Solid Mechanics and Its Applications

Mikhail Z. Zgurovsky
Victor A. Sadovnichiy *Editors*

Continuous and Distributed Systems

Theory and Applications

Springer

Solid Mechanics and Its Applications

Volume 211

Series Editor

G. M. L. Gladwell, Waterloo, Canada

For further volumes:
http://www.springer.com/series/6557

Aims and Scope of the Series

The fundamental questions arising in mechanics are: *Why*?, *How*?, and *How much*? The aim of this series is to provide lucid accounts written by authoritative researchers giving vision and insight in answering these questions on the subject of mechanics as it relates to solids.

The scope of the series covers the entire spectrum of solid mechanics. Thus it includes the foundation of mechanics; variational formulations; computational mechanics; statics, kinematics, and dynamics of rigid and elastic bodies: vibrations of solids and structures; dynamical systems and chaos; the theories of elasticity, plasticity, and viscoelasticity; composite materials; rods, beams, shells, and membranes; structural control and stability; soils, rocks, and geomechanics; fracture; tribology; experimental mechanics; biomechanics and machine design.

The median level of presentation is the first-year graduate student. Some texts are monographs defining the current state of the field; others are accessible to final year undergraduates; but essentially the emphasis is on readability and clarity.

Mikhail Z. Zgurovsky · Victor A. Sadovnichiy
Editors

Continuous and Distributed Systems

Theory and Applications

Springer

Editors
Mikhail Z. Zgurovsky
National Technical University of Ukraine
"Kyiv Polytechnic Institute"
Kyiv
Ukraine

Victor A. Sadovnichiy
Lomonosov Moscow State University
Moscow
Russia

ISSN 0925-0042 ISSN 2214-7764 (electronic)
ISBN 978-3-319-03145-3 ISBN 978-3-319-03146-0 (eBook)
DOI 10.1007/978-3-319-03146-0
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013953260

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Given collected articles have been organized as a result of joint academic panels of research workers from Faculty of Mechanics and Mathematics of Lomonosov Moscow State University and Institute for Applied System Analysis of the National Technical University of Ukraine "Kyiv Polytechnic Institute," devoted to applied problems of mathematics and mechanics, which attracted attention of researchers from leading scientific schools of Europe, the USA, Russia, Ukraine, and other countries.

Modern technological applications require development and synthesis of fundamental and applied scientific areas, with a view to reducing the gap that may still exist between theoretical basis used for solving complicated technical problems and implementation of obtained innovations. To solve these problems, mathematicians, mechanics, and engineers from Lomonosov Moscow State University and National Technical University of Ukraine "Kyiv Polytechnic Institute" worked together, and results of their joint efforts are partially presented here, including abstract mathematical directions (abstract algebra, number theory, nonlinear functional analysis, partial differential equations, methods of nonlinear and multivalued analysis) and its applications in nonlinear mechanics, decision-making theory and control theory. Also modern mathematical modeling methods for numerical solution of complicated engineer problems are presented as well as their applications in hydromechanics, geophysics, mechanics of continua, quantum mechanics, decision-making theory, etc. In fact, serial publication of such collected papers to similar seminars is planed.

The book is addressed to a wide circle of mathematical, mechanical, and engineering readers.

We want to express the special gratitude to Olena L. Poptsova for a technical support of our collection. Finally, we express our gratitude to editors of the "Springer" Publishing House who worked with collection and everybody who took part in preparation of the manuscript.

Moscow, July 2013 — Mikhail Z. Zgurovsky
Kiev — Victor A. Sadovnichiy

International Editorial Board for this Volume

Contents

Part II Mechanics and Numerical Methods

Contributors

G. I. Arkhipov Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, Russian Federation119991,

V. N. Chubarikov Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, Russian Federation119991,

L. P. Dobrovolskaya Institute of Economics and Management, Moscow, Russian Federation

M. N. Dobrovolsky Geophysical center of RAS, Moscow, Russian Federation

N. M. Dobrovol'skii Tolstoy Tula State Pedagogical University, Tula, Russian Federation

N. N. Dobrovolsky Tula State University, Tula, Russian Federation

A. V. Drutsa Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, Russian Federation119991,

G. V. Fedorov Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, Russian Federation119991,

Eugene A. Feinberg Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600, USA, e-mail: eugene.feinberg@sunysb.edu

Anatoly Fomenko Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, Russian Federation119991,

D. V. Georgievskii Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, Russian Federation, 119991, e-mail: georgiev@mech.math.msu.su

N. M. Glazunov Kiev, Ukraine, e-mail: glanm@yahoo.com

O. O. Gluschenko Institute of Physics and Technology, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kiev 03056, Ukraine

I. M. Gorban Institute of Hydromechanics, National Academy of Sciences of Ukraine, 8/4 Zheliabova St., Kiev 03680, Ukraine, e-mail: ivgorban@gmail.com

Nataliia V. Gorban Institute for Applied System Analysis, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kiev 03056, Ukraine, e-mail: nata_gorban@i.ua

O. V. Khomenko Institute for Applied System Analysis, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kiev 03056, Ukraine

K. A. Ilchuk Institute of Physics and Technology, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kiev 03056, Ukraine

V. I. Ivanenko National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, Kiev 03056, Ukraine, e-mail: viktorivanenko@gmail.com

O. V. Kapustyan Taras Shevchenko National University of Kyiv, 64, Volodymyrs'ka St., Kiev 01601, Ukraine, e-mail: alexkap@univ.kiev.ua

V. O. Kapustyan National Technical University of Ukraine "Kyiv Polytechnic Institute", 37 Prospect Peremogy, Kiev 03056, Ukraine, e-mail: kapustyanv@ukr.net

Pavlo O. Kasyanov Institute for Applied System Analysis, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kiev 03056, Ukraine, e-mail: kasyanov@i.ua

G. M. Kobelkov Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, Russian Federation, 119991, , e-mail: kobelkov@dodo.inm.ras.ru

Peter I. Kogut Department of Differential Equations, Dnipropetrovsk National University, Gagarin av., 72, Dnipropetrovsk 49010, Ukraine, e-mail: p.kogut@i.ua

Andrei Konyaev Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, Russian Federation119991,

O. P. Kupenko Department of System Analysis and Control, National Mining University, Karl Marks av., 19, Dnipropetrovsk 49000, Ukraine; Institute for Applied System Analysis, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kiev 03056, Ukraine, e-mail: kogutolga@bk.ru

Guenter Leugering Institut für Angewandte Mathematik, Lehrstuhl II Universität Erlangen-Nürnberg, Martensstr.3, D-91058 Erlangen, Germany

Lion Lokutsievskiy Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow 119991, Russian Federation, e-mail: lion.lokut@gmail.com

I. L. Makeenko Institute of Physics and Technology, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kiev 03056, Ukraine

O. K. Mazur National Technical University of Ukraine "Kyiv Polytechnic Institute", 37 Prospect Peremogy, Kiev 03056, Ukraine, e-mail: okmazur@ukr.net

V. M. Mikhalevich National University of Kyiv-Mohyla Academy, Skovorody St., 2, Kiev 04070, Ukraine, e-mail: mih@ukma.kiev.ua

N. A. Oriekhova V. Glushkov Institute of Cybernetics, National Academy of Sciences of Ukraine, Glushkova ave., 40, Kiev, Ukraine

V. V. Palin Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow 119991, Russian Federation, e-mail: greystranger84@mail.ru

Liliia S. Paliichuk Institute for Applied System Analysis, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kiev 03056, Ukraine, e-mail: kasyanov@i.ua

N. D. Pankratova Institute for Applied System Analysis, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kiev 03056, Ukraine, e-mail: natalidmp@gmail.com

Roman Samulyak Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600, USA, e-mail: rosamu@ams.sunysb.edu

Vladimir V. Semenov Department of Computational Mathematics, Taras Shevchenko National University of Kyiv, Volodimirska str., 64, Kiev 03601, Ukraine, e-mail: semenov.volodya@gmail.com

S. A. Smirnov Institute of Physics and Technology, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kiev 03056, Ukraine, e-mail: smir@pti.kpi.ua

José Valero Centro de Investigación Operativa, Universidad Miguel Hernandez de Elche, Avda. Universidad s/n, 03202 Elche, Spain, e-mail: jvalero@umh.es

Nina V. Zadoianchuk Department of Computational Mathematics, Taras Shevchenko National University of Kyiv, Volodimirska str., 64, Kiev 03601, Ukraine, e-mail: ninellll@i.ua

Mikhail Z. Zgurovsky Institute for Applied System Analysis, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kiev 03056, Ukraine, e-mail: zgurovsm@hotmail.com

Part I
Abstract Algebra and Applications

Chapter 1
Algebra and Geometry Through Hamiltonian Systems

Anatoly T. Fomenko and Andrei Konyaev

Abstract Hamiltonian systems are considered to be the prime tool of classical and quantum mechanics. The proper investigation of such systems usually requires deep results from algebra and geometry. Here we present several results which in some sense go the opposite way: the knowledge about the integrable system enables us to obtain results on geometric and algebraic structures which naturally appear in such problems. All the results were obtained by employees of the Chair of Differential Geometry and Applications in Moscow State University in 2011–2012.

1.1 Introduction

Hamiltonian systems are common in classical and quantum mehanics. Usually the investigation of such system's properties requires deep results from different fields of algebra, geometry, topology etc. Here we present several results which in some sence go the opposite ways. It means that the study of the objects, which naturally appear in such problems gets some extra perspective from the study of the dynamics of the system itself.

The chapter consists of five parts, each related to some topic in study of Hamiltonian systems. All the results presented were obtained by the employees of the Chair of Differential Geometry and Applications in Moscow State Univercity during the period of 2011–2012.

The first part is dedicated to the study of symmetry groups of atoms, which are the main building block of Fomenko-Ziechang invariants (FZ invariants). These

A. T. Fomenko (✉) · A. Konyaev
Faculty of Mechanics and Mathematics, Lomonosov Moscow State University,
GSP-1, Leninskie Gory, Moscow, Russian Federation 119991
e-mail: fomenko@mech.math.msu.su

A. Konyaev
e-mail: maodzund@bk.ru

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*,
Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_1,

invariants provide the classification for the integrable systems with two degrees of freedom. The main result of this section is due to Fomenko and Kudryavtseva: every finite group can be realized as a symmetry group for some atom. Moreover, there are some restriction on atom's topological complexity in terms of the genus of the atom.

The second part is dedicated to so called integer lattice on bifurcation diagrams. The similar objects naturally appear in quantum mechanics, but study of their classical counterparts is a relatively new topic. The results of this section deal with so called Fomenko hypothesis. It states that the properties of the lattice are closely related to the FZ invariants and, as a result, to the property of the system. Kantonistova studies the system called the Spherical Pendulum. The computation of the ZF invariats for this system is relatively simple which allows the direct verification of the hypothesis which, in this case, as Kantonistova showed, is valid.

The third section deals with integrable billiards, that is the motion of the material point inside two-dimensional domain in this case bounded by segments of the quadric from the same family (this section also contains some extra information about FZ inveriants). The computaion and thorough discription of invariants is done by Fokicheva for a large family of such systems. She showed, that the properties of integrable system are closely related to the shape of the domain.

The fourth section is dedicated to the generalization of classical Bertrand's problem, that is the discription of such potentials that the movement of a particle (moving point) in the corresponding field on a surface of bounded revolution has only periodic trajectories. This problem was formulated in nineteenth century and is yet still interesting. The main result of this section is due to Fedoseev et al.: in two theorems a discriprion of the class of so called Bertrand's manifolds (the manifolds that admit the necessary potentials) is presented.

The final part is the only one that has nothing to do with the integrable systems with FZ invariants. It deals with the classification of linear Poisson structures with generic leaves of dimension two. These structures are of an interest as Hamiltonian systems associated with them have the following property: after the restriction on the leaf every such system is integrable. The other aspect is that two-dimentional leaves are the leaves of the smallest possible non-zero dimension while linear brackets are the simplest possible non-trivial structures. Konyaev proved a theorem that provide a full description od such brackets.

1.2 Atoms and Their Symmetries

The notion of atom was introduced by Fomenko in [17, 19]. Atoms encode typical bifurcations of Liouville tori in non-degenerate integrable Hamiltonian systems. Now many notorious integrable systems with two degrees of freedom and the equivalence classes have been described in terms of two-dimensional atoms and molecules (the set of atoms with additional structure). Moreover many bifurcations of integrable systems in higher dimensions can be represented as semidirect product of two-dimensional atoms [34]. Because the notion of semidirect product uses the

symmetry group of atom, the study of such groups becomes essential. Two-dimensional saddle atoms (2-atoms) are a special class of atoms which can be described in many equivalent ways: f-graphs [35], maps also known as abstract polytopes.

Map (abstract polytope) is an equivalence class of particular cellular decomposition of closed two-dimensional surface up to cellular homeomorphisms. We call two decomposition preserving automorphisms of the surface equivalent iff they differ by the homeomorphism, which sends every cell to itself, preserving any orientation on it. It turns out that every finite group can be realized as a symmetry group of some 2-atom, orientable map or chord diagram. All three approaches are equivalent.

In [5] Fomenko and Bolsinov formulated a question: is it true that every finite group can be realized as symmetry group of bifurcation diagram of some integrable Hamiltonian system or in other terms as symmetry group of some 2-atom? For maximally symmetric atoms (the definition is given below) the description of their symmetries was done by Oshemkov and Brailov in [9]. In [10] Brailov and Kudryavtseva discovered an unexpected link between several of the infinite series of maximally symmetric atoms and stable topological non-conjugacy of integrable Hamiltonian systems. Thorough investigation of the group symmetries of the atoms was done by Fomenko et al. [30, 31].

Siran and Skoviera in [39] notice, that there are a number of results about various classes of combinatorial structures saying that every finite group is the automorphism group of some member of the class. Examples are provided by graphs [22], cubic graphs [23], Steiner triple systems [32], "pictures" [1], and others. Results of this type indicate that a given class is, to some extent, rich. On the other hand, there are some very natural classes that do not have this property, for instance, trees [14]. Similar questions have been asked in connection with graph embeddings on surfaces. As was proven in [11], every finite group is the automorphism group of some map on an orientable surface. However, it is by no means obvious that the same holds for non-orientable maps. The main result of the chapter [39] gives an affirmative answer to this question.

In [29] Fomenko and Kudryavtseva prove that every finite group can be realized as a symmetry group of some orientable two-dimensional atom. The method used in the work differs from the ones described above. In particular the algorithm of the construction of the atom by the given group is presented together with formulas for upper estimates of the atom's genus. The proof utilizes the notion of the covering of the atoms.

Let us recollect some of the basic notions of the atoms theory. Let M be a connected closed two-dimensional surface (orientable or non-orientabe) and $f : M \to R$ is a Morse function with exactly three singular values: maximum, minimum and saddle. We call this kind of function proper Morse function. For such a function its level surface for saddle value can be considered a connected graph K with only degree 4 vertices. The complement to K consists of two-dimensional cells homeomorphic to standard two-dimensional discs. Therefore we have degree 4 cellular decomposition. In particular this means, that such decomposition allows chess coloring, e.g. coloring of the discs in black and white, such that every edge borders discs of different colors.

Proper Morse functions f and f on the surfaces M and M respectively are called leaf-wise equivalent in the neighborhoods P and P of critical levels $\{f = c\}$ and $\{f = c\}$ iff there exist small $\varepsilon > 0$ and $\epsilon > 0$ together with diffeomorphism $D : P = \{|f - c| < e\} \to P' = \{|f' - c'| < e'\}$ such that the connected components of level surfaces of f map with D into the connected components of the level surfaces of f. If D preserves the direction of the growth of the functions f and f then we call them leaf-wise equipped equivalent.

Atom (P, K) is a class of leaf-wise equipped equivalence of Morse function f in the neighborhood $P = \{|f| < \varepsilon\}$ of its saddle critical level $K = \{f = 0\}$. We also call an atom any representative of such class, that is the pair of the surface P with the graph K embedded into it. The complement to the K consists of rings, which we denote positive and negative according to the sign of f. To define the symmetry group of an atom we start with the group of all homeomorphisms of (P, K) onto itself, which preserve the structure of an atom and the direction of the growth of th function. Then we take the quotient by the subgroup which sends every edge of the K into itself with preservation of any given orientation on it. This group is oviously descrete and we call it the symmetry group of an atom. In case of orientable surface P we call a symmetry proper if it preserves the orientation on P. Saddle critical values of the f are called atom vertecies and their number is atom's complexity.

As the boundary of P is disjoint union of circles the surface M with the proper Morse function f described in the beginning of this section is obtained from P by gluing the boundary of some two-dimentional disc to every such circle (different discs are taken for different circles). The proper Morse function on M is a continuation of f such that in every glued disc there's exactly one maximum or minimum in its center. The genus of an atom is the genus of M. Orientable atom is called maximally symmetric [30] if its symmetry group acts transitively on its edges.

There's a natural bijection between atoms and maps (abstract polytopes) [30]. Recollect that [30, 33] starting with a given map we can construct 2-atom (same as proper Morse function on M). That's why all the results about atoms can be if needed reformulated equivalently in terms of maps and their symmetry groups.

These are the main results.

Theorem 1.1 (Fomenko and Kudryavtseva [29]) *Every finite group G is a symmetry group of some orientable two-dimensional atom*

It should be noted that the atom $X(G)$ is constructed explicitly. Moreover there are lots of atoms with symmetry group G, but the properties of the $X(G)$ constructed in the the proof of the theorem can be described in great detail, e.g. there is an upper estimation on the genus of M.

Let k, d be a pair of non-negative integers with condition $d \geq 5$ for any $k \geq 0$ and $d \geq 6$ for $k = 0$. For every such a pair it's possible to construct an orientable atom $T(k, d)$ of genus k with boundary consisting of exactly d circles and trivial symmetry group. Let n be a number of $T(k, d)$ vertices. Define d_- as the number of non-positive boundary circles and d_+ as number of positive boundary circles of $T(k, d)$. Obviously $d = d_+ + d_-$.

Theorem 1.2 (Fomenko and Kudryavtseva [29]) *Let k be a number (not necessary least possible) of generators of a given group G. Among all the atoms with symmetry group isomorphic to G there exists orientable atom $X(G)$ of genus $g = (k-1)|G|+1$ where $|G|$ is the order of the group G. At the same time atom $X(G)$ is G-regular covering of $T(k, d)$ and has $n|G|$ vertices, $d_-|G|$ negative boundary circles and $d_+|G|$ positive boundary circles. Every symmetry of sucn atom preserves its orientation.*

We should notice that the symmetry group G of an atom $X(G)$ acts freely on its vertices and boundary circles, that is every symmetry is either trivial or doesn't have any invariant vertices and boundary circles. More accurate estimates in some special cases given in [29].

1.3 Integer Lattices of Action Variables for "Spherical Pendulum" System

The problem of constructing integer lattices of action variables for integrable Hamiltonian systems with two degrees of freedom is relatively new. It appeared due to Fomenko's hypothesis, which states that the structure of integer lattice of action variables (see the definition below) in "typical case" is completely determined by Fomenko-Ziechang invariants (FZ invariants). On the other hand such integer lattice allows one to calculate at least some of the FZ-invariants of such systems, particularly the marks on the ribs of "molecule" (see [4] and also Sect. 1.4). It should be noted that in this case we need not only the lattice but the level lines of action variables.

This hypothesis was proved in case of the Spherical Pendulum by Kantonistova [26]. She obtained analytical description of the action variables, momentum map and bifurcation diagram. It turns out, that in this case it's possible to describe the algorithm for calculating the monodromy matrix of the isolated singular value and, therefore, the marks of the molecule.

Definition 1.1 Let (M^{2n}, ω, H) be an integrable Hamiltonian system with n degrees of freedom, and $F_1, \ldots, F_n$ are its first integrals ($F_1 = H$). The map $\Phi = (F_1, \ldots, F_n) : M^{2n} \to R^n$ is called *momentum map*.

The approach of Kantonistova is as following. From the Liouville theorem (see [4, Sect. 1.5]) it's well-known that regular connected compact level surface T_{ξ_0} of integrals $F_1, \ldots, F_n$ on Hamiltonian integrable system on M^{2n} described in definition is an $n-$dimensional torus T^n. Moreover there exists a set of coordinates $(I_1, \ldots, I_n, \varphi_1, \ldots, \varphi_n)$ in the neighbourhood $U(T_{\xi_0})$ of the T_{ξ_0} such that I_i depend only on the first integrals, and expressed by the formulas

$$I_i(\xi) = \frac{1}{2\pi} \oint\limits_{\gamma_i(\xi)} \alpha, \tag{1.1}$$

where $\gamma_1(\xi), \ldots, \gamma_n(\xi)$ are 1–cycles on the Liouville torus T_ξ, which form the basis of the homology group $H_1(T_\xi)$ continiously depending on $\xi \in U(\xi_0)$, and α being any 1–form in $U(\xi)$, such that $d\alpha = \omega$. This coordinates are called action-angle variables, I_i being the angle variables.

Definition 1.2 Fix 1–form α on the connected open domain P^{2n} in M^{2n}, such that $d\alpha = \omega$. Let all nonsingular Liouville fibers laying in the domain P^{2n} be compact and connected. Than a set of points in $\Phi(P^{2n})\backslash\Sigma \subset R^n$, formed by the intersections of n level-surfaces $\{\xi \in \Phi(P^{2n})\backslash\Sigma \mid I_1(\xi) = c_1, \ldots, I_n(\xi) = c_n,\ c_i \in Z\}$ of functions I_i, defined above, is called *the integer lattice* $\mathscr{R}$ *o*f action variables (or simply *the lattice*).

In the case of spherical pendulum such 1–form α exists on the cotangent bundle $M^4 = T^*S^2$ and we can assume that $P^4 = M^4$.

Definition 1.3 The Spherical Pendulum is the system which describes the movement of the particle with mass m, confined to the surface of the sphere with radius R in a uniform gravitational field of strength g.

The phase space of the system is

$$T^*S^2 \cong \{(\mathbf{x}, \mathbf{p}) \in R^3 \times R^3 \mid x^2 + y^2 + z^2 = 1,\ xp_x + yp_y + zp_z = 0\}. \quad (1.2)$$

For further calculations it is convenient to introduce the following coordinates on the phase space: φ, $p_\varphi = M_z$, θ, p_θ.

The system has two independent integrals $F_1 = E$ (energy antegral) and $F_2 = p_\varphi = M_z$ (cyclic integral). Hence, the phase space is fibered into two-dimensional surfaces, which are according to Liouville theorem the 2−tori (in regular connected case).

Consider the functions $W(E, M_z, z)$ and $M_z(E)$, where:

$$W(E, M_z, z) = 2(E - z)(1 - z^2) - M_z^2, \quad (1.3)$$

where

$$M_z(E) := \frac{29}{(3 - E^2 + E\sqrt{E^2 + 3})}\sqrt{E + \sqrt{E^2 + 3}}. \quad (1.4)$$

Theorem 1.3 (Kantonistova [26]) *The image of the momentum map for given integrals is the variety* $\{(E, M_z) |\ E \geq -1,\ |M_z| \leq M_z(E)\}$.

The only interesting pairs (E, M_z), are the ones laying in the image of the momentum map.

Theorem 1.4 (Kantonistova [26]) *For all pairs* $(E, M_z) \in \Phi(M^4)$ $I_1(E, M_z) = M_z$ *is an action variable. For all pairs* $\{(E, M_z) \in \Phi(M^4) \mid M_z > 0\}$ *(similar for all pairs* $\{(E, M_z) \in \Phi(M^4) \mid M_z < 0\}$*) the action variable* $I_2(E, M_z)$ *is defined by the formula*

$$I_2(E, M_z) = \frac{1}{\pi} \int_{z_1}^{z_2} \frac{\sqrt{2(E-z)(1-z^2) - M_z^2}}{1-z^2} dz, \tag{1.5}$$

where z_1, z_2 are the roots of the equation $W(E, M_z, z) = 2(E-z)(1-z^2) - M_z^2 = 0$ such that $-1 < z_1 < z_2 < 1$ and $W(E, M_z, z) > 0$ for all $z \in (z_1, z_2)$.

It turns out that there exists exactly two singular values of rank 0, namely $(E, M_z) = (\pm 1, 0)$. In the preimage of every singular value there is exactly one singular point of the momentum map on M^4, and its rank is 0. Moreover, the bifurcation diagram of momentum map of the system is composed of two sets: piecewise smooth curve defined by the equation $|M_z| = M_z(E)$, where

$$M_z(E) := \frac{29}{(3 - E^2 + E\sqrt{E^2+3})} \sqrt{E + \sqrt{E^2+3}}, \quad E \geq -1 \tag{1.6}$$

and isolated singular point with the coordinates $(E, M_z) = (1, 0)$. The singular point of rank 0, corresponding to isolated singular value $(E, M_z) = (1, 0)$, is nonsingular and has a focus–focus type. The singular point of rank 0, corresponding to isolated singular value $(E, M_z) = (-1, 0)$, is nonsingular and has a center–center type.

To calculate the lattice for this system the computer program on C++ (with the help of Wolfram Mathematica 7.0 package) was written by Kantonistova. This program numerically solves the system of equations with respect to variables E and M_z:

$$\begin{cases} I_1(E, M_z) = A \\ I_2(E, M_z) = B \end{cases} \tag{1.7}$$

for all possible pairs $(A, B) \in Z \times Z$.

The result, i.e. the pair of numbers (E, M_z), is drawn on the plane $R^2(E, M_z)$. The set of all pairs which are the solutions of this system forms the required integer lattice of action variables.

The order of coordinate axes on $R^2(E, M_z)$ containing the image of momentum map determines the orientation (i.e. defines the positive direction of circuit) on $R^2(E, M_z)$ and on the image of momentum map. Fix the numbering of functions I_1, I_2 in such way that the orientation given by them in every regular point of the image of momentum map coincide with the orientation induced by the numbering of coordinate axes. Then fix the order of level-lines of action variables I_1, I_2 by the method described above. There appears uniquely defined direction of the circuit around the isolated singular value of the momentum map. Start "basis" (e_1, e_2) of the lattice $\mathscr{R}$, which is not passing through the singular point. Go around the singular value in a closed loop in the positive direction. As the result a new "basis" (e_1', e_2') is derived from (e_1, e_2).

Let $\widetilde{M}$ be the transition matrix between the bases (e_1, e_2) and (e_1', e_2'), where $(e_1', e_2') = \widetilde{M}(e_1, e_2)$. $\widetilde{M}$ is related to monodromy matrix. Recall the definition of monodromy matrix.

Definition 1.4 Let (γ_1, γ_2) be the basis cycles on the torus T_ξ (where ξ is the regular value of momentum map), (γ_1', γ_2') be the result of deformation of the cycles around the singularity. Matrix M such that $(\gamma_1', \gamma_2') = M\dot{(}\gamma_1, \gamma_2)$ is called *the monodromy matrix of singular value of the system with respect to the basis* (γ_1, γ_2) *on the torus.*

Recall that there exists the canonical morphism between the set of cycles on torus T_ξ (where the value ξ is regular) and the set of the integer covectors of the lattice in $\mathscr{R}$. For every such cycle on the torus there exists some function (action variable) on the base of Liouville fibration, which diferential is the desired covector.

Let $(\varepsilon^1, \varepsilon^2)$ and $(\varepsilon^{1'}, \varepsilon^{2'})$ be the covector bases corresponding to vector bases (e_1, e_2) and (e_1', e_2'). Using the morphism above we can reformulate the definition of monodromy matrix.

Definition 1.5 Matrix M such that $(\varepsilon_1', \varepsilon_2') = M\dot{(}\varepsilon_1, \varepsilon_2)$ is called the monodromy matrix of singular value of the system with respect to the covector basis $(\varepsilon_1, \varepsilon_2)$ of the lattice $\mathscr{R}$.

There exits a relation between matrices M and $\widetilde{M}$: $M = \widetilde{M}^{t^{-1}}$. Moreover, for the "Spherical Pendulum" system monodromy matrix corresponding to isolated singular value $(E, M_z) = (1, 0)$ belongs to the conjugation class of matrix $M = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ in the group $\mathrm{SL}(2, Z)$. The main result of is the following.

Theorem 1.5 (Kantonistova [26]) *For the "Spherical Pendulum" system for the energy levels* $-1 < E < 1$ *there is a mark* $r = 0$*, and for the energy levels* $E > 1$ *there is a mark* $r = \frac{1}{2}$.

If $E_0 > 1$ than the level surface $Q^3_{E_0} = \{E = E_0\}$ is diffeomorfic to RP^3, and if $-1 < E_0 < 1$ than the $Q^3_{E_0}$ is diffeomorfic to S^3.

1.4 Billiards in Confocal Quadrics

Let Ω be the domain in R^2, bounded by some quadrics from the family:

$$\frac{x^2}{a - \Lambda} + \frac{y^2}{b - \Lambda} = 1, \quad a > 0,\ b > 0, \tag{1.8}$$

where $\Lambda \in R$ is a numerical parameter. Assume that angles of the boundary of the Ω differ from $\frac{3\pi}{2}$. Consider the dynamical system describing the motion of the material point inside this domain with the reflection rule on the boundary $P = \partial\Omega$ (including the vertices) being the equality of angles before and after reflection.

The phase space of this system is the manifold

$$M^4 := \{(x, v) \mid x \in \Omega, v \in T_x R^2, |v| > 0\}/\sim, \tag{1.9}$$

with the following equivalence relation:

$$(x_1, v_1) \sim (x_2, v_2) \quad \Leftrightarrow \quad x_1 = x_2 \in P, \quad |v_1| = |v_2| \quad \text{and} \quad v_1 - v_2 \perp T_{x_1} P. \tag{1.10}$$

Where $T_x P$ stands for a tangent line in the point x to the smooth part of boundary $P = \partial\Omega$ and $|v|$ is a euclidian length of velocity v.

Theorem 1.6 (Jacobi, Shasles) *Tangent lines to the geodesic curve on the quadric $u \in R^n$ (where u belongs to the family of the confocal quadrics U) are tangent to the $n-1$ different confocal quadrics $(u_1, ...u_{n-1}) \in U$. The set $(u_1, ...u_{n-1})$ is the same for all points of the initial curve.*

Due to this theorem in two-dimensional case the tangent line to each points of trajectory of the billiard is tangent to an ellipse or a hyperbola confocal to the family of quadrics that define the boundry of Ω. In other terms this billiard system has two integrals: velocity $|v|$ and the parameter of the family Λ. Fixing $|v|$ restricts the system onto three-dimensional surface $Q^3 \in M^4$.

Definition 1.6 Let $(M_1^4, \omega_1, f_1, g_1)$ and $(M_2^4, \omega_2, f_2, g_2)$ be two integrable Hamiltonian systems on the simplectic manifolds M_1^4 and M_2^4 with integrals f_1, g_1 and f_2, g_2, respectively. Consider restriction of these systems onto the "isoenergetic" manifolds $Q_1^3 = \{\xi \in M_1^4 \mid f_1(\xi) = c_1\}$ and $Q_2^3 = \{\xi \in M_2^4 \mid f_2(\xi) = c_2\}$. Two Liouville foliations on the Q_1^3 and Q_2^3 are Liouville equivalent iff there exists a diffeomorphism that sends the leaves of the first foliation to those of the second one (with two rather technical restrictions on the orientation) [4].

According to the Liouville theorem the manifold Q^3 is foliated into two-dimentional tori (see [4, Sect. 1.5]) and singular leaves. Consider the base of this Liouville foliation. It is a one-dimensional graph W, which is called rough molecule. The "atoms" describing the corresponding neighborhoods of the singular leaves are the vertices of the graph W. In the classical billiards the following atoms appear: A, A^*, B, C_2, D_1 [4].

All these atoms except atom A^* can be described as the cartesian product of the two-dimensional atom and one-dimensional circle S^1. For A this two-dimensional atom is a disk D^2, for atom B it is a neighborhood of the bouquet of two circles and for C_2 this two-dimensional atom is a neighborhood of two intercecting (in two points) circles on the plane. For the atom D_1 we should take a neighborhood of the three circles intersecting in two points. Finally, the atom A^* is similar to the atom B: we start with the bouguet of two circles. The only difference is that when the neighbourhood of the bouquet is multiplied by S^1 it should be cut in one place and then twisted. As a result we obtain three-dimensional manifold with boundary consisting of two (instead of three in the case of B) two-dimensional tori.

Rough molecule W does not describe the topology of the Liouville foliation completely because it does not have information about the gluing of the singular leaves. To save the information we can choose basis in the fundamental group of the boundary tori of all atoms (according to the proper set of rules, see [4, 18, 20, 21]

for the list) and write transformation matrices called the gluing matrices. These matrices are dependent on the choice of the basis, but it's possible to calculate a set of numbers, which are both independent from the choice of basis and at the same time allow to write gluing matrices after fixing the basis. This numbers are called marks. Together with the graph they form marked molecule, which is an invariant of the Liouville equivalence.

Dragovich and Radnovich calculated these marks for some billiard systems [12] in the domain bounded by confocal quadrics. The work was continued by Fokicheva [15, 16], who not only completed calculation, but also found that some of the original results by Dragovich and Radnovich contained errors. Fokicheva also did the calculations for so called "covering billiards". The latter notion was introduced by Oshemkov and Kudryavtseva

Let us start with the classification of all possible domains Ω. We call domain regular if its boundary doesn't contain horizontal line segments and singular otherwise.

Definition 1.7 Domain Ω bounded by quadrics from the family (1.8) is called *equivalent* to domain Ω' bounded by quadrics from the same family (1.8) iff Ω' obtained from Ω by symmetries via axes and/or continuous change of parameter Λ with the only condition $\Lambda \neq b$.

For all pairs of real numbers $a > b > 0$ there are 7 classes of equivalence of the regular compact domains Ω and 6 classes of singular domains. This is the list of regular domains:

Domain and its notation	Boundary
Ω_1	Ellipse
$\Omega_{1.1.1}$	Ellipse and hyperbola
$\Omega_{1.1.2}$	Ellipse and two hyperboles
$\Omega_{1.2}$	Two ellipses
$\Omega_{1.2.1}$	Two ellipses and two hyperbolas
$\Omega_{1.2.2}$	Two ellipses and two hyperbolas
$\Omega_{1.2.3}$	Two ellipses and two hyperbolas

This is the list of singular domains:

Domain and its notation	Boundary
ω_1	Ellipse and horizontal line
ω_2	Ellipse, horizontal line and hyperbola
ω_3	Ellipse, two hyperbolas and horizontal line
ω_4	Two ellipses, one segment of hyperbola and one horizontal segment
ω_5	Two ellipses and two segments of the horizontal line
ω_6	Two ellipses, one hyperbolic segment and one horizontal segment inside domain

Theorem 1.7 (Fokicheva [15, 16]) *For all 1-connected domains the isoenergetic manifold Q^3 is homeomorphic to sphere S^3, for domain bounded by two ellipses this manifold is homeomorphic to $S^1 \times S^2$*

For the domains which boundary is smooth this fact was mentioned in the article [25]. Consider now the integral value $\Lambda < b$ and the trajectory corresponding to it. This trajectory (or its continuation) is a tangent to the ellipse of the family (1.8) with the parameter Λ and lies on the Liouville torus on the lower edge of the molecule. The trajectory corresponding to the integral value $\Lambda > b$ (or its continuation) is a tangent to the hyperbola and lies on the upper edge of the molecule. The trajectory corresponding to the irregular value $\Lambda = b$ (or its continuation) must go through the focuses of the family. The critical circles on the corresponding leaf lie along the horizontal line. The number V of such circles coincides with the number of the horizontal segments inside the domain. If $V \neq 0$ then the neighborhood of such leaf is a saddle atom. Otherwise this is a torus and the molecule of such system is A–A.

Theorem 1.8 (Fokicheva [15, 16]) *The marks and the saddle atom describing the topology of the billiard system in regular domain are written in the following table. The molecule describing the topology of the billiard system in domain $\Omega_{1.2.1}$ is A–A, with marks $r = 1$ and $\varepsilon = \pm 1$.*

Domain	V	Saddle atom	Lower edges	Upper edges	n
Ω_1	1	B	$r = 0\ \varepsilon = 1$	$r = 0\ \varepsilon = 1$	± 1
$\Omega_{1.1.1}$	1	A^*	$r = 0\ \varepsilon = 1$	$r = 0\ \varepsilon = 1$	± 1
$\Omega_{1.1.2}$	1	B	$r = 0\ \varepsilon = 1$	$r = \infty\ \varepsilon = \pm 1$	
$\Omega_{1.2}$	2	C_2	$r = \infty\ \varepsilon = \pm 1$	$r = 0\ \varepsilon = 1$	
$\Omega_{1.2.1}$	0	–	–	–	
$\Omega_{1.2.2}$	1	B	$r = \infty\ \varepsilon = \pm 1$	$r = 0\ \varepsilon = 1$	
$\Omega_{1.2.3}$	2	D_1	$r = \infty\ \varepsilon = \pm 1$	$r = 0\ \varepsilon = 1$	

The following theorem completes the discription of the topology of billiard systems in confocal quadrics.

Theorem 1.9 (Fokicheva [15, 16]) *The molecule describing the topology of the billiard system in singular domain (except the domain ω_6) is A–A, $r = 0$, $\varepsilon = 1$. For ω_6 the molecule coincides with the molecule for the system in the domain $\Omega_{1.2.2}$.*

The billiard systems admit the following generalization. Consider k copies of the domain $\Omega_{1.2}$ (this domain is bounded by two ellipses) and make a cut along the lower segment of the coordinate line Oy. Then glue cuts by the following rule: the left edge of the cut on the i-th copy is glued to the right edge of the cut on the $i + 1$-th copy. This domain is called Δ_k. If we glue the rest of the edges of the cut together we get the domain Δ'_k.

For the first time the problem of the describing this billiard system was proposed by Oshemkov and Kudryavtseva. Fokicheva showed that the isoenergetic manifold Q^3 for the system in the domain Δ_k is homeomorphic to $S^1 \times S^2$. She also computed that the saddle atom describing the topology of the billiard system in Δ_k is A_{2k-1} [30], marks are same as marks for the molecule for domain $\Omega_{1.2}$. The isoenergetic manifold Q^3 for the system in the domain Δ'_k is homeomorphic to S^3. The saddle atom describing the topology of the billiard system in Δ'_k is D_{2k-1} [30], marks are like marks in the molecule for domain $\Omega_{1.2.3}$.

1.5 Bertrand's Manifolds and Their Properties

Consider the movement of a particle (moving point) in a central potential field on a surface of bounded revolution, that is on a manifold $S \approx (a, b) \times S^1$ with the metric of revolution $ds^2 = dr^2 + f^2(r)d\varphi^2$ in polar coordinates $(r, \varphi mod 2\pi)$ for some arbitrary smooth function $f(r)$. Denote the potential by $V(r)$. The system is Hamiltonian with hamiltonian $H = \frac{p_r^2}{2} + \frac{p_\varphi^2}{2f^2(r)} + V(r)$. This construction appears naturally in many physical and mechanical problems.

Originally this problem was formulated by Bertrand in 1873: *if $S = R^2$ and all the bounded trajectories of the particle are closed (regardless of initial conditions) what can be said about the potential $V(r)$?* This problem was solved by Bertrand himself. Later it was generalized as follows: *consider a class $\mathscr{P}$ of central potentials with certain properties (that is all the potentials yielding closed trajectories for a given class of initial conditions) defined on the surface S; the problem is to find all the pairs $(f \in C^\infty, V \in \mathscr{P})$, i.e. classify all manifolds of revolution allowing a potential of the chosen class to exist and to describe all the corresponding potentials.*

Definition 1.8 Manifolds of revolution equipped with a central potential of class $\mathscr{P}$ are called *Bertrand's $\mathscr{P}-$manifolds*.

For Bertrand's manifolds with such $f(r)$ that $f'(r) \neq 0\ \forall r \in (a, b)$ the latter problem was completely solved by Fedoseev et al. in the work [13] which generalizes results obtained by Bertrand [2], Santoprete [37], Darboux, Libman and others.

In their work Fedoseev et al. considered the following classes of potentials:

Definition 1.9 Let $V(r)$ be a central potential on the surface S. It is called *closing*, if

($\exists$) there exists a nonsingular bounded noncircular orbit γ in S;
($\forall$) every nonsingular bounded orbit in S is closed.

Potential $V(r)$ is called *locally closing*, if

$(\exists)^{loc}$ there exists a strongly stable circular orbit $\{r_0\} \times S^1$ in S;
$(\forall)^{loc}$ for every strongly stable circular orbit $\{r_0\} \times S^1$ in S there exists an $\varepsilon > 0$, that every nonsingular bounded orbit in $[r_0 - \varepsilon, r_0 + \varepsilon] \times S^1$ with kinetic

moment in $(K_0 - \varepsilon, K_0 + \varepsilon)$, is closed; K_0 is the kinetic moment value for the corresponding circular trajectory.

Potential $V(r)$ is called *semi-locally closing*, if the conditions $(\exists)$, $(\forall)^{loc}$ are satisfied as well as the following:

$(\forall)$sloc every nonsingular bounded orbit in $U = [a', b'] \times S^1$ with kinetic moment value equal to $\hat{K}$ is closed, where $a' := \inf r|_\gamma$, $b' := \sup r|_\gamma$, γ is the bounded orbit existing due to $(\exists)$, $\hat{K}$ it's kinetic moment value.

Potential $V(r)$ is called *strongly* (*weakly*) *closing*, if the condition $(\forall)^{loc}$ is satisfied (it's analog for every orbitally stable circular orbit) and the following condition is satisfied: every circle $\{r\} \times S^1$ is strongly (orbitally) stable circular orbit.

The following theorem by Fedoseev et al. gives an explicit solution to the stated above generalized Bertrand problem on surfaces (manifolds) of revolution without "equators" (i.e. points $x \in (a, b)$ such that $f'(x) = 0$).

Theorem 1.10 (Fedoseev et al. [13]) *Consider a manifold of revolution $S \approx (a, b) \times S^1$ with the metric $ds^2 = dr^2 + f^2(r)d\varphi^2$ in polar coordinates $(r, \varphi mod 2\pi)$ and $f' \neq 0$ on (a, b). Then*

(a) the above defined classes of potentials coincide (therefore from now on we call the potentials in question "closing" meaning potentials of all the defined types);

(b) if there exists such a $\xi \in Q_{>0}$ that the following equality $-f'^2(r) + f(r)f'(r)' = -\xi^2$ holds for every $r \in (a, b)$ on the corresponding surface (Bertrand's manifold of type I) that there exists exactly two closing potentials and they are of the form $V_i(r) = (-1)^i A|\theta(r)|^{2-i^2}/i + B$, $i = 1, 2$, where $A > 0$, B are arbitrary constants, $\theta(r) = -\frac{f'(r)}{f(r)}$;

(c) if for every $\xi \in Q_{>0}$ the equality $-f'^2(r) + f(r)f'(r)' = -\xi^2$ is not tautological, there exists a smooth function $\theta = \theta(r)$, $\theta(r) \neq 0$ on (a, b) such that in the coordinates $(\theta, \varphi mod 2\pi)$ the metric can be written as $ds^2 = \frac{d\theta^2}{(\theta^2 + c - t\theta^{-2})^2} + \frac{d\varphi^2}{\mu^2(\theta^2 + c - t\theta^{-2})}$ for some constants $c \in R, t \in R \setminus \{0\}, \mu \in Q_{>0}$ and there also exists exactly one closing potential on the corresponding surface (Bertrand's manifold of type II) which is of the form $V_2(r) = \frac{A}{2\theta(r)} + B$, $A, B \in R$.

Bertrand's manifolds of type I can be described explicitly. They are the rational cones and rational coverings of sphere and hyperbolic plane, that is

$$f(r) = \xi f_c(r - r_0) := \begin{cases} \pm\xi(r - r_0), & c = 0, \\ \frac{\xi}{\sqrt{c}} \sin(\sqrt{c}(r - r_0)), & c > 0, \\ \pm\frac{\xi}{\sqrt{-c}} \mathrm{sh}(\sqrt{-c}(r - r_0)), & c < 0, \end{cases} \tag{1.11}$$

Type II Bertrand's manifolds are classified with the pair of parameters (c, t); the parameter μ is irrelevant to the geometry of the manifold. The following rigor definition follows from the Theorem 1.10:

Definition 1.10 Bertrand's manifold is a manifold $S_{\mu,c,t} \approx \cup_{k=1}^{k_{c,t}} I_{k,c,t} \times S^1$, $I_{k,c,t} \subset (-\infty, 0)$, with coordinates θ, $\varphi mod 2\pi$, and the metric of revolution

$$ds^2_{\mu,c,t} = \frac{d\theta^2}{(\theta^2 + c - t\theta^{-2})^2} + \frac{d\varphi^2}{\mu^2(\theta^2 + c - t\theta^{-2})}. \tag{1.12}$$

where $c, t \in R$, $\mu > 0$. The manifold consist of $k_{c,t}$ connected components. The component corresponding to $k = 1$, is called main, corresponding to $k = 2$—additional. The additional component exists only if $t < 0$. Manifolds with $t = 0$ are called type I manifolds, with $t \neq 0$—type II.

Not all the type II Bertrand's manifolds are real *surfaces* of revolution embedded in R^3. The following result holds:

Theorem 1.11 (Fedoseev et al. [13])

1. *Additional component is never realized as a surface of revolution in* R^3*;*
2. *Main component is realized completely as a surface of revolution for the following values of the parameters* (c, t) *and only for them:* $\{t = 0, c \geq 0, \mu \geq 1\} \cup \{t < 0, c = -2\sqrt{-t}, \mu \geq 2\} \cup \{t < 0, c \geq 0, \mu \geq 1\} \cup \{t < 0, -2\sqrt{-t} < c < 0, \mu \in (1, \tilde{\mu})\} \cup \{t < 0, -2\sqrt{-t} < c < 0, \mu \in [2, \infty)\}$*, where* $\tilde{\mu}$ *is a real positive root of the equation* $-256t + 192tx^2 + (27c^2 + 60t)x^4 + 4tx^6 = 0$.

Partial realization of the main component of type II Bertrand's manifolds was also completely studied by the authors.

Movement in a closing potential field on a Bertrand's manifold is an integrable Hamiltonian system with an additional integral p_φ. Therefore a classical hamiltonian analysis can be performed as well as the construction of Fomenko-Zishang invariants of Liouville's equivalence. This was done by Zagryadsky et al. in [13].

It appears that Bertrand's systems are a simple and natural example of Hamiltonian systems with non-compact atoms. Therefore those systems are a good testing case for the non-compact analog to the theory of Fomenko-Zishang invariants.

It is also possible to generalize this problem further to the "Hamiltonian" study of movement on an arbitrary surface of revolution. Some results concerning the connections between the properties of the function $f(r)$ and the bifurcation diagram, the image of the moment map, atoms and molecules were obtained by Zagryadsky et al.

Bertrand's problem for the manifolds with equators is still under consideration as well as the problem on pseudo-Riemannian manifolds of revolution. On those manifolds the study of dynamics and topology of Liouville foliation is also a natural and promising problem.

1.6 Lie Algebras with Generic Coadjoint Orbits of Dimension Two

The Euler top is a textbook example of integrable system in classical mechanics. One of the reasons for its popularity is the simplicity of the analysis of its dynamics. Hamiltonian representation on dual space to Lie algebra $so(3)$ allows to easily see the effects of small perturbations of "stable rotations", for example. In this case we don't need to write the explicit solution in terms of elliptic functions. Instead we can just look at the intersection of two-dimensional symplectic leaves and level surface of system's Hamiltonian. The intersection is one-dimensional and is an integral trajectory of a system or a disjoint union of such trajectories.

The simplicity is mainly due to the fact that in case of Poisson structure with symplectic leaves of dimension two the restriction of every Hamiltonian system on such leaf is integrable as it needs only one integral that is Hamiltonian itself. Therefore the same approach for let's say dynamical analysis holds for two-dimensional leaves in case of a linear Poisson manifold of higher dimension. This posses a natural question: *what are the Possion brackets with generic symplectic leaves of dimension two or, in other terms, what are the Lie algebras with generic coadjoint orbits of dimension two?* This question was formulated in [8] by Bolsinov et al. The complete answer to the question is given by Konyaev in [27] (see also [38] for similar question for complex Lie algebras in terms of homogeneous spaces). It turns out that it is possible to classify up to isomorphism all real Lie algebras with generic coadjoint orbits of dimension two.

Definition 1.11 We call Poisson bracket with Poisson tensor of rank less or equal to 2 *decomposable* if there exist two vector fields v and w such that Poisson tensor equals to $v \wedge w$.

From the properties of Schouten bracket immediately follows that wedge product of two vector fields defines Poisson bivector iff the distribution spanned by these fields is integrable. This construction gives a lot of simple examples of polynomial bivector fields. For example, if v and w are both linear then their wedge product defines quadratic Poisson bracket.

In linear case the decomposable brackets define a dual space to the series of Lie algebras in the form of semidirect sum $R +_\rho R^n$ via representation ρ. These Lie algebras are solvable. The vector fields, that define their Lie-Poisson tensor can be chosen in the form of one linear field and one constant vector field, that also commute and everywhere independent. As both fields are complete, that is their trajectories exist for all times $-\infty < t < \infty$, and tangent to the symplectic leaves, it can be shown that all the generic leaves are diffeomorphic to the two-dimensional plane R^2.

Main tool for the classification of Lie algebras with generic coadjoint orbits of dimension two is the following result by Konyaev, concerning linear vector field.

Theorem 1.12 (Konyaev [27]) *Consider a pair of linear vector fields v and w on affine space R^n, that are everywhere dependent, i.e. $v \wedge w = 0$. Than at least one of the following is true:*

- *$v = \lambda w$, where λ is a constant,*
- *$v = l(x)a$, $w = m(x)a$, where a is a constant vector and $l(x), m(x)$ are linear functions*

The following theorem provides a complete description of the Lie algebras with generic coadjoint orbits of dimension two. It should be noted that the first infinite series of Lie algebras is a result of the first part of the previous theorem, while the exceptional cases are central extensions of three-dimensional Lie algebras that are not isomorphic to the first infinite series.

Theorem 1.13 (Konyaev [27]) *Up to the direct sum with commutative Lie algebra of orbitrary dimension there exists one infinite series of real Lie algebras with generic coadjoint orbits of dimension two and six exeptional Lie algebras. The exeptional Lie algabras are pairwise non-isomorphic and are not isomorphic to any Lie algebra from the infinite series:*
(1) Semidirect sums $R +_\rho R^n$
(2) Three-dimensional simple Lie algebra $so(3)$
(3) Three-dimensional simple Lie algebra $sl(2)$
(4) Four-dimensional solvable Lie algebra $A_{4,8}$. In the special basis e_1, e_2, e_3, e_4 the commutative relations for this Lie algebra have the form (given only non-zero commutators):

$$[e_2, e_3] = e_1, [e_2, e_4] = e_2, [e_3, e_4] = -e_3$$

(5) Four-dimensional solvable Lie algebra $A_{4,10}$. In special basis e_1, e_2, e_3, e_4 the commutative relations for this Lie algebra have the form (given only non-zero commutators):

$$[e_2, e_3] = e_1, [e_2, e_4] = -e_3, [e_3, e_4] = e_2$$

(6) Five-dimensional solvable Lie algebra $A_{5,3}$. In special basis e_1, e_2, e_3, e_4, e_5 the commutative relations for this Lie algebra have the form (given only non-zero commutators):

$$[e_3, e_4] = e_5, [e_3, e_5] = e_1, [e_4, e_5] = e_3$$

(7) Six-dimensional nilpotent Lie algebra $A_{6,3}$. In special basis $e_1, e_2, e_3, e_4, e_5, e_6$ the commutative relations for this Lie algebra have the form (given only non-zero commutators):

$$[e_1, e_2] = e_6, [e_1, e_3] = e_4, [e_2, e_3] = e_5$$

To complete the classification one needs a theorem, that discribes Lie algebras from the infinite series $R +_\rho R^n$ up to the isomorphism.

Definition 1.12 We call two linear operators P and P' equivalent iff for some non-zero constant μ operators P and $\mu P'$ are adjoint, that is have the same Jordan normal forms.

Definition 1.13 We call two linear representations ρ and ρ' of R in $gl(R^n)$ equivalent iff for any $v \in R$ and $v \neq 0$ operators $\rho(v)$ and $\rho'(v)$ are equivalent.

Theorem 1.14 (Konyaev [27]) *Consider a pair of Lie algebras $R+_\rho R^n$ and $R+_{\rho'} R^n$. They are isomorphic iff the representations ρ and ρ' are equivalent.*

In other words the set of semidirect sums is in bijection with the set of equivalence classes of operators.

In [28, 36] the invariants for the exceptional Lie algebras are given. For the infinite series the invariants are the first integrals of some linear vector fields. They have also been computed see [6, 7]. Recently Konyaev found that these invariant admit a simpler description.

Acknowledgments This work was supported by the Government grant of the Russian Federation for support of research projects implemented by leading scientists, in the Federal State Budget Educational Institution of Higher Professional Education Lomonosov Moscow State University under the agreement No. 11.G34.31.0054.

References

1. Behrendt, G.: Automorphism groups of pictures. J. Graph Theor. **14**, 423–426 (1990)
2. Bertrand, J.: Théorème relatif au mouvement d'un point attiré vers un centre fixe. C.R. Acad. Sci. Paris **77**, 849–853 (1873)
3. Biggs, N., White, A.: Permutation Groups and Combinatorial Structures. London Math. Soc. Lect. Notes, Cambridge University Press, Cambridge (1979)
4. Bolsinov, A.V., Fomenko, A.T.: Integrable Hamiltonian Systems: Geometry, Topology and Classification. Taylor & Francis Group, 752 p. (1999)
5. Bolsinov, A.V., Fomenko, A.T.: Some actual unsolved problems on topology of integrable Hamiltonian systems. Topological methods in theory of Hamiltonian systems, pp. 5–23. Factorial, Moscow (1998)
6. Bolsinov, A.V., Taimanov, I.A.: Integrable geodesic flows on suspensions of automorphisms of tori. Proc. Steklov Inst. Math. **231**, 42–58 (2000)
7. Bolsinov, A.V., Taimanov, I.A.: Integrable geodesic flows with positive topological entropy. Invent. Math. **140**, 639–650 (2000)
8. Bolsinov, A.V., Izosimov, A.M., Konyaev, A.Y., Oshemkov, A.A.: Algebra and topology of integrable systems: research problems. Trudy seminara po vectornomu i tenzornomu analizu **28**, 119–191 (2012)

9. Brailov, Y.A: Algebraic properties of atom symmetries. Topological Methods in Theory of Hamiltonian Systems, pp. 24–40. Factorial, Moscow (1998)
10. Brailov, YuA, Kudryavtseva, E.A.: Stable topological non-conjugacy of Hamiltonian systems on two-dimensional surfaces. Mosc. Univ. Math. Bull. **54**(2), 20–27 (1999)
11. Cori, R., Machi, A.: Construction of maps with prescribed automorphism group. Theor. Comp. Sci. **21**, 91–98 (1982)
12. Dragovic, V., Radnovic, M.: Bifurcations of Liouville tori in elliptical billiards. Regul. Chaotic Dyn. **14**(4–5), 479–494 (2009)
13. Fedoseev, D.A., Kudryavtseva, E.A., Zagryadsky, O.A.: Generalization of Bertrand's theorem to surfaces of revolution (in Russian). Sb. Math. **203**(8), 39–78 (2012)
14. Feinberg, V.Z.: Automorphism groups of trees. Dokl. Akad. Nauk BSSR. **13**, 1065–1067 (1969)
15. Fokicheva, V.V.: Description the topology of the hamiltonian integrable system billiard within an ellipse. Vestn. Moscow. Univ. Math. Mech. **5**, 31–35 (2012)
16. Fokicheva, V.V.: Description the topology of the hamiltonian integrable system "billiard in an domain bounded by the segments of the confocal quadrics". Vestn. Moscow. University. Math. Mech (to appear)
17. Fomenko, A.T.: Morse theory of integrable Hamiltonian systems. Soviet Math. Dokl. **33**(2), 502–506 (1986)
18. Fomenko, A.T.: The symplectic topology of completely integrable Hamiltonian systems. Russian Math. Surv. **44**(1), 181–219 (1989)
19. Fomenko, A.T.: The symplectic topology of completely integrable Hamiltonian systems. Russian Math. Surv. **44**(1), 181–219 (1989)
20. Fomenko, A., Zieschang, H.: On typical topological properties of integrable Hamiltonian systems. Math. USSR-Izv. **32**(2), 385-412 (1989)
21. Fomenko, A., Zieschang, H.: A topological invariant and a criterion for the equivalence of integrable Hamiltonian systems with two degrees of freedom. Math. USSR-Izv. **36**(3), 567–596 (1991)
22. Frucht, R.: Herstellung von Graphen mit vorgegebener abstrakten Gruppe. Comp. Math. **6**, 239–250 (1938)
23. Frucht, R.: Graphs of degree three with a given abstract group. Can. J. Math. **1**, 365–378 (1949)
24. Fujii, K.: A note on finite groups which act freely on closed surfaces. Hiroshima Math. J. **5**, 261–267 (1975), II. Hiroshima Math. J. **6**, 457–463 (1976)
25. Gutkin, E.: Billiard dynamics: a survey with the emphasis on open problem. Regul. Chaot. Dyn. **8**(1), 1–13 (2003)
26. Kantonistova, E.O.: Integer lattices of action variables for the generalized Lagrange case. Vestnik MGU. **1**, 54–58 (2012)
27. Konyaev, A.Y.: Classification of Lie algebras with coadjoint orbits of general position of dimension two. Submitted to Sb. Math.
28. Korotkevich, A.A.: Integrable Hamiltonian systems on low-dimensional Lie algebras. Sb. Math. **200**(12), 1731–1766 (2009)
29. Kudryavtseva, E.A., Fomenko, A.T.: Symmetries groups of nice Morse functions on surfaces [in Russian]. Doklady Akademii Nauk. **446**(6), 615–617 (2012)
30. Kudryavtseva, E.A., Nikonov, I.M., Fomenko, A.T.: Maximally symmetric cell decompositions of surfaces and their coverings. Sb. Math. **199**(9), 1263–1359 (2008)
31. Kudryavtseva, E.A., Nikonov, I.M., Fomenko, A.T.: Symmetric and irreducible abstract polyhedra [in Russian]. V. A. Sadovnichiy Anniv. Coll. Articles, Contemporary Problems of Mathematics and Mechanics. **3**(2), 58–97 (2009)
32. Mendelsohn, E.: On the group of automorphisms of Steiner triple and quadruple systems. J. Combination Theor. (A) **25**, 97–104 (1978)
33. Milnor, J.W.: Morse Theory. Princeton University Press, Princeton (1963)
34. Nguyen, T.Z.: Decomposition of nondegenerate singularities of integrable Hamiltonian systems. Lett. Math. Phys. **33**, 187–193 (1995)

35. Oshemkov, A.A.: Morse functions on two-dimensional surfaces. Encoding features. Proc. Steklov Inst. Math. **205**, 119–127 (1995)
36. Patera, J., Sharp, R., Winternitz, P., Zassenhaus, H.: Invariants of real low dimension Lie algebras. J. Math. Phys. **17**(6), 986–994 (1976)
37. Santoprete, M.: Gravitational and harmonic oscillator potentials on surfaces of revolution. J. Math. Phys. (2008). doi:10.1063/1.2912325
38. Shashkov, S.A.: Commutative homogeneous spaces with one-dimensional stabilizer. Izv. RAN. Ser. Mat. **76**(4), 185–206 (2012) doi:10.1070/IM2012v076n04ABEH002605
39. Siran, J., Skoviera, M.: Orientable and non-orientable maps with given automorphism groups. Aust. J. Combination **7**, 47–53 (1993)

Chapter 2
On Hyperbolic Zeta Function of Lattices

L. P. Dobrovolskaya, M. N. Dobrovolsky, N. M. Dobrovol'skii and N. N. Dobrovolsky

Dedicated to the 95th Birth Anniversary of Nikolai Mikhailovich Korobov (23.11.1917–25.10.2004)

Abstract This chapter provides an overview of the theory of hyperbolic zeta function of lattices. A functional equation for the hyperbolic zeta function of Cartesian lattice is obtained. Information about the history of the theory of the hyperbolic zeta function of lattices is provided. The relations with the hyperbolic zeta function of nets and Korobov optimal coefficients are considered.

2.1 Introduction

The introduction contains necessary definitions, results and historical facts about the appearance of the concepts of the hyperbolic zeta functions of nets and lattices, and gives its general theoretical review. The article is partly based on the monographs

L. P. Dobrovolskaya
Institute of Economics and Management, 10, Veresaeva St., Tula, Russia 300041
e-mail: lbocharova6565@mail.ru

M. N. Dobrovolsky (✉)
Geophysical center of RAS, 3, Molodezhnaya St., Moscow, Russia 119296
e-mail: dobrovolsky.michael@gmail.com

N. M. Dobrovol'skii
Leo Tolstoy Tula State Pedagogical University, 125, Lenina pr., Tula, Russia 300026
e-mail: dobrovol@tspu.tula.ru

N. N. Dobrovolsky
Tula State University, 92, Lenina pr., Tula, Russia 300012
e-mail: nikolai.dobrovolsky@gmail.com

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_2,

[8, 15], but it addresses the given problems from a more unified point of view. The article also utilizes the data from Chap. 6 of the monograph [30].

2.1.1 Lattices

First, we will recall some definitions.

Definition 2.1 Let $\boldsymbol{\lambda_1}, \ldots, \lambda_{\mathbf{m}}$, $m \leq s$ be linearly independent system of vectors from $\mathbb{R}^s$. The set Λ of all vectors $a_1\lambda_1 + \cdots + a_m\lambda_{\mathbf{m}}$, where a_i, $1 \leq i \leq m$ independently run through all integers, is called an m-dimensional lattice in $\mathbb{R}^s$, and the vectors $\lambda_1, \ldots, \lambda_{\mathbf{m}}$ are considered its basis.

If $m = s$, then a lattice is considered complete, otherwise it is incomplete. In this chapter we assume all lattices to be complete. Obviously, $\mathbb{Z}^s$ is a lattice. It is also called the fundamental lattice.

A lattice Λ is called an integer lattice in $\mathbb{R}^s$, if Λ is a sublattice of the fundamental lattice $\mathbb{Z}^s$, i.e.

$$\Lambda = \{m_1\lambda_1 + \cdots + m_s\lambda_{\mathbf{s}} | m_1, \ldots, m_s \in \mathbb{Z}\}$$

and $\lambda_1, \ldots, \lambda_{\mathbf{s}}$ is a linearly independent system of integer vectors.

Definition 2.2 For a lattice Λ there is a dual lattice Λ^*, which is the set

$$\Lambda^* = \{\mathbf{y} \mid \forall \mathbf{x} \in \Lambda \ (\mathbf{y}, \mathbf{x}) \in \mathbb{Z}\}. \tag{2.1}$$

Obviously, a dual lattice Λ^* for a lattice Λ is set by the dual basis $\lambda_1^*, \ldots, \lambda_{\mathbf{s}}^*$, determined by the equations

$$\left(\lambda_{\mathbf{i}}^*, \lambda_{\mathbf{j}}\right) = \delta_{ij} = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases} \tag{2.2}$$

It's easy to see that the fundamental lattice $\mathbb{Z}^s$ coincides with its dual lattice and is also a sublattice of a dual lattice of any integer lattice. Moreover, if $\Lambda_1 \subset \Lambda \subset \mathbb{Z}^s$, then $\mathbb{Z}^s \subset \Lambda^* \subset \Lambda_1^*$; thus, for any $C \neq 0$ we see that $(C\Lambda)^* = \Lambda^*/C$. The equality $\det\Lambda^* = (\det\Lambda)^{-1}$ is true for any lattice.

The set of all s-dimensional complete lattices from $\mathbb{R}^s$ will be denoted as PR_s. The set of $\Lambda + \mathbf{x}$, where $\Lambda \in PR_s$ and $\mathbf{x} \in \mathbb{R}^s$ is called a shifted lattice. The set of all shifted lattices $\Lambda + \mathbf{x}$ from $\mathbb{R}^s$ will be denoted as CPR_s.

Concepts of lattices, shifted lattices and lattice projections on coordinate subspaces let us to discuss various issues of number theory in the uniform language.

E.g., if $(a_j, N) = 1 (1 \leq j \leq s)$, then the set $\Lambda = \Lambda(a_1, \ldots, a_s; N)$ of solutions of the linearly homogeneous comparison is the lattice Λ with $\det\Lambda = N$

$$a_1 \cdot x_1 + \cdots + a_s \cdot x_s \equiv 0 \pmod{N}.$$

If F is a totally real algebraic extension of degree s of the field of rational numbers $\mathbb{Q}$ and $\mathbb{Z}_F$ is a ring of algebraic integers of the field F, then the set $\Lambda(F)$, which has been derived in the following way from $\mathbb{Z}_F$, is an s-dimensional lattice

$$\Lambda(F) = \{(\Theta^{(1)}, \ldots, \Theta^{(s)}) \mid \Theta^{(1)} \in \mathbb{Z}_F\}, \tag{2.3}$$

where $(\Theta^{(1)}, \ldots, \Theta^{(s)})$ is a system of algebraic conjugates, and if d is the discriminant of the field F, then $\det\Lambda(F) = \sqrt{d}$.

These two examples, namely the lattice $\Lambda(a_1, \ldots, a_s; N)$ of solutions of a linear equation and the algebraic lattice $\Lambda(F)$, are the focus of this chapter.

A lot of problems of geometry of numbers are defined in terms of shifted lattices $\Lambda + \mathbf{x}$, norms $N(\mathbf{x}) = |x_1 \cdot \ldots \cdot x_s|$, lattice norm minimum and shifted lattice norm minimum.

For an arbitrary lattice $\Lambda \in PR_s$, a norm minimum is the value

$$N(\Lambda) = \inf_{\mathbf{x} \in \Lambda \setminus \{\mathbf{0}\}} N(\mathbf{x}).$$

For an arbitrary shifted lattice $\Lambda + \mathbf{b} \in CPR_s$, a norm minimum is the value

$$N(\Lambda + \mathbf{b}) = \inf_{\mathbf{x} \in (\Lambda + \mathbf{b}) \setminus \{\mathbf{0}\}} N(\mathbf{x}).$$

Littlewood hypothesis has the following formula in these terms:

for $s > 1$ and any non-zero real numbers $\alpha_1, \ldots, \alpha_s$ for the lattice

$$\Lambda(\alpha_1, \ldots, \alpha_s) = \{(q, q \cdot \alpha_1 + p_1, \ldots, q \cdot \alpha_s + p_s) \mid q, p_1, \ldots, p_s \in \mathbb{Z}\}$$

$$N(\Lambda(\alpha_1, \ldots, \alpha_s)) = 0.$$

Oppenheim hypothesis, from which follows the Littlewood hypothesis, states in lattice terms that

for $s > 2$ any s-dimensional lattice Λ $N(\Lambda) > 0$ is similar to an algebraic lattice.

These two hypotheses are closely related to the Korobov's method of optimal coefficients.

A norm minimum is closely connected with a truncated norm minimum, or a hyperbolic lattice parameter. This is the value ([14, 17])

$$q(\Lambda) = \min_{\mathbf{x} \in \Lambda \setminus \{\mathbf{0}\}} q(\mathbf{x}),$$

which has simple geometrical meaning:

the hyperbolic cross $K_s(T)$ does not contain nonzero points of the lattice Λ with $T < q(\Lambda)$.

A hyperbolic cross is the area

$$K_s(T) = \{\mathbf{x} \mid q(\mathbf{x}) \le T\},$$

where $q(\mathbf{x}) = \overline{x}_1 \cdot \ldots \cdot \overline{x}_s$ is a truncated norm of $\mathbf{x}$, and for a real x we will define $\overline{x} = \max(1, |x|)$ ([31], 1963).

Since $\max(1, N(\mathbf{x})) \le q(\mathbf{x})$, it follows that $\max(1, N(\Lambda)) \le q(\Lambda)$ for any lattice Λ, and the Minkowski's theorem on convex bodies states that

$$q(\Lambda) \le \max(\det \Lambda,\ 1).$$

The issue of calculation of the hyperbolic parameter of the lattice of solutions of a linear equation has been addressed in the article [21].

2.1.2 Exponential Sums of Lattices

We will use $G_s = [0; 1)^s$ to denote a s-dimensional half-open cube. A net is an arbitrary nonempty finite set M in G_s. A net with weights is an ordered pair (M, ρ), where ρ is an arbitrary numerical function on M. For the sake of convenicence, we will indentify a net M with an ordered pair $(M, 1)$, that is, with a net with unit weights: $\rho \equiv 1$.

Definition 2.3 A product of two nets with weights (M_1, ρ_1) and (M_2, ρ_2) in G_s is a net with weights (M, ρ):

$$M = \{\,\{\mathbf{x} + \mathbf{y}\} \,|\, \mathbf{x} \in M_1,\ \mathbf{y} \in M_2\,\}, \quad \rho(\mathbf{z}) = \sum_{\substack{\{\mathbf{x}+\mathbf{y}\}=\mathbf{z},\\ \mathbf{x}\in M_1,\, \mathbf{y}\in M_2}} \rho_1(\mathbf{x})\rho_2(\mathbf{y}),$$

where $\{\mathbf{z}\} = (\{z_1\}, \ldots, \{z_s\})$.

The product of nets with weights (M_1, ρ_1) and (M_2, ρ_2) is denoted by

$$(M_1, \rho_1) \cdot (M_2, \rho_2).$$

Moreover, if $(M, \rho) = (M_1, \rho_1) \cdot (M_2, \rho_2)$, then we will write $M = M_1 \cdot M_2$ assuming that a net M is the product of nets M_1 and M_2 (see [23]).

Definition 2.4 An exponential sum of a net with weights (M, ρ) for an arbitrary integer vector $\mathbf{m}$ is

$$S(\mathbf{m}, (M, \rho)) = \sum_{\mathbf{x}\in M} \rho(\mathbf{x}) e^{2\pi i(\mathbf{m},\mathbf{x})}, \tag{2.4}$$

and a normed exponential sum of a net with weights is

$$S^*(\mathbf{m}, (M, \rho)) = \frac{1}{|M|} S(\mathbf{m}, (M, \rho)).$$

Let $\rho(M) = \sum\limits_{j=1}^{N} |\rho_j|$, then the following trivial estimate is true for all normed exponential sums of a net with weights:

$$|S^*(\mathbf{m}, (M, \rho))| \le \frac{1}{|M|} \rho(M).$$

It is easy to see, that for any nets with weights (M_1, ρ_1) and (M_2, ρ_2) the following equality is true:

$$S(\mathbf{m}, (M_1, \rho_1) \cdot (M_2, \rho_2)) = S(\mathbf{m}, (M_1, \rho_1)) \cdot S(\mathbf{m}, (M_2, \rho_2)). \tag{2.5}$$

Definition 2.5 If the following equality is true:

$$(M, 1) = (M_1, 1) \cdot (M_2, 1),$$

then nets M_1 and M_2 are called coprime nets.

Thus, if M_1 and M_2 are coprime nets then the equation $\mathbf{z} = \{\mathbf{x} + \mathbf{y}\}$ does not have more than one solution for $\mathbf{x} \in M_1$ and $\mathbf{y} \in M_2$. That is why the following equality is only true for coprime nets: $|M_1 \cdot M_2| = |M_1| \cdot |M_2|$.

When $\rho \equiv 1$ we obtain a definition of an exponential sum of a net.

Definition 2.6 An exponential sum of a net M for an arbitrary integer vector $\mathbf{m}$ is the value

$$S(\mathbf{m}, M) = \sum_{\mathbf{x} \in M} e^{2\pi i (\mathbf{m}, \mathbf{x})},$$

and a normed exponential sum of a net is

$$S^*(\mathbf{m}, M) = \frac{1}{|M|} S(\mathbf{m}, M).$$

It is easy to see, that for any coprime nets M_1 and M_2 the following equality is true:

$$S(\mathbf{m}, M_1 \cdot M_2) = S(\mathbf{m}, M_1) \cdot S(\mathbf{m}, M_2). \tag{2.6}$$

Let us take for an arbitrary integer lattice Λ, an integer vector $\mathbf{m}$ and an arbitrary vector $\mathbf{x}$ from a dual lattice Λ^* the following values:

$$\delta_\Lambda(\mathbf{m}) = \begin{cases} 1, & \text{if } \mathbf{m} \in \Lambda, \\ 0, & \text{if } \mathbf{m} \in \mathbb{Z}^s \setminus \Lambda, \end{cases} \qquad \delta^*_\Lambda(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in \mathbb{Z}^s, \\ 0, & \text{if } \mathbf{x} \in \Lambda^* \setminus \mathbb{Z}^s. \end{cases}$$

The $\delta_\Lambda(\mathbf{m})$ is the multidimensional generalisation of the famous Korobov's number-theoretical symbol

$$\delta_N(m) = \begin{cases} 1, & \text{if } m \equiv 0 \pmod N, \\ 0, & \text{if } m \not\equiv 0 \pmod N. \end{cases}$$

Definition 2.7 A generalised parallelepipedal net $M(\Lambda)$ is the set $M(\Lambda) = \Lambda^* \cap G_s$.

For an integer lattice Λ its generalised parallelepipedal net $M(\Lambda)$ is a complete system of residues of a dual lattice Λ^* modulo the fundamental sublattice $\mathbb{Z}^s$. Thus, we have the equality $|M(\Lambda)| = \det \Lambda$.

Definition 2.8 A complete linear multiple exponential sum of an integer lattice Λ is

$$s(\mathbf{m}, \Lambda) = \sum_{\mathbf{x} \in M(\Lambda)} e^{2\pi i(\mathbf{m},\mathbf{x})} = \sum_{\mathbf{x} \in \Lambda^*/\mathbb{Z}^s} e^{2\pi i(\mathbf{m},\mathbf{x})},$$

where $\mathbf{m}$ is an arbitrary integer vector.

It is clear, that for a generalised parallelepipedal net $M(\Lambda)$ the following equality is true: $S(\mathbf{m}, M(\Lambda)) = s(\mathbf{m}, \Lambda)$.

Definition 2.9 A complete linear multiple exponential sum of a dual lattice Λ^* of an integer lattice Λ is

$$s^*(\mathbf{x}, \Lambda) = \sum_{\mathbf{m} \in \mathbb{Z}^s/\Lambda} e^{2\pi i(\mathbf{m},\mathbf{x})} = \sum_{j=1}^{N} e^{2\pi i(\mathbf{m}_j,\mathbf{x})},$$

where $\mathbf{x}$ is an arbitrary vector of the dual lattice Λ^* and $\mathbf{m}_1, \ldots, \mathbf{m}_N$ is a complete system of residues of the lattice $\mathbb{Z}^s$ modulo the sublattice Λ.

The following dual statements are true:

Theorem 2.1 *For $s(\mathbf{m}, \Lambda)$ the following equality is true:*

$$s(\mathbf{m}, \Lambda) = \delta_\Lambda(\mathbf{m}) \cdot \det\Lambda.$$

Theorem 2.2 *For any integer lattice Λ with $\det\Lambda = N$ and for an arbitrary $\mathbf{x} \in \Lambda^*$ the following equality is true:*

$$s^*(\mathbf{x}, \Lambda) = \delta^*_\Lambda(\mathbf{x}) \cdot \det\Lambda.$$

2.1.3 Multidimensional Quadrature Formulas and Hyperbolic Zeta Function of a Net

First works by Korobov were published in 1957–1959 [33–35], where the methods of number theory were applied to the problems of numerical integration of multiple integrals. After the class of periodical functions E_s^α had been defined, it has become possible to use methods of harmonic analysis and the theory of exponential sums (an important branch of analytic number theory) to estimate errors of approximate integration. The history of the creation of this method is presented in the chapter [32].

Banach space E_s^α consists of functions $f(\mathbf{x})$, where each of s variables $x_1, \ldots, x_s$ has a period of one, for which their Fourier series

$$f(\mathbf{x}) = \sum_{\mathbf{m}\in\mathbb{Z}^s} C(\mathbf{m}) e^{2\pi i(m_1 x_1 + \cdots + m_s x_s)} \tag{2.7}$$

comply with the conditions

$$\sup_{\mathbf{m}\in\mathbb{Z}^s} |C(\mathbf{m})|(\overline{m}_1 \ldots \overline{m}_s)^\alpha = \|f(\mathbf{x})\|_{E_s^\alpha} < \infty. \tag{2.8}$$

Clearly, such Fourier series are absolutely convergent, since

$$\|f(\mathbf{x})\|_{l_1} = \sum_{\mathbf{m}\in\mathbb{Z}^s} |C(\mathbf{m})| \le \|f(\mathbf{x})\|_{E_s^\alpha} (1 + 2\zeta(\alpha))^s,$$

and thus for any ($\alpha > 1$) they are continuous functions. Here and hereafter, as usual, $\zeta(\alpha)$ is the Riemann zeta function.

A truncated norm surface with parameter $t \ge 1$ is the set $N_s(t) = \{\mathbf{x} \mid q(\mathbf{x}) = t, \mathbf{x} \ne \mathbf{0}\}$, which is the boundary of the hyperbolic cross $K_s(t)$.

For a natural t on a truncated norm surface there is $\tau_s^*(t)$ of integer nonzero points, where[1]

$$\tau_s^*(t) = \sum_{\mathbf{m}\in N(t)}{}' 1 \tag{2.9}$$

is the number of presentations of the natural number t as $t = \overline{m_1} \ldots \overline{m_s}$.

Using new definitions, we can rewrite the expression for the norm $||f(\mathbf{x})||_{E_s^\alpha}$. The following equality is true:

$$||f(\mathbf{x})||_{E_s^\alpha} = \max\left(|C(\mathbf{0})|, \sup_{t\in\mathbb{N}}\left(t^\alpha \max_{\mathbf{m}\in N(t)} |C(\mathbf{m})|\right)\right).$$

It is easy to see, that an arbitrary periodic function $f(\mathbf{x})$ from $E_s^\alpha(C)$ is bounded in absolute value by $C\,(1 + 2\zeta(\alpha))^s$, and this estimate is achieved by the function

[1] Here and hereafter $\sum'$ denotes summation over systems: $(m_1, \ldots, m_s) \ne (0, \ldots, 0)$.

$$f(\mathbf{x}) = \sum_{\mathbf{m}=-\infty}^{\infty} \frac{C}{(\overline{m}_1 \cdot \ldots \cdot \overline{m}_s)^{\alpha}} e^{2\pi i(\mathbf{m},\mathbf{x})}$$

in the point $\mathbf{x} = \mathbf{0}$.

Obviously, $E_s^{\alpha}(C) \subset E_s^{\beta}(C)$ for $\alpha \geq \beta$. For any periodic function

$$f(\mathbf{x}) \in E_s^{\alpha}(C) \subset E_s^{\beta}(C)$$

the following inequality is true

$$||f(\mathbf{x})||_{E_s^{\alpha}} \geq ||f(\mathbf{x})||_{E_s^{\beta}}.$$

The equality is true only for finite exponential polynomials

$$f(\mathbf{x}) = C(\mathbf{0}) + \sum_{\mathbf{m}\in N(1)} C(\mathbf{m})\, e^{2\pi i(\mathbf{m},\mathbf{x})}.$$

Let us take *the quadrature formula with weights*

$$\int_0^1 \ldots \int_0^1 f(x_1, \ldots, x_s) dx_1 \ldots dx_s = \frac{1}{N} \sum_{k=1}^{N} \rho_k f[\xi_1(k), \ldots, \xi_s(k)] - R_N[f]. \quad (2.10)$$

Here, $R_N[f]$ is the error resulting from the replacement of the integral

$$\int_0^1 \ldots \int_0^1 f(x_1, \ldots, x_s) dx_1 \ldots dx_s$$

with the weighted average value of the function $f(x_1, \ldots, x_s)$, calculated in points

$$M_k = (\xi_1(k), \ldots, \xi_s(k)) \qquad (k = 1, \ldots, N).$$

The set M of points M_k is *a net* M, and the points themselves are *the nodes of the quadrature formula*. The values $\rho_k = \rho(M_k)$ are the weights of the quadrature formula. In this chapter we assume all weights to be real-valued.

Definition 2.10 Zeta function of a net M with weights ρ and parameter $p \geq 1$ is the function $\zeta(\alpha, p|M, \rho)$ defined in the right half-plane $\alpha = \sigma + it$ $(\sigma > 1)$ by the Dirichlet series

$$\zeta(\alpha, p|M, \rho) = \sum_{m_1,\ldots,m_s=-\infty}^{\infty}{}' \frac{|S^*(\mathbf{m}, (M, \rho))|^p}{(\overline{m_1} \ldots \overline{m_s})^{\alpha}} = \sum_{n=1}^{\infty} \frac{S^*(p, M, \rho, n)}{n^{\alpha}}, \quad (2.11)$$

where

$$S^*(p, M, \rho, n) = \sum_{\mathbf{m} \in N(n)} |S^*(\mathbf{m}, (M, \rho))|^p. \tag{2.12}$$

The definition provides us with the following inequality:

$$\zeta(p\alpha, p|M, \rho) \le \zeta^p(\alpha, 1|M, \rho). \tag{2.13}$$

When all the weights are 1, we get the zeta function of a net M with parameter p and denote it as $\zeta(\alpha, p|M)$.

The formula (2.11) provides that the zeta function $\zeta(\alpha, p|M, \rho)$ of a net M with weights ρ and parameter $p \ge 1$ is a Dirichlet series, which converges in the right half-plane $\alpha = \sigma + i \cdot t$ ($\sigma > 1$).

The following two Korobov's generalised theorems on errors of quadrature formulas are true:

Theorem 2.3 *Let the Fourier series of a function $f(\mathbf{x})$ absolutely converge, with $C(\mathbf{m})$ being its Fourier coefficients and $S(\mathbf{m}, (M, \rho))$ be an exponential sum of a lattice with weights, then the following equation is true:*

$$R_N[f] = C(\mathbf{0}) \left(\frac{1}{N} S(\mathbf{0}, (M, \rho)) - 1 \right) + \frac{1}{N} {\sum_{m_1, \dots, m_s = -\infty}^{\infty}}' C(\mathbf{m}) S(\mathbf{m}, (M, \rho)) =$$

$$= C(\mathbf{0}) \left(S^*(\mathbf{0}, (M, \rho)) - 1 \right) + {\sum_{m_1, \dots, m_s = -\infty}^{\infty}}' C(\mathbf{m}) S^*(\mathbf{m}, (M, \rho)) \tag{2.14}$$

and with $N \to \infty$ the error $R_N[f]$ will tend to zero only if the weighted nodes of the quadrature formula are evenly distributed in a s-dimensional unit cube.

Theorem 2.4 *If $f(x_1, \dots, x_s) \in E_s^\alpha(C)$, then the following estimate is true for the error of the quadrature formula:*

$$|R_N[f]| \le C \left| \frac{1}{N} S(\mathbf{0}, (M, \rho)) - 1 \right| + \frac{C}{N} {\sum_{m_1, \dots, m_s = -\infty}^{\infty}}' \frac{|S(\mathbf{m}, (M, \rho))|}{(\overline{m_1} \dots \overline{m_s})^\alpha} =$$

$$= C \left| S^*(\mathbf{0}, (M, \rho)) - 1 \right| + C \cdot \zeta(\alpha, 1|M, \rho), \tag{2.15}$$

where the sum $S(\mathbf{m}, (M, \rho))$ is defined by the equality (2.4). On the class $E_s^\alpha(C)$ this estimate cannot be improved.

The Theorem 2.4 can also be formulated as:

For the norm $\|R_N[f]\|_{E_s^\alpha}$ of the linear functional of the error of approximate integration with quadrature formula (2.10) the following equality is true:

$$\|R_N[f]\|_{E_s^\alpha} = \left|\frac{1}{N}S(\mathbf{0},(M,\rho)) - 1\right| + \frac{1}{N}\sum_{m_1,\ldots,m_s=-\infty}^{\infty}{}' \frac{|S(\mathbf{m},(M,\rho))|}{(\overline{m_1}\ldots\overline{m_s})^\alpha} =$$
$$= \left|S^*(\mathbf{0},(M,\rho)) - 1\right| + \zeta(\alpha,1|M,\rho). \tag{2.16}$$

The method of optimal coefficients has proven to be the most productive for construction for the s-dimensional cube $G_s = [0;1)^s$ of multidimensional quadrature formulas with parallelepipedal nets of the from:

$$\iint_{Gs} f(\mathbf{x})d\mathbf{x} = \frac{1}{N}\sum_{k=1}^{N} f\left(\left\{\frac{a_1 k}{N}\right\},\ldots,\left\{\frac{a_s k}{N}\right\}\right) - R_N(f),$$

where $R_N(f)$ is the error of the quadrature formula, and integers a_j $(a_j, N) = 1$ $(j = 1, .., s)$ are the optimal coefficients, chosen in a special way.

The first algorithms for calculation of optimal coefficients were created by Korobov in 1959. He is also the author of the most efficient and high-performance algorithms we use nowadays (see [38]). These algorithms are based on the lemma on hyperbolic parameter of the lattice of solutions of a linear equation by Gelfand (see [13, 28, 37, 38]). Based on the Korobov's suggestion, Dobrovol'skii and Klepikova have made tables of optimal coefficients for dimensions $s \le 30$ and modulo $N = 2^k \quad 3 \le k \le 22$ [11], which is far beyond the scope of the famous tables by Saltykov. The chapter by Bocharova, Van'kova and Dobrovol'skii [2] describes the modification of the Korobov's algorithm, which allows to find not only one optimal net modulo $N = 2^k$, but the whole class of such lattices. One more class of high-performance algorithms for optimal coefficients calculation has been found in the article [3]. Problems of finding optimal coefficients for combined lattices have been addressed in the articles [22, 39].

A series of important articles on applying divisor theory to the optimal coefficients search for parallelepipedal nets have been produced by Voronin and Timergaliyev (see [41–44]). In fact, these articles describe algorithms for the search of integer lattices with high-value hyperbolic lattice parameter.

In the study of the error of approximate integration for quadrature formulas with parallelepipedal nets on the class of periodical functions E_s^α Korobov in his article [34] for the first time mentions a special case of the hyperbolic zeta function of a lattice $\Lambda = \Lambda(a_1,\ldots,a_s;N)$ for real $\alpha > 1$:

$$\zeta_H(\Lambda|\alpha) = \sum_{m_1,\ldots,m_s=-\infty}^{+\infty}{}' \frac{\delta_N(a_1\cdot m_1 + \cdots + a_s\cdot m_s)}{(\overline{m}_1\cdot\ldots\cdot\overline{m}_s)^\alpha}, \tag{2.17}$$

where the Korobov's symbol $\delta_N(m)$ is defined by the following equalities:

$$\delta_N(m) = \begin{cases} 1 & \text{if} \quad m \equiv 0 \pmod N, \\ 0 & \text{if} \quad m \not\equiv 0 \pmod N, \end{cases}$$

and $(a_j, N) = 1 \quad (j = 1, 2, \ldots, s)$.

The hyperbolic zeta function of a lattice $\Lambda = \Lambda(a_1, \ldots, a_s; N)$ is important, because for the parallelepipedal net $M(\mathbf{a}, N)$, defined by the formula

$$M(\mathbf{a}, N) = \left\{ M_k = \left(\left\{ \frac{a_1 k}{N} \right\}, \ldots, \left\{ \frac{a_s k}{N} \right\} \right) \middle| \, k = 0, \ldots, N-1 \right\},$$

there is an equality $\zeta_H(\Lambda|\alpha) = \zeta(\alpha, 1|M(\mathbf{a}, N))$, i.e. the norm of the linear functional of the error of approximate integration with quadrature formulas with parallelpipedal nets equals the hyperbolic zeta function of the corresponding integer lattice of solutions of a linear equation.

The hyperbolic zeta function of the form (2.17) appears in a lot of articles addressing the estimate of errors of multidimensional quadrature formulas with parallelepipedal nets on the class E_s^α. Specifically, Bakhvalov [1] proved the estimate

$$\zeta_H(\Lambda|\alpha) \ll \frac{(\ln q(\Lambda) + 1)^{s-1}}{q(\Lambda)^\alpha}. \tag{2.18}$$

Korobov ([35], 1959) proved, that for such lattices

$$\zeta_H(\Lambda|\alpha) \gg \frac{\ln^{s-1} \det \Lambda}{(\det \Lambda)^\alpha} \tag{2.19}$$

for any integers $a_1, \ldots, a_s$, which are coprime with N.

There are algorithms for finding $a_1, \ldots, a_s$ such that

$$\zeta_H(\Lambda|\alpha) \ll \frac{\ln^{s\alpha} \det \Lambda}{(\det \Lambda)^\alpha} \quad \text{(Korobov 1960)},$$

$$\zeta_H(\Lambda|\alpha) \ll \frac{\ln^{(s-1)\alpha} \det \Lambda}{(\det \Lambda)^\alpha} \quad \text{(Bakhvalov and Korobov)}. \tag{2.20}$$

In its general form the hyperbolic zeta function of lattices appears in works by Frolov [26, 27]. Frolov's thesis [26] states, that for any $\alpha > 1$ and an arbitrary s-dimensional lattice Λ the series

$$\sum_{\mathbf{x} \in \Lambda}{}' (\overline{x}_1 \cdot \ldots \cdot \overline{x}_s)^{-\alpha}$$

absolutely converges.

Having studied an algebraic lattice of the form (2.3), Frolov proved, that for $t > 1$ and the lattice $\Lambda(t, F) = t\Lambda(F)$ with $\det \Lambda(t, F) = t^s \det \Lambda(F)$ the following estimate is true:

$$\zeta_H(\Lambda(t, F)|\alpha) \ll \frac{\ln^{s-1} \det \Lambda(t, F)}{(\det \Lambda(t, F))^\alpha}. \tag{2.21}$$

The Frolov's method is further developed in works by Bykovskii [4, 5] and by Dobrovol'skii [14, 16]. Construction from the chapter [14] shows, that the methods of Korobov and Frolov are two opposite poles of the theory of quadrature formulas with generalised parallelepipedal nets and special weight-function. At the same time, the problem of calculation of errors of approximate integration by these formulas can be turned into a number-theoretic problem of estimating the hyperbolic zeta function of the corresponding lattice once and for all. There's no need to estimate the norm of linear functional of errors of approximate integration for each new type of generalised parallelepipedal nets all over again.

The problems of integration over modified nets have been addressed in chapters [9, 10].

2.1.4 Hyperbolic Zeta Function of Lattices

The term "hyperbolic zeta function of lattice" has been introduces by Dobrovol'skii in 1984 in his works [14, 16], where systematic study of the function $\zeta_H(\Lambda|\alpha)$ has been started.

Specifically, lower estimates for the hyperbolic zeta function of an arbitrary s-dimensional lattice have been obtained:

$$\begin{cases} \zeta_H(\Lambda|\alpha) \geq C_1(\alpha, s)(\det \Lambda)^{-1}, & \text{if } 0 < \det \Lambda \leq 1, \\ \zeta_H(\Lambda|\alpha) \geq C_2(\alpha, s)(\det \Lambda)^{-\alpha} \ln^{s-1} \det \Lambda, & \text{if } \det \Lambda > 1, \end{cases} \tag{2.22}$$

where $C_1(\alpha, s)$, $C_2(\alpha, s) > 0$ are constants depending only on α and s.

An upper estimate for the hyperbolic zeta function of an s-dimensional lattice has been proven:

$$\begin{cases} \zeta_H(\Lambda|\alpha) \leq C_3(\alpha, s)C_1(\Lambda)^s, & \text{if } q(\Lambda) = 1, \\ \zeta_H(\Lambda|\alpha) \leq C_4(\alpha, s)q^{-\alpha}(\Lambda)(\ln q(\Lambda) + 1)^{s-1}, & \text{if } q(\Lambda) > 1. \end{cases} \tag{2.23}$$

This result is a generalisation of the Bakhvalov's theorem, i.e. the inequality (2.18). The estimate (2.23) provides us with the following conclusions. Specifically, it unconditionally provides us with the result, obtained by Frolov (2.21), as the hyperbolic parameter $q(\Lambda(t, F)) = t^s$ for $t > 1$.

Dobrovol'skii has also proven the following theorem: *for any integer lattice Λ and a natural n we have the following presentation:*

$$\zeta_H(\Lambda|2n) = -1 + (\det \Lambda)^{-1} \sum_{\mathbf{x} \in M(\Lambda)} \prod_{j=1}^{s} \left(1 - \frac{(-1)^n (2\pi)^{2n}}{(2n)!} B_{2n}(x_j)\right), \tag{2.24}$$

where $B_{2n}(x)$ is a Bernoulli polynomial of the order $2n$ and $M(\Lambda)$ is the generalised parallelepipedal net of the lattice Λ, which consists of the points of the dual lattice Λ^*, lying in the s-dimensional half-open unit cube $G_s = [0; 1)^s$;

$$\zeta_H(\Lambda \mid 2n+1) = -1 + \frac{1}{\det \Lambda} \sum_{\mathbf{x} \in M(\Lambda)} \prod_{j=1}^{s} \left(1 - (-1)^n \frac{(2\pi)^{2n+1}}{(2n+1)!} \times \right.$$

$$\left. \times \int_0^1 \frac{B_{2n+1}(\{y+x_j\}) + B_{2n+1}(\{y-x_j\})}{2} \operatorname{ctg}(\pi y)\, dy \right).$$

This theorem points out an analogy between the hyperbolic zeta function of a lattice and the Riemann zeta function, for which

$$\zeta(2n) = (-1)^{n-1} \frac{2^{2n-1}\pi^{2n}}{(2n)!} B_{2n},$$

$$\zeta(2n+1) = (-1)^{n+1} \frac{2^{2n}\pi^{2n+1}}{(2n+1)!} \int_0^1 B_{2n+1}(y) \operatorname{ctg}(\pi y)\, dy.$$

Also, the following equality is true:

$$\zeta(\alpha) = \frac{1}{2} \zeta_H(\mathbb{Z}|\alpha) \quad \alpha = \sigma + it \quad \sigma > 1.$$

The presentation (2.24) unconditionally states that for any integer lattice Λ and an even $\alpha = 2n$ the value of $\zeta_H(\Lambda|2n)$ is a transcendental number.

The formula (2.24) allows to utilize $O(ns \det \Lambda)$ of operations to calculate $\zeta_H(\Lambda|2n)$. In their joint article, Dobrovol'skii, Esayan, Pihtilkov, Rodionova and Ustyan [20] have obtained the formula, which allows to calculate $\zeta_H(\Lambda(a; N)|2)$ using $O(\ln N)$ operations.

For the hyperbolic zeta function of the lattice $\Lambda(t, F)$ Dobrovol'skii, Van'kova and Kozlova in their joint article [12] have obtained the asymptotic formula

$$\zeta_H(\Lambda(t,F)|\alpha) = \frac{2(\det \Lambda(F))^\alpha}{R(s-1)!} \left(\sum_{(w)} \frac{1}{|N(w)|^\alpha} \right) \frac{\ln^{s-1} \det \Lambda(t,F)}{(\det \Lambda(t,F))^\alpha} +$$

$$+ O\left(\frac{\ln^{s-2} \det \Lambda(t,F)}{(\det \Lambda(t,F))^\alpha} \right), \tag{2.25}$$

where R is the regulator of a field F, and in the sum $\sum_{(w)} \frac{1}{|N(w)|^\alpha}$ the summation is over all the main ideals of the ring $\mathbb{Z}_F$.

At the first stage of research (1984–1990), the function $\zeta_H(\Lambda|\alpha)$ had been studied only for real $\alpha > 1$. But the joint articles by Dobrovol'skii, Rebrova and Roshchenya in 1995 ([17, 19]) introduced a new stage of research of the hyperbolic zeta function

$\zeta_H(\Lambda|\alpha)$ of a lattice Λ from different aspects: firstly, as a function of a complex argument α, and secondly, as a function on a metric space of lattices.

Thus, we have the following most general definition of the hyperbolic zeta function of a lattice Λ for a complex α.

Definition 2.11 The hyperbolic zeta function of a lattice Λ is the function $\zeta_H(\Lambda|\alpha)$, $\alpha = \sigma + it$ defined for $\sigma > 1$ by the absolutely convergent series

$$\zeta_H(\Lambda|\alpha) = \sum_{\mathbf{x}\in\Lambda}{}' (\overline{x}_1 \cdot \ldots \cdot \overline{x}_s)^{-\alpha}. \tag{2.26}$$

By Abel's theorem ([6], p. 106) the hyperbolic zeta function of lattices can be represented in the following integral form:

$$\zeta_H(\Lambda|\alpha) = \alpha \int_1^\infty \frac{D(t|\Lambda)dt}{t^{\alpha+1}},$$

where $D(T|\Lambda)$ is the number of nonzero points of the lattice Λ in the hyperbolic cross $K_s(T)$.

First, we note that the hyperbolic zeta function of lattices is a Dirichlet series. Let us give some definitions.

The norm spectrum of a lattice Λ is the set of norm values in the nonzero points of the lattice Λ:

$$N_{sp}(\Lambda) = \{\lambda \mid \lambda = N(\mathbf{x}),\ \mathbf{x} \in \Lambda\backslash\{\mathbf{0}\}\}.$$

Correspondingly, the truncated norm spectrum of a lattice Λ is the set of truncated norm values in the nonzero points of the lattice:

$$Q_{sp}(\Lambda) = \{\lambda \mid \lambda = q(\mathbf{x}),\ \mathbf{x} \in \Lambda\backslash\{\mathbf{0}\}\}.$$

The truncated norm spectrum is a discrete numerical set, i.e.

$$Q_{sp}(\Lambda) = \{\lambda_1 < \lambda_2 < \cdots < \lambda_k < \cdots\} \qquad \lim_{k\to\infty} \lambda_k = \infty.$$

Obviously,

$$N(\Lambda) = \inf_{\lambda\in N_{sp}(\Lambda)} \lambda, \quad q(\Lambda) = \min_{\lambda\in Q_{sp}(\Lambda)} \lambda = \lambda_1.$$

The order of a point of the spectrum is the number of lattice points with the given norm value. If the number of such lattice points is infinite, then we assume that the point of the spectrum has an infinite order. The order of a point λ of the norm spectrum is denoted by $n(\lambda)$, and the order of a point λ of the truncated norm spectrum is denoted by $q(\lambda)$ correspondingly.

The concept of the order of a point of the spectrum provides a better understanding of the definition of the hyperbolic zeta function of a lattice. In it instead of the norm of a point $\mathbf{x}$ appears the truncated norm.

Let us give an example of a lattice Λ, for which the series

$$\sum_{\mathbf{x}\in\Lambda}{}' |x_1 \cdot \ldots \cdot x_s|^{-\alpha}$$

diverges for any $\alpha > 1$.

Actually, let $\Lambda = t\Lambda(F)$ be an algebraic lattice, then

$$\sum_{\mathbf{x}\in\Lambda}{}' |x_1 \cdot \ldots \cdot x_s|^{-\alpha} = \sum_{w\in\mathbb{Z}_F}{}' |t^s \cdot N(w)|^{-\alpha}, \tag{2.27}$$

where $N(w)$ is the norm of an algebraic integer from the ring $\mathbb{Z}_F$. By Dirichlet's unit theorem the series on the right side of the equality (2.27) diverges for any $\alpha > 1$, as the ring $\mathbb{Z}_F$ of algebraic integers of a totally real algebraic number field F of the power s has an infinite number of units ε and for them $|N(\varepsilon)| = 1$. Thus, in this case each point of the spectrum has an infinite order, which leads to the series' divergence for any α.

This example shows that the usage of the truncated norm of the vector $q(\mathbf{x}) = \overline{x}_1 \cdot \ldots \cdot \overline{x}_s$ instead of the norm $N(\mathbf{x}) = |x_1 \cdot \ldots \cdot x_s|$ in the definition of $\zeta_H(\Lambda \mid \alpha)$ has substantial meaning, as it provides absolute convergence of the series of the hyperbolic zeta function of any lattice Λ.

The discrete nature of the truncated norm spectrum provides that the hyperbolic zeta function of an arbitrary lattice Λ can be presented as a Dirichlet series:

$$\begin{aligned}\zeta_H(\Lambda|\alpha) &= \sum_{\mathbf{x}\in\Lambda}{}' (\overline{x}_1 \cdot \ldots \cdot \overline{x}_s)^{-\alpha} = \sum_{\mathbf{x}\in\Lambda}{}' q(\mathbf{x})^{-\alpha} = \sum_{k=1}^{\infty} q(\lambda_k)\lambda_k^{-\alpha} = \\ &= \sum_{\lambda\in Q_{sp}(\Lambda)} q(\lambda)\lambda^{-\alpha}.\end{aligned} \tag{2.28}$$

As $D(T|\Lambda) = 0$ for $T < q(\Lambda)$, then

$$\zeta_H(\Lambda|\alpha) = \alpha \int_{q(\Lambda)}^{\infty} \frac{D(t|\Lambda)dt}{t^{\alpha+1}}.$$

The equality (2.28) provides, that for any complex $\alpha = \sigma + it$ in the right half-plane ($\sigma > 1$) there is a regular function of a complex variable, defined by the series (2.26) and the following inequality is true:

$$|\zeta_H(\Lambda|\alpha)| \le \zeta_H(\Lambda|\sigma).$$

A reasonable question arises, whether the hyperbolic zeta function $\zeta_H(\Lambda|\alpha)$ of an arbitrary lattice Λ can be extended to the whole complex plane. In their works, Dobrovol'skii, Rebrova and Roshchenya ([17, 19]) addressed these issues for PZ_s, i.e. the set of all integer lattices, PQ_s, i.e. the set of all rational lattices, PD_s i.e. the set of all lattices with diagonal matrices. It has been proven, that

for any integer lattice $\Lambda \in PZ_s$ the hyperbolic zeta function $\zeta_H(\Lambda|\alpha)$ is a regular function on all α-plane, excluding the point $\alpha = 1$, where it has a pole of order s.

For any lattice $\Lambda \in PQ_s$ the hyperbolic zeta function $\zeta_H(\Lambda|\alpha)$ is also a regular analytic function on all the α-plane, excluding the point $\alpha = 1$, where it has a pole of order s.

The behavior of the hyperbolic zeta function of lattices on the lattice space has been studied. In particular, it was found that

if a sequence of lattices $\{\Lambda_n\}$ converges to the lattice Λ, then the sequence of the hyperbolic zeta functions of lattices $\zeta_H(\Lambda_n|\alpha)$ converges uniformly to the hyperbolic zeta function of the lattice $\zeta_H(\Lambda|\alpha)$ in any half-plane $\sigma \geq \sigma_0 > 1$.

Another result of this kind can be formulated as follows:

for any point α on the α-plane, except of the point $\alpha = 1$, there is neighborhood $|\alpha - \beta| < \delta$ such that for any lattice $\Lambda = \Lambda(d_1, \ldots, d_s) \in PD_s$

$$\lim_{M \to \Lambda, M \in PD_s} \zeta_H(M|\beta) = \zeta_H(\Lambda|\beta),$$

and this convergence is uniform in the neighborhood of the point α.

The derivation of these results is principally based on the asymptotic formula for the number of points of an arbitrary lattice in the hyperbolic cross as a function of the parameter of the hyperbolic cross. The formula has been obtained by Dobrovol'skii and Roshchenya ([18]):

$$D(T \mid \Lambda) = \frac{2^s T \ln^{s-1} T}{(s-1)! \det \Lambda} + \Theta C(\Lambda) \frac{2^s T \ln^{s-2} T}{\det \Lambda},$$

where $C(\Lambda)$ is an effective constant, calculated through the lattice basis, and $|\Theta| \leq 1$.

Gelfond has already pointed out an important relationship between the value of the hyperbolic parameter $q(\Lambda)$ of a lattice $\Lambda(a_1, \ldots, a_{s-1}, 1; N)$ and the valule

$$Q = \min_{k=1,\ldots,N-1} \overline{k} \cdot \overline{k}_1 \cdot \ldots \cdot \overline{k}_{s-1},$$

where integers $k, k_1, \ldots, k_{s-1}$ comply with the system of equations

$$\begin{cases} k_1 \equiv a_1 \cdot k \\ k_2 \equiv a_2 \cdot k \\ \cdots\cdots\cdots \\ k_{s-1} \equiv a_{s-1} \cdot k \end{cases} \pmod{N}$$

with the lattice of solutions $\Lambda^{(p)}(a_1, \ldots, a_{s-1}, 1; N)$. This result is known as the Gelfond's lemma. It turned out that this relationship manifests itself during the analytic continuation into the left half-plane too.

Theorem 2.5 *In the left half-plane $\alpha = \sigma + it$ $(\sigma < 0)$ the following equalities are true:*

$$\zeta_H(\Lambda(a_1, \ldots, a_{s-1}, 1; N) \mid \alpha) =$$

$$= \sum_{t=1}^{s} M_\alpha^t N^{-\alpha t} \sum_{\mathbf{j}_t \in J_{t,s}} N^{t-1} \zeta(\Lambda^{(p)}(a_{j_1}, \ldots, a_{j_t}; N) \mid 1-\alpha),$$

$$\zeta_H(\Lambda^{(p)}(a_1, \ldots, a_{s-1}, 1; N) \mid \alpha) = -1 + \left(1 + \frac{M_\alpha}{N^\alpha} \zeta(\mathbb{Z} \mid 1-\alpha)\right)^s -$$

$$- \frac{M_\alpha^s}{N^{\alpha s}} \zeta^s(\mathbb{Z} \mid 1-\alpha) + \zeta(\Lambda(a_1, \ldots, a_{s-1}, 1; N) \mid 1-\alpha) \frac{M_\alpha^s N}{N^{\alpha s}},$$

where

$$M(\alpha) = \frac{2\Gamma(1-\alpha)}{(2\pi)^{1-\alpha}} \sin\frac{\pi\alpha}{2}.$$

This theorem provides the following result for the values of the hyperbolic zeta function of these lattices in negative odd points:

Theorem 2.6 *For $\alpha = 1 - 2n$, $n \in \mathbb{N}$ the following equalities are true:*

$$\zeta_H(\Lambda(a_1, \ldots, a_{s-1}, 1; N) \mid \alpha) =$$

$$= \sum_{t=1}^{s} \frac{(-1)^t N^{2nt-t}}{n^t} \sum_{\mathbf{j}_t \in J_{t,s}} \sum_{k_1, \ldots, k_{t-1}=0}^{N-1} \prod_{\nu=1}^{t-1} B_{2n}\left(\left\{\frac{k_\nu a_{j_\nu}}{N}\right\}\right) \times$$

$$\times B_{2n}\left(\left\{\frac{-(a_{j_1}k_1 + \cdots + a_{j_{s-1}}k_{s-1})}{N}\right\}\right),$$

$$\zeta_H(\Lambda^{(p)}(a_1, \ldots, a_{s-1}, 1; N) \mid \alpha) = -1 + \left(1 + \frac{N^{2n-1}B_{2n}}{n}\right)^s -$$

$$- \left(\frac{N^{2n-1}B_{2n}}{n}\right)^s + \left(\frac{1}{n}\right)^s \sum_{k=0}^{N-1} \prod_{j=1}^{s} B_{2n}\left(\left\{\frac{a_j k}{N}\right\}\right),$$

and negative even points are trivial zeroes.

2.1.5 Generalised Hyperbolic Zeta Function of Lattices

Based on the analogy between the hyperbolic zeta function of lattices and the Riemann zeta function, Rebrova in the article [40] studied the generalisation of the hyperbolic zeta function of lattices as an s-dimensional analogue of the Hurwitz zeta function. In her research she tried to answer the questions, naturally arising from such an approach: to what extent can the results regarding the hyperbolic zeta function of a lattice be transferred onto a general case? Can we obtain an analytic continuation of the generalised hyperbolic zeta function of a lattice to the whole complex plane? What is the behaviour of the generalised hyperbolic zeta function of a lattice as a function on the metric lattice space?

Definition 2.12 The generalised hyperbolic zeta function of a lattice Λ is the function $\zeta_H(\Lambda + \mathbf{b}|\alpha)$, defined in the right half-plane $\alpha = \sigma + it \quad (\sigma > 1)$ by the absolutely convergent series

$$\zeta_H(\Lambda + \mathbf{b} \mid \alpha) = \sum_{\mathbf{x}\in\Lambda}{}' (\overline{x_1 + b_1} \cdot \ldots \cdot \overline{x_s + b_s})^{-\alpha} = \sum_{\mathbf{x}\in(\Lambda+\mathbf{b})\setminus\{\mathbf{0}\}} q(\mathbf{x})^{-\alpha}, \qquad (2.29)$$

where $\sum'$ means, that the point $\mathbf{x} = -\mathbf{b}$ is excluded from the summation.

From this point of view, we have to examine the place of shifted lattices and explore the possibility to define metrics on them.

Chapter 2 of the monograph [15] (see also [8]) addresses CPR_s i.e. the set of all shifted lattices $\Lambda(\mathbf{x}) = \Lambda + \mathbf{x}$, where $\Lambda \in PR_s$ is an arbitrary s-dimensional real lattice, and $\mathbf{x} \in R^s$ is an arbitrary vector. A metric is defined on this set.

For the construction of an analytic continuation of the generalised hyperbolic zeta function, a fairly broad class of lattices is allocated—Cartesian lattices. We need the following definitions.

Definition 2.13 A simple Cartesian lattice is a shifted lattice $\Lambda + \mathbf{x}$ of the form

$$\Lambda + \mathbf{x} = (t_1\mathbb{Z} + x_1) \times (t_2\mathbb{Z} + x_2) \times \cdots \times (t_s\mathbb{Z} + x_s),$$

where $t_j \neq 0 \quad (j = 1, \ldots, s)$.

In other words, if the lattice $\Lambda + \mathbf{x}$ is a simple Cartesian lattice then it is the result of the stretching of the fundamental lattice along the axes with coefficients $t_1, \ldots, t_s$ followed by a shift by the vector $\mathbf{x}$.

Definition 2.14 A Cartesian lattice is a shifted lattice, which can be presented as a union of a finite number of simple Cartesian lattices.

Definition 2.15 A Cartesian lattice is a shifted lattice with a shifted sublattice which is a simple Cartesian lattice.

Theorem 2.7 *Definitions 2.14 and 2.15 are equivalent.*

Theorem 2.8 *Any shift of a rational lattice is a Cartesian lattice.*

Two lattices Λ and Γ are considered similar, if

$$\Gamma = D(d_1, \ldots, d_s) \cdot \Lambda, \quad \Lambda = D\left(\frac{1}{d_1}, \ldots, \frac{1}{d_s}\right) \cdot \Gamma,$$

where

$$D(d_1, \ldots, d_s) = \begin{pmatrix} d_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & d_s \end{pmatrix}$$

is an arbitrary diagonal matrix, $d_1 \cdot \ldots \cdot d_s \neq 0$.

The set of all nonsingular real diagonal matrices of an order s will be denoted as

$$D_s(\mathbb{R}) = \{D(d_1, \ldots, d_s) \mid d_1 \cdot \ldots \cdot d_s \neq 0\}.$$

Regarding the operation of matrix multiplication $D_s(\mathbb{R})$ is a multiplicative abelian group.

The set of all unimodular real diagonal matrices $DU_s(\mathbb{R})$ is a subgroup of the group $D_s(\mathbb{R})$. Moreover,

$$D_s(\mathbb{R}) \cong DU_s(\mathbb{R}) \times \mathbb{R}^+,$$

where isomorphism φ between $D_s(\mathbb{R})$ and the direct product $DU_s(\mathbb{R}) \times \mathbb{R}^+$ is given by the rule

$$\varphi(D(d_1, \ldots, d_s)) =$$

$$= \left(D\left(\frac{d_1}{\sqrt[s]{|d_1 \cdot \ldots \cdot d_s|}}, \ldots, \frac{d_s}{\sqrt[s]{|d_1 \cdot \ldots \cdot d_s|}}\right), \sqrt[s]{|d_1 \cdot \ldots \cdot d_s|}\right).$$

Theorem 2.9 *An arbitrary Cartesian lattice is similar to a shifted integer lattice.*

Definition 2.16 An integer lattice Λ is simple, if its projections on any axis coincide with $\mathbb{Z}$.

Theorem 2.10 *Any integer lattice Λ is similar to a simple lattice uniquely determined by the lattice Λ.*

Theorem 2.11 *For any Cartesian lattice Λ there is only one presentation:*

$$\Lambda = D(t_1, \ldots, t_s)\Lambda_0, \quad t_1, \ldots, t_s > 0,$$

where Λ_0 is a simple lattice.

Let $M^*(\Lambda)$ be the set of points of the lattice Λ located in the s-dimensional half-open cube $[0; \det \Lambda)^s$. Thus, for any integer lattice Λ the set $M^*(\Lambda)$ is the complete system of residues of the lattice Λ modulo the sublattice $\det \Lambda \times \mathbb{Z}^s$.

Theorem 2.12 *Let*

$$\mathbf{x}(k_1, \ldots, k_{s-1}) = \left(k_1, \ldots, k_{s-1}, N\left\{\frac{-(a_1k_1 + \cdots + a_{s-1}k_{s-1})}{N}\right\}\right),$$

then for the lattice $\Lambda = \Lambda(a_1, \ldots, a_{s-1}, 1; N)$

$$M^*(\Lambda) = \{\mathbf{x}(k_1, \ldots, k_{s-1}) \mid 0 \le k_\nu \le N-1 \ (\nu = 1, \ldots, s-1)\} \tag{2.30}$$

and the following partition is true:

$$\Lambda(a_1, \ldots, a_{s-1}, 1; N) = \bigcup_{\mathbf{x} \in M^*(\Lambda)} (N\mathbb{Z}^s + \mathbf{x}) =$$

$$= \bigcup_{k_1, \ldots, k_{s-1}=0}^{N-1} (N\mathbb{Z}^s + \mathbf{x}(k_1, \ldots, k_{s-1})). \tag{2.31}$$

Corollary 2.1 *The following partition is true:*

$$\Lambda(a_1, \ldots, a_{s-1}, 1; N) =$$

$$= \bigcup_{k_1, \ldots, k_{s-1}=0}^{N-1} \left(\prod_{j=1}^{s-1}(N\mathbb{Z} + k_j)\right) \times (N\mathbb{Z} - a_1k_1 - \cdots - a_{s-1}k_{s-1}).$$

For the lattice $\Lambda(a_1, \ldots, a_{s-1}, 1; N)$ we will examine its combined lattice $\Lambda^{(p)}(a_1, \ldots, a_{s-1}; N)$ of solutions of the system of linear equations

$$\begin{cases} m_1 \equiv a_1 \cdot m_s \\ m_2 \equiv a_2 \cdot m_s \\ \ldots\ldots\ldots \\ m_{s-1} \equiv a_{s-1} \cdot m_s \end{cases} \pmod{N}. \tag{2.32}$$

For $(a_j, N) = 1$ $(j = 1, \ldots, s-1)$ the lattice $\Lambda^{(p)}(a_1, \ldots, a_{s-1}, 1; N)$ is also simple.

Corollary 2.2 *The following partition is true:*

$$\Lambda^{(p)}(a_1, \ldots, a_{s-1}; N) = \bigcup_{k=0}^{N-1} \left(\prod_{j=1}^{s-1}(N\mathbb{Z} + a_jk)\right) \times (N\mathbb{Z} + k).$$

For an arbitrary shifted lattice $\Lambda + \mathbf{b} \in CPR_s$ a truncated norm minimum, or a hyperbolic parameter, is the value

$$q(\Lambda + \mathbf{b}) = \min_{\mathbf{x} \in (\Lambda + \mathbf{b}) \setminus \{\mathbf{0}\}} q(\mathbf{x}).$$

As $\max(1, N(\mathbf{x})) \leq q(\mathbf{x})$, then $\max(1, N(\Lambda + \mathbf{b})) \leq q(\Lambda + \mathbf{b})$, for any lattice Λ.

The norm spectrum of the shifted lattice $\Lambda + \mathbf{b}$ is the set of norm values in the nonzero points of the shifted lattice $\Lambda + \mathbf{b}$:

$$N_{sp}(\Lambda + \mathbf{b}) = \{\lambda \mid \lambda = N(\mathbf{x}),\ \mathbf{x} \in (\Lambda + \mathbf{b}) \setminus \{\mathbf{0}\}\}.$$

Correspondingly, the truncated norm spectrum of the shifted lattice $\Lambda + \mathbf{b}$ is the set of truncated norm values in the nonzero points of the shifted lattice:

$$Q_{sp}(\Lambda + \mathbf{b}) = \{\lambda \mid \lambda = q(\mathbf{x}),\ \mathbf{x} \in (\Lambda + \mathbf{b}) \setminus \{\mathbf{0}\}\}.$$

Obviously,

$$N(\Lambda + \mathbf{b}) = \inf_{\lambda \in N_{sp}(\Lambda + \mathbf{b})} \lambda,$$

$$q(\Lambda + \mathbf{b}) = \min_{\lambda \in Q_{sp}(\Lambda + \mathbf{b})} \lambda.$$

An order of a point of the spectrum is the number of points of the shifted lattice with the given norm value. If the number of such points of the shifted lattice is infinite, then we assume the point of the spectrum to have an infinite order. The order of a point λ of the spectrum is denoted by $n(\lambda)$, and the order of a point λ of the truncated norm spectrum is denoted by $q(\lambda)$.

The following analogue of the Lemma 1 from the article [17] is true.

Lemma 2.1 *For any lattice $\Lambda + \mathbf{b}$ and any point λ of the truncated norm spectrum $Q_{sp}(\Lambda + \mathbf{b})$ the order of the point λ is finite and $Q_{sp}(\Lambda + \mathbf{b})$—discrete.*

The Lemma 2.1 provides, that

$$Q_{sp}(\Lambda + \mathbf{b}) = \{\lambda_1 < \lambda_2 < \cdots < \lambda_n < \cdots\}$$

and

$$q(\Lambda + \mathbf{b}) = \lambda_1, \quad \lim_{n \to \infty} \lambda_n = \infty.$$

That provides, that the hyperbolic zeta function of an arbitrary shifted lattice $\Lambda + \mathbf{b}$ can be presented as a Dirichlet series:

$$\zeta_H(\Lambda + \mathbf{b} \mid \alpha) = \sum_{\mathbf{x} \in (\Lambda + \mathbf{b}) \setminus \{\mathbf{0}\}} q(\mathbf{x})^{-\alpha} = \sum_{k=1}^{\infty} q(\lambda_k) \lambda_k^{-\alpha} = \sum_{\lambda \in Q_{sp}(\Lambda + \mathbf{b})} q(\lambda) \lambda^{-\alpha}.$$

Theorem 2.13 *For any $\alpha = \sigma + it$ in the right half-plane $\sigma > 1$ the Dirichlet series for $\zeta_H(\Lambda + \mathbf{b} \mid \alpha)$ is absolutely convergent; and in the half-plane $\sigma \geq \sigma_0 > 1$ it is uniformly convergent.*

As for $\alpha = \sigma + it$ and $\sigma \geq \sigma_0 > 0$

$$\sum_{k=1}^{\infty} \left| \frac{q(\lambda_k)}{\lambda_k^{\alpha}} \right| \leq \sum_{k=1}^{\infty} \frac{q(\lambda_k)}{\lambda_k^{\sigma_0}} = \zeta_H(\Lambda + \mathbf{b} \mid \sigma_0),$$

then the Theorem 2.13 provides, that for any complex $\alpha = \sigma + it$ in the right half-plane ($\sigma > 1$) there is a regular function of a complex variable, defined by the series (2.29) and the following inequality is true:

$$|\zeta_H(\Lambda + \mathbf{b} \mid \alpha)| \leq \zeta_H(\Lambda + \mathbf{b} \mid \sigma).$$

Theorem 2.14 *The generalised hyperbolic zeta function of the unidimensional fundamental lattice is an analytic function on the whole α-plane, excluding the point $\alpha = 1$, where it has a pole of order 1 with the residue equal to 2.*

Theorem 2.15 *For an arbitrary shifted unidimensional lattice $\Lambda + b = d\mathbb{Z} + b$ the generalised hyperbolic zeta function $\zeta_H(d \cdot \mathbb{Z} + b \mid \alpha)$ is analytic on the whole α-plane, excluding the point $\alpha = 1$, where it has a pole of order 1 with the residue equal to $\dfrac{2}{\det \Lambda}$.*

Theorem 2.16 *The generalised hyperbolic zeta function $\zeta_H(\Lambda \mid \alpha)$ of any simple Cartesian lattice $\Lambda = \prod_{j=1}^{s}(d_j\mathbb{Z} + a_j)$ is analytic on the whole α-plane, excluding the point $\alpha = 1$, where it has a pole of order s.*

Theorem 2.17 *For any Cartesian lattice Λ the generalised hyperbolic zeta function $\zeta_H(\Lambda + \mathbf{b} \mid \alpha)$ is analytic on the whole α-plane, excluding the point $\alpha = 1$, where it has a pole of order s.*

After that the problem of behavior of the generalised hyperbolic zeta function on the orbit of Cartesian lattices is addressed. Again, we start the examination with the unidimensional case.

