Computational Methods in Applied Sciences

Pekka Neittaanmäki
Sergey Repin
Tero Tuovinen *Editors*

# Mathematical Modeling and Optimization of Complex Structures

ECCOMAS
European Community on Computational Methods in Applied Sciences

Springer

# Computational Methods in Applied Sciences

Volume 40

**Series editor**

E. Oñate
CIMNE
Edificio C-1, Campus Norte UPC
Gran Capitán, s/n
08034 Barcelona, Spain
onate@cimne.upc.edu

More information about this series at http://www.springer.com/series/6899

Pekka Neittaanmäki · Sergey Repin
Tero Tuovinen
Editors

# Mathematical Modeling and Optimization of Complex Structures

Springer

*Editors*
Pekka Neittaanmäki
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä
Finland

Tero Tuovinen
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä
Finland

Sergey Repin
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä
Finland

ISSN 1871-3033
Computational Methods in Applied Sciences
ISBN 978-3-319-23563-9 ISBN 978-3-319-23564-6 (eBook)
DOI 10.1007/978-3-319-23564-6

Library of Congress Control Number: 2015947948

Springer Cham Heidelberg New York Dordrecht London


Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

*Dedicated to Prof. Nikolay Banichuk*
*on the occasion of his 70th birthday*

# Foreword

**Professor Nikolay Banichuk**

Professor Nikolay Banichuk is a member of the Russian Engineering Academy, the International Academy of Astronautics and the National Committee on Theoretical and Applied Mechanics (Russia). He is the Head of the Laboratory of Mechanics and Optimization of Structures at the Institute for Problems in Mechanics (IPM) and Professor of the Moscow Institute of Physics and Technology (MPTI). Professor Banichuk is one of the leading scientists in the modern fields of solid mechanics, optimal structural design, computational mechanics, optimization and variational theory, numerical methods, and computational algorithms. He has written 12 books, 260 scientific articles, and authored more than 200 reports for scientific meetings.

Nikolay Banichuk was born in Komsomolsk-on-Amur-river (East Russia) in 1944, the son of Vladimir Banichuk and Iraida Ivanova. His father was a railway engineer and participated in such famous constructions of railways as Moscow-Peking, Baikal-Amur, Sakhalin, and Stalingrad. For these reasons, Nikolay's family moved around various places, and Nikolay received a variety of impressions and accumulated important experience. In spite of often moving to new locations, Nikolay was a good student and received a classical education.

In 1961, Nikolay Banichuk entered the Moscow Institute of Physics and Technology (MPTI, Aeromechanical Faculty), where he became deeply acquainted with physical and mathematical knowledge. Nikolay combined his studies in MPTI with practice at the Institute for Problems in Mechanics (IPM) and at the Computer Center of Russian Academy of Sciences. During his studies at MPTI, Nikolay participated, under my supervision, in the creation of an effective computational algorithm of local variations and carried out research on elastic-plastic and visco-plastic variational problems with unknown boundaries. In 1967 Banichuk received his diploma of Engineer-Physicist-Researcher from MPTI and continued his investigations as a postgraduate student and researcher under my supervision.

Two years later, Nikolay Banichuk defended his dissertation devoted to numerical solution of nonlinear problems with unknown boundaries arising in mechanics of contact interaction, in deformation of nonelastic material, and in

**Fig. 1** Prof. Nikolay Banichuk

fracture mechanics with curvilinear cracks and earned a doctoral degree in Physico-Mathematical Sciences from IPM.

In 1969, Nikolay entered the IPM in the position of a "junior scientific researcher." Among his first tasks was the optimal design of structures interacting with moving gas or fluids. He also initiated pioneering studies on applications of game theory, especially those of differential games to the problems of the structural optimization with uncertainties. He started his teaching career as a Lecturer at the Aerophysics and Applied Mathematics Faculty of MPTI. During these years, his research in the area of structural mechanics and optimization became well known, and in 1979 he defended his dissertation on the shape optimization for elastic bodies and received his second scientific degree (doctor habitaliation) in Phys-Math from IPM.

After defending his dissertation, Nikolay accepted an invitation from the famous mechanician of the twentieth century, Alexander Ishlinsky (Director of IPM now the Institute carries his name), to occupy a position of the head of laboratory, to formulate its scientific thematics, and form the team of the laboratory. In this connection, Ishlinsky recommended the formation of a laboratory collective from the young scientists and mainly from personal students, thus allowing the team to grow in depth but not in extent. From this time onwards Banichuk, as the head of the laboratory and then as the head of the department, very closely interacted with Ishlinsky. Banichuk began concentrating on the development of analytical, computational, and experimental methods for problems of analysis and design of large space structures. He has obtained important results for large space flexible deployable antenna reflectors. For the obtained results, Banichuk was awarded the Gagarin's medal (twice) and the Korolev medal, and he was elected to the International Academy of Astronautics, at first as a correspondent member and then as a full member (academician).

The first seminar around mechanics and optimization of structures was organized by Nikolay Banichuk in IPM in 1980 and attracted many promising students. Starting from 1981, as a professor of MPTI and Moscow Aviation Technology Institute, he delivered lectures devoted to applied mathematics and mechanics, including numerical analysis and optimization theory. He was a supervisor for 21 academic dissertations. He devoted about 20 years to attestation and qualification activities as a member and vice-chairman of the governmental highest attestation commission on mathematics and mechanics.

Banichuk's engineering activities spread to engineering construction in large protection systems, to the earth reflector and structural problems for new aircraft. He was elected as Academician of Russian Engineering Academy, and then as Academician-Secretary of Russian Engineering Academy and the member of its presidium.

International scientific cooperation plays an important role in Professor Banichuk's activity. The most fruitful relations he has are those with scientists from Finland (Jyväskylä), Italy (Cagliary), Germany (Hannover, Braunschweig), Portugal (Lisbon), Denmark (Lyngby), USA (Iowa City), Netherlands (Delft), and UK (London), and many others. In these places, he has received prestigious

scientific grants and delivered invited lectures. He has served also as chairman and as member of organizing committees in many international conferences.

In 1968, Nikolay met his wife, Natalia Evgenievna Shinaeva, a most gracious and lovely lady. Four years later, his son Alexey was born. Now Nikolay is a grandpa to his 16-year-old grandchild. His sister, Natalia Vladimirovna, also became a mathematician.

The science and engineering community looks forward to many more years of Nikolay's active participation, his leadership, and his continued contributions to science and engineering. More importantly, we, his friends, look forward to many, many years of his congenial and helpful personality, his ever-smiling and energetic face, his cautious wisdom, his tremendous sense of humor, and the sheer enjoyment of being with, and learning from a most charming and amazing gentleman!

Moscow
June 2015

Prof. Felix L. Chernousko
Member of the Russian Academy of Sciences

# Preface

This book collects results presented at the *International Conference for Mathematical Modeling and Optimization in Mechanics (MMOM 2014) 6–7 March 2014, Jyväskylä, Finland,* which was dedicated to the 70th jubilee of Professor Nikolay Banichuk. The book consists of three parts: numerical analysis, mathematical modeling in mechanics, and optimization. This structure reflects the three main lines of the conference closely related to the scientific interests of Prof. Nikolay Banichuk and his colleagues.

*Part I* of the book contains four papers related to rather different but important problems in modern numerical analysis.

The first paper, by O. Pironneau, is devoted to highly nonlinear coupled models used for modeling of aortic flow. The paper combines analysis of viscous incompressible flow based on Navier–Stokes equations with ideas of shape optimization. Two next papers present new results on a posteriori error estimation methods for boundary value problems. The paper by O. Mali extends known a posteriori estimates of the functional type to the case of nonsymmetric elliptic operators. Another paper (by M. Nokka and S. Repin) is focused on applications of a posteriori estimates to iteration Uzawa methods for the stationary Bingham flow problem. The last paper in this part (by J.I. Toivanen) considers applications of the parametric level set method (which is one of the key tools of topology optimization) to methods of automatic differentiation. The author uses an adjoint approach to perform sensitivity analysis, but contrary to standard implementations the state problem is differentiated in its discretized form. The paper contains several examples demonstrating the performance of the method.

*Part II* collects the papers associated with mathematical modeling of mechanics.

It starts with the paper by Yuli D. Chashechkin, where the author discusses harmonization of analytical, numerical, and laboratory models of flows. This is mainly an overview paper aimed to present historical development of models and concepts in the theory of fluids. Other papers in this section are concerned with mathematical models of various mechanical and technological objects. Effects of friction in sliding contact of a sphere and a viscoelastic half space are studied in the

paper by I. Goryacheva, F. Stepanov, and E. Torskaya. Multiaxial fatigue criteria are used by N. Burago and I. Nikitin in an analysis of a complicated technical system. The paper by T. Saksa and J. Jeronen is devoted to dynamic analysis of viscoelastic Poynting Thompson beams. A projection approach to analysis of natural vibrations for beams with nonsymmetric cross-sections is presented in the paper by V. Saurin and G. Kostin. In the paper by N. Banichuk, A. Barsuk, J. Jeronen, P. Neittaanmäki, and T. Tuovinen, the authors consider bifurcation type problems arising in the theory of elastic stability.

*Part III* contains publications related to optimization methods.

The papers presented in this section can be classified into two groups. The first group mainly deals with optimization algorithms, while the second is more oriented to optimization and sensitivity analysis of engineering problems.

The first paper, by M.M. Mäkelä, N. Karmitsa and O. Wilppu, presents new algorithms of nonconvex multiobjective optimization based on the proximal bundle method. Parallelization of Nash genetic algorithms for solving inverse problems in structural engineering is discussed in the work of J. Périaux and D. Greiner.

A variant of variational design sensitivity analysis in structural optimisation using rigorous separation of physical quantities into geometry and displacement mappings is exposed in the paper by F.-J. Barthold, N. Gerzen, W. Kijanski, and D. Materna. The paper by G. Kostin and V. Saurin studies dynamics modeling and control design for elastic systems with distributed parameters, with the help of variational methods. Finally, contact optimization problems are considered in the work of I. Páczelt, A. Baksa, and Z. Mróz, and some problems of multipurpose optimization of deformed structures are investigated in the paper by A. Sinitsin, S. Ivanova, E. Makeev, and N. Banichuk.

The articles collected in the volume present only a part of the results of the conference. The editors tried to select contributions that are the most interesting. Some of them contain new results related to concrete mathematical or mechanical problems. Other articles were included by us because they overview the state of the art and discuss open questions related to a certain topic on mechanics, optimization methods, or modern technology. All the papers have been reviewed by two independent reviewers.

Jyväskylä
June 2015

Pekka Neittaanmäki
Sergey Repin
Tero Tuovinen

**Fig. 2** Participants

# Acknowledgments

We would like to thank all the authors for their contributions. All the papers included in the volume have been reviewed by at least two independent reviewers, and all the papers have been modified in accordance with the comments received. We would like to warmly thank all the reviewers for their excellent work, which made it possible to essentially improve the publications.

The editors are grateful to the Federation of Finnish Learned Societies for financial support. We would like to thank Marja-Leena Rantalainen for her careful work on preparing the electronic version of the book and express our sincere thanks to our counterpart in Springer.

# Contents

# Contributors

**Attila Baksa** Institute of Applied Mechanics, University of Miskolc, Miskolc-egyetemváros, Hungary

**Nikolay Banichuk** A. Ishlinsky Institute for Problems in Mechanics of the Russian Academy of Sciences, Moscow, Russia

**Alexander Barsuk** Moldova State University, Kishinev, Moldova

**Franz-Joseph Barthold** Numerische Methoden und Informationsverarbeitung, TU Dortmund, Dortmund, Germany

**Nikolay Burago** Ishlinski Institute for Problems in Mechanics of RAS, Moscow, Russia

**Yuli D. Chashechkin** A. Yu. Ishlinskiy Institute for Problems in Mechanics of the Russian Academy of Sciences, Moscow, Russia

**Nikolai Gerzen** Numerische Methoden und Informationsverarbeitung, TU Dortmund, Dortmund, Germany

**Irina Goryacheva** IPMech RAS, Moscow, Russia

**David Greiner** Institute of Intelligent Systems and Numerical Applications in Engineering (SIANI), Universidad de Las Palmas de Gran Canaria (ULPGC), Las Palmas, Spain

**Svetlana Ivanova** A. Ishlinsky Institute for Problems in Mechanics of the Russian Academy of Sciences, Moscow, Russia

**Juha Jeronen** Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland

**Napsu Karmitsa** Department of Mathematics and Statistics, University of Turku, Turku, Finland

**Wojciech Kijanski** Numerische Methoden und Informationsverarbeitung, TU Dortmund, Dortmund, Germany

**Georgy Kostin** Institute for Problems in Mechanics RAS, Moscow, Russia

**Evgeniy Makeev** A. Ishlinsky Institute for Problems in Mechanics of the Russian Academy of Sciences, Moscow, Russia

**Olli Mali** Department of Mathematical Information Technology, University of Jyvaskyla, Jyvaskyla, Finland

**Daniel Materna** Department of Civil Engineering, Ostwestfalen-Lippe University of Applied Sciences, Detmold, Germany

**Zenon Mróz** Institute of Fundamental Technological Research, Warsaw, Poland

**Marko M. Mäkelä** Department of Mathematics and Statistics, University of Turku, Turku, Finland

**Pekka Neittaanmäki** Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland

**Ilia Nikitin** Institute for Computer Aided Design of RAS, Moscow, Russia

**Marjaana Nokka** Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland

**Olivier Pironneau** Laboratoire Jacques-Louis Lions (LJLL), Sorbonne Universités, UPMC Univ Paris 06, UMR 7598, Paris Cedex 05, France

**István Páczelt** Institute of Applied Mechanics, University of Miskolc, Miskolc-egyetemváros, Hungary

**Jacques Périaux** Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland; International Center for Numerical Methods in Engineering (CIMNE), Universidad Politécnica de Cataluña (UPC), Barcelona, Spain

**Sergey Repin** Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland; St. Petersburg State Polytechnical University, St. Petersburg, Russia

**Tytti Saksa** Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland

**Vasily Saurin** Institute for Problems in Mechanics RAS, Moscow, Russia

**Alexander Sinitsin** A. Ishlinsky Institute for Problems in Mechanics of the Russian Academy of Sciences, Moscow, Russia

**Fedor Stepanov** IPMech RAS, Moscow, Russia

**Jukka I. Toivanen** Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland

**Elena Torskaya** IPMech RAS, Moscow, Russia

**Tero Tuovinen** Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland

**Outi Wilppu** Department of Mathematics and Statistics, University of Turku, Turku, Finland

# Part I
# Numerical Analysis

# Computational Issues for Optimal Shape Design in Hemodynamics

**Olivier Pironneau**

**Abstract** A Fluid-Structure Interaction model is studied for aortic flow, based on Koiter's shell model for the structure, Navier–Stokes equations for the fluid and transpiration for the coupling. It accounts for wall deformation while yet working on a fixed geometry. The model is established first. Then a numerical approximation is proposed and some tests are given. The model is also used for optimal design of a stent and possible recovery of the arterial wall elastic coefficients by inverse methods.

**Keywords** Fluid-structure interaction · Compliant walls · Finite element method · Convergence analysis · Navier–Stokes equations · Blood flow

**Mathematical Subject Classification:** 35Q30 · 74K25 · 65M60 · 65N12

## 1 Introduction

Hemodynamics, a special branch of computational fluid dynamics, poses many problems of modeling, data acquisition, computation and visualization. However, even as of now it is a valuable tool to understand aneurisms, to design stents and heart valves, etc. (see, for example, [6, 10, 11]).

In this paper, we shall focus on algorithms for fluid flows with compliant walls like aortic flow, their modelisation, numerical simulation and inverse techniques. Blood in large vessels like the aorta is Newtonian and flows in a laminar regime with Reynolds number of a few thousands. The Navier–Stokes equation for incompressible fluid is a good model for it. A blood vessel on the other hand is a complex structure for

O. Pironneau (✉)
Laboratoire Jacques-Louis Lions (LJLL), Sorbonne Universités,
UPMC Univ Paris 06, UMR 7598, Boite Courrier 187,
75252 Paris Cedex 05, France
e-mail: Olivier.Pironneau@upmc.fr



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_1

which linear elasticity is only a first crude approximation and for which the Lamé coefficients do not have a universal value and can vary with individuals.

Nevertheless, like many authors ([8, 9], for instance) we shall use Koiter's linear shell theory.

## 2 Koiter's Shell Model for Arteries

The following hierarchy of approximations for the displacement $d$ of the aortic wall will be made:

- Small displacement linear elasticity instead of large displacement (needed for the heart).
- No contact inequalities with the surrounding organs.
- Shell model for the mean surface.
- With reference to the mean surface, normal displacement of the walls only.

Let $\Sigma$ be the shell surface representing the mean position of the blood vessel. Let $n(x)$ be the normal at $x \in \Sigma$. Let $d(x,t)$ be the displacement of the wall at $x$ at time $t$. Normal displacement implies $d = \eta n$.

In [8] it is shown that under such conditions, Koiter's model reduces to the following equation of $\eta$ on $\Sigma$:

$$\rho_s h \partial_{tt}\eta - \nabla\cdot(\mathbf{T}\nabla\eta) - \nabla\cdot(\mathbf{C}\nabla\partial_t\eta) + a\partial_t\eta + b\eta = f^s, \tag{1}$$

where $\rho_s$ is the density and $h$ the thickness of the vessel, $\mathbf{T}$ is the pre-stress tensor, $\mathbf{C}$ is a damping term, $a$, $b$ are viscoelastic terms and $f^s$ the external normal force, i.e. the normal component of the normal stress tensor $-\sigma^{\mathbf{s}}{}_{nn}$. As with all second order wave type equations two conditions must be given at $t = 0$, for instance

$$\eta|_{t=0} = \eta_0, \quad \partial_t\eta|_{t=0} = \eta_0'.$$

*Remark 1* When $[h, T, C, a] \ll b$, the Eq. (1) leads to the so-called *surface pressure model*

$$-\sigma^{\mathbf{s}}{}_{nn} = b\eta, \quad \text{with } b = \frac{Eh\pi}{A(1-\xi^2)}, \tag{2}$$

where $A$ is depends on the geometry of the artery's cross section and equal to the cross section surface when it is circular; $E$ is the Young modulus, $\xi$ the Poisson coefficient.

Some typical values are (in the metric system MKSA) for a heart beat of one pulsation per second:

$$E = 3\,\text{MPa},\ \xi = 0.3,\ A = \pi R^2,\quad R = 0.013,\quad h = 0.001,\quad \rho^f = 9.81\times 10^6,$$

leading to $b = 3.3 \times 10^7\,\mathrm{ms}^{-2}$ and giving displacements in the range of $0.1 \times 10^{-3}$ m and flow rates around $2 \times 10^{-5}\,\mathrm{m}^3\mathrm{s}^{-1}$ for aortic flows.

## 3 Fluid Equations

The *Navier–Stokes equations* in a moving domain $\Omega(t)$ define the velocity $u$ and the pressure $p$,

$$\rho^f \left( \frac{\partial u}{\partial t} + u \cdot \nabla u \right) + \nabla p - \mu \nabla \cdot (\nabla u + \nabla u^T) = 0, \quad \nabla \cdot u = 0, \tag{3}$$

where $\rho^f$ is the density of the fluid and $\mu$ is the viscosity.

Continuity on $\Sigma$ of fluid and solid velocities implies

$$u = \frac{\partial d}{\partial t} := n \frac{\partial \eta}{\partial t}, \quad \text{on } \Sigma.$$

With the surface pressure model, continuity of normal stresses implies

$$\sigma^f_{nn} := n \cdot (\mu(\nabla u + \nabla u^T) - p)n = -\sigma^s_{nn} := b\eta.$$

Notice that as a consequence of the hypothesis of normal displacements only of the structure, there is no provision to write the continuity of the tangential stresses.

For aortic flow there also an inflow and an outflow boundary $\Gamma_i$ and $\Gamma_o$ on which we will prescribe pressure and no tangential velocity. If $S = \Gamma_i \cup \Gamma_o$, then the boundary $\Gamma$ is

$$\Gamma := \partial\Omega(t) = \Sigma \cup S = \Sigma \cup \Gamma_i \cup \Gamma_o.$$

In [8] and many other publications, the matching conditions on $\Sigma$ are written on the boundary of a fixed reference domain $\partial\Omega_0$ because Koiter's shell model works with a fixed mean surface $\Sigma$.

With the notations of [5], assume that the domain of the fluid is $\Omega_t = \mathscr{A}_t(\Omega_0)$ with $\mathscr{A}_t : x_0 \to x_t := \mathscr{A}_t(x_0)$. Let

$$u_\tau(x,t) = u(\mathscr{A}_t(\mathscr{A}_\tau^{-1}(x)), t), \quad \forall x \in \Omega_\tau. \tag{4}$$

Then in $\Omega_t$ at $t = \tau$, the Navier–Stokes equations are in ALE format

$$\rho^f \frac{\partial u_\tau}{\partial t} + (u_\tau - c_\tau) \cdot \nabla u_\tau + \nabla p - \mu \nabla \cdot (\nabla u_\tau + \nabla u_\tau^T) = 0,$$
$$\nabla \cdot u_\tau = 0, \quad \text{with } c_\tau(x) = -\frac{\partial \mathscr{A}_t(\mathscr{A}_\tau^{-1}(x))}{\partial t}|_{t=\tau}.$$

# 4 Transpiration Conditions for the Fluid

## *4.1 Conservation of Energy*

We begin with an important remark on the conservation of energy.

The variational formulation of (3)—divided by $\rho^s$—is, for all $\hat{u}$, $\hat{p}$

$$\int_{\Omega(t)}\Big[\hat{u}\cdot(\partial_t u+u\cdot\nabla u)+\nabla p\cdot\hat{u}-\hat{p}\nabla\cdot u + \frac{\nu}{2}(\nabla u+\nabla u^T):(\nabla\hat{u}+\nabla\hat{u}^T)\Big]=\int_{\Omega(t)} f^s\cdot\hat{u}. \tag{5}$$

An energy balance is obtained by taking $\hat{u}=u$ and $\hat{p}=-p$,

$$\partial_t\int_{\Omega(t)}\frac{u^2}{2}+\frac{\nu}{2}\int_\Omega|\nabla u+\nabla u^T|^2=\int_\Omega f^s\cdot\hat{u}-\int_{\partial\Omega}pu\cdot n, \tag{6}$$

because

$$\partial_t\int_{\Omega(t)}u\cdot w=\int_{\Omega(t)}\partial_t(u\cdot w)+\int_{\partial\Omega}v\,u\cdot w,$$
$$\int_\Omega((u\nabla u)\cdot u)=\int_{\partial\Omega}u\cdot n\frac{u^2}{2}=\int_{\partial\Omega}\frac{v}{2}u\cdot u,$$

where $v=u\cdot n$ is the normal speed of $\partial\Omega$.

With transpiration conditions we intend to work on a fixed domain with zero tangential velocity but non zero normal velocity $u\cdot n=w$. In that case, in order to preserve energy, the relation (5) on a fixed domain $\Omega$ needs to be modified into

$$\int_\Omega\Big[\hat{u}\cdot(\partial_t u+u\cdot\nabla u)+\nabla\tilde{p}\cdot\hat{u}-\hat{p}\nabla\cdot u + \frac{\nu}{2}(\nabla u+\nabla u^T):(\nabla\hat{u}+\nabla\hat{u}^T)\Big]-\int_{\partial\Omega}\frac{w}{2}u\cdot\hat{u}=\int_\Omega f^s\cdot\hat{u}$$

or equivalently into

$$\int_\Omega\Big[\hat{u}\cdot(\partial_t u-u\times\nabla\times u)+\nabla\tilde{p}\cdot\hat{u}-\hat{p}\nabla\cdot u + \frac{\nu}{2}(\nabla u+\nabla u^T):(\nabla\hat{u}+\nabla\hat{u}^T)\Big]=\int_\Omega f^s\cdot\hat{u},$$

where $\tilde{p}=p+\frac{1}{2}|u|^2$ is the dynamic pressure.

*Remark 2* Notice that the difference between $p$ and $\tilde{p}$ is second order with respect to the displacement, so exchanging one for the other in the shell model is a modification well within the small displacement hypothesis. However, it makes a difference on $\Gamma_i$, $\Gamma_o$ and $p_\Gamma$ should be changed accordingly.

From now on we drop the tilde on $p$.

Finally, we recall the identity (see [4], for instance) which holds whenever $u \times n = 0$ and shows that we can use several forms for the viscous terms, namely,

$$\int_\Omega [\nabla \times u \cdot \nabla \times v + \nabla \cdot u \nabla \cdot v] = \int_\Omega \nabla u : \nabla v = \int_\Omega \left[ \frac{1}{2} (\nabla u + \nabla u^T) : (\nabla v + \nabla v^T) - \nabla \cdot u \nabla \cdot v \right].$$

Hence a variational formulation adapted to the problem is to find $u$ with $u \times n = 0$ and, such that for all $\hat{p}$ and all $\hat{u}$ with $\hat{u} \times n = 0$ we have

$$\int_\Omega [\hat{u} \cdot (\partial_t u - u \times \nabla \times u) - p \nabla \cdot \hat{u} - \hat{p} \nabla \cdot u + \nu \nabla \times u \cdot \nabla \times v] + \int_{\partial\Omega} p u \cdot n = \int_\Omega f^s \cdot \hat{u}. \tag{7}$$

## *4.2 Transpiration*

As the wall vessel is $\{x + \eta n : x \in \Sigma\}$ and as, by Taylor,

$$u(x + \eta n) = u(x) + \eta \nabla u \cdot n(x) + o(\eta),$$

matching the velocities of fluid and structure may be presented in the form

$$u + \eta \frac{\partial u}{\partial n} = n \frac{\partial \eta}{\partial t} + o(\eta), \ u \times n = 0 \quad \text{on } \Sigma. \tag{8}$$

For a torus with small radius $r$ and large radius $R$, at a point of coordinates $((R + r\cos\theta)\cos\varphi, (R + r\cos\theta)\cos\varphi, r\sin\theta)$, a straightforward calculation shows that

$$u \times n = 0, \ \nabla \cdot u = 0 \Rightarrow n \cdot \frac{\partial u}{\partial n} = (1 + \frac{r}{R}\cos^2\theta) \frac{u \cdot n}{r}.$$

In view of this, (8) becomes

$$u \cdot n = \partial_t \eta \left(1 + \frac{\eta}{r}\left(1 + \frac{r}{R}\cos^2\theta\right)\right)^{-1}, \quad u \times n = 0. \tag{9}$$

Similarly, the normal component of the normal stress tensor is

$$\sigma_{nn}^f = p + 2\left(1 + \frac{r}{R}\cos^2\theta\right)\frac{\mu}{r}u\cdot n.$$

Therefore, for quasi toroidal geometry and large $R$, the relation (1) reads

$$\rho_s h\partial_{tt}\eta - \nabla\cdot(\mathbf{T}\nabla\eta) - \nabla\cdot(\mathbf{C}\nabla\partial_t\eta) + a\partial_t\eta + b\eta$$
$$= p + 2\left(1 + \frac{r}{R}\cos^2\theta\right)\frac{\mu}{r}\partial_t\eta\left(1 + \frac{\eta}{r}\left(1 + \frac{r}{R}\cos^2\theta\right)\right)^{-1}.$$

So, in fine, the domain $\Omega$ no longer varies in time but on a part of its boundary we have the conditions

$$u\cdot n = \partial_t\eta\left(1 + \frac{\eta}{r}\left(1 + \frac{r}{R}\cos^2\theta\right)\right)^{-1}, \quad u\times n = 0,$$
$$\rho_s h\partial_{tt}\eta - \nabla\cdot(\mathbf{T}\nabla\eta) - \nabla\cdot(\mathbf{C}\nabla\partial_t\eta) + a\partial_t\eta + b\eta = p,$$

where $a$ is a non linear function of $\eta$.

*Remark 3* Notice that $\eta \ll r$, i.e. large vessels, allows us to eliminate $\eta$ and write everything in terms of $\partial_t p$ and $u_n := u\cdot n$. It suffices to differentiate the last equation with respect to t and use the first one and integrate in time,

$$p = p_0 + \mathscr{L}(u\cdot n)$$
$$:= \int_0^t \left(\rho_s h\partial_{tt}u_n - \nabla\cdot(T\nabla u_n) - \nabla\cdot(C\nabla\partial_t u_n) + a\partial_t u_n + bu_n\right). \tag{10}$$

## 5 Variational Formulation and Approximation

Coming back to (7) and using (10), we arrive at the following:

**Continuous Problem** Find $u$ with $u\times n = 0$ and, for all $\hat{p}$ and all $\hat{u}$ with $\hat{u}\times n = 0$

$$\int_\Omega \left[\hat{u}\cdot(\partial_t u - u\times\nabla\times u) - p\nabla\cdot\hat{u} - \hat{p}\nabla\cdot u + \nu\nabla\times u\cdot\nabla\times\nu\right]$$
$$+ \int_\Sigma \left(p_0 + \mathscr{L}(u\cdot n)\right)u\cdot n = -\int_S p_\Gamma\hat{u}\cdot n.$$

### 5.1 Approximation in Time

From now on, for clarity, we consider only the case of the surface pressure model, i.e. $h = T = C = a = 0$, $\mathscr{L}(u \cdot n) = bu \cdot n$. However, everything below extends to the full model.

We define $U(t) = \int_0^t u(s)\mathrm{d}s$ and use the integration rule $U^{m+1} = U^m + u^{m+1}\mathrm{d}t$ and

$$V = \{u \in H^1(\Omega)^d \mid u \times n = 0 \text{ on } \partial\Omega\}, \quad Q = L^2(\Omega).$$

**Time Discrete Problem** Let $p(t) = p_0 + bU(t)$. We seek $u^{m+1} \in V$, $\hat{p}^{m+1} \in Q$, satisfying for all $\hat{u} \in V$ and $\hat{p} \in Q$ the following identity:

$$\int_\Omega \Big[\hat{u} \cdot \left(\frac{u^{m+1} - u^m}{\delta t} - u^{m+\frac{1}{2}} \times \nabla \times u^{m+\theta}\right) - p^{m+1}\nabla \cdot \hat{u} - \hat{p}\nabla \cdot u^{m+\frac{1}{2}} + \nu \nabla \times u^{m+\frac{1}{2}} \cdot \nabla \times \hat{u}\Big] + \int_\Sigma \Big[b\hat{u} \cdot n(u^{m+\frac{1}{2}}\delta t + U^m) \cdot n\Big] = -\int_S p_\Gamma \hat{u} \cdot n, \tag{11}$$

where $u^{m+\frac{1}{2}} = \frac{1}{2}(u^{m+1} + u^m)$ and $\theta = 0$ for a semi-explicit linear scheme of the first order or $\theta = \frac{1}{2}$ for a fully implicit nonlinear scheme of the second order.

### 5.2 Convergence

A convergence analysis was done in [3]; we recall the results. We denote $u_\delta$ the linear in time interpolate of $\{u^m\}_1^M$ on $(0, T) = \cup_1^M[(m-1)\delta t, m\delta t]$. For clarity let's assume that $S = \emptyset$.

**Lemma 1** *If $\Omega$ is $\mathscr{C}^{1,1}$ or polyhedral and $u_0 \in L^2(\Omega)^3$, $p_0 \in H^{1/2}(\Sigma)$, then the weak solution of the continuous problem verifies $u \in L^2(\mathbf{H}^2)$, $\partial_t u \in L^2(\mathbf{L}^2)$, $p \in L^2(H^1)$, and $u \times n = 0$ in $L^2(L^4(\Sigma))$, $\partial_t p = bu \cdot n$ in $L^2(H^{1/2}(\Sigma))$, $p(0) = p_0$.*

**Theorem 1** *The solution of the time discretized variational problem satisfies*

$$\|u_\delta\|_{L^\infty(\mathbf{L}^2)} + \sqrt{\nu}\,\|u_\delta\|_{L^2(\mathbf{H}^1)} + b\,\|\delta t \sum_{k=1}^{n+1} u^k \cdot n\|_{L^\infty(\mathbf{L}^2(\Sigma))} \leq C\left(\|u_0\|_{0,2,\Omega} + \frac{1}{\sqrt{\nu}}\|p_0\|_{L^2(\Sigma)}\right).$$

```
problem bb([u,v,w,p],[uh,vh,wh,ph], solver=UMFPACK)
          = int3d(th)((u*uh+v*vh+w*wh)/dt2
                      + nu*(dx(u)*dx(uh)+dy(u)*dy(uh)+dz(u)*dz(uh)        // rot u.rot uh
                          + dx(v)*dx(vh)+dy(v)*dy(vh) +dz(v)*dz(vh)
                          + dx(w)*dx(wh)+dy(w)*dy(wh) +dz(w)*dz(wh)
                          -(dx(u)+dy(v)+dz(w))*(dx(uh)+dy(vh)+dz(wh))
                       ) - (dx(uh)+dy(vh)+dz(wh))*p - (dx(u)+dy(v)+dz(w))*ph //-ph div u - p div uh
- v*(dx(vold)-dy(uold))*uh -w*(dy(wold)-dz(vold))*vh -u*(dz(uold)-dx(wold))*wh // u x rot u
+ w*(dz(uold)-dx(wold))*uh +u*(dx(vold)-dy(uold))*vh + v*(dy(wold)-dx(vold))*wh
                     )
            - int3d(th)( (uold*uh + vold*vh + wold*wh)/dt2 )              // u^m/dt
            + int2d(th,1)((     (u*N.y-v*N.x)*(uh*N.y-vh*N.x)             // uxn.uhxn/eps
                          + (v*N.z-w*N.y)*(vh*N.z-wh*N.y)
                          + (w*N.x-u*N.z)*(wh*N.x-uh*N.z)   )/eps)
            + int2d(th,1)(b*dt2*(u*N.x+v*N.y+w*N.z)*(uh*N.x+vh*N.y+wh*N.z)) // b dt u.n uh.n
            + int2d(th,1)(b*(Uold*N.x+Vold*N.y+Wold*N.z)*(uh*N.x+vh*N.y+wh*N.z)) // b n. int_0^t u
            - int2d(th,2)(p0*wh ) - int2d(th,3)(p1*vh )
            + on(2,3,u=0) + on(2,v=0) +on(3,w=0);
```

**Fig. 1** Freefem++ implementation for problem (11)

**Theorem 2** *If $\Omega$ is simply connected, there is a subsequence $(u_{\delta'}, p_{\delta'})$ which converges to the continuous problem in $L^2(\mathbf{W}) \times H^{-1}(L^2)$, where*

$$\mathbf{W} = \{w \in L^2(\Omega) \mid \nabla \times w \in L^2(\Omega),\ \nabla \cdot w \in L^2(\Omega),\ n \times \mathbf{w}|_\Sigma = \mathbf{0}\}.$$

### *5.3 Spatial Discretization with Finite Elements*

The easiest way is to use penalization to enforce $u \times n = 0$ by adding to the boundary integral $\frac{1}{\varepsilon}\int_\Sigma u^{m+1} \times n \cdot \hat{u} \times n$. Then we may use conforming triangular or tetrahedral elements $P^2$ or $P^1$ + bubble for the velocities and $P^1$ elements for the pressure.

