_0 \colon \hat{Z}_{app} \subset W_0$. Then find its intersection with the half-space $G_1 z \leqslant q_1$ (G_1 is the first row of G). This intersection itself is also a convex polyhedron - W_1. We continue this procedure of finding intersections with half-spaces $G_i z \leqslant q_i$ until $i = k$, where k is the number of rows in G. The final polyhedron $W_k = \hat{Z}_{app}$.

The polyhedron is defined by its vertices z^i, $i = 1 \dots M$ and the $M \times M$ connectivity matrix C with logical elements

$$C_{i,j} = \begin{cases} 1, & \text{if there is an edge between the vertices } i \text{ and } j, \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

At step i, if a vertex j does not belong to the half-space $G_i z \leqslant q_i$, it is deleted from the list of vertices. For each edge that connects this vertex with another one, which belongs the i-th half-space, we find its intersection with the hyperplane $G_i z = q_i$ and add it to the list of vertices. We do the same for all vertices which do not belong to the i-th half-space.

After that it has to be cleared which pairs of new vertices are connected with an edge. The following criterion holds true ([5]).

Theorem 1. *Let z_1, z_2 and z_3 be vertices of a polyhedron W. Let P_1, P_2 and P_3 be the sets of hyperplanes where z_1, z_2 and z_3 belong to, respectively. Then z_1 and z_2 are connected with an edge in W if and only if for any z_3 (different from z_1 and z_2)*

$$(P_1 \cap P_2) \setminus P_3 \neq \emptyset.$$

When all vertices $\tilde{z}^i$, $i = 1, \ldots, M$ of $\hat{Z}_{app}$ are known, we just need to choose the vertex $\tilde{z}^k$ that provides maximum of the norm $\|\tilde{z} - \tilde{z}^k\|$ for the approximate solution $\tilde{z}$.

The computational complexity of this algorithm depends greatly on the connectivity matrix C. If at step i the number of vertices belonging to the half-space $G_i z \leqslant q_i$ is n_1 and n_2 is the number of vertices outside of this half-space ($n_1 + n_2 = k$ - the number of vertices in the polyhedron W_{i-1}) then there may be up to $n_1 n_2$ new vertices, i.e. the polyhedron W_i may have up to $n_1(n_2 + 1)$ vertices, which is about $k^2/4$. In worst case (if the connectivity matrix is dense) the number of vertices can grow exponentially as the number of steps increases. But if the connectivity matrix is sparse, the growth of the number of vertices is not so grave. We usually worked with polyhedrons defined in a 10-dimensional space (k was about 1000) on an ordinary laptop.

If the number of vertices in the polyhedron W_i is k, we need at most $k(k-1)/2 \cdot (k-2)$ operations to decide, which vertices are connected with an edge, which means that at each step we need $O(k^3)$ operations. We will go in detail in our further publications.

A 3D illustration of this algorithm is provided in Fig 1. The edges of the polyhedron W_{i-1} are shown with a dotted line and the edges of the polyhedron W_i are shown with a solid line.

To find a solution with optimal a posteriori error estimate, we have to minimize a convex function $\varphi(z)$ on a convex set Z_{app}. Standard optimization methods such as projections of conjugate gradients method or conditional gradient method can be applied (see, for example, [2]).

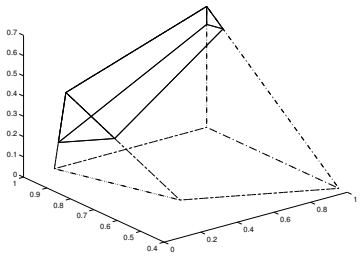


Figure 1. Illustration for the algorithm of finding all vertices of a polyhedron

5 Example

Consider a 1D Fredholm integral equation of the 1st kind

$$\int_0^1 K(\xi, x)z(x)dx = u(\xi), \quad \xi \in [0, 1] \tag{23}$$

with exact kernel $K(\xi, x) = \frac{1}{1+100(\xi-x)^2}$. Let $z \in M \subset Z = L_2[0, 1]$, $u \in U = C[0, 1]$. The set of a priori restrictions M is the set convex non-increasing functions, which is compact in L_p, $p \geqslant 1$. Moreover, natural partial order is introduced in Z and U. Let the exact solution be $\bar{z} = 5 - xe^{1-x}$. The exact right-hand side is $u(\xi)$ (we can calculate it). Suppose that only following estimates for the kernel and the right-hand side are available

$$\begin{gathered} K^l(\xi, x) = \frac{1}{1 + (100 + d)(\xi - x)^2}, \quad K^u(\xi, x) = \frac{1}{1 + (100 - d)(\xi - x)^2}, \\ u^l(\xi) = u(\xi) * 0.999, \quad u^u(\xi) = u(\xi) * 1.001. \end{gathered} \tag{24}$$

where $d > 0$ is the known error in the variable $(\xi - x)^2$ held constant. We used the value $d = 1$ in our computation.

A priori information about the exact solution (boundedness, monotonicity, convexity) can be expressed with the following inequalities:

$$\begin{gathered} 0 \leqslant z_i \leqslant 5, \quad i = 1, \ldots, n, \\ -z_i + z_{i+1} \leqslant 0, \quad i = 1, \ldots, n - 1, \\ -z_{i-1} + 2z_i - z_{i+1} \leqslant 0, \quad i = 2, \ldots, n - 1. \end{gathered} \tag{25}$$

In Fig. 2, the lower and upper estimates z^l and z^u for the exact solution are shown with a dashdotted line, for the case when all a priori information is used. The exact solution is the solid line. The number of segments is 10, the number of right-hand side approximation points is also 10.

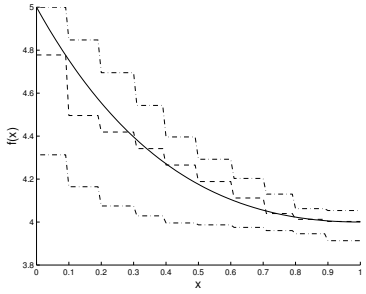


Figure 2. Lower and upper estimations of the exact solution (dashdot line), as well as the solution with optimal a posteriori error estimate (dashed line). the exact solution is the solid line

The dashed line in the same figure is the solution with optimal error estimate. In this example the optimal relative error is $\varphi(z^*)/\|z^*\| = 0,04$. We use the following expression for the norm: $\|z\| = \left(\sum_i z_i^2\right)^{1/2}$.

This example shows the effectiveness of the described methods for error estimations in linear ill-posed problems.

6 Conclusion

We have described methods of a posteriori error estimation of approximates solutions of inverse problems in partially ordered sets. We suppose that some a priori information about the exact solution is available, i.e. the information that it belongs to a compact set. We also have provided numerical algorithms for the computation of the error estimate, as well as some examples.

References

1. K.Yu. Dorofeev and A.G. Yagola. The method of extending compacts and a posteriori error estimates for nonlinear ill- posed problems. *Journal of Inverse and Ill-Posed Problems*, 12(6):627–636, 2004.
2. D.G. Luenberger. *Linear and nonlinear programming.* ADDISON-WESLEY PUBLISHING COMPANY, 2 edition, 1984.

3. Pablo Pedregal. *Introduction to Optimization*. Springer, New York, 2003.
4. A.N. Tikhonov, A.V. Goncharsky, V.V. Stepanov, and A.G. Yagola. *Numerical Methods for the Solution of Ill-Posed Problems*. Kluwer, Dordrecht, 1995.
5. V.N. Titarenko and A.G. Yagola. Method of cutting of convex polyhedrons and its applications in ill-posed problems. *Computational Methods and Programming (in Russian)*, 1(1):10–15, 2000.
6. V.N. Titarenko and A.G. Yagola. The problems of linear and quadratic programming for ill-posed problems on some compact sets. *Journal of Inverse and Ill-Posed Problems*, 11(3):311–328, 2003.
7. A.G. Yagola and V.N. Titarenko. Numerical methods and regularization techniques for the solution of ill-posed problems. In H.R.B. Orlande, editor, *Inverse Problems in Ingeneering: Theory and Practice*, volume 1, pages 49–58. E-papers, Rio de Janeiro, 2002.
8. A.G. Yagola and V.N. Titarenko. Using a priori information about a solution of an ill-posed problem for constructing regularizing algorithms and their applications. *Inverse Problems in Science and Engineering*, 15(1):3–17, January 2007.

A. G. Yagola
Faculty of Physics, Lomonosov Moscow State University, Russia, 119991, Moscow, GSP-1, 1-2 Leninskiye Gory, Email: yagola@physics.msu.ru

Yu. M. Korolev
Faculty of Physics, Lomonosov Moscow State University, Russia, 119991, Moscow, GSP-1, 1-2 Leninskiye Gory, Email: um.korolev@physics.msu.ru

IV.3. Medical mathematics

(Sessions organizers: R. Gilbert, Yu. Rappoport, V. Yakushev)

MATHEMATICAL AND INTERNET TECHNOLOGIES UNDER THE KERATOCONUS TREATMENT

A. A. Kasparov, E. A. Kasparova, J. M. Rappoport

Key words: Mathematical technologies, Internet technologies, keratoconus treatment, morphometry, endothelium

AMS Mathematics Subject Classification: 62P10, 94A08

Abstract. The quantitative and qualitative changes in the transplant endothelium after keratoplasty for keratoconus patients in various periods after surgery are analysed. The website keratoconus.ru created under the technical support of the Russian Academy of Sciences is described. The information about the typical refractive and clinical forms of the disease, its pathogenesis, contemporary methods of surgery and correction is collected and analyzed by means of internet. A forum and FAQ are created also.

1 Matematical-statistical analysis of the corneal endothelium microscopy results after penetrating keratoplasty

The quantitative and qualitative changes in the transplant endothelium after keratoplasty for keratoconus patients in various periods after surgery are analysed. The criteria for the choice of donor material with consideration of endothelial status for keratoconus patients is suggested. The method of reflecting microscopy of the endothelium is used for differential diagnosis of tissue incompatibility reactions and infections recurrence in patients. The influence of some endothelial protectors for the loss of endothelial cells is analysed. The minimal count of cells necessary for the transplant to remain transparent is estimated. The automatic image analysis is used under the endothelium microscopy. The computer graphics is used for analysis of keratoconus dynamics.

Electronic microscopy of corneal endothelium is important for the analysis of donor material for keratoplasty [1,2]. Let's investigate the dynamics of the quantity of endothelium cells in the process of donor graft retention after the penetrating keratoplasty [3-11].

Work was partially supported by the Ministry of Science and Technology of Russian Federation.

The mathematical-statistical processing of the endothelial microscopy results was performed. The purpose of this research was the establishment of the dependence of density of endothelial cells and corneal thickness as surgery outcome index in dependence from type of the surgery and endothelial protector, diagnosis, diameter of donor transplant and endothelial cells loss percent. So the dynamics of the quantity of endothelium cells in the process of donor graft retention was studied for different cases. The averaged characteristics of cells quantity and percentage ratio of opacities and edemas between all cases were evaluated, diagrams of average characteristics (cells quantities) depending on time after the siurgery were constructed. The results were considered satisfactory for the cells quantities more then 1200.

The sampling from 199 operated patients, 200 eys, was taken for conducted analysis. 28 parameters were considered for every patient. So the table from 200 lines and 28 columns was constructed.

The 1st parameter was the consequent number of the patient's line in table, 2nd - patient's family name, 3rd - patient's age in the point of surgery, 4th - sex (m, f), 5th - number of ambulatory card (five digit number), 6th - transplant's diameter in mm, 7th - protector's type ("h" - hansurid, "n" - healon, "0" - absent), 8th - type of surgery ("0" - ordinary penetrating keratoplasty, "1" - penetrating keratoplasty with intraocular lense (IOL) without reconstruction, "2" - penetrating keratoplasty without IOL with reconstruction, "3" - penetrating keratoplasty with IOL with reconstruction), 9th - reaction of tissue incompatibility ("0" -absent, term of onset), 10th - cells quantity of donor eye, 11th - loss percent of donor eye (in percents), 12th - pachymetria (thickness) of donor eye in mm, 13th and 14th - thickness and cells quantity of operated eye in term until 1st week after the keratoplasty respectively, 15th and 16th - in term 1 week - month, 17th and 18th - the same in term of 1 - 3 months, 19th and 20th - the same in term of 3 months - 1 year, 21st and 22nd - in term of 1 - 3 years, 23rd and 24th - in term of 3 - 5 years, 25th and 26th - more then 5 years, 27th and 28th - diagnosis and code of diagnosis.

The additional type of the surgery ((A) - autokeratoplasty, (T) - therapeutic keratoplasty) was indicated in the second column also if it was the case.

Let's note that a part of dates was absent in the columns 14 - 26 in the connection with the impossibility of its measurement and all sums were calculated on the basis of filled positions of the table only.

The following diagnoses were taken into the consideration: 00 - leukoma of nonclear etiology, 01 - herpes leukoma - keratitus outcome, 02 - keratoconus, 03 - inborn dystrophia, 04 - gained dystrophia, 05 - transplant's opacity, 06 - scar, 07 - perforation, 08 - corneal opacity, 09 - corneal ulcer, 10 - keratitus, 11 - herpes keratitus.

Let's clarify that the diameter of the cornea - number expressed decimal fraction with two signes before comma and after comma; cells quantity - four digits integer number.

A big number of samplings from general tables population was done on parameter values as in one column and as at once in two or three columns.

The following samplings were done and their comparative analysis between them and with general population was carried on: on column 2 with parameter (A), with parameter (T); on column 7 with parameters "0", "h" and "n"; on column 8 with parameters "0", "1", "2" and "3", on column 28 with parameters "00"-"11". The mean value of quantities is calculated in columns 10, 14, 16, 18, 20, 22, 24, 26 for every sampling.

Let's introduce the following notations. Let's $a_{ij}^{(s)}$ - value of j-th column for i-th patient of the sampling, $a_j^{(s)}$ - mean value in j-th column for given sampling, $n^{(s)}(n_j^{(s)})$ - number of the patients of given sampling (number of the patients of given sampling for which the dates of observations exist) for every sampling. We take into the account the general population when symbol upwards is absent. Then we obtain for given sampling under the assumption that missing dates supposed equal zero

$$a_j^{(s)} = \frac{1}{n_j^{(s)}} \sum_{i=1}^{n^{(s)}} a_{ij}^{(s)}. \tag{1}$$

We obtain for general population

$$a_j = \frac{1}{n_j} \sum_{i=1}^{n} a_{ij}. \tag{2}$$

So we obtain, for example, for the mean value of cells quantity in term 1 - 3 years after the surgery in the case of sampling on 8th column of patients with ordinary penetrating keratoplasty

$$a_{22}^{(8,"0")} = \frac{1}{n_{22}^{(8,"0")}} \sum_{i=1}^{n^{(8,"0")}} a_{i22}^{(8,"0")}.$$

The samplings on 11th column with loss perecentage from 0 till 3, 4 - 6, 7 - 10, 11 - 15, more then 15 and on 6th column depending on the transplant's diameter 5.0 - 6.5 mm (partial keratoplasty), 7.0 - 8.5 mm (subtotal), equal or more then 9.0 mm (total) were done also.

The samplings were compared between themself in all terms of the dynamics of alteration. The graphs and diagrams were used for the analysis of mentioned alterations. The graphs and diagrams were studied separately for samplings on values of every column and all graphs and diagrams were studied altogether for all samplings from every column.

The percentage of favorable outcomes (without reaction of tissue incompatibility, edema and corneal opacity) was calculated for every sampling by formula

$$P_f^{(s)} = \frac{n_f^{(s)}}{n^{(s)}} \times 100\%, \tag{3}$$

where $n_f^{(s)}$ - number of favorable outcomes in the sampling.

The consideration of mean values samplings graphs on column 2, i.e. therapeutical and autokeratoplasty, shows that their results (on dynamics of cells quantity alteration on time after the surgery) were not satisfactory till previous time though therapeutical keratoplasty gives little better results.

The consideration of the samplings mean values on column 7, on protector's presence and type, shows that the use of healon gives significant advantages, but the use of hansurid practically doesn't increase the cells quantity in the comparison with the protector's absence.

The analysis of cells quantity mean values for the samplings on column 8 confirms the fact that the ordinary penetrating keratoplasty without complications gives the best results. The presence of eye reconstruction has little bigger influence on the surgery outcome then the intraocular lense in the connection with the decreasing of cells quantity though the results remain still good.

The consideration of mean values for samplings on 11th column shows that the surgery result better in sense of cells quantity as percentage loss of donor eye smaller.

Analysis of mean values for samplings on 28th column on illnesses diagnosis certified that the keratoplasty surgery gives the best results for keratoconus. The results were quite satisfactory and just good for all types of leukomas, gained dystrophia and cornea opacities. The number of patients with corneal perforation and ulcer and herpes keratitus was not sufficient for analysis. Less succesful results were obtained for inborn dystrophia and keratitus. The samplings on two parameters on 7th and 8th columns altogether were done also. They confirm the conclusions done before though some from these samplings less informative in the connection with small number of observations.

The mean unbiased quadratic deviation (dispersion) was calculated for all samplings by formulas

$$b_j^{(s)} = \frac{1}{n_j^{(s)}} \sum_{i=1}^{n^{(s)}} (a_{ij}^{(s)})^2,\ \sigma_j^{(s)} = \sqrt{b_j^{(s)} - (a_j^{(s)})^2},\ \sigma_{Nj}^{(s)} = \sigma_j^{(s)} \sqrt{\frac{n_j^{(s)}}{n_j^{(s)} - 1}}. \qquad (4)$$

It was interesting to calculate the confidence boundaries $d_j^{(s)}$ pointed out that the value of cells quantity for the patient from this sampling will be with the probability 95interval $(a_j^{(s)} - d_j^{(s)}, a_j^{(s)} + d_j^{(s)})$. They were evaluated by formulas

$$d_j^{(s)} = \sigma_{Nj}^{(s)} \frac{T(n_j^{(s)} - 1)}{\sqrt{n_j^{(s)}}}. \qquad (5)$$

where $T(k) = T_{1-\alpha}(k)$ - values of STUDENT'S t-distribution for number of degrees of freedom k and probability $1 - \alpha = 0.95$.

It was necessary to perform additional comparative analisis in the connection with small volume of some separate samplings which was necessary for the verification of the comparison's reliability.

Let's You have two samplings on the same column. Then these samplings may be distinguishable with confidence $1 - \alpha$, or they may be indistinguishable on the STUDENT'S criteria in other case. This analysis was used, for example, for the analysis of samplings from 7th column.

All computations were carried on personal computers.

2 Websites for keratoconus patients

The website keratoconus.ru was created by a doctor and a keratoconus patient in 2002 under the technical support of the Russian Academy of Sciences. It was decided to do that due to a great amount of mistakes in diagnostics and treatment of keratoconus. The information about typical refractive and clinical forms of the disease, its pathogenesis, the contemporary methods of surgery and correction was collected and analyzed. A forum and FAQ were created also. The information about other corneal diseases - herpetic keratitis, adenovirus infection, dystrophies, urgent help etc. was presented at the website also. According to the feedback of our patients - the website helped them in seeking professional help for their diseases, explained many facts about pathogenesis of the corneal pathology and demonstrated that there are good outcomes in the situations that were regarded hopeless before.

Later two other websites were launched - crosslinking.ru and keraring.ru that were dedicated to collagen crosslinking and implanting of kerarings for the correction of keratoconus. Now we are working on the website dedicated to herpetic keratitis and adenovirus eye infection. The profiles on facebook and in livejournal are opened also, but it is the task for the future to launch them. It's necessary to mention in the conclusion that internet is a very effective tool in educating of our patients and doctors and that it's a good source of patients for every clinic and doctor.

The creation of English version of the website is planned. The search function exists on the website. The profiles in the social networks vkontakte.ru , odnoklassniki.ru , facebook.com are created also. The website is created on the basis of great experience of ophthalmologists and surgeons of the Scientific Research Institute of Eye Diseases of the Russian Academy of Medical Sciences. It was the purpose to give information for the patient about the modern achievements of ophthalmology more convenient for his concrete case under the creation of this website It's possible only on the basis of instant improvement of applicable technologies, use of the modern and safe equipment, professionals of the highest qualification. The ophthalmology is one of the most rapid developed fields of medicine so the website was created on the basis of the latest improvements and elaborations of Russian and foreign specialists, scientific and clinical research studies, participation in the international meetings and educational courses. The website is designated mostly for patients but may be useful for the ophthalmologists also. The website is not recommended for the diagnostics of the illnesses and determination of the treatment yourself in the connection with danger. But the page "Contacts" exists where you can find the addresses and phone numbers for clinics and doctors. E-mail address for contacts is given. It's planned for the future that the patients from long-distance regions perform their eye tests and observations on the digital medical equipment and send them by E-mail for the investigations and recommendations. It will be the step to the realization of the idea of virtual patient and mobile patient.

You go to the principal page of the website firstly. The following columns exist: adenovirus eye illnesses, bullez keratopaty, herpetic eye illnesses, cornea dystrophia, keratoconus, other cornea illnesses, emergency help. The following subwebpages exist: principal, news, about the website, contacts, doctors with a lot of concrete additional information.

The website keraring.ru describes the purposes of this type of eye surgery. The following columns exist: "what is it keraring?", mechanism of the method's effect, indications, contraindications, complications, FAQ, video and contacts.

The website crosslinking.ru gives information about the method of collagen crosslinking. The following webpages exist: "what is it crosslinking?", history of

the method, mechanism of the method's effect, indications, contraindications, path of the surgery, results, complications, FAQ, riboflavin - vitamin B_2 and contacts.

References

1. S.E.Avetisov, V.R.Mamikonian *Keratorefractive surgery*, Poligran publishing house, Moscow, 1993.-120pp. [in Russian].
2. A.K.Sharma *Morphometry. Applications to medical sciences*, Macmillan India Limited, New Delhi, 1996.-289pp.
3. A.A.Kasparov, N.V.Ermakov and J.M.Rappoport *Endothelium bioprotector protective effect on the transplant in reconstructive surgery on the anterior segment of the eye*, Annals of ophthalmology, **105** (1989), N 4, 13–17 [in Russian].
4. A.A.Kasparov, N.V.Ermakov and J.M.Rappoport *Time course of donor transplant endothelium status in patients after perforating keratoplasty*, Annals of ophthalmology, **106** (1990), N 5, 12–17 [in Russian].
5. M.N.Churkina, J.M.Rappoport *The use of computer graphics for keratoconus estimation*, Proceedings of the 3rd International workshop on automation of scientific investigations, Puschino, (1990), 246–247 [in Russian].
6. A.V.Gubin, V.I.Dymkov, N.V.Ermakov, A.A.Kasparov, S.O.Novikov and J.M.Rappoport *Hand and automatic image analysis under endothelium microscopy*, Proceedings of the 3rd International workshop on automation of scientific investigations, Puschino, (1990), 271–276 [in Russian].
7. A.A.Kasparov, N.V.Ermakov and J.M.Rappoport *The morphometry of donor transplant endothelium after perforating keratoplasty*, Proceedings of the 4th ISAAC Congress. Advances in Analysis. Mathematisches Institut, Muenchen, (1993).
8. A.A.Kasparov, N.V.Ermakov and J.M.Rappoport *On some aspects of keratoplasty analysis*, Informatics and Medicine, "Nauka", Moscow, (1997), 168–188 [in Russian].
9. A.A.Kasparov, N.V.Ermakov and J.M.Rappoport *The morphometry of donor transplant endothelium after perforating keratoplasty*, Imperial College of Science, Technology and Medicine, Department of Mathematics, London, 2001, preprint 01P/002, 8p.
10. N.V.Ermakov, J.M.Rappoport *Infocommunication technologies in ophthalmology clinics*, Medicine in the mirror of informatics, "Nauka", Moscow, 2008, 236–240 [in Russian].
11. J.M.Rappoport Fourier *method and corneal transplant technology*, Abstracts of Satellite Conference of ICM2010 "Mathematics in science and technology",

New Delhi, August 19 - 27 2010, 231–232.

A. A. Kasparov
Scientific Research Institute of Eye Diseases, Russian Academy of Medical Sciences, Rossolimo street, 11, Moscow, Russia

E. A. Kasparova
Scientific Research Institute of Eye Diseases, Russian Academy of Medical Sciences, Rossolimo street, 11, Moscow, Russia

J. M. Rappoport
Institute for Computer Aided Design, Russian Academy of Sciences, Vlasov street, Building 27, Apt.8, 117335 Moscow Russia, E-mail:jmrap@landau.ac.ru

UNIFORM ASYMPTOTICS OF RUIN PROBABILITIES FOR LÉVY PROCESSES

M. Kelbert, F. Avram, I. Sazonov

Key words: Ruin probability, Cramér–Lundberg model, Lévy process, Cremona equation, Saddle-point approximation, Fresnel integral

AMS Mathematics Subject Classification: 60G51, 62P05

Abstract. In this paper we obtain, for spectrally negative Lévy processes X, uniform approximations for the finite time ruin probability

$$\Psi(t,u) = P_u[T \leqslant t], T = \inf\{t \geqslant 0 : X(t) < 0\},$$

when $u = X(0)$ and t tend to infinity such that $v = u/t$ is constant, and the so-called Cramér light-tail conditions are satisfied.

1 Introduction

Let $X(t) = u - Y(t)$, $u > 0$, and $Y(t)$ be a Lévy process on $\mathbf{R}$ with the symbol

$$\kappa_Y(z) = \log \mathbf{E} e^{z(u - X(1))} = \kappa_X(-z). \tag{1}$$

Define the moment of ruin with the initial capital u as

$$\tau_u = \inf\{t : X(t) \leqslant 0\}. \tag{2}$$

Let $\Psi(t,u) = 1 - R(t,u) = \mathbf{P}\{\tau_u < t\}$ be the ruin probability, and $R(t,u)$ be the survival probability. In contrast to [3] we want to study the asymptotics of $\Psi(t,u)$ as $t \to \infty, u \to \infty$ and $u/t \to v$.

Assume that the Cramér light tail condition holds: $\exists \gamma > 0$ such that $\kappa_Y(\gamma) = 0$, $\kappa'_Y(\gamma) < \infty$. Let $v_{cr} = \kappa'_Y(\gamma)$. Define by z_v the so-called Cramér tilt, i.e. the real solution of equation $\kappa'_Y(z_v) = v$. Let $\tilde{z}_v$ be the so-called adjoint tilt defined by

$$\tilde{z}_v = \max[z \in \mathbf{R} : z \neq z_v, \kappa_Y(z) = \kappa_Y(z_v).] \tag{3}$$

Finally, define the Legendre transform

$$\kappa^*_Y(v) = \sup_z [zv - \kappa_Y(z)]. \tag{4}$$

Then for $v < v_{cr}$

$$\Psi(t,u) \approx \frac{e^{-\kappa_Y^*(v)t}}{\sqrt{2\pi t \kappa_Y''(z_v)} z_v} + C_\gamma e^{-\gamma u} \tag{5}$$

with $C_\gamma = |\frac{\kappa'(0)}{\kappa'(\gamma)}|$. The coefficient C_γ coincides with

$$\lim_{u\to\infty} e^{\gamma u}\Psi(u) = -\frac{\kappa_Y'(0)}{\kappa_Y'(\gamma)}$$

where $\Psi(u) = \mathbf{P}(\tau_u < \infty)$. For $v > v_{cr}$

$$\Psi(t,u) \approx \frac{e^{-\kappa_Y^*(v)t}}{\sqrt{2\pi t \kappa_Y''(z_v)}} \Big(\frac{1}{z_v} - \frac{1}{\tilde{z}_v}\Big). \tag{6}$$

This asymptotic expansion breaks down on the Stokes line $v = v_{cr}$.

2 The Schrödingher-Bachelier-Lévy formula for Brownian motion with drift

Let $Y(t) = \sigma B(t) - ct$, $\kappa_Y(z) = \frac{\sigma^2 z^2}{2} - cz$, $\gamma = \frac{2c}{\sigma^2}$, $z_v = \frac{c+v}{\sigma^2}$, $\tilde{z}_v = \frac{c-v}{\sigma^2}$, $\kappa_Y^*(v) = \frac{(c+v)^2}{2\sigma^2}$. In this case the ruin probability may be found explicitly:

$$\Psi(t,u) = \bar{\Phi}\Big(\frac{u+ct}{\sigma\sqrt{t}}\Big) + e^{-\frac{2cu}{\sigma^2}}\Phi\Big(\frac{-u+ct}{\sigma\sqrt{t}}\Big). \tag{7}$$

If $u = vt$ and $t \to \infty$ write down the asymptotic expansion of (7) for $v < v_c$

$$\Psi(t,u) \approx \frac{1}{\sqrt{2\pi t}} e^{-\frac{(c+v)^2}{2\sigma^2}t} \frac{\sigma}{(c+v)} + e^{-\frac{2cvt}{\sigma^2}} \tag{8}$$

and for $v > v_{cr}$

$$\Psi(t,u) \approx \frac{1}{\sqrt{2\pi t}} e^{-\frac{(c+v)^2}{2\sigma^2}t} \Big(\frac{\sigma}{(c+v)} + \frac{\sigma}{(v-c)}\Big). \tag{9}$$

Expressions (8),(9) form a particular case of (5),(6). As before the asymptotic expansion breaks down on the Stokes line $v_{cr} = c$.

3 Smoothing of Stokes discontinuities

Asymptotic analysis of integrals with a pole near the saddle point has roots going back to [7]. Here we present a refined method of obtaining an asymptotic expansion of integrals of the type

$$I = \frac{1}{2\pi i}\int\limits_{\Gamma} f(z,v)e^{tF(z,v)}\mathrm{d}z = \frac{1}{2\pi i}\int\limits_{\Gamma} \frac{g(z,v)}{z-z_p(v)}e^{tF(z,v)}\mathrm{d}z. \tag{10}$$

Let $F(z,v)$ and $g(z,v)$ be holomorphic functions and $F(z,v)$ possesses a single saddle point. Denote by F_p, F_s and g_p, g_s the values of the phase function and out-of-exponent function at the pole and the saddle point, respectively. Suppose that for $v < v_{cr}$ the contour Γ may be deformed into a steepest descent contour Γ_s by crossing only one pole z_s. Using an explicit expression for the modified Fresnel integral ([4]) we get

$$I \approx g_p e^{tF_p}\Phi\Big(sign(v-v_{cr})\sqrt{2t(F_p-F_s)}\Big) + \frac{e^{tF_s}}{2\pi i}\int\limits_{\mathbf{R}} h_1(s)e^{-ts^2}\mathrm{d}s. \tag{11}$$

The standard saddle point method provides an asymptotic expansion of the last integral in (11). Here we use the change of variables $s(z) = \sqrt{F_s - F(z)}$ and define an inverse function $z = z(s)$ in a neighborhood of a saddle point. Next,

$$h(s) = \frac{s-s_p}{z(s)-z_p}g(z(s))z'(s)$$

and

$$h_1(s) = \frac{h(s)-h_p}{s-s_p}. \tag{12}$$

Observe that $h_p = g_p$. The expression $d_{ps} = sign(v-v_{cr})\sqrt{2(F_p-F_s)}$ is usually called the Dingle singulant. Substuting $s=0$ into (12) we obtain

$$h_{1,s} = \frac{h_s-h_p}{-s_p} = \frac{1}{-s_p}\left(-\frac{s_p}{z_s-z_p}g_s z_s' - g_p\right) = \frac{g_s z_s'}{z_s-z_p} + \frac{g_p}{s_p} = f_s\sqrt{\frac{2}{(-F_s'')}} + \frac{g_p}{s_p}.$$

Combining all terms together we obtain

$$I \approx g_p e^{tF_p}\Big[\Phi(d_{ps}\sqrt{t}) + \frac{\varphi(d_{ps}\sqrt{t})}{d_{ps}\sqrt{t}}\Big] + \frac{f_s e^{tF_s}}{\sqrt{2\pi t(-F_s'')}} + O(t^{-3/2}) \tag{13}$$

4 Cramér-Lundberg model.

Next, we derive an asymptotic expansion in the case of the classical Cramér-Lundberg process $Y(t) = -ct + \sum_{j=1}^{N(t)} Y_j$. Here $N(t)$ is the Poisson process of rate λ, the IID jumps $Y_j \sim \mathrm{Exp}(\beta)$ with mean $m = \beta^{-1}$ (cf. [1]). In this case,

$$\kappa_Y(z) = cz\big(\frac{\rho}{1-mz} - 1\big),\ \rho = \frac{\lambda}{\beta c} = \frac{1}{1+\vartheta},\ \gamma = \beta - \frac{\lambda}{c}. \tag{14}$$

The eventual ruin probability has the form $\Psi(u) = \rho e^{-\gamma u}$. The critical velocity $v_c = \kappa_Y'(\gamma) = c(\rho^{-1} - 1) = c\vartheta$. Next, the convex conjugate of the symbol

$$\kappa_Y^*(v) = \Big(\sqrt{\beta(c+v)} - \sqrt{\lambda}\Big)^2.$$

The Cramér tilt equation is $\kappa_Y'(z) = -c + \frac{\lambda\beta}{(\beta - z)^2} = v$, hence $z_v = \beta\big(1 - \frac{1}{v_1\vartheta_1}\big)$, and the adjoint tilt $\tilde{z}_v = \beta\big(1 - \frac{v_1}{\vartheta_1}\big)$. Here $\vartheta_1 = \sqrt{1+\vartheta}$, $v_1 = \sqrt{1 + v/c}$. Finally, the saddle-point constant

$$D_v = \frac{1}{z_v} - \frac{1}{\tilde{z}_v} = \frac{\lambda v}{(\sqrt{\lambda\beta(c+v)} - \lambda)(\sqrt{\lambda\beta(c+v)} - c\beta)}.$$

It is convenient to specify the integral representation of the solution in terms of the so-called Cremona equation (3): $\kappa_Y(\zeta) = \kappa_Y(z)$, we denote its real solution as $z = \chi(\zeta)$. Selecting ζ as a new independent variable we obtain the integral representation of the ruin probability in the form

$$\Psi(t,u) = \frac{1}{2\pi i}\int\limits_{\zeta_0 - i\infty}^{\zeta_0 + i\infty} f(\zeta) e^{tF(\zeta)} \mathrm{d}\zeta \tag{15}$$

with the phase function

$$F(\zeta) = \kappa_Y(\zeta) - \chi(\zeta)v \tag{16}$$

and prefactor

$$f(\zeta) = \frac{\kappa_Y'(\zeta)}{\kappa_Y'(\chi(\zeta))}\Big(\frac{1}{\chi(\zeta)} - \frac{1}{\zeta}\Big). \tag{17}$$

Specifying (15) in the case of the symbol (14) we obtain that $\chi(\zeta) = \beta\frac{\zeta-\gamma}{\zeta-\beta}$, and the phase function

$$F(\zeta) = F_s + c\frac{(\zeta-\zeta_s)^2}{\beta-\zeta}, \tag{18}$$

$$F_s = -\kappa_Y^*(v) = -\lambda(\vartheta_1 v_1 - 1)^2, \zeta_s = \beta(1 - v_1/\vartheta_1). \tag{19}$$

Next, specify the prefactor

$$f(\zeta) = \frac{\rho}{\zeta} + \frac{1}{\zeta-\gamma} + \frac{\lambda/c - (1+\rho)(\zeta-\beta)}{(\zeta-\beta)^2} + \psi(\zeta) \tag{20}$$

where $\psi(\zeta)$ is a regular function. In fact, $F(\zeta)$ has two saddle points $\zeta_{s;1,2} = \beta\big(1 \pm \frac{v_1}{\vartheta_1}\big)$ but the contour can only be deformed to pass through the saddle point $\zeta_s = \zeta_{s;1} = \beta\big(1 - \frac{v_1}{\vartheta_1}\big)$. The uniform asymptotics is significantly different from the "naive saddle point approximation"

$$\Psi(t,u) \approx SP(t,u) = \frac{f_s e^{tF_s}}{\sqrt{2\pi t(-F_s'')}}$$

with

$$f_s = \frac{\rho v}{\lambda v_1^2(v_1\vartheta_1 - 1)(v_1/\vartheta_1 - 1)}, \quad -F_s'' = \frac{2c^2}{\lambda\vartheta_1 v_1}.$$

The saddle point $z_s(v) = \beta(1 - v_1/c_1)$ intersects the pole $z = 0$ when $v_{cr} = c\big(\frac{\lambda\beta}{c} - 1\big) = c\vartheta$, and it never intersects the pole $z = \gamma$. Let us apply (11) with $z_p = 0, g_p = \rho, F_p = -\gamma v$. The uniform appoximation takes the form

$$\Psi(t,u) \approx SP(t,u) + \rho e^{-\gamma v t}\Big(\Phi(d_0\sqrt{t}) - \frac{\varphi(d_0\sqrt{t})}{d_0\sqrt{t}}\Big). \tag{21}$$

Here the Dingle singulant $d_0 = \sqrt{2}\big(\sqrt{c\beta} - \sqrt{\lambda(1+v/c)}\big)$ vanishes at v_{cr}.

5 Laplace transform of survival probability

It is instructive to present an expression for the double Laplace tranform of the survival probability

$$\tilde{R}(q,z) = \int_0^\infty\int_0^\infty R(t,u)e^{-qt-zu}\mathrm{d}u\mathrm{d}t.$$

In terms of the symbol $\kappa_X(z)$ and the inverse function $\kappa_X^{-1}(q)$ (the unique non-negative root is selected)

$$\tilde{R}(q,z) = \frac{\frac{1}{z} - \frac{1}{\kappa_X^{-1}(q)}}{\kappa_X(z) - q}. \tag{22}$$

Finally, we present the derivation of (22) in the case of compound Poisson process $Y(t) = -ct + \sum\limits_{i=1}^{N(t)} Y_i$ where $N(t)$ is a Poisson process of rate λ and IID RVs $\{Y_i\}$ has a CDF $B_Y(y)$. The survival probability $R = R(t,u)$ satisfies the equation (cf. [6])

$$R_t - cR_u + \lambda R = \lambda R \star b \tag{23a}$$

with the boundary condition

$$\lim_{u\to\infty} R(t,u) = 1 \ \forall t \tag{23b}$$

and initial condition

$$R(0,u) \equiv 1 \tag{23c}$$

where $R \star b = \int_0^u R(t,u-y)\mathrm{d}B_Y(y)$ and B_Y stands for the CDF of the claim sizes. First, substitute $t' = \lambda t$ and $u' = \lambda u/c$ to reduce the number of parameters. Performing the Laplace transform with respect to u we obtain the initial value problem

$$\tilde{R}_t - \kappa(z)\tilde{R} = -g(t), \quad \tilde{R}(0,z) = \frac{1}{z}. \tag{24}$$

Here $g(t) = R(t,0)$ is an unknown function and $\kappa(z) = z - 1 + \tilde{b}(z)$. The solution of problem (23) has the form

$$\hat{R}(q,z) = \frac{1/z - \tilde{g}(q)}{q - \kappa(z)}. \tag{25}$$

In the absense of the pole we have $1/z - \tilde{g}(q) = 0$ or $\tilde{g}(q) = \left[(\kappa^{-1}(q))\right]^{-1}$. Thus, relation (25) coincides with (22).

References

1. Arfwedson G., Reseach in collective risk theory I, II, *Skandinavisk Aktuarietidskrift*, V. 37, 1954, 191-223; V.38, 1955, 53-100
2. Asmussen S., *Ruin probabilities*. Word Scientific, Singapore, 2000

3. Bardorff-Nielsen O.E., Schmidli H., Saddlepoint approximation for the probability of ruin in finite time. *Scand. Actuarial Journ.*, V.2, 1995, 169-188
4. Berry M.V., Uniform asymptotic smoothing of Stokes's discontinuities, *Proc.R. Soc. Lond. A*, V. 422, 1989, 7-21
5. Champan S.J. On the non-universality of the error function in the smoothing of Stokes discontinuities. *Proceedings of the Royal Society, Ser. A*, V.452, 1996, 2225-2230
6. Knessl C., Peters C.S., Exact and asymptotic solutions for the time-dependent problem of collective ruin I, *SIAM Journ. Appl. Math.*, V.54, 1994, N 6, 1745-1767
7. Van der Waerden B.L., On the method of saddle point, *Journ. Applied Scientific Research*, V. 2, 1952, N.1, 33-45

M. Kelbert
Dept of Mathematics, Swansea University, Singleton Park, Swansea, SA2 8PP, UK.
International Institute of Earthquake Prediction and Mathematical Geophysics RAS

F. Avram
Dept de Mathémathiques, Université de Pau, France

I. Sazonov
Civil Engineering, Swansea University, Singleton Park, Swansea, SA2 8PP, UK. Institute of Atmospheric Physics RAS

AN OUTBREAK SPREAD AND TRAVELLING WAVES IN SPATIALLY DISTRIBUTED POPULATIONS

M. Ya. Kelbert, I. A. Sazonov, M. B. Gravenor

Key words: SIR model, Travelling waves, Epidemic modeling, Infective disease, Biosurveilance

AMS Mathematics Subject Classification: 92C50, 37N25

Abstract. Mathematical models, based on the SIR (susceptible-infected-removed) process, have long been used to analyse epidemics of infectious disease. We consider spatial aspects of interacting SIR populations by introducing a one-dimensional lattice of SIR nodes. We obtain an accurate approximation for the propagation speed of traveling-wave type solutions. When coupling coefficients are randomly distributed, the average speed of propagation is shown to slow down. The critical reaction time between initial registration of an epidemic and the actual intervention before the number of infected reaches a critical proportion is studied in a stochastic framework. We develop a two-stage model of developed epidemic describing the evolution as a deterministic system with randomized initial conditions linked to the stochastic stage when the number of infected is small and the fluctuations are essential.

1 Travelling waves in a chain of SIR centres

A 1D lattice of susceptible/infected/removed (SIR) epidemic centers with a weak coupling and a finite characteristic migration time is considered numerically and analytically.

Travelling wave-like solutions preserving their shape and speed (see Figure 1) are found over a wide parameter range, and explicit formulae for the speed of these waves are obtained. For a nearest-neighbour interaction when the transport terms $i_{n-1\to n}$ and $i_{n+1\to n}$ are accounted only, the evolution of the system is described by

$$\frac{d}{dt}s_n = -\rho s_n i_n,\ \frac{d}{dt}i_n = (\rho s_n - 1)i_n + \frac{d}{dt}i_{n-1\to n} + \frac{d}{dt}i_{n+1\to n},\ n = 0, \pm 1, \pm 2, \dots \quad (1)$$

where $s_n = S_n/N, i_n = I_n/N$ are shares of susceptible, S_n, and infected, I_n, respectively, in the nth node (N is the total population in every node); $\rho = \beta N/\alpha$ is reproduction number, (here β and α are the contamination and recovery rates,

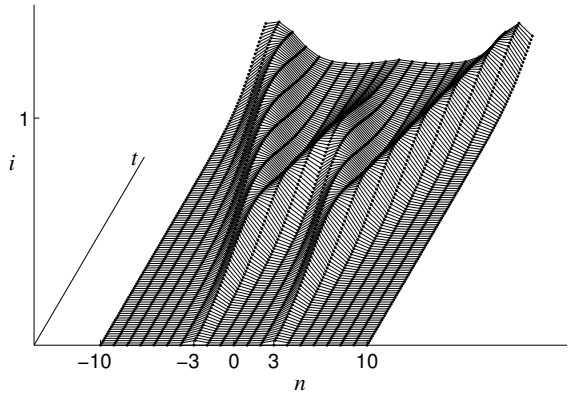


Figure 1. Propagation of waves from nodes $n = -3$ and $n = +3$ with different initial numbers of infected: $i_{-3}(0) = 10^{-2}$, $i_{+3}(0) = 10^{-3}$

respectively); the time is dimensionlized by the recovery rate α; the transport terms can be written in accordance with diffusion-like model of migration:

$$i_{n\pm1\to n}(t) = \int_{-\infty}^{\infty} i_{n\pm1}(t-t')\dot{g}(t')\mathrm{d}t, \quad g(t) = \varepsilon(1-e^{-t/\tau})\,\theta(t). \tag{2}$$

where ε is the coupling coefficient (share of time a given individual spents in the neighbour node), τ is the characteristics migration time.

Travelling wave is a solution to (1)–(2) of the form $i_n(t) = \exp(\lambda t - nT)$. The velocity $v = T^{-1}$ of the travelling wave is defined by the characteristic equation

$$L(T,\lambda) = \lambda - (\rho - 1) - \frac{\varepsilon\tau}{\lambda\tau + 1}\left[e^{\lambda T} + e^{-\lambda T}\right] = 0 \tag{3}$$

via the relations $L(\lambda, T) = 0$, $\frac{\partial L}{\partial \lambda}(\lambda, T) = 0$. These results are extended to a model with a random coupling. It is interesting that in a lattice with the small random fluctuations of the coupling constant ε the travelling wave is slowed down. The asymptotic expansion with respect to the variance $\sigma^2 = \sigma_\varepsilon^2$ has the form

$$\lambda_0\bar{T} = \frac{(W_0-1)^2}{W_0} + \sigma^2\frac{W_0(W_0^2+2)}{8(W_0+1)^5} + O(\sigma^4) \tag{4}$$

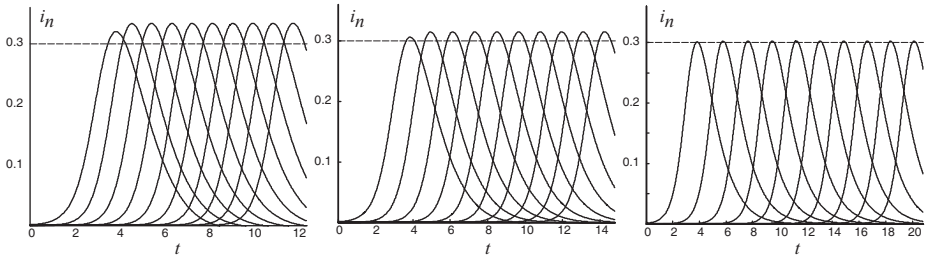
where $W_0 = W_0(e/\bar{\varepsilon})$ is the zeroth branch of the Lambert function ($W(x)$ is a solution to $We^W = x$) and $\bar{\varepsilon}$ is the mean value of the coupling constant. The direct numerical modeling confirms analytical relations in both the deterministic and randomized cases [2, 3].

An important observation is that, for weak coupling, the main part of the travelling wave is well approximated by the limiting SIR solution which describes the epidemic in the limit of the infinitesimally small initial contamination. In this approximation the number of susceptibles and infected in the outbreak are, respectively, $s^{\lim}_{\text{outb}} = \rho^{-1}, i^{\lim}_{\text{outb}} = 1 - \rho^{-1}(\ln\rho + 1)$. Then every solution can be approximated by $i = i^{\lim}(t - t_{\text{outb}}(i_0))$, where the outbreak time t_{outb} can be calculated through the expansion by small i_0 (see [3])

$$t_{\text{outb}} = \int\limits_{s_{\text{outb}}}^{s_0} \frac{\mathrm{d}s}{\rho(s - s^2) + s\ln(s/s0)} \cong -\frac{\ln i_0}{\rho - 1} + \Theta(\rho) + O(i_0),$$

$$\Theta = \int\limits_0^{1-\rho^{-1}} \left[\frac{1}{(1-\xi)\left[\rho\xi + \ln(1-\xi)\right]} - \frac{1}{\xi\,(\rho-1)}\right]\mathrm{d}\xi + \frac{2\ln(1-\rho^{-1})}{\rho-1} \approx \frac{\ln(\rho-1)}{\rho-1}.$$

Numerically results depicted in Figure 2 indicate that the smaller the coupling coefficient ε, the closer the amplitude of the travelling wave to that of the limiting solution.



a b c

Figure 2. Numerical simulation of a lattice (first 10 nodes) for $\rho = 3, i_0 = 0.01$ and $(\rho-1)\tau = 1$: for $\varepsilon = 0.1$ (a), $\varepsilon = 0.05$ (b), $\varepsilon = 0.01$ (c). The dashed line indicates $i^{\lim}_{\text{outb}}$ for $\rho = 3$

2 Stochastic models

2.1 Critical reaction time

Consider a Markov chain on $(S, I), S, I \geqslant 0, S + I = N$ with absorption in $(0, N)$ and the transition rate βSI for transition $(S, I) \to (S - 1, I + 1)$. Let $\beta = \beta'/N$ and $\beta' = 1$ wlog. Define $\tau = \tau(I_1, I_0, N) = \inf\{t : I(t) = I_1\}$. Then with $S_i = N - I_i, i = 1, 2$ and $X = X(\theta, N) = \frac{N}{2}\sqrt{1 - \frac{\theta}{N}} - \frac{N}{2}$

$$\psi(\theta) = \mathbb{E}e^{\theta\tau} = \frac{\Gamma(I_1)\Gamma(S_0 + 1)\Gamma(I_0 + X)\Gamma(S_1 + 1 + X)}{\Gamma(I_0)\Gamma(S_1 + 1)\Gamma(I_1 + X)\Gamma(S_0 + 1 + X)} \tag{5}$$

[5]. Then the scaled process $(S_\Lambda(t), I_\Lambda(t)) = (\Lambda^{-1}S(t), \Lambda^{-1}I(t))$ in a population of size $\lfloor \Lambda N \rfloor$ with the initial conditions $\lfloor \Lambda S_0 \rfloor, \lfloor \Lambda I_0 \rfloor$ and the re-scaled transition rate $\Lambda^{-1}N^{-1}SI$ converges in distribution as $\Lambda \to \infty$ to the solution of ODEs:

$$\frac{d}{d}S = -N^{-1}SI, \quad \frac{d}{dt}I = N^{-1}SI.$$

The solution of (5) with the initial condition I_0 reaches the level I_1 at the moment $\bar{t} = \ln\left(\frac{S_0 I_1}{I_0 S_1}\right)$. Let $S_j = s_j N, I_j = i_j N$, $j = 1, 2$. In the stochastic case we demonstrate that

$$\mathbb{E}\tau = \ln\left(\frac{s_0 i_1}{i_0 s_1}\right) + \frac{1}{2N}\left[\frac{1}{i_0} - \frac{1}{i_1} + \frac{1}{s_0} - \frac{1}{s_1}\right] + O(N^{-2}).$$

Central limit theorem. As $N \to \infty$

$$\sqrt{N}(\tau - \bar{t}) \Rightarrow N(0, \sigma^2)$$

where $\sigma^2 = 2\bar{t} + \frac{1}{i_0} - \frac{1}{i_1} - \frac{1}{s_0} + \frac{1}{s_1}$.

2.2 Two-stage model of the SIR outbreak

We study the standard SIR model in the small initial contamination (SIC) approximation and distinguish two stages of epidemic evolution. Stage 1 is the initial contamination stage when the number of infected is small and the influence of their fluctuations is vital. At this stage the system is randomized, and governed by stochastic equations. Stage 2 is the developed outbreak and the standard SIR model works well as the number of individuals in all components is large. Therefore, we consider a deterministic system with randomized initial conditions linked to the stochastic stage (see Figure 3), On the initial contamination stage SIR model may

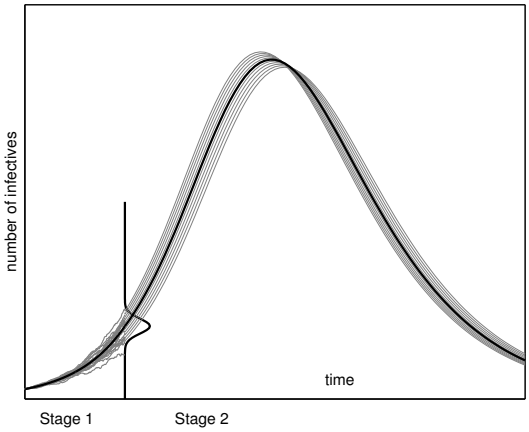


Figure 3. Schematic of proposed model: semi-randomized SIR model with a Small Initial Contagion approximation. The discrete nature of the population and the associated stochastic behaviour is accounted for at Stage 1 when the number of infected is small. This leads to a stochastic initial condition (with PDF depicted between the stages) for the Stage 2 described by deterministic equations

be approximated by an idealised SI model on $\mathbb{Z}_+$ with the transitions $I \to I+1$ and $I \to I-1$ with rates $\beta' I$ and αI and absorption at the origin, the moment generating function (MGF) has the form

$$G(z,t) = \mathbb{E}z^{I(t)} = \Big[\frac{(z-1)e^{\lambda_0 t} - (\rho z - 1)}{\rho(z-1)e^{\lambda_0 t} - (\rho z - 1)}\Big]^{I_0} \tag{6}$$

where $\lambda_0 = \beta' - \alpha, \rho = \beta'/\alpha$ and I_0 is the initial condition [1]. In the stochastic SIR model with transitions $I \to I+1, S \to S-1$ with rate βSI and $I \to I-1, R \to R+1$ with rate αI on the set $S+I+R = N, S, I, R \geqslant 0$ at absorption at points $(0, N, 0)$ and $(0, 0, N)$ the MGF

$$G(z,y,t) = \mathbb{E}\Big[z^{I(t)} y^{S(t)}\Big] \tag{7}$$

satisfies the PDE [4]

$$G_t = \beta z(z-y)G_{zy} - \alpha(z-1)G_z \tag{8}$$

with the initial condition $G(z,y,0) = z^{I_0} y^{N-I_0}$.

2.3 Hydrodynamic limit

In the SI model the scaled random process $I_*(t) = \Lambda^{-1} I(t)$ on $\mathbb{Z}_+$ with the scaled initial condition $\lfloor \Lambda I_0 \rfloor$ converges in distribution as $\Lambda \to \infty$ to a deterministic function $\hat{I}(t)$ satisfying the ODE

$$\frac{d}{dt}\hat{I} = (\beta' - \alpha)\hat{I} \tag{9}$$

with the initial condition $\hat{I}(0) = I_0$.

For the stochastic SIR model the scaled Markov chain $(I_*(t), S_*(t), R_*(t))$

$$I_*(t) = \Lambda^{-1} I(t), S_*(t) = \Lambda^{-1} S(t)$$

in a population of size $\lfloor \Lambda N \rfloor$ defined by the scaling of the transition rate $\beta \to \Lambda^{-1}\beta$ and scaling of the initial conditions

$$I(0) = \lfloor \Lambda N \rfloor, S(0) = \lfloor \Lambda N \rfloor, R(0) = 0$$

converges in distribution as $\Lambda \to \infty$ to the deterministic functions $(\hat{I}, \hat{S})$ described by the ODEs

$$\frac{d}{dt}\hat{S} = -\beta \hat{S}\hat{I}, \frac{d}{dt}\hat{I} = \beta \hat{S}\hat{I} - \alpha \hat{I}, \hat{R} = N - \hat{I} - \hat{S}. \tag{10}$$

with the initial condition $\hat{I}(0) = I_0, \hat{S}(0) = S_0$.

2.4 Contamination of an SI node by external infected

Consider a Markov chain $\{I(t), J(t)\}$ on $\mathbb{Z}_+^2$ with transition rates: $(I, J) \to (I + 1, J)$, the rate $\beta'(I + J)$; $(I, J) \to (I - 1, J)$, the rate αI; $(I, J) \to (I, J - 1)$ the rate $(\alpha + \delta)J$ and $(I, J) \to (I, J + 1)$, the rate $\mu(t)$. Here $I \geqslant 0$ is the number of local infected in the node, $J \geqslant 0$ is the number of infective visitors from other centers. Let $P_{mn}(t) = \mathbb{P}(I(t) = m, J(t) = n)$, introduce the MGF

$$G(t, y, z) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} y^m z^n P_{mn}(t). \tag{11}$$

The function $G(t, y, z)$ satisfies a PDE

$$G_t = (y - 1)(\beta' y - \alpha)G_y + [\beta'(y - 1)z - (\alpha + \delta)(z - 1)]G_z + \mu(t)(z - 1)G \tag{12}$$

with the general solution

$$G = F\left(z, t + \frac{(\alpha+\delta)(z-1)}{\beta'-\alpha}\ln(y-1) + \frac{1+\alpha+\delta-(\beta'+\delta)z}{\beta'-\alpha}\ln(\beta' y-\alpha)\right)\times \\ \times \exp\Big(-\int\limits_C^y \frac{\mathrm{d}y'}{(\beta' y'-1)(y'-1)}\mu(t+\theta)\Big),$$

$$\theta = \frac{1-\alpha-\delta+(\alpha+\delta)z}{\beta'-\alpha}\ln\frac{y'-1}{y-1} - \frac{1+\alpha+\delta-(\beta'+\delta)z}{\beta'-\alpha}\ln\frac{\beta' y'-\alpha}{\beta' y-\alpha}.$$

Here $F(\cdot,\cdot)$ is an arbitrary function and C is an arbitrary constant.

Next, by differentiation of (12) with respect y and z, and substituting $y = z = 1$, we obtain the following equations for $\bar{I}(t) = \mathbb{E}I(t)$ and $\bar{J}(t) = \mathbb{E}J(t)$:

$$\frac{d}{dt}\bar{I}(t) = (\beta'-\alpha)\bar{I}(t) + \beta'\bar{J}(t), \frac{d}{dt}\bar{J}(t) = \mu(t) - (\alpha+\delta)\bar{J}(t).$$

In order to find the travelling wave velocity we make an assumption that $\mu(t) = \varepsilon\bar{I}(t+T)$ where T is the characteristic migration time between two centers. Solving the second equation we obtain a closed equation for $\bar{I}(t)$:

$$\frac{d}{dt}\bar{I}(t) = (\beta'-\alpha)\bar{I}(t) + \beta'\varepsilon\int\limits_{-\infty}^{t}\bar{I}(s+T)e^{(s-t)(\alpha+\delta)}\mathrm{d}s.$$

This equation leads to the characteristic equation of the type (3);

$$L(T,\lambda) = \lambda - (\beta'-\alpha) - \frac{\beta'\varepsilon}{\lambda+\alpha+\delta}e^{\lambda T}.$$

As a result we justified an effective method of computing the travelling wave velocity in a 1D lattice of susceptible/infected/removed (SIR) epidemic centers described in Section 1.

References

1. Daley D.J., Gani J., *Epidemic Modelling*, Cambridge University Press, 1999
2. Sazonov I., Kelbert M., Gravenor M.B., *Travelling waves in a lattice of SIR nodes in approximation of small coupling.* Mathematical Medicine and Biology, **28**(2), 2011, Pp. 165–183.

3. Sazonov I., Kelbert M., Gravenor M.B., *Approximation for epidemic speed in a one-dimensional lattice of SIR models.* Mathematical Modelling of Natural Phenomena, **3**(4), 2008, Pp. 28–47
4. Sazonov I., Kelbert M., Gravenor M.B., *A two-stage model for the SIR outbreak accounting for the discrete nature of epidemic on the contamination stage.* Mathematical Biosciences 2012, (in press). DOI: 10.1016/j.mbs.2011.09.002
5. Kelbert M., Sazonov I., Gravernor M.B., *Critical reaction time during the disease outbreak.* Ecological Complexity, 2012, (in press). DOI: doi:10.1016/j.ecocom.2011.07.002

M. Ya. Kelbert
Swansea University, Singleton Park, Swansea, UK, SA2 8PP; International Institute of Earthquake Prediction and Mathematical Geophysics RAS, Moscow. E-mail: m.kelbert@swansea.ac.uk

I. A. Sazonov
Swansea University, Singleton Park, Swansea, UK, SA2 8PP; Institute of Atmospheric Physics RAS, Moscow. E-mail: i.sazonov@swansea.ac.uk

M. B. Gravenor
Swansea University, Singleton Park, Swansea, UK, SA2 8PP. E-mail: m.b.gravenor@swansea.ac.uk

HUMAN CORNEA MODELING USING ARTIFICIAL COLLAGEN

A. A. Khokhlov, K. P. Lovetskiy, V. I. Bukanina

Key words: human cornea, artificial collagen, spectrophotometric research

AMS Mathematics Subject Classification: 65K10

Abstract. At least 12 million people all over the world are blind because of the damaged or diseased corneas. Till recently keratoplastics was the only way to help these people. This is a surgical procedure where a damaged cornea is replaced by donated corneal tissue. The main problem of this approach is dficiency of donor tissue. More universal method which allows obviating necessity for donors is keratoprosthesis. The main problems of existing plastic keratoprosthesis are high percent of rejection of an artfcial cornea, its keratomalacia and opacity. So with reference to an artificial cornea of an eye collagen is the most perspective material. Merits of collagen keratoprosthesis are: almost absolute biological compatibility; reduction terms of the operations performing and postoperative rehabilitation to a minimum; reproduction biomechanical characteristics of natural cornea; wider patients base. The main problem is that thin membranes made of collagen are optically transparent, but the single-layered samples approached on a thickness to a normal cornea, practically lose this property. This problem can be solved by modeling of a cornea as multilayered structure. Within the limits of this project problems of selection of necessary quantity of lamellas, corners under which collagen lamellas are turned to each other and a number of other parameters. Restoration of complex refraction index was carried out using MorphoVision software, created in laboratory Optics of nanostructures, PFUR. Obtained refractive indices were used in the designing of multilayer model of human cornea, having sufficient transparency.

1 Introduction

Cornea is optically transparent part of the outer shell of the eyeball, which performs approximately three-quarters of the work on refraction and focusing a beam of light on the retina. Various damages caused by mechanical, chemical or microbiological effects, with high probability lead to total or partial loss of vision. Small erosions occurring on the surface of the cornea, close in several days, but in case of damaging deeper layers, the site of the defect appears translucent haze, which can lead to a significant reduction in vision, especially if the injury occurred in the central zone

of the cornea. At least 10 million people all over the world are blind because of the damaged or diseased corneas.

Till recently keratoplasty was the only way to help these people. This is a surgical procedure where a damaged cornea is replaced by donated corneal tissue. The main problem of this approach is deficiency of donor tissue. Each year over 100000 people are waiting for cornea transplants. More universal method which allows obviating necessity for donors is keratoprosthesis. The most widely used keratoprosthesis such as AlphaCor (Lions Eye Institute, Argus Biomedical Pty Ltd), Boston K-Pro, have standard design: all models include plastic optical part and differ from each other by fasting to the cornea. The main problems of these keratoprosthesis are high percent of rejection of an artificial cornea, its keratomalacia and opacity.

Macroscopicaly natural cornea consists of five layers:

— Corneal epithelium: a thin epithelial multicellular tissue layer of fast-growing and easily regenerated cells, kept moist with tears.
— Bowman's layer: a tough layer that protects the corneal stroma.
— Corneal stroma (also substantia propria): a thick, transparent middle layer, consisting of regularly arranged collagen along with sparsely distributed interconnected keratocytes, which are the cells for general repair and maintenance. They are parallel and are superimposed like book pages The corneal stroma consists of approximately 200 layers of mainly type I collagen fibrils. Up to 90% of the corneal thickness is composed of stroma.
— Descemet's membrane: a thin acellular layer that serves as the modified basement membrane of the corneal endothelium.
— Corneal endothelium: a simple squamous or low cuboidal monolayer. These cells are responsible for regulating fluid and solute transport between the aqueous and corneal stromal compartments.

The thickness of each layer of stroma is 1500–2000 nm. These layers are arranged at 90 degrees relative to each other and parallel to the surface of the cornea.

Collagen is one of the most promising biomaterials in medical practice, it has a wide range of applications. And since it is a native material for the human cornea, one of the most important applications of collagen is its use to develop an artificial cornea, so-called keratoprosthesis.

Thus, it is necessary to know the optical properties of thin films of collagen, such as the dependence of reflection and transmission coefficients from medium acidity, properties of light scattering on collagen structures, etc. Spectrophotometric research and multiple solving of inverse problem allow determining these parameters.

The obtained data is used in the mathematical synthesis of collagen structures with predetermined characteristics such as artificial cornea. Merits of collagen keratoprosthesis: almost absolute biological compatibility; reducing terms of the operations performing and postoperative rehabilitation to a minimum; reproducing biomechanical characteristics of natural cornea; wider patient base.

2 Spectrophotometric research of collagen

Specialists of Fibralign Corporation, CA, USA developed the technology of synthesis of collagen type II and the formation of structures with predetermined characteristics, including thin films, which can be used in the problem of creating an artificial cornea of the human eye. For the use of artificially produced collagen films in solving this problem they need to be comprehensive researched.

Research of available samples were made using RPFU Optics Research Laboratory equipment, including spectrophotometer Lambda-950, manufactured by PerkinElmer, USA, and profilometer DekTak 150, manufactured by Veeco Instruments Inc, USA.

As materials for the research were used samples of different thickness and samples consisting of different number of layers: one, two, four and six, as well as samples with different types of fastening: on the glass substrate and the special cassettes. Optical properties of artificial collagen were priori unknown. Measurements of reflection and transmission coefficients were conducted with the help of spectrophotometer Lambda 950. We used the visible spectrum — from 400 to 800 nm as the range of wavelengths. The sample was moistened with a solution with a neutral pH = 7,4. In order to determine optical properties of collagen, the measurements were performed for single layer sample, fixed in the cassette. This approach has several advantages: it is easier to use the results for solving the inverse problem of recovering, and substance by which layers in multilayer samples are fastened may introduce error into the results of spectrophotometric measurements.

The thickness of the collagen film on the substrate was also measured using profilometer DekTak150. Collagen layer covered the substrate not fully, so it was possible to measure the difference in height of the sample and the substrate, which is the film thickness.

To determine the properties of collagen, namely whether it is optically anisotropic or isotropic material, were carried out a series of polarization scans: at a particular wavelength at a certain angle of incidence of light on the sample, was measured the reflection, transmission and absorption coefficients of the thin film. After the analysis of measurement data it was concluded, that there is the dependence of the properties of light propagation in thin film from the direction of

propagation and the polarization of the incident wave, thus collagen is anisotropic material with one optical axis. This conclusion was confirmed by several series of measurements of coefficients of reflection and transmission in the visible light spectrum.

Another series of measurements was aimed at identifying dependence of the optical transparency of the collagen on the acidity of the environment. This problem arose due to the fact that the intraocular fluid of the eye under normal conditions has a definite almost neutral Ph = 7,4, but during the operation on the cornea, it is necessary to use drugs that affect the level of acidity. For measurements were used four-layer and two-layer samples, the incident light was unpolarized. Each sample was placed for a while in solutions with certain pH, and then were carried out it's measurements on the transmission. The measurement results show that the optical transparency of the collagen strongly depends on the acidity of the medium in the blue and violet range of the optical spectrum, in the red range dependence is weaker. It wasn't revealed any dependence of the transmission of acidity on the entire range, but it wasn't a primary goal of this research.

As part of this experiment was also carried out a series of measurements a two-layer collagen film, soaked in a solution with Ph 5,03. The interval between the measurements was 3 minutes, and measurements were carried out until dry. According to the results of measurements, the transmittance of the film increases as it dries, which is an atypical manifestation of the properties of collagen. Visually, it is obvious that the dry film is not optically transparent and moisture of the sample is one of the main conditions of transparency. Nature of the curves remains virtually unchanged as the drying, which suggests that the potential scattering on the droplets of the solution is not happening. The results can be explained from the standpoint that the concentration of salts, which are responsible for acidity, gradually increases due to evaporation of water. This increase of concentration may cause the increase in transmittance of collagen film. Also it was measured the transmission coefficient of a two-layer collagen film, soaked in saline solution with a neutral pH. Measurements were carried out at different angles: 0°, 30°, 50°, 60°. According to the results, the transmission decreases with increasing angle of incidence, as it has been calculated theoretically. After a series of measurements with different acidity — from Ph 9,18 to Ph 4,01 — was re-obtained the transmittance of the film, soaked in a solution with Ph 9,18. According to the results of measurements the repeated and prolonged exposure to fluids with different indices of acidity did not affect the structure of the sample, since the nature of the curves remained almost unchanged. The difference in transparency in the blue and green parts of the spectrum can be explained by the error in the degree of moisture of collagen, as well as by deformation of the film, embodied in the cassette.

The results obtained during the spectrophotometric investigations of collagen, are the basis for restoring the properties of thin films of collagen, in particular the refractive index, as well as for select a mathematical model underlying the numerical experiments on the synthesis of keratoprosthesis.

3 Restoration of collagen optical properties

To solve the problem of mathematical synthesis of collagen keratoprothesis it is necessary to determine the optical properties of synthetic collagen. The problem of restoration of the dielectric tensor and thickness of thin anisotropic films using measurements of reflected and refracted waves is formulated as follows. Using sets of spectrophotometric data about transmission $\tilde{T}(\lambda)$ and reflection $\tilde{R}(\lambda)$ it is necessary to calculate the parameters of the dielectric tensor of the material $\hat{\varepsilon}(\lambda)$ in the range of wavelengths $[\lambda_{start}; \lambda_{end}]$. This is the mathematically ill-posed problem.

Using sets of measured energy coefficients of reflection and transmission and methods of numerical optimization we can recover the dielectric tensor and thickness for each wavelength separately and obtain the indexes of refraction and absorption, which together constitute the complex refractive index: $\tilde{n} = n + ik$. Complex refractive index and dielectric tensor linked as following: $\tilde{n}(\omega) = \sqrt{\hat{\varepsilon}(\omega)}$. However, practice shows that the results of these calculations are unstable — the dielectric tensor is restored in the form of a non-smooth function. Therefore, there is a need to use any method of solving problems for all investigated range of wavelengths, typically an optical range — 400–800 nm. Reconstruction algorithm in this case is much more complicated, but allows using of a priori information about the decision and apply the methods of regularization. As an a priori information there were used the dispersion of the Kramers–Kronig — integral relation between the real and imaginary parts of analytic complex functions [1].

The propagation of polarized light in multilayer anisotropic structure can be described by a matrix equation for the vector of tangential components of reflected and transmitted fields χ_R and χ_T with initial conditions — information about tangential components of incident field χ_I.

The proposed algorithm for solving the inverse problem lies in the approximation of the imaginary part of dielectric tensor — sum of Gaussian functions (or, otherwise, the approximation by radial basis functions) and using the Kramers–Kronig relations to calculate the real part of this function. The objective function, that expresses the sum of squared differences between measured and calculated energy coefficients of reflection and transmission, is minimized by the Nelder — Mead algorithm.

The described algorithm is implemented in software "MorphoVision", developed in the lab "Optics of nanostructures" in People's Friendship University. The initial parameters of the software takes the data sets, each set must contain a sample thickness, the angle of incidence on the sample, the polarization of light (TE or TM), text files with energy coefficients of reflection and transmission.

The calculation procedure using the proposed algorithm is different for different types of materials (isotropic material, a uniaxial anisotropic material, biaxial anisotropic material). In the case of uniaxial anisotropic material, which is the collagen (as identified by polarizing scanning) is sufficient to hold two-cycle recovery. To obtain the refractive and absorption indexes along one direction it is necessary to use data on the energy reflection and transmission coefficients obtained from the measurements when the plane of light incidence was parallel to the optical axis. To obtain the refractive and absorption indexes along the other direction we should use spectrophotometric data obtained when the plane of light incidence on the sample was perpendicular to the optical axis. We carried out a series of calculations and obtained the coefficients of refraction and absorption of artificial collagen.

4 Synthesis algorithm

After the refractive indices of synthetic collagen were recovered, we can solve the problem of synthesis of multilayer optical structure of collagen, which has geometric and optical characteristics similar to the characteristics of human cornea. Problem of mathematical synthesis of optical systems with specified characteristics are a large class of inverse mathematical problems, and they are usually ill-posed. There are different approaches to solve them, but the most effective of these is the method of Tikhonov regularization [2].

As initial data we have recovered optical properties of collagen, maximum and minimum thickness of the simulated system (the thickness of a human cornea is about 0.5–0.7 mm), maximum and minimum thickness of one layer, the desired spectral characteristics of reflection and transmission of the simulated system (the transmission not less than 60% in the optical range, which corresponds to transmission of adult human cornea). Our task is to determine the number of layers of the system, the angles at which they are located relative to each other and the thickness of the layers of the system.

Let $\widehat{T}(\lambda)$ — defined on the wavelength range $[\lambda_1, \lambda_2]$ — be energy coefficient of transmission. We assume $\widehat{T}(\lambda)$ is a general function on $L_2[\lambda_1, \lambda_2]$. The direct problem of modeling of propagation of light in multilayer optical system can be expressed as:

$$M(h, \hat{\varepsilon}, \theta_j)\vec{A} = \vec{D};\ \vec{A} \Rightarrow \vec{R}, \vec{T}. \tag{1}$$

Vector $\vec{D}$ includes amplitudes of reflected and transmitted waves. Energy coefficients of transmission and reflection can be calculated using these amplitudes. Required properties of simulated optical system expressed as:

$$M(d^0, \hat{\varepsilon}, \theta^0)\vec{A}^0 = \vec{D};\ \vec{A}^0 \Rightarrow \vec{R}^0, \vec{T}^0. \tag{2}$$

Thus, to solve the problem it is necessary to minimize the functional:

$$F(\vec{T} - \vec{T}^0, \vec{R} - \vec{R}^0) \to \min, \tag{3}$$

where $\bar{T}_0$ and $\bar{R}_0$ are required transmission and reflection of simulated optical system, and $\bar{T}$ and $\bar{R}$ — calculated transmission and reflection. Problem is solved by Tikhonov regularization. Let

$$\delta_N = \inf \left\| A(x, \lambda) - A^0(\lambda) \right\|_{L_2} \tag{4}$$

be the maximum achievable accuracy of approximation on optical system consists of N layers. Maximum achievable accuracy satisfy the estimates $\delta_1 \geq \delta_2 \geq ... \geq \delta_N$ and together they are bounded below by $\delta = \lim\limits_{N\to\infty} \delta_N$, which we call the maximum possible accuracy.

It is necessary to minimize the residual (2.1), approximating the desired response with a given accuracy. Introduce an error: $S(n, N)$, where n — number of layers in the system, and N — the number of iterations in the computation. If the selected number of layers at step N is achieved the specified accuracy, the algorithm stops, the result has been obtained. If $S(n)$ is more than a specified accuracy, then move on to the system of $(n + 1)$ layers, then — to a system of $(n + 2)$ layers, and so on. If the number of layers exceeds the maximum (or the thickness of the system exceeds the maximum), then should be pointed out the impossibility of achieving the required characteristics to these terms.

5 Conclusions

Using created software we have performed series of calculations and synthesized a variety of structures with desired characteristics [3]. To select the optimal structure, which could be used as a keratoprosthesis, further joint work of many professionals is needed — from engineers involved in the production of stacks of synthetic collagen and ending with scientists in the field of eye medicine.

References

1. L.D. Landau, E.M. Lifshitz, L.P. Pitaevskii (1984). Electrodynamics of Continuous Media. Vol. 8 (2nd ed.). Butterworth–Heinemann. ISBN 978-0-750-62634-7.
2. V. B. Glasko, A. N. Tikhonov, A. V. Tikhonravov. On the synthesis of multilayer coatings. Computational Mathematics and Mathematical Physics. Vol. 14, No1, 1974.
3. V.I. Bukanina, K.P. Lovetskiy, A.A. Khokhlov. Human Cornea Modeling using artificial collagen. Bulletin of Peoples' Friendship University of Russia. Series "Mathematics. Informatiom Sciences. Physics". No3, 2011. — P. 82–85.

A. A. Khokhlov
Laboratory Optics of nanostructures, Russian Peoples' Friendship University, Russia, Moscow, Ordjonikidze st. 3, +7 (926) 231-31-00, alex@ahohlov.name

K. P. Lovetskiy
Laboratory Optics of nanostructures, Russian Peoples' Friendship University, Russia, Moscow, Ordjonikidze st. 3, +7 (916) 157-72-61, lovetskiy@gmail.com

V. I. Bukanina
Laboratory Optics of nanostructures, Russian Peoples' Friendship University, Russia, Moscow, Ordjonikidze st. 3, +7 (985) 459-96-81, vibukanina@gmail.com

MATHEMATICAL MODEL OF LOCAL REGULATION OF BLOOD FLOW AND OXYGEN TRANSPORT

A. V. Kopyltsov

Key words: oxygen transport, blood flow, regulation, modelling

AMS Mathematics Subject Classification: 11B39

Abstract. The mathematical model of a regulation of a blood flow and oxygen transport by products of metabolism is developed. The model takes into account construction and haemodynamic characteristics of a vascular network. Approximate formula of dependence T (an interval of time of transition of a local system of blood flow and oxygen transport from one steady state in another) from oxygen consumption rate by tissue W and a pressure differences on the ends of a vascular network $\triangle P$ is determined. It is shown that at transition from easy physical activity to intensive and back, an interval of time spent for transients, in the second case is more at 2.6-3.4 times, than in the first. At increase W in 2 times a speed of a blood flow increases at 2.2-2.8 times that corresponds to experimental data (in 2-3 times).

1 Introduction

In a lot of papers [1–12] a regulation of a blood flow and oxygen transport is considered. The particular interest represent the local regulation of a blood flow and oxygen transport by metabolism products [5, 10, 11]. It is important to reveal mechanisms of dependence of oxygen consumption rate by tissue from physical activity value of an organism [1, 9–11, 13]. For the decision of this problem the mathematical model has been constructed [5]. With the help of this model the calculations of time intervals of transition of a local system of blood flow and oxygen transport from one steady state in another at various physical activity of organism are carried out.

2 Equations Describing a Transport of Oxygen and Metabolism Product

At modeling of oxygen transport from RBC (red blood cell, erythrocyte) in surrounding tissue it is supposed that RBC moves rectilinearly, it velocity is constant,

and in RBC is an equilibrium reaction of hemoglobin with oxygen, and in a tissue is an equilibrium reaction of myoglobin with oxygen.

Oxygen transport is carried out from RBC through a layer (plasma, endothelium, and interstitial space) in a tissue where oxygen is absorbed. In a tissue are allocated, during biochemical reactions, metabolism products which by diffusion and with a blood flow are deduced from a tissue. Oxygen transport is described by the following system of the differential equations [5]. In RBC:

$$\frac{\partial P}{\partial t} + \mathbf{V}\nabla P = D_{O2E}\nabla^2 P + \frac{\rho(P,SS)}{\alpha_E} \tag{1}$$

$$\frac{\partial SS}{\partial t} + \mathbf{V}\nabla SS = D_{Hb}\nabla^2 SS - \frac{\rho(P,SS)}{C_{HbT}} \tag{2}$$

$$C_{HbT} = C_{Hb} + C_{HbO2},\ SS = \frac{C_{HbO2}}{C_{HbT}}$$

$$\rho(P,SS) = k^+ C_{HbO2} - k^- C_{Hb} C_{O2} = k^+ C_{HbT} SS - k^- C_{HbT}(1-SS)\alpha_E P$$

In the layer (plasma, capillary endothelium, interstitial space):

$$\frac{\partial P}{\partial t} + \mathbf{V}\nabla P = D_{O2C}\nabla^2 P \tag{3}$$

In the tissue:

$$\frac{\partial P}{\partial t} = D_{O2M}\nabla^2 P + \frac{\sigma(P,S)}{\alpha_M} - \frac{W}{\alpha_M} \tag{4}$$

$$\frac{\partial S}{\partial t} = D_{Mb}\nabla^2 S - \frac{\sigma(P,S)}{C_{MbT}} \tag{5}$$

$$C_{MbT} = C_{Mb} + C_{MbO2},\ S = \frac{C_{MbO2}}{C_{MbT}}$$

$$\sigma(P,S) = k1^+ C_{MbO2} - k1^- C_{Mb} C_{O2} = k1^+ C_{MbT} S - k1^- C_{MbT}(1-S)\alpha_M P$$

Oxygen is absorbed by tissue, metabolism products are allocated and are transported by diffusion in a tissue and are transferred with a blood flow to a venous part of a vascular network. Thus, formation and transport of metabolism products in a tissue, in a layer (interstitial space and capillary endothelium), and in blood vessels is described by the following system of the differential equations. In the tissue:

$$\frac{\partial C}{\partial t} = D_{MpM}\nabla^2 C + \beta W^n + \frac{\gamma\ \epsilon}{P+\epsilon} \tag{6}$$

In the layer (interstitial space and capillary endothelium):

$$\frac{\partial C}{\partial t} = D_{MpC}\nabla^2 C \tag{7}$$

In the capillary:

$$\frac{\partial C}{\partial t} + \mathbf{V}\nabla C = D_{MpB}\nabla^2 C \tag{8}$$

where P is oxygen partial pressure, V is local velocity vector of fluid movement, D_{O2E}, D_{O2C}, D_{O2M} are diffusion coefficient of oxygen in RBC, in the layer (consisting of a plasma sleeve, the capillary endothelium, and the interstitial space), and in the tissue (muscle fiber), respectively, D_{Hb}, D_{Mb} are diffusion coefficients of hemoglobin and myoglobin, α_E, α_M are solubility coefficients for oxygen inside RBC and the muscle fiber, C_{HbO2} is oxyhemoglobin concentration in RBC, C_{Hb} is hemoglobin concentration, C_{O2} is oxygen concentration, C_{MbO2} is oxymyoglobin concentration in the muscle fiber, C_{Mb} is myoglobin concentration, W is oxygen consumption rate by muscle tissue, k^-, k^+, $k1^-$, $k1^+$ are constants of speeds of biochemical reactions, t is time, C is concentration of metabolism products, D_{MpM}, D_{MpC}, D_{MpB} are diffusion coefficients of metabolism products in a tissue, in a layer (interstitial space and capillary endothelium), and a capillary, β, γ, ϵ, n are coefficients.

Thus, the equations (1) - (8) describe transport of oxygen and metabolism products in RBC, in plasma, in a capillary endothelium, in an interstitial space and in a tissue.

3 Regulation of Oxygen Transport and Blood Flow

The mathematical model of regulation of blood flow and oxygen transport in a tissue take into account a structure of the vascular network including arterioles (A2, A3, A4), venues (V2, V3, V4), and capillaries (C) between arterioles A4 and venues V4 [5, 9]. Blood flows through arterioles (A2, A3, A4), capillaries (C), and venues (V2, V3, V4). Total quantity of arterioles and venues in branchings 2, 3, and 4 is designated n2, n3, and n4 accordingly. The quantity of capillaries between arteriole A4 and venue V4 is designated n5 (Fig. 1).

The blood flow in vascular system is carried out at the expense of a pressure difference on the ends of a network and described by the law of Poiseuille in arterioles and venues, and in capillaries is used the generalized law of Poiseuille which take into account pressure differences on RBCs and plasma columns between them.

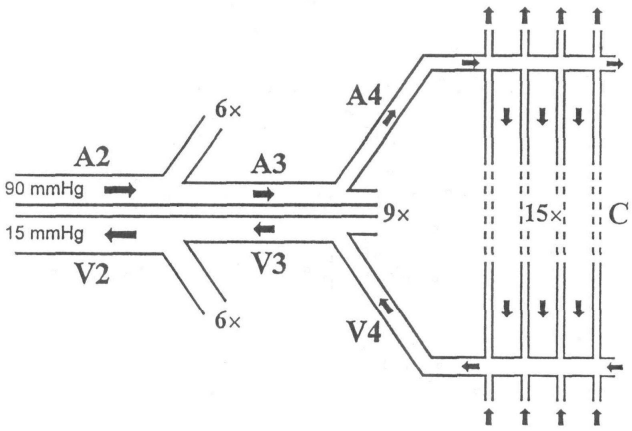


Figure 1. Model of a vascular network. $A2$, $A3$, $A4$ are the arterioles, $V2$, $V3$, $V4$ are the venues, C are the capillaries connecting arterioles $A4$ and venues $V4$ of the fourth order.

According to [8] the pressure gradient on the ends of vessel is equal to

$$\triangle P = Q/G \tag{9}$$

where Q is a volume flow of plasma and RBCs, G is conductance of a vessel and for arterioles and venues

$$G = \Big(\frac{128 mu_B\ L}{\pi d^4}\Big)^{-1}$$

and for capillaries

$$G = \Big(\frac{128 mu_B\ L}{\pi d^4} + \frac{16 mu_P}{\pi d^3} \Sigma_j \triangle PP_j\Big)^{-1}$$

where μ_B and μ_P are viscosity of blood and blood plasma, L and d are lengths and diameters of vessels, $\triangle PP_j$ is dimensionless additional pressure difference on j-th RBC.

At receipt of RBCs incite capillaries network oxygen is allocated from RBCs, it is transported by diffusion in a tissue and is absorbed. During biochemical reactions in tissue there is an allocation of metabolism products which are taken out by diffusion and with a blood flow in a venous part of vascular system. One part of metabolism products arrives with a blood flow in larger veins, and another diffuse in arterioles. The flow of metabolism products from venues in arterioles (on unit of

length of a vessel) is equal to [3]

$$F = K(C_v - C_a)\alpha \tag{10}$$

where K is permeability of a tissue for metabolism products, C_v and C_a are concentration of metabolism products in venue and arteriole, and, according to [3]

$$\alpha = 2\pi ln(\sigma)$$

$$\sigma = \frac{R_v}{2R_a}(cd - 1 - ((c^2 - 1)(d^2 - 1))^{0.5})$$

$$c = 1 + \frac{\delta}{R_v}\ d = 1 + \frac{\delta + 2R_a}{R_v}$$

where δ is distance between arteriole and venue, R_a and R_v are radiuses of arteriole and venue.

The metabolism products receipts from venue in arteriole, influences on smooth muscles of arteriole, and the radius of arteriole R_a is changed as follows

$$R_a = R_{max} - (R_{max} - R_{min})e^{-aC_a} \tag{11}$$

where R_{max} and R_{min} are the maximum and the minimum radius of arteriole, C_a is concentration of a metabolism product in arteriole, a is a constant.

Change of arteriole radius leads to change of blood flow speed in vessels, speed of oxygen delivery in a tissue, to change of partial pressure of oxygen in a tissue, speed of allocation of metabolism products in a tissue, to change of their concentration in venous and arterial vessels and, hence, to change of arteriole radius. Thus, we have the system of the equations (1) - (11). This system of the differential equations (1) - (11) are solved at initial conditions: $P = P0, SS = SS0, S = S0, C = C0$, where P0, SS0, S0, C0 are known functions from z and r.

As boundary conditions we use conditions of symmetry and a condition of continuity of flows of substances on borders: RBC - plasma, plasma - endothelium, endothelium – interstitial space, interstitial space- tissue. The radial component of a flow of all substances on a capillary axis is absent (from symmetry reasons).

The system of the equations is solved by a method of final differences [14]. At the decision of the system of the equations the grid step got out depending on the sizes of a capillary, RBC, and the tissue cylinder, on r from 0.01 to 0.1 μm, on z from 0.01 to 0.1 μm , on t from 0.001 to 0.1 seconds Thus, solving the system of the equations, we have distribution of oxygen and metabolism products in the blood vessel and the surrounding tissue. At study of a blood flow and oxygen transport to tissue the greatest interest is represented by steady state of an organism, because

the most part of life the organism is in a steady state (during a dream, on work, at walking, etc.). Therefore the following section is devoted consideration of steady state of a blood flow and oxygen transport in tissue.

4 Steady-States of Blood Flow and Oxygen Transport

At transition of the organism from one steady state to another (for example, from a state of easy physical activity to intensive or on the contrary) all biochemical and other processes pass from one steady state in another. Because the live organism is not ideal system on these transients some time is required. And, from the experimental data received on human organism and animals it is known that, this time interval is less, then the organism is more adapted physically for environment changes (for example, sportsman). Therefore, it is important to estimate time interval spent for the transients.

During numerical experiments on the computer the time interval necessary for stabilization of processes at transition from entry conditions (absence of oxygen and metabolism products in a tissue) to steady states has been calculated. We have been received appraximate formula for T from $\triangle P$ and W

$$T = a\triangle P\ W - b\ W - c\ \triangle P + d \tag{12}$$

where a=0.0016, b=0.248, c=0.037, d=6.085.