Theorem 2.18 *For any point α on the α-plane, excluding the point $\alpha = 1$, there is neighborhood $|\alpha - \beta| < \delta$ such that for any shifted lattice $\Lambda + b \in CPR_1$*

$$\lim_{\Gamma + g \to \Lambda + b} \zeta_H(\Gamma + g \mid \beta) = \zeta_H(\Lambda + b \mid \beta),$$

and this convergence is uniform in the neighborhood of the point α.

Theorem 2.19 *For any point α on the α-plane, excluding the point $\alpha = 1$, there is neighborhood $|\alpha - \beta| < \delta$ such that for any Cartesian lattice $\Lambda + \mathbf{b} \in CPR_s$*

$$\lim_{D(q_1,\ldots,q_s)\cdot\Lambda+\mathbf{g}\to\Lambda+\mathbf{b}} \zeta_H(D(q_1,\ldots,q_s)\cdot\Lambda+\mathbf{g}\mid\beta)=\zeta_H(\Lambda+\mathbf{b}\mid\beta),$$

and this convergence is uniform in the neighborhood of the point α.

2.2 Functional Equation for Hyperbolic Zeta Function of Integer Lattices

The articles [24, 25] utilized a new approach to obtain the functional equation for the hyperbolic zeta function. Earlier, to prove the existence of an analytic continuation of the hyperbolic zeta function of an arbitrary Cartesian lattice only the method of expansion of the integer lattice Λ on sublattice $\det\Lambda\cdot\mathbb{Z}^s$ was used followed by the Hurwitz functional equation. Now exponential sums of a lattice were used, which allowed to apply the known features of Dirichlet series with periodic coefficients. Moreover, the concept of the zeta function helps to simplify the arguments and formulas.

As usual, we will use $N(\mathbf{x})=|x_1\ldots x_s|$ to denote the multiplicative norm of the vector $\mathbf{x}$. It has non-zero values only in points of general position, i.e. points without zero coordinates. Let us present new definitions using the multiplicative norm.

Definition 2.17 The zeta function of a lattice Λ is the function $\zeta(\Lambda|\alpha)$, $\alpha=\sigma+it$, defined for $\sigma>1$ by the series

$$\zeta(\Lambda|\alpha)=\sum_{\mathbf{x}\in\Lambda,\,N(\mathbf{x})\neq 0}|x_1\ldots x_s|^{-\alpha}. \tag{2.33}$$

Generally speaking, there is no zeta function for certain lattices Λ, as the corresponding series can diverge for any value of $\alpha=\sigma+it$ but for an arbitrary Cartesian lattice Λ it is obviously exist for $\sigma>1$.

Also, the hyperbolic zeta function is not homogeneous (as a function of a lattice), while the zeta function of a lattice is homogeneous:

$$\zeta(T\Lambda|\alpha)=T^{-s\alpha}\zeta(\Lambda|\alpha). \tag{2.34}$$

The concept of the zeta function of a lattice is the special case with $\mathbf{b}=\mathbf{0}$ of the concept of the generalised zeta function of a lattice.

Definition 2.18 A generalised zeta function of a lattice Λ is the function $\zeta(\Lambda+\mathbf{b}|\alpha)$, $\alpha=\sigma+it$, defined for $\sigma>1$ by the series

$$\zeta(\Lambda+\mathbf{b}|\alpha)=\sum_{\mathbf{x}\in\Lambda+\mathbf{b},\,N(\mathbf{x})\neq 0}|x_1\ldots x_s|^{-\alpha}. \tag{2.35}$$

It is easy to see, that the hyperbolic zeta function of a lattice Λ is directly defined by the sum of the zeta function of a lattice Λ and the zeta functions of corresponding integer lattices of smaller dimensions, which are obtained by discarding of zero coordinates.

Let

$$J_{t,s} = \{\mathbf{j}_t = (j_1, \ldots, j_s) \mid 1 \le j_1 < \cdots < j_t \le s,\ 1 \le j_{t+1} < \cdots < j_s \le s,$$

$$\{j_1, \ldots, j_s\} = \{1, 2, \ldots, s\}\}.$$

In other words, the set $J_{t,s}$ consists of integer vectors $\mathbf{j}_t$, coordinates of which form a permutation of numbers from 1 to s, while coordinates from 1 to t and from $t+1$ to s form increasing sequences.

If we denote the coordinate subspace as $\Pi(\mathbf{j}_t)$

$$\Pi(\mathbf{j}_t) = \{\mathbf{x} \mid x_{j_\nu} = 0\ (\nu = t+1, \ldots, s)\},$$

and denote the projection of intersection of $(\Lambda + \mathbf{a}) \bigcap \Pi(\mathbf{j}_t)$ on $\mathbb{R}^t$ as $(\Lambda + \mathbf{a})_{\mathbf{j}_t}$, then for any shifted lattice the following equality is true:

$$\zeta_H(\Lambda + \mathbf{a} \mid \alpha) = \sum_{t=1}^{s} \sum_{\mathbf{j}_t \in J_{t,s}} \zeta((\Lambda + \mathbf{a})_{\mathbf{j}_t} \mid \alpha).$$

2.2.1 Periodized in the Parameter b Hurwitz Zeta Function

Hereafter we will use the periodized in the parameter b Hurwitz zeta function

$$\zeta^*(\alpha; b) = \sum_{0<n+b} (n+b)^{-\alpha} = \begin{cases} \sum\limits_{n=1}^{\infty} n^{-\alpha}, & \{b\} = 0, \\ \sum\limits_{n=0}^{\infty} (n + \{b\})^{-\alpha}, & \{b\} > 0 \end{cases}, \quad (\sigma > 1).$$

It's easy to write out various explicit formulas for analytic continuation on the whole complex plane except the point $\alpha = 1$ of the periodized Hurwitz zeta function. In this point for any real value of b the periodized Hurwitz zeta function has a pole of order 1 with residue equal to 1.

The following formulas cover the whole complex plane and define the explicit analytic continuation of $\zeta^*(\alpha; b)$.

$$\zeta^*(\alpha; b) =$$

$$= \begin{cases} \sum\limits_{0<n+b} (n+b)^{-\alpha}, & \sigma > 1, \\ \dfrac{1}{2} + \dfrac{1}{\alpha - 1} - \alpha(\alpha+1) \int\limits_1^\infty \dfrac{\{x\}^2 - \{x\} dx}{2x^{\alpha+2}}, & \{b\} = 0, \quad \sigma > -1, \\ \dfrac{1}{2\{b\}^\alpha} + \dfrac{1}{(\alpha-1)\{b\}^{\alpha-1}} - \alpha(\alpha+1) \int\limits_1^\infty \dfrac{\{x\}^2 - \{x\}\, dx}{2\,(x+\{b\})^{\alpha+2}}, & \{b\} \neq 0,\ \sigma > -1, \\ 2(2\pi)^{\alpha-1} \Gamma(1-\alpha) \left(\sin\dfrac{\pi\alpha}{2} \sum\limits_{n=1}^\infty \dfrac{\cos 2\pi nb}{n^{1-\alpha}} + \cos\dfrac{\pi\alpha}{2} \sum\limits_{n=1}^\infty \dfrac{\sin 2\pi nb}{n^{1-\alpha}} \right), & \sigma < 0. \end{cases} \tag{2.36}$$

2.2.2 Dirichlet Series with Periodic Coefficients

Let us examine the special case of Dirichlet series with periodic coefficients of the form

$$l\left(\alpha, \frac{b}{n}\right) = \sum_{m=1}^\infty \frac{e^{2\pi i \frac{bm}{n}}}{m^\alpha} \quad (\sigma > 1) \tag{2.37}$$

and prove for them the special case of the general theorem (see [7]) on analytic continuation of Dirichlet series with periodic coefficients on the whole complex plane.

Lemma 2.2 *For $\sigma > 1$ the following equality is true:*

$$l\left(\alpha, \frac{b}{n}\right) = \begin{cases} \zeta(\alpha) & \text{if} \quad \delta_n(b) = 1, \\ \dfrac{1}{n^\alpha} \sum\limits_{j=1}^n e^{2\pi i \frac{bj}{n}} \zeta^*\left(\alpha, \dfrac{j}{n}\right) & \text{if} \quad \delta_n(b) = 0. \end{cases} \tag{2.38}$$

Lemma 2.3 *For $\sigma > 0$ and $\delta_n(b) = 0$ the following equality is true:*

$$\int_1^\infty \frac{e^{2\pi i \frac{b[t]}{n}}}{t^{\alpha+1}} dt = (\alpha+1) \int_1^\infty \frac{\frac{e^{2\pi i \frac{b[t]}{n}} - e^{2\pi i \frac{b}{n}}}{e^{2\pi i \frac{b}{n}} - 1} + e^{2\pi i \frac{b[t]}{n}} \{t\}}{t^{\alpha+2}} dt. \tag{2.39}$$

Theorem 2.20 *For a natural n, an integer b with $\delta_n(b) = 0$ and analytic continuation of the function $l\left(\alpha, \frac{b}{n}\right)$ on the whole complex plane the following presentations are true:*

$$l\left(\alpha,\frac{b}{n}\right)=$$

$$=\begin{cases}\displaystyle\sum_{m=1}^{\infty}\frac{e^{2\pi i\frac{bm}{n}}}{m^{\alpha}}, & \sigma>1,\\ \displaystyle\frac{\alpha e^{2\pi i\frac{b}{n}}}{e^{2\pi i\frac{b}{n}}-1}\int\limits_{1}^{\infty}\frac{e^{2\pi i\frac{b[t]}{n}}}{t^{\alpha+1}}\,dt-\frac{e^{2\pi i\frac{b}{n}}}{e^{2\pi i\frac{b}{n}}-1}, & \sigma>0,\\ \displaystyle\frac{\alpha(\alpha+1)e^{2\pi i\frac{b}{n}}}{e^{2\pi i\frac{b}{n}}-1}\int\limits_{1}^{\infty}\frac{\frac{e^{2\pi i\frac{b[t]}{n}}-e^{2\pi i\frac{b}{n}}}{e^{2\pi i\frac{b}{n}}-1}+e^{2\pi i\frac{b[t]}{n}}\{t\}}{t^{\alpha+2}}\,dt-\frac{e^{2\pi i\frac{b}{n}}}{e^{2\pi i\frac{b}{n}}-1}, & \sigma>-1,\\ (2\pi)^{\alpha-1}\Gamma(1-\alpha)\left(\displaystyle\sum_{m=1}^{\infty}\frac{e^{\frac{\pi i(\alpha-1)}{2}}}{\left(m-\left\{\frac{b}{n}\right\}\right)^{1-\alpha}}+\sum_{m=0}^{\infty}\frac{e^{\frac{-\pi i(\alpha-1)}{2}}}{\left(m+\left\{\frac{b}{n}\right\}\right)^{1-\alpha}}\right), & \sigma<0.\end{cases}\tag{2.40}$$

This result can be applied to another type of Dirichlet series with periodic coefficients. Let

$$l^{*}\left(\alpha,\frac{b}{n}\right)=\sum_{m=-\infty}^{\infty}\frac{e^{2\pi i\frac{bm}{n}}}{\overline{m}^{\alpha}}\quad(\Re\alpha>1).\tag{2.41}$$

The Dirichlet series of the latest form can directly define the hyperbolic zeta function of integer lattices for $\sigma>1$, if we use exponential sums of lattices, and namely, for any integer lattice Λ:

$$\begin{aligned}\zeta_H(\Lambda|\alpha)+1&=\sum_{\mathbf{x}\in\Lambda}{}'(\overline{x}_1\cdot\ldots\cdot\overline{x}_s)^{-\alpha}+1=\sum_{\mathbf{m}\in\mathbb{Z}^s}\frac{\delta_{\Lambda}(\mathbf{m})}{(\overline{m}_1\cdot\ldots\cdot\overline{m}_s)^{\alpha}}=\\&=\frac{1}{\det\Lambda}\sum_{\mathbf{x}\in M(\Lambda)}\sum_{\mathbf{m}\in\mathbb{Z}^s}\frac{e^{2\pi i(\mathbf{m},\mathbf{x})}}{(\overline{m}_1\cdot\ldots\cdot\overline{m}_s)^{\alpha}}=\\&=\frac{1}{\det\Lambda}\sum_{\mathbf{x}\in M(\Lambda)}\prod_{j=1}^{s}\sum_{m_j=-\infty}^{\infty}\frac{e^{2\pi i m_j x_j}}{\overline{m}_j^{\alpha}}\\&=\frac{1}{\det\Lambda}\sum_{\mathbf{x}\in M(\Lambda)}\prod_{j=1}^{s}l^{*}\left(\alpha,\frac{b_j(\mathbf{x})}{\det\Lambda}\right),\end{aligned}\tag{2.42}$$

where $b_j(\mathbf{x})=x_j\det\Lambda$ is an integer $(j=1,\ldots,s)$ for any point $\mathbf{x}=(x_1,\ldots,x_s)\in M(\Lambda)$.

Theorem 2.21 *For a natural n, an integer b with $\delta_n(b)=0$ and analytic continuation of the function $l^{*}\left(\alpha,\frac{b}{n}\right)$ on the whole complex plane the following presentations are true:*

$$l^*\left(\alpha, \frac{b}{n}\right) =$$

$$= \begin{cases} \displaystyle\sum_{m=-\infty}^{\infty} \frac{e^{2\pi i \frac{bm}{n}}}{\overline{m}^{\alpha}}, & \sigma > 1, \\ \displaystyle\frac{\alpha}{e^{2\pi i \frac{b}{n}} - 1} \int_1^{\infty} \frac{e^{2\pi i \frac{b([t]+1)}{n}} - e^{-2\pi i \frac{b[t]}{n}}}{t^{\alpha+1}} dt, & \sigma > 0, \\ \displaystyle\frac{\alpha(\alpha+1)}{e^{2\pi i \frac{b}{n}} - 1} \int_1^{\infty} \frac{g(t,b,n)}{t^{\alpha+2}} dt, & \sigma > -1, \\ \displaystyle 1 + 2(2\pi)^{\alpha-1} \Gamma(1-\alpha) \cos\frac{\pi(\alpha-1)}{2} \cdot n^{1-\alpha} \sum_{m=-\infty}^{\infty} \frac{1}{\left(\overline{nm+b}\right)^{1-\alpha}} & \sigma < 0, \end{cases} \tag{2.43}$$

where

$$g(t,b,n) = \frac{e^{2\pi i \frac{b}{n}} \left(e^{2\pi i \frac{b[t]}{n}} - e^{2\pi i \frac{b}{n}} + e^{-2\pi i \frac{b[t]}{n}} - e^{-2\pi i \frac{b}{n}} \right)}{e^{2\pi i \frac{b}{n}} - 1} + \\ + \left(e^{2\pi i \frac{b([t]+1)}{n}} - e^{-2\pi i \frac{b[t]}{n}} \right) \{t\}.$$

Note 2.1 The latest equality won't change if rewritten as follows

$$l^*\left(\alpha, \frac{b}{n}\right) = 1 + 2(2\pi)^{\alpha-1} \Gamma(1-\alpha) \cos\frac{\pi(\alpha-1)}{2} \cdot n^{1-\alpha} \sum_{\substack{m=-\infty, \\ nm+b \neq 0}}^{\infty} \frac{1}{\left(\overline{nm+b}\right)^{1-\alpha}},$$

which remains true with $\delta_n(b) = 1$:

$$l^*\left(\alpha, \frac{0}{n}\right) = 1 + 2\zeta(\alpha) = 1 + 2(2\pi)^{\alpha-1} \Gamma(1-\alpha) \cos\frac{\pi(\alpha-1)}{2} \sum_{m=1}^{\infty} \frac{1}{m^{1-\alpha}} =$$

$$= 1 + 2(2\pi)^{\alpha-1} \Gamma(1-\alpha) \cos\frac{\pi(\alpha-1)}{2} \cdot n^{1-\alpha} \sum_{\substack{m=-\infty, \\ nm \neq 0}}^{\infty} \frac{1}{(\overline{nm})^{1-\alpha}}.$$

2.2.3 Functional Equation for Hyperbolic Zeta Zunction of Integer Lattices

Let us obtain the explicit form of the $\zeta_H(\Lambda \mid \alpha)$ in the left half-plane for an arbitrary integer lattice Λ. For this, we will need a combined lattice $\Lambda^{(p)}$, which is defined by the following relationship:

$$\Lambda^{(p)} = \det \Lambda \cdot \Lambda^*. \tag{2.44}$$

For any integer lattice Λ its combined latice $\Lambda^{(p)}$ is also integer. As these lattices are special cases of Cartesian lattices, then, as we know, there are analytic continuations

$$\zeta_H(\Lambda \mid \alpha) \quad \text{and} \quad \zeta_H(\Lambda^{(p)} \mid \alpha)$$

on the whole complex α–plane, excluding the point $\alpha = 1$, where they have a pole of order s.

For the sake of convenience, we will use the following notations:

$$N = \det \Lambda, \quad M^{(p)}(\Lambda) = \det \Lambda \cdot M(\Lambda), \quad M^*(\Lambda) = \Lambda \cap [0; \det \Lambda)^s. \tag{2.45}$$

It is clear, that the following expansions are true:

$$\Lambda = \bigcup_{\mathbf{x} \in M^*(\Lambda)} \left(\mathbf{x} + N\mathbb{Z}^s\right), \quad \Lambda^{(p)} = \bigcup_{\mathbf{x} \in M^{(p)}(\Lambda)} \left(\mathbf{x} + N\mathbb{Z}^s\right). \tag{2.46}$$

Let $\mathbf{j}_t \in J_{t,s}$. We will denote the coordinate subspace as $\Pi(\mathbf{j}_t)$

$$\Pi(\mathbf{j}_t) = \{\mathbf{x} \mid x_{j_\nu} = 0 \ (\nu = t+1, \ldots, s)\}.$$

If we assume, that $\mathbf{j}_t^* = (j_{t+1}, \ldots, j_s, j_1, \ldots, j_t)$, then $\mathbf{j}_t^* \in J_{s-t,s}$ and

$$\mathbb{R}^s = \Pi(\mathbf{j}_t) \bigoplus \Pi(\mathbf{j}_t^*)$$

is decomposition into the direct sum of coordinate subspaces. If we denote projections of a shifted lattice on coordinate subspaces $\Pi(\mathbf{j}_t)$ and $\Pi(\mathbf{j}_t^*)$ according to decomposition of the space in the direct sum of these coordinate subspaces as $(\Lambda + \mathbf{a})_{\mathbf{j}_t}^{(1)}$ and $(\Lambda + \mathbf{a})_{\mathbf{j}_t}^{(2)}$; and denote its intersections with coordinate subspaces as $(\Lambda + \mathbf{a})_{\mathbf{j}_t} = (\Lambda + \mathbf{a}) \bigcap \Pi(\mathbf{j}_t)$ and $(\Lambda + \mathbf{a})_{\mathbf{j}_t^*} = (\Lambda + \mathbf{a}) \bigcap \Pi(\mathbf{j}_t^*)$, then, generally speaking, $(\Lambda+\mathbf{a})_{\mathbf{j}_t}^{(1)} \neq (\Lambda+\mathbf{a})_{\mathbf{j}_t}$ and $(\Lambda+\mathbf{a})_{\mathbf{j}_t}^{(2)} \neq (\Lambda+\mathbf{a})_{\mathbf{j}_t^*}$. The equality is possible, if and only if $\Lambda+\mathbf{a} = (\Lambda_1+\mathbf{a}_1)\times(\Lambda_2+\mathbf{a}_2)$, $\Lambda_1+\mathbf{a}_1 = (\Lambda+\mathbf{a})_{\mathbf{j}_t}$ $\Lambda_2+\mathbf{a}_2 = (\Lambda+\mathbf{a})_{\mathbf{j}_t^*}$.

We need to recall that

$$M(\alpha) = \frac{2\Gamma(1-\alpha)}{(2\pi)^{1-\alpha}} \sin\frac{\pi\alpha}{2}$$

and that for an arbitrary integer lattice Λ its zeta function $\zeta(\Lambda \mid \alpha)$ in the right half-plane is defined by the equality

$$\zeta(\Lambda \mid \alpha) = \sum_{\mathbf{x}\in\Lambda,\, N(\mathbf{x})\neq 0} |x_1 \dots x_s|^{-\alpha}.$$

Theorem 2.22 *For the zeta function of an arbitrary integer lattice Λ in the left half-plane $\sigma < 0$ the following functional equation is true:*

$$\zeta(\Lambda \mid \alpha) = \frac{1}{N}\left(M(\alpha)N^{1-\alpha}\right)^s \zeta\left(\Lambda^{(p)}\Big| 1-\alpha\right). \tag{2.47}$$

If we address dual lattices, then this theorem can be rewritten in the following way:

Theorem 2.23 *For the zeta function of an arbitrary integer lattice Λ in the left half-plane $\sigma < 0$ the following functional equation is true:*

$$\zeta(\Lambda \mid \alpha) = \frac{M(\alpha)^s}{N}\zeta\left(\Lambda^*\Big| 1-\alpha\right). \tag{2.48}$$

Proof As we can see, $\Lambda^{(p)} = N \cdot \Lambda^*$, therefore

$$\left(N^{1-\alpha}\right)^s \zeta\left(\Lambda^{(p)}\Big| 1-\alpha\right) = \left(N^{1-\alpha}\right)^s \sum_{\mathbf{x}\in\Lambda^{(p)},\, N(\mathbf{x})\neq 0} |x_1 \dots x_s|^{\alpha-1} =$$

$$= \sum_{\mathbf{x}\in\Lambda^{(p)},\, N(\mathbf{x})\neq 0} \left|\frac{x_1}{N}\cdots\frac{x_s}{N}\right|^{\alpha-1} = \sum_{\mathbf{y}\in\Lambda^*,\, N(\mathbf{y})\neq 0} |y_1 \dots y_s|^{\alpha-1} = \zeta\left(\Lambda^*\Big| 1-\alpha\right),$$

which proves the statement of the theorem.

According to the aforementioned definitions, $(\Lambda)_{\mathbf{j}_t} = \Lambda \bigcap \Pi(\mathbf{j}_t)$ is the intersection of the lattice and the coordinate subspace. Let us denote a t-dimensional lattice derived from the lattice $(\Lambda)_{\mathbf{j}_t}$ by discarding $s-t$ zero coordinates from each point as $\Lambda_{\mathbf{j}_t}$ and denote its determinant as $N_{\mathbf{j}_t}$. Thus, $\Lambda^{(p)}_{\mathbf{j}_t}$ is the "combined" t-dimensional lattice, $N_{\mathbf{j}_t} = \det \Lambda_{\mathbf{j}_t}$ and $N_{\mathbf{j}_t} | N$.

Theorem 2.24 *For the zeta function of an arbitrary integer lattice Λ in the left half-plane $\sigma < 0$ the following functional equation is true:*

$$\zeta_H(\Lambda \mid \alpha) = \sum_{t=1}^{s} M(\alpha)^t \sum_{\mathbf{j}_t\in J_{t,s}} N_{\mathbf{j}_t}^{t(1-\alpha)-1}\zeta\left(\Lambda^{(p)}_{\mathbf{j}_t}\Big| 1-\alpha\right). \tag{2.49}$$

If we use the Theorem 2.23 and denote the t-dimensional dual lattice as $\Lambda^*_{\mathbf{j}_t}$, then we will obtain a new form of the functional equation for the hyperbolic zeta function of an integer lattice.

Theorem 2.25 *For the hyperbolic zeta function of an arbitrary integer lattice Λ in the left half-plane $\sigma < 0$ the following functional equation is true:*

$$\zeta_H(\Lambda \mid \alpha) = \sum_{t=1}^{s} \sum_{\mathbf{j}_t \in J_{t,s}} \frac{M(\alpha)^t}{N_{\mathbf{j}_t}} \zeta\left(\Lambda^*_{\mathbf{j}_t} \middle| 1-\alpha\right). \tag{2.50}$$

Proof The definitions of the hyperbolic zeta function of a lattice and the zeta function of a lattice provide, that

$$\zeta_H(\Lambda \mid \alpha) = \sum_{t=1}^{s} \sum_{\mathbf{j}_t \in J_{t,s}} \zeta\left(\Lambda_{\mathbf{j}_t} \middle| \alpha\right). \tag{2.51}$$

Applying to each term of the right side the Theorem 2.23 we obtain the required result.

2.3 Functional Equation for Hyperbolic Zeta Function of Cartesian Lattices

First of all, we need the main result on the form of an arbitrary Cartesian lattice (see Theorem 2.11). According to this theorem, a Cartesian lattice Λ can be unambiguously presented as

$$\Lambda = D(d_1, \ldots, d_s) \cdot \Lambda_0, \quad d_1, \ldots, d_s > 0,$$

where Λ_0 is a simple lattice, and $D(d_1, \ldots, d_s)$ is a diagonal matrix.

Similarly to the aforementioned definitions, $(\Lambda_0)_{\mathbf{j}_t} = \Lambda_0 \bigcap \Pi(\mathbf{j}_t)$ is the intersection of the lattice and the coordinate space. Let us denote the t-dimensional lattice derived from the lattice $(\Lambda_0)_{\mathbf{j}_t}$ by discarding $s-t$ zero coordinates from each point as $\Lambda_{0,\mathbf{j}_t}$. Thus, $\Lambda^{(p)}_{0,\mathbf{j}_t}$ is the "combined" t-dimensional lattice.

First, let us examine the simpler case, where all the elements $d_j \geq 1$ $(j = 1, \ldots, s)$.

Theorem 2.26 *For the hyperbolic zeta function of a Cartesian lattice Λ of the form $\Lambda = D(d_1, \ldots, d_s) \cdot \Lambda_0$, where Λ_0 is a simple lattice and all its elements $d_j \geq 1$ $(j = 1, \ldots, s)$, in the left half-plane $\sigma < 0$ the following functional equation is true:*

$$\zeta_H(\Lambda \mid \alpha) = \sum_{t=1}^{s} M(\alpha)^t \sum_{\mathbf{j}_t \in J_{t,s}} \prod_{\nu=1}^{t} (d_{j_\nu})^{-\alpha} N_{0,\mathbf{j}_t}^{t(1-\alpha)-1} \zeta\left(\Lambda^{(p)}_{0,\mathbf{j}_t} \middle| 1-\alpha\right), \tag{2.52}$$

where $N_{0,\mathbf{j}_t} = \det \Lambda_{0,\mathbf{j}_t}$.

Proof The definitions of the hyperbolic zeta function of a lattice and the zeta function of a lattice provide that

$$\zeta_H(\Lambda \mid \alpha) = \sum_{t=1}^{s} \sum_{\mathbf{j}_t \in J_{t,s}} \prod_{\nu=1}^{t} (d_{j_\nu})^{-\alpha} \zeta\left(\Lambda_{0,\mathbf{j}_t} \middle| \alpha\right). \tag{2.53}$$

Applying to each term of the right side the Theorem 2.22 we obtain the required result.

Now we will obtain a functional equation using a dual lattice.

Theorem 2.27 *For the hyperbolic zeta function of a Cartesian lattice* Λ *of the form* $\Lambda = D(d_1, \ldots, d_s) \cdot \Lambda_0$, *where* Λ_0 *is a simple lattice and all elements* $d_j \geq 1$ $(j = 1, \ldots, s)$, *in the left half-plane* $\sigma < 0$ *the following functional equation is true:*

$$\zeta_H(\Lambda \mid \alpha) = \sum_{t=1}^{s} \sum_{\mathbf{j}_t \in J_{t,s}} \frac{M(\alpha)^t}{\det \Lambda_{\mathbf{j}_t}} \zeta\left(\Lambda^*_{\mathbf{j}_t} \middle| 1 - \alpha\right). \tag{2.54}$$

Proof First of all, we need to state, that $\Lambda^* = (D(d_1, \ldots, d_s) \cdot \Lambda_0)^* = D\left(\frac{1}{d_1}, \ldots, \frac{1}{d_s}\right) \cdot \Lambda_0^*$ and $\det\left(D(d_1, \ldots, d_s) \cdot \Lambda_0\right) = d_1 \cdots d_s \cdot \det \Lambda_0$.

If we address the projections of $\Lambda_{\mathbf{j}_t}$, then we will obtain that

$$\begin{aligned}
\Lambda_{\mathbf{j}_t} &= D(d_{j_1}, \ldots, d_{j_t}) \cdot \Lambda_{0,\mathbf{j}_t}, \\
\Lambda^*_{\mathbf{j}_t} = \left(D(d_{j_1}, \ldots, d_{j_t}) \cdot \Lambda_{0,\mathbf{j}_t}\right)^* &= D\left(\frac{1}{d_{j_1} N_{0,\mathbf{j}_t}}, \ldots, \frac{1}{d_{j_t} N_{0,\mathbf{j}_t}}\right) \cdot \Lambda^{(p)}_{0,\mathbf{j}_t} = \\
&= D\left(\frac{1}{d_{j_1}}, \ldots, \frac{1}{d_{j_t}}\right) \cdot \Lambda^*_{0,\mathbf{j}_t}, \\
\Lambda^*_{0,\mathbf{j}_t} &= D(d_{j_1}, \ldots, d_{j_t}) \Lambda^*_{\mathbf{j}_t}, \\
\det\left(D(d_{j_1}, \ldots, d_{j_t}) \cdot \Lambda_{0,\mathbf{j}_t}\right) &= d_{j_1} \cdot \ldots \cdot d_{j_t} \cdot \det \Lambda_{0,\mathbf{j}_t} = d_{j_1} \cdot \ldots \cdot d_{j_t} \cdot N_{0,\mathbf{j}_t}, \\
\zeta\left(\Lambda^*_{0,\mathbf{j}_t} \middle| 1 - \alpha\right) &= \left(d_{j_1} \cdot \ldots \cdot d_{j_t}\right)^{\alpha - 1} \zeta\left(\Lambda^*_{\mathbf{j}_t} \middle| 1 - \alpha\right).
\end{aligned}$$

The definitions of the hyperbolic zeta function of a lattice and the zeta function of a lattice provide that

$$\zeta_H(\Lambda \mid \alpha) = \sum_{t=1}^{s} \sum_{\mathbf{j}_t \in J_{t,s}} \prod_{\nu=1}^{t} (d_{j_\nu})^{-\alpha} \zeta\left(\Lambda_{0,\mathbf{j}_t} \middle| \alpha\right). \tag{2.55}$$

Applying to each term of the right side the Theorem 2.23, we obtain that

$$\zeta_H(\Lambda \mid \alpha) = \sum_{t=1}^{s} \sum_{\mathbf{j}_t \in J_{t,s}} \prod_{\nu=1}^{t} (d_{j_\nu})^{-\alpha} \frac{M(\alpha)^t}{N_{0,\mathbf{j}_t}} \zeta\left(\Lambda^*_{0,\mathbf{j}_t} \middle| 1-\alpha\right) =$$
$$= \sum_{t=1}^{s} \sum_{\mathbf{j}_t \in J_{t,s}} \prod_{\nu=1}^{t} (d_{j_\nu})^{-\alpha} \frac{M(\alpha)^t}{N_{0,\mathbf{j}_t}} \left(d_{j_1} \cdots d_{j_t}\right)^{\alpha-1} \zeta\left(\Lambda^*_{\mathbf{j}_t} \middle| 1-\alpha\right) =$$
$$= \sum_{t=1}^{s} \sum_{\mathbf{j}_t \in J_{t,s}} \frac{M(\alpha)^t}{\det \Lambda_{\mathbf{j}_t}} \zeta\left(\Lambda^*_{\mathbf{j}_t} \middle| 1-\alpha\right), \tag{2.56}$$

which proves the statement of the theorem.

Now, let us examine a general case, where the set $D_1 = \{j | 0 < d_j < 1\} \neq \emptyset$. For this, we need to examine one more type of Dirichlet series with periodic coefficients. Let

$$l^{**}\left(\alpha, d, \frac{b}{n}\right) = \sum_{m=-\infty}^{\infty} \frac{e^{2\pi i \frac{bm}{n}}}{\overline{dm}^{\alpha}} \quad (\Re\alpha > 1, \quad d > 0). \tag{2.57}$$

The Dirichlet series of the latest form can directly define the hyperbolic zeta function of Cartesian lattices for $\sigma > 1$, if we use exponential sums of lattices, and namely, for any Cartesian lattice $\Lambda = D(d_1, \ldots, d_s) \cdot \Lambda_0$, where Λ_0 is a simple lattice, and $D(d_1, \ldots, d_s)$ is a diagonal matrix:

$$\zeta_H(\Lambda|\alpha) + 1 = \sum_{\mathbf{x} \in \Lambda}{}' (\overline{x}_1 \cdots \overline{x}_s)^{-\alpha} + 1 =$$
$$= \sum_{\mathbf{m} \in \mathbb{Z}^s} \frac{\delta_{\Lambda_0}(\mathbf{m})}{(\overline{d_1 m_1} \cdots \overline{d_s m_s})^{\alpha}} =$$
$$= \frac{1}{\det \Lambda_0} \sum_{\mathbf{x} \in M(\Lambda_0)} \sum_{\mathbf{m} \in \mathbb{Z}^s} \frac{e^{2\pi i (\mathbf{m}, \mathbf{x})}}{(\overline{d_1 m_1} \cdots \overline{d_s m_s})^{\alpha}} =$$
$$= \frac{1}{\det \Lambda_0} \sum_{\mathbf{x} \in M(\Lambda_0)} \prod_{j=1}^{s} \sum_{m_j=-\infty}^{\infty} \frac{e^{2\pi i m_j x_j}}{\overline{d_j m_j}^{\alpha}} =$$
$$= \frac{1}{\det \Lambda_0} \sum_{\mathbf{x} \in M(\Lambda_0)} \prod_{j=1}^{s} l^{**}\left(\alpha, d_j, \frac{b_j(\mathbf{x})}{\det \Lambda_0}\right), \tag{2.58}$$

where $b_j(\mathbf{x}) = x_j \det \Lambda_0$ is an integer ($j = 1, \ldots, s$) for any point $\mathbf{x} = (x_1, \ldots, x_s) \in M(\Lambda_0)$.

As it was stated above the hyperbolic zeta function of a lattice is not homogeneous, while the zeta function is. Our previous arguments provide, that the homogeneous zeta function of a lattice is crucial for the analytic continuation. In the general case, the hyperbolic zeta function of a lattice can not be presented as a sum of homogeneous components (as it can be done with integer lattices), but in the case of Cartesian lattices we can define $\mathbf{j}_t$-components.

As it has been done above, for a Cartesian lattice Λ we will use $\Lambda_{\mathbf{j}_t}$ to denote the projection of the intersection $\Lambda \bigcap \Pi(\mathbf{j}_t)$ on $\mathbb{R}^t$.

Definition 2.19 The $\mathbf{j}_t$-component of the hyperbolic zeta function of the lattice Λ is the function $\zeta_{H,\mathbf{j}_t}(\Lambda|\alpha)$, $\alpha = \sigma + it$, defined for $\sigma > 1$ by the series

$$\zeta_{H,\mathbf{j}_t}(\Lambda|\alpha) = \sum_{\mathbf{x}\in\Lambda_{H,\mathbf{j}_t},\, N(\mathbf{x})\neq 0} |x_1 \cdots x_t|^{-\alpha}. \tag{2.59}$$

It is easy to see, that for the $\mathbf{j}_t$-component of the hyperbolic zeta function of a lattice Λ the analogue of the formula (2.58) is true.

$$\zeta_{H,\mathbf{j}_t}(\Lambda|\alpha) = \frac{1}{\det \Lambda_{0,\mathbf{j}_t}} \sum_{\mathbf{x}\in M(\Lambda_{0,\mathbf{j}_t})} \prod_{\nu=1}^{t} \left(l^{**}\left(\alpha, d_{j_\nu}, \frac{b_j(\mathbf{x})}{\det \Lambda_{0,\mathbf{j}_t}}\right) - 1\right). \tag{2.60}$$

Moreover, we can see the decomposition into components:

$$\zeta_H(\Lambda|\alpha) = \sum_{t=1}^{s} \sum_{\mathbf{j}_t\in J(t,s)} \zeta_{H,\mathbf{j}_t}(\Lambda|\alpha). \tag{2.61}$$

Definition 2.20 Let the $\mathbf{j}_s$-component of the hyperbolic zeta function of a lattice Λ be called the main component and denoted as $\zeta_{H,s}(\Lambda|\alpha)$.

It is clear, that the following equality is true:

$$\zeta_{H,\mathbf{j}_t}(\Lambda|\alpha) = \zeta_{H,\,t}(\Lambda_{\mathbf{j}_t}|\alpha). \tag{2.62}$$

Theorem 2.28 *For a natural n, an integer b with $\delta_n(b) = 0$, a positive d and the analytic continuation of the function $l^{**}\left(\alpha, d, \frac{b}{n}\right)$ on the whole complex plane the following presentations are true:*

$$l^{**}\left(\alpha, d, \frac{b}{n}\right) = 1 + \frac{1}{d^\alpha}\left(l^{*}\left(\alpha, \frac{b}{n}\right) - 1\right) + f\left(\alpha, d, \frac{b}{n}\right), \tag{2.63}$$

where

$$f\left(\alpha, d, \frac{b}{n}\right) = \sum_{1\le|m|\le\left[\frac{1}{d}\right]} e^{2\pi i \frac{bm}{n}}\left(1 - \frac{1}{|dm|^\alpha}\right)$$

and $f\left(\alpha, d, \frac{b}{n}\right) = 0$ with $d \ge 1$.

Proof For $\sigma > 1$ from the definition follows that

$$l^{**}\left(\alpha, d, \frac{b}{n}\right) = 1 + \sum_{1 \le |m| \le \left[\frac{1}{d}\right]} e^{2\pi i \frac{bm}{n}} + \sum_{|m| > \left[\frac{1}{d}\right]} \frac{e^{2\pi i \frac{bm}{n}}}{|dm|^{\alpha}} =$$

$$= 1 + \sum_{1 \le |m| \le \left[\frac{1}{d}\right]} e^{2\pi i \frac{bm}{n}} \left(1 - \frac{1}{|dm|^{\alpha}}\right) + \sum_{|m| \ge 1} \frac{e^{2\pi i \frac{bm}{n}}}{|dm|^{\alpha}} =$$

$$= 1 + \frac{1}{d^{\alpha}} \left(l^{*}\left(\alpha, \frac{b}{n}\right) - 1\right) + f\left(\alpha, d, \frac{b}{n}\right).$$

As there are analytic functions in the right side of the equality, which are defined on the whole complex α-plane, excluding the point $\alpha = 1$, where is a pole of order 1, then the theorem is proven.

Let us introduce some additional definitions. For $1 \le r \le |D_1|$ and $1 \le t \le s - r$ let us define the set of integer vectors

$$J_{t,r,s}(D_1) = \{\mathbf{j}_{t,r} = (j_1, \ldots, j_s) \mid 1 \le j_1 < \cdots < j_t \le s, \quad 1 \le j_{t+r+1} < \cdots < j_s \le s,$$

$$1 \le j_{t+1} < \cdots < j_{t+r} \le s, \quad \{j_1, \ldots, j_s\} = \{1, 2, \ldots, s\},$$

$$j_{t+\nu} \in D_1 \text{ if } 1 \le \nu \le r\}.$$

In other words, the set $J_{t,r,s}(D_1)$ consists of integer vectors $\mathbf{j}_{t,r}$, coordinates of which form the permutation of numbers from 1 to s, wile coordinates from 1 to t, and from $t+1$ to $t+r$, and from $t+r+1$ to s form increasing sequences. Moreover, all coordinates from $t+1$ to $t+r$ belong to the set D_1. Obviously, $J_{t,r,s}|D_1| = C_{s-r}^{t} C_{|D_1|}^{r}$.

Theorem 2.29 *For the main component of the hyperbolic zeta function of an arbitrary Cartesian lattice Λ of the form $\Lambda = D(d_1, \ldots, d_s) \cdot \Lambda_0$, where Λ_0 is a simple lattice and all its elements $d_j > 0$ $(j = 1, \ldots, s)$, in the left half-plane $\sigma < 0$ the following functional equation is true:*

$$\zeta_{H,s}(\Lambda \mid \alpha) = \frac{M(\alpha)^s}{\det \Lambda} \zeta\left(\Lambda^* \middle| 1 - \alpha\right) + \frac{1}{\det \Lambda_0} \sum_{\mathbf{x} \in M(\Lambda_0)} \sum_{r=1}^{|D_1|} M(\alpha)^{s-r} N_0^{s-r-\alpha(s-r)}$$

$$\cdot \sum_{\mathbf{j}_{s-r,r} \in J_{s-r,r,s}(D_1)} \prod_{\nu=1}^{s-r} (d_{j_\nu})^{-\alpha} \prod_{\nu=s-r+1}^{s} f\left(\alpha, d_{j_\nu}, \frac{b_{j_\nu}(\mathbf{x})}{\det \Lambda_0}\right) \zeta\left(N_0 \mathbb{Z}^{s-r} + \mathbf{b}_{s-r}(\mathbf{x}) \middle| 1 - \alpha\right), \tag{2.64}$$

where $N_0 = \det \Lambda_0$.

Proof According to the equality (2.60) and the Theorem 2.28 for the main component of the hyperbolic zeta function of an arbitrary Cartesian lattice $\Lambda = D(d_1, \ldots, d_s) \cdot \Lambda_0$ on the whole complex α-plane, excluding the point $\alpha = 1$,

which has a pole of order s, the following equality is true:

$$\zeta_{H,s}(\Lambda|\alpha) = \frac{1}{\det \Lambda} \sum_{\mathbf{x} \in M(\Lambda_0)} \prod_{j=1}^{s} \left(l^{**}\left(\alpha, d_j, \frac{b_j(\mathbf{x})}{\det \Lambda_0}\right) - 1 \right). \tag{2.65}$$

For $\sigma < 0$, let us apply the Theorems 2.28 and 2.21, and therefore obtain that

$$\zeta_{H,s}(\Lambda|\alpha) = \frac{1}{\det \Lambda_0} \sum_{\mathbf{x} \in M(\Lambda_0)} \prod_{j=1}^{s} \left(\frac{1}{d_j^{\alpha}} \left(l^{*}\left(\alpha, \frac{b_j(\mathbf{x})}{\det \Lambda_0}\right) - 1 \right) + f\left(\alpha, d_j, \frac{b_j(\mathbf{x})}{\det \Lambda_0}\right) \right) =$$

$$= \frac{1}{\det \Lambda_0} \sum_{\mathbf{x} \in M(\Lambda_0)} \prod_{j=1}^{s} \left(\frac{M(\alpha)}{d_j^{\alpha}} N_0^{1-\alpha} \sum_{\substack{m=-\infty, \\ N_0 \cdot m + b_j(\mathbf{x}) \neq 0}}^{\infty} \frac{1}{\left|N_0 \cdot m + b_j(\mathbf{x})\right|^{1-\alpha}} + f\left(\alpha, d_j, \frac{b_j(\mathbf{x})}{\det \Lambda_0}\right) \right). \tag{2.66}$$

To expand the product in the right side of the formula (2.66) let us use the following equality:

$$\prod_{j=1}^{s} \left(\frac{M(\alpha)}{d_j^{\alpha}} N_0^{1-\alpha} \sum_{\substack{m=-\infty, \\ N_0 \cdot m + b_j(\mathbf{x}) \neq 0}}^{\infty} \frac{1}{\left|N_0 \cdot m + b_j(\mathbf{x})\right|^{1-\alpha}} + f\left(\alpha, d_j, \frac{b_j(\mathbf{x})}{\det \Lambda_0}\right) \right) =$$

$$= \prod_{j \in D_1} \left(\frac{M(\alpha)}{d_j^{\alpha}} N_0^{1-\alpha} \sum_{\substack{m=-\infty, \\ N_0 \cdot m + b_j(\mathbf{x}) \neq 0}}^{\infty} \frac{1}{\left|N_0 \cdot m + b_j(\mathbf{x})\right|^{1-\alpha}} + f\left(\alpha, d_j, \frac{b_j(\mathbf{x})}{\det \Lambda_0}\right) \right) \times$$

$$\times \prod_{j \notin D_1} \left(\frac{M(\alpha)}{d_j^{\alpha}} N_0^{1-\alpha} \sum_{\substack{m=-\infty, \\ N_0 \cdot m + b_j(\mathbf{x}) \neq 0}}^{\infty} \frac{1}{\left|N_0 \cdot m + b_j(\mathbf{x})\right|^{1-\alpha}} \right) =$$

$$= \frac{M(\alpha)^s}{N_0^{(\alpha-1)s}} \prod_{j=1}^{s} (d_j)^{-\alpha} \sum_{\substack{m_j=-\infty (1 \leq j \leq s), \\ N_0 \cdot m_j + b_j(\mathbf{x}) \neq 0}}^{\infty} \frac{1}{\left|(N_0 \cdot m_1 + b_1(\mathbf{x})) \cdots (N_0 \cdot m_s + b_s(\mathbf{x}))\right|^{1-\alpha}} +$$

$$+ \sum_{r=1}^{|D_1|} \left(M(\alpha)^{s-r} N_0^{s-r-\alpha(s-r)} \sum_{\mathbf{j}_{s-r,r} \in J_{s-r,r,s}(D_1)} \prod_{\nu=1}^{s-r} (d_{j_\nu})^{-\alpha} \prod_{\nu=s-r+1}^{s} f\left(\alpha, d_{j_\nu}, \frac{b_{j_\nu}(\mathbf{x})}{\det \Lambda_0}\right) \times \right.$$

$$\times \sum_{\substack{m_{j_\nu}=-\infty(1\le\nu\le s-r),\\ N_0\cdot m_{j_\nu}+b_{j_\nu}(\mathbf{x})\neq 0}}^{\infty} \frac{1}{\left|(N_0\cdot m_{j_1}+b_{j_1}(\mathbf{x}))\cdots(N_0\cdot m_{j_{s-r}}+b_{j_{s-r}}(\mathbf{x}))\right|^{1-\alpha}}\Bigg). \tag{2.67}$$

From (2.66) and (2.67), assuming that $\mathbf{b}_t(\mathbf{x}) = (b_{j_1}(\mathbf{x}), \ldots, b_{j_t}(\mathbf{x}))$, we will obtain that

$$\begin{aligned}
&\zeta_{H,s}(\Lambda|\alpha)\\
&= \frac{1}{\det\Lambda_0}\sum_{\mathbf{x}\in M(\Lambda_0)}\Bigg(M(\alpha)^s N_0^{s-\alpha s}\prod_{j=1}^{s}(d_j)^{-\alpha}\times\\
&\quad\times\sum_{\substack{m_j=-\infty(1\le j\le s),\\ N_0\cdot m_j+b_j(\mathbf{x})\neq 0}}^{\infty}\frac{1}{|(N_0\cdot m_1+b_1(\mathbf{x}))\cdots(N_0\cdot m_s+b_s(\mathbf{x}))|^{1-\alpha}}+\\
&\quad+\sum_{r=1}^{|D_1|}\Bigg(M(\alpha)^{s-r}N_0^{s-r-\alpha(s-r)}\sum_{\mathbf{j}_{s-r,r}\in J_{s-r,r,s}(D_1)}\prod_{\nu=1}^{s-r}(d_{j_\nu})^{-\alpha}\prod_{\nu=s-r+1}^{s}f\left(\alpha,d_{j_\nu},\frac{b_{j_\nu}(\mathbf{x})}{\det\Lambda_0}\right)\times\\
&\quad\times\sum_{\substack{m_{j_\nu}=-\infty(1\le\nu\le s-r),\\ N_0\cdot m_{j_\nu}+b_{j_\nu}(\mathbf{x})\neq 0}}^{\infty}\frac{1}{\left|(N_0\cdot m_{j_1}+b_{j_1}(\mathbf{x}))\cdots(N_0\cdot m_{j_{s-r}}+b_{j_{s-r}}(\mathbf{x}))\right|^{1-\alpha}}\Bigg)\Bigg)=\\
&= \frac{1}{\det\Lambda_0}\sum_{\mathbf{x}\in M(\Lambda_0)}\Bigg(M(\alpha)^s N_0^{s-\alpha s}\prod_{j=1}^{s}(d_j)^{-\alpha}\zeta\left(N_0\mathbb{Z}^s+\mathbf{b}_s(\mathbf{x})\middle|\,1-\alpha\right)+\\
&\quad+\sum_{r=1}^{|D_1|}\Bigg(M(\alpha)^{s-r}N_0^{s-r-\alpha(s-r)}\sum_{\mathbf{j}_{s-r,r}\in J_{s-r,r,s}(D_1)}\prod_{\nu=1}^{s-r}(d_{j_\nu})^{-\alpha}\prod_{\nu=s-r+1}^{s}f\left(\alpha,d_{j_\nu},\frac{b_{j_\nu}(\mathbf{x})}{\det\Lambda_0}\right)\times\\
&\quad\times\zeta\left(N_0\mathbb{Z}^{s-r}+\mathbf{b}_{s-r}(\mathbf{x})\middle|\,1-\alpha\right)\Bigg)\Bigg).
\end{aligned}\tag{2.68}$$

As

$$\begin{aligned}
&\frac{1}{\det\Lambda_0}\sum_{\mathbf{x}\in M(\Lambda_0)}M(\alpha)^s N_0^{s-\alpha s}\prod_{j=1}^{s}(d_j)^{-\alpha}\zeta\left(N_0\mathbb{Z}^s+\mathbf{b}_s(\mathbf{x})\middle|\,1-\alpha\right)=\\
&\quad=\frac{1}{\det\Lambda_0}M(\alpha)^s N_0^{s-\alpha s}\prod_{j=1}^{s}(d_j)^{-\alpha}\zeta\left(\Lambda_0^{(p)}\middle|\,1-\alpha\right)=\frac{M(\alpha)^s}{\det\Lambda}\zeta\left(\Lambda^*\middle|\,1-\alpha\right),
\end{aligned}\tag{2.69}$$

then the statement of the theorem is completely proven.

Theorem 2.30 *For the hyperbolic zeta function of an arbitrary Cartesian lattice Λ of the form $\Lambda = D(d_1, \ldots, d_s) \cdot \Lambda_0$, where Λ_0 is a simple lattice and all elements $d_j > 0$ $(j = 1, \ldots, s)$, in the left half-plane $\sigma < 0$ the following functional equation is true:*

$$\begin{aligned}\zeta_H(\Lambda \mid \alpha) = &\sum_{t=1}^{s} \sum_{\mathbf{j}_t \in J(t,s)} \frac{M(\alpha)^t}{\det \Lambda_{\mathbf{j}_t}} \zeta\left(\Lambda^*_{\mathbf{j}_t} \middle| 1-\alpha\right) + \\ &+ \sum_{t=1}^{s} \sum_{\mathbf{j}_t \in J(t,s)} \frac{1}{\det \Lambda_{0,\mathbf{j}_t}} \sum_{\mathbf{x} \in M(\Lambda_{0,\mathbf{j}_t})} \sum_{r=1}^{|D_{1,\mathbf{j}_t}|} M(\alpha)^{t-r} N_{0,\mathbf{j}_t}^{t-r-\alpha(t-r)} \\ &\times \sum_{\mathbf{j}_{t-r,r} \in J_{t-r,r,t}(D_{1,\mathbf{j}_t})} \prod_{\nu=1}^{t-r} (d_{j_\nu})^{-\alpha} \prod_{\nu=t-r+1}^{t} f\left(\alpha, d_{j_\nu}, \frac{b_{j_\nu}(\mathbf{x})}{\det \Lambda_{0,\mathbf{j}_t}}\right) \\ &\zeta\left(N_{0,\mathbf{j}_t} \mathbb{Z}^{t-r} + \mathbf{b}_{t-r}(\mathbf{x}) \middle| 1-\alpha\right), \end{aligned} \tag{2.70}$$

where $N_{0,\mathbf{j}_t} = \det \Lambda_{0,\mathbf{j}_t}$.

Proof The theorem statement follows from the decomposition into components formula (see (2.61)) and the application of the Theorem 2.29 to each component according to the formula (2.62).

2.4 On Some Unsolved Problems of the Theory of Hyperbolic Zeta Function of Lattices

The article [9] hints at some possible directions of further development of Korobov number-theoretical method in approximate analysis. We are going to examine the problems regarding the theory of the hyperbolic zeta function of lattices in more detail.

The problem of right order The class of algebraic lattices is known for making it possible to achieve the correct order of decreasing hyperbolic zeta function of lattices when increasing the determinant of lattices (see the formulas (2.19) and (2.21)). Moreover, the asymptotic formula (2.25) is true for these lattices. The continuity of the hyperbolic function on the lattice space provides that the correct order of decreasing hyperbolic zeta function of lattices can be achieved on the class of rational lattices. It is enough to take rational lattices from very small neighborhoods of algebraic lattices. A natural question arise: can the correct order of decreasing be achieved in the class of integer lattices, or not? If it can be achieved, we need to provide an algorithm for construction of such optimal parallelepipedal nets, which would have the right order of the error of approximate integration on the classes E_s^α. Otherwise, we will obtain a kind of the theorem, which is analogous to the Liouville-Thue-Siegel-Roth theorem for algebraic lattices, as the impossibility of the right order means that algebraic lattices can not be correctly approximated by integer ones.

The problem of existence of analytic continuation As stated above, any Cartesian lattice has an analytic continuation of the hyperbolic zeta function of an arbitrary Cartesian lattice. Moreover, there's been obtained the functional equation for an arbitrary Cartesian lattice, which explicitly defines this analytic continuation. Naturally, there are questions, whether an analytic continuation of the hyperbolic zeta function exists in the following cases:

for a lattice of joint approximations $\Lambda(\theta_1, \ldots, \theta_s)$, defined by the equality

$$\Lambda(\theta_1, \ldots, \theta_s) = \{(q, q\theta_1 - p_1, \ldots, q\theta_s - p_s) \mid q, p_1, \ldots, p_s \in \mathbb{Z}\},$$

where $\theta_1, \ldots, \theta_s$ are arbitrary irrational numbers.

for an algebraic lattice $\Lambda(t, F) = t\Lambda(F)$, where the lattice $\Lambda(F)$ is defined by the equality (2.3).

for an arbitrary lattice Λ. If the hyperbolic zeta function of an arbitrary lattice can not be continued onto the whole complex plane (and we have strong doubts about that), then we will have to describe a new class, containing all lattices, for which their hyperbolic zeta functions can be analytically continued onto the whole complex plane, excluding the point $\alpha = 1$, which has a pole of order s.

The problem of the critical strip behaviour This problem has been underlined by Korobov. He suggested the hypothesis, according to which the analytic continuation of the hyperbolic zeta function of a lattice into the critical strip from the right half-plane and the analytic continuation of the hyperbolic zeta function of a dual lattice or combined lattices into the critical strip from the left half-plane will allow us to get the constants in the corresponding transfer theorems.

Acknowledgments The authors are grateful to professor G. I. Arkhipov and to professor V. N. Chubarikov for constant attention to this work and for useful discussions. This research was partially supported by the RFBR grant 11-01-00571.

References

1. Bakhvalov, N.S.: On the approximate calculation of integrals. Vestn. Mosk. Univ. Ser. Mat. Mekh. Astron. Fiz. Him. **4**, 3–18 (1959)
2. Bocharova, L.P., Van'kova, V.S., Dobrovol'skii, N.M.: Computation of the optimal coefficients. Math. Notes. **49**(2), 130–134 (1991). doi:10.1007/BF01137541
3. Bocharova, L.P.: Algorithms for finding optimal coefficients. Cheb. Sb. **8**(1), 4–109 (2007)
4. Bykovskii, V.A.: Extreme cubature formulas for anisotropic classes. Preprint, Khabarovsk (1995)
5. Bykovskii, V.A.: On the correct order of the error of optimal cubature formulas in spaces with dominating derivative and standard deviation of nets. Preprint, Vladivostok (1985)
6. Chandrasekharan, K.: Introduction to Analytic Number Theory. Springer, New York (2012)
7. Chudakov, N.G.: Introduction to the Theory of Dirichlet L-Functions. OGIZ, Moscow-Leningrad (1947)

8. Dobrovol'skaya, L.P., Dobrovol'skii, M.N., Dobrovol'skii, N.M., Dobrovol'skii, N.N.: Multidimensional Number-Theoretic Nets and Lattices and Algorithms for Finding Optimal Coefficients. Publishing House of the Tula State Lev Tolstoy Pedagogical University, Tula (2012)
9. Dobrovol'skaya, L.P., Dobrovol'skii, N.M., Dobrovol'skii, N.N., Ogorodnichuk, N.K., Rebrov, E.D., Rebrova, I.Yu.: Some issues of number-theoretic method in approximate analysis. In: Proceedings of the X International Conference "Algebra and Number Theory: Contemporary Issues and Applications". Uch. Zap. Orlov. Gos. Univ. Ser.: Estestv., Teh. i Med. Nauk. 6(2), 90–98 (2012)
10. Dobrovol'skaya, L.P., Dobrovol'skii, N.M., Simonov, A.S.: On the error of numerical integration on modified nets. Cheb. Sb. **9**(1), 185–223 (2008)
11. Dobrovol'skii, N.M., Klepikova, N.L.: Table of optimal coefficients for the approximate calculation of multiple integrals. Preprint, 63. General Physics Institute of the U.S.S.R. Academy of Sciences, Moscow (1990)
12. Dobrovol'skii, N.M., Van'kova, V.S., Kozlova, S.L.: The hyperbolic zeta-function of algebraic lattices. Available from VINITI, Moscow, no. 2327–90
13. Dobrovol'skii, N.M., Van'kova, V.S.: On a lemma by Gelfond, A.O. Available from VINITI, Moscow, no. 1467–87
14. Dobrovol'skii, N.M.: Hyperbolic zeta function of lattices. Available from VINITI, Moscow, no. 6090–84
15. Dobrovolskii, N.M.: Multidimensional Number-Theoretic Nets and Lattices and Their Applications. Publishing House of the Tula State Lev Tolstoy Pedagogical University, Tula (2005)
16. Dobrovol'skii, N.M.: On quadrature formulas on $E_s^{\alpha}(c)$ and $H_s^{\alpha}(c)$. Available from VINITI, Moscow, no. 6091–84
17. Dobrovol'skii, N.M., Roshchenya, A.L.: On the continuity of the hyperbolic zeta function of lattices. Izv. Tul. Gos. Univ. Ser. Mat. Mekh. Inform. **2**(1), 77–87 (1996)
18. Dobrovol'skii, N.M., Roshchenya, A.L.: Number of lattice points in the hyperbolic cross. Math. Notes. **63**(3), 319–324 (1998). doi:10.1007/BF02317776
19. Dobrovolskii, N.M., Roshchenya, A.L., Rebrova, IYu.: Continuity of the hyperbolic zeta function of lattices. Math. Notes. **63**(4), 460–463 (1998). doi:10.1007/BF02311248
20. Dobrovol'skii, N.M., Esayan, A.R., Pihtilkov, S.A., Rodionova, O.V., Ustyan, A.E.: On one algorithm for finding optimal coefficients. Izv. Tul. Gos. Univ. Ser. Mat. Mekh. Inform. **5**(1), 51–71 (1999)
21. Dobrovol'skii, N.M., Esayan, A.R., Rebrova, I.Yu.: On one recursive algorithm for lattices. Izv. Tul. Gos. Univ. Ser. Mat. Mekh. Inform. **5**(3), 38–51 (1999)
22. Dobrovolskii, N.M., Korobov, N.M.: Optimal coefficients for combined nets. Cheb. Sb. **2**, 41–53 (2001)
23. Dobrovol'skii, M.N., Dobrovol'skii, N.M., Kiseleva, O.V.: On the product of generalized parallelepipedal nets of integer lattices. Cheb. Sb. **3**(2), 43–59 (2002)
24. Dobrovol'skii, M.N.: Functional equation for the hyperbolic zeta function of integer lattices. Dok. Math. **75**(1), 53–54 (2007). doi:10.1134/S1064562407010152
25. Dobrovolskii, M.N.: A functional equation for the hyperbolic zeta function of integer lattices. Mosc. Univ. Math. Bull. **62**(5), 186–191 (2007). doi:10.3103/S0027132207050038
26. Frolov, K.K.: Quadrature formulas for classes of functions. Dissertation, Computer Centre of the Academy of Sciencies of the USSR (1971)
27. Frolov, K.K.: The upper bounds of the error of quadrature formulas for classes of functions. Dok. Akad. Nauk SSSR. **231**(4), 818–821 (1976)
28. Korobov, N.M.: An estimate of Gelfond, A.O. Vestn. Mosk. Univ. Ser. 1 Mat. Mekh. **3**, 3–7 (1983)
29. Korobov, N.M.: Exponential Sums and Their Applications. Kluwer Academic Publishers Group, Dordrecht (1992)
30. Korobov, N.M.: Number-Theoretic Methods in Approximate Analysis, 2nd edn. MCNMO, Moscow (2004)
31. Korobov, N.M.: Number-Theoretic Methods in Approximate Analysis. Fizmatgiz, Moscow (1963)

32. Korobov, N.M.: On the number-theoretic methods of numerical integration. In: Ushkevich, A.P. (ed.) Istoriko-Matematicheskie Issledovanija vol. 35, pp. 285–301. Nauka, Moscow (1994)
33. Korobov, N.M.: The approximate calculation of multiple integrals using methods of number theory. Dok. Akad. Nauk SSSR. **115**(6), 1062–1065 (1957)
34. Korobov, N.M.: On the approximate calculation of multiple integrals. Dok. Akad. Nauk SSSR. **124**(6), 1207–1210 (1959)
35. Korobov, N.M.: Calculation of multiple integrals by the optimal coefficients method. Vestn. Mosk. Univ. Ser. Mat. Mekh. Astron. Fiz. Him. **4**, 19–25 (1959)
36. Korobov, N.M.: Properties and calculation of optimal coefficients. Dok. Akad. Nauk SSSR **132**(5), 1009–1012 (1960)
37. Korobov, N.M.: Some problems in the theory of Diophantine approximation. Russ. Math. Surv. **22**(3), 80–118 (1967). doi:10.1070/RM1967v022n03ABEH001220
38. Korobov, N.M.: On the calculation of optimal coefficients. Dok. Akad. Nauk SSSR. **267**(2), 289–292 (1982)
39. Korobov, N.M.: Quadrature formulas with combined grids. Math. Notes. **55**(2), 159–164 (1994). doi:10.1007/BF02113296
40. Rebrova, IYu.: The continuity of the generalized hyperbolic zeta function of lattices and its analytic continuation. Izv. Tul. Gos. Univ. Ser. Mat. Mekh. Inform. **4**(3), 99–108 (1998)
41. Temirgaliev, N.: Application of divisor theory to the numerical integration of periodic functions of several variables. Math. USSR Sb. **69**(2), 527–542 (1991). doi:10.1070/SM1991v069n02ABEH001250
42. Voronin, S.M., Temirgaliev, N.: Quadrature formulas associated with divisors of the field of Gaussian numbers. Math. Notes Acad. Sci. USSR. **46**(2), 597–602 (1989). doi:10.1007/BF01137622
43. Voronin, S.M.: On quadrature formulas. Russ. Acad. Sci. Izv. Math. **45**(2), 417–422 (1995). doi:10.1070/IM1995v045n02ABEH001657
44. Voronin, S.M.: The construction of quadrature formulae. Izv. Math. **59**(4), 665–670 (1995). doi:10.1070/IM1995v059n04ABEH000028

Chapter 3
The Distribution of Values of Arithmetic Functions

G. V. Fedorov

Abstract Let us usual $\tau_k(n)$ denote the number of ways n may be written as a product of k fixed factors. In this chapter there introduce the notation

$$D_k(x) = \sum_{n \le x} \tau_k(n).$$

We show that the asymptotic formula for $D_k(x)$ is changing with growing values of k and present specific values of k, which is a change.

In [1], this author obtained the estimate

$$D_k(x) \le x \sum_{j=0}^{k-1} \binom{k-1}{j} \frac{\ln^j x}{j!}, \tag{3.1}$$

for $D_k(x)$, which is uniform in the parameter k and holds for any real $x \ge 1$ and integer $k \ge 2$.

The value of the quantity $D_k(x)$ equals the number of points in the integer lattice in a domain of the form $1 \le x_1, x_2, \ldots x_k \le x$. Note that if the parameter k grows as $x \to \infty$, then the form of the asymptotic formula for $D_k(x)$ is different from that of the formula for fixed k. In 2001, Pavlov [3] proved the following assertion.

Theorem 3.1 *Suppose that $x \to \infty$, k is an integer, and $C_1(\ln x)^\beta < k < C_2(\ln x)^\alpha$, where $\alpha < \frac{2}{3}$ and $\beta > 6$ are fixed and C_1 and C_2 are positive constant. Then*

$$D_k(x) = x \frac{(\ln x)^{k-1}}{(k-1)!} e^{\gamma \frac{k^2}{\ln x}} \left(1 + O\left(k^{-\rho_0}\right)\right),$$

G. V. Fedorov (✉)
Fedorov Gleb Vladimirovich, Lomonosov Moscow State University, GSP-1,
Leninskie Gory, Moscow, Russian Federation 119991
e-mail: glebonyat@mail.ru

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*,
Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_3,

where γ is the Euler constant and $\rho_0 > 0$ is positive and does not depend on k and x.

In this chapter, we obtain more accurate boundary values of the parameter k in Pavlov's theorem. The following assertion is valid.

Theorem 3.2 (Main Theorem) *Suppose that the integer parameter satisfies the condition $k = k(x) \to \infty$ as $x \to \infty$, and for some fixed $0 < \rho < \frac{1}{3}$, the inequality $k \ll (\ln x)^{\frac{4}{5+\rho}}$ holds. Then, the asymptotic formula*

$$D_k(x) = x\frac{(\ln x)^{k-1}}{(k-1)!}\exp\{q_k(x)\}\,L_k(x)\left(1 + O\left(\frac{k^{5+\rho}}{\ln^4 x}\right) + O\left(k^{3\rho-1}\right)\right),$$

is valid, in which the functions $q_k(x)$ and $L_k(x)$ are defined by

$$q_k(x) = \gamma_0\frac{k^2}{\ln x} - \left(\gamma_0^2 + \gamma_1\right)\frac{k^3}{\ln^2 x} + \left(\frac{5}{3}\gamma_0^3 + 3\gamma_0\gamma_1 + \frac{\gamma_2}{2}\right)\frac{k^4}{\ln^3 x}, \tag{3.2}$$

$$L_k(x) = 1 - \frac{k}{\ln x}\left(\gamma_0 + \frac{3}{2}\right) + \frac{k^2}{\ln^2 x}\left(\frac{3\gamma_0^2}{2} + \gamma_1 + 3\gamma_0 + \frac{7}{4}\right) -$$

$$- \frac{k^3}{\ln^3 x}\left(\frac{21}{4}\gamma_0^2 + \frac{5}{2}\gamma_0^3 + 3\gamma_0\gamma_1 + \frac{3}{2}\gamma_1 + \frac{21}{4}\gamma_0 + \frac{15}{8}\right), \tag{3.3}$$

and the Stieltjes constants are defined by

$$\gamma_n = \lim_{m\to\infty}\left(\sum_{k=1}^{m}\frac{(\ln k)^n}{k} - \frac{(\ln m)^{n+1}}{n+1}\right), \tag{3.4}$$

in particular, $\gamma_0 = \gamma$ is the Euler constant.

The proof of the *main theorem* is based on the following assertion.

Lemma 3.1 *Suppose that $\sigma = 1 + \frac{1}{b}$, $b = \gamma_0 + \frac{\ln x}{k}$ and*

$$I_k(x) = \frac{1}{2\pi i}\int_{\sigma-\frac{i}{2}}^{\sigma+\frac{i}{2}}\zeta^k(s)\frac{x^{s+1}}{s(s+1)}ds.$$

Suppose also that $x \to \infty$ and $k \to \infty$ so that, for some fixed $0 < \rho < \frac{1}{3}$, the inequality $k \ll (\ln x)^{\frac{4}{5+\rho}}$ holds. Then the asymptotic formula

$$I_k(x) = \frac{x^2}{2} \cdot \frac{(\ln x)^{k-1}}{(k-1)!} \exp\{q_k(x)\} L_k(x) \left(1 + O\left(\frac{k^{5+\rho}}{\ln^4 x}\right) + O\left(k^{3\rho-1}\right)\right), \tag{3.5}$$

is valid, in which the functions $q_k(x)$ *and* $L_k(x)$ *are determined from* (3.2) *and* (3.3).

This lemma sharpens the corresponding lemma from Pavlov's chapter ([3], Lemma 1).

The proof of *lemma* 3.1 uses the Laurent expansion of Riemann's zeta function $\zeta(s)$ in the neighborhood of the pole $s = 1$

$$\zeta(s) = \frac{1}{s-1} + \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \cdot \gamma_n \cdot (s-1)^n,$$

where constants γ_n defined from 3.4.

As is known, for $\Re s > 1$

$$\sum_{n=1}^{\infty} \frac{\tau_k(n)}{n^s} = \zeta^k(s),$$

where $\zeta(s)$ is the Riemann zeta function. We have (see [2])

$$\int_1^x D_k(t)dt = \frac{1}{2\pi i} \int_{\sigma-iT}^{\sigma+iT} \zeta^k(s) \frac{x^{s+1}}{s(s+1)} ds + R(x) = J_k(x) + R(x), \tag{3.6}$$

where the parameter σ is the same as in the *lemma* 3.1. Using estimate (3.1), we obtain the following estimate for the remainder:

$$R(x) \ll \frac{x^2}{T} \left(\frac{\ln x}{k}\right)^k \exp\left\{k + (\gamma_0 + 1)\frac{k^2}{\ln x}\right\}$$

$$+ \left(\frac{x^2}{T} + x\frac{\ln T}{\ln x}\right) \sqrt{k} \left(\frac{\ln x}{k}\right)^k \exp\left\{k + \frac{k^2}{\ln x}\right\}. \tag{3.7}$$

For $k \ll (\ln x)^{\frac{5}{7}}$, we deform the interval of integration in $J_k(x)$ as

$$\int_{\sigma-iT}^{\sigma+iT} = \int_{\sigma-iT}^{1-iT} + \int_{1-iT}^{1-\frac{i}{2}} + \int_{1-\frac{i}{2}}^{\sigma-\frac{i}{2}} + \int_{\sigma-\frac{i}{2}}^{\sigma+\frac{i}{2}} + \int_{\sigma+\frac{i}{2}}^{1+\frac{i}{2}} + \int_{1+\frac{i}{2}}^{1+iT} + \int_{1+iT}^{\sigma+iT},$$

by virtue of the estimate $|\zeta(1+it)| \leq C \ln^{\frac{2}{3}} |t|$ where $|t| > 2$ and C is some constant, we have

$$J_k(x) = I_k(x) + O\left(C^k x^2 (\ln T)^{\frac{2k}{3}}\right).$$

We can choose the parameter T so that $\frac{k^2}{\ln x} \ll \ln T \ll \left(\frac{\ln x}{k}\right)^{\frac{3}{2}}$ and the remainders in expressions (3.6) and (3.7) do not exceed those in (3.5). Applying the *lemma* 3.1, we obtain

$$\int_1^x D_k(t)dt =$$

$$= \frac{e^k}{\sqrt{2\pi k}} \frac{x^2}{2} \left(\frac{\ln x}{k}\right)^{k-1} \exp\{q_k(x)\} L_k(x) \left(1 + O\left(\frac{k^{5+\rho}}{\ln^4 x}\right) + O\left(k^{3\rho-1}\right)\right). \tag{3.8}$$

In the case of $k \gg (\ln x)^{\frac{2}{3}}$, we decompose the integral $J_k(x)$ into three integrals as

$$J_k(x) = \frac{1}{2\pi i} \int_{\sigma - iT}^{\sigma - \frac{i}{2}} + \frac{1}{2\pi i} \int_{\sigma - \frac{i}{2}}^{\sigma + \frac{i}{2}} + \frac{1}{2\pi i} \int_{\sigma + \frac{i}{2}}^{\sigma + iT} =$$

$$= I_k(x) + O\left(x^2 \left(\frac{\ln x}{k}\right)^k \exp\left\{k + \frac{\gamma_0^2 k^3}{\ln^2 x}\right\}\right),$$

Let $T = x$; then the remainder $R(x)$ does not exceed the remainders in formula (3.5) given in the *lemma* 3.1; therefore, the relation (3.8) again holds.

The function $D_k(x)$ is nondecreasing. We have

$$\frac{1}{h} \int_{x-h}^{x} D_k(t)dt \le D_k(x) \le \frac{1}{h} \int_x^{x+h} D_k(t)dt.$$

Applying (3.8) and choosing $h = x \frac{k^{5+\rho}}{\ln^4 x}$, we obtain the assertion of the *Main Theorem*.

References

1. Fedorov, G.V.: Vestn. Mosk. Univ. Ser. 1: Mat. Mekh. **2**, 50–53 (2010)
2. Karacuba, A.A.: Izv. Akad. Nauk SSSR. Ser. Mat. **3**, 475–483 (1972)
3. Pavlov, A.I.: Dokl. Math. **63**, 48–51 (2001)

Chapter 4
On the One Method of Constructing Digital Control System with Minimal Structure

V. V. Palin

Abstract We consider the linear digital control system with invariable matrix A. In this report we introduce one method which permit to obtain the characteristic of completely controllability and construct the matrix of control B with minimal structure without calculation of eigenvalues of matrix A.

4.1 The Statement of Problem and Some Familiar Results

Let us discuss stationary open discrete system

$$X_{k+1} = AX_k. \tag{4.1}$$

We will find the full rank matrix B of control actions with $n \times p$ size such that the following closed stationary system

$$X_{k+1} = AX_k + BU_k + F_k \tag{4.2}$$

will be completely controllable.

Definition 4.1 Characteristic of completely controllable for system (4.1) is the minimal number $p \in \mathbb{N}$ such that the system (4.2) can make completely controllable by the choice of full rank matrix B of $n \times p$ size.

On 2010 the article [1] was published in journal Doklady Akademii Nauk. There the structural minimization problem discussed and the following result obtained:

V. V. Palin (✉)
Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow,
Russian Federation 119991
e-mail: grey_stranger84@mail.ru

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*,
Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_4,

Theorem 4.1 *Characteristic of completely controllable of system* (4.1) *is equivalent to the maximal geometric multiplicity of eigenvalues of A.*

More over, in this article the method of constructing the matrix B was established for the case where the Jordan canonical form of A was given.

In this article we obtain the method to find the characteristic of completely controllable for (4.1) and constructing the matrix B without evaluation of eigenvalues of A.

4.2 Definitions and Some Preliminary Transformations

Suppose that A is square matrix with $n \times n$ size, λ_j are eigenvalues of A, $q_A(x)$ is the minimal polynomial for A and $d_A(x) = \det(xE - A)$. We note that the polynomial $q_A(x)$ can be found without calculations of eigenvalues of A; $d_A(x)$ is characteristic polynomial of A, multiplied by (-1) powering relevant (so that the leading coefficient equal to 1) hence, this polynomial can be obtained without calculations of eigenvalues of A. Let

$$q_A(x) = \prod_{j=1}^{m} (x - \lambda_j)^{k_j}.$$

We denote

$$q(x, A, \geq r) = \prod_{j:k_j \geq r} (x - \lambda_j),$$

$$q(x, A, = r) = \prod_{j:k_j = r} (x - \lambda_j).$$

Let us note that the polynomial $q(x, a, \geq r)$ can be found without factorization of $q_A(x)$. For example,

$$q(x, A, \geq 1) = \frac{q_A(x)}{\text{g.c.d.}(q_A(x), q'_A(x))},$$

and $q(x, A, \geq 2)$ can be obtained by the same formula, where $q_A(x)$ changes by g.c.d.$(q_A(x), q'_A(x))$ and so forth. We can evaluate the polynomials $q(x, A, = r)$ by the polynomials $q(x, A, \geq r)$:

$$q(x, A, \geq r) = q(x, A, = r)q(x, A, \geq r + 1).$$

Similarly we define $d(x, A, \geq r)$ and $d(x, A, = r)$.

4.3 The Method to Obtain the Characteristic of Completely Controllable

Let us denote $A_1 = q(A, A, \geq 1)$. Because the polynomial $d_A(x)$ is the divisor for $(q(x, A, \geq 1))^n$ and $d_A(A) = 0$, the following identity holds: $A_1^n = 0$. More over, from the definition of the eigenvector of A and the Jordan canonical form for the matrix A_1 it follows that the eigenvectors $\{v_1, \ldots, v_s\}$ of the matrix A form the basis in the kernel of A_1.

If $A_1 = 0$ then the matrix A has the basis consists of it's eigenvectors. Hence, the geometric multiplicity of any eigenvalue of the matrix A is equivalent to it's algebraical multiplicity. Thus, $p = \max\{t \mid d(x, A, \geq t) \neq 1\}$ in this case.

Suppose that $A_1 \neq 0$, $\mathrm{Ker}(A_1) = \mathrm{Lin}\{v_1, \ldots, v_s\}$. Let us note that we can find vectors of the basis of the kernel of A_1 as orthogonal complement of the linear envelope of the set of rows of A_1. Suppose that $v_{s+1}, \ldots, v_n$ is basis of the set of columns of A_1^T. Let C_1 be the matrix constructed of vectors $v_1, \ldots, v_n$ as columns. Let $j \leq s$ be the fixed index and $e_1, \ldots, e_n$ is a basis consists of unit vectors. By virtue of definition of the matrix C_1 we have $C_1 e_j \in \mathrm{Lin}\{v_1, \ldots, v_s\}$, $AC_1 e_j \in \mathrm{Lin}\{v_1, \ldots, v_s\}$, $C_1^{-1} AC_1 e_j \in \mathrm{Lin}\{e_1, \ldots, e_s\}$. Hence the matrix $C_1^{-1} AC_1$ is sectional upper triangular:

$$C_1^{-1} AC_1 = \begin{pmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{pmatrix}. \tag{4.3}$$

Further from the arguments given above it follows that there is one-to-one correspondence between the eigenvectors of A and the eigenvectors of M_{11}. Hence, the characteristics of completely controllable for matrix A and M_{11} are equivalent. Let us note that M_{11} has the basis consists of eigenvectors.

Remark 4.1 The set $\mathrm{Ker}(d(A, M_{11}, = t))$ is the linear envelope of all eigenvectors of A such that their correspond eigenvalues has geometric multiplicity of exactly t.

4.4 Auxiliary Statements

To describe a method of constructing a matrix B without finding eigenvalues of the matrix A, we need two lemmas. The proof of the first of them is trivial, and we omit it.

Theorem 4.2 *Let $\lambda_1, \ldots, \lambda_s$—eigenvalues of a matrix A, vector $h \in \mathrm{Ker}(\prod_{j=1}^{s}(A - \lambda_j E))$, $h \neq 0$. Then if $q \in \mathbb{N}$ of such that vectors $h,\ Ah,\ A^2h, \ldots, A^{q-1}h$ are linearly independent, and vectors $h,\ Ah,\ A^2h, \ldots, A^q h$ linearly dependent, there will be eigenvectors $z_1, \ldots, z_q \in \mathrm{Ker}(\prod_{j=1}^{s}(A - \lambda_j E))$ such that $h = z_1 + z_2 + \ldots + z_q$.*

Back, if there are eigenvectors $z_1, \ldots, z_q \in \mathrm{Ker}(\prod_{j=1}^{s}(A - \lambda_j E))$ *such that* $h = z_1 + z_2 + \cdots + z_q$, *vectors* $h,\ Ah,\ A^2h, \ldots, A^{q-1}h$ *are linearly independent, and vectors* $h,\ Ah,\ A^2h, \ldots, A^qh$ *linearly dependent.*

Theorem 4.3 *Let* $\pi_k(x)$ *is a polynomial of degree* t_k *such that* $\mathrm{Ker}(\pi_k(A))$ *is the linear envelope of the eigenvectors of the matrix A of geometric multiplicity of exactly k. Then, without finding the eigenvalues of the matrix A can be built vectors* $w_1, \ldots, w_{CC}$ *such that* $\mathrm{Ker}(\pi_k(A)) = \mathrm{Lin}\{w_1, Aw_1, \ldots, A^{t_k-1}w_1, \ldots, A^{t_k-1}w_{kk}\}$.

Proof We give an algorithm that allows each step reduce one of k or t_k by 1. Let us take an arbitrary non-zero vector $h_1 \in \mathrm{Ker}(\pi_k(A))$. There is a $q \in \mathbb{N}$ such that the vectors $h_1, ACAh_1, \ldots, A^{q-1}h_1$ are linearly independent, and vectors $h_1, ACAh_1, p(A^2h_1, \ldots, A^qh_1$ are linearly dependent. On Lemma 1, we obtain that $q \leq t_k$. If $q = t_k$, then $w_1 = h_1$, and, demanding further orthogonality of the vector h_2 to all vectors A^jh_1, we obtain that k has decreased by 1. If $q < t_k$, then, by Lemma 1, there are eigenvectors $z_1, \ldots, z_q$ such that $h_1 = z_1 + \cdots + z_q$. Add orthogonal vectors $v_{q+1}, \ldots, v_n$ in the system of vectors $v_1 = h_1, v_2 = Ah_1, \ldots, v_q = A^{q-1}h_1$ to obtain the basis of all space. We write the matrix C_2, the columns of which are vectors $v_1, \ldots, v_n$. As in the previous section, the matrix $C_2AC_2^{-1}$ is upper triangular. Let us denote by N_{11} its upper the left bloc. Left to note that there exists a polynomial $\tilde{p}(x)$ such that

$$\pi_k(x) = d(x, N_{11}, \geq 1)\tilde{p}(x).$$

Thus, the problem for the polynomial $\pi_k(x)$ is reduced to the problem for polynomials $d(x, N_{11}, \geq 1)$ and $\tilde{p}(x)$, the sum of which degrees is equal to t_k. This means that in this case we have managed to reduce the t_k at least by 1.

4.5 The Absence of Associated Vectors Case

Let us discuss the method of constructing B in the case when the matrix A has a basis consists of the eigenvalues. In this case we construct polynomials $\pi_k(x) = d(x, A, = k)$ and, using lemma 2, we obtain vectors $w_{1k}, \ldots, w_{kk}$ for any of these polynomials. Let us denote

$$b_j = \sum_{k=j}^{p} w_{jk}.$$

Left to notice that the matrix B with the columns $b_1, \ldots, b_p$ is sought-for matrix.

4.6 The Case of General Position

Let τ is the degree of polynomial $q_A(x)$. Let $V_1 = \mathrm{Ker}(q(A, A, \geq 1))$, V_2 are the set of all vectors from $\mathrm{Ker}(q(A, A, \geq 2))$, orthogonal to V_1, V_3 are the set of all vectors from $\mathrm{Ker}(q(A, A, \geq 3))$, orthogonal to $V_1 + V_2$, etc. Let us notice that for finding basis in any V_j it suffices to use orthogonalization method. Let W_1 is the set of orthogonal to AV_2 vectors from V_1, W_2 is the set of orthogonal to AV_3 vectors from V_2, etc.. The basis in each of the spaces W_j can be found by using orthogonalization method. Let us consider the mapping

$$g\colon \sum_{j=1}^{\tau} W_j \to V_1.$$

By this mapping the vector $gw_j \in V_1$ is associated to the vector $w_j \in W_j$ such that $gw_j \in V_1$ is orthogonal projection of vector $A^{j-1}w_j$ on V_1. The mapping g is invertible: it is sufficient to note that g is linear, has zero kernel, and to set the basis in any of W_j, and the result of mapping g on this basis.

Let us describe the method of constructing the matrix B in the case of general position. As well as the case of absence of associated vectors, we construct the polynomials $\pi_k(x) = d(x, M_{11}, = k)$ and, using the lemma 2 for each polynomial, we obtain the vectors $w_{1k}, \ldots, w_{kk}$. Further, we put

$$\tilde{b}_j = \sum_{k=j}^{p} w_{jk},$$

$$b_j = g^{-1}\tilde{b}_j.$$

Left to notice that the matrix B with the columns $b_1, \ldots, b_p$ is sought-for matrix.

Reference

1. Zubov, A.V., Dikusar, V.V., Zubov, N.V.: Kriterii Upravlyaemosti Statsionarnykh Sistem. Doklady Akademii Nauk. **430**(1), 13–14 (2010)

Chapter 5
On Norm Maps and "Universal Norms" of Formal Groups Over Integer Rings of Local Fields

Nikolaj M. Glazunov

To the memory of Oleg Nikolaevich Vvedenskii (1937–1981)

Abstract We review and investigate norm maps and "universal norms" of formal groups over integer ring of local and quasi-local fields. Theorem on triviality of universal norm group of one dimensional fornal groups of reduction height 3 over integer ring of local and quasi-local fields is presented. The theorem on triviality of universal norm group is based on the lemma about function that gives the minimal degree of elements of the subgroup F_K^t of the group F_K that contains the norm group $N_{L/K}(F_L^n)$. In the case of formal groups of elliptic curves the function has used by O. N. Vvedenskii and is denoted as $\mu(n)$. The proof of the lemma is also presented.

5.1 Introduction

Under the construction by Shafarevich [1], Tate [2], Ogg [3], Vvedenskii [4, 5] the analog of local and quasi-local class field theory for elliptic curves and abelian varieties the authors use arithmetic properties of formal groups that corresponds to elliptic curves. Foundations of local and quasi-local class field theories of elliptic curves in the framework were constructed by Vvedenskii [4, 5] in contexts of elliptic curves over local and quasi-local fields. Important statements of these theories were introduced as statements about norm maps of commutative formal groups of elliptic curves.

It is well known that formal groups of elliptic curves over finite fields have height (reduction height) one or two [6–11].

N. M. Glazunov (✉)
National Aviation University, Kiev, Ukraine
e-mail: glanm@yahoo.com

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_5,

Let A be an elliptic curves over quasi-local fields K, $F(x, y)$ its formal group over the ring of integers O_K of K, $\mathscr{D}_K^*$ it's group of universal norms [4, 5]. In the case O.N. Vvedenskii have proved.

Theorem 5.1 [4, 5] $\mathscr{D}_K^* = 0$.

Author extends, following to the advice of O. N. Vvedenskii, Theorem 5.1 and some another results of O. N. Vvedenskii to more general formal groups and present their in papers [8, 10]. Complete proves of the results are contained in author's candidate dissertation that is not published.

Here we present theorem on triviality of universal norm group of one dimensional formal groups of height 3 over integer ring of local and quasi-local fields and present the lemma about function $\mu(n)$ that gives the minimal degree of elements of the subgroup F_K^t of the group F_K that contains the norm group $N_{L/K}(F_L^n)$.

Let K be a complete discrete variation field with the ring of integers O_K and the maximal ideal M_K.

A complete discrete variation field with finite residue field is called a *local* field [12].

A complete discrete variation field K with algebraically closed residue field k is called a *quasi-local* field [5]. Below we will suppose that in the case the characteristic of k satisfies $p > 0$.

Let K be a local or quasi-local field. If K is a local field [12] and has the characteristic 0 then it is a finite extension of the field of p-adic numbers $\mathbf{Q}_p$. Let v_K be the normalized exponential valuation of K. If $[K : \mathbf{Q}_p] = n$ then $n = e \cdot f$, where $e = v_K(p)$ and $f = [k : \mathbf{F}_p]$, where k is the residue field of K (always assumed perfect).

If K has the characteristic $p > 0$ then it isomorphic to the field $k((T))$ of formal power series, where T is uniformizing parameter.

Let L be a finite extension of a local field K, k, l their residue fields, $p = char\ k$ and $e_{L/K}$ ramification index of L over K.

An extension L/K is said to be *unramified* if $e_{L/K} = 1$ and extension l/k is separable.

An extension L/K is said to be *tamely ramified* if p not devides $e_{L/K}$ and the residue extension l/k is separable.

An extension L/K is said to be *totally ramified* if $e_{L/K} = [L : K] = (char\ k)^s, s \geq 1$.

Let L/K be the finite Galois extension of quasi-local field K with Galois group G, $F(x, y)$ one dimensional formal group low over the ring of integers O_K of the field K, $F(M_K)$ be the G-module, that is defined by the group low $F(x, y)$ on the maxilal ideal M_K of the ring O_K, $M_K^t (t \in \mathbf{Z}, t \geq 1)$ be the subgroup of t-th degrees of elements from M_K, $F_K^t := F(M_K^t)$.

Definition 5.1 For $n \in \mathbf{Z}$ the function $\mu(n)$, $N_{L/K}(F_L^n) \subset F_K^{\mu(n)}$ is defined by the condition: $F_K^{\mu(n)}$ is the least of subgroups F_K^t $(t = 1, 2, \ldots)$ containes $N_{L/K}(F_L^n)$.

Below we will suppose that $char\ k > 3$.

5.2 Norm Maps

Here we use results on formal groups from [9–11, 13]. Let $F_L = F(M_L)$ be the G-module that is defined by the n-dimentional group low $F(x, y)$ on the product $(M_L)^n := M_L \times \cdots \times M_L$, ($n$ times) of maximal ideals of the ring O_L of any finite Galois extension L of the field K.

Definition 5.2 The norm map $N : F_L \to F_K$ of the module F_L to F_K is defined by the formula $N(a) = (((a +_F \sigma a) +_F \cdots) +_F \sigma_s a)$, where $a +_F b$ denotes the addition of points in the sense of group structure of the module F_L, $a, b \in M_L$, $G = Gal(L/K)$, $\sigma_s \in G$, $[G : 1] = s$.

Let $p := char\ k$, $e := v_K(p)$, ($e = +\infty$, if characteristic of the field K is equal p and e is positive integer in the opposide case), L/K be the Galois extension of the prime degree q, $F(x, y)$ be the one dimensional group low over O_K. Let $p := char\ k > 0$.

Lemma 5.1 *If* $\Pi_s \in \pi_L^s \cdot O_L$, $s \geq 1$ *then*
$N(\Pi_s) \equiv Tr(\Pi_s) + \sum_{n=1}^{\infty} c_n [Norm\ \Pi_s]^n (mod\ Tr(\pi_L^{2s} \cdot O_L))$
where $c_n \in O_K$ *are coefficients of the p-iteration of the group low. The iteration is defined below.*

(In paper [6] the lemma has proved for one dimensional group lows that correspond to elliptic curves)

Proof At first make two remarks:

1. If $F(x, y)$—one dimensional group low over the ring O_K, then p-iteration $[p]_F(T)$ of the group low F has the form [9]
   $[p]_F(T) = p(T + \cdots) + \sum_{i=1}^{\infty} c_i T^{pi}$,
   where dots denote intermediates of the degree greater than one.
2. If the series expansion of the expression $(((t_1 +_F t_2) +_F \cdots) +_F t_n)$ includes monomial $t_1{}^{\alpha_1} \cdots t_q{}^{\alpha_q}$, then it also includes a monomial that is the result of acting of arbitrary permutation of digits $1, 2, \ldots, q$ on it.

Let us go to the proof of the lemma. Let $G = Gal(L/K)$. If $\omega = r_1 + r_2\sigma + \cdots + r_q\sigma^{q-1}$ is an element of the group algebra $\mathbf{Z}[G]$ (where $\mathbf{Z}$ is the ring of integers). Let
$\Pi_s^{\omega} := \Pi_s^{r_1} (\sigma \Pi_s^{r_2}) \cdots (\sigma^{q-1} \Pi_s^{r_q})$.
We have $N(\Pi_s) = (((\Pi_s +_F \sigma \Pi_s) +_F \cdots) +_F \sigma^{q-1} \Pi_s) = \sum_{(r_1,\ldots,r_q)} d_{r_1,\ldots,r_q} \Pi_s^{\omega}$ where $d_{r_1,\ldots,r_q} \in O_K$, and sum by corresponding ω. By symmetry (see remark 2) in the expansion of $N(\Pi_s)$ with $d_{r_1,\ldots,r_q} \Pi_s^{\omega}$ comes also $d_{r_1,\ldots,r_q} \Pi_s^{\sigma^i \omega}$ ($i = 1, 2, \ldots, q-1$). Since
$\sigma^i \omega = \omega$
(i is one of numbers $i = 1, 2, \ldots, q-1$), so $\omega = n(1 + \sigma + \cdots + \sigma^{q-1})$. Hence

$$N(\Pi_s) = \sum_{n=1}^{\infty} d_n [Norm(\Pi_s)]^n + \sum_{\omega} d_{r_1,\ldots,r_q} Tr(\Pi_s^{\omega}), \tag{5.1}$$

where sum by ω such that do not satisfy the condition $\sigma^i \omega = \omega$.
If $r_1 + \cdots + r_q > 1$ then by ([14], lemma 2)
$Tr(\Pi_s^{\omega}) \subset Tr(\pi_L^{2s} \cdot O_L)$, hence

$$N(\Pi_s) \equiv Tr(\Pi_s) + \sum_{n=1}^{\infty} d_n [Norm\ \Pi_s]^n (mod\ Tr(\pi_L^{2s} \cdot O_L)). \tag{5.2}$$

Demonstrate that as a d_n we may take c_n from the expansion of $[p]_F(T)$. This follow from the fact that as d_n so c_n define to $mod\ p$.

Let $r := v_K(c_1), r_j := v_K(c_j)$, $j > 1$ and let the height of F is $\infty > h \geq 1$; recall that $v_K(c_{p^{h-1}}) = 0$.

By ([14], lemma 2) $Tr(\pi_L^n \cdot O_L) = \pi_K^{y_0(n)}$ where $y_0(n) = \lfloor \frac{(m+1)(p-1)+n}{p} \rfloor$.

Put $y_1(n) = r + n$, $y_2(n) = r_2 + 2n, \ldots, y_{p-1}(n) = r_{p-1} + (p-1)n$, $y_p(n) = r_n + pn, \ldots, y_{p^{h-1}}(n) = r_n + pn$.

Lemma 5.2

$$\mu(n) = \min\{y_0(n), y_1(n), y_p(n), y_{p^2}(n), \ldots, y_{p^{h-1}}(n)\}. \tag{5.3}$$

Proof (we follow to [7]).

Define $\mu_1(n) = \min\{y_0(n), y_1(n), y_2(n), \ldots, y_p(n), y_{p^2}(n), \ldots, y_{p^{h-1}}(n)\}$.

It is clear since the estimation (5.2) that $\mu(n) \geq \mu_1(n)$ ($\mu(n)$ is understood in the sense of the definition 5.1). Choose Π_n such that $v_L(\Pi_n) = n$, $v_K(Tr(\Pi_n)) = y_0$.

Let $d \in O_K$. Consider expression $N(d\Pi_n)$. By (5.1) in the case $d \in O_K$ the terms from $N(d\Pi_n)$ that are included in the ideal $Tr(\pi_L^{2n})$ and have the form

$$Tr(\sigma^{i_1}(d\Pi_n)^{k_1} \cdots \sigma^{i_s}(d\Pi_n)^{k_s}) \tag{5.4}$$

under $k_1 + \cdots + k_s \geq p + 1$ will have the norm in K greater then $y_0(n)$. This follow from the computation by the formula for $y_0(n)$. Hence

$N(d\Pi_n) = \pi_K^{\mu_1(n)}[(\pi_K^{-\mu_1(n)} Tr(\Pi_n))d +$ (summands contain d from 2 to p-s degree, that obtained from terms of N(dΠ_n), that include in

$$Tr(\pi_L^{2n})) + \sum_{i=1}^{ph} \pi_K^{-\mu_1(n)} \times c_i [Norm(\Pi_n)]^i d^{pi} + \cdots] \tag{5.5}$$

where dots denote terms of higher orders.

Term $\pi_K^{\mu_1(n)}$ in (5.5) holds coefficient that is polynomial from d of degree not greater than p^h; if $\mu_1(n) = y_j(n)(j = 0, 2, 3, \ldots, p^{h-1}$; j is different from 1) then the coefficient under d^{pj} is not equal zero $mod\ \pi_K$, hence $\mu(n) = \mu_1(n)$; if

$$\mu_1(n) = y_1(n) < y_0(n), y_2(n), \ldots, y_p(n), y_{p^2}(n), \ldots, y_{p^{h-1}}(n).$$

then terms from $N(d\Pi_n)$ that are included in $Tr(\pi_L^{2n})$, will have in K a norm that is not less then $y_0(n)$, hence only coefficient under d^p will differ from zero under $mod\ \pi_K$, hence again $\mu(n) = \mu_1(n)$. Hence always

$$\mu(n) = \mu_1(n).$$

Demonstrate now that actually

$$\mu_1(n) = \min\{y_0(n), y_1(n), y_p(n), y_{p^2}(n), \ldots, y_{p^{h-1}}(n)\}.$$

We prove this by induction on n. If $n = 1$ and $\mu_1(1) = y_0(1)$ then the lemma is proved, and then

$y_0(1) \le y_i(1)$ $(i = 1, 2, 3, ..., p^{h-1})$ and all $y_i(n)$, $i \ne 0$ grow faster then $y_0(n)$.

If $\mu_1(1) = y_i(1) < y_0(1)$, $1 \le i \le p^{h-1}$ (specifically: $i = r_0$), then demonstrate at first that $\mu_1(n)$ is strictly increasing function

$$\mu_1(1) < \mu_1(2).$$

If $\mu_1(2) = y_{r_0}(2)$ $(r_0 \ne 0)$ then we have

$y_2(1) \le y_{r_0}(2)$, that is $\mu_1(1) < \mu_1(2)$.

But if $\mu_1(2) = y_0(2)$ then $\mu_1(1) = y_r(1) < y_0(1)$, hence $y_2(1) < y_0(1) \le y_0(2)$, and again $\mu_1(1) < \mu_1(2)$.

Thereby the homomorphism

$$F_L^1/F_L^2 \xrightarrow{N_1^*} F_K^{\mu_1(1)}/F_K^{\mu_1(1)+1} \tag{5.6}$$

that is induced by N is defined. Under $\pi_L - \pi_K$ isomorphisms [7] it passes to homomorphism $\overline{N}_1{}^* : G_a(l) \to G_a(k)$ where $G_a(k)$ is the additive group of the field $l = O_L/M_L$ that is defined by polynomial from (5.5) under reduction by $mod\ \pi_K$. But any homomorphism of additive groups of the field of characteristic $p > 0$ is given by the polynomial from $T, T^p, T^{p^2}, \ldots$ (sums of degrees of Frobenius automorphism), hence in the case $n = 1$ the lemma is proved.

Let lemma is true for $n = n_0$. Prove it for $n = n_0 + 1$. If $\mu_1(n_0) = y_{n_0}(n)$ then the lemma is proved. If $\mu_1(n_0) = y_j(n_0)$ $(1 \le j \le p^{h-1}, j \ne 0)$ then we have

$$\mu_1(n_0) < \mu_1(n_0 + 1)$$

Ipso facto the homomorphism

$$F_L^{n_0}/F_L^{n_0+1} \xrightarrow{N_{n_0}^*} F_K^{\mu_1(n_0)}/F_K^{\mu_1(n_0)+1} \tag{5.7}$$

that is induced by N is defined. And again the passage to the homomorphism $\overline{N}_{n_0}{}^*$: $G_a(l) \to G_a(k)$ demonstrates that (5.3) takes place.

5.3 Results

Let $F(x, y)$ be the one dimensional formal groups of height 3 over integer ring of local and quasi-local fields K.

Consider the tower of fields

$$K = L_0 - L_1 - L_2 - \cdots - L_{s-1} - L_s \tag{5.8}$$

where L_i/L_{i-1}, $(i = 1, 2, \ldots, s)$ are Galois extensions with Galois groups $\mathbf{Z}/p\mathbf{Z}$.

Let $\mu_i(n)$ be the function of the definition 5.1 that is computed on the i-s floor of the tower (5.8) and let m_i be the number of the last nontrivial ramification group of the extension L_i/L_{i-1}.

Put $r_1 := \nu_K(c_p)$, $r_2 := \nu_K(c_{p^2})$, $e := \nu_K(p)$.

Lemma 5.3 *Depend on numbers r_1, r_2, e the function $\mu_i(n)$ is computed by the next four formulas:*

(i) *If $r_1, r_2 \geq e$ then the computation of the $\mu_i(n)$ makes by the formula*

$$\mu_i(n) = \begin{cases} p^2 n, n \leq \frac{m_i+1}{p^2+p+1} \\ \lfloor \frac{(m_1+1)(p-1)+n}{p} \rfloor, n > \frac{m_i+1}{p^2+p+1} \end{cases} \tag{A}$$

(ii) *If $\frac{r_2}{p^2} \leq \frac{e}{p^2+p+1} \leq \frac{r_1}{p^2+p}$ then the computation of the $\mu_i(n)$ makes by the formula*

$$\mu_i(n) = \begin{cases} p^2 n, n \leq \frac{r_2 p^{i-1}}{p(p-1)} \\ r_2 p^{i-1} + pn, \frac{r_2 p^{i-1}}{p(p-1)} < n < \left[\frac{(m_i+1)(p-1)+p^i r_2}{p^2-1}\right] \\ \lfloor \frac{(m_1+1)(p-1)+n}{p} \rfloor, n > \left[\frac{(m_i+1)(p-1)+p^i r_2}{p^2-1}\right] \end{cases} \tag{B}$$

(iii) *If $\frac{r_1}{p^2+p} \leq \frac{r_2}{p^2} \leq \frac{e}{p^2+p+1}$ then the computation of the $\mu_i(n)$ makes by the formula*

$$\mu_i(n) = \begin{cases} p^2 n, n \leq \frac{r_2 p^{i-1}}{(p^2-1)} \\ r_1 p^{i-1} + n, \frac{r_1 p^{i-1}}{(p^2-1)} < n < \left[\frac{(m_i+1)(p-1)+p^i r_1}{p-1}\right] \\ \lfloor \frac{(m_1+1)(p-1)+n}{p} \rfloor, n > \left[\frac{(m_i+1)(p-1)+p^i r_1}{p-1}\right] \end{cases} \tag{C}$$

(iv) *If $\frac{r_2}{p^2} \leq \frac{r_1}{p^2+p} \leq \frac{e}{p^2+p+1}$ then the computation of the $\mu_i(n)$ makes by the formula*

$$\mu_i(n) = \begin{cases} p^2 n, n \le \frac{r_2 p^{i-1}}{p(p-1)} \\ r_2 p^{i-1} + pn, \frac{r_2 p^{i-1}}{p(p-1)} < n \le \frac{(r_1 - r_2)p^{i-1}}{p-1} \\ r_1 p^{i-1} + n, \frac{(r_1-r_2)p^{i-1}}{p-1} < n \le \left[\frac{(m_i+1)(p-1)+p^i r_1}{p-1}\right] \\ \lfloor \frac{(m_1+1)(p-1)+n}{p} \rfloor, n > \left[\frac{(m_i+1)(p-1)+p^i r_1}{p-1}\right] \end{cases} \quad \text{(D)}$$

The lemma is proved by direct computation.

Let K be a local or quasi-local field and $F(x, y)$ be the one dimensional formal group over integer ring of K. Let $F_L = F(M_L)$ be the G-module that is defined by the group low $F(x, y)$ on the maximal ideal M_L of the ring O_L of any finite Galois extension L of the field K.

In the case when K is the quasi-local field it is possible, follow to Serre [15], induced on F_L the structure of the proalgebraic group. Denote the group as $\overline{F}_L$. Let $\pi_1(\overline{F}_L)$ be its fundamental group.

Definition 5.3 Let K be a local field, $N_{L/K} : F_L \to F_K$ the norm homomorphism. The subgroup

$$\mathscr{V}_K = \bigcap_K N_{L/K}(F_L)$$

(intersection on all finite Galois extensions L/K) of the group F_K is called the universal norm group of the group F defined over ring O_K.

If K is a quasi-local field, then the subgroup

$$\mathscr{V}_K^* = \bigcap_K N_{L/K}(\pi_1(\overline{F}_L))$$

(intersection on all finite Galois extensions L/K) of the group $\pi_1(\overline{F}_L)$ is called the universal norm group of the group F defined over ring O_K.

Theorem 5.2

$$\mathscr{V}_K (\text{respectively } \mathscr{V}_K^*) = 0.$$

Sketch of the proof We use an extension of the method of Vvedenskii [4] by which he prove the result for one dimensional fornal groups of reduction height 1 and 2 over integer ring of local and quasi-local fields.

If K is a local field, then the prove of the theorem reduced to the prove of the next lemma 5.4. If K is a quasi-local field, then we follow the method that has proposed in the paper [13]. In the case it is sufficient to prove that for any finite Galois extensions L/K the next equality and inclusion take place

$$N_{L/K}(\mathscr{V}_L^*) = \mathscr{V}_K^*$$

$$\mathscr{V}_L^* \subset p\pi_1(\overline{F}_L)$$

Lemma 5.4 *For any integer n, $n \geq 1$ there is such finite Galois extension L/K, that the image $N_{L/K}(F_L)$ (respectively $N_{L/K}(\pi_1(\overline{F}_L))$ of the norm homomorphism*

$$N_{L/K} : F_L \to F_K$$

(respectively $N_{L/K} : \pi_1(\overline{F}_L) \to \pi_1(\overline{F}_K)$) is contained in F_K^n (respectively in $\pi_1(\overline{F}_K)$).

References

1. Shafarevich, I.R.: Principal homogenious spaces over function fields (in Russian). Tr. Math. Steklov Inst. **64**, 316–346 (1961)
2. Tate, J.: Principal homogenious spaces for abelian varieties. J. Reine Angew. Math. **209**, 98–99 (1962)
3. Ogg, A.: Cohomology of abelian varieties over function fields. Ann. Math. **76**, 185–220 (1962)
4. Vvedenskii, O.N.: On local "class fields" of elliptic curves (in Russian). Izv. Akad. Nauk SSSR Math. USSR Izvestija Ser Mat. **37**(1), 20–88 (1973)
5. Vvedenskii, O.N.: On quasi-local "class fields" of elliptic curves I (in Russian). Izv. Akad. Nauk SSSR Math. USSR Izvestija Ser Mat. **40**(5), 969–992 (1976)
6. Vvedenskii, O.N.: Duality in elliptic curves over local fields I (in Russian). Izv. Akad. Nauk SSSR Math. USSR Izvestija Ser Mat. Tom **28**(4), 1091–1112 (1964)
7. Vvedenskii, O.N.: Duality in elliptic curves over local fields II (in Russian). Izv. Akad. Nauk SSSR Math. USSR Izvestija Ser Mat. Tom **30**(4), 891–922 (1966)
8. Glazunov, N.M.: On the "norm subgroups" of one-parameter formal groups over integer ring of local field (in Russian). Dokl. Akad. Nauk Ukr. SSR, Ser. A **11**, 965–968 (1973)
9. Lubin, J.: One parameter formal Lie groups over ρ-adic integer rings. Ann. Math. **80**(3), 464–484 (1964)
10. Glazunov, N.M.: Remarks on n-dimensional commutative formal groups over integer ring of ρ-adic field (in Russian). Ukr. Math. J. **25**(3), 352–355 (1973)
11. Hazewinkel, M.: Formal Groups. Academic Press, New York (1977)
12. Cassels, J., Fröhlich, A. (eds.): Algebraic Number Theory. Academic Press, London (2003)
13. Vvedenskii, O.N.: On "universal norms" of formal groups over integer ring of local fields (In Russian). Izv. Akad. Nauk SSSR Math. USSR Izvestija Ser Mat. Tom **37**, 737–751 (1973)
14. Glazunov, N.M.: Quasi-local class fields of elliptic curves and formal groups I. Proc. IAMM NANU **24**, 87–98 (2012)
15. Serre, J.-P. Géométrie algébrique. In: Proceedings of the International Congress of Mathematicians, pp.190–196. Institut Mittag-Leffler, Djursholm (1963)

Chapter 6
Assignment of Factors Levels for Design of Experiments with Resource Constraints

S. A. Smirnov, O. O. Glushchenko, K. A. Ilchuk, I. L. Makeenko and N. A. Oriekhova

Abstract An optimal procedure for factors levels assignment is proposed. The procedure is based on choice of levels number proportionally to factor significance, guaranteed estimation of entropy, and 1D-parametrization of iteration process for multidimensional mapping fixed point finding. Solution existence and convergence of the procedure is proved.

6.1 Introduction

One of the major planning stages, that predestinates an effectiveness of experiment, is the choice of the most significant parameters of the situation as a factors of experiment, and the appointment some discrete values as their levels. In this case, for the set of admissible values for each factor, it is reasonable to use a *partition* into disjoint subsets covering it completely, and as a factor *level* to choose one *representative* value for each partition element. Naturally the question arises as to effectively select the set partition and the corresponding representative values. Then it is necessary to

S. A. Smirnov (✉) · O. O. Glushchenko · K. A. Ilchuk · I. L. Makeenko
Institute of Physics and Technology, National Technical University of Ukraine,
Kyiv Polytechnic Institute, Peremogy ave., 37, Kyiv 03056, Ukraine
e-mail: smir@pti.kpi.ua

N. A. Oriekhova
V.Glushkov Institute of Cybernetics, National Academy of Sciences of Ukraine,
Glushkova ave., 40, Kyiv, Ukraine
e-mail: natalyaorekhova@yandex.ru

K. A. Ilchuk
e-mail: will14ka@gmail.com

O. O. Glushchenko
e-mail: strokalex@ukr.net

I. L. Makeenko
e-mail: irinamakeenko@mail.ru

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*,
Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_6,

take into account the resource constraints of the experiment, always present, and try to manage them well.

Thus, for a complex system undergoing experimental investigation, to have to build a simplified mathematical model, based on reduced descriptions. It uses not all existing independent variables, but only the most important, which is determined by natural resource constraints assigned for the solution. Those restrictions limit the number of values taken by each of the parameters. Formalization of such problems, with emphasis on the situation of decision-making can be found in monograph [1].

6.2 Hansel Method

Among the methods of formation of the simplified discrete model we single out a method of Hansel, proposed in work [2] and based on the idea of appointment of representative values of all parameters in proportion to their *significance*. Under the *significance* of the independent parameter means the product of its *relevance* to the *entropy*.

Parameters of an experiment, which values we can change, is called factors, and result of an experiment (numerical value) is called observable. *Relevance* (coefficient of impact) R_l is defined as the ratio of the range of observable values for different values of this parameter, to its average value (with nominal values of other parameters). It describes the influence, and usually normalized by all parameters, and that is a weighting coefficient of this parameter. To calculate *relevance* you need to know the dependence of observable from all factors. *Entropy* of a factor is calculated traditionally, as:

$$H_l\,(N_l) = -\sum_{j=1}^{N_l} P_j \ln(P_j),$$

where N_l—number of levels of l-th factor, and P_j—the probability of the j-th value of l-th factor, characterizes informational complexity of representation of corresponding parameter. Discretization of the continuous variables is realized by corresponding values of the factor levels.

Let us assumed that $N = \prod_{l=1}^{L} N_l$ is the number, defined from resource restrictions, that meets the full set of combinations of the levels of all factors. This number characterize the combinatorial size of the problem. Since the *significance* of parameter is the product of its *relevance* to its *entropy,* and the entropy depends on the same number

$$N_l = k R_l H_l\,(N_l),$$

where $1 \le l \le L$ is parameter index, and k—the proportionality coefficient. Then numbers N_l can be find as a fixed point of a multidimensional mapping $\Phi(N)$:

$$N = \Phi(N) \Longleftrightarrow \begin{cases} N_1 = k R_1 H_1\,(N_1) \\ \dots \\ N_L = k R_L H_L\,(N_L) \end{cases}$$

under condition $N = \prod_{l=1}^{L} N_l$.

Author of the method have considered the dependence $H_l\,(N_l)$ for the cases of three probability distributions: normal, Weibull and log-normal and for partitioning into intervals of equal length. For a normal distribution with equidistant partition entropy formula:

$$H_l\,(N_l) = \ln\left(2\sqrt{(2\pi e)}N_l/\sigma\right),$$

for others distributions entropy formulas look more complicated. Then author proposed to calculate a fixed point by sequential approximation method

Note some weaknesses of Hansel method:

1. Equidistant partition and is a very special case.
2. The choice of the initial approximation is arbitrary, there are no recommendations.
3. Convergence of iterations L-dimensional mapping of the specified type is not theoretically justified and in practice is rare.
4. Integer rounding adds complexity for multidimensional mapping.

Thus, the absolute advantage of the method is the construction of an iterative process of finding N_l, and disadvantage—its complexity and the lack of any guarantee of convergence.

6.3 Modification

Here we propose a modification of this method to get rid of its shortcomings. It is based on two considerations.

1. Using of equiprobable partition to select representative value (Fig. 6.1), that working for the probability distributions of any kind. This partition with guarantee has the property of maximum entropy, that is optimum on the information criteria.

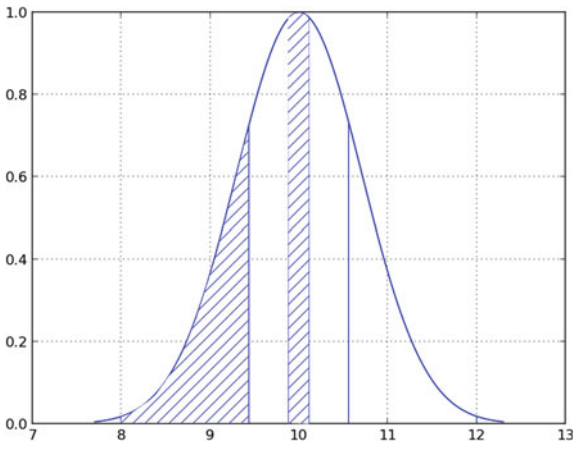


Fig. 6.1 Equiprobable 5-partition on (8; 12) interval for normal distribution

For such case there is an exact expression for the entropy $H_l\,(N_l) = \ln\,(N_l)$. Then we choose representative value from partition element as a level of factor. It is taken as the average value in the corresponding interval, or as the most probable value out of it.

2. The transition from L- to 1-parametric representation of the iterative process fixed point searching. For this we use the expression for the fixed point $N_l/\ln\,(N_l) = kR_l$, from which we conclude that for determination of all numbers N_l it is sufficient to find the coefficient of proportionality k—only one scalar parameter. It is the maximum decomposition, because a multidimensional mapping splits into direct product of independent 1-dimensional mappings.

How to organize an iterative process 1-dimensional search of a value, that correspond a fixed point of multidimensional mapping? We suggest the following procedure. The initial value k_0 is obtained from the obvious inequality for levels number of factor with minimal relevancy $N_1 \geq 2$. Then

$$k_0 = 3/(R_1 \ln 3) \approx 2.73/R_1$$

is the lower bound for the k finding value. From the other hand $N_1 \leq \sqrt[L]{N}$, and then

$$k_1 = L\sqrt[L]{N}/(R_1 \ln N)$$

is the upper bound, $k_0 \leq k \leq k_1$. We can calculate the zero approximation N_l from the equation $x/\ln\,(x) = k_0 R_L$. Note that the function $x/\ln(x)$ is unimodal (one minimum at e point), and the solution is *unique* for $x > e$ (see Fig. 6.2). Let $N^{(0)}$—the product of all zero approximation estimates N_l. It turns $N^{(0)} < N$ (in other case our problem is unsolvable), then calculate the first approximation from the equation $x/\ln\,(x) = k_1 R_L$. If $N^{(1)}$ is the product of all first approximation estimates N_l, it turns $N^{(1)} > N$. Then perform a one-dimensional search in the interval between k_0 and k_1. Second approximation we find from

$$k_2 = (k_0 + k_1)/2,$$

and refine sign of inequality between $N^{(2)}$ and N. For construction of a sequence k_n use a dichotomous iteration. Method of division in halves allows to guarantee a convergence of the iterative process.

In fact, due to the inevitability of the integer rounding, a condition

$$(N_{l+1} - N_l)^2 < 1$$

can serve as iteration stopping criteria, and final assessments of N_l meet the requirements of the fixed point only approximately. But if we have exact solution, we compare different integer rounding by discrepancy evaluation. So, the obtaining computing solution is optimal, due to its integer-valued nature it is the best possible.

Thus, the basic idea of the proposed procedure:

1. Best (on criterion of entropy) approach to the selection of representative values (factor levels), that is universal for all probability distributions;
2. One-dimensional parametrization of the iterative process for finding a fixed point of a multidimensional nonlinear mapping.

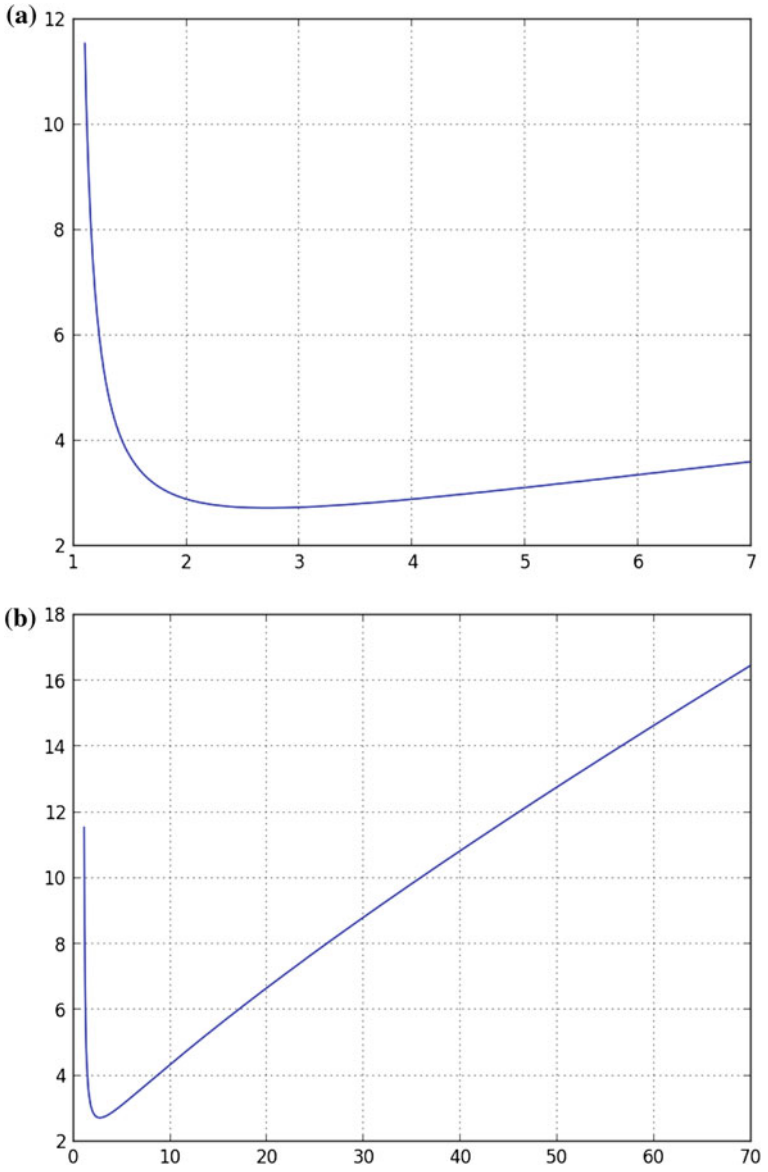


Fig. 6.2 Graph of the function $x/\ln(x)$ in **a** small and **b** large scale

6.4 Example

Let to consider some 3-factors experiment with relevancies $R_1 = 1/5$, $R_2 = 0.3$, $R_3 = 1/2$. Our calculations gives next numbers of factor levels (Table 6.1).