A freefem++ implementation (see [7]) is shown in Fig. 1.

## 6 Optimization and Inverse Problems

### *6.1 Optimal Stents with the Surface Pressure Model*

A stent is a device to reinforce part of a cardiac vessel and/or to change the topology of the flow by its rigidity. This results in a change of the coefficient $b$. So with a first order scheme in time we can consider

$$\min_{b(x)} J = \int_{\Sigma \times (0,T)} F(p)\, \mathrm{d}x\, \mathrm{d}t$$

subject to

$$\int_\Omega \left[\hat{u}\cdot\left(\frac{u^{m+1}-u^m}{\delta t}-u^{m+1}\times\nabla\times u^m\right)-p^{m+1}\nabla\cdot\hat{u}-\hat{p}\nabla\cdot u^{m+1}\right] + \int_\Omega \nu\nabla\times u^{m+1}\cdot\nabla\times\hat{u}\int_\Sigma [b\hat{u}\cdot n(u^{m+\frac{1}{2}}\delta t+U^m)\cdot n] = -\int_S p_\Gamma\hat{u}\cdot n \quad \forall\hat{u}\in V_h,\ \hat{p}\in Q_h. \tag{12}$$

For instance, $F=|p|^4$ will minimize the time averages pressure peak on $\Sigma$.

## 6.2 Inverse Problems

Can we recover the structural parameters of the vessel walls from the observation of the pressure?

Consider the minimization problem

$$\min_{b(x),x\in\Sigma} J(u,p,b) := \frac{1}{2}\int_{\Omega\times(0,T)}(p^m-p_d^m)^2, \tag{13}$$

subject to (12) or to

$$\int_\Omega \left[\hat{u}\cdot\left(\frac{1}{\delta t}\left(u^{m+1}-u^m(x-u^m(x)\delta t)\right)\right)-p^{m+1}\nabla\cdot\hat{u}-\hat{p}\nabla\cdot u^{m+1}\right] + \int_\Omega \nu\nabla\times u^{m+1}\cdot\nabla\times\hat{u}+\int_\Sigma b(u^{m+1}\delta t+U^m)\cdot\hat{u} = -\int_\Gamma p_\Gamma\hat{u}_n \quad \forall\hat{u}\in V_h,\ \hat{p}\in Q_h \quad \text{with } \hat{u}\times n|_\Gamma=0;\ U^{m+1}=U^m+u^{m+1}\delta t. \tag{14}$$

The difference between (12) and (14) is the numerical treatment of the nonlinear term: implicit Euler in the first and Characteristic-Galerkin in the second.

## 6.3 Calculus of Variations

To set up a descent algorithm we must do a sensitivity analysis of the problem. This is done with a "Calculus of Variations".

When a parameter varies it triggers a variation of $u$, $p$, which we call $\delta u$, $\delta p$. To compute them we linearise the Navier–Stokes equations. These written globally over $(0,T)$ in a weak form are as follows:

$$\sum_0^{M-1}\delta t\Big(\int_\Omega\Big[\hat u^{m+1}\cdot\Big(\frac{1}{\delta t}\Big(\delta u^{m+1}-\delta u^m(x-u^m(x)\delta t)\Big)\Big)$$
$$-\delta p^{m+1}\nabla\cdot\hat u^{m+1}-\hat p^{m+1}\nabla\cdot\delta u^{m+1}\Big]$$
$$+\int_\Omega \nu\nabla\times\delta u^{m+1}\cdot\nabla\times\hat u^{m+1}+\int_\Sigma r^{m+1}(\delta U^{m+1}-\delta U^m-\delta u^{m+1}\delta t)$$
$$+\int_\Sigma\Big(b(\delta u^{m+1}\delta t+\delta U^m)+\delta b(u^{m+1}\delta t+U^m)\Big)\cdot\hat u^{m+1}\Big)=0$$
$$\forall\hat u^m\in V_h,\ \hat p^m\in Q_h,\ r^m \text{ with } \hat u^m\times n|_\Gamma=0. \quad (15)$$

If $\hat u^{M+1}=0$, $r^{M+1}=0$, then it can be rearranged and presented in the form

$$\sum_0^{M-1}\delta t\Big(\int_\Omega\Big[\frac{1}{\delta t}\Big(\hat u^{m+1}-\hat u^{m+2}(x+u^m(x)\delta t)\Big)$$
$$\times\ \delta u^{m+1}-\delta p^{m+1}\nabla\cdot\hat u^{m+1}-\hat p^{m+1}\nabla\cdot\delta u^{m+1}\Big]$$
$$+\int_\Omega \nu\nabla\times\delta u^{m+1}\cdot\nabla\times\hat u^{m+1}+\int_\Sigma((r^{m+1}-r^{m+2})\delta U^{m+1}-r^{m+1}\delta u^{m+1}\delta t)$$
$$+\int_\Sigma\Big(b(\delta u^{m+1}\delta t+\delta U^m)+\delta b(u^{m+1}\delta t+U^m)\Big)\cdot\hat u^{m+1}\Big)=0. \quad (16)$$

## 6.4 Adjoint State

To express the variations in terms of $\delta b$, we need to introduce an adjoint state $v$, solution of the following,

$$\sum_0^{M-1}\delta t\Big(\int_\Omega\Big[\frac{1}{\delta t}\Big(v^{m+1}-v^{m+2}(x+u^m(x)\delta t)\Big)$$
$$\times\ \hat v^{m+1}-\hat q^{m+1}\nabla\cdot v^{m+1}-q^{m+1}\nabla\cdot\hat u^{m+1}\Big]$$
$$+\int_\Omega \nu\nabla\times\hat u^{m+1}\cdot\nabla\times v^{m+1}+\int_\Sigma(r^{m+1}-r^{m+2})\hat V^{m+1}-r^{m+1}\hat v^{m+1}\delta t$$
$$+\int_\Sigma b(\hat v^{m+1}\delta t+\hat V^m)\cdot v^{m+1}\Big)=\sum_0^{M-1}\delta t\int_\Omega(p^{m+1}-p_d^{m+1})\hat q^{m+1}, \quad (17)$$

for all $\hat{v}$, $\hat{q}$ such that $\hat{v} \times n = 0$ on $\partial\Omega$. Denote $V^m = r^m \delta t$. Then $V^{m+1} = V^{m+2} - bv^{m+2}\delta t$ and

$$\int_\Omega \left[ \frac{1}{\delta t} \left( v^{m+1} - v^{m+2}(x + u^m(x)\delta t) \right) \cdot \hat{v} - \hat{q} \nabla \cdot v^{m+1} - q^{m+1} \nabla \cdot \hat{v} \right] + \int_\Omega \nu \nabla \times v^{m+1} \cdot \nabla \times \hat{v} + \int_\Sigma (bv^{m+1}\delta t - V^{m+1}) \cdot \hat{v} = \int_\Omega (p^{m+1} - p_d^{m+1}) \hat{q}, \tag{18}$$

for all $\hat{v}$, $\hat{q}$ such that $\hat{v} \times n = 0$ on $\partial\Omega$.

## 6.5 Computation of Gradients with Respect to b

Letting $\hat{v} = \delta u^{m+1}$, $\hat{q} = \delta p^{m+1}$ and summing in $m$, from 1 to M after multiplication by $\delta t$ gives,

$$\begin{aligned}
&\sum_0^{M-1} \delta t \int_\Omega (p^{m+1} - p_d^{m+1}) \delta p^{m+1} \\
&= \sum_0^{M-1} \delta t \Bigg( \int_\Omega \left[ \frac{1}{\delta t} \left( v^{m+1} - v^{m+2}(x + u^m(x)\delta t) \right) \cdot \delta u^{m+1} \right. \\
&\quad \left. - \delta p^{m+1} \nabla \cdot v^{m+1} - q^{m+1} \nabla \cdot \delta u^{m+1} \right] + \int_\Omega \nu \nabla \times \delta u^{m+1} \cdot \nabla \times v^{m+1} \\
&\quad + \int_\Sigma \left( (bv^{m+1} - r^{m+1}) \delta u^{m+1} \delta t - \delta U^{m+1} (r^{m+2} - r^{m+1} - bv^{m+2}) \right) \Bigg) \\
&= \sum_0^{M-1} \delta t \Bigg( \int_\Omega \left[ \frac{1}{\delta t} \left( \delta u^{m+1} - \delta u^m(x - u^m(x)\delta t) \right) \cdot v^{m+1} \right. \\
&\quad \left. - \delta p^{m+1} \nabla \cdot v^{m+1} - q^{m+1} \nabla \cdot \delta u^{m+1} \right] + \int_\Omega \nu \nabla \times \delta u^{m+1} \cdot \nabla \times v^{m+1} \\
&\quad + \int_\Sigma \left( bv^{m+1} \delta u^{m+1} \delta t + \delta U^m b v^{m+1} + (\delta U^{m+1} - \delta U^m - \delta u^{m+1} \delta t) r^{m+1} \right) \Bigg) \\
&= - \sum_0^{M-1} \delta t \int_\Sigma \delta b (u^{m+1} \delta t + U^m) \cdot v^{m+1} = - \int_\Sigma \delta b \left( \delta t \sum_0^{M-1} U^m \cdot v^m \right),
\end{aligned}$$

because $U^0 = v^M = 0$. To minimize in $H^1$-norm, we solve for $g \in H_0^1(\Sigma)$,

$$\int_\Sigma \nabla_s g \cdot \nabla_s w = -\int_\Sigma \left( \delta t \sum_0^{M-1} U^m \cdot v^m \right) w, \quad \forall w \in H_0^1(\Sigma)$$
$$\Rightarrow \quad \delta J = \int_\Sigma \nabla_s g \cdot \nabla_s \delta b. \tag{19}$$

## 6.6 Numerical Tests

We take the test case documented in [2]. It is a 2-d problem for the upper part of a symmetric straight vessel. The geometry is the rectangle $(0, L) \times (0, R)$ with $L = 6$ and $R = 0.5$. Pressure is imposed at both end, zero on the right and $p_1 = \frac{1}{2} p_{\max}(1 - \cos(2\pi \frac{t}{t_{\max}}))$ with $p_{\max} = 2000$ and $t_{\max} = 0.005$.

The mesh is uniform $60 \times 10$. The step size is $\delta t = 2 \times 10^{-4}$ and there are 60 time steps in this simulation, so $T = 0.012 = 2.4 t_{\max}$. The $P^2 \times P^1$ element is used for velocity-pressure. The objective is to see if it is possible to reconstruct $b$ on the upper wall from the pressure in the vessel.

So we first solve the direct problem with $b = b_d := 2 \times 10^5 (1 + 6\frac{x}{L}(1 - \frac{x}{L}))$ approximated with the $P^1$ element. We call the computed pressure $\{p_d^m(x)\}_0^{M-1}$. Then, we solve (13) with 50 iterations of an $H_0^1$-projected gradient method with fixed step size, $\lambda = 10^6$.

Algorithmic Steps

- Compute $p_d$ by a time loop from 0 to $T$ and store on disk.
- Optimization loop:
  1. Compute $u$, $p$ by a time loop from 0 to $T$ and store on disk $u$, $p$, $U$.
  2. Compute $v$, $p$ by a time loop from T down to 0 requiring to read from disk $p_d$, $u$, $p$, $U$.
  3. Compute gradient by solving (19).
  4. Compute cost function and $\|\partial_x g\|_0^2$.
  5. Update $b$ by $b \leftarrow b - \lambda g$.
  6. Modify $b$ by $b \leftarrow \max\{\min(b, b_{\max}), b_{\min}\}$.
- Display results.

We choose $b_{\max} = 2 \times 10^5 (1 + 12\frac{x}{L}(1 - \frac{x}{L}))$, $b_{\min} = 2 \times 10^5 (1 + 2\frac{x}{L}(1 - \frac{x}{L}))$. The results are shown in Figs. 2, 3, and 4.

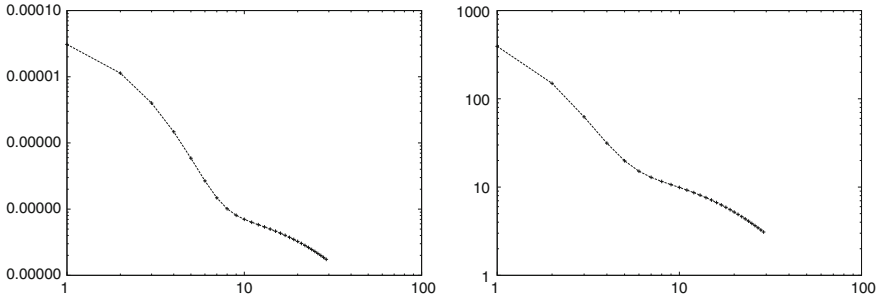



**Fig. 2** $g$ (*left*) and $J$ (*right*) versus iteration number in log-log scale. Initially $J = 1403$ and after 50 gradient iterations $J = 1.27$ while $g$ decreases from $1.2 \times 10^{-4}$ to $3.3 \times 10^{-9}$

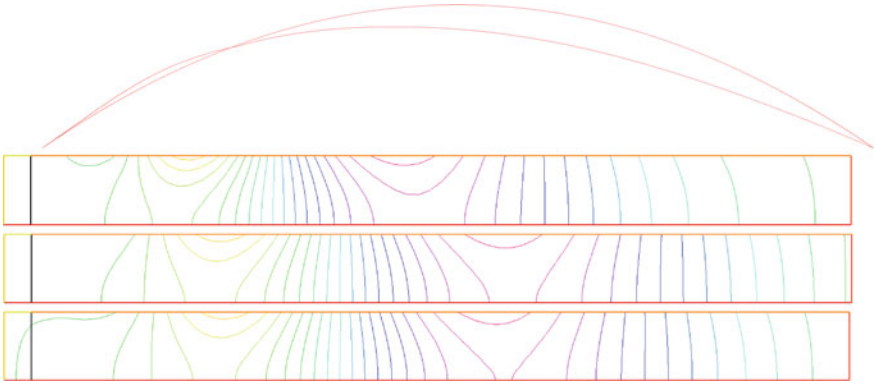


**Fig. 3** Target $b_d$ (*top curve*) and computed $b$ after 50 iterations. Initial Pressure map after one iteration (*top*), final pressure after 50 gradient iterations (*middle*) and target pressure $p_d$ (*bottom*). The *color scales* are linear from $-986$ to 896 except for $p_0$ which has a range from $-680$ to 782

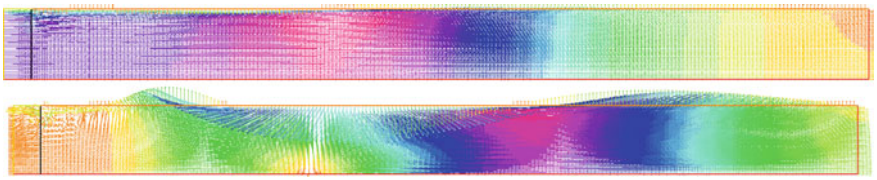


**Fig. 4** Flow velocity vectors $u$ (*middle*) and adjoint flow velocity vectors $v$ (*bottom*) at final time after 50 gradient iterations. The *color scales* are linear from 0 (*saffron*) to 0.03 (*red*) for $u$ and 0 (*saffron*) to 2.9 (*red*) for $v$. The singularity at the *top left corner* is due to a theoretical incompatibility between the normal velocities at this corner

## 6.7 Preliminary 3D Tests

Experiment 1

This is only a feasibility test with $F = p^4$; The geometry is a quarter of a torus with $R = 4$ and $r = 1$. It is discretized with 1395 vertices and 6120 elements. The number of unknown of the coupled system $[u, p]$ is 23940 with the $P^1$-bubble/$P^1$ element and Crank-Nicolson implicit scheme. The viscosity is $\nu = 0.01$; we chose $\varepsilon = \nu$. The final time is $T = 1$, the time step is $\mathrm{d}t = 0.1$ and the pressure difference imposed at $\Gamma_i$ (top) and $\Gamma_o$ bottom is $6\cos^2(\pi t)$.

The flow is stored on disk at every iteration ready to be reused backward in time for the adjoint equations. Starting with $b = 200$, after three iterations of steepest descent with fixed step size, the cost function is decreased from 1200 to 900. But as there is no constraint $b$ is much reduced at the top near $\Gamma_i$. Consequently the vessel wall becomes fragile as shown by a simulated wall motion by $x \to x + \sum u^m \cdot n \mathrm{d}t$ at every time step, as shown in Fig. 5.

Experiment 2

The same computations has been made but now $b$ is constrained to be greater than $b_0/2$. A mesh double the size of the previous one has been used, with 191808 degrees of freedom. The initial value of $b$ is $b_0 = 200$. After 10 iterations, similar to Experiment 1 but with a projected gradient method for the optimization, the results of Fig. 6 are found.

Experiment 3

Finally we run an identification test of $b$ from the observation of the wall displacement, ideally, $u \cdot n$. However, the formulation does not allow it because the extra integral in the adjoint variational formulation is in competition with a similar term from the surface pressure model, so we used $p/b$. For this first test the criteria is

$$J = \int_{\Sigma\times(0,T)} |p - p_d|^2 \mathrm{d}x\mathrm{d}t,$$

where $p_d$ is obtained from a reference computation (introduction of $b$ in the criteria makes the problem harder) with

$$b = 200 + 100\cos x \cos y \cos z.$$

The results are shown in Fig. 7.

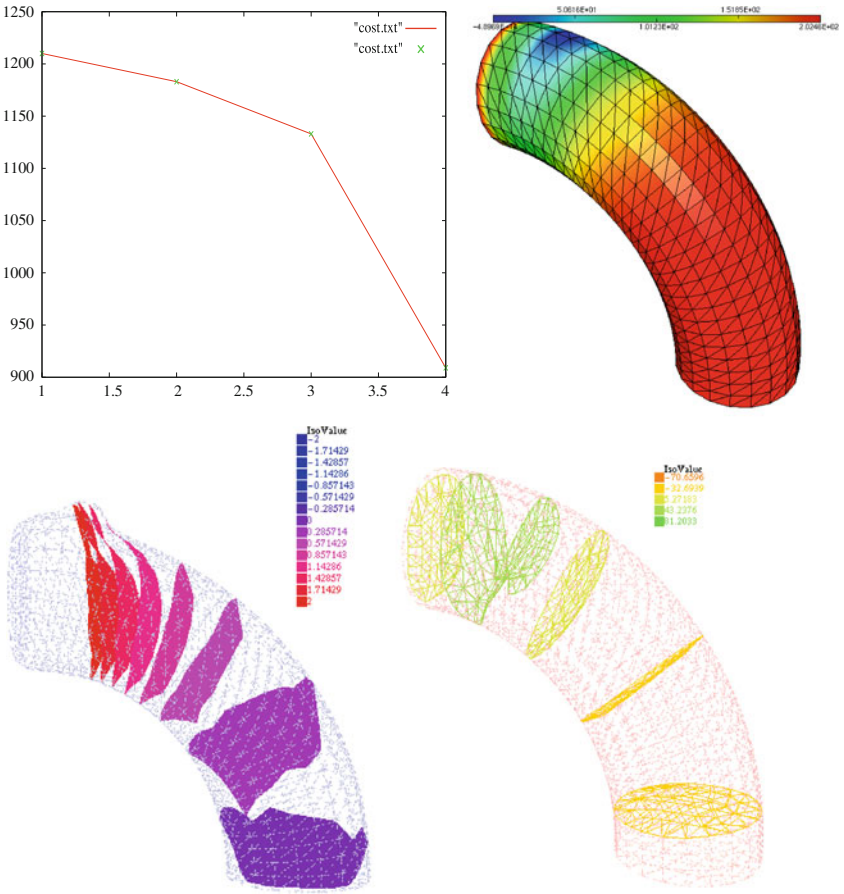



**Fig. 5** *Top left* optimization criteria versus iteration number. *Top right* the coefficient $b(x)$ after 3 iterations. *Bottom left* effect of the change of $b$ on the dilatation of the vessel and some iso surfaces of constant pressure. *Bottom right* a snap shot of the adjoint pressure and some iso surfaces

Because of the computing cost, we made only an initial study; the target is not reached, but 5 iterations go into the right direction. To do better one would have to used a varying step size gradient method and a better computer (this being done on a macbook pro, takes about 15 min).

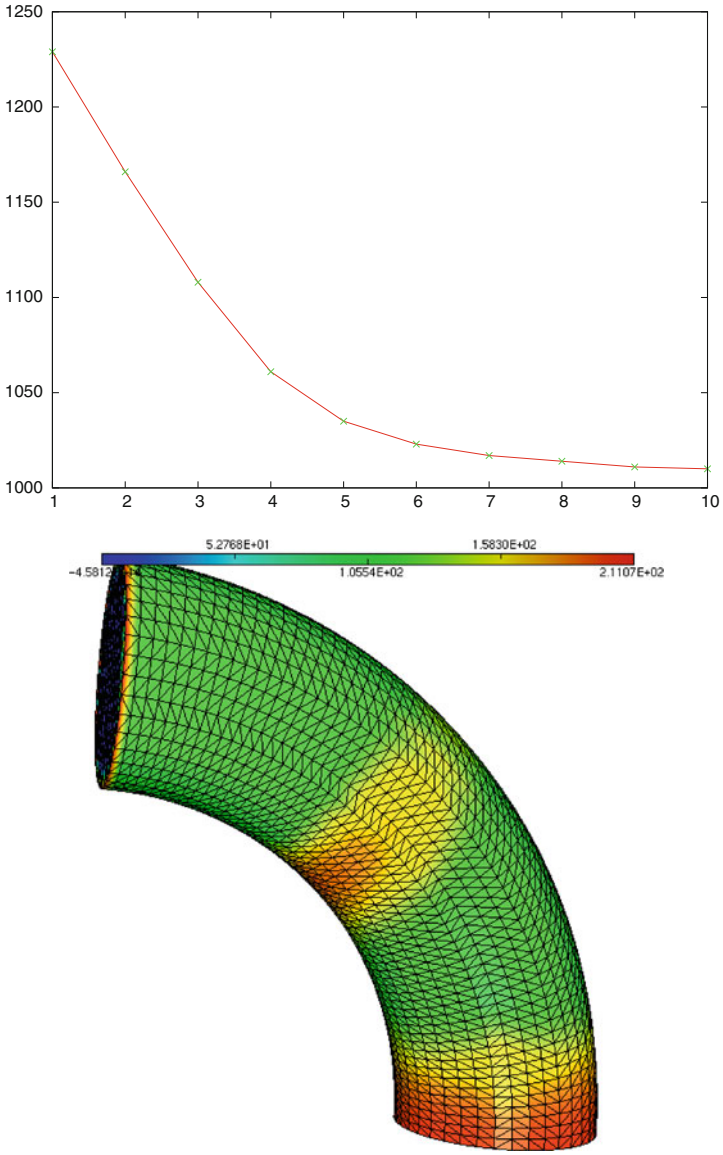



**Fig. 6** *Top* optimization criteria $\int_{\Sigma\times(0,T)} p^4$ versus iteration number. *Bottom* the coefficient $b(x)$ after 4 iterations. *Right* effect of the change of $b$ on the dilatation of the vessel

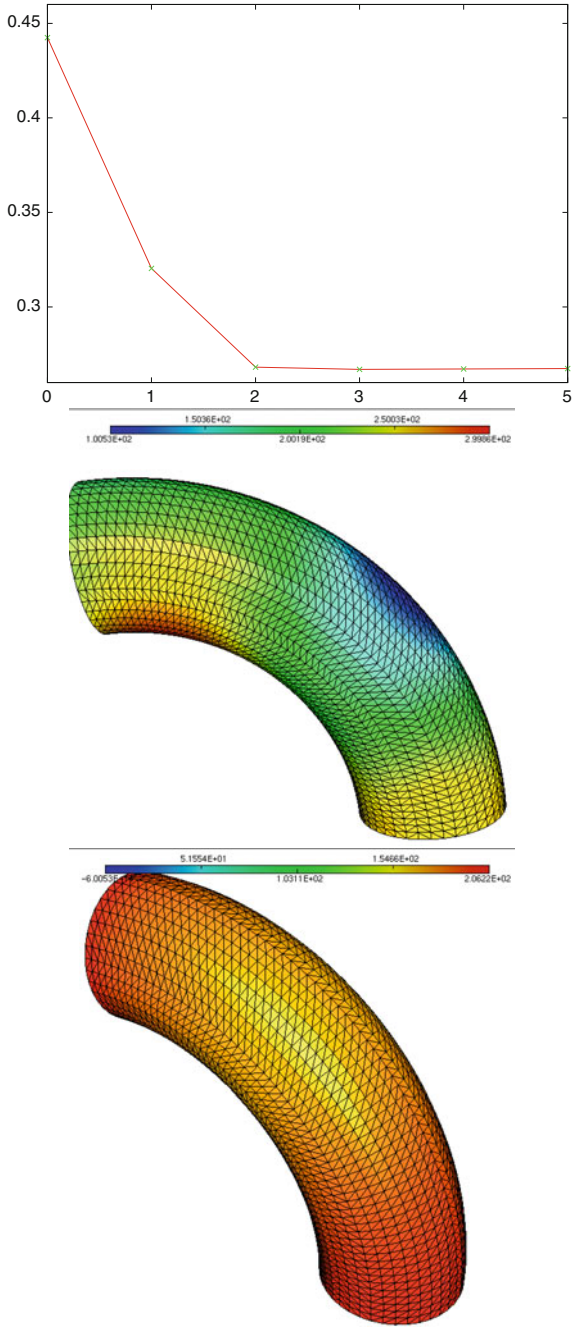



**Fig. 7** *Top* optimization criteria $\int_{\Sigma\times(0,T)}(p-p_d)^2$ versus iteration number. *Middle* the target $b$. *Bottom* the coefficient $b(x)$ after 5 iterations

## 7 Conclusion

In this paper, we have introduced a reduced fluid structure model based on a transpiration condition and applied it on a problem arising from hemodynamics. We have shown that it has good stability property. In [3] a comparison study is made with full fluid-structure models on moving domains; it is shown to give very similar results.

The greatest advantage of this reduced model is its computational speed and unconditional stability. As inverse problems are important in hemodynamics [1], it could be a good idea to use it. This preliminary study shows that it is indeed feasible.

**Acknowledgments** Special thanks to Frédéric Hecht for his help with freefem++ and Marc Thiriet and Sunčica Čanić for helpful discussions.

## References

1. C. Bertoglio, D. Barber, N. Gaddum, I. Valverde, M. Rutten, P. Beerbaum, P. Moireau, R. Hose, J.-F. Gerbeau, Identification of artery wall stiffness: in vitro validation and in vivo results of a data assimilation procedure applied to a 3D fluid-structure interaction model. J. Biomech. **47**(5), 1027–1034 (2014)
2. M. Bukač, S. Čanić, R. Glowinski, J. Tambača, A. Quaini, Fluid-structure interaction in blood flow capturing non-zero longitudinal structure displacement. J. Comput. Phys. **235**, 515–541 (2013)
3. T. Chacon Rebollo, V. Girault, F. Murat, O. Pironneau, Analysis of a simplified coupled fluid-structure model for computational hemodynamics. SIAM J. Numer. Anal. Submitted
4. M. Costabel, M. Dauge, Singularities of electromagnetic fields in polyhedral domains. Preprint 97-19, Université de Rennes 1, 1997. http://www.maths.univ-rennes1.fr/~dauge/
5. A. Decoene, B. Maury, Moving meshes with freefem++. J. Numer. Math. **20**(3–4), 195–214 (2012)
6. L. Formaggia, A. Quarteroni, A. Veneziani (eds.), *Cardiovasuclar Mathematics: Modeling and Simulation of the Circulatory System* (Springer, Milano, 2009)
7. F. Hecht, New development in freefem++. J. Numer. Math. **20**(3–4), 251–265 (2012)
8. F. Nobile, C. Vergara, An effective fluid-structure interaction formulation for vascular dynamics by generalized Robin conditions. SIAM J. Sci. Comput. **30**(2), 731–763 (2008)
9. J. Tambaca, S. Canic, M. Kosor, R.D. Fish, D. Paniagua, Mechanical behavior of fully expanded commercially available endovascular coronary stents. Texas Heart Inst. J. **38**(5), 491–501 (2011)
10. M. Thiriet, *Control of Cell Fate in the Circulatory and Ventilatory Systems*. Biomathematical and Biomechanical Modeling of the Circulatory and Ventilatory Systems, vol. 2 (Springer, New York, 2012)
11. F. Usabiaga, J. Bell, R. Delgado-Buscalioni, A. Donev, T. Fai, B. Griffith, C. Peskin, Staggered schemes for fluctuating hydrodynamics. Multiscale Model. Simul. **10**(4), 1369–1408 (2012)

# Functional A Posteriori Error Estimate for a Nonsymmetric Stationary Diffusion Problem

**Olli Mali**

**Abstract** In this paper, a posteriori error estimates of functional type for a stationary diffusion problem with nonsymmetric coefficients are derived. The estimate is guaranteed and does not depend on any particular numerical method. An algorithm for the global minimization of the error estimate with respect to an auxiliary function over some finite dimensional subspace is presented. In numerical tests, global minimization is done over the subspace generated by Raviart-Thomas elements. The improvement of the error bound due to the $p$-refinement of these spaces is investigated.

## 1 Introduction

In this paper, we derive a posteriori error estimates of the functional type for a class of elliptic problems with nonsymmetric coefficients. Since mid 90's (see [8]), estimates of this type has been derived for a wide range of problems (see, e.g., monographs [5, 6, 9] and references therein). However, the case of a stationary diffusion problem, where coefficients are not symmetric has not been studied before. Problems of this type are not very typical among other elliptic equations but they arise in certain models (see, e.g., [1, 2]). It is shown that the derived estimate has the standard properties of a deviation estimate for a linear problem, i.e., it is guaranteed and computable. The derivation of the estimate is based on the method of integral identities and a special case of Cauchy-Schwartz-Bunyakovsky inequality.

O. Mali (✉)
Department of Mathematical Information Technology, University of Jyvaskyla,
P.O. Box 35, FI-40014 Jyvaskyla, Finland
e-mail: olli.mali@jyu.fi



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_2

Consider the Poisson problem

$$-\operatorname{div} \mathbf{A}\nabla u = f \quad \text{in } \Omega \subset \mathbb{R}^d \tag{1}$$

$$u = 0 \quad \text{on } \Gamma := \partial\Omega, \tag{2}$$

where $\Omega$ a is simply connected domain with a Lipschitz-continuous boundary, $f \in L^2(\Omega)$, and $\mathbf{A} \in L_\infty(\Omega, \mathbb{R}^{d\times d})$ is strictly positive definite, bounded, and has a bounded inverse $\mathbf{A}^{-1} \in \mathbb{R}^{d\times d}$ in $\Omega$. Moreover, $\mathbf{A}$ is positive definite, i.e., there exists constant $\underline{c} > 0$ such that

$$(\mathbf{A}\boldsymbol{\xi}, \boldsymbol{\xi})_{\mathbb{R}^d} \geq \underline{c}\|\boldsymbol{\xi}\|^2_{\mathbb{R}^d}, \quad \forall \boldsymbol{\xi} \in \mathbb{R}^d, \text{ a.e. in } \Omega. \tag{3}$$

The generalized solution $u \in H^1_0(\Omega)$ satisfies the integral identity,

$$(\mathbf{A}\nabla u, \nabla w)_{L^2(\Omega,\mathbb{R}^d)} = (f, w)_{L^2(\Omega)}, \quad \forall w \in H^1_0(\Omega). \tag{4}$$

## 2 Error Majorant

For symmetric problems with $\mathbf{A} \in L_\infty(\Omega, \mathbb{R}^{d\times d}_{\text{sym}})$ the respective guaranteed upper bounds (error majorants) have been presented in [5, 6, 9] and other publications cited therein. It has the form

$$\overline{\mathfrak{M}}(v, \mathbf{y}) := (\mathbf{A}\nabla v - \mathbf{y}, \nabla v - \mathbf{A}^{-1}\mathbf{y})^{1/2}_{L^2(\Omega,\mathbb{R}^d)} + \frac{C_F}{\sqrt{\underline{c}}}\|\operatorname{div}\mathbf{y} + f\|_{L^2(\Omega)},$$

where $v \in H^1_0(\Omega)$, $\mathbf{y} \in H(\operatorname{div}, \Omega)$, and $C_F$ is the constant in the Friedrichs inequality

$$\|w\|_{L^2(\Omega)} \leq C_F\|\nabla w\|_{L^2(\Omega,\mathbb{R}^d)}, \quad \forall w \in H^1_0(\Omega). \tag{5}$$

A special case of the Cauchy-Schwartz-Bunyakovsky inequality presented below is required to obtain an analogous error estimate in the nonsymmetric case.

**Lemma 1** *Let $\mathscr{U}$ be a Hilbert space whose field is real numbers, $A : \mathscr{U} \to \mathscr{U}$ is continuous, bounded, strictly positive definite, and has a continuous inverse $A^{-1}$. Moreover,*

$$B := (\mathrm{Id} + A^T A^{-1})^{-1}$$

*is continuous and bounded. Then,*

$$(y, q)_{\mathscr{U}} \leq 2(Ay, y)^{1/2}_{\mathscr{U}}(A^{-1}Bq, Bq)^{1/2}_{\mathscr{U}}, \quad \forall y, q \in \mathscr{U}. \tag{6}$$

*Proof* Since $A$ is strictly positive definite,

$$0 \le (A(y - \gamma A^{-1}q), y - \gamma A^{-1}q)_{\mathscr{U}} \\ = (Ay, y)_{\mathscr{U}} - \gamma(y, (\mathrm{Id} + A^T A^{-1})q)_{\mathscr{U}} + \gamma^2 (A^{-1}q, q)_{\mathscr{U}}.$$

Selecting (assume $y \neq 0$ and $q \neq 0$, otherwise (6) holds trivially)

$$\gamma = \frac{2(Ay, y)_{\mathscr{U}}}{(y, (\mathrm{Id} + A^T A^{-1})q)_{\mathscr{U}}}$$

yields

$$(y, (\mathrm{Id} + A^T A^{-1})q)^2_{\mathscr{U}} \le 4(Ay, y)_{\mathscr{U}} (A^{-1}q, q)_{\mathscr{U}},$$

where setting $q = Bq = (\mathrm{Id} + A^T A^{-1})^{-1} q$ leads to (6).