Thus, we have dependences (12) for transition from entry conditions (absence of oxygen and metabolism products in a tissue) to a stationary blood flow and transport of metabolism products in a tissue. However, in a human organism there are other entry conditions. In particular, oxygen consumption in tissue is distinct from zero. Therefore calculations on the computer which have allowed to estimate time intervals of stabilization of processes at transition from easy physical activity (oxygen consumption rate by tissue $W = 10ml/100g/min$) to intensive ($W = 20ml/100g/min$) ($T1(min)$) and back ($T2(min)$) at various values of a pressure drop $\triangle P(mmHg)$ on the ends of a vascular network have been carried out. So, we have $T1 = 0.7min$, $T2 = 2.4min$, and $T2/T1 = 3.4$ at $\triangle P = 75mmHg$, we have $T1 = 0.6min$, $T2 = 2.0min$, and $T2/T1 = 3.3$ at $\triangle P = 90mmHg$, and we have $T1 = 0.6min$, $T2 = 1.7min$, and $T2/T1 = 2.8$ at $\triangle P = 105mmHg$. These results show that time intervals of a stabilization of transients from the entry conditions corresponding to easy physical activity ($W = 10ml/100g/min$), to intensive ($W = 20ml/100g/min$) has value $0.6 - 0.7min$ at an arterio-venous difference of pressure $75-105mmHg$. It is needs 1.7-2.4 minutes for transition from intensive to easy physical activity. These results were compared to experimentally

observable time answers in the microvascular network which make about 2-4 mines [1, 4, 9, 10, 12, 13]. Comparison of the time intervals of transition from easy to intensive physical activity and back shows that in the latter case the time intervals of transition is more at 2.8-3.4 times.

Thus, the model gives good approximation of the time intervals of stabilization of processes in a tissue depending on oxygen consumption rate by tissue and pressure differences on the ends of a vascular network.

5 Conclusion

Mathematical model of local regulation of blood flow and oxygen transport by the metabolism products is developed. Model considers a structure of a vascular network, diameters and lengths of vessels, viscosity of blood and plasma, a pressure difference on the ends of a vascular network, a layer between RBC and a tissue (plasma, capillary endothelium, interstitial space), a formation of a metabolism products in tissue, a transport of oxygen from RBCs in a tissue and a metabolism products from a tissue in vascular network, influence of a vasoactive metabolism products on arteriolar muscles and, hence, on the arteriolar diameters, speed of a blood flow, and oxygen transport in a tissue. On the basis of the constructed model the problem about steady state of a blood flow and oxygen transport is considered. The solution of the problem allows numerically estimating speed of a blood flow, arteriolar diameters, and pressure differences in vessels at various values of oxygen consumption rate by tissue which depends on physical activity value on an organism. Approximate formula of the dependence of time interval of transition of oxygen transport system from one steady state in another depending on oxygen consumption rate by tissue and a pressure differences on the ends of a vascular network is definite. Significance of oxygen consumption rate by tissue (W) is determined. It is shown that at transition from easy physical activity ($W = 10 ml O_2/100 g/min$) to intensive ($W = 20 ml O_2/100 g/min$) and back, time spent for transients, in the second case is more in 2.6-3.4 times, than in the first. At increase W in 2 times the speed of a blood flow increases at 2.2-2.8 times that corresponds to experimental data (in 2-3 times).

References

1. Y. C. Fung. *Biodinamics (Circulation)*. Springer-Verlag, 1984. 404 pages.
2. K. Groebe. *An easy-to-use model for oxygen supply to red muscle*. Biophys. J., 1995. Pp. 175-185.

3. C. R. Honig et. al. *Arteriovenous oxygen diffusion shunt is negligible in resting and working gracilis muscles*. Am. J. Physiol. **261**, (Heart Circ. Physiol. 30), 1991. Pp. 2031-2043.
4. J. J. Kisljakov, A. V. Kopyltsov, V. S. Kovalchuk. *Complex system of an estimation of physical working capacity of the person. In: Medical and biologic problems of ecology of Northwest region.* Leningrad: Nauka, 1990. Pp.81-84.
5. A. V. Kopyltsov, K. Groebe. *Mathematical modeling of local regulation of blood flow by venulo-arterioles diffusion of vasoactive metabolites. In: Proceedings of the international congress of International Society of Oxygen Transport into Tissue.* Pittsburgh (USA), 1995. Pp.171-177.
6. I. P. Pavlov. *Lecture on blood circulation physiology*. Moscow: The informative book Plus, 2002. 220 pages.
7. S. A. Regirer. *Lectures on the biological mechanics*. Moscow: Moscow State University, 1980. 180 pages.
8. G. W. Schmid– Schonbein et. al. *Cell distribution on capillary networks.* Microvasc. Res. **19**, 1, 1980. Pp.18-44.
9. B. I. Tkachenko. *Blood circulation physiology: Physiology of vascular system*. Leningrad: Nauka, 1984. 652 pages.
10. B. I. Tkachenko. *Blood circulation physiology: Regulation of blood circulation*. Leningrad: Nauka, 1986. 640 pages.
11. B. I. Tkachenko. *A vascular tonus and its local regulation*. Saint-Petrsburg: Nauka, 1992. 628 pages.
12. J. A. Vlasov, C. M. Smirnov. *From the molecule of hemoglobin to the microcirculation system*. Novosibirsk: Nauka, 1993. 288 pages.
13. J. J. Kisljakov, A. V. Kopyltsov, V. S. Kovalchuk. *An estimation of power maintenance of a human body on experimental data and on mathematical model. In: Functional reserves and adaptation.* Leningrad: Nauka, 1990. Pp.65-68.
14. D. U. Panov. *Handbook of the numerical solution of the differential equations in partial derivatives*. Moscow - Leningrad: GITTL, 1951. 154 pages.

A. V. Kopyltsov
The Herzen State Pedagogical University of Russia, Russia, 191186, St. Petersburg, Moika Emb., 48, 8-921-4019427, kopyl2001@mail.ru

MODELING OF DYNAMIC PROBLEMS IN BIOMECHANICS USING HPC CLUSTERS

I. B. Petrov, A. V. Vasyukov, D. V. Chernikov, Y. V. Bolotskikh

Key words: method of characteristics, biomechanics, cataract extraction, brain injury

AMS Mathematics Subject Classification: 11B39

Abstract. This paper is devoted to solving of dynamic problems in biomechanics that require detailed study of fast processes. Numerical method of characteristics is used to model the temporal development of the processes with high accuracy.

1 Introduction

This paper is devoted to dynamic biomechanical problems concerning traumatologic and surgical processes. These problems specific feature is that the occurring processes often have a wave nature with very small characteristic times.

Unfortunately most papers known to the authors in this area of research give insufficient attention to the temporal development of the processes. It is the final state of the system that comes under close examination in mathematical modeling. The tools employed usually include a variety of well accepted finite elements methods that provide high accuracy for static problems (see, for example [1–3]).

This approach could be improved. Fast mechanical processes in biological tissues, such as shock wave propagation and interaction, are commonly described by systems of hyperbolic equations as shown below. Grid-characteristic numerical methods and difference schemes of the types devised by Godunov, Kogan, Fedorenko can be used for studying the development of fast processes with small characteristic times. These methods and schemes were initially designed for solving systems of hyperbolic equations and take into account specific features of the equations to achieve high accuracy for the solution of problems of wave nature including the wavefront interactions.

Application of grid-characteristic numerical methods to biomechanical problems will be demonstrated on the example of modeling craniocerebral injury and laser cataract extraction. Accurate consideration of shock wave interactions allows to predict areas of maximum damage which appear as a result of interactions of different waves. This process has an intrinsically dynamic nature and its examination by any other means is highly hampered.

2 Mathematical models and methods used

2.1 Mechanical models

Mechanical properties of biological medium under relatively small strain which is considered as a continuous medium under the action of shock loads can be described using common equations for deformable rigid body. This approach should be used carefully, because biological tissues modelling often requires taking into account fluids, cavities and rigid bodies with very specific rheology. However, for many problems it is reasonable and allows to achieve good results.

For this research we used the equations system of the linear elasticity theory [4]:

$$\rho \cdot \dot{v_i} = \bigtriangledown_j \cdot \sigma_{ij}, \tag{2.1}$$

$$\sigma_{ij} = q_{ijkl} \cdot e_{kl} + F_{ij}. \tag{2.2}$$

Here (2.1) are the equations on motion and (2.2) are the rheological relations. In these equations ρ is the medium density, v_i are the displacement velocity components, σ_{ij} and e_{ij} are the components of the stress tensors and the strain velocities, $\bigtriangledown_j$ is the covariance derivative with respect to the jth coordinate and F_{ij} is the right-hand side.

The tensor q_{ijkl} determines the rheology of the medium. In case of a linearly elastic body its components are expressed in terms of two independent Lame constants λ and μ:

$$q_{ijkl} = \lambda \cdot \delta_{ij} \cdot \delta_{kl} + \mu(\delta_{ik} \cdot \delta_{jl} + \delta_{il} \cdot \delta_{jk}).$$

We also used Maxwell's viscoelastic body model.

The density is determined from the equation of state $\rho = \rho_0 e^{(p/K)}$, where $p = -\frac{1}{3}\sum \sigma_{kk}$ is a pressure and $K = \lambda + \frac{2}{3}\mu$ is a uniform compression coefficient.

These equations are suitable for modelling wave processes in continuous medium under relatively small shock loads. Wave processes have small characteristic times, so deformations are small during the numerical experiment. These assumptions allow us to use the model of linear elastic body.

Equations of motion and rheological relations stated above can be written in the matrix form:

$$\frac{\partial}{\partial t}u + A_1\frac{\partial}{\partial x_1}u + A_2\frac{\partial}{\partial x_2}u + A_3\frac{\partial}{\partial x_3}u = f. \tag{2.3}$$

Here $u = (v_1, v_2, v_3, \sigma_{11}, \sigma_{12}, \sigma_{13}, \sigma_{22}, \sigma_{23}, \sigma_{33})^T$ is the vector of variables, f is the right-hand side of the same dimension, A_i are the matrices of ninth order (their

explicit form can be found in [7]), x_i are independent three-dimensional variables and t is the time.

If the matrices A_i have nine real eigenvalues, then the system is called hyperbolic and its solutions correspond to the processes that are usually called wave processes. That is, this equation describes the propagation of perturbations along the characteristic cones in the space (x_1, x_2, x_3, t).

2.2 Grid-characteristic numerical methods

Grid-characteristic numerical methods were used for solving the system of hyperbolic equations (2.3).

One of the widely used approaches to solving three-dimensional system (2.3) is splitting it by space variables. In this case it can be replaced with four one-dimensional systems:

$$\frac{\partial}{\partial t}u + A_1\frac{\partial}{\partial x_1}u = 0,\ \frac{\partial}{\partial t}u + A_2\frac{\partial}{\partial x_2}u = 0,\ \frac{\partial}{\partial t}u + A_3\frac{\partial}{\partial x_3}u = 0,\ \frac{\partial}{\partial t}u = f.$$

These systems should be solved in their course, where a solution of the previous system should be used as the initial state to solve the next one. This approach allows to simplify numerical implementation and obtain better computation performance.

Each one-dimensional system is solved using common grid-characteristic numerical methods. For a system of equations with a single space variable

$$\frac{\partial}{\partial t}u + A\frac{\partial}{\partial x}u = f, \tag{2.4}$$

the solution is sought in the form of a grid function u_m^n defined at the points of the computational grid $x_m = mh$, $t^n = n\tau$, where h and τ are the grid size with respect to space and time.

Monotonous first order scheme (Courant-Isaacson-Rees). This scheme is constructed on the basis of the analysis of system (2.4) characteristics behavior and yields the following formulas (see [5]):

$$u_m^{n+1} = u_m^n - \frac{\tau}{h}\Omega^{-1}\Lambda^+\Omega(u_{m+1}^n - u_m^n) - \frac{\tau}{h}\Omega^{-1}\Lambda^-\Omega(u_m^n - u_{m-1}^n). \tag{2.5}$$

Here Ω is the matrix of left-hand eigenvectors of matrix A, $\Lambda^{\pm}$ is the diagonal matrix of corresponding eigenvalues.

Scheme (2.5) has the order of approximation $O(h, \tau)$, it is monotone and has the minimal approximation viscosity among the first-order monotone schemes, which is very important for the calculation of dynamical processes in inhomogeneous media.

Second order scheme (Lax-Wendroff). This is the only central scheme of the second order of approximation on a three-point stencil (i.e., a standard scheme using the values at the points $m-1, m, m+1$):

$$u_m^{n+1} = u_m^n - \frac{\tau}{2h} A(u_{m+1}^n - u_{m-1}^n) - \frac{\tau^2}{2h^2} A^2(u_{m+1}^n - 2u_m^n + u_{m-1}^n). \tag{2.6}$$

The scheme (2.6) has minimum smearing between schemes on three-point templates. However, this scheme is not monotonous and it leads to non-physical oscillations near discontinuities of the exact solution.

Hybrid scheme. Linear combination of these two schemes allows to overcome the limitations of each scheme. This combination can be written as:

$$u_m^{n+1} = u_m^n - \frac{\tau}{2h} A(u_{m+1}^n - u_{m-1}^n) + \frac{1}{2}((1-a)\frac{\tau}{h}\Omega^{-1}|\Lambda|\Omega + a\frac{\tau^2}{h^2}A^2)(u_{m+1}^n - 2u_m^n + u_{m-1}^n). \tag{2.7}$$

Here $|\Lambda|$ is a diagonal matrix of absolute eigenvalues of matrix A. If $a = 0$ the scheme (2.7) has the first order like the scheme (2.5). If $a = 1$ the scheme (2.7) has the second order like the scheme (2.6). The scheme (2.7) is hybrid one if a depends on the solution local behavior.

In this paper the local smoothness of the solution was determined from the following condition proposed by Fedorenko:

$$\frac{(u_{m+1}^n - 2u_m^n + u_{m-1}^n)}{2} \leqslant K\frac{(u_{m+1}^n - u_{m-1}^n)}{2}. \tag{2.8}$$

In these calculations we assume that $K = 0.5$. Parameter a is 0 when inequation (2.8) is false and 1 when inequation (2.8) is true. The resulting scheme has second order for smooth solution areas and first order near discontinuities. Different advanced schemes comparison (see for instance [8] for Lax-Friedrichs scheme, Rusanov's scheme, Godunov's scheme) shows that hybrid scheme should have good balance between scheme accuracy, implementation simplicity and calculation speed. However, a detailed comparison is a separate interesting task that requires additional study.

Hybrid scheme approach allows to achieve minimum smearing in the smooth solution areas and to avoid simultaneously parasitic oscillations near discontinuities. Figure 1 shows the velocity profile for the one-dimensional problem concerning the propagation of a rectangular pulse obtained using these difference schemes. The pulse width is 40 grid points, by the time depicted the pulse traveled 200 grid

points from its initial location. The calculation was performed for the Courant number $\sigma = 0.7$.

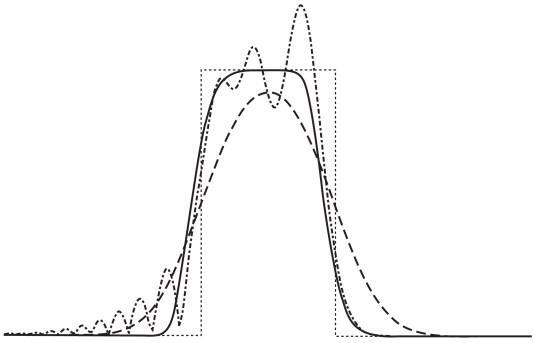


Figure 1. The problem of the rectangular pulse propagation : the dashed line corresponds to the Courant-Isaacson-Rees scheme, the dotted and dashed line corresponds to the Lax-Wendroff scheme, the solid line corresponds to the hybrid scheme and the dotted line shows the exact solution

2.3 Contact boundaries

Modelling of bodies with substantially different rheological and mechanical properties requires accurate solving of the contact boundaries problem. It is important to study wave interactions and reflection from the boundaries. Grid-characteristic numerical methods allow to set contact boundaries conditions explicitly. This approach gives higher precision if compared with pass-through calculation.

The conditions on a contact boundary are set in the form of relations between variables at two adjacent points on the contacting surfaces (see [6]). In this paper full-adhesive

$$v = v^{'}, \sigma_n = \sigma_n^{'}, \sigma_\tau = \sigma_\tau^{'} \tag{2.9}$$

and free-sliding

$$v_n = v_n^{'}, \sigma_n = \sigma_n^{'}, \sigma_\tau = \sigma_\tau^{'} = 0 \tag{2.10}$$

conditions are used.

Here, variables with and without primes correspond to the opposite contacting surfaces. The indices n and τ denote the normal and tangent directions, respectively.

3 Solving different dynamic biomechanical problems

3.1 Modelling brain injury consequences

The main goal of brain injury modelling in this research was to study the formation of the damage areas at early stages of cerebral trauma. It is well known in neurosurgery that damaged areas often do not coincide with the areas adjacent to the impact site. In particular, in case of the skull nape area hitting the damage is often localized in the forehead. This phenomenon referred to as a "countrecoup" can be accounted by means of numerical simulation of skull and brain wave processes.

The scheme of cerebral system that was chosen for this study is represented in Fig. 2. Quadrangular and triangular computational grids used are plotted in Fig. 6.

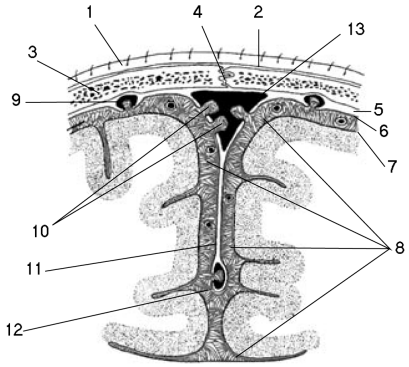


Figure 2. The scheme of cerebral system: (1) skin, (2) periosteum, (3) skull bone, (4) longitudinal seam, (5) dura mater, (6) arachnoid, (7) pia mater, (8) subarachnoidal cavity, (9) venous lacunae, (10) arachnoid granulations, (11) falx cerebrum, (12) the inferior sagittal sinus (the distances between the membranes are enlarged).

This research uses the model of cerebral system containing three-layer skull (two outer layers of compact bone tissue and internal layer of spongy bone tissue), a layer of liquor between skull and brain tissue, ventricles filled with liquor, membranes and gray substance.

The boundary between different layers of skull was fully adhesive. Two types of boundary conditions, free slip and adhesion, were tested for other contacts (i.e. surface between skull and gray substance). The free slip condition suits better to real biomechanical process.

Rheological parameters are summarized in the table below.

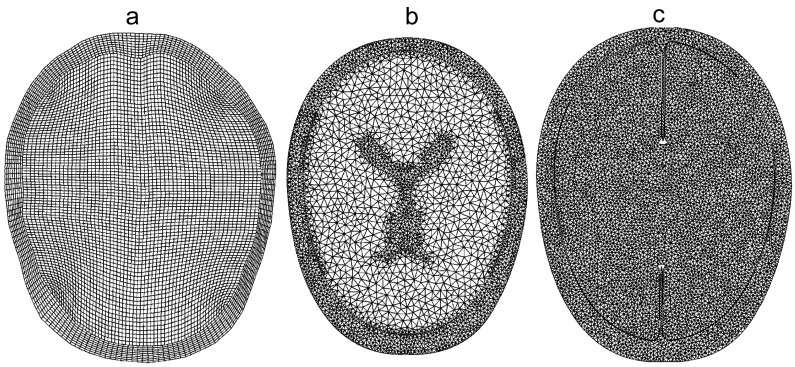


Figure 3. Quadrangular and triangular computational grids that were used in simulation

Tissue	ρ, kg/m^3	λ, MPa	μ, MPa
Compact bone tissue	1600	7900	5270
Spongy bone tissue	1500	3975	2650
Liquor	1000	1700	0.001
Grey substance	1020	1.7	0.23

Areas of positive and negative stresses after a strike from the left obtained as a result of modelling are presented in Fig. 4 a, b.

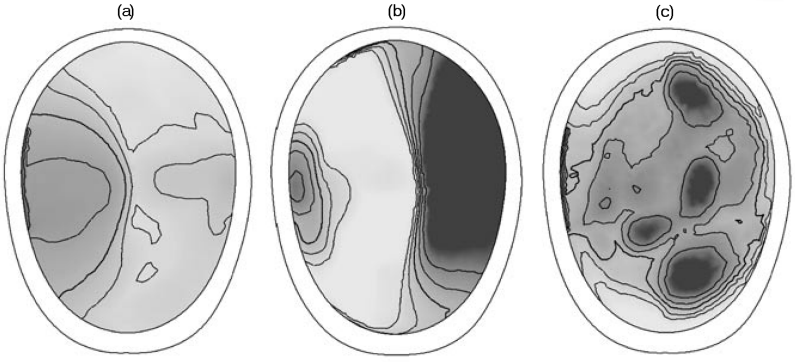


Figure 4. Areas of positive, negative and shear stresses after a strike from the left

The hypothesis is the following: because vessels and gray substance are fibrous they should stand longitudinal strain and pressure well but transversal load can damage them much more easily. So, we can expect that the brain damage arises mostly in the areas of highest shear load.

The comparison of calculated and obtained from MRT maps brain damages is represented in Fig. 5 a, b. Areas of calculated damage refer to the zones of highest shear load. MRT map was provided by Burdenko Principal Military Clinical Hospital.

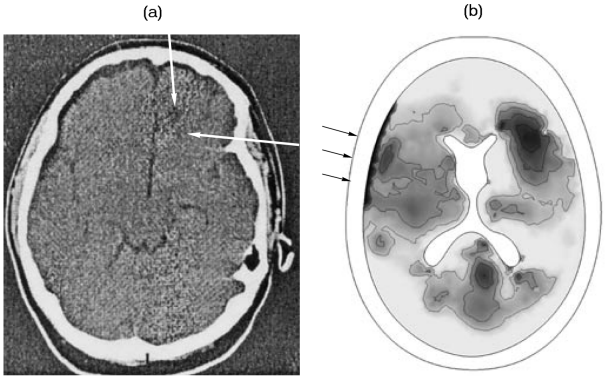


Figure 5. The comparison of computational and obtained from MRT maps brain damages

3.2 Modelling of laser cataract extraction

The surgery technology supposes an intracapsular crush of opaque lens nucleus by an energetic effect (ultrasound or laser radiation) without injuring of a retina or a cornea and then aspiration of masses through a self-sealing incision of width up to 3.5 mm. A detailed understanding of the energy propagation nature in the eye depending on different parameters allows to minimize the intraoperative traumas and patients medical rehabilitation period.

This problem of the dynamic process simulation in the eye during cataract extraction can be conventionally divided into three stages. The first stage includes the computation of impulse action on the lens. The second one supposes the calculation of the acoustic pulse in vitreous body up to the retina and its impact on the retina. This part of the process has practical interest because the retina can exfoliate as a result of the pulse impact on it. The last stage supposes opaque lens masses aspiration from the anterior chamber. This research is focused on the second stage.

The intensity of absorbed laser radiation is calculated according to Lambert-Beer law:

$$Q(r,h) \;=\; I(r)e^{-H/h}, \tag{3.1}$$

$$I(r) \;=\; I_0 e^{-(r/R)^2}. \tag{3.2}$$

Here r is the radius in the cross-section of the lightguide fiber, h is the depth of the laser pulse penetration, H is the characteristic deepness of penetration ($H = 3.2$ mm), I_0 is the intensity at the center of the pulse, R is the lightguide fiber diameter ($R = 0.3$ mm).

Since the first stage of the process was not the subject of modelling, the equivalent stress was applied for modelling pulse impact at the second stage. The laser pulse was a train with a 250 ms duration. The period of micropulses in the train was 12.5 ms and micropulse duration was 3 ms. Only the first 250 ms were taken into account because perturbation intensity is noticeably attenuated after the pulse is over.

Pulse propagation through the lens and the vitreous humour to the retina was modelled as described above. Computational grid and velocity field are shown at Fig. 4.

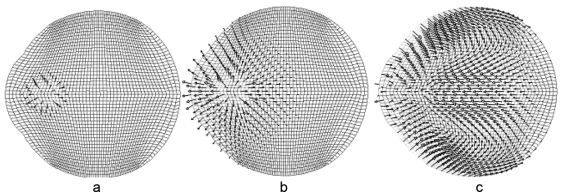


Figure 6. Computational grid and velocity field

The distribution of ocular media particle speeds is shown at Fig. 7 at time moments $t_1 = 6.24$ ms, when the perturbation reached the iris and $t_2 = 23.6$ ms when the perturbation is reflected from the posterior surface of the eye. It can be noted that the movement of ocular media has a rather complicated character. At the initial moment during the pulse an expansion of bio-media occurs, after the pulse is over, the pressure in the lens zone becomes less than in its surrounding areas. It induces changes of speed direction and a subsequent eye contraction as a consequence. So, the eye is periodically expanded and constricted during the operation. Tissue ruptures as well as cavitation processes are possible in the sites of expansion. In addition, reflected waves from ocular materials borders influence the wave picture.

As the result the picture of potential zones of ocular structure damage risk is rather complicated (see Fig. 8).

An eye lens pressure was also modelled using modern FEM software (ANSYS). The results are presented at Fig. 9.

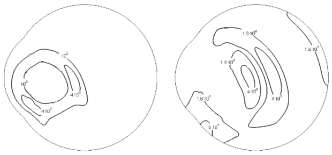


Figure 7. Isolines of speed modulus. Numerical characteristic of speed are indicated in m/sec

Figure 8. Zones of risk of the ocular structure damage

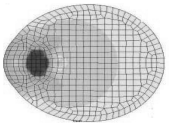


Figure 9. Calculation of pressure in eye lens using FEM and computational grid

It is worth paying attention to the fact that finite-element method identifies the zone of the most probable damage near the lightguide fiber. Method of characteristics allows to find also a possible damage zone near the retina and a number of small additional risk zones in the eye frontal part. These additional zones are caused by wave fronts interactions during initial fast dynamic part of overall process and can not be found using static or quasi-static modelling methods.

References

1. J. H. Adams, D. I. Graham, T. A. Genmarelli. *Head injury in man and experimental animals: neuropathology.* Acta Neurochir., 32 (1975), 15–30.
2. L. Zhang, K. H. Yang, A. I. King. *Comparison of brain responses between frontal and lateral impacts by finite element modelling.* J. Neurotrauma, 18 (2001), 21–30.
3. A. H. Kuijpers, M. H. Claessens, A. A. Sauren. *The influence of different boundary conditions on the response of the head to impact: a two-dimensional*

finite element study. J. Neurotrauma, 12 (1995), 715–724.
4. V. Novatskii. *Elasticity theory*. Nauka, Moscow, 1975 (in Russian).
5. K. M. Magomedov, A. S. Kholodov. *Grid characteristic methods*. Nauka, Moscow, 1988 (in Russian).
6. I. B. Petrov, A. G. Tormasov, A. S. Kholodov. *On the numerical study of unsteady processes in layered solids*. Izv. Akad. Nauk SSSR, Ser. Mekh. Tverd. Tela, 4 (1989), 89–95.
7. I. B. Petrov, A. S. Kholodov. *Numerical study of some dynamic problems in mechanics of deformable solids by the grid characteristic method*. Zh. Vych. Mat. Mat. Fiz., 24 (1984), 722–739.
8. J. A. Trangenstein. *Numerical solution of hyperbolic partial differential equations*. Cambridge University Press, New York, 2007.

I. B. Petrov
Department of Computer Science, Moscow Institute of Physics and Technology, Russia, 141700, Dolgoprudny, Institutskiy 9, +7 495 408 48 00, petrov@mipt.ru

A. V. Vasyukov
Department of Computer Science, Moscow Institute of Physics and Technology, Russia, 141700, Dolgoprudny, Institutskiy 9, +7 495 408 48 00, vasyukov@gmail.com

D. V. Chernikov
Department of Computer Science, Moscow Institute of Physics and Technology, Russia, 141700, Dolgoprudny, Institutskiy 9, +7 495 408 48 00

Y. V. Bolotskikh
Department of Computer Science, Moscow Institute of Physics and Technology, Russia, 141700, Dolgoprudny, Institutskiy 9, +7 495 408 48 00

APPLICATIONS OF EVOLUTIONARY OPTIMALITY IN STRUCTURED SYSTEMS TO MEDICAL AND BIOLOGICAL PROBLEMS

V. N. Razzhevaikin

Key words: models of biological systems, stability, evolutionary optimality, age structure, spatial structure, correlation adaptometry technique

AMS Mathematics Subject Classification: 92D15, 34G20

Abstract. The theory of the relation between evolutionary optimality and the stability of equilibrium states in structured systems is announced. The main result of the theory is stated in the case of quasilinear dynamical systems in normed spaces. Applications of the theory to models of structured biological communities are discussed. Functionals for communities with age or with spatial structure may be computed on the basis of available information about steady-state stationary distributions. The functionals are optimized with respect to the parameters of evolutionary selection. The ways to use the results in several medical problems are indicated.

1 Introduction

Extreme principles arise from the formalization of the idea of evolutionary selection and its consequence – evolutionary optimality, which goes back to Darwin's fundamental work. The fact that the competitive struggle within a limited space leads to elimination of all the species, except for a small number of ones that are most adapted to the given conditions, makes it possible to construct the fundamental principles of functioning of biological systems.

These principles are based on the hypothesis that stationary states of biological communities formed in the course of evolution are stable. Within the framework of the simplest finite-dimensional mathematical model, one can uncover the essence of necessary stability conditions that have the form of extreme relations. The objects to be optimized are the values of Malthusian functions computed over formed stable equilibrium states. Later, these functions may play the role of reference points in the construction of functionals for more complicated systems.

This work was supported by the Russian Foundation for Basic Research, project No 09-07-00398.

In his previously constructed theory, the author proposed a construction that relates the stability of steady states of distributed biological systems to the extremality of values of inherited traits in species surviving in these states. Despite its artificial character (the purely technical hypothesis of quasilinearity i.e., the linearity of unbounded components of the operators on the right-hand sides of the equations leads to rather unnatural constraints in continuous-time problems), this theory yields interesting and practically useful results for many important problems.

Here the most general mathematical statement of the main applied result in the theory for the case of autonomous quasilinear systems with continuous time in Banach spaces with unbounded operators on the right-hand side is given. At this level of abstractness, the main result (a necessary stability condition for a stationary distribution) can be formulated in terms of the localization properties of the spectrum of the explicitly specified operator on the right-hand side of the system calculated over this distribution. Further elaboration associated with the possibility of constructing of functionals to be optimized, requires more specific formulations.

2 Preliminaries of evolutionary optimality

As an example of the most simplest model one can consider a competition model of the finite number biological species $dx_i/dt = x_i f_i(x)$, $i = 1, \ldots, n,\ x = (x_1, \ldots, x_n)$. Here one can find as necessary condition for stability of an equilibrium of the form $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_m, 0, \ldots, 0),\ \bar{x}_i > 0,\ i = 1, \ldots, m$ the equality $f_i(\bar{x}) = \max(f_j(\bar{x}))$, $1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant n$. It has the character of the extremal relationship and is called the evolutionary optimality principle. Its biological sense is in that species, survived in a stable equilibrium, are obliged to have the maximal values of Malthusian parameters among all potentially possible ones, which may be computed at the equilibrium. These factors characterize the "power" of species in its Darwin's understanding if one bears in mind the statement of the principle about survival of the most strong.

Since in the equilibrium the species with senior numbers are absent, they may be considered as virtual ones, i.e. one can add to their collections anyone from other species, which have hypothetical possibility to turn out to be in the initial set. Herewith their distinguishing parameters can have a free nature and in particular can be chosen from a certain area in the space of parameters, so the optimization problem already may be solved with respect to it. Such an expansion allows to find isolated values of parameter, under which the equilibrium turns out to be stable. On this way one can built the methods of the calculation of parameters values

for (quasi)stationary biological systems, for determination of which the natural measurement can turn out to be impossible or difficult.

3 Some results of general theory

Here some of fundamental results concerning the theory relating stability and evolutionary optimality in models of structured biological systems are formulated. The consideration is restricted to continuous-time autonomous systems, which are constructively covered by the description of quasilinear dynamical systems in Banach spaces.

The original autonomous dynamical system has the form

$$d_t x = (h_x + a(x,y))x, \quad d_t y = h_y y + b(x,y)$$

where $t \in J = [0,T]$, $T > 0$, $d_t = d/dt$, $x \in X$, $y \in Y$, X and Y are Banach spaces, $a \in C^1(X \oplus Y, B(X))$, $b \in C^1(X \oplus Y, Y)$, and $B(X)$ is the space of bounded linear operators in X. The linear operators $h_{x,y}$ are infinitesimal generators of strongly continuous semigroups of linear bounded operators acting in X and Y, respectively. with domains $D(h_{x,y})$. Specifically, such operators are closed and their domain are dense.

The variables of the system are initially divided into two groups: evolutionary ones x (their variation vanishes at zero values) and non-evolutionary y (required only for the generality necessary in applications; in mathematical constructions they can be discarded for brevity). The reduced formulation of system has the form $d_t w = hw + K(w)$, $w = (x,y) \in W = X \oplus Y$, $h = \mathrm{diag}\{h_x, h_y\}$.

For the system the stability of a stationary solution $\bar{w} = (\bar{x}, \bar{y})$ is understood in the sense that the spectrum of the Jacobian of the mapping calculated over positive time at $\bar{w}$ lies inside the unit disk.

A projector P, i.e., a linear bounded idempotent ($P^2 = P$) operator in W is called *admissible with respect to* h if the domain of the latter satisfies $PD(h) \subset D(h)$ and, additionally, $hP = PhP$.

A projector P in W is called *admissible with respect to* $w \in W$ if $Pw = w$, P is admissible with respect to h and commutes with I_Y (projector onto Y), and, for some neighborhood $O(w) \subset W$ $v \in PW \cap O(w)$ implies $K(v) \in PW$. The Jacobian of the system calculated at the equilibrium $\bar{w} = (\bar{x}, \bar{y})$ is decomposed into the sum $l(\bar{w}) = l_0(\bar{w}) + l_1(\bar{w})$ where $l_0(\bar{w}) = \mathrm{diag}\{h_x + a(\bar{w}), 0\}$.

Let $\mathbf{C}_{-\delta}, \quad \delta > 0$, be the left complex half-plane shifted to the left by δ. In the autonomous case, the most interesting result of the theory (minus some generalizations aimed at larger adequacy for biological setting) can be formulated as follows.

Theorem. *Let $\bar{w} = (\bar{x}, \bar{y})$ be a linear stable stationary solution to the system. Then, for any projectors P_1, P_2 in W that are admissible with respect to $\bar{w}$ and such that $P_1P_2 = P_2P_1 = P_2$ and $P_2I_Y = P_1I_Y$, there exists $\delta > 0$ such that $\sigma((Ql_0(\bar{w}))_{QW}) \subset \mathbf{C}_{-\delta}$.*

Here, $Q = P_1 - P_2$ is a projector in W, A_V is the restriction of a linear operator $A : W \to W$ to a subspace $V \subset W$ that is invariant with respect to it, and $\sigma(A)$ is the spectrum of the operator A. So, $\bar{x} \in \mathrm{Ker}(h_x + a(\bar{w}))$, therefore, the following result holds in the case $\bar{x} \neq 0$, which is of interest for applications.

Corollary *(extreme principle). Equal to zero maximum of the upper bound for the real part of the spectrum of restrictions of the operator $h_x + a(\bar{w})$ to $QW \oplus \{l\bar{x}\}, \quad l \in \mathbf{R}$ is reached at the vector $\bar{x} \neq 0$ realized in the stable equilibrium $\bar{w} = (\bar{x}, \bar{y})$.*

This assertion is a direct generalization of the necessary condition for external stability (formulated in extreme form) to the structured case.

4 Continuous age-structured model

The original system of equations describing the dynamics of a community of species with a continuous age structure has the form

$$\partial x_\lambda = -\mu_\lambda x_\lambda, \quad \lambda \in \Lambda,$$
$$\partial y_i = b_i, \quad i \in I = \{1, \ldots, J_1\}, \quad \partial_t y_j = b_j, \quad j \in J = \{J_1 + 1, \ldots, J_2\}$$

with the boundary conditions

$$x_\lambda(0,t) = \int_0^\infty \beta_\lambda(a) x_\lambda(a,t)\, da,\ \lambda \in \Lambda, \quad y_i(0,t) = \int_0^\infty g_i(a) y_i(a,t)\, da,\ i \in I$$

and with suitable initial conditions. Here, t is time, a is age, $\partial_t = \partial/\partial t$, $\partial = \partial_t + \partial_a$, λ is the index (possibly from the infinite set Λ) of an evolving species with an age population density $x_\lambda = x_\lambda(a,t)$, and i is the index (from a finite set) of a not evolving species with an age density$y_i = y_i(a,t)$. The difference between the first and the second is that the second can be controlled externally. Moreover, the system can contain external (i.e., not evolving) variables $y_j = y_j(t)$ with no age structure.

The system is assumed to be autonomous in time, so each of the death rates $\mu_\lambda = \mu_\lambda(a, x, y)$ of individuals of the species indexed by λ and the current variations (death rate, migration, etc.) of the non-evolving species are assumed to depend only on the age of these individuals and the values of the distribution vectors $x = x(t) = \{x_\rho(\cdot, t)\}$, $\rho \in \Lambda$, $y = (y_I, y_J)$, $y_I = y_I(t) = \{y_i(\cdot, t)\}$, $i \in I$, $y_J = y_J(t) = \{y_j(t)\}$, $i \in J$, which describe the current state of the community structure (i.e., at a fixed time). Here and below, a dot placed instead of a distribution variable means that the distribution is treated as a whole, i.e., as an element of a suitable function space. More specifically, this means that the death rates are functions of the state of the community as a whole. The birth rates $\beta_\lambda(a)$, $g_i(a)$ are assumed to be independent of the current form of the distributions (quasilinearity condition).

To use the results of the previous section, all the functions are assumed to be twice uniformly continuously differentiable with respect to their arguments. As suitable Banach spaces, one can use $X = l_\infty(\Lambda) \otimes L_1(\mathrm{R}_+)$ and $Y = \mathrm{R}^{|\mathrm{J}_1|} \otimes L_1(\mathrm{R}_+) \oplus \mathrm{R}^{|\mathrm{J}_2|}$. Here, $l_\infty(\Lambda)$ denotes the space of real-valued functions on Λ with a countable support that are summable on the support in the l_∞ norm. Integral summability with respect to $a \in \mathrm{R}_+$ reflects the natural requirement that the total population size be bounded. The operator hx is assumed to be diagonal in the structure of $l_\infty(\Lambda)$, and its nonzero component indexed by Λ is a closed operator of the form $-\partial_a$ with the domain consisting of bounded absolutely continuous functions from $L_1(\mathrm{R}_+)$ that satisfy the first boundary condition. In a similar manner, the second of these conditions defines the domain of the corresponding diagonal components of h_y. The remaining components on the right-hand side (death rate, etc.) are described by bounded operators and functions. Since the general quasilinear theory does not prevent their dependence on the phase variables, this dependence (possibly even on the age distributions overall) is admissible in the application under study.

Let $(\bar{x}, \bar{y})$ be a stationary solution of the system that is stable in the sense of the previous section. Let $\bar{\lambda} \in \mathrm{supp}\bar{x}$ (the subset of those values of $\lambda \in \Lambda$ for which $\bar{x}_\lambda(a)$ does not vanish identically). For the above construction one can verify the conditions of the above theorem. So, by the Lotka theorem, having the maximal real part the eigenvalue of the operator $h_\lambda - M_\lambda$, where M_λ is the operator of pointwise multiplication (with respect to a) by the function $\mu_\lambda(a, \bar{x}, \bar{y})$, is real and the corresponding eigenfunction is positive. This eigenvalue is determined for every $\lambda \in \Lambda$ from the characteristic equation for κ, which has the form $1 = \Phi(\lambda, \kappa, \bar{x}, \bar{y})$, where the righthand side equals to $\int_0^\infty \beta_\lambda(a) \exp(-\kappa a - \int_0^a \mu_\lambda(s, \bar{x}, \bar{y})\, ds) da) da$.

By the theorem its solution satisfies $\kappa_\lambda \leqslant 0$, so $\Phi(\lambda, 0, \bar{x}, \bar{y}) \leqslant \Phi(\lambda, \kappa_\lambda, \bar{x}, \bar{y}) = 1 = \Phi(\bar{\lambda}, 0, \bar{x}, \bar{y})$, and the functional $\varphi(\lambda) = \Phi(\lambda, 0, \bar{x}, \bar{y})$ reaches its maximum value at $\bar{\lambda}$. Hence, in terms of the model of this section, the extreme principle can be formulated as follows.

Theorem. *If the above system has a stable stationary equilibrium* $\bar{x} = \{\bar{x}_\lambda(a)\}$, $\lambda \in \Lambda$, *then* $\bar{\lambda} \in \operatorname{supp}(\bar{x})$ *satisfies the relation* $\varphi\left(\bar{\lambda}\right) = \max\limits_{\lambda\in\Lambda}\left(\varphi\left(\lambda\right)\right)$ *with the functional* $\varphi\left(\lambda\right) = \int\limits_0^\infty b_\lambda(a)\exp\left(-\int\limits_0^a m_\lambda(s,\bar{x})\,ds\right)da$.

From a practical point of view, this means that the set of parameters values λ for an a priori known steady-state stationary distribution $\bar{x}$ can be determined by maximizing the above functional over λ. Note that the theoretical maximum value of this functional is equal to unity.

The biological meaning of this functional which goes back to Lotka's fundamental constructions, is the mean number of the newborn per individual with allowance for the age-specific death rate.

5 Models with continuous spatial structure

Starting here, the nonevolving variables y are ignored.

For spatially distributed biological communities, the most frequently used continuous model is based on systems of second-order quasilinear parabolic equations with homogeneous conditions on the boundary of the considered spatial domain. In the case of an a priori known stationary distribution of biological species, the results of the general theory can be used to construct the minimization problem for a suitable integral functional in order to determine the selection parameters values corresponding to surviving species.

The original system of equations has the form $\partial_t x_\lambda = h_\lambda x_\lambda + \hat{a}_\lambda(x)x_\lambda$, $\lambda \in \Lambda$ with $x_\lambda = x_\lambda(\xi,t)$ – the spatial biomass density of the species indexed by λ at the point $\xi \in \Omega$ at time t (here, $\Omega \subset \mathbf{R^n}$ is a connected bounded domain with a sufficiently smooth boundary $\partial\Omega$ which describes the habitat of the community), $h_\lambda x_\lambda = \operatorname{div}\left(A_\lambda(\xi)\left(\operatorname{grad} x_\lambda + x_\lambda \operatorname{grad} q_\lambda(\xi)\right)\right)$ are the elliptic operators with rather smooth coefficients $\alpha_\lambda^{\iota\kappa}(\xi)$, $q_\lambda(\xi)$, $\iota,\kappa = 1,\ldots,n$, in the closure of Ω, and $A_\lambda\left(\xi\right) = \|\alpha_\lambda^{\iota\kappa}(\xi)\|$ are symmetric matrices that are uniformly positive definite in Ω. Here, the divergence and the gradient are calculated with respect to ξ, and $(u,v) = \sum\limits_{i=1}^{n} u_i v_i$. These operators are used to describe diffusion in the case of an anisotropic space and the presence of a spatial drift defined by the gradient of $q_\lambda(\xi)$ (for example, in chemotaxis problems, this is a species-specific function of the attractant concentration). On the boundary $\partial\Omega$ the homogeneous Dirichlet conditions $x_\lambda|_{\partial\Omega} = 0$ or $\left(\operatorname{grad} x_\lambda + x_\lambda \operatorname{grad} q_\lambda(\xi), A_\lambda(\xi)\nu\right)|_{\partial\Omega} = 0$ – the impermeability conditions, where ν is the normal vector to the boundary at the point $\xi \in \partial\Omega$, are implied. The last condition, in which the projection of the total flux (due to diffusion and drift) of individuals of species λ through the boundary is set equal to

zero, means that, from a biological point of view, the boundary is an insuperable barrier for individuals.

The operator $\hat{a}_\lambda(x)$ defines the pointwise (with respect to ξ) multiplication of function $x_\lambda(\xi,t)$ by $a_\lambda(x(\cdot,t),\xi)$, in which $x = x(\cdot,t) = \{x_\rho(\cdot,t)\}$, $\rho \in \Lambda$. for each spatial point ξ, this operator plays the role of a Malthusian function for the species indexed by λ. The collection of these operators specifies all the intraspecies and interspecies interactions in the community.

As in the previous section, all introduced functions are assumed to be twice uniformly continuously differentiable with respect to their arguments. As $W = X$ we use the space $l_\infty(\Lambda, \{L_2^\lambda(\Omega)\}_{\lambda\in\Lambda})$ of l_∞-normalized finite-dimensional vectors with the λ-th component $x_\lambda(\xi)$ from the Hilbert space $L_2^\lambda(\Omega)$ with the norm $\sqrt{\int_\Omega e^{q_\lambda(\xi)} x_\lambda^2(\xi)\, d\xi}$. The operators$h = h_x$ and a(x) act componentwise as h_λ and $\hat{a}_\lambda(x)$, and the domain of h_λ is the space $W_2^\lambda(\Omega) \subset W_2^2(\Omega)$ of functions having second partial derivatives from $L_2^\lambda(\Omega)$ and satisfying the boundary conditions in the sense of the trace. The operators h_λ thus defined are closed and self-adjoint in $L_2^\lambda(\Omega)$, which means that they are sectorial. Therefore, there exists an analytical (and, hence, strongly continuous) semigroup for which they are infinitesimal generators.

Moreover, one can check that the variational principle holds for $h_\lambda + \hat{a}_\lambda(x)$. This means that the minimum eigenvalue of $h_\lambda + \hat{a}_\lambda(x)$ is simple and, up to the sign, coincides at its eigenfunction $v^\lambda(\xi)$ with the minimum of the functional

$$\Phi(\lambda, x, v) = \left[\int_\Omega e^{q_\lambda(\xi)} \left[(w_\lambda(\xi), A_\lambda(\xi) w_\lambda(\xi)) - a_\lambda(x,\xi) v^2(\xi)\right] d\xi\right] \left[\int_\Omega e^{q_\lambda(\xi)} v^2(\xi)\, d\xi\right]^{-1}$$

where $w_\lambda(\xi) = \operatorname{grad} v(\xi) + v(\xi) \operatorname{grad} q_\lambda(\xi)$. The minimum of functional (8) is calculated over$v(\xi) \neq 0$ on the Friedrichs extension of the domain of $h_\lambda + \hat{a}_\lambda(x)$, which coincides with $H_0^1(\Omega)$ in the Dirichlet case (distributions on Ω with first partial derivatives from $L_2(\Omega)$ vanishing on the domain boundary in the sense of the trace) and with $H^1(\Omega)$ in the impermeability case (the same but without the boundary conditions).

Let $\bar{x}$ be a stationary solution to the system that is stable in sense above, and let $\bar{\lambda} \in \operatorname{supp}\bar{x}$. The last inclusion means that the kernel of $h_{\bar{\lambda}} + \hat{a}_{\bar{\lambda}}(\bar{x})$ is not empty, since it includes the nonzero distribution $\bar{x}_{\bar{\lambda}}$. For it Green's identity gives $\Phi\bar{\lambda}, \bar{x}, \bar{x}_{\bar{\lambda}}) = 0$. The variational principle also implies the inequality $\Phi(\lambda, \bar{x}, v^\lambda) \leqslant \Phi(\lambda, \bar{x}, \bar{x}_{\bar{\lambda}})$ and the equality $\Phi(\lambda, \bar{x}, v^\lambda) = -\sup \sigma(h_\lambda + \hat{a}_\lambda(\bar{x}))$ for $\lambda \in \Lambda$. The application of the first theorem to $\lambda \notin \operatorname{supp}\bar{x}$ yields $\sup \sigma(h_\lambda + \hat{a}_\lambda(\bar{x})) < 0$. Collecting these relations, one obtains the chain $\Phi(\bar{\lambda}, \bar{x}, \bar{x}_{\bar{\lambda}}) = 0 < \Phi(\lambda, \bar{x}, v^\lambda) \leqslant \Phi(\lambda, \bar{x}, \bar{x}_{\bar{\lambda}})$ which implies

the extreme principle for$\varphi(\lambda) = \Phi(\lambda, \bar{x}, \bar{x}_{\bar{\lambda}})$. More specifically, the following result holds.

Theorem. *If the system with one of the boundary conditions has a stable stationary equilibrium* $\bar{x} = \{\bar{x}_\lambda(\xi)\}$, $\lambda \in \Lambda$ *then* $\bar{\lambda} \in \operatorname{supp}(\bar{x})$ *satisfies the relation* $\varphi(\bar{\lambda}) = \min\limits_{\lambda \in \Lambda} \varphi(\lambda)$.

Note that $\varphi(\lambda)$ can be replaced by $\phi(\lambda) = \Phi(\lambda, \bar{x}, v^\lambda)$. First, this follows formally from the first inequality and $\Phi(\bar{\lambda}, \bar{x}, v^{\bar{\lambda}}) \leqslant \Phi(\bar{\lambda}, \bar{x}, \bar{x}_{\bar{\lambda}})$. Second, assuming that $\bar{x}$ is stable with respect to spatial perturbations of the distributions for $\bar{\lambda} \in \operatorname{supp}(\bar{x})$, one obtains $\bar{x}_{\bar{\lambda}} = v^{\bar{\lambda}}$ (otherwise, $\bar{x}_{\bar{\lambda}}$ is unstable with respect to $v^{\bar{\lambda}}$), which makes both formulations of the extreme principle equivalent. However, in the construction of functionals, the second version allows to take into account the form of the steady-state distribution only in the computation of the coefficients. Specifically we do not need to determine its spatial derivatives, since, instead of the latter, we use the functions v^λ computed by minimizing the functional $\Phi(\lambda, \bar{x}, v)$.

6 Conclusions

If there is no need to use unbounded operators to invoke the general theory, one can use "simpler" constructions. The examples include various models for the propagation of epidemics, epiphytoties, etc. Even models with a discrete structure can be formally reduced to the problem addressed in the preliminary approach. The same is true for discrete-time systems. The best-known population model with discrete time and age is that of Leslie. Some of its generalizations associated with the possibility of interage (more exactly, interstage) transitions also have found their reflection in the computation of evolutionary selection functionals.

Concerning applications of these results, an example is the theory of correlation adaptometry, constructed on the basis of the extreme properties of functionals for spatial distributions. This theory uncovers the relation between the level of unfavorable actions on a population and the degree of correlation between the distributions of physiological parameters of its terms.

V. N. Razzhevaikin
Contacts for the author: Dorodnicyn Computing Center, Russian Academy of Sciences, Russia, 119333, Moscow, Vavilov str., 40, razzh@mail.ru

SIMULATION OF THE MEASUREMENT OF INTRAOCULAR PRESSURE

V. L. Yakushev

Key words: intraocular pressure, optical analyzer, nonlinear shell theory, method of finite differences, method of additional viscosity

AMS Mathematics Subject Classification: 74K25

Abstract. The intraocular pressure is an important characteristic of the human eye. For that reason, the elaboration and development of new methods for its measurement is an important direction of research in ophthalmology. In this report the procedure of measuring the intraocular pressure by an optical analyzer is numerically simulated. The cornea and the sclera are considered as axisymmetrically deformable shells of revolution; the space between these shells is filled with incompressible fluid. Nonlinear shell theory is used to describe the stressed and strained state of the cornea and sclera. The spatial problem was decided by a method of finite differences. For a solution of the nonlinear problems of deformation shells a method of additional viscosity was used. The optical system is calculated from the viewpoint of the geometrical optics. Dependences between the pressure in the air jet and the area of the surface reflecting the light into a photo-detector are obtained. The shapes of the regions on the cornea surface are found from which the reflected light falls on the photo-detector. First, the light is reflected from the center of the cornea, but then, as the cornea deforms, the light is reflected from its periphery.

1 Introduction

Presently, the use of mathematical modelling in medical research is expanding. One field of the successful application of such models is the development of mechanical models of the eye based on fluid dynamics. The intraocular pressure is an important characteristic of the human eye. The deviation of this pressure from the norm is a cause of the impairment of vision in many patients. For that reason, the elaboration and development of new methods for its measurement is an important direction of research in ophthalmology. The intraocular pressure has important physiological functions - it smoothes the intraocular shells and gives the eyeball the shape required by the optical eye system. From the level of the intraocular pressure, one can judge the development of such pathological processes as glaucoma or opacity of aqueous humor and vitreous body. The intraocular fluid feeds the internal structures of the

eye. It provides for the exchange processes between the internal structures and the tissue structures. Tonometry is the measurement of the intraocular pressure (IOP) to determine the ability of the eyeball to deform under the influence of an external mechanical action, which can be applied to the cornea and the sclera. The IOP is measured using specially designed devices. There are finger, impression, and applanation methods of tonometry. One variant of the applanation tonometry assumes that a jet of air is directed to the cornea center (air-puff tonometry), and the displacement of the cornea is measured by the reflected beam of light (optical method). This displacement is then used to calculate the IOP.

2 Statement of the problem

The IOP is measured using specially designed devices. In this paper, we numerically simulate the eye deformation when the pressure is measured using ORA (the Ocular Response Analyzer) developed by the USA company Reichert (see [1]). The measurement procedure is as follows. A patient presses his or her forehead to the device. A narrow beam of light is directed to the cornea center (which is depicted in fig. 1 in the coordinates x, y, z as a segment of a sphere) at a certain angle using a special positioning system. A point light source is located at the point O. The light passes through the aperture A; as a result, a part of the light flux is cut off and an illuminated area S emerges on the cornea.

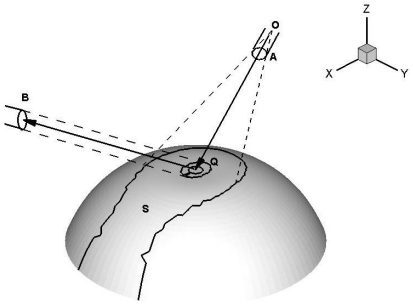


Figure 1. The path of a narrow beam in the device

The photo-detector is located at the point B, and its axis passes through the cornea center. Since the photo-detector has a limited size, it captures only the part of the light flux reflected from the cornea in the region Q. Then, an air jet in which the pressure increases from zero to a certain magnitude is directed to the center of the cornea. As a result, the cornea is deformed, and the reflected light

flux changes depending on the cornea shape. As a result of the measurements, a curve is shown on the device display that plots the reflected light flux against the pressure in the jet. Based on numerous experiments on the comparison of the results provided by the ORA with other methods for the IOP measurement, a method for the interpretation of the result obtained by the ORA was developed that yields the IOP value. A feature of this device is the digital signal processing, which gives more accurate values of the IOP.

However, the experiment-based procedure of the measurement result interpretation does not fully utilize the device capabilities. The mathematical modelling of the measurement process makes it possible to justify the analysis of measurements and calculate a more accurate value of the IOP.

A considerable difficulty in the interpretation of the measurement results is the lack of fundamental knowledge about the bio-mechanical properties of the cornea. Presently, there are no conventional methods for the life-time measurement of these properties. For that reason, simplified mechanical models of the eye are often used in the literature; these models only make it possible to draw qualitative conclusions. From the mechanical point of view, a scheme of the human eye is described in [2,3]. The shell of the eye ball consists of two parts: the cornea and the sclera. Their mechanical properties are quite different. The sclera and the cornea are separated by a thin membrane. The anterior chamber between the cornea and the membrane is filled with aqueous humor, and the posterior chamber is filled with vitreous body.

In the calculations, we assume that the eyeball is filled with incompressible fluid. The influence of the intermediate membrane is neglected. The cornea and the sclera are considered as shells of revolution uniformly loaded by the internal pressure p_i (see fig. 2) and firmly fastened at their common points.

The uniform external pressure p_e is applied at the circle of radius r_p in the center of the cornea; this pressure increases from zero to a certain magnitude. Since the shells are firmly fastened at their common points, they affect each other because a part of the intraocular fluid flows into the sclera as the external pressure exerted by the air jet increases; the sclera is thrust and the intraocular pressure increases. Its value is found from the condition that the internal volume of the eye remains unchanged.

3 Nonlinear equations of shells of revolution theory

We assume that the cornea and the sclera are axisymmetric about the vertical axis z and the pressure at the center is also distributed symmetrically about this axis. Thus, we have an axisymmetric problem for the calculation of the deformation of both shells, and we may consider only the cut of the shell at $y = 0$. The coordinates

$x,\ z$ in this section are denoted by the capital letters $X,\ Z$. The shell surface can be obtained by rotating the plane curve $X = X(s_0),\ Z = Z(s_0)$, around the axis z, where s_0 is the arc length along the cornea surface measured from its center (see [6]). The angle between the tangent to the surface and the axis x is $\varphi(s_0)$. The values $X(s_0)$,$Z(s_0)$ and $\varphi(s_0)$ are obtained by solving the problem of the simultaneous deformation of the cornea and the sclera. Such problems were studied in [4, 5, 7, 8]. The cornea and the sclera are considered as elastic shells whose deformation is described by a geometrically nonlinear theory under finite displacements and rotation angles.

Here, we give only the basic equations of the theory of shells of revolution under the influence of an axisymmetric load. These equations are the same for the cornea and the sclera; only their geometric and mechanical parameters are different. For that reason, we write these equations in the general form.

In the system of coordinates $X,\ Z$, consider (see [6]) an axisymmetrically loaded shell of revolution of thickness h. The coordinates of the midsurface $X,\ Z$, the angle φ between the tangent to the midsurface and the axis, the curvature radii in the meridional r_1 and circumferential direction r_2 are known functions of the arc length of the midline s. The corresponding quantities in the unstrained state are marked by subscript 0; for example, the arc length in the initial state is denoted by s_0. The normal to the midsurface is directed such that the tangent and the normal form the righthanded system of coordinates.

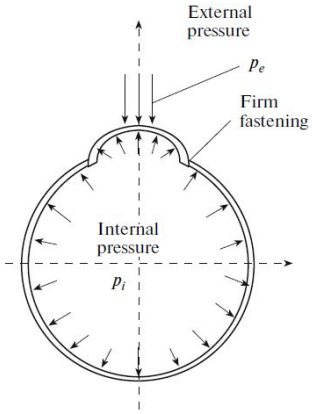


Figure 2. The scheme of the cornea and the sclera

The strains at the distance ζ along the normal to the midsurface in the meridional direction ε_φ and the circumferential direction ε_θ are determined on the basis of the Kirchhoff-Love model as

$$\varepsilon_\phi = \varepsilon_1 - \zeta k_1, \quad \varepsilon_\theta = \varepsilon_2 - \zeta k_2, \tag{3.1}$$

$$\varepsilon_1 = \frac{ds}{ds_0} - 1, \quad k_1 = \frac{1}{r_1} - \frac{1}{r_{10}}, \quad \varepsilon_2 = \frac{X}{X_0} - 1, \quad k_2 = \frac{\sin\varphi - \sin\varphi_0}{X_0}. \tag{3.2}$$

The coordinates of the midsurface in the strained state are determined by the equations

$$\frac{d\varphi}{ds_0} = k_1 + \frac{1}{r_{10}}, \quad \frac{dX}{ds_0} = (1+\varepsilon_1)\cos\varphi, \quad \frac{dZ}{ds_0} = (1+\varepsilon_1)\sin\varphi. \tag{3.3}$$

In addition, the conformity conditions for the strains must be satisfied:

$$\frac{d\varepsilon_2}{ds_0} = \frac{1}{X_0}\left[(1+\varepsilon_1)\cos\varphi - (1+\varepsilon_2)\cos\varphi_0\right], \tag{3.4}$$

$$\frac{dk_2}{ds_0} = \frac{1}{X_0}\left[\left(\frac{d\varphi_0}{ds_0} + k_1\right)\cos\varphi - \left(\frac{d\phi_0}{ds_0} + k_2\right)\cos\varphi_0\right]. \tag{3.5}$$

The equations of equilibrium in the case of considerable displacements and rotation angles are written for the strained state as

$$\frac{dQ_\varphi}{ds_0} + (1+\varepsilon_1)\frac{N_\theta\sin\varphi + Q_\varphi\cos\varphi}{X} + N_\varphi\left(k_1 + \frac{d\varphi_0}{ds_0}\right) + p_z = 0, \tag{3.6}$$

$$\frac{dN_\varphi}{ds_0} + (1+\varepsilon_1)\frac{N_\varphi - N_\theta}{X}\cos\varphi - Q_\varphi\left(k_1 + \frac{d\varphi_0}{ds_0}\right) + p_\varphi = 0, \tag{3.7}$$

$$\frac{dM_\varphi}{ds_0} + (1+\varepsilon_1)\frac{M_\varphi - M_\theta}{X}\cos\varphi - Q_\varphi = 0, \tag{3.8}$$

where Q_φ is the transverse force, N_φ and N_θ are the normal forces, M_φ and are M_θ the bending momentums, p_φ and and p_z are, respectively, the tangent and the normal (with respect to the midsurface) components of the distributed load.

We solve this problem using the additional viscosity technique (see [6]). Introducing the rheological viscosity, we obtain, expressions for the normal forces and moments:

$$N_\varphi = \frac{Eh}{1-v^2}\left[\varepsilon_1 + v\varepsilon_2 + \tau\left(\dot{\varepsilon}_1 + v\dot{\varepsilon}_2\right)\right], \tag{3.9}$$

$$M_\varphi = -\frac{Eh^3}{12\left(1-v^2\right)}\left[k_1 + vk_2 + \tau\left(\dot{k}_1 + v\dot{k}_2\right)\right], \tag{3.10}$$

$$N_\theta = \frac{Eh}{1-v^2}\left[\varepsilon_2 + v\varepsilon_1 + \tau\left(\dot{\varepsilon}_2 + v\dot{\varepsilon}_1\right)\right], \tag{3.11}$$

$$M_\theta = \frac{Eh^3}{12\left(1-v^2\right)}\left[k_2 + vk_1 + \tau\left(\dot{k}_2 + v\dot{k}_1\right)\right]. \tag{3.12}$$

Shell theory is an approximate one, and it does not take into account the relationship between and the shear strain. For that reason, this relation is not included in the constitutive equations. In order to have unified expressions of type Eq. (3.9) - (3.12) for all the force factors needed to write a resolving system of equations, we formally introduce by analogy with Eq. (3.9), the quantity γ for the transverse force Q_φ (see [6]):

$$Q_\varphi = \frac{Eh}{1-v^2}\left(\gamma + \tau\dot{\gamma}\right). \tag{3.13}$$

The resolving system of equations was constructed using the procedure described in [6]. Expressions Eq. (3.9) - (3.12) and (3.13) for the forces and moments were substituted into equilibrium Eq. (3.6) - (3.8). The derivatives $\partial^2\varepsilon_2/\partial s_0\partial t$ and $\partial^2 k_2/\partial s_0\partial t$ were taken from conformity relations Eq. (3.4) differentiated with respect to t. Similarly, $\partial\varepsilon_2/\partial t$ and $\partial k_2/\partial t$ in Eq. (3.2) were determined.

Three other equations were obtained by differentiating the expressions for the coordinates of midsurface Eq. (3.3) with respect to t and combining the original equations with these derivatives.

As a result, a system of six partial differential equations was obtained. It has the form of the canonical hyperbolic system (see [6])

$$\frac{\partial^2\Phi}{\partial s_0\partial t} + \mathrm{A}\frac{\partial\Phi}{\partial t} + \frac{1}{\tau}\left\{\frac{\partial\Phi}{\partial s_0} + \mathrm{B}\right\} = 0. \tag{3.14}$$

The components of Φ are functions of the spatial coordinate s_0 and the time t; they are determined by the relation

$$\Phi = [\gamma, \varepsilon_1, k_1, \varphi, X, Z]^T . \tag{3.15}$$

The matrices A and B of size 6×6 and 6×1 are functions of the components of Φ, which, in turn, are functions of the time t and the coordinate s_0.

The static problem is described by Eq. (3.9), (3.11) and (3.13) for the zero velocities $\dot{\varepsilon}_1 = \dot{\varepsilon}_2 = \dot{k}_1 = \dot{k}_2 = \dot{\gamma} = 0$. Therefore, the static problem is described by the equations in braces:

$$\frac{\partial \Phi}{\partial s_0} + \mathrm{B} = 0.$$

Its solution is obtain from system (3.14) asymptotically when $\dot{\gamma}$, $\dot{\varepsilon}_1$, $\dot{k}_1\dot{\varphi}$, $\dot{x}$, $\dot{y}$ tend to zero.

This problem was solved using the step-by-step method. For every set of load values p_φ, p_z, the iterative process was carried out until stabilization, i.e., until the velocity $|\partial\Phi/\partial t|$ became less than a prescribed value determining the computation error.

The velocities of the unknowns can be approximately replaced with the ratio of the increment $\Delta\Phi_l = \Phi(t_l) - \Phi(t_{l-1})$ to the increment of time $\Delta t_l = t_l - t_{l-1}$, $l = 1, \ldots, N_t$, $t_0 = 0$:

$$\left.\frac{\partial \Phi}{\partial t}\right|_{t=t_l} \approx \frac{\Delta\Phi_l}{\Delta t_l}, \tag{3.16}$$

here, N_t is the number of time steps needed to obtain the solution. Then, system (3.14) can be reduced to the form

$$\frac{d\Delta\Phi_l}{ds_0} + \mathrm{A}|_{t=t_l}\,\Delta\Phi_l + \frac{\Delta t_l}{\tau}\left\{\frac{\partial \Phi}{\partial s_0} + \mathrm{B}\right\}\Big|_{t=t_l} = 0. \tag{3.17}$$

With respect to t, we used a first-order finite difference scheme with a variable step size. In this case, the components of $\Phi(t_l, s_0)$ from the preceding time level must be stored in computer memory for certain points. The procedure of numerical solution is described in [6] in more detail.

The choice of the viscosity parameter τ is not a problem because we can use the dimensionless time $T = t/\tau$, and then the solution is a function of T.

4 Determining the intraocular pressure

We have already mentioned in Section 1.2 that the cornea and the sclera affect each other due to the flow of fluid inside the eye. It follows from the incompressibility condition that

$$V_r + V_s = V_r^0 + V_s^0, \tag{4.1}$$

where V_r and V_s are the volumes of the fluid between the surfaces of the cornea and sclera and the plane passing through their common circle. As before, the subscript 0 marks the quantities corresponding to the state when the pressure in the air jet is zero.

The calculation procedure was as follows. It was assumed that the original shape of the eye is defined when there is neither external nor internal pressure. Then the internal pressure in the eye was increased up to a certain value p_i, while the external pressure was still zero: $p_e = 0$; the eyes shape and the initial volume $V_r^0 + V_s^0$ were determined.

Then, the pressure p_e in the jet was increased step-by-step beginning from zero, and the shape of the cornea was determined. For each p_e , the value of p_i ensuring the fulfillment of condition (4.1) was determined. This value is exactly the intraocular pressure when the pressure in the jet is p_e.

5 Calculation of the optical fluxes in the system

To determine the dependence between the external pressure and the reflected light flux that hits the photo-detector, we should consider the passage of the light rays between the light source and the photo-detector (see fig. 3). The scheme of the optical system is shown in fig. 1. The illuminated area on the cornea is denoted by S, and the area that reflects the light falling on the photo-detector is denoted by Q.

To calculate the passage of the incident and the reflected light beams, we put a grid on the cornea surface that is finer near the center (see [8]). First, each point of the grid was connected by a straight line with the source and it was checked if the ray passes through the aperture A. Thus, we constructed the illuminated region of the cornea S (see fig. 1). From these conditions, we constructed the reflected ray and checked if it hits the photo-detector. As a result, we obtained the region Q. This region can be multiply connected.

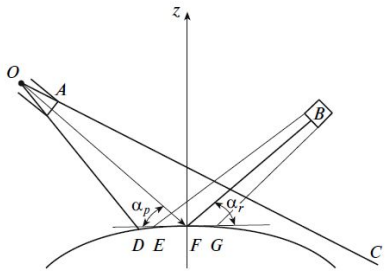


Figure 3. The passage of the light rays between the light source and the photo-detector

6 Deformation of the cornea

Since we solved the problem numerically, the shapes of the cornea and the sclera could be defined by coordinates of some of their points. However, there is not enough data to do so; for that reason, we assumed that their shapes were sphere segments. In our calculations, we used the following parameters: the cornea radius was 0.008 m, the height of the cornea segment was 0.00493 m, the radius of the base was 0.00739 m, the sclera radius was 0.012 m, the Poisson coefficients were 0.45, the modula of elasticity were, respectively, $1.2 \cdot 10^6$ Pa and $6.0 \cdot 10^6$ Pa, the cornea thickness was varied between 0.00045 m, 0.00055 m, and 0.00065 m, and the sclera thickness was the constant equal to 0.001 m. The radius of the region where the pressure was applied was 0.0015 m.

The calculations were performed for the values of the intraocular pressure from p_i=15 to p_i=50 mmHg with the interval of 5 mmHg.

Fig. 4 shows the shapes of the regions in the central part of the cornea corresponding to the unlit part (black), the illuminated part that reflects the light not hitting the photo-detector (gray), and the illuminated part that reflects the light hitting the photo-detector (white).

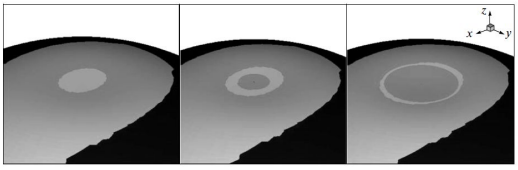


Figure 4. The shapes of the regions in the central part of the cornea

It is seen that the light is first reflected from the center of the cornea; then, as the cornea deforms, it is reflected only from its periphery. This is a very important result

because no experiments on investigating the shape of the region for the reflected light were earlier carried out, while this is necessary for the correct interpretation of measurement results.

Fig. 5, 6 illustrates the dependences between the pressure in the air jet p_e (in mmHg) and the area S (sq. mm) of the region Q that reflects the light hitting the photo-detector for three values of the cornea thickness 0.45 mm, 0.55 mm, and 0.65 mm for the eight values of p_i.

It is seen that the maximum of the dependence $S - p_e$ moves to the right as p_i increases. This fact can be used to interpret the measurement results. The results obtained in this paper provide a new look at the process of measuring the intraocular pressure.

For more details see [8].

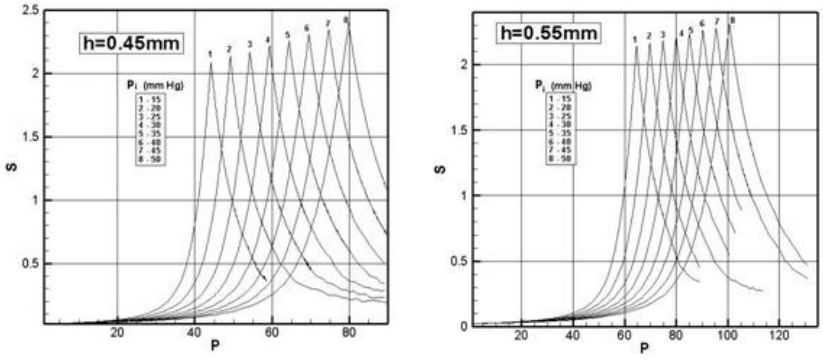


Figure 5. The dependence between the pressure in the air jet p_e (in mHg) and the area S (sq. mm) of the region Q for the cornea thickness 0.45 and 0.55 mm

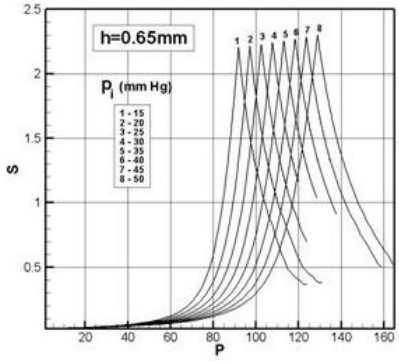


Figure 6. The dependence between the pressure in the air jet p_e (in mmHg) and the area S (sq. mm) of the region Q for the cornea thickness 0.65 mm

References

1. D. Lusce and D. Taylor, Reichert Ocular Response Analyzer Measures Corneal Biomechanical Properties and IOP, Reichert Ophthalmic Instruments (2006).
2. S. M. Bauer, G. A. Lyubimov, and P. E. Tovstik, Mathematical Modeling of Maklakoff's Method for Measuring the Intraocular Pressure, Izv. Ross. Akad. Nauk, Mekh. Zhidkosti Gaza, No. 1, 24-39 (2005) [Fluid Dynamics, 40 (1), 20–33 (2005)].
3. S. M. Bauer, "On the Applanation Methods for Measuring Intraocular Pressure, in Proc. of the Workshop on Computer Methods in Mechanics of Continua (S Peterburg Gos. Univ., Saint Petersburg, 2007), pp. 3-19 [in Russian].
4. V. L. Yakushev, R. R. Khusainov, and V. R. Tsibul'skii, Combined Deformation of Eye Cornea and Sclera with Account for the Flow of Fluid, in Proc. of the Int. Conf. on Nonequilibrium Processes in Nozzles and Jets (NRNJ'2008), Alushta, 2008 (Iosc. Aviatsionnyi Institut, Moscow, 2008), pp. 449-451 [in Russian].
5. V. L. Yakushev, V. R. Tsibul'skii, and R. R. Khusainov, Mathematical Modeling of Nonlinear Deformation of the Eye Cornea and Sclera, in Abstracts of the IV All-Russia Conference on Important Problems of Applied Mathematics and Mechanics, Abrau-Dyurso, 2008, pp. 75-76.
6. V. L. Yakushev, Nonlinear Deformations and Stability of Thin Shells (Nauka, Moscow, 2004) [in Russian].
7. V. L. Yakushev, V. R. Tsibul'skii, and R. R. Khusainov, Numerical Simulation of Light Fluxes in the Measurements of Intraocular Pressure Using an Optical

Method, Vestn. Kibern., No. 9, 74-84 (2010) [in Russian].
8. R. R. Khusainov, V. R. Tsibul'skii, and V. L. Yakushev. Simulation of Eye Deformation in the Measurement of Intraocular Pressure. Computational Mathematics and Mathematical Physics, 2011, Vol. 51, No. 2, pp. 326-338. 2011.

V. L. Yakushev
Institute for Computer Aided Design, Russian Academy of Sciences, Russia, 123056, Moscow, 19/18, 2-nd Brestskaja st., +7 (499) 2500892, yakushev@icad.org.ru

V. Nonlinear analysis, extremal problems, approximation theory and applications

V.1. Optimization theory, variational analysis and nonlinear analysis

(Sessions organizers: A.V. Arutyunov, B.S. Mordukhovich, L.F. Pereira)

SET-VALUED MAPPINGS IN GAME DYNAMIC PROBLEMS

A. A. Chikrii

Key words: conflict-controlled process, Pontryagin's condition, set-valued mapping, superpositional measurability, guaranteed time

AMS Mathematics Subject Classification: 49N70

Abstract. The paper is devoted to quasilinear conflict-controlled processes with a cylindrical terminal set. A specific feature is that, instead of a dynamical system, we start with representation of a solution in a form that allows one to include an additive term with the initial data and a control unit. This makes it possible to consider a broad spectrum of dynamic processes in a unified scheme. Our study is based on the method of resolving functions. We obtain sufficient conditions for the solvability of the pursuit problem at a certain guaranteed time in the class of strategies that use information on the behavior of the opponent in the past, as well as in the class of stroboscopic strategies. We also find conditions under which information on the prehistory of the evader does not matter. The guaranteed times of various schemes of the resolving function method are compared with the guaranteed time of Pontryagin's first direct method.

1 Introduction

In the theory of dynamic games, along with the Krasovskii's extremal aiming principle [1,2] Pontryagin–Pshenichnyi backward procedures [3,4], and Isaacs ideology [5] concerning the basic equation in the theory of differential games, there exist efficient methods that can apparently be classified into a separate group. These are Pontryagin's first direct method [3,6] and the method of resolving functions [7]. They share a common principle of constructing the control of the first player on the basis of the Filippov-Castaing measurable choice theorem [8]. The method of resolving functions, being in this sense a development of Pontryagin's first direct method, stems from the solution of game problems of evading a group of pursuers [9]. In this way, Pshenichnyi formulated in [10] necessary and sufficient conditions for the solvability of a pursuit problem in which one group of objects aims to surround another group of objects with simple motions and equal maximal velocities. This gave impetus to the development of methods for solving group pursuit problems [7,11,13] on the basis of the conception of resolving functions. Later, these studies have been extended to the following problems: problems with phase constraints regarded as

static pursuers [7, 13], game problems of successive pursuit (salesman-type problems) [7], pursuit-evasion games with interacting groups [7, 11–13]. The method has been applied to various dynamic processes, including systems with fractional derivatives [14].

The apparatus of resolving functions, which, as a rule, are represented in examples by large positive roots of quadratic equations, turned out to be very convenient and universal for solving specific problems.

Note also that in the general case resolving functions can be expressed in terms of inverse Minkowski functionals [7] of some set-valued mappings; this provides additional possibilities for the study.

An attractive feature of the method of resolving functions is that it fully justifies the classical rule of parallel approach and allows one to efficiently apply the modern technique of set-valued mappings and their selections [8, 15] when validating game constructions and obtaining meaningful results on the basis of these constructions.

When solving specific game problems with the use of resolving functions, it is important to provide a rigorous substantiation of the method. In the proofs of the corresponding statements, a key role is played by problems related to the properties of special set-valued mappings and their selections. In particular, the $(L \times B)$-measurability of closed-valued mappings [15] related to the process and the compositional measurability of their selections are decisive factors in substantiating the scheme of the method and constructing the control of the first player. These questions in one-to-one games, sufficient termination conditions, and a comparison of the guaranteed times of various schemes of the method constitute the content of this paper. The paper is related to the publications [1, 2, 4] and continues the studies of [7].

2 Scheme of the method

Consider a conflict-controlled process whose evolution is described by the equality

$$z(t) = g(t) + \int_0^t \Omega(t,\tau)\varphi\left(u(\tau), v(\tau)\right) d\tau, \quad t \geqslant 0. \tag{2.1}$$

Here $z(t) \in R^n$, the function $g(t)$, $g : R_+ \to R^n$, $R_+ = \{t : t \geqslant 0\}$, is Lebesgue measurable and bounded for $t > 0$, and the matrix function $\Omega(t,\tau)$, $t \geqslant \tau \geqslant 0$, is measurable in t and integrable in τ for every $t \in R_+$. The control unit is defined by the function $\varphi\left(u, v\right), \varphi : U \times V \to R^n$, which is assumed to be jointly continuous

in its variables on the direct product of nonempty compacts sets U and V, i.e., $U \in K(R^m)$ and $V \in K(R^l)$, where m, l and n positive integers.

The controls of the players, $u(\tau)$, $u : R_+ \to U$, and $v(\tau)$, $v : R_+ \to V$, are measurable functions of time.

In addition to the process (2.1), consider a terminal cylindrical set M^* of the form

$$M^* = M_0 + M, \tag{2.2}$$

where M_0 is a linear subspace in R^n and $M \in K(L)$, with L being the orthogonal complement of M_0 in R^n.

The goals of the first (u) and second (v) players are opposite. The first player tries to bring the trajectory of process (2.1) to the terminal set in the shortest time, whereas the second player tries to maximally put off the instant when the trajectory reaches the set M^*, or even avoid this meeting at all.

The representation of a solution to a dynamical system in the form (2.1) allows one to consider a wide range of functional-differential systems operating under conflict conditions within a unified scheme, including systems of integral, integro-differential, and difference-differential equations, as well as systems of equations with classical Riemann-Liouville fractional derivatives, regularized fractional derivatives in the sense of Caputo, and the Miller–Ross sequential fractional derivatives [14]. A similar representation in the discrete situation makes it possible to analyze multistep processes and impulse systems.

A specific form of the function $g(t)$ and the matrix function $\Omega(t, \tau$ determines the type of a conflict-controlled process.

Let us take the side of the first player and assume that the opponent chooses an arbitrary V-valued measurable function as a control. If the game (2.1), (2.2) occurs on an interval $[0, T]$, then we assume that the first player decides on its control at time t depending on the information about $g(T)$ and $v_t(\cdot)$; i.e., the control of the first player is either a measurable function

$$u(t) = u\left(g\left(T\right), v_t\left(\cdot\right)\right), \quad t \in [0, T], \quad u(t) \in U, \tag{2.3}$$

where $v_t(\cdot) = \{v(s) : s \in [0, t]\}$ is the prehistory of the control of the second player up to instant t, or a countercontrol

$$u(t) = u\left(g\left(T\right), v\left(t\right)\right), \quad t \in [0, T], \quad u(t) \in U, \tag{2.4}$$

The aim of this study is to establish sufficient conditions for the termination of a pursuit game in the guaranteed time of an analog of the method of resolving functions [7] in the class of stroboscopic strategies for the conflict-controlled process

(2.1), (2.2), and to compare them with analogous conditions for Pontryagin's first direct method [3,6].

Denote by π the orthogonal projection from R^n to L. Setting $\varphi(U,v) = \{\varphi(u,v) : u \in U\}$, consider the set-valued mappings

$$W(t,\tau,v) = \pi\Omega(t,\tau)\varphi(U,v), W(t,\tau) = \bigcap_{v\in V} W(t,\tau,v),$$

on the sets $\Delta \times V$ and Δ respectively, where $\Delta = \{(t,\tau) : 0 \leqslant \tau \leqslant t < \infty\}$.

Pontryagin condition. The set-valued mapping $W(t,\tau)$ takes nonempty values on the set Δ.

In view of the properties of the parameters of the conflict-controlled process (2.1), the mapping $\varphi(U,v)$, $v \in V$, is continuous in the Hausdorff metric. Therefore, taking into account the assumptions about the matrix function $\Omega(t,\tau)$, we can conclude that for any fixed $t > 0$ the mapping $W(t,\tau,v)$ is a set-valued mapping that is measurable in τ on the interval $[0,t]$ and closed in v, $v \in V$. Then [15] the mapping $W(t,\tau)$ is a closed-valued mapping measurable in $\tau \in [0,t]$.

Denote by $P(R^n)$ the family of nonempty closed sets of the space R^n. Then it is obvious that $W(t,\tau,v) : Delta \times V \to P(R^n)$ and $W(t,\tau) : Delta \to P(R^n)$. In this case, the set-valued mappings $W(t,\tau,v)$ and $W(t,\tau)$ that are measurable in τ are said to be normal.

The Pontryagin condition and the measurable selection theorem [15] imply that for any $t \geqslant 0$ these exists at least one selection $\gamma(t,\tau)$ measurable in τ such that $\gamma(t,\tau) \in W(t,\tau)$, $(t,\tau) \in \Delta$. Denote the set of such selections by Γ_t and introduce a function

$$\xi(t, g(t), \gamma(t,\cdot)) = \pi g(t) + \int_0^t \gamma(t,\tau)\, d\tau, \tag{2.5}$$

where $\gamma(\cdot,\cdot) \in \Gamma = \cup_{t\geqslant 0}\Gamma_t$. By the assumptions, the selection $\gamma(t,\tau)$ in integrable in τ, $\tau \in [0,t]$, for any $t > 0$.

Consider the set-valued mapping

$$\mathfrak{A}(t,\tau,v) = \{\alpha \geqslant 0 : [W(t,\tau,v) - \gamma(t,\tau)] \cap \alpha[M - \xi(t,g(t),\gamma(t,\cdot))] \neq \emptyset\}, \tag{2.6}$$

$$\mathfrak{A} : \Delta \times V \to 2^{R_+},$$

and its support function in the direction $+1$, $\alpha(t,\tau,v) = \sup\{\alpha : \alpha \in \mathfrak{A}(t,\tau,v)\}$, $(t,\tau) \in \Delta$, $v \in V$. This function is called a resolving function [7].

In order to stress the role of the Minkowski functionals and their inverses in the scheme of the method, we express the function $\alpha(t,\tau,v)$ in a different form. To

this end, we introduce a functions

$$\alpha_X(p) = \sup\{\alpha \geqslant 0 : \alpha p \in X\}, \quad p \in R^n, \quad X \in P(R^n), \quad 0 \in X.$$

Then $\alpha(t,\tau,v) = \sup_{m\in M} \alpha_{W(t,\tau,v)-\gamma(t,\tau)}(m - \xi(t, g(t), \gamma(t,\cdot)))$.

Due to the Pontryagin condition, the set-valued mapping $\mathfrak{A}(t,\tau,v)$ on the set $\Delta \times V$ has a nonempty closed image. Note also that if $\xi(t, g(t), \gamma(t,\cdot)) \in M$, then $\mathfrak{A}(t,\tau,v) = [0,+\infty)$ and, hence $\alpha(t,\tau,v) = +\infty$ for all $\tau \in [0,t]$ and $v \in V$.

Taking into account the properties of the parameters of the conflict-controlled process (2.1), (2.2) and applying the characterization and inverse image theorems, we can show that the set-valued mapping $\mathfrak{A}(t,\tau,v)$ is jointly $(L \times B)$- measurable [15] in the variables τ, v, $\tau \in [0,t]$, $v \in V$; the resolving function $\alpha(t,\tau,v)$ is jointly $(L \times B)$-measurable in the variables τ, v, by the theorem on the support function [16] for $\xi(t, g(t), \gamma(t,\cdot)) \notin M$.

Consider the set

$$T(g(\cdot), \gamma(\cdot,\cdot)) = \left\{ t \geqslant 0 : \inf_{v(\cdot)} \int_0^t \alpha(t,\tau,v(\tau))\, d\tau \geqslant 1 \right\}. \tag{2.7}$$

Since the function $\alpha(t,\tau,v)$ is $(L \times B)$-measurable in τ and v, it is compositionally measurable; i.e., $\alpha(t,\tau,v(\tau)))$ is a measurable function for any measurable function $v(\tau)$, $v(\tau) \in V$.

If $\xi(t, g(t), \gamma(t,\cdot)) \in M$, then $\alpha(t,\tau,v) = +\infty$ for $\tau \in [0,t]$ and $v \in V$; in this case in is natural to set the value of the integral in relation (2.7) equal to $+\infty$, and the related inequality holds automatically. If the inequality in braces in (2.7) fails for all $t > 0$, we set $T(g(\cdot), \gamma(\cdot,\cdot)) = \emptyset$.

Theorem 1. *Suppose that the conflict-controlled process (2.1), (2.2) satisfies the Pontryagin condition, the set M is convex, and the inclusion $T \in T(g(\cdot), \gamma(\cdot,\cdot)) \neq \emptyset$ holds for a given function $g(\cdot)$ and some selection $\gamma(\cdot,\cdot) \in \Gamma$. Then a trajectory of process (2.1) can be brought to the terminal set (2.2) at time T by a control of the form (2.3).*

Proof. Let $v(\tau)$, $v : [0,T] \to V$, be an arbitrary measurable function.

Suppose $\xi(T, g(T), \gamma(T,\cdot)) \overline{\in} M$. Introduce a test function $h(t) = 1 - \int_0^t \alpha(T,\tau,v(\tau))\, d\tau$, $t \in [0,T]$. As pointed out above, the function $\alpha(T,\tau,v)$ is $(L \times B)$-measurable in (τ, v) for $\tau \in [0,T]$ and $v \in V$; hence it is compositionally measurable, and $\alpha(T,\tau,v(\tau))$ is a measurable function of τ. This implies that $h(t)$ is absolutely continuous and, hence, continuous; it does not increase ($\alpha(T,\tau,v) \geqslant 0$

by definition), and $h(0) = 1$. Since, in addition, $h(T) \leqslant 0$, there exists a time instant t_*, $t_* = t(v(\cdot))$, $t_* \in (0, T]$, such that $h(t_*) = 0$.