Table 6.1 3-factors experiment with relevancies $R_1 = 0.2$, $R_2 = 0.3$, $R_3 = 0.5$

N_1	2	3	4	5	6	7	8	9	10	11	12
N_2	10	9	10	11	13	14	16	18	19	21	23
N_3	22	21	22	25	28	31	34	37	40	43	46
N	440	567	880	1375	2184	3038	4352	5508	7600	9933	12696

Table 6.2 2-factor problem with relevancies 3/8 and 5/8

N_1	2	2	2	2	2	2	2	2
N_2	2	3	4	5	6	7	8	9
N_3	12	11	12	13	15	17	19	21
N	48	66	96	130	180	238	304	378

So, on the base of the approach of minimal relevancy we can choose the numbers, that are corresponds in a best way to combinatorial complexity restrictions of experiment.

Some words about fast experiments. For $N < 440$ an optimal planning of a level numbers is uncertain. A good way is to froze minimal value $N_1 = 2$, and to solve a 2-factor problem with relevancies 3/8 and 5/8. Here are the results of calculations (Table 6.2).

6.5 Conclusions

Thanks to the implementation of the proposed modification, the constructed procedure has guaranteed convergence, significantly greater ease of realization and flexibility of reconfiguration at the refinement of the problem statement. For design of experiment problems the proposed procedure can manage the available resources in the best way. Particular attention should be paid to the assessment of the relevance of various factors, since their determination procedures can not be completely formalized. It is reasonable to use expert interval estimates for relevancies of factors, based on the technique, developed in [3]. We also believe that it is useful to make greater employment of entropy criterion for constructing estimates based on subjective measures.

References

1. Muschick, E., Müller, P.H.: Entscheidungspraxis: Ziele, Verfahren, Konsequenzen. VEB Verlag Technik, Berlin (1987)
2. Hansel, V.: Ein allgemeines Entscheidungskozept zur Bearbeitung mehrzielorientierter Informationsmangel-probleme. Dissertation A, IH Zittau (1984)
3. Smirnov, S., Gontarenko, I.: Garantirovanyi sintez skalyarnogo kriteriya dlya resheniya zadachi mnogokriterial'noy optimizatsii. Syst. Res. Inf. Tech. **5**, 99–106 (2006)

Part II
Mechanics and Numerical Methods

Chapter 7
How to Formulate the Initial-Boundary-Value Problem of Elastodynamics in Terms of Stresses?

D. V. Georgievskii

Abstract In case when loadings are given on all the boundary of deformable solid, the initial-boundary-value problem for obtaining stress-strain state seems to be more suitable and effective if it is formulated and investigated in terms of stress tensor components. In this chapter typical peculiarities of some (in chronological order) formulations the initial-boundary-value problems in dynamic theory for isotropic linear elastic solid are discussed.

7.1 The Classic Formulation of the Dynamic Problem and its Peculiarities

As is generally known, the problem of elastodynamics for linear isotropic solid consists of investigation inside a solid domain V of the system

$$\mathbf{S}(\boldsymbol{\sigma}, \mathbf{u}) \equiv \mathrm{Div}\,\boldsymbol{\sigma} + \rho\mathbf{F} - \rho\mathbf{u}_{,tt} = 0 \tag{7.1}$$

$$\boldsymbol{\sigma} = \lambda\theta\mathbf{I} + 2\mu\boldsymbol{\varepsilon}\,, \quad \theta = \mathrm{tr}\,\boldsymbol{\varepsilon}\,, \quad -1 < \nu < \frac{1}{2} \tag{7.2}$$

$$\boldsymbol{\varepsilon} = \mathrm{Def}\,\mathbf{u} \equiv \frac{1}{2}\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^T\right) \tag{7.3}$$

by fulfilment of initial conditions

$$\mathbf{u}(0, \mathbf{x}) = \mathbf{U}(\mathbf{x}), \quad \mathbf{u}_{,t}(0, \mathbf{x}) = \mathbf{V}(\mathbf{x}) \tag{7.4}$$

D. V. Georgievskii (✉)
Mechanical and Mathematical Faculty, Moscow State University,
Vorobjovy Gory, Moscow, Russia 119991
e-mail: georgiev@mech.math.msu.su

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*,
Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_7,

Let the following loadings

$$(\underset{\sim}{\sigma} \cdot \mathbf{n})_{\Sigma} = \mathbf{P}^{\circ} \tag{7.5}$$

are given on all the boundary $\Sigma = \partial V$. Here λ, μ are the Lame constants, ν is the Poisson ratio, $\mathbf{F}$ is mass force (mass acceleration), $\underset{\sim}{\mathbf{I}}$ is two-rank unit tensor, ρ is mass density, $\mathbf{n}$ is unit normal in each point of the boundary Σ.

Cauchy stress tensor $\underset{\sim}{\sigma}(0, \mathbf{x})$, low strain tensor $\underset{\sim}{\varepsilon}(0, \mathbf{x})$, and displacement vector $\mathbf{u}(0, \mathbf{x})$ must be obtained from the system (7.1)–(7.5). A presence of the Cauchy relations (7.3) guarantees an identical fulfilment of well-known six Saint-Venant compatibility equations

$$\underset{\sim}{\eta}(\underset{\sim}{\varepsilon}) \equiv \operatorname{Ink} \underset{\sim}{\varepsilon} = 0 \tag{7.6}$$

where $\underset{\sim}{\eta}$ is the Kröner unit incompatibility tensor. The Eq. (7.6) may also be written in Cartesian coordinates in one of two following forms

$$\eta_{il} = \varepsilon_{ijk}\varepsilon_{lmn}\varepsilon_{km,jn} = 0 \tag{7.7}$$

$$\eta_{\alpha\alpha} \equiv 2\varepsilon_{\beta\gamma,\beta\gamma} - \varepsilon_{\beta\beta,\gamma\gamma} - \varepsilon_{\gamma\gamma,\beta\beta} = 0$$

$$\eta_{\alpha\beta} \equiv \varepsilon_{\alpha\beta,\gamma\gamma} + \varepsilon_{\gamma\gamma,\alpha\beta} - \varepsilon_{\alpha\gamma,\gamma\beta} - \varepsilon_{\beta\gamma,\gamma\alpha} = 0 \tag{7.8}$$

where ε_{ijk} are the Levi-Civita symbols; $(\alpha, \beta, \gamma) = \{(1, 2, 3); (2, 3, 1); (3, 1, 2)\}$. We remind that the summation from 1 to 3 is realized by recurring twice (in each monomial) Latin subscripts. There is no summation by Greek subscripts.

From the point of view of computational mechanics and, generally, mechanics of deformable solid, in case when the conditions (7.5) are given on all the surface Σ it is very convenient to formulate the initial-boundary-value problem in terms of stress tensor [1–5]. This is explained by the fact that namely stress components (but not displacement ones) are the main values being of interest in most practical and engineering applications. These components are contained in various tests and criteria of strength, fracture, and phase transfers.

The problem (7.1)–(7.5) of isotropic elastodynamics in terms of stresses may be formulated in the following classic way:

$$\mathbf{S}\big[\underset{\sim}{\sigma}, \mathbf{u}\big(\underset{\sim}{\varepsilon}(\underset{\sim}{\sigma})\big)\big] \equiv \operatorname{Div} \underset{\sim}{\sigma} + \rho\mathbf{F} - \rho\mathbf{u}\big(\underset{\sim}{\varepsilon}(\underset{\sim}{\sigma})\big)_{,tt} = 0 \tag{7.9}$$

$$\underset{\sim}{\eta}\big(\underset{\sim}{\varepsilon}(\underset{\sim}{\sigma})\big) = 0 \tag{7.10}$$

$$\underset{\sim}{\sigma}(0, \mathbf{x}) = \lambda \operatorname{div} \mathbf{U}\, \underset{\sim}{\mathbf{I}} + 2\mu \operatorname{Def} \mathbf{U}, \quad \underset{\sim}{\sigma}_{,t}(0, \mathbf{x}) = \lambda \operatorname{div} \mathbf{V}\, \underset{\sim}{\mathbf{I}} + 2\mu \operatorname{Def} \mathbf{V} \tag{7.11}$$

The boundary conditions (7.5) should be added here.

We keep in mind that the inverse to (7.2) law

$$\underset{\sim}{\varepsilon} = \frac{1}{E}\big(-3\nu\sigma\underset{\sim}{\mathbf{I}} + (1+\nu)\underset{\sim}{\sigma}\big), \quad 3\sigma = \operatorname{tr} \underset{\sim}{\sigma} \tag{7.12}$$

as well as the Cesaro formulae expressing displacements in terms of strains

$$\mathbf{u} = \mathbf{u}^\circ + (\mathbf{x} - \mathbf{x}^\circ) \cdot \mathcal{Q}^\circ + \int\limits_{M^\circ}^{M} \left[\underline{\varepsilon} + (\mathbf{x} - \mathbf{x}') \cdot \nabla\underline{\varepsilon} - \nabla\underline{\varepsilon} \cdot (\mathbf{x} - \mathbf{x}')\right] \cdot d\mathbf{x}' \tag{7.13}$$

$$\mathcal{Q} = \frac{1}{2}\left(\nabla\mathbf{u} - (\nabla\mathbf{u})^T\right) \tag{7.14}$$

are used in the Eqs. (7.9) and (7.10). Here E is Young modulus, $\mathcal{Q}$ is antisymmetric rotation tensor, M° is some fixed point of solid with coordinates $\mathbf{x}^\circ$ (both displacements $\mathbf{u}^\circ$ and rotations $\mathcal{Q}^\circ$ are known in M°), M is arbitrary moving point with coordinates $\mathbf{x}$.

A presence of the contour integral in (7.13) turns the equations of motion (7.9) into integro-differential ones with respect to $\underline{\sigma}$. This considerably complicates an analytical investigation and makes the classic formulation as a whole inefficient for application of computational methods. Note that this complexity is missing in the corresponding static (quasistatic) problem. In this case one can derive the Beltrami–Michel compatibility equations

$$\Delta\underline{\sigma} + \frac{3}{1+\nu}\nabla\nabla\sigma = -\frac{\rho\nu}{1-\nu}\,\mathrm{div}\,\mathbf{F}\,\underline{\mathbf{I}} - 2\rho\mathrm{Def}\,\mathbf{F} \tag{7.15}$$

using conditions (7.10) after substitution there combinations $\mathrm{Def}\,(\mathrm{Div}\,\underline{\sigma})$.

Equation (7.15) may be derived [3] not resorting to the Saint-Venant identities (7.10) but using the Lame equations and the Hooke law in inverse form (7.12).

7.2 Ignaczak–Nowacki' Formulation

The generalized Beltrami–Michel compatibility equations in the problem of elatodynamics may be obtained by means of any of two methods mentioned above. In fact, applying operator Def to left and right hands of the motion Eq. (7.9) we derive at first

$$(\mathrm{Def}\,\mathbf{S})(\underline{\sigma}) = 0 \tag{7.16}$$

or in detail

$$2\mathrm{Def}\,(\mathrm{Div}\,\underline{\sigma} + \rho\mathbf{F}) = -\frac{3\nu}{(1+\nu)c_2^2}\,\sigma_{,tt}\,\underline{\mathbf{I}} + \frac{1}{c_2^2}\,\underline{\sigma}_{,tt}\left(\equiv 2\rho\underline{\varepsilon}(\underline{\sigma})_{,tt}\right) \tag{7.17}$$

where $c_2 = \sqrt{\mu/\rho}$ is a shear wave velocity in elastic medium. Then on the basis of (7.17) we can write the generalized Beltrami–Michel compatibility Equations

$$\Big(\Delta - \frac{1}{c_2^2}\frac{\partial^2}{\partial t^2}\Big)\boldsymbol{\sigma} + \frac{3}{1+\nu}\Big(\nabla\nabla\sigma + \frac{\nu}{2(1-\nu)c_2^2}\,\sigma_{,tt}\underset{\sim}{\mathbf{I}}\Big) =$$
$$= -\frac{\rho\nu}{1-\nu}\,\mathrm{div}\,\mathbf{F}\,\underset{\sim}{\mathbf{I}} - 2\rho\,\mathrm{Def}\,\mathbf{F} \tag{7.18}$$

For the first time the Eqs. (7.17) and (7.18) are led out in [1, 6]. Significant attention in monograph [4] is devoted to these subjects too. In this way the Ignaczak–Nowacki' formulation consists of six generalized Beltrami–Michel compatibility equations inside domain V as well as three Eq. (7.9) by satisfaction the boundary conditions (7.5) and the initial conditions (7.11). In [6] an uniqueness theorem for the problem in terms of stresses is proved without using of some kind of energetic concepts.

7.3 Konovalov' Formulation

Both formulations of the dynamic problem which are discussed above presuppose a solving in V nine equations with respect to six components of symmetric tensor $\boldsymbol{\sigma}$ by fulfilment of only three boundary conditions on Σ. Furthermore, it is necessary to express displacements $\mathbf{u}$ in terms of $\boldsymbol{\sigma}$ by means of contour integrals (7.13).

It should be noted that initial stresses $\boldsymbol{\sigma}(0,\mathbf{x})$, $\boldsymbol{\sigma}_{,t}(0,\mathbf{x})$ (7.11) are taken in such a way that they are compatible certainly, i. e. initial strains $\boldsymbol{\varepsilon}(0,\mathbf{x})$, $\boldsymbol{\varepsilon}_{,t}(0,\mathbf{x})$ which are obtained using (7.12) turn the Eq. (7.6) into identities. The following tensor relation

$$\rho\boldsymbol{\varepsilon}(\boldsymbol{\sigma}) = \int_0^t \int_0^{t'} \mathrm{Def}\,(\mathrm{Div}\,\boldsymbol{\sigma} + \rho\mathbf{F})(t'',\mathbf{x})\,dt''\,dt' + \rho\boldsymbol{\varepsilon}_{,t}(0,\mathbf{x})\,t + \rho\boldsymbol{\varepsilon}(0,\mathbf{x}) \tag{7.19}$$

turns out after double integration (7.17) by time.

All three terms in right hand of (7.19) comply with the Eqs. (7.6) or (7.10). Therefore, the stress field satisfying with both the Eq. (7.17) inside V and boundary conditions (7.5) on Σ and initial conditions (7.11) is also adjusted with the Eq. (7.10).

The stated formulation of the dynamic problem in terms of stresses as well as a development of corresponding computational methods were realized in the sixties and seventies of twentieth century in works of Konovalov (see his monograph [3]).

The Eq. (7.17) do not yet ensure an identical fulfilment of the motion Eq. (7.9) inside V as the classic formulation demands. For the following consequent statement

$$\mathrm{Def}\,\mathbf{S} = 0 \quad \Longrightarrow \quad \mathbf{S} = 0, \quad \mathbf{x} \in V, \quad t > 0 \tag{7.20}$$

it is necessary and sufficiently to require two additional conditions in any point $\xi \in V$:

$$\mathbf{S} = 0, \quad \nabla\mathbf{S} - (\nabla\mathbf{S})^T = 0, \quad \mathbf{x} = \xi, \quad t > 0 \tag{7.21}$$

where the second condition is equivalent to rot $\mathbf{S} = 0$.

Let us choose in the capacity of point $\boldsymbol{\xi}$ some pole M° with coordinates $\mathbf{x}^\circ$ where both displacements $\mathbf{u}^\circ$ and rotations $\mathcal{Q}^\circ$ are known by $t > 0$ (see (7.13)). Then two additional conditions (7.21) may be written in the following way

$$(\mathrm{Div}\,\boldsymbol{\sigma} + \rho\mathbf{F})(t, \mathbf{x}^\circ) = \rho\mathbf{u}^\circ_{,tt} \tag{7.22}$$

$$\frac{1}{2}\left\{\nabla\left[\mathrm{Div}\,\boldsymbol{\sigma} + \rho\mathbf{F}\right] - \left(\nabla\left[\mathrm{Div}\,\boldsymbol{\sigma} + \rho\mathbf{F}\right]\right)^T\right\}(t, \mathbf{x}^\circ) = \rho\mathcal{Q}^\circ_{,tt} \tag{7.23}$$

The conditions (7.22) and (7.23) also are part of the Konovalov' formulation in which it is not necessary to express unknown inside V displacements in terms of stresses $\boldsymbol{\sigma}$.

7.4 Pobedria' Formulation

In the late seventies of twenteeth century Pobedria proposed [7] a new formulation the problem in terms of stresses in mechanics of solids which is better adjusted to applications of computational algorithms. The classic variational Castigliano' principle is employed bad for a construction of difference schemes by one or another level so long so the question is about conditional extremum of Castiglianian. So the new variational principle had been stated.

The Pobedria' formulation [5] consists of solution inside elastic solid six generalized compatibility equations in terms of stresses by satisfying on all the boundary three motion (equilibrium) equations as well as three boundary conditions. It should be given also initial conditions in case of the problem of elastodynamics.

For the last thirty years this formulation acquires a widespread world fame (see, for example, the papers [8–10] that are devoted to its development). With the Pobedria' formulation help many 2D and 3D quasistaic boundary-value problems in elasticity, plasticity, viscoelasticity, contact problems, heat transfer problems, tasks of computational mechanics of composites have been analyzed by numerical and analytical methods.

In conformity to the problem of isotropic elastodynamics, this formulation requires a solving six generalized compatibility Beltrami–Michel Eq. (7.18) inside the domain V by satisfying in each point of the boundary Σ three Eq. (7.9) as well as three conditions (7.5) and initial conditions (7.11).

7.5 One More Possible Formulation

We require a fulfilment inside V of six generalized compatibility Beltrami–Michel Eq. (7.18) by satisfying on Σ six conditions (7.17) and three boundary conditions (7.5) as well as the additional equalities (7.22) and (7.23) in some point $\mathbf{x}^\circ \in \bar{V}$. Let

us show that such formulation of the dynamic problem is equivalent, for example, to the Ignaczak–Nowacki' formulation in terms of stresses.

In fact, the Ignaczak–Nowacki' formulation involving the motion Eq. (7.9) inside V guarantees a validity of (7.17) inside V and on Σ by virtue of continuity. Furthermore, the equalities (7.22) and (7.23) are realized in some point $\mathbf{x}^\circ \in \bar{V}$ where $\mathbf{u}^\circ$ and $\underset{\sim}{\mathcal{Q}}^\circ$ are known.

A method of proof to another hand consists of two standard stages.

1. We make a full contraction of the generalized compatibility Beltrami–Michel Eq. (7.18) with 2-rank identity tensor $\underset{\sim}{I}$ and express $\Delta\sigma$:

$$\Delta\sigma = -\frac{(1+\nu)\rho}{3(1-\nu)}\operatorname{div}\mathbf{F} + \frac{1-2\nu}{2(1-\nu)c_2^2}\,\sigma_{,tt} \tag{7.24}$$

   It is naturally that by $\nu < 1/2$ a mean stress σ turns out to satisfy with not uniform wave equation with wave spreading velocity

$$\sqrt{\frac{2(1-\nu)}{1-2\nu}}\,c_2 = c_1 \tag{7.25}$$

2. We apply operator Div to both hands of (7.18) and take into account the equality (7.24):

$$\Delta\operatorname{Div}\underset{\sim}{\mathcal{Q}} - \frac{1}{c_2^2}\Big[\operatorname{Div}\underset{\sim}{\mathcal{Q}} - \frac{3}{2(1+\nu)}\nabla\sigma\Big]_{,tt} = -\rho\Delta\mathbf{F} \tag{7.26}$$

An expression in square brackets is equal to $\mu\Delta\mathbf{u}$ (this may be verified by substitution $\underset{\sim}{\mathcal{Q}} = \lambda\operatorname{div}\mathbf{u}\,\underset{\sim}{\mathbf{I}} + 2\mu\operatorname{Def}\mathbf{u}$ consistent with the initial conditions (7.11)).

Thus, vector $\mathbf{S}(\underset{\sim}{\mathcal{Q}})$ as well as tensor $\operatorname{Def}\mathbf{S}(\underset{\sim}{\mathcal{Q}})$ are harmonic inside V. As the conditions (7.17) are realized on all the boundary Σ, i. e. $[\operatorname{Def}\mathbf{S}(\underset{\sim}{\mathcal{Q}})]_\Sigma = 0$, then the equalities (7.16) are correct in any point of the solid V. In addition, if the equalities (7.22) and (7.23) are realized then $[\mathbf{S}(\underset{\sim}{\mathcal{Q}})]_V = 0$, i. e. the motion Eq. (7.9) are correct. A required equivalence have been shown.

A proof withstands a passage to the limit to incompressible material. When $\nu = 1/2$ it is necessary to write (instead of (7.24)) Poisson' equation for mean stress: $\Delta\sigma = -\rho\operatorname{div}\mathbf{F}$ and instead of (7.26):

$$\Delta\operatorname{Div}\underset{\sim}{\mathcal{Q}} - \frac{1}{c_2^2}\big[\operatorname{Div}\underset{\sim}{\mathcal{Q}} - \nabla\sigma\big]_{,tt} = -\rho\Delta\mathbf{F} \tag{7.27}$$

The terms in square brackets in (7.27) as before are equal to $\mu\Delta\mathbf{u}$ what now may be verified in result of substitution $\underset{\sim}{\mathcal{Q}} = -p\,\underset{\sim}{\mathbf{I}} + 2\mu\operatorname{Def}\mathbf{u}$ taking account of $\operatorname{div}\mathbf{u} = 0$. Here

$$p = -\lim_{\nu\to 1/2}(\lambda\operatorname{div}\mathbf{u}) = -\sigma \tag{7.28}$$

is hydrostatic pressure in incompressible elastic solid.

Both in the Konovalov' formulation and in this one we do not require in any point an expression of unknown displacements $\mathbf{u}$ in terms of stresses $\boldsymbol{\sigma}$.

References

1. Ignaczak, J.: Direct determination of stresses from the stress equations on motion in elasticity. Arch. Mech. Stos. **11**, 671–678 (1959)
2. Ilyushin, A.A.: Plasticity. Foundations of General Mathematical Theory [in Russian]. USSR Acad. Sci., Moscow (1963).
3. Konovalov, A.N.: Solving the Problems of Elasticity in Terms of Stresses [in Russian]. Novosibirsk State University, Novosibirsk (1979)
4. Nowacki, W.: Teoria Sprężystości. PAN, Warszawa (1970)
5. Pobedria, B.E.: Numerical Methods in Elasticity and Plasticity [in Russian]. Moscow State University, Moscow (1995)
6. Ignaczak, J.: A completeness problem for stress equations of motion. Arch. Mech. Stos. **15**, 225–234 (1963)
7. Pobedria, B.E.: On the problem in terms of stresses [in Russian]. Doklady USSR Acad. Sci. **240**, 564–567 (1978)
8. Kucher, V.A., Markenscoff, X.: Stress formulation in 3D elasticity and applications to spherically uniform anisotropic solids. Int. J. Solids Struct. **42**(11–12), 295–297 (2005)
9. Kucher, V.A., Markenscoff, X., Paukshto, M.V.: Some properties of the boundary value problem of linear elasticity in terms of stresses. J. Elasticity **74**, 135–145 (2004)
10. Shaofan, L., Gupta, A., Markenscoff, X.: Conservation laws of linear elasticity in stress formulations. Proc. Roy. Soc. London. A **461**(2053), 99–116 (2005)

Chapter 8
Finite-Difference Method of Solution of the Shallow Water Equations on an Unstructured Mesh

G. M. Kobelkov and A. V. Drutsa

Abstract In the chapter we consider a linearized system of shallow water equations. Since this problem should be solved in domains being seas and oceans (or their parts), then solving this problem should use unstructured meshes to approximate domains under consideration properly. This problem was studied in the papers [1–4]. Here we consider finite-difference approximation of these equations, prove convergence of approximate solution to the differential one, and provide a number of numerical experiments confirming theoretical results. We also carried out some numerical experiments for real geographic objects.

8.1 Introduction

In the chapter we consider a linearized system of shallow water equations. Since this problem should be solved in domains being seas and oceans (or their parts), then solving this problem should use unstructured meshes to approximate domains under consideration properly. This problem was studed in the papers [1–4]. Here we consider finite-difference approximation of these equations, prove convergence of approximate solution to the differential one, and provide a number of numerical experiments confirming theoretical results. We also carried out some numerical experiments for real geographic objects.

8.2 Formulation of the Problem

Let us consider the system of shallow water equations in $2D$ Cartesian coordinates (see, e.g., [4–6]):

$$\mathbf{u}_t = g\nabla\zeta - R\mathbf{u} - \lambda\bar{k} \times \mathbf{u} + \mathbf{f}, \tag{8.1}$$

G. M. Kobelkov (✉) · A. V. Drutsa
Faculty of Mechanics and Mathematics, Lomonosov Moscow State University,
GSP-1, Leninskie Gory, Moscow, Russian Federation 119991
e-mail: kobelkov@dodo.inm.ras.ru

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*,
Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_8,

$$\zeta_t = \mathrm{div} H\mathbf{u}; \tag{8.2}$$

here $\mathbf{u} = (u, v)$ is a velocity vector, ζ—height of a tidal wave, $\bar{k}$—a unite vector in Oz direction, R, λ, g—some constants, $H(x, y)$ is a depth—a function of coordinates. These equations are considered in a bounded domain Ω with the boundary $\Gamma_1 \bigcup \Gamma_2$. Boundary conditions are of the form of impermeability conditions or fixed wave height:

$$\mathbf{u} \cdot n \equiv un_1 + vn_2 = 0, \quad \text{on } \Gamma_1; \tag{8.3}$$

$$\zeta = 0, \quad \text{on } \Gamma_1; \tag{8.4}$$

Initial conditions are

$$\begin{aligned} \mathbf{u}(x, y, 0) &= \mathbf{u}_0(x, y), \\ \zeta(x, y, 0) &= \zeta_0(x, y), \end{aligned} \tag{8.5}$$

where $\zeta_0(x, y)$ is some initial distribution of flow level while $\mathbf{u}_0(x, y)$ is some velocity vector field. Our aim is to approximate problem (8.1)–(8.5).

8.3 Mesh and Mesh Operators

Triangulate Ω in such a way that Ω^h (triangulation of the original domain) contains acute triangles only.

A boundary of Ω^h is denoted as $\Gamma^h = \Gamma_1^h \bigcup \Gamma_1^h$. A mesh is constructed in the following way: a cell center is a center of circumference described round a triangle, intersection point of a segment connecting centers of two neighboring cells and their common side is called flow node. Since all the triangles are acute, then centers of cells are inside of each triangle, while a flow node is a middle of the side where it locates. Enumerate cell centers $k = 1, 2, \ldots, K$ and flow nodes $i = 1, \ldots N$. Denote a center of the k-th cell by O^k and the i-th flow node—by X_i. By double upper index k, α we denote various mesh elements associated with the flow node X_i laying on α-th side of the k-th cell (i.e., $X_i = X^{k,\alpha}$). By double lower index i, m we denote elements associated with the k-th cell containing the flow node X_i (i.e., $O_{i,1} = O^k$). A node O^k and basic elements of a cell (side length—$l^{k,\alpha}$, length of the segment connecting centers—$d^{k,\alpha}$, square of the cell—S^k) are illustrated in Fig. 8.1. A flow node X_i and elements of the mesh associated with it (length of i-th side l_i, square of the quadrangle S_i) are illustrated in Fig. 8.2. By $\mathbf{O}$ we denote variety of triangle centers and by $\mathbf{X}$—variety of flow nodes.

Introduce the scalar products as

$$(f, g) = \sum_{k=1}^{K} S^k f^k \cdot g^k, \quad (\zeta, \xi) = \sum_{i=1}^{N} S_i \zeta_i \cdot \xi_i,$$

where functions f, g are defined at cell centers, while ζ, ξ are defined at flow nodes.

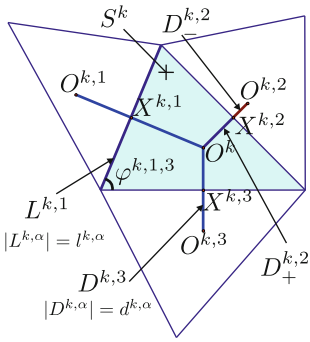


Fig. 8.1 A node O^k and three neighbor nodes $O^{k,\alpha}$, $\alpha = 1, 2, 3$

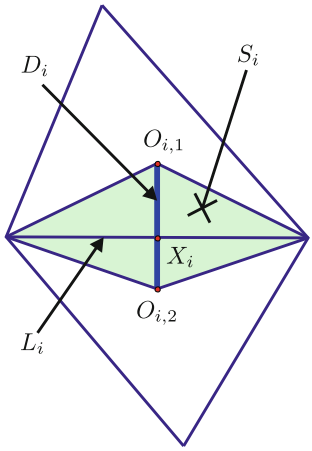


Fig. 8.2 Flow node X_i and two neighbor cells with the centers $O_{i,1}$ and $O_{i,2}$

To approximate the gradient and divergence operators we use the Fryazinov method [7]. The finite-dimensional gradient operator of ζ is defined at the cell center while ζ is defined at flow nodes. Define it by the formula

$$\nabla^h \zeta \Big|_{O^k} = \frac{1}{S^k} \sum_{\alpha=1}^{3} l^{k,\alpha} \zeta^{k,\alpha} \mathbf{n}^{k,\alpha}, \quad k = \overline{1, K}, \tag{8.6}$$

where $n^{k,\alpha}$ is an outer normal to the side $l^{k,\alpha}$ of the triangle O^k, $\zeta^{k,\alpha}$ is a value of the function ζ at $X^{k,\alpha}$. Components of the normal $n^{k,\alpha}$ are denoted by $n_x^{k,\alpha}$ and $n_y^{k,\alpha}$.

Let a function $\mathbf{u}$ be defined on cell centers. Approximate divergence operator of this function in inner flow nodes by the formula

$$\operatorname{div}^h \mathbf{u} \Big|_{X_i} = \frac{l_i}{S_i} \left(-\mathbf{u}_{i,1} \mathbf{n}_{i,1} - \mathbf{u}_{i,2} \mathbf{n}_{i,2} \right), \tag{8.7}$$

where $\mathbf{u}_{i,m}$ is a value of the function $\mathbf{u}$ at $O_{i,m}$. For the nodes laying on the part of boundary where the impermeability condition holds, the approximation of divergence is of the form

$$\operatorname{div}^h \mathbf{u}\Big|_{X_i} = \frac{l_i}{S_i}\left(-\mathbf{u}_{i,1}\mathbf{n}_{i,1}\right), \tag{8.8}$$

In [7] it was shown that the operators introduced above are conjugate (up to a sign), namely, $(\mathbf{u}_h, \nabla^h \zeta_h) = -(\operatorname{div}^h \mathbf{u}_h, \zeta_h)$, and the order of approximation is $O(h)$; hereafter by h we denote the maximal size of cells.

8.4 Finite-Dimensional Problem

For problem (8.1)–(8.5), introduce discretization in time with the time step τ, $t_j = j\tau$, $j = 0, 1, 2, \ldots$. Velocities and wave height on the top layer we shall denote by the same letters but with hat on them: $\hat{\mathbf{u}}$, $\hat{\zeta}$. Velocities are defined on cell centers and wave height—on flow nodes. Write down completely implicit finite-dimensional approximation for (8.1)–(8.5) using (8.6), (8.7). We have

$$\frac{\hat{\mathbf{u}}^k - \mathbf{u}^k}{\tau} = \lambda\overline{k} \times \hat{\mathbf{u}}^k - R\hat{\mathbf{u}}^k + \mathbf{f}|_{O^k} + g\nabla^h\hat{\zeta}, \quad k = \overline{1, K}, \tag{8.9}$$

where $\overline{k}$ is a unit vector in Oz direction, $\mathbf{f} = (f_1, f_2)$. Approximation of the continuity Eq. (8.2) in inner nodes is of the form

$$\frac{\hat{\zeta}_i - \zeta_i}{\tau} = \operatorname{div}^h \hat{\mathbf{u}}, \quad i = \overline{1, N} \quad X_i \in \Omega^h \backslash \Gamma^h. \tag{8.10}$$

In boundary nodes $X_i \in \Gamma_1^h$ the continuity equation is modified in the following way:

$$\frac{\hat{\zeta}_i - \zeta_i}{\tau} = \operatorname{div}^h \hat{\mathbf{u}} \equiv \frac{l_i}{S_i}\left(-\hat{\mathbf{u}}_{i,1}\mathbf{n}_{i,1}\right), \quad i = \overline{1, N} \quad X_i \in \Gamma_1^h. \tag{8.11}$$

For the rest nodes $X_i \in \Gamma_2^h$, the equations look as boundary conditions (8.4):

$$\hat{\zeta}_i = 0, \quad i = \overline{1, N} \quad X_i \in \Gamma_2^h. \tag{8.12}$$

Eqs. (8.9)–(8.12) approximate problem (8.1)–(8.2) and boundary conditions (8.3)–(8.4). In [2] it was proved that the system (8.9)–(8.12) is a system with symmetric positive definite M-matrix under the condition that all mesh triangles are acute ones.

8.5 Convergence

Theorem 8.1 *Let a solution to* (8.1)–(8.5) *be smooth enough and angles of all cells be less than* $\pi/2-\mu$ *for some* $\mu > 0$. *Then a solution to system* (8.9)–(8.12) *converges to a solution to the original problem as* $O(\tau + h)$ *for* $\tau, h \to 0$.

Proof Since approximation of the operators is of order $O(h)$ (see [7]), then the following relations hold

$$\frac{\partial \mathbf{u}}{\partial t}\Big|_{O^k} = \frac{\mathbf{u}(t+\tau)-\mathbf{u}(t)}{\tau}\Big|_{O^k} - \overline{\varepsilon}\Big|_{O^k}, \quad \frac{\partial \zeta}{\partial t}\Big|_{X_i} = \frac{\zeta(t+\tau)-\zeta(t)}{\tau}\Big|_{X_i} - \underline{\varepsilon}\Big|_{X_i},$$
$$\mathrm{div}\mathbf{u}\Big|_{X_i} = \mathrm{div}^h\mathbf{u}\Big|_{X_i} + \underline{\upsilon}\Big|_{X_i}, \quad \nabla\zeta\Big|_{O^k} = \nabla^h\zeta\Big|_{O^k} + \overline{\upsilon}\Big|_{O^k}. \tag{8.13}$$

For the functions $\overline{\varepsilon}$, $\underline{\varepsilon}$, $\overline{\upsilon}$, $\underline{\upsilon}$ the following relation takes place

$$\overline{\varepsilon}\big|_{O^k} = O(\tau), \quad \underline{\varepsilon}\big|_{X_i} = O(\tau), \quad \overline{\upsilon}\big|_{O^k} = O(h^k), \quad \underline{\upsilon}\big|_{X_i} = O(h_i), \tag{8.14}$$

where h^k is a diameter of k-th cell and h_i is a diameter of the quadrangle with the diagonals L_i and D_i. Then

$$\|\underline{\varepsilon}\| = \sqrt{\sum_{i=1}^{N} S_i \left(\underline{\varepsilon}\big|_{X_i}\right)^2} \le C_1 \mathrm{mes}^{1/2}(\Omega)\tau,$$
$$\|\overline{\varepsilon}\| = \sqrt{\sum_{k=1}^{K} S^k\, \overline{\varepsilon}\big|_{O^k} \cdot \overline{\varepsilon}\big|_{O^k}} \le C_2 \mathrm{mes}^{1/2}(\Omega)\tau, \tag{8.15}$$

$$\|\underline{\upsilon}\| = \sqrt{\sum_{i=1}^{N} S_i \left(\underline{\upsilon}\big|_{X_i}\right)^2} \le C_3 \mathrm{mes}^{1/2}(\Omega)h,$$
$$\|\overline{\upsilon}\| = \sqrt{\sum_{k=1}^{K} S^k\, \overline{\upsilon}\big|_{O^k} \cdot \overline{\upsilon}\big|_{O^k}} \le C_4 \mathrm{mes}^{1/2}(\Omega)h, \tag{8.16}$$

where C_1, C_2, C_3, C_4—some constants not depending on the mesh.

Note a solution to the differetial problem (8.1)–(8.5) by $\tilde{\mathbf{u}}$, $\tilde{\zeta}$, i.e.,

$$\tilde{\mathbf{u}}_t = g\nabla\tilde{\zeta} - R\tilde{\mathbf{u}} - \lambda\bar{k}\times\tilde{\mathbf{u}} + \mathbf{f},$$
$$\tilde{\zeta}_t = \mathrm{div}\tilde{\mathbf{u}}. \tag{8.17}$$

Substituting formulas for the operators from (8.13) into (8.17), we have

$$\frac{\tilde{\mathbf{u}}(t+\tau)-\tilde{\mathbf{u}}(t)}{\tau}\Big|_{O^k} = g\nabla^h\tilde{\zeta}\Big|_{O^k} - R\tilde{\mathbf{u}}\Big|_{O^k} - \lambda\bar{k}\times\tilde{\mathbf{u}}\Big|_{O^k} + \mathbf{f}\Big|_{O^k} + \overline{\varepsilon}\Big|_{O^k} + g\overline{\upsilon}\Big|_{O^k},$$
$$\frac{\tilde{\zeta}(t+\tau)-\tilde{\zeta}(t)}{\tau}\Big|_{X_i} = \mathrm{div}^h\tilde{\mathbf{u}}\Big|_{X_i} + \underline{\varepsilon}\Big|_{X_i} + \underline{\upsilon}\Big|_{X_i}. \tag{8.18}$$

Let $\mathbf{u}^n$, ζ^n be a solution to finite-difference problem (8.9), (8.12) for nth time layer. Then the difference

$$\mathbf{w}^n = \tilde{\mathbf{u}}(n\tau) - \mathbf{u}^n, \quad \xi^n = \tilde{\zeta}(n\tau) - \zeta^n \tag{8.19}$$

satisfies the system of equations

$$\frac{\mathbf{w}^{n+1}-\mathbf{w}^n}{\tau}\Big|_{O^k} = g\nabla^h\xi^{n+1}\Big|_{O^k} - R\mathbf{w}^{n+1}\Big|_{O^k} - \lambda\bar{k}\times\mathbf{w}^{n+1}\Big|_{O^k} + \overline{\varepsilon}\Big|_{O^k} + g\overline{\upsilon}\Big|_{O^k},$$
$$\frac{\xi^{n+1}-\xi^n}{\tau}\Big|_{X_i} = \mathrm{div}^h\mathbf{w}^{n+1}\Big|_{X_i} + \underline{\varepsilon}\Big|_{X_i} + \underline{\upsilon}\Big|_{X_i},$$
$$\mathbf{w}^0 = 0,\ \xi^0 = 0. \tag{8.20}$$

From conjugation of mesh operators and the second equation of (8.20) it follows

$$(\nabla^h\xi^{n+1}, \mathbf{w}^{n+1}) = -(\mathrm{div}^h\mathbf{w}^{n+1}, \xi^{n+1}) = -(\xi_t^{n+1}, \xi^{n+1}) + (\underline{\varepsilon}+\underline{\upsilon}, \xi^{n+1}), \tag{8.21}$$

where $\xi_t^{n+1} = \frac{\xi^{n+1}-\xi^n}{\tau}$.

Take a scalar product of the first equation of (8.20) and $2\tau\mathbf{w}^{n+1}$. After some obvious transformations we obtain

$$\begin{aligned}\|\mathbf{w}^{n+1}\|^2 - \|\mathbf{w}^n\|^2 + \tau^2\|\mathbf{w}_t^{n+1}\|^2 + 2R\tau\|\mathbf{w}^{n+1}\|^2 \\ +\ g(\|\xi^{n+1}\|^2 - \|\xi^n\|^2 + \tau^2\|\xi_t^{n+1}\|^2) \\ = 2\tau(\overline{\varepsilon} + g\overline{\upsilon}, \mathbf{w}^{n+1}) + 2g\tau(\underline{\varepsilon}+\underline{\upsilon}, \xi^{n+1}).\end{aligned} \tag{8.22}$$

Estimate the right-hand side with the use of Cauchy-Bunyakovskii-Schwarz inequality and the Young inequality with the parameters (2, 2):

$$\begin{aligned}\|\mathbf{w}^{n+1}\|^2 - \|\mathbf{w}^n\|^2 + \tau^2\|\mathbf{w}_t^{n+1}\|^2 + 2R\tau\|\mathbf{w}^{n+1}\|^2 \\ +\ g(\|\xi^{n+1}\|^2 - \|\xi^n\|^2 + \tau^2\|\xi_t^{n+1}\|^2) \\ \le \tau\|\overline{\varepsilon} + g\overline{\upsilon}\|^2 + \tau\|\mathbf{w}^{n+1}\|^2 + \tau g\|\underline{\varepsilon}+\underline{\upsilon}\|^2 + \tau g\|\xi^{n+1}\|^2.\end{aligned}$$

Sum up the above relation over time layers $n = 0, 1, \ldots, m-1$ and use the initial conditions $w^0 = 0$ and $\xi^0 = 0$:

$$\|\mathbf{w}^m\|^2 + g\|\xi^m\|^2 + \sum_{n=1}^{m} \left(\tau^2\|\mathbf{w}_t^n\|^2 + 2R\tau\|\mathbf{w}^n\|^2 + g\tau^2\|\xi_t^n\|^2\right)$$
$$\leq m\tau(\|\overline{\varepsilon} + g\overline{\upsilon}\|^2 + g\|\underline{\varepsilon} + \underline{\upsilon}\|^2) + \tau\sum_{n=1}^{m}(\|\mathbf{w}^n\|^2 + g\|\xi^n\|^2).$$

The Gronwall lemma [8] gives

$$\|\mathbf{w}^m\|^2 + g\|\xi^m\|^2 + \tau\sum_{n=1}^{m}\left(\tau\|\mathbf{w}_t^n\|^2 + 2R\|\mathbf{w}^n\|^2 + g\tau\|\xi_t^n\|^2\right)$$
$$\leq exp\left(\frac{\tau m}{1-\tau}\right)\left(\tau\sum_{n=1}^{m}(\|\overline{\varepsilon} + g\overline{\upsilon}\|^2 + g\|\underline{\varepsilon} + \underline{\upsilon}\|^2)\right). \tag{8.23}$$

Let $t = m\tau$. Then using (8.15) and (8.16), one has

$$\|\mathbf{w}^m\|^2 + g\|\xi^m\|^2 \leq e^{\frac{t}{1-\tau}}\left(\tau\sum_{n=1}^{m}(\|\overline{\varepsilon} + g\overline{\upsilon}\|^2 + g\|\underline{\varepsilon} + \underline{\upsilon}\|^2)\right)$$
$$\leq C^2 t\,\mathrm{mes}(\Omega)e^{\frac{t}{1-\tau}}(\tau + h)^2, \tag{8.24}$$

where $C = \sqrt{\max\{g(\max\{C_1, C_3\})^2, (\max\{C_2, C_4\})^2\}}$. Since the last inequality is fulfilled for each $m = 0, \ldots, M$, where $T = m\tau$, then

$$\max_{m=0,\ldots,M}(\|\mathbf{w}^m\|^2 + g\|\xi^m\|^2) \leq C^2 T\,\mathrm{mes}(\Omega)e^{\frac{T}{1-\tau}}(\tau + h)^2,$$

Taking into account (8.19), we get

$$\max_{m=0,\ldots,M}\|\tilde{\mathbf{u}}(t_m) - \mathbf{u}^m\| \leq C\sqrt{T\,\mathrm{mes}(\Omega)}e^{\frac{T}{2-2\tau}}(\tau + h),$$
$$\max_{m=0,\ldots,M}\|\tilde{\zeta}(t_m) - \zeta^m\| \leq C\sqrt{T\,\mathrm{mes}(\Omega)}e^{\frac{T}{2-2\tau}}(\tau + h); \tag{8.25}$$

a constant C does not depend on τ and h.

Therefore, a solution to the finite-dimensional problem converges to a solution to the differential one for $\tau, h \to 0$. The theorem is proved.

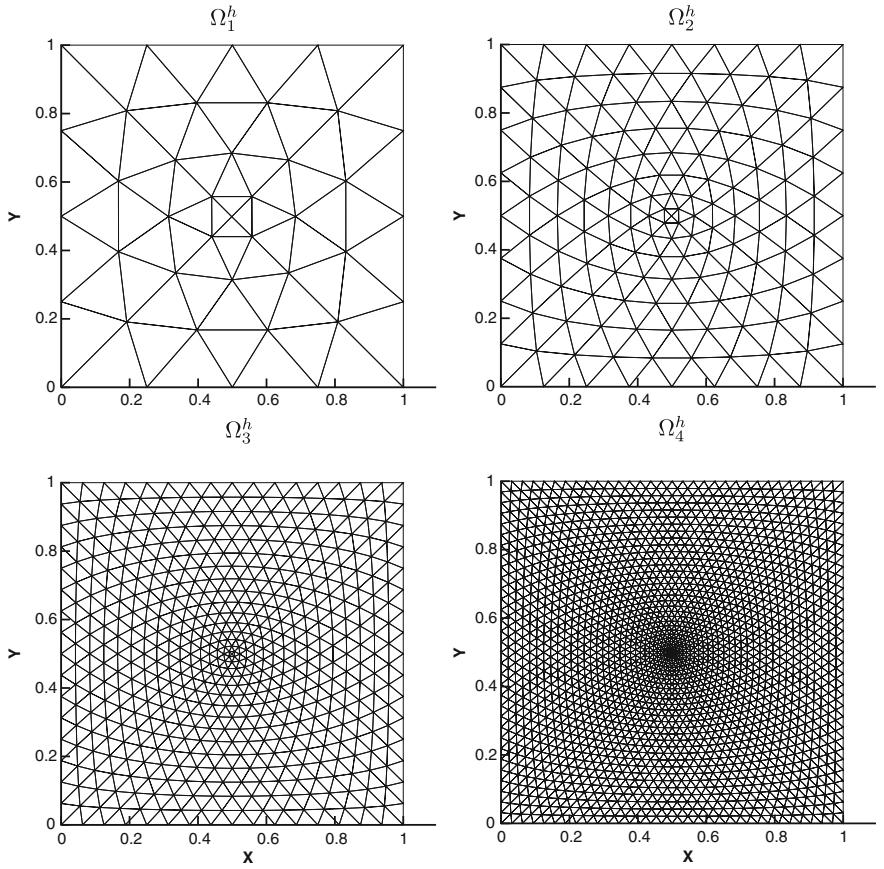


Fig. 8.3 Nets $\Omega_1^h, \Omega_2^h, \Omega_3^h, \Omega_4^h$

8.6 Results of Numerical Experiments

8.6.1 Estimation of Convergence Order

To check order of convergence, a number of numerical experiments was carried out on the domain $\Omega = [0, 1] \times [0, 1]$. An analytical solution to the differential problem for the appropriate right-hand side is of the form:

$$\tilde{u}(x, y, t) = -(\cos(6t) + 2)\sin(2\pi x)\sin(\pi y),$$
$$\tilde{v}(x, y, t) = (\cos(6t) + 2)\sin(\pi x)\sin(2\pi y),$$
$$\tilde{\zeta}(x, y, t) = -2\pi\left(\frac{\sin(6t)}{6} + 2t\right)(\cos(2\pi x)\sin(\pi y) - \sin(\pi x)\cos(2\pi y)).$$

Impermeability boundary condition was put. Parameters and time step were chosen as follows: $R = \frac{1}{100}$, $\lambda = 0$, $T = 1$.

With the use of the mesh generators **gmsh** [9] a number of meshes were constructed each of which was obtained from the previous one by dividing of each cell onto for ones. A diameter of maximal cell for such a variety of meshes diminishes two times when passing from the course mesh to the finer one. There were constructed 4 meshes and for each mesh the appropriate time step was chosen:

- Ω_1^h—number of triangles—64, number of flow nodes—104. Time step $\tau = 0.01$. Number of time steps—100.
- Ω_2^h—number of triangles—256, number of flow nodes—400. Time step $\tau = 0.005$. Number of time steps—200.
- Ω_3^h—number of triangles—1024, number of flow nodes—1568. Time step $\tau = 0.0025$. Number of time steps—400.
- Ω_4^h—number of triangles—4096, number of flow nodes—6208. Time step $\tau = 0.00125$. Number of time steps—800.

In Fig. 8.3 the meshes Ω_1^h, Ω_2^h, Ω_3^h, Ω_4^h are illustrated. In the table given below, the norm of a difference between finite-difference and analytical solutions is presented:

$$F_{\Omega^h}(t) = \|\hat{\zeta}(x, y, t) - \tilde{\zeta}(x, y, t)\|_{2,h}.$$

Domain	τ	$F_{\Omega}(1)$
Ω_1^h	0.01	0.2609552
Ω_2^h	0.005	0.0674637
Ω_3^h	0.0025	0.0173317
Ω_4^h	0.00125	0.0049251

Remark 8.1 As is seen from the table, an order of convergence is $O(\tau^2 + h^2)$.

8.6.2 Computation of the Real Geographic Domain

The aim of this numerical experiment was numerical solution of the problem under consideration on the real domain, namely, Black sea. Initial conditions were chosen in the following way:

$$u_0(x, y) = 0, \quad v_0(x, y) = 0,$$

$$\zeta_0(x, y) = 0.25e^{-0.02(x - 2200)^2 - (y - 2500)^2} + 0.2e^{-0.02(x - 1830)^2 - (y - 2500)^2}.$$

So, ζ_0 is some initial perturbation. We simulated distribution of the wave from the initial perturbation. Impermeability boundary condition was set. An unstructured mesh was constructed by the generator **ani2d** [10]. Number of triangles—7305,

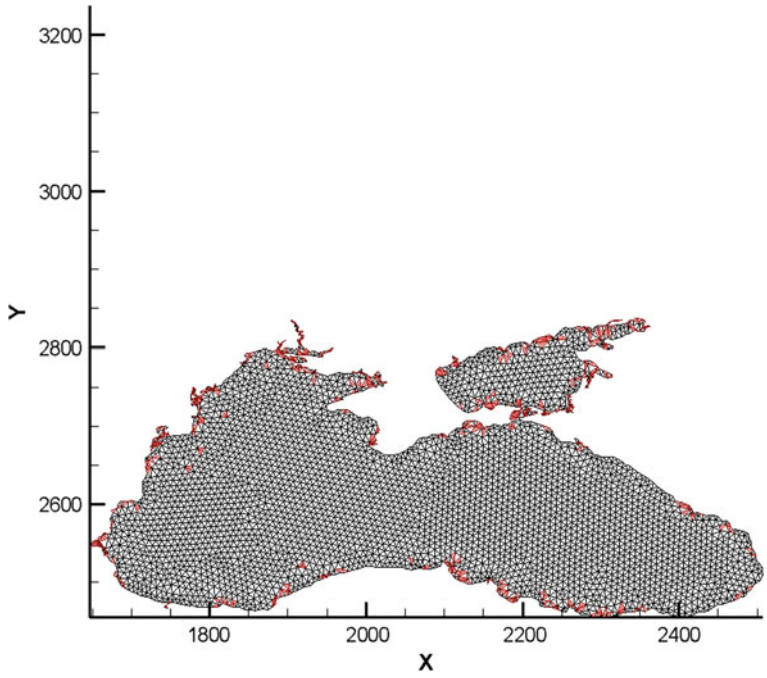


Fig. 8.4 Mesh on Black sea

number of flow nodes—11498. Number of obtusangular triangles is about 8 %. The mesh of Ω is illustrated in Fig. 8.5; obtusangular triangles are marked by red. The parameters and time step were chosen in the following way:

$$R = 0, \qquad \lambda = 0, \qquad T = 100.0, \quad \tau = 0.1.$$

In Figs. 8.5, 8.6, 8.7, 8.8, 8.9 and 8.10 wave heights obtained by computations are presented at various time instants: $t = 0$, $t = 10$, $t = 20$, $t = 40$, $t = 60$, $t = 80$.

Thus, approximation of the problem under consideration proposed and justified in the chapter allows us to simulate propagation of waves in domains of complex form with the use of unstructured meshes and can be applied for modeling tidal wave dynamics of lakes, rivers, seas, oceans, etc.

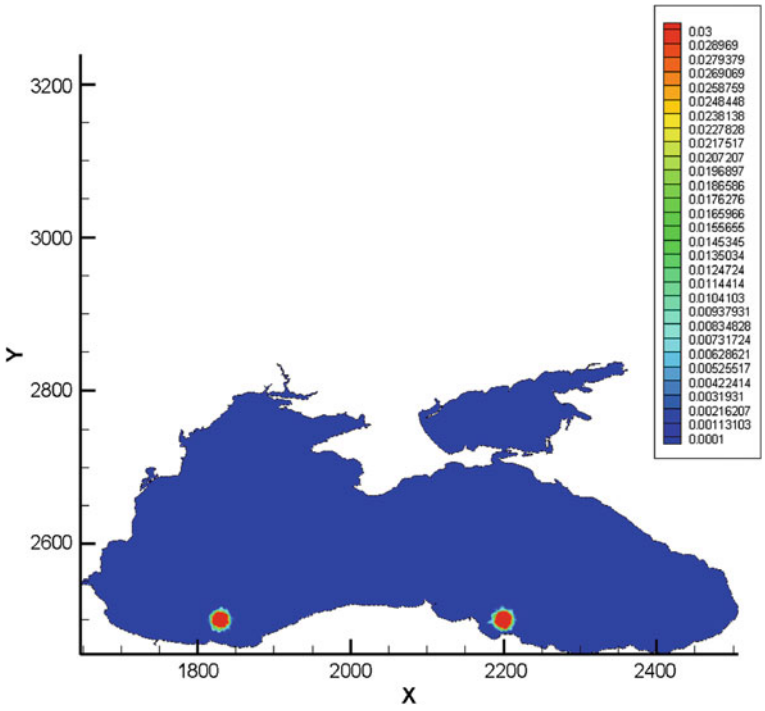


Fig. 8.5 Wave height at initial time instant

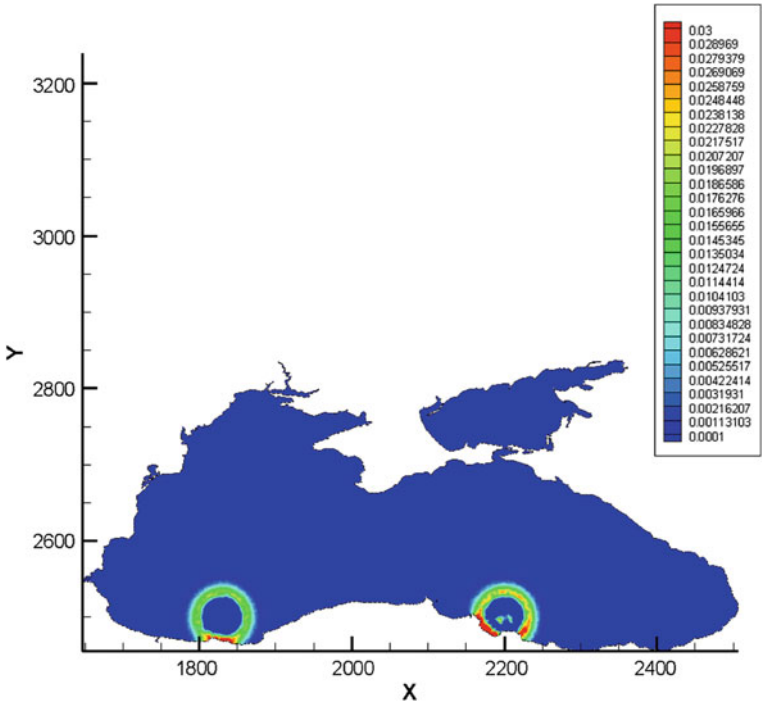


Fig. 8.6 Wave height at time instant $t = 10$

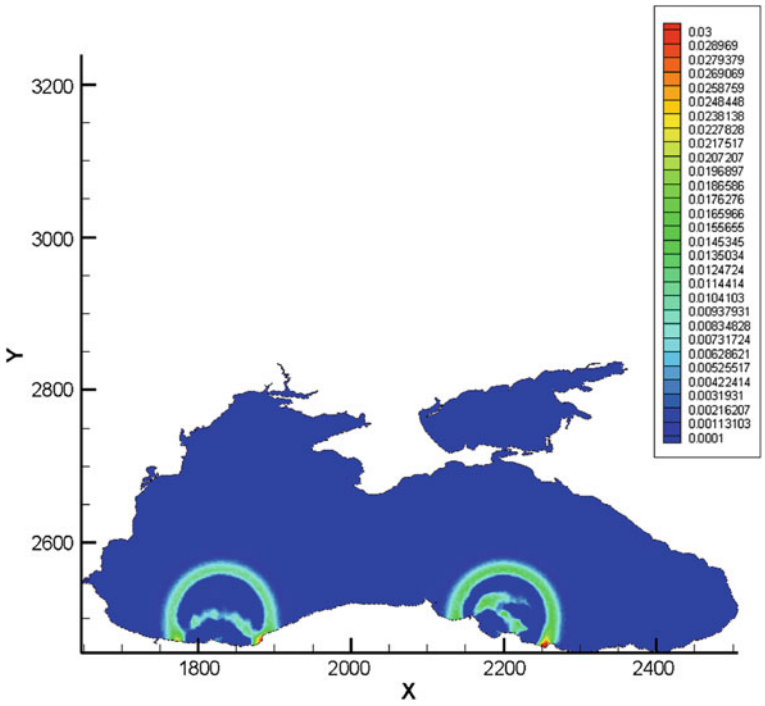


Fig. 8.7 Wave height at time instant $t = 20$

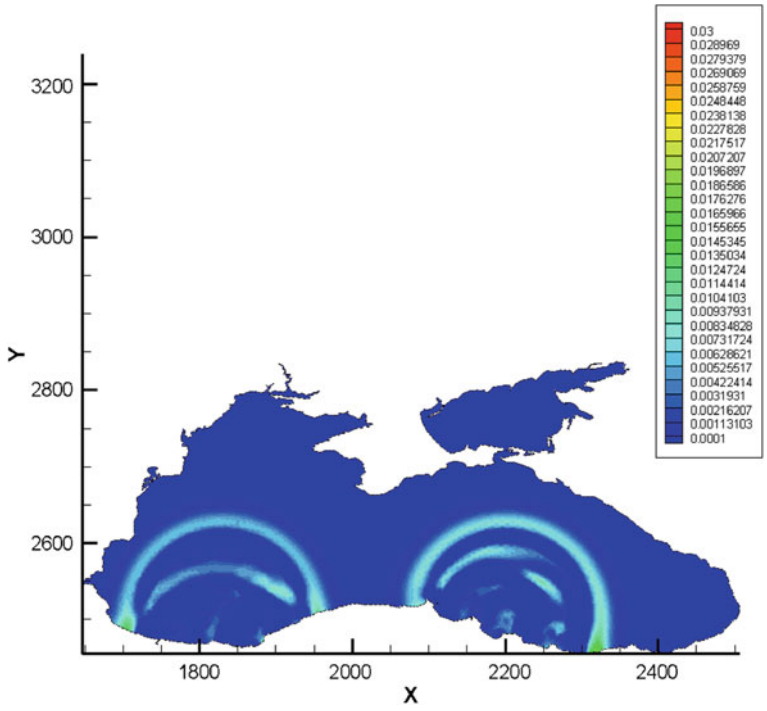


Fig. 8.8 Wave height at time instant $t = 40$

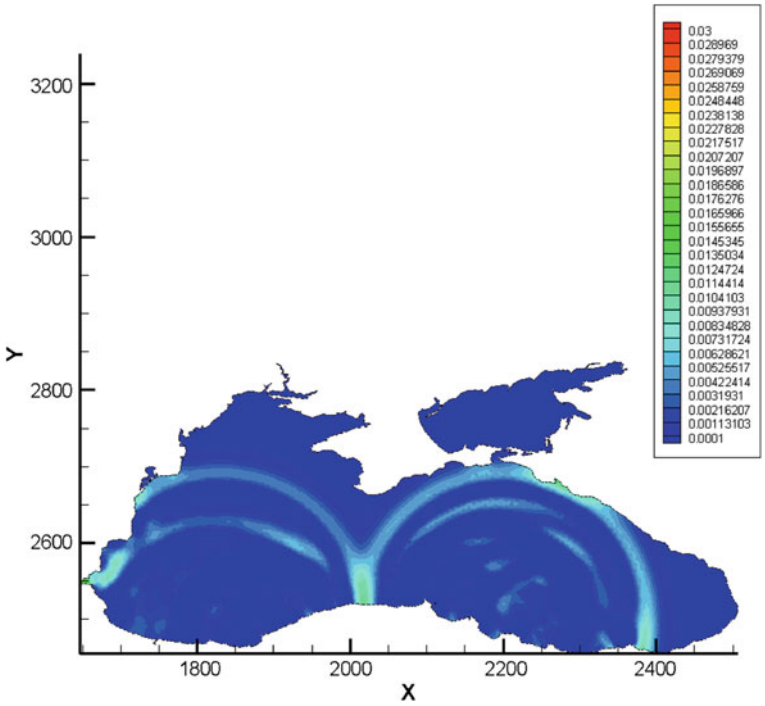


Fig. 8.9 Wave height at time instant $t = 60$

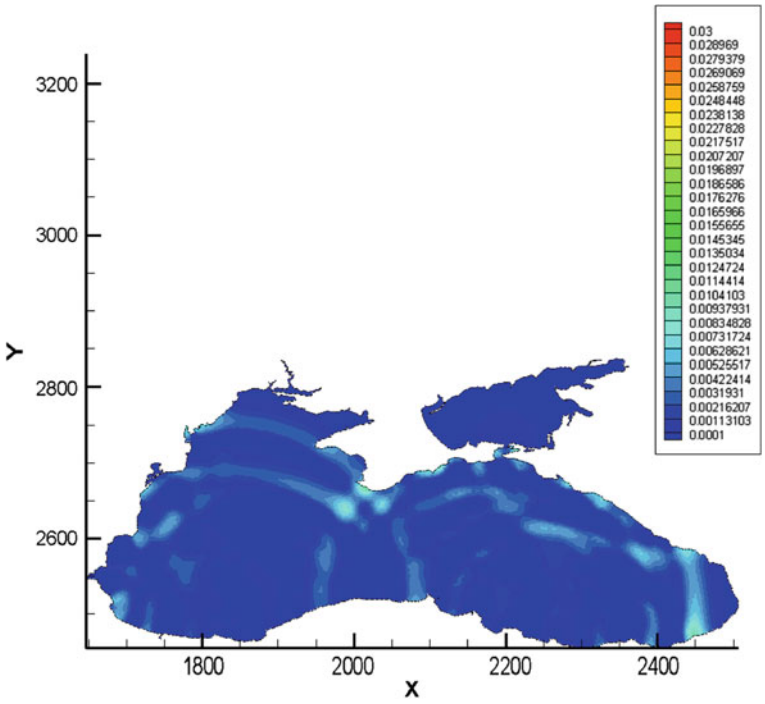


Fig. 8.10 Wave height at time instant $t = 80$

Acknowledgments The authors are grateful to V. Zalesny, Yu. Vasilevskii and A. Gusev for valuable discussions.

References

1. Bogachev, K.Yu., Kobelkov, G.M. Numerical solution of a tidal wave problem. In: Proceedings of "Parallel Computational Fluid Dynamics", vol. 2, pp. 163–173. J.-Wiley Press (2004)
2. Arushanyan, I.O., Drutsa, A.V., Kobelkov, G.M.: Finite-difference method for solution of the system of tidal dynamics equations. Diff. Equat. **45**(7), 965–972 (2009) (in Russian)
3. Kobelkov, G.M., Drutsa, A.V.: Finite difference approximation of tidal wave equations on unstructured grid in spherical coordinates. Russ. J. Numer. Math. Math. Model. **25**(6), 535–544 (2010)
4. Agoshkov, V.I., Botvinovsky, E.A.: Numerical solution of a hyperbolic-parabolic system by splitting methods and optimal control approaches. Comput. Methods Appl. Math. **7**,(3), 193–207 (2007)
5. Zalesny, V.B.: Mathematical model of sea dynamics in a σ-coordinate system. Russ. J. Numer. Anal. Math. Modelling **20**(1), 97–113 (2005)
6. Marchuk, G.I., Kagan, B.A.: Ocean Tides. Gidrometeoizdat, Leningrad (1977) (in Russian)
7. Popov, I.V., Fryazinov, I.V., Stanichenko, M.Yu., Taimanov, A.V.: Construction of a difference scheme for Navier-Stokes equations on unstructured grids. Russ. J. Numer. Anal. Math. Modelling **23**(5), 487–503 (2009)
8. Heywood, J.G., Rannacher, R.: Finite-element approximation of the nonstationary Navier Stokes problem Part IV: error analysis for second-order time discretization, SIAM J. Numer. Anal. **27**(2), 353–384 (1990)
9. Geuzaine, C., Remacle, J.-F.: Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities. Int. J. Numer. Meth. Eng. **79**(11), 1309–1331 (2009)
10. Vassilevski, Yu., Lipnikov, K.: An adaptive algorithm for quasi-optimal mesh generation. Comput. Math. Math. Phys. **39**(9), 1468–1486 (1999)

Chapter 9
Dynamics of Vortices in Near-Wall Flows with Irregular Boundaries

I. M. Gorban and O. V. Khomenko

Abstract Behavior of stationary vortices in near-wall flows with irregular boundaries is investigated. The vortices were shown to locate in the critical points of flow and to be characterized not only by its strength but by the eigenfrequency that specifies precession of the vortex about the flow critical point along the small trajectory. Due to eigenfrequency, the stationary vortex responds selectively on external periodical perturbations. The last cause low-frequency vortex motion with large amplitudes and when the frequency of external perturbations is to be near the vortex eigenfrequency the vortex moves away from the critical point. So, dependency of the amplitude of perturbed vortex motion from the frequency of external perturbations has the resonance character. The resonant perturbations are shown to cause chaotization of local circulation zones generated by stationary vortices.

9.1 Introduction

Vortical structure of fluid flows is a determining factor when moving a body in water or in air as well as when operating hydraulic systems. A lot of important technical problems in fluid dynamics connect with optimal transformation of vortical pattern in the flow area. Artificial separation of flow resulting in generation of the local recirculation zone is the effective way that allows changing as the vortical flow pattern as the flow in whole. One may see the examples when artificial flow separation has been successfully applied in papers [1–6]. This method of control may be considered

I. M. Gorban (✉)
Institute of Hydromechanics, National Academy of Sciences of Ukraine,
8/4 Zheliabova St, Kyiv 03680, Ukraine
e-mail: ivgorban@gmail.com

O. V. Khomenko
Institute for Applied System Analysis, National Technical University of Ukraine
"Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kyiv03056, Ukraine
e-mail: olghomenko@mail.ru

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*,
Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_9,

as a way for regularization of near-wall flows at large Reynolds numbers. Transfer from a turbulent near-wall flow with chaotic motion of small-scale vortices to regular large-scale vortical pattern leads to reducing of energy exchange between the flow and the surface, in particular, to decreasing the body drag [3, 4]. The control strategy in this case is directed on creating the "intellectual" flow of fluid, in which the vortices are formed according to the control scheme and either theoretical or semiempirical model predicting the vortex behavior.

One of the ways to generate large-scale vortices in near-wall flow is artificial change of the surface configuration with help of bulges, grooves, ribs and so on [2–4]. The vortices may be immovable ones, stationary recirculation zones, or moving together with flow along the wall in regular manner. The fundamental requirement when generating the artificial vortex structures is their stability in respect of perturbations of external flow [7]. At the same time, the laboratory experiments testify fast response of the local separation zones to external perturbations, especially with a periodic component. This sensitivity is known to grow when rising the Reynolds number of flow. So, the progress in development of near-wall flow control algorithms connects with researching dynamical properties of the large-scale vortices and nature of their chaotic behavior.

Because of generation of large-scale vortices in near-wall flows is under action of viscous forces its investigation demands development of the mathematical models and numerical algorithms basing on the Navier-Stokes equations. At the same time, dynamical properties of the vortices, their stability and interaction with external flow may be studied within the scope of the model of ideal fluid. The efforts in investigation of the vortex dynamics have led to some understanding of chaotization of fluid flows [8–10].

It has to be noted that one of advantages of the vortex dynamic models, which don't take into consideration viscous effects, is their simplicity. This fact permits use these models for creation of algorithms of flow control in near-wall areas. Discovered recently properties of motion of vortices and fluid particles in near-wall flows have allowed to derive new ways of near-wall flow control [11–13].

It has been mentioned above one of the effective ways to change a near-wall flow pattern is installation of special irregularities on the wall, in particular, cross grooves. For the first time this method was proposed in papers [5–7] for decreasing hydraulic losses in diffusers. Developed by Ringleb [7] the model of standing vortices in the cross grooves of special configuration allowed derive new shapes of diffusers with minimal hydraulic losses. Use of the cross groove as a control element in aerodynamics was demonstrated in papers [3, 14] where an influence of shape, size and location of the groove on wing hydrodynamic characteristics was experimentally investigated.

At the same time, researches noted considerable instability of the flows with stationary recirculation zones and standing vortices [4, 15] that makes difficult its using in engineering. The knowledge about causes of this phenomenon would permit to broaden the practical application of the control schemes with standing vortices.

The analysis shows [15–17] that minimal energy losses for generating and supporting standing vortices will be achieved if one takes into account flow topology in

the region under consideration. Modern methods of near-flow control are connected with creating the needed flow topology that characterized by location of flow critical points, its type, separatrix shape and so on. Note the flow topology governs also chaotic processes in the region.

In the present paper, topology of the flows in the regions with non-regular boundaries and standing vortices is researched on the base of the standing vortex model. It will be shown that the vortex located in the neighborhood of a stable critical point is characterized by eigenfrequency which responsible for dynamical reaction of the vortex with external flow perturbations.

9.2 Model of Standing Vortex

A simplified model that describes dynamic properties of local recirculation zones formed near non-regular flow boundaries is considered. Linear parameters of the surface irregularity are supposed to exceed considerably the boundary layer thickness on the wall. The separation zone is simulated by a vortex that locates in the vorticity center and whose circulation is equal to integral vorticity strength in the region. In spite of simplicity, this model is effective enough for researching dynamic properties of near-wall flows [7].

Two-dimensional flow of ideal incompressible fluid bounded by non-regular wall is considered. Motion of a vortex located in this region is governed by a set of non-linear equations:

$$\frac{dx_v}{dt} = v_x(x_v, y_v, t), \quad \frac{dy_v}{dt} = v_y(x_v, y_v, t), \tag{9.1}$$

where x_v, y_v are the vortex coordinates and v_x, v_y are the components of the vortex velocity.

To determine the right part of system (9.1), one has to solve the Laplace equation for the complex flow potential Φ:

$$\Delta\Phi = 0 \tag{9.2}$$

with boundary conditions on the wall:

$$\left.\frac{\partial\Phi}{\partial n}\right|_{\Sigma} = 0, \tag{9.3}$$

and at infinity:

$$\left.\frac{\partial\Phi}{\partial z}\right|_{z\to\infty} = U_0, \tag{9.4}$$

Here Σ is the flow boundary and U_0 is the flow velocity.

Note if the velocity U_0 does not change in time, system (9.1) will be autonomous one. Analysis of its solutions may be carried out with applying the theory of critical points [18]. According to this theory, critical points of the flow with a vortex are determined from the condition of vortex equilibrium:

$$v_x(x_v, y_v) = 0, \quad v_y(x_v, y_v) = 0. \tag{9.5}$$

The divergence $div v = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y}$ and Jacobean $J = (\frac{\partial v_x}{\partial x})(\frac{\partial v_y}{\partial y}) - (\frac{\partial v_x}{\partial y})(\frac{\partial v_y}{\partial x})$ of set (9.1) specify the type of critical points. The critical point may be a saddle, if $J < 0$, a node, if $J < \pm\frac{div^2 v}{4}$, and a focus, if $J > \pm\frac{div^2 v}{4}$. Saddles are always unstable points, noses and foci may be either stable, when $div v < 0$, or unstable, when $div v > 0$. As we consider conservative flows, without energy supply, the divergence is equal to zero. Then critical points may be either unstable hyperbolic, if $J < 0$, or elliptical, when $J > 0$. For us, the latest points are interesting because they are conditionally stable ones and such a flow may be only realized in practice. The vortex, whose parameters are similar to those of the standing vortex, moves periodically around the elliptical point. For the standing vortex, the precession trajectory is infinitesimal and the precession frequency $\omega_0 = \sqrt{J}$ may be considered as its eigenfrequency. The eigenfrequency is a very important characteristic of the standing vortex. In particular, it governs the vortex reaction to external flow perturbations.

To find the solution of Eq. (9.2), we use the conformal mapping of the flow field in the physical z-plane into an upper half-plane of the auxiliary plane $\zeta(\xi, \eta)$. In ζ-plane, the complex flow potential is:

$$\Phi(\zeta) = \Phi_0(\zeta) + \frac{\Gamma}{2\pi i} \ln \frac{\zeta - \zeta_v}{\zeta - \bar{\zeta}_v}, \tag{9.6}$$

where Γ is the vortex circulation, ζ_v and $\Phi_0(\zeta)$ are the vortex complex coordinate and the non-separated flow potential in ζ-plane respectively.

If the conformal mapping function $\zeta = f(z)$ is known, one has the following expression for the vortex velocity in the physical plane:

$$\overline{v}(x_v, y_v) = \left(\frac{d\Phi_0}{d\zeta} + \frac{\Gamma}{4\pi\eta_v}\right)\frac{df}{dz}\bigg|_{\zeta=\zeta_v} + \frac{\Gamma}{4\pi i}\left(\frac{d^2 f}{dz^2} \Big/ \frac{d\zeta}{dz}\right)\bigg|_{\zeta=\zeta_v}. \tag{9.7}$$

The real and imaginary components of (9.7) are the right-hand sides of (9.1).

The coordinates x_0, y_0 of the critical point are determined from the condition of the flow equilibrium here:

$$\overline{v}|_{z=z_0} = 0, \tag{9.8}$$

where $z_0 = x_0 + i y_0$.

Taking into account that coordinates of the critical point and the standing vortex coincide, we obtain from (9.7) the following equation:

$$\left(\left.\frac{d\Phi_0}{d\zeta}\right|_{\zeta=\zeta_0}+\frac{\Gamma}{4\pi\eta_0}\right)\left[\left(\frac{df}{dz}\right)^2\Big/\frac{d^2f}{dz^2}\right]\Bigg|_{\zeta=\zeta_0}-\frac{i\Gamma}{4\pi}=0. \tag{9.9}$$

From (9.9), two transcendental equations for determining the standing vortex coordinates are derived. To calculate the vortex circulation, this set has to be completed by an equation that follows from physical conditions of the problem under consideration. For example, if the flow boundary has a sharp edge, the unsteady Kutta condition can be involved.

9.3 Standing Vortex in Cross Groove

It was mentioned above cross grooves on the flowed surface are an effective way of near-wall flow control. We make here analysis of dynamic properties of the standing vortex in the uniform flow above the surface with a circular groove. The geometry of interest in the present study is presented in Fig. 9.1a. The mapping function that transforms the half-plane with a cut circular hollow (Fig. 9.1a) into the upper half-plane (Fig. 9.1b) has the following form:

$$f(z)=a\gamma\frac{1+\left(\dfrac{z-a}{z+a}\right)^{\gamma}}{1-\left(\dfrac{z-a}{z+a}\right)^{\gamma}},\quad \gamma=\frac{\beta}{\pi-\beta} \tag{9.10}$$

Here a is the semichord of groove, angle β characterizes the groove depth ($\beta<0$). The dependence of the groove depth on the angle β is shown in Fig. 9.3, curve 1. The semichord a and the free-stream velocity U_0 are characteristic parameters of the problem. The dimensionless circulation is introduced as $\overline{\Gamma}=\Gamma/aU_0$.

The stationary point coordinates x_0, y_0 and standing vortex circulation Γ_0 are determined from (9.9) and Kutta condition in the sharp groove edges. The last requires finiteness of the flow velocity in the groove edge:

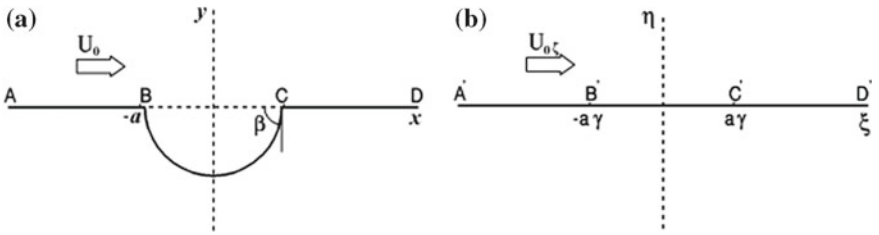


Fig. 9.1 Coordinate system in the physical plane z and the transformed plane ζ. Here $ABCD$ denote points in the physical plane which are mapped to points in the transformed plane $A'B'C'D'$

$$\left.\frac{d\Phi}{dz}\right|_{z=z_*} = const, \tag{9.11}$$

where z_* is the coordinate of the sharp edge in the physical plane.

Using the ratio $\left.\frac{d\Phi}{dz}\right|_{z=z_*} = \left.\frac{d\Phi}{d\zeta}\frac{df}{dz}\right|_{z=z_*}$ and taking into account that the function $f(z)$ has a singularity in the sharp edge, one obtains:

$$\left.\frac{d\Phi}{d\zeta}\right|_{\zeta=\zeta_*} = 0, \tag{9.12}$$

where $\zeta_*(\xi_*, 0)$ is the coordinate of the sharp edge in ζ-plane.

Taking into account symmetry of the flow region, it is sufficient to fulfill condition (9.12) in one groove edge only. As the unity flow in the physical plane transfers into the same flow in the transformed plane, from (9.6) the following equation may be derived:

$$\pi + \frac{\Gamma_0\eta_0}{(\xi_* - \xi_0)^2 + \eta_0^2} = 0, \tag{9.13}$$

where (ξ_0, η_0) is the stationary point image in ζ-plane.

In the present research, Eqs. (9.9 and 9.13) are solved numerically with applying the secant method. The obtained results show that there are three stationary points when a vortex interacts with the stream in the considered region. As seen on the portrait of vortex trajectories (Fig. 9.2a), two points locate near the groove edges. As follows from analysis of their stability they are unstable. So, the flow with such standing vortices does not realize in physical experiment. The elliptical stationary point, which is conditionally stable, lies on the groove axis. The flow pattern corresponding such a standing vortex is shown in Fig. 9.2b.

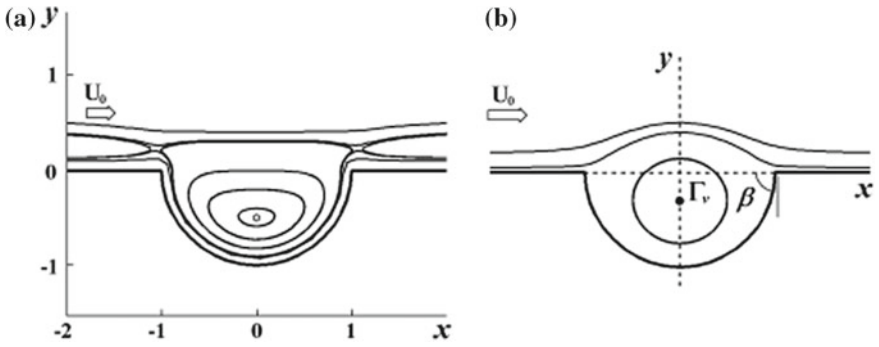


Fig. 9.2 Portrait of vortex trajectories—**a** and streamlines—**b** above the medium-sized groove ($\beta = -90°$)

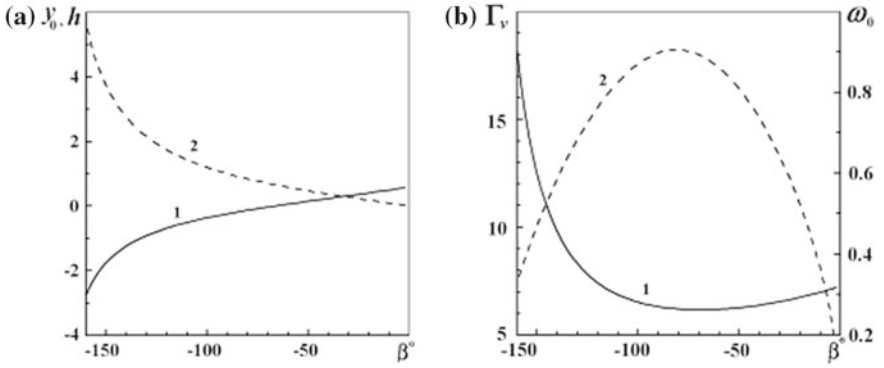


Fig. 9.3 Groove depth h (*curve 2*)—**a**, vertical coordinate y_0 (*curve 1*)—**a**, circulation Γ_v (*curve 1*)—**b** and eigenfrequency ω_0 (*curve 2*)—**b** of the standing vortex against the angle β

The vertical coordinate y_0 of the standing vortex against the angle β characterizing the groove depth is represented in Fig. 9.3a. It follows from this curve, the standing vortex locates above the wall for shallow grooves ($h < 0,2$). So, a very small groove on the flowed surface promotes to stabilization of a vortex here. Because of the vortices placed above the flat wall are always non-stable, shallow grooves may be used for stabilization of vortices in near-wall flows. It is important for development of the control schemes that use stable vortices on the surface ("vortical lubrication" of a wall).

The circulation Γ_v and eigenfrequency ω_0 of standing vortex against the angle β are represented in Fig. 9.3b. These results point out fast reduction ω_0 as with increasing as with decreasing the groove depth. The standing vortex circulation Γ_v is large enough in deep hollows and it grows slightly in shallow grooves due to approaching the vortex to surface $y = 0$ in this case. Minimal circulation and maximal eigenfrequency of the standing vortex are observed in medium-sized hollows ($\beta \approx -90°$).

9.4 Standing Vortex in an Angular Region

In the simple cases, for example, when the fluid flow in an angular region is considered (Fig. 9.4), the standing vortex parameters may be obtained analytically. The following function maps interior of the angle β into a half-plane:

$$\zeta = z^{\frac{\pi}{\beta}}. \tag{9.14}$$

Taking into account that potential of irrotational flow is $\Phi_0(\zeta) = -\zeta$, one has motion equations of a vortex within the angular region in the following form:

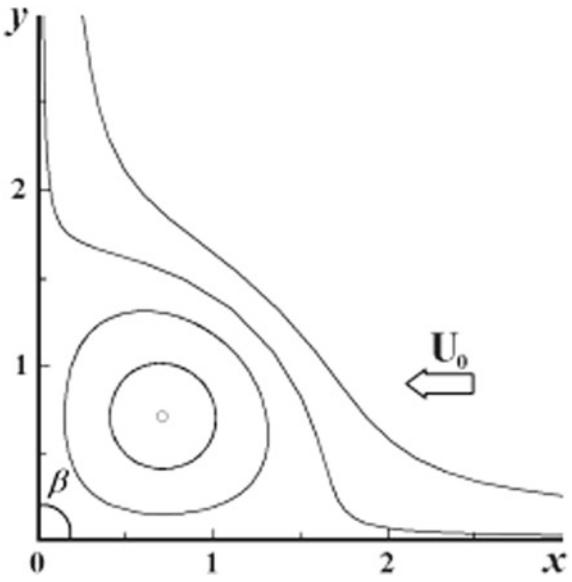


Fig. 9.4 Flow pattern with the standing vortex in an angular region ($\beta = \frac{\pi}{2}$)

$$\begin{aligned}\frac{dx_v}{dt} &= \left(\frac{\Gamma}{4\pi \sin\gamma\varphi} - 1\right)\gamma\cos\varphi(\gamma-1) - \frac{\Gamma}{4\pi}y_v \\ \frac{dy_v}{dt} &= -\left(\frac{\Gamma}{4\pi \sin\gamma\varphi} - 1\right)\gamma\sin\varphi(\gamma-1) + \frac{\Gamma}{4\pi}x_v,\end{aligned} \tag{9.15}$$

where $\varphi = \arctan\dfrac{y_v}{x_v}$, $\gamma = \dfrac{\pi}{\beta}$. The standing vortex circulation and coordinates of flow stationary point are derived by putting to zero the right-hand sides of (9.15):

$$\Gamma_0 = 4\pi\gamma, \quad x_0 = \cos\frac{\beta}{2}, \quad y_0 = \sin\frac{\beta}{2} \tag{9.16}$$

The carried out dynamic analysis shows the stationary point will be conditionally stable elliptic, when $\beta < \pi$. The circulation $\Gamma_v = \Gamma_0/4\pi$ and eigenfrequency ω_0 of the standing vortex against the angle β are represented in Fig. 9.5. Both the characteristics are seen growth when decreasing the angle β. So, the obtained results reveal the conditions when existence of the standing vortex in an angular region is possible and give value of the vortex parameters.

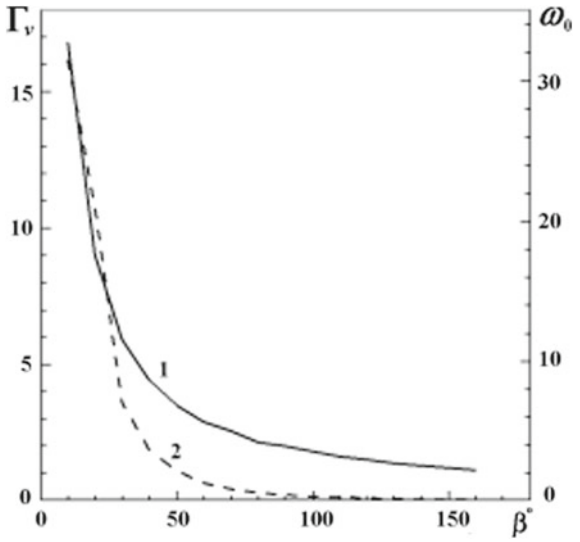


Fig. 9.5 The circulation Γ_v (*curve 1*) and eigenfrequency ω_0 (*curve 2*) of the standing vortex against angle β

9.5 Resonant Properties of Standing Vortices and Their Behavior in Perturbed Flow

In practice, near-wall flows are heterogeneous. There are many factors that entail nonstationarity of an external stream, such as body vibrations, migration of turbulent spots and motion of external vortices. So, it is crucial to investigate how behavior of a standing vortex changes under external flow disturbances. We consider here periodic perturbations of the flow velocity:

$$U = U_0(1 + \varepsilon \sin \Omega t), \quad \varepsilon \ll 1 \tag{9.17}$$

where ε, Ω are the amplitude and frequency of perturbations respectively.

It is supposed that at an initial instance $t = 0$, the vortex of circulation Γ_0 locates in the stable stationary point (x_0, y_0). Reaction of the vortex on perturbations given by (9.17) will be studied. To determine the vortex trajectory in the perturbed flow, (9.1) are integrated numerically by a fourth-order Runge-Kutta method.

The obtained results show the standing vortex begins to move around its stationary position under influence of the external perturbations. Character of this motion depends on ratio between the external frequency Ω and the eigenfrequency ω_0. If the value of external frequency is far from that of eigenfrequency ω_0 or its subharmonics $\frac{\omega_0}{2}$ and $2\omega_0$, the vortex will move periodically on a closed trajectory in the small neighborhood of stationary point. The neighborhood size is proportional to the amplitude of perturbations ε. The vortex trajectory will be much more complicated

when the external frequency Ω tends to the vortex frequency ω_0 or its subharmonics. Then multiperiodic large amplitude motion of the standing vortex is generated.

Motion of the vortex is characterized by its deviation from the stationary point (x_0, y_0):

$$R(t) = \sqrt{(x_v(t) - x_0)^2 + (y_v(t) - y_0)^2}. \tag{9.18}$$

Then the maximum deviation

$$R_{max} = max\{R(t) \mid t = (0, \infty)\} \tag{9.19}$$

gives us the amplitude of vortex motion in the perturbed flow.

As follows from the results obtained, the amplitude R_{max} is finite although the external perturbation is small. It is due to non-linear character of the equations that govern motion of a vortex near complex flow boundaries. Dependence R_{max} on the perturbation frequency Ω has the resonant character. Under $\Omega \to \omega_0$, the amplitude of the vortex precession R_{max} increases rapidly.

The curves characterizing function $R_{max}\left(\dfrac{\Omega}{\omega_0}\right)$ in angular regions are depicted in Fig. 9.6. Three curves there correspond to different values β. These results approve the resonant character of interaction between the standing vortex and periodic perturbations of external flow. The sharpest display of that is observed for blunt angles.

Flow perturbations lead also to significant stimulation of fluid mixing in the recirculation zone. If perturbations are absent, fluid particles of this zone will move along closed trajectories around the standing vortex. Under resonant perturbation, advection of the fluid particles intensifies. To define the character of motion of fluid particles and of standing vortex in the perturbed flow, the corresponding Puincare

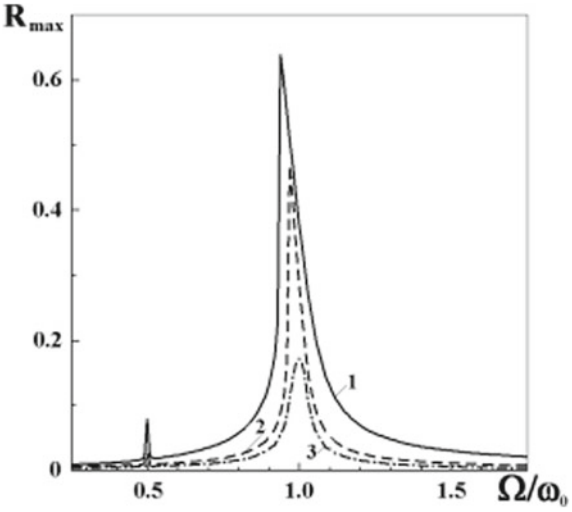


Fig. 9.6 Maximum deviation R_{max} of the standing vortex from the stationary point in an angular region against the relative frequency of external perturbation $\dfrac{\Omega}{\omega_0}$: $\varepsilon = 0, 01$, $1 - \beta = \dfrac{3\pi}{4}$, $2 - \beta = \dfrac{\pi}{2}$, $3 - \beta = \dfrac{\pi}{3}$.

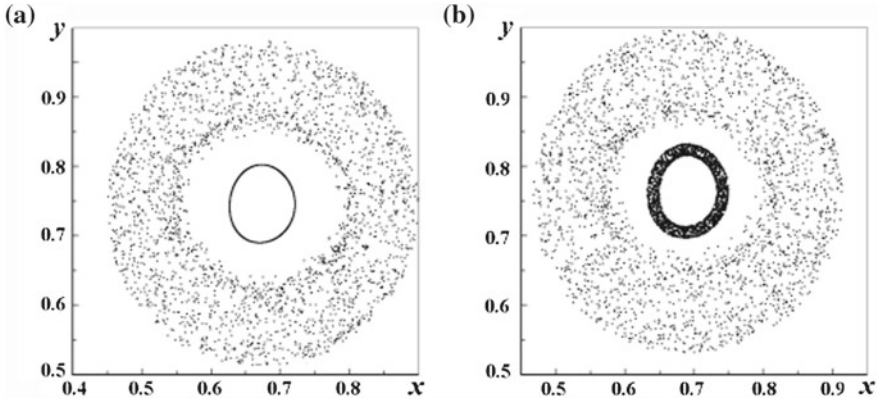


Fig. 9.7 Poincare sections of trajectories of the standing vortex and a fluid particle—**a** and two vortices of circulations $\Gamma_1 = \Gamma_v$ and $\Gamma_2 = \dfrac{\Gamma_v}{20}$—**b** in the angular region ($\beta = \dfrac{\pi}{2}$) at resonant flow perturbations: $\varepsilon = 0,001, \quad \Omega = \omega_0$

sections are computed when positions of particle or of vortex are calculated at the following points of time: $t_n = nT$, where $T = \dfrac{2\pi}{\Omega}$, $n = 1, 2, \ldots$. Then those are plotted in the physical flow region. The resulting Puincare sections in the angular region with $\beta = \dfrac{\pi}{2}$ and $\Omega = \omega_0$, $\varepsilon = 0,01$ are represented in Fig. 9.7a. It denotes the chaotic motion of fluid particles because the points depicting the particle positions after the period fill closely certain area in the physical plane. On the contrary, the points corresponding to the vortex positions dispose along the closed curve that indicates on regular character of the vortex motion.

Figure 9.7b depicts Puincare sections for two vortices placed in the flow with resonant perturbations. One of those is the standing vortex of circulation Γ_v. At an initial instance, it is located in the stationary point (x_0, y_0). Another small vortex, whose circulation is $\dfrac{\Gamma_v}{20}$, moves around the first one. It is obvious that motion of the small vortex has chaotic character. But in this case, the standing vortex positions after the period fill the annulus of the finite thickness. It points out presence of secondary small vortices in the perturbed flow leads to chaotic motions of the large-scale vortex generated in the recirculation zone. Such dynamic reaction of the large vortex on the external perturbations is very important factor that acts on development of flow as a whole. Note it is an example of appearance of chaos in nonautonomous system.

The similar behavior of the standing vortex is observed in the periodically perturbed flow above the surface with a gross groove. Figure 9.8 demonstrates the vortex trajectory and corresponding time dependence of vortex deviation $R(t)$ from the stationary point under condition that the perturbation frequency Ω is close to the vortex eigenfrequency ω_0, $(\Omega = 1,1\omega_0)$. The vortex motion is likely to be multiperiodic one with a small basic frequency and high-frequency pulsations. The amplitude of the vortex oscillations R_{max} is comparable with the groove size. Note

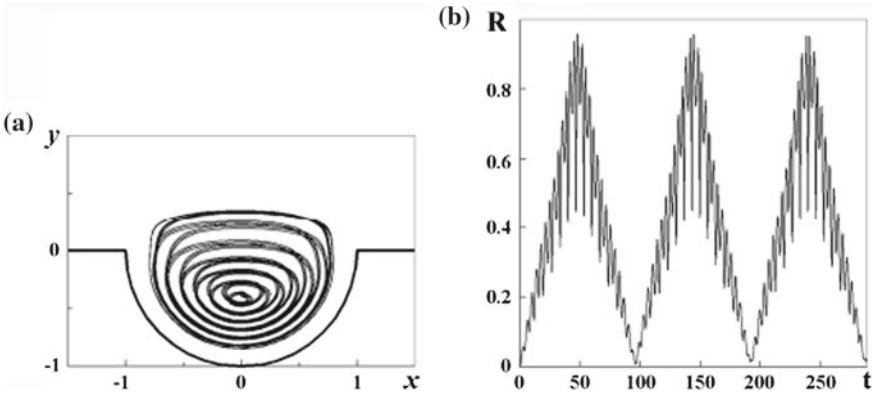


Fig. 9.8 Trajectory of the standing vortex in the perturbed flow near the wall with a groove—**a** and corresponding time dependence of the vortex deviation from the stationary point—**b**: $\varepsilon = 0, 1$, $\dfrac{\Omega}{\omega_0} = 1, 1$

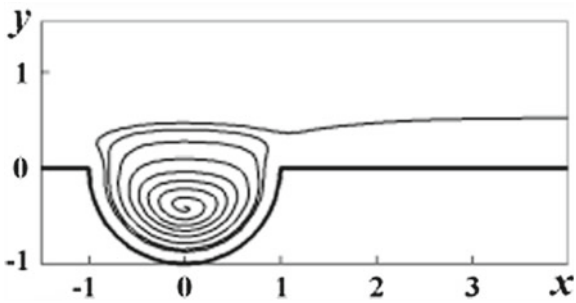


Fig. 9.9 Trajectory of the standing vortex in the perturbed flow when the vortex is carried away from a groove: $\varepsilon = 0, 1$, $\dfrac{\Omega}{\omega_0} = 1, 15$

a Kutta-Joukowski condition satisfies in the sharp edges of the boundary as long as the vortex is in a small neighborhood of the stationary point. With increasing the amplitude of perturbed motion R_{max}, this condition violates and groove edges begin to generate vortex layers.

Other unfavorable outcome is connected with ejection of the vortex into the near-wall region that is possible when the perturbation amplitude grows (Fig. 9.9). From a standpoint of dynamic analysis, the vortex loses its stability and jumps across the separatrix between different trajectories on a phase portrait (Fig. 9.2). In practice, the vortex is carried away by flow. Taking into account continuous generation of vorticity in the upstream edge, one may predict periodical replication of this process that leads to degradation of body hydrodynamic characteristics.

Dependence of the amplitude of vortex perturbed motion R_{max} on relative frequency of external perturbation $\dfrac{\Omega}{\omega_0}$ has the resonant character (Figs. 9.10, 9.11). The size of resonant peak depends on both the amplitude of perturbation ε and the groove depth h (or angle β). Under $\varepsilon > 0, 1$, the secondary peaks of a resonant curve

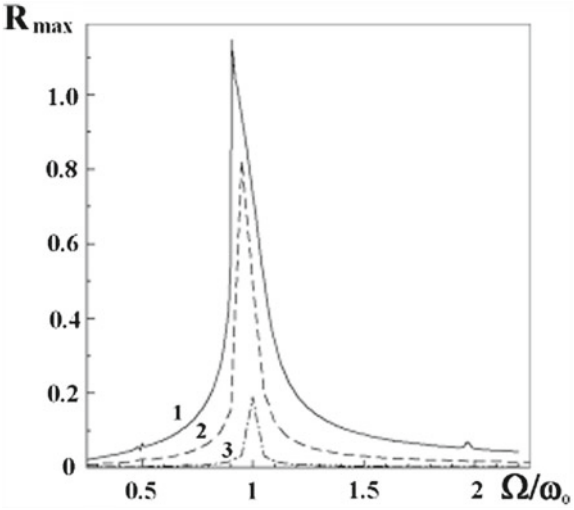


Fig. 9.10 Maximum deviation R_{max} of the standing vortex from the stationary point in different grooves against the relative frequency of external perturbation $\frac{\Omega}{\omega_0}$: $\varepsilon = 0,01$, $1-\beta = -5°$, $2-\beta = -30°$, $3-\beta = -150°$

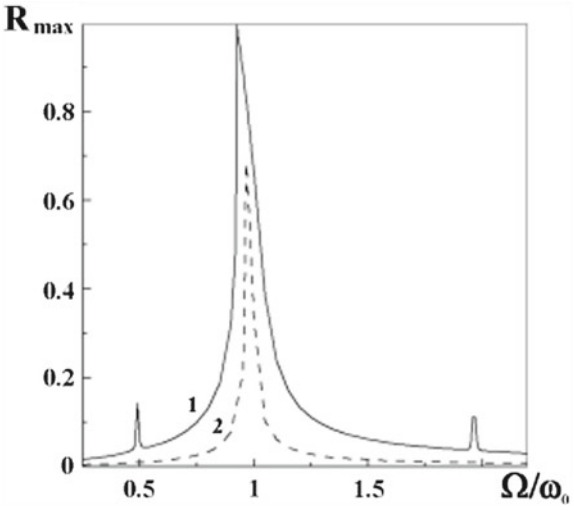


Fig. 9.11 Influence of intensity of external perturbations on dependence $R_{max}\left(\frac{\Omega}{\omega_0}\right)$: $\beta = -\frac{\pi}{6}$, $1-\varepsilon = 0,02$, $2-\varepsilon = 0,005$

near the frequencies $\frac{\omega_0}{2}$ and $2\omega_0$ (Fig. 9.11) take place due to non-linear character of the considered dynamic system.

External periodic perturbations in the flow above a hollow also lead to chaotic motion of fluid particles and small vortices in the field governed by the standing vortex that intensifies fluid mixing.

The obtained results show that instability of the standing vortex generated in near-wall flow with a non-regular boundary is connected with periodic perturbations which are present in the free-stream. Response of the vortex to perturbation is maximal when

the perturbation frequency is close to the vortex eigenfrequency that is display of resonant interaction between the vortex and the perturbed flow.

9.6 Summary

The pattern of the near-wall flow bounded by a non-regular surface is shown to depend on flow topological properties, in particular, on a type of flow critical points and existing of stationary vortices. If the critical point is stable, a strong enough vortex may be generated in the point environment (standing vortex). The vortex stabilizes the near-wall flow due to suppression of vorticity generation in sharp edges of the boundary.

A standing vortex is characterized by its eigenfrequency which governs the dynamic behavior of the vortex in the periodically perturbed flow. Periodic oscillations of the flow velocity cause multiperiodic large amplitude motion of the standing vortex. The maximal amplitude of deviation of the vortex from its stationary point depends on the external perturbation frequency in resonance manner. When the perturbation frequency approaches to the vortex eigenfrequency, the deviation amplitude grows rapidly.

Resonance flow perturbations in the regions bounded non-regular wall cause intensification of fluid mixing in recirculation zones. They stimulate generation of vorticity in sharp boundary edges, lead to chaotization of motion of both fluid particles and small vortices, cause non-regular fluctuations of the flow.

The obtained results are useful for further development of control algorithms in near-wall flows as well as for understanding of chaotization processes in nonautonomous systems.

References

1. Belov, I.A.: Interaction of Nonuniform Flows with Obstacles. Mashinostroenie, Moskow (1983). (in Russian)
2. Chernyshenko, S.I., Galleti, B., Iollo, Z.L.: Trapped vortices and a favourable pressure gradient. J. Fluid. Mech. **482**, 235–255 (2003)
3. Mkhitaryan, A.M., Lukashuk, S.A., Trubenok, V.D., Fridland, V.Ya.: Influence of spoilers on the aerodynamic characteristics of a wing and a solid of revolution. Naukova Dumka, Kyiv, 254–263 (1966) (in Russian)
4. Zheng, P.: Flow Separation Control. Mir, Moskow (1979) (in Russian)
5. Migaj, V.K.: The aerodynamic efficiency of discontinuous surface. Eng. Phys. J. **4**, 20–23 (1962). (in Russian)
6. Migaj, V.K.: The study of a finned diffuser. Teploenergetika. **10**, 55–59 (1962). (in Russian)
7. Ringleb, F.O.: Two-dimensional flow with standing vortex in ducts and diffusers. Trans. ASME. J. Basic. Eng. **10**, 921–927 (1960)
8. Aref, H.: Integrable, chaotic and turbulent vortex motion in two-dimensional flows. Annu. Rev. Fluid. Mech. **15**, 345–389 (1983)

9. Aref, H., Kadtke, J.B., Zawadzki, I.: Point vortex dynamics: recent results and open problems. Fluid. Dyn. Res. **3**, 63–64 (1988)
10. Veretentsev, A.N., Geshev, P.I., Kuibin, P.A., Rudyak, V.Ya.: On the development of the method of vortex particles as applied to the description of detached flows. Zh. Vychisl. Mat. Mat. Fiz. **29**(6), 878–887 (1989) (in Russian)
11. Acton, E., Dhanak, M.R.: The motion and stability of a vortex above a pulsed surface. J. Fluid. Mech. **247**, 231–246 (1993)
12. Cortelezzi, L.: Nonlinear feedback control of the wake past a plate with a suction point on the downstream wall. J. Fluid. Mech. **327**, 303–324 (1996)
13. Cortelezzi, L., Leonard, A., Doyle, J.C.: An example of active circulation control of the unsteady separated flow past a semi-infinite plate. J. Fluid. Mech. **260**, 127–154 (1994)
14. Gorban, V., Gorban, I.: Dynamics of vortices in near-wall flows: eigenfrequencies, resonant properties, algorithms of control. AGARD. Rep. **827**, 1–11 (1998)
15. Gorban, V.O., Gorban, I.M.: Resonant properties of vortices at boundary irregularities. Rep. NAS. Ukraine. **2**, 44–47 (1996). (in Ukrainian)
16. Gorban, V.O., Gorban, I.M.: The Study of dynamics of the vortex structures in an angular area and near by surface with hollow. Appl. Hydromech. **1**(1), 4–11 (1999). (in Ukrainian)
17. Perry, A.E., Chong, M.S.: A description of eddying motion and flow patterns using critical-point concept. Annu. Rev. Fluid. Mech. **19**, 125–155 (1998)
18. Zaslavsky. G.M., Sagdeev R.Z.: Introduction to Nonlinear Physics. Nauka, Moskow (1988)

Chapter 10
Strongly Convergent Algorithms for Variational Inequality Problem Over the Set of Solutions the Equilibrium Problems

Vladimir V. Semenov

Abstract This chapter deals with a variational inequality problem over the set of solutions the equilibrium problem or over the set of solutions the system of equilibrium problems in a real Hilbert space. Several new iterative algorithms are proposed. Strong convergence theorems for algorithms are proved. The convergence of iterative algorithms with the presence of computational errors without traditional summability conditions also studied. To this aim, we use new Mainge's techniques for analysis non–Fejerian iterative processes (Set–Valued Analysis. 16, 899–912, 2008).

10.1 Introduction

Throughout, H is a real Hilbert space with inner product $(\cdot, \cdot)$ and induced norm $\|\cdot\|$. We denote the strongly convergence and the weak convergence of (x_n) to $x \in H$ by $x_n \to x$ and $x_n \rightharpoonup x$, respectively. For operator $A : H \to H$, set $M \subseteq H$, and bifunction $F : H \times H \to \mathbb{R} \cup \{+\infty\}$ we denote by $VI(A, M)$ and $EP(F, M)$ sets $\{x \in M : \ (Ax, y - x) \geq 0 \ \ \forall\, y \in M\}$ and $\{x \in M : \ F(x, y) \geq 0 \ \ \forall\, y \in M\}$, respectively.

In this chapter, we are interested in the approximate solvability of problems

$$\text{find } \ x \in VI(A, EP(F, C)), \tag{10.1}$$

and,

$$\text{find } \ x \in VI(A, EP(F_1, C_1) \cap EP(F_2, C_2)). \tag{10.2}$$

V. V. Semenov (✉)
Department of Computational Mathematics, Taras Shevchenko
National University of Kyiv, Volodimirska str., 64, Kyiv 03601, Ukraine
e-mail: semenov.volodya@gmail.com

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_10,

Problems of form (10.1) or (10.2) are referred as variational inequality over the set of solutions the equilibrium problem or over the set of solutions the system of equilibrium problems. This problems have found applications in a wide array of disciplines, including mechanics, economics, partial differential equations, information theory, approximation theory, signal and image processing, game theory, optimal transport theory, probability and statistics, and machine learning. About the computational aspects mainly, see [5, 6, 12, 13, 15, 18, 22, 24].

Our main objective is to devise iterative algorithms for solving (10.1) and (10.2) and to analyze their asymptotic behavior. We'll use Mainge's techniques for analysis non–Fejerian iterative processes [11]. We are continuing our research published in [1, 7, 10, 12, 16, 17, 20, 22].

For solving the problem (10.1), let us assume that set $C \subseteq H$, bifunction $F : C \times C \to \mathbb{R}$, and operator $A : H \to H$ all satisfy the following set of standard properties:

(A1) $C \subseteq H$ is a nonempty closed convex set;
(A2) $F(x, x) = 0$, for all $x \in C$;
(A3) $F(x, y) + F(y, x) \leq 0$, for all $x, y \in C$ (monotonicity);
(A4) for each $x \in C$, the fuctional $F(x, \cdot)$ is convex and lower semicontinuous;
(A5) for each $x, y, z \in C$, $\limsup_{t \to +0} F(x + t(z - x), y) \leq F(x, y)$;
(A6) $EP(F, C) \neq \emptyset$;
(A7) $A : H \to H$ is a l–strongly monotone and L–Lipschitz continuous operator.

Remark 10.1 $EP(F, C)$ is a closed convex set [6]. Assumptions (A1)–(A7) guarantee the uniqueness and existence of the solution of variational inequality (10.1).

We need the following important notion.

Definition 10.1 ([6]) *The resolvent of a bifunction $F : C \times C \to \mathbb{R}$ is the set–valued operator $J_F : H \to 2^H : x \mapsto J_F x = \{z \in C : F(z, y) + (z - x, y - z) \geq 0 \ \forall y \in C\}$.*

Theorem 10.1 ([6]) *Let $C \subseteq H$ be a nonempty closed convex set, let $F : C \times C \to \mathbb{R}$ be a bifunction satisfying (A2)–(A5). Then, the following statements hold:*

(a) $\operatorname{dom} J_F = \{x \in H : J_F x \neq \emptyset\} = H$;
(b) J_F *single–valued and firmly nonexpansive, i.e.*

$$\|J_F x - J_F y\|^2 \leq (J_F x - J_F y, x - y) \quad \forall x, y \in H;$$

(c) $E(F, C) = \operatorname{Fix} J_F$, *where* $\operatorname{Fix} J_F = \{x \in H : J_F x = x\}$.

Remark 10.2 Equivalently, from the (b) we have

$$\|J_F x - J_F y\|^2 \leq \|x - y\|^2 - \|(x - J_F x) - (y - J_F y)\|^2 \quad \forall x, y \in H. \tag{10.3}$$

Now let us present algorithm for solving problem (10.1) under assumptions (A1)–(A7).

Algorithm 1 Select an arbitrary point $x_1 \in H$ and generates the sequence (x_n) iteratively by

$$\begin{cases} y_n = J_{\lambda_n F} x_n, \\ x_{n+1} = y_n - \alpha_n A y_n, \end{cases}$$

where $\lambda_n, \alpha_n > 0$.

Remark 10.3 This iterative method belongs to the class of the hybrid steepest descent methods [24].

Finally, let us present iterative algorithms for solving variational inequality (10.2). At first, we make the following assumptions throughout this chapter ($i = 1, 2$):

(B1) $C_i \subseteq H$ is a nonempty closed convex set;
(B2) $F_i(x, x) = 0$, for all $x \in C_i$;
(B3) $F_i(x, y) + F_i(y, x) \leq 0$, for all $x, y \in C_i$;
(B4) for each $x \in C_i$, the fuctional $F_i(x, \cdot)$ is convex and lower semicontinuous;
(B5) for each $x, y, z \in C_i$, $\limsup_{t \to +0} F_i(x + t(z - x), y) \leq F_i(x, y)$;
(B6) $EP(F_1, C_1) \cap EP(F_2, C_2) \neq \emptyset$;
(B7) $A : H \to H$ is a l–strongly monotone and L–Lipschitz continuous operator.

Remark 10.4 Assumptions (B1)–(B7) guarantee the uniqueness and existence of the solution of variational inequality (10.2).

Now we introduce two schemes for solving problem (10.2).

Algorithm 2 *(Barycentric)* Select an arbitrary point $x_1 \in H$ and generates the sequence (x_n) iteratively by

$$\begin{cases} y_n = J_{\lambda_n F_1} x_n, \\ z_n = J_{\lambda_n F_2} x_n, \\ v_n = \frac{1}{2} y_n + \frac{1}{2} z_n, \\ x_{n+1} = v_n - \alpha_n A v_n, \end{cases}$$

where $\lambda_n, \alpha_n > 0$.

Remark 10.5 Our Barycentric Algorithm 2 has parallel organization.

Algorithm 3 *(Alternating)* Select an arbitrary point $x_1 \in H$ and generates the sequence (x_n) iteratively by

$$\begin{cases} y_n = J_{\lambda_n F_1} x_n, \\ z_n = J_{\lambda_n F_2} y_n, \\ x_{n+1} = z_n - \alpha_n A z_n, \end{cases}$$

where $\lambda_n, \alpha_n > 0$.

Remark 10.6 Our alternating method is inspired by von Neumann's original alternating projections method [14, 8]. Let us mention that Algorithm 3 can be regarded

as a Halpern–type regularization for alternating method [2]. In 2005 Bauschke, Combettes and Reich [4, 3] studied the alternating resolvents method for finding a common zero of two maximal monotone mappings.

The remainder of the chapter is organized as follows. In Sect. 10.2, we provide technical facts that will be used in subsequent section. In Sect. 10.3, we establish convergence results for methods to solve (10.1) and (10.2). Some final conclusions are given in the last section.

10.2 Preliminaries

The following lemmas will be crucial in proving our main results.

Lemma 10.1 ([23]) *Let (ξ_n) be a sequence of nonnegative real numbers satisfying: $\xi_{n+1} \leq (1-\alpha_n)\xi_n + \alpha_n\beta_n + \gamma_n$ for any $n \in \mathbb{N}$, where (α_n), (β_n) and (γ_n) are real sequences such that: (i) $\alpha_n \in (0,1)$ with $\sum_{n=1}^{\infty} \alpha_n = +\infty$; (ii) $\limsup_{n\to\infty} \beta_n \leq 0$; (iii) $\gamma_n \in [0,+\infty)$ with $\sum_{n=1}^{\infty} \gamma_n < +\infty$. Then $\lim_{n\to\infty} \xi_n = 0$.*

Lemma 10.2 ([11]) *Let (a_n) be a sequence of real numbers such that there exists a subsequence (a_{n_k}) such that $a_{n_k} < a_{n_k+1}$ for all $k \in \mathbb{N}$. Then there exists a nondecreasing sequence (m_k) of $\mathbb{N}$ such that $m_k \to +\infty$ and $a_{m_k} \leq a_{m_k+1}$, $a_k \leq a_{m_k+1}$ for all $k \geq n_1$.*

Remark 10.7 Lemma 10.2 is fundamental tools for the techniques of analysis used through this chapter.

The following lemma put out basic contraction property of Lipschitz continuous and strongly monotone mappings.

Lemma 10.3 ([24]) *Let $A : D(A) \to H$ be a l–strongly monotone and L–Lipschitz continuous operator. An operator $T_\alpha : D(A) \to H$ defined by $T_\alpha x = x - \alpha Ax$, $\alpha \in (0,+\infty)$. Then the following inequality holds*

$$\|T_\alpha x - T_\alpha y\| \leq \left(1 - \tau\mu^{-1}\alpha\right)\|x-y\| \quad \forall x \in D(A)\ \forall y \in D(A),$$

where $\mu \in \left(0, 2lL^{-2}\right)$, $\alpha \in (0,\mu]$, $\tau = 1 - \sqrt{1-2l\mu+L^2\mu^2} \in (0,1]$.

We need the following "demiclosed–type" lemma.

Lemma 10.4 *Suppose that (x_n) is a sequence in H such that $\left(I - J_{\lambda_n F}\right)x_n \to 0$ and $x_n \rightharpoonup x$, where $\lambda_n \geq \lambda > 0$. Then $x \in EP(F,C)$.*

Proof We may assume without loss of generality that $\lambda_n \to \lambda_0 \in [\lambda,+\infty)\cup\{+\infty\}$ $n \to \infty$. We have $\lambda_n F\left(J_{\lambda_n F}x_n, y\right) + \left\langle J_{\lambda_n F}x_n - x_n, y - J_{\lambda_n F}x_n\right\rangle \geq 0$ for any $y \in C$. Since the bifunction F is monotone, we get

$$\lambda_n^{-1}\left\langle J_{\lambda_n F}x_n - x_n, y - J_{\lambda_n F}x_n\right\rangle \geq F\left(y, J_{\lambda_n F}x_n\right) \quad \forall y \in C.$$

We have $J_{\lambda_n F} x_n \rightharpoonup x$. Taking into account (A4), we obtain $0 \geq F(y, x)$ for any $y \in C$. For $t \in (0, 1)$, let $y_t = ty + (1-t)x \in C$. From (A2) and (A4), we have

$$0 = F(y_t, y_t) \leq tF(y_t, y) + (1-t)F(y_t, x) \leq tF(y_t, y).$$

Dividing by t, we get $F(y_t, y) \geq 0$. Letting $t \to 0$ and from (A5), we get $F(x, y) \geq 0$ foa all $\forall y \in C$. Therefore, we obtain $x \in EP(F, C)$. The proof is finished.

Remark 10.8 Let $\lambda_0 < +\infty$. In this case $x \in EP(F, C)$ follows from the well known resolvent identity (see [4]): $J_{\beta F} x = J_{\alpha F}\left(\frac{\alpha}{\beta} x + \left(1 - \frac{\alpha}{\beta}\right) J_{\beta F} x\right)$, $\alpha, \beta > 0$. Indeed, we have

$$\begin{aligned} \left\| J_{\lambda_n F} x_n - J_{\lambda_0 F} x_n \right\| &= \left\| J_{\lambda_0 F}\left(\lambda_0 \lambda_n^{-1} x_n + \left(1 - \lambda_0 \lambda_n^{-1}\right) J_{\lambda_n F} x_n\right) - J_{\lambda_0 F} x_n \right\| \\ &\leq |\lambda_n - \lambda_0|\, \lambda_n^{-1} \left\| J_{\lambda_n F} x_n - x_n \right\| \to 0. \end{aligned}$$

Therefore,

$$\left\| x_n - J_{\lambda_0 F} x_n \right\| \leq \left\| x_n - J_{\lambda_n F} x_n \right\| + \left\| J_{\lambda_n F} x_n - J_{\lambda_0 F} x_n \right\| \to 0 \text{ as } n \to \infty.$$

Since the operator $I - J_{\lambda_0 F}$ is demiclosed, we obtain $x \in \operatorname{Fix} J_{\lambda_0 F} = EP(F, C)$.

10.3 Convergence Analysis

Our first result is a strong convergence result for Algorithm 1. The following two Lemmas are needed for proving our convergence theorem.

Lemma 10.5 *Assume that* $\lim_{n\to 0} \alpha_n = 0$. *Let* (x_n), (y_n) *be sequences generated by Algorithm 1. Then* (x_n), (y_n) *are bounded.*

Proof Let $z \in EP(F, C)$. We have

$$\|x_{n+1} - z\| = \|y_n - \alpha_n A y_n - z\| \leq \|(y_n - \alpha_n A y_n) - (z - \alpha_n A z)\| + \alpha_n \|Az\|.$$

Using Lemma 10.3 with $\mu \in \left(0, \frac{2l}{L^2}\right)$ and $\alpha_n \in (0, \mu)$, we get

$$\|(y_n - \alpha_n A y_n) - (z - \alpha_n A z)\| \leq \left(1 - \alpha_n \mu^{-1} \beta\right) \|y_n - z\|,$$

where $\beta = 1 - \sqrt{1 - 2l\mu + L^2\mu^2} \in (0, 1)$. Since the operator $J_{\lambda F} : H \to C$ is nonexpansive and $x \in EP(F, C) \Leftrightarrow x = J_{\lambda F} x$ whenever $\lambda > 0$, we have

$$\|y_n - z\| = \left\| J_{\lambda_n F} x_n - J_{\lambda_n F} z \right\| \leq \|x_n - z\|. \tag{10.4}$$

Therefore,

$$\begin{aligned}\|x_{n+1}-z\| &\le \left(1-\beta\mu^{-1}\alpha_n\right)\|x_n-z\|+\beta\mu^{-1}\alpha_n\left(\mu\beta^{-1}\|Az\|\right)\\ &\le \max\left\{\|x_n-z\|,\mu\beta^{-1}\|Az\|\right\}.\end{aligned}$$

Since $\lim_{n\to\infty}\alpha_n=0$, we can assume that $\alpha_n\in(0,\mu)$ for all $n\in\mathbb{N}$. Hence,

$$\|x_{n+1}-z\|\le\max\left\{\|x_1-z\|,\mu\beta^{-1}\|Az\|\right\}. \tag{10.5}$$

From (10.5) it follows that the sequence (x_n) is bounded. From (10.4) it follows that the sequence (y_n) is bounded.