**Theorem 1** *Let $v \in H^1_0(\Omega)$ and $u$ be the solution of (4). Then*

$$(\mathbf{A}\nabla(u - v), \nabla(u - v))^{1/2}_{L^2(\Omega, \mathbb{R}^d)} \le \overline{\mathfrak{M}}(v, \mathbf{y}), \quad \forall \mathbf{y} \in H(\mathrm{div}, \Omega),$$

*where*

$$\overline{\mathfrak{M}}(v, \mathbf{y}) := 2(\mathbf{A}^{-1}\mathbf{B}(\mathbf{y} - \mathbf{A}\nabla v), \mathbf{B}(\mathbf{y} - \mathbf{A}\nabla v))^{1/2}_{L^2(\Omega, \mathbb{R}^d)} + \frac{C_F}{\sqrt{\underline{c}}} \| \mathrm{div}\, \mathbf{y} + f \|_{L^2(\Omega)} \tag{7}$$

*and*

$$\mathbf{B} := (\mathbf{I} + \mathbf{A}^T \mathbf{A}^{-1})^{-1}.$$

*The constants $C_F$ and $\underline{c}$ are defined in (5) and (3), respectively.*

*Proof* Subtracting $(\mathbf{A}\nabla v, \nabla w)_{L^2(\Omega, \mathbb{R}^d)}$ from both sides of (4) and applying the integration by parts formula

$$(\mathbf{y}, \nabla w)_{L^2(\Omega, \mathbb{R}^d)} = (-\mathrm{div}\, \mathbf{y}, w)_{L^2(\Omega)}, \quad \forall \mathbf{y} \in H(\mathrm{div}, \Omega),\ w \in H^1_0(\Omega)$$

yields

$$(\mathbf{A}\nabla(u - v), \nabla w)_{L^2(\Omega, \mathbb{R}^d)} = (\mathbf{y} - \mathbf{A}\nabla v, \nabla w)_{L^2(\Omega, \mathbb{R}^d)} + (\mathrm{div}\, \mathbf{y} + f, w)_{L^2(\Omega)}.$$

The first term can be estimated from above by (6), where $\mathscr{U} := L^2(\Omega, \mathbb{R}^d)$ and $A := \mathbf{A}$. The second term is estimated from above by the Hölder inequality, (5), and (3), which leads to

$$
\begin{aligned}
&(\mathbf{A}\nabla(u-v),\nabla w)_{L^2(\Omega,\mathbb{R}^d)}\\
&\qquad\le 2(\mathbf{A}^{-1}\mathbf{B}(\mathbf{y}-\mathbf{A}\nabla v),\mathbf{B}(\mathbf{y}-\mathbf{A}\nabla v))^{1/2}_{L^2(\Omega,\mathbb{R}^d)}(\mathbf{A}\nabla w,\nabla w)^{1/2}_{L^2(\Omega,\mathbb{R}^d)}\\
&\qquad+\frac{C_F}{\sqrt{\underline{c}}}\|\operatorname{div}\mathbf{y}+f\|_{L^2(\Omega)}(\mathbf{A}\nabla w,\nabla w)^{1/2}_{L^2(\Omega,\mathbb{R}^d)}.
\end{aligned}
$$

Setting $w=u-v$ leads to (7).

*Remark 1* Two parts of the majorant are related to the violations of the duality relation and the equilibrium condition, respectively. They are denoted by

$$
\begin{aligned}
\overline{\mathfrak{M}}_{\mathrm{Dual}} &:= (\mathbf{A}^{-1}\mathbf{B}(\mathbf{y}-\mathbf{A}\nabla v),\mathbf{B}(\mathbf{y}-\mathbf{A}\nabla v))^{1/2}_{L^2(\Omega,\mathbb{R}^d)},\\
\overline{\mathfrak{M}}_{\mathrm{Equi}} &:= \|\operatorname{div}\mathbf{y}+f\|_{L^2(\Omega)}.
\end{aligned}
$$

## 3 Global Minimization of the Error Majorant

Squaring and applying the Young's inequality yields a quadratic form of the majorant, which is more suitable for the minimization over $\mathbf{y}$.

**Corollary 1** *Let $v\in H^1_0(\Omega)$ and $u$ be the solution of (4), then,*

$$
(\mathbf{A}\nabla(u-v),\nabla(u-v))_{L^2(\Omega,\mathbb{R}^d)}\le\overline{\mathfrak{M}}^2(v,\mathbf{y},\beta),\quad\forall\mathbf{y}\in H(\operatorname{div},\Omega),\ \beta>0,
$$

*where*

$$
\begin{aligned}
\overline{\mathfrak{M}}^2(v,\mathbf{y},\beta) &:= 4(1+\beta)(\mathbf{A}^{-1}\mathbf{B}(\mathbf{y}-\mathbf{A}\nabla v),\mathbf{B}(\mathbf{y}-\mathbf{A}\nabla v))_{L^2(\Omega,\mathbb{R}^d)}\\
&\quad+\frac{1+\beta}{\beta}\frac{C_F^2}{\underline{c}}\|\operatorname{div}\mathbf{y}+f\|^2_{L^2(\Omega)}.
\end{aligned}
\tag{8}
$$

**Corollary 2** *The minimizers*

$$
\begin{aligned}
\overline{\mathfrak{M}}^2(v,\hat{\mathbf{y}},\beta) &= \min_{\mathbf{y}\in H(\operatorname{div},\Omega)}\overline{\mathfrak{M}}^2(v,\mathbf{y},\beta),\\
\overline{\mathfrak{M}}^2(v,\mathbf{y},\hat{\beta}) &= \min_{\beta>0}\overline{\mathfrak{M}}^2(v,\mathbf{y},\beta)
\end{aligned}
$$

*satisfy*

$$
\begin{aligned}
&\frac{C_F^2}{\underline{c}}(\operatorname{div}\hat{\mathbf{y}},\operatorname{div}\mathbf{q})_{L^2(\Omega)}+2\beta\left((\mathbf{A}^{-1}\mathbf{B}\mathbf{q},\mathbf{B}\hat{\mathbf{y}})_{L^2(\Omega,\mathbb{R}^d)}+(\mathbf{A}^{-1}\mathbf{B}\hat{\mathbf{y}},\mathbf{B}\mathbf{q})_{L^2(\Omega,\mathbb{R}^d)}\right),\\
&=-\frac{C_F^2}{\underline{c}}(f,\operatorname{div}\mathbf{q})_{L^2(\Omega)}+2\beta\left((\mathbf{A}^{-1}\mathbf{B}\mathbf{q},\mathbf{B}\mathbf{A}\nabla v)_{L^2(\Omega,\mathbb{R}^d)}+(\mathbf{A}^{-1}\mathbf{B}\mathbf{A}\nabla v,\mathbf{B}\mathbf{q})_{L^2(\Omega,\mathbb{R}^d)}\right),\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall\mathbf{q}\in H(\operatorname{div},\Omega)
\end{aligned}
\tag{9}
$$

*and*

$$\hat{\beta} = \frac{\frac{C_F}{\sqrt{\underline{c}}} \| \operatorname{div} \mathbf{y} + f \|_{L^2(\Omega)}}{2(\mathbf{A}^{-1}\mathbf{B}(\mathbf{y} - \mathbf{A}\nabla v), \mathbf{B}(\mathbf{y} - \mathbf{A}\nabla v))^{1/2}_{L^2(\Omega, \mathbb{R}^d)}}, \tag{10}$$

*respectively.*

*Proof* The functional $\overline{\mathfrak{M}}^2(v, \mathbf{y}, \beta)$ is quadratic and convex w.r.t. $\mathbf{y}$. Thus the necessary and sufficient condition for the minimizer $\hat{\mathbf{y}}$ is

$$\frac{\mathrm{d}}{\mathrm{d}t} \overline{\mathfrak{M}}^2(v, \hat{\mathbf{y}} + t\mathbf{q}, \beta)\Big|_{t=0} = 0, \quad \forall \mathbf{q} \in H(\operatorname{div}, \Omega),$$

which leads to (9). Similarly,

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \overline{\mathfrak{M}}^2(v, \mathbf{y}, \hat{\beta}) = 0$$

yields (10).

*Remark 2* If $\mathbf{A}$ is symmetric, then (9) reduces to

$$\frac{C_F^2}{\underline{c}} \int_\Omega \operatorname{div} \hat{\mathbf{y}} \operatorname{div} \mathbf{q} d\mathbf{x} + \beta \int_\Omega \mathbf{A}^{-1}\hat{\mathbf{y}} \cdot \mathbf{q} d\mathbf{x} = -\frac{C_F^2}{\underline{c}} \int_\Omega f \operatorname{div} \mathbf{q} d\mathbf{x} + \beta \int_\Omega \nabla v \cdot \mathbf{q} d\mathbf{x} \quad \forall \mathbf{q} \in H(\operatorname{div}, \Omega).$$

There are many alternatives how to compute the value of the majorant (see, e.g., [5, Chap. 3]). Here, the global minimization of the majorant over finite dimensional subspace is presented. The minimization is done iteratively by solving (9) and (10) subsequently.

Let $\mathbf{y} = \sum_{j=1}^N c_j \boldsymbol{\phi}_j$ and $Q_h := \operatorname{span}(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N) \subset H(\operatorname{div}, \Omega)$, i.e., $\boldsymbol{\phi}_j$ ($j \in \{1, \ldots, N\}$) are the global basis functions. Then (9) leads to a system of linear equations

$$\left(\frac{C_F^2}{\underline{c}}\mathbf{S} + 2\beta\mathbf{M}\right)\mathbf{c} = -\frac{C_F^2}{\underline{c}}\mathbf{b} + 2\beta\mathbf{z}, \tag{11}$$

where

$$S_{ij} := (\operatorname{div} \boldsymbol{\phi}_j, \operatorname{div} \boldsymbol{\phi}_i)_{L^2(\Omega)}, \tag{12}$$

$$M_{ij} := (\mathbf{A}^{-1}\mathbf{B}\boldsymbol{\phi}_j, \mathbf{B}\boldsymbol{\phi}_i)_{L^2(\Omega, \mathbb{R}^d)} + (\mathbf{A}^{-1}\mathbf{B}\boldsymbol{\phi}_i, \mathbf{B}\boldsymbol{\phi}_j)_{L^2(\Omega, \mathbb{R}^d)}, \tag{13}$$

$$b_i := (f, \operatorname{div} \boldsymbol{\phi}_i)_{L^2(\Omega)}, \tag{14}$$

$$z_i := (\mathbf{A}^{-1}\mathbf{B}\boldsymbol{\phi}_i, \mathbf{B}\mathbf{A}\nabla v)_{L^2(\Omega,\mathbb{R}^d)} + (\mathbf{A}^{-1}\mathbf{B}\mathbf{A}\nabla v, \mathbf{B}\boldsymbol{\phi}_i)_{L^2(\Omega,\mathbb{R}^d)}, \tag{15}$$

and $\mathbf{c} \in \mathbb{R}^N$ is the (column) vector of unknown coefficients. The natural choice is to generate $Q_h$ using Raviart-Thomas elements (see [7]). The global minimization procedure for $\overline{\mathfrak{M}}^2$ is described in Algorithm 1.

**Algorithm 1**: Computation of the majorant for the problem (1)–(2)

**Input**: $v$ {approximate solution}, $\mathbf{A}$, {diffusion coefficient matrix} $f$, {RHS of the problem}, $C_F$, {Constant in (5)}, $\underline{c}$, {Constant in (3)}, $I_{\max}$ {maximum number of iterations}, $\epsilon$ {stopping criteria for $\overline{\mathfrak{M}}$}
Generate $\mathbf{S}$, $\mathbf{M}$, $\mathbf{b}$, and $\mathbf{z}$ in (12)–(15).
Compute norms $\|f\|$ and $\|\nabla v\|$.
Set $\beta_1 := 1$, $\overline{\mathfrak{M}}_k = \infty$ and $k = 0$. {initialize parameters}
**while** $k < I_{\max}$ **and** $\frac{\overline{\mathfrak{M}}_{k+1} - \overline{\mathfrak{M}}_k}{\overline{\mathfrak{M}}_k} > \epsilon$ **do**
  $k = k + 1$
  Solve $\mathbf{c}_{k+1}$ from $\left(\frac{C_F^2}{\underline{c}}\mathbf{S} + 2\beta_k\mathbf{M}\right)\mathbf{c}_{k+1} = -\frac{C_F^2}{\underline{c}}\mathbf{b} + 2\beta_k\mathbf{z}$.
  $\overline{\mathfrak{M}}_{k+1}^{\text{Equi}} = \sqrt{\mathbf{c}_{k+1}^T\mathbf{S}\mathbf{c}_{k+1} + 2\mathbf{c}_{k+1}^T\mathbf{b} + \|f\|^2}$
  $\overline{\mathfrak{M}}_{k+1}^{\text{Dual}} = \sqrt{\mathbf{c}_{k+1}^T\mathbf{M}\mathbf{c}_{k+1} - 2\mathbf{c}_{k+1}^T\mathbf{z} + \|\nabla v\|^2}$
  $\beta_{k+1} = \frac{C_F\overline{\mathfrak{M}}_{k+1}^{\text{Equi}}}{2\sqrt{\underline{c}}\overline{\mathfrak{M}}_{k+1}^{\text{Dual}}}$
  $\overline{\mathfrak{M}}_{k+1} = 2\overline{\mathfrak{M}}_{k+1}^{\text{Dual}} + \frac{C_F}{\sqrt{\underline{c}}}\overline{\mathfrak{M}}_{k+1}^{\text{Equi}}$
**end while**
$\mathbf{y} = \sum_{j=1}^N c_{kj}\boldsymbol{\phi}_j$
**Output**: $\overline{\mathfrak{M}}_{k+1}$ {Upper bound for the approximation error}, $\mathbf{y}$ {Approximation of the minimizer}

*Remark 3* Note that in Algorithm 1, the global matrices $\mathbf{S}$ and $\mathbf{M}$ have to be assembled only once. The coefficient matrix in (11) is symmetric regardless of the fact that $\mathbf{A}$ is not.

## 4 Numerical Tests

Algorithm 1 is very convenient to implement using any finite element software, e.g., FEniCS [4] and FREEFEM++ [3]), which allows user to define problems using weak forms. This is true for all estimates of the functional type presented in [5, 6, 9]. The following tests are computed using FEniCS finite element package. Here, we apply Algorithm 1 to estimate the error of a finite element approximation for a test example, where the exact solution is known.

*Example 1* Let $\Omega = (0,1) \times (0,1)$, $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $u(x_1, x_2) = \sin(k_1 \pi x_1) \sin(k_2 \pi x_2)$, and

$$f(x_1, x_2) = \pi^2 \Big( (a+d) k_1^2 \sin(k_1 \pi x_1) \sin(k_2 \pi x_2) - (b+c) k_1 k_2 \cos(k_1 \pi x_1) \cos(k_2 \pi x_2) \Big).$$

Select $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$, then $\underline{c} = 2$, $\mathbf{A}^{-1} = \frac{1}{6} \begin{pmatrix} 3 & -1 \\ 0 & 2 \end{pmatrix}$ and $\mathbf{B} = \frac{1}{23} \begin{pmatrix} 11 & 2 \\ -3 & 12 \end{pmatrix}$.

The approximate solution $v \in V_h$ of Example 1 is computed on a mesh $\mathscr{T}_h$, using triangular Courant elements of the order $p_1$. The space $Q_h$ is generated using the Raviart-Thomas elements of order $p_2$ on the same mesh. The amount of global degrees of freedom are denoted by $N_1 = \dim(V_h)$ and $N_2 = \dim(Q_h)$. The efficiency index of the majorant is

$$I_{\text{eff}} := \sqrt{\frac{\overline{\mathfrak{M}}^2(v, \mathbf{y}, \beta)}{(\mathbf{A}\nabla(u-v), \nabla(u-v))_{L^2(\Omega, \mathbb{R}^d)}}}. \tag{16}$$

The majorant is computed for different meshes with $k_1 = 1$, $k_2 = 1$, and $p_1 = 1$ in Table 1.

The efficiency of the majorant and the number of iterations (in Algorithm 1, $\varepsilon = 10^{-6}$) do not depend on the mesh size. For $p_2 = 2$ and $p_2 = 3$, $Q_h$ can

**Table 1** Example 1: $k_1 = 1$, $k_2 = 1$, and $p_1 = 1$

| $N_1$ | $p_2$ | $N_2$ | $k$ | $\overline{\mathfrak{M}}^2(v, \mathbf{y}_k, \beta_k)$ | $\overline{\mathfrak{M}}_k^{\text{Dual}}$ | $\overline{\mathfrak{M}}_k^{\text{Equi}}$ | $I_{\text{eff}}$ |
|---|---|---|---|---|---|---|---|
| 441 | 1 | 1240 | 3 | 1.76E+00 | 2.46E−02 | 2.06E+00 | 6.6480 |
| 441 | 2 | 4080 | 3 | 3.15E−01 | 1.78E−02 | 2.09E−03 | 1.1858 |
| 441 | 3 | 8520 | 4 | 2.68E−01 | 1.78E−02 | 1.07E−06 | 1.0090 |
| 1681 | 1 | 4880 | 3 | 8.85E−01 | 6.23E−03 | 5.17E−01 | 6.6452 |
| 1681 | 2 | 16160 | 3 | 1.45E−01 | 4.44E−03 | 1.31E−04 | 1.0920 |
| 1681 | 3 | 33840 | 4 | 1.33E−01 | 4.44E−03 | 1.68E−08 | 1.0023 |
| 6561 | 1 | 19360 | 2 | 4.43E−01 | 1.56E−03 | 1.29E−01 | 6.6445 |
| 6561 | 2 | 64320 | 3 | 6.97E−02 | 1.11E−03 | 8.20E−06 | 1.0458 |
| 6561 | 3 | 134880 | 3 | 6.66E−02 | 1.11E−03 | 2.62E−10 | 1.0006 |
| 14641 | 1 | 43440 | 2 | 2.95E−01 | 6.95E−04 | 5.75E−02 | 6.6443 |
| 14641 | 2 | 144480 | 3 | 4.58E−02 | 4.93E−04 | 1.62E−06 | 1.0305 |
| 14641 | 3 | 303120 | 3 | 4.44E−02 | 4.93E−04 | 2.30E−11 | 1.0003 |
| 40401 | 1 | 120400 | 2 | 1.77E−01 | 2.50E−04 | 2.07E−02 | 6.6443 |
| 40401 | 2 | 400800 | 3 | 2.71E−02 | 1.78E−04 | 2.10E−07 | 1.0183 |
| 40401 | 3 | 841200 | 3 | 2.67E−02 | 1.78E−04 | 1.07E−12 | 1.0002 |

**Table 2** Example 1: $k_1 = 2$, $k_2 = 3$, and $p_1 = 2$

| $N_1$ | $p_2$ | $N_2$ | $k$ | $\overline{\mathfrak{M}}^2(v, \mathbf{y}_k, \beta_k)$ | $\overline{\mathfrak{M}}_k^{\text{Dual}}$ | $\overline{\mathfrak{M}}_k^{\text{Equi}}$ | $I_{\text{eff}}$ |
|---|---|---|---|---|---|---|---|
| 1681 | 1 | 1240 | 3 | 2.60E+01 | 3.94E−01 | 6.10E+02 | 189.9638 |
| 1681 | 2 | 4080 | 3 | 2.15E+00 | 6.05E−03 | 3.92E+00 | 15.6634 |
| 1681 | 3 | 8520 | 2 | 2.53E−01 | 4.71E−03 | 1.26E−02 | 1.8496 |
| 6561 | 1 | 4880 | 3 | 1.32E+01 | 9.51E−02 | 1.56E+02 | 380.2599 |
| 6561 | 2 | 16160 | 3 | 5.41E−01 | 3.89E−04 | 2.49E−01 | 15.6199 |
| 6561 | 3 | 33840 | 3 | 4.93E−02 | 3.00E−04 | 1.99E−04 | 1.4258 |
| 25921 | 1 | 19360 | 3 | 6.60E+00 | 2.36E−02 | 3.94E+01 | 760.6287 |
| 25921 | 2 | 64320 | 2 | 1.35E−01 | 2.45E−05 | 1.56E−02 | 15.6082 |
| 25921 | 3 | 134880 | 3 | 1.05E−02 | 1.88E−05 | 3.12E−06 | 1.2139 |
| 58081 | 1 | 43440 | 3 | 4.40E+00 | 1.05E−02 | 1.75E+01 | 1140.9677 |
| 58081 | 2 | 144480 | 2 | 6.02E−02 | 4.84E−06 | 3.09E−03 | 15.6060 |
| 58081 | 3 | 303120 | 2 | 4.41E−03 | 3.72E−06 | 2.74E−07 | 1.1430 |

practically present the exact minimizer of the majorant, since the efficiency index is almost one. Note that in this case $\overline{\mathfrak{M}}^{\text{Dual}}$ is almost the exact error and $\overline{\mathfrak{M}}^{\text{Equi}}$ vanishes. Results of a similar experiment in the case $k_1 = 2$, $k_2 = 3$, and $p_1 = 2$ are depicted in Table 2. It is easy to see that lowest order Raviart-Thomas elements are not able to represent the minimizer properly and in the case $p_2 = 1$, the efficiency index of the majorant is poor. Again, in the $p$-refined spaces the estimate improves significantly.

*Example 2* Let $\Omega := (0, 1) \times (0, 1) \times (0, 1)$, $f(x_1, x_2, x_3) = x_1 x_2 x_3$, and

$$\mathbf{A} = \begin{pmatrix} 1000 & 20 & -500 \\ -3 & 30 & 16 \\ 2 & 0 & 3 \end{pmatrix}.$$

Then,

$$\mathbf{A}^{-1} \approx \begin{pmatrix} 7.4490978E-04 & -4.9660652E-04 & 1.2680020E-01 \\ 3.3934779E-04 & 3.3107104E-02 & -1.2001324E-01 \\ -4.9660652E-04 & 3.3107101E-04 & 2.4879987E-01 \end{pmatrix}$$

and

$$\mathbf{B} \approx \begin{pmatrix} 1.0126139 & -0.4980245 & 2.0416897 \\ -0.0160603 & 0.5154516 & -0.0408795 \\ -0.0060666 & 0.009230 & -0.0280656 \end{pmatrix}.$$

In Example 2, the exact solution is not known. Instead a reference solution was computed using third order Courant type elements with 29791 global degrees of freedom. The approximations were computed using linear tetrahedral Courant type elements and the space $Q_h$ is generated using tetrahedral Raviart-Thomas elements

**Table 3** Example 2, $p_1 = 1$

| $N_1$ | $p_2$ | $N_2$ | $k$ | $\overline{\mathfrak{M}}^2(v, \mathbf{y}_k, \beta_k)$ | $\overline{\mathfrak{M}}_k^{\text{Dual}}$ | $\overline{\mathfrak{M}}_k^{\text{Equi}}$ | $I_{\text{eff}}$ |
|---|---|---|---|---|---|---|---|
| 125 | 1 | 864 | 4 | 4.67E−02 | 1.15E−05 | 1.57E−03 | 10.0122 |
| 125 | 2 | 3744 | 3 | 8.47E−03 | 6.99E−06 | 9.43E−06 | 1.8164 |
| 125 | 3 | 9792 | 3 | 5.26E−03 | 6.68E−06 | 8.94E−09 | 1.1284 |
| 343 | 1 | 2808 | 3 | 3.12E−02 | 5.81E−06 | 6.85E−04 | 9.2258 |
| 343 | 2 | 12312 | 3 | 5.24E−03 | 3.65E−06 | 1.86E−06 | 1.5489 |
| 343 | 3 | 32400 | 3 | 3.81E−03 | 3.65E−06 | 7.85E−10 | 1.1241 |
| 729 | 1 | 6528 | 3 | 2.35E−02 | 3.48E−06 | 3.83E−04 | 9.1082 |
| 729 | 2 | 28800 | 3 | 3.78E−03 | 2.21E−06 | 5.88E−07 | 1.4642 |
| 729 | 3 | 76032 | 3 | 2.97E−03 | 2.64E−06 | 1.40E−10 | 1.1527 |
| 1331 | 1 | 12600 | 3 | 1.88E−02 | 2.31E−06 | 2.44E−04 | 7.8120 |
| 1331 | 2 | 55800 | 3 | 2.94E−03 | 1.47E−06 | 2.41E−07 | 1.2208 |

of order $p_2$. The results were depicted on Table 3 and they show similar characteristics as in the two dimensional example.

## 5 Summary

An upper functional deviation estimate (majorant) for nonsymmetric stationary diffusion problem is derived. An algorithm for the global minimization of the majorant over a finite dimensional subspace is presented and tested. The efficiency of the majorant depends on the particular problem (i.e., the exact solution) and the relation of spaces $V_h$ and $Q_h$. The question is that how accurately $V_h$ can represent $u$ (in the energy norm) in comparison with the ability of $Q_h$ to represent the minimizer of the majorant. If $Q_h$ is "better", then the estimate is very accurate and the other way round. The crude overestimation in Table 2 shows that using a "worse" space for the computation of the minimizer.

## References

1. V.V. Denisenko, Variational methods for elliptic boundary value problems that describe transport processes with nonsymmetric tensor coefficients. Zh. Prikl. Mekh. i Tekhn. Fiz. **3**, 69–75 (1989). Translation in J. Appl. Mech. Tech. Phys. **30**(3), 404–410 (1989)
2. V.V. Denisenko, The energy method for three-dimensional elliptic equations with nonsymmetric tensor coefficients. Sibirsk. Mat. Zh., **38**(6), 1267–1281, ii (1997). Translation in Siberian Math. J. **38**(6), 1099–1111 (1997)
3. F. Hecht, New development in freefem++. J. Numer. Math. **20**(3–4), 251–265 (2012)

4. A. Logg, K.-A. Mardal, G. Wells (eds.), *Automated Solution of Differential Equations by the Finite Element Method*. Lecture Notes in Computational Science and Engineering, vol. 84 (Springer, Berlin, 2012)
5. O. Mali, P. Neittaanmäki, S. Repin, *Accuracy Verification Methods: Theory and Algorithms*. Computational Methods in Applied Sciences, vol. 32 (Springer, Dordrecht, 2014)
6. P. Neittaanmäki, S. Repin, *Reliable Methods for Computer Simulation: Error Control and a Posteriori Estimates* (Elsevier, Amsterdam, 2004)
7. P.-A. Raviart, J.M. Thomas, Primal hybrid finite element methods for 2nd order elliptic equations. Math. Comp. **31**(138), 391–413 (1977)
8. S. Repin, A posteriori error estimates for approximate solutions of variational problems with strongly convex functionals. Probl. Mat. Anal. **17**, 227–237 (1997). Translation in J. Math. Sci. **97**(4), 4311–4328 (1999)
9. S. Repin, *A Posteriori Estimates for Partial Differential Equations*. Radon Series on Computational and Applied Mathematics, vol. 4 (Walter de Gruyter, Berlin, 2008)

# Error Estimates of Uzawa Iteration Method for a Class of Bingham Fluids

**Marjaana Nokka and Sergey Repin**

**Abstract** The paper is concerned with fully guaranteed and computable bounds of errors generated by Uzawa type methods for variational problems in the theory of visco-plastic fluids. The respective estimates have two forms. The first form contains global constants (such as the constant in the Friedrichs inequality for the respective domain), and the second one is based upon decomposition of the domain into a collection of subdomains and uses local constants associated with subdomains.

**Keywords** Bingham fluid · Uzawa algorithm · A posteriori estimates

**Mathematical Subject Classification:** 65N15

## 1 Introduction

Models of fluids with nonlinear viscosity are widely used in engineering applications and natural sciences (e.g., for modeling of creamy substances, flow of blood, lymph or waxy crude oil) [4, 11, 12].

Mathematical models of Bingham type fluids started receiving serious attention in the 60s and the 70s in the framework of general studies related to variational inequalities (see, e.g., [7]). Variational posings of stationary models were studied since [17] in many publications. A consequent exposition of results related to regularity of solutions can be found in, e.g., [9, 10]. In [2, 3, 6, 13–15], the reader will find a

M. Nokka (✉) · S. Repin
Department of Mathematical Information Technology, University of Jyväskylä,
P.O. Box 35, 40014 Jyväskylä, Finland
e-mail: marjaana.nokka@jyu.fi

S. Repin
St. Petersburg State Polytechnical University, Polytechnicheskaya 29,
St. Petersburg, Russia
e-mail: serepin@jyu.fi



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_3

deeply elaborated theory of numerical approximations and respective computational algorithms. For non-stationary problems there are also several known approaches (e.g., the operator splitting method [6, 14, 23]).

Uzawa type methods are often used for solving nonlinear problems generated by models of viscous fluids. Each step of this iteration method includes solving a suitable linear problem and redefinition of the Lagrange multiplier (which is typically reduced to projection on a certain convex set). General theory of Uzawa type approximations is well developed (see, e.g., [13]) and the conditions sufficient to guarantee that the respective iteration sequence converges to the exact solution are known. However, in practical computations we need to know and explicitly control the error associated with a particular iteration. In other words, we wish to have explicitly computable and realistic estimates of errors generated by Uzawa iterations. The goal of this paper is to deduce such type estimates for a simple stationary models with Bingham type dissipative potential. Simple forms of the Uzawa iteration algorithm may generate sequences with rather low convergence rate. Therefore, in practice, advanced forms of this algorithm are often used (e.g., one of the most known modifications is known as the augmented Lagrangian method). In this paper we do not consider these algorithms. This analysis will be presented in the next publication, which will use results of the present paper.

## 2 Governing Equations

Evolution of a generalized Newtonian fluid in a bounded Lipschitz domain $\Omega \in \mathbb{R}^d$, $(d = 2, 3)$ is described by the differential equation of motion

$$u_t - \operatorname{div} \sigma + (u \cdot \nabla u) = f - \nabla p \quad \text{in } \Omega, \tag{1}$$

incompressibility condition

$$\operatorname{div} u = 0 \quad \text{in } \Omega, \tag{2}$$

and the differential inclusion

$$\sigma \in \partial \pi(\nabla u) \tag{3}$$

that reflects mechanical properties of the fluid. The system (1)–(3) should be supplied with proper initial and boundary conditions (see, e.g., [7] for a consequent discussion of the mathematical statement). Here that $\partial$ denotes the subdifferential and $\pi$ is the *dissipative potential* of the fluid. Many physically interesting models are described by the dissipative potentials in the form

$$\pi(\varepsilon) = \frac{1}{m} \nu |\varepsilon|^m + k_* |\varepsilon|$$

where $m > 1$ is associated with the energy growth at infinity, $\nu > 0$ is the viscosity parameter, and $k_* \geq 0$ is the plasticity parameter. Models with dissipative potentials of this type are known as models of Bingham fluids. In particular, the most known Bingham fluid model is described by the potential

$$\frac{1}{2}\nu|\varepsilon|^2 + k_*|\varepsilon| \quad k_* > 0.$$

Nowadays it is commonly accepted that the system (1)–(3) adequately describes the behaviour of various nonlinear fluids (in particular, it is used for computer simulation of the blood flow, see, e.g., [4, 5, 8]).

Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded and connected domain with Lipschitz continuous boundary $\partial\Omega$. Below we consider a simple stationary anti-plane problem of the Bingham fluid in a long domain $\Omega \times ]0, L[$, with the potential

$$\pi(\varepsilon) = \frac{1}{2}A\varepsilon : \varepsilon + k_*|\varepsilon|,$$

where $k_* > 0$ and $A$ is a symmetric matrix satisfying the condition

$$c_1^2|\xi|^2 \leq A\xi \cdot \xi \leq c_2^2|\xi|^2 \quad \forall \xi \in \mathbb{R}^d, \; c_2 \geq c_1.$$

Henceforth, we use two equivalent norms generated by the matrix $A$:

$$|||v|||^2 := \int_\Omega A\nabla v \cdot \nabla v \mathrm{d}x \quad \text{and} \quad |||\eta|||_*^2 := \int_\Omega A^{-1}\eta \cdot \eta \, dx.$$

In this case (see, e.g., [7, 13]), the problem can be reduced to the variational inequality: Find the velocity $u \in V_0 + u_0$ such that

$$a(u, w - u) + \int_\Omega (j(\nabla w) - j(\nabla u))\mathrm{d}x \geq \ell(w - u) \quad \forall w \in V_0 + u_0, \tag{4}$$

where

$$V_0 := \left\{ v \in V := H^1(\Omega) \mid v = u_0 \text{ on } \partial\Omega \right\},$$

$u_0 \in V$ is given and defines Dirichlet type boundary conditions, $a : V \times V \to \mathbb{R}$ is a bilinear $V$-elliptic form

$$a(v, w) := \int_\Omega A\nabla v \cdot \nabla w \mathrm{d}x,$$

and $j : \mathbb{R}^d \to \mathbb{R}$ is a convex continuous function, which in our case is defined by the relation

$$j(\nabla w) = k_*|\nabla w|.$$

The inequality (4) is equivalent to the following variational problem: Find $u \in V_0 + u_0$ such that

$$J(u) = \inf_{w \in V_0 + u_0} J(w), \tag{5}$$

where

$$J(w) := \frac{1}{2} a(w, w) + \int_\Omega j(\nabla w) \mathrm{d}x - \ell(w).$$

Due to well-known results in the theory of variational calculus, the problem (5) associated with a strictly convex and lower semicontinuous functional is uniquely solvable (see, e.g., [7]).

## 3 First Form of the Majorant

Our analysis is based on functional type a posteriori estimates. For the considered class of problems (and many other variational inequalities) have been studied in [20–22]. Now, our goal is to obtain somewhat different forms of these estimates adapted to approximations generated by Uzawa method. Similar estimates have been recently presented in [18] for the Oseen problem.