In what follows, we will call the time intervals $[0, t_*)$ active and passive intervals, respectively. Let us describe the method of control of the first player on each of these intervals. To this end, consider a compact-valued mapping

$$U(\tau, v) = \{u \in U : \ \pi\Omega(T, \tau)\varphi(u, v) - \gamma(T, \tau) \in \\ \in \alpha(T, \tau, v)[M - \xi(T, g(T), \gamma(T, \cdot))]\}, \quad \tau \in [0, T], \quad v \in V. \quad (2.8)$$

By the inverse image theorem, this mapping is $(L \times B)$-measurable [15]; hence, according to the measurable selection theorem [15], the set-valued mapping $U(\tau, v)$ has at least one $(L \times B)$-measurable selection $u(\tau, v)$ that is compositionally measurable. Denote $u(\tau) = u(\tau, v(\tau))$. Set the control of the first player on the active interval equal to $u(\tau)$.

Consider the passive time interval $[t_*, T)$. For $\tau \in [t_*, T)$ and $v \in V$, we set the resolving function to be $\alpha(T, \tau, v) \equiv 0$ in expression (2.8). The yields a set-valued mapping

$$U_0(\tau, v) = \{u \in U : \pi\Omega(T, \tau)\varphi(u, v) - \gamma(T, \tau) = 0\}. \quad (2.9)$$

Just as in the previous case, it follows from the measurable selection theorem that the $(L \times B)$-measurable closed-valued mapping $U_0(\tau, v)$ has an $(L \times B)$-measurable selection. Denote this selection by $u_0(\tau, v)$ and choose the control of the pursuer on the passive interval to be $u_0(\tau) = u_0(\tau, v(\tau))$.

When $\xi(T, g(T), \gamma(T, \cdot)) \in M$ we choose the control of the first player on the whole interval $[0, T]$ to be $u_0(\tau) = u_0(\tau, v(\tau))$, where $u_0(\tau, v)$ is a $(L \times B)$-measurable selection of the set-valued mapping $U_0(\tau, v)$. Let us schow that when the control of the first player on the active and passive intervals is chosen in accordance with the above rules, the trajectory of system (2.1) is brought to the terminal set at time instant T for any admissible controls of the second player.

Consider first the case of $\xi(T, g(T), \gamma(T, \cdot)) \notin M$. From (2.1) we find

$$\pi z(T) = \pi g(T) + \int_0^T \pi\Omega(T, \tau)\varphi(u(\tau), v(\tau))\, d\tau. \quad (2.10)$$

Using relations (2.8) and (2.9), we obtain the following inclusion from (2.10):

$$\pi z\left(T\right) \in \xi\left(T, g\left(T\right), \gamma\left(T, \cdot\right)\right) \left[1 - \int_0^{t_*} \alpha\left(T, \tau, v\left(\tau\right)\right) d\tau\right] + \int_0^{t_*} \alpha\left(T, \tau, v\left(\tau\right)\right) M d\tau. \tag{2.11}$$

Since M is a convex compact set, $\alpha\left(T, \tau, v\left(\tau\right)\right)$ is a nonnegative function for $\tau \in [0, t_*)$, and $\int_0^{t_*} \alpha\left(T, \tau, v\left(\tau\right)\right) d\tau = 1$, it follows that $\int_0^{t_*} \alpha\left(T, \tau, v\left(\tau\right)\right) M d\tau = M$. Taking into account these facts, from (2.11) we obtain $\pi z\left(T\right) \in M$, or $z\left(T\right) \in M^*$.

Suppose $\xi\left(T, g\left(T\right), \gamma\left(T, \cdot\right)\right) \in M$. Then, taking into account (2.9), from equality (2.10) we find that $\pi z\left(T\right) = \xi\left(T, g(T), \gamma(T, \cdot)\right) \in M$.

The theorem is proved.

3 Modification of the method and sufficient conditions

The results of the previous section lead to a certain modification of the scheme of the resolving function method. In a sense, this modification gives an exhaustive answer to the question of solvability of pursuit game problems in the class of stroboscopic strategies.

Consider the set-valued mapping

$$\mathfrak{A}(t, \tau) = \bigcap_{v \in V} \mathfrak{A}(t, \tau, v), t \geqslant \tau \geqslant 0,$$

which has a nonempty image because at least $0 \in \mathfrak{A}(t, \tau, v)$ for $t \geqslant \tau \geqslant 0$ and $v \in V$, and its support function in the direction $+1$,

$$\alpha(t, \tau) = \sup\left\{\alpha \geqslant 0 : \ \alpha \in \mathfrak{A}(t, \tau)\right\}.$$

If $\xi\left(t, g(t), \gamma\left(t, \cdot\right)\right) \notin M$, then the mapping $\mathfrak{A}(t, \tau)$ is closed-valued and measurable in τ, $\tau \in [0, t]$; hence, by the support function theorem, the function $\alpha(t, \tau)$ is also measurable in τ.

Introduce the set

$$\Theta\left(g(\cdot), \gamma(\cdot, \cdot)\right) = \left\{t \geqslant 0 : \int_0^t \alpha(t, \tau) d\tau \geqslant 1\right\}. \tag{3.1}$$

If $\xi\left(t, g(t), \gamma\left(t, \cdot\right)\right) \in M$ for some $t > 0$, then obviously $\mathfrak{A}(t, \tau) = [0, +\infty)$, and $\alpha(t, \tau) \equiv +\infty$ for $\tau \in [0, t]$;therefore, in this case it is natural to set the value of the

integral in (3.1) to be $+\infty$, while the corresponding inequality holds automatically. If the inequality in (3.1) does not hold for any $t > 0$, then we set $\Theta\left(g(\cdot), \gamma(\cdot,\cdot)\right) = \emptyset$.

Theorem 2. *Suppose that the conflict-controlled process (2.1), (2.2) satisfies the Pontryagin condition, $M = co\, M$, and the inclusion $\Theta \in \Theta\left(g(\cdot), \gamma(\cdot,\cdot)\right) \neq \emptyset$ holds for a given function $g(\cdot)$ and some selection $\gamma\left(\cdot,\cdot\right) \in \Gamma$. Then a trajectory of the process (2.1) can be brought to the terminal set (2.2) at time Θ by a control of the form (2.4).*

Corollary 1. *If the conflict-controlled process (2.1), (2.2) satisfies the Pontryagin condition, $M = co\, M$, the sets $T\left(g(\cdot), \gamma(\cdot,\cdot)\right)$ and $\Theta\left(g(\cdot), \gamma(\cdot,\cdot)\right)$ are nonempty for a given function $g(\cdot)$ and some selection $\gamma(\cdot,\cdot) \in \Gamma$ and mapping $\mathfrak{A}\left(T, \tau, v\right) = [0, (T, \tau, v)]$, $T \geqslant \tau \geqslant 0$, $v \in V$, then $T\left(g(\cdot), \gamma(\cdot,\cdot)\right) = \Theta\left(g(\cdot), \gamma(\cdot,\cdot)\right)$.*

In the general case, the following inclusion is always valid: $Q\left(g(\cdot), \gamma(\cdot,\cdot)\right) \subset T\left(g(\cdot), \gamma(\cdot,\cdot)\right)$.

In what follows, we assume that there exist minimal elements $T_0\left(g(\cdot), \gamma(\cdot,\cdot)\right)$ and $\Theta_0\left(g(\cdot), \gamma(\cdot,\cdot)\right)$ in the number sets $T\left(g(\cdot), \gamma(\cdot,\cdot)\right)$ and $\Theta\left(g(\cdot), \gamma(\cdot,\cdot)\right)$, respectively.

4 Pontryagin's first direct method

Pontryagin's first direct method is well known from the publications devoted to the theory of diffential games [3, 6, 7]. This method gives sufficient conditions for the termination of a differential pursuit game in a guaranteed time in the class of stroboscopic strategies. We mean the proof of Gusyatnikov and Nikol'skii, which uses the Filippov-Castaing measurable choice theorem to construct a control. In connection with the results obtained above, it is expedient to compare them with the results of Pontryagin's first direct method.

Consider the Pontryagin function for the conflict-controlled process (2.1), (2.2),

$$P\left(g\left(\cdot\right)\right) = \inf\left\{ t \geqslant 0 : \quad \pi g\left(t\right) \in M - \int_0^t W\left(t, \tau\right)\, d\tau \right\}. \tag{4.1}$$

Here, as before, the integral of a set-valued mapping is the Aumann integral [15]. If the inclusion in braces does not hold for any $t \geqslant 0$, then we set $P\left(g\left(\cdot\right)\right) = +\infty$.

Theorem 3. *Suppose that the conflict-controlled process (2.1), (2.2) satisfies the Pontryagin condition, the greatest lower bound in (4.1) is attainable, and $P = P\left(g\left(\cdot\right)\right) < +\infty$. Then a trajectory of the process (2.1) can be brought to the terminal set (2.2) at time P by a control of the form (2.4).*

Theorem 3, which generalizes the first direct method to conflict-controlled processes of the form (2.1), entails the following corollaries.

Corollary 2. *Suppose that the conflict-controlled process (2.1), (2.2) satisfies the Pontryagin condition. Then* $\pi g(t) \in M - \int_0^t W(t,\tau)\, d\tau$, $t \geqslant 0$, *if and only if there exists a selection* $\gamma(t,\cdot) \in \Gamma_t$, *such that*

$$\xi(t, g(t), \gamma(t,\cdot)) \in M. \tag{4.2}$$

Note that in the scheme of the method of resolving functions the fulfillment of inclusion (4.2) leads to the degeneration of the above-mentioned functions; i.e., the values of these functions become $+\infty$. This situation falls within the scope of Pontryagin's first direct method, and the game in this case can be terminated in the guaranteed time of Pontryagin's first direct method in the class of stroboscopic strategies without any assumptions about the parameters of the conflict-controlled process (2.1), (2.2), except, naturally,the Pontryagin conditions.

Corollary 3. *Suppose that the conflict-controlled process (2.1), (2.2) satisfies the Pontryagin condition. Then there exists a selection* $\gamma(\cdot,\cdot)$ *such that* $T_0(g(\cdot), \gamma(\cdot,\cdot)) \leqslant P(g(\cdot))$ *for any measurable function* $g(t)$ *that is bounded for* $t > 0$.

Corollary 4. *Suppose that the conflict-controlled processes (2.1), (2.2) satisfies the Pontryagin condition. Then there exists a selection* $\gamma(\cdot,\cdot) \in \Gamma$, *such that* $\Theta_0(g(\cdot), \gamma(\cdot,\cdot)) \leqslant P(g(\cdot))$ *for any measurable function* $g(t)$ *that is bounded for* $t > 0$.

5 Functional form of the first direct method. Comparison of guaranteed times

Let us express the termination time of the game (2.1), (2.2) provided by Pontryagin's first direct method (4.1) in terms of resolving functions. To this end, consider the set-valued mapping

$$B(t,\tau) = \left\{\beta \geqslant 0 : [W(t,\tau) - \gamma(t,\tau)] \bigcap \beta\, [M - \xi(t, g(t), \gamma(t,\cdot))] \neq \emptyset\right\} \tag{5.1}$$

and its support function in the direction $+1$,

$$\beta(t,\tau) = \sup\{\beta : \quad \beta \in B(t,\tau)\}, t \geqslant \tau \geqslant 0. \tag{5.2}$$

Here $\gamma(t,\tau)$ is a τ-measurable selection of the set-valued mapping $W(t,\tau)$, introduced earlier and the function $\xi(t, g(t), \gamma(t,\cdot))$ is defined by (2.5).

If $\xi(t, g(t), \gamma(t,\cdot)) \bar{\in} M$, then, according to the characterization and inverse image theorems [15], the mapping $B(t,\tau)$ is measurable and closed-valued in τ, $\tau \in [0,t]$. Hence, by the support function theorem [16], the function $\beta(t,\tau)$ is

measurable in τ. If $\xi\ (t, g\ (t)\,,\ \gamma\ (t, \cdot)) \in M$, then $B\ (t, \tau) = [0, +\infty)$, and $\beta\ (t, \tau) = +\infty$ for all $\tau \in [0, t]$.

Introduce a time function

$$P\ (g\ (\cdot)\,,\ \gamma\ (\cdot, \cdot)) = \inf \left\{ t \geqslant 0 : \int_0^t \beta\ (t, \tau)\ d\tau \geqslant 1 \right\}, \tag{5.3}$$

which is assumed to be $+\infty$ if the inequality in braces fails for all $t \geqslant 0$.

Theorem 4. *Suppose that the conflict-controlled process (2.1), (2.2) satisfies the Pontryagin condition, $M = \operatorname{co} M$, and the greatest lower bound in (5.3) is attainable for a given function $g\ (\cdot)$ and a selection $\gamma\ (\cdot, \cdot) \in \Gamma$ with $P = P\ (g\ (\cdot)\,,\ \gamma\ (\cdot, \cdot)) < +\infty$. Then a trajectory of the process (2.1) can be brought to the terminal set (2.2) at time P by a certain countercontrol.*

Theorem 5. *Suppose that the conflict-controlled process (2.1), (2.2) satisfies the Pontryagin condition, $M = \operatorname{co} M$, and the greatest lower bound in (5.3) is attainable. Then, for any function $g\ (t)$ that is measurable and bounded for $t > 0$, we have $\min\limits_{\gamma(\cdot,\cdot)\in\Gamma} P\ (g\ (\cdot)\,,\ \gamma\ (\cdot, \cdot)) = P\ (g\ (\cdot))$.*

For simpler dynamics, this result was proved in [7]. The proof of the present statement is completely analogous.

Thes cheme (5.1)–(5.3) is called [7] a functional form of Pontryagin's first direct method.

Let us establish a relationship between the time functions $T_0\ (g\ (\cdot)\,,\ \gamma\ (\cdot, \cdot))$ and $P\ (g\ (\cdot))$ for the method of resolving functions and Pontryagin's first direct method.

Theorem 6. *Suppose that the conflict-controlled process (2.1), (2.2) satisfies the Pontryagin condition and Condition 1, the terminal set M^* is an affine manifold, i.e., $M = \{m\}$ is a point, and the greatest lower bounds with respect to t in (2.7) and (5.3) are attainable. Then*

$$\min_{\gamma(\cdot,\cdot)\in\Gamma} T_0\ (g\ (\cdot)\,,\ \gamma\ (\cdot, \cdot)) = P\ (g\ (\cdot))$$

for all functions $g\,(t)$ that are measurable and bounded for $t > 0$.

Corollary 5. *Suppose that the conflict-controlled process (2.1), (2.2) satisfies the Pontryagin condition, the terminal set M^* is an affine manifold ($M = \{m\}$), and the greatest lower bounds with respect to t in (3.1) and (5.3) are attainable. Then, for all functions $g\ (t)$ that are measurable and bounded for $t > 0$, we have*

$$\min_{\gamma(\cdot,\cdot)\in\Gamma} \Theta_0\ (g\ (\cdot)\,,\ \gamma\ (\cdot, \cdot)) = P\ (g\ (\cdot))\,.$$

6 Scheme of the method of resolving functions with fixed aiming points in the terminal set

As is clear from the definition of the set-valued mapping $\mathfrak{A}(t,\tau,v)$, the tangency of the sets in the intersection in (2.6) for $\alpha = \alpha(t,\tau,v)$ occurs in general at different points for different values of the arguments. The tangency points of the left set define the control of the pursuer, while the tangency points of the right set are the points of M at which the motion is aimed. In the general scheme of the method, it is assumed $M = \operatorname{co} M$.

Let us present another form of the method of resolving functions with (time-independent) fixed points of the set M, which is not generally convex.

Let $m \in M$ and $\eta(t,m) = \xi(t,g(t),\gamma(t,\cdot)) - m$, $\gamma(t,\cdot) \in \Gamma_t$, $t \geqslant 0$. Introduce a set-valued mapping

$$\mathfrak{A}(t,\tau,v,m) = \{\alpha \geqslant 0 : \ -\alpha\eta(t,m) \in W(t,\tau,v) - \gamma(t,\tau)\}$$

and its support function in the direction $+1$,

$$\alpha(t,\tau,v,m) = \sup\{\alpha : \ \alpha \in \mathfrak{A}(t,\tau,v,m)\}, \ t \geqslant \tau \geqslant 0, \ v \in V.$$

Since the Pontryagin condition is supposed to hold, we have $\operatorname{dom}\mathfrak{A} = \Delta \times V \times M$. Note that if $\eta(t,m) = 0$, then $\mathfrak{A}(t,\tau,v,m) = [0,+\infty)$ for $\tau \in [0,t]$, $v \in V$, and $m \in M$, and so $\alpha(t,\tau,v,m) \equiv +\infty$.

By virtue of the inverse image theorem, the set-valued mapping $\mathfrak{A}(t,\tau,v,m)$ is $(L \times B)$-measurable in τ, v, $\tau \in [0,t]$, $v \in V$, and by the support function theorem, the resolving function $\alpha(t,\tau,v,m)$ is $(L \times B)$-measurable in τ, v. Consider the function

$$\mathfrak{T}(g(\,\cdot\,),m,\gamma(\,\cdot\,,\cdot\,)) = \inf\left\{t \geqslant 0 : \inf_{v(\cdot)} \int_0^t \alpha(t,\tau,v(\tau),m)\, d\tau \geqslant 1\right\}. \tag{6.1}$$

Condition 1. The function $\mathfrak{T}(g(\,\cdot\,),m,\gamma(\,\cdot\,,\cdot\,))$ is lower semicontinuous in m, $m \in M$.

Then, by the Weierstrass theorem, this function generates a marginal function $\mathfrak{T}(g(\,\cdot\,),\gamma(\,\cdot\,,\cdot\,)) = \min\limits_{m\in M} \mathfrak{T}(g(\,\cdot\,),m,\gamma(\,\cdot\,,\cdot\,))$ and a marginal set-valued mapping

$$\mathfrak{M}(g(\,\cdot\,),\gamma(\,\cdot\,,\cdot\,)) = \{m \in M : \mathfrak{T}(g(\,\cdot\,),\gamma(\,\cdot\,,\cdot\,)) = \mathfrak{T}(g(\,\cdot\,),m,\gamma(\,\cdot\,,\cdot\,))\} \subset M.$$

Note [7] that the time functions can be represented as follows:

$$T_0\left(g(\cdot),\gamma\left(\cdot,\cdot\right)\right)=\inf\left\{t\geqslant 0:\ \inf_{v(\cdot)}\int_0^t\max_{m\in M}\alpha\left(t,\tau,v(\tau),m\right)d\tau\geqslant 1\right\},$$

$$\mathfrak{T}\left(g(\cdot),\gamma\left(\cdot,\cdot\right)\right)=\inf\left\{t\geqslant 0:\ \max_{m\in M}\inf_{v(\cdot)}\int_0^t\alpha\left(t,\tau,v(\tau),m\right)d\tau\geqslant 1\right\},$$

moreover, the relation $\alpha\left(t,\tau,v\right)=\max\limits_{m\in M}\alpha\left(t,\tau,v,m\right)$ holds for $t\geqslant\tau\geqslant 0$, $v\in V$, and $\gamma\left(\cdot,\cdot\right)\in\Gamma$. If $\eta(t,m)=0$, then $\inf\limits_{v\in V}\alpha\left(t,\tau,v,m\right)=+\infty$ for $\tau\in[0,t]$, and it is natural to set the value of the untegral in (6.1) to be equal to $+\infty$; then the corresponding inequality will hold automatically. If the inequality in (6.1) fails for all $t>0$ and $m\in M$, we will assume that $\mathfrak{T}\left(g(\,\cdot\,),\gamma(\,\cdot\,,\cdot\,)\right)=+\infty$.

Theorem 7. *Suppose that the conflict-controlled process (2.1), (2.2) satisfies the Pontryagin condition and Condition 1, and the external greatest lower bound is attainable in (6.1) for a given function $g(\cdot)$ and some selection $\gamma(\,\cdot\,,\cdot\,)\in\Gamma$ such that $\mathfrak{T}=\mathfrak{T}\left(g(\,\cdot\,),\gamma(\,\cdot\,,\cdot\,)\right)<+\infty$. Then the projection $\pi z(t)$ of a trajectory of the process (2.1) can be brought to any point of the set $\mathfrak{M}\left(g(\,\cdot\,),\gamma(\,\cdot\,,\cdot\,)\right)$ at time $\mathfrak{T}$ by a control of the first player prescribed by an appropriate quasistrategy.*

The proof of Theorems 2–7 is conducted on the basis of the method of resolving functions [7] and using ideas of the Theorem 1 proof.

Corollary 6. *Suppose that the conflict-controlled process (2.1), (2.2) satisfies the Pontryagin condition. Then $\pi g(t)\in M-\int_0^t W(t,\tau)d\tau$, $t\geqslant 0$, if and only if there exist a measurable selection $\gamma(t,\cdot)\in\Gamma_t$ and an element $m\in M$ such that $\eta(t,m)=0$.*

Corollary 7. *Suppose that the conflict-controlled process (2.1), (2.2) satisfies the Pontryagin condition. Then*

$$\inf_{\gamma(\cdot,\cdot)\in\Gamma}T_0\left(g(\cdot),\gamma(\cdot,\cdot)\right)\leqslant\inf_{\gamma(\cdot,\cdot)\in\Gamma}\mathfrak{T}\left(g(\cdot),\gamma(\cdot,\cdot)\right)\leqslant P(g(\cdot))$$

for any function $g(t)$ that is measurable and bounded for $t>0$.

Corollary 8. *Suppose that the conflict-controlled process (2.1), (2.2) satisfies the Pontryagin condition, the terminal set M^* is an affine manifold, i.e., $M=\{m\}$ is a point, and the greatest lower bounds with respect to t in (2.7) and (6.1) are attainable. Then, for any measurable function $g(t)$ that is bounded for $t>0$, we*

have

$$\min_{\gamma(\cdot,\cdot)\in\Gamma} T_0\left(g(\cdot),\gamma(\cdot,\cdot)\right) = \min_{\gamma(\cdot,\cdot)\in\Gamma} \mathfrak{T}\left(g(\cdot),\gamma(\cdot,\cdot)\right), t > 0.$$

References

1. N.N. Krasovskii *Game Problems of Encounter of Motions*, Nauka, Moscow, 1970, 420 pages. [in Russian].
2. N.N. Krasovskii, A.I. Subbotin *Game-Theoretical Control Problems*, Springer, New York, 1988, 517 pages.
3. L.S. Pontryagin *Selected Scientific Works*, Nauka, Moscow, 1988, **2**, 576 pages. [in Russian].
4. B.N. Pshenitchnyi, V.V. Ostapenko *Differential Games*, Naukova Dumka, Kiev, 1992, 260 pages. [in Russian].
5. R. Isaacs *Differential Games*, J. Wiley and Sons, New York, 1965, 480 pages.
6. M.S. Nikol'skii *Pontryagin's First Direct Method in Differential Games*, Mosk. Gos. Univ., Moskow, 1984, 64 pages. [in Russian].
7. A.A. Chikrii *Conflict-Controlled Processes*, Kluwer Academic Publishers, Dordrecht, 1997, 424 pages.
8. A.F. Filippov *Differential Equations with Diskontinuous Right-Hand Sides* Kluwer, Dordrecht, 1988, 224 pages.
9. A.A. Chikrii *"The Problem of Avoidance for Controlled Dynamic Objects"*, Int. J. Math. Game Theory Algebra **7** (2-3), 81–94 (1998).
10. B.N. Pshenitchnyi *"Simple Pursuit by Several Objects"*, Kibernetika, No. 3, 1976, Pp. 145–146.
11. N.L. Grigorenko *Mathematical Methods of Control for Several Dynamic Processes*, Mosk. Gos. Univ., Moscow, 1990, 198 pages. [in Russian].
12. A.A. Chikrii *"Differential Games with Several Pursuers"*, in Math Control Theory (PWN-Polish Sci. Publ., 1985), Banach Center Publ., **14**, Pp. 81-107.
13. A.I. Blagodatskikh, N.N. Petrov *Conflict Interaction between Groups of Objects*, Udm. Gos. Univ., Izhevsk, 2009, 266 pages. [in Russian].
14. A.A. Chikrii *"Optimization of Game Interaction of Fractional-Order Controlled Systems"*, Optim. Methods Softw. **23** (1), 2008, Pp. 39–72.
15. J.-P. Aubin, H. Frankovska *Set-Valued Analysis*, Birkhauser, Boston, 1990, 461 pages.
16. B.S. Mordukhovich *Variational Analysis and Generalized Differentiation*, I: Basis Theory; II: Applications (Springer, Berlin, 2006), Grundl. Math. Wiss. **330, 331**, 582 pages, 612 pages.

A. A. Chikrii
V.M. Glushkov Institute of Cybernetics, National Academy of Sciences of Ukraine, pr. akademika Glushkova 40, Kiev, 03680 Ukraine, tel: (044)526-2158, E-mail: chik@insyg.kiev.ua

CONTROLLABILITY OF ABSTRACT DEGENERATE DIFFERENTIAL EQUATION

V. E. Fedorov, B. Shklyar

Key words: controllability, degenerate control system, strongly minimal sequence

AMS Mathematics Subject Classification: 93B05, 34G10

Abstract. The exact controllability to the origin for degenerate linear evolution control system is considered. The obtained general results are applied for the investigation of the exact controllability of the equation of free surface evolution of filtered fluid.

Introduction

The large majority of authors investigates abstract non-generate control differential equations. Below we will consider the exact null controllability problem for degenerate abstract control differential equations.

Let $\mathfrak{X}$, $\mathfrak{Y}$, $\mathfrak{U}$ be Hilbert spaces. Denote by $\mathcal{L}(\mathfrak{X};\mathfrak{Y})$ the Banach space of linear continuous operators acting from $\mathfrak{X}$ to $\mathfrak{Y}$. If $\mathfrak{Y}=\mathfrak{X}$, then the denotation will be cutted to $\mathcal{L}(\mathfrak{X})$. The set of linear closed operators with dense domains in $\mathfrak{X}$, acting to $\mathfrak{Y}$ will be denoted by $\mathcal{C}l(\mathfrak{X};\mathfrak{Y})$. The set $\mathcal{C}l(\mathfrak{X};\mathfrak{X})$ of operators will be denoted by $\mathcal{C}l(\mathfrak{X})$.

Consider the abstract degenerate differential equation

$$L\dot{x}(t)=Mx(t)+Bu(t),\quad 0\leqslant t<+\infty, \tag{0.1}$$

with initial conditions

$$x(0)=x_0\in\mathfrak{X}, \tag{0.2}$$

where $L\in\mathcal{L}(\mathfrak{X};\mathfrak{Y})$, $\ker L\neq\{0\}$, $M\in\mathcal{C}l(\mathfrak{X};\mathfrak{Y})$, $B\in\mathcal{L}(\mathfrak{U};\mathfrak{Y})$.

New results on the exact null controllability of the degenerate equation (0.1) in general Hilbert space are presented. The obtained general results are applied for the investigation of the exact controllability of the equation of free surface evolution of filtered fluid.

Because of restrictions for the number of pages the proofs will be published in the full version of the paper.

1 Strongly (L,p)-radial operator

This section contains some auxiliary results. Their proofs can be found in [4].

Denote $\rho^L(M)=\{\mu\in\mathbb{C}:(\mu L-M)^{-1}\in\mathcal{L}(\mathfrak{Y};\mathfrak{X})\}$, $R^L_\mu(M)=(\mu L-M)^{-1}L$, $L^L_\mu(M)=L(\mu L-M)^{-1}$, $\mathbb{N}_0=\mathbb{N}\cup\{0\},\mathbb{R}_+=\{a\in\mathbb{R}:a>0\}$, $\overline{\mathbb{R}}_+=\mathbb{R}_+\cup\{0\}$.

Definition 1. [5]. Let $p\in\mathbb{N}_0$. Operator M is called strongly (L,p)-radial, if
(i) $\exists a\in\mathbb{R}\quad\forall\mu>a\quad\mu\in\rho^L(M)$;
(ii) $\exists K>0\quad\forall\mu>a\quad\forall n\in\mathbb{N}$

$$\max\{\|(R^L_\mu(M))^{(p+1)n}\|_{\mathcal{L}(\mathfrak{X})},\|(L^L_\mu(M))^{(p+1)n}\|_{\mathcal{L}(\mathfrak{Y})}\}\leqslant\frac{K}{(\mu-a)^{(p+1)n}},$$

$$\|(R^L_\mu(M))^{p+1}(\mu L-M)^{-1}\|_{\mathcal{L}(\mathfrak{Y};\mathfrak{X})}\leqslant\frac{K}{(\mu-a)^{p+2}};$$

(iii) there exists a lineal $\overset{\circ}{\mathfrak{Y}}$ dense in $\mathfrak{Y}$ such that

$$\|M(\mu L-M)^{-1}(L^L_\mu(M))^{p+1}y\|_{\mathfrak{Y}}\leqslant\frac{\mathrm{const}(y)}{(\mu-a)^{p+2}}\quad\forall y\in\overset{\circ}{\mathfrak{Y}}$$

for all $\mu>a$.

Theorem 1. *Let $p\in\mathbb{N}_0$, operator M be strongly (L,p)-radial. Then*
(i) $\mathfrak{X}=\mathfrak{X}_1\oplus\mathfrak{X}_2$, $\mathfrak{Y}=\mathfrak{Y}_1\oplus\mathfrak{Y}_2$;
(ii) $L_k\in\mathcal{L}(\mathfrak{X}_k;\mathfrak{Y}_k)$, $M_k\in\mathcal{C}l(\mathfrak{X}_k;\mathfrak{Y}_k)$, *where*

$$L_k=L\Big|_{\mathfrak{X}_k},\ M_k=M\Big|_{\mathrm{dom}M_k},\ \mathrm{dom}M_k=\mathrm{dom}M\cap\mathfrak{X}_k,\ k=1,2;$$

(iii) *there exist operators* $L_1^{-1}\in\mathcal{L}(\mathfrak{Y}_1;\mathfrak{X}_1)$ *and* $M_2^{-1}\in\mathcal{L}(\mathfrak{Y}_2;\mathfrak{X}_2)$;
(iv) *there exists strongly continuous semigroup* $\{X(t)\in\mathcal{L}(\mathfrak{X}):t\in\overline{\mathbb{R}}_+\}$ *for the equation* $L\dot{x}(t)=Mx(t)$;
(v) *infinitesimal generator of* C_0*-continuous semigroup* $\{S(t)=X(t)\Big|_{\mathfrak{X}_1}\in\mathcal{L}(\mathfrak{X}_1):t\in\overline{\mathbb{R}}_+\}$ *is* $A=L_1^{-1}M_1\in\mathcal{C}l(\mathfrak{X}_1)$;
(vi) *operator* $H=M_2^{-1}L_2\in\mathcal{L}(\mathfrak{X}_2)$ *is nilpotent of the power not greater than* p.

Denote by P (Q) the projector along $\mathfrak{X}_2$ $(\mathfrak{Y}_2)$ on $\mathfrak{X}_1$ $(\mathfrak{Y}_1)$. By conditions of Theorem 1 the equalities

$$LP=QL,\quad MPx=QMx\ \forall x\in\mathrm{dom}M \tag{1.1}$$

hold.

Theorem 2. *Let $p \in \mathbb{N}_0$, operator M be strongly (L,p)-radial, and let the function u be such that $(I-Q)Bu \in C^{p+1}([0,T];\mathfrak{Y})$, $QBu \in C^1([0,T];\mathfrak{Y})$. Then for every initial value*

$$x_0 \in \left\{ x \in \mathrm{dom}M : (I-P)x = -\sum_{k=0}^{p} H^k M_2^{-1}((I-Q)Bu)^{(k)}(0) \right\}$$

there exists a unique solution $x \in C^1([0,T];\mathfrak{X}) \cap C([0,T];\mathrm{dom}M)$ of the problem (0.1), (0.2). *Besides,*

$$x(t) = X(t)x_0 + \int_0^t X(t-s)L_1^{-1}QBu(s)ds - \sum_{k=0}^{p} H^k M_2^{-1}((I-Q)Bu)^{(k)}(t). \quad (1.2)$$

A function $x \in C([0,T];\mathfrak{X})$ is called a mild solution of the problem (0.1), (0.2) if it has a form (1.2).

Theorem 3. *Let $p \in \mathbb{N}_0$, operator M be strongly (L,p)-radial, and let the function u be such that $(I-Q)Bu \in C^{p}([0,T];\mathfrak{Y})$, $QBu \in C([0,T];\mathfrak{Y})$. Then for every initial value*

$$x_0 \in \left\{ x \in \mathfrak{X} : (I-P)x = -\sum_{k=0}^{p} H^k M_2^{-1}((I-Q)Bu)^{(k)}(0) \right\}$$

there exists a unique mild solution $x \in C([0,T];\mathfrak{X})$ of the problem (0.1), (0.2).

Definition 2. A number $\mu \in \mathbb{C}$ is called L-eigenvalue of the operator M, if there exists a vector $x \in \mathfrak{X} \setminus \{0\}$ such that $\mu Lx = Mx$. This vector x is called L-eigenvector of the operator M according to the L-eigenvalue μ. The set of all L-eigenvalues of the operator M is called a point (or discrete) L-spectrum $\sigma_d^L(M)$ of the operator M. The discrete spectrum of any operator C will be denoted by $\sigma_d(C)$.

It is easily to see, that the set of all L-eigenvectors of the operator M, corresponding to the same L-eigenvalue, is a linear subspace of $\mathfrak{X}$. If this subspace is one-dimensional, L-eigenvalue will be called a simply eigenvalue.

Definition 3. Vectors $\varphi_1, ..., \varphi_k \in \mathfrak{X} \setminus \{0\}$, $k \in \mathbb{N}$ are called L-generalized eigenvectors, corresponding to L-eigenvalue of the operator M, if $\mu L\varphi_1 = M\varphi_1$, $\mu L\varphi_{j+1} = M\varphi_{j+1} + L\varphi_j$, $j = 1, 2, ..., k-1$.

Theorem 4. *Let $p \in \mathbb{N}$, operator M be strongly (L,p)-radial. Then all the generalized L-eigenvectors of the operator M belong to the subspace $\mathfrak{X}_1 \equiv \overline{\mathrm{im}(R_\mu^L(M))^{p+1}}$, and $\sigma_d^L(M) = \sigma_d(L_1^{-1}M)$.*

Moreover, the vector φ is L-eigenvector (generalized L-eigenvector) of the operator M if and only if it is eigenvector (generalized eigenvector) for the operator $L_1^{-1}M_1$ of the eigenvalue $\mu \in \mathbb{C}$.

The proof of this theorem will be published in the full version of the paper.

The additional assumptions on the operators L and M are listed below.

(A1) The operator M has purely point L-spectrum $\sigma^L(M) = \sigma_d^L(M)$ with no finite limit points. All the L-eigenvalues of the operator M have finite multiplicities, the sequence of multiplicities is bounded from above.

(A2) The family of generalized L-eigenvectors of the operator M produces a Riesz basis of the subspace $\mathfrak{X}_1$ [1, 7].

2 Problem statement

Let $\mathfrak{X}$, $\mathfrak{Y}$, $\mathfrak{U}$ are Hilbert spaces, operator M is strongly (L,p)-radial, $p \in \mathbb{N} \cup \{0\}$. By acting of the operators $L_1^{-1}Q$ and $M_2^{-1}(I-Q)$ on the Cauchy problem (0.1), (0.2), using the equalities (1.1) and Theorem 1 we obtain equivalent system of two problems

$$\dot{x}_1(t) = Ax_1(t) + B_1u(t),\ x_1(0) = x_{10},\ 0 \leqslant t < +\infty, \tag{2.1}$$

$$H\dot{x}_2(t) = x_2(t) + B_2u(t),\ x_2(0) = x_{20},\ 0 \leqslant t < +\infty, \tag{2.2}$$

where $x_1(t)$, $x_{10} \in \mathfrak{X}_1$, $u(t) \in \mathfrak{U}$ for $t \geqslant 0$; $B_1 = L_1^{-1}QB : \mathfrak{U} \to \mathfrak{X}_1$ and $B_2 = M_2^{-1}(I-Q)B : \mathfrak{U} \to \mathfrak{X}_2$ are linear bounded operators, the operator $A = L_1^{-1}M_1$ is an infinitesimal generator of strongly continuous C_0-semigroup $\{S(t) : t \in \overline{\mathbb{R}}_+\}$ in $\mathfrak{X}_1$, $H = M_2^{-1}L_2 : \mathfrak{X}_2 \to \mathfrak{X}_2$ is a linear continuous nilpotent operator of the degree not greater than p (i. e. $H^{p+1} = 0$).

Definition 4. Equation (2.1) is said to be exact null-controllable on $[0,t_1]$, if for each $x_{10} \in \mathfrak{X}_1$ there exists a control $u \in L_2([0,t_1];\mathfrak{U})$, such that $x_1(t_1, x_{10}, u) = 0$.

Definition 5. Equation (2.2) is said to be exact null-controllable on $[0,t_1]$, if for each $x_{20} \in \mathfrak{X}_2$ there exists a control $u \in C^{(p)}([0,t_1];\mathfrak{U})$, such that $u^{(p)} \in L_2([0,t_1];\mathfrak{U})$, $x_2(t_1, x_{20}, u) = 0$.

Definition 6. System (2.1), (2.2) is said to be exact null-controllable on $[0,t_1]$, if for each $x_{10} \in \mathfrak{X}_1$, $x_{20} \in \mathfrak{X}_2$ there exists a control $u \in C^{(p)}([0,t_1];\mathfrak{U})$, such that $u^{(p)} \in L_2([0,t_1];\mathfrak{U})$, $x_1(t_1, x_{10}, u) = 0$, $x_2(t_1, x_{20}, u) = 0$.

Remark 1. Controllability in the sense of Definition 6 means the controllability of equations (2.1) and (2.2) in the sense of Definitions 4 and 5 correspondingly by the same control.

Below necessary and sufficient conditions of exact null-controllability for degenerate linear evolution control system (2.1), (2.2) with scalar control functions ($\mathfrak{U} = \mathbb{R}$) and with bounded input operators B_1 and B_2 are presented.

3 Controllability of equation (2.2)

Denote

$$\operatorname{span}\{B_2, HB_2, \dots, H^pB_2\} = \left\{ x \in \mathfrak{X}_2 : \exists \alpha_0, \alpha_1, \dots, \alpha_p \in \mathfrak{U} : x = \sum_{k=0}^{p} H^k B_2 \alpha_k \right\},$$

Theorem 5. *Let $t_1 > 0$. Equation (2.2) is exact null-controllable on $[0, t_1]$, if and only if*

$$\operatorname{span}\{B_2, HB_2, \dots, H^pB_2\} = \mathfrak{X}_2. \tag{3.1}$$

The complete proof based on Theorem 3 will be published in the full version of the paper.

4 Controllability of equation (2.1) by smooth scalar controls

From the assumptions (A1), (A2) and Theorem 4 assertions (B1), (B2) follows.

(B1) The operator A has purely point spectrum $\sigma(A) = \sigma_d(A)$ with no finite limit points. All the eigenvalues of A have finite multiplicities, the sequence of multiplicities is bounded from above.

(B2) The family of generalized eigenvectors of the operator A produces a Riesz basis of the space $\mathfrak{X}_1$.

Denote by $\sigma(A)$ the spectrum of operator A. Let $\lambda_j \in \sigma(A)$, $j \in \mathbb{N}$, be eigenvalues, and let α_j and q_j be the algebraic and geometric multiplicities[1] of $\lambda_j \in \sigma(A)$ correspondingly.

Let all the geometrical multiplicities q_j, $j \in \mathbb{N}$, be equal to 1. In this case [8] the exact null-controllable equation (2.1) on $[0, t_1]$ by scalar controls ($r = 1$) can

[1]The geometric multiplicity q_j is the number of Jordan blocks corresponding to $\lambda_j \in \sigma(A)$, and β_j^m is the dimension of m-th Jordan block, $m = 1, 2, \dots, q_j$.

be considered, and the operator $B : \mathfrak{U} \to \mathfrak{X}$ is defined by$Bu = bu$, $u \in \mathbb{R}$, where $b = \begin{pmatrix} b^1 \\ b^2 \end{pmatrix} \in \mathfrak{X}$, $b^1 \in \mathfrak{X}_1$, $b^2 = \mathfrak{X}_2$.

Let ψ_{jk}, $j \in \mathbb{N}$, $k = 1, 2, \ldots, \alpha_j$, be the generalized eigenvectors of the adjoint operator A^*, i. e. $A^*\psi_{j\alpha_j} = \bar{\lambda}_j \psi_{j\alpha_j}$, $j \in \mathbb{N}$, $A^*\psi_{jk} = \bar{\lambda}_j \psi_{jk} + \psi_{jk+1}, j \in \mathbb{N}, k = 1, 2, ..., \alpha_j - 1$.

We use the following notations:

$$g_{jk}(t) = \exp(\lambda_j t) \sum_{l=0}^{\alpha_j - k} b_{j(k+l)} \frac{t^l}{l!},\ t \in [0, t_1],\ j \in \mathbb{N},\ k = 1, 2, \ldots \alpha_j. \tag{4.1}$$

Definition 7. The sequence $\{x_j \in \mathfrak{X} : k \in \mathbb{N}\}$ is said to be minimal, if there no element of the sequence belonging to the closure of the linear span of others. By other words, $x_j \notin \overline{\operatorname{span}}\{x_k \in \mathfrak{X} : k \in \mathbb{N} \setminus \{j\}\}$.

Definition 8. The sequence $\{x_j \in \mathfrak{X} : k \in \mathbb{N}\}$ is said to be strongly minimal, if there exists a positive number $\gamma > 0$ such that

$$\gamma \sum_{k=1}^{n} |c_k|^2 \leqslant \left\| \sum_{k=1}^{n} c_k x_k \right\|^2,\ n \in \mathbb{N}. \tag{4.2}$$

The number γ can be found by the procedure described in [8].

Theorem 6. [8]. *Let the sequence of eigenvalues of operator A forms a Riesz basis in* $\mathfrak{X}_1$. *Equation* (2.1) *is exact null-controllable on* $[0, t_1]$, *if and only if the sequence* (4.1) *is strongly minimal in* $L_2([0, t_1]; \mathbb{R})$.

The proof is obtained by previous Theorem 6 and properties of strongly minimal sequences.

Definition 9. Equation (2.1) is said to be exact null-controllable on $[0, t_1]$ by $(p+1)$-smooth controls, if for each $x_{10} \in \mathfrak{X}_1$ and $(\alpha_0, \alpha_1, \ldots, \alpha_p) \in \mathbb{R}^{p+1}$ there exists a control $u \in C^{(p)}[0, t_1]$, such that $u^{(k)}(0) = \alpha_k$, $u^{(k)}(t_1) = 0$, $k = 0, 1, \ldots, p$, $x_1(t_1, x_{10}, u) = 0$.

To find exact null-controllability conditions by $(p+1)$-smooth controls consider the auxiliary evolution equation

$$\dot{z}(t) = \mathcal{A} z(t) + b^{p+1} v(t), \tag{4.3}$$

where $\mathcal{A} = \begin{pmatrix} A & B_{p+1} \\ 0 & E_{p+1} \end{pmatrix}$, $B_{p+1} = \begin{pmatrix} b^1 & 0 & 0 & \dots & 0 \end{pmatrix}$,

$$E_{p+1} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix},$$

$u = \operatorname{col}(u_1, \dots, u_p) \in \mathbb{R}^p$, $z = \begin{pmatrix} x \\ u \end{pmatrix} \in \mathfrak{X} = \mathfrak{X}_1 \times \mathbb{R}^{p+1}$, $b^{p+1} = \begin{pmatrix} 0 \\ e_{p+1} \end{pmatrix} \in \mathfrak{X}$, where $e_{p+1} = \operatorname{col}(0, 0, \dots, 1) \in \mathbb{R}^{p+1}$.

It is proven that the operator $\mathcal{A}$ generates strongly continuous C_0-semigroup, and the exact null-controllability of system (4.3) by $(p+1)$-smooth control is equivalent to exact null-controllability of degenerate system (2.1), (2.2) in accordance with Definition 6.

In the case of simple eigenvalues λ_j of the operator A denote by ψ_j^1, $j \in \mathbb{N}$, eigenvalues of operator A^*.

Theorem 7. *Let* 0 *be a regular point of the operator* A, $t_1 > 0$. *Equation* (2.1) *is exact null-controllable by scalar control on* $[0, t_1]$ *by* $(p+1)$*-smooth controls, if and only if the family*

$$\frac{(-t)^k}{k!},\ k = 0, 1, \dots, p,\ \frac{1}{\lambda_j} e^{-\lambda_j t} (\psi_j^1, b^1),\ j \in \mathbb{N}, \tag{4.4}$$

of generalized exponents is strongly minimal in $L_2([0, t_1]; \mathbb{R})$.

The proof uses Theorem 6 and the equivalence between exact null-controllability of equation (4.3) and exact null-controllability of equation (2.1) by $(p+1)$-smooth controls.

Theorem 8. *System (0.1) is exact null-controllable by scalar control on* $[0, t_1]$ *if and only if*

(i) *the family (4.4) of generalized exponents is strongly minimal,*

(ii) $\operatorname{span}\left\{b^2, Hb^2, \dots, H^p b^2\right\} = \mathfrak{X}_2$.

The proof is obtained by using of Definition 6 and Remark 1.

5 Exact controllability of Dzektser equation

Consider the initial boundary value problem for Dzektser equation, describing free surface evolution of filtered fluid [2],

$$\left(1+\frac{\partial^2}{\partial\xi^2}\right)\frac{\partial v}{\partial t}(\xi,t)=\left(\frac{\partial^2}{\partial\xi^2}+2\frac{\partial^4}{\partial\xi^4}\right)v(\xi,t)+b(\xi)u(t),\ (\xi,t)\in(0,\pi)\times\mathbb{R}_+,\ (5.1)$$

$$v(0,t)=\frac{\partial^2 v}{\partial\xi^2}(0,t)=v(\pi,t)=\frac{\partial^2 v}{\partial\xi^2}(\pi,t)=0,\quad t\in\mathbb{R}_+, \tag{5.2}$$

$$v(\xi,0)=x_0(\xi),\quad \xi\in(0,\pi).$$

By notations

$$\mathfrak{X}=\left\{x\in H^2(0,\pi):x(0)=x(\pi)=0\right\},\mathfrak{Y}=L_2(0,\pi),\ Lx=x+x'',M=x''+2x'''',$$

$$\mathrm{dom}M=\{x\in H^2(0,\pi):x(0)=x''(0)=x(\pi)=x''(\pi)=0\}$$

$$(Bu)(\xi)=b(\xi)u,\ \xi\in(0,\pi),\ u\in\mathfrak{U}=\mathbb{R},\ b\in L_2(0,\pi),$$

this problem can be reduced to the problem (0.1), (0.2).

It is easily to see, that $\ker L\neq\{0\}$. It is shown, that here $H=0$.

The L-eigenvalues of the operator M are calculated, and it is shown that they are simple. The sequence of generalized exponents (4.4) with $p=0$ (because of $H=0$) is also calculated. Using results of [3] and [7] the strong minimality of sequence (4.4) has been proven, and then by Theorem 8 we obtain that

Theorem 9. *Let* $\langle b(\xi),\sin n\xi\rangle\neq 0$ *for* $n\in\mathbb{N}$ *and series*

$$\sum_{n=2}^{\infty}\frac{(2n^4-n^2)^2}{|\langle b(\xi),\sin n\xi\rangle|^2}e^{-n^2\left(1+\frac{n^2}{n^2-1}\right)t_0}$$

converges for $t_0\geqslant 0$. *Then the system* (5.1), (5.2) *is exact null-controllable on* $[0,t_1]$ *for any* $t_1>t_0$.

References

1. N. Bari, Biorthogonal sequences and bases in Hilbert spaces. *Uchen. Zap. Mosk. Univ.*, 148, Nat, **4**(1951), p. 69-107.
2. E. S. Dzektser, Generalization of moving equation for subterranean waters with free surface, *Dokl. Akad. Nauk SSSR,* **202** (1972), no. 5, p. 1031-1033.

3. H. Fattorini, D. Russel, Uniform bounds on biorthogonal functions for real exponents with an application to the control theory of parabolic equations, *Quart. Appl.Math.,*1974, p. 45-69.
4. V. E. Fedorov, Degenerate strongly continuous semigroups of operators, *St. Petersburg Math. J.*, **12** (2001), no. 3, p. 471-489.
5. V. E. Fedorov, Pseudoresolvents properties and existence conditions for degenerate semigroup of operators, *Herald of Chelyabinsk State University. Mathematics, mechanics, informatics,* **11** (2009), p. 12-19.
6. E. Hille, R. Philips, *Functional Analysis and Semi-Groups*, AMS, 1957.
7. S. Kaczmarz, H. Steinhaus, *Theory of Orthogonal Series,* Monographs Mat., Bd. 6, (PWN, Warsaw), 1958.
8. B. Shklyar, Exact null controllability of abstract differential equations by finite-dimensional controls and strongly minimal families of exponentials, *Differential Equations&Applications*, **2**:3 (2011), p. 171-188.

V. E. Fedorov
Russia, Chelyabinsk, Chelyabinsk State University,

B. Shklyar
Israel, Holon, Holon Institute of Technology

REMARKS ON THE LOCAL CASCADE SEARCH FOR ROOTS AND PREIMAGES

T. N. Fomenko

Key words: multicascade, iteration, local search, root, subspace' preimage

AMS Mathematics Subject Classification: 39B12, 54H25

Abstract. In previous author's works so called cascade search principle was suggested. Given a multivalued functional on a metric space X or given a finite collection of multivalued mappings from X to a metric space Y, that principle allows one to construct some another multivalued mapping from X to itself, generating a multicascade (that is a multivalued discrete dynamic system) on X, with its limit set being equal either to the nil-subspace of the given functional or to the set of common preimages (under the actions of the given mappings) of some given closed subspace $H \subset Y$. Some applications were also given, and stability problems for such multicascades were considered. This paper contains an addition to the latest author's results concerning local versions of the cascade search principle and their applications.

1 Introduction

In the author's works [1–3] several versions of so called cascade search principle were suggested. Given a multivalued functional on a metric space X or given a finite collection of multivalued mappings from X to another metric space Y (or into X itself), that principle allows one to construct some another multivalued mapping from X into itself, generating a multicascade (that is a multivalued disctere dynamic system) on X. That multicascade has a nonempty limit set being equal either to the nil-subspace of the given functional or to the set of common preimages (under the actions of the given mappings) of some given closed subspace $H \subset Y$. Then, in [4–6], the stability problems of the cascade search methods were considered. Several applications of the obtained results were given, as well, concerning the existence and approximation problems of common fixed points and common roots of a finite collection of multivalued mappings from X into itself or from X to another metric space Y, respectively. In particular, essential generalizations of several results of the works [7, 8] were obtained.

The latest author's work [9] is devoted to a generalized cascade search principle, its local versions and some applications. These results were partly represented in

the author's talk at the 8-th International ISAAC Congress (see [10]). In particular, a generalization of [11, Theorem 2] was obtained as one of the consequences.

This paper is an addition to the paper [9]. Here we concentrate on the local cascade search (that is the search near the starting point) for common roots and for common preimages of a given closed subspace under actions of n ($n \geqslant 1$) given multivalued mappings between metric spaces.

Now, let us give necessary definitions and notations.

$\mathbb{R}_{\geqslant 0} = \{t \in \mathbb{R} \mid t \geqslant 0\}$ is the set of nonnegative real numbers; $(X, \rho), (Y, d)$ are metric spaces; $P(Y)$ is the totality of all nonempty subsets of the space Y; $C(Y)$ is the totality of all nonempty closed subsets of Y.

$U(M, r) = \{x \in X \mid \rho(x, M) \leqslant r\}$ is a closed ball neighbourhood of the set M of the radius $r > 0$ in a metric space X. In particular, if $M = \{x\}, x \in X$, $U(x, r)$ is just a closed ball of the radius r with the center in the point x.

A metric D in the space Y^n is defined as follows: $D(y, z) := \sum\limits_{i=1}^{n} d(y_i, z_i)$, where $y = (y_1, \ldots, y_n), z = (z_1, \ldots, z_n) \in Y^n$.

Let $\Delta_n(H) = \{\tilde{y} \in Y^n \mid \tilde{y} = (y, \ldots, y), y \in H\}$ stand for the part of the diagonal of Y^n "over H", where H is a closed subspace in Y. In particular, $\Delta_n = \Delta_n(Y)$ is the whole diagonal in Y^n.

We call the set $F^{-1}(H) = \{x \in X \mid F(x) \cap H \neq \emptyset\}$ *full preimage* of the closed subspace H under the action of a multivalued mapping $F : X \to C(Y)$.

We say the graph $G(F) = \{(x, y) \in X \times Y \mid y \in F(x)\}$ of a multivalued mapping $F : X \to C(Y)$ is *H-complete*, if any cauchy sequence $\{(x_m, y_m)\}_{m=0,1,\ldots} \subseteq G(F)$ with $d(y_m, H) \underset{m\to\infty}{\longrightarrow} 0$ has a limit $(\xi, \eta) \in G(F)$, that is $\eta \in F(\xi) \cap H$.

We say the graph $G(F)$ is *H-closed*, if all its limit points of the form of (x, y) with $y \in H$ are contained in $G(F)$.

A multicascade on X is a multivalued discrete dynamic system with the phase space X and the additive translation semigroup $(\mathbb{Z}_{\geqslant 0}, +)$ ($\mathbb{Z}_{\geqslant 0} = \{0, 1, 2, \ldots\}$). In other words, we say a multicascade is given on X if a multivalued self-mapping $\mathcal{G} : X \to P(X)$ is given, which nonnegative iterations make a semigroup $\{\mathcal{G}^n\}_{n=0,1,\ldots}$ clearly representing the semigroup $(\mathbb{Z}_{\geqslant 0}, +)$ where $\mathcal{G}^0 = \mathrm{id}_X$ is the identical mapping of X. The mapping $\mathcal{G} = \mathcal{G}^1 : X \to P(X)$ representing the generative element $1 \in \mathbb{Z}_{\geqslant 0}$ is called *the multicascade generator*. *A trajectory* of a multicascade is any sequence $\{x_n\}_{n=1,2,\ldots}$, where $x_{n+1} \in \mathcal{G}(x_n), n = 1, 2, \ldots$. So, iterations of the generator $\mathcal{G}$ being applied to a point $x \in X$, make trajectories starting with x. The limits of all trajectories (if exist) form *the limit set* of the multicascade.

2 Generally search functionals and local cascade search for preimages and roots.

Definiton 1. Say a multivalued nonnegative functional $\varphi : X \to P(\mathbb{R}_{\geqslant 0})$ is *generally* (α, β)*-search* on a metric space X, $0 < \beta < \alpha$, if for any pair $(x, t) \in G(\varphi)$ there exists a pair $(x', t') \in G(\varphi)$ such that

$$\rho(x, x') \leqslant \frac{t}{\alpha}, t' \leqslant \frac{\beta}{\alpha} \cdot t.$$

Let a nonnegative multivalued functional $\varphi : X \to P(\mathbb{R}_{\geqslant 0})$ be defined on a metric space X. Denote $\varphi_*(x) = \inf\{\varphi | \varphi \in \varphi(x)\}$.

In [9] the following local cascade search principle was proved.

Theorem 1 [9, Theorem 5]. *Let a multivalued functional $\varphi : X \to P(\mathbb{R}_{\geqslant 0})$ be generally (α, β)-search, $0 < \beta < \alpha$, and $\varphi_*(x_0) < (\alpha - \beta)r$ for some point $x_0 \in X$ and some number $r > 0$. Let also either the graph $G(\varphi)$ be 0-complete or X be complete and the graph $G(\varphi)$ be 0-closed. Then the following statements are true:*

1)There is a multicascade on $G(\varphi)$ with limit set $\mathcal{A} \subseteq G(\varphi)$, $\mathcal{A} \neq \emptyset$, $\mathcal{A}_X = Nil(\varphi)$ where $\mathcal{A}_X$ stands for the projection of $\mathcal{A}$ onto X;

2)$\mathcal{A}_X \cap U(x_0, r) \neq \emptyset$. In particular, for any pair $(x_0, \varphi_0) \in G(\varphi)$ with $\varphi_0 < (\alpha - \beta)r$, there exists a limit point $\xi = \xi(x_0, \varphi_0) \in U(x_0, r)$. □

In [9] some applications of Theorem 1 were also given concerning the local cascade search for the set $Coin_H(F_1, \ldots, F_n) = \{x \in X \mid F_1(x) \cap \ldots \cap F_n(x) \cap H \neq \emptyset\}$ of common preimages of a given closed subspace $H \subset Y$ under the actions of given n $(n \geqslant 1)$ multivalued mappings $F_1, \ldots, F_n : X \to C(Y)$ (see [9, Theorems 6,7]).

Moreover, in [9] there was also considered the problem of the local cascade search for the subset of such coincidence points of n multivalued mappings in which the images intersections are not far from a given closed subspace $Q \subset Y$ (see [9, Theorems 8,9 and Statement 1]). In particular, the obtained results imply a generalization of the recent result [11, Theorem 2].

Let us consider here the more general problem of the local cascade search for the subset of such common preimages of a given closed subspace $H \subset Y$ under actions of n $(n \geqslant 1)$ given multivalued mappings from X to Y that the intersection of the images at each of that points is not far from the given closed subspace $Q \subset Y$. In this way we obtain the following statements (Theorem 2 and Corollaries 1,2 below).

Theorem 2.*Let multivalued mappings $F_1, \ldots, F_n : X \to C(Y)$ be given, $F = F_1 \times \ldots \times F_n$, $Q \in C(Y), H \in C(Y)$. Let the graph $G(F)$ be $\Delta_n(H)$-closed and at least one of the graphs $G(F_i)$ $(1 \leqslant i \leqslant n)$ be H-complete. Let also the following 2 conditions be fulfilled:*

1)for some numbers $\alpha, \beta, 0 < \beta < \alpha, \gamma \geqslant 0$, let the mapping F and the multivalued functional ψ, where $\psi(x) := \{D|D = D(y) = D(y, \Delta_n(H))), y \in F(x)\}$, satisfy the following condition: for any pair $(x, y) \in G(F)$ there exists a pair $(x', y') \in G(F)$ such that the following inequalities are fulfilled ($y = (y_1, \ldots, y_n), y' = (y'_1, \ldots, y'_n)$):

$$\rho(x, x') \leqslant \frac{D(y)}{\alpha}, D(y') \leqslant \beta \cdot \rho(x, x'), d(y_n, y'_1) \leqslant \gamma \cdot \rho(x, x'),$$

2)there exists a pair $(x_0, y_0) \in G(F)$, such that $d(y_{01}, Q) \leqslant (\alpha - \beta)R$, and

$$D_0 = D(y_0) \leqslant \frac{\beta(\alpha - \beta)\min\{2r, R\}}{2\alpha + \gamma};$$

Then a multicascade is defined on $G(F)$ with the nonempty limit set $C \subseteq G(F)$, which projection onto X is $C_X = Coin_H(F_1, \ldots, F_n)$, $C_X \cap U(x_0, r) \neq \emptyset$. In particular, for any pair $(x_0, y_0) \in G(F)$ satisfying the condition 2), there exists a limit pair $(\xi, \tilde{\eta}) \in G(F), \tilde{\eta} = (\eta, \ldots, \eta) \in \Delta_n(H)$, reachable from (x_0, y_0) such that $\xi = \xi(x_0, y_0) \in C_X \cap U(x_0, r)$. Moreover, $\eta = \eta(x_0, y_0) \in U(Q, \alpha R)$. $\square$

We don't give here the detailed proof of Theorem 2 because it is quite similar to the one of [9, Theorem 8]. Just note that in [9, Theorem 8] we used the functional φ where $\varphi(x) := \{\varphi|\varphi = \varphi(y) = \sum\limits_{i=1}^{n-1} d(y_i, y_{i+1}), y = (y_1, \ldots, y_n) \in F(x)\}$. And in the above Theorem 2 we use another functional ψ, $\psi(x) := \{D|D = D(y) = D(y, \Delta_n(H))), y \in F(x)\}$. There is the following obvious inequality between values of the functionals:

$$\sum_{i=1}^{n-1} d(y_i, y_{i+1}) \leqslant 2 \cdot D(y, \Delta_n(H))), y = (y_1, \ldots, y_n)$$

Because of that inequality, the estimation of $D(y_0)$ in the condition 2) of Theorem 2 differs a little from the one of [9, Theorem 8].

So, the above Theorem 2 is a modification of [9, Theorem 8]. Nevertheless, it implies useful consequences.

In the case of $n = 1$ we obtain the following result concerning the local cascade search for preimages of a given closed subspace H under the action of one multivalued mapping.

Corollary 1. *Let a multivalued mapping $F : X \to C(Y)$ be given, $Q \in C(Y), H \in C(Y)$. Let either the graph $G(F)$ be H-complete or X be complete and $G(F)$ be H-closed. Let also the following 2 conditions be fulfilled:*

1)for some numbers $\alpha, \beta, 0 < \beta < \alpha, \gamma \geqslant 0$, let the mapping F and the multivalued functional ψ, $\psi(x) := \{D|D = D(y) = D(y, H)), y \in F(x)\}$, satisfy the following condition: for any pair $(x, y) \in G(F)$ there exists a pair $(x', y') \in G(F)$ such that the following inequalities are fulfilled:

$$\rho(x, x') \leqslant \frac{D(y)}{\alpha}, D(y') \leqslant \beta \cdot \rho(x, x'), d(y, y') \leqslant \gamma \cdot \rho(x, x');$$

2)there exists a pair $(x_0, y_0) \in G(F)$, such that $d(y_0, Q) \leqslant (\alpha - \beta)R$, and

$$D(y_0) \leqslant \frac{\beta(\alpha - \beta)\min\{2r, R\}}{2\alpha + \gamma};$$

Then there exists a multicascade on $G(F)$ with the nonempty limit set $C \subseteq G(F)$, which projection onto X is $C_X = F^{-1}(H)$, and $C_X \cap U(x_0, r) \neq \emptyset$. In particular, for any pair $(x_0, y_0) \in G(F)$ satisfying the condition 2), there exists a limit pair $(\xi, \eta) \in G(F)$, reachable from (x_0, y_0) such that $\xi = \xi(x_0, y_0) \in F^{-1}(H) \cap U(x_0, r)$. Moreover, $\eta = \eta(x_0, y_0) \in H \cap U(Q, \alpha R)$. $\square$

The following statement concerns the local cascade search for common roots of n ($n \geqslant 2$) given multivalued mappings.

Corollary 2. *Let multivalued mappings $F_1, \ldots, F_n : X \to C(Y)$ be given, $F = F_1 \times \ldots \times F_n$, $Q \in C(Y)$ and $c \in Y$. Let the graph $G(F)$ be $\tilde{c}$-closed, $\tilde{c} = (c, \ldots, c) \in \Delta_n \subset Y^n$, and at least one of the graphs $G(F_i)$ $(1 \leqslant i \leqslant n)$ be c-complete. Let also the following 2 conditions be fulfilled:*

1)for some numbers $\alpha, \beta, 0 < \beta < \alpha, \gamma \geqslant 0$, the mapping F and the multivalued functional ψ, $\psi(x) = \{D|D = D(y) = \sum_{i=1}^{n} d(y_i, c), y = (y_1, \ldots, y_n) \in F(x)\}$, satisfy the following condition: for any pair $(x, y) \in G(F)$ there exists a pair $(x', y') \in G(F)$ such that the following inequalities are fulfilled:

$$\rho(x, x') \leqslant \frac{D(y)}{\alpha}, D(y') \leqslant \beta \cdot \rho(x, x'), d(y_n, y_1') \leqslant \gamma \cdot \rho(x, x'),$$

where $y = (y_1, \ldots, y_n), y' = (y_1', \ldots, y_n'))$;

2)there exists a pair $(x_0, y_0) \in G(F)$, such that $d(y_{01}, Q) \leqslant (\alpha - \beta)R$, and

$$D(y_0) \leqslant \frac{\beta(\alpha - \beta)\min\{2r, R\}}{2\alpha + \gamma};$$

Then there exists a multicascade on $G(F)$ with the nonempty limit set $C \subseteq G(F)$, which projection onto X is the set $C_X = CR_c(F_1, \ldots, F_n) = \{x \in X \mid F_1(x) \cap \ldots \cap F_n(x) \ni c\}$ of common roots of the mappings $F_1, \ldots, F_n$ corresponding to their

common value c, and $C_X \cap U(x_0, r) \neq \emptyset$. *In particular, for any pair* $(x_0, y_0) \in G(F)$ *satisfying the condition 2), there exists a limit pair* $(\xi, \tilde{c}) \in G(F)$ *reachable from* (x_0, y_0) *such that* $\rho(x_0, \xi) \leqslant r$, $d(\eta, Q) \leqslant \alpha R$. $\square$

Now, let us formulate the following modification of [9, Theorem 9] also (like as Theorem 2 above) concerning the local search for common preimages of n given multivalued mappings.

Theorem 3. *Let mappings* $F_1, \ldots, F_n : X \to C(Y)$ *be given,* $F = F_1 \times \ldots \times F_n$, $Q \in C(Y), H \in C(Y)$, $\psi : X \to P(\mathbb{R}_{\geqslant 0}), \psi(x) = \{D | D = D(y, \Delta_n(H)), y \in F(x)\}, x \in X$. *Let the graph* $G(F)$ *be* $\Delta_n(H)$*-closed and at least one of the graphs* $G(F_i)$, $i = 1, \ldots, n$, *be* H*-complete. Moreover, let for some numbers* $0 < \beta < \alpha, \gamma \geqslant 0$, $r > 0, R > 0$ *the following conditions be fulfilled (where* $y = (y_1, \ldots, y_n), y' = (y'_1, \ldots, y'_n), y = (y''_1, \ldots, y''_n)$*):*
a)for any pair $(x, y) \in G(F)$ *there exists such a pair* $(x', y') \in G(F)$, *that the following inequalities are true:*

$$\psi(y') \leqslant \beta \cdot \rho(x, x'), d(y_n, y'_1) \leqslant \gamma \cdot \rho(x, x');$$

b)for any pair of pairs $((x, y), (x', y')) \in G(F) \times G(F)$ *there exists such a pair* $(x'', y'') \in G(F)$, *that the following inequalities are fulfilled*

$$\psi(y'') \leqslant \beta \cdot \rho(x', x''), d(y'_n, y''_1) \leqslant \gamma \cdot \rho(x', x''), \rho(x', x'') \leqslant \frac{\beta}{\alpha} \rho(x, x');$$

c)there exists a pair of pair $((x_0, y_0), (x_1, y_1)) \in G(F) \times G(F)$ *satisfying the condition a), that*

$$d(y_{0n}, Q) \leqslant (\alpha - \beta)R, \quad \rho(x_0, x_1) \leqslant \frac{\beta(\alpha - \beta)}{\alpha(2\beta + \gamma)} \min\{2r, R\}.$$

Then there is a multicascade on $G(F) \times G(F)$ *with nonempty limit set* $\mathcal{A} \subseteq \Delta$, *where* Δ *is the diagonal in* $G(F) \times G(F)$. *The projection of any trajectory of that multicascade onto the first component is an approximating sequence in* $G(F)$, *converging to a pair* $(\xi, \eta) \in G(F)$, *where* $\xi \in Coin_H(F_1, \ldots, F_n)$. *Besides, if such sequence starts with the pair* (x_0, y_0) *satisfying the condition c), then it converges to such a pair* $(\xi_0, \eta_0) \in G(F)$ *that* $\xi_0 \in Coin_H(F_1, \ldots, F_n) \cap U(x_0, r)$ *and* $\eta_0 \in U(Q, \alpha R)$. $\square$

The proof of Theorem 3 is rather standart and similar to the proof of [9, Theorem 9]. The only difference is that in Theorem 3 we consider $\Delta_n(H)$ instead of Δ_n and use the functional $\psi(x) := \{D | D = D(y, \Delta_n(H)), y \in F(x)\}$ instead of the functional $\varphi(x) := \{\varphi | \varphi = \varphi(y) = \sum_{i=1}^{n-1} d(y_i, y_{i+1}), y = (y_1, \ldots, y_n) \in F(x)\}$. Taking

into account the inequality $\sum_{i=1}^{n-1} d(y_i, y_{i+1}) \leqslant 2 \cdot D(y, H)$ and proceeding by standart way we obtain the required estimations which are slightly different from the ones of [9, Theorem 9].

In the case of $H = Y$, Theorem 3 implies a consequence concerning the local cascade search for such coincidence points of the mappings $F_1, \ldots, F_n$ which have common values lying "not far from Q." That consequence differs from [9, Theorem 9] because of the use of another search functional ψ.

In the case of $H = \{c\}, c \in Y$, Theorem 3 implies a corollary concerning the local cascade search for common roots of the mappings $F_1, \ldots, F_n$ corresponding to their common value c, also "not far from Q."

In conclusion, let us formulate one more consequence from Theorem 3 similar to [9, Statement 1].

Corollary 3. *Let mappings $F_1, \ldots, F_n : X \to C(Y)$ be given, $F = F_1 \times \ldots \times F_n$, $\psi : X \to P(\mathbb{R}_{\geqslant 0}), \psi(x) = \{D | D = D(y) = D(y, \Delta_n(H)), y = (y_1, \ldots, y_n) \in F(x)\}$, $x \in X$, $H \in C(Y)$, $Q \in C(Y)$. Let the graph $G(F)$ be $\Delta_n(H)$-closed and at least one of the graphs $G(F_i)$ $(1 \leqslant i \leqslant n)$ be H-complete. Let for some numbers $0 < \beta < \alpha, \gamma \geqslant 0$, $r > 0, R > 0$, and for some pair $(x_0, y_0) \in G(F)$ it is true that $d(y_{0n}, Q) \leqslant (\alpha - \beta)R$ and there exists a sequence $\{(x_m, y_m)\}_{m=0,1,\ldots} \subseteq G(F)$ starting with (x_0, y_0) and satisfying the following inequalitites (where $m \geqslant 1$, $y_m = (y_{m1}, \ldots, y_{mn})$):*

$$\rho(x_0, x_1) \leqslant \frac{\beta(\alpha - \beta)}{\alpha(2\beta + \gamma)} \min\{2r, R\}, \quad \rho(x_m, x_{m+1}) \leqslant \frac{\beta}{\alpha} \rho(x_{m-1}, x_m),$$
$$\psi(y_{m+1}) \leqslant \beta \cdot \rho(x_m, x_{m+1}), \qquad d(y_{mn}, y_{(m+1)1}) \leqslant \gamma \cdot \rho(x_m, x_{m+1}).$$

Then every such sequence has a limit $(\xi, \eta) \in G(F)$, where $\xi \in Coin_H(F_1, \ldots, F_n) \cap U(x_0, r)$, $\eta \in H \cap U(Q, \alpha R)$. $\square$

Using other suitable functionals and concrete subspaces H and Q one can realize local cascade search for the solving specific problems.

References

1. Fomenko T.N. *Approximation of coincidence points and common fixed points of a collection of mappings of metric spaces.* Mathematical Notes, 86:1, 2009. Pp.107-120.
2. Fomenko T.N. *Cascade search of the coincidence set of collections of multivalued mappings.* Mathematical Notes, 86:2, 2009. Pp.276-281.

3. Fomenko T.N. *Cascade search principle and its applications to the coincidence problem of n one-valued or multi-valued mappings.* Topology and its Applications, 157, 2010. Pp.760-773.
4. Fomenko T.N. *Stability of Cascade Search.* Izvestiya: Mathematics, 74:5, 2010. Pp. 1051-1068.
5. Fomenko T.N. *Cascade search: Stability of reachable limit points.* Moscow University Mathematics Bulletin, vol.65, Number 5/October, 2010. Pp.179-185.
6. Fomenko T.N. *The stability of Cascade Search Principle.* 2010 International Conference on Topology and its Applications, June 26-30, Nafpaktos, Greece. Abstracts. Nafpaktos, 2010. P.99.
7. Arutyunov A.V. *Covering mappings in Metric Spaces and Fixed Points.* Doklady Mathematics, vol.76, No.2, 2007. Pp.665-669.
8. Arutyunov A.V. *Stability of coincidence points and properties of covering mappings.* Mathematical Notes, 86:2, 2009. Pp.153-158.
9. Fomenko T.N. *Cascade search for preimages and coincidences: global and local versions.* Mathematical Notes, to appear.
10. Fomenko T.N. *New developments in the cascade search theory.* 8-th International ISAAC Congress, Moscow, August 22-27, 2011. Abstracts. P.369.
11. Arutyunov A., Avakov E., Gelman B., Dmitruk A., Obukhovskii V. *Locally covering maps in metric spaces and coincidence points.* Journal of Fixed Points Theory and its Applications, 5, №1, 2009, 106-127.

T. N. Fomenko
Department of General Mathematics, Faculty of Computational Mathematics and Cybernetics, Moscow State University, Russia, 119991, Moscow, Leninskie Gory, MGU, phones: (495)939-55-91(off.), (495)939-01-32(h.), (916)401-56-87(mob.), E-mail address: tn-fomenko@yandex.ru

LEVEL SETS OF THE VALUE FUNCTION IN DIFFERENTIAL GAME WITH TWO PURSUERS AND ONE EVADER

S. A. Ganebny, S. S. Kumkov, V. S. Patsko, Stéphane Le Ménec

Key words: pursuit-evasion differential game, linear dynamics, value function

AMS Mathematics Subject Classification: 49N70 49N75

Abstract. An antagonistic differential game is considered where motion occurs in a straight line. Deviations between the first and second pursuers and the evader are computed at the instants T_1 and T_2, respectively. The pursuers act together. Their aim is to minimize the resultant miss, which is equal to the minimum of the deviations taken at the instants T_1 and T_2. Numerical study of value function level sets (Lebesgue sets) for qualitatively different cases is given.

1 Introduction and Problem Formulation

1. In the paper, a model differential game with two pursuers and one evader is studied. Three inertial objects moves in the straight line. The dynamics descriptions for pursuers P_1 and P_2 are

$$\begin{aligned} &\ddot{z}_{P_1} = a_{P_1}, && \ddot{z}_{P_2} = a_{P_2}, \\ &\dot{a}_{P_1} = (u_1 - a_{P_1})/l_{P_1}, && \dot{a}_{P_2} = (u_2 - a_{P_2})/l_{P_2}, \\ &|u_1| \leqslant \mu_1, && |u_2| \leqslant \mu_2, \\ &a_{P_1}(t_0) = 0, && a_{P_2}(t_0) = 0. \end{aligned} \tag{1.1}$$

Here, z_{P_1} and z_{P_2} are the geometric coordinates of the pursuers; a_{P_1} and a_{P_2} are their accelerations generated by the controls u_1 and u_2. The time constants l_{P_1} and l_{P_2} define how fast the controls affect the systems.

The dynamics of the evader E is similar:

$$\ddot{z}_E = a_E, \quad \dot{a}_E = (v - a_E)/l_E, \quad |v| \leqslant \nu, \quad a_E(t_0) = 0. \tag{1.2}$$

Let us fix some instants T_1 and T_2. At the instant T_1, the miss of the first pursuer with respect to the evader is computed, and at the instant T_2, the miss of

This work was supported by the Russian Foundation for Fundamental Research under grants No.10-01-96006, 11-01-12088.

the second one is calculated:

$$r_{P_1,E}(T_1) = |z_E(T_1) - z_{P_1,E}(T_1)|, \quad r_{P_2,E}(T_2) = |z_E(T_2) - z_{P_2,E}(T_2)|. \tag{1.3}$$

Assume that the pursuers act in coordination. This means that we can join them into one player P (which will be called the *first player*). This player governs the vector control $u = (u_1, u_2)$. The evader is counted as the *second player*. The resultant miss is the following value:

$$\varphi = \min\{r_{P_1,E}(T_1),\ r_{P_2,E}(T_2)\}. \tag{1.4}$$

At any instant t, both players know exact values of all state coordinates z_{P_1}, $\dot{z}_{P_1}$, a_{P_1}, z_{P_2}, $\dot{z}_{P_2}$, a_{P_2}, z_E, $\dot{z}_E$, a_E. The vector composed of these components is denoted as z. The first player choosing its feedback control minimizes the miss φ, the second one maximizes it.

Relations (1.1)–(1.4) define a standard antagonistic differential game. One needs to construct the value function $(t, z) \mapsto \mathcal{V}(t, z)$ of this game.

2. Up to now, there are a lot of publications dealing with linear differential games where one group of objects pursues another group; see, for example, works [1, 3, 5, 10]. The problem under consideration has two pursuers and one evader. So, from the point of view of number of objects, it is the simplest one. On the other hand, strict mathematical studies of problems "group-on-group" usually include quite strong assumptions onto the dynamics of objects, dimension of the state vector, and conditions of termination. Conversely, this paper considers the problem without any assumptions of these types.

3. Let us describe a practical problem, whose reasonable simplification gives the model game (1.1)–(1.4). Suppose that two pursuing objects attack the evading one on collision courses. They can be rockets or aircrafts in the horizontal plane. A nominal motion of the first pursuer is chosen such that at the instant T_1 the exact capture occurs. In the same way, a nominal motion of the second pursuer is chosen (the capture is at the instant T_2). But indeed, the real positions of the objects differ from the nominal ones. Moreover, the evader using its control can change its trajectory in comparison with the nominal one (but not principally, without sharp turns). Correcting coordinated efforts of the pursuers are computed during the process by the feedback method to minimize the resultant miss, which is the minimum of absolute values of deviations at the instants T_1 and T_2 from the first and second pursuers, respectively, to the evader.

The passage from the original non-linear dynamics to a dynamics, which is linearized with respect to the nominal motions, gives [11, 12] the problem under consideration.

2 Passage to Two-Dimensional Differential Game

At first, let us pass to relative geometric coordinates

$$y_1 = z_E - z_{P_1}, \quad y_2 = z_E - z_{P_2} \tag{2.1}$$

in dynamics (1.1), (1.2) and payoff function (1.4). After this, we have the following notations:

$$\begin{aligned} &\ddot{y}_1 = a_E - a_{P_1}, && \ddot{y}_2 = a_E - a_{P_2}, \\ &\dot{a}_{P_1} = (u_1 - a_{P_1})/l_{P_1}, && \dot{a}_{P_2} = (u_2 - a_{P_2})/l_{P_2}, \\ &\dot{a}_E = (v - a_E)/l_{P_1}, && |u_2| \leqslant \mu_2, \\ &|u_1| \leqslant \mu_1,\ |v| \leqslant \nu, && \varphi = \min\{|y_1(T_1)|,\ |y_2(T_2)|\}. \end{aligned} \tag{2.2}$$

State variables of system (2.2) are y_1, $\dot{y}_1$, a_{P_1}, y_2, $\dot{y}_2$, a_{P_2}, a_E; u_1 and u_2 are controls of the first player; v is the control of the second one. The payoff function φ depends on the coordinate y_1 at the instant T_1 and on the coordinate y_2 at the instant T_2.

A standard approach to study linear differential games with fixed terminal instant and payoff function depending on some state coordinates at the terminal instant is to pass to new state coordinates (see, for example, [6,7]) that can be treated as values of the target coordinates forecasted to the terminal instant under zero controls. Often, these coordinates are called the *zero effort miss coordinates* [11, 12]. In our case, we have two instants T_1 and T_2, but coordinates computed at these instants are independent; namely, at the instant T_1, we should take into account $y_1(T_1)$ only, and at the instant T_2, we use the value $y_2(T_2)$. This fact allows us to use the mentioned approach when solving the differential game (2.2). With that, we pass to new state coordinates x_1 and x_2, where $x_1(t)$ is the value of y_1 forecasted to the instant T_1 and $x_2(t)$ is the value of y_2 forecasted to the instant T_2.

The forecasted values are computed by formula

$$x_i = y_i + \dot{y}_i\tau_i - a_{P_i}l_{P_i}^2 h(\tau_i/l_{P_i}) + a_E l_E^2 h(\tau_i/l_E), i = 1, 2. \tag{2.3}$$

Here, x_i, y_i, $\dot{y}_i$, a_{P_i}, and a_E depend on t; $\tau_i = T_i - t$. Function h is described by the relation $h(\alpha) = e^{-\alpha} + \alpha - 1$. Emphasize that the values τ_1 and τ_2 are connected

to each other by the relation $\tau_1 - \tau_2 = \text{const} = T_1 - T_2$. It is very important that $x_i(T_i) = y_i(T_i)$. Let $X(t,z)$ be a two-dimensional vector composed of the variables x_1, x_2 defined by formulae (2.1), (2.3).

The dynamics in the new coordinates x_1, x_2 is the following [8]:

$$\begin{aligned} \dot{x}_1 &= -l_{P_1} h(\tau_1/l_{P_1}) u_1 + l_E h(\tau_1/l_E) v, \qquad |u_1| \leqslant \mu_1,\ |u_2| \leqslant \mu_2, \\ \dot{x}_2 &= -l_{P_2} h(\tau_2/l_{P_2}) u_2 + l_E h(\tau_2/l_E) v, \qquad |v| \leqslant \nu. \end{aligned} \tag{2.4}$$

The payoff function is $\varphi\big(x_1(T_1), x_2(T_2)\big) = \min\{|x_1(T_1)|,\ |x_2(T_2)|\}$.

The first player governs the controls u_1, u_2 and minimizes the payoff φ; the second one has the control v and maximizes φ.

Note that the control u_1 (u_2) affects only the horizontal (vertical) component $\dot{x}_1$ ($\dot{x}_2$) of the velocity vector $\dot{x} = (\dot{x}_1, \dot{x}_2)^{\mathrm{T}}$. When $T_1 = T_2$, the second summand in dynamics (2.4) is the same for $\dot{x}_1$ and $\dot{x}_2$. Thus, the component of the velocity vector $\dot{x}$ depending on the second player control is directed at any instant t along the bisectrix of the first and third quadrants of the plane x_1, x_2. When $v = +\nu$, the angle between the axis x_1 and the velocity vector of the second player is 45°; when $v = -\nu$, the angle is 225°. This property simplifies the dynamics in comparison with the case $T_1 \neq T_2$.

Let $x = (x_1, x_2)^{\mathrm{T}}$ and $V(t,x)$ be the value of the value function of game (2.4) at the position (t,x). From general results of the differential game theory, it follows that $\mathcal{V}(t,z) = V\big(t, X(t,z)\big)$. This relation allows to compute the value function of the original game (1.1)–(1.4) using the value function for game (2.4).

For any $c \geqslant 0$, a level set (a Lebesgue set) $W_c = \big\{(t,x) : V(t,x) \leqslant c\big\}$ of the value function in game (2.4) can be treated as the solvability set for the considered game with the result not greater than c, that is, for a differential game with dynamics (2.4) and the terminal set $M_c = \big\{(t,x) : t = T_1,\ |x_1| \leqslant c;\ t = T_2,\ |x_2| \leqslant c\big\}$. When $c = 0$, one has the situation of the exact capture. The exact capture means equality to zero, at least, one of $x_1(T_1)$ and $x_2(T_2)$. Let $W_c(t) = \{x : (t,x) \in W_c\}$ be the time section (t-section) of the set W_c at the instant t. Similarly, let $M_c(t)$ for $t = T_1$ and $t = T_2$ be the t-section of the set M_c at the instant t.

Comparing dynamics capabilities of each of pursuers P_1 and P_2 and the evader E, one can introduce the parameters [8, 12] $\eta_i = \mu_i/\nu$, $\varepsilon_i = l_E/l_{P_i}$, $i = 1, 2$. They define the shape of the solvability sets in the individual games P_1–E and P_2–E. Namely, depending on values of η_i and $\eta_i\varepsilon_i$ (which are not equal 1 simultaneously), there are 4 cases [12] of the solvability set evolution (see Fig. 1):

— expansion in the backward time (a strong pursuer);
— contraction in the backward time (a weak pursuer);
— expansion until some backward time instant and further contraction;

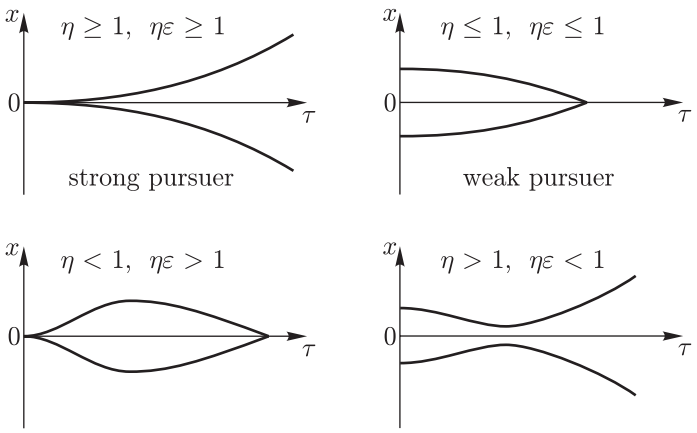


Figure 1. Variants of the solvability set evolution in an individual game

— contraction until some backward time instant and further expansion (if the solvability set still has not broken).

Respectively, given combinations of pursuers' capabilities in individual games and durations T_1, T_2 (equal/different), there are significant number of variants for the problem with two pursuers and one evader.

The ideology of solving the game used by us is the following. Choose the parameters η_i, ε_i, and also the instants T_i, $i = 1, 2$; then, using some fine grid of values of c, we compute level sets W_c of the value function. After that, we can build quasioptimal strategies of the first and second players. But in this paper, we study only the level sets W_c of the value function.

Nowadays, different workgroups suggested many algorithms for numeric solution of differential games of quite general type (see, for example, [2,4,9,13]). Problem (2.4) has the second order on the phase variable and can be rewritten as

$$\dot{x} = \mathscr{D}_1(t)u_1 + \mathscr{D}_2(t)u_2 + \mathscr{E}(t)v, \qquad |u_1| \leqslant \mu_1,\ |u_2| \leqslant \mu_2,\ |v| \leqslant \nu. \tag{2.5}$$

Here, $x = (x_1, x_2)^{\mathrm{T}}$; vectors $\mathscr{D}_1(t)$, $\mathscr{D}_2(t)$, and $\mathscr{E}(t)$ look like

$$\mathscr{D}_1(t) = \big(-l_{P_1}h((T_1 - t)/l_{P_1})^{\mathrm{T}},\, 0\big), \qquad \mathscr{D}_2(t) = \big(0,\, -l_{P_2}h((T_2 - t)/l_{P_2})\big)^{\mathrm{T}},$$
$$\mathscr{E}(t) = \big(l_E h((T_1 - t)/l_E),\, l_E h((T_2 - t)/l_E)\big)^{\mathrm{T}}.$$

The control of the first player has two independent components u_1 and u_2. The vector $\mathscr{D}_1(t)$ ($\mathscr{D}_2(t)$) is directed along the horizontal (vertical) axis. The second

player's control v is scalar. When $T_1 = T_2$, the angle between the axis x_1 and the vector $\mathscr{E}(t)$ equals 45°; when $T_1 \neq T_2$, the angle changes in time.

Due to peculiarity of our problem, we use special methods for constructing level sets of the value function.

3 Maximal Stable Bridge: Control with Discrimination

A level set W_c of the value function V is a maximal stable bridge (MSB), breaking on the terminal set M_c [6, 7].

Let $T_1 = T_2$. Denote $T_f = T_1$. Using the concept of MSB from [6, 7], we can say that W_c is the set maximal by inclusion in the space $t \leqslant T_f$, x such that $W_c(T_f) = M_c(T_f)$ and the *stability* property is hold: for any position $(t_*, x_*) \in W_c(t_*)$, $t_* < T_f$, any instant $t^* > t_*$, $t^* \leqslant T_f$, any constant control v of the second player, which obeys the constraint $|v| \leqslant \nu$, there is a measurable control $t \to \big(u_1(t), u_2(t)\big)$ of the first player, $t \in [t_*, t^*)$, $|u_1(t)| \leqslant \mu_1$, $|u_2(t)| \leqslant \mu_2$, guiding system (2.4) from the state x_* to the set $W_c(t^*)$ at the instant t^*.

The stability property assumes a discrimination of the second player by the first one: the choice of the first player's control in the interval $[t_*, t^*)$ is made after the second player announces his control in this interval.