Lemma 10.6 *Let $z\in EP(F,C)$, and let (x_n), (y_n) be sequences generated by Algorithm 1. Then, for each $n\in\mathbb{N}$, the following inequality holds*

$$\begin{aligned}\|x_{n+1}-z\|^2-\|x_n-z\|^2+\|x_{n+1}-y_n\|^2\\ +\|y_n-x_n\|^2\le-2\alpha_n(Ay_n,x_{n+1}-z).\end{aligned} \tag{10.6}$$

Proof We have

$$\begin{aligned}\|x_{n+1}-z\|^2 &= \|y_n-\alpha_nAy_n-z\|^2=\|y_n-z\|^2-2\alpha_n(Ay_n,y_n-z)\\ &+\alpha_n^2\|Ay_n\|^2=\|y_n-z\|^2-2\alpha_n(Ay_n,x_{n+1}-z)-\|x_{n+1}-y_n\|^2.\end{aligned} \tag{10.7}$$

From (10.3) it follows that

$$\|y_n-z\|^2\le\|x_n-z\|^2-\|y_n-x_n\|^2. \tag{10.8}$$

Substituting right side of (10.8) for $\|y_n-z\|^2$ in (10.7), we obtain the inequality (10.6).

Now, we claim the main result for Algorithm 1.

Theorem 10.2 *Let hypotheses (A1)–(A7) hold. Assume that*

(i) $\alpha_n\in(0,+\infty)$, $\alpha_n\to0$, $\sum_{n=1}^{\infty}\alpha_n=+\infty$;
(ii) $\lambda_n\in[\lambda,+\infty)$, *where* $\lambda>0$.

Then the sequences (x_n), (y_n) generated by Algorithm 1 converge strongly to the unique $\overline{x}\in VI(A,EP(F,C))$.

Proof Let $\overline{x}\in H$ be the unique solution to (10.1). From Lemma 10.5 it follows that there exists $M>0$ such that $|(Ay_n,x_{n+1}-\overline{x})|\le M$ for all $n\in\mathbb{N}$. Using Lemma 10.6 we derive that

$$\|x_{n+1} - \bar{x}\|^2 - \|x_n - \bar{x}\|^2 + \|x_{n+1} - y_n\|^2 + \|y_n - x_n\|^2 \leq 2\alpha_n M. \tag{10.9}$$

Consider the sequence $(\|x_n - \bar{x}\|)$. We have two cases:

(a) there exists $\bar{n} \in \mathbb{N}$ such that

$$\|x_{n+1} - \bar{x}\| \leq \|x_n - \bar{x}\| \quad \forall\, n \geq \bar{n};$$

(b) there exists increasing sequence (n_k) such that

$$\|x_{n_k+1} - \bar{x}\| > \|x_{n_k} - \bar{x}\| \quad \forall\, k \in \mathbb{N}.$$

At first we consider the case (a). It follows that $\lim_{n\to\infty} \|x_n - \bar{x}\| = c \in \mathbb{R}$ exists. We assume $c > 0$ and we show that this latter hypothesis is impossible. Since $\|x_{n+1} - \bar{x}\|^2 - \|x_n - \bar{x}\|^2 \to 0$ and $\alpha_n \to 0$, we have

$$\|x_{n+1} - y_n\| \to 0, \quad \|y_n - x_n\| \to 0. \tag{10.10}$$

Using strong monotonicity of operator A, we get

$$(Ay_n, x_{n+1} - \bar{x}) \geq l\,\|y_n - \bar{x}\|^2 + (A\bar{x}, y_n - \bar{x}) + (Ay_n, x_{n+1} - y_n). \tag{10.11}$$

Since (y_n) is a bounded sequence and (10.10), it follows that

$$\lim_{n\to\infty} (Ay_n, x_{n+1} - y_n) = 0. \tag{10.12}$$

The sequence $((A\bar{x}, y_n - \bar{x}))$ is bounded. It is then immediate that there exists a subsequence (y_{n_k}) of (y_n) such that $y_{n_k} \rightharpoonup \bar{y} \in C$ and

$$\liminf_{n\to\infty} (A\bar{x}, y_n - \bar{x}) = \lim_{k\to\infty} (A\bar{x}, y_{n_k} - \bar{x}). \tag{10.13}$$

Applying Lemma 10.4, we have $\bar{y} \in EP(F, C)$. Therefore, in (10.13) we get

$$\liminf_{n\to\infty} (A\bar{x}, y_n - \bar{x}) = \lim_{k\to\infty} (A\bar{x}, y_{n_k} - \bar{x}) = (A\bar{x}, \bar{y} - \bar{x}) \geq 0. \tag{10.14}$$

Observing that $\|x_n - \bar{x}\| - \|y_n - x_n\| \leq \|y_n - \bar{x}\| \leq \|x_n - \bar{x}\| + \|y_n - x_n\|$, we immediately have

$$\lim_{n\to\infty} \|y_n - \bar{x}\| = c. \tag{10.15}$$

Therefore, by (10.11), (10.12), (10.14), and (10.15), we obtain $\liminf_{n\to\infty}(Ay_n, x_{n+1} - \bar{x}) \geq l \cdot c^2 > 0$. Take the real $\delta \in (0, l \cdot c^2)$. There exists some $n_* \in \mathbb{N}$ such that, for any $n \geq n_*$ there holds $(Ay_n, x_{n+1} - \bar{x}) \geq \delta$. Then, for any $n \geq n_*$ and taking into account (10.6), we deduce $\|x_{n+1} - \bar{x}\|^2 - \|x_n - \bar{x}\|^2 \leq -2\delta\alpha_n$, which leads to

$$\|x_{n+1} - \bar{x}\|^2 \leq \|x_{n_*} - \bar{x}\|^2 - 2\delta \sum_{i=n_*}^{n} \alpha_i.$$

Since $\sum_{n=1}^{\infty} \alpha_n = +\infty$, it follows that $\|x_n - \bar{x}\| \to -\infty$, which is absurd. As a sraightforward consequence, we deduce $c = 0$; namely, $x_n \to \bar{x}$ and $y_n \to \bar{x}$.

Now we consider the case (b). Let (m_k) be the integer sequence as in Lemma 10.2:

(i) $m_k \nearrow +\infty$;
(ii) $\|x_{m_k+1} - \bar{x}\| \geq \|x_{m_k} - \bar{x}\|$ $k \geq n_1$;
(iii) $\|x_{m_k+1} - \bar{x}\| \geq \|x_k - \bar{x}\|$ $k \geq n_1$.

From (10.9) and (ii), it follows that $\|x_{m_k+1} - y_{m_k}\|^2 + \|y_{m_k} - x_{m_k}\|^2 \leq 2\alpha_{m_k} M$. Hence,

$$\lim_{k\to\infty} \|x_{m_k+1} - y_{m_k}\| = 0, \quad \lim_{k\to\infty} \|y_{m_k} - x_{m_k}\| = 0.$$

Let us prove that the sequence (y_{m_k}) convergence strongly to $\bar{x}$ as $k \to \infty$. Since the sequence (y_{m_k}) is bounded, we see that there exists subsequence $(y_{m_{k_j}})$ which converges weakly to some point $\bar{y} \in C$. We have $\bar{y} \in EP(F, C)$. By (ii), (10.6), so that

$$(Ay_{m_k}, x_{m_k+1} - \bar{y}) \leq 0 \quad \forall k \geq n_1. \tag{10.16}$$

For $k \geq n_1$ and using strong monotonicity of operator A, we have

$$(Ay_{m_k} - A\bar{x}, y_{m_k} - \bar{x}) = (Ay_{m_k}, x_{m_k+1} - \bar{x}) + (Ay_{m_k}, y_{m_k} - x_{m_k+1})$$

$$-(A\bar{x}, y_{m_k} - \bar{x}) \geq l\|y_{m_k} - \bar{x}\|^2.$$

Taking into account (10.16), we obtain

$$\|y_{m_k} - \bar{x}\|^2 \leq \{(Ay_{m_k}, y_{m_k} - x_{m_k+1}) - (A\bar{x}, y_{m_k} - \bar{x})\}/l. \tag{10.17}$$

Observing that $\lim_{k\to\infty}(Ay_{m_k}, y_{m_k} - x_{m_k+1}) = 0$, $\lim_{j\to\infty}(A\bar{x}, y_{m_{k_j}} - \bar{x}) = (A\bar{x}, \bar{y} - \bar{x})$, by (10.17), we get $\limsup_{j\to\infty} \|y_{m_{k_j}} - \bar{x}\|^2 \leq -(A\bar{x}, \bar{y} - \bar{x})/l \leq 0$. Therefore, $y_{m_{k_j}} \to \bar{x}$. By the uniqueness of $\bar{x}$, and $\bar{y} = \bar{x}$, we deduce that $\lim_{k\to\infty} \|y_{m_k} - \bar{x}\| = 0$. From $\|x_{m_k+1} - \bar{x}\| \leq \|x_{m_k+1} - y_{m_k}\| + \|y_{m_k} - \bar{x}\|$ it follows that $\lim_{k\to\infty} \|x_{m_k+1} - \bar{x}\| = 0$. Taking into account (iii), we obtain

$$\lim_{n\to\infty} \|x_n - \bar{x}\| = 0. \tag{10.18}$$

From (10.18), and (10.9) it follows that $\lim_{n\to\infty} \|x_n - y_n\| = 0$. Hence,

$$\lim_{n\to\infty} \|y_n - \bar{x}\| = 0.$$

This completes the proof of Theorem 10.2.

Next, we prove strong convergence theorem related to Algorithm 2. At first, we get the boundedness of the sequences generated by Algorithm 2.

Lemma 10.7 *Assume that* $\lim_{n\to 0}\alpha_n = 0$. *Let* (x_n), (y_n), (z_n), *and* (v_n) *be sequences generated by Algorithm 2. Then* (x_n), (y_n), (z_n), *and* (v_n) *are bounded.*

Now, we provide useful estimate needed for proving convergence theorem related to Algorithm 2.

Lemma 10.8 *Let* $z \in EP(F_1, C_1) \cap EP(F_2, C_2)$, *and let* (x_n), (y_n), (z_n), *and* (v_n) *be sequences generated by Algorithm 2. Then, for each* $n \in \mathbb{N}$, *the following inequality holds*

$$\begin{aligned}\|x_{n+1}-z\|^2 \ &-\|x_n-z\|^2+\|x_{n+1}-v_n\|^2+\frac{1}{2}\|y_n-x_n\|^2+\frac{1}{2}\|z_n-x_n\|^2\\ &+\frac{1}{4}\|y_n-z_n\|^2 \le -2\alpha_n(Av_n, x_{n+1}-z). \end{aligned} \tag{10.19}$$

Proof Let $z \in EP(F_1, C_1) \cap EP(F_2, C_2)$. We have

$$\begin{aligned}\|x_{n+1}-z\|^2 &= \|v_n-\alpha_n Av_n-z\|^2 = \|v_n-z\|^2-2\alpha_n(Av_n, v_n-z)\\ &\quad+\alpha_n^2\|Av_n\|^2 = \|v_n-z\|^2-2\alpha_n(Av_n, x_{n+1}-z)-\|x_{n+1}-v_n\|^2. \end{aligned} \tag{10.20}$$

We have

$$\|v_n-z\|^2 = \left\|\frac{1}{2}y_n+\frac{1}{2}z_n-z\right\|^2 = \frac{1}{2}\|y_n-z\|^2+\frac{1}{2}\|z_n-z\|^2-\frac{1}{4}\|y_n-z_n\|^2.$$

From (10.3) it follows that

$$\|y_n-z\|^2 \le \|x_n-z\|^2-\|y_n-x_n\|^2, \quad \|z_n-z\|^2 \le \|x_n-z\|^2-\|z_n-x_n\|^2.$$

Therefore,

$$\|v_n-z\|^2 \le \|x_n-z\|^2-\frac{1}{2}\|y_n-x_n\|^2-\frac{1}{2}\|z_n-x_n\|^2-\frac{1}{4}\|y_n-z_n\|^2. \tag{10.21}$$

Substituting right side of (10.21) for $\|v_n-z\|^2$ in (10.20), we obtain the (10.19).

Now, we claim the main theorem for Algorithm 2.

Theorem 10.3 *Let hypotheses (B1)–(B7) hold. Assume that*

(i) $\alpha_n \in (0, +\infty)$, $\alpha_n \to 0$, $\sum_{n=1}^{\infty}\alpha_n = +\infty$;
(ii) $\lambda_n \in [\lambda, +\infty)$, *where* $\lambda > 0$.

Then the sequence (x_n) *generated by Algorithm 2 converges strongly to the unique* $\overline{x} \in VI(A, EP(F_1, C_1) \cap EP(F_2, C_2))$.

Proof Let $\overline{x} \in H$ be the unique solution to (10.2). From Lemma 10.7 it follows that there exists $M > 0$ such that $|(Av_n, x_{n+1} - \overline{x})| \leq M$ for any $n \in \mathbb{N}$. Using Lemma 10.8, we derive

$$\begin{aligned} \|x_{n+1} - \overline{x}\|^2 \ &- \|x_n - \overline{x}\|^2 + \|x_{n+1} - v_n\|^2 + \frac{1}{2}\|y_n - x_n\|^2 \\ &+ \frac{1}{2}\|z_n - x_n\|^2 + \frac{1}{4}\|y_n - z_n\|^2 \leq 2\alpha_n M. \end{aligned} \tag{10.22}$$

Consider the sequence $(\|x_n - \overline{x}\|)$. We have two cases:

(a) there exists number $\overline{n} \in \mathbb{N}$ such that

$$\|x_{n+1} - \overline{x}\| \leq \|x_n - \overline{x}\| \quad \forall n \geq \overline{n};$$

(b) there exists increasing sequence (n_k) such that

$$\|x_{n_k+1} - \overline{x}\| > \|x_{n_k} - \overline{x}\| \quad \forall k \in \mathbb{N}.$$

At first we consider the case (a). In this case, $\lim_{n\to\infty} \|x_n - \overline{x}\| = c \in \mathbb{R}$ exists. We assume $c > 0$. Since $\|x_{n+1} - \overline{x}\|^2 - \|x_n - \overline{x}\|^2 \to 0$ and $\alpha_n \to 0$, we have

$$\|x_{n+1} - v_n\|^2 \to 0, \tag{10.23}$$
$$\|y_n - x_n\|^2 \to 0, \tag{10.24}$$
$$\|z_n - x_n\|^2 \to 0, \tag{10.25}$$
$$\|y_n - z_n\|^2 \to 0. \tag{10.26}$$

Next, we make use of the fact that operator A is strongly monotone. We have

$$(Av_n, x_{n+1} - \overline{x}) \geq l\|v_n - \overline{x}\|^2 + (A\overline{x}, v_n - \overline{x}) + (Av_n, x_{n+1} - v_n). \tag{10.27}$$

Since (v_n) is a bounded sequence and (10.23), it follows that

$$\lim_{n\to\infty} (Av_n, x_{n+1} - v_n) = 0. \tag{10.28}$$

The sequence $((A\overline{x}, v_n - \overline{x}))$ is bounded. It is then immediate that there exists a subsequence (v_{n_k}) of (v_n) such that $v_{n_k} \rightharpoonup \overline{v} \in H$ and

$$\liminf_{n\to\infty} (A\overline{x}, v_n - \overline{x}) = \lim_{k\to\infty} (A\overline{x}, v_{n_k} - \overline{x}). \tag{10.29}$$

Taking into account (10.23)–(10.26) and applying Lemma 10.4, we obtain $\overline{v} \in EP(F_1, C_1) \cap EP(F_2, C_2)$. Therefore, for (10.29) we get

$$\liminf_{n\to\infty}(A\bar{x}, v_n - \bar{x}) = \lim_{k\to\infty}(A\bar{x}, v_{n_k} - \bar{x}) = (A\bar{x}, \bar{v} - \bar{x}) \geq 0. \tag{10.30}$$

Observing that

$$\|x_{n+1} - \bar{x}\| - \|v_n - x_{n+1}\| \leq \|v_n - \bar{x}\| \leq \|x_{n+1} - \bar{x}\| + \|v_n - x_{n+1}\|,$$

we get

$$\lim_{n\to\infty}\|v_n - \bar{x}\| = c. \tag{10.31}$$

Therefore, by (10.27), (10.28), (10.30), and (10.31), we obtain $\liminf_{n\to\infty}(Av_n, x_{n+1} - \bar{x}) \geq l \cdot c^2 > 0$. Take the real $\delta \in (0, l \cdot c^2)$. There exists some number $n_* \in \mathbb{N}$ such that, for any $n \geq n_*$ there holds $(Av_n, x_{n+1} - \bar{x}) \geq \delta$. Then, for any $n \geq n_*$ and taking into account (10.19), we deduce $\|x_{n+1} - \bar{x}\|^2 - \|x_n - \bar{x}\|^2 \leq -2\delta\alpha_n$. Therefore,

$$\|x_{n+1} - \bar{x}\|^2 \leq \left\|x_{n_*} - \bar{x}\right\|^2 - 2\delta\sum_{i=n_*}^{n}\alpha_i.$$

Since $\sum_{n=1}^{\infty}\alpha_n = +\infty$, it follows that $\|x_n - \bar{x}\| \to -\infty$, which is absurd. As a sraightforward consequence, we deduce $c = 0$, namely, $\lim_{n\to\infty}\|x_n - \bar{x}\| = 0$.

Now finally analyze the case (b). We apply Lemma 10.2 to obtain a sequence (m_k) of positive integers such that:

(i) $m_k \nearrow +\infty$;
(ii) $\left\|x_{m_k+1} - \bar{x}\right\| \geq \left\|x_{m_k} - \bar{x}\right\|$ $k \geq n_1$;
(iii) $\left\|x_{m_k+1} - \bar{x}\right\| \geq \|x_k - \bar{x}\|$ $k \geq n_1$.

From (10.22) and (ii), we get the following estimate:

$$\left\|x_{m_k+1} - v_{m_k}\right\|^2 + \frac{1}{2}\left\|y_{m_k} - x_{m_k}\right\|^2 + \frac{1}{2}\left\|z_{m_k} - x_{m_k}\right\|^2 + \frac{1}{4}\left\|y_{m_k} - z_{m_k}\right\|^2 \leq 2\alpha_{m_k}M.$$

Hence,

$$\lim_{k\to\infty}\left\|x_{m_k+1} - v_{m_k}\right\| = \lim_{k\to\infty}\left\|y_{m_k} - x_{m_k}\right\| = 0,$$

$$\lim_{k\to\infty}\left\|z_{m_k} - x_{m_k}\right\| = \lim_{k\to\infty}\left\|y_{m_k} - z_{m_k}\right\| = 0.$$

Let as show that the sequence (v_{m_k}) convergence strongly to $\bar{x}$. Since the sequence (v_{m_k}) is bounded, we see that there exists subsequence $(v_{m_{k_j}})$ which converges weakly to some point $\bar{v} \in H$. Since $v_{m_{k_j}} = y_{m_{k_j}} + \frac{1}{2}(z_{m_{k_j}} - y_{m_{k_j}}) = z_{m_{k_j}} + \frac{1}{2}(y_{m_{k_j}} - z_{m_{k_j}})$, we get $y_{m_{k_j}} \rightharpoonup \bar{v}$, $z_{m_{k_j}} \rightharpoonup \bar{v}$. Hence $\bar{v} \in EP(F_1, C_1) \cap EP(F_2, C_2)$. By (ii), (10.19), so that

$$(Av_{m_k}, x_{m_k+1} - \bar{v}) \leq 0 \quad \forall k \geq n_1. \tag{10.32}$$

For $k \geq n_1$ and using strong monotonicity of operator A, we get

$$\begin{aligned}(Av_{m_k} - A\overline{x}, v_{m_k} - \overline{x}) &= (Av_{m_k}, x_{m_k+1} - \overline{x}) + (Av_{m_k}, v_{m_k} - x_{m_k+1}) \\ &\quad - (A\overline{x}, v_{m_k} - \overline{x}) \geq l \left\| v_{m_k} - \overline{x} \right\|^2.\end{aligned}$$

Taking into account (10.32), we obtain

$$\left\| v_{m_k} - \overline{x} \right\|^2 \leq \left\{ (Av_{m_k}, v_{m_k} - x_{m_k+1}) - (A\overline{x}, v_{m_k} - \overline{x}) \right\} / l. \tag{10.33}$$

Observing that $\lim_{k\to\infty}(Av_{m_k}, v_{m_k} - x_{m_k+1}) = 0$, $\lim_{j\to\infty}(A\overline{x}, v_{m_{k_j}} - \overline{x}) = (A\overline{x}, \overline{v} - \overline{x})$, by (10.33), we get $\limsup_{j\to\infty} \|v_{m_{k_j}} - \overline{x}\|^2 \leq -(A\overline{x}, \overline{v} - \overline{x})/l \leq 0$. Therefore,

$$\lim_{j\to\infty} \left\| v_{m_{k_j}} - \overline{x} \right\| = 0.$$

By the uniqueness of $\overline{x}$, and $\overline{v} = \overline{x}$, we deduce that $\lim_{k\to\infty} \left\| v_{m_k} - \overline{x} \right\| = 0$. From $\left\| x_{m_k+1} - \overline{x} \right\| \leq \left\| x_{m_k+1} - v_{m_k} \right\| + \left\| v_{m_k} - \overline{x} \right\|$ it follows that $\lim_{k\to\infty} \left\| x_{m_k+1} - \overline{x} \right\| = 0$. Taking into account (iii), we obtain $\lim_{n\to\infty} \|x_n - \overline{x}\| = 0$. The proof is finished.

Now we present results for Alternating Algorithm 3. Proofs are omitted.

Lemma 10.9 *Let $z \in EP(F_1, C_1) \cap EP(F_2, C_2)$, and let (x_n), (y_n), and (z_n) be sequences generated by Algorithm 3. Then, for each $n \in \mathbb{N}$, the following inequality holds*

$$\begin{aligned}\|x_{n+1} - z\|^2 &- \|x_n - z\|^2 + \|x_{n+1} - z_n\|^2 + \|x_n - y_n\|^2 \\ &+ \|y_n - z_n\|^2 \leq -2\alpha_n (Az_n, x_{n+1} - z).\end{aligned}$$

Theorem 10.4 *Let hypotheses (B1)–(B7) hold. Assume that*

(i) $\alpha_n \in (0, +\infty)$, $\alpha_n \to 0$, $\sum_{n=1}^{\infty} \alpha_n = +\infty$;
(ii) $\lambda_n \in [\lambda, +\infty)$, *where* $\lambda > 0$.

Then the sequence (x_n) generated by Algorithm 3 converges strongly to the unique $\overline{x} \in VI(A, EP(F_1, C_1) \cap EP(F_2, C_2))$.

We now describe algorithms with computational errors and analyze its convergence. To solve the problem (10.1), we propose the following iterative procedure.

Algorithm 4 Select an arbitrary point $v_1 \in H$ and generates the sequence (v_n) iteratively by

$$\begin{cases} u_n = J_{\lambda_n F}(v_n + e_n), \\ v_{n+1} = u_n - \alpha_n A u_n, \end{cases}$$

where $\lambda_n, \alpha_n > 0$ and $e_n \in H$ is an error.

Now let us make a convergence analysis on Algorithm 4.

Theorem 10.5 *Let hypotheses (A1)–(A7) hold. Assume that*

(i) $\alpha_n \in (0, +\infty)$, $\alpha_n \to 0$ *with* $\sum_{n=1}^{\infty} \alpha_n = +\infty$;
(ii) $\lambda_n \in [\lambda, +\infty)$ *for some* $\lambda > 0$;
(iii) either $\sum_{n=1}^{\infty} \|e_n\| < +\infty$ *or* $\|e_n\|\alpha_n^{-1} \to 0$.

Then the sequences (v_n), (u_n) *generated by Algorithm 4 converge strongly to the unique* $\overline{x} \in VI(A, EP(F, C))$.

Proof Taking into account the Theorem 10.2, if we can show that $\|v_n - x_n\| \to 0$ as $n \to \infty$, then the sequence (v_n) strongly converges to the solution of (10.1). For sufficiently large number $n \in \mathbb{N}$ we have

$$\|v_{n+1} - x_{n+1}\| = \|(u_n - \alpha_n A u_n) - (y_n - \alpha_n A y_n)\| \leq \left(1 - \alpha_n \mu^{-1}\beta\right) \|u_n - y_n\|,$$

where $\mu \in \left(0, 2lL^{-2}\right)$, $\beta = 1 - \sqrt{1 - 2l\mu + L^2\mu^2} \in (0, 1)$. Since operators $J_{\lambda_n F}$ are nonexpansive, we get $\|v_{n+1} - x_{n+1}\| \leq \left(1 - \alpha_n \mu^{-1}\beta\right) \|v_n - x_n\| + \|e_n\|$. By Lemma 10.1, we obtain that $\|v_n - x_n\| \to 0$ as $n \to \infty$. From $\|u_n - y_n\| \leq \|v_n - x_n\| + \|e_n\|$ it follows that the sequence (u_n) converges strongly to the unique solution of (10.1).

Now we consider the variant of barycentric method with errors for solving problem (10.2).

Algorithm 5 *(Barycentric with errors)* Select an arbitrary point $p_1 \in H$ and generates the sequence (p_n) iteratively by

$$\begin{cases} u_n^1 = J_{\lambda_n F_1}\left(p_n + e_n^1\right), \\ u_n^2 = J_{\lambda_n F_2}\left(p_n + e_n^2\right), \\ w_n = \frac{1}{2}u_n^1 + \frac{1}{2}u_n^2, \\ p_{n+1} = w_n - \alpha_n A w_n, \end{cases}$$

where $\lambda_n, \alpha_n > 0$ and $e_n^1, e_n^2 \in H$ are computational errors.

We have the following theorem.

Theorem 10.6 *Let hypotheses (B1)–(B7) hold. Assume that*

(i) $\alpha_n \in (0, +\infty)$, $\alpha_n \to 0$ *with* $\sum_{n=1}^{\infty} \alpha_n = +\infty$;
(ii) $\lambda_n \in [\lambda, +\infty)$ *for some* $\lambda > 0$.

In addition, if any of the following conditions is satified,

(iii) $\sum_{n=1}^{\infty} \left\|e_n^1\right\| < +\infty$ *and* $\sum_{n=1}^{\infty} \left\|e_n^2\right\| < +\infty$;
(iv) $\sum_{n=1}^{\infty} \left\|e_n^1\right\| < +\infty$ *and* $\lim_{n\to\infty} \left\|e_n^2\right\|\alpha_n^{-1} = 0$;
(v) $\sum_{n=1}^{\infty} \left\|e_n^2\right\| < +\infty$ *and* $\lim_{n\to\infty} \left\|e_n^1\right\|\alpha_n^{-1} = 0$;
(vi) $\lim_{n\to\infty} \left\|e_n^1\right\|\alpha_n^{-1} = 0$ *and* $\lim_{n\to\infty} \left\|e_n^2\right\|\alpha_n^{-1} = 0$;

then the sequence (p_n) *generated by Algorithm 5 converges strongly to the unique* $\overline{x} \in VI(A, EP(F_1, C_1) \cap EP(F_2, C_2))$.

Proof In light of Theorem 10.4, it is enough to show that $\|p_n - x_n\| \to 0$ as $n \to \infty$, where sequence (x_n) generated by Algorithm 2. For sufficiently large number $n \in \mathbb{N}$ we have

$$\|p_{n+1} - x_{n+1}\| = \|(w_n - \alpha_n A w_n) - (v_n - \alpha_n A v_n)\| \leq \left(1 - \alpha_n \mu^{-1} \beta\right) \|w_n - v_n\|.$$

where $\mu \in \left(0, 2lL^{-2}\right)$, $\beta = 1 - \sqrt{1 - 2l\mu + L^2\mu^2} \in (0, 1)$. Further, we have

$$\begin{aligned}\|w_n - v_n\| = \left\| \frac{1}{2}(u_n^1 - y_n) + \frac{1}{2}(u_n^2 - z_n) \right\| &\leq \frac{1}{2} \left\| J_{\lambda_n F_1}\left(p_n + e_n^1\right) - J_{\lambda_n F_1} x_n \right\| \\ &+ \frac{1}{2} \left\| J_{\lambda_n F_2}\left(p_n + e_n^2\right) - J_{\lambda_n F_2} x_n \right\| \leq \|p_n - x_n\| + \frac{1}{2}\left\|e_n^1\right\| + \frac{1}{2}\left\|e_n^2\right\|.\end{aligned}$$

These two inequalities imply that

$$\|p_{n+1} - x_{n+1}\| \leq \left(1 - \beta\mu^{-1}\alpha_n\right) \|p_n - x_n\| + \frac{1}{2}\left\|e_n^1\right\| + \frac{1}{2}\left\|e_n^2\right\|.$$

By Lemma 10.1, we obtain desired fact.

Finally, we consider alternating method with errors.

Algorithm 6 *(Alternating with errors)* Select an arbitrary point $p_1 \in H$ and generates the sequence (p_n) iteratively by

$$\begin{cases} u_n = J_{\lambda_n F_1}\left(p_n + e_n^1\right), \\ v_n = J_{\lambda_n F_2}\left(u_n + e_n^2\right), \\ p_{n+1} = p_n - \alpha_n A p_n, \end{cases}$$

where $\lambda_n, \alpha_n > 0$ and $e_n^1, e_n^2 \in H$ are computational errors.

Theorem 10.7 *Let hypotheses (B1)–(B7) hold. Assume that*

(i) $\alpha_n \in (0, +\infty)$, $\alpha_n \to 0$ *with* $\sum_{n=1}^{\infty} \alpha_n = +\infty$;
(ii) $\lambda_n \in [\lambda, +\infty)$ *for some* $\lambda > 0$.

In addition, if any of the following conditions is satified,

(iii) $\sum_{n=1}^{\infty} \left\|e_n^1\right\| < +\infty$ *and* $\sum_{n=1}^{\infty} \left\|e_n^2\right\| < +\infty$;
(iv) $\sum_{n=1}^{\infty} \left\|e_n^1\right\| < +\infty$ *and* $\lim_{n\to\infty} \left\|e_n^2\right\| \alpha_n^{-1} = 0$;
(v) $\sum_{n=1}^{\infty} \left\|e_n^2\right\| < +\infty$ *and* $\lim_{n\to\infty} \left\|e_n^1\right\| \alpha_n^{-1} = 0$;
(vi) $\lim_{n\to\infty} \left\|e_n^1\right\| \alpha_n^{-1} = 0$ *and* $\lim_{n\to\infty} \left\|e_n^2\right\| \alpha_n^{-1} = 0$;

then the sequence (p_n) *generated by Algorithm 6 converges strongly to the unique solution* $\bar{x} \in H$ *of variational inequality* (10.2).

Proof This result is immediately deduced from Theorem 10.4 and Lemma 10.1.

10.4 Concluding Remarks

This chapter presented several iterative algorithms to solve the variational inequality problem over the solutions set of equilibrium problems. Using Mainge's techniques for analysis non–Fejerian iterative processes, strong convergence theorems for algorithms are proved. This methods have theoretical value mainly. From a practical point of view, main disadvantage of these methods is the calculation of resolvents (generally, hard computational problem). Motivated by the idea of Korpelevich's extragradient method [9] and modern extensions the extragradient method to equilibrium problem [19, 21], we'll introduce a new and maybe efficient method for solving problem (10.2). One version of this method is following.

Algorithm 7 *(Barycentric resolvents free)* Select an arbitrary point $x_1 \in H$ and generates the sequence (x_n) iteratively by

$$\begin{cases} v_n = \operatorname{prox}_{\lambda_n F_1(x_n,\cdot)} x_n, \quad y_n = \operatorname{prox}_{\lambda_n F_1(v_n,\cdot)} x_n, \\ u_n = \operatorname{prox}_{\lambda_n F_2(x_n,\cdot)} x_n, \quad z_n = \operatorname{prox}_{\lambda_n F_2(u_n,\cdot)} x_n, \\ w_n = \frac{1}{2} y_n + \frac{1}{2} z_n, \\ x_{n+1} = w_n - \alpha_n A w_n. \end{cases}$$

where $\lambda_n, \alpha_n > 0$.

This method use the Moreau's proximity operator [4]. Principal difference here, at each iteration, we solve strongly convex programming problems only instead of a auxiliary equilibrium programming problems. On the convergence of similar schemes, see [1, 19, 21]. This will be discussed in a further papers.

Acknowledgments The author has been partially supported by the Ukrainian State Fund for Fundamental Researches under grant GP/F44/042.

References

1. Apostol, R. Ya., Grynenko, A.A., Semenov, V.V.: Iterative algorithms for monotone bilevel variational inequalities. J. Num. Appl. Math. **1**(107), 3–14 (2012) (In Ukrainian).
2. Bauschke, H.H.: The approximation of fixed points of compositions of nonexpansive mappings in Hilbert space. J. Math. Anal. Appl. **202**, 150–159 (1996)
3. Bauschke, H.H., Combettes, P.L., Reich, S.: The asymptotic behavior of the composition of two resolvents. Nonlinear Anal. **60**, 283–301 (2005)
4. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, New York (2011)
5. Blum, E., Oettli, W.: From optimization and variational inequalities to equilibrium problems. Math. Stud. **63**, 123–145 (1994)
6. Combettes, P.L., Hirstoaga, S.A.: Equilibrium Programming in Hilbert Spaces. J. Nonlinear Convex Anal. **6**, 117–136 (2005)
7. Denisov, S.V., Semenov, V.V.: Proximal algorithm for bilevel variational inequalities: strong convergence. J. Num. Appl. Math. **3**(106), 27–32 (2011) (In Ukrainian).

8. Kopecka, E., Reich, S.: A note on the von Neumann alternating projections algorithm. J. Nonlinear Convex Anal. **5**, 379–386 (2004)
9. Korpelevich, G.M.: The extragradient method for finding saddle points and other problems. Ekonomika i Matematicheskie Metody. **12**, 747–756 (1976) (In Russian).
10. Lyashko, S.I., Semenov, V.V., Voitova, T.A.: Low-cost modification of Korpelevich's method for monotone equilibrium problems. Cybern. Syst. Anal. **47**, 631–639 (2011)
11. Mainge, P.-E.: Strong convergence of projected subgradient methods for nonsmooth and non-strictly convex minimization. Set-Valued Analysis. **16**, 899–912 (2008)
12. Malitskyi, I.V., Semenov, V.V.: The new theorems of strong convergence of proximal method for the problem of equilibrium programming. J. Num. Appl. Math. **3**(102), 79–88 (2010)
13. Moudafi, A., Mainge, P.-E.: Towards viscosity approximations of hierarchical fixed-point problems. Fixed Point Theory Appl. Article ID **95453**, 1–10 (2006)
14. von Neumann, J.: On rings of operators. Reduction theory. Annals Math. **50**, 401–485 (1949)
15. Nurminski, E.A.: The use of additional diminishing disturbances in Fejer models of iterative algorithms. Comput. Math. Math. Phys. **48**, 2154–2161 (2008)
16. Semenov, V.V.: About convergence of methods for solving bilevel variational inequalities with monotone operators. J. Num. Appl. Math. **2**(101), 121–129 (2010) (In Russian)
17. Semenov, V.V.: On the parallel proximal decomposition method for solving the problems of convex optimization. J. Autom. Inf. Sci. **42**(4), 13–18 (2010)
18. Takahashi, S., Takahashi, W.: Viscosity approximation methods for equilibrium problems and fixed point problems in Hilbert spaces. J. Math. Anal. Appl. **331**, 506–515 (2007)
19. Tran, D.Q., Muu, L.D., Nguyen, V.H.: Extragradient algorithms extended to solving equilibrium problems. Optimization. **57**(6), 749–776 (2008)
20. Voitova, T.A., Denisov, S.V., Semenov, V.V.: Alternating proximal algorithm for the problem of bilevel convex minimization. Reports of the National Academy of Sciences of Ukraine. No. 2, pp. 56–62 (2012) (In Ukrainian)
21. Voitova, T.A., Denisov, S.V., Semenov, V.V.: Strongly convergent modification of Korpelevich's method for equilibrium programming problems. J. Num. Appl. Math. **1**(104), 10–23 (2011) (In Ukrainian).
22. Voitova, T.A., Semenov, V.V.: Method for solving the bilevel operator inclusions. J. Num. Appl. Math. **3**(102), 34–39 (2010) (In Russian)
23. Xu, H.K.: Iterative algorithms for nonlinear operators. J. London Math. Soc. **2**, 240–256 (2002)
24. Yamada, I.: The hybrid steepest-descent method for the variational inequality problem of the intersection of fixed point sets of nonexpansive mappings. In: Butnariu, D., Censor, Y., Reich, S. (eds.) Inherently Parallel Algorithm for Feasibility and Optimization, pp. 473–604. Elsevier, New York (2001)

Part III
Long-time Forecasting in Multidisciplinary Investigations

Chapter 11
Multivalued Dynamics of Solutions for Autonomous Operator Differential Equations in Strongest Topologies

Mikhail Z. Zgurovsky and Pavlo O. Kasyanov

Abstract We consider nonlinear autonomous operator differential equations with pseudomonotone interaction functions satisfying (S)-property. The dynamics of all weak solutions on the positive time semi-axis is studied. We prove the existence of a trajectory and a global attractor in a strongest topologies and study their structure. As a possible application, we consider the class of high-order nonlinear parabolic equations.

11.1 Introduction: Statement of the Problem

In this chapter, we study the limiting behavior as time $t \to +\infty$ of the solutions of first-order general nonlinear evolution equations of the form

$$u'(t) + A(u(t)) = \bar{0}, \tag{11.1}$$

It is assumed that the nonlinear operator $A : V \to V^*$, acts in a Banach space V, which is reflexive and separable and, for some Hilbert space H, the embeddings $V \Subset H \equiv H \subset V^*$ are valid. Suppose that the nonlinear operator A is *pseudomonotone* and satisfies dissipation conditions of the form

$$\langle A(u), u\rangle_V \geq \alpha \|u\|_V^p - \beta \quad \forall u \in V, \tag{11.2}$$

where $p \geq 2$, and $\alpha, \beta > 0$, and also power growth conditions of the form

M. Z. Zgurovsky · P. O. Kasyanov (✉)
Institute for Applied System Analysis, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kyiv 03056, Ukraine
e-mail: kasyanov@i.ua

M. Z. Zgurovsky
e-mail: zgurovsm@hotmail.com

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_11,

$$\|A(u)\|_{V^*} \le c(1 + \|u\|_V^{p-1}) \quad \forall u \in V, \tag{11.3}$$

for some $c > 0$. Here $\langle \cdot, \cdot \rangle_V : V^* \times V \to \mathbb{R}$ is the pairing in $V^* \times V$ coinciding on $H \times V$ with the inner product $(\cdot, \cdot)$ in the Hilbert space H.

By a *weak solution* of the operator differential equation (11.1) on a closed interval $[\tau, T]$ we mean an element u of the space $L_p(\tau, T; V)$ such that

$$\forall \xi \in C_0^\infty([\tau, T]; V) \quad -\int_\tau^T (\xi'(t), u(t))dt + \int_\tau^T \langle A(u(t)), \xi(t) \rangle_V dt = 0. \tag{11.4}$$

Many evolution partial differential equations in a domain Ω whose leading part is a pth power nonlinear monotone differential operator and which may contain lower (now nonmonotone) summands with subordinate nonlinearity growth can be reduced to the form (11.1). In this case, the space V is a Sobolev space of the corresponding order, while the space H is $H = L_2(\Omega)$. Such equations are very often used to describe complicated evolution processes in various models in physics and mechanics. For equations of the form (11.1), there is a well-developed technique for constructing global (i.e., for all $t \ge 0$) weak solutions $u(t)$, $t \ge 0$, from the space $L_p^{loc}(\mathbb{R}_+; V)$ such that $u'(\cdot) \in L_q^{loc}(\mathbb{R}_+; V^*)$ (here $1/p + 1/q = 1$). It is well known that such weak solutions $u(t)$ are continuous functions with values in H, i.e., $u(\cdot) \in C(\mathbb{R}_+; H)$.

The problem is to study the asymptotic behavior as $t \to +\infty$ of the families of weak solutions $\{u(t)\}$ of Eq. (11.1) in the norm of H under the assumption that the initial data $\{u(0)\}$ constitute a bounded set in H.

Note that, under certain additional conditions on the nonlinear operator $A(u)$ ensuring, for Eq. (11.1), the unique solvability of the Cauchy problem $u|_{t=0} = u_0$ for any $u_0 \in H$, the study of the class of weak solutions under consideration involves the highly fruitful theory of dynamical semigroups and their global attractors in infinite-dimensional phase spaces. This theory has been successfully developed over a period of more than 30 years; its foundations were created by Ladyzhenskaya, Babin, Vishik, Hale, Temam and other well-known mathematicians [4, 11, 14–16, 21].

The problem becomes significantly more complicated if the corresponding Cauchy problem is not uniquely solvable or the proof of the relevant theorem is not known. Such a situation often occurs in complicated mathematical models. In this case, the "classical" method based on unique semigroups and global attractors cannot be applied directly. However, two approaches to the study of the dynamics of the corresponding weak solutions are well known. The first method is based on the theory of multivalued semigroups; it was developed in ground-breaking papers of Babin and Vishik (see, for example, [3]). The second approach uses the method of trajectory attractors; it was proposed in the papers [5, 7] of Chepyzhov and Vishik as well as in the independent work [20] of Sell.

The new result contained in the present chapter consists in the application of these two approaches to the study of the asymptotic behavior of the weak solutions of equations of the form (11.1) with general nonlinear pseudomonotone operator $A(u)$ satisfying (S)-property without any conditions guaranteeing the unique solvability of the Cauchy problem; cf. [1, 2, 4, 9, 14, 16, 19, 21, 23] . In this chapter, we prove a theorem on the existence of a global attractor $\mathscr{A}$ in the space H for the multivalued semigroup corresponding to Eq. (11.1) as well as a theorem on the existence of a trajectory attractor $\mathscr{P}$ in the space $C^{loc}(\mathbb{R}_+; H) \cap L_p^{loc}(\mathbb{R}_+; V)$ for the corresponding translation semigroup in the space of all weak trajectories (weak solutions on the half-line) of Eq. (11.1). We also describe the structure of the global and trajectory attractors as well as establish a simple relation between these attractors.

11.2 Additional Properties of Solutions

For fixed $\tau < T$ let us set

$$X_{\tau,T} = L_p(\tau, T; V), \quad X^*_{\tau,T} = L_q(\tau, T; V^*), \quad W_{\tau,T} = \{u \in X_{\tau,T} \mid u' \in X^*_{\tau,T}\},$$

where u' is a derivative of an element $u \in X_{\tau,T}$ in the sense of the space of distributions $\mathscr{D}([\tau, T]; V^*)$ (see, for example, [10, Definition IV.1.10, p. 168]). We note that

$$A(u)(t) = A(u(t)), \quad \text{for any } u \in X_{\tau,T} \text{ and a.e. } t \in (\tau, T).$$

The space $W_{\tau,T}$ is a reflexive Banach space with the graph norm of a derivative (see, for, example [24, Proposition 4.2.1, p. 291]):

$$\|u\|_{W_{\tau,T}} = \|u\|_{X_{\tau,T}} + \|u'\|_{X^*_{\tau,T}}, \quad u \in W_{\tau,T}. \tag{11.5}$$

Properties of A and (V, H, V^*) provide the existence of a weak solution of Cauchy problem (11.1) with initial data

$$u(\tau) = u_\tau \tag{11.6}$$

on the interval $[\tau, T]$ for an arbitrary $y_\tau \in H$. Therefore, the next result takes place:

Lemma 11.1 Kasyanov [12] *For any $\tau < T$, $y_\tau \in H$ Cauchy problem (11.1), (11.6) has a weak solution on the interval $[\tau, T]$. Moreover, each weak solution $u \in X_{\tau,T}$ of Cauchy problem (11.1), (11.6) on the interval $[\tau, T]$ belongs to $W_{\tau,T} \subset C([\tau, T]; H)$.*

For fixed $\tau < T$ we denote

$$\mathscr{D}_{\tau,T}(u_\tau) = \{u(\cdot) \mid u \text{ is a weak solution of (11.1) on } [\tau, T],\ u(\tau) = u_\tau\}, \quad u_\tau \in H.$$

From Lemma 11.1 it follows that $\mathscr{D}_{\tau,T}(u_\tau) \neq \emptyset$ and $\mathscr{D}_{\tau,T}(u_\tau) \subset W_{\tau,T}$ $\forall \tau < T$, $u_\tau \in H$.

We note that the translation and concatenation of weak solutions is a weak solution too.

Lemma 11.2 Kasyanov [12] *If* $\tau < T$, $u_\tau \in H$, $u(\cdot) \in \mathscr{D}_{\tau,T}(u_\tau)$, *then* $v(\cdot) = u(\cdot + s) \in \mathscr{D}_{\tau-s,T-s}(u_\tau)$ $\forall s$. *If* $\tau < t < T$, $u_\tau \in H$, $u(\cdot) \in \mathscr{D}_{\tau,t}(u_\tau)$ *and* $v(\cdot) \in \mathscr{D}_{t,T}(u(t))$, *then*

$$z(s) = \begin{cases} u(s), \ s \in [\tau, t], \\ v(s), \ s \in [t, T] \end{cases}$$

belongs to $\mathscr{D}_{\tau,T}(u_\tau)$.

As a rule, the proof of the existence of compact global and trajectory attractors for equations of type (11.1) is based on the properties of the set of weak solutions of problem (11.1) related to the absorption of the generated m-semiflow of solutions and its asymptotic compactness (see, for example, [18, 22] and the references therein). The following lemma on a priori estimates of solutions and Theorem 11.1 on the dependence of solutions on initial data will play a key role in the study of the dynamics of the solutions of problem (11.1) as $t \to +\infty$.

Lemma 11.3 Kasyanov [12] *There exist* $c_4, c_5, c_6, c_7 > 0$ *such that for any finite interval of time* $[\tau, T]$ *every weak solution* u *of problem* (11.1) *on* $[\tau, T]$ *satisfies estimates:* $\forall t \geq s$, $t, s \in [\tau, T]$

$$\|u(t)\|_H^2 + c_4 \int_s^t \|u(\xi)\|_V^p d\xi \leq \|u(s)\|_H^2 + c_5(t - s), \tag{11.7}$$

$$\|u(t)\|_H^2 \leq \|u(s)\|_H^2 e^{-c_6(t-s)} + c_7. \tag{11.8}$$

We recall that $A : V \to V^*$ satisfies (S)-*property*, if from $u_n \to u$ weakly in V and $\langle A(u_n), u_n - u\rangle_V \to 0$, as $n \to \infty$, it follows that $u_n \to u$ strongly in V, as $n \to +\infty$.

Further we assume that A satisfies (S)-property.

Theorem 11.1 *Let* $\tau < T$, $\{u_n\}_{n\geq 1}$ *be an arbitrary sequence of weak solutions of* (11.1) *on* $[\tau, T]$ *such that* $u_n(\tau) \to \eta$ *weakly in* H. *Then there exist* $\{u_{n_k}\}_{k\geq 1} \subset \{u_n\}_{n\geq 1}$ *and* $u(\cdot) \in \mathscr{D}_{\tau,T}(\eta)$ *such that*

$$\forall \varepsilon \in (0, T - \tau) \quad \max_{t\in[\tau+\varepsilon,T]} \|u_{n_k}(t) - u(t)\|_H + \int_{\tau+\varepsilon}^{T} \|u_{n_k}(t) - u(t)\|_V^p dt \to 0, \quad k \to +\infty. \tag{11.9}$$

Before the proof of Theorem 11.1 let us provide some auxiliary statements.

Lemma 11.4 *Let $\tau < T$, $y_n \to y$ weakly in $W_{\tau,T}$, and*

$$\overline{\lim_{n\to+\infty}} \langle A(y_n), y_n - y\rangle_{X_{\tau,T}} \le 0. \tag{11.10}$$

Then

$$\lim_{n\to+\infty} \int_\tau^T |\langle A(y_n(t)), y_n(t) - y(t)\rangle_V| \, dt = 0. \tag{11.11}$$

Proof There exists a set of measure zero, $\Sigma_1 \subset (\tau, T)$ such that for $t \notin \Sigma_1$, we have that

$$y_n(t) \in V \text{ for all } n \ge 1.$$

Similarly to [13, p. 7] we verify the following claim.

Claim Let $y_n \to y$ weakly in $W_{\tau,T}$ and let $t \notin \Sigma_1$. Then

$$\underline{\lim_{n\to+\infty}} \langle A(y_n(t)), y_n(t) - y(t)\rangle_V \ge 0.$$

Proof of the claim Fix $t \notin \Sigma_1$ and suppose to the contrary that

$$\underline{\lim_{n\to+\infty}} \langle A(y_n(t)), y_n(t) - y(t)\rangle_V < 0. \tag{11.12}$$

Then up to a subsequence $\{y_{n_k}\}_{k\ge1} \subset \{y_n\}_{n\ge1}$ we have

$$\lim_{k\to+\infty} \langle A(y_{n_k}(t)), y_{n_k}(t) - y(t)\rangle_V = \underline{\lim_{n\to+\infty}} \langle A(y_n(t)), y_n(t) - y(t)\rangle_V < 0. \tag{11.13}$$

Therefore, for all rather large k, growth and dissipation conditions imply

$$\alpha\|y_{n_k}(t)\|_V^p - \beta \le \|A(y_{n_k}(t))\|_{V^*}\|y(t)\|_V \le c(1 + \|y_{n_k}(t)\|_V^{p-1})\|y(t)\|_V.$$

which implies that the sequences $\{\|y_{n_k}(t)\|_V\}_{k\ge1}$ and consequently $\{\|A(y_{n_k}(t))\|_{V^*}\}_{k\ge1}$ are bounded sequences. In virtue of the continuous embedding $W_{\tau,T} \subset C([\tau, T]; H)$ we obtain that $y_{n_k}(t) \to y(t)$ weakly in H. Due to boundedness of $\{y_{n_k}(t)\}_{k\ge1}$ in V we finally have

$$\forall t \in [\tau, T]\backslash\Sigma_1 \; y_{n_k}(t) \to y(t) \text{ weakly in } V, \; k \to +\infty. \tag{11.14}$$

The pseudomonotony of A, (11.12), (11.13) and (11.14) imply that

$$\underline{\lim}_{n\to+\infty}\langle A(y_n(t)), y_n(t)-y(t)\rangle_V \geq \langle A(y(t)), y(t)-y(t)\rangle_V$$
$$= 0 > \underline{\lim}_{n\to+\infty}\langle A(y_n(t)), y_n(t)-y(t)\rangle_V.$$

We obtain a contradiction.

The claim is proved.

Now let us continue the proof of Lemma 11.4. The claim provides that for a.e. $t \in [\tau, T]$, in fact for any $t \notin \Sigma_1$, we have

$$\underline{\lim}_{n\to+\infty}\langle A(y_n(t)), y_n(t)-y(t)\rangle_V \geq 0. \tag{11.15}$$

Dissipation and growth conditions imply that, if $\omega \in X_{\tau,T}$, then

$$\langle A(y_n(t)), y_n(t)-\omega(t)\rangle_V \geq \alpha\|y_n(t)\|_V^p - \beta - c(1+\|y_n(t)\|_V^{p-1})\|\omega(t)\|_V \text{ for a.e. } t\in[\tau,T]\backslash\Sigma_1.$$

Using $p-1=\frac{p}{q}$, the right side of the above inequality equals to

$$\alpha\|y_n(t)\|_V^p - \beta - c\|y_n(t)\|_V^{\frac{p}{q}}\|\omega(t)\|_V - c\|\omega(t)\|_V.$$

Now using Young's inequality, we can obtain a constant $c(c,\alpha)$ depending on c,α such that

$$c\|y_n(t)\|_V^{\frac{p}{q}}\|\omega(t)\|_V \leq \frac{\alpha}{2}\|y_n(t)\|_V^p + \|\omega(t)\|_V^p \cdot c(c,\alpha).$$

Letting $\bar{c} = \max\{\beta + \frac{c}{q}; c(c,\alpha)+\frac{c}{p}\}$ it follows that

$$\langle A(y_n(t)), y_n(t)-\omega(t)\rangle_V \geq -\bar{c}(1+\|\omega(t)\|_V^p) \text{ for a.e. } t\in[\tau,T]. \tag{11.16}$$

Letting $\omega = y$, we can use Fatou's lemma and we obtain

$$\underline{\lim}_{n\to+\infty}\int_0^T [\langle A(y_n(t)), y_n(t)-y(t)\rangle_V + \bar{c}(1+\|y(t)\|_V^p)]dt \geq$$

$$\geq \int_0^T \underline{\lim}_{n\to+\infty}[\langle A(y_n(t)), y_n(t)-y(t)\rangle_V + \bar{c}(1+\|y(t)\|_V^p)]dt \geq \bar{c}\int_0^T (1+\|y(t)\|_V^p)dt.$$

Therefore,

$$0 \geq \overline{\lim}_{n\to+\infty}\langle A(y_n), y_n - y\rangle_{X_{\tau,T}} \geq \underline{\lim}_{n\to+\infty}\int_\tau^T \langle A(y_n(t)), y_n(t)-y(t)\rangle_V dt =$$

$$= \underline{\lim}_{n\to+\infty} \langle A(y_n), y_n - y\rangle_{X_{\tau,T}} \geq \int_\tau^T \underline{\lim}_{n\to+\infty} \langle A(y_n(t)), y_n(t) - y(t)\rangle_V dt = 0,$$

showing that

$$\lim_{n\to+\infty} \langle A(y_n), y_n - y\rangle_{X_{\tau,T}} = 0. \tag{11.17}$$

From (11.16),

$$\forall n \geq 1\ \forall t \notin \Sigma_1\ 0 \leq \langle A(y_n(t)), y_n(t) - y(t)\rangle_V^- \leq \bar{c}(1 + \|y(t)\|_V^p),$$

where $a^- = \max\{0, -a\}$, for $a \in \mathbb{R}$. Due to (11.15) we know that for a.e. t, $\langle A(y_n(t)), y_n(t) - y(t)\rangle_V \geq -\varepsilon$ for all rather large n. Therefore, for such n, $\langle A(y_n(t)), y_n(t) - y(t)\rangle_V^- \leq \varepsilon$, if $\langle A(y_n(t)), y_n(t) - y(t)\rangle_V < 0$ and $\langle A(y_n(t)), y_n(t) - y(t)\rangle_V^- = 0$, if $\langle A(y_n(t)), y_n(t) - y(t)\rangle_V \geq 0$. Therefore, $\lim_{n\to+\infty} \langle A(y_n(t)), y_n(t) - y(t)\rangle_V^- = 0$ and we can apply the dominated convergence theorem and from (11.15) we conclude that

$$\lim_{n\to+\infty} \int_\tau^T \langle A(y_n(t)), y_n(t) - y(t)\rangle_V^- = \int_\tau^T \lim_{n\to+\infty} \langle A(y_n(t)), y_n(t) - y(t)\rangle_V^- dt = 0.$$

Now by (11.17) and the above equation we have

$$\lim_{n\to+\infty} \int_\tau^T \langle A(y_n(t)), y_n(t) - y(t)\rangle_V^+ dt =$$

$$= \lim_{n\to+\infty} \int_0^T \left[\langle A(y_n(t)), y_n(t) - y(t)\rangle_V + \langle A(y_n(t)), y_n(t) - y(t)\rangle_V^-\right] dt =$$

$$= \lim_{n\to+\infty} \langle A(y_n), y_n - y\rangle_{X_{\tau,T}} = 0.$$

Therefore,

$$\lim_{n\to+\infty} \int_\tau^T |\langle A(y_n(t)), y_n(t) - y(t)\rangle_V| dt = 0.$$

The lemma is proved.

Lemma 11.5 *Let $\tau < T$, $y_n \to y$ weakly in $W_{\tau,T}$, and (11.10) holds. Then there exists a subsequence $\{y_{n_k}\}_{k\geq 1} \subset \{y_n\}_{n\geq 1}$ such that for a.e. $t \in (\tau, T)$ we have that $y_{n_k}(t) \to y(t)$ weakly in V, and $\langle A(y_{n_k}(t)), y_{n_k}(t) - y(t)\rangle_V \to 0$, $k \to +\infty$.*

Proof Let $y_n \to y$ weakly in $W_{\tau,T}$ and

$$\overline{\lim_{n\to+\infty}} \langle A(y_n), y_n - y\rangle_{X_{\tau,T}} \leq 0.$$

In virtue of Lemma 11.4 we obtain

$$\lim_{n\to+\infty} \int_\tau^T |\langle A(y_n(t)), y_n(t) - y(t)\rangle_V| dt = 0. \tag{11.18}$$

Due to the continuous embedding $W_{\tau,T} \subset C([\tau, T]; H)$ we have

$$\forall t \in [\tau, T]\ y_n(t) \to y(t) \text{ weakly in } H,\ n \to +\infty. \tag{11.19}$$

From (11.18) it follows that there exists a subsequence $\{y_{n_k}\}_{k\geq 1} \subset \{y_n\}_{n\geq 1}$ such that

$$\text{for a.e. } t \in [\tau, T]\ \langle A(y_{n_k}(t)), y_{n_k}(t) - y(t)\rangle_V \to 0,\ k \to +\infty.$$

Let $\Sigma_1 \subset [\tau, T]$ be a set of measure zero such that for $t \notin \Sigma_1$ $y_{n_k}(t)$, $y(t)$ are well-defined $\forall k \geq 1$, and

$$\langle A(y_{n_k}(t)), y_{n_k}(t) - y(t)\rangle_V \to 0,\ k \to +\infty.$$

In virtue of growth and dissipation conditions we obtain

$$\forall t \notin \Sigma_1\ \forall k \geq 1\ \overline{\lim_{k\to+\infty}} \left(\alpha \|y_{n_k}(t)\|_V^p - \beta - c(1 + \|y_{n_k}(t)\|_V^{p-1})\|y(t)\|_V\right) \leq 0.$$

Thus $\forall t \notin \Sigma_1$

$$\overline{\lim_{k\to+\infty}} \|y_{n_k}(t)\|_V^p \leq c(c, \alpha, \beta, p)(1 + \|y(t)\|_V^p).$$

Therefore, due to (11.19) we obtain that for a.e. $t \in (\tau, T)$ $y_{n_k}(t) \to y(t)$ weakly in V, $k \to +\infty$.

The lemma is proved.

Proof (Proof of Theorem 11.1) Let $\tau < T$, $\{u_n\}_{n\geq 1}$ be an arbitrary sequence of weak solutions of (11.1) on $[\tau, T]$ such that $u_n(\tau) \to \eta$ weakly in H. Theorem 1 from [12] implies the existence of a subsequence $\{u_{n_k}\}_{k\geq 1} \subset \{u_n\}_{n\geq 1}$ and $u(\cdot) \in \mathscr{D}_{\tau,T}(\eta)$ such that

$$\forall \varepsilon \in (0, T-\tau) \quad \max_{t\in[\tau+\varepsilon,T]} \|u_{n_k}(t) - u(t)\|_H \to 0, \quad k \to +\infty. \tag{11.20}$$

Let us prove that

$$\forall \varepsilon \in (0, T-\tau) \quad \int_{\tau+\varepsilon}^{T} \|u_{n_k}(t) - u(t)\|_V^p dt \to 0, \quad k \to +\infty. \tag{11.21}$$

On the contrary, without loss of generality we assume that for some $\varepsilon \in (0, T-\tau)$ and $\delta > 0$ it is fulfilled

$$\int_{\tau+\varepsilon}^{T} \|u_{n_k}(t) - u(t)\|_V^p dt \geq \delta, \quad \forall k \geq 1. \tag{11.22}$$

In virtue of (11.7), without loss of generality we claim that

$$u_{n_k} \to u \text{ weakly in } W_{\tau+\varepsilon,T}, \quad k \to +\infty. \tag{11.23}$$

Moreover, due to (11.20), we have

$$\overline{\lim_{k\to\infty}} \int_{\tau+\varepsilon}^{T} \langle A(u_{n_k}(t)), u_{n_k}(t) - u(t)\rangle_V dt \leq 0. \tag{11.24}$$

Thus, Lemma 11.5 and (S)-property for A imply that up to a subsequence which we denote again as $\{u_{n_k}\}_{k\geq 1}$ for a.e. $t \in (\tau+\varepsilon, T)$ we have that $u_{n_k}(t) \to u(t)$ strongly in V, $k \to +\infty$. Moreover, Lemma 11.4 provides that

$$\lim_{k\to+\infty} \int_{\tau+\varepsilon}^{T} \left|\langle A(u_{n_k}(t)), u_{n_k}(t) - u(t)\rangle_V\right| dt = 0.$$

Dissipation and growth conditions follow the existence a constant $C > 0$ such that

$$\|u_{n_k}(t) - u(t)\|_V^p \leq C(1 + \|u(t)\|_V^p + \left|\langle A(u_{n_k}(t)), u_{n_k}(t) - u(t)\rangle_V\right|)$$

for a.e. $t \in (\tau+\varepsilon, T)$ and any $k \geq 1$. Therefore,

$$\lim_{k\to+\infty} \int_{\tau+\varepsilon}^{T} \|u_{n_k}(t) - u(t)\|_V^p dt = 0.$$

We obtain a contradiction.

The theorem is proved.

11.3 Attractors in Strongest Topologies

First we consider constructions presented in [18]. Denote the set of all nonempty (nonempty bounded) subsets of H by $P(H)$ ($\mathscr{B}(H)$). We recall that the multivalued map $G : \mathbb{R} \times H \to P(H)$ is said to be a *m-semiflow* if:

(a) $G(0, \cdot) = \mathrm{Id}$ (the identity map),
(b) $G(t + s, x) \subset G(t, G(s, x))\ \forall x \in H, t, s \in \mathbb{R}_+$;

m-semiflow *is a strict* one if $G(t + s, x) = G(t, G(s, x))\ \forall x \in H, t, s \in \mathbb{R}_+$.

From Lemmas 11.2 and 11.3 it follows that any weak solution can be extended to a global one defined on $[0, +\infty)$. For an arbitrary $y_0 \in H$ let $\mathscr{D}(y_0)$ be the set of all weak solutions (defined on $[0, +\infty)$) of problem (11.1) with initial data $y(0) = y_0$.

We define the m-semiflow G as $G(t, y_0) = \{y(t) \mid y(\cdot) \in \mathscr{D}(y_0)\}$.

Lemma 11.6 Kasyanov [12] *G is the strict m-semiflow.*

We recall that the set $\mathscr{A}$ is said to be *a global attractor* G, if:

(1) $\mathscr{A}$ is negatively semiinvariant (i.e. $\mathscr{A} \subset G(t, \mathscr{A})\ \forall t \geq 0$);
(2) $\mathscr{A}$ is attracting, that is,

$$\mathrm{dist}(\mathrm{G(t,\,B)}, \mathscr{A}) \to 0, \quad \mathrm{t} \to +\infty \quad \forall \mathrm{B} \in \mathscr{B}(\mathrm{H}), \tag{11.25}$$

where $\mathrm{dist}(\mathrm{C,\,D}) = \sup\limits_{c \in C} \inf\limits_{d \in D} \|\mathrm{c} - \mathrm{d}\|_{\mathrm{H}}$ is the Hausdorff semidistance;
(3) For any closed set $Y \subset H$ satisfying (11.25), we have $\mathscr{A} \subset Y$ (minimality).

The global attractor is said to be *invariant* if $\mathscr{A} = G(t, \mathscr{A})\ \forall t \geq 0$.

Theorem 11.2 Kasyanov [12] *The m-semiflow G has the invariant compact in the phase space H global attractor $\mathscr{A}$.*

Let us consider the family $\mathscr{K}_+ = \cup_{y_0 \in H} \mathscr{D}(y_0)$ of all weak solutions of inclusion (11.1) defined on the semi-infinite interval $[0, +\infty)$. Note that $\mathscr{K}_+$ is *translation invariant* one, i.e. $\forall u(\cdot) \in \mathscr{K}_+, \forall h \geq 0\ u_h(\cdot) \in \mathscr{K}_+$, where $u_h(s) = u(h + s)$, $s \geq 0$. We set the *translation semigroup* $\{T(h)\}_{h \geq 0}$, $T(h)u(\cdot) = u_h(\cdot)$, $h \geq 0$, $u \in \mathscr{K}_+$ on $\mathscr{K}_+$.

We shall construct the attractor of the translation semigroup $\{T(h)\}_{h \geq 0}$ acting on $\mathscr{K}_+$. On $\mathscr{K}_+$ we consider a topology induced from the Fréchet space $C^{loc}(\mathbb{R}_+; H) \cap L_p^{loc}(\mathbb{R}_+; V)$. Note that Π_M is the restriction operator to the interval $[0, M]$. We denote the restriction operator to the semi-infinite interval $[0, +\infty)$ by Π_+.

We recall that the a $\mathscr{P} \subset C^{loc}(\mathbb{R}_+; H) \cap L_\infty(\mathbb{R}_+; H)$ is said to be *attracting* for the trajectory space $\mathscr{K}_+$ of inclusion (11.1) in the topology of $C^{loc}(\mathbb{R}_+; H)$ if for any bounded in $L_\infty(\mathbb{R}_+; H)$ set $\mathscr{B} \subset \mathscr{K}_+$ and any number $M \geq 0$ the following relation holds:

$$\mathrm{dist}_{C([0,M];H)}(\Pi_M T(t)\mathscr{B}, \Pi_M \mathscr{P}) \to 0, \quad t \to +\infty. \tag{11.26}$$

A set $\mathscr{U} \subset \mathscr{K}_+$ is said to be *trajectory attractor* in the trajectory space $\mathscr{K}_+$ with respect to the topology of $C^{loc}(\mathbb{R}_+; H) \cap L_p^{loc}(\mathbb{R}_+; V)$ if

(i) $\mathscr{U}$ is a compact set in $C^{loc}(\mathbb{R}_+; H) \cap L_p^{loc}(\mathbb{R}_+; V)$ and bounded in $L_\infty(\mathbb{R}_+; H)$;
(ii) $\mathscr{U}$ is strictly invariant with respect to $\{T(h)\}_{h\geq 0}$, i.e. $T(h)\mathscr{U} = \mathscr{U}$ $\forall h \geq 0$;
(iii) $\mathscr{U}$ is an attracting set in the trajectory space $\mathscr{K}_+$ in the topology $C^{loc}(\mathbb{R}_+; H) \cap L_p^{loc}(\mathbb{R}_+; V)$.

Let us consider inclusions (11.1) on the entire time axis. Similarly to the space $C^{loc}(\mathbb{R}_+; H)$ the space $C^{loc}(\mathbb{R}; H)$ is equipped with the topology of local uniform convergence on each interval $[-M, M] \subset \mathbb{R}$ (see, for example, [22, p. 198]). A function $u \in C^{loc}(\mathbb{R}; H) \cap L_\infty(\mathbb{R}; H)$ is called a *complete trajectory* of inclusion (11.1) if $\forall h \in \mathbb{R}$ $\Pi_+ u_h(\cdot) \in \mathscr{K}_+$ [22, p. 198]. Let $\mathscr{K}$ be a family all complete trajectories of inclusion (11.1). Note that

$$\forall h \in \mathbb{R},\ \forall u(\cdot) \in \mathscr{K} \quad u_h(\cdot) \in \mathscr{K}. \tag{11.27}$$

Lemma 11.7 *The set $\mathscr{K}$ is nonempty, compact in $C^{loc}(\mathbb{R}; H) \cap L_p^{loc}(\mathbb{R}_+; V)$ and bounded in $L_\infty(\mathbb{R}; H)$. Moreover,*

$$\forall y(\cdot) \in \mathscr{K},\ \forall t \in \mathbb{R} \quad y(t) \in \mathscr{A}, \tag{11.28}$$

where $\mathscr{A}$ is the global attractor from Theorem 11.2.

Proof The statement of lemma follows from [12] and Theorem 11.1.

Lemma 11.8 Kasyanov [12] *Let $\mathscr{A}$ be a global attractor from Theorem 11.2. Then*

$$\forall y_0 \in \mathscr{A} \quad \exists y(\cdot) \in \mathscr{K} : \quad y(0) = y_0. \tag{11.29}$$

Theorem 11.3 *Let $\mathscr{A}$ be a global attractor from Theorem 11.2. Then there exists the trajectory attractor $\mathscr{P} \subset \mathscr{K}_+$ in the space $\mathscr{K}_+$. At that the next formula holds:*

$$\mathscr{P} = \Pi_+ \mathscr{K} = \Pi_+ \{y \in \mathscr{K} \mid y(t) \in \mathscr{A}\ \forall t \in \mathbb{R}\}, \tag{11.30}$$

Proof The statement of theorem follows from Theorem 11.1 and [12].

11.4 Application

Consider an example of the class of nonlinear boundary value problems for which we can investigate the dynamics of solutions as $t \to +\infty$. Note that in discussion we do not claim generality.

Let $n \geq 2, m \geq 1, p \geq 2, 1 < q \leq 2, \frac{1}{p} + \frac{1}{q} = 1, \Omega \subset \mathbb{R}^n$ be a bounded domain with rather smooth boundary $\Gamma = \partial\Omega$. We denote a number of differentiations by x of order $\leq m-1$ (correspondingly of order $= m$) by N_1 (correspondingly by N_2). Let $A_\alpha(x, \eta; \xi)$ be a family of real functions ($|\alpha| \leq m$), defined in $\Omega \times \mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$ and satisfying the next properties:

(C_1) for a.e. $x \in \Omega$ the function $(\eta, \xi) \to A_\alpha(x, \eta, \xi)$ is continuous one in $\mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$;

(C_2) $\forall (\eta, \xi) \in \mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$ the function $x \to A_\alpha(x, \eta, \xi)$ is measurable one in Ω;

(C_3) exist such $c_1 \geq 0$ and $k_1 \in L_q(\Omega)$, that for a.e. $x \in \Omega, \forall (\eta, \xi) \in \mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$

$$|A_\alpha(x, \eta, \xi)| \leq c_1[|\eta|^{p-1} + |\xi|^{p-1} + k_1(x)];$$

(C_4) exist such $c_2 > 0$ and $k_2 \in L_1(\Omega)$, that for a.e. $x \in \Omega, \forall (\eta, \xi) \in \mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$

$$\sum_{|\alpha|=m} A_\alpha(x, \eta, \xi)\xi_\alpha \geq c_2|\xi|^p - k_2(x);$$

(C_5) there exists increasing function $\varphi : \mathbb{R}_+ \to \mathbb{R}$ such that for a.e. $x \in \Omega$, $\forall \eta \in \mathbb{R}^{N_1}, \forall \xi, \xi^* \in \mathbb{R}^{N_2}, \xi \neq \xi^*$ the inequality

$$\sum_{|\alpha|=m} (A_\alpha(x, \eta, \xi) - A_\alpha(x, \eta, \xi^*))(\xi_\alpha - \xi^*_\alpha) \geq (\varphi(|\xi_\alpha|) - \varphi(|\xi^*_\alpha|))(|\xi_\alpha| - |\xi^*_\alpha|)$$

takes place.

Consider such denotations: $D^k u = \{D^\beta u, |\beta| = k\}$, $\delta u = \{u, Du, ..., D^{m-1}u\}$ (see [17, p. 194]).

For an arbitrary fixed interior force $f \in L_2(\Omega)$ we investigate the dynamics of all weak (generalized) solutions defined on $[0, +\infty)$ of such problem:

$$\frac{\partial y(x,t)}{\partial t} + \sum_{|\alpha|\leq m} (-1)^{|\alpha|} D^\alpha \left(A_\alpha(x, \delta y(x,t), D^m y(x,t))\right) = f(x) \text{ on } \Omega \times (0, +\infty), \tag{11.31}$$

$$D^\alpha y(x,t) = 0 \text{ on } \Gamma \times (0, +\infty), \quad |\alpha| \leq m-1 \tag{11.32}$$

as $t \to +\infty$.

Consider such denotations (see for detail [17, p. 195]): $H = L_2(\Omega)$, $V = W_0^{m,p}(\Omega)$ is a real Sobolev space,

$$a(u,\omega)=\sum_{|\alpha|\le m}\int\limits_{\Omega}A_\alpha(x,\delta u(x),D^m u(x))D^\alpha\omega(x)dx,\quad u,\omega\in V.$$

Note that the operator $A:V\to V^*$, defined by the formula $\langle A(u),\omega\rangle_V=a(u,\omega)$ $\forall u,\omega\in V$, satisfies all mentioned assumptions. Hence, we can pass from problem (11.31)–(11.32) to corresponding problem in "generalized" setting (11.1). Note that

$$A(u)=\sum_{|\alpha|\le m}(-1)^{|\alpha|}D^\alpha\left(A_\alpha(x,\delta u,D^m u)\right)\quad\forall u\in C_0^\infty(\Omega).$$

Therefore, all statements from previous sections are fulfilled for weak (generalized) solutions of problem (11.31)–(11.32).

Remark 11.1 As applications we can also consider new classes of problems with degenerations, problems on a manifold, problems with delay, stochastic partial differential equations etc. [6, 8, 17, 20] with differential operators of pseudomonotone type as corresponding choice of the phase space.

11.5 Conclusions

For the class of autonomous differential operator equations with pseudomonotone nonlinear dependence between the defining parameters of the problem, we have studied the dynamics as $t\to+\infty$ in strongest topologies of all global weak solutions defined on $[0,+\infty)$. We have proved the existence of a global compact attractor and a compact trajectory attractor, studied their structure. The results obtained allow one to study the dynamics of solutions for new classes of evolution equations related to nonlinear mathematical models of geophysical and socioeconomic processes and for fields with interaction functions of pseudomonotone type satisfying the (S)-property, the condition of "power growth", and the standard sign condition.

Acknowledgments This work was partially supported by the Ukrainian State Fund for Fundamental Researches under grants GP/F44/076, GP/F49/070, and by the NAS of Ukraine under grant 2273/13.

References

1. Aubin, J.-P., Cellina, A.: Differential inclusions. Set-valued maps and viability theory. Grundlehren Math. Wiss. **264**, 364 p. (1984)
2. Aubin, J.-P., Frankowska, H.: Set-valued analysis. Syst. Control Found. Appl. **2**, 484 p. (1990) (Birkhauser Boston, Boston)
3. Babin, A.V., Vishik, M.I.: Maximal attractors of semigroups corresponding to evolution differential equations. Math. USSR-Sb. **54**(2), 387–408 (1986)

4. Babin, A.V., Vishik, M.I.: Attractors of Evolution Equations [in Russian]. Nauka, Moscow (1989)
5. Chepyzhov V.V., Vishik, M.I.: Evolution equations and their trajectory attractors. J. Math. Pures Appl. **76**(10), 913–964 (1997)
6. Chepyzhov, V.V., Vishik, M.I.: Trajectory attractor for reaction-diffusion system with diffusion coeffcient vanishing in time. Discrete Contin. Dyn. Syst. **27**(4), 1493–1509 (2010)
7. Chepyzhov, V.V., Vishik, M.I.: Trajectory attractors for evolution equations. C. R. Acad. Sci. Paris Ser. I Math. **321**(10), 1309–1314 (1995)
8. Dubinskii, Yu.A.: Higher-order nonlinear parabolic equations. J. Soviet Math. 5**6**(4), 2557–2607 (1991)
9. Eden, A., Foias, C., Nicolaenko, B., Temam, R.: Exponential attractors for dissipative evolution equations. RAM Res. Appl. Math. **37**, 182 p. (1994) (John Wiley & Sons, Chichester)
10. Gajewski, H., Groger, K., Zacharias, K.: Nichtlineare Operatorgleichungen und Operatordifferential-gleichungen. Akademie-Verlag, Berlin (1974)
11. Hale, J.K.: Asymptotic behavior of dissipative systems. Math. Surveys and Monogr. **25**, 198 p. (1988)
12. Kasyanov, P.O.: Multivalued dynamics of solutions of autonomous operator differential equations with pseudomonotone nonlinearity. Math. Notes **92**(2), 57–70 (2012)
13. Kuttler, K.: Non-degenerate implicit evolution inclusions. Electron. J. Differ. Equ. **34**, 1–20 (2000)
14. Ladyzhenskaya, O.: Attractors for semigroups and evolution equations. In: Lezioni Lincee. Cambridge University Press, Cambridge (1991)
15. Ladyzhenskaya, O.A.: The dynamical system that is generated by the Navier-Stokes equations [in Russian]. Zap. Naucn. Sem. Leningrad. Otdel. Mat. Inst. Steklov **27**(6), 91–115 (1972) (Nauka, Leningrad)
16. Ladyzhenskaya, O.A.: The infinite-dimensionality of bounded invariant sets for the Navier-Stokes system and other dissipative systems [in Russian]. Zap. Naucn. Sem. Leningrad. Otdel. Mat. Inst. Steklov **115**(6), 137–155 (1982) (Nauka, Leningrad)
17. Lions, J.-L.: Quelques Methodes de Resolution des Problemes aux Limites Nonlineaires. Dunod, Paris (1969)
18. Melnik, V.S., Valero, J.: On attractors of multivalued semi-flows and generalized differential equations. Set-Valued Anal. **6**(1), 83–111 (1998)
19. Migorski, S.: Boundary hemivariational inequalities of hyperbolic type and applications. J. Global Optim. **31**(3), 505–533 (2005)
20. Sell, G.R.: Global attractors for the three-dimensional NavierStokes equations. J. Dyn. Differ. Equ. **8**(1), 1–33 (1996)
21. Temam, R.: Infinite-dimensional dynamical systems in mechanics and physics. Appl. Math. Sci. **68**, 648 p. (1988)
22. Vishik, M.I., Chepyzhov, V.V.: Trajectory and global attractors of the three-dimensional Navier-Stokes system. Mat. Zametki. **71**(2), 194–213 (2002)
23. Vishik, M.I., Zelik, S.V., Chepyzhov, V.V.: Strong trajectory attractor for a dissipative reaction-diffusion system. Dokl. Ross. Akad. Nauk. **435**(2), 155–159 (2010)
24. Zgurovsky, M.Z., Kasyanov, P.O., Melnik, V.S.: Operator Differential Inclusions and Variational Inequalities in Infinite-dimensional Spaces [in Russian]. Naukova Dumka, Kiev (2008)

Chapter 12
Structure of Uniform Global Attractor for General Non-Autonomous Reaction-Diffusion System

Oleksiy V. Kapustyan, Pavlo O. Kasyanov, José Valero and Mikhail Z. Zgurovsky

Abstract In this paper we study structural properties of the uniform global attractor for non-autonomous reaction-diffusion system in which uniqueness of Cauchy problem is not guarantied. In the case of translation compact time-depended coefficients we prove that the uniform global attractor consists of bounded complete trajectories of corresponding multi-valued processes. Under additional sign conditions on non-linear term we also prove (and essentially use previous result) that the uniform global attractor is, in fact, bounded set in $L^\infty(\Omega) \cap H_0^1(\Omega)$.

12.1 Introduction

In this paper we study the structural properties of the uniform global attractor of non-autonomous reaction-diffusion system in which the nonlinear term satisfy suitable growth and dissipative conditions on the phase variable, suitable translation compact conditions on time variable, but there is no condition ensuring uniqueness of Cauchy problem. In autonomous case such system generates in the general case

O. V. Kapustyan (✉)
Taras Shevchenko National University of Kyiv, 64, Volodymyrs'ka St, Kyiv 01601, Ukraine
e-mail: alexkap@univ.kiev.ua

P. O. Kasyanov · M. Z. Zgurovsky
Institute for Applied System Analysis, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kyiv 03056, Ukraine
e-mail: kasyanov@i.ua

M. Z. Zgurovsky
e-mail: zgurovsm@hotmail.com

J. Valero
Universidad Miguel Hernandez de Elche, Centro de Investigación Operativa, Avda. Universidad s/n, 03202 Elche, Spain
e-mail: jvalero@umh.es

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_12,

a multi-valued semiflow having a global compact attractor (see [1–5]). Also, it is known [1, 2, 6], that the attractor is the union of all bounded complete trajectories of the semiflow. Here we prove the same result for non-autonomous system. More precisely, we prove that the family of multi-valued processes, generated by weak solutions of reaction-diffusion system, has uniform global attractor which is union of bounded complete trajectories of corresponding processes. Using this result, we can prove that under additional restrictions on nonlinear term obtained uniform global attractor is bounded set in the space $L^\infty(\Omega) \cap H_0^1(\Omega)$.

12.2 Setting of the Problem

In a bounded domain $\Omega \subset \mathbb{R}^n$ with sufficiently smooth boundary $\partial\Omega$ we consider the following non-autonomous parabolic problem (named RD-system) [7–17]

$$\begin{cases} u_t = a\Delta u - f(t,u) + h(t,x), \quad x \in \Omega,\ t > \tau, \\ u|_{\partial\Omega} = 0, \end{cases} \tag{12.1}$$

where $\tau \in \mathbb{R}$ is initial moment of time, $u = u(t,x) = (u^1(t,x), ..., u^N(t,x))$ is unknown vector-function, $f = (f^1, ..., f^N)$, $h = (h^1, ..., h^N)$ are given functions, a is real $N \times N$ matrix with positive symmetric part $\frac{1}{2}(a + a^*) \geq \beta I$, $\beta > 0$,

$$h \in L^2_{loc}(\mathbb{R}; (L^2(\Omega))^N), \quad f \in C(\mathbb{R} \times \mathbb{R}^N; \mathbb{R}^N), \tag{12.2}$$

$\exists\, C_1, C_2 > 0, \quad \gamma_i > 0,\ p_i \geq 2,\ i = \overline{1,N}$ such that $\forall\, t \in \mathbb{R}\ \forall v \in \mathbb{R}^N$

$$\sum_{i=1}^N |f^i(t,v)|^{\frac{p_i}{p_i-1}} \leq C_1(1 + \sum_{i=1}^N |v^i|^{p_i}), \tag{12.3}$$

$$\sum_{i=1}^N f^i(t,v)v^i \geq \sum_{i=1}^N \gamma_i |v^i|^{p_i} - C_2. \tag{12.4}$$

In further arguments we will use standard functional spaces

$$H = (L^2(\Omega))^N \text{ with the norm } |v|^2 = \int_\Omega \sum_{i=1}^N |v^i(x)|^2 dx,$$

$$V = (H_0^1(\Omega))^N \text{ with the norm } \|v\|^2 = \int_\Omega \sum_{i=1}^N |\nabla v^i(x)|^2 dx.$$

Let us denote $V' = (H^{-1}(\Omega))^N, q_i = \frac{p_i}{p_i - 1}, p = (p_1, ..., p_N), q = (q_1, ..., q_N)$, $L^p(\Omega) = L^{p_1}(\Omega) \times \cdots \times L^{p_N}(\Omega)$.

Definition 12.1 The function $u = u(t,x) \in L^2_{loc}(\tau, +\infty; V) \bigcap L^p_{loc}(\tau, +\infty; L^p(\Omega))$ is called a (weak) solution of the problem (12.1) on $(\tau, +\infty)$ if for all $T > \tau,\ v \in V \cap L^p(\Omega)$

$$\frac{d}{dt}\int\limits_{\Omega} u(t,x)v(x)dx + \int\limits_{\Omega}\Big(a\nabla u(t,x)\nabla v(x) + f(t,u(t,x))v(x) - h(t,x)v(x)\Big)dx = 0 \tag{12.5}$$

in the sense of scalar distributions on (τ, T).

From (12.3) and Sobolev embedding theorem we see that every solution of (12.1) satisfies inclusion $u_t \in L^q_{loc}(\tau, +\infty; H^{-r}(\Omega))$, where $r = (r_1, ..., r_N), r_k = max\{1, n(\frac{1}{2} - \frac{1}{p_k})\}$. The following theorem is well-known result about global resolvability of (12.1) for initial conditions from the phase space H.

Theorem 12.1 [18, Theorem 2] *or* [8, p.284]. *Under conditions* (12.3), (12.4) *for every* $\tau \in \mathbb{R}, u_\tau \in H$ *there exists at least one weak solution of* (12.1) *on* $(\tau, +\infty)$ *with* $u(\tau) = u_\tau$ *(and it may be non unique) and any weak solution of* (12.1) *belongs to* $C([\tau, +\infty); H)$. *Moreover, the function* $t \mapsto |u(t)|^2$ *is absolutely continuous and for a.a.* $t \geq \tau$ *the following energy equality holds*

$$\frac{1}{2}\frac{d}{dt}|u(t)|^2 + (a\nabla u(t), \nabla u(t)) + (f(t,u(t)), u(t)) = (h(t), u(t)). \tag{12.6}$$

Under additional not-restrictive conditions on function f and h it is known that solution of (12.1) generate non-autonomous dynamical system (two-parametric family of m-processes), which has uniform global attractor. The aim of this paper is to give description of the attractor in terms of bounded complete trajectories and show some regularity property of this set.

12.3 Multi-Valued Processes and Uniform Attractors

Let (X, ρ) be a complete metric space. The Hausdorff semidistance from A to B is given by

$$dist(A, B) = \sup_{x \in A} \inf_{y \in B} \rho(x, y),$$

By $\overline{A}$ and $O_\varepsilon(A) = \{x \in X \mid \inf\limits_{y \in A} \rho(x, y) < \varepsilon\}$we denote closure and ε-neighborhood of the set A. Denote by $P(X)$ $(\beta(X), C(X), K(X))$ the set of all non-empty (not-empty bounded, not-empty closed, not-empty compact) subsets of X,

$$\mathbb{R}_d = \{(t, \tau) \in \mathbb{R}^2 | t \geq \tau\}.$$

Let Σ be some complete metric space $\{T(h)\colon \Sigma \mapsto \Sigma\}_{h\geq 0}$ be a continuous semigroup, acting on Σ. Note, that in most applications $T(h)$ is shift semigroup.

Definition 12.2 Two-parameter family of multi-valued mappings $\{U_\sigma : \mathbb{R}_d \times X \mapsto P(X)\}_{\sigma\in\Sigma}$ is said to be the family of m-processes (family of MP), if $\forall\, \sigma \in \Sigma$, $\tau \in \mathbb{R}$:
(1) $U_\sigma(\tau, \tau, x) = x\ \forall x \in X$,
(2) $U_\sigma(t, \tau, x) \subseteq U_\sigma(t, s, U_\sigma(s, \tau, x)), \forall t \geq s \geq \tau\ \ \forall x \in X$,
(3) $U_\sigma(t + h, \tau + h, x) \subseteq U_{T(h)\sigma}(t, \tau, x)\ \forall t \geq \tau\ \ \forall h \geq 0,\ \forall\, x \in X$,
where for $A \subset X, B \subset \Sigma$ $U_B(t, s, A) = \bigcup\limits_{\sigma\in B}\bigcup\limits_{x\in A} U_\sigma(t, s, x)$, in particular

$$U_\Sigma(t, \tau, x) = \bigcup_{\sigma\in\Sigma} U_\sigma(t, \tau, x).$$

Family of MP $\{U_\sigma|\sigma \in \Sigma\}$ is called strict, if in conditions (2), (3) equality take place.

Definition 12.3 A set $A \subset X$ is called uniformly attracting for the family of MP $\{U_\sigma|\sigma \in \Sigma\}$, if for arbitrary $\tau \in \mathbb{R}$, $B \in \beta(X)$

$$dist(U_\Sigma(t, \tau, B), A) \to 0,\ t \to +\infty, \tag{12.7}$$

that is $\forall\varepsilon > 0$, $\tau \in \mathbb{R}$ and $B \in \beta(X)$ there exists $T = T(\tau, \varepsilon, B)$ such that

$$U_\Sigma(t, \tau, B) \subset O_\varepsilon(A) \forall\, t \geq T.$$

For fixed $B \subset X$ and $(s, \tau) \in \mathbb{R}_d$ let us define the following sets

$$\gamma^\tau_{s,\sigma}(B) = \bigcup_{t\geq s} U_\sigma(t, \tau, B),\ \gamma^\tau_{s,\Sigma}(B) = \bigcup_{t\geq s} U_\Sigma(t, \tau, B),$$

$$\omega_\Sigma(\tau, B) = \bigcap_{s\geq\tau} cl_X(\gamma^\tau_{s,\Sigma}(B)).$$

It is clear that $\omega_\Sigma(\tau, B) = \bigcap\limits_{s\geq p} cl_X(\gamma^\tau_{s,\Sigma}(B))\ \ \forall p \geq \tau$.

Definition 12.4 The family of MP $\{U_\sigma|\sigma \in \Sigma\}$ is called uniformly asymptotically compact, if for arbitrary $\tau \in \mathbb{R}$ and $B \in \beta(X)$ there exists $A(\tau, B) \in K(X)$ such that

$$U_\Sigma(t, \tau, B) \to A(\tau, B),\ t \to +\infty\ \text{in}\ X.$$

It is known [19] that if $\forall\tau \in \mathbb{R},\ \forall B \in \beta(X)\ \exists T = T(\tau, B)\ \gamma^\tau_{T,\Sigma}(B) \in \beta(X)$, then the condition of uniformly asymptotically compactness is equivalent to the following one:

$$\forall\tau \in \mathbb{R}\ \forall B \in \beta(X)\ \forall t_n \nearrow \infty$$

$$\text{every sequence}\ \ \xi_n \in U_\Sigma(t_n, \tau, B)\ \text{is precompact.}$$

Definition 12.5 A set $\Theta_\Sigma \subset X$ is called uniform global attractor of the family of MP $\{U_\sigma|\sigma \in \Sigma\}$, if :

(1) Θ_Σ is uniformly attracting set;
(2) for every uniformly attracting set Y we have $\Theta_\Sigma \subset cl_X Y$.

Uniform global attractor $\Theta_\Sigma \subset X$ is called invariant (semiinvariant), if $\forall\ (t, \tau) \in \mathbb{R}_d$

$$\Theta_\Sigma = U_\Sigma(t, \tau, \Theta_\Sigma)\ \ (\Theta_\Sigma \subset U_\Sigma(t, \tau, \Theta_\Sigma)).$$

If Θ_Σ is compact, invariant uniform global attractor, then it is called stable if $\forall\ \varepsilon > 0 \exists\ \delta > 0 \forall\ (t, \tau) \in \mathbb{R}_d$

$$U_\Sigma(t, \tau, O_\delta(\Theta_\Sigma)) \subset O_\varepsilon(\Theta_\Sigma).$$

The following sufficient conditions we can obtain with slight modifications from [19].

Theorem 12.2 (*I*) *Let us assume that the family of MP* $\{U_\sigma|\sigma \in \Sigma\}$ *satisfies the following conditions:*

(1) $\exists\ B_0 \in \beta(X)\ \forall\ B \in \beta(X)\ \forall \tau \in \mathbb{R}\ \exists\ T = T(\tau, B)$

$$\forall\ t \geq T\ U_\Sigma(t, \tau, B) \subset B_0;$$

(2) $\{U_\sigma|\sigma \in \Sigma\}$ is uniformly asymptotically compact.
Then $\{U_\sigma\}_{\sigma\in\Sigma}$ has compact uniform global attractor

$$\Theta_\Sigma = \bigcup_{\tau\in\mathbb{R}} \bigcup_{B\in\beta(X)} \omega_\Sigma(\tau, B) = \omega_\Sigma(0, B_0) = \omega_\Sigma(\tau, B_0)\ \ \forall \tau \in \mathbb{R}. \tag{12.8}$$

(II) If $\{U_\sigma\}_{\sigma\in\Sigma}$ satisfy (1), (2), Σ is compact and $\forall t \geq \tau$ the mapping

$$(x, \sigma) \mapsto U_\sigma(t, \tau, x) \tag{12.9}$$

has closed graph , then Θ_Σ is semiinvariant.
If, moreover, $\forall h \geq 0\ T(h)\Sigma = \Sigma$ and the family MP $\{U_\sigma|\sigma \in \Sigma\}$ is strict, then Θ_Σ is invariant.

(III) If $\{U_\sigma\}_{\sigma\in\Sigma}$ satisfy (1), (2), Σ is connected and compact, $\forall t \geq \tau$ the mapping (12.9) is upper semicontinuous and has closed and connected values, B_0 is connected set, then Θ_Σ is connected set.

(IV) If $\{U_\sigma|\sigma \in \Sigma\}$ is strict, $T(h)\Sigma = \Sigma$ for any $h \geq 0$, there exists a compact, invariant uniform global attractor Θ_Σ and the following condition hold:

$$\textit{if}\ y_n \in U_\Sigma(t_n, \tau, x_n),\ t_n \to t_0,\ x_n \to x_0,$$

$$\text{then up to subsequence } y_n \to \ y_0 \in U_\Sigma(t_0, \tau, x_0), \tag{12.10}$$

then Θ_Σ is stable.

Proof (I) From conditions (1), (2) due to [19] we have that $\forall \tau \in \mathbb{R}\ \forall B \in \beta(X)\ \omega_\Sigma(\tau, B) \neq \emptyset$, is compact , $\omega_\Sigma(\tau, B) \subset B_0$ and the set

$$\Theta_\Sigma = \bigcup_{\tau\in\mathbb{R}} \bigcup_{B\in\beta(X)} \omega_\Sigma(\tau, B)$$

is uniform global attractor. Let us prove that $\omega_\Sigma(\tau, B) \subset \omega_\Sigma(\tau_0, B_0)\ \forall \tau, \tau_0 \in \mathbb{R}$.

$$U_\sigma(t, \tau, B) \subset U_\sigma(t, \frac{t}{2}, U_\sigma(\frac{t}{2}, \tau, B)) \subset U_{T(\frac{t}{2}-\tau_0)\sigma}(\frac{t}{2} + \tau_0, \tau_0, U_\sigma(\frac{t}{2}, \tau, B)) \subset$$

$$\subset U_\Sigma(\frac{t}{2} + \tau_0, \tau_0, B_0), \text{ if } \frac{t}{2} \geq T(\tau, B) + |\tau_0| + |\tau| := T.$$

So, for $t \geq 2T$

$$U_\Sigma(t, \tau, B) \subset U_\Sigma(\frac{t}{2} + \tau_0, \tau_0, B_0).$$

Then for $s \geq 2T$

$$\bigcup_{t\geq s} U_\Sigma(t, \tau, B) \subset \bigcup_{t\geq s} U_\Sigma(\frac{t}{2} + \tau_0, \tau_0, B_0) = \bigcup_{p\geq \frac{s}{2}+\tau_0} U_\Sigma(p, \tau_0, B_0),$$

$$\bigcap_{s\geq 2T} \overline{\bigcup_{t\geq s} U_\Sigma(t, \tau, B)} = \omega_\Sigma(\tau, B) \subset \bigcap_{s\geq 2T} \overline{\bigcup_{p\geq \frac{s}{2}+\tau_0} U_\Sigma(p, \tau_0, B_0)}$$

$$= \bigcap_{s'\geq T+\tau_0} \overline{\bigcup_{p\geq s'} U_\Sigma(p, \tau_0, B_0)} = \omega_\Sigma(\tau_0, B_0).$$

So we deduce equality (12.8).

(II) Due to (12.8) $\forall \xi \in \Theta_\Sigma = \omega_\Sigma(\tau, B_0)\ \exists t_n \nearrow +\infty,\ \exists \sigma_n \in \Sigma\ \exists \xi_n \in U_{\Sigma_n}(t_n, \tau, B_0)$ such that $\xi = \lim_{n\to\infty} \xi_n$. Then

$$\xi_n \in U_{\sigma_n}(t_n - t - \tau + t + \tau, \tau, B_0) \subset$$

$$\subset U_{\sigma_n}(t_n - t - \tau + t + \tau, t_n - t + \tau, U_{\sigma_n}(t_n - t + \tau, \tau, B_0))$$
$$\subset U_{T(t_n-t)\sigma_n}(t, \tau, \eta_n),$$

where $\eta_n \in U_{\sigma_n}(t_n - t + \tau, \tau, B_0),\ t \geq \tau$ and for sufficiently large $n \geq 1$.

From uniform asymptotically compactness we have that on some subsequence $\eta_n \to \eta \in \omega_\Sigma(\tau, B_0) = \Theta_\Sigma$,

$$T(t_n - t)\sigma_n \to \sigma \in \Sigma.$$

Then from (12.9) we deduce :

$$\xi \in U_\Sigma(t, \tau, \Theta_\Sigma),$$

and therefore $\Theta_\Sigma \subset U_\Sigma(t, \tau, \Theta_\Sigma)$.
Other statements of the theorem are proved analogously to [19]. Theorem is proved.

Corollary 12.1 *If for the family of MP* $\{U_\sigma\}_{\sigma\in\Sigma}$ *we have :*

(1) $\forall h \geq 0\ T(h)\Sigma = \Sigma$;
(2) $\forall (t, \tau) \in \mathbb{R}_d\ \forall h \geq 0\ \forall \sigma \in \Sigma\ \forall x \in X$

$$U_\sigma(t + h, \tau + h, x) = U_{T(h)\sigma}(t, \tau, x),$$

then all conditions of previous theorem can be verified only for $\tau = 0$.

Proof Under conditions (1), (2) $\forall t \geq \tau$ if $\tau \geq 0$ then

$$U_\sigma(t, \tau, x) = U_{T(\tau)\sigma}(t - \tau, 0, x),$$

and if $\tau \leq 0$ then $\exists \sigma' \in \Sigma : \sigma = T(-\tau)\sigma'$, so

$$U_\sigma(t, \tau, x) = U_{T(-\tau)\sigma'}(t, \tau, x) = U_{\sigma'}(t - \tau, 0, x).$$

In the single-valued case it is known [8], that the uniform global attractor consists of bounded complete trajectories of processes $\{U_\sigma\}_{\sigma\in\Sigma}$.

Definition 12.6 The mapping $\varphi : [\tau, +\infty) \mapsto X$ is called trajectory of MP U_σ, if $\forall t \geq s \geq \tau$

$$\varphi(t) \in U_\sigma(t, s, \varphi(s)). \tag{12.11}$$

If for $\varphi : \mathbb{R} \mapsto X$ the equality (12.11) takes place $\forall t \geq s$, then φ is called complete trajectory.

Now we assume that for arbitrary $\sigma \in \Sigma$ and $\tau \in \mathbb{R}$ we have the set K_σ^τ of mappings $\varphi : [\tau, +\infty) \mapsto X$ such that :

(a) $\forall x \in X\ \exists \varphi(\cdot) \in K_\sigma^\tau$ such, that $\varphi(\tau) = x$;
(b) $\forall \varphi(\cdot) \in K_\sigma^\tau\ \ \forall s \geq \tau\ \ \varphi(\cdot)|_{[s,+\infty)} \in K_\sigma^s$;
(c) $\forall h \geq 0\ \forall \varphi(\cdot) \in K_\sigma^{\tau+h}\ \ \varphi(\cdot + h) \in K_{T(h)\sigma}^\tau$.

Let us put

$$U_\sigma(t,\tau,x)=\{\varphi(t)|\varphi(\cdot)\in K_\sigma^\tau,\ \varphi(\tau)=x\}. \tag{12.12}$$

Lemma 12.1 *Formula* (12.12) *defines the family of MP* $\{U_\sigma\}_{\sigma\in\Sigma}$, *and* $\forall\varphi(\cdot)\in K_\sigma^\tau$

$$\forall t\geq s\geq\tau\quad \varphi(t)\in U_\sigma(t,s,\varphi(s)). \tag{12.13}$$

Proof Let us check conditions of the Definition 12.2.

(1) $U_\sigma(\tau,\tau,x)=\varphi(\tau)=x$;

(2) $\forall\xi\in U_\sigma(t,\tau,x)\ \ \xi=\varphi(t)$, where $\varphi\in K_\sigma^\tau,\varphi(\tau)=x$. Then for $s\in[\tau,t]$ $\varphi(s)\in U_\sigma(s,\tau,x)$ and from $\varphi|_{[s,+\infty)}\in K_\sigma^s$ we have $\varphi(t)\in U_\sigma(t,s,\varphi(s))$. So
$$\xi\in U_\sigma(t,s,U_\sigma(s,\tau,x)).$$

(3) $\forall\xi\in U_\sigma(t+h,\tau+h,x)\ \xi=\varphi(t+h)$, where $\varphi\in K_\sigma^{\tau+h},\varphi(\tau+h)=x$. Then $\psi(\cdot)=\varphi(\cdot+h)\in K_{T(h)\sigma}^\tau,\psi(\tau)=x$, so $\xi=\psi(t)\in U_{T(h)\sigma}(t,\tau,x)$. Lemma is proved.

It is easy to show that under conditions (a)–(c), if $\forall s\geq\tau\ \forall\psi\in K_\sigma^\tau,\forall\varphi\in K_\sigma^s$ such that $\psi(s)=\varphi(s)$, we have

$$\theta(p)=\begin{cases}\psi(p),\ p\in[\tau,s]\\ \varphi(p),\ p>s,\end{cases}\in K_\sigma^\tau, \tag{12.14}$$

then in the condition (2) of Definition 12.2 equality takes place.

If $\forall h\geq 0\ \forall\varphi\in K_{T(h)\sigma}^\tau\varphi(\cdot-h)\in K_\sigma^{\tau+h}$, then in the condition (3) of Definition 12.2 equality takes place.

From (12.13) we immediately obtain that if for mapping $\varphi(\cdot):\mathbb{R}\mapsto X$ for arbitrary $\tau\in\mathbb{R}$ we have $\varphi(\cdot)|_{[\tau,+\infty)}\in K_\sigma^\tau$, then $\varphi(\cdot)$ is complete trajectory of U_σ.

The next result is generalization on non-autonomous case results from [20, 21].

Lemma 12.2 *Let the family of MP* $\{U_\sigma\}_{\sigma\in\Sigma}$ *be constructed by the formula* (12.12), $\forall\varphi(\cdot)\in K_\sigma^\tau$ *is continuous on* $[\tau,+\infty)$, *the condition* (12.14) *takes place and the following one: if* $\varphi_n(\cdot)\in K_\sigma^\tau,\ \varphi_n(\tau)=x$, *then* $\exists\varphi(\cdot)\in K_\sigma^\tau,\ \varphi(\tau)=x$ *such that on some subsequence*

$$\varphi_n(t)\to\varphi(t)\ \ \forall t\geq\tau.$$

Then every continuous on $[\tau,+\infty)$ *trajectory of MP* U_σ *belongs to* K_σ^τ.

Proof Let $\psi:[\tau,+\infty)\mapsto X$ be continuous trajectory. Let us construct sequence $\{\varphi_n(\cdot)\}_{n=1}^\infty\subset K_\sigma^\tau$ such that

$$\varphi_n(\tau+j2^{-n})=\psi(\tau+j2^{-n}),\quad j=0,1,...,n2^n.$$

For $\varphi_1(\cdot)$ we have

$$\psi(\tau+\frac{1}{2}) \in U_\sigma(\tau+\frac{1}{2},\tau,\psi(\tau)),$$

$$\psi(\tau+1) \in U_\sigma(\tau+1,\tau+\frac{1}{2},\psi(\tau+\frac{1}{2})).$$

So there exists $\widetilde{\varphi}(\cdot) \in K_\sigma^\tau$, there exists $\widetilde{\widetilde{\varphi}}(\cdot) \in K_\sigma^{\tau+\frac{1}{2}}$ such that

$$\psi(\tau+\frac{1}{2}) = \widetilde{\varphi}(\tau+\frac{1}{2}),\ \ \widetilde{\varphi}(\tau) = \psi(\tau),$$

$$\psi(\tau+1) = \widetilde{\widetilde{\varphi}}(\tau+1),\ \ \widetilde{\widetilde{\varphi}}(\tau+\frac{1}{2}) = \psi(\tau+\frac{1}{2}).$$

Therefore due to (12.14) for function

$$\varphi_1(p) = \begin{cases} \widetilde{\varphi}(p), & \tau \le p \le \tau+\frac{1}{2}, \\ \widetilde{\widetilde{\varphi}}(p), & p > \tau+\frac{1}{2} \end{cases} \quad \text{we have:}$$

$$\varphi_1(\cdot) \in K_\sigma^\tau,\ \ \varphi_1(\tau) = \psi(\tau),\ \ \varphi_1(\tau+\frac{1}{2}) = \psi(\tau+\frac{1}{2}),\ \ \varphi_1(\tau+1) = \psi(\tau+1).$$

Further, using (12.14), we obtain required property for every $n \ge 1$. As $\varphi_n(\tau) = \psi(\tau)$, so $\exists \varphi(\cdot) \in K_\sigma^\tau,\ \varphi(\tau) = \psi(\tau)$ such that on subsequence $\forall t \ge \tau\ \ \varphi_n(t) \to \varphi(t)$. As $\forall t = \tau + j2^{-n}\ \ \varphi(t) = \psi(t)$, so from continuity $\varphi(t) = \psi(t)\ \ \forall t \ge \tau$. Lemma is proved.

The following theorem declare structure of uniform global attractor in terms of bounded complete trajectories of corresponding m-processes. It should be noted that this result is known for single-valued case [8] and in multi-valued case for very special class of strict processes, generated by strict compact semiprocesses, which act in Banach spaces [22].

Theorem 12.3 *Let Σ is compact, $T(h)\Sigma = \Sigma\ \forall\ h \ge 0$, the family of MP $\{U_\sigma\}_{\sigma\in\Sigma}$ satisfies (12.12), in condition (3) of Definition 12.2 equality takes place, the mapping $(x,\sigma) \mapsto U_\sigma(t,0,x)$ has closed graph. Let us assume that there exists Θ_Σ—compact uniform global attractor of the family $\{U_\sigma\}_{\sigma\in\Sigma}$, and one of two conditions hold: either the family of MP $\{U_\sigma\}_{\sigma\in\Sigma}$ is strict, or*

$$\text{for every } \sigma_n \to \sigma_0,\ \ x_n \to x_0 \text{ if } \varphi_n(\cdot) \in K_{\sigma_n}^0,\ \ \varphi_n(0) = x_n,$$

$$\text{so } \exists\, \varphi(\cdot) \in K_{\sigma_0}^0,\ \ \varphi(0) = x_0 \text{ such that on subsequnce } \forall t \ge 0\ \ \varphi_n(t) \to \varphi(t). \quad (12.15)$$

Then the following structural formula holds

$$\Theta_\Sigma = \bigcup_{\sigma\in\Sigma} \mathscr{K}_\sigma(0), \tag{12.16}$$

where $\mathscr{K}_\sigma$ is the set of all bounded complete trajectories of MP U_σ.

Proof First let us consider situation when the family of MP $\{U_\sigma\}_{\sigma\in\Sigma}$ is strict. In this case one can consider multivalued semigroup (m-semiflow) on the extended phase space $X \times \Sigma$ by the rule

$$G(t, \{x, \sigma\}) = \{U_\sigma(t, 0, x), T(t)\sigma\}. \tag{12.17}$$

Then G is strict, has closed graph and compact attracting set $\Theta_\Sigma \times \Sigma$. So G has compact invariant global attractor

$$\mathscr{A} = \bigcap_{s\geq 0} \overline{\bigcup_{t\geq s} G(t, \Theta_\Sigma \times \Sigma)} = \{\gamma(0) | \gamma \text{ is bounded complete trajectories of } G\}.$$

Here under complete trajectory of m-semiflow G we mean the mapping $\mathbb{R} \ni t \mapsto \gamma(t)$ such that

$$\forall\, t \in \mathbb{R}\ \forall\, s \geq 0\ \gamma(t + s) \in G(s, \gamma(t)).$$

Let us consider two projectors Π_1 and Π_2, $\Pi_1(u, \sigma) = u$, $\Pi_2(u, \sigma) = \sigma$. As $T(t)\Sigma = \Sigma$, so $\Pi_2\mathscr{A} = \Sigma$. Let us prove that $\Pi_1\mathscr{A} = \Theta_\Sigma$.

As $\forall\, B \in \beta(X)$ $G(t, B \times \Sigma) \to \mathscr{A}$, $t \to +\infty$, so

$$U_\Sigma(t, \tau, B) \to \Pi_1\mathscr{A},$$

so $\Theta_\Sigma \subset \Pi_1\mathscr{A}$. Let us prove that $\Pi_1\mathscr{A} = \bigcup\limits_{\sigma\in\Sigma} \mathscr{K}_\sigma(0)$. For this purpose we take $(u_0, \sigma_0) \in \mathscr{A}$. Then there exists $\gamma(\cdot) = \{u(\cdot), \sigma(\cdot)\}$, which is bounded complete trajectory of G and such that $\gamma(0) = (u_0, \sigma_0)$. Then $\forall t \geq \tau$

$$u(t) \in U_{\sigma(\tau)}(t - \tau, 0, u(\tau)), \quad \sigma(t) = T(t - \tau)\sigma(\tau).$$

If $\tau \geq 0$, then $\sigma(\tau) = T(\tau)\sigma_0$, that is

$$u(t) \in U_{T(\tau)\sigma_0}(t - \tau, 0, u(\tau)) = U_{\sigma_0}(t, \tau, u(\tau)).$$

If $\tau < 0$, then $\sigma_0 = T(-\tau)\sigma(\tau)$, so

$$u(t) \in U_{\sigma(\tau)}(t - \tau, \tau - \tau, u(\tau)) = U_{T(-\tau)\sigma(\tau)}(t, \tau, u(\tau)) = U_{\sigma_0}(t, \tau, u(\tau)).$$

Therefore $u_0 = u(0) \in \mathscr{K}_{\sigma_0}(0) \subset \bigcup\limits_{\sigma\in\Sigma} \mathscr{K}_\sigma(0)$.

Now let $u_0 = u(0) \in K_{\sigma_0}(0),\ \ u(t) \in U_{\sigma_0}(t, \tau, u(\tau))\ \ \forall t \geq \tau$. As $T(t)\Sigma = \Sigma$, so there exists $\sigma(s), s \in \mathbb{R}$, such that $\sigma(t) = T(t-\tau)\sigma(\tau),\ \ \forall t \geq \tau,\ \ \sigma(0) = \sigma_0$. Then for $s \geq 0$ we have

$$\begin{aligned} G(t, \{u(s), \sigma(s)\}) &= (U_{\sigma(s)}(t, 0, u(s)), T(t)\sigma(s)) \\ &= (U_{T(s)\sigma_0}(t, 0, u(s)), \sigma(t+s)) = (U_{\sigma_0}(t+s, s, u(s)), \sigma(t+s)), \\ &\quad \{u(t+s), \sigma(t+s)\} \in (U_{\sigma_0}(t+s, s, u(s)), \sigma(t+s)). \end{aligned}$$

If $s < 0$, then $\sigma_0 = T(-s)\sigma(s)$, and

$$u(t+s) \in U_{\sigma_0}(t+s, s, u(s)) = U_{T(-s)\sigma(s)}(t+s, s, u(s)) = U_{\sigma(s)}(t, 0, u(s)).$$

Then $u_0 \in \Pi_1 \mathscr{A}$ and $\Pi_1 \mathscr{A} = \bigcup_{\sigma \in \Sigma} \mathscr{K}_\sigma(0)$.

Since for arbitrary attracting set P and for arbitrary bounded complete trajectory $\Gamma = \{u(s)\}_{s \in \mathbb{R}}$ of the process U_σ we have

$$u(0) \in U_\sigma(0, -n, u(-n)) = U_{T(n)\sigma(-n)}(0, -n, u(-n))$$

$$\subset U_\Sigma(n, 0, \Gamma) \to P,\ \ n \to +\infty,$$

so $u(0) \in P$, and we obtain (12.16).

Now let us consider another case, when family of m-processes is not strict, but the condition (12.15) holds. Let us show that $\mathscr{K}_\sigma(0) \subset \Theta_\Sigma$. If $z \in \mathscr{K}_\sigma(0)$, then there exists bounded complete trajectory $\varphi(\cdot)$ of m-process U_σ, such that $\varphi(0) = z$. Let us denote $\Gamma = \bigcup\limits_{t \in \mathbb{R}} \varphi(t) \in \beta(X)$. Then for $z = \varphi(0)$ we have

$$\varphi(0) \in U_\sigma(0, -n, \varphi(-n)) = U_{T(n)\sigma_n}(0, -n, \varphi(-n)) \subset U_\Sigma(n, 0, \Gamma).$$

Since $\forall \varepsilon > 0\ \ \exists n_0\ \ \forall n \geq n_0\ \ U_\Sigma(n, 0, \Gamma) \subset O_\varepsilon(\Theta_\Sigma)$, then $z \in \Theta_\Sigma$ and we obtain required embedding.

Now let $z \in \Theta_\Sigma = \omega_\Sigma(0, B_0)$. Then $z = \lim\limits_{n \to +\infty} \xi_n,\ \ \xi_n \in U_\Sigma(t_n, 0, B_0)$. Therefore on some subsequence

$$z = \lim_{n \to +\infty} \varphi_n(t_n),\ \ \varphi_n(\cdot) \in K^0_{\sigma_n},\ \ \varphi_n(0) \in B_0,\ \ \sigma_n \to \sigma.$$

For $\forall n \geq 1$ let us consider

$$\psi_n(\cdot) := \varphi_n(\cdot + t_n) \in K^{-t_n}_{T(t_n)\sigma_n},$$

that is $\psi_n(\cdot) \in K^{-t_n}_{\tilde{\sigma}_n}$, where $\tilde{\sigma}_n = T(t_n)\sigma_n$. Then $\psi_n(\cdot) \in K^0_{\tilde{\sigma}_n},\ \ \tilde{\sigma}_n \to \tilde{\sigma},\ \ \psi_n(0) = \varphi_n(t_n) \to z$, so there exists $\psi^{(0)}(\cdot) \in K^0_{\tilde{\sigma}},\ \ \psi^{(0)}(0) = z$, such that

$$\forall t \geq 0 \;\; \psi_n(t) = \varphi_n(t + t_n) \to \psi^{(0)}(t).$$

For $\tau = -1 \;\; \forall n \geq n_1 \;\; -t_n < -1$, therefore $\psi_n(\cdot) \in K^{-1}_{\tilde{\sigma}_n}$ and on some subsequence

$$\psi_n(-1) = \varphi_n(t_n - 1) \to z_1.$$

Herewith there exists $\psi^{(-1)}(\cdot) \in K^{-1}_{\tilde{\sigma}}$ such that on subsequence

$$\psi_n(t) = \varphi_n(t + t_n) \to \psi^{(-1)}(t) \;\; \forall t \geq -1,$$

and $\forall t \geq 0 \;\; \psi^{(0)}(t) = \psi^{(-1)}(t)$. By standard diagonal procedure we construct sequence of functions

$$\psi^{(-k)}(\cdot) \in K^{-k}_{\tilde{\sigma}}, \;\; k \geq 0,$$

with $\psi^{(-k+1)}(t) = \psi^{(-k)}(t) \;\; \forall t \geq -k + 1$. Let us put

$$\psi(t) := \psi^{(-k)}(t), \;\; if \;\; t \geq -k.$$

Then the function $\psi(\cdot)$ is correctly defined, $\psi : \mathbb{R} \mapsto X$.

Moreover $\forall \tau < 0 \;\; \exists k$ such that $[\tau, +\infty) \subset [-k, +\infty)$, on $[-k, +\infty) \;\; \psi(\cdot) \equiv \psi^{(-k)}$, so $\psi(\cdot) \in K^{-k}_{\tilde{\sigma}}$, and from this

$$\psi(\cdot) \in K^{\tau}_{\tilde{\sigma}}, \;\; \psi(0) = \psi^{(0)}(0) = z.$$

Since on subsequence

$$\forall t \in \mathbb{R} \;\; \psi(t) = \lim_{n \to +\infty} \varphi_n(t + t_n) \in \omega_{\Sigma}(0, B_0) \in \beta(X),$$

then $z = \psi(0) \in \mathscr{K}_{\tilde{\sigma}}$ and theorem is proved.

12.4 Uniform Global Attractor for RD-System

Definition 12.7 Let Θ be some topological space of functions from $\mathbb{R}$ to topological space E. The function $\xi \in \Theta$ is called translation compact in Θ, if the set

$$H(\xi) = cl_{\Theta}\{\xi(\cdot + s) \mid s \in \mathbb{R}\}$$

is compact in Θ.

To construct family of m-processes for the problem (12.1) we suppose that time-depended functions f and h are translation compact in natural spaces [8]. More precisely, we will assume that

$$h \text{ is translation compact in } L^{2,w}_{loc}(\mathbb{R}; H), \tag{12.18}$$

where $L^{2,w}_{loc}(\mathbb{R}; H)$ is the space $L^2_{loc}(\mathbb{R}; H)$ with the local weak convergence topology, and

$$f \text{ is translation compact in } C(\mathbb{R}; C(\mathbb{R}^N, \mathbb{R}^N)), \tag{12.19}$$

where $C(\mathbb{R}; C(\mathbb{R}^N, \mathbb{R}^N))$ equipped with local uniform convergence topology.

It is known that condition (12.18) is equivalent to

$$|h|^2_+ := \sup_{t\in\mathbb{R}} \int_t^{t+1} |h(s)|^2 ds < \infty \tag{12.20}$$

It is also known that condition (12.19) is equivalent to

$$\begin{gathered}\forall R > 0 \, f \text{ is bounded and uniformly continuous on} \\ Q(R) = \{(t, v) \in \mathbb{R} \times \mathbb{R}^N \mid |v|_{\mathbb{R}^N} \le R\}.\end{gathered} \tag{12.21}$$

If conditions (12.18),(12.19) take place, then the symbol space

$$\Sigma = cl_{C(\mathbb{R};C(\mathbb{R}^N,\mathbb{R}^N))\times L^{2,w}_{loc}(\mathbb{R};H)}\{(f(\cdot + s), h(\cdot + s)) \mid s \in \mathbb{R}\} \tag{12.22}$$

is compact, and $\forall\, s \ge 0\, T(s)\Sigma = \Sigma$, where $T(s)$ is translation semigroup, which is continuous on Σ.

For every $\sigma = (f_\sigma, h_\sigma) \in \Sigma$ we consider the problem

$$\begin{cases} u_t = a\Delta u - f_\sigma(t, u) + h_\sigma(t, x), \quad x \in \Omega, \ t > \tau, \\ u|_{\partial\Omega} = 0. \end{cases} \tag{12.23}$$

It is proved in [19] that $\forall\ \sigma \in \Sigma f_\sigma$ satisfies (12.3), (12.4) with the same constants $C_1, C_2, \gamma_i, |h_\sigma|_+ \le |h|_+$. So we can apply Theorem 2 and obtain that $\forall\ \tau \in \mathbb{R}, u_\tau \in H$ the problem (12.23) has at least one solution on $(\tau, +\infty)$, each solution of (12.23) belongs to $C([\tau, +\infty); H)$ and satisfies energy equality (12.6). For every $\sigma \in \Sigma, \tau \in \mathbb{R}$ we define

$$K^\tau_\sigma = \{u(\cdot) \mid u(\cdot) \text{ is solution of } (12.23) \text{ on } (\tau, +\infty)\} \tag{12.24}$$

and according to (12.12) we put $\forall\, \sigma \in \Sigma, \forall\, t \ge \tau, \forall\, u_\tau \in H$

$$U_\sigma(t, \tau, u_\tau) = \{u(t) | u(\cdot) \in K^\tau_\sigma,\ u(\tau) = u_\tau\}. \tag{12.25}$$

From [19] and Theorem 13 we obtain the following result

Theorem 12.4 *Under conditions* (12.3), (12.4), (12.18), (12.19) *formula* (12.25) *defines a strict family of MP* $\{U_\sigma\}_{\sigma\in\Sigma}$ *which has compact, invariant, stable and con-*

nected uniform global attractor Θ_Σ, which consists of bounded complete trajectories, that is

$$\Theta_\Sigma = \bigcup_{\sigma\in\Sigma} \mathscr{K}_\sigma(0), \tag{12.26}$$

where $\mathscr{K}_\sigma$ is the set of all bounded complete trajectories of MP U_σ.

Now we want to use formula (12.26) for proving that the uniform global attractor of RD-system is bounded set in the space $(L^\infty(\Omega))^N \cap V$.

First let us consider the following conditions:

$$\exists\, M_i > 0,\ i = \overline{1,N} \text{ such that for all } v = (v^1, ..., v^N) \in \mathbb{R}^N \text{ for a.a. } x \in \Omega\ \forall\, t \in \mathbb{R}$$

$$\sum_{i=1}^{N} (f^i(t,v) - h^i(t,x))(v^i - M_i)^+ \geq 0 \tag{12.27}$$

$$\sum_{i=1}^{N} (f^i(t,v) - h^i(t,x))(v^i + M_i)^- \leq 0 \tag{12.28}$$

where $\varphi^+ = max\{0,\varphi\}$, $\varphi^- = max\{0,-\varphi\}$, $\varphi = \varphi^+ - \varphi^-$.