**Theorem 1** *For any $v \in V_0 + u_0$, $\eta \in L^2(\Omega, \mathbb{R}^d)$ and $y \in L^2(\Omega, \mathbb{R}^d)$, the functional $M_{1\oplus}(v, y, \eta, \alpha, \beta)$ gives an upper bound of the deviation from the exact solution $u$ in terms of the energy norm, i.e.*

$$\begin{aligned} |||u - v|||^2 &\le M_{1\oplus}(v, y, \eta, \alpha, \beta) \\ &:= \frac{2}{\gamma} D_j(\nabla v, \eta) + \frac{1}{\alpha\gamma} |||A\nabla v + \eta - y|||_*^2 \\ &\quad + \frac{1}{\beta\gamma} \| \operatorname{Div} y + f \|_{-1,\Omega}^2, \end{aligned}$$

*where $\alpha, \beta > 0$, $\gamma := 2 - \alpha - \beta > 0$,*

$$\| \operatorname{Div} y + f \|_{-1,\Omega} := \sup_{w \in V_0(\Omega)} \frac{\int_\Omega (fw - y \cdot \nabla w) \mathrm{d}x}{\|\nabla w\|},$$

$$\begin{aligned} D_j(\nabla v, \eta) &:= \int_\Omega (j(\nabla v) + j^*(\eta) - \eta \cdot \nabla v) \mathrm{d}x \\ &= \begin{cases} \int_\Omega (k_* |\nabla v| - \eta \nabla v) \mathrm{d}x & \text{if } |\eta| \le k_*, \\ +\infty & \text{otherwise} \end{cases} \end{aligned}$$

*is the compound functional associated with $j$, and $j^* : \mathbb{R}^d \to \mathbb{R}$ is the Young-Fenchel conjugate of $j$.*

*Proof* Substituting $v \in V_0 + u_0$ into (4), we obtain

$$a(u - v, u - v) + \int_\Omega (j(\nabla v) - j(\nabla u))\mathrm{d}x \geq \ell(v - u) - a(v, u - v),$$

that implies

$$|||u - v|||^2 \leq \int_\Omega (j(\nabla v) - j(\nabla u))\mathrm{d}x + \ell(u - v) + a(v, v - u).$$

Let $\eta \in L^2(\Omega, \mathbb{R}^d)$. Due to the Fenchel-Young inequality we have

$$\int_\Omega -j(\nabla u)\mathrm{d}x \leq \int_\Omega (j^*(\eta) - \eta \cdot \nabla u)\mathrm{d}x.$$

Hence,

$$\begin{aligned} |||u - v|||^2 \leq\ & a(v, u - v) + \int_\Omega \eta \cdot \nabla(v - u)\mathrm{d}x \\ & + \int_\Omega (j(\nabla v) + j^*(\eta) - \eta \cdot \nabla v)\mathrm{d}x + \ell(u - v). \end{aligned}$$

For any $y \in L_2(\Omega, \mathbb{R}^d)$ the following identity holds:

$$\int_\Omega (w \operatorname{div} y + y \cdot \nabla w)\mathrm{d}x = 0, \quad \forall w \in V_0.$$

Therefore, we have

$$\begin{aligned} |||u - v|||^2 \leq\ & D_j(\nabla v, \eta) - \int_\Omega (f + \operatorname{div} y)(v - u)\mathrm{d}x \\ & + \int_\Omega (A\nabla v + \eta - y) \cdot \nabla(v - u)\mathrm{d}x. \end{aligned}$$

Notice that

$$\begin{aligned} \int_\Omega (A\nabla v + \eta - y) \cdot \nabla(v - u)\mathrm{d}x &\leq |||A\nabla v + \eta - y|||_* |||v - u||| \\ &\leq \frac{\alpha}{2}|||u - v|||^2 + \frac{1}{2\alpha}|||A\nabla v + \eta - y|||_*^2, \end{aligned} \tag{6}$$

where $\alpha > 0$. Also

$$\left| \int_\Omega (f + \operatorname{Div} y)(v - u)\mathrm{d}x \right| \le \| \operatorname{Div} y + f \|_{-1,\Omega} |||u - v||| \le \frac{\beta}{2} |||u - v|||^2 + \frac{1}{2\beta} \| \operatorname{Div} y + f \|^2_{-1,\Omega}, \quad (7)$$

where $\beta > 0$. By (6) and (7) we obtain the required result. □

## 4 Uzawa Type Algorithm

Numerical analysis of the variational problem (4) can be performed by different methods. Many of them are discussed in [2, 13, 15, 16] (see also numerous publications cited therein). The classical Uzawa algorithm for solving problem (4) has been introduced in [1] and systematically analysed in, e.g., [3, 24]. In the context of Bingham fluids it has been studied in [6]. Below we consider the simplest (classical) form of the Uzawa method.

Define

$$\lambda^0 \in K := \left\{ \lambda \in L^\infty(\Omega) \mid |\lambda| \le 1 \right\}$$

and generate the sequence $\{u^n, \lambda^n\}$, $n = 1, 2, \ldots$, by the following algorithm:

**Algorithm 1**

Step 1. For known $\lambda^n$ and given $\rho > 0$ compute $u^{n+1} \in V_0$ as a generalized solution of the problem:

$$\int_\Omega (A\nabla u^{n+1} \cdot \nabla w + k_* \lambda^n \cdot \nabla w - f w)\mathrm{d}x = 0, \ \forall w \in V_0. \quad (8)$$

Step 2. Define

$$\lambda^{n+1} = \Pi(\lambda^n + \rho k_* \nabla u^{n+1}),$$

where $\Pi : L^\infty \to K$ is the projection operator on the set $K$

$$\Pi(\lambda)(x) = \frac{\lambda(x)}{\max(1, |\lambda(x)|)}, \quad \text{a.e. in } \Omega.$$

Step 3. Set $n = n + 1$ and go to Step 1.

It is well known (see, e.g., [6]) that approximations generated by this algorithm converge to the exact solution (as $n \to \infty$) in the sense that

$$u^n \to u \quad \text{in } V$$

provided that

$$0 < \rho < \bar{\rho} := \frac{2c_1}{k_*^2}.$$

Suppose that the Eq. (8) has a solution $\{u, \lambda\} \in V_0 \times L^\infty(\Omega)$. It is well known that the pair $\{u, \lambda\}$ is a saddle point of the Lagrangian functional $L : V_0 \times L^\infty(\Omega) \to \mathbb{R}$ defined by

$$L(v, \mu) := \frac{1}{2}a(v, v) + \int_\Omega k_*\mu \cdot \nabla v \mathrm{d}x - \int_\Omega f v \mathrm{d}x.$$

We have

$$L(u, \mu) \le L(u, \lambda) \le L(v, \lambda) \quad \forall v \in V_0,\ \mu \in L^\infty(\Omega),$$

and

$$\{u, \lambda\} = \inf_{v \in V_0} \sup_{\mu \in L^\infty(\Omega)} L(v, \mu).$$

In order to deduce computable and realistic estimate of $u^n - u$ in terms of the energy norm, we use results of previous section. Set

$$\begin{aligned} v &= u^n, \\ \eta &= k_*\lambda^{n-1}, \\ y &= A\nabla u^n + k_*\lambda^{n-1}. \end{aligned}$$

In this case,

$$|||A\nabla u^n + k_*\lambda^{n-1} - y|||_*^2 = 0$$

and in view of (8), we find that

$$\| \operatorname{Div} y + f \|_{-1,\Omega} = \sup_{v \in V_0} \frac{\int_\Omega (-A\nabla u^{n+1} \cdot \nabla w - k_*\lambda^n \cdot \nabla w + f w)\mathrm{d}x}{\|\nabla w\|} = 0.$$

We use the estimate in Theorem 1 and let $\alpha, \beta \to 0$. Then, we arrive at the following result.

**Theorem 2** *Let $u^n$ be the exact solution computed the step n of the Uzawa algorithm. Then*

$$\begin{aligned} |||u - u^n|||^2 &\le \int_\Omega (k_*|\nabla u^n| - \lambda^{n-1} \cdot \nabla u^n)\mathrm{d}x \\ &:= M_{1\oplus}^{Uz}(u^n, \lambda^{n-1}). \end{aligned} \tag{9}$$

Since $\lambda^{n-1} \le 1$, we see that $M_{1\oplus}^{Uz}(u^n, \lambda^{n-1})$ is nonnegative.

The estimate (9) would give a complete answer to the question stated if we would have $u^n$ (exact solution of the boundary value problem in the first step of Uzawa

method). In practice, the problem is solved on a certain mesh $\mathcal{T}_h$ (as usual $h$ denotes the characteristic size of cells). For this case, we need an advanced form of the error majorant, which takes into account approximation errors. We are now concerned with the derivation of such a majorant.

Let $V_{0h}$ be finite dimensional subspace of $V_0$ and $u_h^n \in V_{0h} + u_0$ be an approximation of $u^n$ calculated in the $n$th step of Uzawa algorithm. We wish to estimate the difference between $u$ and $u_h^n$. We have

$$\begin{aligned}
|||u - u_h^n||| &\le |||u - u^n||| + |||u^n - u_h^n||| \\
&\le \int_\Omega (k_*|\nabla u^n| - \lambda^{n-1} \cdot \nabla u^n)\mathrm{d}x + |||u^n - u_h^n||| \\
&\le \int_\Omega (k_*|\nabla(u^n - u_h^n + u_h^n)| - \lambda^{n-1} \cdot \nabla(u^n - u_h^n + u_h^n))\, dx \\
&\le \int_\Omega (k_*|\nabla u_h^n| - \lambda^{n-1} \cdot \nabla u_h^n)\mathrm{d}x + 2k_*\|\nabla(u^n - u_h^n)\| + |||u^n - u_h^n||| \\
&\le \int_\Omega k_*(|\nabla u_h^n| - \lambda^{n-1} \cdot \nabla u_h^n)\mathrm{d}x + \left(\frac{2k_*}{c_1} + 1\right)|||u^n - u_h^n|||. \qquad (10)
\end{aligned}$$

In order to find the upper bound for the approximation error $|||u^n - u_h^n|||$ we follow the derivation method presented in [22].

Subtracting $\int_\Omega A\nabla u_h^n \mathrm{d}x$ from the $n$th step of the Uzawa algorithm (8), we have for all $w \in V_0$ and all $y \in Q := \{y \in L^2(\Omega, \mathbb{R}^d) \mid \mathrm{div}(y) \in L^2(\Omega)\}$

$$\begin{aligned}
\int_\Omega A\nabla(u^n - u_h^n) \cdot \nabla w \mathrm{d}x &= \int_\Omega (fw - k_*\lambda^{n-1} \cdot \nabla - A\nabla u_h^n \cdot \nabla w)\mathrm{d}x \\
&= \int_\Omega (fw + w\,\mathrm{div}\, y + (y - A\nabla u_h^n - k_*\lambda^{n-1}) \cdot \nabla w)\mathrm{d}x \\
&= \int_\Omega (r(y)w + d(u_h^n, \lambda^{n-1}, y) \cdot \nabla w)\mathrm{d}x, \qquad (11)
\end{aligned}$$

where

$$r(y) := \mathrm{div}\, y + f$$

and

$$d(v, \lambda, y) := y - A\nabla v - k_*\lambda.$$

If the domain $\Omega$ is simple, we can estimate $|||u^n - u_h^n|||$ by using global constant $C_\Omega$ arriving from Friedrichs inequality. For simple domains upper and lower bounds for $C_\Omega$ can be found analytically. From (11) we find that

$$\begin{aligned}
\int_\Omega r(y)w\mathrm{d}x &\le C_\Omega\|r(y)\|\|\nabla w\| \\
&\le c_1 C_\Omega\|r(y)\|\,|||w|||.
\end{aligned}$$

Set $w := u_h^n - u^n$. Then

$$\int_{\Omega} r(y) w \mathrm{d}x \leq c_1 C_{\Omega} \|r(y)\| \|\!| u^n - u_h^n |\!\|$$

and

$$\int_{\Omega} d(u_h^n, \lambda^{n-1}, y) \cdot \nabla w \mathrm{d}x \leq \ |\!| d(u_h^n, \lambda^{n-1}, y) |\!|_* |\!| u^n - u_h^n |\!|.$$

Now (11) yields the estimate

$$|\!| u^n - u_h^n |\!| \ \leq c_1 C_{\Omega} \|r(y)\|_{\Omega_i} + |\!| d(u_h^n, \lambda^{n-1}, y) |\!|_*. \tag{12}$$

In view of (10) and (12) we obtain the following result.

**Theorem 3** *For any $y \in Q$, we have the estimate*

$$|\!| u - u_h^n |\!| \ \leq (\frac{2k_*}{c_1} + 1) \mathscr{E}^h(u_h^n, \lambda_h^{n-1}, y) + \int_{\Omega} (k_* |\nabla u_h^n| - \lambda^{n-1} \cdot \nabla u_h^n) \mathrm{d}x, \tag{13}$$

*where the first term*

$$\mathscr{E}^h(u_h^n, \lambda_h^{n-1}, y) := c_1 C_{\Omega} \|r(y)\| + |\!| d(u_h^n, \lambda^{n-1}, y) |\!|_*$$

*is related to the approximation error and the second term presents the error associated with the Uzawa method.*

The estimate (13) contains the constant $C_{\Omega}$, which the upper bound can be easily found for the simple domains. If, however, $\Omega$ is a complicated domain (in the sense of geometry) and the problem involves different boundary conditions (e.g., mixed Dirichlet-Neumann conditions), then a guaranteed upper bound may be difficult to find. In order to avoid this, we use another way to estimate the term $\int_{\Omega} r(y) w \mathrm{d}x$.

**Theorem 4** *For any $y \in \tilde{Q} := \{y \in Q \mid \{\mathrm{div}\, y + f\}_{\Omega_i} = 0\}$, we have the estimate*

$$|\!| u - u_h^n |\!| \leq \left(\frac{2k_*}{c_1} + 1\right) \mathscr{E}^h(u_h^n, \lambda_h^{n-1}, y) + \int_{\Omega} \left(k_* |\nabla u_h^n| - \lambda^{n-1} \cdot \nabla u_h^n\right) \mathrm{d}x, \tag{14}$$

*where the first term*

$$\mathscr{E}^h(u_h^n, \lambda_h^{n-1}, y) = c_1 \left(\sum_{i=1}^{N} \mathrm{diam}(\Omega_i) \pi^{-1} \|r(y)\|_{\Omega_i}\right) + |\!| d(u_h^n, \lambda^{n-1}, y) |\!|_*$$

*is related to the approximation error and the second term presents the error associated with the Uzawa method.*

*Proof* We start by using similar methods as we used above. Subtracting $\int_\Omega A\nabla u_h^n \mathrm{d}x$ from the $n$th step of the Uzawa algorithm (8) we have for all $w \in V_0$ and all $y \in Q := \{y \in L^2(\Omega, \mathbb{R}^d) \mid \mathrm{div}(y) \in L^2(\Omega)\}$

$$\int_\Omega A\nabla(u^n - u_h^n) \cdot \nabla w \mathrm{d}x = \int_\Omega (r(y)w + d(u_h^n, \lambda^{n-1}, y) \cdot \nabla w)\mathrm{d}x. \tag{15}$$

Henceforth we set $w = u^n - u_h^n$ and assume that $\Omega$ is represented as a set of nonintersecting convex subdomains $\Omega_i$, $i = 1, \ldots, N$, i.e.,

$$\bar{\Omega} = \sum_{i=1}^{N} \bar{\Omega}_i.$$

Also, we impose additional conditions on $y$:

$$y \in \tilde{Q} := \left\{y \in Q \mid \{\mathrm{div}\, y + f\}_{\Omega_i} = 0\right\},$$

where

$$\{y\}_{\Omega_i} := \frac{1}{|\Omega_i|}\int_{\Omega_i} y \mathrm{d}x$$

denotes the mean value of $y$ in a domain $\Omega_i$. By means of the Payne-Weinberger estimate [19], we obtain

$$\begin{aligned}\int_\Omega r(y)(u^n - u_h^n)\mathrm{d}x &= \sum_{i=1}^{N}\int_{\Omega_i} (\mathrm{div}\, y + f)(u^n - u_h^n - \{u^n - u_h^n\}_{\Omega_i})\mathrm{d}x \\ &= \sum_{i=1}^{N} \|r(y)\|_{\Omega_i}\|u^n - u_h^n - \{u^n - u_h^n\}_{\Omega_i}\|_{\Omega_i} \\ &\le \left(\sum_{i=1}^{N} \mathrm{diam}(\Omega_i)\pi^{-1}\|r(y)\|_{\Omega_i}\right)\|\nabla(u^n - u_h^n)\|. \end{aligned} \tag{16}$$

Also,

$$\int_\Omega d(u_h^n, \lambda^{n-1}, y) \cdot \nabla w \mathrm{d}x \le |||d(u_h^n, \lambda^{n-1}, y)|||_* |||u^n - u_h^n|||. \tag{17}$$

Combining (16) and (17), we find that for any $y \in \tilde{Q}$,

$$|||u^n - u_h^n||| \le c_1\left(\sum_{i=1}^{N} \mathrm{diam}(\Omega_i)\pi^{-1}\|r(y)\|_{\Omega_i}\right) + |||d(u_h^n, \lambda^{n-1}, y)|||_*. \tag{18}$$

Finally, by using (15) and (18) we obtain the estimate (14).

## 5 Concluding Remarks

The paper is devoted to analysis of a class of free boundary problems motivated by models of nonlinear viscous fluids. We presented fully guaranteed and computable bounds of errors for approximations generated by Uzawa type methods. The first of the two models is presented by Theorem 3 and it involves a global constant, which is a constant in the Friedrichs type inequality. The second is presented by Theorem 4 and involves local constants. For the complicated domains the latter one is preferable, because the global constant might be difficult to find.

## References


1. K.J. Arrow, L. Hurwicz, H. Uzawa (eds.), *Studies in Linear and Nonlinear Programming* (Stanford University Press, Stanford, 1958)
2. M.O. Bristeau, R. Glowinski, Finite element analysis of the unsteady flow of a visco-plastic fluid in a cylindrical pipe, in *Finite Element Methods in Flow Problems*, ed. by J.T. Oden, O.C. Zienkiewicz, R.H. Gallagher, C. Taylor (University of Alabama Press, Huntsville, 1974), pp. 471–488
3. J. Cea, R. Glowinski, Méthodes numériques pour l'écoulement laminaire d'un fluide rigide viscoplastique incompressible. Int. J. Comput. Math. **3**, 225–255 (1972)
4. C.M. Colciago, S. Deparis, A. Quarteroni, Comparisons between reduced order models and full 3D models for fluid-structure interaction problems in haemodynamics. J. Comput. Appl. Math. **265**, 120–138 (2014)
5. P. Crosetto, P. Reymond, S. Deparis, D. Kontaxakis, N. Stergiopulos, A. Quarteroni, Fluid-structure interaction simulation of aortic blood flow. Comput. Fluids **43**, 46–57 (2011)
6. E.J. Dean, R. Glowinski, G. Guidoboni, On the numerical simulation of Bingham visco-plastic flow: old and new results. J. Non-Newton. Fluid Mech. **142**, 36–62 (2007)
7. G. Duvaut, J.-L. Lions, *Les inéquations en mécanique et en physique* (Dunod, Paris, 1972)
8. L. Formaggia, D. Lamponi, A. Quarteroni, One-dimensional models for blood flow in arteries. J. Eng. Math. **47**(3–4), 251–276 (2003)
9. M. Fuchs, G. Seregin, Some remarks on non-Newtonian fluids including nonconvex perturbations of the Bingham and Powell-Eyring model for viscoplastic fluids. Math. Models Methods Appl. Sci. **7**(3), 405–433 (1997)
10. M. Fuchs, G. Seregin, Regularity results for the quasi-static Bingham variational inequality in dimensions two and three. Math. Z. **227**(3), 525–541 (1998)
11. L. Fusi, On the stationary flow of a waxy crude oil with deposition mechanisms. Nonlinear Anal. **53**(3–4), 507–526 (2003)
12. L. Fusi, A. Farina, Modelling of Bingham-like fluids with deformable core. Comput. Math. Appl. **53**(3–4), 583–594 (2007)
13. R. Glowinski, *Numerical Methods for Nonlinear Variational Problems* (Springer, New York, 1984)
14. R. Glowinski, Finite element methods for incompressible viscous flow, in *Handbook of Numerical Analysis*, vol. IX, ed. by P.G. Ciarlet, J.-L. Lions (North-Holland, 2003), pp. 3–1176
15. R. Glowinski, J.-L. Lions, R. Trémolières, *Numerical Analysis of Variational Inequalities* (North-Holland, Amsterdam, 1981)
16. J.W. He, R. Glowinski, Steady Bingham fluid flow in cylindrical pipes: a time dependent approach to the iterative solution. Numer. Linear Algebra Appl. **7**(6), 381–428 (2000)
17. P.P. Mosolov, V.P. Miasnikov, Variational methods in the theory of the fluidity of a viscous-plastic medium. J. Appl. Math. Mech. **29**(3), 545–577 (1965)

18. M. Nokka, S. Repin, A posteriori error bounds for approximations of the Oseen problem and applications to Uzawa iteration algorithm. Comput. Methods Appl. Math. **14**(3), 373–383 (2014)
19. L.E. Payne, H.F. Weinberger, An optimal Poincaré inequality for convex domains. Arch. Ration. Mech. Anal. **5**, 286–292 (1960)
20. S. Repin, Estimates of deviations from exact solutions of elliptic variational inequalities. Zapiski Nauchn. Semin. POMI **271**, 188–203 (2000)
21. S. Repin, *A Posteriori Estimates for Partial Differential Equations* (Walter De Gruyter, Berlin, 2008)
22. S. Repin, Estimates of deviations from exact solutions of variational inequalities based upon Payne-Weinberger inequality. J. Math. Sci. (N. Y.) **157**(6), 874–884 (2009)
23. F.J. Sánchez, Application of a first-order operator splitting method to Bingham fluid flow simulation. Comput. Math. Appl. **36**(3), 71–86 (1998)
24. R. Temam, *Navier-Stokes Equations: Theory and Numerical Analysis* (North-Holland, New York, 1977)

# An Automatic Differentiation Based Approach to the Level Set Method

**Jukka I. Toivanen**

**Abstract** This paper discusses an implementation of the parametric level set method. Adjoint approach is used to perform the sensitivity analysis, but contrary to standard implementations, the state problem is differentiated in its discretized form. The required partial derivatives are computed using tools of automatic differentiation, which avoids the need to derive the adjoint problem from the governing partial differential equation. The augmented Lagrangian approach is used to enforce volume constraints, and a gradient based optimization method is used to solve the subproblems. Applicability of the method is demonstrated by repeating well known compliance minimization studies of a cantilever beam and a Michell type structure. The obtained topologies are in good agreement with reference results.

**Keywords** Automatic differentiation · Level set method · Topology optimization

**Mathematical Subject Classification:** 35Q93 · 65D25 · 74P05

## 1 Introduction

The level set method was proposed in [1, 20] for the topology optimization of structures. The basic idea of the method is quite general, and similar techniques can in principle be applied to any problem for which we are able to perform the shape sensitivity analysis. For example, problems of fluid mechanics and electromagnetics are considered in [6, 16] respectively.

The shape sensitivity analysis is usually conducted in the continuous setting, which requires deriving an adjoint equation from the governing partial differential equation, and subsequent discretization in order to numerically evaluate the sensitivity. While this approach is well established for traditional fields of application, such

J.I. Toivanen (✉)
Department of Mathematical Information Technology,
University of Jyväskylä, 35 (Agora), 40014 Jyväskylä, Finland
e-mail: jukka.i.toivanen@gmail.com



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_4

as structural mechanics, new areas of application and multidisciplinary design cases may be problematic. In fluid mechanics, for example, effect of the turbulence model is often neglected during sensitivity analysis to simplify calculations [13].

An alternative approach is to perform the differentiation of the problem *after* discretization. In principle this can be done by manually differentiating all computations, and implementing the corresponding differentiated code. However, the use of automatic differentiation (AD) minimizes the risk of programming errors, and reduces application development time significantly. Moreover, if the code is changed for example to modify source terms, boundary conditions, the objective functional, or the constraints, the gradient computation is updated (almost) automatically. AD tools can be applied even on simulation codes of commercial complexity, as demonstrated in [3].

In this work, the parametric level set method is implemented using automatic differentiation to compute the derivatives of the discrete problem. Dynamic exploitation of sparsity [4] is utilized, and AD is applied only to the assembly process, not on the whole solver. Together with the discrete adjoint approach this technique provides an efficient means to perform the sensitivity analysis, since only the nodes residing near the zero level curve need to be used as independent variables in the differentiation process.

Applicability of the method is demonstrated by solving topology optimization problems related to structural mechanics. The proposed framework is very generic and can be easily extended to other problems.

## 2 Sensitivity Analysis

This paper deals with the level set method [15], where the material region $\Omega$ is defined implicitly using a scalar function $\Psi$:

$$\Omega = \{\mathbf{x} \in D \mid \Psi(\mathbf{x}) > 0\}. \tag{1}$$

Here $D$ is a reference domain containing all admissible geometries. Such definition naturally permits topological changes, such as merging and splitting of material regions. The level set method can also be extended to include mechanism for nucleating new holes away from the boundaries [5], but such extensions are outside the scope of this work.

In parametric level set methods, the scalar function $\Psi$ has an explicit parametrization by means of some design variables $\boldsymbol{\alpha}$. This approach is used in this work, due to the following attractive properties: there is no need for an upwind solution scheme, velocity extension, or reinitialization of the level set function [12].

Unlike the material distribution approach to the topology optimization [2, 14], the level set method is based on the sensitivity analysis with respect to variations in the shape of the material region. The sensitivity analysis is usually conducted in the

continuous setting, which requires deriving an adjoint equation from the governing partial differential equation, and implementing an appropriate discretization. This is a complicated and error-prone task.

It is well known in the shape optimization community that the shape sensitivity analysis can be performed in the discrete setting as well, by differentiating the algebraic form of the problem with respect to the movement of the mesh. This process can be automated with the help of automatic differentiation, or by providing sensitivity computation routines in a finite element library. However, such an approach is not commonly used to implement the level set method. The purpose of this work is to show that the discrete adjoint shape sensitivity analysis is feasible in the context of the level set method, and that all sensitivity computations can be automatized using AD without significant computational overhead.

Let the set of algebraic equations arising from the finite element discretization of the state problem be denoted by

$$\mathbf{r}(\mathbf{q}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) = 0. \tag{2}$$

Here $\mathbf{r}$ is the residual vector, $\boldsymbol{\alpha}$ are the geometrical design variables, and $\mathbf{q}$ is a vector containing the basis function expansion coefficients. Using the discrete adjoint shape sensitivity analysis, the derivative of an objective function $J = J(\mathbf{q}(\boldsymbol{\alpha}), \boldsymbol{\alpha})$ is obtained as

$$\frac{dJ}{d\alpha_i} = \sum_{j,k} \frac{\partial J}{\partial x_j^k} \frac{\partial x_j^k}{\partial \alpha_i} + \gamma^T \left( \sum_{j,k} \frac{\partial \mathbf{r}}{\partial x_j^k} \frac{\partial x_j^k}{\partial \alpha_i} \right), \tag{3}$$

where the adjoint vector $\gamma$ satisfies

$$\left( \frac{\partial \mathbf{r}}{\partial \mathbf{q}} \right)^T \gamma = - \left( \frac{\partial J}{\partial \mathbf{q}} \right)^T. \tag{4}$$

Here $\mathbf{x}_j = (x_j^1, \ldots, x_j^{\dim})$ represents the coordinates of the $j$th mesh node and dim is the dimension of the geometry (dim $= 2$ in this paper).

In the classical shape optimization [10], the geometrical changes are governed by a so called design velocity field, and the sensitivities $\partial \mathbf{x}_j / \partial \boldsymbol{\alpha}$ are known thereof. In practise the mesh is often adapted to the changes of the geometry using some mesh deformation method (see, e.g., [7, 11]), which can be differentiated to obtain the sensitivity information.

In the level set approach [15], however, the mesh is not actually deformed. Instead, a fixed mesh is used, and the boundary of the geometry is given implicitly as

$$\partial \Omega = \{ \mathbf{x} \in D \mid \Psi(\mathbf{x}) = 0 \}. \tag{5}$$

This work proposes to perform the sensitivity analysis on the discretized problem and the use the Eqs. (3) and (4) to compute the gradient of the objective. To this

end, a relation between geometrical changes and the scalar function $\Psi$ needs to be established in order to obtain the design velocity $\partial \mathbf{x}/\partial \boldsymbol{\alpha}$ for mesh nodes residing near the zero level curve. Sensitivity of nodes away from the boundary is assumed to be zero, and they are neglected from the differentiation process.

Let $\mathbf{x}$ be a point residing on the zero level curve $\Psi(\mathbf{x}) = 0$. Assuming that a change in the design variable $\alpha_i$ causes $\mathbf{x}$ to move along the normal vector $\nabla\Psi/|\nabla\Psi|$, we obtain the relation

$$\frac{\partial \mathbf{x}}{\partial \alpha_i} = -\frac{\partial \Psi}{\partial \alpha_i} \frac{\nabla \Psi}{|\nabla \Psi|^2}. \tag{6}$$

Even though the mesh nodes are not actually moving, we use these sensitivities as the design velocity field in (3) to compute an approximate gradient of the objective.

The reference domain is not allowed to change its shape. Thus the nodes on the boundary of $D$ do not move, and we set

$$\frac{\partial \mathbf{x}}{\partial \alpha_i} \cdot \mathbf{n} := 0 \quad \forall i \tag{7}$$

for all nodes $\mathbf{x}$ residing on $\partial D$. Here $\mathbf{n}$ is the boundary normal.

To sum things up, the following approach for the sensitivity analysis in the context of the parametric level set method is proposed:

1. Solve the state problem (2).
2. Compute $\partial J/\partial \mathbf{q}$, $\partial J/\partial \mathbf{x}$, and $\partial \mathbf{r}/\partial \mathbf{x}$ using automatic differentiation.
3. If the problem is not self-adjoint, solve the adjoint problem (4).
4. Compute the gradient of $J$ using (3), where $\partial x_j^k/\partial \alpha_i$ is obtained from the level set function using Eq. (6).

Notice, that we have formally performed the sensitivity analysis without using any information about the state problem, which presents only implicitly through the residual representation (2). This is the main benefit of the proposed approach: since the automatic differentiation is used to compute the shape derivatives, there is no need to manually derive any problem specific sensitivity expressions.

## 3 Parametrization of the Level Set Function

In this paper, the compactly supported $C^2$-continuous radial basis functions [22] are utilized to explicitly parametrize the level set function. Let us consider a set of $N \times M$ basis functions, whose knots are distributed over the domain $D \subset \mathbb{R}^2$ so that coordinates $(b_{ij}^1, b_{ij}^2)$ of the knot $ij$ are

$$b_{ij}^1 = x_{\min}^1 + (i-1) \cdot \delta_{RBF}, \quad b_{ij}^2 = x_{\min}^2 + (j-1) \cdot \delta_{RBF}, \tag{8}$$

where $j = 1, \ldots, M$, $i = 1, \ldots, N$, the point $(x^1_{\min}, x^2_{\min})$ is the lower left corner of the rectangle $D$, and $\delta_{RBF}$ is a given parameter. The radial basis function (RBF) associated with this knot is

$$\psi_{ij}(\mathbf{x}) = \max\left\{0, 1 - r_{ij}(\mathbf{x})\right\}^4 (4r_{ij}(\mathbf{x}) + 1), \tag{9}$$

where

$$r_{ij}(\mathbf{x}) = \frac{\sqrt{(x^1 - b^1_{ij})^2 + (x^2 - b^2_{ij})^2}}{r_s}, \tag{10}$$

and radius of the support $r_s > 0$ is a given parameter.

The parametrized level set function $\psi(\boldsymbol{\alpha})$ is defined as the linear combination

$$\psi(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_{ij} \psi_{ij}. \tag{11}$$

The design variables of the parametrized optimization problem are represented by the vector $\boldsymbol{\alpha} = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{NM})$.

## 4 Automatic Differentiation

We consider a new framework for implementing the parametrized level set method, in which manual derivation of the sensitivity expressions is no longer required. The use of automatic differentiation [9] is proposed for computing the partial derivatives of the discretized problem. Below a brief overview of AD is given.

Automatic differentiation exploits the fact that the computer program can be represented as a sequence of elementary arithmetic operations, and systematically applies the chain rule of differentiation to these operations. There are two main variants of AD: the forward and the reverse modes. Using the reverse mode, the gradient of a single objective function can be computed in a time that is independent on the number of design variables. However, information about every operation that is performed during the computation needs to be stored to a so called tape, and the tape must be traversed in reverse order to compute the gradient. If AD is applied to a complete solver code, the tape may become so large that it needs to be stored to the disk. Since disk access is very slow, this step may actually dominate the computation time [19].

In the so-called forward mode AD, which is used in this work, the derivatives are propagated forward in the execution chain. If applied in a naïve manner, the computation time is directly proportional to the number of derivatives that are computed. Fortunately, this limitation can be overcome by applying AD only to the assembly

process, performing the differentiation with respect to the mesh nodal coordinates, and exploiting sparsity.

Assume that there exists $n$ independent variables $\mathbf{a} = (a_1, \ldots, a_n)$, and that we are interested in the derivatives of some output variables with respect to these independent variables. Now consider an elementary arithmetic operation $\Phi$ of two arguments

$$C = \Phi(A(\mathbf{a}), B(\mathbf{a})). \tag{12}$$

Since the forward mode AD is used, the partial derivatives of the arguments $A$ and $B$ are known, as they have already been computed and stored when the arguments themselves where evaluated. We can proceed and compute the partial derivatives of the result variable $C$ as

$$\frac{\partial C}{\partial a_i} = \frac{\partial \Phi}{\partial A}\frac{\partial A}{\partial a_i} + \frac{\partial \Phi}{\partial B}\frac{\partial B}{\partial a_i}. \tag{13}$$

In the traditional implementations this is done for all $i = 1, \ldots, n$, and the computational cost is therefore proportional to $n$. If the arguments depend only on a few independent variables, then $\partial C/\partial a_i = 0$ for most $i$, and a lot of computation is performed in vain.

Instead, the AD code used in this work performs so called sparse forward propagation [4], where only the non-zero partial derivatives are stored and computed. Let us define the index domain of variable $Y$ as $\chi(Y) = \{i \mid \partial Y/\partial a_i \neq 0\}$. The key idea behind our implementation is to store the index domain of each variable to a vector in increasing order, and the actual derivatives to another vector in the same order. The index domain of the result variable $C$ above is $\chi(C) = \chi(A) \cup \chi(B)$. Exploiting the ordering of the index domains, $\chi(C)$ can be formed, and the corresponding derivatives computed, in a time that is directly proportional to the size of the set $\chi(C)$. In particular, computational cost of the operation does not depend on $n$.

The implementation is based on the operator overloading capabilities of the C++ programming language, and uses a custom data type `addouble` to represent a real variable and its derivatives with respect to the independent variables. Arithmetic operations involving this data type have been redefined so that they implement also the computation of the derivative information. Details of the implementation can be found in [19].

Exploitation of the sparsity is essential for the efficiency of the proposed method. There are as many as 6000 design variables in the numerical examples considered in this paper. Clearly, approaches where the computational cost scales in direct proportion to the number of design variables are infeasible. In this work, such scaling is avoided as follows: like mentioned in Sect. 2, AD is used to compute the derivatives of the residual vector $\mathbf{r}$ with respect to the mesh nodal coordinates. In FEM, each component of the residual vector depends only on the shapes of the elements that belong to the support of one particular test function, which means that $\partial r_i/\partial x_j^k \neq 0$ only for very few $j$. The radial basis functions used to parametrize the level set function are compactly supported, which makes the vectors $\partial x_j^k/\partial \boldsymbol{\alpha}$ sparse as well. We will return to this topic when computation times are discussed in Sect. 7.3.