It is known (see [6, 7]) that any MSB is close. The set $W_c^{(2)}(t) = \mathrm{cl}\left(R^2 \setminus W_c(t)\right)$ (the symbol cl denotes the operation of closure) is the time section of MSB $W_c^{(2)}$ for the second player at the instant t. The bridge terminates at the instant T_f on the set $M_c^{(2)}(T_f) = \mathrm{cl}\big(R^2 \setminus M_c(T_f)\big)$. If the initial position of system (2.4) is in $W_c^{(2)}$ and if the first player is discriminated by the second one, then the second player is able to guide the motion to the set $M_c^{(2)}(T_f)$ at the instant T_f. Thus, $\partial W_c = \partial W_c^{(2)}$. It is proved that for any initial position $(t_0, x_0) \in \partial W_c$, the value c is the best guaranteed result for the first (second) player in the class of feedback controls.

Due to symmetry of dynamics (2.4) and the set $W_c(T_f)$ with respect to the origin, one gets that for any $t \leqslant T_f$ the time section $W_c(t)$ is symmetric also.

If $T_1 \neq T_2$, then there is no any appreciable complication in constructing MSBs for the problem considered in this paper in comparison with the case $T_1 = T_2$. Indeed, let $T_1 > T_2$. Then in the interval $(T_2, T_1]$ in (2.4), we take into account only the dynamics of the variable x_1 when building the bridge W_c backwardly from the instant T_1. With that, the terminal set at the instant T_1 is taken as $M_c(T_1) = \{(x_1, x_2) : |x_1| \leqslant c\}$. When the constructions are made up to the

instant T_2, we add the set $M_c(T_2)$, that is, we take

$$W_c(T_2) = W_c(T_2 + 0) \bigcup \{(x_1, x_2) : |x_2| \leqslant c\},$$

and further constructions are made on the basis of this set.

So, our tool for finding a level set of the value function in game (2.4) corresponding to a number c is the backward procedure for constructing a MSB with the terminal set M_c. Presence of an idealized element (the discrimination of the opponent) allowed us to create effective numerical methods for backward construction of MSBs.

The solvability set with the index equal to c in the individual game $P1$–E ($P2$–E) is MSB built in the coordinates t, x_1 (t, x_2) and terminating at the instant T_1 (T_2) on the set $|x_1| \leqslant c$ ($|x_2| \leqslant c$). Its t-section, if it is non-empty, is a segment in the axis x_1 (x_2) symmetric with respect to the origin. In the plane x_1, x_2, this segment corresponds to a vertical (horizontal) strip of the same width near the axis x_2 (x_1). It is evident that when $t \leqslant T_1$ ($t \leqslant T_2$) such a strip is contained in the section $W_c(t)$ of MSB W_c of game (2.4) with the terminal set M_c.

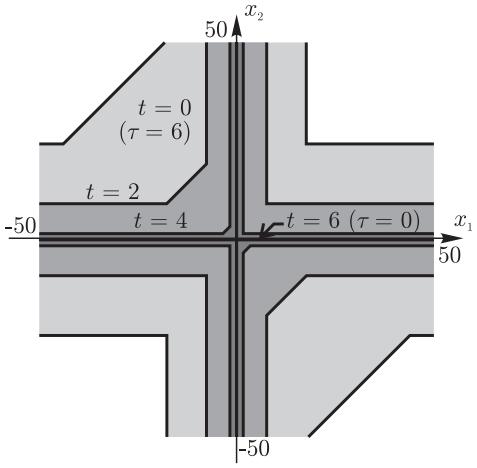


Figure 2. Two strong pursuers, equal terminal instants: time sections of the maximal stable bridge W_0

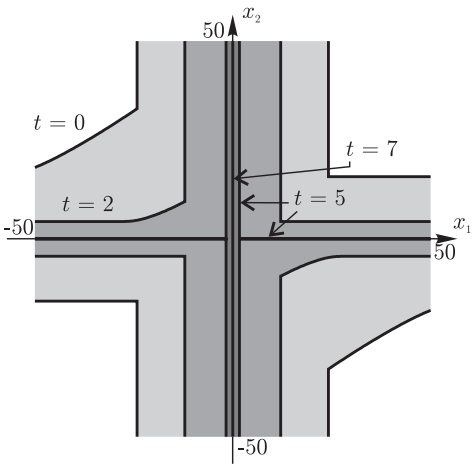


Figure 3. Two strong pursuers, different terminal instants: time sections of the maximal stable bridge W_0

4 Results of Numerical Constructions

Case of strong pursuers. In the case of two strong pursuers, the t-sections of MSBs in individual games $P1$–E and $P2$–E grow with increasing of the backward time. This gives that for any $c \geqslant 0$ and any $t \leqslant \bar{t} = \min\{T_1, T_2\}$ the set $W_c(t)$ includes a cross near the axes x_1, x_2, which expands with decreasing t.

Let us give results of constructing t-sections $W_c(t)$ for the following values of the game parameters: $\mu_1 = 2$, $\mu_2 = 3$, $\nu = 1$, $l_{P_1} = 1/2$, $l_{P_2} = 1/0.857$, $l_E = 1$.

Equal terminal instants. Let $T_1 = T_2 = 6$. Fig. 2 shows results of constructing the set W_0 (that is, with $c = 0$). In the figure, one can see several time sections $W_0(t)$ of this set. The bridge has a quite simple structure. At the initial instant $\tau = 0$ of the backward time (when $t = 6$), its section coincides with the target set, which is the union of two coordinate axes. Further, at the instants $t = 4$, 2, 0, the cross thickens, and two triangles are added to it. The widths of the vertical and horizontal parts of the cross correspond to sizes of MSBs in the individual games with the first and second pursuers. These triangles are located in the II and IV quadrants (where the signs of x_1 and x_2 are different, in other words, when the evader is between the pursuers). They give the zone where the exact capture is possible only under collective actions of both pursuers.

Time sections $W_c(t)$ of other bridges W_c, $c > 0$, have a shape similar to $W_0(t)$.

Different terminal instants. Let $T_1 = 7$, $T_2 = 5$. Results of constructing the set W_0 are given in Fig. 3. When $t < 5$, time sections $W_0(t)$ grow both horizontally and vertically; two additional triangles appear, but in this case they are curvilinear. In Fig. 4, the set W_0 is shown in the three-dimensional space t, x_1, x_2.

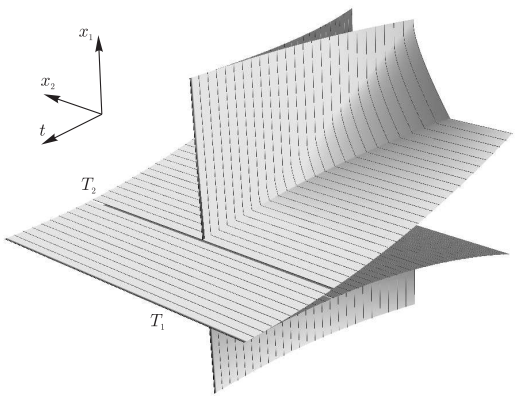


Figure 4. Strong pursuers, different terminal instants: 3D-view of the set W_0

The given results are typical for the case of strong pursuers. When $T_1 = T_2$, the sets $W_c(t)$ can be described analytically. This was done in paper [8]. Also, there the case $T_1 \neq T_2$ was studied. But for it, only an upper approximation of the sets $W_c(t)$ was obtained.

Case of weak pursuers. Since in the case of weak pursuers the t-sections of MSBs in individual games $P1$–E and $P2$–E contract with growth of the backward time and become empty at some instant, the set $W_c(t)$ for any $c \geqslant 0$ with decreasing of t loses infinite sizes along axes x_1 and x_2.

The most surprising fact discovered during the numerical study was that the connected set $W_c(t)$ with decreasing of t loses connectedness and disjoins into two separate parts.

Take the parameters $\mu_1 = 0.9$, $\mu_2 = 0.8$, $\nu = 1$, $l_{P_1} = l_{P_2} = 1/0.7$, $l_E = 1$. Let us show results for the case of different terminal instants only: $T_1 = 9$, $T_2 = 7$. Since in this variant the evader is more maneuverable than the pursuers, the first player cannot guarantee the exact capture.

The set W_c in the space t, x_1, x_2 for $c = 2.0$ is shown in Fig. 5. During evolution of the sections $W_{2.0}(t)$ in t, they change their structure at some instants. These places are marked by drops in the constructed surface of the set.

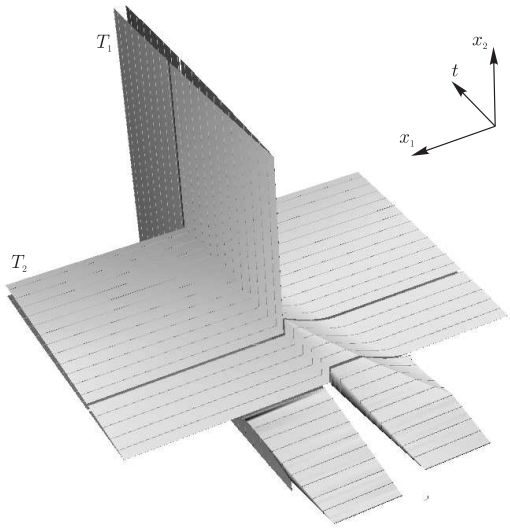


Figure 5. Two weak pursuers, different terminal instants: 3D-view of the set $W_{2.0}$

One strong and one weak pursuers. Let us take the following parameters: $\mu_1 = 2$, $\mu_2 = 1$, $\nu = 1$, $l_{P_1} = 1/2$, $l_{P_2} = 1/0.3$, $l_E = 1$. Now the evader is more

maneuverable than the second pursuer, and an exact capture by this pursuer is unavailable. Assume $T_1 = 5$, $T_2 = 7$.

In Fig. 6, a three-dimensional view of MSB $W_{5.0}$ is shown. The horizontal part of its time section $W_{5.0}(t)$ contracts with decreasing of τ, and breaks further. The vertical part grows. After breaking the individual MSB $P2$–E (and respective collapse of the horizontal part of the cross), there is the vertical strip only with two additional parts determined by the joint actions of both pursuers.

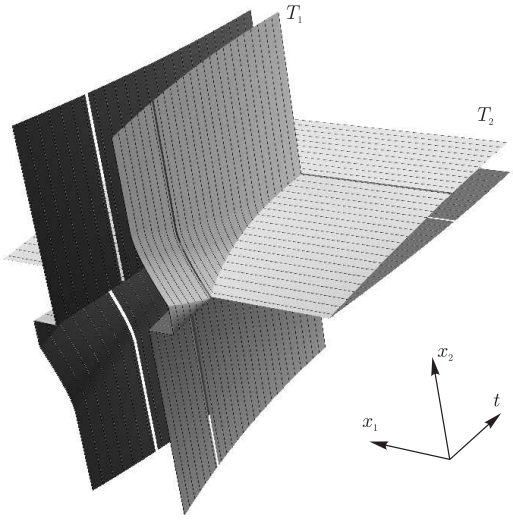


Figure 6. One strong and one weak pursuers, different termination instants: 3D-view of the set $W_{5.0}$

Varying advantage of pursuers. Consider a variant when both pursuers P_1 and P_2 are equal, with that at the beginning of the backward time, the bridges in the individual games contract and further expand. Choose the game parameters in such a way that for some c the section $W_c(t)$ of MSB W_c with decreasing of t disjoins into two parts, which join back with further decreasing of t.

Parameters of the game are $\mu_1 = \mu_2 = 1.1$, $\nu = 1$, $l_{P_1} = l_{P_2} = 1/0.6$, $l_E = 1$. Termination instants are equal: $T_1 = T_2 = 20$.

A three-dimensional view of MSB $W_{0.526}$ is shown in Fig. 7.

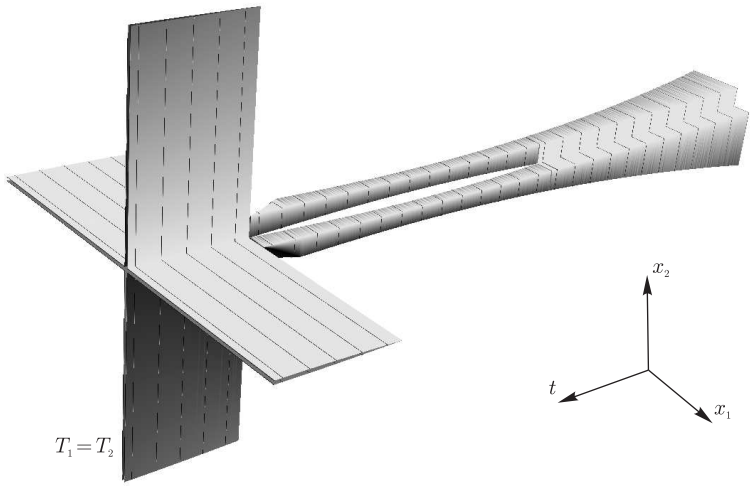


Figure 7. Varying advantage of the pursuers, equal termination instants: 3D-view of the maximal stable bridge $W_{0.526}$

5 Conclusion

The paper deals with numerical investigation of a differential game with two pursuers and one evader. With the help of the standard change of variables, the problem is reduced to a two-dimensional antagonistic game. The difficulty of solution is connected to non-convexity of the terminal payoff function. For typical variants of the game parameters, an analysis of the level sets (Lebesgue sets) of the value function is done. Three-dimensional views of the level sets are given.

References

1. A. I. Blagodatskih, N. N. Petrov. *Conflict Interaction Controlled Objects Groups*. Udmurt State University, Izhevsk, Russia, 2009. 266 pages. (in Russian)
2. P. Cardaliaguet, M. Quincampoix, P. Saint-Pierre. *Set-valued numerical analysis for optimal control and differential games*. Annals of ISDG, Vol. 4: Stochastic and Differential Games — Theory and Numerical Methods, M. Bardi, T. E. Raghavan, T. Parthasarathy (Eds.), Birkhauser, Boston, 1999. Pp. 177–247.
3. A. A. Chikrii. *Conflict-Controlled Processes*. Mathematics and its Applications, vol. 405. Kluwer Academic Publishers Group, Dordrecht, 1997. 424 pages.
4. E. Cristiani, M. Falcone. *Fully-discrete schemes for the value function of pursuit-evasion games with state constraints*. Annals of ISDG, Vol. 10: Advances

in Dynamic Games and Applications, P. Bernhard, V. Gaitsgory, O. Pourtallier (eds.), Birkhauser, Boston, 2009. Pp. 177–206.
5. N. L. Grigorenko. *The problem of pursuit by several objects.* Differential games — developments in modelling and computation (Espoo, 1990), *Lecture Notes in Control and Inform. Sci.*, vol. 156, Springer, Berlin, 1991. Pp. 71–80.
6. N. N. Krasovskii, A. I. Subbotin. *Positional Differential Games.* Nauka, Moscow, 1974. 456 pages. (in Russian)
7. N. N. Krasovskii, A. I. Subbotin. *Game-Theoretical Control Problems.* Springer-Verlag, New York, 1988. 518 pages.
8. S. Le Ménec. *Linear differential game with two pursuers and one evader.* Annals of ISDG, Vol. 11: Advances in Dynamic Games. Theory, Applications, and Numerical Methods for Differential and Stochastic Games, M. Breton, K. Szajowski (eds.), Birkhauser, Boston, 2011. Pp. 209–226.
9. I. Mitchell. *Application of level set methods to control and reachability problems in continuous and hybrid systems.* Ph.D. thesis, Stanford University, 2002. 127 pages.
10. B. N. Pschenichnyi. *Simple pursuit by several objects.* Kibernetika **3**, 1976, Pp.145–146. (in Russian)
11. T. Shima, J. Shinar. *Time-varying linear pursuit-evasion game models with bounded controls.* J. Guid. Control Dynam. **25**(3), 2002. Pp. 425–432.
12. J. Shinar, T. Shima. *Non-orthodox guidance law development approach for intercepting maneuvering targets.* J. Guid. Control Dynam. **25**(4), 2002. Pp. 658–666.
13. A. M. Taras'ev, T. B. Tokmantsev, A. A. Uspenskii, V. N. Ushakov. *On procedures for constructing solutions in differential games on a finite interval of time.* J. Math. Sci. **139**(5), 2006. Pp. 6954–6975.

S. A. Ganebny

S. S. Kumkov

V. S. Patsko
Institute of Mathematics and Mechanics, S.Kovalevskaya str., 16, Ekaterinburg, 620990, Russia. Tel.: +7-343-3753444. Fax: +7-343-3742581. E-mail: patsko@imm.uran.ru

Stéphane Le Ménec
EADS / MBDA France, 1 avenue Réaumur, 92358 Le Plessis-Robinson Cedex, France,
Tel.: +33 (0)1 71 54 14 92. Fax: +33 (0)1 71 54 01 71.
E-mail: stephane.le-menec@mbda-systems.com

AN EXTREMAL PROPERTY OF THE INF- AND SUP-CONVOLUTIONS REGARDING THE STRONG MAXIMUM PRINCIPLE

V. V. Goncharov, T. J. Santos

Key words: strong maximum principle, convex variational problem, convolution, gauge function

AMS Mathematics Subject Classification: 49J10, 49J53, 49N15

Abstract. In this paper we continue investigations started in [6] concerning the extension of the variational Strong Maximum Principle for lagrangeans depending on the gradient through a Minkowski gauge. We essentially enlarge the class of comparison functions, which substitute the identical zero when the lagrangean is not longer strictly convex at the origin.

1 Introduction

The Strong Maximum Principle, a well known property of the elliptic partial differential equations (see, e.g., [5,8] and the bibliography therein), can be formulated in the variational setting as was done by A. Cellina in 2002. Extending the main result of his work [3] we consider the integral functional

$$\int_{\Omega} f\left(\rho_F\left(\nabla u\left(x\right)\right)\right) dx, \tag{1.1}$$

where $\Omega \subset \mathbb{R}^n$ is an open bounded connected domain; $f : \mathbb{R}^+ \to \mathbb{R}^+ \cup \{+\infty\}$, $f\left(0\right) = 0$, is a lower semicontinuous convex function; $F \subset \mathbb{R}^n$ is a convex closed bounded set with $0 \in \mathrm{int}F$ (*interior* of F), and $\rho_F\left(\cdot\right)$ is the *Minkowski functional* (*guage function*) associated to F,

$$\rho_F\left(\xi\right) := \inf\left\{\lambda > 0 : \xi \in \lambda F\right\}. \tag{1.2}$$

In the traditional sense the Strong Maximum Principle (SMP) for Eq. (1.1) means that there is no a nonconstant continuous minimizer of this functional on

The research is fulfilled in the framework of the Project "Variational Analysis: Theory and Applications", PTDC/MAT/111809/2009 financially supported by the portugues institutions FCT, COMPETE, QREN and the European Regional Development Fund (FEDER)..

$u_0(\cdot) + W_0^{1,1}(\Omega)$ (with some Sobolev function $u_0(\cdot)$), admitting its *minimal* (*maximal*) value in Ω. In the case of a *rotationally invariant* lagrangean ($\rho_F(\xi) = \|\xi\|$) and $n > 1$ it was proved in [3] that this property is valid if and only if the function $f(\cdot)$ is strictly convex and smooth at the origin, or, in other words, if the equalities

$$\partial f^*(0) = \{0\} \tag{1.3}$$

and

$$\partial f(0) = \{0\} \tag{1.4}$$

hold. Here, as usual, ∂f stands for the *subdifferential* of the function $f(\cdot)$ in the sense of Convex Analysis, and $f^*(\cdot)$ is the *Legendre-Fenchel tranform* (*conjugate*) of $f(\cdot)$. Observe that in the case $n = 1$ the smoothness of $f(\cdot)$ at zero (the condition Eq. (1.4)) is not necessary, and for validity of the SMP one needs to set only the assumption Eq. (1.3) unless the function $f(\cdot)$ is not affine near the origin.

In [6] we proved that under the same hypotheses on $f(\cdot)$ the Strong Maximum Principle remains valid for a general functional Eq. (1.1), where the gauge F is not assumed to be either rotund or smooth or symmetric. Furthermore, we tried to extend the SMP to the case when the condition Eq. (1.3) fails.

Since the SMP equivalently can be reformulated as a comparison property:

if a continuous nonnegative (nonpositive) minimizer $u(\cdot)$ **of the functional** (1.1) on $u_0(\cdot) + W_0^{1,1}(\Omega)$ **touches zero at some point** $x^* \in \Omega$ **then necessarily** $u(x) \equiv 0$,

(1.5)

it is obviously violated whenever the lagrangean is no longer strictly convex at the origin (see also [2]). Nevertheless, we emphasized a class $\mathfrak{C}$ of continuous functions, which being themselves solutions of the variational problem can substitute in some sense the *identical zero* in the property Eq. (1.5). These functions (further called *test*, or *comparison,* functions) depend certainly on the subdifferential $\partial f^*(0)$ and reduce to the constants when $\partial f^*(0)$ reduces to the singleton $\{0\}$. (In the place of the null-function in Eq. (1.5), clearly, any constant can stand).

In the case $\partial f^*(0) \neq \{0\}$ so extended Strong Maximum Principle for a test function $\hat{u}(\cdot) \in \mathfrak{C}$ can be given as follows:

each continuous minimizer of (1.1) **on** $u_0(\cdot) + W_0^{1,1}(\Omega)$ **such that** $u(x) \geqslant \hat{u}(x)$ (**respectively,** $u(x) \leqslant \hat{u}(x)$), $x \in \Omega$, **having the same points of** *local minimum* (**respectively,** *local maximum*) **as** $\hat{u}(\cdot)$ **should coincide with** $\hat{u}(\cdot)$ **everywhere on** Ω.

(1.6)

Observe that all the functions $\hat{u}(\cdot) \in \mathfrak{C}$ are written in terms of the *polar set* F^0. Namely, in the simplest case of unique local minimum (local maximum) point $x_0 \in \Omega$ the functions $\hat{u}^+_{x_0,\mu}(x) := \mu + a\rho_{F^0}(x - x_0)$ and $\hat{u}^-_{x_0,\mu}(x) := \mu - a\rho_{F^0}(x_0 - x)$ belong to $\mathfrak{C}$ for each real μ. Here $a := \sup \partial f^*(0)$.

If instead $\partial f^*(0) = \{0\}$ (equivalently, $a = 0$) then we can take $x_0 = x^*$ where x^* is an arbitrary "floating" point from Ω (see Eq. (1.5)), and we arrive at the traditional SMP although the property Eq. (1.6) is not formally applicable.

In [6] also a "multipoint" version of the Strong Maximum Principle was established when the comparison function $\hat{u}(\cdot)$ is the lower (upper) envelope of a finite number of the functions $\hat{u}^+_{x_0,\mu}(\cdot)$ (respectively, $\hat{u}^-_{x_0,\mu}(\cdot)$) for various $x_0 \in \Omega$ and $\mu \in \mathbb{R}$. Notice that Eq. (1.6) takes place only for convex domains $\Omega \subset \mathbb{R}^n$ (or, at least, under a kind of *star-shapeness* hypothesis that can not be removed, see [6]). Another restriction, under which validity of the property Eq. (1.6) was proved, is smoothness of the gauge function $\rho_F(\cdot)$, or, equivalently, rotundity of the polar set F^0. In fact, one of the tools we use in the proofs is so named *modulus of rotundity*

$$\begin{aligned}\mathfrak{M}_{F^0}(r;\alpha,\beta) := \inf\{1 - \rho_{F^0}(\xi + \lambda(\eta - \xi)) :\\ \xi,\eta \in \partial F^0,\ \ \rho_{F^0}(\xi - \eta) \geqslant r,\ \ \alpha \leqslant \lambda \leqslant \beta\},\end{aligned} \tag{1.7}$$

which is strictly positive for all $r > 0$ and all $0 < \alpha \leqslant \beta < 1$ whenever F^0 is *rotund*.

In this paper we essentially enlarge the class $\mathfrak{C}$ envolving the infinite (continuous) envelopes of the functions $\hat{u}^\pm_{x_0,\mu}(\cdot)$ by such a way that the generalized SMP gets an *unique extremal extension principle* and unifies both properties Eq. (1.5) and Eq. (1.6). Namely, given an arbitrary function $\vartheta(\cdot)$ defined on a closed subset $\Gamma \subset \Omega$ and satisfying a natural *slope condition* w.r.t. F we prove in Section 3 that the *inf-convolution*

$$u^+_{\Gamma,\vartheta}(x) := \inf_{y\in\Gamma}\{\vartheta(y) + a\rho_{F^0}(x - y)\} \tag{1.8}$$

(respectively, the *sup-convolution*

$$u^{-}_{\Gamma,\vartheta}(x) := \sup_{y\in\Gamma}\{\vartheta(y) - a\rho_{F^0}(y-x)\}) \tag{1.9}$$

is the only continuous minimizer $u(\cdot)$ of the functional Eq. (1.1) on $u_0(\cdot)+W_0^{1,1}(\Omega)$ such that $u(x) = \vartheta(x)$ on Γ and $u(x) \geqslant u^{+}_{\Gamma,\vartheta}(x)$ (respectively, $u(x) \leqslant u^{-}_{\Gamma,\vartheta}(x)$), $x \in \Omega$. The domain Ω is always assumed to be convex.

2 Preliminaries. Auxiliary statements

In what follows we assume that $a := \sup \partial f^*(0) > 0$, and so the second Cellina's hypothesis Eq. (1.4) is automatically fulfilled. Furthermore, we introduce the nondecreasing upper semicontinuous function $\varphi(t) := \sup \partial f^*(t)$. So $\varphi(0) = a$ and $\varphi(t) < +\infty$ on the interior of the domain $\mathrm{dom} f^* := \{t \in \mathbb{R}^+ : f^*(t) < +\infty\}$. The version of SMP we wish to prove is essentially based on the following *a priori* local estimates of continuous minimizers of Eq. (1.1) obtained in [6] by using the dual properties of convex sets (see, e.g., [9] or [7])) being themselves an interesting result of Convex Analysis.

Theorem 1. *Given an open bounded region $\Omega \subset \mathbb{R}^n$, $n \geqslant 1$, and a continuous admissible minimizer $\bar{u}(\cdot)$ of the functional Eq. (1.1) on $u_0(\cdot) + W_0^{1,1}(\Omega)$, assume a point $\bar{x} \in \Omega$ and real numbers $\beta > 0$ and μ to be such that $\bar{u}(x) \geqslant \mu \ \ \forall x \in \bar{x} - \beta F^0 \subset \Omega$ and $\bar{u}(\bar{x}) > \mu + a\beta$. Then for some $\eta > 0$ the inequality*

$$\bar{u}(x) \geqslant \mu + \varphi(\eta)(\beta - \rho_{F^0}(\bar{x} - x)) \tag{2.1}$$

holds for all $x \in \bar{x} - \beta F^0$.

Simmetrically, if a point $\bar{x} \in \Omega$ and numbers $\beta > 0$ and μ are such that

$$\bar{u}(x) \leqslant \mu \ \ \forall x \in \bar{x} + \beta F^0 \subset \Omega$$

and $\bar{u}(\bar{x}) < \mu - a\beta$, then there exists $\eta > 0$ such that

$$\bar{u}(x) \leqslant \mu - \varphi(\eta)(\beta - \rho_{F^0}(x - \bar{x})) \tag{2.2}$$

for all $x \in \bar{x} + \beta F^0$.

Roughly speaking, the statement above means that for each continuous admissible minimizer $\bar{u}(\cdot)$ of Eq. (1.1) and for each point $\bar{x} \in \Omega$, which is not local extremum for $\bar{u}(\cdot)$, the deviation of $\bar{u}(\cdot)$ from the extremal level can be controlled near $\bar{x}$ by an affine transformation of the dual Minkowski gauge (see Eq. (2.1) and

Eq. (2.2)). Recall that *admissible minimizers* are those, which give finite values to the functional Eq. (1.1).

In the case $a > 0$ (it is our standing assumption along with the paper) we have the following simple consequence of this theorem.

Corollary 1. Given $\Omega \subset \mathbb{R}^n$, $n \geqslant 1$, and $\bar{u}(\cdot)$ as in Theorem 1 let us assume that for some $x_0 \in \Omega$ and $\delta > 0$

$$\bar{u}(x) \geqslant \bar{u}(x_0) + a\rho_{F^0}(x - x_0) \ \ \forall x \in x_0 + \delta F^0 \subset \Omega. \tag{2.3}$$

Then

$$\bar{u}(x) = \bar{u}(x_0) + a\rho_{F^0}(x - x_0) \ \ \forall x \in x_0 + \frac{\delta}{\|F\| \, \|F^0\| + 1} F^0.$$

Similarly, if in the place of Eq. (2.3)

$$\bar{u}(x) \leqslant \bar{u}(x_0) - a\rho_{F^0}(x_0 - x) \ \ \forall x \in x_0 - \delta F^0 \subset \Omega \tag{2.4}$$

then

$$\bar{u}(x) = \bar{u}(x_0) - a\rho_{F^0}(x_0 - x) \ \ \forall x \in x_0 - \frac{\delta}{\|F\| \, \|F^0\| + 1} F^0.$$

Here $\|F\| := \sup\{\|\xi\| : \xi \in F\}$.

As the standing hypotheses in what follows we assume that $F \subset \mathbb{R}^n$ is a convex closed bounded set, $0 \in \mathrm{int}F$, with the *smooth boundary* (the latter means that the Minkowski functional $\rho_F(\xi)$ is *Fréchet differentiable* at each $\xi \neq 0$), and that $\Omega \subset \mathbb{R}^n$ is an open convex bounded region.

Let us consider an arbitrary nonempty closed subset $\Gamma \subset \Omega$ and a function $\vartheta : \Gamma \to \mathbb{R}$ satisfying the *slope condition*:

$$\vartheta(x) - \vartheta(y) \leqslant a\sigma_F(x - y) \quad \forall x, y \in \Gamma, \tag{2.5}$$

where

$$\sigma_F(\xi) := \sup_{v \in F} \langle v, \xi \rangle$$

is the *support function* of F ($\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{R}^n$). It is well known that

— $\left(F^0\right)^0 = F$;
— $\sigma_F(\xi) = \rho_{F^0}(\xi)$ whenever $\xi \in F^0$;
— the polar set F^0 is rotund (see Section 1).

We will use also the following property of the gauge function:

$$\frac{1}{\|F\|}\|\xi\| \leqslant \rho_F(\xi) \leqslant \left\|F^0\right\|\|\xi\|,\ \xi \in \mathbb{R}^n. \tag{2.6}$$

Let us define now inf- and sup-*convolutions* of $\vartheta(\cdot)$ with the gauge function $a\rho_{F^0}(\cdot)$ by the formulas Eq. (1.8) and Eq. (1.9). We observe first that the function $u^{\pm}_{\Gamma,\vartheta}(\cdot)$ is the minimizer of Eq. (1.1) on $u^{\pm}_{\Gamma,\vartheta}(\cdot) + W^{1,1}_0(\Omega)$. Indeed, it is obviously Lipschitz continuous on Ω, and for its (classical) gradient $\nabla u^{\pm}_{\Gamma,\vartheta}$ existing by Rademacher's theorem we have that

$$\nabla u^{\pm}_{\Gamma,\vartheta}(x) \in \partial^c u^{\pm}_{\Gamma,\vartheta}(x) \subset aF$$

for a.e. $x \in \Omega$ (see [4, Theorem 2.8.6]). Here ∂^c stands for the *Clarke's subdifferential* of a (locally) Lipschitzean function. Consequently, $f\left(\rho_F\left(\nabla u^{\pm}_{\Gamma,\vartheta}(x)\right)\right) = 0$ a.e. on Ω, and the function $u^{\pm}_{\Gamma,\vartheta}(\cdot)$ gives to Eq. (1.1) the minimal possible value zero. Due to the slope condition Eq. (2.5) it follows also that $u^{\pm}_{\Gamma,\vartheta}(x) = \vartheta(x)$ for all $x \in \Gamma$. Moreover, $u^{\pm}_{\Gamma,\vartheta}(\cdot)$ is the (unique) *viscosity solution* of the *Hamilton-Jacobi equation*

$$\pm\left(\rho_F(\nabla u(x)) - a\right) = 0,\ u|_{\Gamma} = \vartheta,$$

(see, e.g., [1]).

Notice that Γ can be a finite set, say $\{x_1, x_2, ..., x_m\}$, in which case $\vartheta(\cdot)$ associates to each x_i a real number ϑ_i, $i = 1, ..., m$, and the condition Eq. (2.5) slightly strengthened (by assuming that the inequality in Eq. (2.5) is strict for $x_i \neq x_j$) means that all the simplest test functions $\vartheta_i + a\rho_{F^0}(x - x_i)$ (respectively, $\vartheta_i - a\rho_{F^0}(x_i - x)$) are essential (not superfluous) in constructing of the respective lower or upper envelope. Then the extremal property established below is reduced to the extended SMP Eq. (1.6) (see [6, Theorem 6]).

On the other hand, if $\vartheta(\cdot)$ is a Lipschitz continuous function defined on a closed convex set $\Gamma \subset \Omega$ with nonempty interior then Eq. (2.5) holds iff $\nabla\vartheta(x) \in aF$ for almost each (a.e.) $x \in \Gamma$. This immediately follows from Lebourg's mean value theorem (see [4, p. 41]) recalling the properties of the Clarke's subdifferential and from the separability theorem.

Certainly, the mixed (discrete and continuous) case can be considered as well, and all the situations are unified by the hypothesis Eq. (2.5).

In the particular case $\vartheta \equiv 0$ (Eq. (2.5) is trivially fulfilled) the function $u^{+}_{\Gamma,\vartheta}(x)$ is nothing else than the *minimal time* necessary to achieve the closed set Γ from the point $x \in \Omega$ by trajectories of the differential inclusion with the constant convex

right-hand side

$$-a\dot{x}(t) \in F^0, \tag{2.7}$$

while $-u^-_{\Gamma,\vartheta}(x)$ is, contrarily, the *minimal time*, for which trajectories of Eq. (2.7) arrive at x starting from a point of Γ. Furthermore, if $F = \overline{B}$ is the closed unit ball centred at the origin then the gauge function $\rho_{F^0}(\cdot)$ is the euclidean norm in $\mathbb{R}^n$, and we have $u^{\pm}_{\Gamma,\vartheta}(x) = \pm a d_\Gamma(x)$ where $d_\Gamma(\cdot)$ means the *distance* from a point to the set Γ.

3 Generalized Strong Maximum Principle

Now we are ready to deduce the extremal property of the functions $u^{\pm}_{\Gamma,\vartheta}(\cdot)$ announced above.

Theorem 2. *Under all the standing hypotheses formulated in the previous section let us assume that a continuous admissible minimizer $\bar{u}(\cdot)$ of the functional Eq. (1.1) on $u_0(\cdot) + W^{1,1}_0(\Omega)$ is such that*

(i) $\bar{u}(x) = u^+_{\Gamma,\vartheta}(x) = \vartheta(x) \quad \forall x \in \Gamma$;
(ii) $\bar{u}(x) \geqslant u^+_{\Gamma,\vartheta}(x)\ \forall x \in \Omega$.

Then $\bar{u}(x) \equiv u^+_{\Gamma,\vartheta}(x)$ on Ω.

Simmetrically, if a continuous admissible minimizer $\bar{u}(\cdot)$ satisfies the conditions

(i)′ $\bar{u}(x) = u^-_{\Gamma,\vartheta}(x) = \vartheta(x) \quad \forall x \in \Gamma$;
(ii)′ $\bar{u}(x) \leqslant u^-_{\Gamma,\vartheta}(x)\ \forall x \in \Omega$,

then $\bar{u}(x) \equiv u^-_{\Gamma,\vartheta}(x)$ on Ω.

Proof. Let us prove the first part of Theorem only since the respective changements in the symmetric case are obvious.

Given a continuous admissible minimizer $\bar{u}(\cdot)$ satisfying the conditions (**i**) and (**ii**) we suppose, on the contrary, that there exists $x \in \Omega \setminus \Gamma$ with $\bar{u}(x) > u^+_{\Gamma,\vartheta}(x)$. Notice first that without loss of generality one can assume that the latter (strict) inequality holds for all $x \in \Omega \setminus \Gamma \neq \emptyset$.

Indeed, denoting by $\Gamma^+ := \left\{x \in \Omega : \bar{u}(x) = u^+_{\Gamma,\vartheta}(x)\right\}$ we claim that

$$u^+_{\Gamma,\vartheta}(x) = \inf_{y \in \Gamma^+} \left\{\bar{u}(y) + a\rho_{F^0}(x-y)\right\} \tag{3.1}$$

for each $x \in \Omega$. Since $\Gamma^+ \supset \Gamma$ and $\bar{u}(y) = \vartheta(y)$, $y \in \Gamma$, the inequality "$\geqslant$" in Eq. (3.1) is obvious. On the other hand, given $x \in \Omega$ let us take an arbitrary $y \in \Gamma^+$.

Then due to the compactness of Γ we find $y^* \in \Gamma$ such that

$$\bar{u}(y) = \vartheta(y^*) + a\rho_{F^0}(y - y^*), \tag{3.2}$$

and by the triangle inequality

$$\bar{u}(y) + a\rho_{F^0}(x - y) \geqslant \vartheta(y^*) + a\rho_{F^0}(x - y^*) \geqslant u^+_{\Gamma,\vartheta}(x). \tag{3.3}$$

Passing to infimum in Eq. (3.3) we prove the inequality "$\leqslant$" in Eq. (3.1) as well. Furthermore, for arbitrary $x, y \in \Gamma^+$ and for $y^* \in \Gamma$ satisfying Eq. (3.2) we have

$$\bar{u}(x) - \bar{u}(y) = u^+_{\Gamma,\vartheta}(x) - u^+_{\Gamma,\vartheta}(y) \leqslant a\sigma_F(x - y). \tag{3.4}$$

Hence, we can extend the function $\vartheta : \Gamma \to \mathbb{R}$ onto the (closed) set $\Gamma^+ \subset \Omega$ by setting $\vartheta(x) = \bar{u}(x)$, $x \in \Gamma^+$, and all the conditions remain valid (see Eq. (3.4) and Eq. (3.1)).

Notice that the *convex hull* $K := \mathrm{co}\Gamma$ is the compact set contained in Ω (due to the convexity hypothesis). Let us choose now $\varepsilon > 0$ such that $K \pm \varepsilon F^0 \subset \Omega$ and denote by

$$\delta := 2\varepsilon\mathfrak{M}_{F^0}\left(\frac{2\varepsilon}{\Delta}; \frac{\varepsilon}{\varepsilon + \Delta}, \frac{\Delta}{\varepsilon + \Delta}\right) > 0,$$

where $\mathfrak{M}_{F^0}$ is the modulus of rotundity associated to F^0 (see Eq. (1.7)) and

$$\Delta := \sup_{\xi,\eta\in\Omega} \rho_{F^0}(\xi - \eta)$$

is the ρ_{F^0}-diameter of the region Ω. Similarly as in [6] (see Step 1 of the proof of Theorem 5) we show that

$$\rho_{F^0}(y_1 - x) + \rho_{F^0}(x - y_2) - \rho_{F^0}(y_1 - y_2) \geqslant \delta \tag{3.5}$$

whenever $y_1, y_2 \in \Gamma$ and $x \in \Omega \setminus \left[\left(K + \varepsilon F^0\right) \cup \left(K - \varepsilon F^0\right)\right]$. Indeed, we obviously have $\varepsilon \leqslant \rho_1 := \rho_{F^0}(y_1 - x) \leqslant \Delta$ and $\varepsilon \leqslant \rho_2 := \rho_{F^0}(x - y_2) \leqslant \Delta$, and, consequently,

$$\lambda := \frac{\rho_2}{\rho_1 + \rho_2} \in \left[\frac{\varepsilon}{\varepsilon + \Delta}, \frac{\Delta}{\varepsilon + \Delta}\right]. \tag{3.6}$$

Setting $\xi_1 := (y_1 - x)/\rho_1$ and $\xi_2 := (x - y_2)/\rho_2$ we can write

$$\xi_1 - \xi_2 = (1/\rho_1 + 1/\rho_2)\left((\rho_2/(\rho_1 + \rho_2))\, y_1 + (\rho_1/(\rho_1 + \rho_2))\, y_2 - x\right),$$

and hence

$$\rho_{F^0}(\xi_1 - \xi_2) \geqslant (1/\rho_1 + 1/\rho_2)\,\varepsilon \geqslant 2\varepsilon/\Delta. \tag{3.7}$$

On the other hand,

$$\begin{aligned} & \rho_{F^0}(y_1 - x) + \rho_{F^0}(x - y_2) - \rho_{F^0}(y_1 - y_2) \\ = \;& (\rho_1 + \rho_2)\left[1 - \rho_{F^0}\left(\rho_1/(\rho_1 + \rho_2)\,\xi_1 + \rho_2/(\rho_1 + \rho_2)\,\xi_2\right)\right] \geqslant \\ \geqslant \;& 2\varepsilon\left[1 - \rho_{F^0}(\xi_1 + \lambda(\xi_2 - \xi_1))\right]. \end{aligned} \tag{3.8}$$

Combining Eq. (3.6) - Eq. (3.8) and the definition of the rotundity modulus (1.7) we arrive at Eq. (3.5).

Let us fix $\bar{x} \in \Omega \setminus \Gamma$ and $\bar{y} \in \Gamma$ such that

$$u^+_{\Gamma,\vartheta}(\bar{x}) = \vartheta(\bar{y}) + a\rho_{F^0}(\bar{x} - \bar{y}).$$

Then by Lemma 1 [6] the point $\bar{y}$ is also a minimizer on Γ of the function $y \mapsto \vartheta(y) + a\rho_{F^0}(x_\lambda - y)$ where $x_\lambda := \lambda\bar{x} + (1-\lambda)\,\bar{y}$, $\lambda \in [0,1]$, i.e.,

$$u^+_{\Gamma,\vartheta}(x_\lambda) = \vartheta(\bar{y}) + a\rho_{F^0}(x_\lambda - \bar{y}). \tag{3.9}$$

Define now the Lipschitz continuous function

$$\bar{v}(x) := \max\left\{\bar{u}(x), \min\left\{\vartheta(\bar{y}) + a\rho_{F^0}(x - \bar{y}), \vartheta(\bar{y}) + a(\delta - \rho_{F^0}(\bar{y} - x))\right\}\right\} \tag{3.10}$$

and claim that $\bar{v}(\cdot)$ minimizes the functional Eq. (1.1) on the set $\bar{u}(\cdot) + W_0^{1,1}(\Omega)$. In order to prove this we observe first that for each $x \in \Omega$, $x \notin K \pm \varepsilon F^0$, and for each $y \in \Gamma$ by the slope condition Eq. (2.5) and by Eq. (3.5) the inequality

$$\begin{aligned} & \vartheta(y) + a\rho_{F^0}(x - y) - \vartheta(\bar{y}) + a\rho_{F^0}(\bar{y} - x) \\ \geqslant \;& a\left(\rho_{F^0}(\bar{y} - x) + \rho_{F^0}(x - y) - \rho_{F^0}(\bar{y} - y)\right) \geqslant a\delta \end{aligned} \tag{3.11}$$

holds. Passing to infimum in Eq. (3.11) for $y \in \Gamma$ and taking into account the basic assumption (**ii**) we have

$$\bar{u}(x) \geqslant \inf_{y\in\Gamma}\left\{\vartheta(y) + a\rho_{F^0}(x - y)\right\} \geqslant \vartheta(\bar{y}) + a(\delta - \rho_{F^0}(\bar{y} - x)),$$

and, consequently, $\bar{v}(x) = \bar{u}(x)\ \ \forall x \in \Omega \setminus \left[\left(K + \varepsilon F^0\right) \cup \left(K - \varepsilon F^0\right)\right]$. In particular, $\bar{v}(\cdot) \in \bar{u}(\cdot) + W_0^{1,1}(\Omega)$. Furthermore, setting $\Omega' := \{x \in \Omega : \bar{v}(x) \neq \bar{u}(x)\}$ by the well known property of the support function we have $\nabla\bar{v}(x) \in aF$ for a.e. $x \in \Omega'$,

while $\nabla\bar{v}(x) = \nabla\bar{u}(x)$ for a.e. $x \in \Omega \setminus \Omega'$. Then

$$\begin{aligned}\int_{\Omega} f(\rho_F(\nabla\bar{v}(x)))\,dx &= \int_{\Omega\setminus\Omega'} f(\rho_F(\nabla\bar{u}(x)))\,dx \\ &\leqslant \int_{\Omega} f(\rho_F(\nabla\bar{u}(x)))\,dx \leqslant \int_{\Omega} f(\rho_F(\nabla u(x)))\,dx\end{aligned}$$

for each $u(\cdot) \in \bar{u}(\cdot) + W_0^{1,1}(\Omega)$.

Finally, setting $\mu := \min\left\{\varepsilon, \delta/\left(\|F\|\left\|F^0\right\| + 1\right)^2\right\}$ we see that the minimizer $\bar{v}(\cdot)$ satisfies on $\bar{y} + \mu\left(\|F\|\left\|F^0\right\| + 1\right)F^0$ the inequality

$$\bar{v}(x) \geqslant \vartheta(\bar{y}) + a\rho_{F^0}(x - \bar{y}). \tag{3.12}$$

Indeed, it follows from Eq. (2.6) that

$$\rho_{F^0}(x - \bar{y}) + \rho_{F^0}(\bar{y} - x) \leqslant \mu\left(\|F\|\left\|F^0\right\| + 1\right)^2 \leqslant \delta$$

whenever $\rho_{F^0}(x - \bar{y}) \leqslant \mu\left(\|F\|\left\|F^0\right\| + 1\right)$, implying that the minimum in Eq. (3.10) is equal to $\vartheta(\bar{y}) + a\rho_{F^0}(x - \bar{y})$. Since, obviously, $\bar{v}(\bar{y}) = \vartheta(\bar{y})$, applying Corollary 1 we deduce from Eq. (3.12) that

$$\bar{v}(x) = \vartheta(\bar{y}) + a\rho_{F^0}(x - \bar{y})$$

for all $x \in \bar{y} + \mu F^0 \subset K + \varepsilon F^0 \subset \Omega$. Comparing with Eq. (3.10), we have

$$\bar{u}(x) \leqslant \vartheta(\bar{y}) + a\rho_{F^0}(x - \bar{y}), \quad x \in \bar{y} + \mu F^0. \tag{3.13}$$

However, for some $\lambda_0 \in [0, 1]$ the points x_λ, $0 \leqslant \lambda \leqslant \lambda_0$, belong to $\bar{y} + \mu F^0$. Combining with Eq. (3.13) and taking into account the equality Eq. (3.9) we obtain

$$\bar{u}(x_\lambda) \leqslant u^+_{\Gamma,\vartheta}(x_\lambda)$$

and hence (see the hypothesis (**ii**))

$$\bar{u}(x_\lambda) = u^+_{\Gamma,\vartheta}(x_\lambda),$$

$0 \leqslant \lambda \leqslant \lambda_0$. This is contradiction because the inequality in (**ii**) is strict outside the set Γ. □

References

1. P. Cardaliaguet, B. Dacorogna, W. Gangbo and N. Georgy, *Geometric restrictions for the existence of viscosity solutions*, Ann. Inst. Henri Poicaré **16**, 1999, Pp. 189-220.
2. A. Cellina, *On minima of a functional of the gradient: sufficient conditions*, Nonlin. Anal.: Theory, Meth. and Appl. **20**, 1993, Pp. 343-347.
3. A. Cellina, *On the Strong Maximum Principle*, Proc. Amer. Math. Soc. **130**, 2002, 413-418.
4. F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983. 309 pages.
5. D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, New York, 1998. 523 pages.
6. V. V. Goncharov and T. J. Santos, *Local estimates for minimizers of some convex integral functional of the gradient and the Strong Maximum Principle*, Set-Valued and Var. Anal. **19**, 2011, Pp. 179-202.
7. R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, Lecture Notes in Math. **1364**, Springer, New York, 1989. 117 pages.
8. P. Pucci, J. Serrin, *The Strong Maximum Principle*, Progress in Nonlinear Differential Equations and their Applications. **73**, Birkhauser, Switzerland, 2007. 235 pages.
9. R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New York, 1972. 468 pages.

V. V. Goncharov
CIMA-UE (Centro de Investigação em Matemática e Aplicações da Universidade de Évora), rua Romão Ramalho 59, 7000-671, Évora, Portugal; phone +351 936514934; Fax: +351 266745393; e-mail: goncha@uevora.pt

T. J. Santos
CIMA-UE, rua Romão Ramalho 59, 7000-671, Évora, Portugal; e-mail: tjfs@uevora.pt

EXHAUSTERS AND IMPLICIT FUNCTIONS IN NONSMOOTH SYSTEMS

G. Y. Murzabekova

Key words: implicit functions; upper and lower exhausters; positively homogeneous function; directionally differentiable functions

AMS Mathematics Subject Classification: 90C30, 52A20

Abstract. The paper is related to Implicit Functions in Nonsmooth Analysis. Recently Implicit Functions were treated by means of upper and lower exhausters as new tools of Nonsmooth analysis. This notion appeared to be very useful for solution of a number of other problems of nonsmooth analysis as well, including research of nonsmooth implicit functions and nonsmooth equations set. Implicit functions for continuous nondifferentiable functions have been studied by J.Warga, for Lipschitz functions — by F.Clarke, V.Demyanov, A.Ioffe, for quasidifferentiable functions — by V.Demyanov.

1 Introduction

Let $f_i(x, y)$ $(i \in 1 : n)$ be continuous jointly in all variables on $S = S_1 \times S_2 \subset \mathbb{R}^m \times \mathbb{R}^n$, where $S_1 \subset \mathbb{R}^m$ and $S_2 \subset \mathbb{R}^n$ are open sets. Put $f = (f_1, \ldots, f_n)$.

Consider the system $f_i(x, y) = 0 \quad \forall\, i \in 1 : n$.

In the nonsmooth case it makes sense to introduce a directional implicit function. Fix a direction $g \in \mathbb{R}^m$, $g \neq 0$, and consider the system $f_i(x_0+g, y) = 0 \quad \forall\, i \in 1 : n$.

We say that there exists an implicit function in the direction g if $\alpha_0 > 0$ and a vector function $y(\alpha)$ given on $[0, \alpha_0]$ exists such that

$$y(\alpha) \overrightarrow{\alpha \downarrow 0}\, y_0, \quad f(x_0 + \alpha g, y(\alpha)) = 0_n \quad \forall \alpha \in [0, \alpha_0].$$

2 Exhausters of the positively homogenious function

V. Demyanov (see [1]) introduced the notion of exhauster, which is helpful in solving various problems in nonsmooth analysis. It is useful to formulate necessary and sufficient conditions of extremum, to find steepest descent (ascent) directions, calculus of exhausters has been developed.

Since every $\overline{h} \in \Lambda^*$ is a convex positively homogenious (p.h.) function then there exists a unique convex compact set $C(\overline{h}) \in \mathbb{R}^n$ such that

$$\overline{h}(g) = \max_{v \in C(\overline{h})} (v, g) \quad \forall\, g \in \mathbb{R}^n.$$

Therefore (1) can be represented as $h(g) = \inf_{C \in E^*} \max_{v \in C}(v, g) \quad \forall\, g \in \mathcal{K}$, where $E^* = \{C \subset \mathbb{R}^n \mid C = C(\overline{h}), \overline{h} \in \Lambda^*\}$. Then the family of sets $E^* = E^*(h)$ is called an upper exhauster of the function h w.r. to the cone $\mathcal{K}$.

Analogously, if h is lower semicontinuous then there exists a family Λ_* of lower semicontinuous approximation's (l.c.a.) satisfying (2). Since every $\underline{h} \in \Lambda_*$ is a concave p.h. function, hence, there exists a unique convex compact set $C(\underline{h}) \in \mathbb{R}^n$ such that

$$\underline{h}(g) = \min_{w \in C(\underline{h})} (w, g) \quad \forall\, g \in \mathbb{R}^n.$$

Therefore (2) can be represented as $h(g) = \sup_{C \in E_*} \min_{w \in C}(w, g) \quad \forall\, g \in \mathcal{K}$, where $E_* = \{C \subset \mathbb{R}^n \mid C = C(\underline{h}), \underline{h} \in \Lambda_*\}$. The family of sets $E_* = E_*(h)$ is called a lower exhauster of the function h w.r. to the cone $\mathcal{K}$.

If a function h is p.h. and continuous on $\mathcal{K}$ then it is both upper and lower semicontinuous and, hence, both an upper exhauster $E^*(h)$ and a lower one $E_*(h)$ exist. The pair $E(h) = [E^*(h), E_*(h)]$ is called a biexhauster of the function h w.r. to the cone $\mathcal{K}$. Note that each of the sets $E^*(h)$ and $E_*(h)$ is a family of convex compact sets.

M.Castellani (see [2]) demonstrates that, generally speaking, one may express h in the forms

$$h(g) = \min_{C \in E^*} \sup_{v \in C} (v, g) \quad \forall\, g \in \mathbb{R}^n$$

and

$$h(g) = \max_{C \in E_*} \inf_{w \in C} (w, g) \quad \forall\, g \in \mathbb{R}^n,$$

where $E^*(h)$ and $E_*(h)$ are some families of convex compact sets of $\mathbb{R}^n$.

Thus, if $h : \mathbb{R}^n \to \mathbb{R}$ is a positively homogeneous and continuous function then function h can be represented as

$$h(g) = \min_{C \in E^*} \max_{v \in C} (v, g) \quad \forall\ g \in \mathbb{R}^n \tag{2.1}$$

and

$$h(g) = \max_{C \in E_*} \min_{w \in C} (w, g) \quad \forall\ g \in \mathbb{R}^n, \tag{2.2}$$

where $E^* = E^*(h)$ is an upper exhauster of function h, and the family of sets $E_* = E_*(h)$ is a lower exhauster of function h.

3 Implicit function theorems

Assume that all functions $f_i(z)$ are directionally differentiable at a point $z_0 = [x_0, y_0]$ and directional derivative $\tilde{h}_i(\eta) = f_i'(z_0, \eta)$, where $\eta = [g, q] \in \mathbb{R}^{m+n}$, continuous as function of η and bounded from above. Then from (1) the following expansions hold

$$f_i(z_0 + \alpha\eta) = f_i(z_0) + \alpha\tilde{h}_i(\eta) + o_{\eta i}(\alpha) \tag{3.1}$$

where

$$\tilde{h}_i(\eta) = \min_{\tilde{C}_i \in \tilde{E}_i^*} \max_{v \in \tilde{C}_i} (v, \eta), \quad \frac{o_{\eta i}(\alpha)}{\alpha} \underset{\alpha \downarrow 0}{\longrightarrow} 0 \quad \forall \eta \in \mathbb{R}^{m+n}, \forall\, i \in 1:n,$$

$$\tilde{E}_i^* \text{ is an upper exhauster of function } \tilde{h}_i.$$

Put $h_i(q) = \tilde{h}_i(g, q)$, g is fixed.

Thus, in order to solve the problem of existence and to study properties of an implicit function in the direction g one should find all solutions of the following system

$$h_i(q) = \min_{\tilde{C}_i \in \tilde{E}_i^*} \max_{v_i \in \tilde{C}_i} [(v_{i1}, g) + (v_{i2}, q)] = 0 \quad \forall\, i \in 1:n, \tag{3.2}$$

where $v_i \in [v_{i1}, v_{i2}]$. System (4) called quasilinear.

Introduce the function

$$F_i(\alpha, q) = \begin{cases} \frac{1}{\alpha} f_i(x_0 + \alpha g, y_0 + \alpha q), & \alpha > 0, \\ \tilde{h}_i(g, q), & \alpha = 0. \end{cases} \tag{3.3}$$

It follows from (1) that $F_i(\alpha, q) = \tilde{h}_i(g, q) + r_i(\alpha, q)$, where $r_i(\alpha, q) = \frac{o_i(\alpha g, \alpha q)}{\alpha} \underset{\alpha \downarrow 0}{\longrightarrow} 0 \quad \forall\, q \in \mathbb{R}^n$, $r_i(0, q) = 0$.

Put $f = (f_1, \ldots, f_n)$, $\tilde{h} = (\tilde{h}_1, \ldots, \tilde{h}_n)$, $\eta_0 = [g, q_0]$. Let $q_0 \in \mathbb{R}^n$ be a solution of the system (4), i.e.

$$\tilde{h}_i(\eta_0) = 0 \quad \forall\, i \in 1:n \quad (\text{or } \tilde{h}(\eta_0) = 0_n). \tag{3.4}$$

The functions F_i are continuous jointly in all variables when $\alpha \geqslant 0, \quad q \in \mathbb{R}^n$. We have $F_i(0, q_0) = 0$. Denote $q = q_0 + \Delta q$ and consider

$$h_i(q_0 + \Delta q) = \min_{C_{i2} \in E^*_{i2}} \max_{v_i \in C_{i2}} (v_{i2}, \Delta q) + o_i(\Delta q),$$

where $E^*_{i2} = E^*_{i2}(q_0)$ is an upper exhauster of function h at a point q_0 (there is a family of convex compact sets in $\mathbb{R}^n$). Clearly,

Introduce the set of matrices

$$\mathcal{L}(q_0) = \text{cl co} \left\{ A = \begin{pmatrix} a_1^T \\ \vdots \\ a_n^T \end{pmatrix} \;\middle|\; a_i \in C_{i2}, \quad C_{i2} \in E^*_{i2}(q_0), \quad \forall\, i \in 1:n \right\}. \tag{3.5}$$

Here T denotes the transposition.

Since the functions h_i are continuous on q and bounded from above, then they are Lipschitz. Then by the mean-value theorem for Lipschitz functions

$$h_i(q) = h_i(q_0) + (a_i(q), q - q_0), \tag{3.6}$$

where $a_i(q) \in \partial_{Cl} h_i(q_0 + \theta_i(q)(q - q_0)), \;\; \theta_i(q) \in (0, 1), \;\; \partial_{Cl} h_i$ is a Clarke subdifferential of function h_i at considered point (see [4]).

V. Demyanov, V. Roshchina [3] showed that

$$\partial_{Cl} h_i(q) \in \mathcal{L}(q_0) = \text{cl co}\{a_i \in C_i, \; C_i \in E^*(h_i)\}. \tag{3.7}$$

From (7), (3) one has

$$f_i(x_0 + \alpha g, y_0 + \alpha q) = f_i(x_0, y_0) + (a_i(q), q - q_0) + r_i(\alpha, q) + o_i(q - q_0), \tag{3.8}$$

where $r_i(\alpha, q) = \frac{o_i(\alpha q)}{\alpha} \overset{\longrightarrow}{\alpha \downarrow 0} 0 \quad \forall\, q \in \mathbb{R}^n$, $r_i(0, q) = 0$.

Denote by $\mathcal{A}_i(q)$ set of vectors $a_i(q)$ satisfying (6)–(7). At the point q_0 put

$$\mathcal{A}_i(q_0) = \mathcal{L}_i(q_0). \tag{3.9}$$

Introduce the set

$$\mathcal{A}(q) = \left\{ A = \begin{pmatrix} a_1^T \\ \vdots \\ a_n^T \end{pmatrix} \;\middle|\; a_i \in \mathcal{A}_i(q), \, \forall\, i \in 1:n \right\}.$$

(9) and (5) yield

$$F(\alpha, q) = \mathcal{A}(q)(q - q_0) + \gamma(\alpha, q), \tag{3.10}$$

where $F = (F_1, \ldots, F_n)$, $\gamma(\alpha, q) = (r_1(\alpha, q) + o_1(q - q_0), \ldots, r_n(\alpha, q) + o_n(q - q_0))$, $\gamma(\alpha, q) \underset{q \to q_0}{\underset{\alpha \downarrow 0,}{\longrightarrow}} 0$. The mapping $\mathcal{A}$ is upper semicontinuous and convex.

Theorem 1. If

$$|\det A| \geqslant \beta > 0 \quad \forall A \in \mathcal{L}(q_0),$$

then for any $\varepsilon > 0$ there exist $\alpha_0 > 0$ and $q(\alpha) \in \mathbb{R}^n$, such that

$$\|q(\alpha) - q_0\| \leqslant \varepsilon,\ f(x_0 + \alpha g, y_0 + \alpha q(\alpha)) = 0_n \quad \forall\, \alpha \in [0, \alpha_0]. \tag{3.11}$$

Proof. Upper semicontinuity of the mapping $\mathcal{A}$ and (9) imply that there exists $\varepsilon_1 > 0$, such that $\varepsilon_1 \leqslant \varepsilon$:

$$|\det A| \geqslant \frac{\beta}{2} \quad \forall A \in \mathcal{A}(q), \quad \forall q \in S_{2\varepsilon}(q_0). \tag{3.12}$$

Put $\bar{q} = q - q_0$,

$$\Phi_\alpha(\bar{q}) = -\mathcal{A}^{-1}(q)\gamma(\alpha, q) = -\mathcal{A}^{-1}(q_0 + \bar{q})\gamma(\alpha, q_0 + \bar{q}). \tag{3.13}$$

It follows from (13), the continuity of γ and relation (14) that $\varepsilon_2 > 0$ and $\alpha_0 > 0$ can be found such that $\varepsilon_2 \leqslant \varepsilon_1 \leqslant \varepsilon$, $\|v\| \leqslant \varepsilon_2\ \forall v \in \Phi_\alpha(\bar{q})\ \forall \alpha \in [0, \alpha_0]$ $\forall \bar{q} \in S_{1\varepsilon_2} = \{\bar{q} \in \mathbb{R}^n \,|\, \|\bar{q}\| \leqslant \varepsilon_2\}$.

For the mapping Φ_α given by (14) all the conditions of the Kakutani generalized theorem are met (see. [3]), i.e. there exists $\bar{q}(\alpha) \in S_{2\varepsilon_2}$ such that $\bar{q}(\alpha) \in \Phi_\alpha(\bar{q}(\alpha))$. Then, there exists $A \in \mathcal{A}(q_0 = \bar{q}(\alpha))$, such that

$$\bar{q}(\alpha) = -A^{-1}\gamma(\alpha, q_0 + \bar{q}(\alpha)). \tag{3.14}$$

Put $q(\alpha) = q_0 + \bar{q}(\alpha)$. Then from (15) we have $q(\alpha) - q_0 = -A^{-1}\gamma(\alpha, q(\alpha))$, i.e. $A(q(\alpha) - q_0) + \gamma(\alpha, q(\alpha)) = 0_n$. From here and from (15) $f(\alpha, q(\alpha)) = 0_n$ $\forall\, \alpha \in [0, \alpha_0]$, $\|q(\alpha) - q_0\| \leqslant \varepsilon$. From the definition of $F_i(\alpha, q)$

$$f(x_0 + \alpha g, y + \alpha q(\alpha)) = 0_n \quad \forall \alpha \in [0, \alpha_0],$$

which was to be proved.

Theorem 2. Let the functions $f_i(x,y)$ be Lipschitz in y. If the mapping $\mathcal{L}(q_0)$ is upper semicontinuous at the point $[0,q_0]$ and besides

$$|\det A| \geqslant \beta > 0 \quad \forall A \in \mathcal{L}(q_0), \tag{3.15}$$

then for any $\varepsilon > 0$ there exist $\alpha_0 > 0$ and $q(\alpha) \in \mathbb{R}^n$ such that

$$\|q(\alpha) - q_0\| \leqslant \varepsilon,\ f(x_0 + \alpha g, y_0 + \alpha q(\alpha)) = 0_n \quad \forall \alpha \in [0,\alpha_0]. \tag{3.16}$$

Such a $q(\alpha)$ is unique and the function $q(\alpha)$ is continuous on $[0,\alpha_0]$.

Proof. If $f(x,y_0) = 0_n$, then $y(x) = y_0$ satisfies (17). Consider the case $f(x,y_0) \neq 0$. Since the mapping $\mathcal{A}$ is upper semicontinuous and convex-valued on S, then there exist $\delta_1 > 0$ and $\varepsilon_1 \leqslant \varepsilon$ such that

$$|\det A| \geqslant \frac{\beta}{2} \quad \forall A \in \mathcal{A}_x(\bar{y}),\ \forall \bar{y} \in S_{2\varepsilon_1},\ \forall x \in S_{1\delta_1}(x_0).$$

Here $S_{2\varepsilon_1} = \{\bar{y} \in \mathbb{R}^n | \|\bar{y}\| \leqslant \varepsilon\}$.

The continuity of f implies that one can find $\delta_2 > 0$ such that $\delta_2 \leqslant \delta$,

$$\|A^{-1} f(x,y_0)\| \leqslant \varepsilon_1 \quad \forall A \in \mathcal{A}_x(\bar{y}),\ \forall \bar{y} \in S_{2\varepsilon_1},\ \forall x \in S_{1\delta_1}(x_0). \tag{3.17}$$

Then by the definition the mapping Φ_x maps the ball $S_{2\varepsilon_1}$ into a subset of $S_{2\varepsilon_1}$. Since all the hypotheses of the generalized Kakutani theorem are satisfied then there exists a fixed point of the mapping Φ_x, i.e. one can find $\bar{y}(x) \in \mathbb{R}^n$ such that

$$\|\bar{y}(x)\| \leqslant \varepsilon_1, \quad \bar{y}(x) \in -\mathcal{A}_x^{-1}(\bar{y}(x)) f(x,y_0). \tag{3.18}$$

It means that for some $A \in \mathcal{A}_x^{-1}(\bar{y}(x)) = \mathcal{A}(x, y_0 + \bar{y}(x))$, we have $\bar{y}(x) = -A^{-1} f(x,y_0)$, i.e.

$$f(x,y_0) + A\bar{y}(x) = 0_n.$$

Hence and from relation (7) follows

$$f(x, y(x)) = 0_n, \tag{3.19}$$

where $y(x) = y_0 + \bar{y}(x)$. Relations (18)–(19) imply that

$$\|y(x) - y_0\| \leqslant \varepsilon_1, \quad f(x,y(x)) = 0_n \quad \forall x \in S_{1\delta_2}(x_0). \tag{3.20}$$

Thus, a function $y(x)$ satisfying (21) when $\varepsilon = \varepsilon_1$ and $\delta = \delta_2$ exists.

Let us prove the uniqueness of $y(x)$. To be more precise we show that there exists $\delta > 0$ such that $\delta \leqslant \delta_2$ and $y(x)$ satisfying (21) is unique for all $x \in S_{1\delta}(x_0)$.

Assume the contrary. Then there exist sequences $\{x_k\}$, $\{y'_k\}$, $\{y''_k\}$ such that

$$x_k \to x_0,\ y'_k \to y_0,\ y''_k \to y_0,\ y'_k \neq y''_k, \tag{3.21}$$

$$f(x_k, y'_k) = f(x_k, y''_k) = 0_n \quad \forall k.$$

By the mean-value theorem for Lipschitz functions $f_i(x_k, y''_k) = f_i(x_k, y'_k) + (v_{ik}, y''_k - y'_k)\ \forall\, i \in 1:\ n$, where $v_{ik} \in C_i(h(x_k, y'_k + \gamma_{ik}(y''_k - y'_k)))$, $\gamma_{ik} \in (0,1)$. It follows from (22) that

$$(v_{ik}, y''_k - y'_k) = 0. \tag{3.22}$$

Since $y'_k \neq y''_k$ then $y''_k - y'_k \neq 0_n$ and (23) implies

$$(v_{ik}, g_k) = 0, \tag{3.23}$$

where $g_k = \dfrac{y''_k - y'_k}{\|y''_k - y'_k\|}$, $\|g_k\| = 1$. Let $g_k \to g_0$, $\|g_0\| = 1$.

Since E^* is bounder, without loss of generality one can assume that $v_{ik} \to v_{i0}$. From (24)

$$(v_{i0}, g_0) = 0 \quad \forall i \in 1:n, \tag{3.24}$$

and, by the virtue of upper semicontinuity of the mapping $\mathcal{L}(q_0)$ $v_{i0} \in C_{iy}(x_0, y_0)$. Hence $A \in \mathcal{L}(q_0)$. It follows from (23) that $Ag_0 = 0_n$. Since $\|g_0\| = 1$, it is possible only when $\det A = 0$ which contradicts (16). Therefore, the uniqueness is proved.

Now we have to prove the continuity of the function $y(x) = y(x, \alpha) = y_0 + \alpha q(\alpha)$. Assume the contrary. Then there will be found a point $x \in S_{1\delta}$ and a sequence of points $\{x_k\}$ such that

$$x_k \to x, \quad y(x_k) \not\to y(x),$$

$$\|y(x_k) - y_0\| \leqslant \varepsilon,\ f(x_k, y(x_k)) = 0_n \quad \forall\, x_k \in S_{1\delta}(x_0). \tag{3.25}$$

Without loss of generality by virtue of (17) it is possible to assume that

$$y(x_k) \to \bar{y} \neq y(x). \tag{3.26}$$

It follows from the continuity f and (27) that

$$f(x, \bar{y}) = 0_n. \tag{3.27}$$

On the other hand, by the definition of $y(x)$

$$f(x, y(x)) = 0_n. \tag{3.28}$$

Relations (27) – (29) contradict the already established uniqueness of the function $y(x)$ on the set $S_{1\delta}$.

Thus, to demonstrate the existence of an implicit function in the direction g one should find all solutions of system (4), i.e. constract the set $\mathcal{L}(q_0)$ for each of them [4, 5]. If q_0 is a solution of the system and the conditions of the theorem are satisfied, then with sufficiently small $\alpha > 0$ for the given q_0 there exists an implicit function $q(\alpha)$. If the exhauster mapping is upper semicontinuius at the point $[0, q_0]$ and the conditions of the theorem hold, then it is possible to conclude that for sufficiently small $\alpha > 0$ there exists a unique vector-function $q(\alpha)$ and this function is continuous.

If a condition of the theorem is not hold then an additional verification is needed since it can happen that none of the implicit functions in the given direction g conforms such q_0, either it is unique or there are a lot of such functions.

References

1. V.F. Demyanov *Exhausters and Convexificators - New Tools in Nonsmooth Analysis.* Kluwer Academic Publishers, Dordrecht, 2000. Pp. 85 - 137.
2. M. Castellani *Convex analysis and minimization algorithms.* Vol. I, II. Springer Verlad, 1993.
3. V.F. Demyanov and V. Roshchina *Exhausters and subdifferentials in nonsmooth analysis*, Optimization, 2008. 57:1. Pp. 41 - 56.
4. V.F. Demyanov and G.Y. Murzabekova *Convexificators and Implicit functions in nonsmooth systems*, Computational Mathematics and Mathematical Physics, Vol. 39, No. 2, 1999. Pp. 222 - 234.
5. G.Y. Murzabekova *Exhausters and implicit functions in nonsmooth systems.* Optimization, Vol. 59, No. 1, 2010. Pp. 105 - 113.

G. Y. Murzabekova
Kazakh agrotechnical university, Kazakhstan, 010011, Astana, Pobeda Prospect,
87172383959, guldenmur07@gmail.com

THE FUNCTIONAL EQUATION AND IT'S SOLUTION FOR THE ANOMALOUS DIFFUSION WITH BERNOULLI SCALING

Yu. G. Rudoy, O. A. Kotelnikova

Key words: normal vs. anomalous diffusion, functional equation, uniform function

AMS Mathematics Subject Classification: 39B, 60J

Abstract. The problem of one-dimensional two-parametric symmetric diffusion for the case of Bernoulli scaling is studied by means of the Markov random walks concept and the parametric crossover between the normal (Gauss) and anomalous (Levy) cases is considered. It is shown that the conversion of Lyapunov characteristic function from the trigonometric series to the power series is effectively produced by means of the functional equation. The original method of FE solution is suggested which gives the most physically obvious description of the crossover. In particular, it is shown that the FE solution splits into regular and singular parts while the normal diffusion is possible only asymptotically when the singular part fully disappears.

1 Introduction

Many processes in natural system's are not of deterministic but of stochastic character and are described by the Markov random walks (MRW) in the system's phase, or state, space $\{X\}$ with r-dimensional variable $X = \{x^{(1)} \dots x^{(r)}\}$ (here $r = 1$). We will consider RW as "jumps", or steps, discrete in time t (or step numbers n), but continuous in the single state variable x. As is well known (see, e.g., [1]) MRW are fully described by the transition probability $P(x_n, x_{n+1})$, which is taken as $P(|x_n - x_{n+1}|)$, i.e., stationary (independent upon n) and symmetric in x.

Given the function $P(|x_n - x_{n+1}|)$ between adjacent time moments n and $n+1$ (for any n) as well as the initial distribution $p(x)$ at $n = 0$, one may find the final distribution $p_N(x)$ using the Bachelier – Smoluchowski – Chapman – Kolmogorov (BSCK) integral equation (see, e.g., [1]).

Asymptotically at large $N \to \infty$ the solution of BSCK describes the diffusion process $X(t)$ of the system in the phase space $\{X\}$. The nature of trajectories $X(t)$ depends crucially upon the effective radius $0 < R < \infty$ of the transition probability $P(x) > 0$, $\int P(x)dx = 1$; the integral is taken over all accessible x (here $0 < x < \infty$). By definition, $R = \sigma$, where $\sigma^2 = \int y^2 P(y)dy$; if $\sigma^2 < \infty$ and R is finite, or "short",

diffusion is *normal* (ND) (Gauss, Brown, Einstein, Wiener). If $\sigma^2 \to \infty$ and R is infinite, or "long", the diffusion is *anomalous* (AD) (non-Gauss, Levy – Khinchine).

The principal difference between ND and AD is that for ND there exist *unique* natural scale $R = \sigma$, whereas for AD it does not exist. The final distributions $p_N(x)$ at $N \to \infty$ are, in general, not stationary — i.e., they have no definite limit $p_\infty(x)$. The $p_N(x)$ are "narrow" (Gauss, normal, short-tailed), and "broad" (Levy – Khinchine, stable, long-tailed), so their fluctuations are, accordingly, small and large.

In general, transition probability $P = P(y\lambda)$ may contain some control parameter λ, so $\sigma = \sigma(\lambda)$ and it is possible to obtain the transition, or crossover, between ND and AD at some $\lambda = \lambda_{cr}$ where $\sigma(\lambda_{cr})$ diverges. At this point infinitely many scales of one "jump", or arbitrary large fluctuations, become possible. The trajectories X(t) in the phase space $\{X\}$ instead of continuous (though not differentiable!) become discontinuous "Levy flights".

It appears reasonable to calculate not the (normalized) probability density $p_N(x\lambda)$, but it's Fourier transform $G_N(k, \lambda)$

$$G_N(k\lambda) = \int p_N(x\lambda) \exp(ikx) dx \equiv \langle \exp(ikx) \rangle, \tag{1.1}$$

which defines the Lyapunov's characteristic function (CF). The function $G_N(k\lambda)$ possess two important properties: (1.1) CF is the generating function for the probability moments

$$m_{2s}(\lambda) \equiv \langle x^{2s} \rangle = (-i)^{2s} G^{(2s)}(k = 0, \lambda), \quad s = 0, 1, \ldots, \tag{1.2}$$

which may exist not for all r: only $m(\lambda) \equiv 1$ should exist always due to normalization; (1.2) CF obeys the functional equation (FE) which follows from the BSCK integral equation or from the explicit form of CF.

2 Bernoulli Scaling for Markov Random Walk

The interesting problem arises when some parameter λ for $P = P(x\lambda)$ (and thus for $p_N(x, \lambda)$ due to BSCK) comes into play; then index α becomes $\alpha(\lambda)$, and the crossover may arise between normal (Gauss – Wiener) and anomalous (Levy – Khinchine) regimes of diffusion. The simple — but non-trivial — choice for transition probability $P = P(x\lambda)$ is the so-called *Bernoulli scaling*, when the radius of one-dimensional steps $j = 0, 1, \ldots, \infty$, as well as their probabilities, form two concurrent geometric progressions: ascending for step's radius $\pm 1, \pm b, \ldots, \pm b^j \ldots$, where $1 < b < \infty$, and descending for probabilities $p, p/\lambda, \ldots, p/\lambda^j \ldots$, where $1 < \lambda < \infty$,

$0 < p < 1$; from normalization condition it follows that $p(\lambda) = 1/2(1 - 1/\lambda)$:

$$P(x\lambda) = p\sum_{j=0}^{\infty} \lambda^{-j} \left|\delta\left(x - b^j\right) + \delta\left(x + b^j\right)\right|, \tag{2.1}$$

Clearly, the result for $p_N(x\lambda)$ is determined by the relation between b and λ: at fixed b=const, $p_N(x\lambda)$ to be non-Gaussian, $\lambda = \lambda(b)$ should be small "enough" — but the true question is: how small exactly?

By definition, $G(x\lambda)$ smooths all the $P(x\lambda)$ singularities (including the δ-like ones) and, as a rule, CF is continuous. For $P(x\lambda)$ with Bernoulli scaling, $G(x\lambda)$ is given by the *Weierstrass function*, where $G(0, \lambda) \equiv 1$, $G'(0, \lambda) = 0$:

$$G(k, \lambda) = 2p(\lambda)\sum_{j=0}^{\infty} \lambda^{-j} \cos\left(kb^j\right). \tag{2.2}$$

We should recall that if all terms in trigonometric (Fourier) series are present, the oscillations in various terms are smoothed. In lacunar trigonometric series the smoothing is absent and, as a rule,the singularities arise. So, the sum of the series may exist (and even be continuous!), but often appears to be non-differentiable — or, equivalently, non-analytic.

Clearly, Eq. (2.2) gives us the explicit analytic expression for CF, but only in the form of (lacunar) *trigonometric series*, which was studied intensively by Hardy and Littlewood about a century ago and then by Titchmarsh about half a century ago; thus, many properties of the Weierstrass function are known very well.

But this form of $G(x\lambda)$ is not convenient for our goals of the asymptotic diffusion study, and it is desirable to rebuild $G(x\lambda)$ in the form of *power series* in k, valid for all values of $1 < \lambda < \infty$. Clearly, in the limit $\lambda \to \infty$ new form should exactly reproduce the Gauss – Wiener, or normal, diffusion (ND). Physically, we need to know the behavior of diffusion — i.e., $P(x\lambda)$, at large x, or, mathematically — due to *Tauberian theorems* — the behavior of $G(k, \lambda)$ at small k.

Thus, finally, we need the Taylor expansion for $G(k, \lambda)$ in the vicinity of $k = 0$, where $G(0, \lambda) \equiv 1$. The most direct way is the use of Mellin integral transform, which was done by Hughes, Montroll and Schlesinger 30 years ago [2]:

$$G(k, \lambda) = 2p(\lambda)\sum_{j=0}^{\infty} \lambda^{-j} \cos\left(kby^j\right) = \frac{1}{2\pi i}\int g(s)|k|^{-\alpha} ds, \tag{2.3}$$

where integration goes from $c - i\infty$ to $c + i\infty$ with $0 < c = \mathrm{Re}s < 1$; unfortunately, this representation appear to lack physical obviousness.

Instead, we propose to use the property 1) of CF from Eq. (1.1), which is equivalent to the Taylor expansion of CF in all orders of k in the vicinity of $k = 0$:

$$G(k,\lambda) = 1 + \sum_{s=1}^{\infty} \frac{1}{(2s)!} \left(i^{2s}\right) m_{2s}(\lambda) k^{2s}. \tag{2.4}$$

This expansion holds only if the moments $m_{2s}(\lambda)$, or due to Eq. (1.2) the derivatives $G^{(2s)}(k=0,\lambda)$, do exist for all $s \geq 1$; from Eq. (2.2) for CF (with $b > 1$) it follows:

$$|m_{2s}(\lambda)| = 2p(\lambda) \sum_{j=0}^{\infty} \left(\frac{\lambda_{cr}^{(s)}}{\lambda}\right)^{j}, \quad \lambda_{cr}^{(s)} = (\lambda_{cr})^{s}, \quad \lambda_{cr} = b^2. \tag{2.5}$$

Clearly, all $m_{2s}(\lambda)$ are represented by the sums of geometric progressions with denominators $\lambda_{cr}^{(s)}/\lambda$ and exist only if $\lambda_{cr}^{(s)}/\lambda < 1$ — i.e., for large enough λ. Threshold values $\lambda_{cr}^{(s)}$ form the ascending (up to ∞) geometric progression $(\lambda_{cr})^s$, $s = 1, 2, \ldots$ with denominator $\lambda_{cr} = b^2 > 1$.

Thus, all $m_{2s}(\lambda)$ may exist only in the limit $\lambda \to \infty$, when all $m_{2s}(\infty) \equiv 1$ (only the $j = 0$ term contributes then in the sum (2.5)), so $G(k\infty) = \cos k$. This is the "trivial" case of Gauss ND; the non-trivial Levy AD arises only for *finite* values of λ. In the context of Tauberian theorems, the most important term in Eq. (2.4) for CF is the lowest order one with $s = 1$ — i.e., term $\sim k^2$ with coefficient $(-1/2)m_2(\lambda) \equiv (-1/2)\sigma^2(\lambda)$, $\sigma^2(\lambda) \geq 1$ (sign "=" holds only for $1/\lambda = 0$).

From Eq. (2.5), $\sigma^2(\lambda) = 2p(\lambda)[1 - \lambda_{cr}/\lambda]^{-1}$, and there are *two regimes of diffusion* — ND and AD; clearly, the crossover between both takes place exactly at $\lambda = \lambda_{cr} = b^2 > 1$:

1. $\lambda_{cr}/\lambda < 1$, $\sigma^2 < \infty \to$ Gauss–Wiener regime (ND) $G(k,\lambda) \approx 1 - (1/2)\sigma^2(\lambda)k^2$.
2. $\lambda_{cr}/\lambda \geq 1$, $\sigma^2 \to \infty \to$ Levy – Khinchine regime (AD).

3 Functional Equation (FE) for Lyapunov Characteristic Function

Note, that nothing can be said here about the form of $G(k,\lambda)$ in AD-regime — only that the derivative $G^{(2)}(k\lambda)$ diverge. To find this form for $G(k,\lambda)$, the property (1.2) of CF from Eq. (2.3) is of use: the *functional equation* (FE) for CF valid for all values of λ and k. In the standard case (i.e., absence of λ) the FE for $G_N(k)$ is well known (see, e.g., [2]) and possess the general solution:

$$G_N(k) = \exp[-N\psi(k)], \quad \psi(k) = C_\alpha |k|^\alpha, \quad C_\alpha \geq 0, \quad \psi(0) = 0, \quad G_N(0) = 1. \tag{3.1}$$

Due to Tauberian theorems, index $0 < \alpha \leq 2$ gives at $N \to \infty$ for $p_N(x)$ asymptotic for large x: $\alpha = 2$ — Gauss, exponential, $\exp[-x^2/2\sigma^2]$ with $\sigma^2 \equiv \langle x^2 \rangle$; $\alpha < 2$ — Levy–Khinchine, power-like, $C_\alpha N x^{-(1+\alpha)}$. In the case with parameter λ (Eq. (2.3)) the FE is:

$$G(k,\lambda) = \frac{1}{\lambda} G(bk,\lambda) + 2p(\lambda)\cos k. \tag{3.2}$$

The most characteristic cases for FE in various domains of the parameter $1 < \lambda < \infty$ (at b fixed) are not evident from Eq. (3.2); we use below coefficient $\alpha = 2\ln\lambda/\ln\lambda_{cr}$ (for now — formally).

1. $\lambda \to 1$, $p(\lambda) \to 0$, FE holds only for $b = 1$; unphys. limit;
2. $\lambda \to \infty$, $p(\lambda) \to 1/2$, FE: $G(k) = \cos k$, trivial ND regime;
3. $\lambda > \lambda_{cr} > 1$, $\alpha = 2(\ln\lambda)/(\ln\lambda_{cr}) > 2$, non-trivial ND reg.;
4. $1 < \lambda \leq \lambda_{cr}$, $0 < \alpha < 2$, $\lambda_{cr} = b^2$, non-trivial AD reg.

In both cases 3) and 4) of non-trivial ND or AD regimes it appears necessary to solve general FE (3.2) for $G(k\lambda)$, which relative to G is linear and inhomogeneous one. The inhomogeneous part $2p(\lambda)\cos k$ with $2p(\lambda) > 0$ is regular in k in the vicinity of $k = 0$ for all $\lambda > 1$, but the homogeneous $(1/\lambda)G(bk,\lambda)$ part of FE (3.2) may give rise to the singular in k behavior at finite $\lambda < \infty$. Full solution of FE may be represented as the sum:

$$G(k,\lambda) = G_{\rm reg}(k,\lambda) + G_{\rm sign}(k,\lambda). \tag{3.3}$$

According to Taylor expansion (2.4), in lowest-k order one obtains:

$$G_{\rm reg}(k) \approx 1 - \frac{1}{2}\sigma^2 k^2, \quad \text{where} \quad \sigma^2(\lambda) = 2p(\lambda)\left[1 - \frac{\lambda_{cr}}{\lambda}\right]^{-1} \tag{3.4}$$

But it is instructive for the following, to see how such an expression for $\sigma^2(\lambda)$ is generated by the FE (3.2) for CF. Substituting G_{reg} in FE (3.2), from l.h.s. and inhomogeneous term we obtain $(1 - 1/\lambda)(1 - (1/2)k^2)$, where as inhomogeneous term gives $(1/\lambda)(1 - (1/2)\sigma^2 b^2 k^2)$; clearly, the expression (3.4) for $\sigma^2(\lambda)$ holds. Evidently, that by this simple calculation the mathematical mechanism of FE for CF becomes more clear.

The "critical" factor λ_{cr}/λ (here < 1) enters in $\sigma^2(\lambda)$ just due to the FE term homogeneous in G, and this fact appears to be of great importance for the whole problem If $\lambda_{cr}/\lambda < 1$, only the regular part $G_{reg}(k,\lambda)$ from Eq. (3.4) is the desired solution in lowest-k approximation. But if $\lambda_{cr}/\lambda \geq 1$, then both $\sigma^2(\lambda)$ and

$G_{reg}(k,\lambda)$ don't exist, and only the singular part $G_{sing}(k\lambda)$ describes the solution (3.3); then, by analogy with Eq. (3.1) $G_{sing}(k,\lambda)$ should be written in the form:

$$G_{sing}(k,\lambda) = C(\lambda)k^{\alpha(\lambda)}Q(k), \quad \text{where} \quad Q(bk) = Q(k). \tag{3.5}$$

Then, by means of only homogeneous part of FE one obtains readily: $1 = (1/\lambda)b^{\alpha(\lambda)} \to \alpha(\lambda) = \ln\lambda/\ln b \leq 2$. For any given finite value of λ from the open interval $(1,\infty)$ there exist always some (also finite) value of $S(\lambda)$ from the half-open interval $[1,\infty)$, so that $\lambda_{cr}^{(S-1)} < \lambda(S) < \lambda_{cr}^{(S)}$, where $\lambda_{cr}^{(0)} = 1$ for $S = 1$; here $S(\lambda)$ and $\lambda(S)$ are mutually reverse functions.

4 The Solution of FE and its Correspondence with Two Regimes of Diffusion

Then $m_{2s}(\lambda)$ exist only for $s = 1, 2, \ldots, S-1$ and diverge for $s \geq S$; the expansion (2.4) breaks, or "cuts off", at $k^{2(S-1)}$ term. Thus, the regular, or analytic, part of $G(k\lambda)$ takes the form:

$$G_{reg}(k,\lambda(S)) = 1 + \sum_{s=1}^{S-1}(-1)^s \frac{1}{(2s)!} m_{2s}[\lambda(S)]k^{2s}. \tag{4.1}$$

Clearly, the remainder of the expansion (2.4) is fully non-analytic, or singular, because all the terms (starting from the order $2S$) are divergent.

$$G_{sign}(k,\lambda(S)) = \sum_{s=S}^{\infty}(i)^{2s} \frac{1}{(2s)!} m_{2s}[\lambda(S)]k^{2s}. \tag{4.2}$$

The most remarkable is the fact that all this *infinite sum* may be represented by the *single* term, but of fractional order instead of entire (and, moreover, even) order: with the natural replacement $k \to |k|$ (4.2) goes over in (clarify with (3.5)):

$$G_{sign}(|k|,\lambda(S)) = C[\lambda(S)]|k|^{\alpha[\lambda(S)]}Q(|k|). \tag{4.3}$$

The "ingredients" of $G_{sing}(|k|\lambda(S))$ are the following: coefficient $C[\lambda(S)]$, (anomalous) index $\alpha[\lambda(S)]$ and the amplitude function $Q(|k|)$; all quantities are real and positive. The most important is index $\alpha[\lambda(S)]$, which should satisfy the strong inequality $2(S-1) < \alpha[\lambda(S)] < 2S$: only in this case all the derivatives $G_{sing}^{(2s)}(k,\lambda)$ with $s \geq S$ become singular at $k = 0$. Indeed, let $2S - \alpha = \varepsilon$ with $0 < \varepsilon < 2$, and $s = S + r$, $r = 0, 1, \ldots$; then $G_{sing}^{(2s)}(|k|,\lambda) \sim |k|^{-(\varepsilon+2r)} \to \infty$ at

$k = 0$ for any arbitrary small $\varepsilon > 0$. Note, that $Q(|k|)$ may be not differentiated here because, as will be shown, it contains only logarithmic corrections $\sim [\log|k|]^n$.