Let us consider some example, which allow to verify conditions (12.27), (12.28).

Lemma 12.3 *If $N = 1$ (scalar equation), then from (12.3), (12.4) and $h \in L^\infty(\mathbb{R}\times\Omega)$ we have (12.27), (12.28).*

Proof From (12.3) and $h \in L^\infty(\Omega)$ for a.a. $x \in \Omega$ and $u \in \mathbb{R}$,

$$\tilde{\gamma}|u|^p - \tilde{C}_2 \leq g(t,x,u)u \leq \tilde{C}_1|u|^p + \tilde{C}_1,$$

where $g(t,x,u) = f(t,u) - h(t,x)$, $\tilde{\gamma}$ does not depend on t, u, x.

If $u \leq M$, then $g(t,x,u)(u-M)^+ = 0$.

If $u > M$, then

$$\begin{aligned} g(t,x,u)(u-M)^+ &= g(t,x,u)u\frac{(u-M)^+}{u} = g(t,x,u)u(1-\frac{M}{u}) \\ &\geq (\tilde{\gamma}u^p - \tilde{C}_2)(1-\frac{M}{u}) \geq (\tilde{\gamma}M^p - \tilde{C}_2)(1-\frac{M}{u}) \end{aligned}$$

and if we choose $M = \left(\frac{\tilde{C}_2}{\tilde{\gamma}}\right)^{\frac{1}{p}}$, then $g(t,x,u)(u-M)^+ \geq 0$ a.e.

Lemma 12.4 *If for arbitrary $N \geq 1$ $h \equiv 0, f(t,u) = (f^1(t,u), ..f^N(t,u))$, where $f^i(t,u) = (\sum_{i=1}^{N} |u^i|^2 - R^2)u^i, R > 0$ is positive constant, then conditions (12.27), (12.28) hold for $M_i = R$.*

Proof If $\sum\limits_{i=1}^{N}|u^i|^2 < R^2$, so $\forall\, i = \overline{1,N}|u^i| < R$ and

$$\sum_{i=1}^{N} f^i(t,u)(u^i - R)^+ = 0,$$

$$\sum_{i=1}^{N} f^i(t,u)(u^i + R)^- = 0.$$

If $\sum\limits_{i=1}^{N}|u^i|^2 \geq R^2$, then

$$\sum_{i=1}^{N} f^i(t,u)(u^i - R)^+ = (\sum_{i=1}^{N}|u^i|^2 - R^2)\sum_{i=1}^{N} u^i(u^i - R)^+ \geq 0,$$

$$\sum_{i=1}^{N} f^i(t,u)(u^i + R)^- = (\sum_{i=1}^{N}|u^i|^2 - R^2)\sum_{i=1}^{N} u^i(u^i + R)^- \leq 0.$$

Theorem 12.5 *If conditions* (12.3), (12.4), (12.18), (12.19), (12.27), (12.28) *hold and matrix a is diagonal, then the uniform global attractor Θ_Σ is bounded set in the space $(L^\infty(\Omega))^N \cap V$.*

Proof First let us prove that $\forall\sigma \in \Sigma$ functions f_σ, h_σ satisfy (12.27), (12.28). Indeed, there exists sequence $t_n \nearrow \infty$ such that $\forall\ T > 0,\ R > 0,\ \eta \in L^2((-T,T)\times\Omega)$

$$\sup_{|t|\leq T}\sup_{|v|\leq R}\sum_{i=1}^{N}|f^i(t+t_n,v) - f^i_\sigma(t,v)|^2 \to 0,\ n\to\infty,$$

$$\sum_{i=1}^{N}\int_{-T}^{T}\int_{\Omega}(h^i(t+t_n,x) - h^i_\sigma(t,x))\eta(t,x)dxdt \to 0,\ n\to\infty.$$

From (12.27)

$$\sum_{i=1}^{N}(f^i(t+t_n,v) - h^i(t+t_n,x))(v^i - M_i)^+ \geq 0. \tag{12.29}$$

Therefore for fixed v and for arbitrary $\varepsilon > 0$ there exists $N \geq 1$ such that $\forall\, n \geq N$

$$\sum_{i=1}^{N} h^i(t+t_n,x)(v^i-M_i)^+ \le \sum_{i=1}^{N} f^i(t+t_n,v)(v^i-M_i)^+ < \sum_{i=1}^{N} f^i_\sigma(t,v)(v^i-M_i)^+ + \varepsilon.$$

Because

$$\sum_{i=1}^{N} h^i(t+t_n,x)(v^i-M_i)^+ \to \sum_{i=1}^{N} h^i_\sigma(t,x)(v^i-M_i)^+ \text{ weakly in } L^2((-T,T)\times\Omega),$$

from Mazur's Theorem we deduce that

$$\sum_{i=1}^{N} h^i_\sigma(t,x)(v^i-M_i)^+ \le \sum_{i=1}^{N} f^i_\sigma(t,v)(v^i-M_i)^+ + \varepsilon \text{ for a.a. } x\in\Omega.$$

From arbitrary choice of ε we can obtain required result.

It is easy to obtain that for arbitrary weak solution of (12.1) and for every $\eta \in C_0^\infty(\tau,T)$

$$\int_\tau^T (u_t,u^+)\eta dt = -\frac{1}{2}\int_\tau^T |u^+|^2\eta_t dt. \tag{12.30}$$

Then putting $g_\sigma = f_\sigma - h_\sigma$ and for numbers $M_1, ..., M_N$ from condition (12.27) we have

$$\frac{1}{2}\frac{d}{dt}\sum_{i=1}^{N}|(u^i-M_i)^+|^2+\beta\sum_{i=1}^{N}\|(u^i-M_i)^+\|^2+\int_\Omega\sum_{i=1}^{N} g^i_\sigma(t,x,u)(u^i-M_i)^+dx=0.$$

Then from (12.27)

$$\frac{d}{dt}\sum_{i=1}^{N}|(u^i-M_i)^+|^2+2\beta\sum_{i=1}^{N}|(u^i-M_i)^+|^2\le 0$$

and for all $t>\tau$

$$\sum_{i=1}^{N}|(u^i-M_i)^+(t)|^2\le\sum_{i=1}^{N}|(u^i-M_i)^+(\tau)|^2e^{-2\lambda_1\beta(t-\tau)}. \tag{12.31}$$

If $u(\cdot)\in\mathscr{K}_\sigma$ then from (12.31) taking $\tau\to-\infty$ we obtain $u^i(x,t)\le M_i$, $i=\overline{1,N}$, $\forall t\in\mathbb{R}$, for a.a. $x\in\Omega$.

In the same way we will have $u^i(x,t)\ge M_i$ (using $(u^i+M_i)^-$).

Then

$$ess\sup_{x\in\Omega}|z^i(x)|\le M_i\ \ \forall z=(z^1,...,z^N)\in\Theta_\Sigma.$$

So we obtain that Θ_Σ is bounded set in the space $(L^\infty(\Omega))^N$. From the equality $\Theta_\Sigma = U_\Sigma(t,\tau,\Theta_\Sigma)\forall\, t \geq \tau$ we deduce that $\forall\, \sigma \in \Sigma\ U_\sigma(t,\tau,\Theta_\Sigma) \subset \Theta_\Sigma$. Now let us consider arbitrary complete trajectory $u(\cdot) \in \mathscr{K}_\sigma$. Due to definition of weak solution for a.a . $t \in \mathbb{R}\ u(t) \in V$. We take such $\tau \in \mathbb{R}$ that $u(\tau) \in V$ and consider the following Cauchy problem

$$\begin{cases} v_t = a\Delta v - f_\sigma(t,u) + h_\sigma(t,x), \quad x \in \Omega,\ t > \tau, \\ v|_{\partial\Omega} = 0, \\ v|_{t=\tau} = u(\tau). \end{cases} \tag{12.32}$$

Because $\forall\ t \geq \tau\ u(t) \in \Theta_\Sigma$, which is bounded in $(L^\infty(\Omega))^N$, we have that $f_\sigma(t,u(t,x)) \in (L^\infty(\Omega))^N$. Thus for linear problem (12.32) from well-known results one can deduce that $\forall\, T > \tau\ v \in C([\tau,T];V)$. So from uniqueness of the solution of Cauchy problem (12.32) $v \equiv u$ on $[\tau,+\infty)$ and, therefore, $\forall\, t \geq \tau\ u(t) \in V$. It means that $\forall\, t \in \mathbb{R}\ u(t) \in V$ and from the formula (12.26) $\Theta_\Sigma \subset V$.

From the energy equality, applying to function u, and boundness of Θ_Σ in the space H we deduce, that $\exists\, C > 0$, which does not depend on σ, such that $\forall\, t \in \mathbb{R}$

$$\int_t^{t+1} \|u(s)\|^2 ds \leq C(1 + \int_t^{t+1} |h_\sigma(s)|^2 ds).$$

From translation compactness of h we have

$$\int_t^{t+1} \|u(s)\|^2 ds \leq C(1 + |h|_+^2).$$

So for arbitrary $t \in \mathbb{R}$ we find $\tau \in [t,t+1]$ such that $\|u(\tau)\|^2 \leq C(1+|h|_+^2)$. Then for the problem (12.32) we obtain inequality

$$\forall\, t \geq \tau\ \ \|v(t)\|^2 \leq e^{-\delta(t-\tau)}\|u(\tau)\|^2 + D,$$

where positive constants $\delta,\ D$ do not depend on σ. Thus

$$\forall\, t \in \mathbb{R}\ \ \|u(t)\|^2 \leq C(1+|h|_+^2) + D$$

and theorem is proved.

Acknowledgments The first two authors were partially supported by the Ukrainian State Fund for Fundamental Researches under grants GP/F44/076 and GP/F49/070.

References

1. Kapustyan, O.V., Melnik, V.S., Valero, J., Yasinsky, V.V.: Global Attractors of Multivalued Dynamical Systems and Evolution Equations Without Uniqueness. Naukova Dumka, Kyiv (2008)
2. Kapustyan, O.V., Valero, J.: On the connectedness and asymptotic behaviour of solutions of reaction-diffusion systems. J. Math. Anal. Appl. **323**, 614–633 (2006)
3. Melnik, V.S., Valero, J.: On global attractors of multivalued semiprocesses and nonautonomous evolution inclusions. Set-Valued Anal. **8**, 375–403 (2000)
4. Melnik, V.S., Valero, J.: On attractors of multi-valued semi-flows and differential inclusions. Set-Valued Anal. **6**, 83–111 (1998)
5. Zgurovsky, M.Z., Kasyanov, P.O., Kapustyan, O.V., Valero, J., Zadoianchuk, N.V.: Evolution Inclusions and Variation Inequalities for Earth Data Processing III: Long-time Behavior of Evolution Inclusions Solutions in Earth Data Analysis. Springer, Berlin (2012)
6. Kapustyan, O.V., Valero, J.: Comparison between trajectory and global attractors for evolution systems without uniqueness of solutions. Internat. J. Bifur. Chaos. **20**, 2723–2734 (2010)
7. Babin, A.V., Vishik, M.I.: Attractors of Evolution Equations. Nauka, Moscow (1989)
8. Chepyzhov, V.V., Vishik, M.I.: Attractors for Equations of Mathematical Physics. American Mathematical Society, Providence (2002)
9. Brunovsky, P., Fiedler, B.: Connecting orbits in scalar reaction diffusion equations. Dyn. Reported **1**, 57–89 (1988)
10. Rocha, C., Fiedler, B.: Heteroclinic orbits of semilinear parabolic equations. J. Differ. Equ. **125**, 239–281 (1996)
11. Hale, J.K.: Asymptotic Behavior of Dissipative Systems. American Mathematical Society, Providence (1988)
12. Lions, J.L.: Quelques Méthodes de Résolution des Problèmes aux Limites non Linéaires. Gauthier-Villar, Paris (1969)
13. Rocha, C.: Examples of attractors in scalar reaction-diffusion equations. J. Differ. Equ. **73**, 178–195 (1988)
14. Rocha, C.: Properties of the attractor of a scalar parabolic PDE. J. Dyn. Differ. Equ. **3**, 575–591 (1991)
15. Sell, G.R., You, Y.: Dynamics of Evolutionary Equations. Springer, New York (2002)
16. Temam, R.: Infinite-Dimensional Dynamical Systems in Mechanics and Physics. Springer-Verlag, New York (1997)
17. Vishik, M.I., Zelik, S.V., Chepyzhov, V.V.: Strong trajectory attractor of dissipative reaction-diffusion system. Doklady RAN. **435**(2), 155–159 (2010)
18. Anguiano, M., Caraballo, T., Real, J., Valero, J.: Pullback attractors for reaction-diffusion equations in some unbounded domains with an H^{-1}-valued non-autonomous forcing term and without uniqueness of solutions. Discrete Contin. Dyn. Syst. B **14**, 307–326 (2010)
19. Kapustyan, O.V., Valero, J.: On the Kneser property for the complex Ginzburg-Landau equation and the Lotka-Volterra system with diffusion. J. Math. Anal. Appl. **357**, 254–272 (2009)
20. Ball, J.M.: Global attractors for damped semilinear wave equations. Discrete Contin. Dyn. Syst. **10**, 31–52 (2004)
21. Caraballo, T., Marin-Rubio, P., Robinson, J.: A comparison between two theories for multivalued semiflows and their asymptotic behavior. Set-valued anal. **11**, 297–322 (2003)
22. Wang, Y., Zhou, S.: Kernel sections and uniform attactors of multi-valued semiprocesses. J. Differ. Equ. **232**, 573–622 (2007)

Chapter 13
Topological Properties of Strong Solutions for the 3D Navier-Stokes Equations

Pavlo O. Kasyanov, Luisa Toscano and Nina V. Zadoianchuk

Abstract In this chapter we give a criterion for the existence of global strong solutions for the 3D Navier-Stokes system for any regular initial data.

13.1 Introduction

Let $\Omega \subseteq \mathbb{R}^3$ be a bounded open set with sufficiently smooth boundary $\partial\Omega$ and $0 < T < +\infty$. We consider the incompressible Navier-Stokes equations

$$\begin{cases} y_t + (y \cdot \nabla)y = \nu\triangle y - \nabla p + f \text{ in } Q = \Omega \times (0, T), \\ \text{div } y = 0 \text{ in } Q, \\ y = 0 \text{ on } \partial\Omega \times (0, T), \qquad y(x, 0) = y_0(x) \text{ in } \Omega, \end{cases} \tag{13.1}$$

where $\nu > 0$ is a constant. We define the usual function spaces

$$\mathscr{V} = \{u \in (C_0^\infty(\Omega))^3 : \text{div } u = 0\},$$

$$H = \text{closure of } \mathscr{V} \text{ in } (L^2(\Omega))^3, \quad V = \{u \in (H_0^1(\Omega))^3 : \text{ div } u = 0\}.$$

P. O. Kasyanov (✉)
Institute for Applied System Analysis, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kyiv 03056, Ukraine
e-mail: kasyanov@i.ua

L. Toscano
Department of Mathematics and Applications R. Caccioppoli, University of Naples "Federico II", via Claudio 21, 80125 Naples, Italy
e-mail: luisatoscano@libero.it

N. V. Zadoianchuk
Department of Computational Mathematics, Taras Shevchenko National University of Kyiv, Volodimirska Street 64, Kyiv 03601, Ukraine
e-mail: ninellll@i.ua

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_13,

We denote by V^* the dual space of V. The spaces H and V are separable Hilbert spaces and $V \subset H \subset V^*$ with dense and compact embedding when H is identified with its dual H^*. Let $(\cdot,\cdot)$, $\|\cdot\|_H$ and $((\cdot,\cdot))$, $\|\cdot\|_V$ be the inner product and the norm in H and V, respectively, and let $\langle\cdot,\cdot\rangle$ be the pairing between V and V^*. For $u, v, w \in V$, the equality

$$b(u,v,w) = \int\limits_{\Omega} \sum_{i,j=1}^{3} u_i \frac{\partial v_j}{\partial x_i} w_j dx$$

defines a trilinear continuous form on V with $b(u,v,v) = 0$ when $u \in V$ and $v \in (H_0^1(\Omega))^3$. For $u, v \in V$, let $B(u,v)$ be the element of V^* defined by $\langle B(u,v), w\rangle = b(u,v,w)$ for all $w \in V$.

We say that the function y is a *weak solution* of Problem (13.1) on $[0,T]$, if $y \in L^\infty(0,T;H) \cap L^2(0,T;V)$, $\frac{dy}{dt} \in L^1(0,T;V^*)$, if

$$\frac{d}{dt}(y,v) + \nu((y,v)) + b(y,y,v) = \langle f, v\rangle \quad \text{for all } v \in V, \tag{13.2}$$

in the sense of distributions on $(0,T)$, and if y satisfies the energy inequality

$$V(y)(t) \le V(y)(s) \quad \text{for all } t \in [s,T], \tag{13.3}$$

for a.e. $s \in (0,T)$ and for $s = 0$, where

$$V(y)(t) := \frac{1}{2}\|y(t)\|_H^2 + \nu \int\limits_0^t \|y(\tau)\|_V^2 d\tau - \int\limits_0^t \langle f(\tau), y(\tau)\rangle d\tau. \tag{13.4}$$

This class of solutions is called Leray–Hopf or physical one. If $f \in L^2(0,T;V^*)$, and if y satisfies (13.2), then $y \in C([0,T];H_w)$, $\frac{dy}{dt} \in L^{\frac{4}{3}}(0,T;V^*)$, where H_w denotes the space H endowed with the weak topology. In particular, the initial condition $y(0) = y_0$ makes sense for any $y_0 \in H$.

Let $A : V \to V^*$ be the linear operator associated to the bilinear form $((u,v)) = \langle Au, v\rangle$. Then A is an isomorphism from $D(A)$ onto H with $D(A) = (H^2(\Omega))^3 \cap V$. We recall that the embedding $D(A) \subset V$ is dense and continuous. Moreover, we assume $\|Au\|_H$ as the norm on $D(A)$, which is equivalent to the one induced by $(H^2(\Omega))^3$. The Problem (13.1) can be rewritten as

$$\begin{cases} \frac{dy}{dt} + \nu Ay + B(y,y) = f \text{ in } V^*, \\ y(0) = y_0, \end{cases} \tag{13.5}$$

where the first equation we understand in the sense of distributions on $(0,T)$. Now we write

$\mathscr{D}(y_0, f) = \{\, y\colon\ y \text{is a weak solution of Problem (13.1) on } [0, T]\}$.

It is well known (cf. [1]) that if $f \in L^2(0, T; V^*)$, and if $y_0 \in H$, then $\mathscr{D}(y_0, f)$ is not empty.

A weak solution y of Problem (13.1) on $[0, T]$ is called a *strong* one, if it additionally belongs to Serrin's class $L^8(0, T; (L^4(\Omega))^3)$. We note that any strong solution y of Problem (13.1) on $[0, T]$ belongs to $C([0, T]; V) \cap L^2(0, T; D(A))$ and $\frac{dy}{dt} \in L^2(0, T; H)$ (cf. [2, Theorem 1.8.1, p. 296] and references therein).

For any $f \in L^\infty(0, T; H)$ and $y_0 \in V$ it is well known the only local existence of strong solutions for the 3D Navier-Stokes equations (cf. [1–4] and references therein). Here we provide a criterion for existence of strong solutions for Problem (13.1) on $[0, T]$ for any initial data $y_0 \in V$ and $0 < T < +\infty$. Presented results were announced in [5].

13.2 Topological Properties of Strong Solutions

The main result of this note has the following form.

Theorem 13.1 *Let $f \in L^2(0, T; H)$ and $y_0 \in V$. Then either for any $\lambda \in [0, 1]$ there is an $y_\lambda \in C([0, T]; V) \cap L^2(0, T; D(A))$ such that $y_\lambda \in \mathscr{D}(\lambda y_0, \lambda f)$, or the set*

$$\{y \in C([0, T]; V) \cap L^2(0, T; D(A)) \ :\ y \in \mathscr{D}(\lambda y_0, \lambda f), \lambda \in (0, 1)\} \tag{13.6}$$

is unbounded in $L^8(0, T; (L^4(\Omega))^3)$.

In the proof of Theorem 13.1 we use an auxiliary statement connected with continuity property of strong solutions on parameters of Problem (13.1) in Serrin's class $L^8(0, T; (L^4(\Omega))^3)$.

Theorem 13.2 *Let $f \in L^2(0, T; H)$ and $y_0 \in V$. If y is a strong solution for Problem (13.1) on $[0, T]$, then there exist L, $\delta > 0$ such that for any $z_0 \in V$ and $g \in L^2(0, T; H)$, satisfying the inequality*

$$\|z_0 - y_0\|_V^2 + \|g - f\|_{L^2(0,T;H)}^2 < \delta, \tag{13.7}$$

the set $\mathscr{D}(z_0, g)$ is one-point set $\{z\}$ which belongs to $C([0, T]; V) \cap L^2(0, T; D(A))$, and

$$\|z - y\|_{C([0,T];V)}^2 + \frac{\nu}{4}\|z - y\|_{D(A)}^2 \le L\left(\|z_0 - y_0\|_V^2 + \|g - f\|_{L^2(0,T;H)}^2\right). \tag{13.8}$$

Remark 13.1 We note that from Theorem 13.2 with $z_0 \in V$ and $g \in L^2(0, T; H)$ with $\|z_0\|_V^2 + \|g\|_{L^2(0,T;H)}^2$ sufficiently small, Problem (13.1) has only one global strong solution.

Remark 13.2 Theorem 13.2 provides that, if for any $\lambda \in [0,1]$ there is an $y_\lambda \in L^8(0,T;(L^4(\Omega))^3)$ such that $y_\lambda \in \mathscr{D}(\lambda y_0, \lambda f)$, then the set

$$\{y \in C([0,T];V) \cap L^2(0,T;D(A)) \,:\, y \in \mathscr{D}(\lambda y_0, \lambda f), \lambda \in (0,1)\}$$

is bounded in $L^8(0,T;(L^4(\Omega))^3)$.

If Ω is a C^∞-domain and if $f \in C_0^\infty(\overline{(0,T)\times\Omega})^3$, then any strong solution y of Problem (13.1) on $[0,T]$ belongs to $C^\infty((0,T]\times\Omega)^3$ and $p \in C^\infty((0,T]\times\Omega)$ (cf. [2, Theorem 1.8.2, p. 300] and references therein). This fact directly provides the next corollary of Theorems 13.1 and 13.2.

Corollary 13.1 *Let Ω be a C^∞-domain, $f \in C_0^\infty(\overline{(0,T)\times\Omega})^3$. Then either for any $y_0 \in V$ there is a strong solution of Problem* (13.1) *on* $[0,T]$*, or the set*

$$\{y \in C^\infty((0,T]\times\Omega)^3 \,:\, y \in \mathscr{D}(\lambda y_0, \lambda f), \lambda \in (0,1)\}$$

is unbounded in $L^8(0,T;(L^4(\Omega))^3)$ for some $y_0 \in C_0^\infty(\Omega)^3$.

13.3 Proof of Theorem 13.2

Let $f \in L^2(0,T;H)$, $y_0 \in V$, and $y \in C([0,T];V) \cap L^2(0,T;D(A))$ be a strong solution of Problem (13.1) on $[0,T]$. Due to [6], [1, Chap. 3] the set $\mathscr{D}(y_0,f) = \{y\}$. Let us now fix $z_0 \in V$ and $g \in L^2(0,T;H)$ satisfying (13.7) with

$$\delta = \min\left\{1; \frac{\nu}{4}\right\} e^{-2TC}, \;\; C = \max\left\{\frac{27c^4}{2\nu^3}; \frac{7^7 c^8}{2^9\nu^7}\right\}\left(\|y\|^4_{C([0,T];V)} + 1\right)^2, \quad (13.9)$$

$c > 0$ is a constant from the inequalities (cf. [2, 1])

$$|b(u,v,w)| \le c\|u\|_V \|v\|^{\frac{1}{2}}_V \|v\|^{\frac{1}{2}}_{D(A)} \|w\|_H \quad \forall u \in V,\; v \in D(A),\; w \in H; \quad (13.10)$$

$$|b(u,v,w)| \le c\|u\|^{\frac{3}{4}}_{D(A)} \|u\|^{\frac{1}{4}}_V \|v\|_V \|w\|_H \quad \forall u \in D(A),\; v \in V,\; w \in H. \quad (13.11)$$

The auxiliary Problem

$$\begin{cases} \frac{d\eta}{dt} + \nu A\eta + B(\eta,\eta) + B(y,\eta) + B(\eta,y) = g - f \text{ in } V^*, \\ \eta(0) = z_0 - y_0, \end{cases} \quad (13.12)$$

has a strong solution $\eta \in C([0,T];V) \cap L^2(0,T;D(A))$ with $\frac{d\eta}{dt} \in L^2(0,T;H)$, i.e.

$$\frac{d}{dt}(\eta, v)+\nu((\eta, v))+b(\eta, \eta, v)+b(y, \eta, v)+b(\eta, y, v)=\langle g-f, v\rangle \quad \text{for all } v \in V,$$

in the sense of distributions on $(0, T)$. In fact, let $\{w_j\}_{j\geq 1} \subset D(A)$ be a special basis (cf. [7, p. 56]), i.e. $Aw_j = \lambda_j w_j$, $j = 1, 2, ..., 0 < \lambda_1 \leq \lambda_2 \leq \cdots$, $\lambda_j \to +\infty$, $j \to +\infty$. We consider Galerkin approximations $\eta_m : [0, T] \to \text{span}\{w_j\}_{j=1}^m$ for solutions of Problem (13.12) satisfying

$$\frac{d}{dt}(\eta_m, w_j) + \nu((\eta_m, w_j)) + b(\eta_m, \eta_m, w_j) + b(y, \eta_m, w_j) + b(\eta_m, y, w_j) = \langle g-f, w_j\rangle,$$

with $(\eta_m(0), w_j) = (z_0 - y_0, w_j)$, $j = \overline{1, m}$. Due to (13.10), (13.11) and Young's inequality we get

$$2\langle g-f, A\eta_m\rangle \leq 2\|g-f\|_H\|\eta_m\|_{D(A)} \leq \frac{\nu}{4}\|\eta_m\|^2_{D(A)} + \frac{4}{\nu}\|f-g\|^2_H;$$

$$-2b(\eta_m, \eta_m, A\eta_m) \leq 2c\|\eta_m\|_V^{\frac{3}{2}}\|\eta_m\|^{\frac{3}{2}}_{D(A)} \leq \frac{\nu}{2}\|\eta_m\|^2_{D(A)} + \frac{27c^4}{2\nu^3}\|\eta_m\|^6_V;$$

$$-2b(y, \eta_m, A\eta_m) \leq 2c\|y\|_V\|\eta_m\|_V^{\frac{1}{2}}\|\eta_m\|^{\frac{3}{2}}_{D(A)} \leq \frac{\nu}{2}\|\eta_m\|^2_{D(A)} + \frac{27c^4}{2\nu^3}\|y\|^4_{C([0,T];V)}\|\eta_m\|^2_V;$$

$$-2b(\eta_m, y, A\eta_m) \leq 2c\|\eta_m\|_{D(A)}^{\frac{7}{4}}\|\eta_m\|_V^{\frac{1}{4}}\|y\|_V \leq \frac{\nu}{2}\|\eta_m\|^2_{D(A)} + \frac{7^7c^8}{2^9\nu^7}\|y\|^8_{C([0,T];V)}\|\eta_m\|^2_V.$$

Thus,

$$\frac{d}{dt}\|\eta_m\|^2_V + \frac{\nu}{4}\|\eta_m\|^2_{D(A)} \leq C(\|\eta_m\|^2_V + \|\eta_m\|^6_V) + \frac{4}{\nu}\|g-f\|^2_H,$$

where $C > 0$ is a constant from (13.9). Hence, the absolutely continuous function $\varphi = \min\{\|\eta_m\|^2_V, 1\}$ satisfies the inequality $\frac{d}{dt}\varphi \leq 2C\varphi + \frac{4}{\nu}\|g-f\|^2_H$, and therefore $\varphi \leq L(\|z_0-y_0\|^2_V+\|g-f\|^2_{L^2(0,T;H)}) < 1$ on $[0, T]$, where $L = \delta^{-1}$. Thus, $\{\eta_n\}_{n\geq 1}$ is bounded in $L^\infty(0, T; V) \cap L^2(0, T; D(A))$ and $\{\frac{d}{dt}\eta_n\}_{n\geq 1}$ is bounded in $L^2(0, T; H)$. In a standard way we get that the limit function η of $\eta_n, n \to +\infty$, is a strong solution of Problem (13.12) on $[0, T]$. Due to [6], [1, Chap. 3] the set $\mathscr{D}(z_0, g)$ is one-point $z = y + \eta \in L^8(0, T; (L^4(\Omega))^3)$. So, z is strong solution of Problem (13.1) on $[0, T]$ satisfying (13.8).

The theorem is proved.

13.4 Proof of Theorem 13.1

We provide the proof of Theorem 13.1. Let $f \in L^2(0, T; H)$ and $y_0 \in V$. We consider the 3D controlled Navier-Stokes system (cf. [8, 9])

$$\begin{cases} \frac{dy}{dt} + \nu Ay + B(z, y) = f, \\ y(0) = y_0, \end{cases} \tag{13.13}$$

where $z \in L^8(0, T; (L^4(\Omega))^3)$.

By using standard Galerkin approximations (see [1]) it is easy to show that for any $z \in L^8(0, T; (L^4(\Omega))^3)$ there exists an unique weak solution $y \in L^\infty(0, T; H) \cap L^2(0, T; V)$ of Problem (13.13) on $[0, T]$, that is,

$$\frac{d}{dt}(y, v) + \nu((y, v)) + b(z, y, v) = \langle f, v \rangle, \text{ for all } v \in V, \tag{13.14}$$

in the sense of distributions on $(0, T)$. Moreover, by the inequality

$$|b(u, v, Av)| \le c_1 \|u\|_{(L^4(\Omega))^3} \|v\|_V^{\frac{1}{4}} \|v\|_{D(A)}^{\frac{7}{4}} \le \frac{\nu}{2} \|v\|_{D(A)}^2 + c_2 \|u\|_{(L^4(\Omega))^3}^8 \|v\|_V^2, \tag{13.15}$$

for all $u \in (L^4(\Omega))^3$ and $v \in D(A)$, where $c_1, c_2 > 0$ are some constants that do not depend on u, v (cf. [1]), we find that $y \in C([0, T]; V) \cap L^2(0, T; D(A))$ and $B(z, y) \in L^2(0, T; H)$, so $\frac{dy}{dt} \in L^2(0, T; H)$ as well. We add that, for any $z \in L^8(0, T; (L^4(\Omega))^3)$ and corresponding weak solution $y \in C([0, T]; V) \cap L^2(0, T; D(A))$ of (13.13) on $[0, T]$, by using Gronwall inequality, we obtain

$$\begin{gathered} \|y(t)\|_V^2 \le \|y_0\|_V^2 e^{2c_2 \int_0^t \|z(t)\|_{(L^4(\Omega))^3}^8 dt}, \quad \forall t \in [0, T]; \\ \nu \int_0^T \|y(t)\|_{D(A)}^2 dt \le \|y_0\|_V^2 \left[1 + 2c_2 e^{2c_2 \int_0^T \|z(t)\|_{(L^4(\Omega))^3}^8 dt} \|z\|_{L^8(0,T;(L^4(\Omega))^3)}^8 \right]. \end{gathered} \tag{13.16}$$

Let us consider the operator $F : L^8(0, T; (L^4(\Omega))^3) \to L^8(0, T; (L^4(\Omega))^3)$, where $F(z) \in C([0, T]; V) \cap L^2(0, T; D(A))$ is the unique weak solution of (13.13) on $[0, T]$ corresponded to $z \in L^8(0, T; (L^4(\Omega))^3)$.

Let us check that F is a compact transformation of Banach space $L^8(0, T; (L^4(\Omega))^3)$ into itself (cf. [10]). In fact, if $\{z_n\}_{n \ge 1}$ is a bounded sequence in $L^8(0, T; (L^4(\Omega))^3)$, then, due to (13.15) and (13.16), the respective weak solutions y_n, $n = 1, 2, ...$, of Problem (13.13) on $[0, T]$ are uniformly bounded in $C([0, T]; V) \cap L^2(0, T; D(A))$ and their time derivatives $\frac{dy_n}{dt}$, $n = 1, 2, ...$, are uniformly bounded in $L^2(0, T; H)$. So, $\{F(z_n)\}_{n \ge 1}$ is a precompact set in $L^8(0, T; (L^4(\Omega))^3)$. In a standard way we deduce that $F : L^8(0, T; (L^4(\Omega))^3) \to L^8(0, T; (L^4(\Omega))^3)$ is continuous mapping.

Since F is a compact transformation of $L^8(0, T; (L^4(\Omega))^3)$ into itself, Schaefer's Theorem (cf. [10, p. 133] and references therein) and Theorem 13.2 provide the statement of Theorem 13.1. We note that Theorem 13.2 implies that the set

$\{z \in L^8(0,T;(L^4(\Omega))^3) : z = \lambda F(z),\ \lambda \in (0,1)\}$ is bounded in $L^8(0,T;(L^4(\Omega))^3)$ iff the set defined in (13.6) is bounded in $L^8(0,T;(L^4(\Omega))^3)$.

The theorem is proved.

Acknowledgments The authors thank Professors J.M. Ball, V.V. Chepyzhov, and M.Z. Zgurovsky for useful suggestions during the preparation of this manuscript. The first author was partially supported by the Ukrainian State Fund for Fundamental Researches under grants GP/F44/076, GP/F49/070, and by the NAS of Ukraine under grant 2273/13.

References

1. Temam, R.: Navier-Stokes Equations. North-Holland, Amsterdam (1979)
2. Sohr, H.: The Navier-Stokes Equations. An Elementary Functional Analytic Approach. Verlag, Birkhäuser (2001)
3. Zgurovsky, M.Z., Kasyanov, P.O., Kapustyan, O.V., Valero, J., Zadoianchuk, N.V.: Evolution Inclusions and Variation Inequalities for Earth Data Processing III. Springer, Berlin (2012)
4. Ponce, G., Rascke, R., Sideris, T., Titi, E.: Global stability of large solutions to the 3D Navier-Stokes equations. Commun. Math. Phys. **159**, 329–341 (1994)
5. Kasyanov, P.O., Toscano, L., Zadoianchuk, N.V.: A criterion for the existence of strong solutions for the 3D Navier-Stokes equations. Appl. Math. Lett. **26**(1), 15–17 (2013)
6. Serrin, J.: The initial value problem for the Navier-Stokes equations. In: Langer, R.E. (ed.) Nonlinear Problems, pp. 69–98. University of Wisconsin Press, Madison (1963)
7. Temam, R.: Infinite-Dimensional Dynamical Systems in Mechanics and Physics. Springer, New York (1988)
8. Melnik, V.S., Toscano, L.: On weak extensions of extreme problems for nonlinear operator equations. Part I. Weak solutions. J. Automat. Inf. Scien. **38**, 68–78 (2006)
9. Kapustyan, O.V., Kasyanov, P.O., Valero, J.: Pullback attractors for a class of extremal solutions of the 3D Navier-Stokes system. J. Math. Anal. Appl. **373**, 535–547 (2011)
10. Cronin, J.: Fixed Points and Topological Degree in Nonlinear Analysis. American Mathematical Society, Providence (1964)

Chapter 14
Inertial Manifolds and Spectral Gap Properties for Wave Equations with Weak and Strong Dissipation

Natalia Chalkina

Abstract Sufficient conditions for the existence of an inertial manifold for the equation $u_{tt} - 2\gamma_s \Delta u_t + 2\gamma_w u_t - \Delta u = f(u)$, $\gamma_s > 0$, $\gamma_w \geq 0$ are found. The nonlinear function f is supposed to satisfy Lipschitz property. The proof is based on construction of a new inner product in the phase space in which the conditions of a general theorem on the existence of inertial manifolds for an abstract differential equation in a Hilbert space are satisfied.

14.1 Introduction

In the theory of nonlinear evolution partial differential equations, great attention is paid to long-time behavior of dynamic systems. Some way of such description relates with notion of an inertial manifold (see [5, 6, 9]).

Let us consider an initial-value problem for an abstract differential equation in a Hilbert space,

$$\frac{d}{dt}y + \mathbf{A}y = F(y), \qquad y \in \mathscr{H}, \tag{14.1}$$

$$y\big|_{t=0} = y_0 \in \mathscr{H}. \tag{14.2}$$

Here $\mathbf{A}$ is a linear operator and F is a nonlinear operator. Suppose problem (14.1), (14.2) has a unique solution y for any $y_0 \in \mathscr{H}$. Hence, this problem generates a continuous semigroup $\{S(t) \mid t \geq 0\}$, acting in the space $\mathscr{H}$ by the formula $S(t)y_0 = y(t) \in \mathscr{H}$.

Definition 14.1 A Lipschitz finite dimensional manifold $\mathscr{M} \subset \mathscr{H}$ is an *inertial manifold* for the semigroup $S(t)$ if it is invariant (i.e., $S(t)\mathscr{M} = \mathscr{M}$, $\forall t \geq 0$) and it satisfies the following asymptotic completeness property:

N. Chalkina (✉)
Department of Mechanics and Mathematics, Nikulinskaya, 15-2, Moscow 119602, Russia
e-mail: chalkinan@mail.ru

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_14,

$$\forall y_0 \in \mathscr{H} \exists \tilde{y}_0 \in \mathscr{M} \text{ such that } \|S(t)y_0 - S(t)\tilde{y}_0\|_{\mathscr{H}} \le q(\|y_0\|_{\mathscr{H}})e^{-ct}, t \ge 0,$$

where the positive constant c and the monotonic function q are independent of y_0.

Inertial manifolds enable one to reduce the study of the behavior of an infinite-dimensional dynamical system to the investigation of this problem for some finite-dimensional dynamical system generated by original system on an inertial manifold.

For the abstract equation of the form (14.1), there are known sufficient conditions under which there is an inertial manifold in the Hilbert space $\mathscr{H}$ (see [3]). Let us present these conditions. Let $\mathbf{A}$ be a linear closed (possibly unbounded) operator with dense domain $\mathscr{D}(\mathbf{A})$ in $\mathscr{H}$ and let the spectrum $\sigma(\mathbf{A})$ of $\mathbf{A}$ be disjoint from the strip $\{m < \Re\zeta < M\}$, where $M \ge 0$, $M > m$. Denote by P the orthogonal projection to the invariant subspace of $\mathbf{A}$ corresponding to the part of the spectrum $\sigma \cap \{\Re\zeta \le m\}$ and write $Q = \mathrm{Id} - P$. Assume that the space $P(\mathscr{H})$ is finite-dimensional.

Theorem 14.1 *Let the space $\mathscr{H}$ be equiped with an inner product in such a way that the space $P(\mathscr{H})$ and $Q(\mathscr{H})$ are orthogonal and the following relations hold:*

$$\begin{aligned} &(\mathbf{A}y, y) \le m|y|^2 \quad \forall y \in P(\mathscr{H}), \\ &(\mathbf{A}y, y) \ge M|y|^2 \quad \forall y \in Q(\mathscr{H}) \cap \mathscr{D}(\mathbf{A}). \end{aligned} \tag{14.3}$$

Moreover, let $F(y)$ be a nonlinear function such that $F(0) = 0$ and let F satisfy the Lipschitz condition with the constant L, where

$$2L < M - m. \tag{14.4}$$

In this case, there is an inertial manifold $\mathscr{M}$ in the Hilbert space $\mathscr{H}$, and this manifold is the graph of a Lipschitz continuous function $\Phi\colon P(H) \to Q(H)$.

In the present chapter, an initial-boundary value problem for a wave equation with weak and strong dissipation is considered. The nonlinear term depends on the unknown function u, these term is assumed to be Lipschitzian,

$$u_{tt} - 2\gamma_s \Delta u_t + 2\gamma_w u_t - \Delta u = f(u).$$

For this equation, we obtain a condition on the Lipschitz constant of the function f which ensures the existence of an inertial manifold. The result is stated in Theorems 14.2 and 14.3. The proof is based on construction of a new inner product in the phase space in which the conditions of Theorem 14.1 are satisfied.

14.2 Statement of the Problem and Spectrum of the Linear Operator

In a bounded domain Ω, we consider the inertial-boundary value problem for a wave equation with dissipation,

$$u_{tt} - 2\gamma_s \Delta u_t + 2\gamma_w u_t - \Delta u = f(u), \qquad u|_{\partial\Omega} = 0, \tag{14.5}$$

$$u|_{t=0} = u_0(x) \in H_0^1(\Omega), \qquad u_t|_{t=0} = p_0 \in L_2(\Omega). \tag{14.6}$$

Here γ_w and γ_s are positive coefficients of the dissipation, and the nonlinear function f is continuously differentiable and satisfy the global Lipschitz condition,

$$|f(v_1) - f(v_2)| \le l|v_1 - v_2| \qquad \forall v_1, v_2 \in \mathbb{R}, \tag{14.7}$$

Moreover, let $f(0) = g(0) = 0$.

Under these assumptions, problem (14.5), (14.6) has a unique weak solution $u \in C\big([0,T]; H_0^1(\Omega)\big)$, $\partial_t u \in C\big([0,T]; L_2(\Omega)\big)$ for any $T > 0$ (see [7, 8, 10]). Hence, this problem generates a continuous semigroup $\{S(t)\}$, $t \ge 0$, acting in the phase space $\mathscr{H} = H_0^1(\Omega) \times L_2(\Omega)$ by the formula

$$S(t)(u_0(x), p_0(x)) = y(t) \equiv (u(t,x), p(t,x)) \in H,$$

where $u(t,x)$ is a solution of the problem (14.5), (14.6), $p(t,x) = \partial_t u(t,x)$ stands for the derivative of this solution w.r.t. t, and $y = (u, p) \in \mathscr{H}$.

Let us represent the initial-boundary value problem in the form of an ordinary differential equation to find the unknown vector function $y = (u, p) \in \mathscr{H}$,

$$\frac{d}{dt} y(t) + \mathbf{A}y = F(y), \quad \mathbf{A}y = \begin{pmatrix} 0 & -1 \\ -\Delta & 2\gamma_w - 2\gamma_s \Delta \end{pmatrix} y, \quad F(y) = \begin{pmatrix} 0 \\ f(u) \end{pmatrix}.$$

Let $e_k(x)$ and λ_k be the eigenfunctions and the eigenvalues of the operator $-\Delta$ in the domain Ω with the Dirichlet conditions on the boundary,

$$-\Delta e_k(x) = \lambda_k e_k(x), \quad e_k(x)\big|_{\partial\Omega} = 0, \quad e_k(x) \not\equiv 0,$$
$$0 < \lambda_1 < \lambda_2 \le \lambda_3 \le \cdots \to +\infty.$$

Denote by $(\cdot,\cdot)_{\mathscr{H}}$ and $\|\cdot\|$ the standard inner product and the corresponding norm in the space $\mathscr{H}$, namely,

$$(y, \tilde{y})_{\mathscr{H}} = (\nabla u, \nabla \tilde{u}) + (p, \tilde{p}) = \sum_{k=1}^{\infty} (\lambda_k u_k \tilde{u}_k + p_k \tilde{p}_k),$$

where $u_k = (u, e_k)$, $p_k = (p, e_k)$, and $(\cdot,\cdot)$ stands for the inner product in $L_2(\Omega)$.

The two-dimensional subspace $\mathscr{H}_k$ with basis $(e_k, 0)$, $(0, e_k)$ is invariant under the operator $\mathbf{A}$. The restriction of the operator $\mathbf{A}$ to the subspace $\mathscr{H}_k$ has the matrix $A_k = \begin{pmatrix} 0 & -1 \\ \lambda_k & 2(\gamma_w + \gamma_s \lambda_k) \end{pmatrix}$. The eigenvalues of A_k are equal to

$$\mu_k = \gamma_k - \sqrt{\gamma_k^2 - \lambda_k} \quad \text{and} \quad \nu_k = \gamma_k + \sqrt{\gamma_k^2 - \lambda_k}$$

where we denote $\gamma_k = \gamma_w + \gamma_s \lambda_k$. In Figs. 14.1 and 14.2, we show the qualitative displacement of these eigenvalues on the complex plane in two cases, namely, $4\gamma_w\gamma_s < 1$ and $4\gamma_w\gamma_s \geq 1$. In the first case, the operator A has both real and nonreal eigenvalues and, in the other case, all eigenvalues are real.

If the orthogonal projection P satisfies the assumptions of the Theorem 14.1, then the image $P(\mathscr{H})$ (which is finite-dimensional) must correspond to finitely many eigenvalues of $\mathbf{A}$ belonging to the domain $\{\mathrm{Re}\zeta \leq m\}$. However, $\mu_k \to 1/(2\gamma_s)$ and

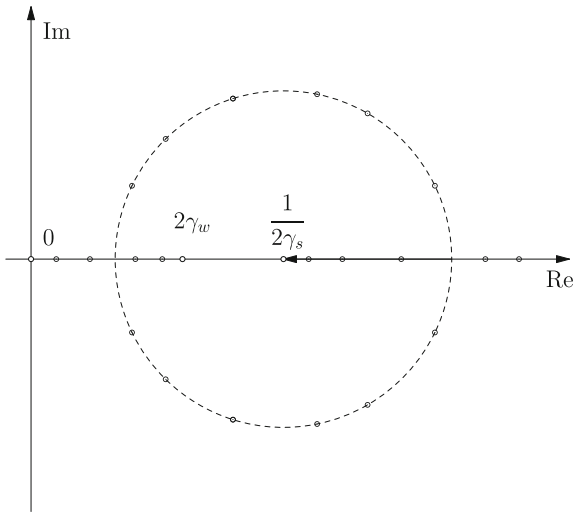


Fig. 14.1 $4\gamma_w\gamma_s < 1$

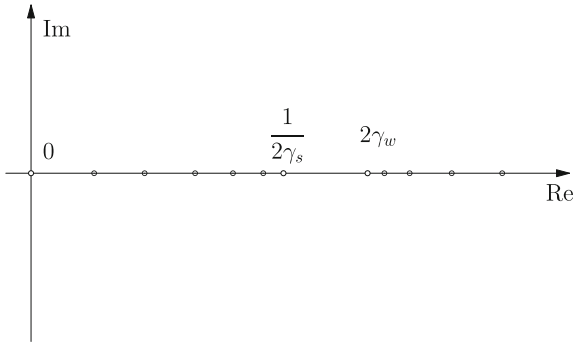


Fig. 14.2 $4\gamma_w\gamma_s > 1$

$\nu_k \to +\infty$ as $\lambda_k \to +\infty$, and thus the quantity m must be less than $1/(2\gamma_s)$. In the case $4\gamma_w\gamma_s < 1$, to the values μ_k and ν_k lying to the left of the accumulation point $1/(2\gamma_s)$ there correspond values $\lambda_k < \frac{1-2\gamma_w\gamma_s}{2\gamma_s^2}$. If $4\gamma_w\gamma_s \geq 1$, then $\mu_k < 1/(2\gamma_s)$ for any k.

14.3 Sufficient Conditions for the Existence of Inertial Manifolds

In this section, we present conditions for the existence of a gap both in the real part (Theorem 14.2) and in the nonreal part (Theorem 14.3) of the spectrum of $\mathbf{A}$.

First let us consider a gap in the real part of the spectrum. Thus, for $4\gamma_w\gamma_s < 1$, the additional condition $m < \frac{1-\sqrt{1-4\gamma_w\gamma_s}}{2\gamma_s}$ is imposed, which corresponds to the inequality $\lambda_k < \frac{1-2\gamma_w\gamma_s-\sqrt{1-4\gamma_w\gamma_s}}{2\gamma_s^2}$.

Remark 14.1 If Eq. (14.5) has not strongly dissipative term (i.e., $\gamma_s = 0$), then the circle to which a part of eigenvalues of the operator $\mathbf{A}$ belongs (see Fig. 14.1) is transformed to the vertical line $\{\Re\zeta = \gamma_w\}$ (see Fig. 14.3), and the condition on m becomes $m < \gamma_w$.

Write

$$\gamma_\star = \begin{cases} \gamma_1, & \text{if } 1 \leq 2\gamma_s\gamma_1; \\ 1/(2\gamma_s), & \text{if } 2\gamma_s\gamma_1 \leq 1 \leq 2\gamma_s\gamma_{N+1}; \\ \gamma_{N+1}, & \text{if } 2\gamma_s\gamma_{N+1} \leq 1; \end{cases} \quad \lambda_\star = \frac{\gamma_\star - \gamma_w}{\gamma_s}.$$

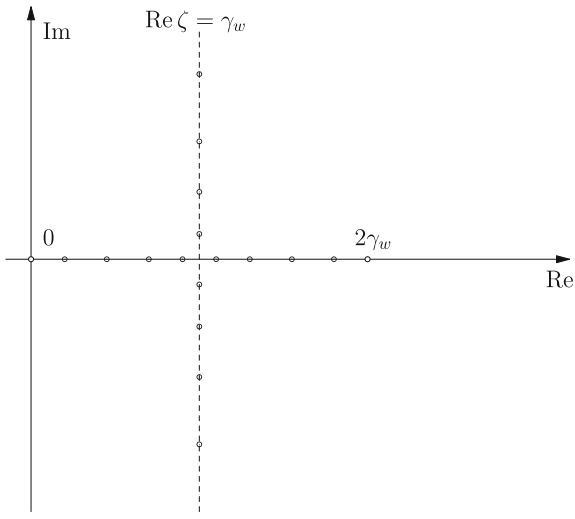


Fig. 14.3 Weak dissipation, $\gamma_s = 0$

Theorem 14.2 *Let f satisfy condition (14.7). Moreover, suppose that there is an N such that the following inequality holds:*

$$2\frac{l}{\sqrt{\gamma_\star^2-\lambda_\star}} < \mu_{N+1}-\mu_N = \gamma_{N+1}-\sqrt{\gamma_{N+1}^2-\lambda_{N+1}}-\gamma_N+\sqrt{\gamma_N^2-\lambda_N}, \quad (14.8)$$

and, if $4\gamma_w\gamma_s < 1$, then the following inequality also holds:

$$\lambda_{N+1} < \frac{1-2\gamma_w\gamma_s-\sqrt{1-4\gamma_w\gamma_s}}{2\gamma_s^2}.$$

In this case, there is an N-dimensional inertial manifold for problem (14.5), (14.6) in the space $\mathscr{H}$.

Remark 14.2 If $\gamma_s = 0$, then condition (14.8) coincides with the similar condition obtained in [4].

Remark 14.3 If there is no weak dissipation, then all real point of the spectrum of the operator $\mathbf{A}$ are located to the right of the number $1/(2\gamma_s)$ (see Fig. 14.4), and Theorem 14.2 cannot be applied to this situation.

Now we consider case of spectral gap in nonreal part of spectrum. Hence we assume that $4\gamma_w\gamma_s < 1$.

Let values m and M be chosen in such a way that

$$\frac{1-\sqrt{1-4\gamma_w\gamma_s}}{2\gamma_s} \le m < M \le \frac{1}{2\gamma_s}, \quad (14.9)$$

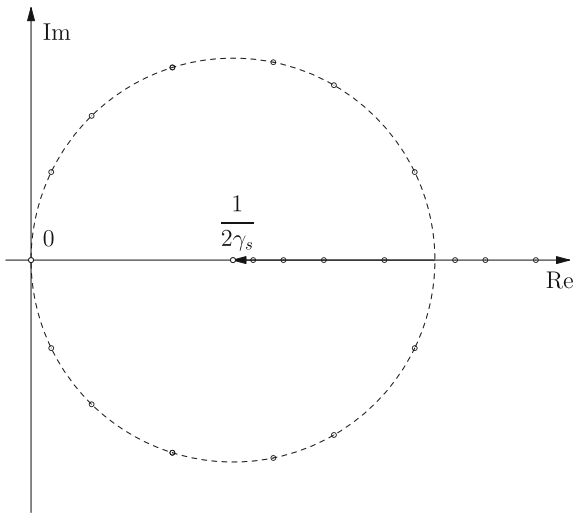


Fig. 14.4 Strong dissipation, $\gamma_w = 0$

and the spectrum $\sigma(\mathbf{A})$ of $\mathbf{A}$ be disjoint from the strip $\{m < \Re\zeta < M\}$, but the set $\sigma(\mathbf{A}) \cap \{\Re\zeta \le m\}$ is not empty.

Let numbers k_1, k_2 are such that values ν_{k_1} and ν_{k_2+1} belong to the domain $\{\Re\zeta \ge M\}$, and numbers ν_{k_1+1} and ν_{k_2} belong to the domain $\{\Re\zeta \le m\}$ (see Fig. 14.5). Thus for $\nu_1 \notin \mathbb{R}$ or $\nu_1 \in \mathbb{R}$, $\nu_1 \le m$ we have $k_1 = 0$; for the converse case we get $\Re\nu_{k_1+1} \le m < M \le \Re\nu_{k_1}$.

If there are not numbers ν_k to the left of the strip, then we have $M \le \Re\nu_{k_2+1}$ and $M \le \Re\nu_{k_1} = \nu_{k_1} = \nu_{k_2}$. Otherwise number k_2 is such that $\Re\nu_{k_2} \le m < M \le \Re\nu_{k_2+1}$.

Denote numbers $\varkappa_I$, $\varkappa_{II}$, $\varkappa_{III}$ and $\varkappa_{IV}$. First if $k_1 = 0$ then formally write $\varkappa_I = +\infty$. In the other case write $\varkappa_I = \sqrt{\gamma_{k_1}^2 - \lambda_{k_1}}$. Secondly if $k_2 = k_1$ then formally write $\varkappa_{II} = \varkappa_{III} = +\infty$. Otherwise denote $\varkappa_{II} = s_{k_1+1}$, $\varkappa_{III} = s_{k_2}$, where

$$s_k = \sqrt{m^2 - 2m\gamma_k + \lambda_k} + m - \gamma_k. \tag{14.10}$$

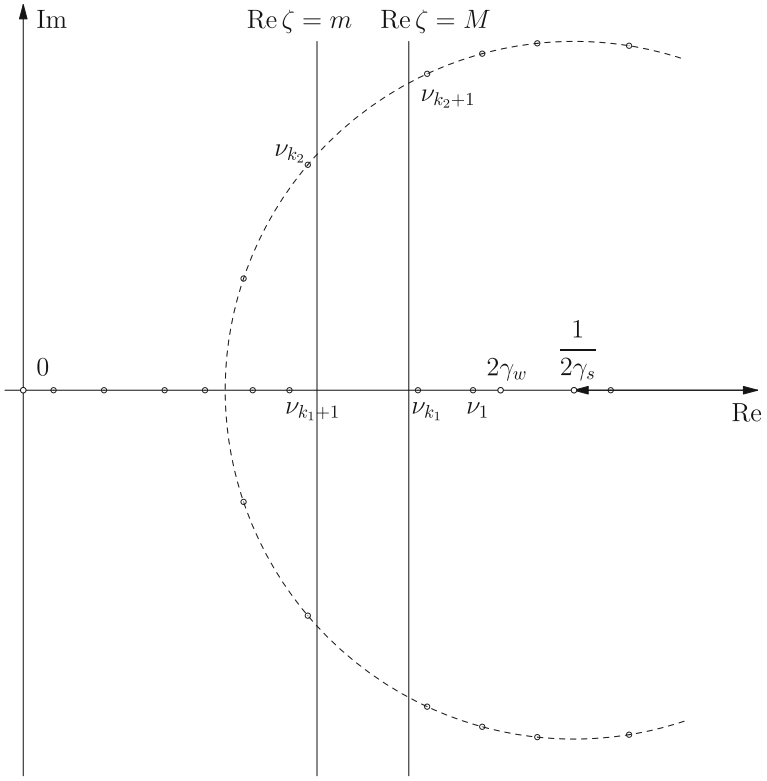


Fig. 14.5 A spectral gap in nonreal part of the spectrum

Finally write $\lambda_M = (M - \gamma_w)/\gamma_s$ and $\varkappa_{IV} = \sqrt{\lambda_M - M^2}$.

Theorem 14.3 *Let nonlinear function f satisfies condition (14.7). Moreover, suppose that the following inequality holds:*

$$2l < (M - m)\min\{\varkappa_I, \varkappa_{II}, \varkappa_{III}, \varkappa_{IV}\}. \tag{14.11}$$

Then there is a $(2k_2 - k_1)$-dimensional inertial manifold for problem (14.5), (14.6) in the space $\mathscr{H}$.

Remark 14.4 It follows from condition (14.11) that there are enough large gaps in the spectrum of operator $-\Delta$ in domain Ω. Actually, we have

$$\varkappa_{IV} = \sqrt{\lambda_M - M^2} = \sqrt{\frac{M - \gamma_w - \gamma_s M^2}{\gamma_s}} =$$
$$= \sqrt{\frac{4\gamma_s M - 4\gamma_w\gamma_s - 4\gamma_s^2 M^2}{4\gamma_s^2}} = \sqrt{\frac{1 - 4\gamma_w\gamma_s - (2\gamma_s M - 1)^2}{4\gamma_s^2}} < \frac{\sqrt{1 - 4\gamma_w\gamma_s}}{2\gamma_s}.$$

Moreover, the inequalities $\gamma_{k_2} \le m$ and $M \le \gamma_{k_2+1}$ hold by definition of the number k_2. Indeed if $\nu_{k_2} \in \mathbb{R}$, then we have $\nu_{k_2} < \frac{1}{2\gamma_s}$, $\gamma_{k_2} < \frac{1-\sqrt{1-4\gamma_w\gamma_s}}{2\gamma_s} \le m$ (see (14.9)); otherwise we have $\gamma_{k_2} = \Re\nu_{k_2} \le m$. Similarly if $\nu_{k_2+1} \in \mathbb{R}$, then we have $\nu_{k_2} > \frac{1}{2\gamma_s}$, $\gamma_{k_2} > \frac{1+\sqrt{1-4\gamma_w\gamma_s}}{2\gamma_s} > M$; otherwise we get $\gamma_{k_2+1} = \Re\nu_{k_2+1} \ge M$.

Thus, by (14.11) it follows the inequality,

$$2l < (\gamma_{k_2+1} - \gamma_{k_2})\frac{\sqrt{1 - 4\gamma_w\gamma_s}}{2\gamma_s} = (\lambda_{k_2+1} - \lambda_{k_2})\frac{\sqrt{1 - 4\gamma_w\gamma_s}}{2}.$$

This means that there are spectral gaps on the order of l:

$$\lambda_{k_2+1} - \lambda_{k_2} > 4l\sqrt{1 - 4\gamma_w\gamma_s}.$$

The proofs of Theorems 14.2 and 14.3 are based on the construction of a new norm in the phase space $\mathscr{H}$, in which the assumptions of Theorem 14.1 are satisfied. Note the schemes of the new inner product construction are essentially different for gaps in the real part and in the nonreal part of the spectrum. Then this two cases are considered separately. In the present chapter we prove Theorem 14.3. The proof of Theorem 14.2 presented in [1].

Remark 14.5 The case of the gap in the nonreal part of the spectrum was partially studied in [2], where a strongly dissipative wave equation (i.e., $\gamma_w = 0$) was considered.

14.4 Proof of Theorem 14.3

Let us decompose the entire phase space $\mathscr{H}$ in direct sum of spaces that are pairwise orthogonal, $\mathscr{H} = \mathscr{H}_1 \oplus \mathscr{H}_2 \oplus \ldots \oplus \mathscr{H}_{k_2} \oplus \mathscr{H}_\infty$, where every subspace $\mathscr{H}_k$, $k = 1, \ldots, k_2$, is two-dimensional and corresponds to the eigenvector e_k with respect to u and p, and $\mathscr{H}_\infty = (\mathscr{H}_1 \oplus \mathscr{H}_2 \oplus \ldots \oplus \mathscr{H}_{k_2})^\perp$ is the subspace of codimension $2k_2$ which corresponds to the eigenvectors $e_{k_2+1}, e_{k_2+2}, \ldots$ of the Laplace operator. Note that the spaces $\mathscr{H}_k$, $k = 1, \ldots, k_2$, and $\mathscr{H}_\infty$ are invariant with respect to the action of the linear operator $\mathbf{A}$.

The new inner product $[\cdot, \cdot]$ introduced below preserves the condition that the spaces $\mathscr{H}_k, k = 1, \ldots, k_2, \infty$, are pairwise orthogonal and modifies the inner product in each of these subspaces. Thus, if $y = (u, p) \in \mathscr{H}$ and the orthogonal projections of y to $\mathscr{H}_k$ are denoted by $y_k = (u_k e_k, p_k e_k) \in \mathscr{H}_k, k = 1, \ldots, k_2, \infty$, then the new norm in $\mathscr{H}$ is defined by the formula

$$|||y|||^2 = \sum_{k=1}^{k_2} |||y_k|||_k^2 + |||y_\infty|||_\infty^2.$$

14.4.1 New Norm in the Spaces $\mathscr{H}_k$, $k = 1, \ldots, k_1$

By definition the number k_1, for $k = 1, \ldots, k_1$ the eigenvalues μ_k and ν_k are real and lie to the different sides of the strip $\{m < \Re\zeta < M\}$. We introduce the new inner product in such a way that the eigenvectors ξ_k and η_k, which correspond to the eigenvalues μ_k and ν_k, are orthogonal with respect to this inner product.

Define a new inner product $[\cdot, \cdot]_k$ of vectors $y = (u, p)$, $\tilde{y} = (\tilde{u}, \tilde{p})$, $y, \tilde{y} \in \mathscr{H}_k$ by the rule

$$[y, \tilde{y}]_k = (2\gamma_k^2 - \lambda_k)(u, \tilde{u}) + \gamma_k(u, \tilde{p}) + \gamma_k(p, \tilde{u}) + (p, \tilde{p}).$$

The following assertions hold.

Lemma 14.1 *The eigenvectors ξ_k and η_k corresponding to the eigenvalues μ_k and ν_k, are orthogonal with respect to the new inner product.*

Proof The eigenvectors of the matrix A_k in the space $\mathscr{H}_k$ are the vectors $\xi_k = (1, -\mu_k)$ and $\eta_k = (1, -\nu_k)$. It follows from $\mu_k + \nu_k = 2\gamma_k$ and $\mu_k \nu_k = \lambda_k$ that

$$[\xi_k, \eta_k]_k = 2\gamma_k^2 - \lambda_k - \gamma_k(\mu_k + \nu_k) + \mu_k \nu_k = 0.$$

Since $\gamma_k^2 > \lambda_k$ for $k \le k_1$, it follows that the new inner product defines the norm

$$|||y|||_k^2 = [y, y]_k = (\gamma_k^2 - \lambda_k)\|u\|^2 + \|\gamma_k u + p\|^2.$$

Let us prove that

Lemma 14.2 *The minimum of the function* $\kappa_1(\gamma) = \gamma^2 - \lambda(\gamma)$, *where* $\lambda(\gamma) = \dfrac{\gamma - \gamma_w}{\gamma_s}$, *on the interval* $\gamma \in [\gamma_1, \gamma_{k_1}]$ *is achieved at the point* $\gamma = \gamma_{k_1}$.

Proof Let us show that the derivative of $\kappa_1(\gamma)$ is negative on the interval $\gamma \in [\gamma_1, \gamma_{k_1}]$. Indeed, by definition of the number k_1 we get $\gamma < \gamma_{k_1} < 1/(2\gamma_s)$. Hence for $\gamma < \gamma_{k_1}$ we have

$$\gamma_s \kappa_{1\gamma}' = 2\gamma\gamma_s - 1 < 0.$$

Thus, the function $\kappa_1(\gamma)$ decreases on the interval $\gamma \in [\gamma_1, \gamma_{k_1}]$, and its minimum is attained at $\gamma = \gamma_{k_1}$.

Since Lemma 14.2 the following estimate of the norm of the vector $y = y_1 + \cdots + y_{k_1}$, $y_k = (u_k, p_k) \in \mathscr{H}_k$, holds

$$|||y|||^2 = \sum_{k=1}^{k_1} |||y_k|||_k^2 \geq \sum_{k=1}^{k_1} (\gamma_k^2 - \lambda_k)\|u_k\|^2 \geq \min_{1 \leq k \leq k_1} \left\{\gamma_k^2 - \lambda_k\right\} \cdot \sum_{k=1}^{k_1} \|u_k\|^2 =$$
$$= (\gamma_{k_1}^2 - \lambda_{k_1})\|u\|^2 = \varkappa_1^2 \|u\|^2. \tag{14.12}$$

14.4.2 New Norm in the Spaces $\mathscr{H}_k$, $k = k_1 + 1, \ldots, k_2$

By definition the numbers k_1, k_2 for $k = k_1 + 1, \ldots, k_2$ the eigenvalues μ_k and ν_k belong to the domain $\{\Re\zeta < m\}$. In this section, we introduce the new inner product $[\cdot, \cdot]_k$ in the spaces $\mathscr{H}_k$, $k = k_1 + 1, \ldots, k_2$, in such a way that $[\mathbf{A}y, y]_k \leq m[y, y]_k$ for any vector $y \in \mathscr{H}_k$.

Define the new inner product $[\cdot, \cdot]_k$ of the vectors $y = (u, p)$, $\tilde{y} = (\tilde{u}, \tilde{p})$, $y, \tilde{y} \in \mathscr{H}_k$ by the rule

$$[y, \tilde{y}]_k = b_k(u, \tilde{u}) + \gamma_k(u, \tilde{p}) + \gamma_k(p, \tilde{u}) + (p, \tilde{p}),$$

where $b_k = \gamma_k^2 + s_k^2$ and the numbers s_k are defined in (14.10).

Define the auxiliary function

$$s(\gamma) = \sqrt{m^2 - 2\gamma m + \lambda(\gamma)} + m - \gamma,$$

where $\lambda(\gamma) = (\gamma - \gamma_w)/\gamma_s$. Then $s(\gamma_k) = s_k$. For $\gamma \in [\gamma_{k_1+1}, \gamma_{k_2}]$ the value $s(\gamma)$ is real. Actually, by the choice of k_1, k_2 we have $m \geq \Re\nu = \Re\left(\gamma + \sqrt{\gamma^2 - \lambda(\gamma)}\right)$ for $\gamma \in [\gamma_{k_1+1}, \gamma_{k_2}]$. Hence $m \geq \gamma$, $m^2 - 2\gamma m + \lambda(\gamma) \geq 0$.

Since the numbers s_k are real, we see that the inner product defines the norm

$$|||y|||_k^2 = [y, y]_k = s_k^2\|u\|^2 + \|\gamma_k u + p\|^2.$$

The following assertions hold.

Lemma 14.3 *For any vector* $y = (u, p) \in \mathscr{H}_k$, $[\mathbf{A}y, y]_k \le m[y, y]$.

Proof Since $\gamma_k = \gamma_w + \gamma_s\lambda_k$, we see that $\mathbf{A}y = (-p, \lambda_k u + 2\gamma_k p)$ and

$$\begin{aligned}[\mathbf{A}y, y]_k &= -b_k(p, u) - \gamma_k(p, p) + \gamma_k(\lambda_k u + 2\gamma_k p, u) + (\lambda_k u + 2\gamma_k p, p) = \\ &= \gamma_k\lambda_k\|u\|^2 + (2\gamma_k^2 - b_k + \lambda_k)(u, p) + \gamma_k\|p\|^2.\end{aligned}$$

Then

$$\begin{aligned}[\mathbf{A}y, y]_k - m[y, y]_k = (\gamma_k\lambda_k - mb_k)\|u\|^2 + \\ + (2\gamma_k^2 - b_k + \lambda_k - 2m\gamma_k)(u, p) + (\gamma_k - m)\|p\|^2.\end{aligned}$$

Simple monomorphisms can show that the determinant of the last quadratic form is equal to

$$\begin{aligned}D &= (2\gamma_k^2 - b_k + \lambda_k - 2m\gamma_k)^2 - 4(\gamma_k\lambda_k - mb_k)(\gamma_k - m) = \\ &= (b_k - \lambda_k - 2(m - \gamma_k)^2)^2 - 4(\gamma_k - m)^2(m^2 - 2\gamma_k m + \lambda_k).\end{aligned}$$

The reader will easily prove that

$$b_k - 2m^2 + 2\gamma_k(2m - \gamma_k) - \lambda_k = 2(m - \gamma_k)\sqrt{m^2 - 2\gamma_k m + \lambda_k}.$$

Thus $D = 0$. Moreover, since $\gamma_k - m \le 0$ then the quadratic form $[\mathbf{A}y, y]_k - m[y, y]_k$ is confluent and nonpositive. This completes the proof of the lemma.

Let us show that $\min_{k_1+1\le k\le k_2}\{s_k\} = \min\{s_{k_1+1}, s_{k_2}\}$.

Lemma 14.4 *The minimum of the function* $s(\gamma)$ *on the closed interval* $I = [\gamma_{k_1+1}, \gamma_{k_2}]$ *is attained at the ends of the closed interval.*

Proof The derivative of $s(\gamma)$ is given by

$$s'_\gamma = \frac{-2\gamma_s m + 1}{2\gamma_s\sqrt{m^2 - 2\gamma m + \lambda(\gamma)}} - 1.$$

Since $2\gamma_s m < 1$ then s'_γ has the same sign as the following expression

$$\begin{aligned}(1 - 2\gamma_s m)^2 - 4\gamma_s^2(m^2 - 2\gamma m + \lambda(\gamma)) = 1 - 4\gamma_s m + 4\gamma_s^2 m^2 - \\ -4\gamma_s^2(m^2 - 2\gamma m) - 4\gamma_s(\gamma - \gamma_w) = 1 - 4\gamma_s m + 4\gamma_s\gamma_w + 4\gamma_s(2\gamma_s m - 1)\gamma.\end{aligned}$$

The last expression is linear with respect to γ and the leading coefficient is negative. Hence, s'_γ may have only one root on the interval I and this root corresponds to the maximum of $s(\gamma)$. We get that the minimum of s is attained at the ends of the closed interval.

By Lemma 14.4 the minimum of s_k for $k_1 + 1 \le k \le k_2$ is achieved either at $k = k_1 + 1$ or at $k = k_2$. This implies the following estimate of the norm of vector $y = y_{k_1+1} + \ldots + y_{k_2}$, $y_k \in \mathscr{H}_k$,

$$|||y|||^2 = \sum_{k=k_1+1}^{k_2} |||y_k|||_k^2 \ge \sum_{k=k_1+1}^{k_2} s_k^2 \|u_k\| \ge \min_{k_1+1\le k\le k_2} \left\{s_k^2\right\} \sum_{k=k_1+1}^{k_2} \|u_k\|^2 \ge$$
$$\ge \min\{s_{k_1+1}^2, s_{k_2}^2\}\|u\|^2 = \min\{\varkappa_{II}^2, \varkappa_{III}^2\}\|u\|^2. \tag{14.13}$$

14.4.3 New Norm in the Space $\mathscr{H}_\infty$

The space $\mathscr{H}_\infty$ is infinitely-dimensional. We introduce the new inner product $[\cdot, \cdot]_\infty$, which is equivalent to the standard one, in such a way that for any vector $y \in \mathscr{H}_\infty$, $[\mathbf{A}y, y]_\infty \ge M[y, y]_\infty$.

Define the inner product of vectors $y = (u, p) \in \mathscr{H}_\infty$, $\tilde{y} = (\tilde{u}, \tilde{p}) \in \mathscr{H}_\infty$, by the rule

$$[y, \tilde{y}]_\infty = (1 - 2M\gamma_s)(\nabla u, \nabla\tilde{u}) + 2M\gamma_s\lambda_M(u, \tilde{u}) + M(u, \tilde{p}) + M(p, \tilde{u}) + (p, \tilde{p}),$$

where $\lambda_M = \frac{M-\gamma_w}{\gamma_s}$. By (14.9) we have $\lambda_M > M^2$. Moreover, for any vector $y = (u, p) \in \mathscr{H}_\infty$,

$$\|\nabla u\|^2 \ge \lambda_{k_2+1}\|u\|^2 = \frac{\gamma_{k_2+1} - \gamma_w}{\gamma_s}\|u\|^2 \ge \lambda_M\|u\|^2. \tag{14.14}$$

Corresponding norm is defined by the formula

$$|||y|||_\infty^2 = (1 - 2M\gamma_s)\|\nabla u\|^2 + M(2\gamma_s\lambda_M - M)\|u\|^2 + \|Mu + p\|^2.$$

Lemma 14.5 *The norms $|||y|||_\infty$ and $\|y\|_H$ are equivalent on the space H_∞.*

Proof Since $2\gamma_s M < 1$ and

$$|||y|||_\infty^2 \le (1 - 2M\gamma_s)\|\nabla u\|^2 + M(2\gamma_s\lambda_M - M)\|u\|^2 + (M\|u\| + \|p\|)^2,$$

it follows that the quantity $|||y|||_\infty^2$ is bounded above by a quantity depending on $\|\nabla u\|^2$ and $\|p\|^2$.

Let us find a lower bound for $|||y|||_\infty^2$. For some $\varepsilon > 0$, we have $\lambda_M(1 - \varepsilon) > M^2$. With regard to (14.14), we have

$$\begin{aligned}\|y\|_\infty^2 &= (1-2M\gamma_s)\|\nabla u\|^2 + (2M\gamma_s\lambda_M - M^2)\|u\|^2 + \|Mu+p\|^2 \geq \\ &\geq \varepsilon\|\nabla u\|^2 + (1-2M\gamma_s-\varepsilon)\|\nabla u\|^2 + \lambda_M(2M\gamma_s - 1 + \varepsilon)\|u\|^2 + \\ &+ \|Mu+p\|^2 \geq \varepsilon\|\nabla u\|^2 + \|Mu+p\|^2 \geq \frac{\varepsilon}{2}\|\nabla + u\|^2 + \frac{\varepsilon\lambda_M}{2}\|u\|^2 + \\ &+ (M\|u\| - \|p\|)^2.\end{aligned}$$

The expression on the right-hand side is a positive-defined quadratic form in $\|\nabla u\|$, $\|u\|$ and $\|p\|$, which can be estimated below by multiple of $\|\nabla u\|^2 + \|p\|^2$.

Lemma 14.6 *For any vector* $y = (u, p) \in \mathscr{H}_\infty$,

$$\|y\|_\infty \geq \sqrt{\lambda_M - M^2}\|u\| = \varkappa_{IV}\|u\|. \tag{14.15}$$

Proof By (14.14) we have

$$\begin{aligned}\|y\|_\infty^2 &\geq ((1-2M\gamma_s)\lambda_M + 2M\gamma_s\lambda_M)\|u\|^2 - 2M\|u\|\|p\| + \|p\|^2 = \\ &= (\lambda_M - M^2)\|u\|^2 + (M\|u\| - \|p\|)^2 \geq (\lambda_M - M^2)\|u\|^2.\end{aligned}$$

Lemma 14.7 *For any vector* $y = (u, p) \in \mathscr{H}_\infty \cap \mathscr{D}(\mathbf{A})$, $[\mathbf{A}y, y]_\infty \geq M[y, y]_\infty$.

Proof With regard to $M = \gamma_w + \gamma_s\lambda_M$, we have

$$\begin{aligned}[y, y]_\infty &= (1-2M\gamma_s)\|\nabla u\|^2 + 2M(M-\gamma_w)\|u\|^2 + 2M(u, p) + \|p\|^2; \\ \mathbf{A}y &= (-p, -\Delta u + 2\gamma_w p - 2\gamma_s\Delta p);\end{aligned}$$

$$\begin{aligned}[\mathbf{A}y, y]_\infty = &- (1-2M\gamma_s)(\nabla p, \nabla u) + 2M(M-\gamma_w)(-p, u) + M(-p, p) + \\ &+ (-\Delta u + 2\gamma_w p - 2\gamma_s\Delta p, Mu + p) = M\|\nabla u\|^2 + 4M\gamma_s(\nabla p, \nabla u) + \\ &+ 2\gamma_s\|\nabla p\|^2 + 2M(2\gamma_w - M)(u, p) + (2\gamma_w - M)\|p\|^2.\end{aligned}$$

It follows that

$$\begin{aligned}[\mathbf{A}y, y]_\infty - M[y, y]_\infty &= 2M^2\gamma_s\|\nabla u\|^2 + 4M\gamma_s(\nabla p, \nabla u) + 2\gamma_s\|\nabla p\|^2 - \\ &- 2M^2(M-\gamma_w)\|u\|^2 + 2M(2\gamma_w - 2M)(u, p) + \\ &+ (2\gamma_w - 2M)\|p\|^2 = \\ &= 2\gamma_s\|M\nabla u + \nabla p\|^2 - 2\gamma_s\lambda_M\|Mu + p\|^2.\end{aligned}$$

The last expression is nonnegative by (14.14).

14.4.4 End of the Proof of Theorem 14.3

Denote $\mathscr{H}^{\eta} = \langle \eta_1 e_1, \ldots, \eta_{k_1} e_{k_1} \rangle$, $\mathscr{H}^{\xi} = \langle \xi_1 e_1, \ldots, \xi_{k_1} e_{k_1} \rangle$, $\mathscr{H}^{I} = \mathscr{H}^{\eta} \oplus \mathscr{H}_{k_1+1} \oplus \ldots \mathscr{H}_{k_2}$, $\mathscr{H}^{II} = \mathscr{H}^{\xi} \oplus \mathscr{H}_{\infty}$. The spaces $\mathscr{H}^{I}$ and $\mathscr{H}^{II}$ are orthogonal to each other with respect to the new inner product.

Since $\mathbf{A}(\xi_k e_k) = \mu_k(\xi_k e_k), \mathbf{A}(\eta_k e_k) = \nu_k(\eta_k e_k)$ for $k = 1, \ldots, k_1$, it follows that

$$[\mathbf{A}y, y] \le \max_{1 \le k \le k_1} \mu_k \cdot [y, y] = \mu_{k_1}[y, y] \qquad \forall y \in \mathscr{H}^{\xi}, \tag{14.16}$$

$$[\mathbf{A}y, y] \ge \min_{1 \le k \le k_1} \nu_k \cdot [y, y] = \nu_{k_1}[y, y] \qquad \forall y \in \mathscr{H}^{\eta}. \tag{14.17}$$

It follows from condition (14.16), Lemma 14.3, and the inequality $m > \mu_{k_1}$ that

$$[\mathbf{A}y, y] \le m[y, y] \qquad \forall y \in \mathscr{H}^{I}. \tag{14.18}$$

Also, condition (14.17), Lemma 14.7, and the inequality $M < \nu_{k_1}$ imply that

$$[\mathbf{A}y, y] \ge M[y, y] \qquad \forall y \in \mathscr{H}^{II} \cap \mathscr{D}(\mathbf{A}). \tag{14.19}$$

Since the vector $F(y)$ has zero u-component, it follows that

$$|||F(y_1) - F(y_2)||| = \|F(y_1) - F(y_2)\|_{\mathscr{H}} = \|f(u_1) - f(u_2)\| \le l\|u_1 - u_2\|. \tag{14.20}$$

By estimates (14.12), (14.13), (14.15) of the vector $y = y_1 - y_2 = y_1 + \ldots + y_{k_2} + y_{\infty}$, $y_k \in \mathscr{H}_k$, $y_{\infty} \in \mathscr{H}_{\infty}$, we obtain

$$|||y|||^2 = \sum_{k=1}^{k_1} |||y_k|||_k^2 + \sum_{k=k_1+1}^{k_2} |||y_k|||_k^2 + |||y_{\infty}|||_{\infty}^2 \ge \min\{\varkappa_I^2, \varkappa_{II}^2, \varkappa_{III}^2, \varkappa_{IV}^2\}\|u\|^2. \tag{14.21}$$

It follows from inequalities (14.20) and (14.21) that

$$|||F(y_1) - F(y_2)||| \le l\|u_1 - u_2\| \le \frac{l\,|||y_1 - y_2|||}{\min\{\varkappa_I, \varkappa_{II}, \varkappa_{III}, \varkappa_{IV}\}}.$$

Thus the global Lipschitz constant L for the function $F(y)$ is equal to

$$L = \frac{l}{\min\{\varkappa_I, \varkappa_{II}, \varkappa_{III}, \varkappa_{IV}\}}.$$

Let us define the orthogonal projection to the $(2k_2 - k_1)$-dimensional space $\mathscr{H}^{I} = P(\mathscr{H})$ and denote it by P and define the orthogonal projection $Q = \mathrm{Id} - P$ to $\mathscr{H}^{II} \oplus \mathscr{H}_{\infty} = Q(\mathscr{H})$. Then the inequalities (14.18) and (14.19) acquire the form (14.3), and the spectral gap condition (14.4) is equivalent to condition (14.11).

Thus, all conditions of Theorem 14.1 are satisfied, and thus the space $\mathscr{H}$ contains an integral manifold which dimension is equal to that of the subspace $\mathscr{H}^1$, i.e., to $2k_2 - k_1$. This completes the proof of the theorem.

Acknowledgments The author express her gratitude to A. Yu. Goritsky and V. V. Chepyzhov for setting the problem and permanent attention to the research.

References

1. Chalkina, N.A.: Sufficient condition for the existence of an inertial manifold for a hyperbolic equation with weak and strong dissipation. Russ. J. Math. Phys. (2012). doi:10.1134/S1061920812010025
2. Chalkina, N.A., Goritsky, A.Yu.: Inertial manifolds for Weakly and Strongly Dissipative hyperbolic equations [in Russia]. Tr. Semin. Im. I. G. Petrovskogo. **29** (2012)
3. Chepyzhov, V.V., Goritsky, AYu.: Global integral manifolds with exponential tracking for nonautonomous equations. Russ. J. Math. Phys. **5**(1), 9–28 (1997)
4. Chepyzhov, V.V., Goritsky, AYu.: The dichotomy property of solutions of semilinear equations in problems on inertial manifolds. Mat. Sb. **196**(4), 23–50 (2005)
5. Chueshov, I.D.: Introduction to the Theory of Infinite-Dimensional Dissipative Systems [in Russia]. Acta, Kharkiv (2002)
6. Constantine, P., Foias, C., Nicolaenko, B., Temam, R.: Integral Manifolds and Inertial Manifolds for Dissipative Partial Differential Equations, Applied Mathematics Sciences, vol. 70. Springer, New York (1989)
7. Dell'Oro, F., Pata, V.: Long-term analysis of strongly damped nonlinear wave equations. Nonlinearity **24**, 3413–3435 (2011)
8. Kalantarov, V.K.: Global Behavior of Solutions to Nonlinear Problems of Mathematical Physics of Classical and Nonclassical Type. Postdoc Thesis, Leningrad (1988)
9. Mora, X.: Finite-dimensional attracting invariant manifolds for damped semilinear wave equations. Res. Notes in Math. **155**, 172–183 (1987)
10. Pata, V., Zelik, S.: Smooth attractors for strongly damped wave equations. Nonlinearity **19**, 1495–1506 (2006)

Chapter 15
On Regularity of All Weak Solutions and Their Attractors for Reaction-Diffusion Inclusion in Unbounded Domain

Nataliia V. Gorban and Pavlo O. Kasyanov

Abstract We consider the reaction-diffusion equation with multivalued function of interaction in an unbounded domain. Conditions on the parameters of the problem can not guarantee the uniqueness of the solution of the Cauchy problem. In this work we focus on the study of long-term forecasts of the state functions of reaction-diffusion equation with use of the theory of global attractors for multivalued semiflows. It is obtained the results of the existence and properties of all weak solutions. We obtain the standard a priori estimates for weak solutions of the investigated problem, prove the existence of weak solutions, the existence of global and trajectory attractors for the problem in phase and extended phase spaces respectively. We provide the regularity properties for all globally defined weak solutions and their global and trajectory attractors. The results can be used for the investigation of specific physical models including combustion models in porous media, conduction models of electrical impulses into the nerve endings, climate models.

15.1 Introduction

Let $N \geq 1$, $\underline{f}, \overline{f} : \mathbb{R}^{N+1} \to \mathbb{R}$ are some real functions. We consider the semilinear reaction-diffusion inclusion

$$u_t - \triangle u + [\underline{f}(x,u), \overline{f}(x,u)] \ni 0 \text{ in } \mathbb{R}^N \times (\tau, T), \quad (-\infty < \tau < T < +\infty), \tag{15.1}$$

with initial conditions

N. V. Gorban (✉) · P. O. Kasyanov
Institute for Applied System Analysis, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kyiv 03056, Ukraine
e-mail: nata_gorban@i.ua

P. O. Kasyanov
e-mail: kasyanov@i.ua

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_15,

$$u(\tau) = u_\tau \in L^2(\mathbb{R}^N), \tag{15.2}$$

where u is unknown function, $u_t = \partial u/\partial t$,

$$[a, b] = \{\alpha a + (1-\alpha)b \,|\, \alpha \in [0,1]\}, \quad a, b \in \mathbb{R}.$$

Note that $f = [\underline{f}, \overline{f}] : \mathbb{R}^{N+1} \to 2^{\mathbb{R}} \setminus \{\emptyset\}$. Let us specify the conditions for parameters of the problem:

(α_1) $\underline{f}, \overline{f} : \mathbb{R}^{N+1} \to \mathbb{R}$ are measurable functions such that for a.e. $x \in \mathbb{R}^N$ $\underline{f}(x, \cdot)$ is lower semi-continuous (l.s.c.), and $\overline{f}(x, \cdot)$ is upper semi-continuous (u.s.c.);

(α_2) there exist $C_1 \in L^1(\mathbb{R}^N)$ and $\alpha > 0$ such that for a.e. $x \in \mathbb{R}^N$, $\forall u \in \mathbb{R}$

$$\begin{aligned} \overline{f}(x,u)u &\geq \alpha|u|^2 - C_1(x), \ u \leq 0; \\ \underline{f}(x,u)u &\geq \alpha|u|^2 - C_1(x), \ u \geq 0; \end{aligned} \tag{15.3}$$

(α_3) there exist $C_2 \in L^1(\mathbb{R}^N)$, $C_2 \geq 0$, and $\beta > 0$ such that for a.e. $x \in \mathbb{R}^N$, $\forall u \in \mathbb{R}$

$$\begin{aligned} |\overline{f}(x,u)|^2 &\leq C_2(x) + \beta|u|^2, \\ |\underline{f}(x,u)|^2 &\leq C_2(x) + \beta|u|^2, \\ \underline{f}(x,u) &\leq \overline{f}(x,u). \end{aligned} \tag{15.4}$$

Further we use the following standard notations: $H = L^2(\mathbb{R}^N)$, $V = H_0^1(\mathbb{R}^N)$, V' is the dual space of V. Let us consider real spaces H, V and V' with corresponding norms $\|\cdot\|$, $\|\cdot\|_V$ and $\|\cdot\|_{V'}$. The norm in $\mathbb{R}^N$, inner product in H and in $\mathbb{R}^N$ we will denote by $|\cdot|$, $(\cdot,\cdot)_H$, $(\cdot,\cdot)$ respectively. The function $u(\cdot) \in L^2(\tau, T; V)$ is a *weak solution* of Problem (15.1) on $[\tau, T]$, if there exists a measurable function $d : \mathbb{R}^N \times (\tau, T) \to \mathbb{R}$ such that

$$d(x,t) \in [\underline{f}(u(x,t)), \overline{f}(u(x,t))] \quad \text{for a.e. } (x,t) \in \mathbb{R}^N \times (\tau, T); \tag{15.5}$$

$$-\int_\tau^T \left\langle u, \frac{d\xi}{dt} \right\rangle dt + \int_\tau^T \int_{\mathbb{R}^N} (\nabla u, \nabla \xi)\, dx dt + \int_\tau^T \int_{\mathbb{R}^N} (d, \xi)\, dx dt = 0 \tag{15.6}$$

for all $\xi \in C_0^\infty(\mathbb{R}^N \times (\tau, T))$, where $\langle \cdot, \cdot \rangle$ denotes the pairing in the space V.

We note that Problem (15.1) arises in many important models for distributed parameter control problems and the large class of identification problems enter this formulation. Let us indicate a problem which is one of motivations for the study of the autonomous evolution inclusion (15.1) (cf. [19, 31] and references therein). We consider the nonstationary heat conduction equation

$$\frac{\partial y}{\partial t} - \triangle y = f \text{ in } \mathbb{R}^3 \times (0, +\infty)$$

with initial conditions and suitable boundary ones. Here $y = y(x,t)$ represents the temperature at the point $x \in \mathbb{R}^3$ and time $t > 0$. It is supposed that $f = \bar{f} + \bar{\bar{f}}$, where $\bar{\bar{f}}$ is given and $\bar{f}$ is a known function of the temperature of the form

$$-\bar{f}(x,t) \in \partial j(x, y(x,t)) \text{ a.e. } (x,t) \in \mathbb{R}^3 \times (0, +\infty).$$

Here $\partial j(x,\xi)$ denotes generalized gradient of Clarke (cf. [7]) with respect to the last variable of a function $j : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ which is assumed to be locally Lipschitz in ξ (cf. [19] and references therein). The multivalued function $\partial j(x,\cdot) : \mathbb{R} \to 2^{\mathbb{R}}$ is generally nonmonotone and it includes the vertical jumps. In a physicist's language it means that the law is characterized by the generalized gradient of a nonsmooth potential j (cf. [17, 22, 29, 30]).

Other motivations connected with parabolic equations with a discontinuous nonlinearity. In [25] it is considered the case, when f is the difference of maximal monotone maps. Global attractor in phase space H for such type equations is considered there. Obtained inclusion is a particular case of an abstract differential inclusion generated by a difference of subdifferential maps of proper convex lower semicontinuous functionals [21]. Models of physical interest includes also the next (cf. [1] and references therein):

- a model of combustion in porous media;
- a model of conduction of electrical impulses in nerve axons;
- a climate energy balance model;

etc. The dynamics in H and topological properties (but not regularity) of attractors for all weak solutions of such type differential equations and inclusions were examined (cf. [1, 25] and references therein). We note that for any $u_\tau \in H$ there exists at least one weak solution of Problem (15.1) on $[\tau, T]$ with initial condition $u(x,\tau) = u_\tau(x)$ in $\mathbb{R}^N$. Moreover, each weak solution $u(\cdot)$ of Problem (15.1) on $[\tau, T]$ belongs to $C([\tau, T]; H) \cap L^2(\tau, T; V)$ and $u_t(\cdot) \in L^2(\tau, T; V')$ (cf. [10, 28], chap. 2 [31] and references therein).

In general case Problem (15.1) on $[\tau, T]$ with initial condition $u(x,\tau) = u_\tau(x)$ in $\mathbb{R}^N$ does not have a unique weak solution with $u_\tau \in H$ (cf. [1, p. 2600] and references therein). Thus, for investigation of the long-time behavior as $t \to +\infty$ of all weak solutions of Problem (15.1) with initial data from H, the results for global and trajectory attractors of multivalued semiflows in infinite-dimensional spaces were applied (cf. [1–12, 18–27] and references therein). 2[31] implies the existence of compact in the phase space H invariant global attractor $\mathscr{A}$ for multivalued (in the general case) semiflow G, constructed on all weak solutions of reaction-diffusion system in a bounded domain with continuous interaction function both in autonomous and nonautonomous case. In [27, Theorem 2.3, 2] it was specified the trajectory attractor for translation semigroup acting on the trajectory space of main problem in a bounded domain with the topology of strong local convergence of the sequences $\{u_m(\cdot)\}_{m\geq 1}$ as $m \to +\infty$ in the norm on the Banach spaces $L^\infty(0, M; H) \cap C([0, M]; H)$ for each $M > 0$. The constructions on the-

ory of trajectory and global attractors presented in [1–6, 11–16, 18, 23–27, 31] are sufficiently used there. In paper [11] the equality of global attractors in the sense of [18, Definition 6] as well as [6, Definition 2.2] is proved. The pointwise behavior of complete trajectories studied in [18, Definition 6] under the existence of Lyapunov function on a phase space. The regularity properties of trajectory attractors for systems of reaction-diffusion equations with continuous nonlinearity of an arbitrary polynomial grows considered in [27]; for reaction-diffusion inclusion in a bounded domain in [13].

The main purpose of this paper is to investigate regularity properties of all globally defined weak solutions and their attractors for Problem (15.1) with initial data $u_\tau \in H$ under listed above assumptions.

15.2 On Compact Global Attractor for Reaction-Diffusion Inclusion in Unbounded Domain

The conditions α_1–α_3 do not provide the uniqueness of the solution of the Problems (15.1–15.2), so let us introduce the definition of multivalued, in the general case, semiflow and its global attractor (see for example [31]), that describe the dynamics of the solutions of initial problem as $t \to +\infty$. We set $P(H) = 2^H \setminus \{\emptyset\}$.

Definition 15.1 A map $G : \mathbb{R}_+ \times H \to P(H)$ is called the multivalued semiflow (m-semiflow) on H, if

(1) $G(0, \cdot) = I_H$ is identical motion H;
(2) $G(t + s, x) \subset G(t, G(s, x)) \ \forall\, t, s \in \mathbb{R}_+, \forall\, x \in H$.

M-semiflow is called the strict, if $G(t+s, x) = G(t, G(s, x)) \ \forall\, t, s \in \mathbb{R}_+, \forall\, x \in H$.

Definition 15.2 M-semiflow G is asymptotically compact, if for any nonempty bounded set $B \in P(H)$ such, that

$$\gamma_T^+(B) = \bigcup_{t \geq T} G(t, B)$$

is bounded for some $T = T(B) \geq 0$, an arbitrary sequence $\{\xi_n\}_{n \geq 1}$, $\xi_n \in G(t_n, B)$, $t_n \to +\infty$, is precompact in H.

Definition 15.3 A set $\mathscr{A} \subset H$ that satisfies the next properties:

(1) $\mathscr{A}$ is absorbing set, i.e.,

$$\operatorname{dist}(G(t, B), \mathscr{A}) \to 0, \quad \text{as} \quad t \to +\infty,$$

for any bounded set B, where $\operatorname{dist}(C, A) = \sup\limits_{c \in C} \inf\limits_{a \in A} \|c - a\|$;

(2) $\mathscr{A}$ is semi-invariant, i.e.,

$$A \subset G(t, A), \quad \text{for every} \quad t \geq 0;$$

(3) $\mathscr{A}$ is minimal closed absorbing set (i.e. for any closed absorbing set C we have, that $A \subset C$)
is called the global attractor $\mathscr{A}$ for the m-semiflow G.

The global attractor is called invariant, if $A = G(t, A)$, for every $t \geq 0$.

Let now $\Omega \subset \mathbb{R}^N$ is a bounded domain, $T > 0$, $Q = \Omega \times (0, T)$, $\mathscr{Y} = L^2(Q)$. Further by $\|\cdot\|_E$ we denote the norm in a real Banach space E. The next lemma is necessary for the proof of the main theorem.

Lemma 15.1 *Let f satisfies assumption α_1, and $\{u_n, d_n\}_{n\geq 0} \subset \mathscr{Y}$ satisfies such conditions*

(1) for a.e. $(x, t) \in Q$ $u_n(x, t) \to u_0(x, t)$ as $n \to +\infty$,
(2) $d_n \to d_0$ weakly in $\mathscr{Y}$ as $n \to +\infty$,
(3) $\forall n \geq 1$ for a.e. $(x, t) \in Q$ $d_n(x, t) \in f(x, u_n(x, t))$.

Then for a.e. $(x, t) \in Q$ $d_0(x, t) \in f(x, u_0(x, t))$.

Proof Let $\{u_n, d_n\}_{n\geq 1} \subset \mathscr{Y}$ satisfy the lemma conditions. Let us select the complete measure set $Q_1 \subset Q$ such, that

$$\forall (x, t) \in Q_1 \quad u_n(x, t) \to u_0(x, t) \quad \text{as} \quad n \to \infty. \tag{15.7}$$

The space $L_2(Q)$ is a Hilbert space. So, in virtue of [9, Remark I.6.2] $L_2(Q)$ is uniformly convex space. (see for example [9, Definition I.5.9]). From the proof of [8, Theorem 1, p. 64–66] it follows that any weakly convergent to $\bar{0}$ sequence $\{d_n - d_0\}_{n\geq 1}$ in $\mathscr{Y}$ has a subsequence $\{d_{n_k} - d_0\}_{k\geq 1} \subset \{d_n - d_0\}_{n\geq 1}$, which arithmetical means converge strongly to $\bar{0}$ in $L_2(Q)$ (in [8, Theorem 1, p. 64–66] it is proved stronger statement than the Banach-Saks property), i.e.

$$\|\frac{1}{k}\sum_{j=1}^{k}(d_{n_j} - d_0)\|_{\mathscr{Y}} \to 0 \quad \text{as} \quad k \to +\infty.$$

It means that

$$\frac{1}{k}\sum_{j=1}^{k} d_{n_j} \to d_0 \quad \text{strongly in} \quad L_2(Q) \quad \text{as} \quad k \to +\infty. \tag{15.8}$$

Further, $\exists Q_2 \subset Q_1$ such, that Q_2 is measurable, $\text{meas}(Q_1 \setminus Q_2) = 0$ and $\forall (x, t) \in Q_2$ $\forall k \geq 1$

$$\underline{f}(x, u_{n_k}(x, t)) \leq d_{n_k}(x, t) \leq \overline{f}(x, u_{n_k}(x, t)).$$

So, $\forall k \geq 1, \forall (x,t) \in Q_2$

$$\frac{1}{k}\sum_{j=1}^{k} \underline{f}(x, u_{n_j}(x,t)) \leq \frac{1}{k}\sum_{j=1}^{k} d_{n_j}(x,t) \leq \frac{1}{k}\sum_{j=1}^{k} \overline{f}(x, u_{n_j}(x,t)). \tag{15.9}$$

From (15.8) there exists a subsequence $\left\{\frac{1}{k_l}\sum_{j=1}^{k_l} d_{n_j}\right\}_{l\geq 1} \subset \left\{\frac{1}{k}\sum_{j=1}^{k} d_{n_j}\right\}_{k\geq 1}$ and a complete measure set $Q_3 \subset Q_2$:

$$\forall (x,t) \in Q_3 \quad \frac{1}{k_l}\sum_{j=1}^{k_l} d_{n_j}(x,t) \to d_0(x,t) \text{ as } l \to +\infty. \tag{15.10}$$

For a.e. $(x,t) \in Q_3$ let us set $a_k = \overline{f}(x, u_{n_k}(x,t))$, $k \geq 1$, $a_0 = \overline{f}(x, u_0(x,t))$. From α_1 and (15.7) it follows, that $\varlimsup_{k\to\infty} a_k \leq a_0$. Thus,

$$\varlimsup_{k\to+\infty} \frac{1}{k}\sum_{j=1}^{k} \overline{f}(x, u_{n_j}(x,t)) \leq \overline{f}(x, u_0(x,t)).$$

Similarly,

$$\varliminf_{k\to+\infty} \frac{1}{k}\sum_{j=1}^{k} \underline{f}(x, u_{n_j}(x,t)) \geq \underline{f}(x, u_0(x,t)).$$

Taking into account (15.9–15.10), we obtain that $d_0(x,t) \in f(x, u_0(x,t))$ for a.e. $(x,t) \in Q$.

Provide the standard a priori estimates for solutions.

Lemma 15.2 [10] *Let assumptions α_1–α_3 hold. Then, for any weak solution u of Problems (15.1–15.2) on $[\tau, T]$ we have*

$$\|u\|_{X(\tau,T)} \leq K_1(\|u_\tau\|, T-\tau), \tag{15.11}$$

$$\|u_t\|_{U(\tau,T)} \leq K_2(\|u_\tau\|, T-\tau), \tag{15.12}$$

where K_i are nondecreasing by each variable functions, $X(\tau,T) = L^2(\tau,T;V) \cap C([\tau,T],H)$ and $U(\tau,T) = L^2(\tau,T;V')$.

Theorem 15.1 [10] *Let assumptions α_1–α_3 hold. Then for any $\tau < T$, and each $u_\tau \in L^2(\mathbb{R}^N)$, Problems (15.1–15.2) has at least one weak solution on $[\tau, T]$.*

Since in Theorem 15.1 $T > 0$ is an arbitrary and the concatenation of weak solutions is the weak solution, then similarly to [20, p. 119], each weak solution can be continued to the global one, defined on $[0, +\infty)$.

Let us denote the family of all global solutions of Problems (15.1–15.2) by $\mathscr{D}(u_0)$. Note that $\mathscr{D}(u_0) \subset L^2_{loc}(0,+\infty;V) \cap C([0,+\infty),H)$. Let us provide that $\mathscr{D}(u_0) \subset L^\infty(0,+\infty;H)\ \forall u_0 \in L^2(\mathbb{R}^N)$.

Lemma 15.3 [10] *Let assumptions α_1–α_3 hold. If u is a globally defined weak solution of Problems (15.1–15.2), then*

$$\forall t \geq 0 \quad \|u(t)\|^2 + 2\int_0^t e^{-2\alpha(t-s)}\|\nabla y\|^2 ds \leq \|u(0)\|^2 e^{-2\alpha t} + D, \tag{15.13}$$

where $D = \|C_1\|_{L^1(\mathbb{R}^N)}/\alpha$.

Define the m-semiflow map $G : \mathbb{R}_+ \times H \to P(H)$:

$$G(t,u_0) = \{z \in H \mid \exists u \in \mathscr{D}(u_0) :\ u(0) = u_0,\ u(t) = z\}.$$

Theorem 15.2 *Let assumptions α_1–α_3 hold. Then Problems (15.1–15.2) defines the m-semiflow in the phase space H, that possesses the invariant global attractor.*

Proof 1° Prove that G is the strict m-semiflow. The proof of $G(t+s,x) \subset G(t,G(s,x))$ repeats the proof of similar inclusion from [20, Lemma 7]. Let us check that $G(t,G(s,x)) \subset G(t+s,x)$. Let $z \in G(t,G(s,x))$. Then there exist z_1, $u_1(\cdot) \in \mathscr{D}(x)$, $u_2(\cdot) \in \mathscr{D}(z_1)$, d_1, d_2 such that

$$u_1(0) = x, \quad u_1(s) = z_1,$$

$$u_2(0) = z_1, \quad u_2(t) = z,$$

$$d_1 = \triangle u_1 - \frac{\partial u_1}{\partial t}, \quad d_1(\xi,\zeta) \in f(\xi,u_1(\xi,\zeta)) \text{ for a.e. } (\xi,\zeta) \in \mathbb{R}^N \times \mathbb{R}_+,$$

$$d_2 = \triangle u_2 - \frac{\partial u_2}{\partial t}, \quad d_2(\xi,\zeta) \in f(\xi,u_2(\xi,\zeta)) \text{ for a.e. } (\xi,\zeta) \in \mathbb{R}^N \times \mathbb{R}_+.$$

Show that there exists $u(\cdot) \in \mathscr{D}(u_0)$: $u(0) = x$, $u(t+s) = z$. Let us define u by:

$$u(r) = \begin{cases} u_1(r), & 0 \leq r \leq s, \\ u_2(r-s), & s \leq r. \end{cases}$$

For a.e. $(\xi,\zeta) \in \mathbb{R}^N \times \mathbb{R}_+$ let us set

$$d(\xi,\zeta) = \begin{cases} d_1(\xi,\zeta), & 0 \leq \zeta \leq s, \\ d_2(\xi,\zeta-s), & s \leq \zeta. \end{cases}$$

Remark that

$$d(\xi,\zeta)\in f(\xi,u(\xi,\zeta))\text{ for a.e. }(\xi,\zeta)\in\mathbb{R}^N\times\mathbb{R}_+.$$

Fianlly,

$$\begin{aligned}
&\int_0^T\left\langle\frac{\partial u}{\partial r},v\right\rangle_V dr+\int_0^T\left[(\nabla u,\nabla v)dr+\int_{\mathbb{R}^N}d(x,r)v(x,r)dx\right]dr+\\
=&\int_0^s\left\langle\frac{\partial u_1}{\partial r},v\right\rangle_V dr+\int_0^s\left[(\nabla u_1,\nabla v)dr+\int_{\mathbb{R}^N}d_1(x,r)v(x,r)dx\right]dr+\\
&+\int_s^T\left\langle\frac{\partial u_2(r-s)}{\partial r},v\right\rangle_V dr+\int_s^T[(\nabla u_2(r-s),\nabla v)dr+\\
&+\int_{\mathbb{R}^N}d_2(x,r-s)v(x,r-s)dx\Bigg]dr=\\
=&0+\int_0^{T-s}\left\langle\frac{\partial u_2}{\partial r},v\right\rangle_V dr+\int_0^{T-s}\left[(\nabla u_2,\nabla v)dr+\int_{\mathbb{R}^N}d_2(x,r)v(x,r)dx\right]dr=0
\end{aligned}$$

$\forall T>s+t,\forall v\in C_0^\infty([0,T]\times\mathbb{R}^N)$.

2° For any fixed $k>0$ we denote by Ω_k the ball of radius k with the center at 0. Let us prove that for an arbitrary nonempty bounded set $B\subset H$, each $u_0\in B$, any $u\in\mathscr{D}(u_0)$, and all $\varepsilon>0$ there exist $T(\varepsilon,B)$, $K(\varepsilon,B)$ such that

$$\forall t\geq T,\ k\geq K\qquad\int_{|x|\geq\sqrt{2}k}|u(x,t)|^2dx\leq\varepsilon.$$

Indeed, let $s\in\mathbb{R}_+$. Let us consider a smooth function

$$\theta(s)=\begin{cases}0, & 0\leq s\leq 1,\\ 0\leq\theta(s)\leq 1, & 1\leq s\leq 2,\\ 1, & s\geq 2\end{cases}$$

such that $|\theta'(s)|\leq C\ \forall s\in\mathbb{R}_+$. Moreover, suppose that $\sqrt{\theta}$ is smooth too.

Let us apply [20, Lemma 3] to $\rho(x)=\sqrt{\theta(\frac{|x|^2}{k^2})}$. From the definition of the weak solution of Eq. (15.1) it follows that

$$\text{for a.e. } t \geq 0 \qquad \frac{1}{2}\frac{d}{dt}\int\limits_{\mathbb{R}^N} \theta\left(\frac{|x|^2}{k^2}\right)|u|^2 dx = \langle u_t, \rho^2 u\rangle_V =$$

$$= \langle \triangle u, \rho^2 u\rangle_V - \int\limits_{\mathbb{R}^N} \theta\left(\frac{|x|^2}{k^2}\right) d(x,t)u(x,t)dx, \tag{15.14}$$

where

$$d = \triangle u - \frac{\partial u}{\partial t}, \quad d(\xi,\zeta) \in f(\xi, u(\xi,\zeta)) \text{ for a.e. } (\xi,\zeta) \in \mathbb{R}^N \times \mathbb{R}_+,$$

Similarly to [20, p. 122–123], the first term in the right part of the last relation is estimated by the next way:

$$\langle \triangle u, \rho^2 u\rangle_V \leq \varepsilon'(1 + \|\nabla u\|^2) \tag{15.15}$$

for an arbitrary $k \geq K_1(\varepsilon')$, where $\varepsilon' > 0$ is an arbitrary and rather small.

For the second term from (15.14), due to assumptions α_2 and α_3, we obtain the estimates

$$-\int\limits_{\mathbb{R}^N} \theta\left(\frac{|x|^2}{k^2}\right) d(x,t)u(x,t)dx \leq -\alpha \int\limits_{\mathbb{R}^N} \theta\left(\frac{|x|^2}{k^2}\right)|u(x,t)|^2 dx +$$

$$+\int\limits_{\mathbb{R}^N} \theta\left(\frac{|x|^2}{k^2}\right) C_1(x)dx \leq -\alpha \int\limits_{\mathbb{R}^N} \theta\left(\frac{|x|^2}{k^2}\right)|u(x,t)|^2 dx + 2\varepsilon', \tag{15.16}$$

as soon as $k \geq K_2(\varepsilon')$. Let us set

$$Y(t) = \int\limits_{\mathbb{R}^N} \theta\left(\frac{|x|^2}{k^2}\right)|u(x,t)|^2 dx.$$

Then from (15.14–15.16) it follows that

$$\frac{1}{2}\frac{d}{dt}Y(t) + \alpha Y(t) \leq 3\varepsilon' + \varepsilon\|\nabla u\|^2,$$

as soon as $k \geq \max\{K_1, K_2\}$. By using the Gronwall-Bellman inequality and Lemma 15.3, we obtain

$$Y(t) \leq Y(0)e^{-2\alpha t} + \frac{3}{\alpha}\varepsilon' + \frac{\varepsilon'}{2}(\|u_0\|^2 + D).$$

Choosing ε', $T(\varepsilon, B)$ such that

$$\frac{3}{\alpha}\varepsilon' + \frac{\varepsilon'}{2}(\|u_0\|^2 + D) \le \frac{\varepsilon}{2}, \quad Y(0)e^{-2\alpha t} \le \frac{\varepsilon}{2}, \quad \forall u_0 \in B,\ t \ge T,$$

we obtain $Y(t) \le \varepsilon$ and

$$\int\limits_{|x|\ge\sqrt{2}k} |u(x,t)|^2 dx \le \int\limits_{\mathbb{R}^N} \theta\left(\frac{|x|^2}{k^2}\right)|u(x,t)|^2 dx \le \varepsilon.$$

3° For a bounded set $B \subset H$ and $T \in \mathbb{R}_+$ let us consider

$$\gamma_T^+(B) = \bigcup_{t\ge T} G(t, B).$$

Following by the proof of [20, Lemma 8] and the proof of Theorem 15.1 we obtain the next result, that is necessity for the proof of asymptotic compactness of the m-semiflow G. Namely, the graph of $G(t,\cdot)$ is weakly closed. This means that if $\xi_n \to \xi_\infty$, $\beta_n \to \beta_\infty$ weakly in H as $n \to \infty$, where $\xi_n \in G(t, \beta_n)$ $\forall n \ge 1$, then $\xi_\infty \in G(t, \beta_\infty)$.

4° Show that m-semiflow G is asymptotically compact. Let $\xi_n \in G(t_n, v_n)$ $v_n \in B$, $n \ge 1$, B be a bounded set in H. Since $\gamma^+_{T(B)}(B)$ is bounded and $\xi_n \in G(t_n, v_n) \subset \gamma^+_{T(B)}(B)$ for $n \ge n_0$, then there exists the weakly convergent in H subsequence (let us denote it by $\{\xi_n\}_{n\ge1}$ again) to some ξ as $n \to \infty$. Let $T_0 > 0$ be an arbitrary number. Using 1° we get that $\xi_n \in G(t_n, v_n) = G(T_0, G(t_n - T_0, v_n))$ for every $n \ge 1$. Then for every $n \ge 1$ there exists $\beta_n \in G(t_n - T_0, v_n)$ such that $\xi_n \in G(T_0, \beta_n)$. Let us choose $N(B, T_0)$ such that $\forall n \ge N(B, T_0)$ $t_n - T_0 \ge T(B)$ and $G(t_n - T_0, v_n) \subset \gamma^+_{T(B)}(B)$ is bounded. $\beta_n \to \xi_{T_0}$ weakly in H as $n \to \infty$. From 3° it follows, that the graph $G(T_0, \cdot)$ is weakly closed. So, $\xi \in G(T_0, \xi_{T_0})$ and $\varliminf\limits_{n\to\infty} \|\xi_n\| \ge \|\xi\|$. Show that up to subsequence, $\varlimsup\limits_{n\to\infty} \|\xi_n\| \le \|\xi\|$ as $n \to \infty$.

Any weak solution u satisfies

$$\frac{1}{2}\frac{d}{dt}\|u\|^2 + \frac{1}{2}\|u\|^2 + \|\nabla u\|^2 = -\int\limits_{\mathbb{R}^N} d \cdot u\, dx + \frac{1}{2}\|u\|^2, \text{ a.e. on } [0, T],$$

where $d \in L^2(0, T; H)$: $d(x,t) \in f(x, u(x,t))$ for a.e. $(x,t) \in \mathbb{R}^N \times (0,T)$. Let $\{u_n(\cdot)\}_{n\ge1}$ is the sequence of weak solutions such that for any $n \ge 1$ $u_n(T_0) = \xi_n$ and $u_n(0) = \beta_n$. In view of the Gronwall-Bellman lemma, $\forall n \ge 1$

$$\|\xi_n\|^2 = e^{-T_0}\|\beta_n\|^2 - 2\int_0^{T_0} e^{-(T_0-s)}\|\nabla u_n\|^2 ds$$
$$- 2\int_0^{T_0}\int_{\mathbb{R}^N} e^{-(T_0-s)} d_n \cdot u_n \, dx ds + \int_0^{T_0} e^{-(T_0-s)}\|u_n\|^2 ds, \tag{15.17}$$

where $d_n \in L^2(0,T;H)$: $d_n(x,t) \in f(x,u_n(x,t))$ for a.e. $(x,t) \in \mathbb{R}^N \times (0,T)$. From Lemma 15.2 and the Banach-Alaoglu theorem, up to subsequence (we denote it again by $\{u_n, d_n\}_{n\geq 1}$), $\{u_n\}_{n\geq 1}$ converges to some weak solution u in the following sense

$$\begin{array}{ll} u_n \to u & \text{weakly in } L^2(0,T;V), \ \ n\to\infty \\ u_n \to u & \text{weakly star in } L^\infty(0,T;H), \ \ n\to\infty \\ d_n \to d & \text{weakly in } L^2(0,T;V'), \ \ n\to\infty \\ \frac{\partial u_n}{\partial t} \to \frac{\partial u}{\partial t} & \text{weakly in } L^2(0,T;V'), \ \ n\to\infty. \end{array} \tag{15.18}$$

From 3°, $u(0) = \xi_{T_0}$, $u(T_0) = \xi$.

Since the sequence $\{\beta_n\}_{n\geq 1}$ is bounded in H, then

$$\forall n \quad e^{-T_0}\|\beta_n\|^2 \leq e^{-T_0} M. \tag{15.19}$$

Further,

$$\overline{\lim_{n\to\infty}}\left(-2\int_0^{T_0} e^{-(T_0-s)}\|\nabla u_n\|^2 ds\right) \leq -2\int_0^{T_0} e^{-(T_0-s)}\|\nabla u\|^2 ds. \tag{15.20}$$

On the other hand,

$$\int_0^{T_0} e^{-(T_0-s)}\|u_n\|^2 ds = \int_0^{T_0}\int_{\Omega_k} e^{-(T_0-s)}|u_n|^2 dx ds + \int_0^{T_0} e^{-(T_0-s)}\int_{|x|\geq k}|u_n|^2 dx ds.$$

From 1° it follows, that $u_n(s) \in G(s, G(t_n - T_0, v_n)) = G(s + t_n - T_0, v_n)$. From 2°, for any $\varepsilon > 0$ there exist such $T(\varepsilon, B)$, $K_1(\varepsilon, B) > 0$, that

$$\int_{|x|\geq k} |u_n(s)|^2 dx \leq \varepsilon,$$

as soon as $k \geq K_1$, $t_n - T_0 \geq T$. Repeating the respective steps from the proof of Theorem 15.1, we obtain that (up to subsequence) $L_k u_n \to L_k u$ strongly in $L^2(0,T;H_k)$ as $n\to\infty$. So,

$$\overline{\lim_{n\to\infty}} \int_0^{T_0} e^{-(T_0-s)}\|u_n\|^2 ds \le \int_0^{T_0} e^{-(T_0-s)}\|u\|^2 dxds + \varepsilon. \tag{15.21}$$

Let us consider the "nonlinear term". Note that assumption α_3 provides

$$-2\int_0^{T_0}\int_{|x|\ge k} e^{-(T_0-s)} d_n \cdot u_n \, dxds \le 4\varepsilon \int_0^{T_0} e^{-(T_0-s)} ds \le 4\varepsilon,$$

as soon as $k \ge K_2(\varepsilon)$. Since $u_n \to u$ strongly in $L^2(0,T;H_k)$ as $n \to \infty$, then up to a subsequence, $u_n(t,x) \to u(t,x)$, $n \to \infty$ for a.e. $(t,x) \in (0,T_0)\times\Omega_k$. Lemma 15.1 and (15.18) imply

$$\lim_{n\to\infty}\left(-2\int_0^{T_0}\int_{\Omega_k} e^{-(T_0-s)} d_n \cdot u_n \, dxds\right) = -2\int_0^{T_0}\int_{\Omega_k} e^{-(T_0-s)} d\cdot u \, dxds.$$

Thus,

$$\overline{\lim_{n\to\infty}}\left(-2\int_0^{T_0}\int_{\Omega_k} e^{-(T_0-s)} d_n \cdot u_n \, dxds\right) \le -2\int_0^{T_0}\int_{\Omega_k} e^{-(T_0-s)} d\cdot u \, dxds + 4\varepsilon. \tag{15.22}$$

Passing to the limit as $k \to \infty$ in (15.22) and using (15.17) and (15.19–15.22) we find, that

$$\begin{aligned}\overline{\lim_{n\to\infty}}\|\xi_n\|^2 &\le e^{-T_0}M - 2\int_0^{T_0}\int_{\mathbb{R}^N} e^{-(T_0-s)}|\nabla u|^2 dxds + \\ &+ \int_0^{T_0}\int_{\mathbb{R}^N} e^{-(T_0-s)}|u|^2 dxds - 2\int_0^{T_0}\int_{\mathbb{R}^N} e^{-(T_0-s)} d\cdot u \, dxds + 5\varepsilon = \\ &= \|\xi\|^2 + e^{-T_0}M - e^{-T_0}\|\xi_{T_0}\|^2 + 5\varepsilon. \end{aligned}\tag{15.23}$$

Passing to the limit as $T_0 \to +\infty$, and then, directing $\varepsilon \to 0$, we finally obtain the inequality

$$\overline{\lim_{n\to\infty}}\|\xi_n\|^2 \le \|\xi\|^2.$$

So, up to subsequence, $\xi_n \to \xi$ strongly in H as $n \to \infty$.

5° Let us prove the semi-continuity of the m-semiflow G [20, p. 126]. Namely, let us prove that the map $G(t,\cdot)$ is upper semi-continuous and has compact values

for any $t \geq 0$. Indeed, let $\xi_n \in G(t, x_n)$ for every $n \geq 1$ and $x_n \to x_0$ as $n \to \infty$. Let us prove, that the sequence $\{\xi_n\}_{n\geq 1}$ is pre-compact in H. From Lemma 15.3, the sequence $\{\xi_n\}_{n\geq 1}$ is bounded, so, up to the subsequence, $\{\xi_n\}_{n\geq 1}$ is weakly convergent to some ξ. Supposing analogically to the proof of 4°, there exist weak solutions u_n, $n \geq 1$, u such that $\forall n \geq 1$ $u_n(t) = \xi_n$, $u_n(0) = x_n$, $u(t) = \xi$, $u(0) = x_0$ and u_n converges to u in the sense of (15.18). Repeating the suppositions from 4° we obtain that $\overline{\lim\limits_{n\to\infty}} \|\xi_n\|^2 \leq \|\xi\|^2$. Thus, $\xi_n \to \xi$ strongly in H as $n \to \infty$. So, taking into account 3°, $G(t, x_0)$ is compact.

Suppose that $G(t, \cdot)$ is not upper semi-continuous. Then there exists the point x_0, the neighborhood $\mathscr{O}$ of the set $G(t, x_0)$ and the sequence $\{\xi_n\}_{n\geq 1}$ such that $\xi_n \in G(t, x_n)$ $\forall n \geq 1$, $\|x_n - x_0\| \to 0$ as $n \to +\infty$, $\xi_n \notin \mathscr{O}$ $\forall n \geq 1$. Passing to the subsequences we obtain that $\xi_{n_k} \to \xi$, $x_{n_k} \to x_0$ strongly in H as $k \to \infty$. From 3° it follows, that $\xi \in G(t, x_0)$. We obtain the contradiction.

Properties 1–5° imply the existence of the global compact invariant attractor for G (see [18, Theorem 3, Remark 8]), that is minimal closed absorbing set.

The theorem is proved.