Such sparsity could be exploited in other ways as well, for example by applying AD elementwise during the assembly. However, a lot of implementation effort is saved by the fact that sparsity is exploited in the automatic differentiation.

## 5 Augmented Lagrangian Method

Topology optimization problems typically include a volume constraint

$$0 = \ell(\boldsymbol{\alpha}) := \left( \int_D H(\psi(\boldsymbol{\alpha})) \right) - V_{\max}, \tag{14}$$

where $H$ is the Heaviside function and $V_{\max}$ is given. This constraint is handled using the augmented Lagrangian approach.

The $k$th iteration of the method consists of solving the subproblem

$$\min_{\boldsymbol{\alpha}} \mathscr{L}_A(\boldsymbol{\alpha}, \lambda, \nu) = \min_{\boldsymbol{\alpha}} \left( J(\boldsymbol{\alpha}) - \lambda^k \ell(\boldsymbol{\alpha}) + \frac{1}{2\nu^k} \ell^2(\boldsymbol{\alpha}) \right), \tag{15}$$

where $\lambda^k$ is a Lagrange multiplier, and $\nu^k$ is a penalty parameter. These parameters are then updated according to rules

$$\lambda^{k+1} = \lambda^k - \frac{\ell(\boldsymbol{\alpha}^k)}{\nu^k}, \qquad \nu^{k+1} = \delta \nu^k, \tag{16}$$

where $\delta \in ]0, 1[$, and $\boldsymbol{\alpha}^k$ is the approximate solution to the subproblem. Initially $\lambda^0 = 0$ and $\nu^0$ is a given parameter.

## 6 Optimization Method

Gradient based methods are obviously preferred for solving the subproblems (15) due to their efficiency, but there are some complications involved. Namely, the optimization problem has a very large number of variables, which calls for a lightweight method. Moreover, there is some numerical noise in the function due to approximations made in the level set approach. Therefore, the optimization method must not be too sensitive to accuracy of the supplied gradient information.

A class of optimization methods based on conservative convex separable approximations was proposed in [18]. These methods construct a convex approximation $\tilde{f}^k$ of the objective function $f(\boldsymbol{\alpha}) : X \subset \mathbb{R}^m \to \mathbb{R}$ at the current iterate $\boldsymbol{\alpha}^k$. The approximation is modified until it becomes conservative in the sense that

$$\tilde{f}^k(\boldsymbol{\alpha}^{k*}) \geq f(\boldsymbol{\alpha}^{k*}), \tag{17}$$

where $\boldsymbol{\alpha}^{k*}$ is the minimizer of the approximating function. The point $\boldsymbol{\alpha}^{k*}$ then becomes the new iterate $\boldsymbol{\alpha}^{k+1}$. For such process, global convergence towards a Karush-Kuhn-Tucker point was proved in [18] under suitable assumptions.

In [17], it was proposed to use spherical quadratic approximation having the form

$$\tilde{f}^k(\boldsymbol{\alpha}) = f(\boldsymbol{\alpha}^k) + \nabla f(\boldsymbol{\alpha}^k)^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^k) + \frac{c^k}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^k)^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^k).$$

Also, convergence of the method was proved, when it is applied to general positive-definite quadratic functions. This approximation is strictly convex if $c^k > 0$. The minimizer can be readily obtained without performing any line searches, and it is given by the relation

$$\boldsymbol{\alpha}^{k*} = \boldsymbol{\alpha}^k - \frac{\nabla f(\boldsymbol{\alpha}^k)}{c^k}. \tag{18}$$

Since there is a certain numerical noise, maximum step length control is added to the optimization method to improve robustness. With this modification, the algorithm used in this work can be written as follows:

1. Select $\xi > 0$, $\rho > 1$ and $c^1 > 0$. Set $k = 1$.
2. Compute $\boldsymbol{\alpha}^{k*}$ using (18).
3. If $\boldsymbol{\alpha}^{k*}$ is acceptable (the condition (17) holds), goto step 5.
4. Set $c^k = \rho c^k$. Goto step 2.
5. Set $\boldsymbol{\alpha}^{k+1} = \boldsymbol{\alpha}^{k^*}$.
6. If $\|\boldsymbol{\alpha}^{k+1} - \boldsymbol{\alpha}^k\| < \xi$ or $k = k_{\max}$ STOP.
7. Compute $c^{k+1}$. Set $k = k + 1$ and goto step 2.

To compute the scalar $c^{k+1}$, we follow essentially the approach presented in [23] and set

$$c^{k+1} = \max\left(\frac{(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k+1})^T(\nabla f(\boldsymbol{\alpha}^k) - \nabla f(\boldsymbol{\alpha}^{k+1}))}{(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k+1})^T(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k+1})}, \beta, \|\nabla f(\boldsymbol{\alpha}^k)\|/\eta\right). \tag{19}$$

The first part of this expression is obtained by matching gradient vectors in the least squares sense, the condition $c^{k+1} \geq \beta$ is enforced to keep the approximation strictly convex, and the condition $c_{k+1} \geq \|\nabla f(\boldsymbol{\alpha}^k)\|/\eta$ enforces maximum step length $\eta$.

## 7 Numerical Examples

This section contains the results of numerical computations, which were performed on a HP ProLiant DL585 server equipped with 4 AMD Opteron 885 2.6 GHz dual core processors and 64 GB memory. Parallelization was not exploited. The version

4.0 of the SuperLU library [8] was used to solve the linear systems representing the state problems. In all cases, linear finite elements and unstructured meshes were used in order to maintain maximum flexibility of meshing.

We consider the well known compliance minimization problem, where the objective functional is

$$J(\mathbf{u}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) = \int_D E_{ijkl}(\boldsymbol{\alpha})\varepsilon_{ij}(\mathbf{u})\varepsilon_{kl}(\mathbf{u}) \tag{20}$$

and the state problem in the variational form reads: find $\mathbf{u} \in V$ such that

$$\int_D E_{ijkl}(\boldsymbol{\alpha})\varepsilon_{ij}(\mathbf{u})\varepsilon_{kl}(\mathbf{v}) - \int_{\Gamma_1} \mathbf{g} \cdot \mathbf{v} = 0 \quad \forall \mathbf{v} \in V. \tag{21}$$

Here $\mathbf{u}$ is the displacement, $E_{ijkl}$ is the elasticity tensor, $\varepsilon_{ij}$ is the strain tensor, $\mathbf{g}$ denotes the surface load and $V = \{\mathbf{v} \in H^1(D)^d \mid \mathbf{v}|_{\Gamma_0} = 0\}$. By performing the standard discretization with linear finite elements the problem (21) is converted to the discrete residual form present in Eq. (2).

Only for the sake of clarity we consider two-dimensional problems and assume that the plane stress conditions hold. The so called ersatz material approach is used, in which the void regions are governed by weak material, and the problem is posed over the entire reference domain. It is assumed that the material is isotropic and the material parameters are constant within each element. The design variables affect the state problem through the tensor $E_{ijkl}(\boldsymbol{\alpha})$ as follows. The Young's modulus has the values $2.1 \times 10^{11}$ and 1.0 in material and void regions respectively, and if the zero level set cuts through the element, it is interpolated between these values accordingly. The Poisson's ratio has the value 0.3.

The radius of support of the radial basis functions was related to the parameter $\delta_{RBF}$ as $r_s = 4\delta_{RBF}$ in all examples. This guarantees sufficient overlap between the basis functions, but on the other hand, provides significant sparsity. The parameters used in the optimization algorithm had the following values in all examples: $c^1 = 1$, $\beta = 10^{-4}$, $\rho = 4$, $\xi = 10^{-6}$, $k_{\max} = 50$. The number of outer iterations in the augmented Lagrangian framework was 5.

The proposed approach is not extremely sensitive to the choice of parameters, but couple of them do play some role in the process. Namely, the initial value of the penalty parameter $\nu$ and the value of the parameter $\delta$ can affect the final topology. If $\nu$ is initially too large, the penalty term will have little effect, and the material portion of the domain will grow. During this phase some of the holes present in the initial guess can disappear. Since there is no nucleation mechanism, the holes can not reappear at later iterations, and the amount of holes in the final topology is reduced. On the other hand, if $\nu$ is initially too small or it is decreased too quickly ($\delta$ is too small) the opposite can happen, and some material regions disappear. The correct choices of $\nu^0$ and $\delta$ depend on the relative magnitudes of the objective and the penalty terms.

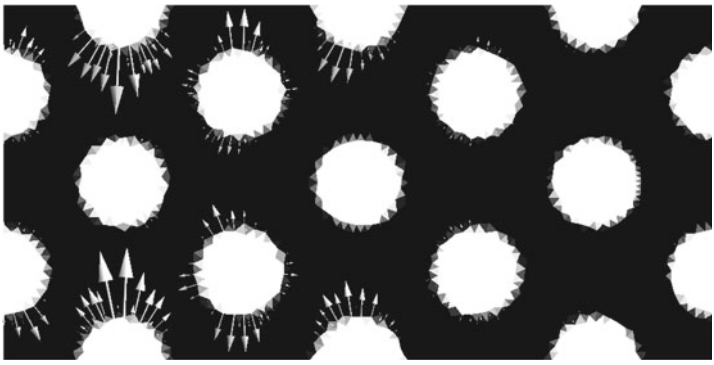


**Fig. 1** Initial guess for the cantilever beam example

## *7.1 Cantilever Beam*

In the first example, the reference domain is a rectangle of size $4 \times 2$, from which a fraction of 0.5 is allowed to be occupied by material in the final design. A vertical point load of 40 kN is applied at the middle of the right edge, and zero displacement constraints are specified on the left edge. The parameter $\delta_{RBF}$ is 0.05, resulting in $80 \times 40 = 3200$ design variables. The mesh had 3645 nodes and 7529 elements.

The initial design is as shown in Fig. 1. The grey regions in the figure denote elements that are cut by the zero level curve, and in which the Young's modulus is interpolated between the extreme values. The vectors appearing in the figure denote the sensitivity of the objective with respect to movement of the mesh nodes, so that the length of the vector is proportional to the magnitude of the sensitivity. Figure 2 shows the geometry at some intermediate stages during the optimization. Notice that the holes with the largest sensitivities quickly start to shrink during the optimization, whereas the holes with little effect on the objective tend to grow in order to meet the volume constraint.

Since movement of nodes inside material or void regions only affects the solution of the problem through discretization error, they are excluded from the sensitivity analysis. In other words, only the nodes belonging to the grey elements are declared as independent variables of the automatic differentiation. This significantly increases efficiency of the proposed approach, since the AD implementation inherently exploits such sparsity.

Values of the objective and the constraint terms during the optimization are shown in Fig. 3. The maximum step length during the optimization was $\eta = 0.025$. The initial penalty parameter $\nu^0$ was 20.0, and the parameter $\delta$ was 0.5. The iterations of the augmented Lagrangian method are clearly visible in Fig. 3, as after each update of $\nu$ and $\lambda$ the value of the constraint violation starts to decrease more rapidly.

The parameters were not specifically chosen to minimize the number of optimization iterations. This could be achieved by decreasing $k_{\max}$ or choosing a larger

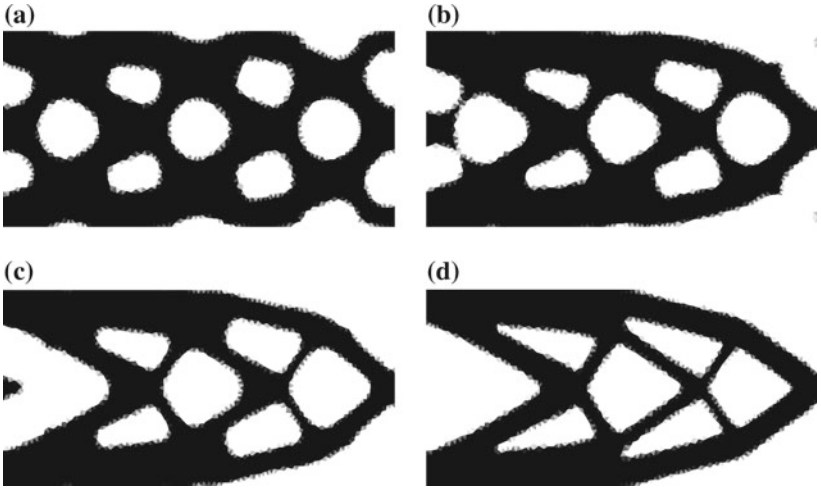



**Fig. 2** Geometry of the cantilever during the optimization. **a** After 20 iterations. **b** After 40 iterations. **c** After 60 iterations. **d** Final geometry

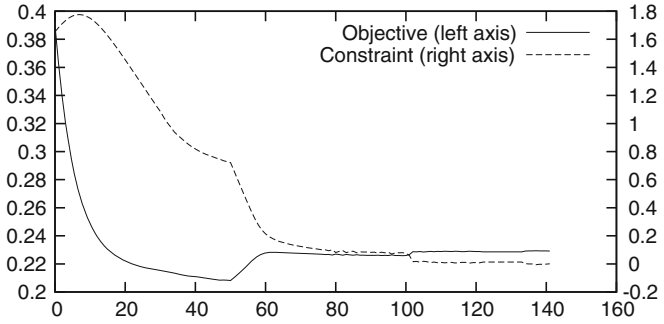



**Fig. 3** Optimization history of the cantilever beam

tolerance parameter $\xi$. The total number of function and gradient evaluations was 233, which is only moderately larger than the number of optimization iterations. This means that the optimization method was able to accept most of the design candidates without the need to adjust $c_k$ to make the approximation conservative.

The final design, shown in Fig. 2, is consistent with the results presented, for example, in [1, 12].

## *7.2 Michell Type Structure*

The next example considers a Michell type structure. The reference domain has the dimensions $2 \times 1.2$. The bottom corners have pinned supports, and vertical point load

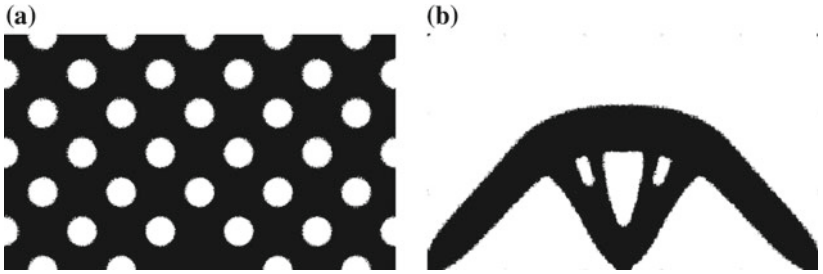



**Fig. 4** Initial and final geometry of the Michell type structure. **a** Initial guess. **b** Final geometry

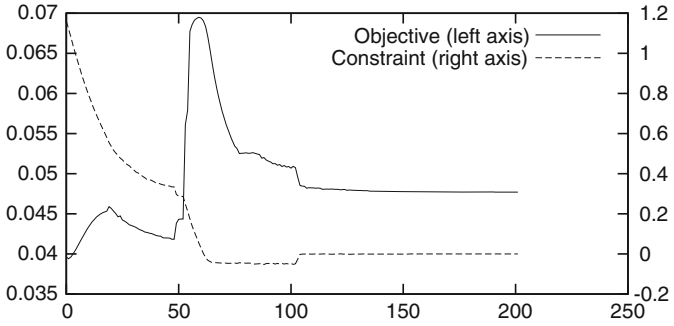



**Fig. 5** Optimization history of the Michell type structure

of 40 kN is applied at the middle of the bottom edge. The initial guess is presented in Fig. 4. In this example $\delta_{RBF} = 0.02$, resulting in $100 \times 60 = 6000$ design variables. The mesh had 16637 nodes and 33747 elements. The material volume fraction in the final design was specified to be 0.3. The parameters $\nu^0$, $\delta$ and $\eta$ had the values 10, 0.2, and 0.025, respectively. The total number of function and gradient evaluations was 334. The final design shown in Fig. 4 has the same topology than the result presented in [21], and it is consistent with the analytical solution of the problem. Figure 5 shows the progress of objective and constraint values during the optimization.

## *7.3 Computation Times*

In this section, the computational efficiency of the proposed method is discussed. To this end, the CPU times required for various tasks were recorded during the optimization runs, and are shown in Table 1. The numbers represent average values over all iterations of the optimization.

Every time the design variables change, the value and the gradient of the scalar function at each mesh node are computed and stored. The average time required to do so is shown in the first row of Table 1.

**Table 1** Average durations of tasks in CPU seconds

| | Cantilever | Michell |
|---|---|---|
| Scalar function evaluation | **0.334** | **1.491** |
| State problem solution | **0.252** | **2.310** |
| Assembly | 0.054 | 0.258 |
| Linear system solution | 0.189 | 2.011 |
| Gradient evaluation | **0.097** | **0.398** |
| Computation of $\partial \mathbf{r}/\partial \mathbf{x}$ | 0.072 | 0.290 |

The state problem corresponding to the updated design must also be solved. The total time required to solve the state problem is shown in the second row of Table 1. This includes assembling the system matrix, and solving the linear system. The durations of these tasks are shown separately in the table.

Finally, Table 1 represents the time required to evaluate the full gradient of the objective function. This includes computing the partial derivatives $\partial \mathbf{r}/\partial \mathbf{x}$, using relation (6) to compute those terms $\partial \mathbf{x}_j/\partial \alpha_i$ that are not zero due to the compact supports of the radial basis functions, and using the relation (3) to compute the final values.

The time required to perform one optimization iteration is used mostly in scalar function evaluation, state problem solution, and gradient evaluation. The time required for other operations is almost negligible.

In both cases, computing the gradient takes less time than the solution of the state problem. Notice that both problems are self-adjoint, and thus there is no need to solve an adjoint problem. On the other hand, majority of the time required to solve the state problem is actually spent in the linear system solver.

Computing $\partial \mathbf{r}/\partial \mathbf{x}$ takes slightly more time than the assembly of the state problem. This is no surprise, since this phase is implemented simply by performing the assembly again, this time using AD to compute $\partial \mathbf{r}/\partial \mathbf{x}$. This process could be optimized: since only some nodes take part in the sensitivity analysis, it would suffice to go through only those elements that share some of these nodes.

Table 1 shows that evaluation of the scalar function takes a significant amount of time. To avoid going through all radial basis functions while evaluating $\Psi$ at given $\mathbf{x}$, a quadtree data structure is exploited to exclude RBFs that can not have a non-zero value at that point. This process could be optimized as well, by exploiting the fact that the mesh is fixed.

In conclusion, the timings show that the proposed combination of discrete adjoint approach, and automatic differentiation performing the sparse forward propagation, is an efficient means to perform the sensitivity analysis. Even though both problems have thousands of design variables, the sensitivity analysis is computationally cheap compared to the state problem solution.

## 8 Conclusions

A new approach for implementing the parametrized level set method is proposed. Contrary to the traditional implementations, the state problem is differentiated in its discretized form, which enables the use of automatic differentiation (AD) to perform the shape sensitivity analysis. The cumbersome derivation of the adjoint problem from the governing partial differential equation is avoided altogether, which is the main benefit of the proposed method.

AD is used in the forward mode, and storing the computation to a 'tape' is therefore not required. Nevertheless, the time required to compute the gradient does not grow in the number of design variables, since AD is not applied to the whole code. Instead the discrete adjoint approach is used to efficiently compute the gradient of the objective function. To this end the assembly process is differentiated, so that the output variables are the components of the residual vector, and the input variables are the mesh nodal coordinates. The resulting vectors of partial derivatives are extremely sparse, which is exploited by the AD implementation. Namely, the code performs so called sparse forward propagation, in which only the non-zero partial derivatives are computed and stored.

Volume constraints were enforced using the augmented Lagrangian approach, and a gradient based optimization method was used to solve the resulting subproblems. Two well known topology optimization studies were used as test cases, and the classical results were recovered. However, the proposed method is quite generic, and could be extended to other problems as well.

**Acknowledgments** The author was financially supported by Academy of Finland, grant #257589.

## References

1. G. Allaire, F. Jouve, A.-M. Toader, Structural optimization using sensitivity analysis and a level-set method. J. Comput. Phys. **194**(1), 363–393 (2004)
2. M.P. Bendsøe, Optimal shape design as a material distribution problem. Struct. Multidiscip. Optim. **1**(4), 193–202 (1989)
3. C.H. Bischof, H.M. Bücker, A. Rasch, Sensitivity analysis of turbulence models using automatic differentiation. SIAM J. Sci. Comput. **26**(2), 510–522 (2004)
4. C.H. Bischof, P.M. Khademi, A. Buaricha, C. Alan, Efficient computation of gradients and Jacobians by dynamic exploitation of sparsity in automatic differentiation. Optim. Methods Softw. **7**(1), 1–39 (1996)
5. M. Burger, S.J. Osher, A survey on level set methods for inverse problems and optimal design. Eur. J. Appl. Math. **16**(2), 263–301 (2005)
6. V.J. Challis, J.K. Guest, Level set topology optimization of fluids in Stokes flow. Int. J. Numer. Methods Eng. **79**(10), 1284–1308 (2009)
7. A. de Boer, M.S. van der Schoot, H. Bijl, Mesh deformation based on radial basis function interpolation. Comput. Struct. **85**(11–14), 784–795 (2007)
8. J.W. Demmel, S.C. Eisenstat, J.R. Gilbert, X.S. Li, J.W.H. Liu, A supernodal approach to sparse partial pivoting. SIAM. J. Matrix Anal. Appl. **20**(3), 720–755 (1999)

9. A. Griewank, A. Walther, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, 2nd edn. (SIAM, Philadelphia, 2008)
10. J. Haslinger, R.A.E. Mäkinen, *Introduction to Shape Optimization: Theory, Approximation, and Computation* (SIAM, Philadelphia, 2003)
11. B.T. Helenbrook, Mesh deformation using the biharmonic operator. Int. J. Numer. Methods Eng. **56**(7), 1007–1021 (2003)
12. Z. Luo, M.Y. Wang, S. Wang, P. Wei, A level set-based parameterization method for structural shape and topology optimization. Int. J. Numer. Methods Eng. **76**(1), 1–26 (2008). doi:10.1002/nme.2092
13. J.E.V. Peter, R.P. Dwight, Numerical sensitivity analysis for aerodynamic optimization: a survey of approaches. Comput. Fluids **39**(3), 373–391 (2010)
14. G.I.N. Rozvany, M. Zhou, T. Birker, Generalized shape optimization without homogenization. Struct. Multidiscip. Optim. **4**(3–4), 250–252 (1992)
15. J.A. Sethian, A. Wiegmann, Structural boundary design via level set and immersed interface methods. J. Comput. Phys. **163**(2), 489–528 (2000)
16. H. Shim, V.T.T. Ho, S. Wang, D.A. Tortorelli, Level set-based topology optimization for electromagnetic systems. IEEE Trans. Magn. **45**(3), 1582–1585 (2009)
17. J.A. Snyman, A.M. Hay, The spherical quadratic steepest descent (SQSD) method for unconstrained minimization with no explicit line searches. Comput. Math. Appl. **42**(1–2), 169–178 (2001)
18. K. Svanberg, A class of globally convergent optimization methods based on conservative convex separable approximations. SIAM J. Optim. **12**(2), 555–573 (2002)
19. J.I. Toivanen, R.A.E. Mäkinen, Implementation of sparse forward mode automatic differentiation with application to electromagnetic shape optimization. Optim. Methods Softw. **26**(4–5), 601–616 (2011)
20. M.Y. Wang, X. Wang, D. Guo, A level set method for structural topology optimization. Comput. Methods Appl. Mech. Eng. **192**(1–2), 227–246 (2003). doi:10.1016/S0045-7825(02)00559-5
21. S.Y. Wang, K.M. Lim, B.C. Khoo, M.Y. Wang, An extended level set method for shape and topology optimization. J. Comput. Phys. **221**(1), 395–421 (2007). doi:10.1016/j.jcp.2006.06.029
22. H. Wendland, Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. Adv. Comput. Math. **4**(1), 389–396 (1995)
23. D.N. Wilke, S. Kok, A.A. Groenwold, The application of gradient-only optimization methods for problems discretized using non-constant methods. Struct. Multidiscip. Optim. **40**, 433–451 (2010)

# Part II
# Mathematical Modeling in Mechanics

# Differential Fluid Mechanics—Harmonization of Analytical, Numerical and Laboratory Models of Flows

**Yuli D. Chashechkin**

**Abstract** Concepts of a "solid body motion" and "fluid flow" are compared taking into account the condition of inobservability of a "fluid particle". General properties of the fundamental set of fluid mechanics equations, accepted for describing fluid flows, are analyzed taking into account the compatibility condition. Hierarchy of periodic flows is classified basing on the order of linearized set of governing equations. Results of theoretical analysis of infinitesimal periodic flows in a stably stratified fluid including periodic internal waves and accompanied family of small scale components are given. Calculations of periodic internal waves propagation and generation in a fluid with arbitrary stable profile of buoyancy are compared with data of schlieren observations in laboratory. Fine flow structure observed behind uniformly towing strip is discussed in context of a given model. Some conclusions and recommendations on improvement techniques of a fluid dynamics experiment are presented.

**Keywords** Dynamics · Solid motion · Fluid flow · Multiscale structure · Asymptotic solution · Laboratory experiment · Internal waves

**Mathematical Subject Classification:** 76M45

## 1 Introduction

Each historical period is characterized by its own requirements to the theoretical and experimental fluid mechanics. During antic and medieval time hydrological problems were solved on the basis of elementary geometry and common sense. New approaches based on theoretical and experimental modeling started arising

Y.D. Chashechkin (✉)
A. Yu. Ishlinskiy Institute for Problems in Mechanics of the Russian Academy of Sciences, Moscow, Russia
e-mail: yulidch@gmail.com



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_5

simultaneously in many countries, and it was G. Galiley who complemented the development of the general principles by consideration of applied tasks, selection of physical quantities and creating of instruments for experimental determination of their values. Later a decisive influence on the experiments was provided by theoretical works based on the ideas of Newton, d'Alembert and Euler to use partial differential equations for calculations of liquid and gas flows. At the end of the XIX century O. Reynolds introduced statistical methods in the fluid mechanics that predetermined methodology of experimental design, data processing and development of theories for all the XX century.

Beginning of the XXI century was marked by several important features. The concentration of the population on the shorelines has increased sensitivity to adverse effects of natural variability of the atmosphere and hydrosphere. Use as intermediates products poisonous and biologically hazardous substances, preserving its negative effects even at low concentrations stimulated further development of the theory of flows and transport of substances.

At the same time, the development of remote sensing instruments has shown that flows of different length scales are characterized by a fine internal structure that can regular or chaotic. Many data indicate that the fine structure is the intrinsic inseparable property of the fluid flow. Current theories are oriented on description of individual flow components like waves, vortices, jets, boundary layers and so on. Accordingly, the future theory should be directed on calculation of all flow properties in their completeness, including both the dynamics and geometry of flows over the global and fine scales. To find a right way for development of the theory at first it is necessary to come back to the foundations of mechanics to discuss in modern language basic principles taking into account long history of fluid dynamics.

The first closed set of governing equations for compressible and incompressible homogeneous fluid was published by Euler [1]. The influence of Euler's ideas was very strong and had predetermined selection of physical variables and mathematical methods to describe fluid flows. In derivation of equations Euler used concept of a "fluid particle" the motion of which can be traced and characterized by velocity on analogy of a solid body movement. However, in a homogeneous fluid deformable "fluid particle" having no distinguished boundaries cannot be identified and traced experimentally. So the principle of selection of parameters characterizing fluid flows must be discussed more carefully taking into account necessary condition of a theory proving by its comparison with experiment and following the concept of "observability" of handled physical quantities.

At the end of XVIII century B. Franklin wrote short paper [2] presenting results of observations of behavior of interface between oil and water in a lighting ship lamp in the rough sea and its modeling by placement lamp on a children's swing. He argued to study dynamics of non-homogeneous fluids but his ideas were ignored up to the end of XX century.

The next important step in the theory of homogeneous fluid flows was done by C.M.H. Navier. Basing on de Laplace hypothesis of the molecular structure of matter he supplemented the Euler's equations by an additional term describing action of shear stresses in the moving viscous fluid [3]. New equations were met skeptically

by contemporaries, who classified them as a hypothesis that needs in experimental confirmation. Existence and uniqueness of 3D solutions of the Navier equations is still unproven.

The equations of viscous homogeneous fluids motions were several times re-derived until G.G. Stokes gave them a modern interpretation and constructed a number of key solutions [4]. Hydrodynamics at the end of XIX century was criticized by a number of scientists, including the great D.I. Mendeleeff who indicated a weakness of experimental methods [5] and ignorance individual properties of fluids which are described by the state equations which he experimentally constructed for gases, pure fluids and water solutions [6, 7].

The accuracy of measurements of fluid flows by many orders worse than in the mechanics of solids. One of the reasons of current situation is the loss of *observability principle* control for physical quantities involved into the theory. It is well known that only invariants are observable quantities which can be measured with prescribed accuracy and their values do not depend on selection of coordinate frame. The list of invariants in the solid body mechanics includes distances between bodies in space, time intervals between events and mass of the body. What are appropriate physical quantities for fluids and invariants of fluid flows? To answer this question it is necessary to come back to axiomatic of fluid mechanics.

## 2 Kinematics of Motion

Theoretical mechanics is based on the concepts of *number*, *space*, discrete or *continuous medium* which is immersed in the space and *motions* or *flows* that are changes the dynamic state and values of the physical parameters of solids, liquids or gases. Mathematical basis of the theory is the notion of non-dimensional *real number*, the properties of which are given here axiomatically [8].

The concepts of "*space*" and "*time*" in classical mechanics are introduced a priori and characterized by their own dimensions. They are assumed to be homogeneous, isotropic and independent on each other or immersed bodies and the physical processes involved.

The main parameter of the physical body in the mechanics is its *mass M* characterized by its own dimension. The body mass is a positive scalar measure of inertia, gravity interaction and the amount of a substance. The mass is invariant and conserved when the body is moving in a space or contracted into a "*material point*". In the physical mechanics is assumed that space where the bodies are placed, is metric (Euclidean) and three-dimensional, allowing the use of a Cartesian coordinate frame.

Mathematical description of motion is based on vector sets axiomatics with different operations and important rule of *external composition*, which means that multiplication of vector by scalar remains the product in the same vector space [8]. This rule ensures the internal unity of theoretical solid body mechanics, abstract and applied mathematics.

The unification is based on the identity of the two physical and two mathematical definitions of a "material point" motion. Parameter of the continuous motion is time $t$ that is also a scalar quantity of its own dimension. The first physical definition of the mechanical motion (displacement) that is change of mutual positions of bodies is based on the registration of distances between the objects. In the second *dynamic definition* the motion of the body with mass $M$ is characterized by the momentum $\mathbf{p}(x_1, x_2, x_3)$ and energy $E = \mathbf{p}^2/2M$ which are also observable quantities.

The *mathematical definition* of the motion requires the introduction of the abstract absolute stationary coordinate frame with center at the point $O$ in the space $\mathbb{R}^3$. The position of body center in the initial and subsequent times is determined by the radius vector.

In kinematic definition the motion it is characterized by trajectory that is an envelope of all positions of radius vectors $\mathbf{r}$ for the material point, and is described by the velocity

$$\mathbf{v} = \frac{dS_t}{dt}\frac{d\mathbf{r}}{dS} = \boldsymbol{\tau}\frac{dS_t}{dt}$$

and the acceleration

$$\mathbf{w} = \frac{d\mathbf{v}}{dt}$$

where $\boldsymbol{\tau}$ is a unit vector in the direction of the tangent to the trajectory $S_t$.

Motion as a *mathematical* (*geometric*) concept is defined as a continuous operation of transformation of the space $\mathbb{R}^3$ into itself with parameter $t$, saving the distances between points and the relative arrangement of the objects [8]. Under these conditions, determinant of the matrix of the transformation coefficients $\alpha_{ik}$ is equal $\|\alpha_{ik}\| = +1$. Orthogonal transformation with determinant $\|\alpha_{ik}\| = -1$, not saving the orientation and mutual locations of the figures, characterizes reflection relative to certain axis.

Decomposition of mechanical motions comprises mutually independent rectilinear displacement (shift) at a velocity $\mathbf{v}_t$ and/or rotation about the instantaneous center $O$ with an angular velocity $\boldsymbol{\Omega}$

$$\delta\mathbf{r} = \mathbf{v}_t\delta t + \boldsymbol{\Omega} \times \delta\mathbf{r}. \tag{1}$$

Motion is characterized by a group of transformations, including independent subgroups of translations and rotations (commutative special orthogonal group SO(2) in two-dimensional space and special non-commutative group SO(3) in the 3D space). Transformations of the space described by the group of motions are studied by elementary geometry.

From properties of the external composition in axiomatic of the vector spaces follows that the velocity $\mathbf{v}$ and momentum $\mathbf{p}$, which differ by a scalar multiplier $M$, belong to the same vector space. Consequently, the velocities defined physically or mathematically (both cinematically and geometrically) are identical and four definitions of the body motion are equivalent. Excess of related invariants of motion

(any value from the set $M$, $\mathbf{p}$, $E$ is expressed in terms of the other two) provides a chance for construction of various forms of the solid mechanics (Newtonian, Lagrangian, Hamiltonian) and selecting the most convenient representation for specific problems.

## 3 Concept of "Fluid Flow"

Macroscopic hydrodynamics is based on the conception of "continuous medium", which occupies the whole space or a part of it and prescribes the use of continuous functions for the description of the entire range of scales. "Continuous medium", including substances in different phase states (liquid, gas, plasma), is described by the physical quantities of the dual nature that are mechanical and atomic-molecular. Mechanical quantity is the same density which characterizes the inertial and gravitational properties of the media, atomic-molecular properties determine dissipative (kinetic) and thermodynamic parameters (pressure, temperature, density).

An important property of a continuous medium is its fluidity that is the ability to deform or move under an arbitrarily small forcing. Decomposition of the fluid velocity, by definition of Cauchy-Helmholtz [9], in addition to translation and rotation operators, allows changing the shape of the medium element

$$v_i(r_r + \delta r_k) = v_i(r_k) + \varepsilon_{ijk}\Omega_j \delta r_k + \frac{\partial v_i}{\partial x_l}\delta r_l, \tag{2}$$

where $\varepsilon_{ijk}$ is the unit antisymmetric tensor of the third rank. Additional deformation term $\frac{\partial v_i}{\partial x_l}\delta r_l$ eliminates the independence of the translation and rotation operators, changes the group properties of operator of a fluid motion as a whole.

Conditions of homogeneity, continuity and deformability of a continuous medium are incompatible with demand of identifying individual "fluid particle", having no physically distinguishable boundaries. As the size of "fluid particle" tends to zero its mass also decreases indefinitely, and the object of study is lost in contrast with the solid body preserving its mass.