Index $\alpha[\lambda(S)]$ may be calculated exactly as in Eq. (3.5) by means of only homogeneous part of FE — i.e., scaling, or self-similar, or generalized uniform FE $G(|k|,\lambda) = (1/\lambda)G(b|k|,\lambda)$. Assume additionally FE $Q(b|k|) = Q(|k|)$ for the amplitude function, then the FE (3.2) gives:

$$1 = [(1/\lambda(S)]b^{\alpha[\lambda(S)]}, \quad \text{and, finally,} \quad \alpha[\lambda(S)] = \ln\lambda(S)/\ln b. \tag{4.4}$$

Apply now the monotonous log-operation to the "input"chain inequality $(\lambda_{cr})^{S-1} < \lambda(S) < (\lambda_{cr})^S$ (all terms here exceed unity, so their log's > 0) and divide by $\ln\lambda_{cr} > 0$, then $(S-1) < \ln\lambda(S)/\ln\lambda_{cr} < S$. Recalling that $\lambda_{cr} = b^2$, one comes at last to most important inequality: $2(S-1) < \alpha[\lambda(S)] < 2S$, accounting for the non-analyticity. Note also, that $2(1-1/S) < \alpha[\lambda(S)]/S < 2$, so $\alpha/S \to 2$ at $S \to \infty$, $\lambda(S) \to \infty$.

Coefficient $C[\lambda(S)]$, introduced in Eq. (4.3), may be determined by the "heuristic" considerations which nevertheless coincide exactly with precise Mellin transform results from [2]. Transition from the first divergent $2S$-term in $G_{reg}(k\lambda)$ (4.1) to corresponding term in $G_{sing}(|k|,\lambda)$ is merely the change in power of k, namely $2S \to \alpha < 2S$. The same change is applied to the coefficient $1/(2S)!(i)^{2S} \to C(\alpha)$:

$$1/(2S)! \equiv 1/\Gamma[1+(2S)!] \to 1/\Gamma[1+\alpha] = -(1/\pi)\Gamma(-\alpha);$$
$$(i)^{2S} \equiv \exp[i(\pi/2)]^{2S} \to \exp[i(\pi/2)]^{\alpha} \to \cos[(\pi/2)\alpha].$$

Amplitude function $Q(|k|)$, introduced in Eq. (4.3), should satisfy only the functional equation $Q(b|k|) = Q(|k|)$, which possess, e.g., $Q = \text{const}$ as a possible solution. But, more generally, this FE defines the so-called log-periodic functions, which belong to the class of *slowly varying* functions (in Karamata's sense). By the "log" change of variable: $|k| \to \log(|k|)$, $b \to \log b$ (recall that $b > 1$), the FE for $Q(|k|)$ goes over to FE for periodic functions: $\mathbb{Q}(\log|k|) = \mathbb{Q}(\log|k| + \log b)$. This new FE has the solution as trigonometric (Fourier) series in $\log(|k|)$ with period $\log b$ and amplitudes q_m and phases ϕ_m(in general, undefined):

$$Q(|k|) = \frac{2}{\log b}\sum_{m=0}^{\infty} q_m \cos\left[2\pi m\frac{\log|k|}{\log b} + \varphi_m\right]. \tag{4.5}$$

5 Conclusion

The problem of one-dimensional symmetric diffusion and the crossover between the normal (Gauss) and anomalous (Levy) cases were considered by means of the Markov random walks concept. The two-parametric case (Bernoulli scaling) was considered and the functional equation (FE) for the Lyapunov characteristic function was formulated. The original method of FE solution was suggested which gives the most physically obvious description of the crossover.

References

1. Gardiner C. W. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, 2^{nd} ed., Springer Series in Synergetics. — Vol. 13. — Springer, 1985.
2. Montroll E. W., Shlesinger M. F. *On the Wonderful World of Random Walks*, in Studies in Statistical Mechanics. — Vol. XI (ed. by E. W. Montroll, J. L. Lebowitz). — Amsterdam: North-Holland PC, 1984. — Pp. 46–121.
3. Feller W. *An Introduction to Probability Theory and its Applications*, 3^{d} ed. — N.Y.: Wiley, 1984.

Yu. G. Rudoy
Russian Federation, 117198, Moscow, PFUR, rudikar@mail.ru

O. A. Kotelnikova
Russian Federation, 119999, Moscow, Lomonosov MSU, olga@magn.ru

PARACONVEXITY AS A GENERALIZED CONVEXITY

P. V. Semenov

Key words: convexity, continuous selections, uniform retractions, function of nonconvexity

AMS Mathematics Subject Classification: 54C60, 54C65, 41A6

Abstract. We show that for a family Ω of pairwise disjoint closed subsets P of a Banach space Y whose functions of nonconvexity α_P, $P \in \Omega$ admit a common majorant $\alpha : (0; \infty) \to [0; 0, 5)$, such that the function $2\alpha(\cdot)$ is geometrically summable, there exists a generalized convex structure ω on Y such that all members of the family Ω are convex with respect on ω. The proof is based on continuous choice of an uniform retractions of the entire space Y onto elements of Ω.

1 Introduction

In general, there exists an entire branch of mathematics devoted to various versions of the notion of convexity. Even if one simply lists the titles of "generalized convexities" one will find at least nearly 20 different notions.

Typically, a creation of "generalized convexities", is usually related to an extraction of several principal properties of the classical convexity which are used in one of the key mathematical theorems or theories and, consequently deals with analysis and generalization of these properties in maximally possible general settings. D. Repovs and the author, based on E. Michael's ingenious idea, [4], systematically studied another approach to weakening or controlled omission of convexity on a set of principal theorems of multivalued analysis and topology. Roughly speaking, to each closed subset $P \subset B$ of a Banach space we have associated a numerical function, say $\alpha_P : (0, +\infty) \to [0, 2)$, the so-called function of nonconvexity of P. The identity $\alpha_P \equiv 0$ is equivalent to the convexity of P and the more α_P differs from zero the "less convex" is the set P.

Such classical results about multivalued mappings as the Michael selection theorem, the Cellina approximation theorem, the Kakutani-Glicksberg fixed point theorem, the von Neumann - Sion minimax theorem, etc. are valid with the replacement of the convexity assumption for values $F(x)$, $x \in X$ of a mapping $F : X \to Y$ by some appropriate control of their functions of nonconvexity, see [7, 10]. Briefly, all

The author was supported in part by the RFBR grant № 11-01-00822.

functions $\alpha_{F(x)},\ x \in X$, of nonconvexity should be "less than 1". In such a situation all values $F(x),\ x \in X$ are called paraconvex subsets of Y. A typical way for creation paraconvex set is a Lipshitz perturbation of a closed convex set along an additional transversal direction, see [10]

In comparison with usual ideas of "generalized convexities", we never define in this approach, for example, a "generalized segment" joining $a \in P$ and $b \in P$. We look only for the distances between points c of the classical segment $[a, b]$ and the set and look for the ratio of these distances and the size of the segment. But, maybe paraconvexity simply is a specific case of an appropriate generalized convex structure, i.e. paraconvexity of a set with respect to the classical convexity structure coincides with convexity under some generalized convexity structure?

Clearly, the formal answer is negative because for example intersection of two paraconvex sets can be even disconnected. But for a family of pairwise disjoint paraconvex sets the answer is affirmative whenever all functions of nonconvexity simultaneously are less than a constant, say $c < 0,5$, see [8].

At this note an analogous (partial) answer is given for the case of "good" functional (non-constant) common majorant of the family of all functions of nonconvexity.

2 Preliminaries and statements

For a numerical function $\alpha : (0; +\infty) \to (0; +\infty)$ the geometric progression with the (functional) coefficient α is defined by setting for every positive t

$$t, \quad \alpha(t) \cdot t, \quad \alpha(\alpha(t) \cdot t) \cdot \alpha(t) \cdot t, ..., \alpha_{n+1}(t) = \alpha(\alpha_n(t)) \cdot \alpha_n(t), ...$$

and a numerical function $\alpha : (0; +\infty) \to (0; +\infty)$ is said to be **geometrically summable** if the series

$$t + \alpha(t) \cdot t + ... + \alpha_{n+1}(t) + ... = \alpha_\infty(t) \cdot t$$

is convergent over all positive reals.

It is not too hard to check that every numerical function which all right upper limits over $[0; +\infty)$ are less than 1 provides an example of geometrically summable function, see [9]. Finally we say that a family $\{\alpha_\gamma(\cdot)\}_{\gamma\in\Gamma}$ of functions **is "less than 1"** and use the notation $\{\alpha_\gamma(\cdot)\}_{\gamma\in\Gamma} \prec 1$ whenever $\alpha^+_\gamma(t) < \alpha(t),\ t > 0,\ \gamma \in \Gamma$ for some geometrically summable function $\alpha(\cdot)$, where for a function g we denote by g^+ the function of right upper limits of g.

For the standard notions of multivalued analysis and topology see, for example, books [1, 2, 5].

Below denote by D_r an arbitrary open ball with the radius r in a metric space. For a nonempty closed subset $P \subset Y$ of a normed space Y its *function of nonconvexity* $\alpha_P(\cdot)$ associates to each positive number r the following nonnegative number

$$\alpha_P(r) = \sup\left\{ \quad \sup\left\{ \frac{dist(q,P)}{r} \mid q \in conv(P \cap D_r) \right\} \mid D_r \text{ is an open } r\text{-ball} \right\}.$$

Clearly, $\alpha_P(r) = 0$ implies $q \in P$, or $conv(P \cap D_r) \subset P$. Hence the identity $\alpha_P(\cdot) \equiv 0$ is equivalent to the convexity of the closed set P. A closed subset $P \subset Y$ is said to be **paraconvex (functionally paraconvex)** if its function of nonconvexity $\alpha_P(\cdot)$ doesn't exceed some constant $C < 1$ (is "less than 1", i.e. $\alpha_P(\cdot) \prec 1$).

Theorem 0, [9, 10]. *Let $F : X \to Y$ be a lower semicontinuous mapping from a paracompact space X into a Banach space Y and let $\left\{\alpha_{F(x)}\right\} \prec 1$, $x \in X$. Then F admits a continuous singlevalued selection $f : X \to Y$, $f(x) \in F(x)$.*

In the case $\alpha_{F(x)}(\cdot) \equiv 0$, $x \in X$ Theorem 0 coincides with the classical Michael's selection theorem for convexvalued mappings and in the case $\sup\left\{\alpha_{F(x)}(r) |\ x \in X, r > 0\right\} < 1$ it coincides with Michael's selection theorem for paraconvexvalued mappings, see [3].

Having in mind that each retraction problem is a partial case of an extension problem and each extension problem in turn is a partial case of a selection problem, Theorem 1 guarantees an existence of some continuous retraction, say $R : Y \to P$ with $R(p) = p$, $p \in P$, for each functionally paraconvex subset $P \subset Y$. Such a retraction induced on P some kind of a generalized convex structure by setting

$$conv_R\{p_1, ..., p_n\} \quad = \quad R(conv\{p_1, ..., p_n\}).$$

Unfortunately, such a convex structure in P is in general uncomplete (in the convex sense). Namely, closed $R-$convex hull of a subcompacta in P fails to be a compact set. So, to achieve such a key point as completeness we need some special type of a retraction onto P.

Theorem 1. *Each functionally paraconvex subset of a Banach space is its uniform retract.*

Recall that a retraction $R : Y \to P$ is said to be *uniform*, see [4], (or *regular*, see [11]) at P if for each $\varepsilon > 0$ there is $\delta > 0$, such that $dist(x, R(x)) < \varepsilon$ whenever $dist(x, P) < \delta$. So, the set $U(P)$ of all uniform retraction onto an arbitrary functionally paraconvex subset P is nonempty. In the case of *bounded* P the set $U(P)$ is closed subset of another (exponential type) Banach space $B = CB(Y, Y)$ of all continuous bounded selfmappings of Y.

It appears that the $\alpha-$paraconvexity of a set implies $\beta-$paraconvexity of the set of all uniform retractions onto this set with some $\beta = \beta(\alpha)$.

Theorem 2. *If $\alpha_P(\cdot)$ is (strongly) majorated by $\alpha(\cdot)$ then $\alpha_{U(P)}$ is (strongly) majorated by $\alpha_\infty(\cdot) - 1$.*

For a function $\alpha(\cdot) \prec 1$ denote $\mathcal{P}_\alpha(Y)$ the family of all $P \subset Y$ with $\alpha_P(\cdot) < \alpha(\cdot)$. With respect to the Hausdorff distance $\mathcal{P}_\alpha(Y)$ is a metric and, hence, is a paracompact space.

Theorem 3. *Multivalued mapping $U : \mathcal{P}_\alpha(Y) \to CB(Y;Y)$ which associates to each $P \in \mathcal{P}_\alpha(Y)$ the subset $U(P) \subset CB(Y;Y)$ is lower semicontinuous.*

Note, Theorem 3 is one more reason to work with uniform retractions because exactly uniformity implies lower semicontinuity of U.

Corollary 4. *Let $2\alpha(\cdot) \prec 1$. Then the mapping $U : \mathcal{P}_\alpha(Y) \to CB(Y;Y)$ admits a continuous singlevalued selection.*

Corollary 5. *Let $2\alpha(\cdot) \prec 1$ and let Ω be an arbitrary family of pairwise disjoint closed subsets P of a Banach space Y such that $\alpha_P < \alpha,\ P \in \Omega$. Then there exists a generalized complete convex structure ω on Y such that all members of the family Ω are convex with respect on ω.*

3 Proofs

Corollary 4 $\Longrightarrow$ Corollary 5.

Pick any continuous singlevalued selection, say u, of $U : \mathcal{P}_\alpha(Y) \to CB(Y;Y)$, $u(P) \in U(P)$. So, $u(P)$ is uniform retraction of the entire space Y onto P and, moreover, $u(P)$ as an element of $CB(Y;Y)$ continuously depends on P. Hence, setting

$$conv_\omega\{p_1, ..., p_n\} = u(P)(conv\{p_1, ..., p_n\})$$

for every finite set of points $p_1, ..., p_n$ from an arbitrary $P \in \Omega$ one defines a needed generalized convex complete structure ω.

Theorems 0-3 $\Longrightarrow$ Corollary 4.

The assumption $2\alpha(\cdot) \prec 1$ implies that $\alpha_\infty(\cdot) - 1 \prec 1$. Hence the mapping $U : \mathcal{P}_\alpha(Y) \to CB(Y;Y)$ is multivalued mapping which is lower semicontinuous (see Theorem 3) and whose values are nonempty (see Theorem 1) and $\beta(\cdot)-$ paraconvex sets with $\beta = \alpha_\infty - 1 \prec 1$ (see Theorem 2). That is why Theorem 0 is applicable to this multivalued mapping and so it gives a desired selection.

Theorems.

Fix a functionally paraconvex subset $P \subset Y$ with $\alpha_P < \alpha \prec 1$. To stress directly and immediately the effect of paraconvexity let us consider the case of *continuous* majorant $\alpha(\cdot) < 1$ with $\alpha(0) = lim_{t\to 0}\, \alpha(t) < 1$.

Let us denote by $d(x) = dist(x, P)$ the distance between a point $x \in Y$ and P. Clearly $d(\cdot)$ is continuous (Lipshitz, in fact) numerical function.

For every $x \in Y \setminus P$ consider the intersection $P \cap D(x, 2d(x))$ and define the convexvalued mapping $H_1 : Y \setminus P \to Y$ by setting

$$H_1(x) = \overline{conv}\{P \cap D(x, 2d(x))\}.$$

This mapping is lower semicontinuous (continuous, in fact) mapping defined on the paracompact domain $Y \setminus P$ with nonempty closed convex values in Banach space Y. So the classical Michael's selection theorem guarantees the existence of a continuous singlevalued selection, say $h_1 : Y \setminus P \to Y,\ h_1(x) \in H_1(x)$.

The inequality $\alpha_P < \alpha$ implies the inequalities

$$dist(h_1(x), P) \leqslant \alpha_P(2d(x)) \cdot 2d(x) < \alpha(2d(x)) \cdot 2d(x) = \alpha_1(2d(x)),$$

$$dist(x, h_1(x)) \leqslant 2d(x) = \alpha_0(2d(x)), \quad x \in Y \setminus P.$$

Similarly, define the convexvalued and closedvalued lower semicontinuous mapping $H_2 : Y \setminus P \to Y$ by setting

$$H_2(x) = \overline{conv}\{P \cap D(h_1(x), \alpha_1(2d(x)))\}, \quad x \in Y \setminus P.$$

and find its continuous singlevalued selection $h_2 : Y \setminus P \to Y,\ h_2(x) \in H_2(x)$. Analogously we see that

$$dist(h_2(x), P) \leqslant \alpha_P(\alpha_1(2d(x))) \cdot \alpha_1(2d(x)) < \alpha_2(2d(x)),$$

$$dist(h_2(x), h_1(x)) \leqslant \alpha_1(2d(x)), \quad x \in Y \setminus P.$$

once again due to the inequality $\alpha_P < \alpha$ for the function of nonconvexity α_P.

Arguing inductively we find a sequence $\{h_n\}_{n=1}^{\infty}$ of continuous singlevalued mappings $h_n : Y \setminus P \to Y$ such that

$$dist(h_{n+1}(x), P) < \alpha_{n+1}(2d(x)), \quad dist(h_{n+1}(x), h_n(x)) \leqslant \alpha_n(2d(x)).$$

Due to the geometric summability of $\alpha(\cdot)$ we see that the sequence $\{h_n\}_{n=1}^{\infty}$ is locally uniformly fundamental and hence its pointwise limit $h(x) = lim_{n\to\infty}\, h_n(x)$ is well-defined and continuous. Moreover, $h(x) \in P,\ x \in Y \setminus P$ due to the closedness of P and convergency of $\{h_n\}_{n=1}^{\infty}$.

Hence the mapping $R : B \to P$ defined by $R(x) = h(x),\ x \in Y \setminus P$ and $R(x) = x,\ x \in P$ is a retraction of B onto P which is continuous over the set $Y \setminus P$ just by its construction.

Moreover for every $x \in Y \setminus P$:

$$dist(x, h(x)) \leqslant dist(x, h_1(x)) + \sum_{n=1}^{\infty} dist(h_n(x), h_{n+1}(x)) \leqslant$$
$$\leqslant \alpha_0(2d(x)) + \alpha_1(2d(x)) + \alpha_2(2d(x)) + .. = \alpha_\infty(2d(x)) \cdot 2d(x),$$

So for $x_0 \in P$ and for $x \in Y \setminus P$ we have

$$dist(R(x_0), R(x)) = dist(x_0, h(x)) \leqslant dist(x_0, x) + dist(x, h(x)) \leqslant$$
$$\leqslant dist(x_0, x) + \alpha_\infty(2d(x)) \cdot 2d(x) \leqslant (1 + 2\alpha_\infty(2d(x))) \cdot dist(x_0, x).$$

Pick a numbers $0 < q < 1$ and $\delta > 0$ such that $\sup\{\alpha(t) \mid t < 2\delta\} \leqslant q$. Then for all x which are $\delta-$close to x_0 we see that $2d(x) \leqslant 2d(x_0, x) < 2\delta$ and $\alpha_\infty(2d(x)) \leqslant \frac{1}{1-q}$. Hence

$$dist(R(x_0), R(x)) = dist(x_0, h(x)) \leqslant C \cdot dist(x_0, x).$$

for some constant $C > 0$ and R is continuous at x_0. The uniformity of R can be proved similarly.

For the case of an arbitrary $\alpha(\cdot)$ the proof above doesn't work because there are no chances to guarantee an appropriate type of continuity of the maps H_n. So, in such a situation the proof is too much complicated and a resulting retraction can be constructed as an uniform limit of some sequence of singlevalued δ_n-continuous ε_n-selections, based on a simultaneous using of the method of outside and the method of inside approximations, see [6].

The proofs of Theorems 2 and 3 in general follow to proofs of Propositions 2.2 and 2.3 from [8] but with some additional technique as in the proof of Theorem 1, above.

4 Open questions

Q1. Does an analytical characterization of the class of all geometrically summable functions exist?

Q2. Are Corollaries 4 and 5 true with the substitution $\alpha(\cdot) \prec 1$ instead of $2\alpha(\cdot) \prec 1$?

Q3. Is it true that the graph of an arbitrary Lipshitz mapping between Euclidean spaces is a paraconvex subsets of their Cartesian product?

References

1. J.-P. Aubin and H. Frankowska, *Set-Valued Analysis.* Birkhäuser, Basel, 1990.
2. R. Engelking, *General Topology.Revised and Completed edition.* Heldermann Verlag, Berlin, 1989.
3. E. Michael, *Paraconvex sets.* Math. Scand. **7**, 1959. pp 372-376.
4. E. Michael, *Uniform AR's and ANR's.* Compositio Math. **39**, 1979. pp. 129-139.
5. D. Repovš and P. V. Semenov, *Continuous Selections of Multivalued Mappings. Mathematics and Its Applications.* **455**, Kluwer, Dordrecht, 1998.
6. D. Repovš and P. Semenov, *Selections as an uniform limits of δ-continuous ε-selections.* Set-Valued Analysis. **7**, 1999. pp. 239-254.
7. D. Repovš and P. Semenov, *A minimax theorem for functions with possibly nonconnected intersections of sublevel sets.* Journal.of Math. Analysis and Appl. **314**, 2006. pp. 537-545 .
8. D. Repovš and P. Semenov, *On continuous choice of retractions onto noconvex subsets.* Topology and its Appl. **157**, 2010. pp. 1510-1517.
9. P. V. Semenov, *Functionally paraconvex sets* . Math. Notes. **54**, 1993. pp.1236-1240.
10. P. V. Semenov, *Nonconvexity in problems of multivalued calculus.* J. Math. Sci (N.Y.). **19**, 2000. pp.2682-2699 .
11. H. Torunczyk *Absolute retracts as factors of normed linear spaces.* Fund. Math. **86**, 1974. pp.53-67.

P. V. Semenov
Department of Mathematics, Moscow City Pedagogical University, pavels@orc.ru

OPTIMAL CONTROL THEORY TO ANALYSIS OF NONLINEAR PDE'S OF THE I-ST ORDER

N. N. Subbotina, E. A. Kolpakova

Key words: Hamilton-Jacobi-Bellman equations, minimax/viscosity solution, method of characteristics, subdifferentials, Dini semiderivatives

AMS Mathematics Subject Classification: 35D40

Abstract. The 1-st order PDE's concave relative impulse variables are considered as local Hamilton–Jacobi–Bellman equations. The theory of optimal control and the nonsmooth analysis tools are applied to describe the structure of the minimax or viscosity solutions of the PDE's in the boundary Cauchy problem under Lipschitz conditions on the boundary terminal function and the right-hand side of the corresponding Hamiltonian ODE's.

1 Introduction

Nonsmooth global generalized solutions to PDE's are vital to practice. The functions arise in optimal control theory and differential games as the value functions which are generalized solutions of the Bellman–Isaacs equations (see, for example works by L.S. Pontryagin, R. Bellman, R. Isaacs, N.N. Krasovskii). The function described the evolution of the front of a light wave in nonhomogeneous environment is a generalized solution of the eikonal equation. The functions described the evolution of some market indexes and the functions described the evolution of some genetic indexes are generalized solutions of corresponding Hamilton-Jacobi equations and so on.

The foundation of the theory of the generalized solutions to PDE's was builded in the 50-th 60-th of the XX century in the works by O.A. Oleinik, O.A. Ladyzhenskaya, E. Hopf, R. Courant, I.M. Gelfand, A.N. Tikhonov, A.A. Samarskii, S.L. Sobolev, N.S. Bakhvalov, P. Lax, W.H. Fleming, C.M. Dafermos and others.

Later in 70-th S.N. Kruzhkov suggested the notion of entropy solution, based on the integral relations.

In 80-th M. Crandall, P.L. Lions introduce viscosity solutions, A.I. Subbotin suggested minimax solutions.

The work was supported by the Russian Foundation for Basic Researches (project No. 11-01-00214) and by the Program of Presidium RAS for Basic Researches on Mathematical Control Theory.

The paper presents results obtained in the framework of the theory of minimax solutions to PDE's of the 1-st order.

We consider the 1-st order PDE's concave relative impulse variables in the boundary Cauchy problem under Lipschitz conditions on the boundary terminal function and the right-hand side of the corresponding Hamiltonian ODE's. It is proven that the equations can be considered locally as the Hamilton-Jacobi-Bellman equations. It gives possibilities to apply the theory of optimal control to describe the minimax solutions to the PDE's. Necessary optimality conditions (L.S. Pontryagin's maximum principle [2]) in the Hamiltonian form (following to F.H. Clarke [9]) play a key part in our researches.

The paper continues the works [10, 11]. Using results by authors on structure of the minimax solutions of HJBE's to optimal control problems, properties of the minimax solutions of the considered PDE's of the 1-st order are described. New tools of the nonsmooth analysis (directional subdifferentials) are suggested to get new relations for the structure. The new tools can be reduced by application of approach suggested by B. Mordukhovich [8].

2 Statement

The paper deals with the boundary Cauchy problem for HJBE

$$\frac{\partial \varphi(t,x)}{\partial t} + H(t,x,D_x\varphi(t,x)) = 0, \quad \varphi(T,x) = \sigma(x), \tag{2.1}$$

Here $(t,x) \in \Pi_T = [0,T] \times \mathbb{R}^n$, $D_x\varphi = \left(\frac{\partial \varphi}{\partial x_1}, \ldots, \frac{\partial \varphi}{\partial x_n}\right) = s \in \mathbb{R}^n$.

The problem is considered under the following assumptions.

A1 The Hamiltonian $H(t,x,s)$ is continuous relative to all variables in $\Pi_T \times \mathbb{R}^n$ and concave relative to s.

A2 There exist $D_xH(t,x,s), D_sH(t,x,s)$ satisfied the Lipschitz conditions in x,s with a constant $L > 0$:

$$||D_sH(t,x_1,s_1) - D_sH(t,x_2,s_2)|| \leqslant L(||x_1 - x_2|| + ||s_1 - s_2||),$$

$$||D_xH(t,x_1,s_1) - D_xH(t,x_2,s_2)|| \leqslant L(||x_1 - x_2|| + ||s_1 - s_2||);$$

A3 The terminal function $\sigma(x)$ is local Lipschitz continuous in $\mathbb{R}^n$:

$$|D_x\sigma(x_1) - D_x\sigma(x_2)| \leqslant L(M)(||x_1 - x_2||,$$

$$x_1, x_2 \in M \subset \mathbb{R}^n, \quad 0 < L(M)text - - const.$$

The problem doesn't have a global classical solution, as a rule. We consider the minimax solution of problem (2.1) following to A.I. Subbotin [3] and investigate its structure. Note, that the solution is equivalent to the viscosity solution introduced by M. Crandall and P.L. Lions [4]. Recall necessary notions.

Definition 1. The lower Dini semiderivative $\frac{d^-\varphi(y)}{h}$ of a function $\varphi : \mathbb{R}^m \to R$ at a point y in a direction $(h) \in \mathbb{R}^m$ is defined as follows:

$$\frac{d^-\varphi(y)}{h} = \liminf_{\delta\to 0, h'\to h} \frac{\varphi(y+\delta h') - \varphi(y)}{\delta}.$$

Similarly the upper Dini semiderivative $\frac{d^+\varphi(y)}{h}$ is defined as lim sup.

Definition 2. A function $\varphi(\cdot,\cdot) : \Pi_T \to R$ is called the minimax solution of (2.1) iff

$$\varphi(T,x) = \sigma(x), \quad \forall x \in \mathbb{R}^n,$$

$$\sup_{s\in\mathbb{R}^n} \inf_{h\in\mathbb{R}^n} \left\{ \frac{d^-\varphi(t,x)}{1,h} - \langle s,h\rangle + H(t,x,s) \right\} \leqslant 0,$$

$$\inf_{s\in\mathbb{R}^n} \sup_{h\in\mathbb{R}^n} \left\{ \frac{d^+\varphi(t,x)}{1,h} - \langle s,f\rangle + H(t,x,s) \right\} \geqslant 0,$$

for all $(t,x) \in (0,T)\times\mathbb{R}^n$.

3 Preliminaries

The section contains some assertions followed from the theory of generalized solutions to PDE's [3, 5] and useful tools to the nonsmooth analysis.

Assertion 1. If assumptions $A1$–$A3$ are true, then there exists a local Lipschitz continuous minimax solution to problem (2.1). The minimax solution $\varphi(\cdot)$ is unique and coincides with the viscosity solution to the problem.

Definition 3. The set $\partial\psi \subset R^m$ is called the subdifferential of the function $\psi(\cdot) : \mathbb{R}^m \to R$ at the point $y \in R^m$ if it has the form

$$\partial\psi(y) = \mathrm{co}\{q \in R^m : \ q = \lim_{y_k\to y} D\psi(y_k)\}.$$

Here $D\psi(y_k)$ are the gradients of the function $\psi(\cdot)$ at points y_k. The symbol co denotes the convex hull.

Consider the Hamiltonian system to problem (2.1)

$$\dot{\tilde{x}} = D_{\tilde{s}}H(t,\tilde{x},\tilde{s}), \quad \dot{\tilde{s}} = -D_{\tilde{x}}H(t,\tilde{x},\tilde{s}), \quad \dot{\tilde{z}} = \langle D_{\tilde{s}}H(t,\tilde{x},\tilde{s}),\tilde{s}\rangle - H(t,\tilde{x},\tilde{s}) \quad (3.1)$$

$$\tilde{x}(T,\xi)=\xi,\ \tilde{s}(T,\xi)\in\partial\sigma(\xi),\ \tilde{z}(T,\xi)=\sigma(\xi),\quad \xi\in\mathbb{R}^n. \tag{3.2}$$

Solutions $\tilde{x}(\cdot,\xi),\tilde{s}(\cdot,\xi),\tilde{z}(\cdot,\xi)$ satisfied (3.1)-(3.2) are called the characteristics to problem (2.1).

Assertion 2. If assumptions $A1$–$A3$ are true, then for any $\xi\in\mathbb{R}^n$ there exists the unique solution of the characteristic system (3.1)-(3.2), and it is extendable on the interval $[0,T]$.

4 Applications of the theory of optimal control

Let us to introduce an auxiliary optimal control problem

$$\dot{x}=D_pH(t,x,p),\ p\in P_0,\quad x(t_0)=x_0,\quad (t_0,x_0)\in\Pi_T. \tag{4.1}$$

We consider the set of admissible controls

$$U_0=\{\forall p(\cdot):[0,T]\to P_0 \text{ are measurable }\}$$

and the cost functional:

$$I_{t_0,x_0}(p(\cdot))=\sigma(x(T))-\int_{t_0}^{T}\langle p(\tau),D_pH(\tau,x(\tau),p(\tau))\rangle-H(\tau,x(\tau),p(\tau))d\tau=$$

$$=\sigma(x(T))-\int_{t_0}^{T}H^*(\tau,x(\tau),D_pH(\tau,x(\tau),p(\tau)))d\tau. \tag{4.2}$$

The symbol $\langle\cdot,\cdot\rangle$ denotes the inner product, $x(\cdot)=x(\cdot;t_0,x_0,p(\cdot))$ is a trajectory of (4.1) under an admissible control $p(\cdot)\in U_0$, the symbol H^* denotes

$$H^*(t,x,l)=\inf_{p\in R^n}[\langle l,p\rangle-H(t,x,p)].$$

The goal of controls $p(\cdot)$ in problem (4.1)-(4.2) is to minimize the cost. The map

$$(t_0,x_0)\to V(t_0,x_0)=\inf_{p(\cdot)\in U_0}I_{t_0,x_0}(p(\cdot))$$

is called the value function.

The following assertion follows from results of the theory of minimax/viscosity solutions to HJE's [1,3,4].

Assertion 3. If assumptions $A1$–$A3$ are true, then the value function $V(t,x)$ to optimal control problem (4.1)-(4.2) coincides with the unique minimax/viscosity solution to the problem

$$\frac{\partial V(t,x)}{\partial t} + \mathbb{H}(t,x,D_xV(t,x)(t,x)) = 0, \quad (t,x) \in \Pi_T, \tag{4.3}$$

$$V(T,x) = \sigma(x), \quad x \in R^n. \tag{4.4}$$

The Hamiltonian to problem optimal control (4.1)-(4.2) has the form

$$\mathbb{H}(t,x,s) = \min_{p\in P_0}\langle s, D_pH(t,x,p)\rangle - H^*(t,x(t),D_pH(t,x(t),p)). \tag{4.5}$$

Note, that the Hamiltonian to problem (2.1) is of the view

$$H(t,x,s) = \min_{p\in R^n}\langle s, D_pH(t,x,p)\rangle - H^*(t,x(t),D_pH(t,x(t),p). \tag{4.6}$$

Let us recall the Pontryagin's maximum principle [2] in the Hamiltonian form by Clarke [9].

Assertion 4 If assumptions $A1$–$A3$ are true, $(t_0,x_0) \in (0,T)\times R^n$, $p^0(\cdot) \in U_0$ and

$$I_{t_0,x_0}(p^0(\cdot)) = V(t_0,x_0),$$

then there exists such a function $s^*(\cdot) : [t_0,T] \to R^n$, that the following conditions are valid for all $t \in [t_0,T]$

$$\dot{x^0}(t) = D_pH(t,x^0(t),p^0(t)) \in \partial_p\mathbb{H}(t,x^0(t),s^*(t)), \quad x^0(t_0) = x_0; \tag{4.7}$$

$$\dot{s^*}(t) \in -\partial_x\mathbb{H}(t,x^0(t),s^*(t)), \quad s^*(T) \in \partial\sigma(x^0(T)); \tag{4.8}$$

$$\mathbb{H}(t,x^0(t),s^*(t)) = \langle s^*(t), \dot{x^0}(t))\rangle - H^*(t,x^0(t),\dot{x^0}(t)). \tag{4.9}$$

Using schemes of proofs in [5] for analogous results obtained under stronger assumptions we get the theorem.

Theorem 1. If assumptions $A1$–$A3$ are true, $(t_0,x_0) \in (0,T)\times R^n$, $p^0(\cdot) \in U_0$, then there exist such compact sets $P_0 \in R^n$, $\Omega_0 \in \Pi_T$, that the relations hold

$$(t,x(t;t_0,x_0,p(\cdot))) \in \Omega_0, \quad \forall\ t \in [t_0,T], \forall\ p(\cdot) \in U_0; \tag{4.10}$$

$$\tilde{s}(t,\xi) \in P_0, \quad \forall\, t \in [t_0,T], \forall \xi = x(T;t_0,x_0,p(\cdot)); \tag{4.11}$$

$$H(t,x,p) = \mathbb{H}(t,x,p), \quad \forall\, (t,x,p) \in \Omega_0 \times P_0; \tag{4.12}$$

$$\partial_p \mathbb{H}(t,x,p) = \{\partial_p H(t,x,p)\}, \quad \partial_x \mathbb{H}(t,x,p) = \{\partial_x H(t,x,p)\}, \tag{4.13}$$

for all $(t,x,p) \in \Omega_0 \times P_0$.

Remark 1. If $A1$–$A3$ are true then Theorem 1 and Assertion 4 provide the equalities

$$\dot{x^0}(t) = D_p H(t,x^0(t),p^0(t)) = D_p \mathbb{H}(t,x^0(t),s^*(t)) = D_p H(t,x^0(t),s^*(t)); \tag{4.14}$$

$$\begin{gathered} \dot{s^*}(t) = -D_x \mathbb{H}(t,x^0(t),s^*(t)) = -D_x H(t,x^0(t),s^*(t)); \\ x^0(t_0) = x_0, \quad s^*(T) \in \partial\sigma(x^0(T)); \end{gathered} \tag{4.15}$$

$$\begin{aligned} \mathbb{H}(t,x^0(t),s^*(t)) &= \langle s^*(t), \dot{x^0}(t)) \rangle - H^*(t,x^0(t),\dot{x^0}(t)) = H(t,x^0(t),s^*(t)) = \\ &= \langle s^*(t), D_p H(t,x^0(t),s^*(t)) \rangle - H^*(t,x^0(t),D_p H(t,x^0(t),s^*(t))). \end{aligned} \tag{4.16}$$

Corollary 1. If assumptions $A1$–$A3$ are true, $(t_0,x_0) \in (0,T) \times R^n$, $p^0(\cdot) \in U(D)$ and

$$I_{t_0,x_0}(p^0(\cdot)) = V(t_0,x_0),$$

then there exist such characteristics (3.1) (3.2) $\tilde{x}(\cdot,\xi), \tilde{s}(\cdot,\xi), \tilde{z}(\cdot,\xi)$, that the following relations hold

$$x^0(t) = x^0(t;t_0,x_0,p^0(\cdot)) = \tilde{x}(t,\xi), \quad \forall t \in [t_0,T];$$

$$s^*(\cdot) = \tilde{s}(\cdot,\xi), \quad x^0(\cdot) = x^0(\cdot;t_0,x_0,\tilde{s}(\cdot,\xi));$$

$$I_{t_0,x_0}(\tilde{s}(\cdot,\xi)) = \tilde{z}(t_0,\xi).$$

Using equality (4.12) for the Hamiltonians $H(t,x,p)$ and $\mathbb{H}(t,x,p)$ in $\Omega_0 \times P_0$, Remark 1 and Assertion 3, one can obtain equivalence for the value function to problem (4.1)-(4.2) and the minimax solution to problem (2.1) in the domain Ω_0.

Following schemes of proofs in [5] we obtain necessary and sufficient conditions for the minimax solution $\varphi(t,x)$ to problem (2.1).

Theorem 2. If assumptions $A1$–$A3$ are true, $(t_0, x_0) \in \Omega_0$, then $\varphi(t_0, x_0) = V(t_0, x_0)$

$$\min_{(f,g)\in F(t_0,x_0)} \left[\frac{d^{\pm}\varphi(t_0, x_0)}{1, f} + g \right] = 0, \tag{4.17}$$

where

$$F(t_0, x_0) = \text{co } \{(D_p H(t_0, x_0, p), H^*(t_0, x_0, D_p H(t_0, x_0, p))) : p \in P_0\}.$$

Corollary 1 and the necessary optimality conditions in Assertion 4 provide the representative formula for the minimax solution $\varphi(\cdot, \cdot)$ in Ω_0.

Theorem 3. If assumptions $A1$–$A3$ are true, $(t_0, x_0) \in \Omega_0$, then

$$\varphi(t_0, x_0) = V(t_0, x_0) = \min_{\xi:\tilde{x}(t_0,\xi)=x_0} \tilde{z}(t_0, \xi), \tag{4.18}$$

where $\tilde{x}(\cdot, \xi), \tilde{s}(\cdot, \xi), \tilde{z}(\cdot, \xi)$ are characteristics (3.1)-(3.2).

5 Structure of the minimax solution

The section contains some assertions followed from the theory of generalized solutions to HJB's [10, 11] and new tools to the nonsmooth analysis.

Assertion 5. Let assumptions $A1$–$A3$ be true, $(t_0, x_0) \in (0, T) \times R^n$. The minimax solution $\varphi(t, x)$ to problem (2.1) is not differentiable at (t_0, x_0), iff there exist such $\xi_1, \xi_2 \in \mathbb{R}^n$, $\xi_1 \neq \xi_2$ that

$$\tilde{x}(t, \xi_1) = \tilde{x}(t, \xi_2) = x, \ \tilde{z}(t, \xi_1) = \tilde{z}(t, \xi_2) = \varphi(t, x), \ \tilde{s}(t, \xi_1) \neq \tilde{s}(t, \xi_2). \tag{5.1}$$

Assertion 6. If assumptions $A1$-$A3$ are valid and the state space is one-dimensional, then all points of nondifferentiability of the minimax solution $\varphi(t, x)$ lie on not more than countable number of lines $t \to x_*(t) : 0 \leqslant t_* < t \leqslant T$ satisfying the Rankine-Hugoniot condition:

$$\frac{dx_*(t)}{dt} = \frac{H(t, x_*(t), D_+\varphi(t, x_*(t))) - H(t, x_*(t), D_-\varphi(t, x_*(t)))}{D_+\varphi(t, x_*(t)) - D_-\varphi(t, x_*(t))}, \tag{5.2}$$

$$D_+\varphi(t, x_*(t)) = \lim_{x\to x_*(t)+0} D\varphi(t, x),$$

$$D_-\varphi(t, x_*(t))) = \lim_{x\to x_*(t)-0} D\varphi(t, x),$$

and the inequality

$$D_-\varphi(t, x_*(t))) < D_+\varphi(t, x_*(t)))$$

is valid.

We introduce the new tool to nonsmooth analysis.

Definition 4. The set $\partial\psi_h \subset R^m$ is called the partial subdifferential of the function $\psi(\cdot)$ at the point y in the direction h if the set has the form

$$\partial\psi_h(y) = \mathrm{co}\{\xi \in R^m : \ \xi = \lim_{y_k \to y} D\psi(y_k)\}.$$

Here $D\psi(y_k)$ are the gradients of the function $\psi(\cdot)$ at points y_k :

$$\limsup_{\delta_k \downarrow 0} \frac{y_k - y - h\delta_k}{\delta_k} = 0.$$

Theorem 4. If assumptions $A1$–$A3$ are true, $(t_0, x_0) \in (0,T) \times R^n$, and there exist such characteristics $\tilde{x}(\cdot,\xi_*), \tilde{s}(\cdot,\xi_*), \tilde{z}(\cdot,\xi_*)$ to problem (2.1) that $\tilde{x}(t_0,\xi_*) = x_0$, and the following relations hold for all $t \in [t_0, T]$

$$\Big(- H(t, \tilde{x}(t,\xi_*), \tilde{s}(t,\xi_*)), \quad \tilde{s}(t,\xi_*)\Big) \in \partial V_{h(t)}(t, \ \tilde{x}(t,\xi_*)), \tag{5.3}$$

$$h(t) = (1, \ D_x H(t, \tilde{x}(t,\xi_*), \tilde{s}(t,\xi_*))),$$

then the minimax solution to problem (2.1) satisfies the equalities

$$\varphi(t_0,x_0) = \tilde{z}(t_0,\xi_*) = \min_{\xi : \tilde{x}(t_0,\xi) = x_0} \tilde{z}(t_0,\xi).$$

Proof. According to Theorem 1 and 2, for any $(t_0,x_0) \in (0,T) \times R^n$, the minimax solution $\varphi(t,x)$ to problem (2.1) coinsides with the value function $V(t,x)$ to an auxiliary optimal control problem (4.1)-(4.2) in a domain $\Omega_0 \ni (t_0,x_0)$. We show, that the derivative in the direction $D_s H(t,x,s)$ of the function V equals 0. It is essential and sufficient condition of optimality.

According to the formula, proved in the work [6], the inequalities are valid

$$\min_{(\alpha,p) \in \partial_{1,h} V(t,x)} \langle (\alpha,p), (1,h) \rangle \leqslant \frac{d^- V(t,x)}{1,h} \leqslant \frac{d^+ V(t,x)}{1,h} \leqslant \max_{(\alpha,p) \in \partial_{1,h} V(t,x)} \langle (\alpha,p), (1,h) \rangle.$$

Consider the expression

$$\langle (\alpha,p), (1,h) \rangle, \ (\alpha,p) \in \partial_{1,h} V(t,x), \quad h = D_s H(t,x,s).$$

Remind that

$$\alpha = \lim_{k \to \infty} -H(t_k, x_k, D_x V(t_k,x_k)), p = \lim_{k \to \infty} D_x V(t_k,x_k).$$

Then

$$\langle(\alpha,p),(1,h)\rangle = \lim_{k\to\infty} -H(t_k,x_k,D_xV(t_k,x_k)) + \langle D_xV(t_k,x_k),h\rangle.$$

Note that $-H(t_k,x_k,D_xV(t_k,x_k)) + \langle D_xV(t_k,x_k),h\rangle = 0$, because t_k, x_k are the points of differentiability of function V.

This theorem gives sufficient optimality conditions for the case of lipschitz continuous data.

Theorem 2 and theorem 4 imply the assertion.

Corollary 2. If assumptions $A1$–$A3$ are true, then a local lipschits continuous function $\tilde{\varphi}(\cdot) : \Pi_T \to R$ coincides with the minimax solution $\varphi(t,x)$ to problem (2.1) iff

$$\tilde{\varphi}(T,x) = \sigma(x), \quad \forall x \in R^n,$$

$$\min_{f\in R^n} \min_{(p,\alpha)\in\partial_h\tilde{\varphi}(t,x)} \langle f,p\rangle - \alpha - H^*(t,x,f) = 0,$$

where $h = (1,f)$.

The presented results are suitable for numerical methods of constructing the minimax solution φ to problem (2.1).

6 Example

Consider boundary Cauchy problem for Hamilton—Jacobi—Bellman equation:

$$u_t - (\sqrt{1+u_x^2}) = 0, \quad u(2,x) = -\frac{x^2}{2},\ t\in[0,2].$$

We construct the characteristic system for this problem

$$\dot{\tilde{x}}(t) = -\frac{\tilde{s}}{\sqrt{1+\tilde{s}^2}}, \dot{\tilde{s}}(t) = 0, \quad \dot{\tilde{z}}(t) = \frac{1}{\sqrt{1+\tilde{s}^2}},$$

with boundary condition

$$\tilde{x}(T,\xi) = \xi, \quad \tilde{s}(T,\xi) = -\xi, \quad \tilde{z}(T,\xi) = -\xi^2/2.$$

The solution of the characteristic system has the form

$$\tilde{x}(t) = -\frac{\tilde{s}(t-T)}{\sqrt{1+\tilde{s}^2}} + \xi, \tilde{s}(t) = -\xi,\ \tilde{z}(t) = \frac{(t-T)}{\sqrt{1+\tilde{s}^2}} - \frac{\xi^2}{2}.$$

The singular set for the minimax solution

$$(t, x) = \{x = 0,\ 0 \leqslant t < 1\}.$$

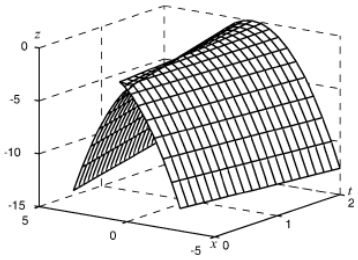


Figure 1. The surface base the characteristics $\tilde{x}(t), \tilde{z}(t)$

References

1. N.N. Krasovskii. *Theory of control of motions.* M.: Nauka, 1968. 476 pages.
2. L.S.Pontryagin, V.G.Boltyanskii, R.V.Gamkrelidze, E.F.Mishchenko. *Mathematical Theory of Optimal Processes.* New York: Interscience, 1962. 394 pages.
3. A.I.Subbotin. *Generalized Solutions of First Order of PDEs: The Dynamical Optimization Perspectives.* Boston: Birkhauser, 1995. 336 pages.
4. M.G. Crandall, P.L. Lions *Viscosity Solutions of Hamilton-Jacobi Equations.* Trans. Am. Math. Soc., **277**, 1, 1983. Pp. 1-42.
5. N.N. Subbotina. *The method of characteristics for Hamilton-Jacobi equations and applications to dynamical optimization.* Journal of Mathematical Sciences **135**, 3, 2006. Pp. 2955-3091.
6. N.N. Subbotina. *The generalized method of characteristics in the optimal control problem with lipschitz data.* Proceedings of the ural state university, **46**, 2006. Pp. 177-186.
7. R.T. Rockafellar, R.J-B. Wets. *Variational Analysis.* Springer-Verlag, 1997. 735 pages.
8. B. Mordukhovich. *Variational Analysis and Generalized Differentiation: Applications II.* Birkhauser, 2006. 610 pages.
9. F.N.Clarke. *Optimization and nonsmooth analysis.* New York: Wiley, 1983. 308 pages.

10. E.A.Kolpakova. *A generalized method of characteristics in the theory of Hamilton—Jacobi equations and conservation laws.* Proceedings of the Intstitute of Mathematics and Mechanics of Ural Branch of RAS **16**, 5, 2010. Pp. 95-102.
11. N.N.Subbotina, E.A.Kolpakova. *On the structure of locally Lipshitz minimax solutions of the Hamilton—Jacobi equations in the terms of classical characteristics.* Proceedings of the Intstitute of Mathematics and Mechanics of Ural Branch of RAS **15**, 3, 2009. Pp. 202-218.

N. N. Subbotina
Institute of mathematics and Mechanics of UrB RAS; Ural Federal University, Russia, 620990, Ekaterinburg, S.Kovalevskoy str.,16, +7(343)3628175, subb@uran.ru

E. A. Kolpakova
Institute of mathematics and Mechanics of UrB RAS; Ural Federal University, Russia, 620990, Ekaterinburg, S.Kovalevskoy str.,16, +7(343)3628176, eakolpakova@gmail.com

ON STABILITY OF METRIC REGULARITY UNDER PERTURBATIONS OF GENERAL TYPE

A. Uderzo

Key words: Metric regularity, Lipschitz-likeness, open covering, perturbations, radius of regularity

AMS Mathematics Subject Classification: 49J53, 46A30, 47H04, 49J52

Abstract. Metric regularity is a property for single as well as for set-valued mappings, which is connected with a certain Lipschitz behaviour of the solution map to generalized equations. Even though such property possesses a purely metric nature, it captures a so deep circle of phenomena to become soon a reference condition in a wide veriety of topics in optimization, control theory, nonlinear and variational analysis. The present note reports on some recent attempts to study robustness properties of metric regularity, i.e. its persistence in the presence of perturbations, in a purely metric setting.

1 Introduction

According to [1, 2], a mapping $F : X \longrightarrow 2^Y$ between metric spaces (X, d) and (Y, d) is said to be *metrically regular* on $U \times V$, with $U \subseteq X$ and $V \subseteq Y$, if there exist constants $\kappa > 0$ and $\gamma > 0$ such that

$$\operatorname{dist}(x, F^{-1}(y)) \leqslant \kappa \operatorname{dist}(y, F(x)), \quad \forall (x, y) \in U \times V : \operatorname{dist}(y, F(x)) \leqslant \gamma, \quad (1.1)$$

where $\operatorname{dist}(x, S)$ denotes the distance from a point x to a set S. Clearly inequality (1.1) provides an estimation of the distance from x to the solution set to the generalized equation $F(x) \ni y$. A triggering achievement of modern variational analysis was the understanding that the phenomenon captured by metric regularity has deep connections with open covering and Lipschitz behaviour (alias Aubin continuity) of multifunctions. In one form or another, such phenomenon arises in a wide variety of situations, not only as a desirable feature that mappings may happen to possess, but also as a requirement enabling the employment of powerful techniques of analysis such as implicit function theorems and penalization methods in constrained optimization, constraint qualification in deriving optimality conditions, qualification conditions in subdifferential calculus, controllability of systems in control theory (see [1, 2]). Recently metric regularity has been also exploited

to study differential equations (see [3]) and linked to fixed and coincidence point theory (see [4]).

The present contribution is focussed on the analysis of stability properties of metric regularity in the presence of perturbations. Such theme of research, which is motivated both by theoretical reasons and real-world applications (data of generalized equations are often subject to imprecise measurements and error/approximations) can be traced back to the very foundation of functional analysis. Indeed, the so-called Banach lemma on the invertible linear operators can be clearly regarded as a stability result, subsequently extended in several directions.

The aim of the investigations here reported is to tackle a crucial problem arisen in the current theory: since metric regularity, despite its relevant role in more structured settings, has a purely metric nature, it seems to be natural to conduct a perturbation stability analysis free from any reference to linear structures. The approach here adopted has been inspired by a research line proposed by A.V. Arutyunov et al. in [3].

2 Global results

Definition 1 A mapping $F : X \longrightarrow 2^Y$ between metric spaces is said to be *globally metrically regular* if there exists $\kappa > 0$ such that

$$\operatorname{dist}(x, F^{-1}(y)) \leqslant \kappa \operatorname{dist}(y, F(x)), \quad \forall (x, y) \in X \times Y.$$

The infimum over all values κ satisfying the above inequality, called *modulus of global metric regularity* of F, is denoted by $\operatorname{reg} F$, with the convention that $\operatorname{reg} F = +\infty$ indicates the failure of such property.

It turns out that global metric regularity can be equivalently reformulated in terms of *global openness at a linear rate*, in the sense that F is globally metrically regular iff there exists $a > 0$ such that $F(B(x, r)) \supseteq B(F(x), ar), \quad \forall x \in X,\ r > 0$. Another equivalent reformulation of the same property can be given in terms of *Lipschitz continuity* for the inverse F^{-1} mapping to F, requiring the existence of $l > 0$ such that $\operatorname{haus}(F^{-1}(y_1), F^{-1}(y_2)) \leqslant l\, d(y_1, y_2), \quad \forall y_1,\ y_2 \in Y$, where $\operatorname{haus}(A, B)$ stands for the Hausdorff distance of sets A and B. The infimum of all values l satisfying the above inequality is called *modulus of Lipschitz continuity* of F^{-1} on Y and will be denoted by $\operatorname{lip}(F^{-1})$. To complement the equivalence between metric regularity and Lipschitz continuity for the inverse in the global case one has the following quantitative relation $\operatorname{reg} F = \operatorname{lip}(F^{-1})$. According to the Banach-Schauder

theorem (alias the open mapping principle) and its modern generalizations, important examples of globally regular mappings acting in Banach spaces are for instance

— all bounded epimorphisms (linear and onto operators);
— all convex processes, i.e. set-valued mappings whose graph is a convex cone, which are onto.

In order to analyze stability properties of metric regularity in a metric space setting, general perturbations of mappings are formalized by means of mappings $\mathcal{H} : X \times Y \longrightarrow 2^Y$ acting on F, in such a way that the correspondingly perturbed mapping $F_{\mathcal{H}} : X \longrightarrow 2^Y$ is defined by

$$F_{\mathcal{H}}(x) = \mathcal{H}(x, F(x)) = \bigcup_{y \in F(x)} \mathcal{H}(x, y). \tag{2.1}$$

By proper specializations of $\mathcal{H}$ in more structured contexts, it possible to obtain various known types of perturbations, such as additive perturbations, product perturbations, enlargement perturbations, composition perturbations. Given a closed-valued mapping F, the class of all $\mathcal{H}$ such that the perturbed mapping $F_{\mathcal{H}}$ is both closed-valued and upper semicontinuous (for short, u.s.c.) is denoted by $\mathfrak{A}(F)$. Below a perturbation stability result for global metric regularity is provided, which needs no reference to any linear structure.

Theorem 1 Let $F : X \longrightarrow 2^Y$ be a mapping between metric spaces and let $\mathcal{H} \in \mathfrak{A}(F)$. Suppose that:

i) (X, d) is complete;

ii) F is globally regular with modulus $\operatorname{reg} F$;

iii) $\mathcal{H}(\cdot, y)$ is Lipschitz on X with modulus $\operatorname{lip}_X \mathcal{H}$, uniformly in $y \in Y$;

iv) $\mathcal{H}(x, \cdot)$ is globally regular with modulus $\operatorname{reg}_Y \mathcal{H}$, uniformly in $x \in X$.

Then, if $\operatorname{reg} F \cdot \operatorname{reg}_Y \mathcal{H} < 1/\operatorname{lip}_X \mathcal{H}$, also $F_{\mathcal{H}}$ is globally metrically regular and

$$\operatorname{reg} F_{\mathcal{H}} \leqslant \frac{\operatorname{reg} F \cdot \operatorname{reg}_Y \mathcal{H}}{1 - \operatorname{reg} F \cdot \operatorname{reg}_Y \mathcal{H} \cdot \operatorname{lip}_X \mathcal{H}}.$$

Theorem 1 has been established in its open covering reformulation in [5] (see Theorem 3.1). Its proof is based on a variational technique relying on the Ekeland principle, which allows to avoid the use of iterative schemes connected with Newton type methods.

Among the consequences of Theorem 1, it is worth mentioning the following set-valued generalization of Milyutin theorem, a fundamental result in the analysis of perturbation stability of metric regularity for nonlinear mappings (see [6]):

Corollary 1 Let $F : X \longrightarrow 2^Y$ and $H : X \longrightarrow 2^Y$ be set-valued mappings between metric spaces. Suppose that:

i) (X, d) is complete and (Y, d) is a linear metric space, whose metric is invariant with respect to translations;

ii) F is u.s.c. on X and metrically regular;

iii) H is compact-valued and it is Lipschitz continuous on X, with modulus such that $\operatorname{lip} H \cdot \operatorname{reg} F < 1$.

Then the sum mapping $F + H$ is still globally regular and it holds

$$\operatorname{reg}(F + H) \leqslant \frac{\operatorname{reg} F}{1 - \operatorname{lip} H \cdot \operatorname{reg} F}.$$

Another result stemming from Theorem 1 is the following:

Corollary 2 Let $F : X \longrightarrow 2^Y$ and $H : Y \longrightarrow 2^Y$ be set-valued mappings between metric spaces. Suppose that:

i) (X, d) is complete;

ii) F and H are both u.s.c. on X and Y, respectively;

iii) F and H are both globally metricaly regular;

Then their composition $H \circ F$ is globally metrically regular with modulus $\operatorname{reg}(H \circ F) \leqslant \operatorname{reg} F \cdot \operatorname{reg} H$.

Theorem 1 allows to expand in a metric space setting the concept of radius of regularity.

Definition 2 Given a set-valued mapping $F : X \longrightarrow 2^Y$, let $\mathfrak{S} \subseteq \mathfrak{A}(F)$. The quantity

$$\operatorname{rad}_{\mathfrak{S}}(F) = \inf\{\nu(\mathcal{H}) : mathcalH \in \mathfrak{S}, nu(\mathcal{H}) < \infty,\ F_{\mathcal{H}} \text{ not globally regular}\},$$

where

$$\nu(\mathcal{H}) = \begin{cases} \operatorname{lip}_X(\mathcal{H}) \cdot \operatorname{reg}_Y(\mathcal{H}), & \text{if } \operatorname{reg}_Y(\mathcal{H}), textrmlip_X(\mathcal{H}) < \infty, \\ +\infty, & \text{otherwise,} \end{cases}$$

is called *radius of global regularity* of F with respect to perturbations in $\mathfrak{S}$.

The next result, again stemming from Theorem 1, provides a one-side estimation for the radius of global regularity in terms of modulus of metric regularity.

Proposition 1 Let $F : X \longrightarrow 2^Y$ be a set-valued mapping between metric spaces, with (X, d) complete, and let $\mathfrak{S} \subseteq \mathfrak{A}(F)$. If F is globally regular, then

$$\operatorname{rad}_{\mathfrak{S}}(F) \geqslant \frac{1}{\operatorname{reg} F} = \frac{1}{\operatorname{lip}(F^{-1})}.$$

3 Local results

When considering in a local sense the property defined by inequality (1.1), as it will be done in the next definition, the stability analysis can not be conducted simply by localizing results achieved for global metric regularity. For instance, a counterexample is known, which makes vain any effort to extend Milyutin theorem to the case of multifunctions additively perturbed by set-valued locally Lipschitz mappings. In what fallows, only single-valued mappings will be considered.

Definition 3 A mapping $F : X \longrightarrow Y$ between metric spaces is said to be *metrically regular* near $\bar{x} \in X$ if there exist positive constants κ, δ and ζ such that

$$\operatorname{dist}(x, F^{-1}(y)) \leqslant \kappa \operatorname{dist}(y, F(x)), \quad \forall x \in B(\bar{x}, \delta),\ \forall y \in B(F(\bar{x}), \zeta).$$

The infimum of κ over all such combinations of κ, δ and ζ satisfying the above inequality is called *local modulus of metric regularity* of F near $\bar{x}$ and denoted by $\operatorname{reg} F(\bar{x})$.

As one expexts, such property can be characterized in terms of local open covering of F. What changes with respect to the global case is the corresponding Lipschitz behaviour of F^{-1}, that turns out to be essentially weakened. It requires the introduction of a generalization of Lipschitz continuity for set-valued mappings, known as Lipschitz-likness (or *Aubin continuity*), which plays a major role in several topics of variational analysis (see [1, 2]). In fact a mapping $F : X \longrightarrow Y$ is metrically regular near $\bar{x}$ iff F^{-1} is *Lipschitz-like* near $(F(\bar{x}), \bar{x})$, namely there exist positive constants δ, ζ and l such that

$$F^{-1}(y_2) \cap B(\bar{x}, \delta) \subseteq B(F^{-1}(y_1), ld(y_1, y_2)), \quad \forall y_1,\ y_2 \in B(F(\bar{x}), \zeta).$$

Given a mapping $F : X \longrightarrow Y$ and a point $\bar{x}$, the set of all perturbations $\mathcal{H} : X \times Y \longrightarrow Y$ such that the resulting perturbed mapping $F_{\mathcal{H}} : X \longrightarrow Y$, defined as in (2.1), is continuous in a neighbourhood of $\bar{x}$ is denoted by $\mathfrak{A}(F, \bar{x})$. Then a local counterpart of Theorem 1 can be stated as follows:

Theorem 2 Let $F : X \longrightarrow Y$ be a mapping between metric spaces, and let $\bar{x} \in X$ and $\bar{y} = F(\bar{x})$. Given any $\mathcal{H} \in \mathfrak{A}(F, \bar{x})$, suppose that:

i) (X, d) is metrically complete;

ii) F is continuous at $\bar{x}$ and it is metricaly regular near $\bar{x}$;

iii) $\mathcal{H}(\cdot, y)$ is locally Lipschitz near $\bar{x}$ with modulus $\operatorname{lip} \mathcal{H}(\bar{x})$, uniformly in $y \in B(\bar{y}, \zeta)$, for some $\zeta > 0$;

iv) $\mathcal{H}(x, \cdot)$ is globally metrically regular with modulus $\operatorname{reg}_Y \mathcal{H}$, uniformly in $x \in B(\bar{x}, \delta)$, for some $\delta > 0$.

Then, if $\operatorname{reg} F(\bar{x}) \cdot \operatorname{reg}_Y \mathcal{H} < 1/\operatorname{lip}_X \mathcal{H}(\bar{x})$, also $F_{\mathcal{H}}$ is metrically regular near $\bar{x}$ and

$$\operatorname{reg} F_{\mathcal{H}}(\bar{x}) \leqslant \frac{\operatorname{reg} F(\bar{x}) \cdot \operatorname{reg}_Y \mathcal{H}}{1 - \operatorname{reg} F(\bar{x}) \cdot \operatorname{reg}_Y \mathcal{H} \cdot \operatorname{lip}_X \mathcal{H}(\bar{x})}.$$

Let us mention some interesting applications of Theorem 2.

Corollary 3 Let $F : X \longrightarrow Y$ be a mapping between metric spaces, let $h : X \longrightarrow (0, +\infty)$ be a positive functional, and let $\bar{x} \in X$. Suppose that:

i) (X, d) is complete and (Y, d) is a linear metric space, whose metric is invariant with respect to translations and is positively homogeneous of degree one;

ii) F is metrically regular near $\bar{x}$ and it is continuous around $\bar{x}$;

iii) h is locally Lipschitz near $\bar{x}$, with modulus $\operatorname{lip} h(\bar{x})$, and there exists $\delta_h > 0$ such that

$$\inf_{x \in B(\bar{x}, \delta_h)} h(x) > \operatorname{lip} h(\bar{x}) \cdot \operatorname{reg} F(\bar{x}) d(F(\bar{x}), \mathbf{0}),$$

where $\mathbf{0}$ denotes the null element of Y. Then, the product mapping hF is metrically regular near $\bar{x}$ with modulus

$$\operatorname{reg}(hF)(\bar{x}) \leqslant \frac{\operatorname{reg} F(\bar{x})}{\inf_{x \in B(\bar{x}, \delta_h)} h(x) - \operatorname{reg} F(\bar{x}) \operatorname{lip} h(\bar{x}) d(F(\bar{x}), \mathbf{0})}.$$

As a further application of Theorem 2, one can easily obtain the following Lyusternik-Graves type theorem, which is valid in linear metric spaces:

Corollary 4 Let $F : X \longrightarrow Y$ be a mapping between metric spaces. Suppose that:

i) (X, d) is complete and (Y, d) is a linear metric space, whose metric is invariant with respect to translations;

ii) F is continuous near $\bar{x}$;

iii) $G : X \longrightarrow Y$ is metrically regular near $\bar{x}$ and such that

$$\operatorname{lip}(F - G)(\bar{x}) < \epsilon, \quad \text{with} \quad \epsilon \in (0, 1/\operatorname{reg} G(\bar{x})).$$

Then F is metrically regular near $\bar{x}$.

From Corollary 4 several nonsmooth versions of the celebrated Lyusternik theorem[1] follow as a special case. They provide full characterizations of metric regularity of mappings acting in Banach spaces in terms of covering properties of certain first-order approximations (generalized derivatives) (see [2]). The concept of radius of regularity can be adapted to the local case as follows:

Definition 4 Given a mapping $F : X \longrightarrow Y$ and $\bar{x} \in X$, let $\mathfrak{S} \subseteq \mathfrak{A}(F, \bar{x})$. The quantity $\operatorname{rad}_{\mathfrak{S}} F(\bar{x}) = \inf\{\nu(\mathcal{H}, \bar{x}) : \mathcal{H} \in \mathfrak{S}, \nu(\mathcal{H}, \bar{x}) < \infty,\ F_{\mathcal{H}} \text{ not regular near } \bar{x}\}$, where

$$\nu(\mathcal{H}, \bar{x}) = \begin{cases} \operatorname{lip}_X \mathcal{H}(\bar{x}) \cdot \operatorname{reg}_Y \mathcal{H}, & \text{if } \operatorname{reg}_Y \mathcal{H}, \operatorname{lip}_X \mathcal{H}(\bar{x}) < \infty, \\ +\infty, & \text{otherwise,} \end{cases}$$

is called *radius if regularity* of F near $\bar{x}$ with respect to perturbations in $\mathfrak{S}$.

As a local counterpart of the estimation expressed by Proposition 1, one has:

Proposition 2 Let $F : X \longrightarrow Y$ be a mapping between metric spaces, with (X, d) complete, let $\bar{x} \in X$, and let $\mathfrak{S} \subseteq \mathfrak{A}(F, \bar{x})$. If F is regular near $\bar{x}$ and it is continuous at $\bar{x}$, then

$$\operatorname{rad}_{\mathfrak{S}} F(\bar{x}) \geqslant \frac{1}{\operatorname{reg} F(\bar{x})}. \tag{3.1}$$

Remark 1 Note that, whenever F is a mapping between Banach spaces and $\mathfrak{S}$ is the class of all additive perturbations defined by locally Lipschitz (single-valued) mappings, then $\operatorname{rad}_{\mathfrak{S}} F(\bar{x})$ reduces to the radius of regularity as introduced in [7]. If, in particular, both the spaces are finite-dimensional, it has been proved that

[1] a nonlinear local generalization of the Banach-Schauder theorem to smooth mappings, which allows to get an algebraic characterization of the tangent space to smooth manifolds.

estimation (3.1) becomes an equality, whereas this is not generally true for infinite-dimensional spaces. Nevertheless, it is worth mentioning that results restoring the validity of the exact estimation for $\mathrm{rad}_{\mathfrak{S}} F(\bar{x})$ has been achieved for set-valued mappings acting from Asplund[1] to finite-dimensional spaces under the so-called coderivative normality assumption (see, for more details, [2]).

References

1. A.D. Ioffe *Metric regularity and subdifferential calculus*, Russian Math. Surveys **55**, 3, 2000, Pp. 501-558.
2. B. S. Mordukhovich *Variational Analysis and Generalized Differentiation, I: Basic Theory*, Springer, 2006. 579 pages.
3. E.R. Avakhov et al. *Covering mappings and their applications to differential equations not solved with respect to the derivative*, Differ. Equ. **45**, 5, 2009, Pp. 627-649.
4. A.V. Arutyunov *Covering mappings in metric spaces, and fixed points*, Dokl. Akad. Nauk **416**, 2, 2007, Pp. 151–155.
5. A. Uderzo *A metric version of Milyutin Theorem*, forthcoming on Set-Valued Var. Anal., DOI 10.1007/s11228-011-0193-9.
6. A. V. Dmitruk et al. *Lyusternik's theorem and the theory of extrema*, Russian Math. Surveys **35**, 6, 1980, Pp. 11-51.
7. A.L. Dontchev et al. *The radius of metric regularity*, Trans. Amer. math. Soc. **355**, 2, 2003, Pp. 493-517.

A. Uderzo
University of Milano-Bicocca, Italy, 20126, Milano, Via Bicocca degli Arcimboldi 8,
Tel. +39 02 6448 5871, e-mail: amos.uderzo@unimib.it

[1] i.e. Banach spaces where every convex continuous function is Fréchet differentiable in a G_δ dense subset of its domain.

ATTRACTORS OF CONFORMAL FOLIATIONS

N. I. Zhukova

Key words: attractor, global attractor, conformal foliation, holohomy group, Riemannian foliation

AMS Mathematics Subject Classification: 57R30, 34B41, 53A30

Abstract. We investigated conformal foliations (M, F) of codimension $q \geqslant 3$ and proved a criterion for them to be Riemannian. In particular, the application of this criterion allowed us to proof the existence of an attractor that is a minimal set for each non-Riemannian conformal foliation. Moreover, if foliated manifold is compact then non-Riemannian conformal foliation (M, F) is $(Conf(S^q), S^q)$-foliation with finitely many minimal sets. They are all attractors, and each leaf of the foliation belongs to the basin of at least one of them. The specificity of the proper conformal foliations is indicated. Special attention is given to complete conformal foliations.

1 Introduction

The goal of this work is to present recent results of the author [1, 2] on the investigation of the structure of conformal foliations.

1.1 The Lichnerowicz Conjecture

Remind that two Riemannian metrics h and g on a manifold M are called conformally equivalent if there exists a positive smooth function f on M with $h = fg$. Each class $[g]$ of conformally equivalent Riemannian metrics is named a conformal structure on M, and the pair $(M, [g])$ is said to be a conformal manifold.

Recall that a diffeomorphism $f : N_1 \to N_2$ between Riemannian manifolds (N_1, g_1) and (N_2, g_2) is named conformal if there exists a smooth function λ on N_1 with $f^* g_2 = \lambda g_1$. A conformal diffeomorphism f from a Riemannian manifold (N, g) to (N, g) is also said to be a conformal transformation.

The group of conformal transformations of a Riemannian manifold (M, g) is called *inessential* if it is a group of isometries of a Riemannian manifold (M, h) with $h \in [g]$. Otherwise, the group is called *essential*.

The work was partially supported by the Russian Foundation of Basic Research (grant 10-01-00457).

Lichnerowicz put forth the conjecture that for $n \geqslant 3$ every n-dimensional compact Riemannian manifold admitting an essential group of conformal transformations is the standard n-dimensional sphere S^n.

The articles by Obata [3], Alekseevskii [4,5], Ferrand [6] and others are devoted to proving this conjecture.

It was established also that if the group of conformal transformations of a noncompact Riemannian manifold M is essential then M is the n-dimensional Euclidean space. In 1996 Ferrand [6] gave a complete proof of the Lichnerowicz conjecture including the case of noncompact manifolds.

1.2 Conformal Foliations

Vaisman [7] introduced the conformal foliations (M, F) as foliations admitting a transversal conformal structure.

Suppose that given are:

1) n-dimensional manifold M and a possibly disconnected q-dimensional manifold N, where $0 < q < n$;

2) an open cover $\{U_i \,|\, i \in J\}$ of M;

3) submersions $f_i : U_i \to V_i$ with connected fibres, $V_i \subset N$,

and if $U_i \cap U_j \neq \emptyset$, then there exists conformal diffeomorphism $\gamma_{ij} : f_j(U_i \cap U_j) \to f_i(U_i \cap U_j)$ such that $f_i = \gamma_{ij} \circ f_j$ on $U_i \cap U_j$.

Maximal, with respect to inclusion, N-cocycle $\{U_i, f_i, \gamma_{ij}\}_{i,j \in J}$ enjoying these properties determines a new topology on M, whose base is the set of leaves of all submersions f_i. This topology is called the leaf topology, denoted by τ.

The path-connected components of the topological space (M, τ) form a partition $F = \{L_\alpha \,|\, \alpha \in A\}$ of M, and (M, F) is named the foliation with leaves L_α determined by the N-cocycle $\{U_i, f_i, \gamma_{ij}\}_{i,j \in J}$.

Definition 1. A codimension $q \geqslant 3$ smooth foliation (M, F) is called conformal if (M, F) is determined by an N-cocycle $\{U_i, f_i, \gamma_{ij}\}_{i,j} \in J$, and N admits a Riemannian metric g such that each γ_{ij} is the local conformal diffeomorphism of the corresponding open subsets.

If each γ_{ij} is an isometry, then (M, F) is named a *Riemannian foliation.*

Let $Conf(S^q)$ be the Lie group of all the conformal transformations of the q-dimensional sphere S^q.

Definition 2. If a foliation (M, F) is defined by N-cocycle $\{U_i, f_i, \gamma_{ij}\}_{i,j \in J}$, where

— $N = S^q$ and

— each γ_{ij} is a restriction of a transformation $f \in Conf(S^q)$,
then refer to (M, F) as *a* $(Conf(S^q), S^q)$*-foliation.*

1.3 The Tarquini – Frances Question

Tarquini [8] and then Frances and Tarquini [9] posed the following question about conformal foliations:

Is every codimension $q \geqslant 3$ *conformal foliation on a compact manifold either a Riemannian foliation or a* $(Conf(S^q), S^q)$*-foliation?*

They refer to the positive answer to this question as the foliated analogous of the Lichnerowicz Conjecture.

As known, for $q \geqslant 3$ a conformal foliation is a $(Conf(S^q), S^q)$-foliation if and only if it is transversally conformally flat.

Frances and Tarquini [9] gave a positive answer to this question under some additional assumptions.

1.4 Attractors and Minimal Sets of Foliations

Let (M, F) be a foliation. A *saturated set* is a union of leaves.

Definition 3. A nonempty closed saturated subset $\mathcal{M}$ of M is said to be an attractor of a foliation (M, F) if there exists a saturated open neighbourhood $Attr(\mathcal{M})$ of $\mathcal{M}$ such that the closure of every leaf from $Attr(\mathcal{M})$ includes $\mathcal{M}$. The set $Attr(\mathcal{M})$ is named a basin of this attractor. If an addition $Attr(\mathcal{M}) = M$ then $\mathcal{M}$ is called a global attractor.

Definition 4. A minimal set of a foliation (M, F) is a nonempty closed saturated subset of M without proper subsets enjoying these properties.

In mathematical encyclopedia Anosov [10] said that the study of minimal sets is one of the fundamental problems in the topological dynamics, hence also in the qualitative theory of foliations.

1.5 Results

We consider conformal foliations as Cartan foliations and use the construction of principal foliated bundle. Thanks this we apply the results of our previous works [11, 12]. In the foliation theory the germ holonomy groups are usually used. By analogy to [12] we gave different interpretations for the germ holonomy group of a leaf of a conformal foliation (Theorem 1).

We proved a criterion for conformal foliation to be Riemannian (Theorem 2). Application of this criterion and some results on local conformal geometry allowed us to prove the existence of an attractor for every non-Riemannian conformal foliation of codimension $q > 2$ (Theorem 4).

As known ([13, 14]), there are smooth foliations on non-compact manifolds without minimal sets. We proved that any non-Riemannian conformal foliation of codimension $q > 2$ on noncompact manifold has the minimal set that is an attractor of this foliation (Corollary 1).

We also described the structure of conformal foliations (M, F) in the cases when:
1) the foliated manifold M is compact (Theorem 5);
2) these foliations are complete (Theorem 6);
3) foliations (M, F) are proper (Theorem 7).

1.6 The Positive Answer to the Frances — Tarquini Question and a Proof of the Conjecture of Ghys

— Theorem 5 implies the positive answer to the Frances — Tarquini question about conformal foliations on compact manifolds.

— Deroin and Kleptsyn [15] indicated that all known examples of a transversely conformal foliation having a diffuse transversely invariant measure have a transverse metric which is transversely invariant. According [15], for transversely conformal foliations this has been conjectured by Ghys. A proof of this conjecture for conformal foliations of codimension $q \geqslant 3$ (in the conforming class of the smoothness) follows from Theorem 5.

Notations

— Let $G = Conf(S^q)$ be the Lie group of all conformal transformations of the q-dimensional sphere S^q and H be the stabilizer in G of an arbitrary point of S^q. Then H is a semidirect product of a conformal group $CO(q) = R^+ \cdot O(q)$ and the group R^q.

— The Lie group H is isomorphic to the the Lie group $Sim(E^q)$ of all conformal transformations of the q-dimensional Euclidean space E^q.

— We assume here that $q \geqslant 3$.

2 A criterion for a conformal foliation to be Riemannian

2.1 The Foliated Bundle of a Conformal Foliation

The construction of a foliated bundle was essentially used. Foliated bundles were introduced in works of Molino and Kamper – Tondeur.

Remark that any conformal foliation may be considered as Cartan foliation of the type (G, H), where $G = Conf(S^q)$ is the Lie group of all conformal transformations of S^q and H be the stabilizer in G of an arbitrary point of S^q.

Let $\mathfrak{g}$, $\mathfrak{h}$ are the Lie algebras of the Lie groups G and H, respectively. Then we have the following statement [11].

Proposition 1. Let (M, F) be a conformal foliation of codimension $q \geqslant 3$. Then the following objects are defined:

1) a principal H-bundle $\pi : \mathcal{R} \to M$;

2) an H-invariant foliation $(\mathcal{R}, \mathcal{F})$ which π transforms into (M, F);

3) a $\mathfrak{g}$-valued 1-form ω on $\mathcal{R}$ with the properties:

(i) $\omega(A^*) = A$ for any $A \in \mathfrak{h}$, where A^* is the fundamental vector field corresponding to A;

(ii) $R_a^*\omega = Ad_G(a^{-1})\omega$ for any $a \in H$, where Ad_G is the adjoint representation of the Lie group G in its Lie algebra $\mathfrak{g}$;

(iii) the Lie derivative $L_X\omega$ vanishes for every vector field X tangent to the leaves of $(\mathcal{R}, \mathcal{F})$.

The H-bundle $\pi : \mathcal{R} \to M$ is said to be *foliated*. The foliation $(\mathcal{R}, \mathcal{F})$ is named a *lifted* foliation, and $(\mathcal{R}, \mathcal{F})$ is a transversally parallelizable foliation, i.e., an e-foliation.

Definition 5. A conformal foliation (M, F) of codimension $q \geqslant 3$ is said to be complete, if so is the associated lifted e-foliation $(\mathcal{R}, \mathcal{F})$, i.e., if complete is every vector field X on $\mathcal{R}$ such that $\omega(X) = constant$.

2.2 Interpretations of the Holonomy Groups of a Conformal Foliation

An application of the foliated H-bundle over a conformal foliation (M, F) allowed us gave the following important interpretations of the holonomy groups of leaves.

Theorem 1. *Let (M, F) be an arbitrary conformal foliation of codimension $q \geqslant 3$ and $\pi : \mathcal{R} \to M$ be the projection of the foliated H-bundle over (M, F) with lifted foliation $(\mathcal{R}, \mathcal{F})$. For each leaf $L = L(x)$ of (M, F) consider the leaf $\mathcal{L} = \mathcal{L}(u)$, where $u \in \mathcal{R}$, $\pi(u) = x$, of the lifted foliation $(\mathcal{R}, \mathcal{F})$. Then the germ holonomy group $\Gamma(L, x)$ of L is isomorphic to the following groups:*

- *the subgroup $H(\mathcal{L}) := \{a \in H \mid R_a(\mathcal{L}) = \mathcal{L}\}$ of H;*
- *the group of deck transformations of the regular covering $\pi|_{\mathcal{L}} : \mathcal{L} \to L$.*

If in conditions of the Theorem 1 we consider an other point $u' \in \pi^{-1}(x)$ and the leaf $\mathcal{L}' = \mathcal{L}'(u')$, then the group $H(\mathcal{L}')$ must be conjugated to $H(\mathcal{L})$ in H. Therefore, the following definition makes sense.

Definition 6. Refer to the holonomy group of a leaf L of a conformal foliation as *relatively compact* or *inessential* if the corresponding subgroup $H(\mathcal{L})$ of the Lie group H is relatively compact. Otherwise the holonomy group of a leaf is called *essential*.

2.3 When a Conformal Foliation is a Riemannian one?

Thanks to well-known works of Reinhart, Molino, Haefliger, Salem, Carriere and other authors, now Riemannian foliations form the most investigated class of foliations with transverse geometric structure. Hence it is very significant to know when a smooth foliation is Riemannian.

We established the following criterion for a conformal foliation to be Riemannian.

Theorem 2. *If (M, F) is a codimension $q \geqslant 3$ conformal foliation modeled on a conformal geometry $(N, [g])$, then there exists of a Riemannian metric $d \in [g]$ such that (M, F) is a Riemannian foliation modeled on (N, d) if and only if every holonomy group of this foliation be relatively compact.*

Corollary 1. If a conformal foliation (M, F) is not Riemannian then it has a leaf with essential holonomy group.

3 The Existence of Attractors of Conformal Foliations

3.1 A closure of a Leaf with Essential Holonomy Group

The following two theorems were proved without the assumptions of compactness of foliated manifolds and completeness of conformal foliations. In proof of Theorem 3 we considered and applied a conformal geometry on non-Hausdorff manifolds.

Theorem 3. *If (M, F) is a non-Riemannian conformal foliation of codimension $q \geqslant 3$, then:*

(i) for each leaf L with essential holonomy group the closure $\bar{L} = \mathcal{M}$ is an attractor, while

- *either $\mathcal{M}$ is a minimal set*
- *or $\mathcal{M}$ includes a closed leaf that is also an attractor;*

(ii) the union K of the closures of all leaves with essential holonomy group is a closed saturated subset of M, and $(M \backslash K, F_{M\backslash K})$ is a Riemannian foliation.

According to Corollary 1 any conformal non-Riemannian foliation has a leaf L with essential holonomy group. Therefore by Theorem 3 the closure $\mathcal{M} := \bar{L}$ is an attractor of conformal foliation (M, F). Thus we have the following assertion.

Theorem 4. *Each codimension $q \geqslant 3$ conformal foliation (M, F)*

- *either is Riemannian*
- *or has an attractor $\mathcal{M}$ that is the closure $\mathcal{M} = \bar{L}$ of a leaf L with essential holonomy group, and the restriction of the foliation to the attraction basin $(\mathcal{A}ttr(\mathcal{M}), F)$ is a $(Conf(S^q), S^q)$-foliation.*

Theorems 3 and 4 imply the following statement.

Corollary 2. Each codimension $q \geqslant 3$ non-Riemannian conformal foliation has a minimal set that is an attractor of this foliation.

3.2 Conformal Foliations on Compact Manifolds

The notion of Ehresmann connection was introduced by Blumenthal and Hebda [16]. It belongs to differential topology. Using an Ehresmann connection we constructed a "trap for the leaves" in the proof of the following assertion for conformal foliations.

Proposition 2. Let $\mathcal{M}$ be a compact minimal set of a conformal foliation (M, F), and all leaves from $\mathcal{M}$ have inessential holonomy group. Then each open neighbourhood $\mathcal{W}$ of $\mathcal{M}$ includes a saturated open neighbourhood V consisting of leaves with inessential holonomy group.

This proposition was essentially used in the proof of Theorem 5.

Theorem 5. *Every codimension $q \geqslant 3$ conformal foliation (M, F) on a compact manifold M is*

- *either a complete Riemannian foliation, and the closure of every leaf is a minimal set which is an embedded submanifold of M,*
- *or a $(Conf(S^q), S^q)$-foliation with finitely many minimal sets. They are all attractors formed by the closures of leaves with essential holonomy group, and each leaf of the foliation belongs to the basin of at least one of them.*

4 Global Attractors of Complete Conformal Foliations

Denote by $Sim(E^q)$ the Lie group of all conformal transformations of the Euclidean space E^q. The group $Sim(E^q)$ is equal to a semidirect product of the conformal Lie group $CO(q)$ and the normal subgroup R^q.

Definition 7. A foliation (M, F) defined by N-cocycle $\{U_i, f_i, \gamma_{ij}\}_{i,j\in J}$ is said to be transversally similar or a $(Sim(E^q), E^q)$-foliation if $N = E^q$ and each γ_{ij} is a restriction of a similar transformation from the group $Sim(E^q)$.

Theorem 6. *Let (M, F) be a complete conformal foliation of codimension $q \geqslant 3$. Then one of the following three possibilities is realized:*

1) the foliation (M, F) is Riemannian, and the closure of each its leaf forms a minimal set that is an embedded submanifold of M;

2) (M, F) is a transversally similar foliation. It has a global attractor $\mathcal{M}$ that is a minimal set containing all leaves with essential holonomy group;

3) (M, F) is a $(Conf(S^q), S^q)$-foliation with a global attractor $\mathcal{M}$, and $\mathcal{M}$ is either one or two leaves of this foliation or else $\mathcal{M}$ is nontrivial minimal set coincided with the closure of every leaf having essential holonomy group.

Moreover, in the cases 2) and 3) the restriction (M_0, F_0) of F onto $M_0 := M \backslash \mathcal{M}$ is a Riemannian foliation, and the closure of any leaf $L \subset M_0$ in M is equal to the union $\mathcal{M} \cup \mathcal{L}$ of $\mathcal{M}$ and a closed submanifold $\mathcal{L}$ coincided with the closure of L in M_0.

Corollary 3. If a complete conformal non-Riemannian foliation has a minimal set $\mathcal{M}$ different from a closed leaf, then $\mathcal{M}$ is a global attractor of this foliation.

Remark 1. Minimal sets of complete transversally similar foliations were investigated by the author in [11]. In particular, there we found conditions guaranteed for the global attractor of a complete transversally similar foliation (M, F) to be a smooth submanifold of M.

5 Specificity of Proper Conformal Foliations

Definition 8. A foliation (M, F) is called proper if all its leaves are embedded submanifolds of M. A leaf L is called closed if L is a closed subset of M.

Emphasize that every minimal set of a proper foliation coincides with a closed leaf. Hence as application of Theorem 3 we have the following assertion.

Corollary 4. Each proper of codimension $q \geqslant 3$ non-Riemannian conformal foliation has a closed leaf with essential holonomy group that is an attractor of this foliation.

Theorem 7. *Any complete proper conformal foliation (M, F) of codimension $q \geqslant 3$ has a structure of one of the following types:*

- *(M, F) is a transversally complete proper Riemannian foliation with closed leaves and its leaf space is a smooth q-dimensional orbifold;*
- *(M, F) is a complete proper non-Riemannian transversally similar foliation. There exists a unique closed leaf L_0, and L_0 is a global attractor and has an essential holonomy group;*

— *the foliation (M, F) is not transversally similar. There exists a global attractor $\mathcal{M}$ that coincides with one or two leaves of (M, F). The restriction (M_0, F_0) of F onto the dense open subset $M_0 := M \backslash \mathcal{M}$ is a Riemannian foliation, and the leaf space M_0/F_0 admits a structure of q-dimensional smooth orbifold. The closure of any leaf $L \subset M_0$ is equal to the union $\mathcal{M} \cup L$.*

Remark 2. As Weil foliations form a subclass of conformal foliations, so in the case of codimension $q \geqslant 3$ the main results for Weil foliations [17] follow from Theorems 6 and 7.