15.3 Regularity of All Weak Solutions and Their Attractors

Further we need to consider the restriction of $v : [\tau, T] \to V'$ on $[s, T]$, $s \in (\tau, T)$, $\tau < T$. To simplify conclusions denote it by the same symbol v.

Theorem 15.3 *Let assumptions α_1–α_3 hold, u be an arbitrary weak solution of Problem (15.1) on $[\tau, T]$. Then for any $\varepsilon \in (0, T - \tau)$ $u \in C([\tau + \varepsilon, T]; V) \cap L^2(\tau + \varepsilon, T; H^2(\mathbb{R}^N) \cap V)$ and $u_t \in L^2(\tau + \varepsilon, T; H)$.*

Proof Let u be an arbitrary weak solution of Problem (15.1) on $[\tau, T]$. Then there exists a measurable function $d : \mathbb{R}^N \times (\tau, T) \to \mathbb{R}$ such that u and d satisfy (15.5–15.6). As $u \in L^2(\mathbb{R}^N \times (\tau, T))$ and the growth condition (15.4) holds, then $d \in L^2(\mathbb{R}^N \times (\tau, T))$. The set

$$\mathscr{D} := \{s \in (\tau, T) \,|\, u(s) \in V\}$$

is dense in $[\tau, T]$. For any arbitrary fixed $s \in \mathscr{D}$ we note that u is a unique weak solution on $[s, T]$ of the problem

$$\begin{cases} z_t - \triangle z = -d(x, t) \text{ in } \mathbb{R}^N \times (s, T), \\ z(x, s) = u(x, s) \text{ in } \mathbb{R}^N. \end{cases} \tag{15.24}$$

Moreover, $u \in L^2(s, T; H^2(\mathbb{R}^N) \cap V) \cap C([s, T]; V)$ and $u_t \in L^2(s, T; H)$, $s \in \mathscr{D}$ (cf. [23, Chapter 4.I], [24, Chapter III] and references therein). Thus for any $\varepsilon \in (0, T - \tau)$ $u \in C([\tau + \varepsilon, T]; V) \cap L^2(\tau + \varepsilon, T; H^2(\mathbb{R}^N) \cap V)$ and $u_t \in L^2(\tau + \varepsilon, T; H)$.

The theorem is proved.

Let us consider the family $\mathscr{K}_+ = \cup_{u_0 \in H} \mathscr{D}(u_0)$ of all weak solutions of Problem (15.1) defined on the semi-infinite interval $[0, +\infty)$. We note that $\mathscr{K}_+$ is *translation invariant* one, i.e. $\forall u \in \mathscr{K}_+$, $\forall h \geq 0$ $u_h \in \mathscr{K}_+$, where $u_h(s) = u(h+s)$, $s \geq 0$. Let us consider Problem (15.1) on the entire time axis. A function $u \in L^\infty(\mathbb{R}; H)$ is called a *complete trajectory* of Problem (15.1) , if $\forall h \in \mathbb{R}$ $\Pi_+ u_h \in \mathscr{K}_+$, where Π_+ is the restriction operator to the interval $[0, +\infty)$. Let $\mathscr{K}$ be a family of all complete trajectories of Problem (15.1) . We note that

$$\forall h \in \mathbb{R},\ \forall u \in \mathscr{K} \quad u_h \in \mathscr{K}. \tag{15.25}$$

Let $\{T(h)\}_{h \geq 0}$ be the translation semigroup acting on $\mathscr{K}_+$, i.e. $T(h)u = u(\cdot + h)$, $h \geq 0$, $u \in \mathscr{K}_+$. On $\mathscr{K}_+$ we consider the topology induced from the Fréchet space $C^{loc}(\mathbb{R}_+; H)$. We note that

$$f_n \to f \text{ in } C^{loc}(\mathbb{R}_+; H) \iff \forall M > 0\ \Pi_M f_n \to \Pi_M f \text{ in } C([0, M]; H),$$

where Π_M is the restriction operator to the interval $[0, M]$ [6, p. 18]. We denote the restriction operator to the semi-infinite interval $[0, +\infty)$ by Π_+.

We recall that a set $\mathscr{P} \subset C^{loc}(\mathbb{R}_+; H) \cap L^\infty(\mathbb{R}_+; H)$ is said to be *the attracting* one for the trajectory space $\mathscr{K}_+$ of Problem (15.1) in the topology of $C^{loc}(\mathbb{R}_+; H)$, if for any bounded in $L^\infty(\mathbb{R}_+; H)$ set $\mathbf{B} \subset \mathscr{K}_+$ and any $M \geq 0$ the following relation holds:

$$\mathrm{dist}_{C([0,M];H)}(\Pi_M T(t)\mathbf{B}, \Pi_M \mathscr{P}) \to 0, \quad t \to +\infty. \tag{15.26}$$

A set $\mathscr{U} \subset \mathscr{K}_+$ is said to be *the trajectory attractor* in the trajectory space $\mathscr{K}_+$ with respect to the topology of $C^{loc}(\mathbb{R}_+; H)$ (cf. [6, Definition 1.2, p. 197]) if

(i) $\mathscr{U}$ is a compact set in $C^{loc}(\mathbb{R}_+; H)$ and bounded in $L^\infty(\mathbb{R}_+; H)$;
(ii) $\mathscr{U}$ is strictly invariant with respect to $\{T(h)\}_{h \geq 0}$, i.e. $T(h)\mathscr{U} = \mathscr{U}$ $\forall h \geq 0$;
(iii) $\mathscr{U}$ is an attracting set in the trajectory space $\mathscr{K}_+$ in the topology of $C^{loc}(\mathbb{R}_+; H)$.

Theorem 15.4 *cf.* [31, Theorem 2.3, p. 65] *Let $\mathscr{A}$ be the global attractor from Theorem 15.2. Then in the space $\mathscr{K}_+$ there exists the trajectory attractor $\mathscr{U} \subset \mathscr{K}_+$. Moreover, the next formula takes place*

$$\mathscr{U} = \Pi_+ \mathscr{K} = \{y \in \mathscr{K}_+ \mid y(t) \in \mathscr{A}\ \forall t \in \mathbb{R}_+\}. \tag{15.27}$$

Theorem 15.5 *Let $\mathscr{A}$ be the global attractor from Theorem 15.2, $\mathscr{U}$ be the trajectory attractor from Theorem 15.4. Then*

- *$\mathscr{A}$ is a bounded subset of V;*
- *$\mathscr{U}$ is a bounded subset of $C^{loc}(\mathbb{R}_+; V)$;*
- *$\mathscr{K}$ is a bounded subset of $C^{loc}(\mathbb{R}; V)$.*

Proof Theorems 15.2, 15.3, and 15.4 imply that $\mathscr{A} \subset V$, $\mathscr{U} \subset C^{loc}(\mathbb{R}_+; V)$, and $\mathscr{K} \subset C^{loc}(\mathbb{R}; V)$. To finish the proof we note that Theorem 15.3 provides $\mathscr{K} \subset C^{loc}(\mathbb{R}; V) \cap L_2^{loc}(\mathbb{R}; H^2(\mathbb{R}^N) \cap V)$. There exists a constant $C > 0$, that does not depend on u and t, such that $\frac{d}{dt} \int\limits_{\mathbb{R}^N} |\nabla u|^2 dx \leq C$ for a.e. $t \in \mathbb{R}$. Thus, $\mathscr{K}$ is a bounded subset of $C^{loc}(\mathbb{R}; V)$. Therefore, $\mathscr{A}$ is a bounded subset of V, and $\mathscr{U}$ is a bounded subset of $C^{loc}(\mathbb{R}_+; V)$.

The theorem is proved.

Acknowledgments The work was partially supported by the Ukrainian State Fund for Fundamental Researches under grant GP/F44/076.

References

1. Balibrea, F., Caraballo, T., Kloeden, P.E., Valero, J.: Recent developments in dynamical systems: three perspectives. Int. J. Bifurcat. Chaos. **20**, 2591 (2010). doi:10.1142/S0218127410027246
2. Ball, J.M.: Continuity properties and global attractors of generalized semiflows and the Navier-Stokes equations. Nonlinear Sci. **7**, 475–502 (1997)
3. Ball, J.M.: Global attractors for damped semilinear wave equations. D.C.D.S. **10**, 31–52 (2004)
4. Babin, A.V., Vishik, M.I.: Attractors of Evolution Equations. Nauka, Moscow (1989)
5. Chepyzhov, V.V., Vishik, M.I.: Evolution equations and their trajectory attractors. J. Math. Pur. Appl. **76**(10), 913–964 (1997). doi:10.1016/S0021-7824(97)89978-3
6. Chepyzhov, V.V., Vishik, M.I.: Trajectory and global attractors of three-dimensional Navier-Stokes systems. Math. Notes **71**, 177–193 (2002). doi:10.1023/A:1014190629738
7. Clarke, F.H.: Optimization and Nonsmooth Analysis. Wiley, New York (1983)
8. Diestel, J.: Geometry of Banach Spaces, Selected Topics. Springer, New York (1980)
9. Gajewski, H., Gröger, K., Zacharias, K.: Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen. Akademie-Verlag, Berlin (1974)
10. Gorban, N.V., Stanzhitsky, A.N.: On the dynamics of solutions for autonomous reaction-diffusion equation in $\mathbb{R}^N$ with multivalued nonlinearity. Ukr. Math. Bull. **6**(2), 235–251 (2009)
11. Kapustyan, O.V., Valero, J.: Comparison between trajectory and global attractors for evolution systems without uniqueness of solutions. Int. J. Bifur. Chaos. **20**, 1–12 (2010). doi:10.1142/S0218127410027313
12. Kasyanov, P.O.: Multivalued dynamics of solutions of an autonomous differential-operator inclusion with pseudomonotone nonlinearity. Cybern. Syst. Anal. **47**(5), 800–811 (2011). doi:10.1007/s10559-011-9359-6
13. Kasyanov, P.O., Toscano, L., Zadoianchuk, N.V.: Regularity of weak solutions and their attractors for a parabolic feedback control problem. Set-Valued Var. Anal. **21**(2), 271–282 (2013)
14. Ladyzhenskaya, O.A.: Dynamical system, generated by Navier-Stokes equations. Zap. Nauch. Sem. LOMI. **27**, 91–115 (1972)
15. Ladyzhenskaya, O.A.: Some comments to my papers on the theory of attractors for abstract semigroups. Zap. Nauchn. Sem. LOMI. **188**, 102–112 (1990)
16. Ladyzhenskaya, O.A.: Attractors for Semigroups and Evolution Equations. University Press, Cambridge (1991)
17. Lions, J.L.: Quelques Méthodes de Résolution des Problémes aux Limites non Linéaires. Dunod, Paris (1969)
18. Melnik, V.S., Valero, J.: On attractors of multivalued semiflows and differential inclusions. Set Valued. Anal. **6**, 83–111 (1998). doi:10.1023/A:1008608431399

19. Migórski, S., Ochal, A.: Optimal control of parabolic hemivariational inequalities. J. Global Optim. **17**, 285–300 (2000)
20. Morillas, F., Valero, J.: Attractors for reaction-diffusion equation in $\mathbb{R}^n$ with continuous nonlinearity. Asymptotic Anal. **44**, 111–130 (2005)
21. Otani, M., Fujita, H.: On existence of strong solutions for $\frac{du}{dt}(t)+\partial\varphi^1(u(t))-\partial\varphi^2(u(t)) \ni f(t)$. J. Fac. Sci. Univ. Tokyo. Sect. 1. A **24**(3), 575–605 (1977)
22. Panagiotopoulos, P.D.: Inequality Problems in Mechanics and Applications: Convex and Nonconvex Energy Functions. Birkhauser, Basel (1985)
23. Sell, G.R., You, Y.: Dynamics of Evolutionary Equations. Springer, New York (2002)
24. Temam, R.: Infinite-Dimensional Dynamical Systems in Mechanics and Physics. Springer, New York (1988)
25. Valero, J.: Attractors of parabolic equations without uniqueness. J. Dyn. Diff. Equat. **13**(4), 711–744 (2001). doi:10.1023/A:1016642525800
26. Valero, J., Kapustyan, A.V.: On the connectedness and asymptotic behaviour of solutions of reaction-diffusion systems. J. Math. Anal. Appl. **323**, 614–633 (2006). doi:10.1016/j.jmaa.2005.10.042
27. Vishik, M.I., Zelik, S.V., Chepyzhov, V.V.: Strong trajectory attractor for dissipative reaction-diffusion system. Doclady Math. **82**(3), 869–873 (2010). doi:10.1134/S1064562410060086
28. Zgurovsky, M.Z., Kasyanov, P.O., Valero, J.: Noncoercive evolution inclusions for Sk type operators. Int. J. Bifurcat. Chaos **20**, 2823–2834 (2010). doi:10.1142/S0218127410027386
29. Zgurovsky, M.Z., Mel'nik, V.S., Kasyanov, P.O.: Evolution Inclusions and Variation Inequalities for Earth Data Processing I. Springer, Berlin (2010a). doi:10.1007/978-3-642-13837-9
30. Zgurovsky, M.Z., Mel'nik V.S., Kasyanov P.O.: Evolution Inclusions and Variation Inequalities for Earth Data Processing II. Springer, Berlin (2010b). doi:10.1007/978-3-642-13878-2
31. Zgurovsky, M.Z., Kasyanov, P.O., Kapustyan, O.V., Valero, J., Zadoianchuk, N.V.: Evolution Inclusions and Variation Inequalities for Earth Data Processing III. Springer, Berlin (2012). doi:10.1007/978-3-642-28512-7

Chapter 16
On Global Attractors for Autonomous Damped Wave Equation with Discontinuous Nonlinearity

Nataliia V. Gorban, Oleksiy V. Kapustyan, Pavlo O. Kasyanov and Liliia S. Paliichuk

Abstract We consider autonomous damped wave equation with discontinuous nonlinearity. The long-term prognosis of the state functions when the conditions on the parameters of the problem do not guarantee uniqueness of solution of the corresponding Cauchy problem are studied. We prove the existence of a global attractor and investigate its structure. It is obtained that trajectory of every weak solution defined on $[0; +\infty)$ tends to a fixed point.

16.1 Introduction

This manuscript is devoted to the research of asymptotical behavior of the autonomous damped wave equation with discontinuous nonlinearity. The investigated problem is considered in a bounded domain Ω with a sufficiently regular boundary $\partial\Omega$. The interaction function $f\colon \mathbb{R} \to \mathbb{R}$ satisfies the standard growth and sign conditions. Wave equation with a non-smooth nonlinearity f can be interpreted as the mathematical model of the controlled piezoelectric fields or processes. The asymptotic behavior of solutions for such problems were studied by Ball [1, 2], Sell [11], Zgurovsky et al. [17–19] and many others. The case of the continuous function f is well-known [2]. The case of the non-autonomous equation with continuous non-

N. V. Gorban (✉) · O. V. Kapustyan · P. O. Kasyanov · L. S. Paliichuk
Institute for Applied System Analysis NAS, Kyiv Polytechnic Institute, National Technical University, Peremogy ave., 37, Kyiv 03056, Ukraine
e-mail: nata_gorban@i.ua

O. V. Kapustyan
e-mail: alexkap@univ.kiev.ua

P. O. Kasyanov
e-mail: kasyanov@i.ua

L. S. Paliichuk
e-mail: lili262808@gmail.com

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_16,

linearity was investigated by Kapustyan [6], Melnik [8, 10], Valero [13]. The case when extension of f admits the maximal monotone graph was studied by Zgurovsky and his scholars [6, 7, 15, 16].

Here we provide sufficient conditions for existence of compact in natural phase space global attractor for the nonlinear damped equation with discontinuous non-monotone in general case interaction function.

16.2 Setting of the Problem

Let $\beta > 0$ be a constant, $\Omega \subset \mathbb{R}^n$ be a bounded domain with sufficiently smooth boundary $\partial\Omega$. Consider the problem

$$\begin{cases} u_{tt} + \beta u_t - \triangle u + f(u) = 0, \\ u|_{\partial\Omega} = 0, \end{cases} \tag{16.1}$$

where $u(x,t)$ is unknown state function defined on $\Omega \times \mathbb{R}_+$; $f : \mathbb{R} \to \mathbb{R}$ is an interaction function such that

$$\lim_{|u|\to\infty} \frac{f(u)}{u} > -\lambda_1, \tag{16.2}$$

where λ_1 is the first eigenvalue for $-\triangle$ in $H_0^1(\Omega)$;

$$\exists\ D \geq 0 : \quad |f(u)| \leq D(1 + |u|), \quad \forall u \in \mathbb{R}. \tag{16.3}$$

Further, we use such denotation

$$\overline{f}(s) := \overline{\lim_{t\to s}} f(t), \quad \underline{f}(s) := \underline{\lim_{t\to s}} f(t), \quad G(s) := [\underline{f}(s), \overline{f}(s)], \quad s \in \mathbb{R}.$$

Let us set $V = H_0^1(\Omega)$ and $H = L^2(\Omega)$. The space $X = V \times H$ is a phase space of Problem (16.1). For the Hilbert space X as $(\cdot,\cdot)_X$ and $\|\cdot\|_X$ denote the inner product and the norm in X respectively.

Definition 16.1 Let $T > 0$, $\tau < T$. The function $\varphi(\cdot) = (u(\cdot), u_t(\cdot))^T \in L^\infty(\tau, T; X)$ is called *a weak solution* of Problem (16.1) on (τ, T) if for a.e. $(x,t) \in \Omega \times (\tau, T)$, there exists $l = l(x,t) \in L^2(\tau, T; L^2(\Omega))$ $l(x,t) \in G(u(x,t))$, such that $\forall \psi \in H_0^1(\Omega)$, $\forall \eta \in C_0^\infty(\tau, T)$,

$$-\int_\tau^T (u_t, \psi)_H \eta_t dt + \int_\tau^T (\beta(u_t, \psi)_H + (u, \psi)_V + (l, \psi)_H)\eta dt = 0. \tag{16.4}$$

The main goal of the manuscript is to obtain the existence of the global attractor generated by the weak solutions of Problem (16.1) in the phase space X.

16.3 Preliminaries

Lemma 16.1 Zgurovsky et al. [19] *For any* $\varphi_\tau = (u_0, u_1)^T \in X$ *and* $\tau < T$ *there exists a weak solution* $\varphi(\cdot)$ *of Problem* (16.1) on (τ, T) such that $\varphi(\tau) = \varphi_\tau$.

Show that in the general case, when the interaction function f is typically multivalued, the m-semiflow generated by all solutions of Problem (16.1) have no a compact global attractor.

Example 16.1 Consider the problem

$$\begin{cases} u_{tt} + \beta u_t - \triangle u + [-\varepsilon, \varepsilon] \ni 0, \quad (x,t) \in (0,\pi) \times \mathbb{R}_+, \\ u(0,t) = u(\pi,t) = 0, \\ u(x,0) = \frac{\varepsilon}{\beta}\varphi_n(x), \quad u_t(x,0) = 0, \quad |\varphi_n'(x)| \le 1. \end{cases} \tag{16.5}$$

There exists a solution $u_n(x,t)$ of Problem (16.5) such that $\{u_n(\cdot,t_n)\}_{n\ge1}$ is not precompact set in $H_0^1(0,\pi)$ for some $\{t_n\}_{n\ge1}$, $t_n \to \infty$, and some bounded in $H_0^1(0,\pi)$ sequence $\{\varphi_n\}$.

D'Alembert's formula implies that Problem (16.5) has the solution of the form

$$u_n(x,t) = \frac{\varepsilon}{2\beta}\left(\varphi_n(x+t) - \varphi_n(t-x)\right)$$

for any sufficiently smooth $\varphi_n : \mathbb{R} \to \mathbb{R}$ such that $\varphi_n(x) = -\varphi_n(-x) = -\varphi_n(2\pi - x)$. Indeed, $u_{n_{tt}} - \triangle u_n = 0$ and

$$\beta u_{n_t}(x,t) = \beta\frac{\varepsilon}{2\pi}\left(\varphi_n'(x+t) - \varphi_n'(t-x)\right) \in [-\varepsilon, \varepsilon].$$

Let $\varphi_n(x) = \frac{1}{n}\sin nx$, $x \in (0,\pi)$. Then

$$u_n(x,t) = \frac{1}{n}\frac{\varepsilon}{2\beta}\left(\sin n(x+t) - \sin n(t-x)\right), \quad (x,t) \in (0,\pi)\times\mathbb{R}_+;$$

$$u'_{n_x}(x,t) = \frac{\varepsilon}{2\beta}\left(\cos n(x+t) + \cos n(t-x)\right), \quad (x,t) \in (0,\pi)\times\mathbb{R}_+.$$

Let $\{t_n\}_{n\ge1} \subset \mathbb{R}_+$ be the sequence such that $t_n = \frac{2\pi}{n} + 2\pi n$, $\forall n \ge 1$. Then

$$\|u_n(\cdot,t_n) - u_m(\cdot,t_m)\|^2_{H_0^1(0,\pi)} =$$

$$= \frac{\varepsilon^2}{4\beta^2} \int_0^\pi (\cos n(x+t_n) + \cos n(t_n - x) - \cos m(x+t_m) - \cos m(t_m - x))^2 \, dx =$$

$$= \frac{\varepsilon^2}{\beta^2} \int_0^\pi (\cos nx - \cos mx)^2 dx = \frac{\pi \varepsilon^2}{\beta^2}, \quad \forall n, m \geq 1.$$

Thus $\{u_n(\cdot, t_n)\}_{n\geq 1}$ is not precompact set in $H_0^1(0,\pi)$, $n \to +\infty$.

Further, we assume that

$$f(s) = f_1(s) - f_2(s), \quad s \in \mathbb{R},$$

where $f_i : \mathbb{R} \to \mathbb{R}$, $i = 1, 2$, are nondecreasing functions.

We remark that

$$[\underline{f}(s), \overline{f}(s)] \subseteq [\underline{f_1}(s), \overline{f_1}(s)] - [\underline{f_2}(s), \overline{f_2}(s)], \quad s \in \mathbb{R}.$$

Thus we consider more general evolution inclusion

$$\begin{cases} u_{tt} + \beta u_t - \triangle u + [\underline{f_1}(u), \overline{f_1}(u)] - [\underline{f_2}(u), \overline{f_2}(u)] \ni 0, \\ u|_{\partial\Omega} = 0. \end{cases} \tag{16.6}$$

Let us set

$$G_i(s) := \int_0^s f_i(\xi) d\xi, \quad J_i(u) := \int_\Omega G_i(u(x)) dx, \quad J(u) = J_1(u) - J_2(u), \quad u \in H, \ i = 1, 2.$$

The functionals G_i and J_i are locally Lipschitz and regular; Clarke [3, Chap. I]. Thus the next result holds.

Lemma 16.2 Kasyanov et al. [9] *Let $u \in C^1([\tau, T]; H)$. Then for a.e. $t \in (\tau, T)$, the functions $J_i \circ u$ are classically differentiable at the point t. Moreover,*

$$\frac{d}{dt}(J_i \circ u)(t) = (p, u_t(t)) \ \forall p \in \partial J_i(u(t)), \ \ i = 1, 2,$$

and $\frac{d}{dt}(J_i \circ u)(\cdot) \in L_1(\tau, T)$.

Consider $W_\tau^T = C([\tau, T]; X)$. Lebourgue's mean value theorem (see Clarke [3, Chap. 2]) provides the existence of constants $c_1, c_2 > 0$ and $\mu \in (0, \lambda_1)$ such that

$$|J(u)| \leq c_1(1 + \|u\|_H^2), \quad J(u) \geq -\frac{\mu}{2}\|u\|_H^2 - c_2 \quad \forall u \in H. \tag{16.7}$$

The weak solution of the Problem (16.1) with initial data

$$u(\tau) = a, \quad u'(\tau) = b \tag{16.8}$$

on the interval $[\tau, T]$ exists for any $a \in V$, $b \in H$. It follows from Zadoianchuk and Kasyanov [15, Theorem 1.4]. Thus the next lemma holds true (see Kasyanov et al. [9, Lemma 3.2]).

Lemma 16.3 Kasyanov et al. [9, Lemma 3.2] *For any $\tau < T$, $a \in V$, $b \in H$, Cauchy Problem (16.1), (16.8) has the weak solution $(u, u_t)^T \in L_\infty(\tau, T; X)$. Moreover, each weak solution $(u, u_t)^T$ of Cauchy Problem (16.1), (16.8) on the interval $[\tau, T]$ belongs to the space $C([\tau, T]; X)$ and $u_{tt} \in L_2(\tau, T; V^*)$.*

16.4 Properties of Solutions

For any $\varphi_\tau = (a, b)^T \in X$, denote

$$\mathscr{D}_{\tau,T}(\varphi_\tau) = \left\{ (u(\cdot), u_t(\cdot))^T \;\middle|\; \begin{array}{l} (u, u_t)^T \text{is a weak solution of Problem (16.1) on } [\tau, T], \\ u(\tau) = a, u_t(\tau) = b \end{array} \right\}.$$

From Lemma 16.3 it follows that $\mathscr{D}_{\tau,T}(\varphi_\tau) \subset C([\tau, T]; X) = W_\tau^T$. Let us check that translation and concatenation of weak solutions are weak solutions too.

Lemma 16.4 *If $\tau < T$, $\varphi_\tau \in X$, $\varphi(\cdot) \in \mathscr{D}_{\tau,T}(\varphi_\tau)$, then $\forall s$ $\psi(\cdot) = \varphi(\cdot + s) \in \mathscr{D}_{\tau-s,T-s}(\varphi_\tau)$. If $\tau < t < T$, $\varphi_\tau \in X$, $\varphi(\cdot) \in \mathscr{D}_{\tau,t}(\varphi_\tau)$ and $\psi(\cdot) \in \mathscr{D}_{t,T}(\varphi_\tau)$, then*

$$\theta(s) = \begin{cases} \varphi(s), & s \in [\tau, t], \\ \psi(s), & s \in [t, T] \end{cases} \in \mathscr{D}_{\tau,T}(\varphi_\tau).$$

Proof The proof is trivial (see Kasyanov et al. [9, Lemma 4.1]).

Let $\varphi = (a, b)^T \in X$ and

$$\mathscr{V}(\varphi) = \frac{1}{2}\|\varphi\|_X^2 + J_1(a) - J_2(a). \tag{16.9}$$

Lemma 16.5 *Let $\tau < T$, $\varphi_\tau \in X$, $\varphi(\cdot) = (u(\cdot), u_t(\cdot))^T \in \mathscr{D}_{\tau,T}(\varphi_\tau)$. Then $\mathscr{V} \circ \varphi : [\tau, T] \to \mathbb{R}$ is absolutely continuous and for a.e. $t \in (\tau, T)$, $\frac{d}{dt}\mathscr{V}(\varphi(t)) = -\beta\|u_t(t)\|_H^2$.*

Proof Let $-\infty < \tau < T < +\infty$, $\varphi(\cdot) = (u(\cdot), u_t(\cdot))^T \in W_\tau^T$ be an arbitrary weak solution of Problem (16.1) on (τ, T). Since $\partial J(u(\cdot)) \subset L_2(\tau, T; H)$, from Temam [12] and Zgurovsky et al. [19, Chap. 2] we obtain that the function $t \to \|u_t(t)\|_H^2 + \|u(t)\|_V^2$ is absolutely continuous and for a.e. $t \in (\tau, T)$,

$$\begin{aligned}\tfrac{1}{2}\tfrac{d}{dt}\left[\|u_t(t)\|_{\mathrm{H}}^2+\|u(t)\|_{\mathrm{V}}^2\right]&=(u_{tt}(t)-\triangle u(t),u_t(t))_{\mathrm{H}}=\\&=-\beta\|u_t(t)\|_{\mathrm{H}}^2-(d_1(t),u_t(t))_{\mathrm{H}}+(d_2(t),u_t(t))_{\mathrm{H}},\end{aligned}\tag{16.10}$$

where $d_i(t)\in\partial J_i(u(t))$ for a.e. $t\in(\tau,T)$ and $d_i(\cdot)\in L_2(\tau,T;H)$, $i=1,2$. As $u(\cdot)\in C^1([\tau,T];H)$ and $J_i:H\to\mathbb{R}$, $i=1,2$ is regular and locally Lipschitz, due to Lemma 16.2 we obtain that for a.e. $t\in(\tau,T)$, $\exists\frac{d}{dt}(J_i\circ u)(t)$, $i=1,2$. Moreover, $\frac{d}{dt}(J_i\circ u)(\cdot)\in L_1(\tau,T)$, $i=1,2$ and for a.e. $t\in(\tau,T)$, $\forall p\in\partial J_i(u(t))$,

$$\frac{d}{dt}(J_i\circ u)(t)=(p,u_t(t))_H,\quad i=1,2.$$

In particular for a.e. $t\in(\tau,T)$, $\frac{d}{dt}(J_i\circ u)(t)=(d_i(t),u_t(t))_H$. Taking into account (16.10) we finally obtain the necessary statement.

This completes the proof.

Lemma 16.6 *Let $T>0$. Then any weak solution of Problem* (16.1) on $[0,T]$ can be extended to a global one defined on $[0,+\infty)$.

Proof The statement of this lemma follows from Lemmas 16.3–16.5, (16.7) and from the next estimates

$$\forall\tau<T,\ \ \forall t\in[\tau,T],\ \ \forall\varphi_\tau\in X,\ \ \forall\varphi(\cdot)=(u(\cdot),u_t(\cdot))^T\in\mathscr{D}_{\tau,T}(\varphi_\tau),$$

$$\begin{aligned}2c_1+\left(1+\tfrac{2c_1}{\lambda_1}\right)\|u(\tau)\|_V^2+\|u_t(\tau)\|_H^2&\ge 2\mathscr{V}(\varphi(\tau))\ge 2\mathscr{V}(\varphi(t))=\\=\|u(t)\|_V^2+\|u_t(t)\|_H^2+2J(u(t))&\ge\left(1-\tfrac{\mu}{\lambda_1}\right)\|u(t)\|_V^2+\|u_t(t)\|_H^2-2c_2.\end{aligned}$$

The lemma is proved.

For an arbitrary $\varphi_0\in X$ let $\mathscr{D}(\varphi_0)$ be the set of all weak solutions (defined on $[0,+\infty)$) of Problem (16.1) with initial data $\varphi(0)=\varphi_0$. We remark that from the proof of Lemma 16.6 we obtain the next corollary.

Corollary 16.1 *For any $\varphi_0\in X$ and $\varphi\in\mathscr{D}(\varphi_0)$, the next inequality is fulfilled*

$$\|\varphi(t)\|_X^2\le\frac{\lambda_1+2c_1}{\lambda_1-\mu}\|\varphi(0)\|_X^2+\frac{2(c_1+c_2)\lambda_1}{\lambda_1-\mu}\quad\forall t>0.\tag{16.11}$$

From Corollary 16.1 in a standard way we obtain such statement.

Theorem 16.1 *Let $\tau<T$, $\{\varphi_n(\cdot)\}_{n\ge1}\subset W_\tau^T$ be an arbitrary sequence of weak solutions of Problem* (16.1) on $[\tau,T]$ such that $\varphi_n(\tau)\to\varphi_\tau$ weakly in X, $n\to+\infty$, and let $\{t_n\}_{n\ge1}\subset[\tau,T]$ be a sequence such that $t_n\to t_0$, $n\to+\infty$. Then there exists $\varphi\in\mathscr{D}_{\tau,T}(\varphi_\tau)$ such that up to a subsequence $\varphi_n(t_n)\to\varphi(t_0)$ weakly in X, $n\to+\infty$.

Proof We prove this theorem in several steps.

Step 1. Let $\tau < T$, $\{\varphi_n(\cdot) = (u_n(\cdot), u_n'(\cdot))\}_{n\geq 1} \subset W_\tau^T$ be an arbitrary sequence of weak solutions of Problem (16.1) on $[\tau, T]$ and $\{t_n\}_{n\geq 1} \subset [\tau, T]$ such that

$$\varphi_n(\tau) \to \varphi_\tau \text{ weakly in } X, \quad t_n \to t_0, \quad n \to +\infty. \tag{16.12}$$

In virtue of Corollary 16.1 we have that $\{\varphi_n(\cdot)\}_{n\geq 1}$ is bounded on $W_\tau^T \subset L_\infty(\tau, T; X)$. Therefore up to a subsequence $\{\varphi_{n_k}(\cdot)\}_{k\geq 1} \subset \{\varphi_n(\cdot)\}_{n\geq 1}$ we have

$$\begin{gathered}
u_{n_k} \to u \text{ weakly star in } L_\infty(\tau, T; V), \ k \to +\infty, \\
u'_{n_k} \to u' \text{ weakly star in } L_\infty(\tau, T; H), \ k \to +\infty, \\
u''_{n_k} \to u'' \text{ weakly star in } L_\infty(\tau, T; V^*), \ k \to +\infty, \\
d_{n_k,i} \to d_i \text{ weakly star in } L_\infty(\tau, T; H), \ i = \overline{1,2}, \ k \to +\infty, \\
u_{n_k} \to u \text{ in } L_2(\tau, T; H), \ k \to +\infty, \\
u_{n_k}(t) \to u(t) \text{ in } H \text{ for a.e. } t \in [\tau, T], \ k \to +\infty, \\
u'_{n_k}(t) \to u'(t) \text{ in } V^* \text{ for a.e. } t \in (\tau, T), \ k \to +\infty, \\
\triangle u_{n_k} \to \triangle u \text{ weakly in } L_2(\tau, T; V^*), \ k \to +\infty,
\end{gathered} \tag{16.13}$$

where $\forall n \geq 1$ $d_{n,i} \in L_2(\tau, T; H)$ and

$$\begin{gathered}
u_n''(t) + \beta u_n'(t) + d_{n,1}(t) - d_{n,2}(t) - \triangle u_n(t) = \bar{0}, \\
d_{n,i}(t) \in \partial J_i(u_n(t)), \quad i = 1, 2, \quad \text{for a.e. } t \in (\tau, T).
\end{gathered} \tag{16.14}$$

Step 2. ∂J_i, $i = 1, 2$ are demiclosed. So, by a standard way we get that $d_i(\cdot) \in \partial J_i(u(\cdot))$, $i = 1, 2$, $\varphi := (u, u') \in \mathscr{D}_{\tau,T}(\varphi_\tau) \subset W_\tau^T$.

Step 3. From (16.13) it follows that for arbitrary fixed $h \in V$ the sequences of real functions $(u_{n_k}(\cdot), h)_H$, $(u'_{n_k}(\cdot), h)_H : [\tau, T] \to \mathbb{R}$ are uniformly bounded and equipotentionally continuous. Taking into account (16.13), (16.11) and density of the embedding $V \subset H$ we obtain that $u'_{n_k}(t_{n_k}) \to u'(t_0)$ weakly in H and $u_{n_k}(t_{n_k}) \to u(t_0)$ weakly in V as $k \to +\infty$.

The theorem is proved.

Theorem 16.2 *Let $\tau < T$, $\{\varphi_n(\cdot)\}_{n\geq 1} \subset W_\tau^T$ be an arbitrary sequence of weak solutions of Problem* (16.1) on $[\tau, T]$ such that $\varphi_n(\tau) \to \varphi_\tau$ strongly in X, $n \to +\infty$, then up to a subsequence $\varphi_n(\cdot) \to \varphi(\cdot)$ in $C([\tau, T]; X)$, $n \to +\infty$.

Proof Let $\tau < T$, $\{\varphi_n(\cdot) = (u_n(\cdot), u_n'(\cdot))^T\}_{n\geq 1} \subset W_\tau^T$ be an arbitrary sequence of weak solutions of Problem (16.1) on $[\tau, T]$ and $\{t_n\}_{n\geq 1} \subset [\tau, T]$:

$$\varphi_n(\tau) \to \varphi_\tau \text{ strongly in } X, \ n \to +\infty. \tag{16.15}$$

From Theorem 16.1 we have that there exists $\varphi \in \mathscr{D}_{\tau,T}(\varphi_\tau)$ such that up to the subsequence $\{\varphi_{n_k}(\cdot)\}_{k\geq 1} \subset \{\varphi_n(\cdot)\}_{n\geq 1}$ $\varphi_n(\cdot) \to \varphi(\cdot)$ weakly in X, uniformly on $[\tau, T]$, $k \to +\infty$. Let us prove that

$$\varphi_{n_k} \to \varphi \text{ in } W_\tau^T, \quad k \to +\infty. \tag{16.16}$$

By contradiction, suppose the existence of $L > 0$ and the subsequence $\{\varphi_{k_j}\}_{j\geq 1} \subset \{\varphi_{n_k}\}_{k\geq 1}$ such that $\forall j \geq 1$,

$$\max_{t\in[\tau,T]} \|\varphi_{k_j}(t) - \varphi(t)\|_X = \|\varphi_{k_j}(t_j) - \varphi(t_j)\|_X \geq L.$$

Without loss of generality we suggest that $t_j \to t_0 \in [\tau, T]$, $j \to +\infty$. Therefore by virtue of a continuity of $\varphi : [\tau, T] \to X$ we have

$$\underline{\lim}_{j\to+\infty} \|\varphi_{k_j}(t_j) - \varphi(t_0)\|_X \geq L. \tag{16.17}$$

On the other hand, we prove that

$$\varphi_{k_j}(t_j) \to \varphi(t_0) \text{ in } X, \quad j \to +\infty. \tag{16.18}$$

First we remark that

$$\varphi_{k_j}(t_j) \to \varphi(t_0) \text{ weakly in } X, \quad j \to +\infty \tag{16.19}$$

(see Theorem 16.1 for details). Secondly let us prove that

$$\overline{\lim}_{j\to+\infty} \|\varphi_{k_j}(t_j)\|_X \leq \|\varphi(t_0)\|_X. \tag{16.20}$$

Since J is sequentially weakly continuous, $\mathscr{V}$ is sequentially weakly lower semi-continuous on X. Hence we obtain

$$\begin{aligned} \mathscr{V}(\varphi(t_0)) &\leq \underline{\lim}_{j\to+\infty} \mathscr{V}(\varphi_{k_j}(t_j)), \\ \int_\tau^{t_0} \|u'(s)\|_H^2 ds &\leq \underline{\lim}_{j\to+\infty} \int_\tau^{t_j} \|u'_{k_j}(s)\|_H^2 ds \end{aligned} \tag{16.21}$$

and

$$\mathscr{V}(\varphi(t_0)) + \beta \int_\tau^{t_0} \|u'(s)\|_H^2 ds \leq \underline{\lim}_{j\to+\infty} \left(\mathscr{V}(\varphi_{k_j}(t_j)) + \beta \int_\tau^{t_j} \|u'_{k_j}(s)\|_{\mathrm{H}}^2 ds \right). \tag{16.22}$$

Since by the energy equation both sides of (16.22) equal $\mathscr{V}(\varphi(\tau))$ (see Lemma 16.5), it follows from (16.21) that $\mathscr{V}(\varphi_{k_j}(t_j)) \to \mathscr{V}(\varphi(t_0))$, $j \to +\infty$ and (16.20). Convergence (16.18) directly follows from (16.19), (16.20) and Gajewski et al. [5, Chap. I]. To finish the proof of the theorem we remark that (16.18) contradicts (16.17). Therefore (16.16) holds.

The theorem is proved.

Define the m-semiflow $\mathscr{G}$ as

$$\mathscr{G}(t, \xi_0) = \{\xi(t) \mid \xi(\cdot) \in \mathscr{D}(\xi_0)\}, \quad t \geq 0.$$

Denote the set of all nonempty (nonempty bounded) subsets of X by $P(X)(\beta(X))$. Note that the multivalued map $\mathscr{G} : \mathbb{R}_+ \times X \to P(X)$ is *a strict m-semiflow*, i.e., (see Lemma 16.4)

1. $\mathscr{G}(0, \cdot) = \mathrm{Id}$ (the identity map);
2. $\mathscr{G}(t + s, x) = \mathscr{G}(t, \mathscr{G}(s, x))\ \forall x \in X, t, s \in \mathbb{R}_+$.

Further, $\varphi \in \mathscr{G}$ means that $\varphi \in \mathscr{D}(\xi_0)$ for some $\xi_0 \in X$.

Definition 16.2 $\mathscr{G}$ is called *an asymptotically compact* m-semiflow if for any sequence $\{\varphi_n\}_{n\geq 1} \subset \mathscr{G}$ with $\{\varphi_n(0)\}_{n\geq 1}$ bounded, and for any sequence $\{t_n\}_{n\geq 1}$: $t_n \to +\infty$, $n \to \infty$, the sequence $\{\varphi_n(t_n)\}_{n\geq 1}$ has a convergent subsequence Ball [2, p. 35].

Theorem 16.3 *$\mathscr{G}$ is an asymptotically compact m-semiflow.*

Proof Let $\xi_n \in \mathscr{G}(t_n, v_n)$, $v_n \in B$, $B \in \beta(X)$, $n \geq 1$, $t_n \to +\infty$, $n \to +\infty$. Let us check a precompactness of $\{\xi_n\}_{n\geq 1}$ in X. Without loss of the generality, we extract a convergent in X subsequence from $\{\xi_n\}_{n\geq 1}$. From Corollary 16.1 we obtain that there exists $\{\xi_{n_k}\}_{k\geq 1}$ and $\xi \in X$ such that $\xi_{n_k} \to \xi$ weakly in X, $\|\xi_{n_k}\|_X \to a \geq \|\xi\|_X$, $k \to +\infty$. Show that $a \leq \|\xi\|_X$.

Let us fix an arbitrary $T_0 > 0$. Then for rather big $k \geq 1$, $\mathscr{G}(t_{n_k}, v_{n_k}) \subset \mathscr{G}(T_0, \mathscr{G}(t_{n_k} - T_0, v_{n_k}))$. Hence $\xi_{n_k} \in \mathscr{G}(T_0, \beta_{n_k})$, where $\beta_{n_k} \in \mathscr{G}(t_{n_k} - T_0, v_{n_k})$ and $\sup\limits_{k\geq 1} \|\beta_{n_k}\|_X < +\infty$ (see Corollary 16.1). From Theorem 16.1 for some $\{\xi_{k_j}, \beta_{k_j}\}_{j\geq 1} \subset \{\xi_{n_k}, \beta_{n_k}\}_{k\geq 1}$, $\beta_{T_0} \in X$, we obtain

$$\xi \in \mathscr{G}(T_0, \beta_{T_0}), \quad \beta_{k_j} \to \beta_{T_0} \text{ weakly in } X, \quad j \to +\infty. \tag{16.23}$$

From the definition of $\mathscr{G}$ we set $\forall j \geq 1$, $\xi_{k_j} = (u_j(T_0), u_j'(T_0))^T$, $\beta_{k_j} = (u_j(0), u_j'(0))^T$, $\xi = (u_0(T_0), u_0'(T_0))^T$, $\beta_{T_0} = (u_0(0), u_0'(0))^T$, where $\varphi_j = (u_j, u_j')^T \in C([0, T_0]; X)$, $u_j'' \in L_2(0, T_0; V^*)$, $d_j \in L_\infty(0, T_0; H)$,

$$u_j''(t) + \beta u_j'(t) - \triangle u_j(t) + d_{j,1}(t) - d_{j,2}(t) = \bar{0},$$

$$d_{j,i}(t) \in \partial J_i(u_j(t)), \quad i = 1, 2 \quad \text{for a.e.} \quad t \in (0, T_0).$$

Let for every $t \in [0, T_0]$,

$$I(\varphi_j(t)) := \frac{1}{2}\|\varphi_j(t)\|_X^2 + J_1(u_j(t)) - J_2(u_j(t)) + \frac{\beta}{2}(u_j'(t), u_j(t))_H.$$

Then in virtue of Lemma 16.2, Gajewski et al. [5, Chap. IV], Temam [12] and Zgurovsky et al. [19]

$$\frac{dI(\varphi_j(t))}{dt} = -\beta I(\varphi_j(t)) + \beta\mathscr{H}(\varphi_j(t)), \text{ for a.e. } t \in (0, T_0),$$

where

$$\mathscr{H}(\varphi_j(t)) = J_1(u_j(t)) - \frac{1}{2}(d_{j,1}(t), u_j(t))_H - J_2(u_j(t)) + \frac{1}{2}(d_{j,2}(t), u_j(t))_H.$$

From (16.11), (16.23) we have $\exists \bar{R} > 0 : \forall j \geq 0, \forall t \in [0, T_0]$,

$$\|u_j'(t)\|_{\mathrm{H}}^2 + \|u_j(t)\|_{\mathrm{V}}^2 \leq \bar{R}^2.$$

Moreover,

$$\begin{gathered} u_j \to u_0 \text{ weakly in } L_2(0, T_0; V), \ \ j \to +\infty, \\ u_j' \to u_0' \text{ weakly in } L_2(0, T_0; H), \ \ j \to +\infty, \\ u_j \to u_0 \text{ in } L_2(0, T_0; H), \ \ j \to +\infty, \\ d_{j,i} \to d_i \text{ weakly in } L_2(0, T_0; H), \ \ i = 1, 2, \ \ j \to +\infty, \\ u_j'' \to u_0'' \text{ weakly in } L_2(0, T_0; V^*), \ \ j \to +\infty, \\ \forall t \in [0, T_0] \ \ u_j(t) \to u_0(t) \text{ in } H, \ \ j \to +\infty. \end{gathered} \tag{16.24}$$

For every $j \geq 0$ and $t \in [0, T_0]$,

$$I(\varphi_j(t)) = I(\varphi_j(0))e^{-\beta t} + \int_0^t \mathscr{H}(\varphi_j(s))e^{-\beta(t-s)}ds.$$

In particular $I(\varphi_j(T_0)) = I(\varphi_j(0))e^{-\beta T_0} + \int_0^{T_0} \mathscr{H}(\varphi_j(s))e^{-\beta(T_0-s)}ds.$

From (16.24) and Lemma 16.2 we have

$$\int_0^{T_0} \mathscr{H}(\varphi_j(s))e^{-\beta(T_0-s)}ds \to \int_0^{T_0} \mathscr{H}(\varphi_0(s))e^{-\beta(T_0-s)}ds, \ \ j \to +\infty.$$

Therefore

$$\begin{aligned} \varlimsup_{j\to+\infty} I(\varphi_j(T_0)) \leq \ & \varlimsup_{j\to+\infty} I(\varphi_j(0))e^{-\beta T_0} + \int_0^{T_0} \mathscr{H}(\varphi_0(s))e^{-\beta(T_0-s)}ds = \\ = I(\varphi_0(T_0)) + & \left[\varlimsup_{j\to+\infty} I(\varphi_j(0)) - I(\varphi_0(0))\right] e^{-\beta T_0} \leq I(\varphi_0(T_0)) + c_3 e^{-\beta T_0}, \end{aligned}$$

where c_3 does not depend on $T_0 > 0$.

On the other hand, from (16.24) we have

$$\overline{\lim_{j\to+\infty}} I(\varphi_j(T_0)) \geq \frac{1}{2} \lim_{j\to+\infty} \|\varphi_j(T_0)\|_X^2 + J(u_0(T_0)) + \frac{\beta}{2}(u_0'(T_0), u_0(T_0))_H.$$

Therefore we obtain $\frac{1}{2}a^2 \leq \frac{1}{2}\|\xi\|_X^2 + c_3 e^{-\beta T_0}\ \forall T_0 > 0$.

Thus, $a \leq \|\xi\|_X$.

The Theorem is proved.

Let us consider the family $\mathscr{K}_+ = \cup_{u_0\in X}\mathscr{D}(u_0)$ of all weak solutions of Problem (16.1) defined on $[0, +\infty)$. Note that $\mathscr{K}_+$ is *translation invariant one*, i.e., $\forall u(\cdot) \in \mathscr{K}_+$, $\forall h \geq 0$, $u_h(\cdot) \in \mathscr{K}_+$, where $u_h(s) = u(h+s)$, $s \geq 0$. On $\mathscr{K}_+$ we set the *translation semigroup* $\{T(h)\}_{h\geq 0}$, $T(h)u(\cdot) = u_h(\cdot)$, $h \geq 0$, $u \in \mathscr{K}_+$. In view of the translation invariance of $\mathscr{K}_+$ we conclude that $T(h)\mathscr{K}_+ \subset \mathscr{K}_+$ as $h \geq 0$.

On $\mathscr{K}_+$ we consider a topology induced from the Fréchet space $C^{loc}(\mathbb{R}_+; X)$. Note that

$$f_n(\cdot) \to f(\cdot) \text{ in } C^{loc}(\mathbb{R}_+; X) \iff \forall M > 0,\ \Pi_M f_n(\cdot) \to \Pi_M f(\cdot) \text{ in } C([0, M]; X),$$

where Π_M is the restriction operator to the interval $[0, M]$; Vishik and Chepyzhov [14, p. 179]. We denote the restriction operator to $[0, +\infty)$ by Π_+.

Let us consider Problem (16.1) on the entire time axis. Similarly to the space $C^{loc}(\mathbb{R}_+; X)$ the space $C^{loc}(\mathbb{R}; X)$ is endowed with the topology of a local uniform convergence on each interval $[-M, M] \subset \mathbb{R}$ (cf. Vishik and Chepyzhov [14, p. 180]). A function $u \in C^{loc}(\mathbb{R}; X) \cap L_\infty(\mathbb{R}; X)$ is said to be *a complete trajectory* of Problem (16.1) if $\forall h \in \mathbb{R}$, $\Pi_+ u_h(\cdot) \in \mathscr{K}_+$; Vishik and Chepyzhov [14, p. 180].

Let $\mathscr{K}$ be a family of *all complete trajectories* of Problem (16.1). Note that $\forall h \in \mathbb{R}$, $\forall u(\cdot) \in \mathscr{K}$ $u_h(\cdot) \in \mathscr{K}$. We say that the complete trajectory $\varphi \in \mathscr{K}$ is *stationary* if $\varphi(t) = z$ for all $t \in \mathbb{R}$ for some $z \in X$. Following Ball [1, p. 486] we denote by $Z(\mathscr{G})$ the set of all rest points of $\mathscr{G}$. Note that

$$Z(\mathscr{G}) = \{(\bar{0}, u) \,|\, u \in V,\ -\triangle(u) + \partial J(u) \ni \bar{0}\}.$$

Lemma 16.7 *$Z(\mathscr{G})$ is an bounded set in X.*

The existence of a Lyapunov function for $\mathscr{G}$ follows from Lemma 16.5 (see Ball [1, p. 486]).

Lemma 16.8 *A functional $\mathscr{V} : X \to \mathbb{R}$ defined by (16.9) is a Lyapunov function for $\mathscr{G}$.*

16.5 The Existence of a Global Attractor

At first we consider constructions presented in Ball [1], Mel'nik and Valero [10]. We recall that the set $\mathscr{A}$ is said to be a *global attractor* $\mathscr{G}$ if

(1) $\mathscr{A}$ is negatively semiinvariant (i.e., $\mathscr{A} \subset \mathscr{G}(t, \mathscr{A})$ $\forall t \geq 0$);

(2) $\mathscr{A}$ is attracting set, i.e.,

$$\operatorname{dist}(\mathscr{G}(t, B), \mathscr{A}) \to 0, \quad t \to +\infty, \quad \forall B \in \beta(X), \tag{16.25}$$

where $\operatorname{dist}(C, D) = \sup\limits_{c \in C} \inf\limits_{d \in D} \|c - d\|_X$ is the Hausdorff semidistance;

(3) for any closed set $Y \subset H$ satisfying (16.25), we have $\mathscr{A} \subset Y$ (minimality).

The global attractor is said to be *invariant* if $\mathscr{A} = \mathscr{G}(t, \mathscr{A})$, $\forall t \geq 0$.

Note that by definition a global attractor is unique.

We prove the existence of an invariant compact global attractor.

Theorem 16.4 *The m-semiflow $\mathscr{G}$ has an invariant compact in the phase space X global attractor $\mathscr{A}$. For each $\psi \in \mathscr{K}$ the limit sets*

$$\alpha(\psi) = \{z \in \mathrm{X} |\ \psi(t_j) \to z \ \text{ for some sequence } \ t_j \to -\infty\},$$

$$\omega(\psi) = \{z \in \mathrm{X} |\ \psi(t_j) \to z \ \text{ for some sequence } \ t_j \to +\infty\}$$

are connected subsets of $Z(\mathscr{G})$ on which $\mathscr{V}$ is constant. If $Z(\mathscr{G})$ is totally disconnected (in particular if $Z(\mathscr{G})$ is countable) the limits

$$z_- = \lim_{t \to -\infty} \psi(t), \quad z_+ = \lim_{t \to +\infty} \psi(t)$$

exist and z_-, z_+ are rest points; furthermore, $\varphi(t)$ tends to a rest point as $t \to +\infty$ for every solution $\varphi \in \mathscr{K}_+$.

Proof The existence of a global attractor for Second Order Evolution Inclusions directly follows from Lemmas 16.3, 16.4, 16.7, 16.8, Theorems 16.1–16.3 and Ball [2, Theorem 2.7].

16.6 Global Attractors for Typically Discontinuous Interaction Functions

Let $\beta > 0$ be a constant, $\Omega \in \mathbb{R}^n$ be a bounded domain with sufficiently smooth boundary $\partial\Omega$. Consider the problem

$$\begin{cases} u_{tt} + \beta u_t - \triangle u \in -f(u) + G(u) + h, \\ u|_{\partial\Omega} = 0, \end{cases} \tag{16.26}$$

where $u(x, t)$ is unknown state function defined on $\Omega \times \mathbb{R}_+$, $h \in L^2(\Omega)$, $f : \mathbb{R} \to \mathbb{R}$ is an interaction function such that

$$f \in \mathbf{C}(\mathbb{R}), \quad G = [g_1, g_2], \quad g_i \in \mathbf{C}(\mathbb{R}), \quad i = 1, 2. \tag{16.27}$$

There exist a small constant $C \geq 0$ ($C < \min\{\beta, \lambda_1\}$), and $D_i \geq 0$, $i = 1, 2$ such that

$$\underline{\lim}_{|u|\to\infty} \frac{f(u)}{u} > -\lambda_1, \tag{16.28}$$

where λ_1 is the first eigenvalue for $-\triangle$ in $H_0^1(\Omega)$,

$$|g_i(u)| \leq C|u| + D_1, \quad \forall u \in \mathbb{R},\ i = 1, 2, \tag{16.29}$$

$$|f(u)| \leq D_2(1 + |u|^{\frac{n}{n-2}}), \quad \forall u \in \mathbb{R}. \tag{16.30}$$

Remark 16.1 The case of ε–neighborhood of $f(u)$ satisfies conditions (16.27)–(16.30), i.e., if $\underline{\lim}_{|u|\to\infty} \frac{f(u)}{u} > -\lambda_1$, $G(u) = [-\varepsilon, \varepsilon]$.

Let us set $V = H_0^1(\Omega)$ and $H = L^2(\Omega)$. The space $X = V \times H$ is a phase space of Problem (16.26).

Definition 16.3 Let $T > 0$. The function $\varphi(\cdot) = (u(\cdot), u_t(\cdot))^T \in L^\infty(0, T, X)$ is called *a weak solution* of Problem (16.26) on $(0, T)$ if for a.e. $(x, t) \in \Omega \times (0, T)$, there exists $l = l(x, t) \in L^2(0, T : L^2(\Omega))$, $l(x, t) \in G(u(x, t))$ such that $\forall \psi \in H_0^1(\Omega)$, $\eta \in \mathbf{C}_0^\infty(0, T)$

$$-\int_0^T (u_t, \psi)_H \eta_t \, dt + \int_0^T [(\beta(u_t, \psi)_H +$$

$$+(u, \psi)_V + (f(u), \psi)_H - (l, \psi)_H - (h, \psi)_H)\eta] \, dt = 0.$$

Lemma 16.9 *For all $\varphi_0 = (u_0, u_1)^T \in X$, $T > 0$, there exists a weak solution $\varphi(\cdot)$ of Problem (16.26) such that $\varphi(0) = \varphi_0$. Moreover, if $\varphi(\cdot) = (u(\cdot), u_t(\cdot))^T$ is a weak solution of Problem (16.26) with respective $l \in L^2(0, T; L^2(\Omega))$, then $\varphi \in \mathbf{C}([0, T]; X)$, functions*

$$t \mapsto \|u_t(t)\|_H^2 + \|u(t)\|_V^2, \quad t \mapsto (F(u(t)), 1)_H$$

are absolutely continuous on $[0, T]$, and for $t, s \in [0, T]$, $s \leq t$,

$$\frac{1}{2}\frac{d}{dt}(\|u_t(t)\|_H^2 + \|u(t)\|_V^2 + (F(u(t)), 1)_H) =$$

$$= -\beta\|u_t(t)\|_H^2 + (l(t), u_t(t))_H + (h, u_t(t))_H, \tag{16.31}$$

$$\|u_t(t)\|_H^2 + \|u(t)\|_V^2 \leq e^{-\delta(t-s)} \left(\|u_t(s)\|_H^2 + \|u(s)\|_V^{\frac{2n-2}{n-2}} \right) + D_3, \tag{16.32}$$

where $F(u) = \int_0^u f(s)ds$, $u \in \mathbb{R}$ and constants $\delta > 0$, $D_3 > 0$ do not depend on φ.

Proof Let us deduce condition (16.32). Consider

$$Y(t) = \frac{1}{2}\|u_t(t)\|_H^2 + \frac{1}{2}\|u(t)\|_V^2 + (F(u(t)), 1)_H + \alpha(u_t(t), u(t))_H, \ t \in [0, T],$$

where $\alpha > 0$. Then for sufficiently small $C > 0$ and $\delta > 0$

$$\begin{aligned}
\tfrac{dY(t)}{dt} &= (u_{tt}(t), u_t(t))_H - (\triangle u, u_t(t))_H + (f(u(t)), u_t(t))_H + \\
&+ \alpha(u_{tt}(t), u(t))_H + \alpha\|u_t(t)\|_H^2 = \\
&= (-\beta u_t(t) + l(t) + h, u_t(t))_H + \alpha\|u_t(t)\|_H^2 + \\
&+ \alpha(-\beta u_t(t) + \triangle u - f(u(t)) + l(t) + h, u(t))_H = \\
&= -(\beta - \alpha)\|u_t(t)\|_H^2 + (l(t), u_t(t))_H + (h, u_t(t))_H - \\
&- \alpha\beta(u_t(t), u(t))_H - \alpha\|u(t)\|_V^2 - \\
&- \alpha(f(u(t)), u(t))_H + \alpha(l(t) + h, u(t))_H \leq \\
&\leq -(\beta - \alpha - \varepsilon)\|u_t(t)\|_H^2 + C\|u(t)\|_H\|u_t(t)\|_H \\
&\leq -\alpha\|u(t)\|_V^2 - \alpha(-\lambda_1 + C + \varepsilon)\|u(t)\|_H^2 + \\
&+ \alpha C\|u(t)\|_H^2 + K \leq -\delta Y(t) + \tilde{K}.
\end{aligned}$$

Therefore the inequalities

$$F(u) \geq \left(-\frac{\lambda_1}{2} + \varepsilon\right)u^2 + L, \quad F(u) \leq M\left(1 + |u|^{\frac{2u-2}{u-2}}\right), \quad \forall u \in \mathbb{R}, \qquad (16.33)$$

imply (16.32). All the other statements follow from Ball [2], Temam [12]. The existence of a solution follows from the existence of a continuous selector for G.

Remark 16.2 The set of solutions of Problem (16.26) is not covered by all continuous selectors of $G : \mathbb{R} \mapsto 2^{\mathbb{R}}$.

Indeed, let $f \equiv 0$, $G(u) \equiv [-\varepsilon, \varepsilon]$, $h \equiv 0$. Consider solutions of the problem

$$\begin{cases} \triangle u \in [-\varepsilon, \varepsilon], & \text{in } \Omega = (0, \pi), \\ u(0) = u(\pi) = 0, \end{cases}$$

i.e., consider stationary solutions of Problem (16.26). Then the function

$$u(x) = \frac{\varepsilon}{2}\sin x + \frac{\varepsilon}{8}\sin 2x, \quad x \in (0, \pi),$$

is a solution of the given problem but there is no $g \in \mathbf{C}(\mathbb{R})$ such that $g(u) \in [-\varepsilon, \varepsilon]$, $\forall u \in \mathbb{R}$, and $\triangle u(x) = g(u(x))$, $x \in (0, \pi)$. Indeed, assume the converse. Suppose that such function exists. The equation

$$\frac{\varepsilon}{2}\sin x + \frac{\varepsilon}{8}\sin 2x = \frac{\varepsilon}{2}$$

has two solutions

$$x = \frac{\pi}{2} \text{ and } x = x^* \neq \frac{\pi}{2} \in (0, \pi).$$

If $x = \frac{\pi}{2}$, then

$$g(\frac{\varepsilon}{2}) = u''(\frac{\pi}{2}) = -\frac{\varepsilon}{2}.$$

If $x = x^*$, then

$$g(\frac{\varepsilon}{2}) = -\frac{\varepsilon}{2}\sin x^* - \frac{\varepsilon}{2}\sin 2x^* = -\frac{\varepsilon}{2} - \frac{3\varepsilon}{8}\sin 2x^* \neq -\frac{\varepsilon}{2}.$$

This contradiction concludes the example.

Remark 16.3 If $G(u) \equiv g(u)$ is a single-valued function, then the existence of a global attractor was proved in Ball [2].

Select the class of solutions for which there exists a global attractor. For this purpose we use the notion of "energy" equation Ball [2], which describes the conservation laws of energy.

Let $\varphi \in C([0, +\infty); X)$ is a solution of Problem (16.26). Denote

$$I(\varphi) = \frac{1}{2}\|u_t(t)\|_H^2 + \frac{1}{2}\|u(t)\|_V^2 + (F(u(t)), 1)_H + \frac{\beta}{2}(u_t(t), u(t))_H,$$

$$g_\lambda(u) = \lambda g_1(u) + (1 - \lambda g_2(u)), \quad G_\lambda(u) = \int_0^u g_\lambda(s)ds, \quad \lambda \in [0, 1],$$

$$H(\varphi) = \beta(F(u), 1)_H - \frac{\beta}{2}(f(u), u)_H + \frac{\beta}{2}(h, u)_H + (h, u_t)_H.$$

Definition 16.4 A weak solution φ of Problem (16.26) with the corresponding function l is called an *energy solution* if there exists $\lambda \in [0, 1]$ $(\lambda = \lambda(\varphi))$ such that $\forall t \geq 0$,

$$\frac{d}{dt}I(\varphi(t)) + \beta I(\varphi(t)) - \frac{d}{dt}(G_\lambda(u(t)), 1)_H = \frac{\beta}{2}(l(t), u(t))_H + H(\varphi(t)). \quad (16.34)$$

Remark 16.4 Any solution satisfies the equation

$$\frac{d}{dt}I(\varphi(t)) + \beta I(\varphi(t)) - (l(t), u_t(t))_H = \frac{\beta}{2}(l(t), u(t))_H + H(\varphi(t)).$$

Any "selector" solution satisfies the equation

$$\frac{d}{dt}I(\varphi(t)) + \beta I(\varphi(t)) - (g(u(t)), u_t(t))_H = \frac{\beta}{2}(g(u(t)), u(t))_H + H(\varphi(t)).$$

Remark 16.5 Any stationary solution $u(t)$ obviously satisfies (16.34). So, the set of all "selector" solutions (solutions of Problem (16.26) with $l(x,t) = g(u(x,t))$, $g \in G$) does not include the set of energy solutions. Moreover, the set of all energy solutions is wider than the set of all solutions of (16.26) with $l(x,t) = g_\lambda(u(x,t))$.

Let us set

$$\mathscr{G}(t, \varphi_0) = \{\varphi(t) \mid \varphi(\cdot) \quad \text{is an energy solution of(16.26)}, \varphi(0) = \varphi_0\} \tag{16.35}$$

Theorem 16.5 *The m-semiflow $\mathscr{G}$ has an invariant compact in the phase space X global attractor.*

Proof $\mathscr{G}$ is the m-semiflow (but not strict; it will be strict if in the definition 16.4 $[0, +\infty)$ is divided into intervals with different λ). Note that $\mathscr{G}$ is dissipative; $\mathscr{G}$ has a closed graph (it is necessary to pass to the limit in (16.34)); $\mathscr{G}$ is asymptotically semicompact m-semiflow. Indeed, similarly to Ball [2] we obtain the equation

$$\begin{aligned} &I(\varphi_j(t_j)) - (G_{\lambda_j}(u_j(t_j)), 1)_H = \\ &= \left(I(\varphi_j(t_j - M)) - (G_{\lambda_j}(\varphi_j(t_j - M)), 1)_H\right) e^{-\beta M} + \int_0^M e^{\beta(t-M)} \cdot \\ &\cdot \left(H(\varphi_j(t)) + \tfrac{\beta}{2}(l_j(t), u_j(t))_H - \beta(G_{\lambda_j}(\varphi_j(t)), 1)_H\right) dt. \end{aligned} \tag{16.36}$$

Since up to a subsequence $\lambda_j \to \lambda$, $\varphi_j(t_j) \to \chi$ weakly in $H_0^1(\Omega)$, we obtain

$$\left(G_{\lambda_j}(\varphi_j(t_j)), 1\right)_H \to (G_\lambda(\chi), 1)_H$$

and similarly Ball [2] we have

$$I(\varphi_j(t_j)) \to I(\chi).$$

Remark 16.6 It is possible to build another multivalued semiflow generated by selector solutions, i.e.,

$$\mathscr{G}(t, \varphi_0) = \left\{ \varphi(t) \;\middle|\; \begin{array}{l} \varphi(\cdot) \quad \text{is a solution of (16.26)}, \\ \varphi(0) = \varphi_0, \\ \exists g \in \mathrm{G}: \ \varphi(\cdot) \text{ is a solution of the resp. equation with } g \end{array} \right\}.$$

However in this case, for the sequence $\{\varphi_j\}_{j=1}^\infty$, we have $\{g_j\}_{j=1}^\infty$, $g_j(u) \in G(u)$, $\forall u \in \mathbb{R}$. In order to $g_j(u) \to g(u)$ $\forall u \in \mathbb{R}, g \in G$, it is necessary to strengthen the conditions for G. But in this case, the question about solvability of Problem (16.26) arises.

Acknowledgments This work was partially supported by the Ukrainian State Fund for Fundamental Researches under grants GP/F44/076, GP/F49/070, and by the NAS of Ukraine under grant 2273/13.

References

1. Ball, J.M.: Continuity properties and global attractors of generalized semiflows and the Navier-Stokes equations. J. Nonlinear Sci. **7**(5), 475–502 (1997)
2. Ball, J.M.: Global attractors for damped semilinear wave equations. DCDS. **10**, 31–52 (2004)
3. Clarke, F.H.: Optimization and Nonsmooth Analysis. Wiley, New York (1983)
4. Dubinskii, Yu.A: High order nonlinear parabolic equations. J. Math. Sci. **56**(4), 2557–2607 (1991)
5. Gajewski, H., Gröger, K., Zacharias, K.: Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen. Akademie, Berlin (1974)
6. Kapustyan, O.V., Valero, J.: Comparison between trajectory and global attractors for evolution systems without uniqueness of solutions. Int. J. Bifurcat. Chaos **20**(9), 2723–2734 (2010)
7. Kasyanov, P.O.: Multivalued dynamic of solutions of autonomous differential-operator inclusion with pseudomonotone nonlinearity. Cybern. Syst. Anal. **47**(5), 800–811 (2011)
8. Kasyanov, P.O., Mel'nik, V.S., Toscano, S.: Solutions of Cauchy and periodic problems for evolution inclusions with multi-valued w_{λ_0}-pseudomonotone maps. J. Diff. Equat. **249**(6), 1258–1287 (2010)
9. Kasyanov, P.O., Toscano, L., Zadoianchuk, N.V.: Long-time behavior of solutions for autonomous evolution hemivariational inequality with multidimensional "reaction-displacement" law. Abstr. Appl. Anal. (2012). doi:10.1155/2012/450984
10. Mel'nik, V.S., Valero, J.: On attractors of multivalued semi-flows and differential inclusions. Set-Valued Anal. **6**(1), 83–111 (1998)
11. Sell, G.R., You, Yu.: Dynamics of Evolutionary Equations. Springer, New York (2002)
12. Temam, R.: Infinite-Dimensional Dynamical Systems in Mechanics and Physics. Springer, New York (1988)
13. Valero, J., Kapustyan, A.V.: On the connectedness and asymptotic behaviour of solutions of reactiondiffusion systems. J. Math. Anal. Appl. **323**(1), 614–633 (2006)
14. Vishik, M., Chepyzhov, V.V.: Trajectory and global attractors of three-dimensional Navier-Stokes systems. Math. Notes **71**(1–2), 177–193 (2002)
15. Zadoyanchuk, N.V., Kas'yanov, P.O.: FaedoGalerkin method for second-order evolution inclusions with W_λ-pseudomonotone mappings. Ukrainian Math. J. **61**(2), 236–258 (2009)
16. Zadoyanchuk, N.V., Kasyanov, P.O.: Analysis and control of second-order differential-operator inclusions with +-coercive damping. Cybern. Syst. Anal. **46**(2), 305–313 (2010)
17. Zgurovsky, M.Z., Mel'nik, V.S., Kasyanov, P.O.: Evolution Inclusions and Variation Inequalities for Earth Data Processing I. Springer, New York (2010). doi:10.1007/978-3-642-13837-9
18. Zgurovsky, M.Z., Mel'nik, V.S., Kasyanov, P.O.: Evolution Inclusions and Variation Inequalities for Earth Data Processing II. Springer, New York (2010). doi:10.1007/978-3-642-13878-2
19. Zgurovsky, M.Z., Kasyanov, P.O., Kapustyan, O.V., Valero, J., Zadoianchuk, N.V.: Evolution Inclusions and Variation Inequalities for Earth Data Processing III. Springer, New York (2012). doi:10.1007/978-3-642-28512-7

Part IV
Control Theory and Decision Making

Chapter 17
On the Regularities of Mass Random Phenomena

Victor I. Ivanenko and Valery A. Labkovsky

Abstract This note presents a not very well known result concerning the frequentist origins of probability. This result provides a positive answer to the question of existence of statistical regularities of so called *random in a broad sense* mass phenomena, using the terminology of A. N. Kolmogorov [20]. It turns out, that some closed in weak-$*$ topology family of finitely-additive probabilities plays the role of the statistical regularity of any such phenomenon. If the mass phenomenon is stochastic, then this family degenerates into a usual countably-additive probability measure. The note provides precise definitions, the formulation and the proof of the theorem of existence of statistical regularities, as well as the examples of their application.

17.1 Introduction

This presentation conveys the main result of our study of the regularities of mass random phenomena (MRP). This study started in the 60s of the twentieth century and was stimulated, first of all, by the necessity to stabilize a generator of random processes [11], extremely sensible to influence of external factors and, hence, statistically unstable. In the 70s, Valery A. Labkovsky, who graduated from the Mechanics and Mathematics Department of the Moscow State University and was recommended to me by A. M Yaglom, joined our team. In a way, my participation in this seminar has a supplementary historical justification. Unfortunately, V. A. Labkovsky got ill and died at the end of the very difficult 90s. What I am about to present is the result of our joint work with Valery Labkovsky.

Valery A. Labkovsky—deceased.

V. I. Ivanenko (✉) · V. A. Labkovsky
Kyiv Polytechnical Institute, Prospekt Peremogy, 37, Kyiv 03056, Ukraine
e-mail: viktorivanenko@gmail.com

Mechanics and Mathematics Department, Moscow State University, Moscow, Russia

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_17,

In the 60s of the twentieth century, there were a lot of reasons to study the MRP, beside the aforementioned technical problem. So in [2, 17] the authors pointed out the difficulties arising in the process of modeling of the social MRP. In particular, in [2] one reads: "Some contemporary theoreticians considered the law of large numbers as a simple tautology, since they thought that the probability could be defined as frequency for a very large number of trials. If for a very big number of trials this frequency does not tend to a limit, but fluctuates more or less between different limits, one needs to affirm that probability p does not remain constant and changes in the process of trials. This concerns, for example, human mortality rate in the course of centuries, since the progress of medicine and hygiene leads to the increase of life duration." The problem of establishing of the regularities of MRP becomes more and more important, especially in relation to the instability of financial markets and other economic objects [21, 23, 26], that makes forecasting in this area very unreliable.

It is relevant to mention here the following remark by A.N. Kolmogorov [20]: "Speaking of randomness in the ordinary sense of this word, we mean those phenomena in which we do not find regularities allowing us to predict their behavior. Generally speaking, there are no reasons to assume that random in this sense phenomena are subject to some probabilistic laws. Hence, it is necessary to distinguish between randomness in this *broad sense* and *stochastic* randomness (which is the subject of probability theory)".

However, what do the words "do not find regularities allowing us to predict their behavior" mean? Hardly these words should be understood in the sense that such regularities do not exist at all. More likely, these words point out to the problem of finding of the statistical regularities of *random in a broad sens mass phenomena* (RBSMP), that is the regularities of asymptotic behavior of different average values that characterize these phenomena. For instance, it can be frequencies of hitting in given subsets, arithmetic averages of some functionals, and so on. Recall that MRP are called *statistically stable* or *stochastic*, if with the increase of the number of "trials" all these averages tend to limits (and if some other conditions are verified as well, see details in [20]). Unlike this, it is natural to consider as RBSMP those MRP, whose behavior is studied to within statistical regularities. In other words, this definition combines in RBSMP stochastic as well as *nonstochastic*[1] random phenomena.

In [19] the question was risen whether the MRP (or, as we say now, RBSMP) posses the properties that are necessary in order to apply the probability theory to their description. It turns out that the answer to this question is positive, but, as we shall see later, under some complementary conditions.

There are several approaches to modeling of the MRP. So, there is the algorithmic approach to randomness [27, 28] as well as the game-theoretic approach to randomness in finance [24]. An alternative approach was studied in [12], where sequences were constructed only with the requirement of the so called Γ_1 indepen-

[1] Remark that the term "nonstochastic" appeared in [27] in the context of Kolmogorov's complexity, meaning "more complex than stochastic". In this chapter the meaning of this term is "more random than stochastic".

dence. In mathematical finance diverse extensions of stochastic models have been popular [1, 3, 6].

At the same time, when modelling the possible properties of the RBSMP it is natural to consider families of probability distributions. In scientific literature, the families of probability distributions appear more and more often. So, they were considered in game theory [25] in order to study non-additive set functions. In the so called subjective decision theory these families appear as consequence of the axioms of rational choice [5, 8, 13], where, similarily to robust statistics [7], they were interpreted as families of a priori distributions.

It turned out that specifically families of probability distributions are necessary for the description of statistical (frequentist) regularities of a rather wide class of RBSMP. The theorem of existence of such statistical regularities was proven and published a quarter of a century ago [14, 15].

Let me pass to precise formulations.

17.2 Theorem of Existence of Statistical Regularities

An ordinary sequence is the simplest mathematical model of a mass phenomenon. In order to construct, on the basis of a sequence, a model of a random phenomenon, it is necessary to identify sequences that have identical statistical properties.

Definition 17.1 *Let X be an arbitrary set. Two sequences $\overline{x}^{(1)}$ and $\overline{x}^{(2)}$ of elements of the set X are called statistically equivalent (S-equivalent) if and only if for any natural number m and any bounded mapping $\gamma \in (X \to \mathbb{R}^m)$ the set of limit points of the sequence*

$$\left\{\overline{y}_n^{(k)};\ n \in \mathbb{N}\right\},\ y_n^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \gamma\left(\overline{x}_i^{(k)}\right)$$

does not depend on $k \in \{1,\ 2\}$.

The class of S-equivalence of the sequence $\overline{x} \in X^{\mathbb{N}}$ will be denoted as $S(\overline{x})$. Our nearest goal is to find the invariant of the relation of S-equivalence. Introduce several notions.

Let M be a Banach space of bounded real functions, defined on the set X, M^* be the dual space of the space M, and τ—is a weak-$*$ topology in M^*. Let, further, $PF(X)$ be the subspace of the topological space $(M^*,\ \tau)$ defined by the formula

$$PF(X) = \left\{p \in M^* : p(f) \geq 0 \text{ if } f \geq 0,\ p(\mathbf{1}_X) = 1\right\},$$

where $\mathbf{1}_A(\cdot)$ is the characteristic function of the set A.

In what follows, instead of $p(\mathbf{1}_A)$ we shall often write $p(A)$ identifying, by the same token, the elements of the set $PF(X)$ with the finitely additive and normed mea-

sures on 2^X. Obviously, $p(f)$ in this case is simply the integral $p(f) = \int f(x)p(dx)$, defined naturally due to boundedness of function f.

Associate to an arbitrary sequence $\overline{x} = \{\overline{x}_n;\ n \in \mathbb{N}\} \in X^{\mathbb{N}}$ the sequence of measures from $PF(X)$ defined as

$$\left\{\overline{p}_{\overline{x}}^{(n)}(\cdot);\ n \in \mathbb{N}\right\},\ \overline{p}_{\overline{x}}^{(n)}(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_A(\overline{x}_i),\ \forall A \subseteq X.$$

Due to compactness of the set $PF(X)$ (as of a bounded closed set in (M^*, τ)), the sequence $\left\{\overline{p}_{\overline{x}}^{(n)}(\cdot);\ n \in \mathbb{N}\right\}$ will have a non-empty closed set of limit points, which we denote as $P_{\overline{x}}$ and call *the regularity* of this sequence. Therefore introduce the following definition.

Definition 17.2 *Any non-empty closed subset of the space $(PF(X), \tau)$ is called* ***a regularity*** *on X. Denote the set of all regularities on X as $\mathbb{P}(X)$ and associate to any sequence $\overline{x} \in X^{\mathbb{N}}$ its regularity $P_{\overline{x}}$. Finally, for $m \in \mathbb{N}$, $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_m) \in (X \to \mathbb{R}^m)$ and $P \in \mathbb{P}(X)$, the symbol $P(\gamma)$ denotes the set*

$$\left\{(r_1, r_2, \ldots, r_m) \in \mathbb{R}^m : \exists p \in P \forall i \in \overline{1, m},\ r_i = p(\gamma_i)\right\},$$

and, in particular, $p(\gamma) = (p(\gamma_1),\ p(\gamma_2),\ \ldots,\ p(\gamma_m))$ for $p \in PF(X)$.

Consider the following proposition.

Proposition 17.1 *The mapping $\overline{x} \mapsto P_{\overline{x}}$ is the invariant of the relation of S-equivalence on $X^{\mathbb{N}}$.*

This statement will be proved below in a more general form. So far, however, let us agree to call the classes of S-equivalence of sequences *the simplest random phenomena*, and their regularities—*statistical regularities* of the corresponding phenomena. Any sequence $\overline{x} \in X^{\mathbb{N}}$ is considered as a realization of a simplest random phenomenon $S(\overline{x})$.

Connection of the notions introduced above with the probabilistic notions follows directly from the enforced law of large numbers.

Proposition 17.2 *Let X be a finite set, μ—a probability distribution on X, and $\overline{\xi} = \left\{\overline{\xi}_n; n \in \mathbb{N}\right\}$—a sequence of independent (in the usual sense) random elements, taking values in X with distribution μ. Then with probability* 1 *the sequence $\overline{x}$ of the values of the sequence $\overline{\xi}$ will be a realization of the simplest random phenomenon with statistical regularity $P_{\overline{x}} = \{\mu\}$, i.e. consisting of the single distribution μ.*

However, when the set X is infinite everything becomes considerably more difficult. In this case, the capabilities of sequences, generally speaking, are insufficient in order to guarantee that the frequencies of hitting in all measurable sets would tend to their limits simultaneously. Moreover, it is easy to see that the regularities of sequences, since they are concentrated only on a countable subset of the set X, constitute only a small part of the set of all regularities on X. This seems to reflect

the fact that sequences constitute only a small part of all mass phenomena. A more general notion of *sampling net* is, as we shall see further, already sufficient for our goals.

Definition 17.3 *A **sampling net** (s.n.) in X is any net $\varphi = \{\varphi_\lambda,\ \lambda \in \Lambda,\ \geq\}$ taking values in the sampling space*

$$X^\infty = \bigcup_{n=1}^{\infty} X^n, \quad X^n = \underbrace{X \times \cdots \times X}_{n}.$$

Moreover, if $\lambda \in \Lambda, \varphi_\lambda \in X^n$ then we denote $n = n_\lambda$, $\varphi_\lambda = (\varphi_{\lambda 1}, \varphi_{\lambda 2}, \ldots, \varphi_{\lambda n_\lambda})$ and associate to this λ the measure $p_\varphi^{(\lambda)} \in PF(X)$ defined as

$$p_\varphi^{(\lambda)}(A) = \frac{1}{n_\lambda} \sum_{i=1}^{n_\lambda} \mathbf{1}_A(\varphi_{\lambda i}), \quad A \subseteq X.$$

*The set P_φ of limit points of the net $p_\varphi = \left\{p_\varphi^\lambda,\ \lambda \in \Lambda,\ \geq\right\}$ will be called **the regularity** of the s.n. φ. The class of all s.n. in X will be denoted as $\Phi(X)$.*

Extend now the relation of S-equivalence on the whole $\Phi(X)$.

Definition 17.4 *Sampling nets $\varphi^{(k)} \in \Phi(X), k = 1, 2$ are considered as S-equivalent if and only if for any $m \in \mathbb{N}$ and any bounded mapping $\gamma \in (X \to \mathbb{R}^m)$ the set of limit points of the net of averages*

$$\left\{y_\lambda^{(k)}, \lambda \in \Lambda, \geq\right\}, \quad y_\lambda^{(k)} = \frac{1}{n_\lambda} \sum_{i=1}^{n_\lambda} \gamma\left(\varphi_{\lambda i}^{(k)}\right) \tag{17.1}$$

does not depend on $k \in \{1, 2\}$.

We can now formulate the main theorem in the following way.

Theorem 17.1 *(i) For any s.n. $\varphi \in \Phi(X)$, any $m \in \mathbb{N}$ and any bounded mapping $\gamma \in (X \to \mathbb{R}^m)$, the set of limit points of the net (17.1) can be written as $P_\varphi(\gamma)$.*

(ii) The mapping $\varphi \mapsto P_\varphi$, defined on $\Phi(X)$, is the invariant of the relation of S-equivalence.

(iii) This mapping is a mapping on the whole set $\mathbb{P}(\mathbb{X})$, i.e. the set $\Phi(X)/S$ of classes of S-equivalence and the set $\mathbb{P}(X)$ of regularities are put by this mapping into one-to-one correspondence.

This theorem justifies the following definition.

Definition 17.5 *Any class of S-equivalence of sampling nets in X is called random in a broad sense mass phenomenon in X. The regularity P_φ is called the statistical regularity of the random phenomenon $S(\varphi)$. Any s.n. $\varphi' \in S(\varphi)$ is called a*

realization of the random phenomenon $S(\varphi)$. The random phenomenon, having statistical regularity P, is called μ-stochastic if and only if there exists a non-trivial σ-algebra $\mathscr{A} \subseteq 2^X$, on which μ is a σ-additive probability, and $p(A) = \mu(A)$ for all $p \in P$, $A \in \mathscr{A}$.

17.3 The Proof

Denote the set of limit points of an arbitrary net $g = \{g_\alpha,\ \alpha \in A,\ \succcurlyeq\}$ with values in X as $LIM(g)$ or $LIM\{g_\alpha,\ \alpha \in A,\ \succcurlyeq\}$. Denote the set of bounded mappings from X into $\mathbb{R}^m$ as M^m. We need to establish the three following facts:

(i) The relation $LIM\{y_\lambda,\ \lambda \in \Lambda,\ \succcurlyeq\} = P_\varphi(\gamma)$ is true for all $m \in \mathbb{N}$, $\gamma \in M^m$, $\varphi \in \Phi(X)$.
(ii) If $P_1,\ P_2 \in \mathbb{P}(X),\ P_1 \neq P_2$, then there exist such $m \in \mathbb{N}$ and such $\gamma \in M^m$, that $P_1(\gamma) \neq P_2(\gamma)$.
(iii) For any regularity $P \in \mathbb{P}(X)$ there exist such s.d. $\varphi \in \Phi(X)$, that $P = P_\varphi$.

Begin with the proof of the proposition (i). Let $r \in LIM(y)$, where $y = \{y_\lambda,\ \lambda \in \Lambda,\ \succcurlyeq\}$. Then there exists a subnet of the net y converging to r, i.e. there exists (see [18]) a directed set $(A, \succcurlyeq)$ and a function $f : A \to \Lambda$ such that the net $\overline{y} = y \circ f$ converges to r, and, in addition, for any $\lambda \in \Lambda$ there exists such $\alpha_1 \in A$ that $f(\alpha) \succcurlyeq \lambda$ for all $\alpha \succcurlyeq \alpha_1$.

Consider now the net of measures $\overline{p}_\varphi = p_\varphi \circ f$, where $p_\varphi = \left\{p_\varphi^{(\lambda)}, \lambda \in \Lambda, \succcurlyeq\right\}$. By virtue of compactness of the space $(PF(X), \tau)$ this has at least one limit point. Denote it as p_0 and consider a subnet $\overline{\overline{p}}_\varphi$ of the net $\overline{p}_\varphi$, converging to p_0. Let it be $\overline{\overline{p}}_\varphi = \overline{p}_\varphi \circ g = p_\varphi \circ f \circ g$, $g : B \to A$. Then the net $\overline{\overline{y}} = y \circ f \circ g$, on the one hand, converges to r, and, on the other hand, $\overline{\overline{y}}_\beta = \overline{\overline{p}}_\varphi^{(\beta)}(\gamma)$, $\beta \in B$, so that

$$r = \lim_\beta \overline{\overline{p}}_\varphi^{(\beta)}(\gamma) = p_0(\gamma) \in P_\varphi(\gamma).$$

By the same token, it is proved that $LIM(y) \subseteq P_\varphi(\gamma)$.

Conversely, if $p_0 \in P_\varphi$, $r = p_0(\gamma)$, then there exists a subnet $\widetilde{p}_\varphi = \{\widetilde{p}_\varphi^\alpha, \alpha \in A, \succcurlyeq\}$ of the net p_φ, converging to p_0. But in this case $\lim_\alpha \widetilde{p}_\varphi^{(\alpha)}(\gamma_i) = p_0(\gamma_i)$ for all $i \in \overline{1, m}$. It means that $\lim_\alpha \widetilde{p}_\varphi^{(\alpha)}(\gamma) = p_0(\gamma)$. And, since $\widetilde{p}_\varphi^{(\alpha)}(\gamma) = y_\lambda$ for $\lambda = f(\alpha)$, this proves (i).

In order to prove (ii) assume that there exists $p_1 \in P_1 \setminus P_2$. Since the set P_2 is closed, there exists a vicinity of the point p_1 that does not cross with P_2 and it means that there exist such $\varepsilon > 0$, $\gamma_1, \gamma_2, \ldots, \gamma_m \in M$ that

$$\forall p_2 \in P_2, \exists i \in \overline{1, m}, |p_1(\gamma_i) - p_2(\gamma_i)| > \varepsilon.$$

So that if $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_m)$, then $p_1(\gamma) \notin P_2(\gamma)$.

The complete proof of (iii) can be found in [10, 15, 16]. Here we shall outline the main ideas of the proof. Let Q be the set of all such measures $q \in PF(X)$ that each one of them is concentrated on a finite set $X_q \subseteq X$, and in addition all numbers $q(x), x \in X_q$ are rational. One can show that the set Q is everywhere dense in $(PF(X), \tau)$.

Now, to an arbitrary regularity $P \in \mathbb{P}(X)$ we put into correspondence the directed set $(\Lambda, \succcurlyeq)$ such that

$$\Lambda = \mathbb{R}^+ \times M^\infty \times P, \quad \mathbb{R}^+ = [0, \infty], \quad M^\infty = \bigcup_{m=1}^{\infty} M^m,$$

and the relation $(\succcurlyeq)$ is given by the formula

$$(\varepsilon_1, \gamma_{11}, \gamma_{12}, \ldots, \gamma_{1n_1}, p_1) \succcurlyeq (\varepsilon_2, \gamma_{21}, \gamma_{22}, \ldots, \gamma_{2n_2}, p_2) \Leftrightarrow$$
$$(\varepsilon_1 \leq \varepsilon_2, \{\gamma_{11}, \gamma_{12}, \ldots, \gamma_{1n_1}\} \supseteq \{\gamma_{21}, \gamma_{22}, \ldots, \gamma_{2n_2}\}),$$

where no condition is imposed on p_1 and p_2.

Finally, to any $\lambda = (\varepsilon, \gamma_1, \gamma_2, \ldots, \gamma_m, p) \in \Lambda$ we put into correspondence some

$$q_\lambda \in Q \bigcap \left\{ p' \in PF(X) : \forall i \in \overline{1, m}, \left| p(\gamma_i) - p'(\gamma_i) \right| < \varepsilon \right\}.$$

It is proven further that with any $\lambda \in \Lambda$ one can associate simultaneously a sequence of points $x_1^{(\lambda)}, x_2^{(\lambda)}, \ldots, x_{n_\lambda}^{(\lambda)} \in X_q$ satisfying the condition

$$q_\lambda(A) = \frac{1}{n_\lambda} \sum_{i=1}^{n_\lambda} \mathbf{1}_A(x_i^{(\lambda)}), \quad \forall A \subseteq X.$$

It remains to chose $\varphi_\lambda = (x_1^{(\lambda)}, x_2^{(\lambda)}, \ldots, x_{n_\lambda}^{(\lambda)})$ and we obtain a s.n. $\varphi : \lambda \mapsto \varphi_\lambda$ that has the regularity $P_\varphi = P$.

17.4 Applications in Decision Theory

Statistical regularities of the general form find their application in decision theory [5, 8, 10, 13, 15, 16, 22] and its applications [9].

Considering decision problems, assume that we need to make a decision u from the set U of possible decisions, knowing that the result of making a decision depends on some uncontrolled parameter θ from the set Θ of possible values of this parameter and is described by the bounded real loss function $L : \Theta \times U \to \mathbb{R}$. If nothing is known about the behavior of the parameter, then we cannot, strictly speaking, exclude that scenario, where the value of θ is chosen in the worst possible for us way. In this case, the quality of decision u is evaluated by means of the loss function

$$L_1^*(u) = \sup_{\theta \in \Theta} L(\theta,\, u), \quad u \in U,$$

a so called "minmax" criterion.

If it is known, that parameter θ is stochastic with the given distribution μ, then, trying to minimize the average losses, one makes use of the Bayes criterion

$$L_2^*(u) = \int L(\theta,\, u)\mu(d\theta), \quad u \in U.$$

Suppose now that parameter θ is random in a broad sense with the statistical regularity $P \in \mathbb{P}(\Theta)$. Let us show that in this case it is natural to chose the criterion in the form of

$$L_3^*(u) = \sup_{p \in P} \int L(\theta,\, u)p(d\theta), \quad u \in U, \tag{17.2}$$

Indeed, let $r_1 < L_3^*(u) < r_2$. The following statement is straightforward.

Proposition 17.3 *Let $\{\varphi_\lambda, \lambda \in \Lambda, \succcurlyeq\}$—be a sampling net in Θ with the regularity P. Then for any $\lambda_1 \in \Lambda$ there is such $\lambda \succcurlyeq \lambda_1$ that*

$$\frac{1}{n_\lambda} \sum_{i=1}^{n_\lambda} L(\varphi_{\lambda i},\, u) > r_1$$

and, at the same time, there is such λ_2, that for all $\lambda \succcurlyeq \lambda_2$ there will be

$$\frac{1}{n_\lambda} \sum_{i=1}^{n_\lambda} L(\varphi_{\lambda i},\, u) < r_2.$$

In other words, $L_3^*(u)$—is that natural border, that separates the average losses, that can happen for a given u for an arbitrary "large" λ, from those average losses that are not "dangerous" to us, when λ is sufficiently "large".

It is easy to see that $L_3^*(u)$ becomes $L_1^*(u)$, when $P = PF(\Theta)$ (strictly nothing is known about θ, save the set Θ where it takes values), and that it becomes $L_2^*(u)$, when $P = \mu$ is stochastic regularity and function $L(\theta, u)$ is measurable relatively to the corresponding σ-algebra.

The inverse result appears as somewhat surprising. It turns out that if one subordinates a criterion choice rule to some natural conditions of consistency with the triplet $(\Theta,\, U,\, L)$, then any rule, satisfying these conditions, leads to the criterion of the form (17.2), where P—is some (not known beforehand) regularity on Θ. In particular, this result justifies the heuristic definition of random in a road sense phenomena introduced above. Therefore, one can conclude that regularity on Θ is, in a certain sense, the most general form of information about the behavior of θ. One can find details in [8, 10, 13, 15, 16].

17.5 Concluding Remarks

In conclusion, let me note that statistical regularities in the form of families of probability distributions attract all the more attention. In particular, in [4][2] one already finds : "We present methods... to estimate the model from finite time series data. The estimation of the set of probability measures is based on the analysis of a set of relative frequencies of events taken along subsequences selected by a collection of rules. In particular, we provide a universal methodology for finding a family of subsequence selection rules that can estimate any set of probability measures with high probability."

In view of the informal character of our seminar, let me note that I have been surprised by reluctance of Westren researchers, who work in similar areas, to make themselves familiar either with the English translations or with the original journal publications of the Soviet scientific school.

References

1. Avellaneda, M., Levy, A., Paras, A.: Pricing and hedging derivative securities in markets with uncertain volatilities. Appl. Math. Financ. **2**, 73–88 (1995)
2. Borel, E.: Probabilité et Cértitude. Presse Universitaire de France, Paris (1956)
3. Calvet, L.E., Fisher, A.J., Thompson, S.B.: Volatility comovement: a multifrequency approach. J. Econometrics. **131**, 179–215 (2006)
4. Fierens, P., Rego, L., Fine, T.: A frequentist understanding of sets of measures. J. Stat. Plan. Infer. **139**, 1879–1892 (2009)
5. Gilboa, I., Schmeidler, D.: Maxmin expected utility with nonunique prior. J. Math. Econ. **18**, 141–153 (1989)
6. Heston, S.L.: A closed-form solution for options with stochastic volatility with applications to bond and currency Options. Rev. Financ. Stud. **6**(2), 327–343 (1993)
7. Huber, P.J.: Robust Statistics. Wiley, New York (1981)
8. Ivanenko, V.I., Labkovskii, V.A.: On the functional dependence between the available information and the chosen optimality principle. Proceedings of the International conference on Stochastic Optimisation. In: Lecture Notes in Control and Information SciencesKiev, pp. 388–392. Springer-Verlag, Berlin (1986)
9. Ivanenko, Y., Munier, B.: Price as a choice under nonstochastic randomness in finance. Risk and Decision Analysis (2012) (forthcoming)
10. Ivanenko, V.I.: Decision systems and nonstochastic randomness. Springer, Dordrecht (2010)
11. Ivanenko, V.I., Khokhel, O.A.: Problems of stabilization of the parameters of artificially generated random processes. Avtomatika i telemehanika. **6**, 32–41 (1968)
12. Ivanenko, V.I., Labkovskii, V.A.: On one kind of uncertainty. Sov. Phys. Dokl. **24**(9), 705–706 (1979)
13. Ivanenko, V.I., Labkovskii, V.A.: A class of criterion-choosing rules. Sov. Phys. Dokl. **31**(3), 204–205 (1986)
14. Ivanenko, V.I., Labkovskii, V.A.: A model of non-stochastic randomness. Sov. Phys. Dokl. **35**(2), 113–114 (1990)

[2] I am thankful to professor Vladimir Vovk who made me familiar with the works of professor Terrence Fine and, in particular, with this chapter.

15. Ivanenko, V.I., Labkovsky, V.A.: Uncertainty problem in decision making [in Russian]. Naukova Dumka, Kyiv (1990)
16. Ivanenko, V.I., Munier, B.: Decision Making in "Random in a Broad Sense" Environments. Theor. decis. **49**(2), 127–150 (2000)
17. Jarvik, M.E.: Probability learning and a negative recency effect in the serial anticipation of alternative symbols. J. Exp. Psychol. **41**, 291–297 (1951)
18. Kelley, J.L.: General Topology. D. Van Nostrand Company. Inc. Princeton, New Jersey (1957)
19. Khinchin, A.Y.: The frequentist theory of Richard von Mises and contemporary ideas in probability theory. I. Vop. philosophii. **1**, 91–102 (1961)
20. Kolmogorov, A.N.: On the logical foundation of probability theory. Probability Theory and Mathematical Statistics, pp. 467–471. Nauka, Moscow (1986)
21. Mandelbrot, B., Hudson, R.: The (mis) behavior of markets. Basic Books, New York (2006)
22. Mikhalevich, V.M.: Parametric decision problems with financial losses. Cybern. Syst. Anal. **47**(2), 286–295 (2011)
23. Munier, B.: Global Uncertainty and the Volatility of Agricultural Commodity Prices. IOS Press, Amsterdam (2012)
24. Shafer, G., Vovk, V.: Probability and Finance: It's Only a Game!. Wiley, New York (2001)
25. Shapley, S.: Notes on n-person games. Chap. VII. Cores of Convex Games. RAND Corp, Santa Monica (1955)
26. Taleb, N.N.: Fooled by Randomness. W. W. Norton, New York (2001)
27. Vyugin, V.V.: On nonstochastic objects. Prob. Inf. Transm. **21**(2), 3–9 (1985)
28. Zvonkin, A.K., Levin, L.A.: Complexity of finite objects and justification of notions of information and and randomness by means of the theory of algorithms. Uspehi Matematicheskih Nauk. **25**(6), 85–127 (1970)

Chapter 18
Optimality Conditions for Partially Observable Markov Decision Processes

Eugene A. Feinberg, Pavlo O. Kasyanov and Mikhail Z. Zgurovsky

Abstract This note describes sufficient conditions for the existence of optimal policies for Partially Observable Markov Decision Processes (POMDPs). The objective criterion is either minimization of total discounted costs or minimization of total nonnegative costs. It is well-known that a POMDP can be reduced to a Completely Observable Markov Decision Process (COMDP) with the state space being the sets of believe probabilities for the POMDP. Thus, a policy is optimal in POMDP if and only if it corresponds to an optimal policy in the COMDP. Here we provide sufficient conditions for the existence of optimal policies for COMDP and therefore for POMDP.

18.1 Introduction

Partially Observable Markov Decision Processes (POMDPs) play an important role in electrical engineering, computer science, and operations research. They have a broad range of applications including sensor networks, artificial intelligence, control and maintenance of complex systems, and medical decision making. In principle, by ignoring complexity issues, it is known how to solve POMDPs. A POMDP can be reduced to a Completely Observable Markov Decision Process (COMDP) with the state space being the sets of believe probabilities for the POMDP [2, 6, 9, 10]. After an optimal policy for the COMDP is found, it can be used to compute an optimal

E. A. Feinberg (✉)
Department of Applied Mathematics and Statistics,
Stony Brook University, Stony Brook, NY 11794-3600, USA
e-mail: eugene.feinberg@sunysb.edu

P. O. Kasyanov (✉) · M. Z. Zgurovsky
Institute for Applied System Analysis, National Technical University of Ukraine
"Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35, Kyiv 03056, Ukraine
e-mail: kasyanov@i.ua

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*,
Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_18,

policy for the POMDP. However, except the case of problems with finite state and action sets and a large variety of particular problems considered in the literature, little is known regarding the existence and properties of optimal policies for COMDPs in terms of the original POMDP.

This problem is studied in Hernández-Lerma [6, Chap. 4], where sufficient conditions for the existence of optimal policies for discounted POMDPs with Borel state spaces, compact action sets, weakly continuous transition and observation probabilities, and bounded continuous cost functions are provided. It is shown there that the weak continuity of the transition kernel in the filtration equation is sufficient for the existence of optimal policies for COMDPs and therefore for COMDPs. A sufficient condition for the case of a countable observation case is also provided in Hernández-Lerma [6, p. 92]. This condition is that the probability of observations depend continuously on the state-action pairs. Since this is the condition for a countable observation space, in the case of general Borel observation spaces, there are three continuity conditions on the observation probabilities that are equivalent to this condition, when the observation space becomes countable. These conditions are weak continuity, setwise continuity, and continuity in the total variation of observation probabilities (also called kernels or stochastic kernel).

In this paper, we a we study either minimization of expected total nonnegative costs or discounted costs with the one-step cost functions bounded below for POMDPs with Borel state spaces. The goal is to obtain sufficient conditions for the existence and characterization of optimal policies for COMDPs with possibly non-compact action sets, unbounded cost functions (they are assumed bounded below), and uncountable observation sets. The one-step cost functions are K-infcompact. The notion of K-infcompactness was introduced recently in Feinberg, Kasyanov, and Zadoianchuk [3]. As shown in Feinberg, Kasyanov, and Zadoianchuk [4], this mild condition and weak continuity of transition probabilities are sufficient for the existence of optimal policies and their characterization for fully observable Markov Decision Processes (MDPs) with the expected total costs.

Of course, for the existence of optimal policies for a POMDP, additional conditions are required for the transition observation probability. Here we show that the sufficient condition is its continuity in the total variation of the observation transition probability. We also provide a general criterion for the existence optimal policies for weakly continuous transition observation probabilities, which is different from the weak continuity of the filtration kernel considered in Hernández-Lerma [6, p. 90, Assumption 4.1(d)].

18.2 Model Description

For a metric space $\mathbb{S}$, let $\mathscr{B}(\mathbb{S})$ be its Borel σ-field, that is, the σ-field generated by all open sets of the metric space $\mathbb{S}$. For a Borel subset $E \subset \mathbb{S}$, we denote by $\mathscr{B}(E)$ the σ-field whose elements are intersections of E with elements of $\mathscr{B}(\mathbb{S})$. Observe that E is a metric space with the same metric as on $\mathbb{S}$, and $\mathscr{B}(E)$ is its Borel σ-field.

The space E is a *Borel space*, if E is a Borel subset of a Polish (complete separable metric) space $\mathbb{S}$. On E consider the induced metrizable topology. For a metric space $\mathbb{S}$, we denote by $\mathbb{P}(\mathbb{S})$ the *set of probability measures* on $(\mathbb{S}, \mathscr{B}(\mathbb{S}))$. A sequence of probability measures $\{\mu_n\}$ from $\mathbb{P}(\mathbb{S})$ *converges weakly (setwise)* to $\mu \in \mathbb{P}(\mathbb{S})$ if for any bounded continuous (bounded Borel-measurable) function f on $\mathbb{S}$

$$\int_{\mathbb{S}} f(s)\mu_n(ds) \to \int_{\mathbb{S}} f(s)\mu(ds) \qquad \text{as} \quad n \to \infty.$$

A sequence of probability measures $\{\mu_n\}$ from $\mathbb{P}(\mathbb{S})$ *converges in the total variation* to $\mu \in \mathbb{P}(\mathbb{S})$ if

$$\sup_{f \in \mathbb{F}_1(\mathbb{S})} \left\{ \int_{\mathbb{S}} f(s)\mu_n(ds) - \int_{\mathbb{S}} f(s)\mu(ds) \right\},$$

where $F_1(\mathbb{S})$ is the set of Borel-measurable functions on $\mathbb{S}$ such that $|f(s)| \le 1$ for all $s \in \mathbb{S}$.

Note that $\mathbb{P}(\mathbb{S})$ is a Polish space with respect to the weak convergence topology for probability measures; Parthasarathy [8, Chap. 2]. For Borel spaces $\mathbb{S}_1$ and $\mathbb{S}_2$, a (Borel-measurable) *transition kernel* $R(ds_1|s_2)$ on $\mathbb{S}_1$ given $\mathbb{S}_2$ is a mapping $R(\cdot\,|\,\cdot) : \mathscr{B}(\mathbb{S}_1) \times \mathbb{S}_2 \to [0, 1]$, such that $R(\cdot\,|s_2)$ is a probability measure on $\mathbb{S}_1$ for any $s_2 \in \mathbb{S}_2$, and $R(B|\cdot)$ is a Borel-measurable function on $\mathbb{S}_2$ for any Borel set $B \in \mathscr{B}(\mathbb{S}_1)$. A transition kernel $R(ds_1|s_2)$ on $\mathbb{S}_1$ given $\mathbb{S}_2$ defines a Borel measurable mapping $s_2 \to R(\cdot|s_1)$ of $\mathbb{S}_2$ to the metric space $\mathbb{P}(\mathbb{S}_1)$ endowed with the topology of weak convergence. A transition kernel $R(ds_1|s_2)$ on $\mathbb{S}_1$ given $\mathbb{S}_2$ is called *weakly continuous (setwise continuous, continuous in the total variation)*, if $R(\cdot\,|x_n)$ converges weakly (setwise, in the total variation) to $R(\cdot\,|x)$ whenever x_n converges to x in $\mathbb{S}_2$.

Let $\mathbb{X}$, $\mathbb{Y}$, and $\mathbb{A}$ be Borel spaces, $P(dx'|x, a)$ is a transition kernel on $\mathbb{X}$ given $\mathbb{X} \times \mathbb{A}$, $Q(dy|a, x)$ is a transition kernel on $\mathbb{Y}$ given $\mathbb{A} \times \mathbb{X}$, $Q_0(dy|x)$ is a transition kernel on $\mathbb{Y}$ given $\mathbb{X}$, p_0 is a probability distribution on $\mathbb{X}$, $c : \mathbb{X} \times \mathbb{A} \to \overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ is a bounded from below Borel function on $\mathbb{X} \times \mathbb{A}$.

Partially observable Markov decision process (*POMDP*) *is specified by* $(\mathbb{X}, \mathbb{Y}, \mathbb{A}, P, Q, c)$, *where* $\mathbb{X}$ *is the state space*, $\mathbb{Y}$ *is the observation set*, $\mathbb{A}$ *is the action set*, $P(dx'|x, a)$ *is the state transition law*, $Q(dy|a, x)$ *is the observation kernel*, $c : \mathbb{X} \times \mathbb{A} \to \overline{\mathbb{R}}$ *is the one-step cost*.

The partially observable Markov decision process evolves as follows:

- at time $t = 0$, the initial unobservable state x_0 has a given prior distribution p_0;
- the initial observation y_0 is generated according to the initial observation kernel $Q_0(\cdot\,|x_0)$;
- at each time epoch $n = 0, 1, 2, \ldots$, if the state of the system is $x_n \in \mathbb{X}$ and the decision-maker chooses an action $a_n \in \mathbb{A}$, then the cost $c(x_n, a_n)$ is incurred;
- the system moves to state x_{n+1} according to the transition law $P(\cdot\,|x_n, a_n)$;
- the observation $y_{n+1} \in \mathbb{Y}$ is generated by the observation kernels $Q(\cdot\,|a_n, x_{n+1})$, $n = 0, 1, \ldots$, and $Q_0(\cdot\,|x_0)$.

Define the *observable histories*: $h_0 := (p, y_0) \in H_0$ and $h_n := (p, y_0, a_0, \ldots, y_{n-1}, a_{n-1}, y_n) \in H_n$ for all $n = 1, 2, \ldots$, where $\mathbb{H}_0 := \mathbb{P}(\mathbb{X}) \times \mathbb{Y}$ and $\mathbb{H}_n := \mathbb{H}_{n-1} \times \mathbb{A} \times \mathbb{Y}$ if $n = 1, 2, \ldots$. Then a *policy* for the POMDP is defined as a sequence $\pi = \{\pi_n\}$ such that, for each $n = 0, 1, \ldots$, π_n is a transition kernel on $\mathbb{A}$ given $\mathbb{H}_n$. Moreover, π is called *nonrandomized*, if each probability measure $\pi_n(\cdot|h_n)$ is concentrated at one point. A nonrandomized policy is called *Markov*, if all of the decisions depend on the current state and time only. A Markov policy is called *stationary*, if all the decisions depend on the current state only. The *set of all policies* is denoted by Π. The Ionescu Tulcea theorem (Bertsekas and Shreve [1, pp. 140–141] or Hernández-Lerma and Lassere [7, p. 178]) implies that a policy $\pi \in \Pi$ and an initial distribution $p_0 \in \mathbb{P}(\mathbb{X})$, together with the transition kernels P, Q and Q_0 determine a unique probability measure $P^{\pi}_{p_0}$ on the set of all trajectories $\mathbb{H}_\infty = \mathbb{P}(\mathbb{X}) \times (\mathbb{Y} \times \mathbb{A})^\infty$ endowed with the product of σ-field defined by Borel σ-field of $\mathbb{P}(\mathbb{X})$, $\mathbb{Y}$, and $\mathbb{A}$ respectively. The expectation with respect to this probability measure is denoted by $E^{\pi}_{p_0}$.

Let us specify a performance criterion. For a finite horizon $N = 0, 1, \ldots$, and for a policy $\pi \in \Pi$, let us define the *expected total discounted costs*

$$v^{\pi}_{N,\alpha}(p) := \mathbb{E}^{\pi}_{p} \sum_{n=0}^{N-1} \alpha^n c(x_n, a_n), \ p \in \mathbb{P}(\mathbb{X}), \tag{18.1}$$

where $\alpha \geq 0$ is the discount factor, $v^{\pi}_{0,\alpha}(p) = 0$. When $N = \infty$, we always assume that at least one of the following two assumptions holds:

Assumption (D) c is bounded below on $\mathbb{X} \times \mathbb{A}$ and $\alpha \in [0, 1]$.
Assumption (P) c is nonnegative on $\mathbb{X} \times \mathbb{A}$ and $\alpha \in [0, 1]$.

In the both cases (18.1) defines an *infinite horizon expected total discounted cost*, and we denote it by $v^{\pi}_{\alpha}(p)$. By using notations (**D**) and (**P**), we follow Bertsekas and Shreve [1, p. 214]. However, our Assumption (**D**) is weaker than the corresponding assumption in [1], because c was assumed to be bounded under Assumption (**D**) in [1].

Since the function c is bounded below on $\mathbb{X} \times \mathbb{A}$, a discounted model can be converted into a positive model by shifting the cost function. In particular, let $c(x, a) \geq -K$ for any $(x, a) \in \mathbb{X} \times \mathbb{A}$. Consider a new cost function $\hat{c}(x, a) := c(x, a) + K$ for any $(x, a) \in \mathbb{X} \times \mathbb{A}$. Then the corresponding total discounted reward is equal to

$$\hat{v}^{\pi}_{\alpha}(p) := v^{\pi}_{\alpha}(p) + \frac{K}{1-\alpha}, \qquad \pi \in \Pi, \ p \in \mathbb{P}(\mathbb{X}).$$

Thus, optimizing v^{π}_{α} and $\hat{v}^{\pi}_{\alpha}$ are equivalent problems, but $\hat{v}^{\pi}_{\alpha}$ is the objective function for the positive model. Though positive models are more general, discounted models are met in lager classes of applications. Thus we formulate the results for either of these models.

For any function $g^\pi(p)$, including $g^\pi(p) = v^\pi_{N,\alpha}(p)$ and $g^\pi(p) = v^\pi_\alpha(p)$ define the *optimal cost*

$$g(p) := \inf_{\pi \in \Pi} g^\pi(p), \qquad p \in \mathbb{P}(\mathbb{X}),$$

where Π is *the set of all policies*. A policy π is called *optimal* for the respective criterion, if $g^\pi(p) = g(p)$ for all $p \in \mathbb{P}(\mathbb{X})$. For $g^\pi = v^\pi_{n,\alpha}$, the optimal policy is called *n-horizon discount-optimal*; for $g^\pi = v^\pi_\alpha$, it is called *discount-optimal*.

We recall that a function c defined on $\mathbb{X} \times \mathbb{A}$ is inf-compact (or lower semi-compact) if the set $\{(x, a) \in \mathbb{X} \times \mathbb{A} : c(x, a) \le \lambda\}$ is compact for any finite number λ. A function c defined on $\mathbb{X} \times \mathbb{A}$ is called K-inf-compact on $\mathbb{X} \times \mathbb{A}$, if for any compact subset K of $\mathbb{X}$, the function c is inf-compact on $K \times \mathbb{A}$; Feinberg, Kasyanov, and Zadoianchuk [3, Definition 11]. K-inf-inf-compactness is a mild assumption that is weaker than inf-compactness. Essentially, K-inf-compactness of the cost function c is almost equivalent to lower-semicontinuity of c in the state variable x and lower semi-continuity in the action variable a. This property holds for many applications including inventory control and various problems with least square criteria. According to Feinberg, Kasyanov, and Zadoianchuk [3, Lemma 2.5], a bounded below function c is K-inf-compact on the product of metric spaces $\mathbb{X}$ and $\mathbb{A}$ if and only if it satisfies the following two conditions:

(a) c is lower semi-continuous;
(b) if a sequence $\{x_n\}_{n=1,2,\ldots}$ with values in $\mathbb{X}$ converges and its limit x belongs to $\mathbb{X}$ then any sequence $\{a_n\}_{n=1,2,\ldots}$ with $a_n \in \mathbb{A}$, $n = 1, 2, \ldots$, satisfying the condition that the sequence $\{\overline{c}(x_n, a_n)\}_{n=1,2,\ldots}$ is bounded above, has a limit point $a \in \mathbb{A}$.

As an POMDP $(\mathbb{X}, \mathbb{Y}, \mathbb{A}, P, Q, c)$, consider the classical MDP $(\mathbb{X}, \mathbb{A}, P, c)$, when all the states are observable. An MDP can be viewed as a particular POMDPs with $\mathbb{Y} = \mathbb{X}$ and $Q(B|a, x) = Q(B|x) = \mathbf{I}\{x \in B\}$ for all $x \in \mathbb{X}$, $a \in A$, and $\in \mathscr{B}(\mathbb{X})$. In fact, this POMP possesses a special property that action sets at all the states are equal. For MDPs, Feinberg, Kasyanov, and Zadoianchuk [4] the following general general conditions for the existence of optimal policies, validity of optimality equations, and convergence of value iterations. Here we formulate these conditions for an MDP whose action sets at different states are equal.

Assumption (W*) (cf. Feinberg, Kasyanov, and Zadoianchuk [4] and Lemma 2.5 in [3]).

(i) c is K-inf-compact on $\mathbb{X} \times \mathbb{A}$;
(ii) the transition probability $P(\cdot\,|x, a)$ is weakly continuous in $(x, a) \in \mathbb{X} \times \mathbb{A}$.

Theorem 18.1 (cf. Feinberg, Kasyanov, and Zadoianchuk [4, Theorem 2]). *Let MDP* $(\mathbb{X}, \mathbb{A}, P, c)$ *satisfies Assumption* $(\mathbf{W}^*)$. *Consider either positive or discounted model. Then:*

(*i*) *the functions* $v_{n,\alpha}$, $n = 0, 1, 2, \ldots$, *and* v_α *are lower semi-continuous on* $\mathbb{X}$, *and* $v_{n,\alpha}(x) \to v_\alpha(x)$ *as* $n \to \infty$ *for all* $x \in \mathbb{X}$;

(*ii*) *for any* $x \in \mathbb{X}$, *and* $n = 0, 1, \ldots,$

$$v_{n+1,\alpha}(x) = \min_{a \in A(x)} \left\{ c(x,a) + \alpha \int_{\mathbb{X}} v_{n,\alpha}(y) P(dy|x,a) \right\}, \tag{18.2}$$

where $v_{0,\alpha}(x) = 0$ *for all* $x \in \mathbb{X}$, *and the nonempty sets*

$$A_{n,\alpha}(x) := \{a \in \mathbb{A} : v_{n+1,\alpha}(x) = c(x,a) + \alpha \int_{\mathbb{X}} v_{n,\alpha}(y) P(dy|x,a)\}$$

satisfy the following properties: (a) the graph $\mathrm{Gr}(A_{n,\alpha}) = \{(x,a) : x \in \mathbb{X}, a \in A_{n,\alpha}(x)\}$, $n = 0, 1, \ldots,$ *is a Borel subset of* $\mathbb{X} \times \mathbb{A}$, *and (b) if* $v_{n+1,\alpha}(x) = \infty$, *then* $A_{n,\alpha}(x) = A(x)$ *and, if* $v_{n+1,\alpha}(x) < \infty$, *then* $A_{n,\alpha}(x)$ *is compact;*

(*iii*) *for any* $N = 1, 2, \ldots,$ *there exists a Markov optimal* N*-horizon policy* $(\phi_0, \ldots, \phi_{N-1})$ *and if, for an* N*-horizon Markov policy* $(\phi_0, \ldots, \phi_{N-1})$ *the inclusions* $\phi_{N-1-n}(x) \in A_{n,\alpha}(x)$, $x \in \mathbb{X}$, $n = 0, \ldots, N-1$, *hold then this policy is* N*-horizon optimal;*

(*iv*) *for* $\alpha \in [0, 1]$

$$v_\alpha(x) = \min_{a \in A(x)} \{c(x,a) + \alpha \int_{\mathbb{X}} v_\alpha(y) P(dy|x,a)\}, \qquad x \in \mathbb{X}, \tag{18.3}$$

and the nonempty sets

$$A_\alpha(x) := \{a \in \mathbb{A} : v_\alpha(x) = c(x,a) + \alpha \int_{\mathbb{X}} v_\alpha(y) P(dy|x,a)\}, \quad x \in \mathbb{X},$$

satisfy the following properties: (a) the graph $\mathrm{Gr}(A_\alpha) = \{(x,a) : x \in \mathbb{X}, a \in A_\alpha(x)\}$ *is a Borel subset of* $\mathbb{X} \times \mathbb{A}$, *and (b) if* $v_\alpha(x) = \infty$, *then* $A_\alpha(x) = A(x)$ *and, if* $v_\alpha(x) < \infty$, *then* $A_\alpha(x)$ *is compact;*

(*v*) *for an infinite-horizon there exists a stationary discount-optimal policy* ϕ_α, *and a stationary policy is optimal if and only if* $\phi_\alpha(x) \in A_\alpha(x)$ *for all* $x \in \mathbb{X}$;

(*vi*) (Feinberg and Lewis [5, Proposition 3.1(iv)]) *if* c *is inf-compact on* $\mathbb{X} \times \mathbb{A}$, *then the functions* $v_{n,\alpha}$, $n = 1, 2, \ldots,$ *and* v_α *are inf-compact on* $\mathbb{X}$.

18.3 Reduction of POMDPs to COMDPs and Optimality Results

In this section, we formulate the known reduction of a POMDP to the completely observable Markov decision process (COMDP). Based on general results for MDPs (Feinberg, Kasyanov, Zadoianchuk [4, Theorem 4.1], Theorem 18.2 states sufficient

conditions for the validity of the following results for the COMDP: the existence of stationary optimal policies, the validity of optimality equations, the characterization of optimal policies via optimality equations, and the convergence of value iterations. Then, we formulate the main result of this paper, Theorem 18.3, that states sufficient conditions of these properties in terms of the parameters of the original POMDP.