This contradiction is resolved by extending the list of physical parameters of the flow by introducing of the density field, making their own value at every point of the medium, which varies with time $\rho = \rho(\mathbf{x}, t)$ as independent physical quantity defined dynamically.

Incorporating the concept of variable density in theory in contrast with constant mass of the material point eliminates the condition of identity of vector spaces of velocity, determined by cinematically and momentum of the media since the density of a real fluid is independent variable depending on coordinates and time $\rho = \rho(x, y, z, t)$.

Difference in the decomposition of operators of the "motion" and the "fluid flow" reflects the contrast between empty space which is equivalent to the space with countable discrete material points of masses $M_i$, whose motion is the transformation

into itself and is characterized by a group, and continuum of a deformed medium immersed in the space.

Thus, to describe the body motion (or self-transformation of the space with mass $M_l$ in points $(i, j, k)$) is sufficient to perform transformations of three-dimensional Euclidean space $\mathbb{R}^3$. For flows description, in addition to space coordinates, it is necessary to address additional physical quantities characterizing the thermodynamic parameters of the medium, and consider their changes. It means that the dimension of the extended space of hydrodynamics problems is higher than 3D metric space. The unique physically reasonable criterion to select the physical quantities in the theory of fluid flows is the *observability* condition.

Conventionally continuous medium (fluid) is characterized by density $\rho(x_1, x_2, x_3)$, pressure $P(x_1, x_2, x_3)$, temperature $T(x_1, x_2, x_3)$, concentration of dissolved or suspended matters $S_i(x_1, x_2, x_3)$, thermodynamic potentials and their derivatives that are variables having a clear physical sense and being accessible to observation. This duality of physical quantities significantly complicates the description of fluid flows that are the self-consistent changes in the physical fields of different nature.

In this approach the main parameter characterizing a fluid flow is *momentum per unit volume of medium*, the flow invariant which is manifested in the forcing action of the flow on the body immersed in the liquid, or the flow rate through the selected cross-section.

For identifying fluid flows in the experiment different markers (solid particles, droplets of immiscible liquids or gas bubbles) having some unique features which allow to distinguish them from surrounding fluid and to trace the fluid velocity are used.

But introducing of an additional physical object (marker) is transformed a pure fluid into a new, more complex multicomponent medium. Moreover, every marker is a dynamically independent object with its own behavior which is distinguished from the flow of the basic fluid both in wavy [10] and vortex flows [11].

As an example of independent behavior of the carrying and marker fluids, deformation of the round spot of solvable dye placed in the center of cavern produced by the compound vortex in cylindrical tank is shown [11]. The round dye spot in the centre of a surface vortex cavern is deformed into spiral arm oriented opposite to the main stream (Fig. 1).

Solid-state marker in the same vortex flow is transported in the direction of the main stream slightly slowly than the carrying fluid, so that the soluble dye, washable out from its surface is mainly transported forward (Fig. 2a). The marker also slides over the surface of the vortex cavern, leaving behind the coloured wake (Fig. 2b). At the same time the marker twists about its own axis, disturbing the flow in its vicinity [12]. The combined effect of introduced perturbation of fluids and displacement of the dye separation point on the marker surface vastly complicates the form of the colored area, which is similar to turbulent flow structure. However, the dye from the liquid spot in the same flow forms the smooth spiral structures in Fig. 3, indicating homogeneity of the flow and absence of intensive transverse fluctuations.

Even more complex flow pattern is observed when the initial dye spot offsets from the center of the vortex (Fig. 3). In this case, the dye is mainly transported in

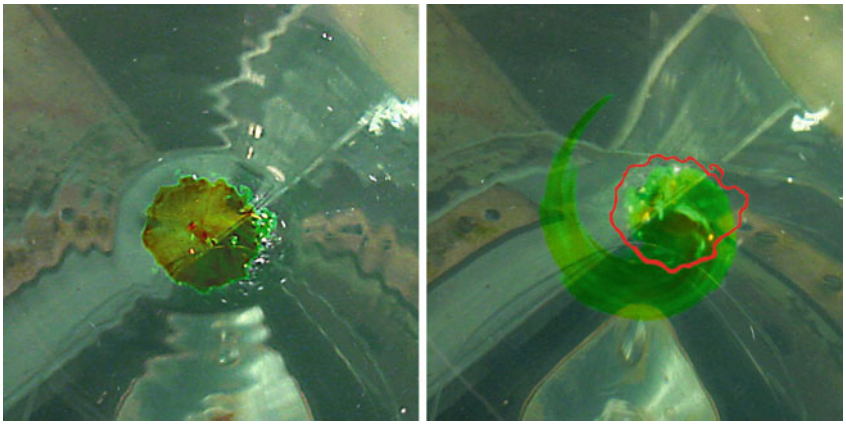


**Fig. 1** The transformation of the *round spot* of Uranil dye solution in the *centre* of cavern on the free surface of the compound vortex into one spiral arm (*red curve* in Fig. 4b is contour of initial dye spot [11])

**Fig. 2** Complex structure of coloured upstream and downstream ink dye wakes washed out from the plastic square marker free drifting on the surface of compound vortex cavern

the direction of the main surface flow to the cavern center [13]. The spot itself slides relative to its initial position and leaves the colored wake behind. At the same time all the dye bands split up into individual fibers. This behavior of the dye eliminates the possibility of rational determination of the position of the colored area center of mass necessary for calculating the fluid velocity.

The universal character of the simultaneous rotation of the solid body around a vortex center and twisting around its own axis was for the first time noticed by Decartes [14].

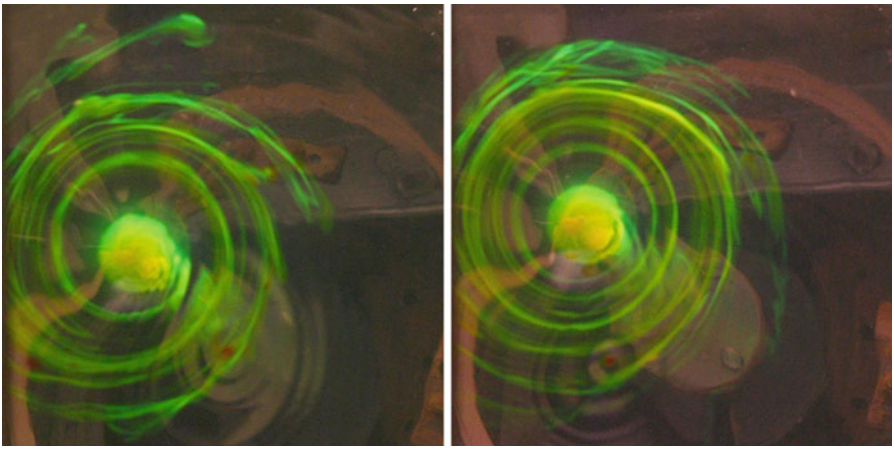


**Fig. 3** Spinning of noncentral Uranil dye spot on the free surface of compound vortex into pair of upstream and downstream spiral arms and splitting into individual fibers [13]

So, the present of any markers disturb the flow of the carrying fluid. Twisting of the solid body additionally perturbs the local environment. Too small markers are involved in Brownian motion. Diffusion effects impact on the transport of soluble impurities. The dynamics of droplets of immiscible liquids or gas bubbles affect surface tension effects and diffusion.

The combined action of many factors causes uncontrolled movements of markers different from the fluid flows, which they are submerged into. Conventional hypothesis of "passive admixture" of impurities is not proved mathematically or confirmed in experiments with a precision control of positions or concentrations of marker fields.

Absence of criteria of "liquid particles" identification means that the "fluid velocity" is a derivative mathematical parameter, but not physically observable physical characteristics of the flow. The necessity to satisfy the criterion of observability for physical quantities was stated by Stokes, Maxwell and Reynolds, and many others scientists, but analysis of the conditions of observability of the "fluid particle" was not previously held.

Observable parameter of the fluid flow is momentum $\mathbf{p}$ that can be evaluated measuring the forcing action of the fluid on the obstacle and the flow rate through the selected cross-section of the flow. Then the fluid velocity is defined as the ratio of the two observed variables that are momentum and density of the fluid $\mathbf{v} = \mathbf{p}(\mathbf{x}, t)/\rho(\mathbf{x}, t)$ in the given points of the absolute coordinate frame. Two independent methods of moment determination allow developing of the procedure for the direct estimation of the measurement error.

Hence, the conventional definition "Fluid flow is an intuitive physical notion which is represented mathematically by a continuous transformation of three-dimensional Euclidean space into itself..." [15], cannot be applied to fluid flows which should be interpreted as the momentum flux by the fluid, supplemented by variations of others

physical parameters. Attempt to trace the difference between concepts "motion" and "fluid flow" was made in [16], but the analysis was limited to the consideration of phenomena in 3D Euclidean space without employing condition of the physical quantities observability.

Once more important consequence of the differences in the concepts of "motion" and "fluid flows" is the next. In solid mechanics, there is only one macroscopic way to transfer mechanical energy and the time scale of the process is determined by the ratio of the linear scale (distance or body size) to velocity.

In fluid flows the energy of mechanical motion can transform into the internal energy and vice versa. Depending on the nature of the flow energy exchange can be defined as external (mechanical, as in the case of solid mechanics) and internal (atomic-molecular) degrees of freedom. Additional time scales can be large when the influence of diffusion effects is important, and are sufficiently small when a rapid release of the available potential energy takes place, for example, due to annihilation of free liquid surface when different fluids merge. Diffusion effects generate long-living structural features, and energetic short time phenomena produce high frequency flow components (for example, short capillary waves and sound when falling drops impact on a liquid surface [17, 18]). The additional free parameters complicate the picture of the process and make it difficult to develop universal methods for flow calculations.

So the principle of physical parameter observation prescribes the choice of the physical quantities, characterizing the fluid flow. The list of them includes momentum vector per unit volume of the medium **p**, as the indicator of acting force, matter and/or the energy transport and quantities of dual nature characterizing the mechanical and thermodynamic properties of the fluid that are density $\rho$, pressure $P$, temperature $T$, concentration of the $i$th impurity $S_i$, entropy $s$, total energy $E$. The equation of state is given as function of the density dependence on other thermodynamic quantities $\rho = \rho(p, T, S_i)$ and one of the thermodynamic potentials—internal energy $e = e(p, T, S_i)$, enthalpy, free energy, or Gibbs potential, as parameter defining the energy transformation. Accounting mutual impact of simultaneously occurring large-scale processes of the mechanical nature and atomic–molecular processes on small scales is necessary for self-consistent calculation of dynamics and structure of flows both at the global and fine scales. The basis for flow calculations constitutes conservation laws of observables physical quantities presented in the local differential forms.

## 4 Basic Equations and Local Symmetries

Set of fundamental conservation laws for the given fluid characterized by the equation of state $\rho = \rho(p, T, S_i)$ includes the continuity equation preserved in the form given by d'Alembert and Euler [1]

$$\frac{\partial \rho}{\partial t} + \nabla_i(p_i) = 0 \tag{3}$$

and balance equations representing laws of conservations of momentum, energy and substances re-wrote for open dissipative media that are Navier–Stokes equations in vector notation

$$\frac{\partial p_i}{\partial t} + \frac{\partial \Pi_{\alpha i}}{\partial x_\alpha} = \rho f_i, \tag{4}$$

Fourier equation describing the transfer of the total energy of an elementary volume of fluid including internal energy, the kinetic energy of motion and potential energy in the fields of external forces

$$\frac{\partial \rho \varepsilon}{\partial t} + \frac{\partial}{\partial x_i}(\varepsilon p_i + J^{i(\varepsilon)}) = 0. \tag{5}$$

Balance of components concentration is determined by generalized Fick's equation

$$\frac{\partial \rho(n)}{\partial t} + \frac{\partial}{\partial x_i}(p_{i,n} + J^i_{(n)}) = 0, \quad \rho = \sum \rho(n). \tag{6}$$

Here $\Pi_{ij} = p_i p_j/\rho + P\delta_{ij} - \sigma_{ij}$ is the momentum flux tensor; $\sigma_{ij}$ is the viscous stress tensor, $p_{i,n}$ is momentum of $i$th impurity, $f_i$ is the density of external forces (including Coriolis and buoyancy forces).

Calculations of the flow energy when it is necessary to take into account the change of internal energy are used on basis of the second law of thermodynamics for irreversible processes that is the condition of positive definiteness of the entropy production $P^{(s)}$ whose differential form is

$$\frac{\partial \rho s}{\partial t} + \frac{\partial}{\partial x_i}(s p_i + J^{(s)}) = P^{(s)}, \tag{7}$$

where $s$ is the entropy per unit mass, $\mathbf{J}^{(\varepsilon)}$, $\mathbf{J}_{(n)}$, and $\mathbf{J}^{(s)}$ are fluxes of energy lows, $n$th impurities and entropy.

The system (2)–(6) is written under the assumption of the existence of local thermodynamic equilibrium, supposing that the characteristic atomic-molecular processes are fast and equilibration times are substantially shorter than the characteristic times of mechanical processes forming gradients of thermodynamic quantities.

In the assumption of the smallness of the gradients, the set of equations (2)–(6) takes the traditional form firstly presented in the first edition of [19] published in 1944

$$\rho = \rho(P, S, T), \tag{8}$$

$$\frac{\partial \rho}{\partial t} + \nabla_i(p_i) = 0, \tag{9}$$

$$\frac{\partial(p_i)}{\partial t} + \left(\frac{p_j}{\rho}\nabla_j\right) p_i = -\nabla_i P + \rho g_i + \nu \Delta(p_i) + 2\varepsilon_{ijk} p_j \Omega_k + f_i, \tag{10}$$

$$\frac{\partial \rho T}{\partial t} + \nabla_j \cdot (p_j T) = \Delta(\kappa_T \rho T) + Q_T, \tag{11}$$

$$\frac{\partial \rho S_i}{\partial t} + \nabla_j \cdot (p_j S_i) = \Delta(\kappa_S \rho S_i) + Q_{S_i}, \tag{12}$$

where $\Omega_k$ is angular velocity of the global rotation of the liquid, $g_i$ is acceleration of gravity, $\nu$, $\kappa_T$, $\kappa_S$ are the kinematic viscosity, thermal conductivity and diffusion, $Q_T$, $Q_{S_i}$ are densities of salt and temperature sources, $\nabla$ and $\Delta$ are Hamilton and Laplace operators, respectively. Here velocity of fluid is determined dynamically as the ratio of two invariant quantities $v_i = p_i/\rho$. Type of the state equation is chosen taking into account the composition of the medium and the nature of the studied flows.

System (8)–(12) involves the dissipation of momentum, but does not consider the impact of processes with rapid changes in internal energy. Forcing and source terms in the right-hand sides of the set (8)–(12) can be permanent, slow with characteristic time $T_M \approx L/U$ and rather short ($\tau_m \ll T_M$) but intensive when describe action of strong atomic-molecular forces. Short time "shocks" which can locally perturb smooth flow fields manifest itself in generation of sound and another high frequency waves or in change of the fine flow structure.

Condition of compatibility applied to the set of the governing equations (8)–(12) or similar, determines the *rank* of the nonlinear set, the *order* of the linear version, the *degree* of the characteristic (dispersion) equation and, consequently, the number of independent functions that constitute the complete solution of the linearized problem. Rank and degree of the set can be defined as the order of the highest derivative if the set can be re-written for one of selected variables. Due to the historical tradition the set (8)–(12) is usually considerably simplified by omitting some terms (mostly linearized) or converted into constitutive set with the replacement of basic equations by new ones. Degree of correspondence for basic and transformed sets is not usually evaluated.

To search a specific problem solution of the basic set of equations is supplemented by physically reasonable initial and boundary conditions, representing a decay of the flow in dissipative medium with time and distance, condition of impermeability and no-slip of liquid on solid surfaces, constant stress on different sides of the contacting fluids interface.

System (8)–(12) is well-posed and self-consistent. The compatibility condition indicates that the set has high rank and high dimension of the extended space of the problem. General properties of the set are characterized by the family of local symmetries.

The calculations showed that the set (8)–(12) is invariant with respect to group of continuous symmetries for the operators [20]

$$
\begin{aligned}
&X_1 = \partial_t, \quad X_2 = \partial_x, \quad X_3 = \partial_y, \quad X_4 = \partial z,\\
&X_5 = y\partial_x - x\partial_y + v\partial_u - u\partial_v,\\
&X_6 = \left(z + \frac{gt^2}{2}\right)\partial_x - x\partial_z + (w + gt)\partial_u - u\partial_w,\\
&X_7 = \left(z + \frac{gt^2}{2}\right)\partial_y - y\partial_z + (w + gt)\partial_v - v\partial_w,\\
&X_8 = t\partial_x + \partial_u, \quad X_9 = t\partial_y + \partial_v, \quad X_{10} = t\partial_z + \partial_w.
\end{aligned}
\tag{13}
$$

The set (13) contains operators of shifts in time and space $X_1, \ldots, X_4$; rotations in the horizontal plane $X_5$; rotations in free fallen coordinate frame in a uniform gravity field $X_6$, $X_7$, and the last line represents generators of Galilean transformation groups $X_8$, $X_9$, $X_{10}$. Symmetries of the set (8)–(12) exactly correspond to the "first principles" of physics reflecting the invariant properties of the time and space and equivalence of all inertial coordinate frames which were used to construct the governing equations set [19].

Nonidentity transformations, which are widely used for simplification of the basic set or derivation of new sets of constitutive equations, change the form of symmetries. For example, neglecting compressibility, temperature and diffusion reduces the set (8)–(12) to the conventional Navier–Stokes equations for homogeneous fluid [19]

$$
\frac{\partial v_i}{\partial x_i} = 0 \quad \rho\frac{dv_i}{dt} = -\frac{\partial P}{\partial x_i} + \rho\nu\frac{\partial^2 v_i}{\partial x_l \partial x_l} + \rho\delta_{i3}g_i. \tag{14}
$$

In uniform gravity field $\mathbf{g} = \nabla\Phi$ the set (14) is transformed to the standard form by redefinition of pressure ($P' = \frac{P}{\rho} - \Phi$) and is characterized by the next rich family of the group symmetries generators:

$$
\begin{aligned}
&Y_1 = y\partial_x - x\partial_y + v\partial_u - u\partial_v,\\
&Y_2 = z\partial_x - x\partial_z + w\partial_u - u\partial_w, \quad Y_3 = z\partial_y - y\partial_z + w\partial_v - v\partial_W,\\
&Y_4 = \partial_t, \quad Y_5 = 2t\partial_t + \mathbf{r}\partial_{\mathbf{r}} - \mathbf{v}\partial_{\mathbf{v}} - 2P\partial_P,\\
&Y_{\chi_1} = \chi_1\partial_x + \dot{\chi}_1\partial_u - \rho\ddot{\chi}_1 x\partial_P, \quad Y_{\chi_2} = \chi_2\partial_y + \dot{\chi}_2\partial_v - \rho\ddot{\chi}_2 y\partial_P,\\
&Y_{\chi_3} = \chi_3\partial_z + \dot{\chi}_3\partial_w - \rho\ddot{\chi}_3 z\partial_P,
\end{aligned}
\tag{15}
$$

where $(\chi_1, \chi_2, \chi_3)$ are arbitrary functions of time (describing laws of coordinate frames motions).

In the family of generators (15) infinite-dimensional sub-algebras $Y_{\chi_1}$, $Y_{\chi_2}$, $Y_{\chi_3}$ appear, which are transformed into operators of shift if $\chi_1 = \chi_2 = \chi_3 = 1$ and into operators of Galilean transformations if $\chi_1 = \chi_2 = \chi_3 = 1 + t$. For arbitrary functions $(\chi_1, \chi_2, \chi_3)$ infinite-dimensional sub-algebras $Y_{\chi_1}, Y_{\chi_2}, Y_{\chi_3}$ generate transformations expanding Galilean relativity principle to coordinate frames moving with arbitrary rectilinear accelerations.

The family (15) contains operator of the group of expansion $Y_5$, which has no analogues in the set (13). The presence of the expansion operator explains the wide

spreading of the boundary layer approximations in hydro- and aerodynamics of homogeneous fluid [19]. Symmetries of sets (8)–(12) and (13) manifest deep difference in structures of solutions describing flows of homogeneous and stratified fluids. Constitutive models are mostly characterized by poor set of symmetries [20].

Currently, both in techniques of small approximations or constitutive models constructions the non-identity transformations of the basic set are used as a rule. Hence every constitutive or particular truncated system of governing equations is characterized by its own specific group of symmetry and a family of conserved parameters. In result the sense of the same symbols written in similar equations is changed and does not correspond to the sense defined by the fundamental set (8)–(12). Absence of universal description of fluid motion blocks development of experimental hydrodynamics.

Since mathematical methods for constructing complete solutions of such complex systems as (8)–(12) are not designed up to date, of practical interest are the results of qualitative analysis of equations, in particular, evaluation of their own temporal and spatial scales for stratified flows. In environmental and technological conditions the density of the liquid $\rho = \rho(x, y, z, t)$ is variable because of temperature and concentration fields inhomogeneities. The density gradient $d\rho/dz$ defines the stratification scale $\Lambda = |d \ln \rho/dz|^{-1}$, buoyancy frequency $N = \sqrt{g/\Lambda}$ and period $T_b = 2\pi/N$ which change within wide range. In the field of mass forces buoyancy effects lead to the formation of stable stratification, damping displacements of fluid layers in the direction of gravity.

However, changes the of real fluids density are usually small and produces small impact on the inertial properties of the flows. Nevertheless, conservation of terms describing the stratification effects in the governing equations set is important since gravity acceleration is large. In this regard, it is useful to consider three types of medium: *stratified fluids* when buoyancy scale frequency and period are in the list of main parameters; then weakly stratified fluids, when the scale of buoyancy substantially exceeds the values of other length scales of the problem (so called *potentially homogeneous fluid*); and *actually homogeneous liquid* whose density is assumed to be constant in the entire space. Using a weak but variable density helps to save the rank of complete non-linear set and order of the linearized set of governing equations and analyzes additional solutions which are lost in approximation of homogeneous fluid. Some of general properties of the basic set (8)–(12) solutions can be evaluated from analysis of intrinsic length scale.

## 5 Characteristic Length Scales of the Fluid Flows

The Eqs. (8)–(12) supplemented by conventional initial and boundary conditions is characterized by a number of distinguished length scales significantly different in magnitudes. Part of them is defined by parameters of the studied stratified medium and flow geometry. The rest depends on flow dynamics that is by characteristic momentum or velocity.

Large scales characterizing initial or boundary conditions includes the scale of stratification $\Lambda_\rho = |d \ln \rho / dz|^{-1}$, geometric lengths, e.g., the size of the obstacle $L$, and typical scales of a flow component like width of a jet, distinguished vortices or wave lengths of attached surface $\lambda_g = 2\pi U^2/g$ and internal gravity waves $\lambda_i = 2\pi U\sqrt{\Lambda/g}$ ($U$ is a relative velocity of the wave source with respect to surrounding fluid).

Set of small scales is defined by dissipative properties of media describing by kinematic coefficients in the set (8)–(12) and characteristic frequencies (buoyancy $N$ or global rotation $\Omega$) or velocity $U$. The list includes small scales of Stokes type $\delta_N^\nu = \sqrt{\nu/N}$, $\delta_N^{\kappa T} = \sqrt{\kappa_T/N}$, $\delta_N^{\kappa S} = \sqrt{\kappa_S/N}$ which are similar to the Stokes length scale $\delta_\omega^\nu = \sqrt{\nu/\omega}$ on the oscillating plane [19], Ekman type $\delta_\Omega^\nu = \sqrt{\nu/\Omega}$, $\delta_\Omega^{\kappa T} = \sqrt{\kappa_T/\Omega}$, $\delta_\Omega^{\kappa S} = \sqrt{\kappa_S/\Omega}$ or Prandtl type $\delta_U^\nu = \nu/U$, $\delta_N^{\kappa T} = \kappa_T/U$ and $\delta_U^{\kappa S} = \kappa_S/U$. The last group of scales characterizes fine structure in jets and wakes past the moving obstacle. Since kinematic coefficients are usually small the values of large and small length scales are distinguished on several orders of magnitude.

Besides the basic scales the fundamental set can be categorized by secondary scales which are important in some particular problems. For example, dissipative-gravity length scale $L_\nu = \sqrt[3]{\Lambda(\delta_N^\nu)^2} = \sqrt[3]{g\nu}/N$ defines critical conditions of the source size producing unimodal or bi-modal periodic internal waves [21]. Multiplicity of intrinsic scales reflects the complex internal flow structure of fluid flows. Large number of intrinsic scales is also associated with a high dimension in the extended space of the problem. In experiments macro scales of the length determine the area size of the visualized flow, which should contain all flows components and micro scales prescribing temporal and spatial resolution of the measuring and recording instruments.

Macro- and micro scale relationships define traditional dimensionless complexes that are Reynolds $\mathrm{Re} = L/\delta_U^\nu = UL/\nu \gg 1$ and Peclet numbers described effects of temperature and salinity $\mathrm{Pe}_T = L/\delta_U^{\kappa T} = UL/\kappa_T \gg 1$, $\mathrm{Pe}_S = L/\delta_U^{\kappa S} = UL/\kappa_S \gg 1$. These ratios are large in the environment and laboratory experiments. In most cases, the change in density of the flow scale is small and ratios of length scale $C = \Lambda/L = \rho_o/\delta\rho \gg 1$ and $C_N^\nu = L/\delta_N^\nu = \sqrt{L^2N/\nu} \gg 1$ with the kinematic viscosity or $C_N^{\kappa T}$, $C_N^{\kappa S}$ with thermal or substance diffusivity are large. The presence of large relationships in the set with small coefficients in the terms along with the highest derivatives gives a room for theory of singular perturbations to calculate a wide range of processes, primarily slow flows such as diffusion induced on topography or small amplitude internal waves describing by the linearized set of governing equations.

As calculations show, all the flow components propagate coherently and fill the entire space. Mechanical energy is transported by large-scale components. Energy dissipation and vorticity are associated with the fine-structure components, which form the flow structure. Contrast of the flow pattern is underlined by contaminants which are accumulated on interfaces. Loci of fine flow components depend on geometry and energy of the processes.

## 6 Classification of Infinitesimal Components of Periodic Flows

Periodic, (or more exactly, almost periodic) flows, such as compact vortices and waves are traditional objects in fluid mechanics. Studies of known forms of oscillations and waves unify different branches of science, vortices are specific to liquids and gas flows. Conventionally, many types of waves are studied on the basis of special approximate equations describing (more or less completely) basic features of particular phenomenon such as, e.g., acoustic waves in compressible media, gravity surface or internal waves in stratified fluid or inertial waves in rotating environment [19]. However, since the symmetry of wave equations and the fundamental set of the Eqs. (8)–(12) are significantly different [20], solutions of model systems can not describe all the properties of periodic flows, or reflect wrongly some of them. In this regard, it is necessary to study periodic solutions of the non-linear fundamental set, taking into account the condition of compatibility of its constituent equations.

Due to complexity of the fundamental nonlinear set, at the first stage the complete solution of the linearized version is constructed. Attention is paid to study the wave processes in a viscous fluid described by orthogonal functions. To select exponential functions for investigation of elementary waves, the problem was analyzed in an infinite medium filling the whole space.

Complete solutions were constructed using singular perturbation theory that is an expansion on direct and inverse value of a small parameter $\varepsilon$ of the problem. In unbounded stratified media all variables of small periodic motions with fixed real positive frequency $\omega$ and complex wave number $\mathbf{k} = \mathbf{k}_1 + i\mathbf{k}_2$ (describing wave energy dissipation) can be found in form of products of individual amplitude factor $\mathbf{v} = \mathbf{v}_0\tau(r,t)$, $\overline{p} = p_0\tau(r,t)$, $\overline{\rho} = \rho_0\tau(r,t)$ by integrals of plane waves $\tau(r,t) = \exp(i(\mathbf{kr} - \omega t))$. Stationary waves are searched in form of Fourier expansions

$$A = \sum_j \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} a_j(k_x, k_y) \exp(i(k_{z_j}(k_x, k_y)z + k_x x + k_y y - \omega t))dk_x\, dk_y, \tag{16}$$

where $A$ is pressure, density, components of velocity, temperature or salinity. Here summation is performed over all roots of dispersion relation, following after substitution of expansion (16) into the linearized set (8)–(12), supplemented by boundary conditions on solid boundaries and at infinity.

Dispersion relation follows from compatibility condition after substitution (16) into linearized set of the Eqs. (8)–(12) and is presented below in factorized form of the algebraic equation of tenth power [22]

$$D_v(k,\omega) \cdot F_\omega(k,\omega) = 0, \tag{17}$$

where

$$F_\omega(k,\omega) = -D_\nu(k,\omega)D_{\kappa_T}(k,\omega)D_{\kappa_S}(k,\omega)\left(k^2 + i\frac{k_z(\Lambda_T+\Lambda_S)}{\Lambda_T\Lambda_S}\right)$$
$$+\ D_{\kappa_T}(k,\omega)\left(\frac{\omega k_z}{\Lambda_S}D_\nu(k,\omega) - N_S^2k_\perp^2\right) + D_{\kappa_S}(k,\omega)(\frac{\omega k_z}{\Lambda_T}D_\nu(k,\omega)$$
$$-\ N_T^2k_\perp^2),$$

$$D_\nu(k,\omega) = -i\omega + \nu k^2,\ \ D_{\kappa_T}(k,\omega) = -i\omega + \kappa_T k^2,\ \ D_{\kappa_S}(k,\omega) = -i\omega + \kappa_S k^2,$$
$$k^2 = \sum_i k_i^2,\quad k_\perp^2 = k^2 - k_z^2.$$

Regular solutions of the equation (17) have typical wave forms with large real and small imaginary part $|\mathbf{k}_1| \gg |\mathbf{k}_2|$. Another part of solution (17) has typical singular perturbed form and can be expand to series where exponential factor $\gamma$ is defined after substitution in governing equation

$$k_z = \varepsilon^{-\gamma}(k_0 + \varepsilon k_1 + \varepsilon^2 k_2 + \cdots),\quad \gamma > 0.$$

Imaginary parts of these roots are not small $|\mathbf{k}_1| \approx |\mathbf{k}_2|$ and inverse proportional to dissipative coefficient $|\mathbf{k}| \approx \sqrt{\omega/\nu}$. They describe small scale flow components.

Hierarchy of fluid mechanics fundamental equations sets is presented in Fig. 4. Complete set of the governing equations (8)–(12) on the Level A for 7 variables (vector of momentum, density, pressure, temperature and salinity) has tenth order. Among its periodic solutions there are two of regular types while the rest eight are of singular perturbed type. Part of solutions can be omitted due to boundary conditions on infinity [23]. Regular disturbed solutions in of zero buoyancy frequency limit match continuously with solutions of Navier–Stokes and Euler sets.

The linearized Navier–Stokes equations have 6th order for both stratified and homogeneous fluids (level C). System is solvable and all small scale flow components are different in stratified environment (left column). One of them is universal and does not depend on geometry (Stokes type). The second fine viscose component is specific internal and depends on the angle $\varphi$ of a solid boundary slope to the horizon corresponding to terms containing $k_\perp^2$ in Eq. (17). In limiting case of homogeneous fluid both singular solutions become identical, so the problem becomes degenerated and ill-posed [22].

In the right column of Fig. 4 degenerate sets are placed on levels B, C and D. Degeneration can be caused by geometrical and physical reasons. In the first case due to symmetry of the boundary (round body with vertical axis) the reduction of configuration space dimension of the problem occurs and initially 3D problem transforms into 2D [23] or even 1D problems. These approximations lead to loss of some solutions or even whole classes of solutions (for example, for 3D motions on long cylinder).

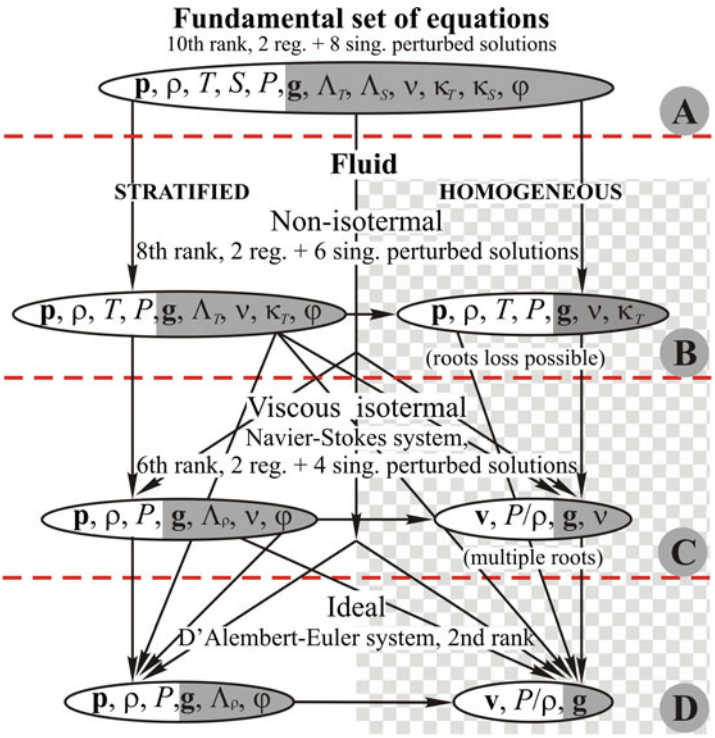



**Fig. 4** Hierarchy of fluid mechanics fundamental equations sets

Approximation of homogeneous fluid results in another type of degeneration (right column, level C). In this case, different solutions of viscous fine structure flow become identical and the 3D problem becomes ill-posed. Only particular solutions containing part of the whole family of complete solutions are calculated in this case. In the ideal fluid approximation (level D) only regular part of solutions is described.

The arrows connecting the individual elements of the scheme indicate the directions of possible transitions between different levels which can be regular or asymptotic. Models of the highest order (level A) allow the transition to the lower-order model when the dissipative coefficient or density gradient tend to zero. In such transitions two final states are formed. In one of them, in the case of *potentially homogeneous fluid*, density variations are considered to be negligibly small and the stratification is essentially unaffected by the dynamics of processes. However, the conservation of system rank leaves the set of equations well-posed and solvable. In another *actually homogeneous fluid* the density of the liquid assumed to be constant from the beginning. In this case the system of 3D equations degenerates and becomes ill-posed.

All properties of the complete and reduced solutions illustrate the theory of periodic internal wave generation in a continuously stratified liquid when the source of

waves is oscillating part of an infinite plane oriented under arbitrary angle to horizon. This problem is 2D and 3D generalizations of classical Stokes problem of oscillating plane in a homogeneous viscous fluid [19]. Calculations are compared with the data of schlieren visualization of the wave fields in tank filled with continuously stratified brine [24].