Examples of conformal foliations with different kinds of attractors are constructed by the method of suspension of a group homomorphism.

References

1. Zhukova N.I. *Attractors and an analog of the Lichnerowicz conjecture for conformal foliations.* Siberian Math. J. **52**, 3, 2011, Pp. 436-450.
2. Zhukova N.I. *Global attractors of complete conformal foliations.* Math. Sb. **203**, 2012.
3. M. Obata *The conjectures on conformal transformations of Riemannian manifolds.* J. Differential Geom. **6**, 2, 1971. Pp. 247-258.
4. D.V. Alekseevskii *Groups of conformal transformations of Riemannian spaces.* Math. Sb.**89**, 2, 1972. Pp. 280-286.
5. D.V. Alekseevskii *S^n and E^n are the only Riemannian spaces that admit an essential conformal transformation.* Uspekhi Mat. Nauk. **28**, 5, 1973. Pp. 225-226.
6. J. Ferrand *The action of conformal transformations on a Riemannian manifold.* Math. Ann. **304**, 2, 1996. Pp. 277-291.
7. I. Vaisman *Conformal foliations,* Kodai Math. J. **2**, 1, 1979, Pp. 26-37.
8. C. Tarquini *Feuilletages conformes.* Ann. Inst. Fourier. **52**, 2, 2004. Pp. 453-480.
9. Frances C. and Tarquini C. *Autour du theoreme de Ferrand.Obata.* Ann. Global Anal. Geom. **21**, 1, 2007. Pp. 51-62.
10. D.V. Anosov *Minima set* in Mathematical Encyclopedia. **3**, 1982. Pp. 690-691.
11. Zhukova N. *Minimal sets of Cartan foliations.* Proc. Steklov Inst. of Math. **256**, 2007. Pp. 105-135.
12. Zhukova N.I. *Complete foliations with transversal rigid geometries and their basis automorphisms.* Vestnik RUDN Ser. Mat. Fiz.Inform. **2**, 2009, Pp. 14-35.
13. Beniere J.-C., Meigniez G. *Flows without minimal set.* Erg. Theory and Dyn. Syst. **19**, 1, 1999. Pp. 21-30.

14. Inaba T. *An example of a flow on a non-compact surface without minimal set.* Erg. Theory and Dyn. Syst. **19**, 1, 1999. Pp. 31-33.
15. B.Deroin and V. Kleptsyn *Random conformal dynamical systems.* Geom. Funct. Anal.**17**, 4, 2007. Pp. 1043-1105.
16. R. A. Blumenthal and J. J. Hebda *Ehresmann connections for foliations.* Indiana Univ. Math. J. **33**, 4, 1984. Pp. 597-611.
17. Zhukova N.I. *Weil foliations.* Nonlinear Dynam. **6**, 1, 2010. Pp. 219-231.

N. I. Zhukova
Nizhny Novgorod State University, Russia, 603095, Gagarin ave., 23, Department of Mechanics and Mathematics, n.i.zhukova@rambler.ru

THE COMPARISON OF THE TWO CRITERIA OF COMPLETE OBSERVABILITY

S. P. Zubova, E. V. Raetskaya

Key words: Differential-algebraic system of observation, complete observability, the method of cascade decomposition

AMS Mathematics Subject Classification: 519.710

Abstract. Article is devoted to comparing the two criteria for complete observability of linear differential-algebraic system. The method of cascade splitting the original space to the subspace are used. The formula for finding the state vector are derived. The relation between input and output functions are obtained. The article also involves the specify the conditions of the well-known criteria of observability.

1 Introduction

Here is considered the differential-algebraic system of observation:

$$\frac{dx(t)}{dt} = Ax(t) + f(t), \tag{1.1}$$

$$F(t) = Bx(t), \tag{1.2}$$

where $B : R^n \to R^m$, $A : R^n \to R^n$, $x(t) \in R^n$, $f(t) \in R^n$, $F(t) \in R^m$.

The vector-function $x(t)$ called **the state of the system**, $f(t)$ **input and** $F(t)$ **output functions**, respectively.

System (1.1), (1.2) is called **completely observable** if the system state (by known input $f(t)$ and output $F(t)$ functions) at any time is uniquely determined.

Formulation of problem of the complete observability problem for the dynamical system are connected with the name of R. Kalman.

The complete observability of the different systems were studied in the works of Krasovskii N.N, Popov V.M., Lee E.B. and Markus L.M., D'Anzhelo, Andreev J.N., Gurman V.I., Kvanernaak X. and Sivan R., Asmykovich I.K. and MarchenkoV.M., Campbell S.L., Cobb J.D., Koumboulis F.N. и Mertzios B.G., Yip E.L. и Sincovec R.F., Paraskevopoulos P.N., Sheglova A.A., etc.

For linear time-invariant systems as a rule consider the case of a regular pencil $(B - \lambda I)$.

At the moment are known a lot of criteria complete observability for the system (1.1), (1.2).

One of them [1]:

Criterion (A).

The system (1.1), (1.2) *is completely observable if and only if the system:*

$$BA^i z = 0, \quad i = \overline{0, n} \tag{1.3}$$

has only the trivial solution z.

It's known ([1], [2]) that the system (3) contains excessive amounts of relations. However, the exact number (l) hasn't been determined.

Here will be formulated new criterion (B) of the complete observability of the system;

two criteria ((A) and (B)) will be compared;

the exact number of relations in the system (3) of the well-known criterion (A) will be set.

2 Proof of the criterion (B)

In the study of complete observability of system (1.1), (1.2) we use the method of cascade decomposition of the original space. This method was used by the authors to investigate the full observability and complete controllability of dynamical systems ([3] – [5]). From initial system (1.1), (1.2) we turn to equivalent systems in the subspaces.

Next decompositions are correspond to the matrix B:

$$R^n = KerB \dot{+} ImB^*, \quad R^m = KerB^* \dot{+} ImB, \tag{2.1}$$

where:

ImB - is a set of values B in R^m,

$KerB$ - is a set of solutions of eq. $Bx(t) = 0$ in ImR^n,

ImB^* - is a direct complement to the subspace $KerB$,

$KerB^*$ - is a direct complement to the subspace ImB.

Through $P(B)$ and $Q(B)$ denote the projections on the subspaces $KerB$ and $KerB^*$, respectively.

Through $(I - P(B))$ and $(I - Q(B))$ denote the projections on the subspaces ImB^* and ImB, respectively. I - identity matrix in the appropriate space.

The restriction $\tilde{B}$ to the subspace ImB^* carries a one-one correspondence between subspaces ImB^* and ImB, respectively.

We introduce B^- - the semi-inverse matrix: $B^- = \tilde{B}^-(I - Q(B))$.

Equation (1.2) of the original system is equivalent to the system:

$$Q(B)F(t) = 0, \tag{2.2}$$

$$x(t) = B^{-}F(t) + x_1(t), \tag{2.3}$$

with an arbitrary vector function $x_1(t) = P(B)x(t) \in KerB$.

Here are three possible cases:

1) $B = 0$. Equation (1.2) has the form: $F(t) = 0$.

The state-function $x(t)$ (here it's the solution of the differential equation (1.1)) is not uniquely.

System (1.1), (1.2) is unobservable.

2) The matrix B - is injective ($P(B) = 0$).

The state-function $x(t)$ is uniquely defined by the formula:

$$x(t) = B^{-}F(t). \tag{2.4}$$

The system (1.1), (1.2) is completely observable.

Taking into account the formula (2.4), the equation (1.1) takes the form:

$$\frac{B^{-}F(t)}{dt} = AB^{-}F(t) + f(t). \tag{2.5}$$

This is the ratio of "input-output" – the $f(t)$ input and $F(t)$ output functions should satisfy of this connection.

3) $B \neq 0$ and $KerB \neq \{0\}$.

We substitute (2.3) in (1.1) and "split" it into 2 equations.

The first equation in the subspace $KerB$:

$$\frac{dx_1(t)}{dt} = P(B)Ax_1(t) + P(B)(AB^{-}F(t) + f(t)). \tag{2.6}$$

The second equation in the subspace ImB^{*}:

$$\frac{B^{-}F(t)}{dt} - (I - P(B))(AB^{-}F(t) + f(t)) = (I - P(B))Ax_1(t). \tag{2.7}$$

Denote:

$$A_1 = P(B)AP(B) : KerB \to KerB,$$

$$B_1 = (I - P(B))AP(B) : KerB \to ImB^{*},$$

$$f_1(t) = P(B)(AB^{-}F(t) + f(t)) \in KerB,$$

$$F_1(t) = \frac{dB^{-}F(t)}{dt} - (I - P(B))(AB^{-}F(t) + f(t)) \in ImB^{*}.$$

The system (2.6), (2.7) now takes the form:

$$\frac{dx_1(t)}{dt} = A_1x_1(t) + f_1(t), \tag{2.8}$$

$$F_1(t) = B_1x_1(t). \tag{2.9}$$

System (1.1), (1.2) is equivalent to (2.2), (2.3), (2.8) (2.9).

Next decompositions are correspond to the matrix $B1 : KerB \to ImB^{*}$:

$$KerB = KerB_1 \dot{+} ImB^{*}, \quad ImB^{*} = KerB_1^{*} \dot{+} ImB,$$

where:

ImB_2 - is a set of values B_1 in ImB^{*},

$KerB_1$ - is a set of solutions of eq. $B_1x_1(t) = 0$,

ImB_1^{*} - is a direct complement to the subspace $KerB_1$,

$KerB_1^{*}$ - is a direct complement to the subspace ImB_1.

Through $P(B_1)$ and $Q(B_1)$ denote the projections on the subspaces $KerB_1$ and $KerB_1^{*}$, respectively.

Through $(I - P(B_1))$ and $(I - Q(B_1))$ denote the projections on the subspaces ImB_1^{*} and ImB_1, respectively.

I - identity matrix in the appropriate space.

The restriction $\tilde{B}_1$ to the subspace ImB_1^{*} carries a one-one correspondence between subspaces ImB_1^{*} and ImB_1, respectively.

We introduce B_1^{-}- the semi-inverse matrix: $B_1^{-} = \tilde{B}_1^{-}(I - Q(B_1))$.

Equation (2.9) is equivalent to the system:

$$Q(B_1)F_1(t) = 0, \tag{2.10}$$

$$x_1(t) = B_1^{-}F(t) + x_2(t), \tag{2.11}$$

with an arbitrary vector function $x_2(t) = P(B_1)x_1(t) \in KerB_1$.

Here are three possible cases:

1) $B1 = 0$. Equation (2.9) has the form: $F_1(t) = 0$.

The state-function $x_1(t)$ (here it's the solution of the differential equation (2.8) is not uniquely. System (2.8), (2.9) is unobservable.

The state-function $x(t)$ of the system (1.1), (1.2) is not uniquely too.

The system (1.1), (1.2) is also unobservable.

2) The matrix B_1 - injective ($P(B_1) = 0$).

Function of the state $x_1(t)$ is uniquely defined by the formula:

$$x_1(t) = B_1^- F(t). \tag{2.12}$$

The system (2.8) ,(2.9) is completely observable.

The state-function $x(t)$ of the system (1.1), (1.2) is uniquely defined by the formula:

$$x(t) = B^- F(t) + B_1^- F_1(t). \tag{2.13}$$

The system (1.1), (1.2) is completely observable too.

Taking into account the formula (2.13), the equation (1.1) takes the form of ratio "input-output" between the input and output functions of the system (1.1), (1.2).

3) $B_1 \neq 0$ and $KerB_1 \neq \{0\}$.

Continue the process of splitting cascade.

Denote:

$$\begin{aligned} &A_2 = P(B_1)A_1P(B_1) : KerB_1 \to KerB_1, \\ &B_2 = (I - P(B_1))A_1P(B_1) : KerB_1 \to ImB_1^*, \\ &f_2(t) = P(B_1)(A_1B_1^-F_1(t) + f_1(t)) \in KerB_1, \\ &F_1(t) = \tfrac{dB_1^-F_1(t)}{dt} - (I - P(B_1))(A_1B_1^-F_1(t) + f_1(t)) \in ImB_1^*. \end{aligned} \tag{2.14}$$

From the system (2.8), (2.9) goes to an equivalent set of conditions (2.10), (2.11) and the system:

$$\frac{dx_2(t)}{dt} = A_2x_2(t) + f_2(t), \tag{2.15}$$

$$F_2(t) = B_2x_2(t). \tag{2.16}$$

Here again, only three cases:

1) $B_2 = 0$;

2) $KerB_2 = \{0\}$;

3) $B_2 \neq 0$ and $KerB_2 \neq \{0\}$.

The original space has finite dimension, so the process of splitting cascade is fully realized in a finite (equal to p, $p \leqslant n$) the number of steps.

Thus, for p steps from the system (1.1), (1.2) we goes to an equivalent set of conditions:

$$x(t) = \sum_{i=0}^{i-2} B_i^- F_i(t) + x_{i-1}(t), \tag{2.17}$$

$$x_{i-1}(t) = B_{i-1}^- F_{i-1}(t) + x_i(t), \tag{2.18}$$

$$Q(B_{i-1})F_{i-1}(t) = 0, \quad i = \overline{0,p} \tag{2.19}$$

and the system:

$$\frac{x_p(t)}{dt} = A_p x_p(t) + f_p(t), \tag{2.20}$$

$$F_p(t) = B_p x_p(t). \tag{2.21}$$

Here, the functions and the matrix coefficients are given by (2.14), replacing the corresponding indices and $B_0 = B$, $B_0^- = B^-$, $F_0(t) = F(t)$.

In this latter p-th step can only be two possibilities:

1) $B_p = 0$. Equation (24) has the form: $F_p(t) = 0$.

The state-function $x_p(t)$ (here it's the solution of the differential equation (2.20) is not uniquely. System (2.20), (2.21) is unobservable.

The state-function $x(t)$ of the system (1.1), (1.2) is not uniquely too.

The system (1.1), (1.2) is also unobservable.

2) $KerB_p = \{0\}$. Equation (24) is equivalent to the system (2.18), (2.19) with $i = p + 1$ and $x_{p+1}(t) = 0$.

The state-function $x_p(t)$ is uniquely defined by the formula:

$$x_p(t) = B_p^- F_p(t). \tag{2.22}$$

The system (2.20), (2.21) is completely observable.

The state-function $x(t)$ of the system (1.1), (1.2) is uniquely defined by the formula:

$$x(t) = \sum_{i=0}^{p} B_i^- F_i(t). \tag{2.23}$$

The system (1.1), (1.2) is completely observable too.

Thus we prove

Criterion (B). *The system* (1.1), (1.2) *is completely observable if and only if there exists* p *such that* $KerB_p = \{0\}$.

Taking into account the formula (2.23), the equation (1.1) takes the form of ratio "input-output" between the $f(t)$ input and $F(t)$ output functions of the system (1.1), (1.2).

3 The comparison of the two criteria

Theorem 1. *Solutions of the system:*

$$\begin{aligned} &Bz = 0, \\ &BAz = 0, \\ &BA^2z = 0, \\ &... \\ &BA^lz = 0, \quad l \leqslant n \end{aligned} \tag{3.1}$$

are the elements of $z \in KerB_l$ *, and only them.*

Proof. Solution of the equation $B_z = 0$ is element $z = P(B)z \in KerB$.

From the equation $BAz = 0$ we obtain $Az = P(B)Az$ and $(I - P(B))Az = 0$, them $(I - P(B))AP(B)z = 0$ or $B_1z = 0$. Therefore: $z = P(B_1)z \in KerB_1$.

From the equatin $BA^2z = 0$ we obtain $A^2z = P(B)A^2z$ and$(I - P(B))A^2z = 0$, then $(I - P(B))AP(B)Az = 0$ or $B_1Az = 0$. Therefore $Az = P(B_1)Az$, then $(I-P(B_1))Az = 0$ and $(I-P(B_1))P(B)AP(B)P(B_1)z = 0$ or $B_2z = 0$. Therefore: $z = P(B)z = P(B_1)z = P(B_2)z$.

And so on.

Just, we obtain: $z = P(B)z = P(B_1)z = ... = P(B_l)z$ so $z \in KerB_l$.

Transformations are equivalent.

It's true: the element $z \in KerB_l$ is a solution of the system (3.1).

That is, the condition of $KerB_l = \{0\}$ is equivalent to the condition: the system (3.1) has a unique solution $z = 0$.

Consequently, we prove the equivalence criterion (B) and the criterion (A).

The system (3.1) contains $l + 1$ - conditions, the remaining $m - l$ conditions are redundant.

Criterion (A) specifies the following.

Theorem 2.

Let $KerBp = 0$.

The system (1.1), (1.2) *is completely observable if and only if from the condition* $BA^iz = 0$ $(i = \overline{0, p})$ *should be* $z = 0$.

That is: $l = p$.

The system (3.1) contains $p + 1$ - conditions.

References

1. Boyarintsev Y.E. Regular and singular system of linear ordinary differential equations / Y.E. Boyarintsev - Novosibirsk: Nauka. Sib. div-e. 1980.-222 p.
2. Kvakernaak H. Linear optimal control systems / H. Kvakernaak, R. Sivan .- Springer-Verlag, 1977 .- 650 p.
3. On polynomial solutions of the linear stationary control system/ S.P. Zubova, L.H. Trung, E.V. Raetskaya // Automation and Remote Control. 2008. T.69. 11. – C. 1852-1858.
4. Full observability of time-dependent differential-algebraic systems / S.P. Zubova, E.V. Rajeckaya, Pham Tuan Cuong / / Bulletin of Voronezh State Technical University. Voronezh. - 2010. Vol/ 6. Number 82 010, the - S. 82-86.
5. On the invariance of non-stationary observations regarding some disturbances / / S.P. Zubova, E.V. Rajeckaya, Pham Tuan Cuong / / Vestnik Tambov University. Series: Natural and Technical Sciences. Tambov .- 2010. Volume 15, Issue 6. - S. 1678-1679.

S. P. Zubova
Contacts for the first author: Voronezh State University, Russia, 394 053, Voronezh, st. 60 Army d.25 kv.17, 8-951-851-36-30, spzubova@mail.ru

E. V. Raetskaya
Contacts for the second author: State Forestry Academy, Russia, 394 088, Voronezh, Victory Boulevard d.43 kv.312, 8-910-340-68-61, raetskaya@inbox.ru

V.2. Approximation theory and Fourier analysis

(Sessions organizers: Z. Ditzian, B. Kashin, S. Tikhonov)

ON J_K-LACUNARY SEQUENCES OF RECTANGULAR PARTIAL SUMS OF MULTIPLE FOURIER SERIES

I. Bloshanskii, O. Lifantseva

Key words: multiple trigonometric Fourier series, weak generalized localization almost everywhere, lacunary sequence, sets of convergence and divergence

AMS Mathematics Subject Classification: 42B05

Abstract. We study multiple trigonometric Fourier series of functions f in the classes $L_p(\mathbb{T}^N), p > 1$, which equal zero on some set $\mathfrak{A}$, $\mathfrak{A} \subset \mathbb{T}^N = [-\pi, \pi]^N$, $N \geq 3$, $\mu\mathfrak{A} > 0$. We consider the case when rectangular partial sums of the indicated Fourier series $S_n(x; f)$ have index $n = (n_1, \dots, n_N) \in \mathbb{Z}^N$, in which k ($k \geq 1$) components on the places $\{j_1, \dots, j_k\} = J_k \subset \{1, \dots, N\}$ are elements of (single) lacunary sequences. A correlation is found of the number k and location (the "sample" J_k) of lacunary sequences in the index n with structural and geometric characteristics of $\mathfrak{A}$. This correlation determines possibility of convergence almost everywhere of the considered series on some subset of positive measure $\mathfrak{A}_1$ of the set $\mathfrak{A}$.

1 Discussion of the problem

Let $\mathfrak{A}$ be an arbitrary measurable set, $\mathfrak{A} \subset \mathbb{T}^N = [-\pi, \pi]^N$, $\mu\mathfrak{A} > 0$ ($\mu = \mu_N$ is the N-dimensional Lebesgue measure), and let $f(x) = 0$ on $\mathfrak{A}$.

I. L. Bloshanskii [1, 2] obtained the necessary and sufficient conditions on structure and geometry of $\mathfrak{A}$ that guarantee convergence almost everywhere (a.e.) on some subset of positive measure $\mathfrak{A}_1$ of the set $\mathfrak{A}$ of multiple trigonometric Fourier series (of functions $f \in L_p(\mathbb{T}^N)$, $p \geqslant 1$, $N \geqslant 2$, $f(x) = 0$ on $\mathfrak{A}$) which are "classically" summed over rectangles.

In the present paper we are investigating the same problem but in the case when rectangular partial sums $S_n(x; f)$ of the indicated Fourier series have index $n = (n_1, \dots, n_N) \in \mathbb{Z}^N$, in which k ($k \geqslant 1$) components on the places $\{j_1, \dots, j_k\} = J_k \subset \{1, \dots, N\}$ are elements of (single) lacunary ($\{n^{(s)}\}$, $n^{(s)} \in \mathbb{Z}^1$, such that $\frac{n^{(s+1)}}{n^{(s)}} \geqslant q > 1$, $s = 1, 2, \dots$) sequences (i.e. we consider multiple Fourier series with J_k-lacunary sequence of partial sums).

This work is supported by grant 11-01-00321 of the Russian Foundation for Basic Research.

2 Convergence of Fourier series

An a priori possibility to obtain new results in the case under investigation is connected with the following results on convergence of one-dimensional and multiple Fourier series with "lacunary sequence of partial sums".

In the one-dimensional case A.N. Kolmogorov even in 1922 established: for any function $f \in L_2(\mathbb{T}^1)$ the sequence of partial sums $S_{n^{(k)}}(x;f)$ (where $\{n^{(k)}\}$, $n^{(k)} \in \mathbb{Z}^1$, $k = 1,\ 2, \dots$, is a lacunary sequence) converges a.e. on $\mathbb{T}^1$. In 1931 this result was extended by J. Littlewood and R. Paley on the classes $L_p(\mathbb{T}^1), p > 1$.[1] Later R. Gosselin and V. Totik established that in $L_1(\mathbb{T}^1)$ this result is not valid.

For multiple series (i.e. for $N \geq 2$) the first result concerning "lacunary sequences of partial sums" was obtained by P. Sjölin in 1971 in his paper [3], where he proved that if $f \in L_p(\mathbb{T}^2)$, $p > 1$, and $\{n_1^{(\nu_1)}\}$, $n_1^{(\nu_1)} \in \mathbb{Z}^1, \nu_1 = 1, 2, \dots$, is a single lacunary sequence, then $\lim\limits_{\nu_1,\, n_2 \to \infty} S_{n_1^{(\nu_1)},\, n_2}(x;f) = f(x)$ a.e. on $\mathbb{T}^2$. In 1977 M. Kojima [4] extended P. Sjölin's result and proved that if $f \in L_p(\mathbb{T}^N)$, $p > 1$, $N \geqslant 2$, and $\{n_j^{(\nu_j)}\}$, $n_j^{(\nu_j)} \in \mathbb{Z}^1$, $\nu_j = 1, 2, \dots, j = 1, \dots, N-1$, are single lacunary sequences, then $\lim\limits_{\nu_1, \dots, \nu_{N-1},\, n_N \to \infty} S_{n_1^{(\nu_1)}, \dots, n_{N-1}^{(\nu_{N-1})},\, n_N}(x;f) = f(x)$ a.e. on $\mathbb{T}^N$. In the same paper M. Kojima (using Ch. Fefferman's counterexample) ascertained that the result formulated above can not be improved.

3 Setting of the problem

Let us return to the problem (mentioned above) to find the structural and geometric characteristics of the set $\mathfrak{A}$, which guarantee convergence on some measurable subset $\mathfrak{A}_1 \subset \mathfrak{A}$ of multiple Fourier series (of $f \in L_p(\mathbb{T}^N)$, $p \geqslant 1$, $f(x) = 0$ on $\mathfrak{A}$). It is convenient to formulate and solve this problem in terms of weak generalized localization a.e. (WGL), which was introduced and examined by I.L. Bloshanskii (see, e.g., [1, 2]).

Definition 1. *Let $\mathfrak{A}$, $\mathfrak{A} \subset \mathbb{T}^N$, $N \geqslant 2$, be an arbitrary set of positive measure. We will say, that for multiple Fourier series of functions in the class $L_p(\mathbb{T}^N)$, $p \geqslant 1$,* **weak generalized localization** *almost everywhere* (**WGL**) *is valid on the set $\mathfrak{A}$ if for any function $f \in L_p(\mathbb{T}^N)$, $f(x) = 0$ on $\mathfrak{A}$, there exists a subset $\mathfrak{A}_1 \subset \mathfrak{A}$, $\mu\mathfrak{A}_1 > 0$, such that $\lim\limits_{n \to \infty} S_n(x;f) = 0$ almost everywhere on $\mathfrak{A}_1$.*

Let us introduce the following notation.

[1] Here we must, naturally, mention results of 1966 by L. Carleson and 1967 by R. Hunt that one-dimensional Fourier series of any function in $L_p(\mathbb{T}^1), p > 1$, converges a.e. on $\mathbb{T}^1$.

Let $M = \{1, \dots, N\}$ and $k \in M$. Denote: $J_k = \{j_1, ..., j_k\}$, $j_s < j_l$ for $s < l$, and (in the case $k < N$) $M \setminus J_k = \{m_1, \dots, m_{N-k}\}$, $m_s < m_l$ for $s < l$, are nonempty subsets of the set M. Let us consider also that $J_0 = \emptyset$ and $M \setminus J_N = \emptyset$. We expand the space $\mathbb{R}^N$ into the sum of two subspaces $\mathbb{R}[J_k]$ and $\mathbb{R}[M \setminus J_k]$, where $\mathbb{R}[J_k] = \{x = (x_1, \dots, x_N) \in \mathbb{R}^N : x_j = 0 \text{ for } j \in M \setminus J_k\}$, and $\mathbb{R}[M \setminus J_k] = \{x \in \mathbb{R}^N : x_j = 0 \text{ for } j \in J_k\}$. Denote also $\mathbb{T}[J_k] = \{x \in \mathbb{R}[J_k] : -\pi \leqslant x_j \leqslant \pi \text{ for } j \in J_k\}$ and $\mathbb{T}[M \setminus J_k] = \{x \in \mathbb{R}[M \setminus J_k] : -\pi \leqslant x_j \leqslant \pi \text{ for } j \in M \setminus J_k\}$.

Let Ω, $\Omega \subset \mathbb{T}^N$, be an arbitrary (nonempty) open set, and $\Omega[J_2] = pr_{(J_2)}\{\Omega\}$ be an orthogonal projection of Ω on the space $\mathbb{R}[J_2]$, $J_2 \subset M$.

Let $N \geqslant 3$. Assume $W[J_2] = \Omega[J_2] \times \mathbb{T}[M \setminus J_2]$, $J_2 \subset M$.[1] The sets $W[J_2]$ we will call the "N-dimensional bars". Further, fix an arbitrary $J_k = J_k^0$, $0 \leqslant k \leqslant N-2$, and consider the following sets (see also [5]): the set

$$W = W(J_k) = W(\Omega, J_k) = \bigcup_{J_2 \subset M \setminus J_k^0} W[J_2] \tag{1}$$

(which we will call the "complete N-dimensional cross", if $J_k = \emptyset$, and "incomplete N-dimensional cross" if $J_k \neq \emptyset$) and the set

$$W^0 = W^0(J_k) = W^0(\Omega, J_k) = \bigcap_{J_2 \subset M \setminus J_k^0} W[J_2] \tag{2}$$

(which we will call the "center" of the corresponding "N-dimensional cross").

Definition 2. *We will say that a set $\mathcal{A}$ is inscribed almost everywhere in a set $\mathcal{B}$ if $\mu(\mathcal{A} \setminus \mathcal{B}) = 0$.*

Definition 3. *Let $\mathfrak{A} \subset \mathbb{T}^N$, $N \geqslant 3$, and $J_k \subset M$, $1 \leqslant k \leqslant N-2$, or $J_0 = \emptyset$, $k = 0$.*

1. *We will say that the set $\mathfrak{A}$ possesses the property $\mathbb{B}_2^{(J_k)}$ if there exists a set $W = W(J_k)$ of the type (1), which is inscribed a.e. in $\mathfrak{A}$, moreover, the property $\mathbb{B}_2^{(J_k)}$ is the property $\mathbb{B}_2^{(J_k)}(W^0)$ if $W = W(W^0)$.*

2. *The property $\mathbb{B}_2^{(J_k)}(W^0)$ of the set $\mathfrak{A}$ we will call the maximal property $\mathbb{B}_2^{(J_k)}$ of the set $\mathfrak{A}$, if for any set $\widetilde{W}^0 = \widetilde{W}^0(J_k)$ of the type (2) such that $\mu(\widetilde{W}^0 \setminus W^0) > 0$, the set $\mathfrak{A}$ does not possess the property $\mathbb{B}_2^{(J_k)}(\widetilde{W}^0)$.*

Note that for $k = 0$ the property $\mathbb{B}_2^{(J_0)} \equiv \mathbb{B}_2^{(\emptyset)}$ and the maximal property $\mathbb{B}_2^{(J_0)}(W^0(J_0)) \equiv \mathbb{B}_2^{(\emptyset)}(W^0(\emptyset))$ coincide, correspondingly, with the property $\mathbb{B}_2$ and

[1] In this case any vector $z = (z_1, \dots, z_{2N}) \in A \times B$, where $A \subset \mathbb{R}[J_k]$, $B \subset \mathbb{R}[M \setminus J_k]$, we identify with vector $x = (x_1, \dots, x_N) \in \mathbb{R}^N$ by formula: $x_s = z_s$ as $s \in J_k$ and $x_s = z_{N+s}$ as $s \in M \setminus J_k$.

the maximal property $\mathbb{B}_2(W^0)$, which were introduced and examined earlier in the works of I.L. Bloshanskii (see, e.g., [1, 2]). Thus, in [2, Theorem 2]) the following criterion for validity of WGL was obtained for multiple Fourier series of functions in $L_p(\mathbb{T}^N)$, $p > 1$, in terms of $\mathbb{B}_2$ property.

Denote as $intP$ the set of interior points of the set $P \subset \mathbb{R}^N$; as $\overline{P}$ the closure of the set P and as $Fr\, P$ the boundary of P.

Let $\mathfrak{A}$ be an arbitrary measurable set, $\mathfrak{A} \subset \mathbb{T}^N$, $N \geqslant 2$, $0 < \mu\mathfrak{A} < (2\pi)^N$, $\mathfrak{B} = \mathbb{T}^N \setminus \mathfrak{A}$. Consider the following conditions on the boundary of $\mathfrak{A}$:

1. $\mu(\mathfrak{B} \setminus \overline{int\mathfrak{B}}) = 0$; (3)
2. $\mu_2 Fr\, pr_{(J_2)}\{int\mathfrak{B}\} = 0$ for all $J_2 \subset M$, (4)

where μ_2 is the measure on the plane.

Theorem A. *Let $\mathfrak{A}$ be an arbitrary measurable set, $\mathfrak{A} \subset \mathbb{T}^N$, $N \geqslant 2$, $\mu\mathfrak{A} > 0$, and let $\mathfrak{A}$ satisfy conditions* (3), (4). *Then on the set $\mathfrak{A}$ in the class $L_p(\mathbb{T}^N)$, $p > 1$, weak generalized localization almost everywhere is valid if and only if this set possesses the property $\mathbb{B}_2$.*

Remark 1. As it was established in [2], in the part of sufficiency Theorem A is true without restrictions (3) and (4).

Remark 2. The set $\mathfrak{A} \subset \mathbb{T}^2$, possessing the property $\mathbb{B}_2$ ($\mathbb{B}_2^{(\emptyset)}$ in terms of Definition 3), is the set for which there exists an open set Ω, $\Omega \subset \mathbb{T}^2$, such that $\mu(\Omega \setminus \mathfrak{A}) = 0$ (see [1, 2]).

The question arises: what structural and geometric characteristics must an arbitrary set $\mathfrak{A}$, $\mathfrak{A} \subset \mathbb{T}^N$, possess in order WGL be valid on this set for multiple Fourier series summed over rectangles, in the case when some of the components n_j, $j \in J_k$, of the index n, $n \in \mathbb{Z}^N$, of the rectangular partial sum $S_n(x; f)$ are elements of (single) lacunary sequences. In particular, to what extent these structural and geometric characteristics of the set $\mathfrak{A}$ remain "stable" when we vary the number k and the location (i.e. the "sample" J_k) of the indicated lacunary sequences in the index n?

4 Formulation of the basic result

Let $\alpha = \alpha(J_k) = (\alpha_{j_1}, \ldots, \alpha_{j_k}) \in \mathbb{Z}^k$, $j_s \in J_k$, $s = 1, \ldots, k$, $1 \leqslant k \leqslant N - 2$. By the symbol $n^{(\alpha)} = n^{(\alpha)}[J_k] = (n_1, \ldots, n_N) \in \mathbb{Z}^N$ let us denote the N-dimensional vector whose components n_j with the numbers $j \in J_k$ are elements of *some* (single infinitely large) *sequences of numbers* (for $j \in J_k : n_j = n_j^{(\alpha_j)}$ and $n_j^{(\alpha_j)} \to \infty$ as $\alpha_j \to \infty$). In particular, by the symbol $n^{(\lambda)} = n^{(\lambda)}[J_k] \in \mathbb{Z}^N$ (where $\lambda = \lambda(J_k) = (\lambda_{j_1}, \ldots, \lambda_{j_k}) \in \mathbb{Z}^k$, $j_s \in J_k$, $s = 1, \ldots, k$) we will denote vector whose components n_j, $j \in J_k$, are elements of some (single) *lacunary* sequences.

Theorem 1. *Let $\mathfrak{A}$ be an arbitrary set, $\mathfrak{A} \subset \mathbb{T}^N$, $N \geqslant 3$, $0 < \mu\mathfrak{A} < (2\pi)^N$, and let J_k be an arbitrary "sample" from M, $1 \leqslant k \leqslant N-2$. If the set $\mathfrak{A}$ satisfies conditions (3), (4′), where*

$$\mu_2 Frpr_{(J_2)}\{int\mathfrak{B}\} = 0 \text{ for all } J_2 \subset M \setminus J_k, \tag{4'}$$

then on the set $\mathfrak{A}$ in the class $L_p(\mathbb{T}^N)$, $p > 1$, for multiple Fourier series whose rectangular partial sums $S_n(x; f)$ have index $n = n^{(\lambda)}[J_k]$ weak generalized localization almost everywhere is valid if and only if the set $\mathfrak{A}$ possesses the property $\mathbb{B}_2^{(J_k)}$.

Remark 3. In the part of sufficiency Theorem 1 is true without restrictions (3), (4′).

Theorem 1 gives no information about those subsets $\mathfrak{A}_1 \subset \mathfrak{A}$ on which there exists a limit of "J_k-lacunary sequence of rectangular partial sums" $S_{n^{(\lambda)}[J_k]}(x; f)$ under the hypothesis $f(x) = 0$ on $\mathfrak{A}$, and about those subsets on which such is not the case. Therefore it is expedient to give a more extended formulation of this theorem (adding the case $k = 0$).

Theorem 1′. *Let $\mathfrak{A}$ be an arbitrary set, $\mathfrak{A} \subset \mathbb{T}^N$, $N \geqslant 3$, $0 < \mu\mathfrak{A} < (2\pi)^N$, and let $J_k \subset M$, $1 \leqslant k \leqslant N-2$, or $J_k = \emptyset$ for $k = 0$.*

1. *If there exists a set $W^0 = W^0(J_k)$ of the type (2) such that the set $\mathfrak{A}$ possesses the property $\mathbb{B}_2^{(J_k)}(W^0)$, then for any function $f \in L_p(\mathbb{T}^N)$, $p > 1$, such that $f(x) = 0$ on $\mathfrak{A}$,*

$$\lim_{\substack{\lambda_j \to \infty, j \in J_k, \\ n_j \to \infty, j \in M \setminus J_k}} S_{n^{(\lambda)}[J_k]}(x; f) = 0 \text{ almost everywhere on } W^0.$$

Let, in addition, the set $\mathfrak{A}$ satisfy conditions (3), (4′), then

2. *If the property $\mathbb{B}_2^{(J_k)}(W^0)$ of the set $\mathfrak{A}$ is the maximal property $\mathbb{B}_2^{(J_k)}$, then there exists a function $f_1 \in L_\infty(\mathbb{T}^N)$ such that $f_1(x) = 0$ on $\mathfrak{A}$, but for any k sequences of numbers $\{n_j^{(\alpha_j)}\}$, $j \in J_k$, $n_j^{(\alpha_j)} \to \infty$ as $\alpha_j \to \infty$, [1]*

$$\overline{\lim_{\substack{\alpha_j \to \infty, j \in J_k, \\ n_j \to \infty, j \in M \setminus J_k}}} |S_{n^{(\alpha)}[J_k]}(x; f_1)| = +\infty \text{ almost everywhere on } \mathbb{T}^N \setminus W^0.$$

3. *In particular, if the set $\mathfrak{A}$ does not possess the property $\mathbb{B}_2^{(J_k)}$ at all, then there exists a function $f_2 \in L_\infty(\mathbb{T}^N)$ such that $f_2(x) = 0$ on $\mathfrak{A}$, but for any k sequences*

[1] In particular, all sequences $\{n_j^{(\alpha_j)}\}$, $j \in J_k$, can be lacunary (in this case in our notation the index of partial sums $n = n^{(\lambda)}[J_k]$) or, for example (if $N \geqslant 4$ and $k \geqslant 2$), can be termwise equal (i.e. $n_{j_1}^{(\alpha_{j_1})} = \cdots = n_{j_k}^{(\alpha_{j_k})} = n_0$).

of numbers $\{n_j^{(\alpha_j)}\}$, $j \in J_k$, $n_j^{(\alpha_j)} \to \infty$ *as* $\alpha_j \to \infty$,

$$\overline{\lim_{\substack{\alpha_j \to \infty, j \in J_k, \\ n_j \to \infty, j \in M \setminus J_k}}} |S_{n^{(\alpha)}[J_k]}(x; f_2)| = +\infty \ \textit{almost everywhere on} \ \mathbb{T}^N.$$

Remark 4. Theorem 1′ for $k = 0$ coincides with the "extended" formulation of Theorem A (see [2, Theorem 2′]).

So, we see that for any k, $1 \leqslant k \leqslant N-2$, validity or invalidity of WGL for multiple Fourier series (summed over rectangles) in the classes L_p, $p > 1$, on the set $\mathfrak{A} \subset \mathbb{T}^N$ are defined by the structure and geometry of the set $\mathfrak{A}$, which, in its turn, are defined by the property $\mathbb{B}_2^{(J_k)}$, where parameter k is the number of "lacunary components" of the vector $n \in \mathbb{Z}^N$ (the index of the partial sum $S_n(x; f)$). [1]

Remark 5. Comparing Theorem A (Theorem 1′ for $k = 0$) and Theorem 1 (Theorem 1′ for $1 \leqslant k \leqslant N-2$), we see that for validity on the measurable set $\mathfrak{A} \subset \mathbb{T}^N$, $N \geqslant 3$, of WGL (for summed over rectangles multiple Fourier series of function $f \in L_p$, $p > 1$, $f(x) = 0$ on $\mathfrak{A}$) in the case when all components of the vector $n \in \mathbb{Z}^N$ – the index of the partial sum $S_n(x; f)$ – are "free", some "more severe" constraints must be posed on the set $\mathfrak{A}$, described by the property $\mathbb{B}_2$ ($\equiv \mathbb{B}_2^{(\emptyset)}$), more severe than constraints on the same set in the case when some components of the vector n are lacunary.

Further, let us note that while (under the growth of k, $1 \leqslant k \leqslant N-2$) the constraints on structural and geometric characteristics of the set $\mathfrak{A}$ (described in Theorem 1′ by the property $\mathbb{B}_2^{(J_k)}$ of the set $\mathfrak{A}$) become "more mild", the constraints on the sequences of partial sums $S_{n^{(\lambda)}[J_k]}(x; f)$ become "more severe", remaining "free" (non-lacunary) less and less components in the vector $n = n^{(\lambda)}[J_k]$ – the index of the partial sum of the multiple Fourier series under consideration. And, finally, in the "limiting case" (when only two components of the vector n remain free) "only" the following constraint must be imposed on the structure and geometry of the set $\mathfrak{A}$: there must exist the "N - dimensional bar" $W[J_2]$ which is inscribed a.e. in $\mathfrak{A}$.

Let us emphasize that if we decrease more the number of "free" components of the vector $n = n^{(\lambda)}[J_k]$ (thus, reducing this number to one and, naturally, remaining lacunary all the rest components), then, as it follows from the mentioned above results of M. Kojima [4] and P. Sjolin [3], for validity of WGL on the set $\mathfrak{A}$ in the

[1] Let us emphasize, that for the fixed $J_k \subset M$ the "N-dimensional bars" $W[J_2]$, $J_2 \subset M \setminus J_k$, constructing the "incomplete cross" $W(J_k)$ - (1) (which satisfies condition $\mu(W(J_k) \setminus \mathfrak{A}) = 0$), have the "bases" $\Omega[J_2]$ in those planes $\mathbb{R}[J_2]$, for which the corresponding components of the index $n \in \mathbb{Z}^N$ – components n_j, $j \in M \setminus J_k$, are "free" (i.e., in particular, are not components of any lacunary sequences).

classes L_p, $p > 1$, no restrictions should be imposed on $\mathfrak{A}$ (in view of its structure and geometry), except measurability.

The "expanded" Theorem $1'$ permits to find a correlation between the structural and geometric characteristics of the set $\mathfrak{A}$ (described by the property $\mathbb{B}_2^{(J_k)}$ of the set $\mathfrak{A}$) and the "lacunarity" of sequence of partial sums $S_n(x;f)$ (described by the number and location of the lacunary components in the vector n), which determine possibility of convergence a.e. of the considered multiple Fourier series.

Remark 6. So, let the set $\mathfrak{A}$ possess the property $\mathbb{B}_2^{(J_k)}(W^0(\Omega, J_k))$, $0 \leqslant k \leqslant N-2$, $f \in L_p(\mathbb{T}^N)$, $p > 1$, $f(x) = 0$ on $\mathfrak{A}$, and let a partial sum $S_n(x;f)$ have an index n, in which the components $j \in J_m$, $0 \leqslant m \leqslant N-2$, are lacunary, i.e. $n = n^{(\lambda)}[J_m]$. Then

1) if $J_m = J_k$, then the sequence $S_{n^{(\lambda)}[J_m]}(x;f)$ converges a.e. on $W^0(\Omega, J_k)$, i.e.

$$\lim_{\substack{\lambda_j \to \infty, j \in J_m, \\ n_j \to \infty, j \in M \setminus J_m}} S_{n^{(\lambda)}[J_m]}(x;f) = 0 \text{ a.e. on } W^0(\Omega, J_k)$$

(lacunarity "corresponds" to the given structural and geometric characteristics of $\mathfrak{A}$);

2) if J_m such that $J_k \not\subseteq J_m$, then the sequence $S_{n^{(\lambda)}[J_m]}(x;f)$ would not converge (generally speaking) a.e. on $W^0(\Omega, J_k)$ (lacunarity is "insufficient" for the given structural and geometric characteristics of $\mathfrak{A}$);

3) if J_m such that $J_k \subset J_m$, $k < m$, then the limit of the sequence $S_{n^{(\lambda)}[J_m]}(x;f)$ exists, generally speaking, on the more wide set, i.e.

$$\lim_{\substack{\lambda_j \to \infty, j \in J_m, \\ n_j \to \infty, j \in M \setminus J_m}} S_{n^{(\lambda)}[J_m]}(x;f) = 0 \text{ a.e. on } W^0(\Omega, J_m) \supseteq W^0(\Omega, J_k)$$

(lacunarity is "excessive" for the given structural and geometric characteristics of $\mathfrak{A}$).

And finally, let us indicate some generalization of Theorem 1 with the object of weakening restrictions on the set $\mathfrak{A}$.

Theorem 2. *Let $\mathfrak{A}$ be an arbitrary measurable set, $\mathfrak{A} \subset \mathbb{T}^N$, $N \geqslant 3$, $\mu\mathfrak{A} > 0$, $\mathfrak{B} = \mathbb{T}^N \setminus \mathfrak{A}$, and let J_k be an arbitrary "sample" from M, $1 \leqslant k \leqslant N-2$. If there exist a subset $\mathfrak{B}_1 \subset \mathfrak{B}$ and an open set Ω such that*

1. $\mu(\mathfrak{B}_1 \triangle \Omega) = 0$; (5)
2. $\mu_2 Frpr_{(J_2)}\{\Omega\} = 0$ *for all* $J_2 \subset M \setminus J_k$; (6)
3. $\mu_2 pr_{(J_2)}\{\Omega\} = \mu_2 pr_{(J_2)}\{\mathfrak{B}\}$ *for all* $J_2 \subset M \setminus J_k$, (7)

then on the set $\mathfrak{A}$ in the class $L_p(\mathbb{T}^N)$, $p > 1$, for multiple Fourier series whose rectangular partial sums $S_n(x;f)$ have the index $n = n^{(\lambda)}[J_k]$, weak generalized localization almost everywhere is valid if and only if the set $\mathfrak{A}$ possesses the property $\mathbb{B}_2^{(J_k)}$.

Remark 7. Restrictions on the set $\mathfrak{B}$ in the form of conditions (5) - (7) appeared for the first time (for $N \geqslant 2$ and $k = 0$) in [2]. These conditions are satisfied, for example, for the sets $\mathfrak{B}$, which can be represented in the form $\mathfrak{B} = \mathfrak{B}^{(1)} \bigcup \mathfrak{B}^{(2)}$, $\mathfrak{B}^{(1)} \bigcap \mathfrak{B}^{(2)} = \emptyset$, where $\mathfrak{B}^{(1)} = \varphi(S_1)$, $\mathfrak{B}^{(2)} \subset \varphi(S_2)$, $S_1 = \{x \in \mathbb{T}^N : 0 < r_1 < |x| < r_2 < \pi\}$, $S_2 = \{x \in \mathbb{T}^N : |x| < r_1\}$ and φ is a homeomorphism, $\varphi : \mathbb{T}^N \to \mathbb{T}^N$; in this case a measurable set $\mathfrak{B}^{(2)}$ can have an arbitrarily "bad" geometry and structure.

References

1. Bloshanskii I.L. *On criteria for weak generalized localization in N-dimensional space.* Dokl. Akad. Nauk SSSR. **271**, 6, 1983. P. 1294-1298; Eng. tr. in Soviet Math. Dokl. **28**, 1983.
2. Bloshanskii I.L. *Two criteria for weak generalized localization for multiple trigonometric Fourier series of functions in L_p, $p \geqslant 1$.* Izv. Akad. Nauk, Ser. Matem. **49**, 2, 1985. P. 243-282; Eng. tr. in Math. USSR Izv. **26**, 2, 1986.
3. Sjölin P. *Convergence almost everywhere of certain singular integrals and multiple Fourier series.* Arkiv Matem. **9**, 1, 1971. P. 65-90.
4. Kojima M. *On the almost everywhere convergence of rectangular partial sums of multiple Fourier series.* Sci. Repts. Kanazava Univ. **22**, 2, 1977. P. 163-177.
5. Bloshanskii I.L., Lifantseva O.V. *A weak generalized localization criterion for multiple Fourier series whose rectangular partial sums are considered over a subsequence.* Dokl. Akad. Nauk. **423**, 4, 2008. P. 439-442; Eng. tr. in Doklady Mathematics **78**, 3, 2008. P. 864-867.

I. Bloshanskii
Moscow Regional State University, Russia, 105005, Moscow, Radio st., 10a,
ig.bloshn@gmail

O. Lifantseva
Moscow Regional State University, Russia, 105005, Moscow, Radio st., 10a,
ov-lifantseva@yandex.ru

APPROXIMATION PROPERTIES OF SYSTEMS OF ROOT FUNCTIONS OF WELL-POSED BOUNDARY VALUE PROBLEMS FOR THE TWO-FOLD DIFFERENTIATION OPERATOR

B. E. Kanguzhin, D. B. Nurakhmetov

Key words: systems of root functions, well-posed boundary value problems, completeness of systems of root functions, resolvent of the operator

AMS Mathematics Subject Classification: 30E10, 30E15, 30E20

Abstract. In this paper, we consider the two-fold differentiation operator $L_{\sigma_1\sigma_2}$ in the function space $\mathbf{L}_2(0,\,1)$ corresponding to nonlocal problem

$$l(y) \equiv -y''(x) = f(x), \quad 0 < x < 1,$$

$$U_\nu(y) \equiv y^{(\nu-1)}(0) - \int\limits_0^1 (-y''(x))\overline{\sigma_\nu(x)}dx = 0, \quad \nu = 1, 2,$$

where $\sigma_\nu(x)$ is the boundary function from the space $\mathbf{L}_2(0,1)$, $\overline{\sigma_\nu}$ denotes the complex conjugate, $i^2 = -1$. We investigate the approximation properties of systems of root functions of the operator $L_{\sigma_1\sigma_2}$. We obtain a sufficient condition for the completeness of systems of root functions of the operator $L_{\sigma_1\sigma_2}$ in terms of the boundary function.

1 Introduction

Let $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ be arbitrary functions from the function space $\mathbf{L}_2(0,\,1)$. We introduce the entire functions with respect to λ

$$\Delta(\lambda) = 1 - \lambda\, A + \lambda^2 \int\limits_0^1 \frac{\sin\sqrt{\lambda}t}{\sqrt{\lambda}} K(t)dt, \tag{1}$$

$$\kappa_1(x,\lambda) = \cos\sqrt{\lambda}x - \lambda \int\limits_0^1 \frac{\sin\sqrt{\lambda}(t-x)}{\sqrt{\lambda}}\, \overline{\sigma_2(t)}dt,$$

$$\kappa_2(x,\lambda)=\frac{\sin\sqrt{\lambda}x}{\sqrt{\lambda}}-\lambda\int\limits_0^1\frac{\sin\sqrt{\lambda}(x-t)}{\sqrt{\lambda}}\overline{\sigma_1(t)}dt,$$

where $A=\int\limits_0^1\overline{\sigma_1(x)}dx+\int\limits_0^1 x\,\overline{\sigma_2(x)}dx,$

$$K(t)=\int\limits_t^1\left(\overline{\sigma_1(x)}+(x-t)\overline{\sigma_2(x)}-\overline{\sigma_1(x)}\,\overline{\sigma_2(x-t)}\right)dx+\int\limits_0^{1-t}\overline{\sigma_1(x)}\,\overline{\sigma_2(x+t)}dx.$$

Denote by $\Lambda=\{\lambda_1,\,\lambda_2,\,\dots\}$ the sequence of zeros of function $\Delta(\lambda)$. Each zero λ_s has a some multiplicity m_s. Then we introduce three systems of functions

$$Y^{(\nu)}_{\sigma_1\sigma_2}=\left\{\lim_{\lambda\to\lambda_s}\frac{1}{j!}\frac{\partial^j\kappa_\nu(x,\lambda)}{\partial\lambda^j}:\nu=1,\,2,\,j=\overline{0,m_s-1},\lambda_s\in\Lambda\right\},\ Y_{\sigma_1\sigma_2}=Y^{(1)}_{\sigma_1\sigma_2}\bigcup Y^{(2)}_{\sigma_2\sigma_2}.$$

The main problem: *under which conditions on $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ from $\mathbf{L}_2(0,\,1)$ the systems of functions $Y^{(1)}_{\sigma_1\sigma_2}$, $Y^{(2)}_{\sigma_2\sigma_2}$, and $Y_{\sigma_1\sigma_2}$ form complete systems in $\mathbf{L}_2(0,\,1)$?*

Note that $Y_{\sigma_1\sigma_2}$ is a system of root functions of the two-fold differentiation operator, where the functions $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ are boundary functions. Details are described in section 2 below. In the case of the usual differentiation operator, the system of root functions is the system of exponentials; (see [5]).

Let us formulate own main results.

Theorem 1. *The system of function $Y^{(1)}_{\sigma_1\sigma_2}$ is complete in $\mathbf{L}_2(0,\,1)$ if the following conditions hold*
1) family of functions $\{D(t,\,\cdot):t\in(0,\,1)\}$ is dense in $\mathbf{L}_2(0,\,1)$, where $D(t,\cdot)$ is defined by

$$D(t,x)=\begin{cases}\sigma_2(x+t), & \text{for}\quad 0\leqslant x\leqslant t;\\ 1-\sigma_2(x-t)+\sigma_2(x+t), & \text{for}\quad t<x\leqslant 1-t;\quad \text{for } 0\leqslant t\leqslant\dfrac{1}{2}\\ 1-\sigma_2(x-t), & \text{for}\quad 1-t<x\leqslant 1.\end{cases}\tag{2}$$

$$D(t,x)=\begin{cases}\sigma_2(x+t), & \text{for}\quad 0<x\leqslant 1-t;\\ 0, & \text{for}\quad 1-t<x\leqslant t;\quad \text{for } 1\geqslant t\geqslant\dfrac{1}{2}\\ 1-\sigma_2(x-t), & \text{for}\quad t<x\leqslant 1.\end{cases}$$

2) for some $\varepsilon>0$ the boundary function $\sigma_1(\cdot)\in\mathbf{W}^1_2([0,\,\varepsilon]\cup[1-\varepsilon,\,1])\cap\mathbf{L}_2(0,\,1)$, the boundary function $\sigma_2(\cdot)\in\mathbf{W}^2_2([0,\,\varepsilon]\cup[1-\varepsilon,\,1])\cap\mathbf{W}^1_2[0,\,1]$, and $-\overline{\sigma_1(1)}+\overline{\sigma_1(1)}\,\overline{\sigma_2(0)}-\overline{\sigma_1(0)}\,\overline{\sigma_2(1)}\neq 0$.

Theorem 2. *The system of function* $Y^{(2)}_{\sigma_1\sigma_2}$ *is complete in* $\mathbf{L}_2(0,\,1)$ *if the following conditions hold*
1) family of functions $\{B(t,\,\cdot) : t \in (0,\,1)\}$ *is dense in* $\mathbf{L}_2(0,\,1)$, *where* $B(t,\cdot)$ *is defined by*

$$B(t,x) = \begin{cases} -\sigma_1(x-t) + \sigma_1(x+t), & for \quad 0 \leqslant x \leqslant t; \\ x - t + \sigma_1(x+t), & for \quad t < x \leqslant 1-t; \quad for\ 0 \leqslant t \leqslant \dfrac{1}{2} \\ x - t, & for \quad 1-t < x \leqslant 1. \end{cases} \tag{3}$$

$$B(t,x) = \begin{cases} \sigma_1(x+t) - \sigma_1(x-t), & for \quad 0 \leqslant x \leqslant 1-t; \\ -\sigma_1(x-t), & for \quad 1-t < x \leqslant t; \quad for\ \ 1 \geqslant t \geqslant \dfrac{1}{2} \\ x - t, & for \quad t < x \leqslant 1. \end{cases}$$

2) for some $\varepsilon > 0$ *the boundary function* $\sigma_1(\cdot) \in \mathbf{W}^1_2([0,\,\varepsilon] \cup [1-\varepsilon,\,1]) \cap \mathbf{L}_2(0,\,1)$, *the boundary function* $\sigma_2(\cdot) \in \mathbf{W}^2_2([0,\,\varepsilon] \cup [1-\varepsilon,\,1]) \cap \mathbf{W}^1_2[0,\,1]$, *and* $-\overline{\sigma_1(1)} + \overline{\sigma_1(1)}\,\overline{\sigma_2(0)} - \overline{\sigma_1(0)}\,\overline{\sigma_2(1)} \neq 0$.

Theorem 3. *If the following conditions hold*
1) family of functions $\{B(t,\cdot), D(t,\cdot) : 0 \leqslant t \leqslant 1\}$ *is dense in* $\mathbf{L}_2(0,1)$.
2) for some $\varepsilon > 0$ *the boundary function* $\sigma_1(\cdot) \in \mathbf{W}^1_2([0,\,\varepsilon] \cup [1-\varepsilon,\,1]) \cap \mathbf{L}_2(0,\,1)$, *the boundary function* $\sigma_2(\cdot) \in \mathbf{W}^2_2([0,\,\varepsilon] \cup [1-\varepsilon,\,1]) \cap \mathbf{W}^1_2[0,\,1]$, *and* $-\overline{\sigma_1(1)} + \overline{\sigma_1(1)}\,\overline{\sigma_2(0)} - \overline{\sigma_1(0)}\,\overline{\sigma_2(1)} \neq 0$, *then system of functions* $Y_{\sigma_1\sigma_2}$ *is dense in the function space* $\mathbf{L}_2(0,1)$

All theorems are proved using the same scheme. For example, the proof of Theorem 1 consists of two stages:
1) first, we find the criteria for density in $\mathbf{L}_2(0,1)$ of the family of functions $\{\kappa_1(\cdot,\,\mu), \forall \mu \in \mathbb{C}\}$,
2) second, we find sufficient conditions for a function $\kappa_1(\cdot,\,\mu)$ to be approximated by linear combination of elements of $Y^{(1)}_{\sigma_1\sigma_2}$.

In fact, in the second part of the proof can be obtained by a stronger statement than the completeness.

Statement 1. *Under conditions of Theorem 1, there exists* $\{R_N\}_{N=1}^{\infty}$, $R_N \to \infty$ *such that we have*

$$\kappa_1(x,\,\mu) = \lim_{\lambda \to \lambda_s} \lim_{N \to \infty} \sum_{|\lambda_S| < R_N} \sum_{\nu=1}^{2} \sum_{j=0}^{m_s-1} c^{(\nu)}_{s,\,m_s-1-j} \frac{1}{j!} \frac{\partial^j \kappa_\nu(x,\,\lambda)}{\partial \lambda^j},$$

where $c^{(\nu)}_{s,\,m_s-1-j}$ *are analogies of the Fourier coefficients of* $\kappa_1(x,\,\mu)$ *with respect to the system* $Y_{\sigma_1\sigma_2}$. Similar facts are true also in the case of Theorem 2 and 3.

2 Boundary problems and the necessary notation

In [3] it is proved the following statement

Theorem 4 (M. Otelbaev)

(a) For any choice of functions $\sigma_\nu(x)$, $\nu = 1, 2$ from the space $\mathbf{L}_2(0,1)$, we consider the nonlocal boundary value problem

$$-y''(x) = f(x), \quad 0 < x < 1, \tag{4}$$

$$y^{(\nu-1)}(0) - \int_0^1 (-y''(x))\overline{\sigma_\nu(x)}dx = 0, \quad \nu = 1, 2. \tag{5}$$

in the space $\mathbf{L}_2(0,1)$, which corresponds to the operator L. Then L has completely continuous inverse L^{-1}.

(b) Assume that a nonhomogeneous equation (4) with some additional conditions for any right-hand side $f(x) \in \mathbf{L}_2(0,1)$ has a unique solution $y(x)$ in the space $\mathbf{W}_2^2[0,1]$, where $y(x)$ satisfies the a priori estimate $\| y \|_{L_2(0,1)} \leqslant c \| f \|_{L_2(0,1)}$. Then there exists a unique set of functions $\{\sigma_\nu(x)\}, \nu = 1, 2$ from the space $\mathbf{L}_2(0,1)$ such that any additional condition is equivalent to (5).

Definition 1 *The following function with respect to λ*

$$\Delta(\lambda) \equiv \begin{vmatrix} 1 - \lambda \int_0^1 \cos\sqrt{\lambda}x\overline{\sigma_1(x)}dx & -\lambda \int_0^1 \frac{\sin\sqrt{\lambda}x}{\sqrt{\lambda}}\overline{\sigma_1(x)}dx \\ -\lambda \int_0^1 \cos\sqrt{\lambda}x\overline{\sigma_2(x)}dx & 1 - \lambda \int_0^1 \frac{\sin\sqrt{\lambda}x}{\sqrt{\lambda}}\overline{\sigma_2(x)}dx \end{vmatrix} \tag{6}$$

is called the characteristic determinant of boundary value problem (4) and (5).

It is convenient to consider the following entire functions with respect to λ

$$\kappa_1(x,\lambda) \equiv \begin{vmatrix} \cos\sqrt{\lambda}x & \frac{\sin\sqrt{\lambda}x}{\sqrt{\lambda}} \\ -\lambda \int_0^1 \cos\sqrt{\lambda}\tau\overline{\sigma_2(\tau)}d\tau & 1 - \lambda \int_0^1 \frac{\sin\sqrt{\lambda}\tau}{\sqrt{\lambda}}\overline{\sigma_2(\tau)}d\tau \end{vmatrix},$$

$$\kappa_2(x,\lambda) \equiv \begin{vmatrix} 1 - \lambda \int_0^1 \cos\sqrt{\lambda}\tau\overline{\sigma_1(\tau)}d\tau & -\lambda \int_0^1 \frac{\sin\sqrt{\lambda}\tau}{\sqrt{\lambda}}\overline{\sigma_1(\tau)}d\tau \\ \cos\sqrt{\lambda}x & \frac{\sin\sqrt{\lambda}x}{\sqrt{\lambda}} \end{vmatrix}.$$

The functions $\kappa_1(x,\lambda)$ and $\kappa_2(x,\lambda)$ are called principal solutions of the equation $-y''(x) = \lambda y(x)$. Further on we need certain important properties of the principal solutions $\kappa_1(x,\lambda)$ and $\kappa_2(x,\lambda)$.

Lemma 1. *For any complex numbers λ the following relations hold:*
1) $\kappa_1(0,\lambda) - \lambda \int\limits_0^1 \kappa_1(x,\lambda)\overline{\sigma_1(x)}dx = \Delta(\lambda)$, *2)* $\kappa_1'(0,\lambda) - \lambda \int\limits_0^1 \kappa_1(x,\lambda)\overline{\sigma_2(x)}dx = 0$,
3) $\kappa_2(0,\lambda) - \lambda \int\limits_0^1 \kappa_2(x,\lambda)\overline{\sigma_1(x)}dx = 0$, *4)* $\kappa_2'(0,\lambda) - \lambda \int\limits_0^1 \kappa_2(x,\lambda)\overline{\sigma_2(x)}dx = \Delta(\lambda)$.

The following theorem gives an integral representation of the resolvent of the operator $L_{\sigma_1\sigma_2}$.

Theorem 5. *The resolvent of $L_{\sigma_1\sigma_2}$ is given by*

$$(L_{\sigma_1\sigma_2} - \lambda I)^{-1} f(x) = (L_{00} - \lambda I)^{-1} f(x) + \frac{\kappa_1(x,\lambda)}{\Delta(\lambda)} < f(\cdot), M_1(\cdot,\overline{\lambda}) > + \\ + \frac{\kappa_2(x,\lambda)}{\Delta(\lambda)} < f(\cdot), M_2(\cdot,\overline{\lambda}) >,$$

where $M_\nu(t,\overline{\lambda}) = L_{00}^*(L_{00}^* - \overline{\lambda}I)^{-1}\sigma_\nu, \nu = 1, 2$. L_{00}^* is adjoint to L_{00}. Note that L_{00} is the operator corresponding to problem (4), (5) as $\sigma_1 \equiv \sigma_2 \equiv 0$.

The proof is straightforward. A complete proof of Theorem 5 can be found in [1].

3 Properties of some families of functions

We will need the following lemma.

Lemma 2. *For any complex numbers λ and μ, we have the matrix identity*

$$\begin{bmatrix} < \kappa_1(\cdot,\lambda), M_1(\cdot,\overline{\mu}) > & < \kappa_1(\cdot,\lambda), M_2(\cdot,\overline{\mu}) > \\ < \kappa_2(\cdot,\lambda), M_1(\cdot,\overline{\mu}) > & < \kappa_2(\cdot,\lambda), M_2(\cdot,\overline{\mu}) > \end{bmatrix} = \frac{\Delta(\mu) - \Delta(\lambda)}{\lambda - \mu} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \\ + \Delta(\mu) \begin{bmatrix} \kappa_1'(0,\lambda) & \kappa_1(0,\lambda) \\ \kappa_2'(0,\lambda) & \kappa_2(0,\lambda) \end{bmatrix} \frac{\begin{bmatrix} \kappa_1'(0,\mu) & \kappa_1(0,\mu) \\ \kappa_2'(0,\mu) & \kappa_2(0,\mu) \end{bmatrix}^{-1} - \begin{bmatrix} \kappa_1'(0,\lambda) & \kappa_1(0,\lambda) \\ \kappa_2'(0,\lambda) & \kappa_2(0,\lambda) \end{bmatrix}^{-1}}{\lambda - \mu}$$

We omit the proof.

Since the spectrum of $L_{\sigma_1\sigma_2}$ is discrete then there exists indefinitely increasing sequence $\{R_N\}$ of radii such that points of the spectrum of operator do not lie on appropriate the circles $|\lambda| = R_N$. Let $A_N = \{\lambda \in C : |\lambda| = R_N\}$ and $\sigma(L) = \{\lambda_1, \lambda_2, \dots\}$. Further on assume that R_N are chosen so that $dist(A_N, \sigma(L)) > \delta > 0$ for all N. Consider the subsequence of partial sums corresponding to the selected

circles

$$(S_N f)(x) = \sum_{|\lambda_s|<R_N} P_s f(x) = -\frac{1}{2\pi i} \oint\limits_{|\lambda|=R_N} (L-\lambda I)^{-1} f(x) d\lambda$$

for arbitrary function $f(\cdot) \in \mathbf{L}_2(0,1)$ and P_s is projector (see [4]).

Let λ_s be eigenvalue of $L_{\sigma_1\sigma_2}$ of algebraic multiplicity m_s. Then

$$\Delta(\lambda_s) = \Delta'(\lambda_s) = \cdots = \Delta^{(m_s-1)}(\lambda_s) = 0, \Delta^{(m_s)}(\lambda_s) \neq 0.$$

For sufficiently small $\delta > 0$ we define by P_s

$$P_s f(x) \equiv -\frac{1}{2\pi i} \oint\limits_{|\lambda-\lambda_s|=\delta} (L_{\sigma_1\sigma_2} - \lambda I)^{-1} f(\cdot) d\lambda =$$

$$= -\sum_{\nu=1}^{2} \frac{1}{(m_s-1)!} \lim_{\lambda\to\lambda_s} \frac{\partial^{m_s-1}}{\partial\lambda^{m_s-1}} \left(\kappa_\nu(\cdot,\lambda) \frac{(\lambda-\lambda_s)^{m_s}}{\Delta(\lambda)} < f, M_\nu(\cdot,\overline{\lambda}) > \right) =$$

$$= -\sum_{j=0}^{m_s-1} \sum_{\nu=1}^{2} \lim_{\lambda\to\lambda_s} \frac{1}{j!} \frac{\partial^j \kappa_\nu(\cdot,\lambda)}{\partial\lambda^j} < f(\cdot),\, h^{(\nu)}_{s,\,m_s-1-j}(\cdot) > .$$

Thus, the projector P_s is a finite-dimensional integral operator

$$P_s f(\cdot) = \sum_{\nu=1}^{2} \sum_{j=0}^{m_s-1} < f(\cdot),\, h^{(\nu)}_{s,\,m_s-1-j}(\cdot) > y^{(\nu)}_{s,\,j}(\cdot),$$

where $y^{(\nu)}_{s,\,j}(\cdot) = \frac{1}{j!} \lim_{\lambda\to\lambda_s} \frac{\partial^j}{\partial\lambda^j} \kappa_\nu(\cdot,\lambda)$ and

$$h^{(\nu)}_{s,\,m_s-1-j}(\cdot) = \frac{1}{(m_s-s-j)!} \lim_{\lambda\to\lambda_s} \frac{\partial^{m_s-1-j}}{\partial\overline{\lambda}^{m_s-1-j}} \left(\frac{(\overline{\lambda}-\overline{\lambda_s})^{m_s}}{\overline{\Delta(\lambda)}} \cdot M_\nu(\cdot,\,\overline{\lambda}) \right).$$

Note that for $\nu = 1,\, 2$ collection of functions $\{y^{(\nu)}_{s,j}(\cdot) : j = 0, \ldots,\, m_s - 1\}$ satisfies condition (5) as well as differential relations

$$-\frac{d^2}{dx^2} y^{(\nu)}_{s,\,j}(x) = \lambda_s \cdot y^{(\nu)}_{s,\,j}(x) + y^{(\nu)}_{s,\,j-1}(x), \quad j \geqslant 1,$$

$$-\frac{d^2}{dx^2} y^{(\nu)}_{s,\,0}(x) = \lambda_s \cdot y^{(\nu)}_{s,\,0}(x).$$

If $y^{(\nu)}_{s,0}(\cdot) \neq 0$ in $\mathbf{L}_2$ sense, then the indicated collection is a system of root functions of $L_{\sigma_1\sigma_2}$. Own main goal is to study completeness of $Y_{\sigma_1\sigma_2}$ in $\mathbf{L}_2(0,\,1)$. $Y_{\sigma_1\sigma_2}$ is called a system of root functions generated by the operator $L_{\sigma_1\sigma_2}$.

Denote $Q_N f(\cdot) = f(\cdot) - S_N f(\cdot)$ and then we call it the remainder of the function $f(\cdot)$. When $f(\cdot) = \kappa_\nu(\cdot, \lambda), \nu = 1,\,2$, it is possible to obtain an integral representation of the remainder. We have

Lemmma 3. *For any complex μ such that $|\mu| < R_N$ holds*

$$Q_N\kappa_\nu(x,\mu) = \frac{1}{2\pi i}\oint\limits_{|\lambda|=R_N} \frac{\Delta(\mu)}{\Delta(\lambda)(\lambda-\mu)}\kappa_\nu(x,\mu)d\lambda,\ for\ \nu = 1,\,2.$$

The proof is straightforward.

In the sequel we shall investigate the density of families of functions $\{\kappa_\nu(x,\mu), \forall\mu \in \mathbb{C}\}, \nu = 1, 2$, $\{\kappa_1(x,\mu), \kappa_2(x,\mu), \forall\mu \in \mathbb{C}\}$ in $\mathbf{L}_2(0,1)$. To answer this question, we shall study property of the family of functions $\{D(t,\cdot) : 0 \leqslant t \leqslant 1\}$, where $D(t,\cdot)$ is defined by (2).

Theorem 6. *If the family of functions $\{D(t,\cdot) : 0 \leqslant t \leqslant 1\}$ is dense in $\mathbf{L}_2(0,1)$, the system of functions $\{\kappa_1(\cdot,\mu), \forall\mu \in \mathbb{C}\}$ is dense in $\mathbf{L}_2(0,1)$. The converse statement is also true.*

Proof. Let $h(\cdot) \in \mathbf{L}_2(0,1)$ be orthogonal to the system of functions $\{\kappa_1(x,\mu), \forall\mu \in \mathbb{C}\}$. We transform the integral

$$\int\limits_0^1 \kappa_1(x,\mu)\overline{h(x)}dx = \begin{vmatrix} \int\limits_0^1 \cos\sqrt{\mu}x\overline{h(x)}dx & \int\limits_0^1 \frac{\sin\sqrt{\mu}x}{\sqrt{\mu}}\overline{h(x)}dx \\ -\mu\int\limits_0^1 \cos\sqrt{\mu}\tau\overline{\sigma_2(\tau)}d\tau & 1-\mu\int\limits_0^1 \frac{\sin\sqrt{\mu}\tau}{\sqrt{\mu}}\overline{\sigma_2(\tau)}d\tau \end{vmatrix} \tag{7}$$

Using the properties of the determinant and the trigonometric formula: $\sin\sqrt{\mu}(\tau - -x) = \sin\sqrt{\mu}\tau\cos\sqrt{\mu}x - \cos\sqrt{\mu}\tau\sin\sqrt{\mu}x$, (7) can be written in the form $\int\limits_0^1 \kappa_1(x,\mu)\overline{h(x)}dx = \int\limits_0^1 \cos\sqrt{\mu}x\overline{h(x)}dx - \mu\int\limits_0^1 \overline{h(x)}\left(\int\limits_0^1 \overline{\sigma_2(\tau)}\frac{\sin\sqrt{\mu}(\tau-x)}{\sqrt{\mu}}d\tau\right)dx$. Given the fact that $\cos\sqrt{\mu}x = 1 - \sqrt{\mu}\int\limits_0^x \sin\sqrt{\mu}t\,dt$, we obtain from the conditions of orthogonality

$$\int\limits_0^1 \kappa_1(x,\mu)\overline{h(x)}dx = \int\limits_0^1 \overline{h(x)}dx - \sqrt{\mu}\int\limits_0^1 \sin\sqrt{\mu}t\left(\int\limits_t^1 \overline{h(x)}dx\right)dt -$$

$$-\sqrt{\mu}\int_0^1 \sin\sqrt{\mu}t\left(\int_0^{1-t}\overline{h(x)}\cdot\overline{\sigma_2(t+x)}dx\right)dt+$$
$$+\sqrt{\mu}\int_0^1 \sin\sqrt{\mu}t\left(\int_t^1\overline{h(x)}\cdot\overline{\sigma_2(x-t)}dx\right)dt\equiv 0$$

for all complex μ.

If $\mu=0$, then $\int_0^1\overline{h(x)}dx=0$. Since $\{\sin\sqrt{\mu}t:\forall\mu\in\mathbb{C}\}$ is dense in $\mathbf{L}_2(0,1)$, for almost all $t\in(0,1)$ following relations hold

$$\int_t^1\overline{h(x)}(1-\overline{\sigma_2(x-t)})dx+\int_0^{1-t}\overline{h(x)}\cdot\overline{\sigma_2(t+x)}dx\equiv 0. \tag{8}$$

Formula (8) implies direct and inverse statement, of theorem 3.1.

The proof is complete.

Corollary 1. *For $\forall\mu\in\mathbb{C}$ we have the representation*

$$\int_0^1\kappa_1(x,\mu)\overline{h(x)}dx=\int_0^1\overline{h(x)}dx-\sqrt{\mu}\int_0^1\sin\sqrt{\mu}t\left(\int_0^1 D(x,t)\overline{h(x)}dx\right)dt,$$

where $D(t,\cdot)$ is defined by (2).

Example 1 (the not dense set of functions $\{D(t,\cdot):0<t<1\}$). If $\sigma_2\equiv\frac{1}{2}$, then family of functions with respect to x

$$D(t,x)=\begin{cases}\frac{1}{2}, & \text{for } 0\leqslant x\leqslant t;\\ 1, & \text{for } t<x\leqslant 1-t;\\ \frac{1}{2}, & \text{for } 1-t<x\leqslant 1\end{cases}\quad\text{for } 0<t<\frac{1}{2},$$

$$D(t,x)=\begin{cases}\frac{1}{2}, & \text{for } 0\leqslant x\leqslant 1-t;\\ 0, & \text{for } 1-t<x\leqslant t;\\ \frac{1}{2}, & \text{for } t<x\leqslant 1\end{cases}\quad\text{for } 1>t>\frac{1}{2}$$

is symmetric with respect to $\frac{1}{2}$, i.e $D(t,x)\equiv D(t,1-x)$. Therefore, any function having the property $h(x)\equiv -h(1-x)$ is orthogonal to the family $\{D(t,\cdot):0<t<1\}$. Consequently, the set of functions $\{D(t,\cdot):0<t<1\}$ is not dense in the functional space $L_2(0,1)$. Thus, the system of functions $\{\kappa_1(x,\mu),\forall\mu\in\mathbb{C}\}$ is

not dense in $L_2(0,1)$ with respect to Theorem 6. Note that in this case $\kappa_1(x,\lambda) = \frac{1}{2}\cos\sqrt{\lambda}x + \frac{1}{2}\cos\sqrt{\lambda}(1-x)$.

Also we study the density of a family of functions $\{\kappa_2(\cdot,\mu), \forall\mu \in \mathbb{C}\}$ in the function space $\mathbf{L}_2(0,1)$. To answer this question, we shall investigate property of the family of functions $\{B(t,\cdot) : 0 \leqslant t \leqslant 1\}$, where $B(t,\cdot)$ defined by (3).

Theorem 7. *If the family of functions $\{B(t,\cdot) : 0 \leqslant t \leqslant 1\}$ is dense in $\mathbf{L}_2(0,1)$, the system of functions $\{\kappa_2(x,\mu), \forall\mu \in \mathbb{C}\}$ is dense in $\mathbf{L}_2(0,1)$. The converse statement is also true.*

Proof of Theorem 7 is similar to that of Theorem 6.

Corollary 2. *For $\forall\mu \in \mathbb{C}$ has the representation*

$$\int_0^1 \kappa_2(x,\mu)\overline{h(x)}dx = \int_0^1 x\,\overline{h(x)}dx - \sqrt{\mu}\int_0^1 \sin\sqrt{\mu}t\left(\int_0^1 B(x,t)\overline{h(x)}dx\right)dt,$$

where $B(t,\cdot)$ is defined by (3).

Example 2 (the dense set of functions $\{B(t,\cdot) : 0 \leqslant t \leqslant 1\}$). If $\sigma_1 \equiv 0$, then the system of functions

$$B(t,x) = \begin{cases} 0, & 0 \leqslant x \leqslant 1-t; \\ 0, & 1-t \leqslant x \leqslant t; \\ x-t, & t \leqslant x \leqslant 1 \end{cases} \quad \text{for} 1 > t > \frac{1}{2},$$

$$B(t,x) = \begin{cases} 0, & 0 \leqslant x \leqslant t; \\ x-t, & t \leqslant x \leqslant 1-t; \\ x-t, & 1-t \leqslant x \leqslant 1 \end{cases} \quad \text{for } 0 < t < \frac{1}{2}.$$

Let the function $h(\cdot) \in L_2(0,1)$ be orthogonal to the family of functions $\{B(t,\cdot) : 0 \leqslant t \leqslant 1\}$. Then for $1 > t > 0$ we obtain the following integral relation:

$$\int_t^1 \overline{h(x)}(x-t)dx = 0 \tag{9}$$

Relation (9) can be differentiated with respect to t: $-\int_t^1 \overline{h(x)}dx = 0$. Let us find the second derivative of (9) with respect to t: $h(t) = 0$ for almost all $t \in (0,1)$. Thus, the family of function $\{B(t,x) : 1 \geqslant t \geqslant 0\}$ is dense in $L_2(0,1)$. That is, the system of functions $\{\kappa_2(x,\mu), \forall\mu \in \mathbb{C}\}$ is dense in $L_2(0,1)$ with respect to Theorem 7.

Note that in the case we have $\kappa_2(x,\,\lambda) = \frac{\sin\sqrt{\lambda}x}{\sqrt{\lambda}}$.

Combining Theorems 6 and 7, we have to following result.

Theorem 8 *System of functions* $\{B(t,\cdot), D(t,\cdot) : 0 \leqslant t \leqslant 1\}$ *is dense in function space* $\mathbf{L}_2(0,1)$ *if and only if of the system of functions* $\{\kappa_1(x,\mu),\, \kappa_2(x,\mu), \forall \mu \in \mathbb{C}\}$ *is dense in function space* $\mathbf{L}_2(0,1)$.

4 Approximation of principal solutions system of root function

In this section we investigate approximation properties of principal solutions system of root functions Y_{σ_1,σ_2} of the operator $L_{\sigma_1\sigma_2}$.

Lemma 4. *Let for some* $\varepsilon > 0$ *the functions* $\sigma_1(\cdot) \in \mathbf{W}_2^1([0,\,\varepsilon] \cup [1-\varepsilon,\,1]) \cap \mathbf{L}_2(0,\,1)$ *and* $\sigma_2(\cdot) \in \mathbf{W}_2^2([0,\,\varepsilon] \cup [1-\varepsilon,\,1]) \cap \mathbf{W}_2^1[0,\,1]$. *Assume that* $-\overline{\sigma_1(1)}\,\overline{\sigma_2(0)} - \overline{\sigma_1(0)}\,\overline{\sigma_2(1)} \neq 0$. *Then for any complex number* μ *we have*

$$\lim_{R_N\to\infty} \|\kappa_1(\cdot,\mu) - \sum_{|\lambda_s|<R_N} \sum_{\nu=1}^{2} \sum_{j=0}^{m_s-1} < \kappa_1(\cdot,\mu),\, h^{(\nu)}_{s,\,m_s-1-j}(\cdot) > y^{(\nu)}_{s,\,j}(\cdot)\| = 0,$$

where $y^{(\nu)}_{s,\,j}(\cdot) = \frac{1}{j!}\lim_{\mu\to\lambda_s} \frac{\partial^j}{\partial\mu^j}\kappa_\nu(\cdot,\mu), \nu = 1,\,2.$

Proof. Consider the norm in $\mathbf{L}_2(0,\,1)$ of remainder

$$\|Q_N^{(1)}(x,\mu)\| = \sup_{\|g\|=1} |\, < Q_N^{(1)}(\cdot,\mu); g(\cdot) > |\, =$$

$$= \sup_{\|g\|=1} \left| \frac{1}{2\pi i} \oint\limits_{|\lambda|=R_N} \frac{\Delta(\mu)}{\Delta(\lambda)(\lambda-\mu)} \left(\int_0^1 \kappa_1(x,\lambda)\overline{g(x)}dx \right) d\lambda \right| =$$

$$= \sup_{\|g\|=1} \left| \frac{1}{2\pi i} \oint\limits_{\substack{|\rho|=\sqrt{R_N} \\ Im\rho>0}} \frac{\Delta(\mu)e^{|Im\rho|}}{\Delta(\rho^2)(\rho^2-\mu)} \left(e^{-|Im\rho|} \int_0^1 \kappa_1(x,\rho^2)\overline{g(x)}dx \right) 2\rho d\rho \right|,$$

where $g(\cdot) \in \mathbf{L}_2(0,1)$ and $\Delta(\lambda)$ are entire functions. Applying the Cauchy-Schwarz inequality, we obtain

$$\|Q_N^{(1)}(x,\mu)\| \leqslant \frac{\Delta(\mu)|}{\pi} \left(\oint\limits_{\substack{|\rho|=\sqrt{R_N} \\ Im\rho>0}} \left| \frac{\rho e^{|Im\rho|}}{\Delta(\rho^2)(\rho^2-\mu)} \right|^2 |d\rho| \right)^{\frac{1}{2}} \cdot$$

$$\cdot \left(\oint\limits_{\substack{|\rho|=\sqrt{R_N} \\ Im\rho>0}} \left| e^{-|Im\rho|} \int\limits_0^1 \kappa_1(x,\rho^2)\overline{g(x)}dx \right|^2 |d\rho| \right)^{\frac{1}{2}} \tag{10}$$

Assume that for some $\varepsilon > 0$ functions $K(t) \in C^2[1-\varepsilon, 1]$ and $K'(1) \neq 0$. Let us estimate $|\Delta(\rho)|$ on circles $|\rho| = \sqrt{R_N}$. Consider $|\Delta(\rho^2)| = |1 - \rho^2 A + \rho^3 \int\limits_0^{1-\varepsilon} \sin\rho t\, K(t)dt + \rho^3 \int\limits_{1-\varepsilon}^1 \sin\rho t\, K(t)dt|$. We transform the right-hand side of the last relation by using a formula of integration by parts. As a result we have $|\Delta(\rho^2)| = = |1 - \rho^2 A + \rho^3 \int\limits_0^{1-\varepsilon} \sin\rho t\, K(t)dt - \rho^2 \cos\rho\, t\, K(t) r|_{t=1-\varepsilon}^1 + \rho^2 \int\limits_{1-\varepsilon}^1 \cos\rho\, t\, K'(t)dt|$. Note that $K(1) = 0$. By applying to last the integral integration by parts, we obtain

$$|\Delta(\rho^2)| = |1 - \rho^2 A + \rho^3 \int\limits_0^{1-\varepsilon} \sin\rho t\, K(t)dt + \rho^2 \cos\rho\,(1-\varepsilon)\, K(1-\varepsilon) +$$

$$+\rho \sin\rho\, K'(1) - \rho \sin\rho(1-\varepsilon)\, K'(1-\varepsilon) - \rho \int\limits_{1-\varepsilon}^1 \sin\, \rho\, t\, K''(t)dt| \geqslant |\rho \sin\rho\, K'(1)| -$$

$$-|1-\rho^2 A+\rho^3 \int\limits_0^{1-\varepsilon} \sin\rho t\, K(t)dt + \rho^2 \cos\rho\,(1-\varepsilon)\, K(1-\varepsilon) - \rho \sin\rho(1-\varepsilon)\, K'(1-\varepsilon) -$$

$$-\rho \int\limits_{1-\varepsilon}^1 \sin\, \rho\, t\, K''(t)dt| \geqslant c\,|\rho| e^{|Im\rho|} \tag{11}$$

as $|\rho| \to \infty$. We estimate

$$\Phi_1(\mu) = \oint\limits_{\substack{|\rho|=\sqrt{R_N} \\ Im\rho>0}} \left| \frac{\rho e^{|Im\rho|}}{\Delta(\rho^2)(\rho^2-\mu)} \right|^2 |d\rho| \leqslant \frac{\sqrt{R_N}}{c} \int\limits_0^\pi \frac{d\vartheta}{|R_N e^{2i\vartheta} - \mu|^2} \to 0 \tag{12}$$

as $R_N \to \infty$. Here we take into account the estimate of (11).

Now we introduce a function $F_1(\rho) = e^{-|Im\rho|}\int\limits_0^1 \kappa_1(x,\rho^2)g(x)\,dx$. We write the function $F_1(\rho)$ given by corollary 1 in the following form

$$F_1(\rho) = e^{-|Im\rho|}\left(\int\limits_0^1 \overline{g(x)} - \rho\int\limits_0^1 \sin\rho t\left(\int\limits_0^1 D(t,x)\overline{g(x)}dx\right)dt\right)$$

Let $\sigma_2(\cdot)\in \mathbf{W}_2^1[0,1]$. Taking into account the formula for $D(t,x)$, the last expression can be transformed to

$$F_1(\rho) = e^{-|Im\rho|}\Big(\int\limits_0^1 \overline{g(x)}\,dx - \rho\int\limits_0^1 \sin\rho\, t\Big(\int\limits_t^1 \overline{g(x)}\,dx + \int\limits_0^{1-t} \overline{g(x)}\,\overline{\sigma_2(t+x)}\,dx - \int\limits_t^1 \overline{g(x)}\,\overline{\sigma_2(x-t)}dx\Big)dt\Big).$$

Applying the formula for integration by parts to the last integral, we obtain

$$F_1(\rho) = e^{-|Im\rho|}\int\limits_0^1 \cos\rho t\left(\overline{g(t)} + \overline{g(1-t)}\,\overline{\sigma_2(1)} + \overline{g(t)}\,\overline{\sigma_2(0)}\right)dt + e^{-|Im\rho|}\int\limits_0^1 \cos\rho t\left(\int\limits_t^1 \overline{g(x)}\,\overline{\sigma_2'(x-t)}\,dx - \int\limits_0^{1-t} \overline{g(x)}\,\overline{\sigma_2'(x+t)}dx\right)dt$$

As $Im\rho > 0$ the function $F_1(\rho)$ belongs to the space $\mathbf{H}_+^2$. Then with respect to Lemma 1.2.5 of [5] it follows that

$$|\rho|\int\limits_0^\pi |F_1(|\rho|e^{i\vartheta})|^2 d\vartheta \leqslant c_1\left(\sup_{Im\rho>0}\int\limits_{-\infty}^{\infty}|F(Re\rho + iIm\rho)|^2 dRe\rho\right)^{\frac{1}{2}}, \qquad (13)$$

where c_1 is independent of $|\rho|$ and $g(x)$.