First, we formulate the well-known reduction of a POMDP to the COMDP ([1, 2, 6, 9, 11]). To simplify notations, we drop sometimes the time parameter. Given a posterior distribution z of the state x at time epoch $n = 0, 1, \ldots$ and given an action a selected at epoch n, denote by $R(B \times C|z, a)$ the joint probability that the state at time $(n + 1)$ belongs to the set $B \in \mathscr{B}(\mathbb{X})$ and the observation at time n belongs to the set $C \in \mathscr{B}(\mathbb{Y})$,

$$R(B \times C|z, a) := \int_{\mathbb{X}} \int_{B} Q(C|a, x')P(dx'|x, a)z(dx), \tag{18.4}$$

where R is a transition kernel on $\mathbb{X} \times \mathbb{Y}$ given $\mathbb{P}(\mathbb{X}) \times \mathbb{A}$; see Bertsekas and Shreve [1]; or Dynkin and Yushkevich [2]; or Hernández-Lerma [6]; or Yushkevich [11] for details. Therefore, the probability $R'(C|z, a)$ that the observation y at time n belongs to the set $C \in \mathscr{B}$ is

$$R'(C|z, a) = \int_{\mathbb{X}} \int_{\mathbb{X}} Q(C|a, x')P(dx'|x, a)z(dx), \tag{18.5}$$

where R' is a transition kernel on $\mathbb{Y}$ given $\mathbb{P}(\mathbb{X}) \times \mathbb{A}$. By Bertsekas and Shreve [1, Proposition 7.27], there exist a transition kernel H on $\mathbb{X}$ given $\mathbb{P}(\mathbb{X}) \times \mathbb{A} \times \mathbb{Y}$ such that

$$R(B \times C|z, a) = \int_{C} H(B|z, a, y)R'(dy|z, a), \tag{18.6}$$

The transition kernel $H(\cdot\,|z, a, y)$ defines a measurable mapping $H : \mathbb{P}(\mathbb{X}) \times \mathbb{A} \times \mathbb{Y} \to \mathbb{P}(\mathbb{X})$, where $H(z, a, y)[\,\cdot\,] = H(\cdot\,|z, a, y)$. For each pair $(z, a) \in \mathbb{P}(\mathbb{X}) \times \mathbb{A}$, the mapping $H(z, a, \cdot) : \mathbb{Y} \to \mathbb{P}(\mathbb{Y})$ is defined $R'(\cdot\,|z, a)$-a.s. uniquely in y; Dynkin and Yushkevich [2, p. 309]. It is known that for a posterior distribution $z_n \in \mathbb{P}(\mathbb{X})$, action $a_n \in A(x)$, and an observation $y_{n+1} \in \mathbb{Y}$, the posterior distribution $z_{n+1} \in \mathbb{P}(\mathbb{X})$ is

$$z_{n+1} = H(z_n, a_n, y_{n+1}). \tag{18.7}$$

However, the observation y_{n+1} is not available in the COMDP model, and therefore y_{n+1} is a random variable with the distribution $R'(\cdot\,|z_n, a_n)$, and (18.7) is a stochastic equation that maps $(z_n, a_n) \in \mathbb{P}(\mathbb{X}) \times \mathbb{A}$ to $\mathbb{P}(\mathbb{P}(\mathbb{X}))$. The stochastic kernel that defines the distribution of z_{n+1} on $\mathbb{P}(\mathbb{X})$ given $\mathbb{P}(\mathbb{X}) \times \mathbb{X}$ is defined uniquely as

$$q(D|z,a) := \int_{\mathbb{Y}} \mathbf{1}_D[H(z,a,y)]R'(dy|z,a), \tag{18.8}$$

where

$$\mathbf{1}_D[u] = \begin{cases} 1, u \in D \in \mathscr{B}(\mathbb{P}(\mathbb{X})), \\ 0, u \notin D \in \mathscr{B}(\mathbb{P}(\mathbb{X})); \end{cases}$$

Hernández-Lerma [7, p. 87]. The measurable particular choice of stochastic kernel H from (18.6) does not effect on the definition of q from (18.8), since for each pair $(z,a) \in \mathbb{P}(\mathbb{X}) \times \mathbb{A}$, the mapping $H(z,a,\cdot) : \mathbb{Y} \to \mathbb{P}(\mathbb{Y})$ is defined $R'(\cdot\,|z,a)$-a.s. uniquely in y; Dynkin and Yushkevich [2, p. 309].

The COMDP is defined as an MDP with parameters $(\mathbb{P}(\mathbb{X}), \mathbb{A}, q, \overline{c})$, where

(i) $\mathbb{P}(\mathbb{X})$ is the state space;
(ii) $\mathbb{A}$ is the action set available at all state $z \in \mathbb{P}(\mathbb{X})$;
(iii) the one-step cost function $\overline{c} : \mathbb{P}(\mathbb{X}) \times \mathbb{A} \to \overline{\mathbb{R}}$, defined as

$$\overline{c}(z,a) := \int_{\mathbb{X}} c(x,a)z(dx), \quad z \in \mathbb{P}(\mathbb{X}),\ a \in \mathbb{A}; \tag{18.9}$$

(iv) transition probabilities q on $\mathbb{P}(\mathbb{X})$ given $\mathbb{P}(\mathbb{X}) \times \mathbb{A}$ defined in (18.8).

see Bertsekas and Shreve [1, Corollary 7.27.1, p. 139] or Dynkin and Yushkevich [2, p. 215], or Hernández-Lerma [6] for details.

If a stationary optimal policy for the COMDP exists and found, it allows the decision maker to compute an optimal policy for the COMDP. First, we recall how the initial state distribution $z_0 \in \mathbb{P}(\mathbb{P}(X))$ can be computed for the COMDP. Similarly to transition kernels R, R', and H, consider a transition kernel

$$R_0(B \times C|p) := \int_B Q_0(C|x)p(dx), \quad B \in \mathscr{B}(\mathbb{X})$$

on $\mathbb{X} \times \mathbb{Y}$ given $\mathbb{P}(\mathbb{X})$. It can be decomposed as

$$R_0(B \times C|p) = \int_C H_0(B|p,y)R_0'(dy|p), \tag{18.10}$$

where

$$R_0'(C|p) = \int_{\mathbb{X}} Q_0(C|x)p(dx), \quad C \in \mathscr{B}(\mathbb{Y}),\ p \in \mathbb{P}(\mathbb{X}),$$

is a transition kernel on $\mathbb{Y}$ given $\mathbb{P}(\mathbb{X})$ and $H_0(\cdot|\cdot,\cdot)$ is a transition kernel on $\mathbb{P}(\mathbb{X})$ given $\mathbb{P}(\mathbb{X}) \times \mathbb{Y}$ that for any initial prior distribution $p_0 \in \mathbb{P}(\mathbb{X})$ and the inital observation y_0 sets the initial posteriori distribution $z_0 = H_0(p_0, y_0)$. Similarly to

(18.7), the observation y_0 is not available in the COMDP, and this equation is a stochastic equation. In addition, $H_0(p, y)$ is defined $R_0'(dy|p)$-a.s. uniquely in y for each $p \in \mathbb{P}(X)$.

Similarly to (18.8), the transition kernel

$$q_0(D|p) := \int_{\mathbb{Y}} \mathbf{1}_D[H_0(p, y)]R_0'(dy|p), \tag{18.11}$$

on $\mathbb{P}(\mathbb{X})$ given $\mathbb{P}(\mathbb{X})$ defines the *initial posterior distribution*. In particular,

$$z_0 := q_0(D|p_0), \qquad D \in \mathbb{P}(\mathbb{X}). \tag{18.12}$$

Define a sequence of *information vectors*

$$i_n := (z_0, a_0, \dots, z_{n-1}, a_{n-1}, z_n) \in I_n, \quad n = 0, 1, \dots,$$

where $z_0 \in \mathbb{P}(\mathbb{X})$ is defined in (18.12), $z_n \in \mathbb{P}(\mathbb{X})$ is recursively defined by Eq. (18.7), $I_n := \mathbb{P}(\mathbb{X}) \times (\mathbb{A} \times \mathbb{P}(\mathbb{X}))^n$ for all $n = 0, 1, \dots$, with $I_0 := \mathbb{P}(\mathbb{X})$. An *information policy* (I-policy) is a policy in a new COMDP, i.e. I-policy is a sequence $\delta = \{\delta_n : n = 0, 1, \dots\}$ such that, for each $n = 0, 1, \dots$, $\delta_n(\cdot\,|i_n)$ is a transition kernel on $\mathbb{A}$ given I_n; Hernández-Lerma [6, p. 88]. Denote by $\triangle$ the set of all I-policies. Identify the set of all Markov I-policies with a subset of Δ.

Consider Δ as a subset of Π; Hernández-Lerma [6, p. 89]. The correspondence of policies in a new COMDP (I-policies) $\delta = \{\delta_n : n = 0, 1, \dots\}$ in $\triangle$ with respective policies $\pi^\delta = \{\pi_n^\delta : n = 0, 1, \dots\}$ in Π is given; Dynkin and Yushkevich [2, pp. 251, 238] and references therein. Moreover, for all $n = 0, 1, \dots$,

$$\pi_n^\delta(\cdot\,|h_n) := \delta_n(\cdot\,|i_n(h_n)) \text{ for all } h_n \in H_n. \tag{18.13}$$

where $i_n(h_n) \in I_n$ is the information vector determined by the observable history h_n via (18.7). Thus δ and π^δ are equivalent in the sense that, for every $n = 0, 1, \dots$, π_n^δ assigns the same conditional probability on $\mathbb{A}$ as that assigned by δ_n for any observable history h_n; Dynkin and Yushkevich [2, pp. 251, 238]; Hernández-Lerma [6, p. 89]. Equality (18.13) yields that I-policy in COMDP is optimal, then the respective policy in initial POMDP is optimal too. For optimality of policy $\pi \in \Pi$ with initial distribution p necessary and sufficient the optimality of respective $\delta^\pi \in \triangle$ with respective initial distribution z^p from (18.12). If δ is stationary, then respective π is stationary too. Therefore, consider an I-policy $\delta \in \triangle$ as a policy $\pi \in \Pi$; see, for example, Dynkin and Yushkevich [2, p. 251], Sawaragi and Yoshikawa [10], Rhenius [9], Yushkevich [11]. The set of policies for the COMDP $(\mathbb{P}(\mathbb{X}), \mathbb{A}, q, q_0, \overline{c})$ is the set $\triangle$ of I-policies; Sawaragi and Yoshikawa [10], Rhenius [9], Yushkevich [11].

This reduction holds for measurable transition kernels P, Q, Q_0. The measurability of these kernels and cost function c lead to the measurability of transition probabilities for the corresponding COMDP. However, it is well known that, except the case of

finite action sets, measurability of transition probabilities is not sufficient for the existence of optimal policies in COMDPs. In spite of this certain properties hold if COMDP satisfies stronger measurability conditions. These properties are provide the validity of optimality equations

$$v_\alpha(z) = \inf_{a\in\mathbb{A}} \left\{ \bar{c}(z,a) + \alpha \int_{\mathbb{P}(\mathbb{X})} v_\alpha(s) q(ds|z,a) \right\},$$

where $z \in \mathbb{P}(\mathbb{X})$, and the property that v_α is a minimal solution of this equation. In addition if the function $\bar{c}$ is bounded on $\mathbb{P}(\mathbb{X})\times\mathbb{A}$, and $\alpha \in [0,1]$, v_α is unique bounded solution of the optimality equation and can be found by value iterations. However, if c is just bounded below on $\mathbb{X}\times\mathbb{A}$, value iterations cannot be applied; Bertsekas [1]. For COMDPs there are sufficient conditions for the existence of stationary optimal policies. If the equivalent COMDP satisfies these conditions, then the optimal policy exists, the value function can be computed by value iterations, the infimum can be substituted with minimum in the optimality equations, and the optimal policy can be derived from the optimality equations. We show below that, if POMDP satisfies these conditions then the COMDP also satisfies them.

For the COMDP, Assumption ($\mathbf{W}^*$) can be rewritten in the following form:

(i) $\bar{c}$ is K-inf-compact on $\mathbb{P}(\mathbb{X})\times\mathbb{A}$;
(ii) the transition probability $q(\cdot|z,a)$ is weakly continuous in $(z,a)\in\mathbb{P}(\mathbb{X})\times\mathbb{A}$.

Theorem 18.1 has the following form for the COMDP $(\mathbb{P}(\mathbb{X}), \mathbb{A}, q, \bar{c})$:

Theorem 18.2 (cf. Feinberg, Kasyanov, and Zadoianchuk [4, Theorem 2]). *Let COMDP $(\mathbb{P}(\mathbb{X}), \mathbb{A}, q, \bar{c})$ satisfy Assumption ($\mathbf{W}^*$) and, in addition, either Assumption ($\mathbf{D}$) or Assumption ($\mathbf{P}$) holds. Then:*

(*i*) *the functions $v_{n,\alpha}$, $n = 0, 1, 2, \ldots$, and v_α are lower semi-continuous on $\mathbb{P}(\mathbb{X})$, and $v_{n,\alpha}(z) \to v_\alpha(z)$ as $n\to\infty$ for all $z\in\mathbb{P}(\mathbb{X})$;*
(*ii*) *for any $z\in\mathbb{P}(\mathbb{X})$, and $n = 0, 1, \ldots,$*

$$\begin{aligned} v_{n+1,\alpha}(z) &= \min_{a\in\mathbb{A}} \{\bar{c}(z,a) + \alpha \int_{\mathbb{P}(\mathbb{X})} v_{n,\alpha}(z') q(dz'|z,a)\} \\ &= \min_{a\in\mathbb{A}} \Big\{ \int_{\mathbb{X}} c(x,a) z(dx) + \int_{\mathbb{X}}\int_{\mathbb{X}}\int_{\mathbb{Y}} v_{n,\alpha}(H(z,a,y)) \\ &\quad \times \alpha Q(dy|a,x') P(dx'|x,a) z(dx) \Big\}, \end{aligned} \tag{18.14}$$

where $v_{0,\alpha}(z) = 0$ for all $z\in\mathbb{P}(\mathbb{X})$, and the nonempty sets

$$\begin{aligned} A_{n,\alpha}(z) := \Big\{ a\in\mathbb{A} :\ & v_{n+1,\alpha}(z) \\ & = c(z,a) + \alpha \int_{\mathbb{P}(\mathbb{X})} v_{n,\alpha}(z') q(dz'|z,a) \Big\}, \end{aligned}$$

where $z \in \mathbb{P}(\mathbb{X})$ *, satisfy the following properties: (a) the graph* $\mathrm{Gr}(A_{n,\alpha}) = \{(z,a) : z \in \mathbb{P}(\mathbb{X}), a \in A_{n,\alpha}(z)\}$, $n = 0, 1, \ldots$, *is a Borel subset of* $\mathbb{P}(\mathbb{X}) \times \mathbb{A}$, *and (b) if* $v_{n+1,\alpha}(z) = \infty$, *then* $A_{n,\alpha}(z) = \mathbb{A}$ *and, if* $v_{n+1,\alpha}(z) < \infty$, *then* $A_{n,\alpha}(z)$ *is compact;*

(*iii*) *for any* $N = 1, 2, \ldots$, *there exists a Markov optimal N-horizon I-policy* $(\phi_0, \ldots, \phi_{N-1})$ *and if, for an N-horizon Markov I-policy* $(\phi_0, \ldots, \phi_{N-1})$ *the inclusions* $\phi_{N-1-n}(z) \in A_{n,\alpha}(z)$, $z \in \mathbb{P}(\mathbb{X})$, $n = 0, \ldots, N-1$, *hold then this I-policy is N-horizon optimal;*

(*iv*) *for* $\alpha \in [0, 1]$

$$
\begin{aligned}
v_\alpha(z) &= \min_{a \in \mathbb{A}} \left\{ \overline{c}(z,a) + \alpha \int_{\mathbb{P}(\mathbb{X})} v_\alpha(z') q(dz'|z,a) \right\} \\
&= \min_{a \in \mathbb{A}} \Bigg\{ \int_{\mathbb{X}} c(x,a) z(dx) + \alpha \int_{\mathbb{X}} \int_{\mathbb{X}} \int_{\mathbb{Y}} v_\alpha(H(z,a,y)) \\
&\qquad \times Q(dy|a,x') P(dx'|x,a) z(dx) \Bigg\}, \quad z \in \mathbb{P}(\mathbb{X}),
\end{aligned}
$$

and the nonempty sets

$$
\begin{aligned}
A_\alpha(z) := \{ a \in \mathbb{A} :\ & v_\alpha(z) = \overline{c}(z,a) \\
& + \alpha \int_{\mathbb{P}(\mathbb{X})} v_\alpha(z') q(dz'|z,a) \}, \quad z \in \mathbb{P}(\mathbb{X}),
\end{aligned}
$$

satisfy the following properties: (a) the graph $\mathrm{Gr}(A_\alpha) = \{(z,a) : z \in \mathbb{P}(\mathbb{X}), a \in \mathbb{A}_\alpha(z)\}$ *is a Borel subset of* $\mathbb{P}(\mathbb{X}) \times \mathbb{A}$, *and (b) if* $v_\alpha(z) = \infty$, *then* $A_\alpha(z) = \mathbb{A}$ *and, if* $v_\alpha(z) < \infty$, *then* $A_\alpha(z)$ *is compact.*

(*v*) *for an infinite horizon there exists a stationary discount-optimal I-policy* ϕ_α, *and a stationary I-policy is optimal if and only if* $\phi_\alpha(z) \in A_\alpha(z)$ *for all* $z \in \mathbb{P}(\mathbb{X})$.

(*vi*) *if the function c is inf-compact, the functions* $v_{n,\alpha}$, $n = 1, 2, \ldots$, *and* v_α *are inf-compact on* $\mathbb{P}(\mathbb{X})$.

Note that statement (vi) of Theorem 18.2 follows from Feinberg and Lewis [5, Proposition 3.1(iv)].

Hernández-Lerma [6, Sect. 4.4] provided the following conditions for the existence of optimal policies for the COMDP: (a) $\mathbb{A}$ is compact, (b) the cost function c is bounded and continuous, (c) the transition probability $P(\cdot|x,a)$ and the observation kernel $Q(\cdot|a,x)$ are weakly continuous transition kernels; (d) there exists a weakly continuous $H : \mathbb{P}(\mathbb{X}) \times \mathbb{A} \times \mathbb{Y} \to \mathbb{P}(\mathbb{X})$ satisfying (18.6). Consider the following relaxed version of Assumption (d).

Assumption (H) There exists a transition kernel H on $\mathbb{X}$ given $\mathbb{P}(\mathbb{X}) \times \mathbb{A} \times \mathbb{Y}$ satisfying (18.6) such that: if a sequence $\{z_n\} \subseteq \mathbb{P}(\mathbb{X})$ converges weakly to $z \in \mathbb{P}(\mathbb{X})$, and $\{a_n\} \subseteq \mathbb{A}$ converges to $a \in \mathbb{A}$, $n \to \infty$, then there exists a subsequence $\{(z_{n_k}, a_{n_k})\}_{k \geq 1} \subseteq \{(z_n, a_n)\}_{n \geq 1}$ such that

$$H(z_{n_k}, a_{n_k}, y) \text{ converges weakly to } H(z, a, y), \ n \to \infty,$$

and this convergence takes place $R'(\cdot|z, a)$ almost surely for all $y \in \mathbb{Y}$.

The following theorem relaxes assumptions (a), (b), and (d) in Hernández-Lerma [6, Sect. 4.4].

Theorem 18.3 *Under the following four conditions:*

(*a*) *either Assumption* (**D**) *or Assumption* (**P**) *holds;*
(*b*) *Assumption* (**W***) *holds for the MDP* $(\mathbb{X}, \mathbb{A}, P, c)$*;*
(*c*) *either the stochastic kernel* $R'(dy|z, a)$ *on* $\mathbb{Y}$ *given* $\mathbb{P}(\mathbb{X}) \times \mathbb{A}$ *is setwise continuous and Assumption* (**H**) *holds, or the stochastic kernel* $Q(dy|a, x)$ *on* $\mathbb{Y}$ *given* $\mathbb{A} \times \mathbb{X}$ *is weakly continuous and there exists a weakly continuous* $H : \mathbb{P}(\mathbb{X}) \times \mathbb{A} \times \mathbb{Y} \to \mathbb{P}(\mathbb{X})$ *satisfying* (18.6)*;*

the COMDP $(\mathbb{P}(\mathbb{X}), \mathbb{A}, q, \overline{c})$ *satisfies Assumption* (**W***) *and therefore statements* (*i*)–(*vi*) *of Theorem 18.2 hold.*

If transition kernel $Q(dy|a, x)$ on $\mathbb{Y}$ given $\mathbb{A} \times \mathbb{X}$ is continuous in the total variation, then Assumption **(H)** holds, and this leads to the following theorem.

Theorem 18.4 *Let the transition kernel* $P(dx'|x, a)$ *on* $\mathbb{X}$ *given* $\mathbb{X} \times \mathbb{A}$ *be weakly continuous and let the transition kernel* $Q(dy|a, x)$ *on* $\mathbb{Y}$ *given* $\mathbb{A} \times \mathbb{X}$ *be continuous in the total variation. Then:* (*i*) *the transition kernel* $R'(dy|z, a)$ *on* $\mathbb{Y}$ *given* $\mathbb{P}(\mathbb{X}) \times \mathbb{A}$ *is setwise continuous, Assumption* (**H**) *holds, and* (*iii*) *the transition kernel* q *on* $\mathbb{P}(\mathbb{X})$ *given* $\mathbb{P}(\mathbb{X}) \times \mathbb{A}$ *is setwise continuous.*

Theorems 18.3 and 18.4 imply the following result.

Theorem 18.5 *Let assumptions of* (*a*) *and* (*b*) *from Theorem 18.3 hold and let the transition kernel* $Q(dy|a, x)$ *on* $\mathbb{Y}$ *given* $\mathbb{A} \times \mathbb{X}$ *be continuous in the total variation. Then statements (i)–(vi) of Theorem 18.2 hold.*

18.4 Example

Let $\mathbb{X}$, $\mathbb{A}$ and $\mathbb{Y}$ are nonempty Borel subsets of $\mathbb{R}$, $\{\xi_n\}_{n\geq 1}$ is a sequence of independent and identically distributed random vectors with values in some Borel subset $\mathbb{S}$ of a Polish space. Assume that the generic disturbance ξ has a distribution μ on $\mathbb{S}$. Let also $\{\eta_n\}_{n\geq 1}$ is a sequence of independent and identically distributed random variables, that uniformly distributed on [0, 1]. The goal is to minimize the expected discounted total costs over the infinite time horizon.

Consider a stochastic partially observable control system of the form

$$x_{n+1} = F(x_n, a_n, \xi_n), \ n = 0, 1, \ldots, \tag{18.15}$$

$$y_{n+1} = G(a_n, x_{n+1}, \eta_n), \ \ n = 0, 1, \ldots, \tag{18.16}$$

where F and G are given measurable function from $\mathbb{X} \times \mathbb{A} \times \mathbb{S}$ to $\mathbb{X}$ and from $\mathbb{A} \times \mathbb{X} \times [0, 1]$ to $\mathbb{Y}$ respectively. The states x_n are not observable, while the states y_n are observable.

The transition law of the system can be written as

$$P(B|x, a) = \int_{\mathbb{S}} \mathbf{1}\{F(x, a, s) \in B\}\mu(ds).$$

The observation kernel is given by

$$Q(C|a, x) = \int_{[0,1]} \mathbf{1}\{G(a, x, s) \in C\}\lambda(ds),$$

where $\lambda \in \mathbb{P}([0, 1])$ is a Lebesgue measure on $[0, 1]$.

It is clear that, if $(x, a) \to F(x, a, s)$ is continuous mapping on $\mathbb{X} \times \mathbb{A}$ for every $s \in \mathbb{S}$, then stochastic kernel $P(dx'|x, a)$ on $\mathbb{X}$ given $\mathbb{X} \times \mathbb{A}$ is weakly continuous.

Assume that G is a continuous mapping on $\mathbb{A} \times \mathbb{X} \times [0, 1]$, its derivative by the last variable exists (we denote it by g) is a continuous mapping on $\mathbb{A} \times \mathbb{X} \times [0, 1]$ and it has a fixed sign, i.e. for some constant $\beta > 0$ we have $|g(a, x, s)| \geq \beta$ for any $a \in \mathbb{A}, x \in \mathbb{X}, s \in G(a, x, [0, 1])$, where $G(a, x, [0, 1]) = \{G(a, x, s') : s' \in [0, 1]\}$. Then it is possible to show that that the observation transition kernel Q on $\mathbb{Y}$ given $\mathbb{A} \times \mathbb{X}$ is continuous in the total variation.

Finally, we assume that one-period cost $c : \mathbb{X} \times \mathbb{A} \to \overline{\mathbb{R}}$ is K-inf-compact function (see for details Feinberg, Kasyanov, and Zadoianchuk [3]), it is bounded from below. Then the MDP satisfies Assumption ($\mathbf{W}^*$), that is, K-inf-compactness of the cost function c and weak continuity of the transition kernel P that describes transition probabilities for the MDP. In addition, the observation transition kernel Q is continuous in the total variation. Therefore, the corresponding COMDP satisfies Assumption ($\mathbf{W}^*$). Thus, in view of Theorems 18.3–18.5 for the COMDP there exist a stationary optimal, they satisfy optimality equations, and the value function can be computed via value iterations. By using the standard known procedures [6, Chap. 4], an optimal policy for the COMDP can be used to construct an optimal policy for the initial problem, which is typically nonstationary.

18.5 Conclusions

This presentation studies POMDPs with Borel state, action, and observation spaces satisfying mild continuity assumptions that guarantee the following properties for the underlying fully observable MDP: (i) the existence of stationary optimal policies, (ii) validity of optimality equations, and (iii) convergence of value iterations for the expected total discounted costs as well as for the expected total costs, when the one-

step cost function is nonnegative. This presentation provides additional sufficient conditions under which the COMDP possesses the same continuity assumptions as the underlying MDP and, therefore, properties (i)–(iii) are also satisfied for the COMDP. One of such sufficient conditions is the continuity of the observation transition kernel in the total probability; see Theorem 18.5. Therefore, this paper provides theoretical foundations to analyze POMDPs with general state and action spaces and with expected total cost criteria.

Acknowledgments The authors thank Dr. Huizhen Janey Yu and Dr. N.V. Zadoianchuk for their useful remarks. Research of the first coauthor was partially supported by NSF grants CMMI-0928490 and CMMI-1335296.

References

1. Bertsekas, D.P., Shreve, S.E.: Stochastic Optimal Control: The Discrete-Time Case. Academic Press, New York (1978) (reprinted by Athena Scientific, Belmont, 1996)
2. Dynkin, E.B., Yushkevich, A.A.: Controlled Markov Processes. Springer, New York (1979)
3. Feinberg, E.A., Kasyanov, P.O., Zadoianchuk, N.V.: Berge's theorem for noncompact image sets. J. Math. Anal. Appl. **397**(1), 255–259 (2013)
4. Feinberg, E.A., Kasyanov, P.O., Zadoianchuk, N.V.: Average-cost Markov decision processes with weakly continuous transition probabilities. Math. Oper. Res. **37**(4), 591–607 (2012)
5. Feinberg, E.A., Lewis, M.E.: Optimality inequalities for average cost Markov decision processes and the stochastic cash balance problem. Math. Oper. Res. **32**(4), 769–783 (2007)
6. Hernández-Lerma, O.: Adaptive Markov Control Processes. Springer, New York (1989)
7. Hernández-Lerma, O., Lassere, J.B.: Discrete-Time Markov Control Processes: Basic Optimality Criteria. Springer, New York (1996)
8. Parthasarathy, K.R.: Probability Measures on Metric Spaces. Academic Press, New York (1967)
9. Rhenius, D.: Incomplete information in Markovian decision models. Ann. Statist. **2**, 1327–1334 (1974)
10. Sawaragi, Y., Yoshikawa, T.: Descrete-time markovian decision processes with incomplete state observations. Ann. Math. Statist. **41**, 78–86 (1970)
11. Yushkevich, A.A.: Reduction of a controlled Markov model with incomplete data to a problem with complete information in the case of Borel state and control spaces. Theor. Probab. Appl. **21**, 153–158 (1976)

Chapter 19
On Existence of Optimal Solutions to Boundary Control Problem for an Elastic Body with Quasistatic Evolution of Damage

Peter I. Kogut and Günter Leugering

Abstract We study an optimal control problem for the mixed boundary value problem for an elastic body with quasistatic evolution of an internal damage variable. We use the damage field $\zeta = \zeta(t, x)$ as an internal variable which measures the fractional decrease in the stress-strain response. When $\zeta = 1$ the material is damage-free, when $\zeta = 0$ the material is completely damaged, and for $0 < \zeta < 1$ it is partially damaged. We suppose that the evolution of microscopic cracks and cavities responsible for the damage is described by a nonlinear parabolic equation, whereas the model for the stress in elastic body is given as $\boldsymbol{\sigma} = \zeta(t, x)A\mathbf{e}(\mathbf{u})$. The optimal control problem we consider in this paper is to minimize the appearance of micro-cracks and micro-cavities as a result of the tensile or compressive stresses in the elastic body.

19.1 Introduction

The damage modeling in the context of industrial applications is in its infancy—corrosion, multi-micro cracking etc. This makes this problem extremely complex. The main idea of a novel approach to modeling material damage is to use the so-called damage field $\zeta = \zeta(t, x)$ as an internal variable which measures the fractional decrease in the stress-strain response. The evolution of the damage field is derived from the principle of virtual work under appropriate assumptions on the

P. I. Kogut (✉)
Department of Differential Equations, Dnipropetrovsk National University, Gagarin av. 72, Dnipropetrovsk 49010, Ukraine
e-mail: p.kogut@i.ua

G. Leugering
Institüt für Angewandte Mathematik Lehrstuhl II Universität, Erlangen-Nürnberg Martensstr.3, 91058 Erlangen, Germany
e-mail: Guenter.Leugering@am.uni-erlangen.de

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_19,

system's free energy, the dissipation pseudopotential, and the spatial interactions of the microcracks. In this approach the damage field ζ varies between one and zero at each point in the body. When $\zeta = 1$ the material is damage-free, when $\zeta = 0$ the material is completely damaged, and for $0 < \zeta < 1$ it is partially damaged. The evolution of the damage field is usually described by a parabolic inclusion or equation with a damage source function ϕ which depends on the mechanical compression or tension [7]. At the same time, the model for the stress is given as $\boldsymbol{\sigma} = \zeta(t, x)A\mathbf{e}(\mathbf{u})$. Without the damage parameter ζ, this is the classical model of elastic material. However, if parameter ζ varies in the interval [0, 1], the corresponding elasticity system

$$-\text{div } (\zeta A\mathbf{e}(\mathbf{u})) = \mathbf{f}$$

becomes degenerate.

In this paper we assume that the elastic body under consideration occupies the domain Ω and is clamped on the part S of its boundary, and the rest part of the boundary $\Gamma = \partial\Omega \setminus S$ is the influence zone of a Neumann control. Therefore, the control variable is the density of a surface traction $\mathbf{p}$ acting on Γ. The optimal control problem we consider in this paper aims at two objectives. On the one hand we try to minimize the discrepancy between a given displacement field $\mathbf{u}_d$ and the solution of the initial-boundary value problem by choosing an appropriate surface traction $\mathbf{p} \in \mathscr{P}_{ad}$. On the other hand, we wish to minimize the appearance of micro-cracks and micro-cavities as a result of the tensile or compressive stresses in the elastic body. To the best knowledge of authors the existence of optimal solutions for the above problem is an open question. Moreover, only few papers deal with optimal control problems for degenerate partial differential equations (see for example [1–3, 5, 6]).

19.2 Notation and Preliminaries

Let Ω be a bounded open connected subset of $\mathbb{R}^N$ ($N \geq 2$) with Lipschitz boundary. We assume that Ω is occupied by some elastic body and its outer surface $\partial\Omega$ is divided into two disjoint measurable parts $\partial\Omega = \Gamma \cup S$. Let the sets S and Γ have positive $(N-1)$-dimensional measures and let S be closed.

For any subset $E \subset \mathbb{R}^N$ we denote by $|E|$ its N-dimensional Lebesgue measure $\mathscr{L}^N(E)$. Let χ_E be the characteristic function of a subset $E \subset \mathbb{R}^N$, i.e. $\chi_E(x) = 1$ if $x \in E$, and $\chi_E(x) = 0$ if $x \notin E$.

We will often use the Lebesgue spaces of vector-value functions. For example, for the L^2-space of vector-valued functions $\mathbf{u}(x) = (u_1(x), \ldots, n_N(x))^t \in \mathbb{R}^N$ we use the notation $L^2(\Omega)^N = L^2(\Omega, \mathbb{R}^N)$. At the same time, $L^2(\Omega)^{\frac{N(N-1)}{2}} = L^2\big(\Omega; \mathbb{R}^{\frac{N(N-1)}{2}}\big)$ is the space of square-summable functions whose values are symmetric matrices. We denote by $\mathbb{S}^N := \mathbb{R}^{\frac{N(N-1)}{2}}$ the set of all symmetric matrices $\xi = [\xi_{ij}]_{i,j=1}^N$, $(\xi_{ij} = \xi_{ji})$. We suppose that $\mathbb{S}^N$ is endowed with the euclidian

scalar product $\boldsymbol{\xi} \cdot \boldsymbol{\eta} = \mathrm{tr}(\boldsymbol{\xi}\,\boldsymbol{\eta}) = \xi_{ij}\eta_{ij}$ and with the corresponding euclidian norm $\|\boldsymbol{\xi}\|_{\mathbb{S}^N} = (\xi \cdot \xi)^{1/2}$. Hereinafter, we adopt the convention regarding summation with respect to repeating indices. In particular, $\boldsymbol{\xi}^2 = \xi_{ij}\xi_{ij}$.

We denote by $A(x) = \left[A^{kl}(x)\right]_{k,\,l=1}^{N} = \left\{a_{ij}^{kl}(x)\right\}$ an elasticity tensor at a material point $x \in \Omega$. The action of the elasticity tensor $A(x)$ on the matrix $\boldsymbol{\xi} \in \mathbb{S}^N$ is defined by $A(x)\boldsymbol{\xi} = \left\{a_{ij}^{kl}(x)\xi_{kl}\right\}$. Then, $A(x)\boldsymbol{\xi}\cdot\boldsymbol{\xi} = a_{ij}^{kl}(x)\xi_{kl}\xi_{ij}$ is the elastic energy density. It is assumed that $A(x)$ satisfies the usual symmetry conditions:

$$a_{ij}^{kl}(x) = a_{ji}^{lk}(x) = a_{il}^{kj}(x), \quad \forall\, i, j, k, l = 1, 2, \ldots, N. \tag{19.1}$$

Let κ_1 and κ_2 be two fixed constants such that $\kappa_2 > \kappa_1$. We define $\mathscr{A}_{\kappa_1}^{\kappa_2}(\Omega)$ as the set of all symmetric elasticity tensors $A(x) = \left\{a_{ij}^{kl}(x)\right\}$ such that the positive definiteness condition holds:

$$\kappa_1\boldsymbol{\xi}^2 \le \mathscr{A}(x)\boldsymbol{\xi}\cdot\boldsymbol{\xi} \le \kappa_2\boldsymbol{\xi}^2 \quad \text{a.e. in } \Omega \quad \forall\, \boldsymbol{\xi} \in \mathbb{S}^N. \tag{19.2}$$

In order to describe a quasistatic evolution of damage in the elastic body Ω, we denote by $\mathbf{u}(x) = (u_1(x), \ldots, u_N(x))$ the displacement field, $\sigma(x) = \left\{\sigma_{ij}(x)\right\}$ the stress tensor, and $\mathbf{e}(\mathbf{u}) = \left\{\mathbf{e}_{ij}(\mathbf{u})\right\}$ the strain tensor. We assume that for every smooth vector $\mathbf{u}(x) = (u_1(x), \ldots, u_N(x))$ the formula for the strain tensor $\mathbf{e}_{ij}(\mathbf{u})$ is provided by the Cauchy law of small deformations

$$\mathbf{e}_{ij}(\mathbf{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \quad \forall\, i, j = 1, \ldots, N. \tag{19.3}$$

It is clear that $\mathbf{e}(\mathbf{u}) \in \mathbb{S}^N$ and $\mathbf{e}(\mathbf{u})$ is the symmetric part of the gradient of a displacement $\mathbf{u}$. Thus $\mathbf{e}(\mathbf{u}) = \frac{1}{2}\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^t\right)$, where the gradient of a displacement $\mathbf{u} \in \mathbb{R}^N$ is the $(N \times N)$ matrix $\nabla\mathbf{u}$ the entries of which are defined by $(\nabla\mathbf{u})_{ij} := \frac{\partial u_i}{\partial x_j}$.

Hence, for any symmetric tensor $A \in \mathscr{A}_{\kappa_1}^{\kappa_2}(\Omega)$, we have $A\mathbf{e}(\mathbf{u}) = A\nabla\mathbf{u}$. Therefore, we will use indifferently both expressions. Note also that the divergence of a smooth matrix $\boldsymbol{\sigma}(x)$ is the vector $\mathrm{div}\,(\boldsymbol{\sigma}) \in \mathbb{R}^N$ the components of which are defined by $(\mathrm{div}\,(\boldsymbol{\sigma}))_i := \sum_{j=1}^{N} \frac{\partial \sigma_{ij}}{\partial x_j}$.

Let $\Omega_T = (0, T) \times \Omega$ for some $T > 0$. Let ζ denote a damage field in Ω_T and measures the fractional decrease in the strength of the material. Usually, for an isotropic material, the damage field $\zeta = \zeta(t, x)$ is defined as the ratio $\zeta = \zeta(t, x) = \frac{E_{eff}}{E}$ between the effective modulus of elasticity E_{eff} and that of the damage-free material E. It follows from this definition that the damage field should only have values between 0 and 1. Since every damage $\zeta : \Omega_T \to [0, 1]$ gives rise to a measure on the measurable subsets of Ω_T through integration, we will denote this measure by ζ. Thus $\zeta(E) = \int_E \zeta\, dz$ for measurable sets $E \subset \Omega_T$. We will use the standard notation $L^2(\Omega_T, \zeta\, dz)$ for the set of measurable functions f on Ω_T such that

$$\|f\|_{L^2(\Omega_T,\,\zeta\,dz)} = \left(\int_{\Omega_T} f^2\,\zeta\,dz\right)^{1/2} = \left(\int_0^T \int_\Omega f^2\,\zeta\,dxdt\right)^{1/2} < +\infty.$$

Let $C_0^\infty(\mathbb{R}^N;\,S) = \left\{\varphi \in C_0^\infty(\mathbb{R}^N)\ :\ \varphi = 0 \text{ on } S\right\}$ be the set of smooth damages in Ω. We define the space $H^1(\Omega;\,S)$ as the closure of $C_0^\infty(\mathbb{R}^N;\,S)$ with respect to the norm

$$\left(\int_\Omega \left[y^2 + |\nabla y|^2_{\mathbb{R}^N}\right] dx\right)^{1/2}.$$

Let $\mathscr{Z} = L^2(0,T;\,H^1(\Omega))$, $\mathscr{V} = L^2(0,T;\,H^1(\Omega;\,S))$. Let $\mathscr{Z}' = L^2(0,T;\,H^1(\Omega)')$ and $\mathscr{V}' = L^2(0,T;\,H^1(\Omega;\,S)')$ be their dual. The following theorem plays an important role in the study of an quasistatic evolution of damage in an elastic bodies (see Simon [10]).

Theorem 19.1 *Let us define the Banach space*

$$\mathscr{W} = \left\{\zeta\ :\ \zeta \in \mathscr{Z},\quad \frac{\partial \zeta}{\partial t} \in \mathscr{Z}'\right\},$$

equipped with the norm of the graph. Then, the following properties hold true:

1. *the embedding $\mathscr{W} \hookrightarrow L^2(0,T;\,L^2(\Omega))$ is compact;*
2. *one has the embedding*

$$\mathscr{W} \hookrightarrow C([0,T];\,L^2(\Omega)), \tag{19.4}$$

   where, $C([0,T];\,L^2(\Omega))$ denotes the space of measurable functions on $[0,T]\times\Omega$ such that $\zeta(t,\cdot) \in L^2(\Omega)$ for any $t \in [0,T]$ and such that the map $t \in [0,T] \mapsto \zeta(t,\cdot) \in L^2(\Omega)$ is continuous;
3. *for any $\zeta, v \in \mathscr{W}$*

$$\frac{d}{dt}\int_\Omega \zeta(t,x)v(t,x)\,dx = \left\langle \zeta'(t,\cdot),\, v(t,\cdot)\right\rangle_{\mathscr{Z}',\mathscr{Z}} + \left\langle v'(t,\cdot),\, \zeta(t,\cdot)\right\rangle_{\mathscr{Z}',\mathscr{Z}}. \tag{19.5}$$

Definition 19.1 We say that a damage $\zeta : \Omega_T \to [0,1]$ is substantial in Ω, if

$$\zeta^{-1} \notin L^\infty(\Omega_T) \quad \text{and} \quad \zeta^{-1} \in L^1(\Omega_T). \tag{19.6}$$

Note that in this case the functions in $L^2(\Omega_T,\,\zeta\,dxdt)$ are Lebesgue integrable on Ω_T.

Let W be the closure of the set of pairs $\left\{(\mathbf{u},\mathbf{e}(\mathbf{u}))\ :\ \mathbf{u} \in C_0^\infty(R^N;\,S)^N\right\}$ in the product of spaces $L^1(\Omega)^N \times L^1(\Omega)^{\frac{N(N+1)}{2}}$. Thus the elements of W are pairs $(\mathbf{u},\mathbf{z})$, where $\mathbf{u}$ is a vector and $\mathbf{z} = \mathbf{e}(\mathbf{u})$ is the symmetric gradient of the vector $\mathbf{u}$. In what follows, we define the space $\mathscr{W}^{1,1}(\Omega;\,S)$ as the union of the first components $\mathbf{u}$ of

W. Following standard technique, it is easy to show that $\mathscr{W}^{1,1}(\Omega; S)$ is a Banach space with respect to the norm $\|\mathbf{u}\|_{\mathscr{W}^{1,1}(\Omega;S)} = \int_\Omega \left[|\mathbf{u}|_{\mathbb{R}^N} + \|\mathbf{e}(\mathbf{u})\|_{\mathbb{S}^N}\right] dx$. To each damage field $\zeta(t, x)$ we may associate two weighted spaces

$$W_\zeta(\Omega \times (0, T); S) \text{ and } H_\zeta(\Omega \times (0, T); S),$$

where $W_\zeta(\Omega \times (0, T); S)$ is the set of vector-functions $\mathbf{u} \in L^1(0, T; \mathscr{W}^{1,1}(\Omega; S))$ for which the norm

$$\|\mathbf{u}\|_\zeta = \Big(\int_0^T \int_\Omega \big(\mathbf{u}^2 + \mathbf{e}^2(\mathbf{u})\zeta\big)\, dxdt \Big)^{1/2} \tag{19.7}$$

is finite, and $H_\zeta(\Omega \times (0, T); S)$ is the closure of the set

$$\Big\{ \psi(t)\boldsymbol{\varphi}(x) \ : \ \psi \in C_0^\infty(0, T), \ \boldsymbol{\varphi} \in C_0^\infty(\mathbb{R}^N; S)^N \Big\} \tag{19.8}$$

in the $W_\zeta(\Omega \times (0, T); S)$-norm. Note that due to the estimates

$$\int_0^T \int_\Omega |\mathbf{u}|_{\mathbb{R}^N}\, dxdt \le \Big(\int_0^T \int_\Omega \mathbf{u}^2\, dxdt \Big)^{1/2} \sqrt{T|\Omega|} \le C\|\mathbf{u}\|_\zeta, \tag{19.9}$$

$$\begin{aligned} \int_0^T \int_\Omega \|\mathbf{e}(\mathbf{u})\|\, dxdt &:= \int_0^T \int_\Omega (\mathbf{e}(\mathbf{u}) \cdot \mathbf{e}(\mathbf{u}))^{1/2}\, dxdt \\ &\le \Big(\int_0^T \int_\Omega \mathbf{e}^2(\mathbf{u})\zeta\, dxdt \Big)^{1/2} \Big(\int_0^T \int_\Omega \zeta^{-1}\, dxdt \Big)^{1/2} \le C\|\mathbf{u}\|_\zeta, \end{aligned} \tag{19.10}$$

the space $W_\zeta(\Omega \times (0, T); S)$ is complete with respect to the norm $\|\cdot\|_\zeta$. Moreover, it is clear that $H_\zeta(\Omega \times (0, T); S) \subseteq W_\zeta(\Omega \times (0, T); S)$, and $W_\zeta(\Omega \times (0, T); S)$, $H_\zeta(\Omega \times (0, T); S)$ are Hilbert spaces endowed with the scalar product

$$(\mathbf{u}, \mathbf{v})_\zeta = \int_0^T \int_\Omega [\mathbf{u} \cdot \mathbf{v} + \mathbf{e}(\mathbf{u}) \cdot \mathbf{e}(\mathbf{v})\zeta]\, dxdt. \tag{19.11}$$

If the damage field $\zeta = \zeta(t, x)$ is bounded between two positive constants, then it is easy to verify that

$$W_\zeta(\Omega \times (0, T); S) = H_\zeta(\Omega \times (0, T); S). \tag{19.12}$$

However, for a "substantial" damage ζ in the sense of Definition 19.1, the set of smooth functions (19.8) is not dense in $W_\zeta(\Omega \times (0, T); S)$. Hence the identity (19.12) is not always valid.

19.3 Radon Measures and Convergence in Variable Spaces

By a nonnegative Radon measure on Ω_T we mean a nonnegative Borel measure which is finite on every compact subset of Ω_T. The space of all nonnegative Radon measures on Ω_T will be denoted by $\mathscr{M}_+(\Omega_T)$. According to the Riesz Representation Theorem, each Radon measure $\mu \in M_+(\Omega_T)$ can be interpreted as element of the dual of the space $C_0(\Omega_T)$ of all continuous functions vanishing at infinity. If μ is a nonnegative Radon measure on Ω_T, we will use $L^r(\Omega_T, d\mu)$, $1 \le r \le \infty$, to denote the usual Lebesgue space with respect to the measure μ with the corresponding norm $\|f\|_{L^r(\Omega_T,\, d\mu)} = \left(\int_{\Omega_T} |f(x)|^r \, d\mu\right)^{1/r}$.

Let $\{\mu_k\}_{k\in\mathbb{N}}$, μ be Radon measures such that $\mu_k \overset{*}{\rightharpoonup} \mu$ in $\mathscr{M}_+(\Omega_T)$, i.e.,

$$\lim_{k\to\infty} \int_{\Omega_T} \psi\varphi \, d\mu_k = \int_{\Omega_T} \psi\varphi \, d\mu \quad \forall\, \psi \in C_0(\mathbb{R}),\ \forall\, \varphi \in C_0(\mathbb{R}^N), \tag{19.13}$$

where $C_0(\mathbb{R}^N)$ is the space of all compactly supported continuous functions. A typical example of such measures is

$$d\mu_k = \zeta_k(t,x)\, dxdt, \quad d\mu = \zeta(t,x)\, dxdt, \ \text{where}\ 0 \le \zeta_k \rightharpoonup \zeta \ \text{in}\ L^1(\Omega_T). \tag{19.14}$$

Let us recall the definition and main properties of convergence in the variable L^2-space.

1. A sequence $\left\{\mathbf{v}_k \in L^2(\Omega_T, d\mu_k)^N\right\}_{k\in\mathbb{N}}$ is called bounded if

$$\limsup_{k\to\infty} \int_{\Omega_T} |\mathbf{v}_k|^2_{\mathbb{R}^N} \, d\mu_k < +\infty.$$

2. A bounded sequence $\left\{\mathbf{v}_k \in L^2(\Omega_T, d\mu_k)^N\right\}_{k\in\mathbb{N}}$ converges weakly to $\mathbf{v} \in L^2(\Omega_T, d\mu)^N$ if $\lim_{k\to\infty} \int_{\Omega_T} \mathbf{v}_k \cdot \boldsymbol{\varphi}\, d\mu_k = \int_{\Omega_T} \mathbf{v} \cdot \boldsymbol{\varphi}\, d\mu$ for any $\boldsymbol{\varphi} \in C_0^\infty(\Omega_T)^N$, which is denoted as $\mathbf{v}_k \rightharpoonup \mathbf{v}$ in $L^2(\Omega_T, d\mu_k)^N$.
3. The strong convergence $\mathbf{v}_k \to \mathbf{v}$ in $L^2(\Omega_T, d\mu_k)^N$ means that $\mathbf{v} \in L^2(\Omega_T, d\mu)^N$ and

$$\lim_{k\to\infty} \int_{\Omega_T} \mathbf{v}_k \cdot \mathbf{z}_k \, d\mu_k = \int_{\Omega_T} \mathbf{v} \cdot \mathbf{z}\, d\mu \ \text{ as } \ \mathbf{z}_k \rightharpoonup \mathbf{z} \ \text{ in } \ L^2(\Omega_T, d\mu_k)^N. \tag{19.15}$$

The following convergence properties in variable spaces hold:

(a) *Compactness*: if a sequence is bounded in $L^2(\Omega_T, d\mu_k)^N$, then this sequence is compact in the sense of the weak convergence in $L^2(\Omega, d\mu_k)^N$;
(b) *Lower semicontinuity*: if $\mathbf{v}_k \rightharpoonup \mathbf{v}$ in $L^2(\Omega_T, d\mu_k)^N$, then

$$\liminf_{k\to\infty}\int_{\Omega_T}|\mathbf{v}_k|^2_{\mathbb{R}^N}\,d\mu_k \geq \int_{\Omega_T}|\mathbf{v}|^2_{\mathbb{R}^N}\,d\mu; \tag{19.16}$$

(c) *Strong convergence*: $\mathbf{v}_k \to \mathbf{v}$ if and only if $\mathbf{v}_k \rightharpoonup \mathbf{v}$ in $L^2(\Omega_T, d\mu_k)^N$ and

$$\lim_{k\to\infty}\int_{\Omega_T}|\mathbf{v}_k|^2_{\mathbb{R}^N}\,d\mu_k = \int_{\Omega_T}|\mathbf{v}|^2_{\mathbb{R}^N}\,d\mu. \tag{19.17}$$

For our further analysis we make use the following concept.

Definition 19.2 We say that a bounded sequence

$$\Big\{(\zeta_n, \mathbf{u}_n) \in L^2(\Omega_T) \times W_{\zeta_n}(\Omega \times (0,T); S)\Big\}_{n\in\mathbb{N}} \tag{19.18}$$

w-converges to $(\zeta, \mathbf{u}) \in L^2(\Omega_T) \times L^1(0,T; \mathscr{W}^{1,1}(\Omega; S))$ as $n \to \infty$, if

$$\zeta_n \rightharpoonup \zeta \quad \text{in } L^2(\Omega_T), \tag{19.19}$$

$$\mathbf{u}_n \rightharpoonup \mathbf{u} \quad \text{in } L^2(\Omega_T)^N, \tag{19.20}$$

$$\mathbf{e}(\mathbf{u}_n) \rightharpoonup \mathbf{e}(\mathbf{u}) \quad \text{in the variable space } L^2\big(0,T; L^2(\Omega, \zeta_n\,dx)^{\frac{N(N+1)}{2}}\big), \tag{19.21}$$

that is,

$$\lim_{n\to\infty}\int_0^T\int_\Omega \zeta_n\eta\,dxdt = \int_0^T\int_\Omega \zeta\eta\,dxdt \quad \forall\,\eta \in L^2(\Omega_T), \tag{19.22}$$

$$\lim_{n\to\infty}\int_0^T\int_\Omega \mathbf{u}_n\cdot\boldsymbol{\lambda}\,dxdt = \int_0^T\int_\Omega \mathbf{u}\cdot\boldsymbol{\lambda}\,dxdt \quad \forall\,\boldsymbol{\lambda} \in L^2(\Omega_T)^N, \tag{19.23}$$

and

$$\lim_{n\to\infty}\int_0^T\int_\Omega \zeta_n\mathbf{e}(\mathbf{u}_n)\cdot\boldsymbol{\xi}(x)\phi(t)\,dxdt = \int_0^T\int_\Omega \zeta\mathbf{e}(\mathbf{u})\cdot\boldsymbol{\xi}(x)\phi(t)\,dxdt$$
$$\forall\,\psi \in C_0^\infty(0,T),\ \forall\,\boldsymbol{\xi} \in C_0^\infty(\Omega; \mathbb{S}^N). \tag{19.24}$$

In order to verify the correctness of this definition, we give the following result.

Lemma 19.1 *Let* $\big\{(\zeta_n, \mathbf{u}_n) \in L^2(\Omega_T) \times W_{\zeta_n}(\Omega \times (0,T); S)\big\}_{n\in\mathbb{N}}$ *be a sequence such that*

(i) this sequence is bounded, i.e.

$$\sup_{n\in\mathbb{N}}\left[\int_0^T\int_\Omega \left(\zeta_n^2+\mathbf{u}_n^2+\mathbf{e}^2(\mathbf{u}_n)\zeta_n\right)dxdt\right]<+\infty; \tag{19.25}$$

(ii) there exists an element $\zeta\in L^1(\Omega_T)$ *such that*

$$\zeta_n\to\zeta \quad \textit{and} \quad \zeta_n^{-1}\to\zeta^{-1} \quad \textit{in} \ L^1(\Omega_T) \ \textit{as} \ n\to\infty, \tag{19.26}$$

(iii) $\zeta_n:\Omega_T\to[0,1]$ *for all* $n\in\mathbb{N}$.

Then, this sequence is relatively compact with respect to w-convergence. Moreover, each w-limit pair $(\zeta,\mathbf{u})$ *belongs to the corresponding space* $L^2(\Omega_T)\times W_\zeta(\Omega\times(0,T);S)$.

Proof To begin with, we note that the condition (19.25) and estimates (19.9)–(19.10) immediately imply the boundedness of the sequence in $L^2(\Omega_T)\times L^1(0,T;\mathscr{W}^{1,1}(\Omega;S))$. The uniform boundedness of $\{\zeta_n\}_{n\in\mathbb{N}}$ in $L^2(\Omega_T)$ and property (19.26) ensure that the limit damage field ζ belongs to $L^2(\Omega_T)$ as well. Moreover, we have (see the property (19.14)): $d\zeta_n:=\zeta_n\,dxdt\overset{*}{\rightharpoonup}\zeta\,dxdt=:d\zeta$ in $\mathscr{M}_+(\Omega_T)$.

Then, the compactness criterium of the weak convergence in variable spaces (see property (a)) and condition (19.25) leads us to the existence of a pair $(\mathbf{u},\mathbf{v})\in L^2\big(0,T;L^2(\Omega)^N\big)\times L^2\big(0,T;L^2(\Omega,\zeta\,dx)^{\frac{N(N+1)}{2}}\big)$ such that, within a subsequence of $\{\mathbf{u}_n\}_{n\in\mathbb{N}}$,

$$\mathbf{u}_n\rightharpoonup\mathbf{u} \ \text{ in } L^2(\Omega_T)^N, \tag{19.27}$$

$$\mathbf{e}(\mathbf{u}_n)\rightharpoonup\mathbf{v} \ \text{ in variable space } \ L^2\big(0,T;L^2(\Omega,\zeta_n\,dx)^{\frac{N(N+1)}{2}}\big). \tag{19.28}$$

Our aim is to show that $\mathbf{v}=\mathbf{e}(\mathbf{u})$ and $\mathbf{u}\in W_\zeta(\Omega\times(0,T);S)$. Indeed, for any $\varphi\in C_0^\infty(\Omega)$ and $\psi\in C_0^\infty(0,T)$, we have

$$\int_0^T\int_\Omega \zeta_n^{-1}\varphi\psi\zeta_n\,dxdt=\int_0^T\int_\Omega\varphi\psi\,dxdt=\int_0^T\int_\Omega\zeta^{-1}\varphi\psi\zeta\,dxdt,$$

i.e. $\zeta_n^{-1}\rightharpoonup\zeta^{-1}$ in $L^2(\Omega_T,d\zeta_n)$. Moreover, the strong convergence in $(19.26)_2$ implies the relation

$$\lim_{n\to\infty}\int_0^T\int_\Omega\zeta_n^{-2}\zeta_n\,dxdt=\lim_{n\to\infty}\int_0^T\int_\Omega\zeta_n^{-1}\,dxdt=\int_0^T\int_\Omega\zeta^{-2}\zeta\,dxdt.$$

Hence, $\zeta_n^{-1}\to\zeta^{-1}$ strongly in $L^2(\Omega_T,d\zeta_n)$ (see property (c)), and therefore,

$$\psi\boldsymbol{\xi}\zeta_n^{-1}\to\psi\boldsymbol{\xi}\zeta^{-1} \ \text{ strongly in } \ L^2\big(0,T;L^2(\Omega,\zeta_n\,dx)^{\frac{N(N+1)}{2}}\big) \tag{19.29}$$

for each $\psi\in C_0^\infty(0,T)$ and $\boldsymbol{\xi}\in C_0^\infty(\Omega;\mathbb{S}^N)$. Further, we note that for every measurable subset $K\subset\Omega_T$, the estimate

$$\int_K \|\mathbf{e}(\mathbf{u}_n)\|_{\mathbb{S}^N}\,dz \le \left(\int_K \zeta_n^{-1}\,dz\right)^{1/2}\left(\int_K \|\mathbf{e}(\mathbf{u}_n)\|^2_{\mathbb{S}^N}\zeta_n\,dz\right)^{1/2} \le C\left(\int_K \zeta_n^{-1}\,dz\right)^{1/2}$$

implies the equi-integrability of the family $\left\{\|\mathbf{e}(\mathbf{u}_n)\|_{\mathbb{S}^N}\right\}_{n\in\mathbb{N}}$. Hence, the sequence $\left\{\|\mathbf{e}(\mathbf{u}_n)\|_{\mathbb{S}^N}\right\}_{n\in\mathbb{N}}$ is weakly compact in $L^1(\Omega_T)$, which means the weak compactness of the matrix-valued sequence $\{\mathbf{e}(\mathbf{u}_n)\}_{n\in\mathbb{N}}$ in $L^1(0,T;L^1(\Omega;\mathbb{S}^N))$. As a result, by the properties of the strong convergence in variable spaces, we obtain

$$\begin{aligned}\int_0^T\int_\Omega \mathbf{e}(\mathbf{u}_n)\cdot\boldsymbol{\xi}(x)\phi(t)\,dxdt &= \int_0^T\int_\Omega \mathbf{e}(\mathbf{u}_n)\cdot\left(\boldsymbol{\xi}(x)\phi(t)\zeta_n^{-1}\right)\zeta_n\,dxdt\\ &\overset{\text{by (19.15), (19.28), and (19.29)}}{\longrightarrow} \int_0^T\int_\Omega \mathbf{v}\cdot\left(\boldsymbol{\xi}(x)\phi(t)\zeta^{-1}\right)\zeta\,dxdt\\ &= \int_0^T\int_\Omega \mathbf{v}\cdot\boldsymbol{\xi}(x)\phi(t)\,dxdt \quad \forall\,\psi\in C_0^\infty(0,T),\ \forall\,\boldsymbol{\xi}\in C_0^\infty(\Omega;\mathbb{S}^N).\end{aligned}$$

Thus, in view of the weak compactness property of the sequence $\{\mathbf{e}(\mathbf{u}_n)\}_{n\in\mathbb{N}}$ in $L^1(0,T;L^1(\Omega;\mathbb{S}^N))$, we conclude

$$\mathbf{e}(\mathbf{u}_n) \rightharpoonup \mathbf{v} \text{ in } L^1(0,T;L^1(\Omega;\mathbb{S}^N)) \quad \text{as } n\to\infty. \tag{19.30}$$

Since $\mathbf{u}_n \in L^1(0,T;\mathscr{W}^{1,1}(\Omega;S))$ for all $n\in\mathbb{N}$ and the space $L^1(0,T;\mathscr{W}^{1,1}(\Omega;S))$ is complete, the conditions (19.27) and (19.30) imply $\mathbf{e}(\mathbf{u}) = \mathbf{v}$, and consequently $\mathbf{u}\in L^1(0,T;\mathscr{W}^{1,1}(\Omega;S))$. To end the proof, it remains to observe that the conditions (19.27)–(19.28) guarantee the finiteness of the norm $\|\mathbf{u}\|_\zeta$ (see (19.7)). Hence $\mathbf{u}\in W_\zeta(\Omega\times(0,T);S)$ and this concludes the proof.

As an obvious consequence of this lemma, we have the following result.

Corollary 19.1 *The main statement of Lemma 19.1 remains true if we replace the condition (ii) by the following one: there exists an element $\zeta\in L^1(\Omega_T)$ such that*

$$\zeta_n\to\zeta \text{ in } L^1(\Omega_T), \quad \text{and } \zeta_n^{-1}\to\zeta^{-1} \text{ in } L^2(\Omega_T,d\zeta_n).$$

19.4 The Model of Quasistatic Evolution of Damage in an Elastic Material

In this section we describe the model for the control process in an elastic body, present its variational formulation, and discuss the questions on existence and uniqueness of weak solution.

We consider an elastic body which occupies the domain Ω. We assume that the body is clamped on the surface S and so the displacement field vanishes there. We suppose that the remaining part of the boundary $\Gamma = \partial\Omega\setminus S$ is the influence zone

of a Neumann control. So, the control variable is the density of surface traction $\mathbf{p}$ acting on Γ. Let $\mathbf{f}$ be a given density of volume forces acting in $\Omega_T = (0, T) \times \Omega$ for some $T > 0$.

For a simplicity, we assume that an initial displacement (when $t = 0$) and an initial stress tensor are equal to zero. Then, having assumed that the Hooke law $\boldsymbol{\sigma}_{ij} = a_{ijkl}\mathbf{e}_{kl}, \forall i, j = 1, \ldots, N$ holds true for the elastic body Ω, we adopt the following relation for the stress $\boldsymbol{\sigma} : \Omega_T \to \mathbb{S}^N$ in the body with damage (see [4, 9] for the details):

$$\boldsymbol{\sigma}(t, x) = \zeta(t, x) A\mathbf{e}(\mathbf{u}(t, x)) \quad \text{a.e. in } \Omega_T, \tag{19.31}$$

where $\zeta = \zeta(t, x)$ is a damage field in Ω_T.

Following the motivation in Kuttler [7], the evolution of the microscopic cracks and cavities responsible for the damage can be described by the equation

$$\zeta' - \kappa \Delta \zeta = \phi(x, \mathbf{e}(\mathbf{u}), \zeta).$$

Here the prime denotes the time derivative, Δ is the Laplace operator, $\kappa > 0$ is a damage diffusion constant, ϕ is the damage source function. Usually, it is assumed that the damage source term $\phi : \Omega \times \mathbb{S}^N \times \mathbb{R}$ satisfies some Lipschitz continuity property and is such that whenever $\zeta > 1, \phi(\mathbf{e}(\mathbf{u}), \zeta) \leq 0$. This assumption makes sense because there should be no way that the source term for the damage produces damage greater than 1.

Let $\zeta_{ad} : \Omega \to [0, 1]$ be a given $L^1(\Omega)$-function satisfying the properties

$$\zeta_{ad}^{-1} \in L^1(\Omega), \quad \zeta_{ad}^{-1} \notin L^\infty(\Omega).$$

Let Ψ_* be a nonempty compact subset of $L^1(\Omega)$ such that the conditions

$$\zeta_{ad} \leq \zeta \leq 1 \text{ a.e. in } \Omega, \tag{19.32}$$

$$\zeta : \Omega \to [0, 1] \text{ is smooth function on the surface } \Gamma, \tag{19.33}$$

$$\zeta = 1 \quad \text{on} \quad \Gamma. \tag{19.34}$$

hold true for every $\zeta \in \Psi_*$. So, each element $\zeta : \Omega \to [0, 1]$ of Ψ_* can be interpreted as a substantial time-independent damage field in the sense of Definition 19.1.

The characteristic feature of this set is the following property.

Proposition 19.1 *Let* $\{\zeta_{*,n}\}_{n\in\mathbb{N}}$ *and* ζ_* *be such that* $\zeta_{*,n} \to \zeta_*$ *in* $L^1(\Omega_T)$ *as* $n \to \infty$, *and* $\{\zeta_{*,n}(t, \cdot)\}_{n\in\mathbb{N}} \subset \Psi_*$ *and* $\zeta_*(t, \cdot) \in \Psi_*$ *for all* $t \in [0, T]$. *Then*

$$\zeta_{*,n}^{-1} \to \zeta_*^{-1} \text{ in } L^1(\Omega_T), \text{ and } \zeta_{*,n}^{-1} \to \zeta_*^{-1} \text{ in } L^2(\Omega_T, d\zeta_{*,n}). \tag{19.35}$$

Proof In view of the initial assumptions, we may assume that $\zeta_{*,n}^{-1} \to \zeta_*^{-1}$ almost everywhere in Ω_T. Since $\zeta_{*,n} \to \zeta_*$ in $L^1(\Omega_T)$ and $\zeta_*^{-1} \le \zeta_{ad}^{-1} \in L^1(\Omega)$, it follows that the sequence $\{\zeta_{*,n}^{-1}\}_{n\in\mathbb{N}}$ is equi-integrable on Ω_T. Hence the property $(19.35)_1$ is a direct consequence of Lebesgue's Theorem. As for the property $(19.35)_2$, it was proved in Lemma 19.1. The proof is complete.

As a result, we adopt the following model for the controlled process in Ω: for a given body force $\mathbf{f} \in L^2\big(0,T;L^2(\Omega)^N\big)$, a surface traction $\mathbf{p} \in \mathscr{P}_{ad}$, the set Ψ_*, and an initial damage field $\zeta_0 \in L^2(\Omega)$ for which

$$\exists\, \zeta_*^0 \in \Psi_* \text{ such that } \zeta_*^0 \le \zeta_0 \le 1 \quad \text{a.e. in } \Omega, \tag{19.36}$$

a displacement field $\mathbf{u} : \Omega_T \to \mathbb{R}^N$, a stress field $\boldsymbol{\sigma} : \Omega_T \to \mathbb{S}^N$, and a damage field $\zeta : \Omega_T \to \mathbb{R}$ satisfy the relations

$$-\operatorname{div}\boldsymbol{\sigma} = \mathbf{f} \quad \text{in} \quad \Omega_T, \tag{19.37}$$

$$\boldsymbol{\sigma} = \zeta A\mathbf{e}(\mathbf{u}) \quad \text{in} \quad \Omega_T, \tag{19.38}$$

$$\boldsymbol{\sigma} = 0 \quad \text{on} \quad (0,T)\times S, \tag{19.39}$$

$$\boldsymbol{\sigma}\boldsymbol{\nu} = \mathbf{p} \quad \text{on} \quad (0,T)\times\Gamma,\ \mathbf{p} \in \mathscr{P}_{ad}, \tag{19.40}$$

$$\zeta' - \kappa\Delta\zeta = \phi(\mathbf{e}(\mathbf{u}),\zeta) \quad \text{in} \quad \Omega_T, \tag{19.41}$$

$$\zeta(0,\cdot) = \zeta_0 \quad \text{in} \quad \Omega, \tag{19.42}$$

$$\zeta = 1 \quad \text{on } (0,T)\times\Gamma, \quad \partial\zeta/\partial n = 0 \quad \text{on } (0,T)\times S, \tag{19.43}$$

$$\exists\, \zeta_* \in \Psi_* \text{ such that } \zeta_* \le \zeta(t,x) \le 1 \quad \text{a.e. in} \quad \Omega_T. \tag{19.44}$$

Here $\boldsymbol{\nu}$ is the outward unit normal to Γ, $\partial/\partial n = n_i\partial/\partial x_i$, n_i denotes i^{th}-component of the unit outward normal vector to S, and $\mathscr{P}_{ad}$ is the set of admissible controls to the process (19.37)–(19.44). For simplicity, we suppose that $\mathscr{P}_{ad}$ is defined as

$$\mathscr{P}_{ad} = \left\{\mathbf{p} \in L^2\big(0,T;L^2(\Gamma)^N\big) \ :\ \|\mathbf{p}\|_{L^2\left(0,T;L^2(\Gamma)^N\right)} \le C_{\mathbf{p}}\right\}. \tag{19.45}$$

To begin with, we note that, to the best knowledge of the authors, the existence of a global weak solution to the initial-boundary value problem (19.37)–(19.44) in an open question. There are several reasons for this. First, this problem is restricted by the state constraints (19.44). It means that without the implication of the truncation operators in the model, the initial conditions (19.42) with properties (19.36) and parabolic equation (19.41) with boundary conditions (19.43), do not guarantee the fulfilment of the inequality (19.44). Secondly, even if a damage field is admissible,

i.e. ζ remains between some $\zeta_* \in \Psi_*$ and 1, the properties (19.32)–(19.34), and (19.44) imply that the original problem (19.37)–(19.40) is a mixed boundary value problem for the degenerate elasticity system

$$-\text{div } (\zeta A\mathbf{e}(\mathbf{u})) = \mathbf{f} \quad \text{in} \quad \Omega_T,$$

This means that for some damage field $\zeta(t, x)$ this problem can exhibit non-uniqueness of weak solutions [11], the Lavrentieff phenomenon, and other surprising consequences.

In view of this, we adopt the following concept:

Definition 19.3 We say that a vector-valued function $\mathbf{u} = \mathbf{u}(\mathbf{p}, \mathbf{f}, \zeta)$ is a weak solution to the boundary value problem (19.37)–(19.40) for a fixed control $\mathbf{p} \in \mathscr{P}_{ad}$, a given body force $\mathbf{f} \in L^2\big(0, T; L^2(\Omega)^N\big)$, and a given damage field $\zeta : \Omega_T \to [0, 1]$ satisfying the condition (19.44), if $\mathbf{u} \in W_\zeta(\Omega \times (0, T); S)$ and the integral identity

$$\int_0^T \int_\Omega \big[\zeta(t, x)A(x)\mathbf{e}(\mathbf{u}) \cdot \mathbf{e}(\boldsymbol{\varphi})\big]\psi \, dxdt = \int_0^T \int_\Omega \mathbf{f} \cdot \boldsymbol{\varphi}\psi \, dxdt + \int_0^T \int_\Gamma \mathbf{p} \cdot \boldsymbol{\varphi}\psi \, d\mathscr{H}^{N-1}dt \tag{19.46}$$

holds for any $\boldsymbol{\varphi} \in C_0^\infty(\mathbb{R}^N; S)^N$ and $\psi \in C_0^\infty(0, T)$.

As was mentioned in Sect. 19.2, the set of smooth functions (19.8) is not dense in the weighted space $W_\zeta(\Omega \times (0, T); S)$. Hence, we can not assert that a weak solution to the degenerate elasticity problem (19.37)–(19.40) is unique. Further, we make use the following result:

Proposition 19.2 *Let Γ be a Lipschitz continuous part of the boundary $\partial\Omega$. Let $\zeta : \Omega_T \to [0, 1]$ be a damage field satisfying the estimate* (19.44)*. Then there exists a bounded linear operator*

$$\gamma_\Gamma : W_\zeta(\Omega \times (0, T); S) \to L^2\big(0, T; H^{1/2}(\Gamma)^N\big) \tag{19.47}$$

such that

(i) $\gamma_\Gamma(\mathbf{u}) = \mathbf{u}|_\Gamma$ *if* $\mathbf{u} \in W_\zeta(\Omega \times (0, T); S) \cap C\big([0, T]; C(\overline{\Omega})^N\big)$,

(ii) $\|\gamma_\Gamma(\mathbf{u})\|_{L^2(0,T;H^{1/2}(\Gamma)^N)} \le C\|\mathbf{u}\|_{W_\zeta(\Omega\times(0,T);S)}$ *for each vector-valued function* $\mathbf{u} \in W_\zeta(\Omega \times (0, T); S)$ *with the constant C independent of Γ.*

Corollary 19.2 *Under the assumptions of Proposition 19.2, the space $W_\zeta(\Omega \times (0, T); S)$ does not contain rigid displacements. In other words, if $\widehat{\mathbf{u}} \neq 0$ is a vector-valued function for which there exists a sequence of smooth functions $\big\{\boldsymbol{\varphi} \in C_0^\infty\big(0, T; C_0^\infty(\mathbb{R}^N)^N\big)\big\}_{n\in\mathbb{N}}$ such that*

$$\boldsymbol{\varphi}_n \to \widehat{\mathbf{u}} \ \ in \ \ L^2(\Omega_T)^N, \quad \mathbf{e}(\boldsymbol{\varphi}_n) \to \mathbf{0} \ \ in \ \ L^2\big(0,T;L^2(\Omega,\zeta\,dx)^{\frac{N(N+1)}{2}}\big),$$

then $\widehat{\mathbf{u}} \notin W_\zeta(\Omega\times(0,T);S)$.

We now give the variational formulation of the initial boundary value problem (19.41)–(19.43).

Definition 19.4 Let $\mathbf{p}\in\mathscr{P}_{ad}, \mathbf{f}\in L^2\big(0,T;L^2(\Omega)^N\big)$, and $\zeta_0\in L^2(\Omega)$ be given functions. We say that a pair $(\zeta,\mathbf{u})\in\mathscr{Z}\times W_\zeta(\Omega\times(0,T);S)$ is a corresponding weak variational solution to the initial-boundary value problem (19.37)–(19.44) with a nonlinear source for the damage $\phi: L^1(0,T;\mathscr{W}^{1,1}(\Omega;S))\times\mathscr{Z}\to L^2(\Omega_T)$, if

$$\frac{\partial\zeta}{\partial t}\in\mathscr{Z}', \quad \zeta-1\in\mathscr{V}, \tag{19.48}$$

and there is an element $\zeta_*\in\Psi_*$ such that the following relations hold true

$$\int_0^T\int_\Omega\big[\zeta(t,x)A(x)\mathbf{e}(\mathbf{u})\cdot\mathbf{e}(\boldsymbol{\varphi})\big]\psi\,dxdt=\int_0^T\int_\Omega\mathbf{f}\cdot\boldsymbol{\varphi}\psi\,dxdt$$
$$+\int_0^T\int_\Gamma\mathbf{p}\cdot\boldsymbol{\varphi}\psi\,d\mathscr{H}^{N-1}dt \quad \forall\,\boldsymbol{\varphi}\in C_0^\infty(\mathbb{R}^N;S)^N, \quad \forall\,\psi\in C_0^\infty(0,T), \tag{19.49}$$

$$\big\langle\zeta',\varphi\psi\big\rangle_{\mathscr{Z}',\mathscr{Z}}+\kappa\int_0^T\int_\Omega\nabla\zeta\cdot\nabla\varphi\,\psi\,dxdt$$
$$=\int_0^T\int_\Omega\phi(\zeta,\mathbf{e}(\mathbf{u}))\varphi\psi\,dxdt \quad \forall\,\varphi\in C_0^\infty(\mathbb{R}^N;\Gamma), \quad \forall\,\psi\in C_0^\infty(0,T), \tag{19.50}$$

$$\zeta(0,\cdot)=\zeta_0(\cdot) \quad \text{in} \quad \Omega, \tag{19.51}$$

$$\zeta_*\le\zeta(t,x)\le 1 \quad \text{for all} \ \ t\in[0,T] \ \ \text{and a.e.} \ \ x\in\Omega. \tag{19.52}$$

Remark 19.1 As follows from Theorem 19.1, the condition (19.52) is reasonable. It means that the initial damage field $\zeta_0\in L^2(\Omega)$ must also be restricted by this inequality.

Remark 19.2 It is worth to notice that the original initial-boundary value problem (19.37)–(19.44) is ill-possed, in general. This means that there are no reasons to suppose that for every admissible initial data $\mathbf{p}\in\mathscr{P}_{ad}, \mathbf{f}\in L^2\big(0,T;L^2(\Omega)^N\big)$, $\zeta_0\in L^2(\Omega)$, and $\zeta_*\in\Psi_*$ this system admits at least one weak variational solution $(\zeta,\mathbf{u})\in\mathscr{Z}\times W_\zeta(\Omega\times(0,T);S)$ in the sense of Definition 19.4. At the same time, by analogy with [5, 6] it can be shown that this system may have an infinitely many weak solutions $(\zeta,\mathbf{u})$ for some fixed admissible control $\mathbf{p}\in\mathscr{P}_{ad}$.

19.5 Setting of the Optimal Control Problems and Existence Theorem for Optimal Traction

The optimal control problem we consider in this paper is twofold. On the one hand we try to minimize the discrepancy between a given displacement field $\mathbf{u}_d \in L^2\big(0,T;L^2(\Omega;\mathbb{R}^N)\big)$ and the solution of the problem (19.37)–(19.44) by choosing an appropriate surface traction $\mathbf{p} \in \mathscr{P}_{ad}$. On the other hand, we wish to minimize the appearance of micro-cracks and micro-cavities as a result of the tensile or compressive stresses in the elastic body. More precisely, we are concerned with the following optimal control problem

$$\text{Minimize } \Big\{ I(\mathbf{p},\mathbf{u},\zeta) = \int_0^T \int_\Omega |\mathbf{u}-\mathbf{u}_d|^2_{\mathbb{R}^N}\, dxdt + \int_0^T \int_\Omega |\zeta - 1|\, dxdt + \int_0^T \int_\Omega \|\mathbf{e}(\mathbf{u})\|^2_{\mathbb{S}^N} \zeta\, dxdt \Big\} \tag{19.53}$$

subject to the constraints (19.37)–(19.45).

We introduce the set of admissible solutions to the original optimal control problem as follows:

$$\begin{aligned} \Xi := \big\{ (\mathbf{p},\zeta,\mathbf{u}) \,\big|\, & \mathbf{p} \in \mathscr{P}_{ad},\ \zeta \in \mathscr{Z},\ \mathbf{u} \in W_\zeta(\Omega\times(0,T);S), \\ & (\zeta,\mathbf{u}) \text{ is a weak variational solution to 19.37–19.44} \\ & \text{in the sense of Definition 19.4} \big\}. \end{aligned} \tag{19.54}$$

We say that a triplet $(\mathbf{p}^0,\zeta^0,\mathbf{u}^0) \in L^2\big(0,T;L^2(\Gamma)^N\big) \times \mathscr{Z} \times W_{\zeta^0}(\Omega\times(0,T);S)$ is optimal for problem (19.53), (19.37)–(19.45) if

$$(\mathbf{p}^0,\zeta^0,\mathbf{u}^0) \in \Xi \ \text{ and } \ I(\mathbf{p}^0,\zeta^0,\mathbf{u}^0) = \inf_{(\mathbf{p},\zeta,\mathbf{u})\in\Xi} I(\mathbf{p},\zeta,\mathbf{u}). \tag{19.55}$$

Remark 19.3 Note that due to the estimates (19.9) and (19.10), we have the following obvious inclusion for the set of admissible solutions

$$\Xi \subset L^2\big(0,T;L^2(\Gamma)^N\big) \times L^2(0,T;H^1(\Omega)) \times L^1\big(0,T;\mathscr{W}^{1,1}(\Omega;S)\big).$$

However, the characteristic feature of this set is the fact that for different admissible controls $\mathbf{p} \in \mathscr{P}_{ad}$ and, therefore, for different admissible damage fields $\zeta : \Omega_T \to [0,1]$ with properties prescribed above, the corresponding admissible solutions $(\mathbf{p},\zeta,\mathbf{u})$ of optimal control problem (19.53), (19.37)–(19.45) belong to different weighted spaces. It is a non-typical situation from the point of view of the classical optimal control theory.

Definition 19.5 We say that the mapping

$$\phi : L^1(0,T;\mathscr{W}^{1,1}(\Omega;S)) \times \mathscr{Z} \to L^2(\Omega_T) \tag{19.56}$$

possesses the property $(\mathfrak{M})$ on Ξ, if

(M_1) for any open bounded domain $Q \subset \mathbb{R}^N$ with a Lipschitz boundary such that $\Omega \subseteq Q$ and $S \subset \partial Q$, this mapping can be extended to the following one

$$\widetilde{\phi} : L^1(0,T;\mathscr{W}^{1,1}(Q;S)) \times L^2(0,T;H^1(Q)) \to L^2(0,T;L^2(Q))$$

which is weakly-$*$ continuous with respect to the w-convergence, i.e. for any sequence $\{(\mathbf{p},\widetilde{\zeta}_n,\widetilde{\mathbf{u}}_n)\}_{n\in\mathbb{N}} \subset \Xi \subset L^2(0,T;H^1(Q)) \times L^1(0,T;\mathscr{W}^{1,1}(Q;S))$ such that

$$\widetilde{\mathbf{u}}_n \in W_{\widetilde{\zeta}_n}(Q\times(0,T);S) \quad \forall\, n \in \mathbb{N}, \tag{19.57}$$

$$(\widetilde{\zeta}_n,\widetilde{\mathbf{u}}_n) \overset{w}{\to} (\widetilde{\zeta},\widetilde{\mathbf{u}}) \ \text{ as } \ n\to\infty \ \text{ in the sense of Definition 19.2} \tag{19.58}$$

(where instead of Ω we have to put Q), the equality

$$\lim_{n\to\infty} \big(\widetilde{\phi}\big(\mathbf{e}(\widetilde{\mathbf{u}}_n),\widetilde{\zeta}_n\big),\varphi\psi\big)_{L^2(0,T;L^2(Q))} = \big(\widetilde{\phi}\big(\mathbf{e}(\widetilde{\mathbf{u}}),\widetilde{\zeta}\big),\varphi\psi\big)_{L^2(0,T;L^2(Q))}$$

holds $\forall\,\varphi \in C_0^\infty(\mathbb{R}^N;\Gamma)$ and $\forall\,\psi \in C_0^\infty(0,T)$;

(M_2) the mapping (19.56) is locally bounded in the following sense: for any constants $C_1, C_2 > 0$ there is a constant $C_3 = C_3(C_1,C_2) > 0$ such that

$$\Big|\,(\phi(\mathbf{e}(\mathbf{u}),\zeta),\zeta-1)_{L^2(\Omega_T)}\,\Big| \le C_3 \tag{19.59}$$

provided $(\mathbf{u},\zeta) \in W_\zeta(\Omega\times(0,T);S)\times\mathscr{Z}$, $\zeta - 1 \in \mathscr{V}$, and

$$(\|\mathbf{e}(\mathbf{u})\|_{L^2\left(0,T;L^2(\Omega,\zeta\,dx)^{\frac{N(N+1)}{2}}\right)} \le C_1, \quad \|\zeta\|_{L^2(\Omega_T)} \le C_2. \tag{19.60}$$

Remark 19.4 Note that for any admissible initial damage field $\zeta_0 \in L^2(\Omega)$, the verification of the regularity property $\Xi \neq \emptyset$ for the original optimal control problem (19.53), (19.37)–(19.45) is a non-trivial problem, in general. In the particular case, when the damage field $\zeta(t,x)$ is assumed to be strictly separated from 0, the regularity property follows from results of Kuttler & Shillor, where the solvability of a similar initial-boundary value problem with a fixed surface traction $\mathbf{p}$ is studied).

Since our prime interest in this section deals with the solvability of optimal control problem (19.53), (19.37)–(19.45), we begin with the study of the topological properties of the set of admissible solutions Ξ.

Definition 19.6 A sequence $\{(\mathbf{p}_n, \zeta_n, \mathbf{u}_n) \in \Xi\}_{n\in\mathbb{N}}$ is called bounded if

$$\sup_{n\in\mathbb{N}} \left[\|\mathbf{p}_n\|_{L^2(0,T;L^2(\Gamma)^N)} + \|\zeta_n\|_{\mathscr{Z}} + \|\mathbf{u}_n\|_{\zeta_n} \right] < +\infty.$$

Definition 19.7 We say that a bounded sequence $\{(\mathbf{p}_n, \zeta_n, \mathbf{u}_n) \in \Xi\}_{n\in\mathbb{N}}$ of admissible solutions τ-converges to a triplet $(\mathbf{p}, \zeta, \mathbf{u}) \in L^2\big(0, T; L^2(\Gamma)^N\big) \times L^2(0, T; H^1(\Omega)) \times L^1\big(0, T; \mathscr{W}^{1,1}(\Omega; S)\big)$ if

(S_1) $\mathbf{p}_n \rightharpoonup \mathbf{p}$ in $L^2\big(0, T; L^2(\Gamma)^N\big)$,
(S_2) $\zeta_n \rightharpoonup \zeta$ in $\mathscr{Z} := L^2(0, T; H^1(\Omega))$,
(S_3) $\mathbf{u}_n \rightharpoonup \mathbf{u}$ in $L^2\big(0, T; L^2(\Omega)^N\big)$,
(S_4) $\mathbf{e}(\mathbf{u}_n) \rightharpoonup \mathbf{e}(\mathbf{u})$ in the variable space $L^2\big(0, T; L^2(\Omega, \zeta_n\, dx)^{\frac{N(N+1)}{2}}\big)$.

Due to the estimates like (19.9)–(19.10), the inclusion $\mathbf{u} \in L^1\big(0, T; \mathscr{W}^{1,1}(\Omega; S)\big)$ is obvious.

Remark 19.5 As immediately follows from Definition 19.2 and Rellich-Kondrashov Theorem (see also Theorem 19.1), if $(\mathbf{p}_n, \zeta_n, \mathbf{u}_n) \xrightarrow{\tau} (\mathbf{p}, \zeta, \mathbf{u})$ then $(\zeta_n, \mathbf{u}_n) \xrightarrow{w} (\zeta, \mathbf{u})$.

Lemma 19.2 *Let $\{(\mathbf{p}_n, \zeta_n, \mathbf{u}_n) \in \Xi\}_{n\in\mathbb{N}}$ be a bounded sequence. Then there exists a triplet*

$$(\mathbf{p}, \zeta, \mathbf{u}) \in L^2\big(0, T; L^2(\Gamma)^N\big) \times L^2(0, T; H^1(\Omega)) \times L^1\big(0, T; \mathscr{W}^{1,1}(\Omega; S)\big)$$

such that, up to a subsequence, $(\mathbf{p}_n, \zeta_n, \mathbf{u}_n) \xrightarrow{\tau} (\mathbf{p}, \zeta, \mathbf{u})$ and $\mathbf{u} \in W_\zeta(\Omega \times (0, T); S)$.

Proof To begin with, we note that by the compactness criterium of the weak convergence in Banach reflexive spaces, there exist a subsequence of $\{(\mathbf{p}_n, \zeta_n)\}_{n\in\mathbb{N}}$, still denoted by the same indices, and $\mathbf{p} \in L^2\big(0, T; L^2(\Gamma)^N\big)$, $\zeta \in L^2(0, T; H^1(\Omega))$ are such that the conditions (S_1)–(S_2) hold true. In order to check the rest conditions (S_3)–(S_4) of Definition 19.7, we make use the following observation.

Since $(\mathbf{p}_n, \zeta_n, \mathbf{u}_n) \in \Xi$ for all $n \in \mathbb{N}$, it follows that there is a sequence $\{\zeta_{*,n}\}_{n\in\mathbb{N}}$ in Ψ_* such that (see Definition 19.4)

$$\zeta_{*,n}(x) \le \zeta_n(t, x) \le 1 \quad \text{for all } t \in [0, T] \text{ and a.e. } x \in \Omega. \tag{19.61}$$

Moreover, by L^1-compactness property of the set Ψ_*, there exists an element $\widehat{\zeta}_* \in \Psi_*$ such that $\zeta_{*,n} \to \widehat{\zeta}_*$ in $L^1(\Omega_T)$ as $n \to \infty$. Then Proposition 19.1 implies the strong convergence

$$\zeta_{*,n}^{-1} \to \widehat{\zeta}_*^{-1} \text{ in } L^1(\Omega_T). \tag{19.62}$$

Hence, in view of (19.61), we have: $\zeta_n \to \zeta$, $\zeta_n^{-1} \to \zeta^{-1}$ in $L^1(\Omega_T)$ as $n \to \infty$, and the inequality $\widehat{\zeta}_* \le \zeta \le 1$ holds a.e. in Ω_T. Thus, by Remark 19.5, all suppositions of Lemma 19.1 are fulfilled. As a result, the fulfilment of the rest conditions (S_3)–(S_4) and the inclusion $\mathbf{u} \in W_\zeta(\Omega \times (0,T); S)$ for w-limiting component of the sequence $\{(\zeta_n, \mathbf{u}_n)\}_{n\in\mathbb{N}}$, are ensured by Lemma 19.1. The proof is complete.

Our next step deals with the study of topological properties of the set of admissible solutions Ξ to the problem (19.53), (19.37)–(19.45).

Theorem 19.2 *Assume that $\Xi \neq \emptyset$ and the damage source term $\phi : \mathbb{S}^N \times \mathbb{R} \to \mathbb{R}$ possesses the property $(\mathfrak{M})$. Then for every force $\mathbf{f} \in L^2\big(0,T; L^2(\Omega)^N\big)$ and every initial damage field $\zeta_0 : \Omega \to [0,1]$ satisfying the condition (19.44), the set of admissible solutions Ξ is sequentially closed with respect to the τ-convergence.*

Proof Let $\{(\mathbf{p}_n, \zeta_n, \mathbf{u}_n) \in \Xi\}_{n\in\mathbb{N}}$ be a bounded τ-convergent sequence of admissible solutions to the optimal control problem (19.53), (19.37)–(19.45). Let $(\widehat{\mathbf{p}}, \widehat{\zeta}, \widehat{\mathbf{u}})$ be its τ-limit. Our aim is to prove that $(\widehat{\mathbf{p}}, \widehat{\zeta}, \widehat{\mathbf{u}}) \in \Xi$.

By the definition of the set of admissible controls $\mathscr{P}_{ad}$, we have $\widehat{\mathbf{p}} \in \mathscr{P}_{ad}$, i.e. the limit function $\widehat{\mathbf{p}}$ is an admissible control. Closely following the proof arguments of Lemma 19.2, it can be shown there exists a compact in $L^1(\Omega_T)$ sequence of separating functions $\{\zeta_{*,n}\}_{n\in\mathbb{N}} \subset \Psi_*$ with properties (19.61)–(19.62). By Theorem 19.1 we have

$$\zeta_n \to \widehat{\zeta} \ \text{ strongly in } \ L^2(0,T; L^2(\Omega)) \ \text{ and } \ \widehat{\zeta} \in C([0,T]; L^2(\Omega)). \tag{19.63}$$

Hence, $\zeta_n(t,x) \to \widehat{\zeta}(t,x)$ for all $t \in [0,T]$ and a.e. $x \in \Omega$. Then passing to the limit in (19.61) and in the relation $\zeta_n(0,\cdot) = \zeta_0(\cdot)$, we deduce: $\widehat{\zeta}(0,\cdot) = \zeta_0(\cdot)$ in Ω, and the inequality $\widehat{\zeta}_*(x) \le \widehat{\zeta}(t,x) \le 1$ holds for all $t \in [0,T]$ and a.e. $x \in \Omega$. Thus the limit damage field $\widehat{\zeta} = \widehat{\zeta}(t,x)$ satisfies the conditions (19.50)–(19.51).

In what follows, we note that in view of the boundedness of the sequence $\{(\mathbf{p}_n, \zeta_n, \mathbf{u}_n) \in \Xi\}_{n\in\mathbb{N}}$ there exist constants $C_1 > 0$ and $C_2 > 0$ such that the estimates (19.60) hold true for each pair $(\zeta_n, \mathbf{u}_n)$ with $n \in \mathbb{N}$. Hence, the (M_2)-property implies

$$\sup_{n\in\mathbb{N}} \Big| \big(\phi(\mathbf{e}(\mathbf{u}_n), \zeta_n), \zeta_n - 1\big)_{L^2(\Omega_T)} \Big| \le C_3.$$

Since the set $\big\{\varphi(x)\psi(t) \ \big| \ \forall\, \varphi \in C_0^\infty(\mathbb{R}^N; \Gamma), \ \forall\, \psi \in C_0^\infty(0,T)\big\}$ is dense in $\mathscr{V} \subset \mathscr{Z}$, by the completeness arguments and formula (19.5), we come to the energy identity

$$\begin{aligned} &\|\zeta_n(t) - 1\|^2_{L^2(\Omega)} + \kappa \int_0^t \|\nabla(\zeta_n(s) - 1)\|^2_{L^2;\mathbb{R}^N)} ds \\ &\qquad = \|\zeta_0 - 1\|^2_{L^2(\Omega)} + \int_0^t \int_\Omega \phi(\zeta_n, \mathbf{e}(\mathbf{u}_n))(\zeta_n(s) - 1)\, dx ds \quad \forall\, t \in [0,T]. \end{aligned} \tag{19.64}$$

As a result, following a standard technique (see, for instance, Lions [8]) it can be shown that the sequence $\{\zeta_n\}_{n\in\mathbb{N}}$ is bounded in $\mathscr{W} = \left\{\zeta \,:\, \zeta \in \mathscr{Z},\quad \frac{\partial \zeta}{\partial t} \in \mathscr{Z}'\right\}$. Thus, without lost of generality, we may suppose that for the $\mathscr{Z}$-weak limiting damage field ζ the conditions (19.48) are valid, and

$$\zeta_n' \rightharpoonup \widehat{\zeta}' \quad \text{in} \quad \mathscr{Z}'. \tag{19.65}$$

It remains to show that the triple $(\widehat{\mathbf{p}}, \widehat{\zeta}, \widehat{\mathbf{u}})$ is related by the integral identities (19.49) and (19.50) for all $\boldsymbol{\varphi} \in C_0^\infty(\mathbb{R}^N; S)^N$, $\psi \in C_0^\infty(0, T)$, and $\varphi \in C_0^\infty(\mathbb{R}^N; \Gamma)$. To do so, we note that for every $n \in \mathbb{N}$ the integral identities (19.50) and (19.51) (with $\mathbf{p}_n$, ζ_n, and $\mathbf{u}_n$ instead of $\mathbf{p}$, ζ, and $\mathbf{u}$, respectively), have to fulfil for the test functions $\boldsymbol{\varphi} \in C_0^\infty(\mathbb{R}^N; S)^N$ and $\varphi \in C_0^\infty(\mathbb{R}^N; \Gamma)$. In this case $\mathbf{e}(\boldsymbol{\varphi}) \in C_0^\infty\big(\mathbb{R}^N; S\big)^{\frac{N(N+1)}{2}}$ and $\boldsymbol{\xi}\varphi \in C_0^\infty(\mathbb{R}^N; \Gamma)^{\frac{N(N+1)}{2}}$ for any $\boldsymbol{\xi} \in \mathbb{S}^N$. However, these classes are essentially wider than the space $C_0^\infty(\Omega)^{\frac{N(N+1)}{2}}$ in the definition of the weak convergence in variable space $L^2(\Omega, \zeta_n\, dx)^{\frac{N(N+1)}{2}}$ (see (19.24)). Therefore, in order to pass to the limit in that integral identities as $n \to \infty$, we make use the following trick (see Buttazzo and Kogut [3]).

Let $(\widetilde{\zeta}_n, \widetilde{\mathbf{u}}_n) \in L^2(0, T; H^1_{loc}(\mathbb{R}^N)) \times L^1\big(0, T; \mathscr{W}^{1,1}_{loc}(\mathbb{R}^N; S)\big)$ be an extension of the functions $(\zeta_n, \mathbf{u}_n)$ to the whole of space $\mathbb{R}^N$ such that the sequence $\{(\widetilde{\zeta}_n, \widetilde{\mathbf{u}}_n)\}_{n\in\mathbb{N}}$ satisfies the properties:

$$\widetilde{\zeta}_n \in L^2(0, T; H^1(Q)), \quad \widetilde{\zeta}_n' \in L^2(0, T; (H^1(Q))') \tag{19.66}$$

$$\xi_* \le \widetilde{\zeta}_n \le 1 \text{ a.e. in } \ Q_T := (0, T) \times Q, \tag{19.67}$$

$$\sup_{n\in\mathbb{N}} \Big[\|\widetilde{\zeta}_n\|_{L^2(0,T;H^1(Q))} + \|\widetilde{\mathbf{u}}_n\|_{L^2(0,T;L^2(Q)^N)} + \|\mathbf{e}(\widetilde{\mathbf{u}}_n)\|_{L^2\left(0,T;L^2(Q,\widetilde{\zeta}_n\, dx)^{\frac{N(N+1)}{2}}\right)} \Big] < +\infty \tag{19.68}$$

for any bounded domain Q in $\mathbb{R}^N$. Here $\xi_* \in L^1(Q_T)$ is a non negative function such that $\xi_*^{-1} \in L^1(Q_T)$ and $\xi_*|_{\Omega_T} \in \Psi_*$.

Then by analogy with Lemma 19.2 (see also the property (19.63)) it can be proved that for every bounded domain $Q \subset \mathbb{R}^N$ there exist functions $\widetilde{\zeta} \in L^2(0, T; H^1(Q))$ and $\widetilde{\mathbf{u}} \in W_{\widetilde{\zeta}}(Q \times (0, T); S)$ such that

$$\widetilde{\zeta}_n \rightharpoonup \widetilde{\zeta} \text{ in } L^2(0, T; H^1(Q)), \qquad \widetilde{\mathbf{u}}_n \rightharpoonup \widetilde{\mathbf{u}} \text{ in } L^2\big(0, T; L^2(Q)^N\big), \tag{19.69}$$

$$\zeta_n \to \widehat{\zeta} \ \text{ strongly in } \ L^2(0, T; L^2_{loc}(\mathbb{R}^N)), \tag{19.70}$$

$$\mathbf{e}(\widetilde{\mathbf{u}}_n) \rightharpoonup \mathbf{e}(\widetilde{\mathbf{u}}) \in L^2\big(0, T; L^2(Q, \widetilde{\zeta}\, dx)^{\frac{N(N+1)}{2}}\big) \tag{19.71}$$

in the variable space $L^2\big(0,T;L^2(Q,\widetilde{\zeta}_n\,dx)^{\frac{N(N+1)}{2}}\big)$.

It is important to note that in this case we have

$$\widetilde{\mathbf{u}}=\widehat{\mathbf{u}}\ \text{ and }\ \widetilde{\zeta}=\widehat{\zeta}\quad \text{a.e. in } \Omega_T. \tag{19.72}$$

Taking this fact and (M_1)-property of the source term ϕ into account, we can rewrite the integral identities (19.49)–(19.50) in the equivalent form

$$\int_0^T\int_{\mathbb{R}^N}\big[\widetilde{\zeta}_n(t,x)A(x)\mathbf{e}(\widetilde{\mathbf{u}}_n)\cdot\mathbf{e}(\boldsymbol{\varphi})\big]\psi\,\chi_\Omega(x)\,dxdt=\int_0^T\int_{\mathbb{R}^N}\mathbf{f}\cdot\boldsymbol{\varphi}\psi\,\chi_\Omega(x)\,dxdt$$
$$+\int_0^T\int_\Gamma\mathbf{p}\cdot\boldsymbol{\varphi}\psi\,d\mathscr{H}^{N-1}dt\quad\forall\,\boldsymbol{\varphi}\in C_0^\infty(\mathbb{R}^N;S)^N,\quad\forall\,\psi\in C_0^\infty(0,T), \tag{19.73}$$

$$\big\langle\zeta_n',\varphi\psi\big\rangle_{\mathscr{Z}',\mathscr{Z}}+\kappa\int_0^T\int_{\mathbb{R}^N}\nabla\widetilde{\zeta}_n\cdot\nabla\varphi\,\psi\,\chi_\Omega(x)\,dxdt$$
$$=\int_0^T\int_{\mathbb{R}^N}\widetilde{\phi}(\widetilde{\zeta}_n,\mathbf{e}(\widetilde{\mathbf{u}}_n))\varphi\psi\,\chi_\Omega\,dxdt\quad\forall\,\varphi\in C_0^\infty(\mathbb{R}^N;\Gamma),\quad\forall\,\psi\in C_0^\infty(0,T). \tag{19.74}$$

In what follows, we note that due to the property (19.70) and the continuity of the embedding $L^2(Q_T)\hookrightarrow L^1(Q_T)$ for every bounded $Q\subset\mathbb{R}^N$, we have $\widetilde{\zeta}_n\to\widetilde{\zeta}$ strongly in $L^1(0,T;L^1_{loc}(\mathbb{R}^N))$. Hence

$$\int_0^T\int_{\mathbb{R}^N}\chi_\Omega^2\widetilde{\zeta}_n\,dxdt=\int_0^T\int_{\mathbb{R}^N}\chi_\Omega\widetilde{\zeta}_n\,dxdt$$
$$\longrightarrow\int_0^T\int_{\mathbb{R}^N}\chi_\Omega\widetilde{\zeta}\,dxdt=\int_0^T\int_{\mathbb{R}^N}\chi_\Omega^2\widetilde{\zeta}\,dxdt. \tag{19.75}$$

As follows from convergence properties (19.15) and (19.17), the equality (19.75) implies the strong convergence $\chi_\Omega\to\chi_\Omega$ in the variable space $L^2(0,T;L^2(\mathbb{R}^N,\widetilde{\zeta}_n\,dx))$. Taking this fact, properties (19.65), (19.69) and (19.71), (M_1), and Remark 19.5 into account, we can pass to the limit in (19.73) and (19.74) as $n\to\infty$. As a result, we obtain

$$\int_0^T\int_{\mathbb{R}^N}\big[\widetilde{\zeta}(t,x)A(x)\mathbf{e}(\widetilde{\mathbf{u}})\cdot\mathbf{e}(\boldsymbol{\varphi})\big]\psi\,\chi_\Omega(x)\,dxdt=\int_0^T\int_{\mathbb{R}^N}\mathbf{f}\cdot\boldsymbol{\varphi}\psi\,\chi_\Omega(x)\,dxdt$$
$$+\int_0^T\int_\Gamma\mathbf{p}\cdot\boldsymbol{\varphi}\psi\,d\mathscr{H}^{N-1}dt\quad\forall\,\boldsymbol{\varphi}\in C_0^\infty(\mathbb{R}^N;S)^N,\quad\forall\,\psi\in C_0^\infty(0,T),$$

$$\langle \widehat{\zeta}', \varphi\psi\rangle_{\mathscr{Z}',\mathscr{Z}} + \kappa \int_0^T \int_{\mathbb{R}^N} \nabla\widetilde{\zeta}\cdot\nabla\varphi\,\psi\,\chi_\Omega(x)\,dxdt$$
$$= \int_0^T \int_{\mathbb{R}^N} \widetilde{\phi}(\widetilde{\zeta}, \mathbf{e}(\widetilde{\mathbf{u}}))\varphi\psi\,\chi_\Omega\,dxdt \quad \forall\,\varphi \in C_0^\infty(\mathbb{R}^N; \Gamma), \quad \forall\,\psi \in C_0^\infty(0,T)$$

which, due to the equalities (19.72), are equivalent to

$$\int_0^T \int_\Omega \big[\widehat{\zeta}(t,x)A(x)\mathbf{e}(\widehat{\mathbf{u}})\cdot\mathbf{e}(\boldsymbol{\varphi})\big]\psi\,dxdt = \int_0^T \int_\Omega \mathbf{f}\cdot\boldsymbol{\varphi}\psi\,dxdt$$
$$+ \int_0^T \int_\Gamma \mathbf{p}\cdot\boldsymbol{\varphi}\psi\,d\mathscr{H}^{N-1}dt \quad \forall\,\boldsymbol{\varphi} \in C_0^\infty(\mathbb{R}^N; S)^N, \quad \forall\,\psi \in C_0^\infty(0,T),$$

$$\langle \widehat{\zeta}', \varphi\psi\rangle_{\mathscr{Z}',\mathscr{Z}} + \kappa \int_0^T \int_\Omega \nabla\widehat{\zeta}\cdot\nabla\varphi\,\psi\,dxdt$$
$$= \int_0^T \int_\Omega \phi(\widehat{\zeta}, \mathbf{e}(\widehat{\mathbf{u}}))\varphi\psi\,dxdt \quad \forall\,\varphi \in C_0^\infty(\mathbb{R}^N; \Gamma), \quad \forall\,\psi \in C_0^\infty(0,T).$$

Hence, the pair $(\widehat{\zeta}, \widehat{\mathbf{u}}) \in \mathscr{Z} \times W_{\widehat{\zeta}}(\Omega \times (0,T); S)$ is a weak solution to the initial-boundary value problem (19.37)–(19.44) under $\mathbf{p} = \widehat{\mathbf{p}}$ in the sense of Definition **??**. Thus, the τ-limit triplet $(\widehat{\mathbf{p}}, \widehat{\zeta}, \widehat{\mathbf{u}})$ belongs to set Ξ, and this concludes the proof.

We are now in a position to state the existence of weak optimal solution to the problem (19.53), (19.37)–(19.45).

Theorem 19.3 *Let* $\mathbf{u}_d \in L^2\big(0,T; L^2(\Omega;\mathbb{R}^N)\big)$, $\mathbf{f} \in L^2\big(0,T; L^2(\Omega)^N\big)$, *and* $\zeta_0 \in L^2(\Omega)$ *be given functions. Assume that* $\Xi \neq \emptyset$, *the damage source term* $\phi : \mathbb{S}^N \times \mathbb{R} \to \mathbb{R}$ *possesses the property* $(\mathfrak{M})$, *and the initial damage field* $\zeta_0 : \Omega \to [0,1]$ *satisfies the condition* (19.44). *Then the optimal control problem* (19.53), (19.37)–(19.45) *admits at least one solution* $(\mathbf{p}^0, \zeta^0, \mathbf{u}^0) \in L^2(0,T; H^1(\Omega)) \times \mathscr{W} \times W_{\zeta^0}(\Omega \times (0,T); S)$.

Proof Since the cost functional $I = I(\mathbf{p}, \mathbf{u}, \zeta)$ is bounded below and $\Xi \neq \emptyset$, it provides the existence of a minimizing sequence $\{(\mathbf{p}_n, \zeta_n, \mathbf{u}_n) \in \Xi\}_{n\in\mathbb{N}}$ to the problem (19.53). From the inequality

$$\inf_{(\mathbf{p},\zeta,\mathbf{u})\in\Xi} I(\mathbf{p},\zeta,\mathbf{u}) = \lim_{n\to\infty}\Big[\int_0^T \int_\Omega |\mathbf{u}_n - \mathbf{u}_d|^2_{\mathbb{R}^N}\,dxdt$$
$$+ \int_0^T \int_\Omega |\zeta_n - 1|\,dxdt + \int_0^T \int_\Omega \|\mathbf{e}(\mathbf{u}_n)\|^2_{\mathbb{S}^N}\zeta\,dxdt\Big] < +\infty, \tag{19.76}$$

there is a constant $C > 0$ such that

$$\sup_{n\in\mathbb{N}} \|\mathbf{e}(\mathbf{u}_n)\|_{L^2\left(0,T;L^2(\Omega,\zeta_n dx)^{\frac{N(N+1)}{2}}\right)} \le C, \tag{19.77}$$

$$\sup_{n\in\mathbb{N}} \|\mathbf{u}_n\|_{L^2(0,T;L^2(\Omega)^N)} \le C, \quad \sup_{n\in\mathbb{N}} \|\zeta_n\|_{L^1(0,T;L^1(\Omega))} \le C. \tag{19.78}$$

Since the sequence $\{\zeta_n\}_{n\in\mathbb{N}}$ is restricted by inequalities (19.61), the estimate $(19.78)_2$ implies

$$\sup_{n\in\mathbb{N}} \|\zeta_n\|^2_{L^2(0,T;L^2(\Omega))} \le \sup_{n\in\mathbb{N}} \|\zeta_n\|_{L^1(0,T;L^1(\Omega))} \le C. \tag{19.79}$$

Then, by energy equality (19.64) and (M_2)-property of the source term ϕ, we arrive at the estimate

$$\begin{aligned}
\kappa\|\nabla\zeta_n\|^2_{L^2(0,T;L^2(\Omega)^N)} &\le 2\kappa\int_0^T \|\nabla(\zeta_n-1)\|^2_{L^2;\mathbb{R}^N)}dt + 2\kappa T|\Omega\| \\
&= 2\|\zeta_0-1\|^2_{L^2(\Omega)} + 2\int_0^T\int_\Omega \phi(\zeta_n,\mathbf{e}(\mathbf{u}_n))(\zeta_n-1)\,dxdt + 2\kappa T|\Omega\| \\
&\qquad \text{(by (19.77), (19.79), and property } (\mathrm{M}_2)) \\
&\le 2\|\zeta_0-1\|^2_{L^2(\Omega)} + 2C_3 + 2\kappa T|\Omega| < +\infty.
\end{aligned}$$

Hence, $\sup_{n\in\mathbb{N}}\|\zeta_n\|_{\mathscr{Z}} < +\infty$, and in view of the definition of the class of admissible controls $\mathscr{R}_{ad}$, the minimizing sequence $\{(\mathbf{p}_n,\zeta_n,\mathbf{u}_n)\in\Xi\}_{n\in\mathbb{N}}$ is bounded in the sense of Definition 19.6. Hence, by Lemma 19.2 there exist functions $\mathbf{p}^0\in L^2\big(0,T;L^2(\Gamma)^N\big)$, $\zeta^0\in L^2(0,T;H^1(\Omega))$, and $\mathbf{u}^0\in W_{\zeta^0}(\Omega\times(0,T);S)$ such that, up to a subsequence, $(\mathbf{p}_n,\zeta_n,\mathbf{u}_n)\xrightarrow{\tau}(\mathbf{p}^0,\zeta^0,\mathbf{u}^0)$. Moreover, by Theorem 19.1 we have $\zeta_n\to\zeta^0$ strongly in $L^2(0,T;L^2(\Omega))$. Hence

$$\zeta_n\to\widehat{\zeta} \text{ strongly in } L^1(0,T;L^1(\Omega)). \tag{19.80}$$

Since the set Ξ is sequentially closed with respect to the τ-convergence (see Theorem 19.2), it follows that the τ-limit triplet $(\mathbf{p}^0,\zeta^0,\mathbf{u}^0)$ is an admissible solution to the optimal control problem (19.53), (19.37)–(19.45) (i.e. $(\mathbf{p}^0,\zeta^0,\mathbf{u}^0)\in\Xi$). To conclude the proof it is enough to observe that by properties (19.16) and (19.80), the cost functional I is sequentially lower τ-semicontinuous. Thus

$$I(\mathbf{p}^0,\zeta^0,\mathbf{u}^0)\le\liminf_{n\to\infty} I(\mathbf{p}_n,\zeta_n,\mathbf{u}_n) = \inf_{(\mathbf{p},\zeta,\mathbf{u})\in\Xi} I(\mathbf{p},\zeta,\mathbf{u}).$$

Hence $(\mathbf{p}^0,\zeta^0,\mathbf{u}^0)$ is an optimal solution, and we come to the required conclusion.

References

1. Bouchitte, G., Buttazzo, G.: Characterization of optimal shapes and masses through Monge-Kantorovich equation. J. Eur. Math. Soc. **3**, 139–168 (2001)
2. Buttazzo, G., Varchon, N.: On the optimal reinforcement of an elastic membrane. Riv. Mat. Univ. Parma. **4**(7), 115–125 (2005)
3. Buttazzo, G., Kogut, P.I.: Weak optimal controls in coefficients for linear elliptic problems. Revista Matematica Complutense **24**, 83–94 (2011)
4. Han, W., Sofonea, M.: Quasistatic Contact Problems in Viscoelasticity and Viscoplasticity. American Mathematical Society, Providence, RI (2002)
5. Kogut, P.I., Leugering, G.: Optimal L^1-control in coefficients for Dirichlet elliptic problems: H-optimal solutions. ZAA **31**(1), 31–53 (2011)
6. Kogut, P.I., Leugering, G.: Optimal L^1-control in coefficients for Dirichlet elliptic problems: W-optimal solutions. J. Optim. Theory Appl. **150**(2), 205–232 (2011)
7. Kuttler, K.L.: Quasistatic evolution of damage in an elastic-viscoplastic material. Electron. J. Differ. Eqns. **147**, 1–25 (2005)
8. Lions, J.-L.: Quelques Méthodes de Résolution des Problèms aux Limites Non Linéares. Dunon, Paris (1969)
9. Shillor, M., Sofonea, M., Telega, J.J.: Models and Analysis of Quasistatic Contact. Lecture Notes in Physics, vol. 655. Springer, Berlin (2004)
10. Simon, J.: Compact sets in the space $L^p(0, T; B)$. Ann. Mat. Pura. Appl. **146**, 65–96 (1987)
11. Zhikov, V.V., Pastukhova, S.E.: Homogenization of degenerate elliptic equations. Siberian Math. J. **49**(1), 80–101 (2006)

Chapter 20
On Existence and Attainability of Solutions to Optimal Control Problems in Coefficients for Degenerate Variational Inequalities of Monotone Type

Olga P. Kupenko

Abstract In this chapter we study an optimal control problem for a nonlinear monotone variational inequality with degenerate weight function and with the coefficients which we adopt as controls in $L^{\infty}(\Omega)$. Since these types of variational inequalities can exhibit the Lavrentieff phenomenon, we consider the optimal control problem in coefficients in the so-called class of H-admissible solutions. Using a special version of celebrated Compensated Compactness Lemma and the direct method of Calculus of Variations we discuss the solvability of the above optimal control problem and prove attainability of H-optimal pairs via optimal solutions of some non-degenerate perturbed optimal control problems.

20.1 Introduction

The aim of this chapter is to study optimal control problems (OCPs) associated to nonlinear degenerate elliptic variational inequalities. The control is a matrix of coefficients in the main part of nonlinear elliptic operator. Mainly, we are interested about solvability of degenerate optimal control problems of this type and attainability of H-optimal solutions to degenerate problems via optimal solutions of non-degenerate problems. In particular, since considered degenerate inequalities may exhibit the Lavrentieff phenomena, which leads to non-uniqueness of weak solutions for such objects, we will mostly focus on existence and attainability properties of the so-called H-optimal solutions to the initial OCP.

O. P. Kupenko (✉)
Department of System Analysis and Control, Dnipropetrovsk Mining University, Karl Marks ave., 19, Dnipropetrovsk 49005, Ukraine
e-mail: kogut_olga@bk.ru

O. P. Kupenko
Institute for Applied and System Analysis, National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build 35, Kyiv 03056, Ukraine

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_20,

More precisely, we consider the following OCP

$$I(\mathscr{U}, y) = \int_{\Omega} |y(x) - z_{\partial}(x)|^p \, dx \to \inf, \tag{20.1}$$

$$\mathscr{U} \in M_p^{\alpha,\beta}(\Omega), \;\; y \in K, \tag{20.2}$$

$$\langle -\mathrm{div}\left(\mathscr{U}(x)\rho(x)\,[(\nabla y)^{p-2}]\nabla y\right) + |y|^{p-2}y, v - y\rangle_W \geq \langle f, v - y\rangle_W, \;\; \forall v \in K \tag{20.3}$$

where $[\eta^{p-2}] = \mathrm{diag}\{|\eta_1|^{p-2}, |\eta_2|^{p-2}, \ldots, |\eta_N|^{p-2}\}\ \forall \eta \in \mathbb{R}^N$.

Here, Ω is a bounded open subset of $\mathbb{R}^N$ ($N \geq 1$) with Lipschitz boundary, $\rho > 0$ is a weight function, $z_{\partial} \in L^p(\Omega)$ and $f \in L^q(\Omega)$ are fixed elements, $M_p^{\alpha,\beta}(\Omega) \subset L^{\infty}(\Omega; \mathbb{R}^{N\times N})$ is a class of admissible controls, K is a closed convex subset of W, where $W = W(\Omega, \rho\, dx)$ is a set of functions $y \in W_0^{1,1}(\Omega)$ for which the norm

$$\|y\|_{\rho} = \left(\int_{\Omega}\left(|y|^p + \rho\sum_{i=1}^{N}\left|\frac{\partial y}{\partial x_i}\right|^p\right)dx\right)^{1/p} \tag{20.4}$$

is finite.

Let p be a real number such that $2 \leq p < \infty$ and let q be its conjugate, namely $p^{-1} + q^{-1} = 1$. We say that a weight function $\rho = \rho(x)$ is degenerate in $\mathbb{R}^N$ if

$$\rho(x) > 0 \quad \text{a.e. in } \mathbb{R}^N \quad \text{and} \quad \rho + \rho^{-1/(p-1)} \in L^1_{loc}(\mathbb{R}^N), \tag{20.5}$$

and the sum $\rho + \rho^{-1/(p-1)}$ does not belong to $L^{\infty}(\Omega)$, in general.

Dealing with degenerate problems leads us to the concept of weighted Sobolev spaces such as $W(\Omega, \rho\, dx)$. In general, these spaces are not new in the literature (see [5, 6]). They allow to enlarge the class of boundary value problems and variational inequalities which are solvable by functional-analytical methods. In fact, we consider variational inequality (20.3) with degenerate weight ρ which is not bounded away from zero and infinity but only satisfying local integrability conditions (20.5). Under these assumptions the nonlinear differential operator in (20.3) is not coercive in the classical sense. Here we encounter non-uniqueness of a particular kind: the smooth functions are, in general, not dense in the weighted Sobolev space $W(\Omega, \rho\, dx)$; that is, if $H(\Omega, \rho\, dx)$ is the closure of $C_0^{\infty}(\Omega)$ with respect to the norm (20.4) then $H(\Omega, \rho\, dx) \neq W(\Omega, \rho\, dx)$. In literature this fact is called the Lavrentieff phenomenon and it leads to surprising consequences like non-uniqueness of solutions to problem (20.2)–(20.3) (see [14]) and, hence, to several possible alternative settings of OCPs, depending on the choice of solution space.

As François Murat [12] showed for OCPs in coefficients for elliptic equations, even if the weight function ρ is non-degenerate, such problems have no solution, in general. The main reason of non-existence of optimal solutions is a lack of continuous dependence of solutions for such elliptic equations on controls with respect to the cor-

responding weak topologies in the state and control spaces. To provide solvability for OCPs in coefficients one should either impose certain additional control restrictions (see [4]) or relax the initial optimal control problem (see, for instance, [2]). The same phenomena takes place for variational elliptic inequalities (see [8]).

In view of this we propose to restrict problem (20.1)–(20.3) by introducing some additional control constraints (see, for instance, [7]).

The chapter is organized as follows. Section 20.2 contains some notation and preliminaries. In Sect. 20.3 we introduce additional control constrains like **div**-conditions of a certain type. After that we discuss the classification of admissible solutions to problem (20.1)–(20.3). In particular, we define the class of W-admissible solutions and the class of so-called H-admissible solutions. However, we restrict our analysis with the later one. Section 20.4 contains a refinement of celebrated div-curl lemma of Murat and Tartar for the case of variable Lebesgue and Sobolev spaces. In Sect. 20.5 using the direct method of Calculus of Variations, we prove the existence of H-optimal solutions to the problem (20.1)–(20.3). In Sect. 20.6, we deal with attainability of H-optimal solutions via the optimal solutions to the special perturbed problems for non-degenerate variational inequalities. In applications a degenerate weight ρ occurs as the limit of a sequence of non-degenerate weights ρ_ε for which the corresponding approximate OCP is solvable (see [8]). The results of this section answer the following question: if limit points of the family of admissible solutions $(\mathscr{U}_\varepsilon, y_\varepsilon)$ to the perturbed problems appear to be H-admissible solutions to the original problem (20.1)–(20.3), whether all H-optimal solutions are attainable in this sense? Note that for the above OCP the attainability and approximability questions remain in the focus of attention. In particular, similar questions were raised by Zhikov and Pastukhova in [13, 15] for the degenerate boundary value problems without controls.

The paper contains a brief review of results obtained by the author in [9, 10] with correspondent citations.

20.2 Notation and Preliminaries

Weighted Sobolev spaces. For any subset $E \subset \Omega$ we denote by $|E|$ its N-dimensional Lebesgue measure $\mathscr{L}^N(E)$. The space $W_0^{1,1}(\Omega)$ is the closure of $C_0^\infty(\Omega)$ in the classical Sobolev space $W^{1,1}(\Omega)$. Let ρ be a degenerate weight in the sense of (20.5). For a given $\Omega \subset \mathbb{R}^N$ we associate to this function two weighted Sobolev spaces $W = W(\Omega, \rho\, dx)$ and $H = H(\Omega, \rho\, dx)$, where H is the closure of $C_0^\infty(\Omega)$ in W.

Note that the spaces W and H are reflexive Banach spaces with respect to the norm $\|\cdot\|_\rho$ due to the estimate

$$\int_\Omega |\nabla y|\, dx \le \left(\int_\Omega \rho |\nabla y|_p^p\, dx\right)^{1/p} \left(\int_\Omega \rho^{-1/(p-1)}\, dx\right)^{p/p-1} \le C\|y\|_\rho,$$

where $|\eta|_p = \left(\sum_{k=1}^{N} |\eta_k|^p\right)^{1/p}$ is a Hölder norm of order p in $\mathbb{R}^N$. It is clear that $H \subseteq W$.