## 7 Diffusion Induced Flows on Obstacles

Diffusion-induced flows (DiF) are the most universal forms of flows in the environmental and laboratory conditions because for their generation only stable stratification and non-trivial geometry of a solid boundary are required. They play a key role in the processes of passive substances transport in the atmosphere and the hydrosphere. Diffusion induced flows are manifested as intensive valley and mountain winds in the stably stratified atmosphere and density flows in oceans. A great number of physical processes are essentially influenced by diffusion-induced flows, such as melting of icebergs, migration of tectonic plates and transport of minerals and bio-spices. It may also trigger a propulsion mechanism of self-movement of neutral buoyancy solids in stably stratified fluid. The problem represents an advance in knowledge of the fundamental aspects of the generic problem of non-equilibrium processes in fluids as an example of the successful integration of analytics, experiments and numerical simulations to explore physical mechanisms.

Numerical solution of the 2D truncated non-linear set (8)–(12) where temperature effects were neglected was constructed to describe both large and small scale flow components

$$\begin{aligned} &\rho = \rho_{00}(\exp(-z/\Lambda) + s), \quad \operatorname{div}\mathbf{v} = 0, \\ &\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\nabla)\mathbf{v} = -\frac{1}{\rho_{00}}\nabla P + \nu\Delta\mathbf{v} - s\mathbf{g}, \\ &\frac{\partial s}{\partial t} + \mathbf{v}\cdot\nabla s = \kappa_S \Delta s + \frac{v_z}{\Lambda}. \end{aligned} \tag{18}$$

Here, $s$ is the salinity disturbance including constant salt contraction coefficients, $\mathbf{v} = (v_x, v_y, v_z)$ is vector of the induced velocity, $P$ is the pressure except for the hydrostatic one. The consideration is conducted in the laboratory coordinate frame directed by the gravity force (undisturbed fluid is at the state of rest). The problem corresponds to the level B in the general scheme (Fig. 4).

At the initial moment $t = 0$ a thin impermeable sloping plate of length $L$ is placed into a quiescent stratified fluid free of any mechanical disturbances. Physically valid initial and boundary conditions, which are unperturbed flow fields before the initial moment, no-slip for velocity components and no-flux for substance on the plate's surface and attenuation of all perturbations at infinity were used

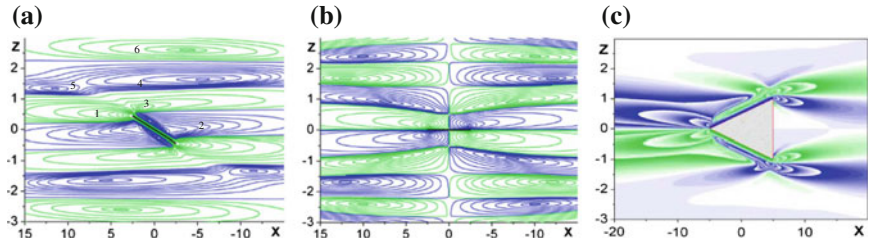



**Fig. 5** Pattern of streamlines of diffusion induced flows: **a**, **b** on the sloping and horizontal strip of length $L = 5\,\mathrm{cm}$; **c** on the wedge

$$\mathbf{v}, s|_{t\le 0} = 0, \tag{19}$$

$$v_\xi|_\Sigma = v_\zeta|_\Sigma = 0, \tag{20}$$

$$\left[\frac{\partial s}{\partial \mathbf{n}}\right]\Big|_\Sigma = \frac{1}{\Lambda}\frac{\partial z}{\partial \mathbf{n}}, \quad \mathbf{v}, s|_{\xi,\zeta\to\infty} = 0, \tag{21}$$

where $\mathbf{n}$ is external normal to the plate's surface $\Sigma$.

The problem (18)–(21) was analyzed numerically using an open software package OpenFOAM [25]. The developed algorithm works in all ranges of the flow parameters corresponding to laboratory, atmosphere and hydrosphere conditions, including zero angles of inclination of the impermeable surface to horizon when conditions of existence of stationary asymptotic solutions are violated. Calculations showed individual patterns of various physical parameters of the problem, part of which is represented in Figs. 5 and 6.

Due to interaction of the buoyancy and dissipation effects pattern of induced flow is rather complex and includes vortex cells with size of the obstacle length and sequence of short waves near the tips. In the flow patterns of streamlines thin interfaces separate a number of regular cells (positive direction by rotating marked green). Above the horizontal plate cells of different signs are located oppositely relative to the principal planes. Asymptotic calculations of stationary flow cannot be

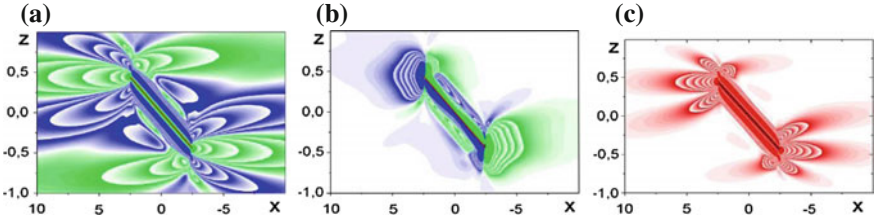



**Fig. 6** Fields of disturbances in the diffusion induced flows around the plate inclined under angle $\varphi = 10°$ to horizon ($N = 1.26\,\mathrm{s}^{-1}$, $L = 5\,\mathrm{cm}$): **a** dynamic vorticity $\boldsymbol{\Omega}$; **b** tempo of vorticity baroclinic generation $\dot{\boldsymbol{\Omega}}$; **c** the rate of the mechanical energy dissipation $\varepsilon$ (different scales are on the axes)

performed by methods proposed in [26] in the whole space because of alternative convergence and divergence of flows on the centerline above and below the strip.

There exists a noticeable difference in fields of stream functions (for multicellular agreed rotation directions in adjacent cells, typical for systems of internal waves), vorticity and velocity dissipation of mechanical energy presented in Fig. 6.

Complex fine structure of a flow is expressed in the dynamic vorticity field $\boldsymbol{\Omega} = \mathrm{rot}\,\mathbf{v}$ (Fig. 6a). Due to the line of isopycnics and isobars crossing an additional vorticity is generated baroclinically with the tempo $\dot{\boldsymbol{\Omega}} = d\boldsymbol{\Omega}/dt = \nabla P \times \nabla \rho^{-1}$ both in the close vicinity and at some distance from the obstacle (Fig. 6b). The formation of new fine components in the vicinity of the edges of the plate is caused by the joined action of buoyancy, limiting the height of elevation or sinking of separating jets, viscosity and diffusivity effects (Fig. 6b). Complex picture of mechanical energy dissipation rate (Fig. 6c) differs significantly from the smooth field of streamlines. Field of the mechanical energy dissipation rate $\varepsilon$ has specific "rosettes" shape, which is typical for pattern of dissipative gravity waves or "zero frequency" waves [27].

The calculated field of the density gradient perturbation in the diffusion induced flow on the horizontal or inclined plate, and the wedge, manifesting large-scale components, defined by the obstacle size, and thin interfaces with transverse scales $\delta_N^{\nu} = \sqrt{\nu/N}$ or $\delta_N^{\kappa_S} = \sqrt{\kappa_S/N}$, at long times is consistent with the schlieren image of the flow. The "natural rainbow method" with horizontal slit and regular grating [24] was used for visualization the of the refractive index gradient near the plate in a laboratory tank (Fig. 7a, b).

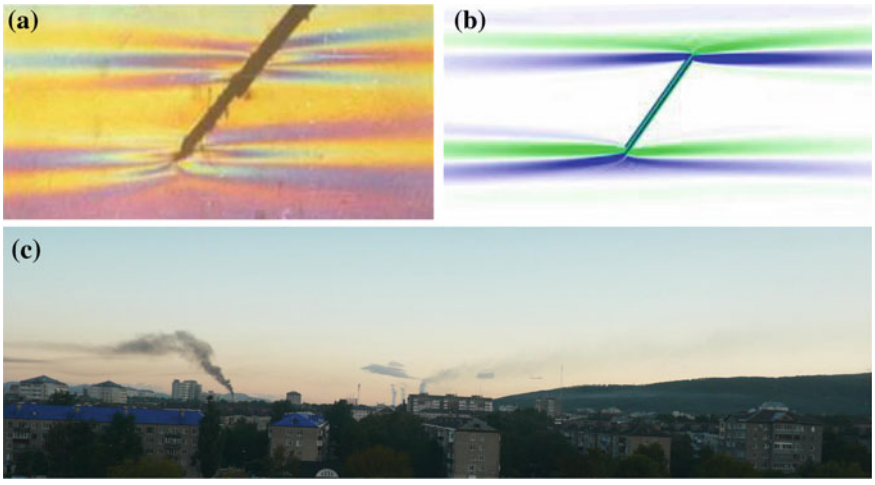



**Fig. 7** Images of diffusion induced flows: **a**, **b** schlieren and numeric visualization of density gradient perturbations on sloping motionless strip ($L = 5\,\mathrm{cm}$, $N = 0.84\,\mathrm{s}^{-1}$, $T_b = 7,5\,\mathrm{s}$, $\varphi = 40°$); **c** smokes and water vapor in Yuzhno-Sakhlinsk valley

In the images in Fig. 7, one can see standing out extensive streaky structures which are directly adjacent to the extreme points of the obstacles. The length of the interfaces increases with the sensitivity of the registration method. Flow induced by diffusion lead to self-induced motion of free neutral buoyancy obstacles in a stratified medium which is absent in a homogeneous fluid. Calculated velocities of fluid, the forces and angular moments acting on the wedge, consistent with the data of direct measurements of the velocity of self-motion of free neutral buoyancy wedge in a laboratory tank [28].

Detailed calculations of velocity profiles in the middle of the sloping plate is consistent with the formulas derived by Prandtl [26], but the thickness of salinity perturbation layer is less than for the velocity layer.

Diffusion induced flows, often are observed in the atmosphere with on relief with topography inhomogeneities (mountain or valley winds). Their intensity can reach high values in a relatively thin surface layer, especially on glaciers or in time of a bright sunlight, as well as in marine environments, where are the determining factors are gravity and rotation. Details of the flow patterns depend on the profiles of the buoyancy frequency and geometry of the underlying surface. The overall pattern of the flow in the cavity (double vortex and sinking jet over the center and outgoing stripe flows along the valley slopes) is often presented among the atmospheric processes [29]. As an example of the diffusion induced flow pattern in the atmosphere the photographs at first glance paradoxical pictures of industrial smokes in Yuzhno-Sakhalinsk (Fig. 7c) is given.

## 8 Theory of Periodic Internal Wave Propagation

The last fifty years, much attention has been paid to internal waves—a special kind of gravitational waves inside a stably stratified fluid. Internal waves are generated in the atmosphere and ocean during restructuring of large-scale processes, flows around topography, under the action of periodic and random force factors and transport data of the source parameters over long distances. The waves accelerate jet streams, intensify mixing, and heat upper atmosphere. In the ocean depths, these waves control the variability of physical fields and can destroy technical installations.

Under natural conditions, the structure, spectra, and amplitudes of wave fields can be determined. Since environmental variability makes it difficult to identify sources of the generation and paths of propagation of internal waves, the observations of natural systems are complemented by laboratory and mathematical modeling. The calculations of beams of periodic waves in a viscous stratified medium are in agreement with experimental data for different types of stratification.

The problems of internal wave generation are usually solved in the approximation of an ideal fluid or by introducing mass and force sources. The resulting wave patterns are generally consistent with laboratory data; however, the parameters of model sources depend on experimental conditions. Inclusion of dissipative factors gives a room for more complete description of the dynamics and structure of internal waves.

In the analytical description of waves, the Boussinesq approximation for an incompressible exponentially stratified fluid is mostly used. When considering the infinitesimal wave, the system of equations (8)–(12) is usually truncated, only diffusion of the stratified component is counted (in the calculations—salt in water) and have the form

$$\begin{aligned} \rho_0 \frac{\partial \mathbf{v}}{\partial t} &= -\nabla P + \nu \rho_0 \Delta \mathbf{v} - \sigma \rho_0 g \mathbf{e}_z, \quad \nabla \cdot \mathbf{v} = 0, \\ \frac{\partial \sigma}{\partial t} - \frac{\nu_z}{\Lambda} &= \kappa_S \Delta \sigma, \quad \rho_0(z) = \rho_0(S(z)) = \rho_{00}(z) e^{\frac{z}{\Lambda}}, \end{aligned} \tag{22}$$

where $P, \sigma$ and $\mathbf{v}$ are the disturbances of fluid pressure, concentration (with included salt contraction coefficient) and the velocity vector, respectively, $\nu$ and $\kappa_S$ are the constant coefficients of kinematic viscosity and diffusion, respectively; the unitary vector $\mathbf{e}_z$ is directed upward. In a general case the buoyancy scale $\Lambda(z)$, frequency $N(z)$ and period $T_b(z)$ are functions of the depth $z$. The problem correspond to level B in scheme presented in Fig. 4.

From dispersion relation for elementary periodic waves $\sin\vartheta = \omega/N$ it follows that waves propagate from the source along a radius-vector whose slope to the horizon is determined by the ratio of the wave $\omega$ to buoyancy $N(z)$ frequency, and became vertical at the critical height where both frequencies become equal [30]. The width of the wave beam and the amplitude of vertical displacement $h(x, z)$ which are also changed presented in calculations [31] in the spectral form $h(x, z) = \int_0^\infty f(z, k) \exp(ikx) dk$, $k$ is wave number.

Solution of the equations (22) in the coordinate frame $(q, p)$ associated with the wave beam the vertical displacement of fluid particles is described by the formulae [31]

$$h(q, p) = \sqrt{\frac{N_\omega(z_0)}{N_\omega(z_1)}} \int_0^\infty A_0(k) \exp\left\{ikp - \frac{\nu_d k^3 [Q(z_1) + q]}{2N(z_1)\cos\vartheta}\right\} dk, \tag{23}$$

where $A_0(k)$ is spectral amplitude of the source, $\nu_d = \nu + \kappa_S$ is sum of dissipative coefficients, $N_\omega^2(z) = \frac{N^2(z)-\omega^2}{\omega^2}$ is difference between buoyancy and wave frequency square normalized by the wave frequency.

*Local length scale* along the center line of the wave beam is

$$Q(z) = \frac{N_\omega(z)}{\sqrt{(1 + N_\omega^2(z))^3}} \int_{z_0}^{z_1} \frac{(1 + N_\omega^2(z'))^2}{N_\omega(z')} dz',$$

where the points $z_0, z_1$ are positions of wave source and observations.

Length scale $Q(z)$ associated with the geometric length of the center line $L(x) = \int_{z_0}^{z_1} \sqrt{1 + N_\omega^2(z')} dz'$ by the integral relation

$$Q(z) = \frac{\sqrt{L_z^2(z) - 1}}{L_z^3(z)} \int_{z_0}^{z_1} \frac{L_z^4(z')}{\sqrt{L_z^2(z') - 1}} dz', \quad L_z(z) = dL/dz.$$

It stratification is exponential $N(z) = \text{const}$, $Q(z) = L(z)$ the expression for wave amplitudes takes the conventional form

$$h(q, p) = \int_0^\infty A_1(k) \exp\left\{ikp - \frac{\nu_d k^3 [L_1 + q]}{2N_1 \cos\vartheta}\right\} dk$$

of the general internal wave type [30].

In regions where buoyancy frequency decreases with height and become less than the frequency of the propagating wave, the beam is reflected from critical level and partially leaks into non-wave zone. For this case, typical wave beams shape is shown in Fig. 8 (different sensors of the wave induced displacements of fluid layers are shown too). Amplitude attenuations in reflected and leakage through critical level 2D internal waves beams calculated in [31] were tested in special experiments, where velocity field was measured by particle image velocimeter (PIV) instrument. The vertical velocity fields of the incoming and reflected waves agree within few percent with theory of beams in an arbitrary smooth stratification [32].

Unfortunately PIV instruments are characterized by low spatial resolutions, in experiments [32] $\delta x = 3.9\,\text{mm}$ and $\Delta z = 2.9\,\text{mm}$. These values exceed the

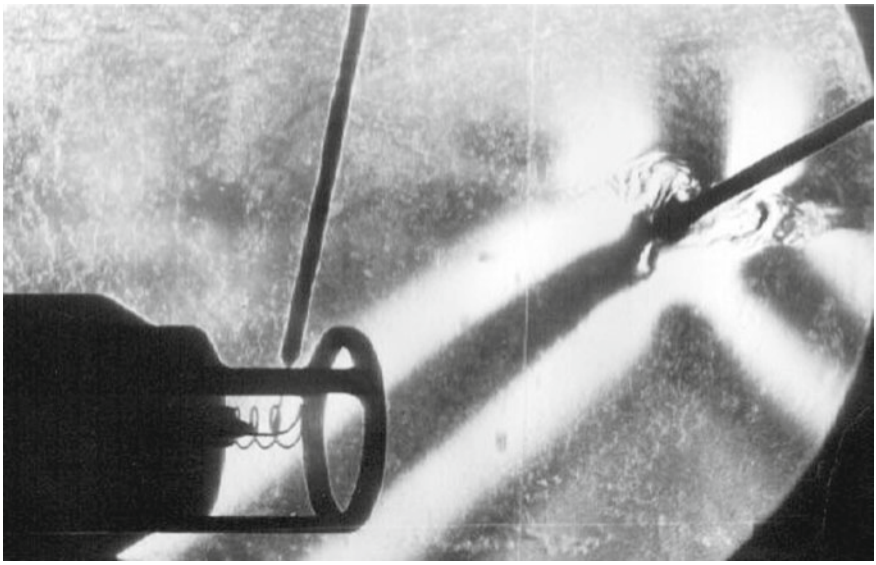


**Fig. 8** Periodic internal waves beams near a critical level in a stratified fluid with the non-uniform profile of buoyancy frequency

microscales $\delta_N^{\nu}$, $\delta_N^{\kappa S}$, $\delta_U^{\nu}$, $\delta_U^{\kappa S}$ so the instrument [32] cannot resolve fine scale flow components accompanying the periodic internal wave beam and forming on critical level, as it was calculated in [31], too.

## 9 Generation of Periodic Internal Waves

In studying of the periodic internal wave generation problem the linearized set of governing equations (22) is supplemented by no-slip and no-flux boundary conditions on all moving and stagnant solid boundaries. Detailed calculations were performed for oscillation strip [33] or disk placed on horizontal [21] or sloping infinite plane, or moving part of vertical cylinder [23]. The problem corresponds to level C in the hierarchy in Fig. 4.

All periodic flow components propagating along with internal waves in the stratified viscous fluids were calculated analytically for a circular piston performed periodic oscillations on horizontal plane. Regular components occupy the whole space (weak background in Fig. 12a, b) and are the most profound in the wave cone sloping under the angle $\vartheta$ to horizon ($\sin\vartheta = \pm\omega/N$). The ratio of disc diameter to viscous wave length scale ($L_{\nu}^{\kappa S} = \sqrt[3]{(\nu+\kappa_S)g}/N$) defines modal structure of the beam. The fine (singular) components are formed on emitting surface and create thin envelopes of the wave cone. Transverse scales of the singular components are characterized by two universal microscales of velocity $\delta_N^{\nu} = \sqrt{\nu/N}$ and salinity (concentration) $\delta_N^{\kappa S} = \sqrt{\kappa_S/N}$ fields, respectively.

The singular components are characterized by a high level of vorticity and rate of energy dissipation. During the wave period, the envelopes are being gradually distributed across the whole beam and finally concentrated into thin edges interfaces. Outer and inner envelopes are shown in Fig. 9a, b in convergence phases. Locations of singular components of periodic internal waves demonstrate pattern of the second derivatives of velocity in Fig. 9c.

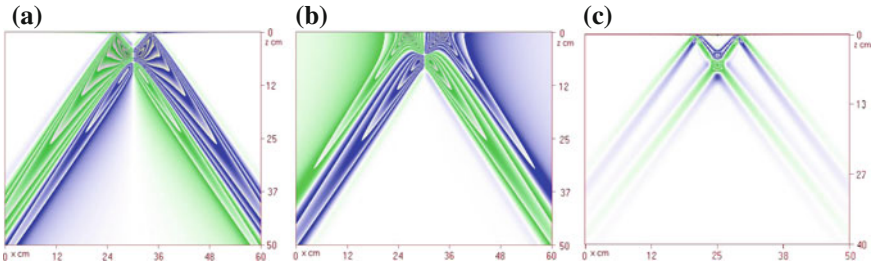



**Fig. 9** Calculated pattern of flow in the central cross section of the conical periodic wave beam produced by a horizontal disc of radius $R = 4$ cm oscillating in the vertical direction ($\omega = 1\,\mathrm{s}^{-1}$, velocity amplitude $U = 0.25$ cm/c, $T_b = 5.2$ s); **a**, **b** the velocity horizontal component $v_r$; $t = 0; 0.25T_b$; **c** its second derivative $\partial^2 v_r/\partial z^2$, $t = 0$ [21]

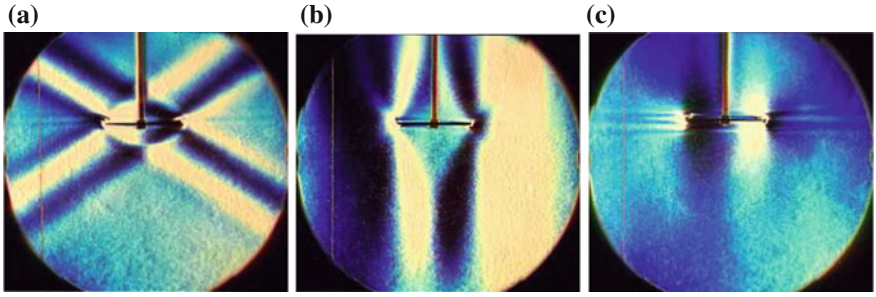



**Fig. 10** Schliren image of internal waves produced by vertically oscillating disc in a stratified brine ($N = 0.88\,\mathrm{s}^{-1}$, $D = 5\,\mathrm{cm}$, $A = 0.25\,\mathrm{cm}$): **a–c** $\omega/N = 0.55; 0.97; 1.27$

Schlieren images of periodic internal wave beams of small amplitude produced by vertically oscillating disc in continuously stratified brine are given in Fig. 10. Due to axial symmetry of the flow only central cross section is visualized. All four wave beams are well outlined in Fig. 10a when $\omega < N$. Images of the wave beams are antisymmetric since motions of the fluid particles occur in opposite directions above and below the disc. Right and left parts of the image in Fig. 10 are antisymmetric also as horizontal displacements in beams occur in opposite directions.

Thin horizontal strips in the central part in Fig. 10a represent the flows, induced by diffusion on the disc. Elliptic domain around the disc shows a near field region where non-linear effects are significant. Small edge vortex rings are formed in vicinity of the disc trajectory turning points.

Internal waves crests and troughs are oriented vertically when the frequency of the disc oscillations reaches the buoyancy frequency ($\vartheta = \pi/2$). In this case, the singular disturbances become weaker, and the near field region in Fig. 10b is extended in vertical direction. Weak secondary waves are visible in upper left part of the picture. Both regular and singular elements of fluid motions are gradually shrinking to the source if the frequency of oscillations exceeds the buoyancy frequency ($\omega > N$) and only edge vortices and diffusion induced flows still exist in Fig. 10c.

While regular and singular components are emitted by the disc edge along all available directions non-linear effects are most expressed inside the conical domain located directly under the disc where the waves are intersected and the inner envelopes are converged.

More intensive internal waves are produced by the vertically oscillating sphere. In this case high gradient beam envelopes are visualized by schlieren instruments. Thin dashed lines in the central part in Fig. 11a, b mark the body turning horizons. Elliptic domain around the sphere and double dark vertical lines near the sphere poles in Fig. 11 show the near field region where non-linear effects are significant.

Dark sloping lines bounds internal wave beams of large amplitude. Sharp mushroom-like interfaces are formed in domains of convergences of high gradient envelopes of the wave beam directly in the fluid body without any contact with solid boundaries. Short horizontal interfaces do not destroy the stratification in Fig. 11c.

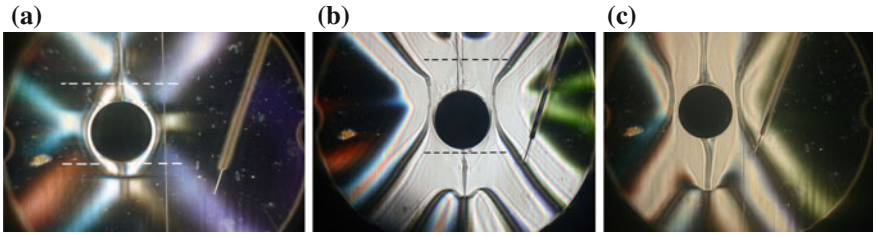



**Fig. 11** Schlieren images of periodic flows induced by oscillating sphere ($D = 4.5$ cm), (method "slit-thread"): **a–c** $T_b = 11.2, 7.3, 11.2$ s; $H = 1, 2.8, 2.8$ cm, $\omega/N = 0.73, 0.8, 0.8$

Gradually, high gradient interfaces outlines the near field domain around moving sphere. Their vertical extension exceeds the size $H$ of the sphere displacement. Forming vortices covered by high gradient interfaces moves to the sphere in contrast with local disturbances moving in opposite direction [34].

Calculations show that wave and fine flow components propagate coherently, and fill the entire space. Mechanical energy is transported by large-scale components. Energy dissipation and vorticity are associated with the fine-structure components, which form the flow structure. Contrast of the flow pattern is underlined by contaminants which are accumulated on the interfaces. The need for registration of multiscale flow components imposes additional requirements on experimental methods that need to visualize large-scale and resolve fine components. In conventional experiments, fine components are not resolved and manifesting themselves as random fluctuations of the measured parameters.

## 10 Fine Structure of the Flow Field Behind Uniformly Moving Obstacle

Uniform motion of an obstacle in the stratified fluid radically change the image of the flow. Periodic internal waves are transformed into attached waves stationary moving with the source [30] and a rich family of fine flow components is observed. The real flow structure depends on parameters of the fluid and stratification, shape of the obstacle and its velocity. Here the examples of the flow structure evolution are given for the uniformly towing body of the most simple shape that is for vertical strip [35].

On the parameter range under current study the optical visualization made it possible to reveal the flows of three types: laminar flow with clearly expressed upstream blocked fluid and split downstream density wake bounded by two plane density envelopes (Fig. 12a); transitional flow with imbedded oblique streaky structures and attached internal waves in the process of their formation (Fig. 12b); stratified flow with developed field of attached internal waves, and bubble vortex inside the downstream wake with pronounced fine structure (Fig. 12c).

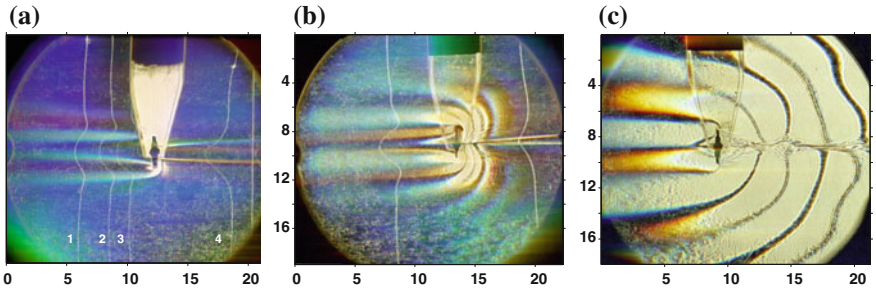



**Fig. 12** Schlieren images of flow *side view* around uniformly towing vertical strip of height 2.5 cm in a continuously stratified fluid. *Curved vertical lines* are density markers used for precise measurements of buoyancy frequency and horizontal component of velocity: **a** upstream and bounded by high gradient envelope downstream wake ($T_b = 17.4$ s, $U = 0.033$ cm/s, Fr $= 0.036$, Re $= 8.25$); **b** sloping streaky structures in the downstream wake ($T_b = 12.5$ s, $U = 1.06$ cm/s, Fr $= 0.084$, Re $= 26.5$); **c** fine structure of downstream vortex bubbles synchronized with pronounced pattern of attached internal waves ($T_b = 12.5$ s, $U = 0.37$ cm/s, Fr $= 0.3$, Re $= 92.5$)

At low velocities introduced disturbances are concentrated in a relatively narrow layer near the body motion horizon. In this layer a smooth upstream disturbance and a downstream density wake bounded by high-gradient envelopes are clearly expressed in Fig. 12a. Ahead of the strip, two groups of dissipation-gravity waves (dark and light long bands) proceed from the plate edges; they enclose the central region of completely blocked fluid which moves together with the plate. The vertical dimension of the blocked fluid decreases with the distance from the plate.

The contours of density markers 1–4 (vertical curved lines in Fig. 12a) illustrate the profile of the horizontal velocity component (the rightmost marker is the vertical reference line). The shapes of the density markers 1–3 ahead of the strip demonstrate the influence of oppositely-directed jets, namely, the flow between the plate horizons and the counter-flows above and below its edges. The internal density marker 3 in the blocked fluid retains its original vertical shape, which indicates the motion of the fluid as the whole, at the velocity equal to that of the body.

The density marker 4 past the body has single maximum at the axis of the motion. Within the velocity shear layers there are thin high-gradient envelopes bounding the density wake. The envelopes do not touch the plate edges but approach its back surface at distances $\pm 0.45$ cm from its center. Upstream dissipation-gravity waves are formed at the strip edges. The flow velocity maximum at the downstream wake axis (marker 4) exceeds the strip velocity, which is due to the participation of the buoyancy forces in the wake flow formation.

A small increase in the strip velocity qualitatively changes the general flow pattern: group of attached internal waves is formed behind the body, while the system of oblique extended high-gradient interfaces appears above and below the narrow density wake in Fig. 12b. The blocked fluid region shrinks and its apex comes closer to the body. In the leading disturbances, along with the dissipation-gravity waves, there are observable upstream transient waves, or extended oblique rays which smoothly

go into attached internal waves past the strip. The central high-gradient interface is clearly expressed within the wake.

Behind the strip, in the region of maximum velocity shear outside the density wake covered by own high gradient envelops, there are set of fine-structure interfaces, 0.5–2.5 cm in length and 0.05–0.12 cm in thickness, whose inclinations to the horizontal line vary from $\pm 55$ to $\pm 16°$ with the wake age. The velocity profiles remain smooth, despite a considerable complication of the density gradient field.

With increase in the velocity, these are internal waves filling the entire field of observation in Fig. 12c, become the main structure element of the stratified flow pattern. In the slot-thread schlieren method, the images of the internal wave crests and troughs are different: black curves correspond to the crests and double gray curves to the troughs. The deviation of the phase surface shapes from the circle, which is typical attached, compact-source-generated waves behind a body, is due to the Doppler-effect-induced variation in the wave frequency. The distance between the two last depressions is 4.7 cm, which corresponds to the theoretical value of the attached internal wavelength defined by the body velocity $U$ and the buoyancy period $\lambda = UT_b = 4.6$ cm. Attached waves are transformed into transient waves ahead of the body; their crests and troughs (oblique bands in Fig. 12c) are gradually widen.

The blocked fluid region ahead of the plate shrinks to the small triangle, 2.25 cm in height, whose base, 1.3 cm in length, does not reach the plate edges. The blocked fluid region is adjoined by almost horizontal phase surfaces of the leading dissipation-gravity waves, which (Fig. 12b) are formed on the density wake envelopes, near the points of their separation—at 0.6 cm from the plate edges.

The back surface of the strip is adjoined with a wedge-shaped bottom vortex, filled with system of interfaces inclined to the axis of the motion. The density wake is split into separate high-gradient interfaces, which are placed horizontally in the compression (thin) regions of the wake, where depression is at the top and crest is at the bottom, whereas in the shape of the "vortex rotor" in the wake expansion regions the crest is at the top of the wake and the depression of the attached internal wave is at its bottom.

Typical thickness of the interfaces is 0.08 cm, the distance between the center of the first separate vortex bubble and the strip is 5.31 cm, and the distance between the first and second bubbles is 6.19 cm. The variation of the distances between the bubbles and their difference from the attached internal wavelength $\lambda = 4.6$ cm is due to the inhomogeneity of the horizontal component of the wake flow velocity and change of the local value of buoyancy frequency inside the wake.

Apart from the attached internal waves, whose scale $\lambda = UT_b$ is determined by the body velocity $U$ and the buoyancy period, the flow pattern in Fig. 12c contains fine-structured disturbances, whose mean thickness (0.1 cm) are close to value of the universal viscous scale $\delta_N^\nu = \sqrt{\nu/N} = 0.11$ cm. Under these conditions the Prandtl scale $\delta_U^\nu = \nu/U = 0.03$ cm.

With the formation of high-gradient interfaces is associated increase in the backscattering coefficient of sound waves takes place, which provides high acoustic contrast of the downstream wake whose efficient height is significantly larger than

the size of the density wake bounded by own high-gradient interfaces. Thickness of the interfaces tends closer to the universal viscous scale $\delta_N^{\nu}$ with increasing distance from the strip [35].

## 11 Conclusion

Observations of the environment in a wide range of length scales and data of laboratory experiments have shown that all kinds of flows are characterized by pronounced structure including distinguished large scale and rich family of small scale components that are a set of high gradient interfaces near boundaries and inside the fluid body. Large scale flow components are responsible for transport of substances, momentum and energy. Fine flow components affect flow energy dissipation, vorticity generation and vortex pattern, transport, separation and structurization of substances.

The analysis of Axioms of Mechanics showed that all four definitions of "a solid body motion"—two physical (through variations of distances between bodies, momentum or energy) and two mathematical (kinematic and geometric) are identical and not equivalent to the concept of "fluid flow" which is the transport of momentum of continuous medium. Concept of "fluid flow" is more complex than the "motion" due to the ability of a "fluid particle" to change its shape (deformation) and split into separate fibers.

The mathematical basis of the flow description is complete set of fundamental governing equations: continuity for density, momentum, energy and concentrations of dissolved or suspended components together with the equation of state for density. The fundamental set defines the fluid flow as the momentum flux accompanied with self-consistent changes of thermodynamic parameters and as a process which occurs in a functional space of a high dimension.

The fundamental set is well posed, self-consistent and forms a basis for adequate modeling, forecasting and controlling of the flow.

The condition of compatibility for set of governing equations determines the *rank* of the complete system, the *order* of its linearized form and the *degree* of the characteristics (dispersion) equation that prescribes the number of independent structural component of flows and their linear length scales. Complete solutions of the fundamental system that describe dynamics and structure of flows are consistent with data from laboratory experiments and allow direct transfer laboratory data to natural conditions.