Using (12) and (13) inequality (10) implies the limit relation

$$\lim_{R_N\to\infty}\|Q_N^{(1)}(\cdot,\,\mu)\| = 0.$$

Lemma 4 is proved.

Similarly, for $f(\cdot) = \kappa_2(\cdot, \mu)$ we prove the following lemma.

Lemma 5. *Let some $\varepsilon > 0$ the functions $\sigma_1(\cdot) \in \mathbf{W}_2^1([0, \varepsilon] \cup [1-\varepsilon, 1]) \cap \mathbf{L}_2(0, 1)$ and $\sigma_2(\cdot) \in \mathbf{W}_2^2([0, \varepsilon] \cup [1-\varepsilon, 1]) \cap \mathbf{W}_2^1[0, 1]$. Assume that $-\overline{\sigma_1(1)} + \overline{\sigma_1(1)}\,\overline{\sigma_2(0)} - -\overline{\sigma_1(0)}\,\overline{\sigma_2(1)} \neq 0$. Then for any complex number μ we have*

$$\lim_{R_N \to \infty} \|\kappa_2(\cdot, \mu) - \sum_{|\lambda_s| < R_N} \sum_{\nu=1}^{2} \sum_{j=0}^{m_s-1} < \kappa_2(\cdot, \mu),\, h^{(\nu)}_{s,\, m_s-1-j}(\cdot) > y^{(\nu)}_{s,\, j}(\cdot)\| = 0.$$

5 Completeness of the systems of root functions

In this section we prove of the theorems on the completeness systems of root functions in $\mathbf{L}_2(0, 1)$. These theorems are a direct corollary of results of section 4.

Proof of Theorem 1. The fact that the family of functions $\{D(t, \cdot) : 0 \leqslant t \leqslant 1\}$ is dense in $\mathbf{L}_2(0, 1)$ with respect to Theorem 6 follows from the density of the systems of functions $\{\kappa_1(\cdot, \mu), \forall \mu \in \mathbb{C}\}$ in $\mathbf{L}_2(0, 1)$. By Lemma 4, the elements of the system $\{\kappa_1(\cdot, \mu), \forall \mu \in \mathbb{C}\}$ can be arbitrarily closely approximated by a system of root functions Y_{σ_1, σ_2}. Therefore the system of functions $Y^{(1)}_{\sigma_1, \sigma_2}$ is complete in the function space $\mathbf{L}_2(0, 1)$. Theorem 1 is proved.

Proofs of Theorem 2 and 3 are similar to the proof of Theorem 1.

References

1. B.E. Kanguzhin and D.B. Nurakhmetov *Estimates of resolvents for well-posed differential operators on the interval.* Vestnik KarGU. Seria Mat. no.4, 2010. Pp. 60-72.
2. M.A. Naimark *Linear Differential operators,* Moscow, Nauka, 1969, pages 528
3. M.O. Otelbaev and A.N. Shynybekov *Well-posed problems of Bitsadze-Samarskii type.* Dokl. Akad. Nauk SSSR **265,** no.4, 1982. Pp. 815-919
4. F. Riesz and B.Sz.-Nagy *Functional analysis,* Blackie and Son Limited., London and Glasgow, 1956, pages 478
5. A.M. Sedletskii *Biorthogonal expansions of functions in exponential series on intervals of the real axis.* Uspekhi Mat. Nauk, T.37, V.5(227), 1982. Pp.51-95 [English translation: Russian Mathematical Surveys, 37:5, 1982, 57-108]

B. E. Kanguzhin
Al-Farabi Kazakh National University, Kazakhstan, 050012, Almaty, 71 Al-Farabi ave.,
+7 777 817 82 90, kanbalta@mail.ru

D. B. Nurakhmetov
Al-Farabi Kazakh National University, Kazakhstan, 050012, Almaty, 71 Al-Farabi ave.,
+7 705 870 45 79, dauletkaznu@gmail.com

GROUP SYMMETRY BIFURCATION PROBLEM WITH SCHMIDT SPECTRUM IN THE LINEARIZATION

B. V. Loginov, I. V. Konopleva

Key words: Stationary bifurcation problems, E. Schmidt spectrum, group symmetry, G-invariant implicit operator theorem, variational type branching equations and branching equations in the root-subspaces

AMS Mathematics Subject Classification: 47J07, 58E09

Abstract. With the aim of applications in electromagnetic oscillations theory G-invariant implicit operator theorem and theorem about reduction of variational type branching equations and branching equations in the root-subspaces on the number of equations are proved for bifurcational problems with E.Schmidt spectrum in the linearization.

1 Introduction

In cycle of works at the origin of XX century on linear and nonlinear integral equations E.Schmidt had introduced [1] eigenvalues λ_k of the operator $B : H \to H$ in a Hilbert space H taking into account their multiplicities and eigenelements $\{u_k\}_1^\infty, \{v_k\}_1^\infty$ satisfying the relations $Bu_k = \lambda_k v_k, \quad B^* v_k = \lambda_k u_k$. This allows to extend Hilbert-Schmidt theory on nonsymmetric completely continuous operators in abstract separable Hilbert spaces [2, 3]. Some physical applications of E.Schmidt spectral problems are indicated in [4], in [5] the development of pseudoperturbiteration method for generalized E. Schmidt spectral problems is given, in [6] the Fredholm property was proved for the problem on electromagnetic oscillation eigenfrequencies in resonators without loss [7], which is typical E. Schmidt spectral problem

$$\begin{array}{ll} \mathrm{rot}\overrightarrow{E} = i\omega\mu\overrightarrow{H}, & \mathrm{rot}\overrightarrow{H} = -i\omega\varepsilon\overrightarrow{E}, \\ \mathrm{div}\overrightarrow{E} = 0, & \mathrm{div}\overrightarrow{H} = 0 \end{array} \quad \text{in } V \subset \mathbb{R}^3$$

with boundary conditions ($S = \partial V$ is ideal conductor) $\left[\overrightarrow{n}, \overrightarrow{E}\right]|_S = 0$, $(\overrightarrow{n}, \overrightarrow{H})_S = 0$, ε and μ are dielectric and magnetic permeabilities of the medium filling the resonator.

The work is supported by the SPPIR Goscontract No. 1122 Ministry of Education and Science of Russia and enter to project No. 12-01-00270 of RFBR..

This work is inspired by nonlinear problems on resonators filling by nonlinear medium, in particular, by nonmagnetic medium when the dielectric permeability ε in the layer $0 < x < h$ is determined by Kerr law $\varepsilon = \varepsilon_2 + a|\overrightarrow{E}|^2$, $a > 0$, $\varepsilon_2 > \max(\varepsilon_1, \varepsilon_3)$ - is constant component of the dielectric permeability of ε, $\varepsilon_1 \geqslant \varepsilon_0$, $(\varepsilon_3 \geqslant \varepsilon_0)$ are permeabilities of semispaces $x < 0$ $(x > h)$, ε_0 – is vacuum dielectric permeability.

The indicated applications state the problem about solutions bifurcation and their stability for nonlinear equations under group symmetry conditions in the linearization of which generalized E.Schmidt eigenvalue problems are contained. The terminilogy, notitions and some results of [8]– [10] are used.

2 Branching Equation in the Root Subspaces and Group Symmetry Inheritance Theorem

In Banach spaces E_1 and E_2, $E_1 \subset E_2 \subset H$ (H- Hilbert space) the nonlinear system is considered

$$\begin{aligned}
&F_1(x,y,\lambda) = 0, \quad F_2(x,y,\lambda) = 0,\ F_k(x_0,y_0,\lambda) \equiv 0,\ \lambda = \lambda_0 + \varepsilon;\ k = 1,2;\\
&F_{1x}'(x_0,y_0,\lambda) = B_0 + B(\varepsilon), \quad -F_{1y}'(x_0,y_0,\lambda) = A_0 + A(\varepsilon),\\
&-F_{2x}'(x_0,y_0,\lambda) = A_0^* + A^*(\varepsilon), \quad F_{2y}'(x_0,y_0,\lambda) = B_0^* + B^*(\varepsilon),
\end{aligned} \tag{2.1}$$

A_0, B_0 are densely defined closed operators, $\overline{D}_A = E_1$, $D_A = D_{A(\varepsilon)} \subset D_{A_0}$, $\overline{D}_B = E_1, D_B = D_{B(\varepsilon)} \subset D_{B_0}$.

The system (2.1) allows the local presentation

$$\begin{aligned}
&B_0X - \lambda_0A_0Y = A(\varepsilon)Y - B(\varepsilon)X + R_1(x_0,y_0,X,Y,\varepsilon),\\
&B_0^*Y - \lambda_0A_0^*X = A^*(\varepsilon)X - B^*(\varepsilon)Y + R_2(x_0,y_0,X,Y,\varepsilon),\\
&R_j(x_0,y_0,0,0,\varepsilon) \equiv 0, R_{jX}'(x_0,y_0,0,0,\varepsilon) \equiv 0, R_{jY}'(x_0,y_0,0,0,\varepsilon) \equiv 0,\\
&X = x - x_0,\ \ Y = y - y_0,\ \ j = 1,2,
\end{aligned} \tag{2.2}$$

Let n-multiple E.Schmidt eigenvalue λ_0 be Fredholm point of the corresponding to (2.2) matrix-operator $(\mathbf{B} - \lambda_0\mathbf{A}) = \begin{pmatrix} -\lambda_0A_0^* & B_0^* \\ B_0 & -\lambda_0A_0 \end{pmatrix}$ in the direct sum $\mathcal{H}$ of two Hilbert spaces H with eigenelements $\Phi_k^{(1)} = (u_k^{(1)}, v_k^{(1)})^T$ and $\Psi_k^{(1)} = (\widetilde{u}_k^{(1)}, \widetilde{v}_k^{(1)})^T$, i.e. $(\mathbf{B}^* - \lambda_0\mathbf{A}^*)\Psi_k^{(1)} = \begin{pmatrix} -\lambda_0A_0 & B_0^* \\ B_0 & -\lambda_0A_0^* \end{pmatrix} \begin{pmatrix} \widetilde{u}_k^{(1)} \\ \widetilde{v}_k^{(1)} \end{pmatrix} = 0$. Setting

$\mathbf{A}(\varepsilon) = \begin{pmatrix} A^*(\varepsilon) & -B^*(\varepsilon) \\ -B(\varepsilon) & A(\varepsilon) \end{pmatrix}$ write the system (2.1) in the form

$$(\mathbf{B} - \lambda_0\mathbf{A}) \begin{pmatrix} X \\ Y \end{pmatrix} = \mathbf{A}(\varepsilon) \begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} R_2(x_0, y_0, X, Y, \varepsilon) \\ R_1(x_0, y_0, X, Y, \varepsilon) \end{pmatrix} = \\ = \mathbf{A}(\varepsilon) \begin{pmatrix} X \\ Y \end{pmatrix} + \mathbf{R}\left(x_0, y_0, \begin{pmatrix} X \\ Y \end{pmatrix}, \varepsilon\right). \quad (2.3)$$

Here it is supposed that the vectorial nonlinear operator can be presented in the form of sufficiently smooth on $(X, Y)^T$ operator $\mathbf{R}$.

Definition 1. The elements $\Phi_k^{(s)} = (u_k^{(s)}, v_k^{(s)})^T, s = 1, \ldots, p_k, k = 1, \ldots, n$ form the complete canonical generalized Jordan set (GJS$\equiv \mathbf{A}(\varepsilon)$-JS), if

$$\begin{gathered} (\mathbf{B} - \lambda_0\mathbf{A})\Phi_k^{(s)} = \sum_{j=1}^{s-1} \mathbf{A}_j \Phi_k^{(s-1)}, \\ \mathbf{A}(\varepsilon) = \mathbf{A}_1\varepsilon + \mathbf{A}_2\varepsilon^2 + \ldots, \langle \Phi_k^{(s)}, \Gamma_l^{(1)} \rangle_{\mathcal{H}} = 0, \\ D_p = \det \left[\sum_{j=1}^{p_k} \langle \mathbf{A}_j \Phi_k^{(p_k+1-j)}, \Psi_l^{(1)} \rangle_{\mathcal{H}} \right] \neq 0, \ \Psi_l^{(1)} = (\widetilde{u}_l^{(1)}, \widetilde{v}_l^{(1)})^T, \\ k, l = 1, \ldots, n; \ \ s = 2, \ldots, p_k; \end{gathered} \quad (2.4)$$

This JS bicanonical, if GJS of elements $\{\Psi_l^{(1)}\}_1^n$ for the conjugate operator $(\mathbf{B}^* - \lambda_0\mathbf{A}^*) - \mathbf{A}^*(\varepsilon)$ is also canonical, and three-canonical if in addition

$$\begin{gathered} \langle \Phi_i^{(j)}, \Gamma_k^{(l)} \rangle_{\mathcal{H}} = \delta_{ik}\delta_{jl}, \Gamma_k^{(l)} = \sum_{s=1}^{p_k+1-l} \mathbf{A}_s^* \Psi_k^{(p_k+2-l-s)}; \\ \langle Z_i^{(j)}, \Psi_k^{(l)} \rangle_{\mathcal{H}} = \delta_{ik}\delta_{jl}, \quad Z_i^{(j)} = \sum_{s=1}^{p_k+1-j} \mathbf{A}_s \Phi_i^{(p_k+2-j-s)}, \\ \Phi = \Phi(x_0, y_0) = (\Phi_1^{(1)}, \ldots, \Phi_1^{(p_1)}, \ldots, \Phi_n^{(1)}, \ldots, \Phi_n^{(p_n)}) \end{gathered} \quad (2.5)$$

$\Phi_i^{(j)} = \Phi_i^{(j)}(x_0, y_0)$, the vectors $\Gamma = \Gamma(x_0, y_0), \Psi = \Psi(x_0, y_0), Z = Z(x_0, y_0)$ are analogously determined, $K = p_1 + \ldots + p_n$ is the root-number.

Lemma 1. Let the Fredholm operator-function $(\mathbf{B} - \lambda_0\mathbf{A}) - \mathbf{A}(\varepsilon)$ has complete three-canonical GJS. Then the projectors are defined

$$\mathbf{P} = \mathbf{P}(x_0, y_0) = \sum_{i=1}^{n} \sum_{j=1}^{p_i} \langle \cdot, \Gamma_i^{(j)} \rangle_{\mathcal{H}} \Phi_i^{(j)} = \langle \cdot, \Gamma \rangle_{\mathcal{H}} \Phi : \mathcal{H} \to \mathcal{H}^K =$$

$$= K(\mathbf{B}-\lambda_0\mathbf{A};\mathbf{A}(\varepsilon)) = \mathrm{span}\{\Phi_i^{(j)}(x_0,y_0)\}, \quad (2.6)$$

$$\mathbf{Q}=\mathbf{Q}(x_0,y_0)=\sum_{i=1}^{n}\sum_{j=1}^{p_i}\langle\cdot,\Psi_i^{(j)}\rangle_{\mathcal{H}}Z_i^{(j)}=\langle\cdot,\Psi\rangle_{\mathcal{H}}Z:\mathcal{H}\to\mathcal{H}_K=\mathrm{span}\{Z_i^{(j)}(x_0,y_0)\}$$

allowing to expand the Hilbert space $\mathcal{H}$ into direct sums in the point (x_0,y_0)

$$\mathcal{H}=\mathcal{H}^K\dot{+}\mathcal{H}^{\infty-K},\ \mathcal{H}=\mathcal{H}_K\dot{+}\mathcal{H}_{\infty-K}. \quad (2.7)$$

Projectors $\mathbf{P}=\mathbf{P}(x_0,y_0)$ and $\mathbf{Q}=\mathbf{Q}(x_0,y_0)$ are intertwining the operator $\mathbf{B}-\lambda_0\mathbf{A}$: $(\mathbf{B}-\lambda_0\mathbf{A})\mathbf{P}=\mathbf{Q}(\mathbf{B}-\lambda_0\mathbf{A})$ on $D_{\mathbf{B}-\lambda_0\mathbf{A}}$, $(\mathbf{B}-\lambda_0\mathbf{A})\Phi=\mathfrak{A}_0 Z$, $(\mathbf{B}^*-\lambda_0\mathbf{A}^*)\Psi=\mathfrak{A}_0\Gamma$, $\mathfrak{A}_0=\mathrm{diag}(B_1,\dots,B_n)$, $B_i-(p_i\times p_i)$–matrix with units on subsidiary subdiagonal and zeros on other places. Operator $(\mathbf{B}-\lambda_0\mathbf{A}):D_{\mathbf{B}-\lambda_0\mathbf{A}}\cap\mathcal{H}^{\infty-K}(x_0,y_0)\to\mathcal{H}_{\infty-K}(x_0,y_0)$ is isomorphism.

Lemma 2. For linear by ε operator-function $(\mathbf{B}-\lambda_0\mathbf{A})-\varepsilon\mathbf{A}_1$ three-canonical GJS always exists and intertwining properties can be added by the following ones

$$\mathbf{A}_1\mathbf{P}=\mathbf{Q}\mathbf{A}_1 \quad \text{on} \quad D_{\mathbf{A}_1},\ \ \mathbf{A}_1\Phi=\mathfrak{A}_1\Phi,\ \ \mathbf{A}_1^*\Psi=\mathfrak{A}_1\Gamma,$$

where $\mathfrak{A}_1=\mathrm{diag}(B^1,\dots,B^n)$- cell-diagonal matrix, $B^i-(p_i\times p_i)$-matrices with units on subsidiary diagonal and zeros on other places. Here the operators $(\mathbf{B}-\lambda_0\mathbf{A}):D_{\mathbf{B}-\lambda_0\mathbf{A}}\cap\mathcal{H}^{\infty-K}\to\mathcal{H}_{\infty-K}$ and $\mathbf{A}_1:\mathcal{H}^K\to\mathcal{H}_K$ are isomorphisms.

Theorem 1. *Let to the bifurcation point $(x_0,y_0,0)$ be correspond of complete three-canonical GJS of the operator-function $(\mathbf{B}-\lambda_0\mathbf{A})-\mathbf{A}(\varepsilon)$. The problem of the finding of small solutions to the system (2.1) (or (2.3)) in a neighborhood of the point $(x_0,y_0,0)$ is equivalent to the finding of small solutions to the A.Lyapounov branching equation in the root subspace (BEqR)*

$$\begin{aligned}f(x_0,y_0,\mathbf{v}(x_0,y_0,\xi),\varepsilon)=f(x_0,y_0,\mathbf{v}(x_0,y_0,\xi)+\mathbf{u}(x_0,y_0,\mathbf{v}(x_0,y_0,\xi),\varepsilon))\equiv\\ \equiv\mathfrak{A}_0\xi-\langle\mathbf{A}(\varepsilon)(\mathbf{v}(x_0,y_0,\xi)+\mathbf{u}(x_0,y_0,\mathbf{v}(x_0,y_0,\xi),\varepsilon))+\\ +\mathbf{R}(x_0,y_0,\mathbf{v}(x_0,y_0,\xi)+\mathbf{u}(x_0,y_0,\mathbf{v}(x_0,y_0,\xi),\varepsilon),\Psi(x_0,y_0)\rangle_{\mathcal{H}}=0\end{aligned} \quad (2.8)$$

and E.Schmidt BEqR

$$\begin{aligned}\mathbf{t}_{s1}(x_0,y_0,\mathbf{v}(x_0,y_0,\xi),\varepsilon)\equiv-\sum_{j=1}^{n}\xi_{j1}\langle(\mathbf{I}-\mathbf{A}(\varepsilon)\mathbf{\Gamma_0})^{-1}\mathbf{A}(\varepsilon)\Phi_j^{(1)}(x_0,y_0),\Psi_s^{(1)}(x_0,y_0)\rangle_{\mathcal{H}}-\\ -\langle(\mathbf{I}-\mathbf{A}(\varepsilon)\mathbf{\Gamma_0})^{-1}\mathbf{R}(x_0,y_0,\mathbf{v}(x_0,y_0,\xi)+\mathbf{w}(x_0,y_0,\mathbf{v}(x_0,y_0,\xi),\varepsilon)),\Psi_s^{(1)}(x_0)\rangle_{\mathcal{H}}=0\end{aligned} \quad (2.9)$$

$$\mathbf{t}_{s\sigma}(x_0,y_0,\mathbf{v}(x_0,y_0,\xi),\varepsilon)\equiv$$
$$\equiv\xi_{s\sigma}-\sum_{j=1}^{n}\xi_{j1}\langle(\mathbf{I}-\mathbf{A}(\varepsilon)\mathbf{\Gamma_0})^{-1}\mathbf{A}(\varepsilon)\Phi_j^{(1)}(x_0,y_0),\Psi_s^{(p_s+2-\sigma)}(x_0,y_0)\rangle_{\mathcal{H}}-$$
$$-\langle(\mathbf{I}-\mathbf{A}(\varepsilon)\mathbf{\Gamma_0})^{-1}\mathbf{R}(x_0,y_0,\mathbf{v}(x_0,y_0,\xi)+\mathbf{w}(x_0,y_0,\mathbf{v}(x_0,y_0,\xi),\varepsilon)),\Psi_s^{(p_s+2-\sigma)}(x_0)\rangle_{\mathcal{H}}=0.$$

Here

$$\mathbf{\Gamma}_0=\mathbf{\Gamma}_0(x_0,y_0)=\widetilde{(\mathbf{B}-\lambda_0\mathbf{A})}^{-1},$$
$$\widetilde{(\mathbf{B}-\lambda_0\mathbf{A})}=\mathbf{B}-\lambda_0\mathbf{A}+\sum_{i=1}^{n}\langle\cdot,\Gamma_j^{(1)}(x_0,y_0)\rangle_{\mathcal{H}}Z_j^{(1)}(x_0,y_0)$$

is E.Schmidt regularizator [8].

Everywhere below it is supposed that the system (2.1) allows the group symmetry

$$K_gF_j(x,y,\lambda)=F_j(L_gx,L_gy,\lambda),\quad k=1,2$$

where $L_g(K_g)$ is the representation of the group G in $E_1(E_2)$, expanded on H. Here the bifurcation point (x_0,y_0,λ_0) moves along its trajectory $(L_gx_0,L_gy_0,\lambda_0)$.

At the presence of continuous group symmetry Lie group $G_l=G_l(a)$, $a=(a_1,\ldots,a_l)$ are its essential parameters, is supposed to be l-dimensional differentiable manifold, satisfying the conditions [11], [12]: $(\mathbf{c_1})$ the representation $a\mapsto\begin{pmatrix}L_{g(a)}x_0\\L_{g(a)}y_0\end{pmatrix}$, acting from a neighborhood of $G_l(a)$ unit element into the space $\mathcal{H}$ belongs to the class C^1, so that $\mathcal{X}(x_0,y_0)\in\mathcal{H}$ for all generating operators $\mathcal{X}$ in tangent to $L_{g(a)}$ manifold $T^l_{g(a)}$; $(\mathbf{c_2})$ stationary subgroup G_s of element $(x_0,y_0)\in\mathcal{H}$ determines the representation $L(G_s)$ of local Lie group $G_s\subset G_l$, $s<l$, with s-dimensional subalgebra $T^s_{g(a)}$ of generators. This means that elements $\mathcal{X}_k(x_0,y_0),\mathcal{X}_k\in T^l_{g(a)}$ from the zero-subspace of the matrix-operator $(\mathbf{B}-\lambda_0\mathbf{A})$ form $\kappa=(l-s)$-dimensional subspace and the bases in it and in the algebra $T^l_{g(a)}$ can be ordered so that $\mathcal{X}_k(x_0,y_0)=\Phi_k(x_0,y_0),1\leqslant k\leqslant\kappa$, and $\mathcal{X}_j(x_0,y_0)=(0,0)$ for $j\geqslant\kappa+1$; $(\mathbf{c_3})$ dense embeddings $E_1\subset E_2\subset H$ in Hilbert space H are supposed with estimates $\|u\|_H\leqslant\alpha_2\|u\|_{E_2}\leqslant\alpha_1\|u\|_{E_1}$, and $\mathcal{X}:\mathcal{H}\to\mathcal{H}$ is bounded in $L(\mathcal{H})$ topology. Further the auxiliary constructions are introduced:

$\mathbf{1^0}$. $K_g[\mathbf{B}(x_0,y_0)-\lambda_0\mathbf{A}(x_0,y_0)]=[\mathbf{B}(L_gx_0,L_gy_0)-\lambda_0\mathbf{A}(L_gx_0,L_gy_0)]L_g$.
$\mathbf{2^0}$. $K_g\mathbf{A}(\varepsilon)=K_g\mathbf{A}(x_0,y_0,\varepsilon)=[\mathbf{A}(L_gx_0,L_gy_0,\varepsilon)]L_g$;
$\mathbf{3^0}$. $K_g\mathbf{R}(x_0,y_0,x-x_0,y-y_0,\varepsilon)=\mathbf{R}(L_gx_0,L_gy_0,L_g(x-x_0),L_g(y-y_0),\varepsilon)$.

$$\begin{gathered}\Phi_i(L_g x_0, L_g y_0) = L_g \Phi_i(x_0, y_0) = (L_g u_k^{(1)}(x_0, y_0), L_g v_k^{(1)}(x_0, y_0))^T, \\ \Gamma_j(L_g x_0, L_g y_0) = {L^*}_g^{-1} \Gamma_j, \quad \Gamma_j = \Gamma_j(x_0, y_0), \quad i, j = 1, \ldots, n\end{gathered} \tag{2.10}$$

and for the range of operators F'_{k_x}, F'_{k_y} one has

$$\begin{gathered}\mathcal{R}(F'_{k_x}(L_g x_0, L_g y_0, \lambda_0)) = \mathcal{R}(K_g F'_{k_x}(x_0, y_0, \lambda_0) L_g^{-1}) = K_g \mathcal{R}(F'_{k_x}(x_0, y_0, \lambda_0)), \\ \mathcal{R}(F'_{k_y}(L_g x_0, L_g y_0, \lambda_0)) = \mathcal{R}(K_g F'_{k_y}(x_0, y_0, \lambda_0) L_g^{-1}) = K_g \mathcal{R}(F'_{k_y}(x_0, y_0, \lambda_0)).\end{gathered}$$

Then for the kernel of adjoint operator

$$\begin{gathered}N^*(\mathbf{B} - \lambda_0 \mathbf{A}) = \operatorname{span}\{\Psi_k^{(1)}\}_1^n = \operatorname{span}\{\widetilde{u}_k^{(1)}(x_0, y_0), \widetilde{u}_k^{(1)}(x_0, y_0)\}_1^n \Longrightarrow \\ N^*(\mathbf{B}(L_g x_0, L_g y_0) - \lambda_0 \mathbf{A}(L_g x_0, L_g y_0)) = \operatorname{span}\{{K^*}^{-1}{}_g \Psi_1^{(1)}, \ldots {K^*}_g^{-1} \Psi_n^{(1)}\},\end{gathered} \tag{2.11}$$

Analogously to [13], [14] it can be proved that the elements of ordered by increasing lengths GJChs of the operator-function $(\mathbf{B} - \lambda_0 \mathbf{A}) - \mathbf{A}(\varepsilon)$ and biorthogonal to them systems are transformating according to formulae

$$\begin{gathered}\Phi_k^{(s)}(L_g x_0, L_g y_0) = L_g \Phi_k^{(s)}(x_0, y_0) = (L_g u_k^{(s)}(x_0, y_0), L_g v_k^{(s)}(x_0, y_0))^T, \\ \Psi_k^{(s)}(L_g x_0, L_g y_0) = {K^*}_g^{-1} \Psi_k^{(s)}(x_0, y_0) = ({K^*}_g^{-1} u_k^{(s)}(x_0, y_0), {K^*}_g^{-1} v_k^{(s)}(x_0, y_0))^T, \\ \Gamma_k^{(s)}(L_g x_0, L_g y_0) = {L^*}_g^{-1} \Gamma_k^{(s)}(x_0, y_0), \quad Z_k^{(s)}(L_g x_0, L_g y_0) = K_g Z_k^{(s)}(x_0, y_0).\end{gathered} \tag{2.12}$$

Lemma 3. Projectors (2.6) satisfy the intertwining conditions

$$\begin{gathered}\mathbf{P}(L_g x_0, L_g y_0) = L_g \mathbf{P}(x_0, y_0) L_g^{-1} \text{ or } L_g \mathbf{P}(x_0, y_0) = \mathbf{P}(L_g x_0, L_g y_0) L_g \\ \mathbf{Q}(L_g x_0, L_g y_0) = K_g \mathbf{Q}(x_0, y_0) K_g^{-1} \text{ or } K_g \mathbf{Q}(x_0, y_0) = \mathbf{Q}(K_g x_0, K_g y_0) K_g\end{gathered} \tag{2.13}$$

with the expansions of the space $\mathcal{H}$ in the bifurcation point $(L_g x_0, L_g y_0, \lambda_0)$

$$\begin{gathered}\mathcal{H} = \mathcal{H}^K(L_g x_0, L_g y_0) \dot{+} \mathcal{H}^{\infty - K}(L_g x_0, L_g y_0) = \\ = \mathcal{H}_K(L_g x_0, L_g y_0) \dot{+} \mathcal{H}_{\infty - K}(L_g x_0, L_g y_0),\end{gathered} \tag{2.14}$$

and relations

$$\begin{gathered}\mathcal{H}^K(L_g x_0, L_g y_0) = L_g \mathcal{H}^K(x_0, y_0), \mathcal{H}^{\infty - K}(L_g x_0, L_g y_0) = L_g \mathcal{H}^{\infty - K}(x_0, y_0), \\ \mathcal{H}_K(L_g x_0, L_g y_0) = L_g \mathcal{H}_K(x_0, y_0), \mathcal{H}_{\infty - K}(L_g x_0, L_g y_0) = L_g \mathcal{H}_{\infty - K}(x_0, y_0).\end{gathered} \tag{2.15}$$

Theorem 4. (Group symmetry inheritance theorem.) *At three-canonical GJS existence for the operator-function* $(\mathbf{B} - \lambda_0 \mathbf{A}) - \mathbf{A}(\varepsilon)$ *A.Lyapounov* (2.8) *and*

E.Schmidt (2.9) *BEqR inherit the group symmetry of the system* (2.1)

$$f(L_g x_0, L_g y_0, L_g \mathbf{v}(x_0, y_0, \xi), \varepsilon) = f(L_g x_0, L_g y_0, \mathbf{v}(L_g x_0, L_g y_0, \xi), \varepsilon) = \\ = K_g f(x_0, y_0, \mathbf{v}(x_0, y_0, \xi), \varepsilon), \quad (2.16)$$

$$t(L_g x_0, L_g y_0, L_g \mathbf{v}(x_0, y_0, \xi), \varepsilon) = t(L_g x_0, L_g y_0, \mathbf{v}(L_g x_0, L_g y_0, \xi), \varepsilon) = \\ = L_g t(x_0, y_0, \mathbf{v}(x_0, y_0, \xi), \varepsilon). \quad (2.17)$$

3 Basic Results

Theorem 5. (Implicit operator theorem.) *Let at group symmetry conditions for the system* (2.1) *the requirements* $(\mathbf{c_1})$–$(\mathbf{c_3})$ *be fulfilled, in the condition* $(\mathbf{c_2})$ $\kappa = n$ *and* $G_s, s < l$ *is the normal divisor of* G_l *with the relevant ideal* $T^s_{g(a)}$ *of generators, and for the operator-function* $(\mathbf{B} - \lambda_0 \mathbf{A}) - \mathbf{A}(\varepsilon)$ *in the Fredholm point* λ_0 *there correspond the complete three-canonical GJS to elements of* $N(\mathbf{B} - \lambda_0 \mathbf{A})$. *Then there exists the continuous function* $\mathbf{v}(x_0, y_0, \xi, \varepsilon) = \mathbf{v}(x_0, y_0, \xi) + \mathbf{u}(x_0, y_0, \mathbf{v}(x_0, y_0, \xi), \varepsilon)$: $T^n_{g(a)} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \times (-\delta, \delta) \to \mathcal{H}$, *invariant with respect to the factor-group* $G_\kappa = G_n = G_l / G_s$ *on* $T^n_{g(a)} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$, *such that*

$$F_1(x_0, y_0, v_1(x_0, y_0, \xi, \varepsilon), v_2(x_0, y_0, \xi, \varepsilon)) = 0, \\ F_2(x_0, y_0, v_1(x_0, y_0, \xi, \varepsilon), v_2(x_0, y_0, \xi, \varepsilon)) = 0$$

at $\mathbf{v}(x_0, y_0, \xi) \in T^n_{g(a)} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$, $|\varepsilon| < \delta$.

Corollary. Theorem 5 is true for semisimple bifurcation points, i.e. at the absence of GJS. Then here we have BEq.

Definition 2. BEqR (2.8) (resp. (2.9)) is the BEqR of potential type if in a neighborhood of the point $(x_0, y_0; 0)$ for the vector $\mathbf{f}(x, y, \mathbf{v}(x, y, \xi), \varepsilon) = (f_{11}, \ldots, f_{1p_1}, \ldots, f_{n1}, \ldots, f_{np_n})$ the equality

$$\mathbf{f}(x, y, \mathbf{v}(x, y, \xi), \varepsilon) = d \cdot \operatorname{grad}_{x,y} U(x, y, \xi, \varepsilon), \quad (3.1)$$

is satisfied, where d is an invertible operator. Then the functional $U(x, y, \xi, \varepsilon)$ is the potential of BEqR (2.8) (resp. (2.9)) and the operator $\mathbf{f}$ (resp. $\mathbf{t}$) is pseudogradient of the functional U.

Theorem 6. *(BEqR reduction.) Let in suppositions* $(\mathbf{c_1})$–$(\mathbf{c_3})$ *A.Lyapounov BEqR (E.Schmidt BEqR) is potential type one, its potential* $U(x, y, \xi, \varepsilon)$ *is invariant of the representation* $L_{g(a)}$ *of the group* $G_l(a)$ *and belongs to the class* C^2 *in some neighborhood of the bifurcation point* $(x_0, y_0; 0)$*,* s *– the dimension of stationary subgroup of the element* (x_0, y_0) *and* $\kappa = l - s > 0$*. Then:*

1. *if* $\kappa = n$*, then for all* $(\xi(\varepsilon), \varepsilon)$ *or* $(\mathbf{v}(x_0, y_0, \xi(\varepsilon), \varepsilon))$ *from some neighborhood of zero in* $\mathbb{R}^{K+1}$ *BEqR (8) (or (9)) is identically fulfilled. i.e. the situation of the Theorem 3 arises;*
2. *if* $\kappa < n$ *and* $n \geqslant 2$*, then the partial reduction has place: the first* $K_\kappa = p_1 + \ldots + p_\kappa$ *equations are linear combinations of the others* $p_{\kappa+1} + \ldots + p_n$*.*

Corollary. In the case of invariant kernel the BEqR reduction realizes on complete Jordan chains with the aid of complete system of functionally independent invariants of the group G_l action in $\Xi^K = \{\xi_{11}, \ldots, \xi_{1p_1}, \ldots, \xi_{n1}, \ldots, \xi_{1n_p}\}$.

References

1. Schmidt E. *Zur Theorie linearen und nichtlinearen Integralgleichungen. Teilen 1-3. Mathematische Annallen.* Bd. 63-65, 1905-1908.
2. Goursat E. *Course d'Analyse Mathematique.* Paris, Gautier-Villars, 1933.
3. Mogilevskii Sch. *On representation of completely continuous operators acting in abstract separable Hilbert space.* Izv. VUZ. Mathematics. 3(4), 1958. Pp.183-186.
4. Loginov B.V. Pospeev V.E. *On eigenvalues and eigenfunctions of perturbed operators.* Izv. Acad. sci. UzbekSSR, phys.-math. ser. 6. 1967. Pp. 29-35.
5. Loginov B.V., Makeeva O.V. *Pseudoperturbation method in generalized eigenvalue problems.* Doklady Mathematics, **77**, 2, 2009. Pp. 160-163.
6. Loginov B.V., Makeeva O.V. *E. Schmidt spectral problem on oscillation eigenfrequencies in resonators without loss.* Proc. Mid.-Volga Math. Soc. **9**, 1. 2007. Pp. 31-38.
7. Il'insky A.S., Slepyan G.A. *Oscillations and waves in electrodynamical systems with losses.* Moscow: Moscow State Univ. Publ., 1988.
8. M. M. Vainberg, V. A. Trenogin *Theory of Branching of Solutions of Non-Linear Equations.* Nauka, Moscow, 1969; Wolter Noordorf, Leiden. 1974.
9. Loginov B.V. *Theory of Branching of Solutions of Nonlinear Equations under Group Symmetry Invariance.* Tashkent, Fan, Akad.Nauk UzSSR, 1985.
10. Sidorov N. et.al. *Lyapunov-Schmidt Methods in Nonlinear Analysis and Applications, MIA.* **550**. Kluwer Acad. Publ. Dordrecht. 2002.
11. N. I. Makarenko. *On solutions branching of invariant variational problems.* Dokl. Math. **348**, 1996, Pp. 369–371.

12. N. I. Makarenko. *Symmetry and cosymmetry of variational problems in the waves theory.* Proc. Int. Workshop "Applications of Symmetry and Cosymmetry to the Theory of Bifurcations and Phase Transitions", Sochi, Russia 2001 (Rostov. Univ., 2001), Pp. 109–120.
13. Konopleva I.V. et.al. *Symmetry and potentiality of branching equations in the root subspaces in implicitely given stationary and dynamical bfurcation problems.* Izv. VUZ, Severo-Kaukaz. Region. Nat. Sci. Special Volume. Actual Problems of Mathematical Hydrodynamics. 2009, Pp. 115-124.
14. Loginov B.V. et. al. *Implicit operator theorems under group symmetry conditions.* Doklady RAN. Mathematics. **440**, 1, 2011, Pp. 15-20.

B. V. Loginov
Contacts for the first author: Ulyanovsk State Technical University, Russia, 432027, Ulyanovsk, Severny Venets, 32, (8422)431547, <loginov@ulstu.ru>

I. V. Konopleva
Contacts for the second author: Ulyanovsk State Technical University, Russia, 432027, Ulyanovsk, Severny Venets, 32, 89603900407, <i.konopleva@ulstu.ru>

MULTIRESOLUTION ANALYSIS ON PRODUCT OF P-ADIC NUMBER FIELDS

S. F. Lukomskii

Key words: zero-dimensional groups, multiresolution analysis, dilation operator, wavelet bases

AMS Mathematics Subject Classification: 42C40, 42A25

Abstract. We describe a dilation operator on a product of locally compact zero-dimensional groups $(G, \dot{+})$, and build the corresponding multiresolution analysis. Bibliography: 10 titles.

1 Introduction

In recent years there has been a considerable interest in the problem of constructing wavelet bases on Vilrnkin groups G_V and p-adic number fields $\mathbb{Q}_p$. Yu. Farkov [2,3] pointed out a method for constructing compactly supported orthogonal wavelets on a locally compact Vilenkin group G_V with a constant generating sequence, and derived necessary and sufficient conditions for a solution of the refinement equation to generate a multiresolution analysis (MRA in the sequel) of $L_2(G_V)$.

A good deal of studies was devoted to the construction of an MRA on the field $\mathbb{Q}_p$ of all p-adic numbers. A.Khrennikov, V.Shelkovich, and M.Skopina [5]– [6] considered the refinement equation

$$\varphi(x) = \sum_{j=0}^{p^s-1} \beta_j \varphi\left(p^{-1}x \dot{-} \frac{r}{p^s}\right)$$

introduced the concept of a p-adic MRA with orthogonal refinable function, and described a general scheme for their creation.

The problem of constructing multidimensional MRA is moor difficult. In [10] multidimensional 2-adic orthogonal wavelet bases for $L_2(\mathbb{Q}_2^d)$ was constructed by means of the tensor product of one-dimensional MRA. Let us mention also that in [7]

This research was carried out with the financial support of the Programme for Support of Leading Scientific Schools of the President of the Russian Federation (grant no. -4383.2010.1) and the Russian Foundation for Basic Research (grant no. 10-01-00097).

E.J. King and M.A. Skopina constructed a wavelet basis in $L_2(\mathbb{Q}_2^2)$. For constructing this bases they used dilation operator with quincunx matrix $\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$.

But the general method of constructing of p-adic multidimensional MRA and multidimensional wavelets missing. In the present paper we examine the problem of construction of MRA on product $\mathbb{Q}_p^d$. The product $\mathbb{Q}_p^d$ is not a field of q-adic numbers, therefore it is impossible to solve this problem within framework p-adic analysis. But the product $\mathbb{Q}_p^d$ is a zero-dimensional locally compact abelian group with condition $pg_n = g_{n+d}$ and conversely: any zero-dimensional locally compact abelian group with condition $pg_n = g_{n+d}$ is the product $\mathbb{Q}_p^d$ of groups of all p-adic numbers [8]. Using this fact we will to construct MRA in $L_2(G^d)$, where G – is an arbitrary zero-dimensional group. We find a condition for matrix A_d under which the operator $\mathcal{A}_d\mathbf{x} = A_dX$ is a dilation operator in G^d. Using dilation operator we construct MRA for $L_2(\mathbb{G}^d)$. Taking $G = \mathbb{Q}_p$ we get MRA for $L_2(\mathbb{Q}_p^d)$.

2 Locally compact zero-dimensional groups, topology and characters

We proceed to give basic notions and facts in the analysis on zero-dimensional groups. A more detailed account may be found in [1].

A topological group in which the connected component of 0 is 0 is usually referred to as a *zero-dimensional group*. If a separable locally compact group $(\mathfrak{G}, \dot{+})$ is zero-dimensional, then topology on it can be generated by means of a descending sequence of subgroups. The converse statement holds for all topological groups (see [1, Ch. 1, § 3]). So, for a locally compact group, we are going to say "zero-dimensional group" instead of saying "a group with topology generated by a sequence subgroups".

Let $(\mathfrak{G}, \dot{+})$ be a locally compact zero-dimensional Abelian group with topology generated by a countable system of open subgroups

$$\cdots \supset \mathfrak{G}_{-n} \supset \cdots \supset \mathfrak{G}_{-1} \supset \mathfrak{G}_0 \supset \mathfrak{G}_1 \supset \cdots \supset \mathfrak{G}_n \supset \cdots$$

where

$$\bigcup_{n=-\infty}^{+\infty} \mathfrak{G}_n = \mathfrak{G}, \qquad \bigcap_{n=-\infty}^{+\infty} \mathfrak{G}_n = \{0\}$$

(0 is the null element in the group $\mathfrak{G}$). Given any fixed $N \in \mathbb{Z}$, the subgroup $\mathfrak{G}_N$ is a compact Abelian group with respect to the same operation $\dot{+}$ under the topology generated by the system of subgroups $\mathfrak{G}_N \supset \mathfrak{G}_{N+1} \supset \cdots \supset \mathfrak{G}_n \supset \cdots$.

As each subgroup $\mathfrak{G}_n$ is compact, it follows that each quotient group $\mathfrak{G}_n/\mathfrak{G}_{n+1}$ is finite (say, of order p_n). We may always assume that all p_n are prime numbers, for in fact, by Sylow's theorem (see [4]), the chain of subgroups can be refined so that the quotient groups $\mathfrak{G}_n/\mathfrak{G}_{n+1}$ will be of prime order. We will name such chain as *basic chain*. In this case, a base of the topology is formed by all possible cosets $\mathfrak{G}_n \dot{+} g$, $g \in \mathfrak{G}$.

We further define the numbers $(\mathfrak{m}_n)_{n=-\infty}^{+\infty}$ as follows:

$$\mathfrak{m}_0 = 1, \quad \mathfrak{m}_{n+1} = \mathfrak{m}_n \cdot p_n.$$

The collection of all such cosets $\mathfrak{G}_n \dot{+} g$, $n \in \mathbb{Z}$, along with the empty set form the semiring $\mathcal{K}$. On each coset $\mathfrak{G}_n \dot{+} g$ we define the measure μ by $\mu(\mathfrak{G}_n \dot{+} g) = \mu\mathfrak{G}_n = 1/\mathfrak{m}_n$. So, if $n \in \mathbb{Z}$ and $p_n = p$, we have $\mu\mathfrak{G}_n \cdot \mu\mathfrak{G}_{-n} = 1$. The measure μ can be extended from the semiring $\mathcal{K}$ onto the σ-algebra (for example, by using Carathéodory's extension). This gives the translation invariant measure μ, which agrees on the Borel sets with the Haar measure on $\mathfrak{G}$. Further, let $\int\limits_{\mathfrak{G}} f(x)\, d\mu(x)$ be the absolutely convergent integral of the measure μ.

Given an $n \in \mathbb{Z}$, consider an element $g_n \in \mathfrak{G}_n \setminus \mathfrak{G}_{n+1}$ and fix it. Then any $x \in \mathfrak{G}$ has a unique representation of the form

$$x = \sum_{n=-\infty}^{+\infty} a_n g_n, \qquad a_n = \overline{0, p_n - 1}, \tag{2.1}$$

the sum (2.1) containing finite number of terms with negative subscripts; that is,

$$x = \sum_{n=N}^{+\infty} a_n g_n, \quad a_n = \overline{0, p_n - 1}, \quad a_N \neq 0.$$

We will name system $(g_n)_{n\in\mathbb{Z}}$ as *a basic system*.

Let X be the collection of the characters of a group $(\mathfrak{G}, \dot{+})$; it is a group with respect to multiplication too. Also let $\mathfrak{G}_n^{\perp} = \{\chi \in X : \forall\, x \in \mathfrak{G}_n \ , \chi(x) = 1\}$ be the annihilator of the group $\mathfrak{G}_n$. Each annihilator $\mathfrak{G}_n^{\perp}$ is a group with respect to multiplication, and the subgroups $\mathfrak{G}_n^{\perp}$ form an increasing sequence

$$\cdots \subset \mathfrak{G}_{-n}^{\perp} \subset \cdots \subset \mathfrak{G}_0^{\perp} \subset \mathfrak{G}_1^{\perp} \subset \cdots \subset \mathfrak{G}_n^{\perp} \subset \cdots \tag{2.2}$$

with

$$\bigcup_{n=-\infty}^{+\infty} \mathfrak{G}_n^{\perp} = X \quad \text{and} \quad \bigcap_{n=-\infty}^{+\infty} \mathfrak{G}_n^{\perp} = \{1\},$$

the quotient group $\mathfrak{G}_{n+1}^{\perp}/\mathfrak{G}_n^{\perp}$ having order p_n. The group of characters X may be equipped with the topology using the chain of subgroups (2.2), the family of the cosets $\mathfrak{G}_n^{\perp}\cdot\chi$, $\chi\in X$, being taken as a base of the topology. The collection of such cosets, along with the empty set, forms the semiring $\mathcal{X}$. Given a coset $\mathfrak{G}_n^{\perp}\cdot\chi$, we define a measure ν on it by $\nu(\mathfrak{G}_n^{\perp}\cdot\chi)=\nu(\mathfrak{G}_n^{\perp})=\mathfrak{m}_n$ (so that always $\mu(\mathfrak{G}_n)\nu(\mathfrak{G}_n^{\perp})=1$). The measure ν can be extended onto the σ-algebra of measurable sets in the standard way (for example, using Caratheodory's extension theorem). One then forms the absolutely convergent integral $\int\limits_X F(\chi)\,d\nu(\chi)$ of this measure.

The value $\chi(g)$ of the character χ at an element $g\in\mathfrak{G}$ will be denoted by (χ,g). The Fourier transform $\widehat{f}$ of an $f\in L_2(\mathfrak{G})$ is defined as follows

$$\widehat{f}(\chi)=\int\limits_{\mathfrak{G}} f(x)\overline{(\chi,x)}\,d\mu(x)=\lim_{n\to+\infty}\int\limits_{\mathfrak{G}_{-n}} f(x)\overline{(\chi,x)}\,d\mu(x),$$

the limit being in the norm of $L_2(X)$.

3 Dilation operator

In this section we will consider a locally-compact zero-dimensional Abelian groups $(\mathfrak{G},\dot{+})$ with the basic chain of subgroups

$$\cdots\supset\mathfrak{G}_{-n}\supset\cdots\supset\mathfrak{G}_{-1}\supset\mathfrak{G}_0\supset\mathfrak{G}_1\supset\cdots\supset\mathfrak{G}_n\supset\cdots$$

We will assume that $(\mathfrak{G}_n/\mathfrak{G}_{n+1})^{\sharp}=p$ for any $n\in\mathbb{Z}$. As regards the operation $\dot{+}$, we assume additionally that

$$pg_n=\gamma_1 g_{n+1}\dot{+}\gamma_2 g_{n+2}\dot{+}\cdots\dot{+}\gamma_\tau g_{n+\tau};\tag{3.1}$$

here, $\gamma_1,\gamma_2,\ldots,\gamma_\tau=\overline{0,p-1}$ are fixed numbers. We set

$$H_n=\left\{q\in\mathfrak{G}:\ q=\sum_{j=N}^{n-1}a_j g_j,\ N\in\mathbb{Z},\ a_j=\overline{0,p-1}\right\}.$$

If $\mathfrak{G}$ is a Vilenkin group, then H_n is a group. This is not so in the general case (for example, if $\mathfrak{G}$ is the group of all p-adic numbers).

Definition 1. *We define the mapping $\mathcal{A}\colon\mathfrak{G}\to\mathfrak{G}$ by $\mathcal{A}x:=\sum_{n=-\infty}^{+\infty}a_n g_{n-1}$, where $x=\sum_{n=-\infty}^{+\infty}a_n g_n\in\mathfrak{G}$. As any element $x\in\mathfrak{G}$ can be uniquely expanded*

as $x = \sum a_n g_n$, the mapping $A\colon \mathfrak{G} \to \mathfrak{G}$ is one-to-one onto. The mapping $\mathcal{A}$ is called a dilation operator if $\mathcal{A}(x\dot{+}y) = \mathcal{A}x\dot{+}\mathcal{A}y$ for all $x, y \in \mathfrak{G}$.

We note that if $\mathfrak{G}$ is a Vilenkin group ($p \cdot g_n = 0$) or is the group of all p-adic numbers ($p \cdot g_n = g_{n+1}$), then A is an additive operator and hence a dilation operator. Moreover, the operator A is additive if the condition (3.1) is satisfied. It is also clear that $\mathcal{A}\mathfrak{G}_n = \mathfrak{G}_{n-1}, \mathcal{A}^{-1}\mathfrak{G}_n = \mathfrak{G}_{n+1}$.

Our main objective is to construct MRA for $L_2(\mathfrak{G})$ and $L_2(\mathfrak{G}^d)$. For this we will use a multiresolution analysis on the group $\mathfrak{G}$ as follows [9].

Definition 2. *A family of closed subspaces V_n, $n \in \mathbb{Z}$, is said to be a multiresolution analysis of $L_2(\mathfrak{G})$ if the following axioms are satisfied:*

1) $V_n \subset V_{n+1}$;
2) $\overline{\bigcup_{n\in\mathbb{Z}} V_n} = L_2(\mathfrak{G})$ *and* $\bigcap_{n\in\mathbb{Z}} V_n = \{0\}$;
3) $f(x) \in V_n \iff f(\mathcal{A}x) \in V_{n+1}$ *($\mathcal{A}$ is a dilation operator);*
4) $f(x) \in V_0 \implies f(x\dot{-}h) \in V_0$ *for all $h \in H_0$ (H_0 is a set of shifts);*
5) *there exists a function $\varphi \in L_2(\mathfrak{G})$ such that the system $(\varphi(x\dot{-}h))_{h\in H_0}$ is an orthonormal basis for V_0.*

A function φ occurring in Axiom 5) is called a refinable function.

Next we will follow the conventional approach. Let $\varphi(x) \in L_2(\mathfrak{G})$, and suppose that $(\varphi(x\dot{-}h))_{h\in H_0}$ is an orthonormal system in $L_2(\mathfrak{G})$. With the function φ and the dilation operator A, we define the linear subspaces $L_j = (\operatorname{span}\varphi(\mathcal{A}^j x\dot{-}h))_{h\in H_0}$ and closed subspaces $V_j = \overline{L_j}$. If the subspaces V_j form an MRA, then the function φ is said to *generate* an MRA in $L_2(\mathfrak{G})$. We will look up a function $\varphi \in L_2(\mathfrak{G})$, which generates an MRA in $L_2(\mathfrak{G})$, as a solution the refinement equation

$$\varphi(x) = \sum_{h\in H} c_h \varphi(Ax\dot{-}h), \tag{3.2}$$

where $H \subset H_0$ is a finite set.

Next theorem was prowed in [9]

Theorem 1. *Let $\varphi \in L_2(\mathfrak{G})$ be a solution of the equation* (3.2). *Suppose that $|\widehat{\varphi}(\chi)| = \mathbf{1}_{\mathfrak{G}_0^\perp}(\chi)$. Then φ generates an MRA in $L_2(\mathfrak{G})$.*

4 The dilation operator on the product of zero-dimensional group

Let $(\mathfrak{G}, \dot{+})$ be a locally compact zero-dimensional abelian group with a basic chain

$$\cdots \subset \mathfrak{G}_n \subset \cdots \subset \mathfrak{G}_1 \subset \mathfrak{G}_0 \subset \mathfrak{G}_{-1} \subset \cdots \subset \mathfrak{G}_{-n} \subset \ldots, \quad (\mathfrak{G}_n/\mathfrak{G}_{n+1})^\sharp = p,$$

$(g_n)_{n\in\mathbb{Z}}$ – be a basic system, i.e. $g_n \in \mathfrak{G}_n \setminus \mathfrak{G}_{n+1}$.

We denote by $G = \mathfrak{G}^d = \mathfrak{G} \times \mathfrak{G} \times \cdots \times \mathfrak{G}$ the direct sum of d copies of group $\mathfrak{G}$. The base of neighborhood of zero in $\mathfrak{G}^d$ consist of all products $\mathfrak{G}_{n_1} \times \mathfrak{G}_{n_2} \times \cdots \times \mathfrak{G}_{n_d}$. We can take the chain of d-dimensional cubes $\mathfrak{G}_n \times \mathfrak{G}_n \times \cdots \times \mathfrak{G}_n$ as a base of neighborhood of zero in $\mathfrak{G}^d$. We note that the chain

$$\cdots \subset \mathfrak{G}_n^d \subset \cdots \subset \mathfrak{G}_1^d \subset \mathfrak{G}_0^d \subset \mathfrak{G}_{-1}^d \subset \cdots \subset \mathfrak{G}_{-n}^d \subset \ldots \tag{4.1}$$

is not a basic chain, since $(\mathfrak{G}_n^d/\mathfrak{G}_{n+1}^d)^\sharp = p^d$ is not a prime number. Denote $\mathfrak{G}_n^d$ as G_{nd}. Using the Silovs theorem we will refine the chain (4.1) to obtain a basic chain G_n with condition $(G_n/G_{n+1})^\sharp = p$. Let $\mathbf{g}_n = (g^{(n_1)}, g^{(n_2)}, \ldots, g^{(n_d)}) \in (G_n/G_{n+1})^d$ be a basic system in $G = \mathfrak{G}^d$. Using this basic system we can define the dilation operator $\mathcal{A}_d$ as

$$\mathcal{A}_d\mathbf{x} = \sum_{n\in\mathbb{Z}} a_n \mathbf{g}_{n-1} \;\text{ if }\; \mathbf{x} = \sum_{\mathrm{n}\in\mathbb{Z}} \mathrm{a_n}\mathbf{g_n}.$$

If the operator $\mathcal{A}_d$ is additive, then $\mathcal{A}_d$ is a dilation operator. We wont to write the operator $\mathcal{A}_d$ in the form

$$(\mathcal{A}_d\mathbf{x})^T = A\mathbf{x}^T,$$

where A is $d \times d$ matrix, $\mathbf{x}^T$ is a vector-column $(x^{(1)}, x^{(2)}, \ldots, x^{(d)})^T$. Let Z_p be a residue-class field on modulo p i.e. $Z_p = \{0, 1, \ldots, p-1\}$ with operations $m + n = (m+n) \bmod \mathrm{p}$, $m \cdot n = \underbrace{m + m + \cdots + m}_{n}$. By A denote a $d \times d$ matrix with elements $a_{i,j} \in Z_p$ $(i, j = \overline{1, p})$, by $E_{\mathcal{A}}$ – the matrix

$$E_{\mathcal{A}} = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 & \mathcal{A} \\ 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 & 0 \end{pmatrix}$$

where $\mathcal{A}$ is one dimensional dilation operator in group $(\mathfrak{G}, \dot{+})$. Using this matrix A we define vectors $\mathbf{g}_{(n+1)d-l}$ and subgroups $G_{(n+1)d-l}$ as

$$\mathbf{g}_{(n+1)d-l} = (a_{1,l}g_n, a_{2,l}g_n, \ldots, a_{d,l}g_n), \;\; l = \overline{1, d},$$

$$G_{(n+1)d-l} = \bigsqcup_{j=0}^{p-1} (G_{(n+1)d-l+1} \dot{+} j\mathbf{g}_{(n+1)d-l}).$$

Lemma 1. *Let A – be a nonsingular matrix over the field Z_p. Define the operator the operator $\mathcal{A}_d$ as*

$$\mathcal{A}_d\left(\sum_{n,l} a_{(n+1)d-l}\mathbf{g}_{(n+1)d-l}\right) = \sum_{n,l} a_{(n+1)d-l}\mathbf{g}_{(n+1)d-l-1}.$$

Then
1) $G_{(n+1)d} \subset G_{(n+1)d-1} \subset \cdots \subset G_{(n+1)d-d+1} \subset G_{nd}$,
2) sets $G_{(n+1)d-l}$ are subgroups for any $l = \overline{1,d}$,
3) $(G_{(n+1)d-l}/G_{(n+1)d-l+1})^{\sharp} = p$,
4) $\mathbf{g}_{(n+1)d-l} \in G_{(n+1)d-l} \setminus G_{(n+1)d-l+1}$,
5) the operator $\mathcal{A}_d$ is additive,
6) if $\mathfrak{X} = (x^{(1)}, x^{(2)}, \ldots, x^{(d)})^T = \mathbf{x}^T$, then $(\mathcal{A}_d(\mathbf{x}))^T = AE_{\mathcal{A}}A^{-1}\mathfrak{X}$.

From lemma 1 follow

Theorem 2. *Let $\mathcal{A}$ – be an one dimensional dilation operator in $(\mathfrak{G}, \dot{+})$. The equality*

$$(\mathcal{A}_d(\mathbf{x}))^T = AE_{\mathcal{A}}A^{-1}\mathbf{x}^T$$

define a dilation operator in $\mathfrak{G}^d$ for any nonsingular matrix $A = A_{d\times d}$ over the residue-class field Z_p.

Theorem 3. *Let p be a prime number and let*

$$E_p = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 & \frac{1}{p} \\ 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 & 0 \end{pmatrix}$$

be a $d \times d$ matrix. Then the equality

$$(\mathcal{A}_d(\mathbf{x}))^T = AE_pA^{-1}\mathbf{x}^T$$

define a dilation operator in $\mathbb{Q}_p^d$.

Let as denote

$$H_0^{(S)} = \{\mathbf{x} = a_{-1}\mathbf{g}_{-1}\dot{+}a_{-2}\mathbf{g}_{-2}\dot{+}\ldots\dot{+}a_{-S}\mathbf{g}_{-S}\}, \quad H_0 = \bigcup_{S\in\mathbb{N}} H_0^{(S)}.$$

H_0 is a set of shifts in group $G = \mathfrak{G}^d$.

Theorem 4. *Let φ be a solution of the refinement equation*

$$\varphi(\mathbf{x}) = \sum_{\mathbf{h}\in H_0^{(S)}} \beta_{\mathbf{h}}\varphi(\mathcal{A}_d\mathbf{x}\dot{-}\mathbf{h}).$$

Suppose $|\hat{\varphi}(\chi)| = \mathbf{1}_{(\mathfrak{G}_0^d)^\perp}(\chi)$. Then φ generate a MRA in $L_2(\mathfrak{G}^d)$. If $\mathfrak{G} = \mathbb{Q}_p$ we get MRA in $L_2(\mathbb{Q}_p^d)$.

References

1. G.N.Agaev, N.Ya.Vilenkin, G.M.Dzhafarli, A.I.Rubinstein. *Multiplicative systems of functions and harmonic analysis on zero-dimensional groups.* Baku, Elm, 1981.(in Russian)
2. Yu. A. Farkov. *Orthogonal wavelets with compact support on locally compact Abelian groups.*Izv. RAN, Ser. Matem., 69:3, 2005, 193–220. in Russian. (engl. trasl. Yu. A. Farkov *Orthogonal wavelets with compact support on locally compact Abelian groups.* Izv. Math., 69:3, 2005, 623–650)
3. Yu. A. Farkov. *Orthogonal wavelets on direct products of cyclic groups* Matem. Zametki, 82:6, 2007, 934–962. in Russian (engl. transl. Yu. A. Farkov. *Orthogonal wavelets on direct products of cyclic groups.* Math. Notes, 82:5–6, 2007)
4. M. I. Kargapolov, Ju. I. Merzljakov. *Fundamentals of the theory of groups.* Grad. Texts in Math. v.62, New York–Berlin, Springer-Verlag, 1979.
5. A. Yu. Khrennikov, V. M. Shelkovich, M. Skopina. *p-Adic orthogonal Wavelet Bases.* P-adic numbers, Ultrametric Analysis and Applications, 1:2, 2009,145-156.
6. A. Yu. Khrennikov, V. M. Shelkovich, M. Skopina *p-Adic refinable functions and MRA-based wavelets.*J.Approx.Theory. 161:1, 2009,226-238.
7. Emily J.King and Maria A.Skopina.*Quincunx Multiresolutiom Analysis for $L_2(\mathbb{Q}_2^2)$, P-adic Numbers, Ultrametric Analysis and Applications, 2010, 2:3, 222–231.*
8. S. F. Lukomskii. *On Haar system on product of groupsof p-adic integers* Matem. Zametki, 90:4, 2011, 541–557. in Russian (engl. transl. Math. Notes, 90:4,2011,517-532)
9. S.F.Lukomskii. *Multiresolution analysis on zero-dimensional Abelian groups and wavelets bases.*Matem. Sbornik, 201:5, 2010, 41-64, in Russian. (engl. transl. S.F.Lukomskii. *Multiresolution analysis on zero-dimensional Abelian groups and wavelets bases.*Sb. Math.,201:5,2010, 669–691)
10. V. M. Shelkovich, M. Skopina. *p-adic Haar multiresolution analysis and pseudo-differential operators.*J. Fourier Anal. Appl., 15:3, 2009.

S. F. Lukomskii
Department of Mathematics, Saratov University, Saratov, Russia Astrakhanskaya 83, Saratov 410012, Russia, (8452)515-537, LukomskiiSF@info.sgu.ru

APPLICATIONS OF THE FUNCTION $MUP_S(X)$

V. A. Makarichev

Key words: generalized Taylor series, basic functions, Kolmogorov width, function with compact support, nonstationary wavelet system, functional differential equation

AMS Mathematics Subject Classification: 41A30

Abstract. We consider applications of the solution with a compact support of some functional differential equation to the theory of generalized Taylor series, approximation theory and wavelet theory.

1 Introduction

In this paper we assume that $s = 2, 3, 4, \ldots$ and $\alpha \in (1, 2s)$.

Consider the function $mup_s(x) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} e^{-itx} \cdot \prod\limits_{k=1}^{\infty} \frac{\sin^2\left(\frac{st}{(2s)^k}\right)}{s^2 \cdot \frac{t}{(2s)^k} \cdot \sin\frac{t}{(2s)^k}} dt$. This function, which is a solution with a compact support of the equation

$$y'(x) = 2 \cdot \sum_{k=1}^{s} \left(y(2s \cdot x + 2s - 2k + 1) - y(2s \cdot x - 2k + 1)\right),$$

was used by V.A. Rvachev and G.A. Starets in [1] to construct basic functions of the generalized Taylor series for some classes of differentiable functions.

By $H_{s,\alpha}$ denote a class of functions $f \in C^{\infty}_{[-1,1]}$ such that

$$\left\| f^{(n)}(x) \right\|_{C_{[-1,1]}} \leqslant c(f) \cdot \alpha^n \cdot 2^n \cdot (2s)^{\frac{n(n-1)}{2}}, \ n = 0, 1, 2, \ldots.$$

Let $N_{s,0} = \{-1, 0, 1\}$ and $x_{s,0,k} = k$ for $k \in N_{s,0}$;

$$N_{s,n} = \left\{-s \cdot (2s)^{n-1}, -s \cdot (2s)^{n-1} + s, \ldots, s \cdot (2s)^{n-1}\right\} \text{ and}$$

$$x_{s,n,k} = \frac{k}{s \cdot (2s)^{n-1}} \text{ for } k \in N_{s,n} \text{ and } n \neq 0;$$

$$I_{s,n} = \left\{1, 2, \ldots, (2s)^{n+1} - 1\right\} \text{ for } n = 0, 1, 2, \ldots;$$

$$D_{s,n} = \{p \in I_n : \ p \neq 0 \ (\mathrm{mod}\, s)\} \text{ and}$$

$$x^*_{s,n,p} = -1 + \frac{p}{s \cdot (2s)^n} \text{ for } p \in D_{s,n} \text{ and } n = 0, 1, 2, \ldots;$$

$$\triangle_h^2(f; x) = f(x+h) - 2 \cdot f(x) + f(x-h).$$

It was shown in [1] that if $f(x)$ belongs to the class $H_{s,\alpha}$, then $f(x)$ expands in generalized Taylor series, which is uniformly convergent on $[-1, 1]$:

$$f(x) = \sum_{n=0}^{\infty} \left(\sum_{k \in N_{s,n}} a_{s,n,k} \cdot \varphi_{s,n,k}(x) + \sum_{p \in D_{s,n}} b_{s,n,p} \cdot \psi_{s,n,p}(x) \right),$$

where $a_{s,n,k} = f^{(n)}(x_{s,n,k})$, $b_{s,n,p} = \triangle^2_{\frac{1}{s \cdot (2s)^n}} \left(f^{(n)}; x^*_{s,n,p} \right)$. Basic functions $\varphi_{s,n,k}(x)$ and $\psi_{s,n,p}(x)$ are the finite linear combinations of translations of the function $mup_s(x)$.

Thus the following problems are of interest:

1) asymptotic behavior of basic functions $\varphi_{s,n,k}(x)$ and $\psi_{s,n,p}(x)$ as $n \to \infty$;

2) approximation properties of translations of the function $mup_s(x)$.

In this paper we consider these two problems and applications of $mup_s(x)$ to wavelet theory.

The author is grateful to professor V.A. Rvachev for attention to this work.

2 Asymptotics of the basic functions of the generalized Taylor series

Consider the following functions:

$$\Phi_s(z) = \sum_{k=0}^{\infty} mup_s \left(-1 + \frac{1}{s \cdot (2s)^k} \right) \cdot z^k,$$

$$\Lambda_s(z) = \sum_{k=0}^{\infty} mup_s \left(-1 + \frac{1}{(2s)^{k+1}} \right) \cdot z^k,$$

$$T_s(z) = \sum_{k=0}^{\infty} mup_s \left(-1 + \frac{1}{s \cdot (2s)^{k+1}} \right) \cdot z^k.$$

Theorem 1. *The function $\Phi_s(z)$ has a unique root λ_s in the domain $G_s = \left\{ z \in C : \ |z| < 4 \cdot s^2 \right\}$. Besides, λ_s is real and belongs to the interval $\left(-3 \cdot s^2, -1 \right)$. Furthermore, $\Lambda_s(\lambda_s) \neq 0$ and $\Lambda_s(\lambda_s) - s \cdot T_s(\lambda_s) \neq 0$.*

Consider the function

$$ab_s(x) = \sum_{j=0}^{\infty} \lambda_s^j \cdot mup_s\left(x - 1 + \frac{1}{s \cdot (2s)^j}\right),$$

which is defined on $(-\infty, 0]$.

Let

$$\Phi_{s,0,0}(x) = \begin{cases} ab_s(x), & x \leqslant 0 \\ ab_s(-x), & x > 0 \end{cases}, \ \Phi_{s,1,0}(x) = \begin{cases} ab_s(x), & x \leqslant 0 \\ -ab_s(-x), & x > 0 \end{cases}$$

$$\text{and } \Psi_{s,0}(x) = \begin{cases} ab_s(x), & x \leqslant 0 \\ 0, & x > 0 \end{cases}.$$

Using these functions we can formulate the following statement:

Theorem 2. *For any nonnegative integer i functions $\frac{2^{2n}\cdot(2s)^{n(2n-1)}}{c_{s,2n}} \cdot \varphi^{(i)}_{s,2n,0}(x)$, $\frac{2^{2n-1}\cdot(2s)^{(2n-1)(n-1)}}{c_{s,2n-1}} \cdot \varphi^{(i)}_{s,2n-1,0}(x)$ and $\frac{2^n\cdot(2s)^{\frac{n(n-1)}{2}}}{b_{s,n}} \cdot \psi^{(i)}_{s,n,s\cdot(2s)^n-1}(x)$ converge uniformly on $[-1,1]$ respectively to functions $\Phi^{(i)}_{s,0,0}(x)$, $\Phi^{(i)}_{s,1,0}(x)$ and $\Psi^{(i)}_{s,0}(x)$, where $c_{s,n} = \frac{Res\left(\frac{\Lambda_s(z)}{\Phi_s(z)},\lambda_s\right)}{\lambda_s^n}$ and $b_{s,n} = \frac{Res\left(\frac{s\cdot T_s(z)-\Lambda_s(z)}{\Phi_s(z)},\lambda_s\right)}{\lambda_s^n}$. Moreover, for any $\rho \in [3s^2, 4s^2)$ the following inequalities are true:*

1) $\left\| \frac{2^{2n}\cdot(2s)^{n(2n-1)}}{c_{s,2n}} \cdot \varphi^{(i)}_{s,2n,0}(x) - \Phi^{(i)}_{s,0,0}(x) \right\|_{C_{[-1,1]}} \leqslant M_1 \cdot \frac{(2n-i)\cdot|\lambda_s|^{2n}}{\rho^{2n}}$ *for* $2n \geqslant i+5$;

2) $\left\| \frac{2^{2n-1}\cdot(2s)^{(2n-1)(n-1)}}{c_{s,2n-1}} \cdot \varphi^{(i)}_{s,2n-1,0}(x) - \Phi^{(i)}_{s,1,0}(x) \right\|_{C_{[-1,1]}} \leqslant M_2 \cdot \frac{(2n-1-i)\cdot|\lambda_s|^{2n-1}}{\rho^{2n-1}}$ *for* $2n-1 \geqslant i+5$;

3) $\left\| \frac{2^n\cdot(2s)^{\frac{n(n-1)}{2}}}{b_{s,n}} \cdot \psi^{(i)}_{s,n,s\cdot(2s)^n-1}(x) - \Psi^{(i)}_{s,0}(x) \right\| \leqslant M_3 \cdot \frac{(n-i)\cdot|\lambda_s|^n}{\rho^n}$ *for* $n \geqslant i+5$.

Let

$$\Phi_{s,i,\frac{h}{s(2s)^q}}(x) = \Phi_{s,i,0}\left(x - \frac{h}{s\cdot(2s)^q}\right) - \sum_{l=0}^{q} \frac{\sum\limits_{j\in N_{s,l}} \Phi^{(l)}_{s,i,0}\left(x_{s,l,j} - \frac{h}{s\cdot(2s)^q}\right)\cdot\widehat{\varphi}_{s,l,j}(x)}{2^l\cdot(2s)^{\frac{l(l-1)}{2}}} -$$

$$- \sum_{l=0}^{q} \frac{\sum\limits_{p\in D_{s,l}} \triangle^2_{\frac{1}{s\cdot(2s)^l}}\left(\Phi^{(l)}_{s,i,0}; x^*_{s,l,p} - \frac{h}{s\cdot(2s)^q}\right)\cdot\widehat{\psi}_{s,l,p}(x)}{2^l\cdot(2s)^{\frac{l(l-1)}{2}}}$$

for $i \in \{1;\ 2\}$ and integer h such that $h \neq 0 \pmod{2s}$,

$$\Psi_{s,\frac{h}{s(2s)^q}}(x) = \Psi_{s,0}\left(x - \frac{h}{s\cdot(2s)^q}\right) - \sum_{l=0}^{q} \frac{\sum\limits_{j\in N_{s,l}} \Psi_{s,0}^{(l)}\left(x_{s,l,j} - \frac{h}{s\cdot(2s)^q}\right)\cdot \widehat{\varphi}_{s,l,j}(x)}{2^l\cdot(2s)^{\frac{l(l-1)}{2}}} - \\ - \sum_{l=0}^{q} \frac{\sum\limits_{p\in D_{s,l}} \triangle^2_{\frac{1}{s\cdot(2s)^l}}\left(\Psi_{s,0}^{(l)}; x^*_{s,l,p} - \frac{h}{s\cdot(2s)^q}\right)\cdot \widehat{\psi}_{s,l,p}(x)}{2^l\cdot(2s)^{\frac{l(l-1)}{2}}}$$

for integer h such that $h \neq 0 \pmod{2s}$.

Theorem 3. *Let $x^* = \frac{h}{s\cdot(2s)^q} \in [-1,1]$, where q is a nonnagative integer and h is an integer such that $h \neq 0 \pmod{2s}$. Then*

– *if $\frac{k}{s\cdot(2s)^{2n-1}} = x^*$ and $2n-1 \geqslant q$, then*

$$\frac{2^{2n}\cdot(2s)^{n(2n-1)}}{c_{s,2n}}\cdot\varphi_{s,2n,k}(x) = \Phi_{s,0,\frac{h}{s(2s)^q}}(x) + R_{s,2n,q}(x),$$

– *if $\frac{k}{s\cdot(2s)^{2n-2}} = x^*$ and $2n-2 \geqslant q$, then*

$$\frac{2^{2n-1}\cdot(2s)^{(n-1)(2n-1)}}{c_{s,2n-1}}\cdot\varphi_{s,2n-1,k}(x) = \Phi_{s,1,\frac{h}{s(2s)^q}}(x) + R_{s,2n-1,q}(x),$$

– *if $-1 + \frac{k}{(2s)^n} = x^*$ and $n-1 \geqslant q$, then*

$$\frac{2^n\cdot(2s)^{\frac{n(n-1)}{2}}}{b_{s,n}}\cdot\psi_{s,n,s\cdot k-1}(x) = \Psi_{s,\frac{h}{s(2s)^q}}(x) + r_{s,n,q}(x),$$

where $c_{s,n}$ and $b_{s,n}$ are defined in the previous theorem,

$$|R_{s,n,q}(x)| \leqslant M_1(s,\rho,q)\cdot\frac{n\cdot|\lambda_s|^n}{\rho^n} \text{ and } |r_{s,n,q}(x)| \leqslant M_2(s,\rho,q)\cdot\frac{n\cdot|\lambda_s|^n}{\rho^n}$$

for any $\rho \in \left[3s^2, 4s^2\right)$.

These theorems were proved in [2, 3] (for the case in which $s = 2$) and [4] (for the general case in which $s = 2, 3, 4, \ldots$).

3 Approximation properties of $mup_s(x)$

In this section we assume that n is a nonnegative integer. By $MUP_{s,n}$ denote the space of functions $f(x)$ such that

$$f(x) = \sum_k c_k \cdot mup_s\left(x - \frac{k}{(2s)^n}\right),\ x \in [-1, 1].$$

Theorem 4. *For any $k = 0, 1, \ldots, n$ the function $f(x) = x^n$ belongs to $MUP_{s,n}$.*

Consider the function

$$Fmup_{s,n}(x) = \frac{1}{2\pi} \cdot \int_{-\infty}^{\infty} e^{-itx} \cdot \left(\frac{\sin\frac{t}{2\cdot(2s)^n}}{\frac{t}{2\cdot(2s)^n}}\right)^n \cdot \prod_{k=n+1}^{\infty} \frac{\sin^2\left(\frac{st}{(2s)^k}\right)}{s^2 \cdot \frac{t}{(2s)^k} \cdot \sin\frac{t}{(2s)^k}} dt.$$

The support of this function is $\left[-\frac{n+2}{s\cdot(2s)^n}, \frac{n+2}{s\cdot(2s)^n}\right]$.

Theorem 5. *The system of functions*

$$\left\{Fmup_{s,n}\left(x - \frac{j}{(2s)^n} + 1 + \frac{n+2}{2\cdot(2s)^n}\right)\right\}_{j=1}^{2\cdot(2s)^n+n+1}$$

is a basis of the linear space $MUP_{s,n}$.

Thus the space $MUP_{s,n}$ combines the good approximation properties and the existence of local basis.

Let $\widetilde{W}_2^r$ be a class of functions $f \in C^{r-1}_{[-\pi,\pi]}$, such that $f^{(k)}(-\pi) = f^{(k)}(\pi)$ for any $k = 0, 1, \ldots, r-1$, $f^{(r-1)}(x)$ is absolutely continuous and $\left\|f^{(r)}\right\|_{L_2[-\pi,\pi]} \leqslant 1$.

And let $\widetilde{MUP}_{s,n} \subset C_{[-\pi,\pi]}$ be a space of functions of the following type:

$$\varphi(x) = \sum_k c_k \cdot mup_s\left(\frac{x}{\pi} - \frac{k}{(2s)^n}\right),$$

such that $\varphi^{(l)}(-\pi) = \varphi^{(l)}(\pi)$ for any $l = 0, 1, 2, \ldots$. It is easy to prove that $\dim \widetilde{MUP}_{s,n} = 2 \cdot (2s)^n$.

Theorem 6. *For any natural r there exists a constant $C \geqslant 0$ and an integer $n_0 \geqslant 0$ such that for any $n \geqslant n_0$*

$$E_{L_2}\left(\widetilde{W}_2^r, \widetilde{MUP}_{s,n}\right) \leqslant d_{2\cdot(2s)^n}\left(\widetilde{W}_2^r, L_2\right) \cdot \sqrt{1 + C \cdot 2^{-n}},$$

where $E_X(A,L) = \sup\limits_{\varphi\in A}\inf\limits_{\psi\in L}\|\varphi-\psi\|_X$ *and* $d_N(K,X) = \inf\limits_{dim\ L=N} E_X(K,L)$ *is the Kolmogorov width.*

That is spaces $\widetilde{MUP}_{s,n}$ are asymptotically extremal for approximation of classes $\widetilde{W}_2^r$ in norm of L_2.

4 Applications of $mup_s(x)$ to wavelet theory

Let $V_{s,n}$ be a space of funcions $f(x)$ such that

$$f(x) = \sum_{k\in I(f)} c_k \cdot Fmup_{s,n}\left(x - \frac{2k-n}{2\cdot(2s)^n}\right),$$

where $I(f)$ is a finite subset of the integers.

It is true that $V_{s,0} \subset V_{s,1} \subset \ldots \subset V_{s,n} \subset \ldots$.

Define the inner product of two functions $f, g \in L_2(R)$ as the integral

$$\int\limits_R f(x)\cdot g(x)dx.$$

Let $W_{s,n} = \{f \in V_{s,n} :\ f \perp V_{s,n-1}\}$ for any natural n.

Theorem 7. *For any natural* n *there exists functions* $\varphi_{s,n,1}(x), \ldots, \varphi_{s,n,2s-1}(x)$ *such that*

1) the system of functions

$$\left\{\varphi_{s,n,p}\left(x - \frac{j}{(2s)^{n-1}}\right)\right\}_{p=1,\ldots,2s-1, j\in Z}$$

is a basis of the linear space $W_{s,n}$*;*

2) supp $\varphi_{s,n,k} \subseteq \left[-\frac{k+1}{(2s)^n}, \frac{n+1}{(2s)^{n-1}} + \frac{n+1}{(2s)^n}\right]$ *for any* $k = 1, \ldots, 2s-1$*;*

3) for any $k = 1, \ldots, 2s-1$ *and* $p = 0, 1, \ldots, n-1$ *the following equality is true:*

$$\int\limits_R x^p \cdot \varphi_{s,n,k}(x)dx = 0.$$

As can be seen from this theorem the basis of $W_{s,n}$ consists of translations of $2s-1$ functions. It would be more convenient, if there exists a basis consisting of translations of one function. It can be proved that this function does not exist.

Consider the case $s = 2$.

Let

$$f_n(x) = \frac{1}{2\pi} \cdot \int\limits_{-\infty}^{\infty} e^{itx} \cdot F_n(t)dt \text{ and } \ g_n(x) = \frac{1}{2\pi} \cdot \int\limits_{-\infty}^{\infty} e^{itx} \cdot G_n(t)dt,$$

where

$$F_n(t) = \left(\frac{\sin \frac{2t}{4^{n+1}}}{\frac{2t}{4^{n+1}}} \right)^n \cdot F\left(\frac{t}{4^n} \right), G_n(t) = \left(\frac{\sin \frac{t}{4^{n+1}}}{\frac{t}{4^{n+1}}} \right)^{n+1} \cdot F\left(\frac{t}{4^{n+1}} \right)$$

and $F(t) = \prod\limits_{k=1}^{\infty} \frac{\sin \frac{2t}{4^k}}{\frac{2t}{4^k}} \cdot \cos \frac{t}{4^k}, n = 0, 1, 2, \ldots$.

Note that $f_0(x) = mup_2(x)$.

Functions $f_n(x)$ and $g_n(x)$ are smooth functions with a compact support.

Let $v_{2k}(x) = f_k\left(x - \frac{k+2}{2 \cdot 4^k}\right)$ and $v_{2k+1}(x) = g_k\left(x - \frac{k+2}{4^{k+1}}\right)$ for any $k = 0, 1, 2, \ldots$.

By V_n denote the space of funcions $f(x)$ such that

$$f(x) = \sum_{k \in I(\varphi)} c_k \cdot v_n\left(x - \frac{k}{2^n}\right), \ x \in R,$$

where $I(f)$ is a finite subset of the integers.

It can be shown that $V_0 \subset V_1 \subset \ldots \subset V_n \subset \ldots$.

Let $W_n = \{f \in V_n : \ f \perp V_{n-1}\}$.

Theorem 8. *For any natural n there exists a function $w_n(x)$ such that*

1) the system of functions $\left\{w_n\left(x - \frac{j}{2^{n-1}}\right)\right\}_{j \in Z}$ is a basis of the linear space W_n;

2) supp $w_n(x) \subseteq \left[0, \frac{n+2}{2^{n-1}}\right]$;

3) for any $m = 0, 1, \ldots, \left[\frac{n+1}{2}\right] - 1$ the following equality is true:

$$\int\limits_R x^m \cdot w_n(x)dx = 0.$$

The linear space W_n is a space of wavelets. The system of functions

$$\left\{mup_2(x - j), w_n\left(x - \frac{j}{2^{n-1}}\right)\right\}_{n \in N, j \in Z}$$

is a system of nonstationary smooth wavelets with a compact support.

The same method can be used for construction of nonstationary smooth wavelets with a compact support for the case $s = 2^k$.

References

1. V.A. Rvachev and G.A. Starets, *Some atomic functions and their applications*, Proc. Ukr. SSR Acad. Sci., **11** (1983), 22–24 [in Ukrainian].
2. V.A. Makarichev, *The asymptotics of the basic functions of a generalized Taylor series for the $H_{\rho,2}$ function class*, Visn. Khark. Univ., Ser. Mat. Prykl. Mat. Mekh., **826** (2008), 67–86 [in Russian].
3. V.A. Makarichev, *Asymptotics of the basis functions of generalized Taylor series for the class $H_{\rho,2}$*, Math. Notes, **89** (2011), no. 5, 689–705.
4. V.A. Makarichev, *On the asymptotics of the basic functions of a generalized Taylor series for some classes of infnitely differentiable functions*, Far Eastern Mathematical Journal, **11** (2011), no. 1, 56–75 [in Russian].

V. A. Makarichev
Contacts for the author: N.Ye. Zhukovsky National Aerospace University "Kharkov Aviation Institute", Ukraine, 61070, Kharkov, Chkalov str. 17, +38(057)788-40-39, victor.makarichev@gmail.com

FEICHTINGER'S CONJECTURE AND SIMPLE FRAMES

S. Y. Novikov, I. S. Ryabtsov

Key words: Feichtinger's Conjecture, Paving Conjecture, Parseval Frames, Simple Frames

AMS Mathematics Subject Classification: 42C15, 42A38

Abstract. Several recent results will be presented on the study of Feichtinger's conjecture: each bounded frame is a finite union of Riesz bases. This conjecture is equivalent to many unsolved problems in various areas of Mathematical Analysis. and naturally connected with the new concept of the so-called simple frames.

A *frame* for a Hilbert space $\mathcal{H}$ is a family of vectors $\{f_i\}_{i\in I}$ in $\mathcal{H}$ so that there are constants $A, B > 0$ satisfying:

$$A\|f\|^2 \leqslant \sum_{i\in I} |\langle f, f_i\rangle|^2 \leqslant B\|f\|^2, \quad \text{for all} \quad f \in \mathcal{H}.$$

The constants A and B are called *lower* and *upper frame bounds*, respectively. If we can choose $A = B$ we say that $\{f_i\}_{i\in I}$ is a *B-tight frame*. *Parseval Frame* is a 1-tight frame. If only the upper frame condition is satisfied we call $\{f_i\}_{i\in I}$ a *Bessel sequence*, with *Bessel constant* B. A sequence $\{f_i\}_{i\in I}$ in $\mathcal{H}$ is *bounded* if $0 < \inf_{i\in I} \|f_i\| \leqslant \sup_{i\in I} \|f_i\| < \infty$.

A bounded unconditional basis for $\mathcal{H}$ is called a *Riesz basis* for $\mathcal{H}$. A sequence $\{f_i\}_{i\in I}$ which is a Riesz basis for its closed linear span in $\mathcal{H}$ is called a *Riesz basic sequence* in $\mathcal{H}$. $\{f_i\}_{i\in I}$ is a Riesz basis for $\mathcal{H} \iff \exists A, B > 0$: for all finite families of scalars $\{a_i\}_{i\in I'\subset I}$

$$A\sum_{i\in I'} |a_i|^2 \leqslant \left\|\sum_{i\in I'} a_i f_i\right\|^2 \leqslant B\sum_{i\in I'} |a_i|^2.$$

In this inequalities we call A a *lower Riesz basis bound* of $\{f_i\}_{i\in I}$ and B an *upper Riesz basis bound.*

Original conjecture by H. Feichtinger is as following;

Conjecture 1. *Every bounded frame can be written as a finite union of Riesz basic sequences.*

Given $N \in \mathbf{N}$, let ℓ_2^N denote $\mathbf{C}^N$ equipped with ℓ_2-norm. The next conjecture concerns frames for ℓ_2^N and is usually called as *finite Feichtinger conjecture.*

Conjecture 2 (Finite Feichtinger Conjecture). *For every $B, C > 0$ there is a natural number $M = M(B,C)$ and an $A = A(B,C) > 0$: whenever $\{f_i\}_{i\in I}$ is a frame for ℓ_2^N with upper frame bound B and $\|f_i\| \geqslant C$ for all $i \in I$, then I can be partitioned into $\{I_j\}_{j=1}^M$ so that for each $1 \leqslant j \leqslant M$, $\{f_i\}_{i\in I_j}$ is a Riesz basic sequence with lower Riesz basis bound A and upper Riesz basis bound B.*

Conjecture 2 PF. *For every N there are $\varepsilon > 0$ and $M \in \mathbb{N}$: whenever $\{f_i\}_{i=1}^{2N}$ is an equal norm Parseval frame for ℓ_2^N, then the set $\{1, 2, \ldots, 2N\}$ can be partitioned into $\{I_j\}_{j=1}^M$ so that for each $1 \leqslant j \leqslant M$, $\{f_i\}_{i\in I_j}$ has the Bessel constant $\leqslant 1 - \varepsilon$ for all $j = 1, 2, \ldots, M$.*

The corresponding conjectures for Bessel sequences are

Conjecture 3. *Every bounded Bessel sequence can be written as a finite union of Riesz basic sequences.*

It's obvious, that Conjecture 3 implies Conjecture 1.