Since the smooth functions are in general not dense in the weight Sobolev space W, it follows that $H \neq W$; that is, for a "typical" degenerate weight ρ the identity $W = H$ is not always valid (for the corresponding examples we refer to [14]). However, if ρ is a non-degenerate weight function, that is, ρ is bounded between two positive constants, then it is easy to verify that $W = H = W_0^{1,p}(\Omega)$. We recall that the dual space of H is $H^* = W^{-1,-p/(p-1)}(\Omega, \rho^{-1/(p-1)}\,dx)$ (for more details see [5]).

Remark 20.1 Assume that there exists a value $\nu \in \left(\frac{N}{p}, +\infty\right) \cap \left[\frac{1}{p-1}, +\infty\right)$ such that $\rho^{-\nu} \in L^1(\Omega)$. Then the following result takes place (see [5, pp. 46]): condition $(20.5)_2$ implies that $|||y|||_{\rho,\Omega} = \left[\int_\Omega \sum_{i=1}^N \left|\frac{\partial y}{\partial x_i}\right|^p \rho\,dx\right]^{1/p}$ is a norm of the space $H(\Omega, \rho\,dx)$ which is equivalent to (20.4) and he embedding $H(\Omega, \rho\,dx) \hookrightarrow L^p(\Omega)$ is compact and dense.

Monotone operators. Let α and β be constants such that $0 < \alpha \le \beta < +\infty$. We define $M_p^{\alpha,\beta}(\Omega)$ as a set of all symmetric matrices $\mathscr{U}(x) = \{a_{i\,j}(x)\}_{1\le i,j\le N}$ in $L^\infty(\Omega; \mathbb{R}^{N\times N})$ such that the following conditions of growth, monotonicity, and strong coercivity are fulfilled:

$$|a_{i\,j}(x)| \le \beta \quad \text{a.e. in } \Omega\ \forall\, i,\ j \in \{1, \dots, N\}, \tag{20.6}$$

$$\Big(\mathscr{U}(x)([\zeta^{p-2}]\zeta - [\eta^{p-2}]\eta), \zeta - \eta\Big)_{\mathbb{R}^N} \ge 0 \quad \text{a.e. in } \Omega\ \forall\, \zeta, \eta \in \mathbb{R}^N, \tag{20.7}$$

$$\Big(\mathscr{U}(x)[\zeta^{p-2}]\zeta, \zeta\Big)_{\mathbb{R}^N} = \sum_{i,\,j=1}^{N} a_{ij}(x)|\zeta_j|^{p-2}\,\zeta_j\,\zeta_i \ge \alpha\,|\zeta|_p^p \quad \text{a.e. in } \Omega. \tag{20.8}$$

Remark 20.2 It is easy to see that $M_p^{\alpha,\beta}(\Omega)$ is a nonempty subset of the space $L^\infty(\Omega; \mathbb{R}^{N\times N})$ and its typical representatives are diagonal matrices of the form $\mathscr{U}(x) = \operatorname{diag}\{\delta_1(x), \delta_2(x), \dots, \delta_N(x)\}$, where $\alpha \le \delta_i(x) \le \beta$ a.e. in $\Omega\ \forall\, i \in \{1, \dots, N\}$.

Considered properties of matrices from $M_p^{\alpha,\beta}(\Omega)$ imply the following result.

Lemma 20.1 [9] *For every fixed control $\mathscr{U} \in M_p^{\alpha,\beta}(\Omega)$, the operator $A_{\mathscr{U}} : H \to H^*$ defined as*

$$\langle A_{\mathscr{U}}(y), v\rangle_H = \sum_{i,\,j=1}^{N} \int_\Omega \left(a_{ij}(x)\left|\frac{\partial y}{\partial x_j}\right|^{p-2}\frac{\partial y}{\partial x_j}\right)\frac{\partial v}{\partial x_i}\rho\,dx + \int_\Omega |y|^{p-2} y\,v\,dx,$$

is strictly monotone, coercive and semicontinuous (here by the semicontinuity property we mean that the scalar function $t \to \langle A_{\mathscr{U}}(y+tv), w\rangle_H$ *is continuous for all* $y,\ v,\ w \in H$*).*

Elliptic Variational Inequalities. Following Lions [11], let us cite some well known results concerning solvability, solution uniqueness and smoothness for non-degenerate nonlinear variational inequalities which will be useful in the sequel.

Theorem 20.1 [11, Theorem 8.2] *Let* V *be a Banach space and* $K \subset V$ *be a closed convex subset. Suppose also that* $A : K \to V^*$ *is a nonlinear operator and* $f \in V^*$ *is a given element of the dual space. The following variational problem: to find an element* $y \in K$ *such that*

$$\langle Ay, v-y\rangle_V \geq \langle f, v-y\rangle_V, \quad \forall v \in K, \tag{20.9}$$

admits at least one solution provided the following conditions:

1. *operator* A *is pseudomonotone, i.e. it is bounded and if* $y_k \to y$ *weakly in* V, $y_k,\ y \in K$ *and* $\limsup_{k\to\infty}\langle A(y_k), y_k - y\rangle_V \leq 0$*, then*

$$\liminf_{k\to\infty}\langle A(y_k), y_k - v\rangle_V \geq \langle Ay, y-v\rangle_V,\ \forall v \in V.$$

2. *operator* A *is coercive, i.e. there exists an element* $v_0 \in K$ *such that*

$$\frac{\langle Ay, y-v_0\rangle_V}{\|y\|_V} \to +\infty \text{ as } \|y\|_V \to \infty,\ y \in K$$

Theorem 20.2 [11, Theorem 8.3] *If the operator* $A : K \to V^*$ *in* Theorem 20.1 *is strictly monotone on* K *then variational inequality* (20.9) *admits a unique solution.*

The pseudomonotony property plays the key role in solvability of the problem (20.9). The following result concerns conditions sufficient for fulfillment of this property.

Proposition 20.1 [11, Proposition 2.5] *For a nonlinear operator* $A : V \to V^*$ *the following implication takes place:* A *is a bounded monotone semicontinuous operator* $\Rightarrow$ A *is a pseudomonotone operator.*

Referring to Lions [11], we make use of the following assumptions, necessary for obtaining the main results of the paper.

Hypothesis 1. There exists a reflexive Banach space X such that $X \subset V^*$, the imbedding $X \hookrightarrow V^*$ is continuous, and X is dense in V^*.

Hypothesis 2. There can be found a duality mapping $J : X \to X^*$ such that $\forall\, y \in K, \forall\, \varepsilon > 0$ there exists an $y_\varepsilon \in K$ such that $A(y_\varepsilon) \in X$ and

$$y_\varepsilon + \varepsilon J(A(y_\varepsilon)) = y.$$

Theorem 20.3 [11, Theorem 8.7] *Assume that Hypotheses 1 and 2 hold true*[1]. *Let operator* $A : V \to V^*$ *be monotone, semicontinuous, bounded and satisfy assumption 2 of theorem 20.1. Then for any solution* y *of variational inequality* (20.9) *the inclusion* $Ay \in X$ *takes place provided* $f \in X$.

Smoothing. Throughout the paper ε denotes a small parameter which varies within a strictly decreasing sequence of positive numbers converging to 0. When we write $\varepsilon > 0$, we consider only the elements of this sequence, while writing $\varepsilon \geq 0$, we also consider its limit $\varepsilon = 0$.

Definition 20.1 We say that a weight function ρ with properties (20.5) is approximated by non-degenerate weight functions $\{\rho^\varepsilon\}_{\varepsilon>0}$ on Ω if:

$$\rho^\varepsilon(x) > 0 \quad \text{a.e. in } \Omega, \quad \rho^\varepsilon + \left(\rho^\varepsilon\right)^{-1} \in L^\infty(\Omega), \quad \forall\, \varepsilon > 0, \tag{20.10}$$

$$\rho^\varepsilon \to \rho, \quad \left(\rho^\varepsilon\right)^{-1/(p-1)} \to \rho^{-1/(p-1)} \text{ in } L^1(\Omega) \text{ as } \varepsilon \to 0. \tag{20.11}$$

Remark 20.3 The family $\{\rho^\varepsilon\}_{\varepsilon>0}$ satisfying properties (20.10)–(20.11) is called the non-degenerate perturbation of the weight function ρ.

Examples of such perturbations can be constructed using the classical smoothing. For instance, let Q be some positive compactly supported function such that $Q \in L^\infty(\mathbb{R}^N)$, $\int_{\mathbb{R}^N} Q(x)\,dx = 1$, and $Q(x) = Q(-x)$. Then, for a given weight function $\rho \in L^1_{loc}(\mathbb{R}^N)$, we can take $\rho^\varepsilon = (\rho)_\varepsilon$, where

$$(\rho)_\varepsilon(x) = \frac{1}{\varepsilon^N} \int_{\mathbb{R}^N} Q\left(\frac{x-z}{\varepsilon}\right) \rho(z)\,dz = \int_{\mathbb{R}^N} Q(z)\rho(x+\varepsilon z)\,dz. \tag{20.12}$$

In this case, we say that the perturbation $\{\rho^\varepsilon = (\rho)_\varepsilon\}_{\varepsilon>0}$ of the original degenerate weight function ρ is constructed by the "direct" smoothing scheme.

Lemma 20.2 ([13]) *If* ρ, $\rho^{-1/(p-1)} \in L^1_{loc}(\mathbb{R}^N)$ *then the "direct" smoothing* $\{\rho^\varepsilon = (\rho)_\varepsilon\}_{\varepsilon>0}$ *possesses properties* (20.10)–(20.11).

Radon measures and convergence in variable spaces. By a nonnegative Radon measure on Ω we mean a nonnegative Borel measure which is finite on every compact subset of Ω. The space of all nonnegative Radon measures on Ω will be denoted by $\mathscr{M}_+(\Omega)$. If μ is a nonnegative Radone measure on Ω, we will use $L^r(\Omega, d\mu)$, $1 \leq r \leq \infty$, to denote the usual Lebesgue space with respect to the measure μ with the corresponding norm $\|f\|_{L^r(\Omega,d\mu)} = \left(\int_\Omega |f(x)|^r\,d\mu\right)^{1/r}$.

Let $\{\mu_\varepsilon\}_{\varepsilon>0}$, μ be Radon measures such that $\mu_\varepsilon \overset{*}{\rightharpoonup} \mu$ in $\mathscr{M}_+(\Omega)$; that is,

$$\lim_{\varepsilon\to 0} \int_\Omega \varphi\,d\mu_\varepsilon = \int_\Omega \varphi\,d\mu \quad \forall\, \varphi \in C_0(\mathbb{R}^N), \tag{20.13}$$

[1] (see the example for $V = H^1_0(\Omega)$ and $X = L^2(\Omega)$ in [11, Theorem 8.8.])

where $C_0(\mathbb{R}^N)$ is the space of all compactly supported continuous functions. A typical example of such measures is $d\mu_\varepsilon = \rho^\varepsilon(x)\,dx$, $d\mu = \rho(x)\,dx$, where $0 \le \rho^\varepsilon \rightharpoonup \rho$ in $L^1(\Omega)$. Let us recall the definition and main properties of convergence in the variable L^p-space [14].

1. A sequence $\{v_\varepsilon \in L^p(\Omega, d\mu_\varepsilon)\}$ is called bounded if $\limsup\limits_{\varepsilon\to 0} \int_\Omega |v_\varepsilon|^p\, d\mu_\varepsilon < +\infty$.
2. A bounded sequence $\{v_\varepsilon \in L^p(\Omega, d\mu_\varepsilon)\}$ converges weakly to $v \in L^p(\Omega, d\mu)$ if $\lim_{\varepsilon\to 0} \int_\Omega v_\varepsilon \varphi\, d\mu_\varepsilon = \int_\Omega v\varphi\, d\mu$ for any $\varphi \in C_0^\infty(\Omega)$ and we write $v_\varepsilon \rightharpoonup v$ in $L^p(\Omega, d\mu_\varepsilon)$.
3. The strong convergence $v_\varepsilon \to v$ in $L^p(\Omega, d\mu_\varepsilon)$ means that $v \in L^p(\Omega, d\mu)$ and

$$\lim_{\varepsilon\to 0} \int_\Omega v_\varepsilon z_\varepsilon\, d\mu_\varepsilon = \int_\Omega vz\, d\mu \ \text{ as } \ z_\varepsilon \rightharpoonup z \ \text{ in } \ L^q(\Omega, d\mu_\varepsilon). \tag{20.14}$$

The following convergence properties in variable spaces hold:

(a) *Compactness criterium:* if a sequence is bounded in $L^p(\Omega, d\mu_\varepsilon)$, then this sequence is compact with respect to the weak convergence.
(b) *Property of lower semicontinuity:* if $v_\varepsilon \rightharpoonup v$ in $L^p(\Omega, d\mu_\varepsilon)$, then

$$\liminf_{\varepsilon\to 0} \int_\Omega |v_\varepsilon|^p\, d\mu_\varepsilon \ge \int_\Omega v^p\, d\mu. \tag{20.15}$$

(c) *Criterium of strong convergence:* $v_\varepsilon \to v$ if and only if $v_\varepsilon \rightharpoonup v$ in $L^p(\Omega, d\mu_\varepsilon)$ and

$$\lim_{\varepsilon\to 0} \int_\Omega |v_\varepsilon|^p\, d\mu_\varepsilon = \int_\Omega v^p\, d\mu. \tag{20.16}$$

Concluding this section, we recall some well-known results concerning the convergence in the variable space $L^p(\Omega, \rho^\varepsilon dx)$.

Lemma 20.3 ([14]) *If $\{\rho^\varepsilon\}_{\varepsilon>0}$ is a non-degenerate perturbation of the weight function $\rho(x) \ge 0$, then: (A_1) $(\rho^\varepsilon)^{-1} \to \rho^{-1}$ in $L^q(\Omega, \rho^\varepsilon dx)$. (A_2) $[\,v_\varepsilon \rightharpoonup v$ in $L^p(\Omega, \rho^\varepsilon dx)\,] \Longrightarrow [\,v_\varepsilon \rightharpoonup v$ in $L^1(\Omega)\,]$. (A_3) If a sequence $\{v_\varepsilon \in L^p(\Omega, \rho^\varepsilon dx)\}_{\varepsilon>0}$ is bounded, then the weak convergence $v_\varepsilon \rightharpoonup v$ in $L^p(\Omega, \rho^\varepsilon dx)$ is equivalent to the weak convergence $\rho^\varepsilon v_\varepsilon \rightharpoonup \rho v$ in $L^1(\Omega)$. (A_4) If $a \in L^\infty(\Omega)$ and $v_\varepsilon \rightharpoonup v$ in $L^p(\Omega, \rho^\varepsilon dx)$, then $av_\varepsilon \rightharpoonup av$ in $L^p(\Omega, \rho^\varepsilon dx)$.*

Variable Sobolev spaces. Let $\rho(x)$ be a degenerate weight function and let $\{\rho^\varepsilon\}_{\varepsilon>0}$ be a non-degenerate perturbation of the function ρ in the sense of Definition 20.1. We denote by $H(\Omega, \rho^\varepsilon\, dx)$ the closure of $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_{\rho^\varepsilon}$. Since for every ε the function ρ^ε is non-degenerate, the space $H(\Omega, \rho^\varepsilon\, dx)$ coincides with the classical Sobolev space $W_0^{1,p}(\Omega)$.

Definition 20.2 We say that a sequence $\{y_\varepsilon \in H(\Omega, \rho^\varepsilon dx)\}_{\varepsilon>0}$ converges weakly to an element $y \in W$ as $\varepsilon \to 0$, if the following hold: (i) This sequence is bounded. (ii) $y_\varepsilon \rightharpoonup y$ in $L^p(\Omega)$. (iii) $\nabla y_\varepsilon \rightharpoonup \nabla y$ in $L^p(\Omega, \rho^\varepsilon dx)^N$.

The following result plays an important role in results concerning attainability properties of optimal solutions.

Theorem 20.4 *Let $\rho^\varepsilon = (\rho)_\varepsilon$ be a direct smoothing of a degenerate weight $\rho \in L^1_{loc}(\mathbb{R}^N)$ and let $y^\varepsilon \in H(\Omega, \rho^\varepsilon\, dx)$, $y^\varepsilon \rightharpoonup y$ in $L^p(\Omega)$, $\nabla y^\varepsilon \rightharpoonup v$ in $L^p(\Omega, \rho^\varepsilon\, dx)^N$. Then $y \in H$ and $v = \nabla y$.*

20.3 Setting of the Optimal Control Problem

The OCP, we consider in this paper, is to minimize the discrepancy between a given distribution $z_\partial \in L^p(\Omega)$ and the solution $y = y_{\mathscr{U},f}$ of the degenerate variational inequality by choosing an appropriate matrix $\mathscr{U} \in L^\infty(\Omega; \mathbb{R}^{N\times N})$. In fact, we deal with the minimization problem in the form (20.1)–(20.3).

Definition 20.3 We say that a matrix $\mathscr{U} = [a_{i\,j}]$ is an admissible control to degenerate problem (20.2)–(20.3) if $\mathscr{U} \in U_{ad}$, where the set U_{ad} is defined as follows

$$U_{ad} = \Big\{ \mathscr{U} = [\mathbf{a}_1, \dots, \mathbf{a}_N] \in M_p^{\alpha,\beta}(\Omega) \Big| \quad |\mathrm{div}\ (\rho\, \mathbf{a}_i)| \le \gamma_i,\ \text{a.e. in } \Omega,\ \forall i = 1, \dots, N \Big\}. \tag{20.17}$$

Here, $\gamma = (\gamma_1, \dots, \gamma_N) \in \mathbb{R}^N$ is a strictly positive vector.

In what follows, depending on the choice of solution space, we introduce the main types of solutions to the above elliptic variational inequality.

Definition 20.4 We say that a function $y = y(\mathscr{U}, f) \in K$ is a W-solution to degenerate variational inequality (20.2)–(20.3) if

$$\langle -\mathrm{div}\left(\mathscr{U}(x)\rho(x)\,[(\nabla y)^{p-2}]\nabla y\right) + |y|^{p-2}y, v - y\rangle_W \ge \langle f, v - y\rangle_W, \tag{20.18}$$

holds for any $v \in K$.

Definition 20.5 Let $\widetilde{K}$ be a closure in the space $C_0^\infty(\Omega)$ of the set $K \bigcap C_0^\infty(\Omega)$, supposing this intersection nonempty. We say that a function $y = y(\mathscr{U}, f) \in \widetilde{K}$ is an H-solution to variational inequality (20.2)–(20.3) if

$$\langle -\mathrm{div}\left(\mathscr{U}(x)\rho(x)\,[(\nabla y)^{p-2}]\nabla y\right) + |y|^{p-2}y, v - y\rangle_H \ge \langle f, v - y\rangle_H, \tag{20.19}$$

holds for any $v \in \widetilde{K}$.

Remark 20.4 It is easy to see that the set $\widetilde{K} \subset H$ is closed and convex.

Proposition 20.2 [9] *For every control $\mathscr{U} \in M_p^{\alpha,\beta}(\Omega)$ and every $f \in L^q(\Omega)$ there exists a unique H-solution to degenerate elliptic variational inequality* (20.2)–(20.3).

Remark 20.5 Note that the uniqueness property in Proposition 20.2 immediately follows from strict monotonicity of the operator $A_{\mathscr{U}} : H \to H^*$ (see Lemma 20.1).

Remark 20.6 In a similar manner we can show the existence and uniqueness of W-solution to problem (20.2)–(20.3).

Taking this fact into account we can introduce two sets of admissible pairs to the optimal control problem (20.1)–(20.3), (20.17):

$$\Xi_W = \{(\mathscr{U}, y) \in U_{ad} \times W \mid y \in K, (\mathscr{U}, y) \text{ are related by (20.18)}\} \quad (20.20)$$

$$\Xi_H = \left\{(\mathscr{U}, y) \in U_{ad} \times H \;\middle|\; y \in \widetilde{K}, (\mathscr{U}, y) \text{ are related by (20.19)}\right\}. \quad (20.21)$$

Hence for the given control object described by relations (20.2)–(20.3) with both fixed control constraints ($\mathscr{U} \in U_{ad}$) and fixed cost functional (20.1), we have two different statements of the original optimal control problem, namely

$$\left\langle \inf_{(\mathscr{U}, y) \in \Xi_W} I(\mathscr{U}, y) \right\rangle \text{ and } \left\langle \inf_{(\mathscr{U}, y) \in \Xi_H} I(\mathscr{U}, y) \right\rangle.$$

As a matter of fact, there is no comparison between these problems, in general. Indeed, having assumed that $W \neq H$ for a given degenerate weight function $\rho \geq 0$, we can come to the effect which is usually called the Lavrentieff phenomenon. It means that for some $\mathscr{U} \in U_{ad}$ and $f \in L^q(\Omega)$ an H-solution $y_H(\mathscr{U}, f)$ to problem (20.2)–(20.3) does not coincide with its W-solution $y_W(\mathscr{U}, f)$ [14]. In this paper we deal with H-solutions to problem (20.2)–(20.3).

Remark 20.7 In view of proposition 20.2, the set Ξ_H is always nonempty.

Taking this observation into account, we adopt the following concept.

Definition 20.6 We say that a pair $(\mathscr{U}^0, y^0) \in L^\infty(\Omega; \mathbb{R}^{N \times N}) \times H$ is an H-optimal solution to problem (20.1)–(20.3), (20.17) if $(\mathscr{U}^0, y^0) \in \Xi_H$ and $I(\mathscr{U}^0, y^0) = \inf_{(\mathscr{U}, y) \in \Xi_H} I(\mathscr{U}, y)$.

20.4 Compensated Compactness Lemma in Variable Lebesgue and Sobolev Spaces

Let $\{\rho^\varepsilon\}_{\varepsilon > 0}$ be a non-degenerate perturbation of a weight function ρ. We associate to every ρ^ε the space

$$X(\Omega, \rho^\varepsilon dx) = \left\{\mathbf{f} \in L^q(\Omega, \rho^\varepsilon dx)^N \mid \operatorname{div}\left(\rho^\varepsilon \mathbf{f}\right) \in L^q(\Omega)\right\} \quad \forall\, \varepsilon > 0 \quad (20.22)$$

and endow it with the norm

$$\|\mathbf{f}\|_{X(\Omega,\rho^\varepsilon dx)} = \left(\|\mathbf{f}\|^q_{L^q(\Omega,\,\rho^\varepsilon dx)^N} + \|\mathrm{div}\,\left(\rho^\varepsilon \mathbf{f}\right)\,\|^q_{L^q(\Omega)}\right)^{1/q}.$$

We call a sequence $\{\mathbf{f}_\varepsilon \in X(\Omega,\,\rho^\varepsilon dx)\}_{\varepsilon>0}$ bounded if $\limsup\limits_{\varepsilon\to 0} \|\mathbf{f}_\varepsilon\|_{X(\Omega,\rho^\varepsilon dx)} < +\infty$.

In order to discuss the existence and attainability of H-optimal solutions to the problem (20.1)–(20.3), (20.17), we use the following result (for comparison, we refer the reader to the Compensated Compactness Lemma in [1, 12]).

Lemma 20.4 [9] *Let $\{\rho^\varepsilon\}_{\varepsilon>0}$ be a non-degenerate perturbation of a weight function $\rho(x) > 0$. Let $\left\{\boldsymbol{f}_\varepsilon \in L^q(\Omega,\,\rho^\varepsilon dx)^N\right\}_{\varepsilon>0}$ and $\{g_\varepsilon \in H(\Omega,\,\rho^\varepsilon dx)\}_{\varepsilon>0}$ be such that $\left\{\boldsymbol{f}_\varepsilon\right\}_{\varepsilon>0}$ is bounded in the variable space $X(\Omega,\,\rho^\varepsilon dx)$, $\boldsymbol{f}_\varepsilon \rightharpoonup \boldsymbol{f}$ in $L^q(\Omega,\,\rho^\varepsilon dx)^N$, $\{g_\varepsilon\}_{\varepsilon>0}$ is bounded in the variable space $H(\Omega,\,\rho^\varepsilon dx)$, $g_\varepsilon \rightharpoonup g$ in $L^p(\Omega)$, and $\nabla g_\varepsilon \rightharpoonup \nabla g$ in $L^p(\Omega,\,\rho^\varepsilon dx)^N$. Then*

$$\lim_{\varepsilon\to 0}\int_\Omega \varphi\,\left(\boldsymbol{f}_\varepsilon, \nabla g_\varepsilon\right)_{\mathbb{R}^N}\rho^\varepsilon dx = \int_\Omega \varphi\,(\boldsymbol{f}, \nabla g)_{\mathbb{R}^N}\,\rho\,dx, \quad \forall\,\varphi \in C_0^\infty(\Omega). \quad (20.23)$$

Remark 20.8 As follows from the arguments given above, we can replace the supposition of Lemma 20.4 "let $\{\rho^\varepsilon\}_{\varepsilon>0}$ be a non-degenerate perturbation of a weight function $\rho(x) > 0$" by the following one: "Let $\{\rho^\varepsilon\}_{\varepsilon>0}$ be a sequence with properties: (1) $\rho^\varepsilon(x) > 0$, $\forall\,\varepsilon > 0$; (2) $\rho^\varepsilon \to \rho$, $(\rho^\varepsilon)^{-1/(p-1)} \to \rho^{-1/(p-1)}$ in $L^1(\Omega)$ as $\varepsilon \to 0$; (3) for every $\varepsilon > 0$ the subspace $C_0^\infty(\Omega)$ is dense in $H(\Omega,\,\rho^\varepsilon\,dx)$ with respect to the norm $\|\cdot\|_{\rho^\varepsilon}$".

20.5 Existence of H-Optimal Solutions

Our prime interest of the paper deals with the solvability of OCP (20.1)–(20.3), (20.17) in the class of H-optimal solutions. To this end, we will use the so-called "direct method" in the Calculus of Variations which, roughly speaking, intends to construct a minimizing sequence $\{(\mathscr{U}_k, y_k) \in \Xi_H\}_{k\in\mathbb{N}}$.

First we prove the result concerning topological properties of the set of H-admissible solutions $\Xi_H \subset L^\infty(\Omega;\mathbb{R}^{N\times N}) \times H$. Let τ be the topology on $L^\infty(\Omega;\mathbb{R}^{N\times N}) \times H(\Omega,\,\rho\,dx)$ which we define as the product of the weak-$*$ topology of the space $L^\infty(\Omega;\mathbb{R}^{N\times N})$ and the weak topology of $H(\Omega,\,\rho\,dx)$.

Additional div-constrains put on admissible controls and Compensated Compactness Lemma play the key role in obtaining this result.

Theorem 20.5 [9] *Let $\rho(x) > 0$ be a degenerate weight function and let $\widetilde{K}$ be such that Hypothesis 2 holds true for $X = L^q(\Omega)$. Then for every $f \in L^q(\Omega)$ the set Ξ_H is sequentially τ-closed.*

Theorem 20.6 *Let $\rho(x)$ be a degenerate weight function. Then the set of H-optimal solutions to the problem (20.1)–(20.3), (20.17) is non-empty for every $f \in L^q(\Omega)$.*

Proof First, we note that the cost functional I_Ω is lower τ-semicontinuous on Ξ_H. Let $\{(\mathscr{U}_k, y_k) \in \Xi_H\}_{k\in\mathbb{N}}$ be an H-minimizing sequence to the problem (20.1)–(20.3), (20.17); that is, $\lim_{k\to\infty} I_\Omega(\mathscr{U}_k, y_k) = \inf_{(\mathscr{U},y)\in\Xi_H} I_\Omega(\mathscr{U}, y) < +\infty$. Hence (see (20.1), (20.17)), this sequence is bounded in $L^\infty(\Omega;\mathbb{R}^{N\times N}) \times H$ and we may suppose that, within a subsequence, there exists $(\mathscr{U}^*, y^*) \in L^\infty(\Omega;\mathbb{R}^{N\times N}) \times H$ such that $\mathscr{U}_k \rightharpoonup \mathscr{U}^*$ weakly-$*$ in $L^\infty(\Omega;\mathbb{R}^{N\times N})$, $y_k \rightharpoonup y^*$ in H. Since Ξ_H is sequentially τ-closed, the pair $(\mathscr{U}^*, y^*)$ is H-admissible to the problem (20.1)–(20.3), (20.17). In view of lower τ-semicontinuity of the cost functional I_Ω we obtain that $I_\Omega(\mathscr{U}^*, y^*) \le \liminf_{k\to\infty} I_\Omega(\mathscr{U}_k, y_k) = \inf_{(\mathscr{U},y)\in\Xi_H} I_\Omega(\mathscr{U}, y)$. Hence, $(\mathscr{U}^*, y^*)$ is an H-optimal pair. The proof is complete.

Therefore, considered optimal control problem (20.1)–(20.3) for degenerate elliptic monotone variational inequality is regular in the class of H-admissible solutions. Imposing additional control constrains (20.17) and using the special version of compensated compactness lemma we proved that the set of H-admissible solutions for problem (20.1)–(20.3) is sequentially closed. And using the direct method of Calculus of Variations we proved existence of H-optimal solutions for considered problem.

20.6 Attainability of H-Optimal Solutions

The aim of this section is to propose an appropriate non-degenerate perturbation for the original degenerate OCP (20.1)–(20.3), (20.17) and to show that H-optimal solutions of (20.1)–(20.3), (20.17) can be attained by optimal solutions of perturbed problems. Hereinafter in this section we assume that the set of H-optimal solutions to the problem (20.1)–(20.3), (20.17) is non-empty.

Let ρ be a degenerate weight function with properties (20.5), and let $\{\rho^\varepsilon\}_{\varepsilon>0}$ be a non-degenerate perturbation of ρ in the sense of Definition 20.1.

Definition 20.7 We say that a bounded sequence

$$\left\{(\mathscr{U}_\varepsilon, y_\varepsilon) \in \mathbb{Y}(\Omega, \rho^\varepsilon dx) = L^\infty(\Omega;\mathbb{R}^{N\times N}) \times H(\Omega, \rho^\varepsilon dx)\right\}_{\varepsilon>0}$$

w-converges to $(\mathscr{U}, y) \in L^\infty(\Omega;\mathbb{R}^{N\times N}) \times W$ in the variable space $\mathbb{Y}(\Omega, \rho^\varepsilon dx)$ as $\varepsilon \to 0$ (in symbols, $(\mathscr{U}_\varepsilon, y_\varepsilon) \overset{w}{\rightharpoonup} (\mathscr{U}, y)$), if $\mathscr{U}_\varepsilon \overset{*}{\rightharpoonup} \mathscr{U}$ in $L^\infty(\Omega;\mathbb{R}^{N\times N})$, $y_\varepsilon \rightharpoonup y$ in $L^p(\Omega)$, and $\nabla y_\varepsilon \rightharpoonup \nabla y$ in $L^p(\Omega, \rho^\varepsilon dx)^N$.

Definition 20.8 We say that a minimization problem

$$\left\langle \inf_{(\mathscr{U},y)\in\Xi_H} I(\mathscr{U}, y) \right\rangle \tag{20.24}$$

is a weak variational limit (or variational w-limit) of the sequence

$$\left\{\left\langle \inf_{(\mathscr{U}_\varepsilon, y_\varepsilon)\in \Xi_\varepsilon} I_\varepsilon(\mathscr{U}_\varepsilon, y_\varepsilon)\right\rangle;\quad \Xi_\varepsilon \subset \mathbb{Y}(\Omega, \rho^\varepsilon dx),\quad \varepsilon > 0\right\}, \tag{20.25}$$

with respect to w-convergence in the variable space $\mathbb{Y}(\Omega, \rho^\varepsilon dx)$, if the following conditions are satisfied:

(1) if $\{\varepsilon_k\}$ is a subsequence of $\{\varepsilon\}$ such that $\varepsilon_k \to 0$ as $k \to \infty$, and a sequence $\left\{(\mathscr{U}_k, y_k) \in \Xi_{\varepsilon_k}\right\}_{\varepsilon>0}$ w-converges to a pair $(\mathscr{U}, y)$, then

$$(\mathscr{U}, y) \in \Xi_H;\quad I(\mathscr{U}, y) \le \liminf_{k\to\infty} I_{\varepsilon_k}(\mathscr{U}_k, y_k); \tag{20.26}$$

(2) for every pair $(\mathscr{U}, y) \in \Xi_H$ and any value $\delta > 0$ there exists a realizing sequence $\left\{(\widehat{\mathscr{U}}_\varepsilon, \widehat{y}_\varepsilon) \in \mathbb{Y}(\Omega, \rho^\varepsilon dx)\right\}_{\varepsilon>0}$ such that

$$(\widehat{\mathscr{U}}_\varepsilon, \widehat{y}_\varepsilon) \in \Xi_\varepsilon\ \forall\, \varepsilon > 0,\quad (\widehat{\mathscr{U}}_\varepsilon, \widehat{y}_\varepsilon) \overset{w}{\to} (\widehat{\mathscr{U}}, \widehat{y}), \tag{20.27}$$

$$\|\mathscr{U} - \widehat{\mathscr{U}}\|_{L^\infty(\Omega;\mathbb{R}^{N\times N})} + \|y - \widehat{y}\|_\rho \le \delta,\ \text{ and }\ I(\mathscr{U}, y) \ge \limsup_{\varepsilon\to 0} I_\varepsilon(\widehat{\mathscr{U}}_\varepsilon, \widehat{y}_\varepsilon) - \delta. \tag{20.28}$$

Definition 20.8 is motivated by the following property of variational w-limits (for the details we refer to [3]).

Theorem 20.7 *Assume that* (20.24) *is a weak variational limit of the sequence* (20.25)*, and the constrained minimization problem* (20.24) *has a solution. Suppose* $\left\{(\mathscr{U}^0_\varepsilon, y^0_\varepsilon) \in \Xi_\varepsilon\right\}_{\varepsilon>0}$ *is a sequence of optimal pairs to* (20.25)*. Then there exists a pair* $(\mathscr{U}^0,\ y^0) \in \Xi_H$ *such that* $(\mathscr{U}^0_\varepsilon, y^0_\varepsilon) \overset{w}{\longrightarrow} (\mathscr{U}^0,\ y^0)$*, and*

$$\inf_{(\mathscr{U}, y)\in \Xi_H} I(A, y) = I\left(\mathscr{U}^0,\ y^0\right) = \lim_{\varepsilon\to 0} \inf_{(\mathscr{U}_\varepsilon, y_\varepsilon)\in \Xi_\varepsilon} I_\varepsilon(\mathscr{U}_\varepsilon,\ y_\varepsilon).$$

Let us consider the sequence $\{K_\varepsilon\}_{\varepsilon>0}$ of non-empty closed and convex subsets, which sequentially converges to the set $\widetilde{K}$ in the sense of Kuratovski as $\varepsilon \to 0$ with respect to weak topology of the space $H(\Omega, \rho^\varepsilon dx)$ and let Hypothesis 2 hold true for $X = L^q(\Omega)$ and $V = H(\Omega, \rho^\varepsilon dx)\ \forall\, \varepsilon > 0$. Taking into account Theorem 20.7, we consider the following collection of perturbed OCPs in coefficients for non-degenerate elliptic variational inequalities:

$$\text{Minimize}\quad \left\{I_\varepsilon(\mathscr{U}, y) = \int_\Omega |y(x) - z_\partial(x)|^p\, dx\right\}, \tag{20.29}$$

$$\mathscr{U} \in U^\varepsilon_{ad},\ y \in K_\varepsilon \tag{20.30}$$

$$\langle -\mathrm{div}\left(\rho^\varepsilon \mathscr{U}[(\nabla y)^{p-2}]\nabla y\right) + |y|^{p-2}y, v - y\rangle_{H(\Omega,\rho^\varepsilon dx)} \ge \langle f, v - y\rangle_{H(\Omega,\rho^\varepsilon dx)}\ \forall v \in K_e, \tag{20.31}$$

$$U_{ad}^{\varepsilon} = \Big\{ \mathscr{U} = [\mathbf{a}_1, \ldots, \mathbf{a}_N] \in M_p^{\alpha,\beta}(\Omega) \Big| \\ \big|\mathrm{div}\,\big(\rho^{\varepsilon}\,\mathbf{a}_i\big)\big| \le \gamma_i,\ \text{a.e. in}\ \Omega,\ \forall\, i = 1, \ldots, N \Big\}, \tag{20.32}$$

where the elements $z_\partial \in L^p(\Omega)$, $f \in L^q(\Omega)$ and $\gamma = (\gamma_1, \ldots, \gamma_N) \in \mathbb{R}^N$ are the same as for the original problem (20.1)–(20.3), (20.17). For every $\varepsilon > 0$ we define Ξ_ε as a set of all admissible pairs to the problem (20.29)–(20.32), namely $(\mathscr{U}, y) \in \Xi_\varepsilon$ if and only if the pair $(\mathscr{U}, y)$ satisfies (20.30)–(20.32).

Note that each of perturbed OCPs (20.29)–(20.32) is solvable provided $\{\rho^{\varepsilon}\}_{\varepsilon>0}$ is a non-degenerate perturbation of $\rho \ge 0$ (see [8]).

Remark 20.9 Let us recall that sequential K-upper and K-lower limits of a sequence of sets $\{E_k\}_{k\in\mathbb{N}}$ are defined as follows, respectively:

$$K_s\text{-}\overline{\lim}\ E_k = \{y \in X : \exists \sigma(k) \to \infty,\ \exists y_k \to y,\ \forall k \in \mathbb{N} :\ y_k \in E_{\sigma(k)}\}$$
$$K_s\text{-}\underline{\lim}\ E_k = \{y \in X : \exists y_k \to y,\ \exists k \ge k_0 \in \mathbb{N} :\ y_k \in E_k\}.$$

The sequence $\{E_k\}_{k\in\mathbb{N}}$ sequentially converges in the sense of Kuratovski to the set E (shortly, K_s-converges), if $E = K_s\text{-}\underline{\lim}\ E_k = K_s\text{-}\overline{\lim}\ E_k$.

Two following results give the attainability property of optimal solutions to considered degenerate problem via optimal solutions of perturbed non-degenerate problems. For details see [10].

Lemma 20.5 *Let $\{\rho^{\varepsilon} = (\rho)_{\varepsilon}\}_{\varepsilon>0}$ be a "direct" smoothing of a degenerate weight function $\rho(x) \ge 0$. Let $\{(\mathscr{U}_\varepsilon, y_\varepsilon) \in \Xi_\varepsilon\}_{\varepsilon>0}$ be a sequence of admissible pairs to the problem (20.29)–(20.32). Then there exist a pair $(\mathscr{U}^*, y^*)$ and a subsequence $\big\{(\mathscr{U}_{\varepsilon_k}, y_{\varepsilon_k})\big\}_{k\in\mathbb{N}}$ of $\{(\mathscr{U}_\varepsilon, y_\varepsilon) \in \Xi_\varepsilon\}_{\varepsilon>0}$ such that $(\mathscr{U}_{\varepsilon_k}, y_{\varepsilon_k}) \overset{w}{\to} (\mathscr{U}^*, y^*)$ as $k \to \infty$ and $(\mathscr{U}^*, y^*) \in \Xi_H$.*

As an evident consequence of this lemma and the lower semicontinuity property of the cost functional (20.29) with respect to w-convergence in variable space $\mathbb{Y}(\Omega, \rho^{\varepsilon} dx)$, we have the following conclusion.

Corollary 20.1 *Let $\{\varepsilon_k\}$ be a subsequence of indices $\{\varepsilon\}$ such that $\varepsilon_k \to 0$ as $k \to \infty$, and let $\big\{(\mathscr{U}_k, y_k) \in \Xi_{\varepsilon_k}\big\}_{k\in\mathbb{N}}$ be a sequence of admissible solutions to corresponding perturbed problems (20.29)–(20.32) such that $(\mathscr{U}_k, y_k) \overset{w}{\rightharpoonup} (\mathscr{U}, y)$.*
Then properties (20.26) are valid.

To discuss properties (20.27)–(20.28), we give a result which is reciprocal in some sense to Lemma 20.5.

Lemma 20.6 *Let $\{\rho^{\varepsilon} = (\rho)_{\varepsilon}\}_{\varepsilon>0}$ be a "direct" smoothing of a degenerate weight function $\rho(x) \ge 0$ and let $(\mathscr{U}, y) \in \Xi_H$ be any admissible pair. Then there exists a realizing sequence $\big\{(\widehat{\mathscr{U}}_\varepsilon, \widehat{y}_\varepsilon) \in \mathbb{Y}(\Omega, \rho^{\varepsilon} dx)\big\}_{\varepsilon>0}$ such that*

$$(\widehat{\mathscr{U}}_\varepsilon, \widehat{y}_\varepsilon) \in \Xi_\varepsilon \ \forall\, \varepsilon > 0, \quad \widehat{\mathscr{U}}_\varepsilon \overset{*}{\rightharpoonup} \mathscr{U} \ \ in \ \ L^\infty(\Omega; \mathbb{R}^{N\times N}); \tag{20.33}$$

$$\operatorname{div}\left(\rho^\varepsilon \widehat{\boldsymbol{a}}_{i\,\varepsilon}\right) \rightharpoonup \operatorname{div}\left(\rho\, \boldsymbol{a}_i\right) \ \ in \ \ L^q(\Omega) \quad \forall\, i \in \{1, \ldots, N\}, \tag{20.34}$$

$$\widehat{y}_\varepsilon \to y \ \ strongly \ in \ L^p(\Omega), \ \ \nabla y_\varepsilon \rightharpoonup \nabla y \ in \ L^p(\Omega, \rho^\varepsilon\, dx)^N. \tag{20.35}$$

Corollary 20.2 *Lemma 20.6 implies the equality* $I(\mathscr{U}, y) = \lim_{\varepsilon\to 0} I_\varepsilon(\widehat{\mathscr{U}}_\varepsilon, \widehat{y}_\varepsilon)$.

As an obvious consequence of Definition 20.8, and Lemmas 20.5–20.6 with their Corollaries, we can give the following conclusion.

Theorem 20.8 *Let* $\{\rho^\varepsilon = (\rho)_\varepsilon\}_{\varepsilon>0}$ *be a "direct "smoothing of a degenerate weight function* $\rho(x) > 0$. *Then the minimization problem* (20.1)–(20.3), (20.17) *is a weak variational limit of the sequence* (20.29)–(20.32) *as* $\varepsilon \to 0$ *with respect to the w-convergence in the variable space* $\mathbb{Y}(\Omega, \rho^\varepsilon dx)$.

As follows from results given above, by Lemma 20.6 each optimal solution to the problem (20.1)–(20.3), (20.17) can be attained by admissible solutions to perturbed problems (20.29)–(20.32), however there exists at least one optimal solution $(\mathscr{U}_0, y_0) \in \Xi_H$ which can be attained by optimal solutions to perturbed problems (20.29)–(20.32).

References

1. Briane, M., Casado-Díaz, J.: Two-dimensional div-curl results. Application to the lack of nonlocal effects in homogenization. Com. Part. Differ. Equ. **32**(4–6), 935–969 (2007)
2. Buttazzo, G., Dal Maso, G., Garroni, A., Malusa, A.: On the relaxed formulation of some shape optimization problems. Adv. Math. Sci. Appl. **1**(7), 1–24 (1997)
3. D'Apice, C., De Maio, U., Kogut, P.I.: Suboptimal boundary control for elliptic equations in critically perforated domains. Ann. Inst. H. Poincare' Anal. Non Line'aire. **25**, 1073–1101 (2008)
4. D'Apice, C., De Maio, U., Kogut, O.P.: On shape stability of Dirichlet optimal control problems in coefficients for nonlinear elliptic equations. Adv. Differ. Equ. **15**(7–8), 689–720 (2010)
5. Drabek, P., Kufner, A., Nicolosi, F.: Non linear elliptic equations, singular and degenerate cases. University of West, Bohemia (1996)
6. Heinonen, J., Kilpelainen, T., Martio, O.: Nonlinear Potential Theory of Degenerate Elliptic Equations. Clarendon Press, London (1993)
7. Kapustjan, V.Ye., Kogut, O.P.: Solenoidal controls in coefficients of nonlinear elliptic boundary value problems. Comput. math. **12**(1), 138–143 (2010) [in Russian].
8. Kogut, O.P.: On optimal control problem in coefficients for nonlinear elliptic variational inequalities. Visnik DNU. Ser.: Probl. Math. Model. Differ. Equ. Theo. **3**(8), 86–98 (2011)
9. Kupenko, O.P.: Optimal control problems in coefficients for degenerate variational inequalities of monotone type. I. Existence of optimal solutions. J. Num. Appl. Math. **106**(3), 88–104 (2011)
10. Kupenko, O.P.: Optimal control problems in coefficients for degenerate variational inequalities of monotone type. II. Attainability problem. J. Num. Appl. Math. **107**(1), 15–34 (2012)
11. Lions, J.-L.: Some Methods of Solving Non-Linear Boundary Value Problems. Dunod-Gauthier-Villars, Paris (1969)
12. Murat, F.: Compacité par compensation. Ann. Sc. Norm. Sup. Pisa. **5**, 489–507 (1978)

13. Pastukhova, S.E.: Degenerate equations of monotone type: Lavrentev phenomenon and attainability problems. Sbornik: Math. **198**(10), 1465–1494 (2007)
14. Zhikov, V.V.: Weighted Sobolev spaces. Sbornik: Math. **189**(8), 27–58 (1998)
15. Zhikov, V.V., Pastukhova, S.E.: Homogenization of degenerate elliptic equations. Siberian Math. J. **49**(1), 80–101 (2006)

Chapter 21
Distributed Optimal Control in One Non-Self-Adjoint Boundary Value Problem

V. O. Kapustyan, O. A. Kapustian and O. K. Mazur

Abstract We prove the solvability of the optimal control problem for elliptic equation with nonlocal boundary conditions in a circular sector with terminal quadratic cost functional in the class of distributed controls.

21.1 Introduction

The theory of linear-quadratic optimal control problems for distributed systems is well researched [1, 2]. In many cases the original problem can be decomposed with the help of Fourier method [3–5]. In this chapter we consider the control problem for elliptic equation with non-local boundary conditions in circular sector [6] with terminal quadratic cost functional. This problem does not allow total decomposition and using of L^2-theory. To resolve this problem in the class of distributed controls we use apparatus of specially constructed biorthonormal basis systems of functions [7] and then we analyze the solutions of Fredholm matrix equations.

V. O. Kapustyan (✉) · O. K. Mazur
National Technical University of Ukraine "Kyiv Polytechnic Institute", 37 Prospect Peremogy, Kyiv 03056, Ukraine
e-mail: kapustyanv@ukr.net

O. K. Mazur
e-mail: okmazur@ukr.net

O. A. Kapustian
Taras Shevchenko National University of Kyiv, 64, Volodymyrs'ka Street, Kyiv 01601, Ukraine
e-mail: olena.kap@gmail.com

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_21,

21.2 Setting of the Problem

In a circular sector $Q = \{(r, \theta) | r \in (0, 1),\ \theta \in (0, \pi)\}$ we consider the optimal control problem

$$\begin{cases} \Delta y := \frac{1}{r}\frac{\partial}{\partial r}(r\frac{\partial y}{\partial r}) + \frac{1}{r^2}\frac{\partial^2 y}{\partial \theta^2} = u(r, \theta),\ (r, \theta) \in Q, \\ y(1, \theta) = p(\theta),\ p(0) = 0, \\ y(r, 0) = 0,\ r \in (0, 1), \\ \frac{\partial y}{\partial \theta}(r, 0) = \frac{\partial y}{\partial \theta}(r, \pi),\ r \in (0, 1), \end{cases} \tag{21.1}$$

$$J(y, u) = \|y(\alpha)\|_D^2 dr + \int_0^1 \|u^2(r)\| dr \to \inf, \tag{21.2}$$

where $p \in C^1([0, \pi])$ is given function, $\alpha \in (0, 1)$ is fixed number, $\|\cdot\|_D$ is a norm in $L^2(0, \pi)$, which is equivalent to standard one and is given by the equality

$$\|v\|_D = \left(\sum_{n=1}^{\infty} v_n^2\right)^{1/2},$$

where $\forall n \geq 1$, $v_n = \int_0^\pi v(\theta)\psi_n(\theta)d\theta$, $\psi_0(\theta) = \frac{2}{\pi^2}$, $\psi_{2n}(\theta) = \frac{4}{\pi^2}(\pi - \theta)\sin 2n\theta$, $\psi_{2n-1}(\theta) = \frac{4}{\pi^2}\cos 2n\theta$.

The aim of this paper is to establish classical solvability of the problem (21.1)–(21.2), that is to find optimal one among admissible processes $\{u, y\} \in C(\bar{Q}) \times \left(C(\bar{Q}) \bigcap C^2(Q)\right)$. For the application of the spectral method we use biorthonormal and complete in $L^2(0, \pi)$ well-known Samarsky-Ionkin systems of functions [7]

$$\Psi = \{\psi_n\}_{n=1}^{\infty} \text{ and}$$

$$\Phi = \{\varphi_0(\theta) = \theta,\ \varphi_{2n}(\theta) = \sin 2n\theta,\ \varphi_{2n-1}(\theta) = \theta\cos 2n\theta\}_{n=1}^{\infty}. \tag{21.3}$$

Then $\forall u \in L^2(Q)$

$$u(r, \theta) = \sum_{n=0}^{\infty} u_n(r) \cdot \varphi_n(\theta), \tag{21.4}$$

where $u_n(r) = \int_0^\pi u(r, \theta)\psi_n(\theta)d\theta$. So, we will seek for the solution of the problem (21.1) in form

$$y(r, \theta) = y_0(r)\theta + \sum_{n=1}^{\infty} \left(y_{2n-1}(r)\theta\cos 2n\theta + y_{2n}(r)\sin 2n\theta\right), \tag{21.5}$$

where the functions $\{y_k(r)\}_{k=0}^{\infty}$ are solutions of the system of ordinary differential equations

$$\frac{d}{dr}(r\frac{dy_0}{dr}) = r \cdot u_0(r),\ y_0(1) = p_0, \tag{21.6}$$

$$r \cdot \frac{d}{dr}\left(r \cdot \frac{dy_{2k-1}}{dr}\right) - (2k)^2 y_{2k-1} = r^2 \cdot u_{2k-1}(r),\ y_{2k-1}(1) = p_{2k-1}, \tag{21.7}$$

$$r\frac{d}{dr}\left(r \cdot \frac{dy_{2k}}{dr}\right) - (2k)^2 y_{2k} - 4k \cdot y_{2k-1} = r^2 \cdot u_{2k}(r),\ y_{2k}(1) = p_{2k}, \tag{21.8}$$

where $p_k = \int_0^{\pi} p(\theta) \cdot \psi_k(\theta) d\theta$.

Thus the original problem (21.1)–(21.2) is reduced to the following one: among admissible pairs $\{u_n(r), y_n(r)\}_{n=0}^{\infty}$ of the problem (21.6)–(21.8) one should minimize the cost functional

$$J(y,u) = y_0^2(\alpha) + \int_0^1 u_0^2(r)dr + \sum_{k=1}^{\infty}(y_{2k-1}^2(\alpha) + y_{2k}^2(\alpha) + \\ + \int_0^1 (u_{2k-1}^2(r) + u_{2k}^2(r))dr) = J_0 + \sum_{k=1}^{\infty} J_k. \tag{21.9}$$

Herewith the optimal process $\{\tilde{u}_n(r), \tilde{y}_n(r)\}_{n=0}^{\infty}$ should be such that the formula (21.4) defines function from $C(\bar{Q})$, and the formula (21.5) defines function from $C(\bar{Q}) \bigcap C^2(Q)$.

21.3 Main Results

A structure of the problem (21.6)–(21.8) allows to reduce it to sequence of the following problems:

On the solutions of (21.6) one should minimize the cost functional

$$J_0 = J_0(u_0), \tag{21.10}$$

on the solutions of (21.7), (21.8) one should minimize the cost functional

$$J_k = J_k(u_{2k-1}, u_{2k}),\ k \geq 1. \tag{21.11}$$

For fixed $\{u_k(r)\}_{k=0}^{\infty} \subset C([0,1])$ solutions of the problem (21.6)–(21.8) have form

$$y_0(r) = p_0 - \int_r^1 \left(\frac{1}{s}\int_0^s \xi u_0(\xi)d\xi\right)ds = p_0 + \int_0^1 G_0(r,s)u_0(s)ds, \qquad (21.12)$$

where

$$G_0(r,s) = \begin{cases} s\ln r, & s \in [0,r], \\ s\ln s, & s \in [r,1], \end{cases}$$

$$y_{2k-1}(r) = p_{2k-1}\cdot r^{2k} + \frac{1}{4k}\int_0^1 s\cdot G_k(r,s)u_{2k-1}(s)ds, \qquad (21.13)$$

where

$$G_k(r,s) = \begin{cases} s^{2k}(r^{2k} - r^{-2k}), & s \in [0,r], \\ r^{2k}(s^{2k} - s^{-2k}), & s \in [r,1], \end{cases}$$

$$y_{2k}(r) = p_{2k}\cdot r^{2k} + p_{2k-1}\cdot r^{2k}\cdot \ln r + \frac{1}{4k}\int_0^1 s\cdot G_k(r,s)u_{2k}(s)ds + \\ + \frac{1}{4k}\int_0^1 s\cdot \bar{G}_k(r,s)u_{2k-1}(s)ds, \qquad (21.14)$$

where

$$\bar{G}_k(r,s) = \int_0^1 p^{-1}\cdot G_k(r,p)G_k(p,s)ds$$
$$= \begin{cases} \frac{1}{2k}\left((\frac{s}{r})^{2k} - (rs)^{-2k}\right) + r^{2k}s^{2k}\ln(rs) - (\frac{s}{r})^{2k}\ln(\frac{s}{r}), & s \in [0,r], \\ \frac{1}{2k}\left((\frac{r}{s})^{2k} - (rs)^{-2k}\right) + r^{2k}s^{2k}\ln(rs) - (\frac{r}{s})^{2k}\ln(\frac{r}{s}), & s \in [r,1]. \end{cases}$$

Lemma 21.1 *For any* $k \geq 0$ *the formulas* (21.12)–(21.14) *define the solutions of the problem* (21.6)–(21.8) $y_k \in C([0,1]) \bigcap C^2(0,1)$.

Proof Since y_k are the solutions of the problem (21.6)–(21.8), then it suffices to show that $\forall k \geq 0$ $y_k \in C([0,1])$. We denote $\prod = [0,1]\times[0,1]$. Then $G_0 \in C(\prod)$, $\max\limits_{\prod} |G_0(r,s)| = e^{-1}$, so, $y_0 \in C([0,1])$.

For $k \geq 1$ $G_k \in C(\prod \setminus \{0,0\})$, $\max\limits_{\prod} |G_k(r,s)| \leq 1$, so, $y_{2k-1} \in C([0,1])$. Since $x^k \ln x \in C([0,1])$, $\max\limits_{x\in[0,1]} |x^k \ln x| = e^{-1}\cdot k^{-1}$, then for $\bar{G}_k \in C(\prod \setminus \{0,0\})$ we have: $\max\limits_{\prod} |\bar{G}_k(r,s)| \leq \frac{1}{k}$, so, $y_{2k} \in C([0,1])$. Lemma is proved.

Theorem 21.1 *The problems* (21.10), (21.11) *have the unique solution* $\{\tilde{u}_k\}_{k=0}^{\infty}$, *moreover* $\forall k \geq 0$ $\tilde{u}_k \in C([0,1])$.

Proof From the formulas (21.12)–(21.14) it follows that the functionals $J_0 : L^2(0,1) \mapsto \mathbf{R}$, $J_k : L^2(0,1)\times L^2(0,1) \mapsto \mathbf{R}$ are strictly convex, continuous

and coercive, which means under [1] that the problems (21.10), (21.11) have the unique solution in the spaces $L^2(0,1)$ and $L^2(0,1)\times L^2(0,1)$ correspondingly.

Equating to zero Frechet derivatives of J_0, J_k, we obtain the following Fredholm integral equations:

$$u_0(s) = -\int_0^1 G_0(\alpha,s)G_0(\alpha,p)u_0(p)dp - p_0\cdot G_0(\alpha,s), \tag{21.15}$$

$$\begin{aligned} u_{2k-1}(s) = &-\frac{1}{2}\cdot\frac{1}{(4k)^2}\int_0^1 [\big(s\cdot G_k(\alpha,s)p\cdot G_k(\alpha,p)+s\cdot\bar G_k(\alpha,s)p\cdot\bar G_k(\alpha,p)\big)u_{2k-1}(p)+\\ &+2s\cdot\bar G_k(\alpha,s)p\cdot G_k(\alpha,p)u_{2k}(p)]dp - p_{2k}\alpha^{2k}\frac{1}{4k}s\cdot G_k(\alpha,s) -\\ &-\Big(p_{2k}\alpha^{2k}+p_{2k-1}\alpha^{2k}\ln\alpha\Big)\frac{1}{4k}\cdot s\cdot\bar G_k(\alpha,s), \end{aligned} \tag{21.16}$$

$$\begin{aligned} u_{2k}(s) = &-\frac{1}{2}\cdot\frac{1}{(4k)^2}\int_0^1 [\big(2s\cdot G_k(\alpha,s)p\cdot\bar G_k(\alpha,p)\big)\,u_{2k-1}(p)+\\ &+s\cdot G_k(\alpha,s)p\cdot\bar G_k(\alpha,p)u_{2k}(p)]dp -\\ &-\Big(p_{2k}\alpha^{2k}+p_{2k-1}\alpha^{2k}\ln\alpha\Big)\frac{1}{4k}\cdot s\cdot G_k(\alpha,s). \end{aligned} \tag{21.17}$$

Since $\max\limits_{(p,s)\in\prod}|G_0(\alpha,p)G_0(\alpha,s)| \le e^{-2} < 1$, then the Eq. (21.15) has the unique solution $\tilde u_0\in C([0,1])$.

Put

$$A_k(p,s) = \begin{pmatrix} s\cdot G_k(\alpha,s)p\cdot G_k(\alpha,p)+s\cdot\bar G_k(\alpha,s)p\cdot\bar G_k(\alpha,p) & 2s\cdot\bar G_k(\alpha,s)p\cdot G_k(\alpha,p)\\ 2s\cdot G_k(\alpha,s)p\cdot\bar G_k(\alpha,p) & s\cdot G_k(\alpha,s)p\cdot\bar G_k(\alpha,p)\end{pmatrix},$$

$$f_k(s) = \begin{pmatrix} -p_{2k}\alpha^{2k}\frac{1}{4k}s\cdot G_k(\alpha,s) - \big(p_{2k}\alpha^{2k}+p_{2k-1}\alpha^{2k}\ln\alpha\big)\frac{1}{4k}\cdot s\cdot\bar G_k(\alpha,s)\\ -\big(p_{2k}\alpha^{2k}+p_{2k-1}\alpha^{2k}\ln\alpha\big)\frac{1}{4k}\cdot s\cdot G_k(\alpha,s)\end{pmatrix}.$$

Then from the Eqs. (21.16), (21.17) we have that vector

$$z_k(s) = \begin{pmatrix} u_{2k-1}(s)\\ u_{2k}(s)\end{pmatrix}$$

satisfies the equation

$$z_k(s) = -\frac{1}{2}\cdot\frac{1}{(4k)^2}\int_0^1 A_k(p,s)z_k(p)dp + f_k(s). \tag{21.18}$$

Under estimates from Lemma 21.1 we obtain

$$\max_{\prod} \|A_k(p,s)\| \le 4,\ \max_{s\in[0,1]} \|f_k(s)\| \le \frac{\alpha^{2k-1}}{2k}\big(|p_2k| + |p_{2k-1}|\big).$$

Then $\forall k \ge 1$ the equation (21.18) has the unique solution

$$\tilde{z}_k(s) = \begin{pmatrix} u_{2k-1}(s) \\ u_{2k}(s) \end{pmatrix} \in C([0,1]),$$

herewith $\forall r \in [0,1]$

$$|u_{2k-1}(r)| \le \frac{\alpha^{2k-1}}{k}\big(|p_{2k-1}| + |p_{2k}|\big),\ |u_{2k}(r)| \le \frac{\alpha^{2k-1}}{k}\big(|p_{2k-1}| + |p_{2k}|\big). \quad (21.19)$$

The theorem is proved.

From the estimates (21.19) it follows that the series $\sum\limits_{n=0}^{\infty} \tilde{u}_n(r)\varphi_n(\theta)$ converges uniformly on $\bar{Q}$ and it defines the function $\tilde{u}(r,\theta) \in C(\bar{Q})$ by the formula (21.4).

Theorem 21.2 *Series*

$$\tilde{y}_0(r)\theta + \sum_{n=1}^{\infty}\big(\tilde{y}_{2n-1}(r)\theta \cdot \cos 2n\theta + \tilde{y}_{2n}(r)\sin 2n\theta\big),$$

defines the function $\tilde{y}(r,\theta) \in C(\bar{Q}) \bigcap C^2(Q)$ *by the formula* (21.5), *where* $\{\tilde{y}_n\}_{n=0}^{\infty}$ *are the solutions of the system* (21.6)–(21.8) *with controls* $\{\tilde{u}_n\}_{n=1}^{\infty}$.

Proof By the formulas (21.12)–(21.14) desired series has the form

$$p_0 \cdot \theta + \theta \cdot \int_0^1 G_0(r,s)u_0(s)ds + \sum_{n=1}^{\infty}\Big(p_{2k-1} \cdot r^{2k} \cdot \theta\cos 2n\theta +$$

$$+ \big(p_{2k} \cdot r^{2k} + p_{2k-1} \cdot r^{2k} \cdot \ln r\big)\sin 2n\theta\Big) + \sum_{n=1}^{\infty} \theta\cos 2n\theta \cdot \frac{1}{4n}\int_0^1 sG_n(r,s)\tilde{u}_{2n-1}(s)ds +$$

$$+ \sum_{n=1}^{\infty} \sin 2n\theta\Big(\frac{1}{4n}\int_0^1 sG_n(r,s)\tilde{u}_{2n}(s)ds + \frac{1}{4n}\int_0^1 s\bar{G}_n(r,s)\tilde{u}_{2n-1}(s)ds\Big). \quad (21.20)$$

The functions $r^{2n}\sin 2n\theta$ and $r^{2n}(\ln r \cdot \sin 2n\theta + \theta\cos 2n\theta)$ are harmonic, $p \in C^1([0,\pi])$, $p(0) = 0$, so, from [6] the first series in 21.20) is the function from the class $C(\bar{Q}) \bigcap C^2(Q)$.

From Lemma 21.1 and the estimates (21.19) we have under Weierstrass theorem that $\tilde{y} \in C(\bar{Q})$.

On $\forall[a,b]\times[c,d]\subset(0,1)\times(0,\pi)$ it remains to investigate the uniform convergence of the series from the first and second-order derivatives on r,θ of functions

$$B_n(r,\theta)=\frac{1}{4n}\int\limits_0^1 sG_n(r,s)\tilde{u}_{2n-1}(s)ds\cdot\theta\cos 2n\theta=b_n(r)\cdot\theta\cos 2n\theta,$$

$$C_n(r,\theta)=\frac{1}{4n}\int\limits_0^1 sG_n(r,s)\tilde{u}_{2n}(s)ds\cdot\sin 2n\theta=c_n\cdot\sin 2n\theta,$$

$$D_n(r,\theta)=\frac{1}{4n}\int\limits_0^1 s\bar{G}_n(r,s)\tilde{u}_{2n-1}(s)ds\cdot\sin 2n\theta=d_n\cdot\sin 2n\theta.$$

From the estimates (21.19) we obtain that the series from derivatives $\frac{\partial}{\partial\theta}$, $\frac{\partial^2}{\partial\theta^2}$ converge on $\bar{Q}$ uniformly under Weierstrass theorem.

For $\forall r\in[a,b]$, $\forall n>1$

$$b_n(r)=\frac{1}{4n}\Big((r^{2n}-r^{-2n})\int\limits_0^r s^{2n+1}\tilde{u}_{2n-1}(s)ds+r^{2n}\int\limits_r^1(s^{2n+1}-s^{1-2n})\tilde{u}_{2n-1}(s)ds\Big),$$

$$b_n'(r)=\frac{1}{2}(r^{2n-1}+r^{-2n-1})\int\limits_0^r s^{2n+1}\tilde{u}_{2n-1}(s)ds+ \tag{21.21}$$

$$+\frac{1}{2}r^{2n-1}\int\limits_r^1(s^{2n+1}-s^{1-2n})\tilde{u}_{2n-1}(s)ds,$$

(summands which do not contain integrals are mutually canceled)

$$b_n''(r)=\frac{1}{2}\Big((2n-1)r^{2n-2}+(-2n-2)r^{-2n-2}\Big)\int\limits_0^r s^{2n+1}\tilde{u}_{2n-1}(s)ds+$$

$$+\frac{1}{2}(2n-1)r^{2n-2}\int\limits_r^1(s^{2n+1}-s^{1-2n})\tilde{u}_{2n-1}(s)ds+\tilde{u}_{2n-1}(r). \tag{21.22}$$

Since $\int\limits_0^r s^{2n+1}ds=\frac{r^{2n+1}}{2n+2}$,

$$\int_r^1 (s^{2n+1} - s^{1-2n})ds = -\frac{n}{1-n^2} - \frac{r^{2n+1}}{2n+2} + \frac{r^{2-2n}}{2-2n},$$

then $\exists C_1 > 0$ such that

$$|b_n'(r)| \le \frac{C}{n} \cdot \frac{\alpha^{2n-1}}{n}(|p_{2n-1}| + |p_{2n}|),$$

so, the series $\sum\limits_{n=2}^{\infty} \frac{\partial}{\partial r} B_n(r,\theta)$, $\sum\limits_{n=2}^{\infty} \frac{\partial}{\partial r} C_n(r,\theta)$, $\sum\limits_{n=2}^{\infty} \frac{\partial^2}{\partial r \partial \theta} B_n(r,\theta)$, $\sum\limits_{n=2}^{\infty} \frac{\partial^2}{\partial r \partial \theta} C_n(r,\theta)$ converge uniformly on $[a,b] \times [c,d]$.

From the same estimates $|b_n''(r)| \le C_2 \cdot \frac{\alpha^{2n-1}}{n}(|p_{2n-1}| + |p_{2n}|)$ and, thereby the series $\sum\limits_{n=2}^{\infty} \frac{\partial^2}{\partial r^2} B_n(r,\theta)$, $\sum\limits_{n=2}^{\infty} \frac{\partial^2}{\partial r^2} C_n(r,\theta)$ converge uniformly on $[a,b] \times [c,d]$.

For the function $d_n(r)$ we have $\forall r \in [a,b]$:

$$\begin{aligned}
d_n(r) = {} & \frac{1}{8n^2} r^{-2n} \cdot \int_0^r s^{2n+1} \tilde{u}_{2n-1}(s)ds - \frac{1}{8n^2} \cdot r^{2n} \int_0^r s^{2n+1} \tilde{u}_{2n-1}(s)ds + \\
& + \frac{1}{4n} r^{2n} \int_0^r s^{2n+1} \ln s \tilde{u}_{2n-1}(s)ds + \frac{1}{4n} r^{2n} \ln r \int_0^r s^{2n+1} \tilde{u}_{2n-1}(s)ds - \\
& - \frac{r^{-2n}}{4n} \int_0^r s^{2n+1} \ln s \tilde{u}_{2n-1}(s)ds + \frac{r^{-2n} \ln r}{4n} \int_0^r s^{2n+1} \tilde{u}_{2n-1}(s)ds + \\
& + \frac{1}{8n^2} r^{2n} \int_r^1 s^{-2n+1} \tilde{u}_{2n-1}(s)ds - \frac{1}{8n^2} r^{2n} \int_r^1 s^{2n+1} \tilde{u}_{2n-1}(s)ds + \\
& + \frac{1}{4n} r^{2n} \int_r^1 s^{2n+1} \ln s \tilde{u}_{2n-1}(s)ds + \frac{1}{4n} r^{2n} \ln s \int_r^1 s^{2n+1} \tilde{u}_{2n-1}(s)ds - \\
& - \frac{1}{4n} r^{2n} \ln s \int_r^1 s^{-2n+1} \tilde{u}_{2n-1}(s)ds + \frac{1}{4n} r^{2n} \int_r^1 s^{-2n+1} \ln s \tilde{u}_{2n-1}(s)ds, \\
d_n'(r) = {} & -\frac{1}{4} \frac{1}{n} r^{-2n-1} \int_0^r s^{2n+1} \tilde{u}_{2n-1}(s)ds - \frac{1}{4n} r^{2n-1} \int_0^r s^{2n+1} \tilde{u}_{2n-1}(s)ds + \\
& + \frac{1}{2} r^{2n-1} \int_0^r s^{2n+1} \ln s \tilde{u}_{2n-1}(s)ds + \frac{1}{4n} \left(2nr^{2n-1} \ln r + r^{2n-1}\right) \int_0^r s^{2n+1} \tilde{u}_{2n-1}(s)ds + \\
& + \frac{1}{2} r^{-2n-1} \int_0^r s^{2n+1} \ln s \tilde{u}_{2n-1}(s)ds + \frac{1}{4n} \left(-2nr^{-2n-1} \ln r + r^{-2n-1}\right) \int_0^r s^{2n+1} \tilde{u}_{2n-1}(s)ds +
\end{aligned}$$

$$+\frac{1}{4n}r^{2n-1}\int_r^1 s^{-2n+1}\tilde{u}_{2n-1}(s)ds-\frac{1}{4n}r^{2n-1}\int_r^1 s^{2n+1}\tilde{u}_{2n-1}(s)ds+$$

$$+\frac{1}{2}r^{2n-1}\int_r^1 s^{2n+1}\ln s\tilde{u}_{2n-1}(s)ds+\frac{1}{4n}\left(2nr^{2n-1}\ln r+r^{2n-1}\right)\int_r^1 s^{2n+1}\tilde{u}_{2n-1}(s)ds-$$

$$-\frac{1}{4n}\left(2nr^{2n-1}\ln r+r^{2n-1}\right)\int_r^1 s^{-2n+1}\tilde{u}_{2n-1}(s)ds+\frac{1}{2}r^{2n-1}\int_r^1 s^{-2n+1}\ln s\tilde{u}_{2n-1}(s)ds.$$

Since

$$\int_0^r s^{2n+1}\ln s ds=\frac{1}{2n+1}r^{2n+1}\ln r-\frac{r^{2n+1}}{(2n+1)^2},$$

then $\exists C_2>0$ such that

$$|d_n'(r)|\le\frac{C_2}{n}\cdot\frac{\alpha^{2n-1}}{n}(|p_{2n-1}|+|p_{2n}|),$$

so, the series $\sum_{n=2}^{\infty}\frac{\partial}{\partial r}D_n(r,\theta)$, $\sum_{n=2}^{\infty}\frac{\partial^2}{\partial r\partial\theta}D_n(r,\theta)$ converge uniformly on $[a,b]\times[c,d]$.

It is easy to see that $\exists C_3>0$ such that

$$|d_n''(r)|\le C_3\cdot\frac{\alpha^{2n-1}}{n}(|p_{2n-1}|+|p_{2n}|).$$

Hence, the series $\sum_{n=2}^{\infty}\frac{\partial^2}{\partial r^2}D_n(r,\theta)$ converges uniformly on $[a,b]\times[c,d]$.

Thereby, $\tilde{y}\in C(\bar{Q})\bigcap C^2(Q)$ and Theorem is proved.

Remark 21.1 If $u(r,\theta)\in C(\bar{Q})$ and for some constant $C>0$ $\forall n\ge 1$ $|u_n(r)|\le\frac{C}{n^2}$, then the control u is admissible in the problem (21.1)–(21.2), that is the corresponding function $y(r,\theta)$ from (21.5) defines classical solution of (21.1).

21.4 Conclusions

In this paper we proved a solvability of the optimal control problem on the classical solutions of elliptic boundary value problem in a circular sector with equality of flows on radiuses and equality of the solution on the one from radiuses to zero in distributed control class for quadratic cost functional.

References

1. Lions, J.-L.: Optimal Problem in PDE Systems. Mir, Moscow (1972)
2. Egorov, A.I.: Optimal Control in Heat and Diffusion Processes. Nauka, Moscow (1978)
3. Belozerov, V.E., Kapustyan, V.E.: Geometrical Methods of Modal Control. Naukova Dumka, Kyiv (1999)
4. Kapustyan, V.E.: Optimal stabilization of the solutions of a parabolic boundary-value problem using bounded lumped control. J. Autom. Inf. Sci. **31**(12), 45–52 (1999)
5. Kapustyan, E.A., Nakonechny, A.G.: Optimal bounded control synthesis for a parabolic boundary-value problem with fast oscillatory coefficients. J. Autom. Inf. Sci. **31**(12), 33–44 (1999)
6. Moiseev, E.I., Ambarzumyan, V.E.: About resolvability of non-local boundary-value problem with equality of fluxes. Differ. Equ. **46**(5), 718–725 (2010)
7. Ionkin, N.I.: Solution of boundary-value problem from heat theory with non-classical boundary conditions. Differ. Equ. **13**(2), 294–304 (1977)

Chapter 22
Guaranteed Safety Operation of Complex Engineering Systems

Nataliya D. Pankratova and Andrii M. Raduk

Abstract A system strategy to estimation of guaranteed survivability and safety operation of complex engineering systems (CES) is proposed. The strategy is based on timely and reliable detection, estimation, and forecast of risk factors and, on this basis, on timely elimination of the causes of abnormal situations before failures and other undesirable consequences occur. The principles that underlie the strategy of the guaranteed safety operation of CES provide a flexible approach to timely detection, recognition, forecast, and system diagnostic of risk factors and situations, to formulation and implementation of a rational decision in a practicable time within an irremovable time constraint. The system control of complex objects is realized. The essence of such control is a systemically coordinated evaluation and adjustment of the operational survivability and safety during the functioning process of an object. The diagnostic unit, which is the basis of a safety control algorithm for complex objects in abnormal situations, is developed as an information platform of engineering diagnostics. By force of systematic and continuous evaluation of critical parameters of object's functioning in the real time mode, the reasons, which could potentially cause the object' tolerance failure of the functioning in the normal mode, are timely revealed.

The practice of the last decades of the last century suggests that the risks of man-made and natural disasters with the consequences of regional, national and global scale are continuously increasing [1], that is due to various objective and subjective conditions and factors [2]. Analysis of accidents and catastrophes can identify the most important causes and weaknesses of control principles for survivability and safety of complex engineering objects (CEO). One of such reasons is the peculiarities of the functioning of the diagnostic systems aimed to identify failures and malfunctions. This approach to security precludes a possibility of a priori prevention

N. D. Pankratova (✉) · A. M. Raduk
Institute for Applied System Analysis National Technical University of Ukraine "Kyiv Polytechnic Institute", Peremogy ave., 37, build, 35,
Kyiv 03056, Ukraine
e-mail: natalidmp@gmail.com

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0_22,

of abnormal modes and as a consequence, there is the possibility of its subsequent transition into an accident and catastrophe. Therefore, it is necessary to develop a new strategy to solve security problems of modern CEO for various purposes. Here we propose a strategy that is based on the conceptual foundations of systems analysis, multicriteria estimation and forecasting of risk [3]. The essence of the proposed concept is the replacement of a standard principle of identifying the transition from operational state of the object into inoperable one on the basis of detection of failures, malfunctions, defects, and forecasting the reliability of an object by a qualitatively new principle. The essence of this principle is the timely detection and elimination of the causes of an eventual transition from operational state of the object into inoperable one on the basis of systems analysis of multifactorial risk of abnormal situations, a reliable estimation of margin of permissible risk of different modes of operation of complex technical objects, and forecast the key indicators of the object survivability in a given period of its operation.

22.1 Introduction

The processes of CEO functioning and processes of ensuring their safety are principally different. The first is focused on achieving the main production target of CES, so they are focused on all stages of a product's life cycle. The second is regarded as secondary by the defined category of specialists, because in their view, all the major issues of efficiency and reliability and, consequently, the security of the products are resolved at the stages of its development, refinement, handling, testing. As a result, there are precedents when the developments of goals, objectives and requirements for security and, above all, for a technical diagnostics system have not proper justification. As a consequence, it turns out that the figures and properties of the created security system do not correspond to real necessities of complex objects, which they must satisfy.

Thus, there is a practical necessity to qualitatively change the principles and the structure of operational-capability controls and the safety of modern engineering systems in real conditions of multifactor risk influence. First of all, the control of complex objects should be systemized which means that there should be system coordination of operability control and safety control not merely by the corresponding goals, tasks, resources, and expected results but also, importantly, by the immediacy and effectiveness of interaction in real conditions of abnormal situations. Such coordination should provide immediate and effective interaction between the mentioned control systems. On the one hand, the effectiveness of the safety system should be provided for timely detection of abnormal situations, evaluation of risk degree and level, and the definition of an permissible risk margin during the process of forming recommendations about immediate actions given to the decision maker. On the other hand, the system of operational capability control after receiving a signal about abnormal situations should, in an effective and operative manner, make a complex object ready for an emergency transition to an offline state and should make it possible to

effect this transition within the limits of permissible risk. This can be achieved only under the condition when the system of technical diagnostics fully complies with the timeliness and efficiency of personnel actions in case of emergencies. Namely: Diagnosis should provide such level of completeness, accuracy and timeliness of information about the state and changing of technologically hazardous processes, which will allow staff to prevent the transition of abnormal situation to an accident and catastrophe in time.

It must be noted that the requirement of timeliness is a priority, as the most accurate, most reliable information becomes unnecessary when it comes to staff after an accident or catastrophe. So there is a practical need of systemic coherence of diagnostic rates with the pace of work processes in different modes of complex engineering systems operation. Such coherence can be one of the most important conditions for ensuring the guaranteed security for the objects with increasing the risk [4].

22.2 Information Platform of Engineering Diagnostics of the Complex Object Operation

The strategy of system control of complex objects survivability and safety is realized as an information platform of engineering diagnostics (IPED) of the complex objects The diagnostic unit, which is the basis of a safety control algorithm for complex

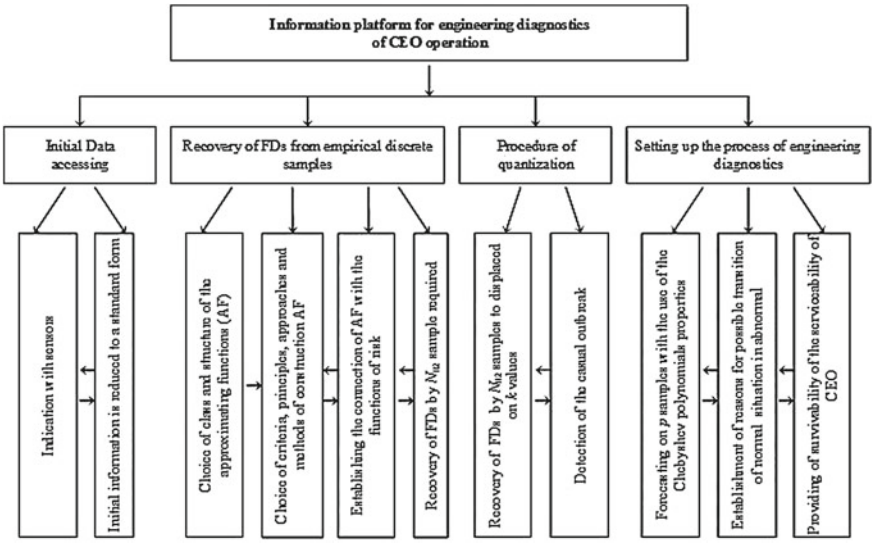


Fig. 22.1 Structural diagram of information platform for engineering diagnostics

objects in abnormal situations, is developed as an IPED (Fig. 22.1). Let us detail some of these modules of the IPED.

Data accessing of the Initial Information during CEO operation. By a CEO we mean an engineering object consisting of several multi-type subsystems that are system-consistent in tasks, problems, resources, and expected results. Each subsystem has functionally interdependent parameters, measured with sensors. With this purpose, groups of sensors are connected to each subsystem, which different parameters (time sampling, resolution, etc.), depending on what there nature is.

The engineering diagnostics during the CEO operation requires samples of size N_{01} and N_{02}, where $N_{01}(N_{01} >> 200)$ is the total sample size during the CEO real-mode operation; $N_{02}(N_{02} << N_{01}; N_{02} = 40 \div 70)$ is the size of the basic sample required to estimate the functional dependences (FD's). The initial information is reduced to a standard form, which makes it possible to form FD's from discrete samples. In view of the proposed methodology, Chebyshev polynomials are taken as basic approximating functions, which normalize all the initial information to the interval [0, 1].

Recovery of Functional Dependences based on Discrete Samples. In the general case, the initial information is specified as a discrete array [5].

$$
\begin{gathered}
M_0 = \langle Y_0, X_1, X_2, X_3 \rangle, \\
Y_0 = (Y_i | i = \overline{1, m}), \;\; Y_i = (Y_i[q_0] | q_0 = \overline{1, k_0}), \\
X_1 = (X_{1j_1} | j_1 = \overline{1, n_1}), \;\; X_{1j_1} = (X_{1j_1}[q_1] | q_1 = \overline{1, k_1}), \\
X_2 = (X_{2j_2} | j_2 = \overline{1, n_2}), \;\; X_{2j_2} = (X_{2j_2}[q_2] | q_2 = \overline{1, k_2}), \\
X_3 = (X_{3j_3} | j_3 = \overline{1, n_3}), \;\; X_{3j_3} = (X_{3j_3}[q_3] | q_3 = \overline{1, k_3})
\end{gathered}
$$

where the set Y_0 determines the numerical values

$$
Y_i[q_0] \Rightarrow \langle X_{1j_1}[q_1], X_{2j_2}[q_2], X_{3j_3}[q_3] \rangle
$$

of the unknown continuous functions $y_i = f_i(x_1, x_2, x_3)$, $i = \overline{1, m}$, $x_1 = (x_{1j_1} \mid j_1 = \overline{1, n_1})$, $x_2 = (x_{2j_2} \mid j_2 = \overline{1, n_2})$, $x_3 = (x_{3j_3} \mid j_3 = \overline{1, n_3})$. To each value of $q_0 \in [1, k_0]$ corresponds a certain set $q_0 \Leftrightarrow (q_1, q_2, q_3)$ of values $q_1 \in [1, k_1]$, $q_2 \in [1, k_2]$, $q_3 \in [1, k_3]$. The set Y_0 consists of k_0 different values $Y_i\,[q_0]$. In the sets X_1, X_2, X_3 a certain part of values $X_{1j_1}\,[q_1]$, $X_{2j_2}\,[q_2]$, $X_{3j_3}\,[q_3]$, for some values $q_1 = \hat{q}_1 \in \hat{Q}_1 \subset [1, k_1]$, $q_2 = \hat{q}_2 \in \hat{Q}_2 \subset [1, k_2]$, $q_3 = \hat{q}_3 \in \hat{Q}_3 \subset [1, k_3]$, repeats each, but there are no completely coinciding sets $\left\langle X_{1j_1}\,[q_1]\,,\; X_{2j_2}\,[q_2]\,,\; X_{3j_3}\,[q_3] \right\rangle$ for different $q_0 \in [1, k_0]$. We have also $n_1 + n_2 + n_3 = n_0, n_0 \leq k_0$. It is known that $x_1 \in D_1$, $x_2 \in D_2$, $x_3 \in D_3, X_1 \in \hat{D}_1$, $X_2 \in \hat{D}_2, X_3 \in \hat{D}_3$, where

$$
D_s = \langle x_{s\,j_s} | d^-_{sj_s} \leq x_{sj_s} \leq d^+_{sj_s}, \;\; j_s = \overline{1, n_s} \rangle, \; s = \overline{1, 3};
$$

$$\hat{D}_s = \langle X_{sj_s} | \hat{d}^-_{sj_s} \le X_{sj_s} \le \hat{d}^+_{sj_s},\ j_s = \overline{1, n_s} \rangle,\ s = \overline{1, 3};$$

$$d^-_{sj_s} \le \hat{d}^-_{sj_s},\ d^+_{sj_s} \ge \hat{d}^+_{sj_s}.$$

It is required to find approximating functions $\Phi_i(x_1, x_2, x_3), i = \overline{1, m}$, that characterize the true functional dependences $y_i = f_i(x_1, x_2, x_3),\ i = \overline{1, m}$, on the set D_s with a practicable error.

Since the initial information is heterogeneous as well as the properties of the groups of factors under study, which are determined, respectively, by the vectors x_1, x_2, x_3, the degree of the influence of each group of factors on the properties of approximating functions should be evaluated independently. With this purpose, the approximating functions are formed as a hierarchical multilevel system of models. At the upper level, the model of determination of the approximating functions dependence on the variables x_1, x_2, x_3 is realized. Such a model in the class of additive functions, where the vectors x_1, x_2, x_3 are independent, is represented as the superposition of functions of the variables x_1, x_2, x_3:

$$\Phi_i(x_1, x_2, x_3) = c_{i1}\Phi_{i1}(x_1) + c_{i2}\Phi_{i2}(x_2) + c_{i3}\Phi_{i3}(x_3), i = \overline{1, m}. \tag{22.1}$$

At the second hierarchical level, models that determine the dependence $\Phi_{is}(s = 1, 2, 3)$ on the components of the variables x_1, x_2, x_3, respectively, and represented as

$$\begin{gathered}\Phi_{i1}(x_1) = \sum_{j_1=1}^{n_1} a^{(1)}_{ij_1}\Psi_{1j_1}(x_{1j_1}),\ \Phi_{i2}(x_2) = \sum_{j_2=1}^{n_2} a^{(2)}_{ij_2}\Psi_{2j_2}(x_{2j_2}),\\ \Phi_{i3}(x_3) = \sum_{j_3=1}^{n_3} a^{(3)}_{ij_3}\Psi_{3j_3}(x_{3j_3}).\end{gathered} \tag{22.2}$$

are formed.

At the third hierarchical level, models that determine the functions $\Psi_{1j_1}, \Psi_{2j_2}, \Psi_{3j_3}$ are formed, choosing the structure and components of the functions $\Psi_{1j_1}, \Psi_{2j_2}, \Psi_{3j_3}$ being the major problem. The structures of these functions are similar to (22.2) and can be represented as the following generalized polynomials:

$$\Psi_{sj_s}(x_{j_s}) = \sum_{p=0}^{P_{j_s}} \lambda_{j_s p}\varphi_{j_s p}(x_{sj_s}), \quad s = 1, 2, 3. \tag{22.3}$$

In some cases, forming the structure of the models, it should be taken into account that the properties of the unknown functions $\Phi_i(x_1, x_2, x_3), i = \overline{1, m}$, are influenced not only by a group of components of each vector x_1, x_2, x_3 but also by the interaction of their components. In such a case, it is expedient to form the dependence of the approximating functions on the variables x_1, x_2, x_3 in a class of multiplicative functions, where the approximating functions are formed by analogy with (22.1)–(22.3) as a hierarchical multilevel system of models

$$[1+\Phi_i(x)]=\prod_{s=1}^{S_0}[1+\Phi_{is}(x_s)]^{c_{is}};\quad [1+\Phi_{is}(x_s)]=\prod_{j_s=1}^{n_s}[1+\Psi_{sj_s}(x_{sj_s})]^{a^s_{ij_s}};$$
$$[1+\Psi_{sj_s}(x_{sj_s})]=\prod_{p=1}^{P_{j_s}}[1+\varphi_{j_sp}(x_{sj_s})]^{\lambda_{j_sp}}. \tag{22.4}$$

The Chebyshev criterion will be used and for the functions φ_{j_sp}, biased Chebyshev polynomials $T_{j_sp}(x_{j_sp})\in[0,1]$ will be used. Then the approximating functions based on the sequence $\Psi_1,\Psi_2,\Psi_3\to\Phi_{i1},\Phi_{i2},\Phi_{i3}\to\Phi_i$ which will allow obtaining the final result by aggregating the corresponding solutions are found. Such an approach reduces the procedure of forming the approximating functions to a sequence of Chebyshev approximation problems for inconsistent systems of linear equations.

Due to the properties of Chebyshev polynomials, the approach to forming the functional dependences makes it possible to extrapolate the approximating functions set up for the intervals $[\hat{d}^-_{j_s},\hat{d}^+_{j_s}]$ to wider intervals $[\hat{d}^-_{j_s},\hat{d}^+_{j_s}]$, which allow forecasting the analyzed properties of a product outside the test intervals.

Quantization of Discrete Numerical Values. The quantization is applied in order to reduce the influence of the measurement error of various parameters on the reliability of the formed solution. The procedure of quantization of discrete numerical values is implemented as follows.

As the base reference statistic for each variable $x_1,\dots,x_n,y_1,\dots,y_m$, the statistic of random samples in these variables of size $N_{01}\geq 200$ is taken.

As the base dynamic statistic in the same variables, the statistic of the sample of the dynamics of the object for the last N_{02} measurements is taken. Therefore, the very first measurement of the original sample should be rejected and measurements should be renumbered in the next measurement $N_{02}+N_2$. Figure 22.2 schematizes the sample for the instant of time $t=t_0$, $N_{02}=40$ and $t=t_0+\Delta t\,(t=1,2,3,\dots,t_k,\dots,T)$.

For the current dynamic parameters, we take the statistics of samples of size $N_{02}+N_2$ biased by N_2 with respect to the statistics of samples of size N_{02}.

Forecasting Nonstationary Processes. The models for forecasting nonstationary processes are based on the original sample of the time series for the initial interval D_0 and base dynamic model of processes (22.1)–(22.3). To this end, we will use the well-known property of Chebyshev polynomials that functions are uniformly approximated on the interval [0, 1]. The essence of the approach is as follows. The initial data are normalized for the interval $D=\{t|t_0^-\leq t\leq t^+\}$, $D=D_0\cup D_0^+$, which includes the initial observation interval $D_0=\{t|t_0^-\leq t\leq t^+\}$ and the prediction interval $D_0^+=\{t|t_0^+<t\leq t^+\}$. Then, to determine the dynamic model of the processes as the estimated approximating functions (22.1) or (22.4), based on the initial data, the system of equations is formed for the interval D_0. The dynamic

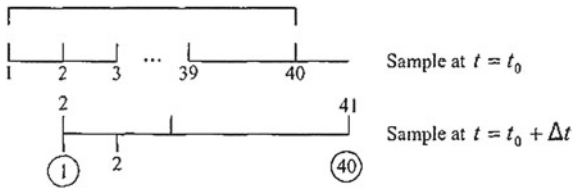


Fig. 22.2 Sample at $t=t_0$ and $t=t_0+\Delta t$

forecasting model is based on the extrapolation of approximating functions for the interval D_0 to the interval D_0^+ [5].

Setting up the Process of Engineering Diagnostics. We will use the system of CES operation models to describe the normal operation mode of the object under the following assumptions and statements.

- Each stage of CEO operation is characterized by the duration and by the initial and final values of each parameter y_i determined at the beginning and the end of the stage, respectively. The variations of y_i within the stage are determined by the corresponding model.
- All the parameters y_i are dynamically synchronous and in phase in the sense that they simultaneously (without a time delay) increase or decrease under risk factors.
- The control $U = (U_j | j = \overline{1, m})$ is inertialless, i.e., there is no time delay between the control action and the object's response.
- The risk factors $\rho_{q_k}^{\tau} | q_k = \overline{1, n_k^{\tau}}$ change the effect on the object in time; the risk increases or decreases with time.
- The control can slow down the influences of risk factors or stop their negative influence on the controlled object if the rate of control exceeds the rate of increase in the influence of risk factors. The negative influence of risk factors is terminated provides the decision making prior implementation to the critical time T_{cr}. At this moment the risk factors cause negative consequences such as an accident or a catastrophe.

To analyze an abnormal mode, let us introduce additional assumptions according to the formation of the model and conditions of recognition of an abnormal situation.

- The risk factors $\rho_{q_k}^{\tau} | q_k = \overline{1, n_k^{\tau}}$ are independent and randomly vary in time with a priori unknown distribution.
- The risk factors can influence on several or all of the parameters y_i simultaneously. A situation of the influence of risk factors is abnormal if at least two parameters y_i are simultaneously changed, without a control, their values are synchronous and are in phase during several measurements (in time).
- The influence of risk factors will be described as a relative change of the level of control. The values of each risk factor are varied discretely and randomly.

Based on acceptable assumptions, let us present additional models and conditions to detect an abnormal situation. Denote by $\tilde{y}_i$ the value of the parameter y_i is influenced by the risk factors; $F_i(\rho_{q_k})$ is the function that takes into account the level of influence of the risk factors on the i parameter y_i; ρ_{q_k} is the value of the q risk factor at the instant of time t_k.

According to item 8, it is assumed that the value of $\tilde{y}_i[t_k]$ at the instant of time t_k is determined by

$$\tilde{y}_i[t_k] = \frac{1}{m} \sum_{j=1}^{m} \tilde{b}_{ij} \sum_{r=0}^{R_j} a_{jr} T_r^*(U_j); \quad \tilde{b}_{ij} = b_{ij} \cdot F_i(\rho_{q_k}), \tag{22.5}$$

where the function $F_i(\rho_{q_k})$ should correspond to the condition where $\tilde{y}_i = y_i$ in the absence of the influence of risk factors (i.e., for $\rho_{q_k} = 0$). Therefore, one of the elementary forms of the function $F_i(\rho_{q_k})$ is

$$F_i(\rho_{q_k}) = 1 - \prod_{q_k=1}^{n_{q_k}} (1 - c_{iq_k}\rho_{q_k}).$$

Note that risk factors can vary in time continuously (for example, pressure continuously changes as an aircraft lifts) or abruptly (for example, during cruise flight at a certain height, pressure may be changed abruptly at the cyclone-anticyclone interface). The most complex is the case where one risk factor varies continuously and others vary abruptly.

We will recognize risk situations by successively comparing $\tilde{y}_i[t_k]$ for $\tilde{y}_i[t_k]$ several successive values of t_k, $k = \overline{1, k_0}$, where $k_0 = 3 \div 7$. As follows from item 2 of the assumptions, the condition of a normal situation is synchronous and in phase changes of $\tilde{y}_i$ for several (in the general case, for all) parameters, whence follows a formula for different instants of time t_k for all of the values of i and for the same instants of time t_k for different values of i (different parameters):

$$\operatorname{sign}\Delta\tilde{y}_i[t_1, t_2] = \ldots = \operatorname{sign}\Delta\tilde{y}_i[t_k, t_{k+1}] = \ldots = \operatorname{sign}\Delta\tilde{y}_i[t_{k_0-1}, t_{k_0}], \tag{22.6}$$

$$\operatorname{sign}\Delta\tilde{y}_1[t_k, t_{k+1}] = \ldots = \operatorname{sign}\Delta\tilde{y}_i[t_k, t_{k+1}] = \ldots = \operatorname{sign}\Delta\tilde{y}_n[t_k, t_{k+1}], i = \overline{1, n}. \tag{22.7}$$

As follows from (22.6) and (22.7), given an abnormal situation on the interval $[t_1, t_{k_0}]$, the following inequalities hold simultaneously:

- the inequality of the signs of increment $\Delta\tilde{y}_i$ for all the adjacent intervals $[t_k, t_{k+1}]$ for $k = \overline{1, k_0}$ for each parameter $\tilde{y}_i$, $i = \overline{1, n}$;
- the inequality of the signs of increment $\tilde{y}_i$, $i = \overline{1, n}$, for all of the parameters $\tilde{y}_i$ for each interval $[t_k, t_{k+1}]$, $k = \overline{1, k_0}$.

Conditions (22.6) and (22.7) are rigid; for practical purposes, it will enough to satisfy the conditions for the representative number (22.3)–(22.5), which determine the parameters $\tilde{y}_i$ but not for all parameters i. The corresponding quantities in (22.6) and (22.7) are defined by

$$\Delta\tilde{y}_i[t_k, t_{k+1}] = \tilde{y}_i[t_{k+1}] - \tilde{y}_i[t_k], \tag{22.8}$$

where $\tilde{y}_i[t_k]$ are defined by (22.5); it is assumed that $\rho_{q_k}[t_{k+1}] > \rho_{q_k}[t_k]$ i.e., the dependence of each risk factor is a function of time, which increases, or $\rho_{q_k}[t_{k+1}] < \rho_{q_k}[t_k]$ i.e., the dependence is a decreasing function.

The practical importance of recognizing an abnormal situation based on (22.6) and (22.7) is in the minor alteration of $\tilde{y}_i[t_k]$ subject to risk factors since the "indicator" of the change is the sign of the difference in (22.6) and (22.7) rather than the

value defined by (22.8). In other words, such an approach is much more sensitive than typical approaches used in diagnostics. Moreover, it allows "filtering" random changes and random measurement errors $\tilde{y}_i$ for separate i according to (22.8) or for individual $[t_k, t_{k+1}]$ according to (22.7).

22.3 Diagnostic of Reanimobile's Functioning

Contensive statement of a problem. The work of reanimobile, which moves in the operational mode, i.e. with the patient on board, is considered. Patient's life is provided with medical equipment, which is powered from the reanimobile's onboard electrical [6].

Basic equipment includes:

- ICE1—basic internal combustion engine (ICE), which causes the car to move and rotate the main generator of G1;
- G1—the main generator, with the capacity of 1.1 kW that generates electricity when the angular velocity of crankshaft rotation is above 220 rad/s (when the speed is above 220 rad/s generator is switched on, when falls down 210 rad/s is off);
- TGB—transmission—gearbox (gear ratio: 1—4.05; 2—2.34; 3—1.39; 4—1; 5—0.85; main transmission—5.125);
- ICE2 and T2—auxiliary engine with a generator power of 1.1 kW, which is used in emergency situations to provide power (standby ICE2 consumes fuel ICE2 0.5 l/h);
- RB—rechargeable battery that provides power to the equipment when the generators do not generate electricity;
- PD—power distribution unit, which provides: battery charge, users' power from one of the generators, or from the battery, or the combination mode.

Tension in the on-board network depends on the generators and the level of battery charge. In the normal mode all equipment power is provided from the main generator and RB.

The main consumers, which are considered during the simulation:

- medical equipment, which consumes about 500 W;
- illumination of the main cabin—120 W;
- outdoor lighting (lights)—110 W;
- car's own needs—100 W.

Charge current is limited at the level that corresponds to the power extracted from the generator, equal to 200 W. Reanimobile must travel a distance of 70 km with a specific schedule of speed, which is formed by road situation.

It is required to ensure electric power for medical equipment, which is located in the main cabin. Since the motion is carried out at night, it is needed to provide additional coverage of the inner and outer. Kinematics parameters approximately correspond to the ambulances, based on GAZ.

Depending on the speed transmission, ratio is changed, therefore, the frequency of crankshaft rotation of the main internal combustion engine is changed (ICE1). At the beginning of the way there are 471 of fuel in the tank. Nutrition ICE1 and ICE2 are from the same tank. In normal situation, the car safely drives patient for 11,700 s (3 h and 15 min). In this case, the battery voltage does not decrease less than 11.85 V. At the end of the way there are 4.11 of fuel in the tank.

Transition into abnormal mode is caused by malfunction of the charger, voltage sensor RB. It is assumed that the sensor gives out false information that the battery is fully charged. Since recharging RB is not done, then with the lapse of time the battery is discharged, and, consequently, the voltage on-board network on the intervals of generator outages (while switching gears, ICE1 is idling) will also be decreased. Due to deep discharge the mode is occurred when the output voltage RB is not enough to maintain the medical equipment operability and this is an emergency situation.

The recognition of an abnormal situation. The recognition of an abnormal situation occurs in accordance with prescribed critical values.

1) For stress in the on-board network: abnormal is 11.7 V, emergency is 10.5 V
2) For the amount of fuel: abnormal is 21, and emergency is 11.
3) For the voltage at the rechargeable battery: an abnormal situation −11.5 V. Thus, while reducing the value of the function below one of the set values, the operation of reanomobile goes to an abnormal mode of functioning.

In other words, if $Y_t <$ H critical exists, at the moment of time t CES functioning goes to an abnormal mode. Where Y_t is a predicted value for the recovered functional dependence. On the diagrams, this process can be observed in the form of decreasing a prediction level (pink curve) below the threshold of the abnormal mode (blue line).

Critical variables:

- Board voltage (depending on the parameters of the RB, the generators condition, the load current). This option could lead directly to an emergency, if the board voltage drops below trip level of medical equipment
- Fuel level depends on the power, which is taken off from the main engine (made in proportion to rotation speed). Decline below a certain point can lead to abnormal (when you can call another car or refueling, and catering equipment from RB) or emergency mode (when the car made a stop for a long time without charging).
- Voltage RB (depending on the generators condition, the total electricity consumption).

Real-time monitoring of the technical diagnostics is conducted in the reanimobile operation process with the purpose of timely exposure of potentially possible abnormal situations and guaranteeing the survivability of the system's functioning. In compliance with the developed methodology of the guaranteed CTO functioning safety at the starting phase $t = t_0$, functional recovery $y_i = f_i(x_1, \ldots, x_j, \ldots)$ is performed using $N_{02} = 50$ given discrete samples of values y_1, y_2, y_3 and their arguments. Here $y_1 = Y_1(x_{11}, x_{12}, x_{13}, x_{14})$, $y_2 = Y_2(x_{21}, x_{22})$, and $y_3 = Y_3(x_{31}, x_{32}, x_{33})$, where x_{11} is the measured voltage RB; x_{12} is the velocity of crankshaft rotation; x_{13} is

power, which is provided by auxiliary generator; x_{14} is the total power consumption; x_{21} is the velocity of crankshaft rotation; x_{22} is power, which is provided by auxiliary generator; x_{31} is the velocity of crankshaft rotation; x_{32} is power, which is provided by auxiliary generator; x_{33} is the total power consumption. All data on the variables Y_i, $i = 1, 2, 3$ and their arguments x_i, $i = 1, 2, 3$ are given as samples during the reanimobile's motion within 50,000 s.

In this case, the voltage sensor gives false information about the voltage RB. When the voltage drops below 11.7 V the diagnostic system provides a driver with the signal about an abnormal situation which can be developed into an emergency. The driver stops the car ($t = 7{,}323$ s), switches on a standby generator ($t = 7{,}414$ s) and eliminates the failure ($t = 7{,}863$ s). Having recharged the battery from a standby generator when $t = 8{,}533$ s, the driver turns off the standby generator and resumes the motion ($t = 8{,}623$ s). Due to low battery, voltage at its terminals starts to decrease rapidly. The diagnostic system warns about abnormal situation again, to solve the problem the driver forcefully supports ICE1 speed at 250 rad/s, thus ensuring continued operation of the main generator.

As a result, fuel consumption is increased, which leads to the abnormal situation ($t = 13{,}000$ s) when the amount of fuel is reduced to 1 l. At this moment of time the car is forcibly stopped by the signal of the diagnostics system (before reaching their destination) and a standby generator is switched on to provide the electric power supply (one liter of fuel is enough for 2 h operation of standby generator that allows refuel the car or call for help).

The Risk Detection Procedure. Taking into account the specifics of operation of the system, following risk detection procedures were constructed.

When reanimobile is functioning, possibility of abnormal situation is calculated with the formula

$$F(\rho_k) = 1 - (1 - \rho_{Gv})(1 - \rho_{Av})(1 - \rho_F),$$

where ρ_{Gv} is the probability that the board voltage drops below the emergency level; ρ_{Av} is the probability that the battery voltage drops below the emergency level; ρ_F is a probability that the fuel level drops below the emergency level. ρ_{Gv}, ρ_{Av} and ρ_F are calculated in the following way:

$$\begin{aligned}
\rho_{Gv} &= 1 - \left|(H_{1es} - y_{1pr})\right| / \left|1,75 * (H_{1es} - H_{1a})\right|; H_{1es} \neq H_{1a};\\
\rho_{Av} &= 1 - \left|(H_{3es} - y_{3pr})\right| / \left|1,75 * (H_{3es} - H_{3a})\right|; H_{3es} \neq H_{3a};\\
\rho_F &= 1 - \left|(H_{2es} - y_{2pr})\right| / \left|1,75 * (H_{2es} - H_{2a})\right|; H_{2es} \neq H_{2a},
\end{aligned}$$

where H_{1es} is board voltage in emergency situations ($Y_{1r} \Leftarrow 11.7$ V); y_{1pr} is the current board voltage (recovery functional dependence using forecast); H_{1a} is board voltage in an emergency ($Y_{1r} \Leftarrow 10.5$ V); H_{2es} is the level of fuel in emergency situations ($Y_{2r} \Leftarrow 1$ L); y_{2pr} is the current value of the fuel (recovery functional dependence using forecast); H_{2a} is the level of fuel in an emergency ($Y_{2r} = 0$); H_{3es} is a battery voltage in the abnormal mode ($Y_{3r} \Leftarrow 11.7$ B); y_{3pr} is the current battery

voltage ((recovery functional dependence using forecast); H_{3a} is a board voltage in an emergency ($Y_{3r} \Leftarrow 10.5$ V).

This structure of risk was taken on the basis of the normalization behavior of the process in the interval (0,1). Create the formula repelled by conditions: the risk during the emergency must be equal to 1, the risk at the border of abnormal mode should be equal to 0.4. In the result, the risks on all fronts are taken into account. The overall risk is 1 during the damage 0.5–0.6 at the border of the abnormal mode.

Some results of reanimobile's functioning during the first 7,000 s. are shown in Fig. 22.3 as the diagrams of stress distribution of the on-board network, the amount of fuel in the tank, the rechargeable battery voltage. The transition into abnormal mode happens due to failure of the sensor battery voltage. So far as the battery recharging is not conducted, the battery is discharged with the lapse of time and, consequently, the voltage in the on-board network in the period of 6,500–7,400 s is also decreased and transits into abnormal mode. The fuel level, which depends on the capacity of the ICE, is also reduced.

At any time of the program operation user has the ability to look at the operator scoreboard (Fig. 22.4), which displays a series of indicators that reflects the character of the state of CEO of the reanimobile functioning. These are such indicators as: indicators of sensors accumulator battery voltage, fuel quantity in the tank, the voltage on-board network, the state of the system, the risk of the damage, the causes of the abnormal or emergency mode, as well as the indicator of the danger level of the system operation and possible failure of sensors.

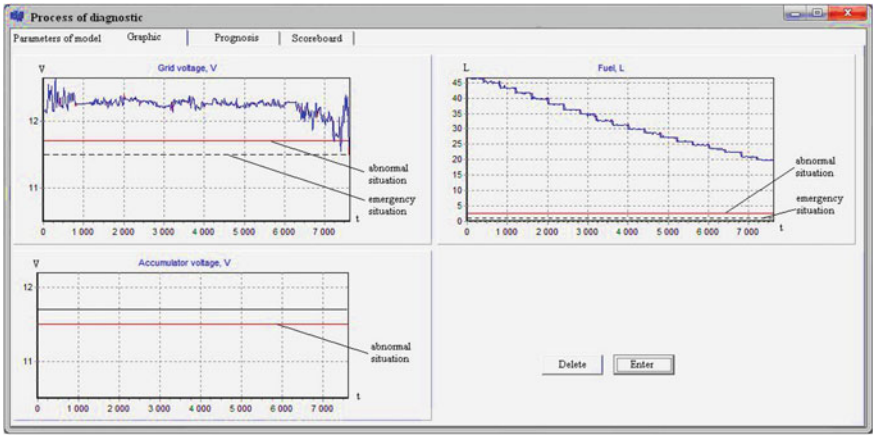
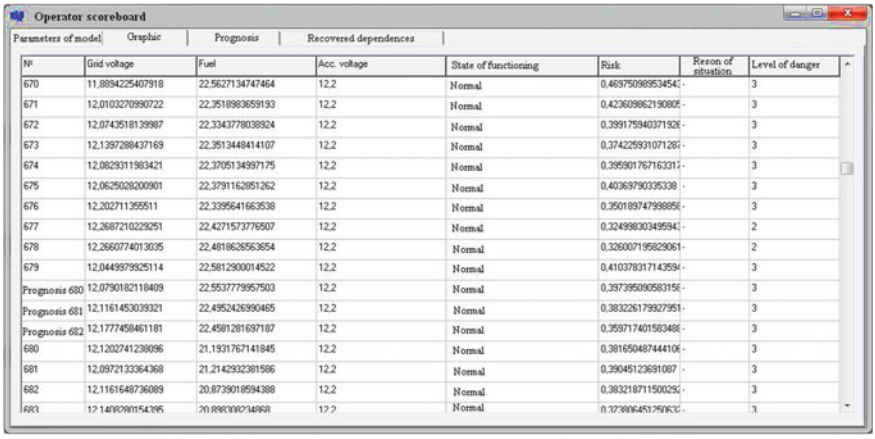
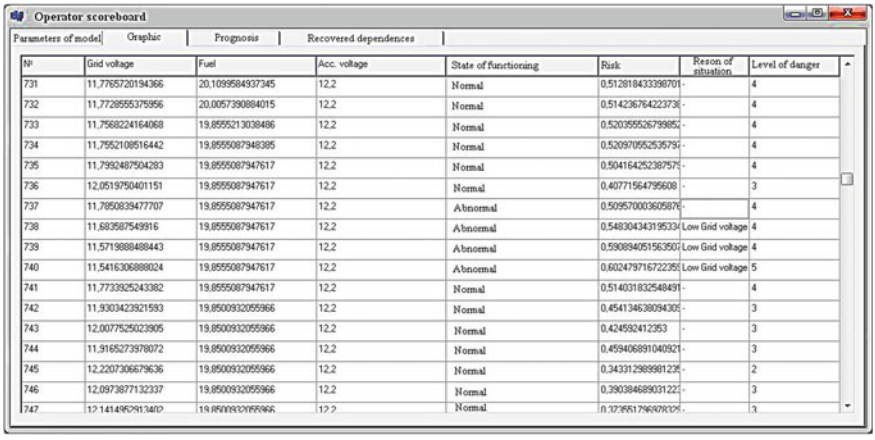


Fig. 22.3 Distribution of the on-board network, the amount of fuel in the tank, the rechargeable battery voltage in accordance of time *t*



№	Grid voltage	Fuel	Acc. voltage	State of functioning	Risk	Reson of situation	Level of danger
670	11,8894225407918	22,5627134747464	12,2	Normal	0,469750989534543		3
671	12,0103270990722	22,3518983659193	12,2	Normal	0,423609862190805		3
672	12,0743518139987	22,3343778038924	12,2	Normal	0,399175940371926		3
673	12,1397288437169	22,3513448414107	12,2	Normal	0,374225931071287		3
674	12,0829311983421	22,3705134997175	12,2	Normal	0,395901767163317		3
675	12,0625028200901	22,3791162851262	12,2	Normal	0,40369790335338		3
676	12,2027113555511	22,3395641663538	12,2	Normal	0,350189747998959		3
677	12,2687210229251	22,4271573776507	12,2	Normal	0,324998303495943		2
678	12,2660774013035	22,4818626563654	12,2	Normal	0,326007195829061		2
679	12,0449979925114	22,5812900014522	12,2	Normal	0,410378317143594		3
Prognosis 680	12,0790182118409	22,5537779957503	12,2	Normal	0,397395090583159		3
Prognosis 681	12,1161453039321	22,4952426990465	12,2	Normal	0,383226179927951		3
Prognosis 682	12,1777458461181	22,4581281697187	12,2	Normal	0,359717401583488		3
680	12,1202741238096	21,1931767141845	12,2	Normal	0,381650487444106		3
681	12,0972133364368	21,2142932381586	12,2	Normal	0,39045123691087		3
682	12,1161648736089	20,8739018594388	12,2	Normal	0,383218711500292		3
683	12,1469891154395	20,8883082344868	12,2	Normal	0,373890451259632		3

Fig. 22.3 (continued)



№	Grid voltage	Fuel	Acc. voltage	State of functioning	Risk	Reson of situation	Level of danger
731	11,7765720194366	20,1099584937345	12,2	Normal	0,512818433398701		4
732	11,7728555375956	20,0057390884015	12,2	Normal	0,514236764223736		4
733	11,7568224164068	19,8555213038486	12,2	Normal	0,520355526799852		4
734	11,7552108516442	19,8555087948385	12,2	Normal	0,520970552535797		4
735	11,7992487504283	19,8555087947617	12,2	Normal	0,504164252387575		4
736	12,0519750401151	19,8555087947617	12,2	Normal	0,40771564795608		3
737	11,7850839477707	19,8555087947617	12,2	Abnormal	0,509570003605876		4
738	11,683587549916	19,8555087947617	12,2	Abnormal	0,548304343195334	Low Grid voltage	4
739	11,5719888488443	19,8555087947617	12,2	Abnormal	0,590894051563507	Low Grid voltage	4
740	11,5416306888024	19,8555087947617	12,2	Abnormal	0,602479716722359	Low Grid voltage	5
741	11,7733925243382	19,8555087947617	12,2	Normal	0,514031832548491		4
742	11,9303423921593	19,8500932055966	12,2	Normal	0,454134638094305		3
743	12,0077525023905	19,8500932055966	12,2	Normal	0,424592412353		3
744	11,9165273978072	19,8500932055966	12,2	Normal	0,459406891040921		3
745	12,2207306679636	19,8500932055966	12,2	Normal	0,343312989981235		2
746	12,0973877132337	19,8500932055966	12,2	Normal	0,390384689031223		3
747	12,1414952913402	19,8500932055966	12,2	Normal	0,373651296878354		3

Fig. 22.4 Scoreboard of diagnostic process

22.4 Conclusion

System coordination of survivability and safety control on the goals, objectives, resources and expected results, as well as by efficiency and effectiveness of interaction in the real conditions of abnormal situations allows to provide the effective and efficient interaction of these control systems. On the one hand, it is ensured the efficiency and effectiveness of security systems according to timely detection of abnormal situations, estimation of its degree and level of risk, definition of the margin of permissible risk in the process of forming the recommendations for the prompt actions of the DM. On the other hand, the survivability control system must effectively and efficiently operate after receiving a signal about the abnormal situation to

ensure the availability of a complex object for the emergency transition into abnormal mode and provide its realization within a margin of permissible risk.

The proposed strategy of system coordination of survivability and safety engineering objects operation, implemented as a tool of information platform of engineering diagnostics of the complex objects, ensures the prevention of inoperability and the danger of object's functioning. By force of systematic and continuous evaluation of critical parameters of object's functioning in the real time mode, the reasons, which could potentially cause the object' tolerance failure of the functioning in the normal mode, are timely revealed. For situations, development of which leads to possible deviations of parameters from the normal mode of the object's functioning, it is possible to make a timely decision about the change of the operation mode of the object, or an artificial correction of the parameters to prevent the transition from the normal mode into the abnormal one, accident and catastrophe.

The principles, which are included in the implementation of the guaranteed safety of CES operation strategy, provide a flexible approach to timely detection, identification, forecasting and system diagnosis of factors and risk situations, formation and implementation of sustainable solutions during the acceptable time within the fatal time limit.

References

1. Frolov, K.V.: Catastrophe Mechanics [in Russian]. Internship Institute for Safety of Complex Engineering System, Moscow (1995)
2. Troshchenko, V.T.: Resistance of Materials to Deformation and Fracture: A Reference Book. Pts. 1, 2 [in Russian]. Naukova Dumka, Kyiv (1993, 1994)
3. Pankratova, N., Kurilin, B.: Conceptual foundations of the system analysis of risks in dynamics of control of complex system safety. P. 1: basic statements and substantiation of approach. Autom. Inform. Sci. **33**(2), 15–31 (2001)
4. Zgurovsky, M.Z., Pankratova, N.D.: System Analysis: Theory and Applications. Springer, Berlin (2007)
5. Pankratova, N.D.: System strategy for guaranteed safety of complex engineering systems. Cybern. Syst. Anal. **46**(2), 243–251 (2010)
6. Raduk, A.M.: System evaluation of the complex technical systems functioning. Syst. Res. Inf. Technol. **1**, 81–94 (2010)

Appendix A
To the Arithmetics of the Bose–Maslov Condensate Statistics

G. I. Arkhipov and V. N. Chubarikov

Abstract The Bose–Maslov condensate statistics is connected with partitions of natural numbers on natural summands. The arithmetics of this phenomenon is discussed. Authors proved that the asymptotical formulae of P. Erdös and J. Lehner for $p_k(n)$ of a form

$$p_k(n) \sim \frac{1}{k!}\binom{n-1}{k-1}$$

is valid uniformly on k for $k \ll n^{1/3}$.

G. I. Arkhipov · V. N. Chubarikov
Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, Russian Federation 119991

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0,

Appendix B
Numerical Algorithms for Multiphase Flows and Applications

Roman Samulyak

Abstract New mathematical models, numerical algorithms, and computational software for the study of multiphase/free surface hydrodynamic and magnetohydrodynamic flows of conducting liquids and partially ionized gases in the presence of phase transitions and external energy sources have been developed. The governing system of equations include a coupled hyperbolic–elliptic system in geometrically complex, evolving domains and equations for phase transitions and external sources. Numerical algorithms use the method of front tracking for material interfaces, high resolution hyperbolic solvers, the embedded boundary method for the elliptic problem in evolving domains, new EOS models, and kinetic models for external sources. They have been implemented as an MHD extension of Frontier, a hydrodynamic code with free interface support. Development of a new MHD code based on smoothed particle hydrodynamics will also be briefly discussed. The software has been applied to a variety of problems including liquid mercury jet targets for future accelerators, pellet fueling of tokamaks, and plasma jet induced magneto inertial fusion (PJMIF). Our main results for the pellet fueling include first calculation of pellet ablation rates in magnetic fields, and studies of the channeling of weakly ionized gases in fusion plasmas by the toroidal magnetic field. 3D simulations of the formation and implosion of plasma liners for the plasma jet induced magneto inertial fusion have been performed. In the PJMIF concept, a plasma liner, formed by merging of a large number of radial, highly supersonic plasma jets, implodes on the target in the form of two compact plasma toroids, and compresses it to conditions of the nuclear fusion ignition. Simulations of accelerator targets explore free surface liquid mercury jets interacting with powerful proton beams in 15 T magnetic fields within the DOE Muon Acceleration Program.

R. Samulyak
Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600, USA
e-mail: rosamu@ams.sunysb.edu

Brookhaven National Laboratory, Computational Science Center, Upton, NY 11973, USA
e-mail: rosamu@bnl.gov

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0,

Appendix C
Singular Trajectories of the First Order in Problems with Multidimensional Control Lying in a Polyhedron

Lion Lokutsievskiy

Abstract In this article control hamiltonian systems are studied. Control is assumed to belong to a polyhedron Ω. Usually, singular trajectories and geometry of their neighbourhoods play the main role in the investigation of global behaviour of trajectories of the system. Theorem about structure of entering (and leaving) the first order singular trajectory in its neighbourhood is proved for holonomic case. In this case the LaGrange surface is woven in a special manner from singular trajectories on facets of the polyhedron Ω. Also a clear method of constructing first order singular trajectories on some facet of Ω is described.

L. Lokutsievskiy
Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, Russian Federation 119991
e-mail: lion.lokut@gmail.com

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0,

Appendix D
The Guaranteed Result Principle in Decision Problems

V. M. Mikhalevich

Abstract The solution of the uncertainty issue for choice problems in Bayesian form requiring a utility function preserving decision and consequence preferences extends to decision problems in generalized neo-Bayesian form allowing randomness to a wide extent for consequences, on the assumption of the utility function's linearity. This solution is based on the transition to multiple choice problems. As a result for multiple decision-making systems, the following models were obtained: non-reducible multi-prior SEU models for choice problems in generalized neo-Bayesian form, which axiomatize the guaranteed and best result principles in statistical form, correspondingly; non-reducible SEU, CEU models and a multi-prior SEU model, all introduced by behaviorist traditions and generalizing the corresponding models by Anscombe–Aumann, Schmeidler and Gilboa–Schmeidler. The indicated models have proof of their corresponding necessary and sufficient criterion replacement conditions.

V.M. Mikhalevich
National University of Kyiv-Mohyla Academy, Skovorody St., 2, Kyiv 04070, Ukraine
e-mail: mih@ukma.kiev.ua

M. Z. Zgurovsky and V. A. Sadovnichiy (eds.), *Continuous and Distributed Systems*, Solid Mechanics and Its Applications 211, DOI: 10.1007/978-3-319-03146-0,