Conjecture 4. *For every $B > 0$ there is a natural number $M = M(B)$ and an $A = A(B) > 0$: every Bessel sequence $\{f_i\}_{i=1}^n$ with Bessel constant $B > 0$ and $\|f_i\| = 1$, for all $1 \leqslant i \leqslant n$, can be written as a union of M Riesz basic sequences each with lower Riesz basis bound A.*

Conjecture 1 implies Conjecture 4.

Let $\{e_i\}$ be an orthonormal basis for Hilbert space we are working in.

Kashin [2], Bourgain and Tzafriri [3] proved the following result known as the *Restricted-Invertibility Theorem*:

Theorem 1. *There is a universal constant $c > 0$: whenever $T : \ell_2^n \longrightarrow \ell_2^n$ is a linear operator for which $\|Te_i\| = 1$ for $1 \leqslant i \leqslant n$, then there exists a subset $\sigma \subset \{1, 2, \ldots, n\}$ of cardinality $|\sigma| \geqslant \frac{cn}{\|T\|^2}$ so that*

$$\Big\| \sum_{j\in\sigma} a_j T e_j \Big\|^2 \geqslant c \sum_{j\in\sigma} |a_j|^2 ,$$

for all choices of scalars $\{a_j\}_{j\in\sigma}$.

Theorem 1 gave rise to the following conjecture which is still open:

Conjecture 5. *For every $B > 0$ there is a natural number $M = M(B)$ and an $A = A(B) > 0$ so that if $T : \ell_2^n \longrightarrow \ell_2^n$ is a linear operator for which $\|Te_i\| = 1$ for $1 \leqslant i \leqslant n$, and $\|t\| \leqslant \sqrt{B}$, then there exists a partition $\{I_j\}_{j=1}^M$ of $\{1, 2, \ldots . n\}$ so that for each $1 \leqslant j \leqslant M$ and all choices of scalars $\|a_i\}_{i\in I_j}$ we have:*

$$\Big\| \sum_{i\in I_j} a_i T e_i \Big\|^2 \geqslant A \sum_{i\in I_j} |a_i|^2 .$$

It was proved in [7] that conjectures 1, 2 and 5 are equivalent in the sense that all three have positive answers or allthree have negative answers. All three conjectures are true if the well known *Paving Conjecture* holds. For given subset I of the integers we denote by P_I the orthogonal projection in ℓ_2 onto the subspace spanned by $\{e_i\}_{i\in I}$.

Conjecture 6 (The Paving Conjecture [1]). *For any $\varepsilon > 0$, there is a constant $M = M(\varepsilon)$ such that for every integer n and every linear operator S on ℓ_2^n whose matrix with respect to $\{e_i\}_{i=1}^n$ has zero diagonal, one can find a partition $\{\sigma_j\}_{j=1}^M$ of $\{1, 2, \dots, n\}$, such that*

$$\left\|P_{\sigma_j} S P_{\sigma_j}\right\| \leqslant \varepsilon \|S\| \quad \text{for all} \quad j = 1, 2. \dots, M.$$

The paving conjecture is known to be equivalent to the Kadison–Singer conjecture [1]. Deep analysis of the paving conjecture was made in [5].

The long history of these conjectures will most likely lead us to negative answers to all of them in general. Nevertheless there are rather wide classes of frames for which these conjectures are true.

The equivalence of all these conjectures was proved in [7].

The Paving Conjecture (Conjecture 6) implies Conjecture 4. The conjecture is known to be true for various classes of operators on ℓ_2^n; in particular it is proved for the operators whose matrices have small entries, $O\left(1/\log^{1+\gamma} n\right)$ for some $\gamma > 0$.

Theorem 2 (Bourgain-Tzafriri, 1991). *Let $\varepsilon > 0$ and S be a linear operator on ℓ_2^n whose matrix has zero diagonal and all enries are bounded by $1/\log^{1+\gamma} n$ for some $\gamma > 0$. Then S satisfies the conclusion of the Paving Conjecture: there exists a partition $\{\sigma_k\}_{k\leqslant M}$ of $\{1, 2, \dots, n\}$, where $M = M(\gamma, \varepsilon)$, and such that*

$$\|P_{\sigma_k} S P_{\sigma_k}\| \leqslant \varepsilon \|S\| \quad \textit{for all} \quad k = 1, 2. \dots, M.$$

The partition $\{\sigma_k\}$ constructed by Bourgain and Tzafriri is random, the conclusion holds with probability close to one. Non-probabilistic proof of the Restricted-Invertibility Theorem was found recently by D.Spielman and N. Srivastava [9].

Theorem 2 implies the positive answer to Conjecture 4 for sequences which are "well separated".

Corollary 1. Let $\{f_i\}_{i=1}^n$ be a Bessel sequence with Bessel constant $B > 0$ and with $\|f_i\| = 1$ for all i. If

$$|\langle f_i, f_j\rangle| \leqslant \frac{1}{\log^{1+\gamma} n} \quad \text{for all} \quad i \neq j,$$

then the sequence $\{f_i\}_{i=1}^n$ can be written as a union of $M = M(B, \gamma)$ Riesz basic sequences each with lower Riesz basis bound 1/2 and upper Riesz basis bound 3/2.

Let's continue some positive results in this direction. First, it is proved in [7], that bounded Bessel sequences can be decomposed into a finite union of linearly independent sets.

For this, we need a result of Christensen and Lindner.

Proposition 1. Let $M \in \mathbf{N}$, I a finite subset of $\mathbf{N}$ and let $\{f_i\}_{i\in I}$ be a sequence of nonzero elements in a Hilbert space. The following are equivalent:

(1) I can be partitioned into M disjoint sets $I_1, I_2, \ldots, I_M$ so that each family $\{f_i\}_{i\in I_j}$ $(j = 1, 2, \ldots, M)$ is linearly independent.

(2) For any nonempty subset $J \subset I$ we have

$$\frac{|J|}{dim\ span\{f_j\}_{j\in J}} \leqslant M.$$

Theorem 3. *Every Bessel sequence $\{f_i\}_{i\in I}$ with Bessel bound B and $\|f_i\| \geqslant C > 0$, for every $i \in I$, can be decomposed into $\lceil B/C^2 \rceil$ linearly independent sets.*

In the same paper the generalized Bourgain–Tzafriri invertibility theorem was proved up to a logarithmic factor, which leads to the following result concerning finite Feichtinger Conjecture.

Theorem 4. *There is a universal constant $c > 0$ and a $D = D(B)$ so that whenever $\{f_i\}_{i=1}^k$ is a Bessel sequence in an n-dimensional Hilbert space $\mathcal{H}$ with $\|f_i\| = 1$ for all $1 \leqslant i \leqslant k$ and Bessel constant B, then there is a partition of so that for each is a Riesz basic sequence with lower Riesz basis bound c.*

Feichtinger Conjecture is true for certain Weyl–Heisenberg frames.

If $g \in L^2(\mathbf{R})$, $a, b > 0$ we define for all $m, n \in \mathbf{Z}$:

$$E_{mb}g(t) = e^{2\pi imbt}g(t) \quad \text{and} \quad T_{na}g(t) = g(t - na).$$

If $\{E_{mb}T_{na}g\}_{m,\,n\in\mathbf{Z}}$ is a frame for $L^2(\mathbf{R})$, it is called a *Weyl–Heisenberg* or *Gabor frame.*

Theorem 5. *If $\{E_{mb}T_{na}g\}_{m,\,n\in\mathbf{Z}}$ is a frame for $L^2(\mathbf{R})$ and $0 < ab < 1$ with ab rational, then it can be written as a finite union of Riesz basic sequences.*

Another possibilities of such a representation of Gabor frames are presented in the K. Grochenig paper [6].

Definition 1. A Parseval frame $F = \{f_i\}_{i=1}^M$ in ℓ_2^N, $(M \geqslant N)$ is called *composite*, if there exists a set of non-negative constants $\{\alpha_i\}_{i=1}^M$, with at least one zero, such that the system of vectors $F_\alpha = \{\alpha_i f_i\}_{i=1}^M$ forms a Parseval frame again. The set $\{\alpha_i\}_{i=1}^M$ from this definition is called *transformation coefficients*.

Parseval frames $F = \{f_i\}_{i=1}^M$, which are not composite, will be called *simple*.

Property 1. *Simple frames are invariant under orthogonal transformations.*

The *reduction* transformation (RT) T for an arbitrary frame $F = \{f_i\}_{i=1}^M$ is defined in the following way. Suppose there exists a pair of vectors

$$f_k \in F, \quad f_p \in F, \quad |\langle f_k, f_p \rangle| = \|f_k\| \, \|f_p\| \,, \quad k \neq p \,.$$

Then the RT maps the frame by the following rule:

$$T(F) = F \setminus \{f_k,\, f_p\} \cup \left\{ \sqrt{\|f_k\|^2 + \|f_p\|^2}\, \frac{f_k}{\|f_k\|} \right\}.$$

If such pair of vectors does not exist, then $T(F) = F$. The action of the RT on a frame can be simplified

$$T(F) = F \setminus \{f_k,\, f_p\} \cup T(\{f_k,\, f_p\}) \,.$$

It is easy to see that this operator does not change the boundaries of the frame and actually replaces a pair of collinear vectors by a single vector.

When applying RT T sufficiently many times we get a frame without collinear vectors. We call such transformation a *full reduction transform (FRT)* and denote by T_∞. So,

$$T_\infty = T \circ T \circ T \circ \ldots \circ T.$$

The number of such compositions depends on the numser of collinear vectors in the frame. The FRT allows to transform a frame with collinear vectors to the frame without them maintaining its boundaries.

The *weight-union* of two Parseval frames $F_1 = \left\{ f_i^{F_1} \right\}_{i=1}^{M_1}$ and $F_2 = \left\{ f_i^{F_2} \right\}_{i=1}^{M_2}$ is, by definition, the following system of vectors

$$w\mathbb{U}(F_1, F_2) = T_\infty \left(\left\{ \lambda_{F_1}\, f_i^{F_1} \right\}_{i=1}^{M_1} \cup \left\{ \lambda_{F_2}\, f_i^{F_2} \right\}_{i=1}^{M_2} \right),$$

provided that $\lambda_{F_1}^2 + \lambda_{F_2}^2 = 1$. The number of vectors in the resulting frame does not exceed $M_1 + M_2$.

Similarly, we can determine the weight-union of a finite number of Parseval frames.

$$\underset{k=1}{\overset{K}{w\mathbb{U}}}\left(F_k\right) = T_\infty\left(\bigcup_{k=1}^{K}\left\{\lambda_{F_k} f_i^{F_k}\right\}_{i=1}^{M_k}\right), \quad F_k = \bigcup_{k=1}^{K}\left\{f_i^{F_k}\right\}_{i=1}^{M_k}, \quad \sum_{k=1}^{K}\lambda_{F_k}^2 = 1\,.$$

The weight-union of Parseval frames is the Parseval frame again.

Theorem 6. *Any Parseval frame $F = \{f_i\}_{i=1}^{M}$ can be written as the finite weight-union of simple Parseval frames.*

Proof. We begin by proving that any composite Parseval frame can be represented as a sum of two simple or composite frames. Indeed, according to definition of composite frame there exists a set of constants $\{\alpha_i\}_{i=1}^{M}$, such that $\exists k : 1 \leqslant k \leqslant M$, for which $\alpha_k = 0$, and the system of vectors $\{\alpha_i f_i\}_{i=1}^{M}$ is a Parseval frame.

Consider the *dual* set $\{\beta_i\}_{i=1}^{M}$ for transformation coefficients

$$\beta_i = \sqrt{\frac{\alpha_{\max}^2 - \alpha_i^2}{\alpha_{\max}^2 - 1}}, \quad \alpha_{\max} = \max_{1\leqslant i\leqslant M}|\alpha_i|\,.$$

The dual system $F_\beta = \{\beta_i f_i\}_{i=1}^{M}$ is a Parseval frame. Indeed,

$$\sum_{i=1}^{M}\langle\, x, \beta_i f_i\,\rangle\,\beta_i f_i = \sum_{i=1}^{M}\langle\, x, \sqrt{\tfrac{\alpha_{\max}^2-\alpha_i^2}{\alpha_{\max}^2-1}}\, f_i\,\rangle\sqrt{\tfrac{\alpha_{\max}^2-\alpha_i^2}{\alpha_{\max}^2-1}}\, f_i =$$

$$= \tfrac{1}{\alpha_{\max}^2-1}\sum_{i=1}^{M}\left(\alpha_{\max}^2 - \alpha_i^2\right)\langle\, x, f_i\,\rangle\, f_i = \tfrac{1}{\alpha_{\max}^2-1}\left(\alpha_{\max}^2 x - x\right) = x\,.$$

We show now that a set $\{\beta_i\}_{i=1}^{M}$ is well-defined. To see this, we remark, that $\forall k : 1 \leqslant k \leqslant M$ a constant β_k has sense and is a real number, and the set $\{\alpha_i\}_{i=1}^{M}$ does not coincide with the set $\{\beta_i\}_{i=1}^{M}$. It suffices to prove that $\alpha_{\max} > 1$ for any choice of $\{\alpha_i\}_{i=1}^{M}$, since the numerator is non-negative by definition.

Suppose the opposite, that there exists a set of constants $\{\alpha_i\}_{i=1}^{M}$, such that $\exists k : 1 \leqslant k \leqslant M$, for which $\alpha_k = 0$, and the system $\{\alpha_i f_i\}_{i=1}^{M}$ is a Parseval frame, but $\alpha_{\max} \leqslant 1$. We'll estimate the upper boundary of the frame $\{\alpha_i f_i\}_{i=1}^{M}$

$$\sum_{i=1}^{M}|\langle\, x, \alpha_i f_i\,\rangle|^2 = \sum_{i=1}^{M}|\alpha_i|^2\,|\langle\, x, f_i\,\rangle|^2 < \alpha_{\max}^2\sum_{i=1}^{M}|\langle\, x, f_i\,\rangle|^2 = \alpha_{\max}^2\,.$$

Strict inequality is ensured by the presence of zero components in the set $\{\alpha_i\}_{i=1}^M$. From here we see, that $\alpha_{\max} > 1$.

It's rather important to notice the following property of dual sets. For any pair of dual sets of $\{\alpha_i\}_{i=1}^M$ and $\{\beta_i\}_{i=1}^M$ there exist numbers $1 \leqslant k \leqslant M$ and $1 \leqslant p \leqslant M$, such that $k \neq p$ and thus

$$\alpha_k = 0,\ \beta_k \neq 0,\ \alpha_p \neq 0,\ \beta_p = 0\,.$$

This property follows from the definition of the appropriate sets. So the numbers $\{\beta_i\}_{i=1}^M$ are well-defined.

Any composite frame can be decomposed into a weight-union of Parseval frames $\{\alpha_i f_i\}_{i=1}^M$ и $\{\beta_i f_i\}_{i=1}^M$. Indeed, we take specific values λ_α и λ_β

$$\lambda_\alpha = \frac{1}{\alpha_{\max}},\quad \lambda_\beta = \frac{\sqrt{\alpha_{\max}^2 - 1}}{\alpha_{\max}},\quad \lambda_\alpha^2 + \lambda_\beta^2 = \frac{1}{\alpha_{\max}^2} + \frac{\alpha_{\max}^2 - 1}{\alpha_{\max}^2} = 1\,,$$

and calculate $w\mathbb{U}(F_\alpha, F_\beta)$. Since the frame F contains no collinear vectors, then the only pairs of collinear vectors in the sum can only be an $\alpha_k f_k$ and $\beta_k f_k$ для $1 \leqslant k \leqslant M$. We apply RT T to this pair.

$$T\left(\{\lambda_\alpha\alpha_i f_i, \lambda_\beta\beta_i f_i\}\right) = \sqrt{\|\lambda_\alpha\alpha_i f_i\|^2 + \|\lambda_\beta\beta_i f_i\|^2}\,\frac{\lambda_\alpha\alpha_i f_i}{\|\lambda_\alpha\alpha_i f_i\|} = \sqrt{\lambda_\alpha^2\alpha_i^2 + \lambda_\beta^2\beta_i^2}\,f_i = f_i\,,$$

$$\lambda_\alpha^2\alpha_i^2 + \lambda_\beta^2\beta_i^2 = \frac{\alpha_i^2}{\alpha_{\max}^2} + \frac{\alpha_{\max}^2 - 1}{\alpha_{\max}^2}\cdot\frac{\alpha_{\max}^2 - \alpha_i^2}{\alpha_{\max}^2 - 1} = \frac{\alpha_i^2 + \alpha_{\max}^2 - \alpha_i^2}{\alpha_{\max}^2} = 1\,.$$

Thus we have a representation of frame F in the form $F = w\mathbb{U}(F_\alpha, F_\beta)$. If the received frames F_α и F_β are simple, than the decomposition process is completed. Otherwise, to obtain a representation in the sum of simple Parseval frames, you need to use the above method several times.

We introduce the operator D by the following recursive formula

$$D\left(F, k\right) = \begin{cases} w\mathbb{U}\left(D\left(F_\alpha, k+1\right), D\left(F_\beta, k+1\right)\right) & ,\quad F \text{ — composite frame,} \\ F & ,\quad F \text{ — simple frame.} \end{cases}$$

It remains to prove that the recursion depth k does not exceed a certain constant. According to Property 1, with an increasing k on the unit, so reduces the number of vectors in each frame F_α and F_β

$$|F_\alpha| \leqslant |F| - 1\,,\quad |F_\beta| \leqslant |F| - 1\,.$$

Based on the fact that the Parseval frames with the smallest number of vectors are orthonormal bases, we obtain $k \leqslant M - N$, where M is the volume (quantity of the elements) of the original frame, and N is the dimension of the space.

A trivial example of a simple frame is an orthonormal basis. Now we'll give more interesting examples, such as equiangular tight frames.

A frame $F = \{f_i\}_{i=1}^{M}$, $\|f_i\| = 1$, $i = 1' \ldots, M$ is *equiangular*, if there exists a constant $c \in [\,0, 1)$, such that for all $i \leqslant j$ holds the following equation

$$|\langle f_i, f_j \rangle| = \begin{cases} 1, & i = j\,, \\ \pm\, c, & i \neq j\,. \end{cases}$$

In the paper [8] proved that a equiangular system is a tight frame if and only if

$$c = \sqrt{\frac{M - N}{N\,(M - 1)}}\,.$$

Known examples of equiangular tight frames are an orthonormal basis, of Mercedes-Benz and others [8].

Any equiangular tight frame $F = \{f_i\}_{i=1}^{M}$ after renormalization $\left\{\sqrt{\frac{N}{M}} f_i\right\}_{i=1}^{M}$ becomes a Parseval frame, and this frame is simple. However, there are simple frames which are not equiangular. □

References

1. R. Kadison and I. Singer, *Extensions of pure states,* Amer. Jour. Math. **81.** (1959), 547-564.
2. B.S. Kashin, *On some properties of random matrices...* Izv. Armenia Acad. Sci. Math. **15** no.5 (1980), 379–394.
3. J. Bourgain, L. Tzafriri, *Invertibility of "large" submatricies with applications to the geometry of Banach spaces ana harmonic analysis.* Israel J. Math., **57** (1987), 137-223.
4. A.A. Lunin, *Operator norms of submatrices.* Mat. Zametki, **45.** No 3 (1989), 94–100.
5. J. Bourgain, L. Tzafriri, *On a problem of Kadison and Singer.* J. Reine Angew. Math., **420** (1991), 1-43.
6. K. Gröchenig, *Localized frames are finite unions of Riesz sequences.* Adv. Comp. Math. **18**, no. 2-4. (2003), 149–157.

7. P.G. Casazza, O. Christensen, A.M. Lindner, R. Vershynin, *Frames and the Feichtinger conjecture.* Proceedings of the AMS, **133** (2005), 1025–1033.
8. P.G. Casazza, D. Redmond and J.C. Tremain, *Real Equiangular Frames*, CISS Meeting Information Sciences and Systems, Princeton, NJ, 2008.
9. D. A. Spielman, N. Srivastava, *An Elementary Proof of the Restricted Invertibility Theorem.* arXiv 0911.1114v4 [math FA] 2 Oct 2010, 1-7.

S. Y. Novikov
Samara State University, Russia, 443011, Samara, Academica Pavlova Str., 1,
+7-846-3379931, mostvil53@gmail.com

I. S. Ryabtsov
Samara State University, Russia, 443011, Samara, Academica Pavlova Str., 1,
+7-9276968254, tinnulion@gmail.com

CONSTRUCTIVE CHARACTERISTICS OF MIXED MODULI OF SMOOTHNESS OF POSITIVE ORDERS

M. K. Potapov, B. V. Simonov, S. Yu. Tikhonov

Key words: fractional mixed moduli of smoothness, K-functionals, constructive characteristics

AMS Mathematics Subject Classification: 41A17

Abstract. In this paper we prove equivalence between mixed modulus of smoothness, its constructive characteristics, and the corresponding K-functional.

1 Introduction

In [1] and [2], S.M. Nikolskii and N.S. Bakhvalov introduced classes of functions with dominating mixed modulus of smoothness of integer order. Since then the topic of embedding theorems for these classes as well as properties of mixed moduli of smoothness was extensively investigated (see e.g. [3]).

Recently it turned out that moduli of smoothness of *positive orders* play an important role in the theory of embedding theorems and some other problems (see e.g. [4]).

In this paper we prove equivalence between mixed modulus of smoothness of an L_p-function, $1 \leqslant p \leqslant \infty$, and its constructive characteristics (section 3). In particular, this allows to show equivalence between mixed modulus of smoothness and the corresponding K-functional (section 4). Finally, we list the main properties of mixed modulus of smoothness in section 5.

2 Notations and auxiliary results for functions on $\mathbb{T}^2$

We define by

— $L_p(\mathbb{T}^2), 1 \leqslant p \leqslant \infty$, the set of measurable functions $f(x,y)$, 2π-periodic in each variable, such that $\|f\|_{L_p(\mathbb{T}^2)} = \left(\int\limits_0^{2\pi}\int\limits_0^{2\pi} |f(x,y)|^p dxdy\right)^{\frac{1}{p}} < \infty$ for $1 \leqslant p < \infty$ and for $p = \infty$ f is continuous, $\|f\|_{L_p(\mathbb{T}^2)} = \max\limits_{0\leqslant x,y\leqslant 2\pi} |f(x,y)|$;

The paper was partially supported by RFFI 12-01-00169, 12-01-00170, NSH 979.2012.1, RYC-2011-09302, and MTM 2011-27637.

— $L_p^0(\mathbb{T}^2)$ the set of functions $f \in L_p(\mathbb{T}^2)$ such that $\int\limits_0^{2\pi} f(x,y)dy = 0$ for a.e. x, and $\int\limits_0^{2\pi} f(x,y)dx = 0$ for a.e. y;

— $V_{m_1,\infty}(f), V_{\infty,m_2}(f), V_{m_1,m_2}(f)$ de la Vallée Poussin sums of the Fourier series of f, i.e.,

$$V_{m_1,\infty}(f) = \frac{1}{\pi}\int\limits_0^{2\pi} f(x+t_1,y)V_{m_1}^{2m_1}(t_1)dt_1,$$

$$V_{\infty,m_2}(f) = \frac{1}{\pi}\int\limits_0^{2\pi} f(x,y+t_2)V_{m_2}^{2m_2}(t_2)dt_2,$$

$$V_{m_1,m_2}(f) = \frac{1}{\pi^2}\int\limits_0^{2\pi}\int\limits_0^{2\pi} f(x+t_1,y+t_2)V_{m_1}^{2m_1}(t_1)V_{m_2}^{2m_2}(t_2)dt_1dt_2,$$

where $V_0^0(t) = D_0(t)$, $V_n^{2n}(t) = \frac{D_n(t)+\ldots+D_{2n-1}(t)}{n}$, $n = 1, 2\ldots$,
$D_m(t) = \frac{\sin(m+\frac{1}{2})t}{2\sin\frac{t}{2}}$, $m = 0, 1, 2\ldots$;

— $f^{(\rho_1,\rho_2)}$ — derivative in the sense of Weyl of the function f of order $\rho_1 \geqslant 0$ with respect to x and of order $\rho_2 \geqslant 0$ with respect to y,

— $Y_{m_1,m_2}(f)_{L_p(\mathbb{T}^2)}$ the best approximation by a two-dimensional angle of the function $f \in L_p(\mathbb{T}^2)$, i.e., $Y_{m_1,m_2}(f)_{L_p(\mathbb{T}^2)} = \inf\limits_{T_{m_1,\infty},T_{\infty,m_2}} \|f - T_{m_1,\infty} - T_{\infty,m_2}\|_{L_p(\mathbb{T}^2)}$, where the function $T_{n_1,\infty}(x,y) \in L_p(\mathbb{T}^2)$ is a trigonometric polynomial of degree at most n_1 in x, and the function $T_{\infty,n_2}(x,y) \in L_p(\mathbb{T}^2)$ is a trigonometric polynomial of degree at most n_2 in y.

For the function $f \in L_p(\mathbb{T}^2)$ we define the difference of order $\alpha_1 > 0$ with respect to the variable x and the difference of order $\alpha_2 > 0$ with respect to the variable y as follows:

$$\Delta_{h_1}^{\alpha_1}(f) = \sum_{\nu_1=0}^{\infty}(-1)^{\nu_1}\binom{\alpha_1}{\nu_1} f(x+(\alpha_1-\nu_1)h_1, y),$$

$$\Delta_{h_2}^{\alpha_2}(f) = \sum_{\nu_2=0}^{\infty}(-1)^{\nu_2}\binom{\alpha_2}{\nu_2} f(x, y+(\alpha_2-\nu_2)h_2),$$

where $\binom{\alpha}{\nu} = 1$ for $\nu = 0$, $\binom{\alpha}{\nu} = \alpha$ for $\nu = 1$, $\binom{\alpha}{\nu} = \frac{\alpha(\alpha-1)\ldots(\alpha-\nu+1)}{\nu!}$ for $\nu \geqslant 2$.

Denote by $\omega_{\alpha_1,\alpha_2}(f,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}$ the mixed modulus of smoothness of a function $f\in L_p(\mathbb{T}^2)$ of orders α_1 и α_2 with respect to the variables x and y, i.e.,

$$\omega_{\alpha_1,\alpha_2}(f,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}=\sup_{|h_i|\leqslant\delta_i,i=1,2}\|\Delta_{h_1}^{\alpha_1}(\Delta_{h_2}^{\alpha_2}(f))\|_{L_p(\mathbb{T}^2)}.$$

Define by

— $W_p^{(\alpha_1,0)}$ the set of functions $f\in L_p^0(\mathbb{T}^2)$ such that $f^{(\alpha_1,0)}\in L_p^0(\mathbb{T}^2)$,

— $W_p^{(0,\alpha_2)}$ the set of functions $f\in L_p^0(\mathbb{T}^2)$ such that $f^{(0,\alpha_2)}\in L_p^0(\mathbb{T}^2)$,

— $W_p^{(\alpha_1,\alpha_2)}$ the set of functions $f\in L_p^0(\mathbb{T}^2)$ such that $f^{(\alpha_1,\alpha_2)}\in L_p^0(\mathbb{T}^2)$.

The K-functional of the function $f\in L_p^0(\mathbb{T}^2)$ is given by

$$K_{\alpha_1,\alpha_2}(f,t_1,t_2)_{L_p(\mathbb{T}^2)}\equiv K(f,t_1,t_2,\alpha_1,\alpha_2)_{L_p(\mathbb{T}^2)}=\inf_{g_1\in W_p^{(\alpha_1,0)},g_2\in W_p^{(0,\alpha_2)},g\in W_p^{(\alpha_1,\alpha_2)}}$$
$$\Big[\|f-g_1-g_2-g\|_{L_p(\mathbb{T}^2)}+t_1^{\alpha_1}\|g_1^{(\alpha_1,0)}\|_{L_p(\mathbb{T}^2)}+t_2^{\alpha_2}\|g_2^{(0,\alpha_2)}\|_{L_p(\mathbb{T}^2)}$$
$$+t_1^{\alpha_1}t_2^{\alpha_2}\|g^{(\alpha_1,\alpha_2)}\|_{L_p(\mathbb{T}^2)}\Big].$$

If $F(f,\delta_1,\delta_2),G(f,\delta_1,\delta_2)>0$, then $F(f,\delta_1,\delta_2)\ll G(f,\delta_1,\delta_2)$ means that there exists a constant C, independent of f,δ_1,δ_2 such that $F(f,\delta_1,\delta_2)\leqslant CG(f,\delta_1,\delta_2)$. If $F(f,\delta_1,\delta_2)\ll G(f,\delta_1,\delta_2)$ and $G(f,\delta_1,\delta_2)\ll F(f,\delta_1,\delta_2)$, then $F(f,\delta_1,\delta_2)\asymp G(f,\delta_1,\delta_2)$.

Lemma 1 (see [5]). *Let $f\in L_p^0(\mathbb{T}^2)$, $1\leqslant p\leqslant\infty$, $k_i\in N$, $n_i=0,1,2\ldots$, $i=1,2$. Then $\|f-V_{n_1,\infty}(f)-V_{\infty,n_2}(f)+V_{n_1,n_2}(f)\|_{L_p(\mathbb{T}^2)}\ll Y_{n_1,n_2}(f)_{L_p(\mathbb{T}^2)}\ll\omega_{k_1,k_2}(f,\frac{\pi}{n_1+1},\frac{\pi}{n_2+1})_{L_p(\mathbb{T}^2)}$.*

Lemma 2 (see [6]). *Let $1\leqslant p\leqslant\infty$, $\alpha_i>0$, $n_i=0,1,2\ldots$, $i=1,2$. Then*

$$\|T_{n_1,\infty}^{(\alpha_1,0)}\|_{L_p(\mathbb{T}^2)}\ll(n_1+1)^{\alpha_1}\|T_{n_1,\infty}\|_{L_p(\mathbb{T}^2)},$$

$$\|T_{\infty,n_2}^{(0,\alpha_2)}\|_{L_p(\mathbb{T}^2)}\ll(n_2+1)^{\alpha_2}\|T_{\infty,n_2}\|_{L_p(\mathbb{T}^2)},$$

$$\|T_{n_1,n_2}^{(\alpha_1,\alpha_2)}\|_{L_p(\mathbb{T}^2)}\ll(n_1+1)^{\alpha_1}(n_2+1)^{\alpha_2}\|T_{n_1,n_2}\|_{L_p(\mathbb{T}^2)}.$$

3 Notations and auxiliary results for functions on $\mathbb{T}$

Define by

— $L_p(\mathbb{T})$, $1\leqslant p\leqslant\infty$, the set of 2π-periodic measurable functions f such that for $1\leqslant p<\infty$, $\|f\|_{L_p(\mathbb{T})}=\left(\int\limits_0^{2\pi}|f(x)|^p\,dx\right)^{1/p}<\infty$, and for $p=\infty$ f is continuous and $\|f\|_{L_p(\mathbb{T})}=\max\limits_{0\leqslant x\leqslant 2\pi}|f(x)|$;

— $L_p^0(\mathbb{T})$ the set of functions $f \in L_p(\mathbb{T})$ such that $\int\limits_0^{2\pi} f(x)dx = 0$;

— $V_n(f)$ de la Vallée Poussin sums of function $f \in L_p(\mathbb{T})$, i.e.,
$V_n(f) = \frac{1}{\pi} \int\limits_0^{2\pi} f(x+t)V_n^{2n}(t)dt$;

— $f^{(\rho)}$ the Weyl derivative of f of order $\rho(\rho > 0)$.

For the function $f \in L_p$ let us define the difference of positive order α as follows

$$\Delta_h^\alpha(f) = \sum_{\nu=0}^{\infty} (-1)^\nu \binom{\alpha}{\nu} f(x + (\alpha - \nu)h).$$

Lemma 3 (see [3]). *Let $f \in L_p^0$, $1 \leqslant p \leqslant \infty$. Then $\|V_n(f)\|_{L_p(\mathbb{T})} \ll \|f\|_{L_p(\mathbb{T})}$.*

Lemma 4 (see [7]). *Let $\alpha > 0$. Then $\sum\limits_{\nu=0}^{\infty} (-1)^\nu \binom{\alpha}{\nu} = 0$.*

Lemma 5 (see [7]). *Let $f \in L_p^0, g \in L_p^0$, $1 \leqslant p \leqslant \infty$, $\alpha > 0, \beta > 0$. Then*

(a) $\triangle_h^\alpha(f+g) = \triangle_h^\alpha f + \triangle_h^\alpha g$;
(b) $\triangle_h^\alpha(\triangle_h^\beta f) = \triangle_h^{\alpha+\beta} f$;
(c) $\|\triangle_h^\alpha f\|_{L_p(\mathbb{T})} \ll \|f\|_{L_p(\mathbb{T})}$.

Lemma 6 (see [7]). *Let $1 \leqslant p \leqslant \infty$, $\alpha > 0$, and T_n be a trigonometric polynomial of degree at most $n \in N$. Then*

(a) *for any $h : 0 < |h| \leqslant \frac{\pi}{n}$, we have $\|\triangle_h^\alpha T_n\|_{L_p(\mathbb{T})} \ll n^{-\alpha} \|T_n^{(\alpha)}\|_{L_p(\mathbb{T})}$;*
(b) $\|T_n^{(\alpha)}\|_{L_p(\mathbb{T})} \ll n^\alpha \|\triangle_{\frac{\pi}{n}}^\alpha T_n\|_{L_p(\mathbb{T})}$.

Lemma 7. *Let $f \in L_p^0(\mathbb{T})$, $1 \leqslant p \leqslant \infty, \alpha > 0$. Then*

$$\|V_{2^{m+1}}(f) - V_{2^m}(f)\|_{L_p(\mathbb{T})} \ll 2^{-m\alpha} \|V_{2^{m+1}}^{(\alpha)}(f) - V_{2^m}^{(\alpha)}(f)\|_{L_p(\mathbb{T})}.$$

This lemma follows from lemmas 3.3 and 3.4 of the paper [8].

4 Constructive characteristics of mixed moduli of smoothness

Theorem 1. *Let $f \in L_p^0(\mathbb{T}^2), 1 \leqslant p \leqslant \infty, \alpha_i > 0, n_i \in \mathbb{N}, i = 1, 2$. Then*

$$\omega_{\alpha_1,\alpha_2}\Big(f, \frac{\pi}{2n_1 - 1}, \frac{\pi}{2n_2 - 1}\Big)_{L_p(\mathbb{T}^2)} \asymp n_1^{-\alpha_1} n_2^{-\alpha_2} \big\|V_{n_1,n_2}^{(\alpha_1,\alpha_2)}(f)\big\|_{L_p(\mathbb{T}^2)} +$$

$$+ n_1^{-\alpha_1} \left\| V_{n_1,\infty}^{(\alpha_1,0)}(f - V_{\infty,n_2}(f)) \right\|_{L_p(\mathbb{T}^2)} + n_2^{-\alpha_2} \left\| V_{\infty,n_2}^{(0,\alpha_2)}(f - V_{n_1,\infty}(f)) \right\|_{L_p(\mathbb{T}^2)} + \\ + \left\| f - V_{n_1,\infty}(f) - V_{\infty,n_2}(f) + V_{n_1,n_2}(f) \right\|_{L_p(\mathbb{T}^2)}.$$

Proof. For any h_i и $n_i \in N, i = 1, 2$ we have

$$\|\triangle_{h_1}^{\alpha_1}(\triangle_{h_2}^{\alpha_2}(f))\|_{L_p(\mathbb{T}^2)} \leqslant \|\triangle_{h_1}^{\alpha_1}(\triangle_{h_2}^{\alpha_2}(f - V_{n_1,\infty}(f) - V_{\infty,n_2}(f) + V_{n_1,n_2}(f)))\|_{L_p(\mathbb{T}^2)} + \\ + \|\triangle_{h_1}^{\alpha_1}(\triangle_{h_2}^{\alpha_2}(V_{n_1,\infty}(f - V_{\infty,n_2})))\|_{L_p(\mathbb{T}^2)} + \|\triangle_{h_1}^{\alpha_1}(\triangle_{h_2}^{\alpha_2}(V_{\infty,n_2}(f - V_{n_1,\infty})))\|_{L_p(\mathbb{T}^2)} + \\ + \|\triangle_{h_1}^{\alpha_1}(\triangle_{h_2}^{\alpha_2}(V_{n_1,n_2}(f)))\|_{L_p(\mathbb{T}^2)} \equiv I_1 + I_2 + I_3 + I_4.$$

Let us first estimate I_1 from above. Denote $\varphi(x, y) = f - V_{n_1,\infty}(f) - V_{\infty,n_2}(f) + V_{n_1,n_2}(f)$. By lemma 5 c), we have:

$$\left(\int_0^{2\pi} |\triangle_{h_1}^{\alpha_1}(\triangle_{h_2}^{\alpha_2}(\varphi))|^p dx \right)^{\frac{1}{p}} \ll \left(\int_0^{2\pi} |\triangle_{h_2}^{\alpha_2}(\varphi)|^p dx \right)^{\frac{1}{p}} \text{ a.e. } y \text{ and } 1 \leqslant p < \infty;$$

and

$$\max_{0 \leqslant x \leqslant 2\pi} |\triangle_{h_1}^{\alpha_1}(\triangle_{h_2}^{\alpha_2}(\varphi))| \ll \max_{0 \leqslant x \leqslant 2\pi} |\triangle_{h_2}^{\alpha_2}(\varphi)| \text{ a.e. } y \text{ and } p = \infty.$$

Then for $1 \leqslant p \leqslant \infty$ we get

$$\|\triangle_{h_1}^{\alpha_1}(\triangle_{h_2}^{\alpha_2}(\varphi))\|_{L_p(\mathbb{T}^2)} \ll \|\triangle_{h_2}^{\alpha_2}(\varphi)\|_{L_p(\mathbb{T}^2)}.$$

Therefore, $I_1 \ll \|\triangle_{h_2}^{\alpha_2}(\varphi)\|_{L_p(\mathbb{T}^2)} \equiv I_5$. Using similar arguments we have $\|\triangle_{h_2}^{\alpha_2}(\varphi)\|_{L_p(\mathbb{T}^2)} \ll \|\varphi\|_{L_p(\mathbb{T}^2)}$, $1 \leqslant p \leqslant \infty$. Then $I_5 \ll \|\varphi\|_{L_p(\mathbb{T}^2)}$ and hence

$$I_1 \ll \|f - V_{n_1,\infty}(f) - V_{\infty,n_2}(f) + V_{n_1,n_2}(f)\|_{L_p(\mathbb{T}^2)}.$$

Estimating I_2, we denote $\psi = f - V_{\infty,n_2}(f)$. Using lemma 5 (c), we have for $1 \leqslant p \leqslant \infty$

$$\|\triangle_{h_1}^{\alpha_1}(\triangle_{h_2}^{\alpha_2}(V_{n_1,\infty}(\psi)))\|_{L_p(\mathbb{T}^2)} \ll \|\triangle_{h_1}^{\alpha_1}(V_{n_1,\infty}(\psi))\|_{L_p(\mathbb{T}^2)}.$$

This yields

$$I_2 \ll \|\triangle_{h_1}^{\alpha_1}(V_{n_1,\infty}(\psi))\|_{L_p(\mathbb{T}^2)} = I_6.$$

Using lemma 6 (a), we get:

$$\left(\int_0^{2\pi} |\triangle_{h_1}^{\alpha_1}(V_{n_1,\infty}(\psi))|^p dx \right)^{\frac{1}{p}} \ll n_1^{-\alpha_1} \left(\int_0^{2\pi} |V_{n_1,\infty}^{(\alpha_1,0)}(\psi)|^p dx \right)^{\frac{1}{p}}$$

for a.e. y, all h_1 such that $0 < |h_1| \leqslant \frac{\pi}{2n_1-1}$ and $1 \leqslant p < \infty$; and

$$\max_{0\leqslant x\leqslant 2\pi} |\triangle_{h_1}^{\alpha_1}(V_{n_1,\infty}(\psi))| \ll n_1^{-\alpha_1} \max_{0\leqslant x\leqslant 2\pi} |V_{n_1,\infty}^{(\alpha_1,0)}(\psi)|$$

for a.e. y, all h_1 such that $0 < |h_1| \leqslant \frac{\pi}{2n_1-1}$ and $p = \infty$.

Then for $1 \leqslant p \leqslant \infty$ we get $\|\triangle_{h_1}^{\alpha_1}(V_{n_1,\infty}(\psi))\|_{L_p(\mathbb{T}^2)} \ll n_1^{-\alpha_1}\|V_{n_1,\infty}^{(\alpha_1,0)}(\psi)\|_{L_p(\mathbb{T}^2)}$. Therefore, $I_6 \ll n_1^{-\alpha_1}\|V_{n_1,\infty}^{(\alpha_1,0)}(\psi)\|_{L_p(\mathbb{T}^2)}$ and for $0 < |h_1| \leqslant \frac{\pi}{2n_1-1}$ and $1 \leqslant p \leqslant \infty$ we get

$$I_2 \ll n_1^{-\alpha_1}\|V_{n_1,\infty}^{(\alpha_1,0)}(f - V_{\infty,n_2}(f))\|_{L_p(\mathbb{T}^2)}.$$

Similarly we have for $0 < |h_1| \leqslant \frac{\pi}{2n_1-1}$, $0 < |h_2| \leqslant \frac{\pi}{2n_2-1}$ and $1 \leqslant p \leqslant \infty$ we write

$$I_3 \ll n_2^{-\alpha_2}\|V_{\infty,n_2}^{(0,\alpha_2)}(f - V_{n_1,\infty}(f))\|_{L_p(\mathbb{T}^2)};\ I_4 \ll n_1^{-\alpha_1}n_2^{-\alpha_2}\|V_{n_1,n_2}^{(\alpha_1,\alpha_2)}(f)\|_{L_p(\mathbb{T}^2)}.$$

Hence,

$$\begin{aligned}\omega_{\alpha_1,\alpha_2}(f, \frac{\pi}{2n_1-1}, \frac{\pi}{2n_2-1})_{L_p(\mathbb{T}^2)} \ll \|f - V_{n_1,\infty}(f) - V_{\infty,n_2}(f) + V_{n_1,n_2}(f)\|_{L_p(\mathbb{T}^2)} + \\ + n_1^{-\alpha_1}\|V_{n_1,\infty}^{(\alpha_1,0)}(f - V_{\infty,n_2}(f))\|_{L_p(\mathbb{T}^2)} + \\ + n_2^{-\alpha_2}\|V_{\infty,n_2}^{(0,\alpha_2)}(f - V_{n_1,\infty}(f))\|_{L_p(\mathbb{T}^2)} + n_1^{-\alpha_1}n_2^{-\alpha_2}\|V_{n_1,n_2}^{(\alpha_1,\alpha_2)}(f)\|_{L_p(\mathbb{T}^2)}.\end{aligned}$$

Thus, the proof of the above estimate in theorem 1 is complete.

Let us estimate $\omega_{\alpha_1,\alpha_2}(f, \frac{\pi}{2n_1-1}, \frac{\pi}{2n_2-1})_{L_p(\mathbb{T}^2)}$ from below. Using lemma 1 and properties of moduli of smoothness of integer order, we get

$$\begin{aligned}A_1 \equiv \|f - V_{n_1,\infty}(f) - V_{\infty,n_2}(f) + V_{n_1,n_2}(f)\|_{L_p(\mathbb{T}^2)} \ll \\ \ll \omega_{[\alpha_1]+1,[\alpha_2]+1}(f, \frac{\pi}{n_1+1}, \frac{\pi}{n_2+1})_{L_p(\mathbb{T}^2)} \ll \\ \ll \omega_{[\alpha_1]+1,[\alpha_2]+1}\Big(f, \pi/(2n_1-1), \pi/(2n_2-1)\Big)_{L_p(\mathbb{T}^2)}.\end{aligned}$$

By lemma 5 (b),

$$A_1 \leqslant \sup_{|h_i|\leqslant \frac{\pi}{2n_i-1}, i=1,2} \|\triangle_{h_1}^{[\alpha_1]+1-\alpha_1}(\triangle_{h_2}^{[\alpha_2]+1-\alpha_2}(\triangle_{h_1}^{\alpha_1}(\triangle_{h_2}^{\alpha_2}(f))))\|_{L_p(\mathbb{T}^2)}.$$

Also, be lemma 5 (c), we get

$$A_1 \leqslant \sup_{|h_i|\leqslant \frac{\pi}{2n_i-1}, i=1,2} \|\triangle_{h_1}^{\alpha_1}(\triangle_{h_2}^{\alpha_2}(f))\|_{L_p(\mathbb{T}^2)} = \omega_{\alpha_1,\alpha_2}\Big(f, \frac{\pi}{2n_1-1}, \frac{\pi}{2n_2-1}\Big)_{L_p(\mathbb{T}^2)}.$$

Let us estimate

$$A_2 = \|V^{(\alpha_1,0)}_{n_1,\infty}(f - V_{\infty,n_2}(f))\|_{L_p(\mathbb{T}^2)}.$$

Denote $\gamma(x,y) = f(x,y) - V_{\infty,n_2}(f)$. Then using lemma 6 (b), we have

$$\left(\int_0^{2\pi} |V^{(\alpha_1,0)}_{n_1,\infty}(\gamma)|^p dx\right)^{\frac{1}{p}} \ll n_1^{\alpha_1} \left(\int_0^{2\pi} |\triangle^{\alpha_1}_{\frac{\pi}{2n_1-1}}(V_{n_1,\infty}(\gamma))|^p dx\right)^{\frac{1}{p}}$$

for a.e. y and $1 \leqslant p < \infty$ and

$$\max_{0\leqslant x\leqslant 2\pi} |V^{(\alpha_1,0)}_{n_1,\infty}(\gamma)| \ll n_1^{\alpha_1} \max_{0\leqslant x\leqslant 2\pi} |\triangle^{\alpha_1}_{\frac{\pi}{2n_1-1}}(V_{n_1,\infty}(\gamma))|$$

for any y and $p = \infty$. Then for $1 \leqslant p \leqslant \infty$ we have

$$\|V^{(\alpha_1,0)}_{n_1,\infty}(\gamma)\|_{L_p(\mathbb{T}^2)} \ll n_1^{\alpha_1} \|\triangle^{\alpha_1}_{\frac{\pi}{2n_1-1}}(V_{n_1,\infty}(\gamma))\|_{L_p(\mathbb{T}^2)}.$$

This gives $A_2 \ll n_1^{\alpha_1} \|V_{n_1,\infty}(\triangle^{\alpha_1}_{\frac{\pi}{2n_1-1}}(\gamma))\|_{L_p(\mathbb{T}^2)}$. Also, lemma 3 implies

$$A_2 \ll n_1^{\alpha_1} \|\triangle^{\alpha_1}_{\frac{\pi}{2n_1-1}}(f - V_{\infty,n_2}(f))\|_{L_p(\mathbb{T}^2)}.$$

Denoting $\triangle^{\alpha_1}_{\frac{\pi}{n_1}}(f) \equiv F$, we have $A_2 \ll n_1^{\alpha_1} \|F - V_{\infty,n_2}(F)\|_{L_p(\mathbb{T}^2)}$. Since $V_{0,\infty}(F) = V_{0,n_2}(F) = 0$, then

$$A_2 \ll n_1^{\alpha_1} \|F - V_{0,\infty}(F) - V_{\infty,n_2}(F) + V_{0,n_2}(F)\|_{L_p(\mathbb{T}^2)}.$$

Using lemma 2.1 and properties of the modulus of smoothness, we get

$$A_2 \ll n_1^{\alpha_1} \omega_{[\alpha_1]+1,[\alpha_2]+1}(F, \pi, \frac{\pi}{2n_2-1})_{L_p(\mathbb{T}^2)}.$$

By lemma 5 (b), we get

$$A_2 \ll n_1^{\alpha_1} \sup_{|h_1|\leqslant\pi, |h_2|\leqslant\frac{\pi}{2n_2-1}} \|\triangle^{[\alpha_1]+1}_{h_1}(\triangle^{[\alpha_2]+1-\alpha_2}_{h_2}(\triangle^{\alpha_2}_{h_2}(F)))\|_{L_p(\mathbb{T}^2)}.$$

Further, lemma 5 (c) gives

$$A_2 \ll n_1^{\alpha_1} \sup_{|h_2|\leqslant\frac{\pi}{2n_2-1}} \|\triangle^{\alpha_2}_{h_2}(F)\|_{L_p(\mathbb{T}^2)} = n_1^{\alpha_1} \sup_{|h_2|\leqslant\frac{\pi}{2n_2-1}} \|\triangle^{\alpha_2}_{h_2}(\triangle^{\alpha_1}_{\frac{\pi}{2n_1-1}}(f))\|_{L_p(\mathbb{T}^2)}$$

$$\ll n_1^{\alpha_1}\omega_{\alpha_1,\alpha_2}(f,\frac{\pi}{2n_1-1},\frac{\pi}{2n_2-1})_{L_p(\mathbb{T}^2)}.$$

Similarly we can show

$$\|V_{\infty,n_2}^{(0,\alpha_2)}(f-V_{n_1,\infty}(f))\|_{L_p(\mathbb{T}^2)}\ll n_2^{\alpha_2}\omega_{\alpha_1,\alpha_2}(f,\frac{\pi}{2n_1-1},\frac{\pi}{2n_2-1})_{L_p(\mathbb{T}^2)},$$

$$\|V_{n_1,n_2}^{(\alpha_1,\alpha_2)}(f)\|_{L_p(\mathbb{T}^2)}\ll n_1^{\alpha_1}n_2^{\alpha_2}\omega_{\alpha_1,\alpha_2}(f,\frac{\pi}{2n_1-1},\frac{\pi}{2n_2-1})_{L_p(\mathbb{T}^2)}.$$

Thus,

$$\begin{aligned}\|f-V_{n_1,\infty}(f)-V_{\infty,n_2}(f)+V_{n_1,n_2}(f)\|_{L_p(\mathbb{T}^2)}+n_1^{-\alpha_1}\|V_{n_1,\infty}^{(\alpha_1,0)}(f-V_{\infty,n_2}(f))\|_{L_p(\mathbb{T}^2)}+\\ +n_2^{-\alpha_2}\|V_{\infty,n_2}^{(0,\alpha_2)}(f-V_{n_1,\infty}(f))\|_{L_p(\mathbb{T}^2)}+\\ +n_1^{-\alpha_1}n_2^{-\alpha_2}\|V_{n_1,n_2}^{(\alpha_1,\alpha_2)}(f)\|_{L_p(\mathbb{T}^2)}\ll\omega_{\alpha_1,\alpha_2}\left(f,\frac{\pi}{2n_1-1},\frac{\pi}{2n_2-1}\right)_{L_p(\mathbb{T}^2)},\end{aligned}$$

i.e., the proof of theorem 1 is complete.

5 Equivalence between mixed moduli of smoothness and K-functionals

Theorem 2. *Let $f\in L_p^0(\mathbb{T}^2)$, $1\leqslant p\leqslant\infty$, $\alpha_i>0, 0<\delta_i\leqslant\pi$, $i=1,2$. Then*

$$\omega_{\alpha_1,\alpha_2}(f,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}\asymp K_{\alpha_1,\alpha_2}(f,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}. \tag{5.1}$$

Proof. For any $\delta_i\in(0,\pi]$ there exist integers n_i such that $\frac{\pi}{2n_i+1}<\delta_i\leqslant\frac{\pi}{2n_i-1}, i=1,2$. If $f\in L_p^0(\mathbb{T}^2)$, then $V_{n_1+1,n_2+1}(f)\in W_p^{(\alpha_1,\alpha_2)}$,

$$(V_{n_1+1,\infty}(f)-V_{n_1+1,n_2+1}(f))\in W_p^{(\alpha_1,0)};\quad (V_{\infty,n_2+1}(f)-V_{n_1+1,n_2+1}(f))\in W_p^{(0,\alpha_2)}.$$

Therefore,

$$\begin{aligned}K(f,\delta_1,\delta_2,\alpha_1,\alpha_2)_{L_p(\mathbb{T}^2)}\leqslant\\ \leqslant\leqslant\|f-(V_{n_1+1,\infty}(f)-V_{n_1+1,n_2+1}(f))-(V_{\infty,n_2+1}(f)-V_{n_1+1,n_2+1}(f))-\\ -V_{n_1+1,n_2+1}(f)\|_{L_p(\mathbb{T}^2)}+\delta_1^{\alpha_1}\|V_{n_1+1,\infty}^{(\alpha_1,0)}(f)-V_{n_1+1,n_2+1}^{(\alpha_1,0)}(f)\|_{L_p(\mathbb{T}^2)}+\\ +\delta_2^{\alpha_2}\|V_{\infty,n_2+1}^{(0,\alpha_2)}(f)-V_{n_1+1,n_2+1}^{(0,\alpha_2)}(f)\|_{L_p(\mathbb{T}^2)}+\delta_1^{\alpha_1}\delta_2^{\alpha_2}\|V_{n_1+1,n_2+1}^{(\alpha_1,\alpha_2)}(f)\|_{L_p(\mathbb{T}^2)}=\\ =\|f-V_{n_1+1,\infty}(f)-V_{\infty,n_2+1}(f)+V_{n_1+1,n_2+1}(f)\|_{L_p(\mathbb{T}^2)}+\end{aligned}$$

$$+\delta_1^{\alpha_1}\|V^{(\alpha_1,0)}_{n_1+1,\infty}(f-V_{\infty,n_2+1}(f))\|_{L_p(\mathbb{T}^2)}+\delta_2^{\alpha_2}\|V^{(0,\alpha_2)}_{\infty,n_2+1}(f-V_{n_1+1,\infty}(f))\|_{L_p(\mathbb{T}^2)}+$$
$$+\delta_1^{\alpha_1}\delta_2^{\alpha_2}\|V^{(\alpha_1,\alpha_2)}_{n_1+1,n_2+1}(f)\|_{L_p(\mathbb{T}^2)}\ll$$
$$\ll\|f-V_{n_1+1,\infty}(f)-V_{\infty,n_2+1}(f)+V_{n_1+1,n_2+1}(f)\|_{L_p(\mathbb{T}^2)}+$$
$$+n_1^{-\alpha_1}\|V^{(\alpha_1,0)}_{n_1+1,\infty}(f-V_{\infty,n_2+1}(f))\|_{L_p(\mathbb{T}^2)}+$$
$$+n_2^{-\alpha_2}\|V^{(0,\alpha_2)}_{\infty,n_2+1}(f-V_{n_1+1,\infty}(f))\|_{L_p(\mathbb{T}^2)}+n_1^{-\alpha_1}n_2^{-\alpha_2}\|V^{(\alpha_1,\alpha_2)}_{n_1+1,n_2+1}(f)\|_{L_p(\mathbb{T}^2)}.$$

By theorem 1, we then arrive at

$$K(f,\delta_1,\delta_2,\alpha_1,\alpha_2)_{L_p(\mathbb{T}^2)}\ll\omega_{\alpha_1,\alpha_2}\Big(f,\frac{\pi}{2n_1+1},\frac{\pi}{2n_2+1}\Big)_{L_p(\mathbb{T}^2)}\ll$$
$$\ll\omega_{\alpha_1,\alpha_2}\big(f,\delta_1,\delta_2\big)_{L_p(\mathbb{T}^2)},$$

which gives the estimate from below in (5.1).

Let us show estimate from above. Considering $g_1\in W_p^{(\alpha_1,0)}$, $g_2\in W_p^{(0,\alpha_2)}$ and $g\in W_p^{(\alpha_1,\alpha_2)}$, by lemma 5 (a) we get:

$$\omega_{\alpha_1,\alpha_2}(f,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}\ll$$
$$\ll\omega_{\alpha_1,\alpha_2}(f-g_1-g_2-g,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}+\omega_{\alpha_1,\alpha_2}(g_1,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}+$$
$$+\omega_{\alpha_1,\alpha_2}(g_2,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}+\omega_{\alpha_1,\alpha_2}(g,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}\equiv J_1+J_2+J_3+J_4.$$

First, Lemma 5 (c) implies $J_1\ll\|f-g_1-g_2-g\|_{L_p(\mathbb{T}^2)}$.

Second, estimating J_2, for any $\delta_i\in(0,\pi]$ we find integers n_i such that $\frac{\pi}{2^{n_i+2}-1}<\delta_i\leqslant\frac{\pi}{2^{n_i+1}-1}$, $i=1,2$. Consider $B_2=\omega_{\alpha_1,\alpha_2}\big(g_1,\frac{\pi}{2^{n_1+1}-1},\frac{\pi}{2^{n_2+1}-1}\big)_{L_p(\mathbb{T}^2)}$.

Lemmas 5 (a) and (b) give

$$B_2\ll\omega_{\alpha_1,\alpha_2}\big(g_1-V_{2^{n_1},\infty}(g_1),\frac{\pi}{2^{n_1+1}-1},\frac{\pi}{2^{n_2+1}-1}\big)_{L_p(\mathbb{T}^2)}+$$
$$+\omega_{\alpha_1,\alpha_2}\big(V_{2^{n_1},\infty}(g_1),\frac{\pi}{2^{n_1+1}-1},\frac{\pi}{2^{n_2+1}-1}\big)_{L_p(\mathbb{T}^2)}\ll$$
$$\ll\|g_1-V_{2^{n_1},\infty}(g_1)\|_{L_p(\mathbb{T}^2)}+\sup_{|h_1|\leqslant\frac{\pi}{2^{n_1+1}-1}}\|\Delta_{h_1}^{\alpha_1}(V_{2^{n_1},\infty}(g_1))\|_{L_p(\mathbb{T}^2)}=J_{21}+J_{22}.$$

Using lemma 6 (a) and then lemma 3, for a.e. y and $h_1:0<h_1\leqslant\frac{\pi}{2^{n_1+1}-1}$, we have

$$\Big(\int\limits_0^{2\pi}\big|\Delta_{h_1}^{\alpha_1}V_{2^{n_1},\infty}(g_1)\big|^p dx\Big)^{1/p}\ll 2^{-n_1\alpha_1}\Big(\int\limits_0^{2\pi}\big|V_{2^{n_1},\infty}^{(\alpha_1,0)}(g_1)\big|^p dx\Big)^{1/p}\ll$$

$$\ll 2^{-n_1\alpha_1}\Big(\int\limits_0^{2\pi}\big|g_1^{(\alpha_1,0)}\big|^p dx\Big)^{1/p},\quad 1\leqslant p<\infty.$$

Also,

$$\max_{0\leqslant x\leqslant 2\pi}\big|\Delta_{h_1}^{\alpha_1}V_{2^{n_1},\infty}(g_1)\big|\ll 2^{-n_1\alpha_1}\max_{0\leqslant x\leqslant 2\pi}\big|V_{2^{n_1},\infty}^{(\alpha_1,0)}(g_1)\big|\ll 2^{-n_1\alpha_1}\max_{0\leqslant x\leqslant 2\pi}\big|g_1^{(\alpha_1,0)}\big|.$$

Thus, for $0<h_1\leqslant\frac{\pi}{2^{n_1+1}-1}$, $J_{22}\ll 2^{-n_1\alpha_1}\big\|g_1^{(\alpha_1,0)}\big\|_{L_p(\mathbb{T}^2)}$, $\quad 1\leqslant p\leqslant\infty$.

Using Lemmas 3 and 7, and similar arguments, we get for $1\leqslant p\leqslant\infty$

$$J_{21}\ll\sum_{m_1=n_1}^{\infty}\|V_{2^{m_1},\infty}(g_1)-V_{2^{m_1+1},\infty}(g_1)\|_{L_p(\mathbb{T}^2)}\ll$$

$$\ll\sum_{m_1=n_1}^{\infty}2^{-m_1\alpha_1}\|\big(V_{2^{m_1},\infty}(g_1)-V_{2^{m_1+1},\infty}(g_1)\big)^{(\alpha_1,0)}\|_{L_p(\mathbb{T}^2)}\ll$$

$$\ll 2^{-n_1\alpha_1}\big\|g_1^{(\alpha_1,0)}\big\|_{L_p(\mathbb{T}^2)}.$$

Collecting estimates for J_{21} and J_{22} yields $B_2\ll 2^{-n_1\alpha_1}\|g_1^{(\alpha_1,0)}\|_{L_p(\mathbb{T}^2)}$. By definition of the moduli of smoothness, we have $\omega_{\alpha_1,\alpha_2}\big(g_1,\delta_1,\delta_2\big)_{L_p(\mathbb{T}^2)}\ll\omega_{\alpha_1,\alpha_2}\big(g_1,\frac{\pi}{2^{n_1+1}-1},\frac{\pi}{2^{n_2+1}-1}\big)_{L_p(\mathbb{T}^2)}$, we have $J_2\ll 2^{-n_1\alpha_1}\|g_1^{(\alpha_1,0)}\|_{L_p(\mathbb{T}^2)}$. Similarly, $J_3\ll 2^{-n_2\alpha_2}\|g_2^{(0,\alpha_2)}\|_{L_p(\mathbb{T}^2)}$ and $J_4\ll 2^{-n_1\alpha_1-n_2\alpha_2}\|g^{(\alpha_1,\alpha_2)}\|_{L_p(\mathbb{T}^2)}$. Collecting estimates for J_1,J_2,J_3 и J_4, we get

$$\omega_{\alpha_1,\alpha_2}\big(f,\delta_1,\delta_2\big)_{L_p(\mathbb{T}^2)}\ll\|f-g_1-g_2-g\|_{L_p(\mathbb{T}^2)}+\delta_1^{\alpha_1}\|g_1^{(\alpha_1,0)}\|_{L_p(\mathbb{T}^2)}+$$

$$+\delta_2^{\alpha_2}\|g_2^{(0,\alpha_2)}\|_{L_p(\mathbb{T}^2)}+\delta_1^{\alpha_1}\delta_2^{\alpha_2}\|g^{(\alpha_1,\alpha_2)}\|_{L_p(\mathbb{T}^2)}.$$

Since the last inequality holds for any $g_1\in W_p^{(\alpha_1,0)}$, $g_2\in W_p^{(0,\alpha_2)}$, and $g\in W_p^{(\alpha_1,\alpha_2)}$, then

$$\omega_{\alpha_1,\alpha_2}\big(f,\delta_1,\delta_2\big)_{L_p(\mathbb{T}^2)}\ll K(f,\delta_1,\delta_2,\alpha_1,\alpha_2)_{L_p(\mathbb{T}^2)},\tag{5.2}$$

which is the above estimate in (5.1). The proof is now complete.

Remark that for integers $\alpha_i\,(i=1,2)$ theorem 5.1 was proved in [9] when $1 \leqslant p \leqslant \infty$ and in [10] when $p=\infty$.

6 Basic properties of the mixed moduli of smoothness

Theorem 3. *Let* $f \in L_p^0(\mathbb{T}^2)$, $g \in L_p^0(\mathbb{T}^2)$, $1 \leqslant p \leqslant \infty$, $\beta_i \geqslant \alpha_i > 0$, $n_i \in N$, $i=1,2$. *Then*

(1) $\omega_{\alpha_1,\alpha_2}(f,\delta_1,0)_{L_p(\mathbb{T}^2)} = \omega_{\alpha_1,\alpha_2}(f,0,\delta_2)_{L_p(\mathbb{T}^2)} = \omega_{\alpha_1,\alpha_2}(f,0,0)_{L_p(\mathbb{T}^2)} = 0.$

(2) $\omega_{\alpha_1,\alpha_2}(f+g,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)} \ll \omega_{\alpha_1,\alpha_2}(f,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)} + \omega_{\alpha_1,\alpha_2}(g,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}.$

(3) $\omega_{\alpha_1,\alpha_2}(f,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)} \ll \omega_{\alpha_1,\alpha_2}(f,t_1,t_2)_{L_p(\mathbb{T}^2)}$, *for* $0<\delta_i \leqslant t_i, i=1,2.$

(4) $\frac{\omega_{\alpha_1,\alpha_2}(f,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}}{\delta_1^{\alpha_1}\delta_2^{\alpha_2}} \ll \frac{\omega_{\alpha_1,\alpha_2}(f,t_1,t_2)_{L_p(\mathbb{T}^2)}}{t_1^{\alpha_1}t_2^{\alpha_2}}$, *for* $0<t_i\leqslant\delta_i\leqslant\pi, i=1,2.$

(5) $\omega_{\alpha_1,\alpha_2}(f,\lambda_1\delta_1,\lambda_2\delta_2)_{L_p(\mathbb{T}^2)} \ll (\lambda_1+1)^{\alpha_1}(\lambda_2+1)^{\alpha_2}\,\omega_{\alpha_1,\alpha_2}(f,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}$, *for* $\lambda_i > 0, i=1,2.$

(6) $Y_{n_1-1,n_2-1}(f)_{L_p(\mathbb{T}^2)} \ll \omega_{\alpha_1,\alpha_2}(f,\frac{1}{n_1},\frac{1}{n_2})_{L_p(\mathbb{T}^2)}$

$$\ll \frac{1}{n_1^{\alpha_1}}\frac{1}{n_2^{\alpha_2}}\sum_{v_1=1}^{n_1+1}\sum_{v_2=1}^{n_2+1} v_1^{\alpha_1-1}v_2^{\alpha_2-1}Y_{v_1-1,v_2-1}(f)_{L_p(\mathbb{T}^2)}.$$

(7) $\omega_{\beta_1,\beta_2}(f,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)} \ll \omega_{\alpha_1,\alpha_2}(f,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}.$

(8) $\frac{\omega_{\alpha_1,\alpha_2}(f,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}}{\delta_1^{\alpha_1}\delta_2^{\alpha_2}} \ll \frac{\omega_{\beta_1,\beta_2}(f,\delta_1,\delta_2)_{L_p(\mathbb{T}^2)}}{\delta_1^{\beta_1}\delta_2^{\beta_2}}$, *for* $0<\delta_i\leqslant\pi$, $i=1,2.$

The proof of theorem 3 uses theorem 1 and 3.

Remark that using properties (3) and (4) of the mixed moduli of smoothness, in the statement of theorem1 $\omega_{\alpha_1,\alpha_2}\left(f,\frac{\pi}{2n_1-1},\frac{\pi}{2n_2-1}\right)_{L_p(\mathbb{T}^2)}$ can be replaced by $\omega_{\alpha_1,\alpha_2}\left(f,\frac{1}{n_1},\frac{1}{n_2}\right)_{L_p(\mathbb{T}^2)}$.

References

1. S. M. Nikol'skii, *Functions with dominating composite derivatives satisfying multiple Holder conditions*, Sib. Mat. Zh., **4**, No. 4, 1963. Pp. 1342-1364. [Am. Math. Soc. Transl., Ser. 2, 102, 27–51 (1973)].
2. N.S. Bakhvalov, *Embedding theorems for classes of functions with several bounded derivatives*, Vestn. Mosk. Univ., Ser. I **18**, No.3, 1963. Pp. 7-16.
3. S.M. Nikol'skii, *Approximation of functions of several variables and imbedding theorems*. Springer-Verlag, 1975.

4. M. K. Potapov, B.V. Simonov, S. Tikhonov, *Relations between the mixed moduli of smoothness and embedding theorems for Nikolskii classes*, Proceedings of the Steklov Institute of Mathematics, Vol. 269, 197–207, 2010; translation from Russian: Trudy Matem. Inst. V. A. Steklova, Vol. 269, 204–214, 2010.
5. M. K. Potapov, *Approximation by angle and imbedding theorems*, Math. Balkanica, No. 2, 1972. Pp. 183-198.
6. A. Zygmund, *Trigonometric Series*. Cambridge Univ. Press, Cambridge, 1959.
7. R. Taberski, *Differences, moduli and derivatives of fractional orders*, Comment. Math. Prace Mat. 19, no. 2, 1976/77. Pp. 389–400.
8. B.V. Simonov, S.Yu.Tikhonov, *Embedding theorems in constructive approximation*, Sb. Math. 199, No. 9, 2008. Pp. 1367-1407; translation from Mat. Sb. 199, No. 9, 2008. Pp. 107-148.
9. K.V. Runovskii, *Several questions of approxiamtion theory*, Disser. Cand. Nauk, Moscow, MGU, 1989.
10. C. Cottin, *Mixed K-Functionals: A Measure of Smoothness for Blending-type Approximation*, Math. Z. 204, 1990. Pp. 69-83.

M. K. Potapov
Faculty of Mechanics and Mathematics, Moscow State University, Moscow, 119991 Russia, mkpotapov@mail.ru

B. V. Simonov
Volgograd State Technical University, pr. Lenina 28, Volgograd, 400131 Russia, simonov-b2002@yandex.ru

S. Yu. Tikhonov
ICREA, CRM, 08193 Bellaterra, Barcelona, Spain, MSU, tikhonov.work@gmail.com

VIII. Other

SPECTRAL CLUSTERING APPLIED TO HURRICANE TRACK PREDICTION

Maximilian F. Hasler

Key words: Spectral clustering, hurricane track prediction

AMS Mathematics Subject Classification: 62H30, 86A10

Abstract. We apply the method of spectral clustering to the problem of prediction of tracks of Atlantic hurricanes. An emerging trajectory is classified, using this method, among a large number of earlier hurricanes from a historical database [1]. The "similarity" of tracks is determined without imposing ad hoc criteria. We use a Thick-Restart variant of Lanczos' method due to Wu & Simon for the principal component analysis. A few of the closest tracks are selected and finally adjusted by a geometrical transformation, to provide a channel of confidence for the evolution of the given nascent hurricane.

1 Introduction

This project aims at making predictions about the evolution of the trajectory for hurricanes of the atlantic ocean. It is conducted in collaboration with Richard Nock, professor of computer science at Université des Antilles et de la Guyane (University of French West Indies).

The idea is to identify similarities between the tracks of past hurricanes. To achieve this goal, we apply methods of data mining to a large "historical" database of hurricane tracks.

The similarities are detected by spectral clustering, *i.e.* , data partitioning by principal component analysis (PCA), a method of unsupervised classification for statistical data analysis.

1.1 About atlantic hurricanes

Atlantic (or "Cape Verde") hurricanes begin their existence as areas of low pressure or *tropical depressions* near Cape Verde to the west of Africa. They move along the equator to the west and may reinforce to become tropical storms and subsequently hurricanes of various strengths.

The author is grateful to Prof. Richard Nock for introducing him into the subject of unsupervised classification by spectral clustering. We also wish to thank the organizers of the 2011 ISAAC Congress for the invitation to present these results in a stimulating interdisciplinary atmosphere.

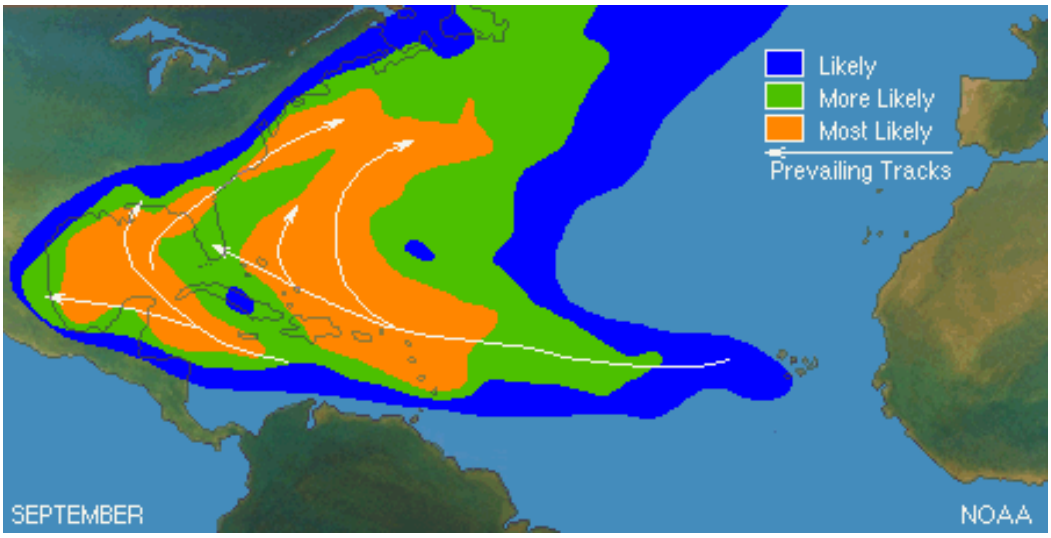


Figure 1. Typical North Atlantic Tropical Hurricanes in September

Then they usually turn to the north, but it makes obviously a big difference whether this happens before they approach the Caribbean islands and/or East coast of the U.S., or after, *i.e.* , over the Caribbean sea, with possible landfall on Haiti, Cuba, Florida or the Mississippi region.

2 Spectral clustering

2.1 The data

We have a collection of data $\{T_j\}_{j\in J}$ which are the hurricane tracks

$$T_j = \{(t_{jk}, x_{jk}, y_{jk});\ k = 1, ..., k_j\}$$

i.e. , sequences of values of time, longitude and latitude, and possibly also other data as wind speed, central pressure, etc.

We constitute our database using as source a file of consolidated data of about 2000 hurricanes ranging back to the 1850's and up to the present, elaborated by specialists of the National Hurricane Center of NOAA [1].

We make a selection of hurricanes from the database to be used for the study: We prefer to retain only the trajectories where the wind (or, equivalently, the central pressure) exceedes a certain threshold, since it cannot be expected that tropical

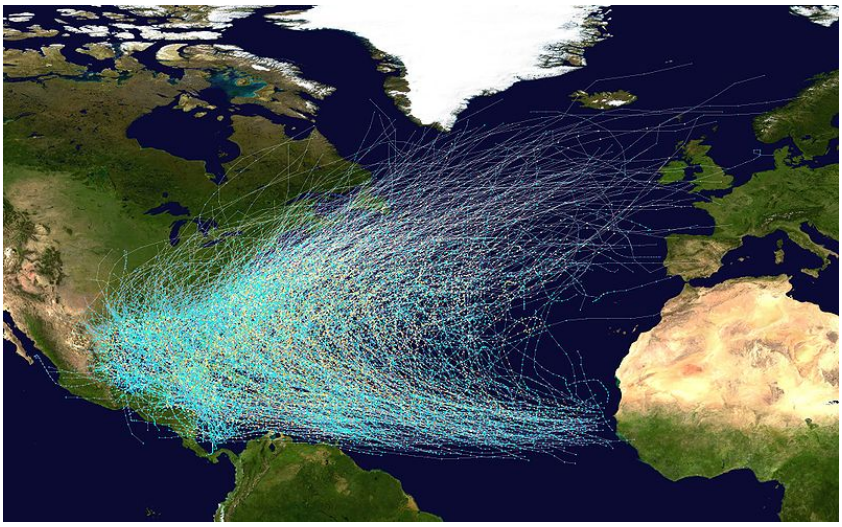


Figure 2. Our database of Atlantic Tropical Hurricanes

storms have dynamics similar to hurricanes, so they would only pollute the results. At the same time, this reduces the size of data to be processed.

2.2 The similarity matrix

We define the similarity matrix

$$S_{ij} = \ f(\, d(T_i, T_j)\,)$$

where $d(\cdot,\cdot)$ measures a distance between two tracks, and f is the so-called *kernel function*. For both functions, different choices are possible.

We chose the kernel function of the form

$$f(x) = e^{-kx^2} \quad \text{or} \quad f(x) = \frac{1}{1 + k\,x^2}\ .$$

The parameter k can be adjusted interactively in our software implementation, which also allows to chose among the first and second of these functions.

For the distance, we initially made the simple, most natral and unbiased choice to use the Euclidean distance of the measured points of the trajectories,

$$d(T_i, T_j) = \sum_{k=1}^{\min(k_i,k_j)} \|(x_{ik} - x_{jk}, y_{ik} - y_{jk})\|^2.$$

2.3 Principal component analysis

Once the similarity matrix is computed, we determine eigenvalues and eigenvectors of the matrix. After discarding the largest eigenvalue, always equal to 1, we consider the position of the objects of our collection, *i.e.* , the hurricane tracks, in the eigenspaces associated to other dominant eigenvalues.

We assume that the trajectories with a neighboring position in this space are of similar kind in nature, and will use this to predict the evolution of an emerging trajectory. To check this hypothesis, we calculate the coordinates of the tracks in the eigenbasis, and create a graphical visualization.We choose 3 among the coordinates $(x_2, x_3, x_4, ...)$ to be kept for this visualization, since is difficult to represent a higher-dimensional space graphically.

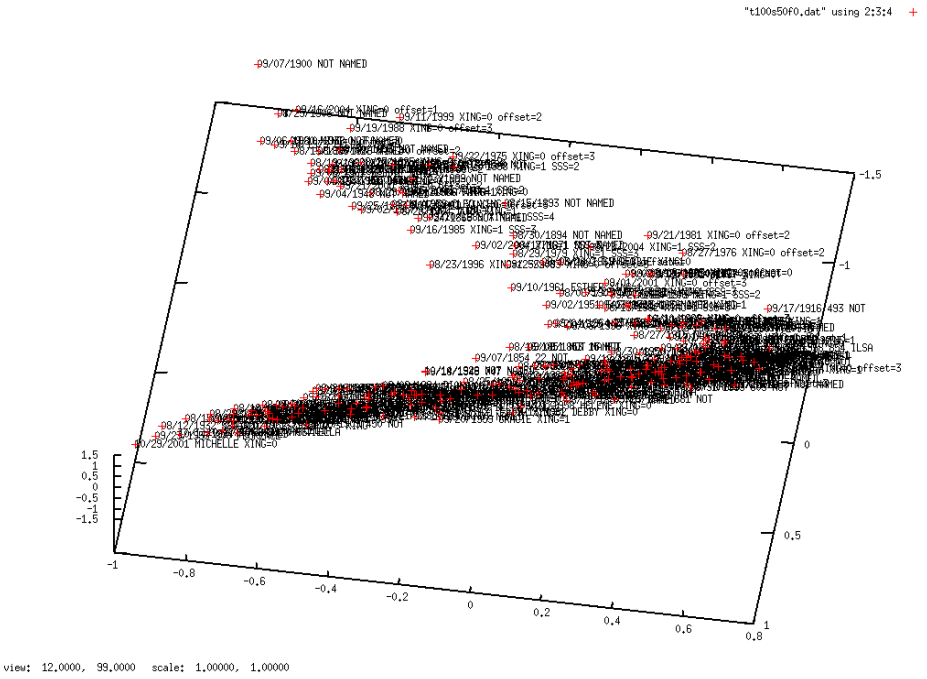


Figure 3. Graphical representation of the database in the space spanned by eigenvectors of principal components of the similarity matrix

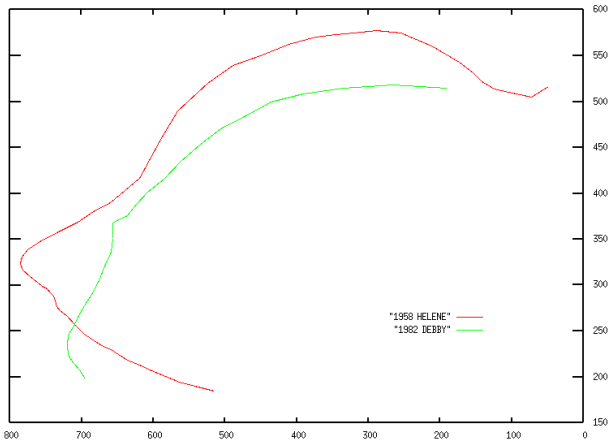


Figure 4. Two tracks located closely together with regard to their principal components

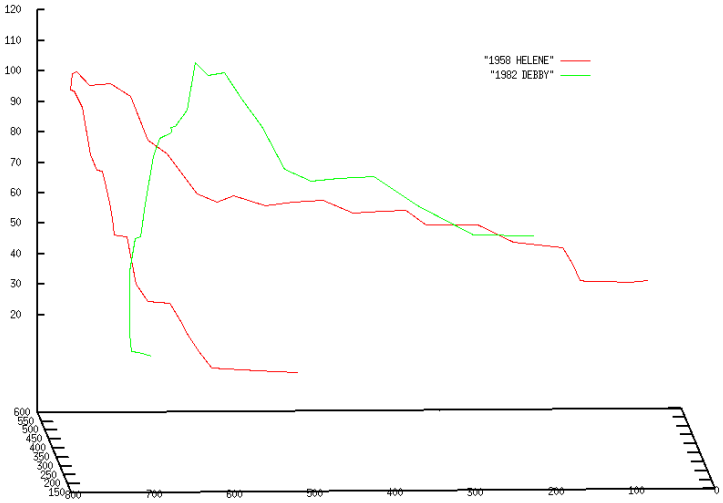


Figure 5. "Similar" tracks turn out not to be only spatially close, but also to have a similar evolution, e.g. concerning the wind speed (plotted as z-values here).

2.4 Lanczos' method

To compute the similarity matrix using the given formula, our program allows to choose interactively among different kernel functions $f(x)$, and to adjust the associated parameters k. It allows also to truncate the data of the individual tracks,

in order to consider it only starting from the point where the wind speed exceeds some adjustable threshold.

It is a nontrivial problem to calculate the eigenvalues and eigenvectors of this similarity matrix of considerable size, $\sim 10^3 \times 10^3$, but there are efficient algorithms for this. We used a variant of Lanczos' method (based on the power method) with "Thick-Restart" [2], which needed, however, to be implemented / interfaced, and adapted to the given problem.

2.5 Normalization of the matrix S

The S matrix must also be normalized, as to get a markovian one,

$$W_{ij} = D_{jj}^{-1} S_{ij} \ , \quad D_{jj} = \sum_j S_{ij} \ ,$$

but then it is no more symmetric, and Lanczos' method cannot be applied directly. Therefore we switch to

$$W' = D^{1/2} W \, D^{-1/2} = D^{-1/2} \, S \, D^{-1/2}$$

and get the eigenvectors as $V = D^{1/2} \, V'$.

In view of applying the method to the prediction of an emerging but still incomplete track, I implemented a procedure to see the evolution of a trajectory in this abstract space, when more and more points are included.

We can see that the chosen hurricane quickly approaches its definitive position in the cloud of all other hurricanes, when its track is less and less truncated. This is obviously important and promising for further consideration.

3 The final step: The prediction

To complete the last step, which is the prediction of a given, supposed to be incomplete track, we have implemented an algorithm which selects a given numbers of tracks "close" to the one to be extrapolated.

In a future version of our algorithm, it is planned to consider rather those close to the point (in the space of principal components) towards which the not yet completely known hurricane appears to be moving.

Usually the trajectories found in the same region are indeed similar. If not, we have several competing hypotheses for the path to expect. For the known paths believed to be relevant, we will then use a geometric transformation that best

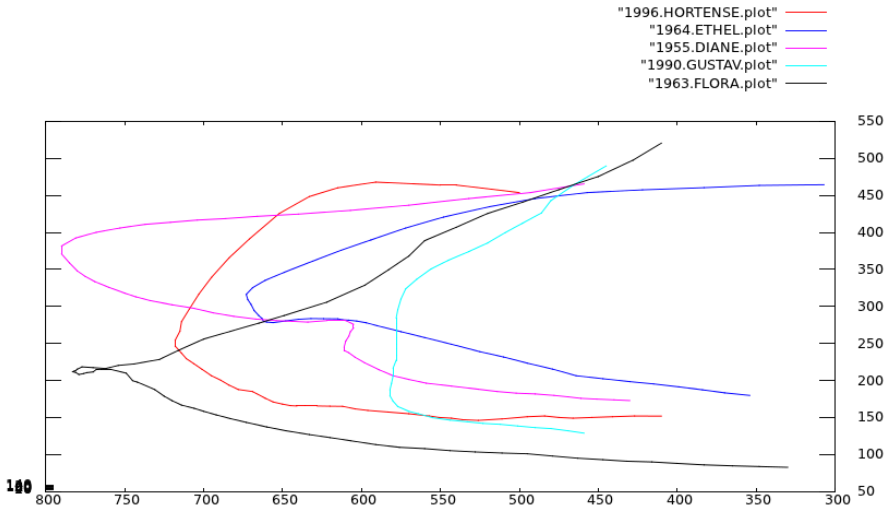


Figure 6. "Similar" tracks before geometrical transformation

approximates the trajectory of the nascent, in order to obtain, finally, a prediction about the spatial evolution of the latter.

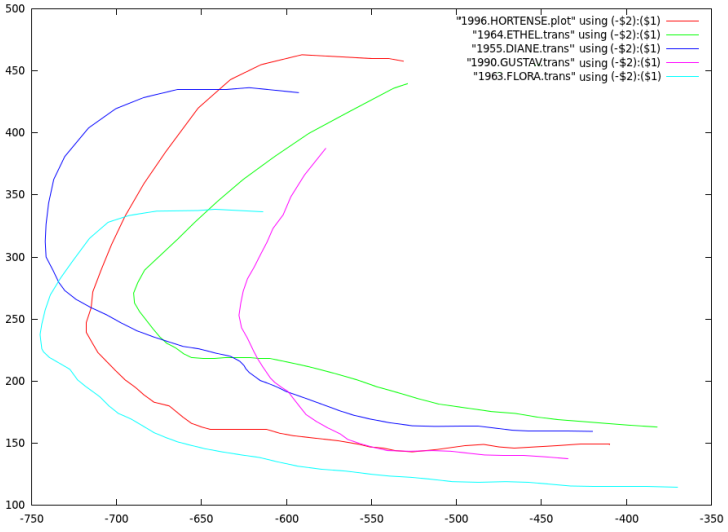


Figure 7. Prediction given as channel spanned by "similar" tracks after geometrical transformation

3.1 Other presentations of this work

Parts of this work have already been presented at other international conferences as follows:

— Mathematical methods for modeling natural risks. V$^{\text{th}}$ Interdisciplinary Congress of Scientific Research, (Santo Domingo, República Dominicana), June 1–5, 2009

— Propagation of singularities in nonlinear PDE and forecast of hurricane tracks. META 2007: Mathematical modelling of tropical and amazonian ecosystems (Kourou, Guyane), Oct. 29–31, 2007.

4 Summary & Conclusions

In this project we apply methods of data mining (PCA; spectral clustering) to a database of trajectories of hurricanes. The project combines conceptual and obvious practical points of interest.

One conceptual novelty is to apply this method to this kind of natural phenomena. The challenge is to try to classify objects for which the "correct" criteria for classification are not known.

The practical interest, namely the prediction of hurricane tracks, is obvious. However, the model is still a bit oversimplified, and several directions should be explored to make it phenomenologically competitive. In particular, probably not only the position (x_i, y_i) of the hurricane should be considered, but also other data (pressure, winds, sea temperature, geography,...). Nevertheless, in spite of the simplicity of the model, the first results are sufficiently remarkable to encourage us to pursue investigating this promising interdisciplinary project.

References

1. "HURDAT". National Hurricane Center, 2010. `http://www.nhc.noaa.gov/pastall.shtml#hurdat`
2. Kesheng Wu and Horst Simon (2000). "Thick-Restart Lanczos Method for Large Symmetric Eigenvalue Problems". SIAM. doi:10.1137/S0895479898334605.

Maximilian F. Hasler
Université des Antilles et de la Guyane, D.S.I. & CEREGMIA,
B.P. 7209, F–97275 Schoelcher (Martinique, F.W.I.)
Phone: +596 696 41 99 41; E-mail: mhasler@univ-ag.fr

AUTHOR INDEX

The 8th Congress of the International Society for Analysis, its Applications, and Computation

Технический редактор *Н. А. Ясько*
Дизайн обложки *М. В. Рогова*
Компьютерная вёрстка *Д. С. Кулябов, А. В. Корольков*

Подписано в печать 20.09.2012 г. Формат 70×100/16.
Бумага офсетная. Печать офсетная. Гарнитура Таймс.
Усл. печ. л. 39,06. Тираж 600 экз. Заказ 1253.

Российский университет дружбы народов
115419, ГСП-1, г. Москва, ул. Орджоникидзе, д. 3

Типография РУДН
115419, ГСП-1, г. Москва, ул. Орджоникидзе, д. 3, тел. 952-04-41