The list of observable parameters defined by the fundamental set of equations includes basic physical quantities (density, momentum, energy and concentration of dissolved or suspended components), thermodynamic parameters, kinetic coefficients and characteristic of physical fields and wave propagation. Velocity of the fluid due to the uncertainty boundaries of the "deforming and splitting fluid particles" is unobservable parameter.

New instruments and protocols must be constructed for simultaneous observation of large-scale flow structure and identification of fine components together defining scenarios of environmental systems evolution and allowing measuring their physical parameters.

Requirement of the experiment completeness demands the simultaneous recording of all fundamental physical variables fields, visualization of large scale and resolution of fine flow components with the ability of direct estimation of the measurement accuracy.

**Acknowledgments** The work was partly financially supported by the Russian Academy of Sciences (Program OE13 "Vortices and waves in complex fluids") and the Russian Foundation for Basic Research (grant 12-01-00128). Experiments were performed at setup USF "HPhC IPMech RAS" and supported by Ministry of Education and Science of Russian Federation.

## References

1. L. Euler, Principes généraux du mouvement des fluides. Mémoires de lacadémie des sciences de Berlin **11**, 274–315 (1757)
2. B. Franklin, Behavior of oil on water: Letter to John Pringle, Philadelphia, Dec 1, 1762, in *The ingenious Dr. Franklin: Selected scientific letters of Benjamin Franklin*, ed. by N.G. Goodman (University of Pennsylvania Press, Philadelphia, 1931), pp. 142–145
3. C.L.M.H. Navier, Mémoire sur les lois du mouvement des fluides. Mem. Acad. Sci. Inst. Fr. **6**, 389–440 (1823)
4. G.G. Stokes, On the theories of the internal friction of fluids in motion, and of the equilibrium and motion of elastic bodies. Trans. Camb. Philos. Soc. **8**, 287–305 (1845)
5. D.I. Mendeleeff, On fluid drag and aeronautics (1880) (In Russian)
6. D.I. Mendeleeff, On elasticity of gases. SPb (1875) (In Russian)
7. D.I. Mendeleeff, Studying of water solutions on specific gravity (1887) (In Russian)
8. O.V. Manturov, Yu.K. Solntsev, Yu.I. Sorkin, N.G. Fedin, *Encyclopedic Dictionary of Mathematical Terms* (Prosveshchenie, Moscow, 1965)
9. H. Helmholtz, Über Integrale der hydrodynamischen Gleichungen, welche den Wirbelbewegungen entsprechen. J. für die reine und angew Math. **55**, 25–55 (1858)
10. Yu.D. Chashechkin, V.A. Kalinichenko, Topographic patterns in the suspension structure in standing waves. Dokl. Phys. **57**(9), 363–366 (2012)
11. Yu.D. Chashechkin, E.V. Stepanova, Formation of a single spiral arm from a central marking-admixture spot on a compound-vortex surface. Dokl. Phys. **55**(1), 43–46 (2010)
12. A.A. Budnikov, Yu.D. Chashechkin, Marker transfer in a settled composite vortex. Moscow Univ. Phys. Bull. **69**(3), 270–274 (2014)
13. E.V. Stepanova, Yu.D. Chashechkin, Marker transport in a composite vortex. Fluid Dyn. **45**(6), 843–858 (2010)
14. R. Descartes, *Principia philosophiae* (Apud Ludovicum Elzevirium Publisher, Amstelodami, 1644)
15. J. Serrin, Mathematical principles of classical fluid mechanics. In *Handbuch der Physik*, vol. VIII/1, Chap. 2, pp. 125–263 (1959)
16. G.E. Mase, *Theory and Problems of Continuum Mechanics* (McGraw Inc., New York, 1970)
17. Yu.D. Chashechkin, V.E. Prokhorov, Drop-impact hydrodynamics: short waves on a surface of the crown. Dokl. Phys. **58**(7), 296–300 (2013)
18. V.E. Prokhorov, Yu.D. Chashechkin, Sound generation as a drop falls on a water surface. Acoust. Phys. **57**(6), 807–818 (2011)

19. L.D. Landau, E.M. Lifshitz, *Fluid Mechanics*, Course of Theoretical Physics, vol. 6, 2nd edn. (Pergamon Press, Oxford, 1987)
20. V.G. Baidulov, Yu.D. Chashechkin, Invariant properties of systems of equations of the mechanics of inhomogeneous fluids. J. Appl. Maths Mech. **75**(4), 390–397 (2011)
21. R.N. Bardakov, Yu.A. Vasilev, Yu.D. Chashechkin, Calculation and measurement of conical beams of three-dimensional periodic internal waves excited by a vertically oscillating piston. Fluid Dyn. **42**(4), 612–626 (2007)
22. Yu.D. Chashechkin, Hierarchy of the models of classical mechanics of inhomogeneous fluids. Phys. Oceanogr. **20**(5), 317–324 (2011)
23. A.V. Kistovich, Yu.D. Chashechkin, Fine structure of a conical beam of periodical internal waves in a stratified fluid. Atm. Ocean. Phys. **50**(1), 103–110 (2014)
24. Yu.D. Chashechkin, Schlieren visualization of a stratified flow around a cylinder. J. Vis. **1**(4), 345–354 (1999)
25. Yu.D. Chashechkin, Ya.V. Zagumennyi, Non-equilibrium processes in non- homogeneous fluids under the action of external forces. Phys. Scripta **T155**, 014010 (2013)
26. L. Prandtl, *Essentials of Fluid Dynamics*, 2nd edn. (Blakie and Son, London, 1952)
27. A.V. Kistovich, Yu.D. Chashechkin, Dissipative-gravity waves in subcritical regimes of multicomponent convection. Izv. Atmos. Ocean. Phys. **37**(4), 476481 (2001)
28. M.R. Allshouse, M.F. Barad, T. Peacock, Propulsion generated by diffusion-driven flow. Nat. Phys. **6**, 516–519 (2010)
29. J. Oerlemans, B. Grisogono, Glacier winds and parameterization of the related heat fluxes. Tellus **54a**, 440–452 (2002)
30. J. Lighthill, *Waves in Fluids* (CUP, Cambridge, 1978)
31. Yu.V. Kistovich, Yu.D. Chashechkin, Linear theory of beams internal wave propagation an arbitrarily stratified liquid. J. Appl. Mech. Tech. Phys. **39**(5), 729–737 (1998)
32. M.S. Paoletti, H.L. Swinney, Propagation and evanescent internal wave in a deep ocean model. J. Fluid Mech. **706**, 571–583 (2012)
33. Yu.V. Kistovich, Yu.D. Chashechkin, Generation of monochromatic internal waves in a viscous fluid. J. Appl. Mech. Tech. Phys. **40**(6), 1020–1028 (1999)
34. Yu.D. Chashechkin, Visualization of singular components of periodic motions in a continuously stratified fluid. J. Vis. **10**(1), 17–20 (2007)
35. V.E. Prokhorov, Yu.D. Chashechkin, Visualization and acoustic sounding of the fine structure of a stratified flow behind a vertical plate. Fluid Dyn. **48**(6), 722–733 (2013)

# Effect of Friction in Sliding Contact of a Sphere Over a Viscoelastic Half-Space

**Irina Goryacheva, Fedor Stepanov and Elena Torskaya**

**Abstract** Imperfect elasticity of contacting solids results in hysteretic losses during the deformation. In rolling/sliding contact, the losses cause the resistant force, which is called the mechanical component of the friction force. Another cause of the friction is related to the energy losses in formation and breaking of the adhesive bridges between the contacting bodies (adhesive component of friction). In this study the combined effect of the adhesive and mechanical components of friction is analysed based on the consideration of the 3-D contact problem for the spherical indenter sliding with a constant velocity at the boundary of the viscoelastic half-space. The material properties are characterized by the linear viscoelastic solid with one relaxation time. The Coulomb-Amonton law of friction is used to describe the adhesive friction inside the contact region. A numerical-analytical method is developed to solve the contact problem and to find the contact stress distribution. The dependence of the mechanical component of friction force on the adhesive friction coefficient for various load-velocity conditions is studied.

**Keywords** Viscoelastic half-space · Sliding contact · Friction · Contact stress · Boundary elements

**Mathematical Subject Classification:** 74D05

I. Goryacheva (✉) · F. Stepanov · E. Torskaya
IPMech RAS, Prosp. Vernadskogo 101, Block 1, Moscow 119526, Russia
e-mail: goryache@ipmnet.ru

F. Stepanov
e-mail: stepanov_ipm@mail.ru

E. Torskaya
e-mail: torskaya@mail.ru



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_6

## 1 Introduction

There are two main sources of energy dissipation in sliding contact of deformable bodies. One is related to the formation and breaking of the adhesive bridges at the real contact spots within the nominal contact region (adhesive component of friction). The other one arises due to hysteretic losses in deformation cycles of imperfect elastic materials (mechanical component of friction). Due to these sources the friction force occurs if one body moves over the other.

To study the effect of the imperfect elasticity, the 2-D sliding contacts of the rigid or viscoelastic cylinder and the viscoelastic half space were considered in [3–8, 11–13]. It was indicated in these studies that even for the case of the absence of the shear stress within the contact region (adhesion effect is neglected) the contact pressure is distributed non symmetrically and the shift of the contact region occurs. Hence the resistance force (mechanical component of friction force) arises due to imperfect elasticity of contacting bodies. The model of a linear viscoelastic solid is usually used to analyze effects of the imperfect elasticity in sliding contacts of deformable bodies.

In [5] the authors studied the effect of the shear stress $\tau$ within the contact due to the adhesive friction in sliding contact of the viscoelastic cylinder over the viscoelastic half space with a constant velocity $V$ (2-D problem formulation). The adhesive friction within the contact region was taken into account by the relation similar to the Amonton law of friction, i.e.,

$$\tau = \mu p \, \mathrm{sign}(V). \tag{1}$$

Here, $p$ is the contact pressure, $\mu$ is the adhesive friction coefficient.

The effect of the adhesive interaction of the contacting surfaces outside of the contact zone in sliding of the rigid cylinder over the viscoelastic half space was studied in [10] in the case where the shear stress is absent within the contact region. 3-D contact problem for a rigid sphere sliding over the viscoelastic half-space was considered in [2] under the assumption that there is no shear stress within the contact region. The influence of the imperfect elasticity of the base on the shape of the contact region and the contact pressure distribution was analyzed for the various load-velocity conditions.

In this study, the effect of the shear contact stress on the contact characteristics and the mechanical component of the friction force is analyzed based on consideration of the 3-D contact problem for the smooth rigid indenter sliding with a constant velocity over the boundary of the viscoelastic half-space. Contact shear and normal stresses are related by the Amonton law (1).

## 2 Problem Formulation

Consider the case where the rigid indenter with smooth surface moves with a constant velocity $V$ over the viscoelastic half-space in the direction of the axis $0x$ (Fig. 1). The indenter shape is described by the function $f(x, y)$. The indenter is loaded by a vertical force $Q$. It is assumed that the shear stresses within the contact region (which arises due to friction satisfy the Amonton law (1)) are opposite to the direction of sliding. This situation corresponds to the case of anisotropic friction [15]. To provide the equilibrium conditions, tangential force $T$ directed at $Ox$-axis is also applied to indenter, i.e., $T = \mu Q$. The half-space is described by the coordinates $|x| < \infty$, $|y| < \infty$, and $z \leq 0$. In the moving Cartesian system of coordinates $(x, y, z)$ related to the indenter, the following boundary and equilibrium conditions ($z = 0$) are considered:

$$\begin{gathered} w(x, y) = f(x, y) + D, \quad (x, y) \in \Omega, \\ \tau_{xz} = \mu\sigma_z, \quad (x, y) \in \Omega, \\ \sigma_z = 0, \tau_{xz} = 0, \quad (x, y) \notin \Omega, \\ \tau_{yx} = 0, \quad -\infty < x < +\infty, -\infty < y < +\infty, \\ Q = \int\int_{\Omega} p(x, y)\, dx\, dy, \\ p(x, y) = -\sigma_z(x, y). \end{gathered} \tag{2}$$

Here, $\Omega$ is the unknown contact region, $w(x, y)$ is the normal displacement of the half-space surface, $D$ is the indenter penetration, and $\sigma_z$, $\tau_{xz}$, and $\tau_{yz}$ are the normal and shear stresses, respectively.

Under the assumption that the shear modulus $G(t)$ is a time-dependent operator and the Poisson ratio is constant, the mechanical behavior of the viscoelastic half-space is described by the following constitutive equation [14]:

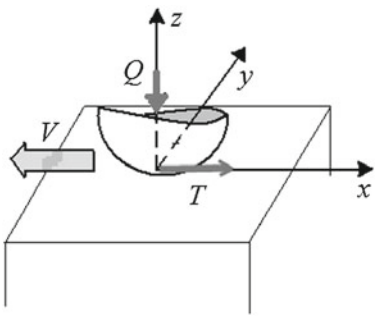



**Fig. 1** Scheme of contact

$$\gamma(t) = \frac{1}{G}\sigma(t) + \frac{1}{G}\int_{-\infty}^{t} \sigma(\tau)K(t-\tau)\,d\tau, \tag{3}$$

$$K(t') = k\exp\left(-\frac{t'}{\lambda}\right).$$

Here, $\gamma(t)$ and $\sigma(t)$ are the shear deformation and shear stress components, respectively, $G$ is the instantaneous shear modulus, $\lambda$ is the retardation time and $1/k$ is the relaxation time.

## 3 Solution Method

We first consider sliding of the concentrated force with a constant velocity $V$ along the elastic half-space boundary at the $x$-axis direction. The force has a normal component $P$ and a tangential component $T$ directed to the $x$-axis. The moving Cartesian system of coordinates $(x, y, z)$ is related to the force. In this system, the point where the force is applied has the coordinates $(\xi, \eta, 0)$. In the case of neglecting inertial terms, the normal displacements at the surface $(z = 0)$ are determined by the relation

$$w(x, y) = \frac{(1-2\upsilon)T(\xi - x)}{4\pi G r^2} - \frac{(1-\upsilon)P}{2\pi G r}, \tag{4}$$

where $G$ and $\upsilon$ are the shear modulus and the Poisson ratio of the elastic material, $r = \sqrt{(\xi - x)^2 + (\eta - y)^2}$ (see [14]).

Relation (4) can be used for determining normal displacements of the viscoelastic half-space boundary under similar load conditions. With the assumption that the shear modulus is a time-dependent operator (3) and the Poisson ratio is constant, we obtain the following relations for the vertical displacements of the viscoelastic surface provided that the concentrated load moves at the $x$-axis direction with the constant velocity $V$:

$$w(x, y) = \frac{(1-2\upsilon)T}{4\pi G}\left[\frac{\xi - x}{r^2} + \int_{-\infty}^{0} K(-\tau)\frac{\xi - x - \tau}{r_1^2}d\tau\right] - \frac{(1-\upsilon)P}{2\pi G}\left[\frac{1}{r} + \int_{-\infty}^{0} K(-\tau)\frac{1}{r_1}d\tau\right]. \tag{5}$$

Here $K(t)$ is the creep kernel and $r_1 = \sqrt{(\xi - x - V\tau)^2 + (\eta - y)^2}$ (see [1, 14]). In the case of the normal stress $p(x, y)$ and the tangential stress $\tau_{xz}(\xi, \eta)$ distributed over the half-space ($z = 0$) within the loading region $\Omega$, we can deduce from (5) the following relationship for the vertical displacement of the viscoelastic half-space surface ($z = 0$) (see [1]):

$$w(x, y) = -\frac{1-\upsilon}{2\pi G}\int\int_{\Omega} p(x, \eta)\left[\frac{1}{r} + \int_{-\infty}^{0} K(-\tau)\frac{1}{r_1}d\tau\right]d\xi\, d\eta$$
$$+ \frac{1-2\upsilon}{4\pi G}\int\int_{\Omega} \tau_1(x, \eta)\left[\frac{\xi - x}{r^2} + \int_{-\infty}^{0} K(-\tau)\frac{\xi - x - V\tau}{r_1^2}d\tau\right]d\xi\, d\eta.$$

Taking into account that the creep kernel $K(t')$ is the exponential function (3) and introducing the notation

$$u = \frac{\xi - x - V\tau}{\lambda V},$$

we deduce the following integral relationship between the normal displacement $w(x, y)$ of the surface of the viscoelastic half-space (considering in the moving system of coordinates) loaded by the distributed normal and tangential stress within the region $\Omega$ (moving with the constant velocity $V$):

$$w(x, y) = -\frac{1-\upsilon}{2\pi G}\int\int_{\Omega} p(x, \eta)\left[\frac{1}{r} + \frac{1}{V}kI_1\left(\frac{\xi - x}{\lambda V}, \frac{\eta - y}{\lambda V}\right)\right]d\xi\, d\eta$$
$$+ \frac{1-2\upsilon}{4\pi G}\int\int_{\Omega} \tau_{xz}(x, \eta)\left[\frac{\xi - x}{r^2} + \frac{1}{V}kI_2\left(\frac{\xi - x}{\lambda V}, \frac{\eta - y}{\lambda V}\right)\right]d\xi\, d\eta, \tag{6}$$

$$I_1(\alpha, \beta) = e^{\alpha}\int_{\alpha}^{\infty}\frac{e^{-u}u\,du}{u^2 + \beta^2},$$

$$I_2(\alpha, \beta) = e^{\alpha}\beta\int_{\alpha}^{\infty}\frac{e^{-u}du}{u^2 + \beta^2}.$$

The Eq. (6) with the boundary conditions (2) is used to solve the contact problem.

A boundary element method is used to find the contact pressure distribution. We use a mesh of square elements which covers the unknown contact region. Amounts of elements along the axes $0x$ and $0y$ are $N_1$ and $N_2$, respectively. The pressure is assumed to be constant in each element. A normal displacement of the surface in the center of an arbitrary element is obtained by summation of displacements

caused by the pressure applied to each element (influence coefficients). The influence coefficients are determined by the following relationship:

$$k_i^j = \frac{2}{\pi^2 c} \int_{-\Delta/2}^{\Delta/2} \int_{-\Delta/2}^{\Delta/2} \left[ \frac{1}{\sqrt{(\xi' - x_{ij})^2 - (\eta' - y_{ij})^2}} + \sum_{i=1}^{n} B e^{A(\xi' - x_{ij})} \int_{A(\xi' - x_{ij})}^{\infty} \frac{e^{-u} du}{\sqrt{u^2 + A^2(\eta' - y_{ij})}} \right] d\xi' \, d\eta'.$$

Here, $c = G/G_l$, $G_l$ is the relation of instant and longitudinal shear modulus, $A = (QR/G_l)^{1/3}/(\lambda V)$ and $B = k(QR/G_l)^{1/3}/V$ are dimensionless parameters, which include the material characteristics and sliding contact parameters, $(x', y', \xi', \eta') = (x, y, \xi, \eta)/(QR/G_l)^{1/3}$ are dimensionless coordinates, $\Delta$ is the element size and $(x_{ij}^2 + y_{ij}^2)^{1/2}$ is distance between the centers of the elements. The matrix equation which is needed to calculate the contact pressure, follows from the relations (2):

$$\begin{pmatrix} \Delta^2 & \dots & \Delta^2 & 0 \\ k_1^1 & \dots & k_N^1 & -1 \\ \vdots & \ddots & \vdots & \vdots \\ k_1^N & \dots & k_N^N & -1 \end{pmatrix} \times \begin{pmatrix} p_1 \\ \vdots \\ p_N \\ D \end{pmatrix} = \begin{pmatrix} Q \\ f_1 \\ \vdots \\ f_N \end{pmatrix}. \tag{7}$$

Here, $p_1 \dots p_N$ are unknown constant pressures inside each square element, and $f_1 \dots f_N$ are the indenter shape in the centers of the elements. The matrix in (7) has the dimension of $N = (N_1 \times N_2)^2$ and it is solved by an iteration method.

The mechanical component of friction force $T^*$, appeared due to hysteretic losses, is calculated by the relation

$$T^* = \frac{1}{R} \int \int_{\Omega} x p(x, y) \, dx \, dy.$$

The mechanical component of the friction coefficient is determined by the relation $\mu^* = T^*/Q$.

## 4 Results and Discussion

The method described above is used to find the contact characteristics, i.e., pressure and shear stress distributions and the contact region as well as the friction force for the sliding contact of a spherical indenter over the viscoelastic half-space. The shape of the indenter is described by the following function:

$$f(x, y) = \frac{x^2 + y^2}{2R}. \tag{8}$$

Analysis of the Eq. (6) with the boundary conditions (2) allows us to conclude that the dimensionless contact pressure $p' = p(x, y)/G_l R^2$ and other contact characteristics depend on the dimensionless parameters $A$ and $B$, which include the material characteristics and sliding contact parameters, relation of instant and longitudinal shear modulus $c$, dimensionless velocity $V' = V\lambda/R$, relative load $Q' = Q/G_l R^2$ and friction coefficient $\mu$.

Figure 2 represents contact pressure distributions for elastic (a) and viscoelastic (b) half-spaces under the same loading-velocity conditions and the friction coefficient $\mu = 0.5$. For the mechanical parameters used in calculations, the distribution of the contact pressure is non-symmetrical, and it is effected by the sliding velocity and the friction coefficient. The pressure maximum shifts in the direction of motion and the contact region is not circular. The cross sections of the pressure distributions are compared in Fig. 2c. The contact zone is smaller and the maximum contact pressure is larger in the case of viscoelastic material (curve 2 in Fig. 2c) compared to elastic one (curve 1 in Fig. 2c). The shapes of pressure distributions are essentially different.

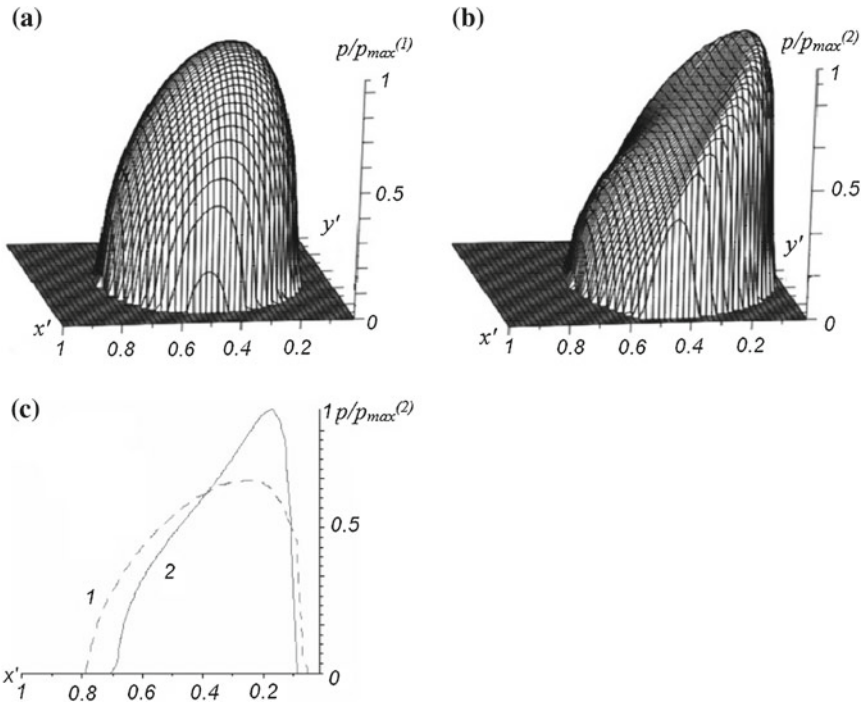



**Fig. 2** Contact pressure distribution for the elastic solution (**a**), viscoelastic solution (**b**), and comparison of the two (**c**): $V' = 0.75$, $\nu = 0.47$, $Q' = 20$, $\mu = 0.5$, $A = 2.31$, $B = 11.58$, $c = 1$ (*curve 1*), $c = 5$ (*curve 2*)

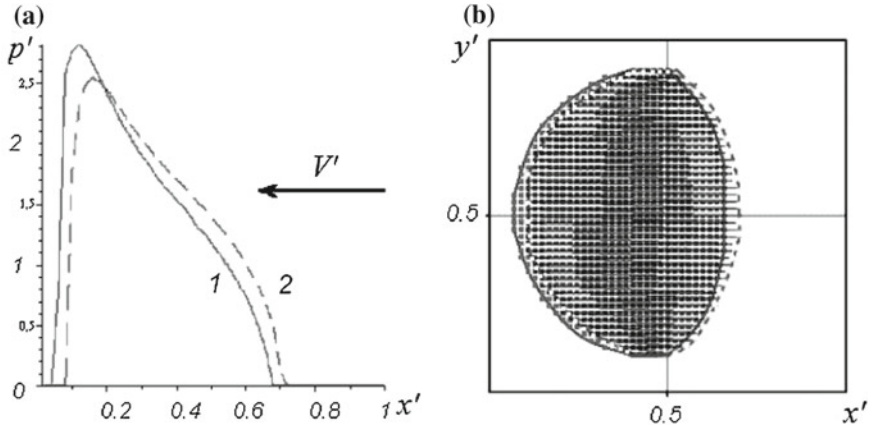



**Fig. 3** The effect of tangential force on pressure distribution (**a**) and contact spot (**b**) $V' = 0.75$, $\nu = 0.3$, $Q' = 20$, $A = 2.31$, $B = 23.17$, $c = 5$; $\mu = 0.5$ (*curve 1* in (**a**), *solid line* in (**b**)), $\mu = 0$ (*curve 2* in (**a**), *dashed line* in (**b**))

The effect of shear stresses within the contact region on the shape of contact spot and a pressure distribution can be analyzed using the results presented in Fig. 3. Figure 3a illustrates the cross-section $y = 0$ of the pressure distributions in the presence (curve 1) or in the absence (curve 2) of the shear stress within the contact region for the same normal load and a sliding velocity. Existence of the shear stress leads to the increase of a maximum pressure and the shift of the contact region to the direction of sliding. The contact zone configurations for both cases are presented in Fig. 3b.

The effect of a sliding velocity on contact characteristics has also been analyzed (see Fig. 4). The contact region configurations for the same viscoelastic material, the same normal and tangential load and for three sliding velocities are presented in Fig. 4a–c. The results indicate that increase of the velocity leads to decrease of the size of the contact zone ("flowing-up" effect) and to increase of the contact shift. One can see that the back side of the a contact spot being closer to the indenters center with the increase of velocity. The shape of a contact spot is also different for different velocities and this occurs because of material viscosity. Figure 4d illustrates the pressure distributions in the cross section $y = 0$ for the cases considered in Fig. 4a–c. The maximum value of the pressure within the contact region increases as the velocity increases.

The nonsymmetrical distribution of the contact pressure and the contact shift affect the mechanical component of the friction coefficient $\mu^*$. Figure 5 illustrates the dependence of the mechanical component of the friction coefficient on the dimensionless velocity for the case of no shear stress within the contact region (curve 1) and for the case of the existence of the adhesive friction (curve 2). The results indicate that

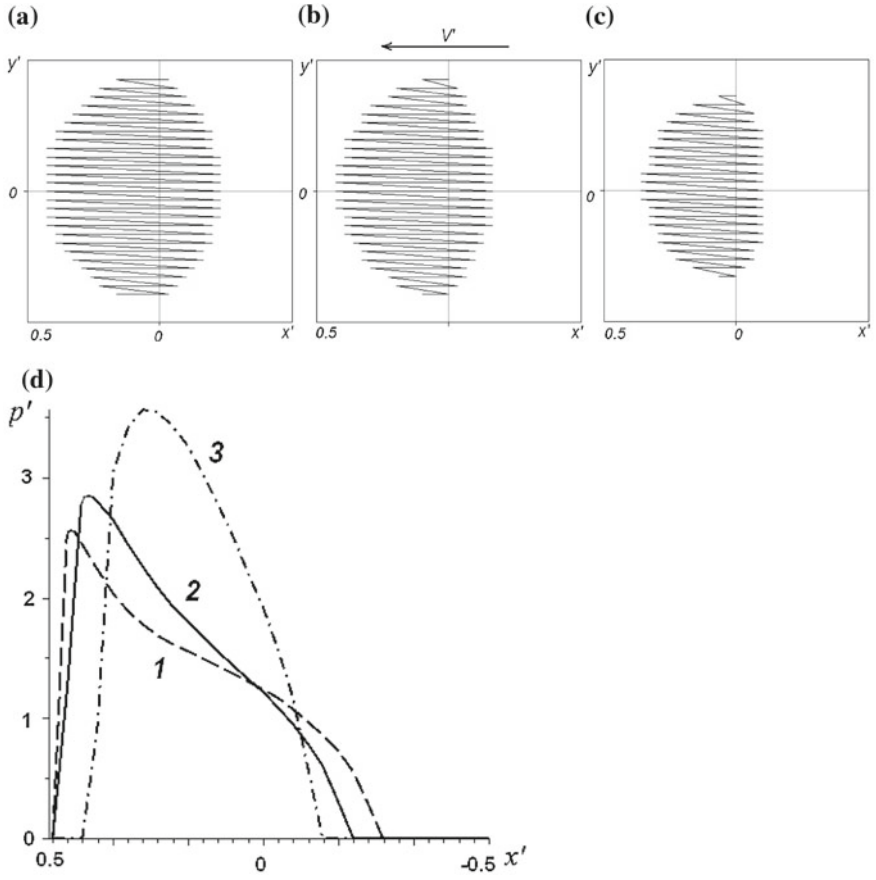



**Fig. 4** The effect of velocity on contact spot and pressure distribution $\nu = 0.47$, $Q' = 20$, $\mu = 0.5$; **a** $V' = 0.5$, *curve 1* in (**d**), **b** $V' = 0.75$, *curve 2* in (**d**), **c** $V' = 2.0$, *curve 3* in (**d**)

the mechanical component of the friction coefficient for both cases is non monotonical function of velocity. The resistant force and the mechanical component of the friction coefficient grows with the increase of velocity until they reach some maximum values and then decrease. The results of calculations also indicate that the adhesive friction within the contact region influences on the mechanical component of the friction force which arises due to hysteretic losses in the contacting bodies. The presence of the shear stress within the contact region leads to the increase of the mechanical component of the friction coefficient $\mu^*$.

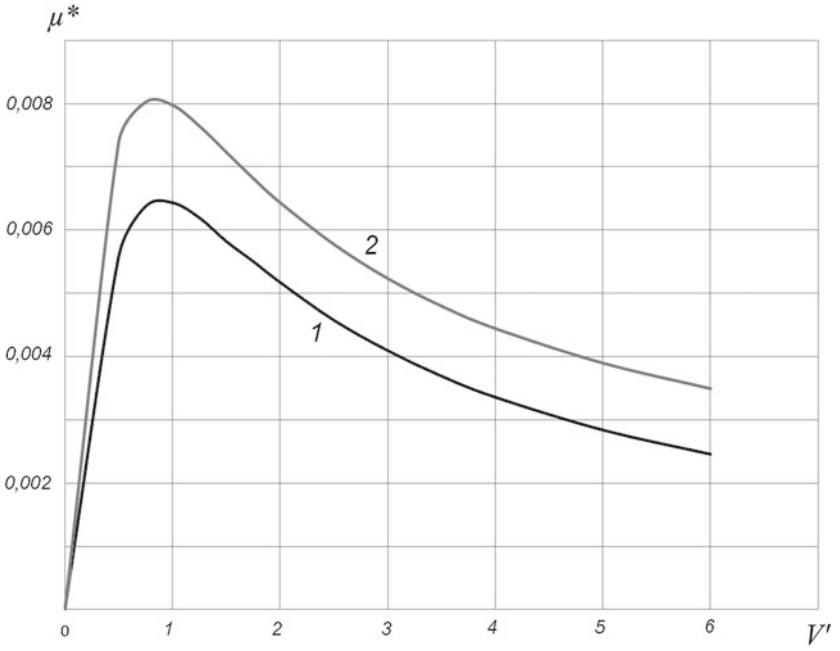



**Fig. 5** The effect of velocity on hysteretic losses $\mu = 0$ (*curve 1*) and $\mu = 0.7$ (*curve 2*), $\nu = 0.3$, $Q' = 20$, $c = 5$

## 5 Conclusions

The numerical-analytical method is developed to solve the friction contact problem for a spherical slider and a viscoelastic half-space. The method is used to analyze the contact pressure distribution for various load-velocity conditions and to study the effect of the shear contact stress due to adhesive friction on the contact pressure distribution and on the mechanical component of the friction force arising due to hysteretic losses inside the contacting bodies.

The results of calculations indicate that the shear contact stresses increase the maximum contact pressure value and the shift of the contact region to the front of the indenter. The mechanical component of friction force arisen due to hysteretic losses inside the contacting bodies also increases with increase of the adhesive friction coefficient. It means, that the mechanical component of friction depends on the adhesive friction within the contact region. Commonly used superposition of friction forces of different nature is not correct for the case of boundary (adhesive) and hysteretic friction.

**Acknowledgments** This work is financially supported by Russian Scientific Foundation (grant 14-29-00198).

## References


1. V.M. Aleksandrov, I.G. Goryacheva, Mixed problems of mechanics of deformable solid, in *Proceedings of V Russian Conference with International Participation*, Izd. Sarat. University in Russian, Saratov, 2005, pp. 23–25
2. V.M. Aleksandrov, I.G. Goryacheva, E.V. Torskaya, Sliding contact of a smooth indenter and a viscoelastic half-space (3D problem). Dokl. Phys. **55**(2), 77–80 (2010)
3. L.A. Galin, *Contact Problems of the Theory of Elasticity and Viscoelasticity* (Nauka, Moscow, 1980). (in Russian)
4. I.G. Goryacheva, Contact problem of rolling of a viscoelastic cylinder on a base of the same material. J. Appl. Math. Mech. **37**(5), 877–885 (1973)
5. I.G. Goryacheva, The limiting case of rolling of a cylinder on a viscoelastic base, in *Proceedings of the Postgraduate Students of Moscow State University*, Moscow, MSU (1973). (in Russian)
6. I.G. Goryacheva, *Contact Mechanics in Tribology* (Kluwer Academic Publishers, Dordrecht, 1998)
7. S.C. Hunter, The rolling contact of a rigid cylinder with a viscoelastic half space. J. Appl. Mech. **28**(4), 611–617 (1961)
8. R. Ya, N. Ivanova, The rolling viscoelastic cylinder on the base of the same material. J. Appl. Mech. Tech. Phys. **3**, 179–184 (1964)
9. A.I. Lur'e, *Spatial Problems of the Theory of Elasticity* (Gostekhizdat, Moscow,1955). (in Russian)
10. Y.Y. Makhovskaya, The sliding of viscoelastic bodies when there is adhesion. J. Appl. Math. Mech. **69**(2), 305–314 (2005)
11. L.W. Morland, A plane problem of rolling contact in linear viscoelasticity theory. J. Appl. Mech. **29**(2), 345–352 (1962)
12. L.W. Morland, Rolling contact between dissimilar viscoelastic cylinders. Quart. Appl. Math. **25**(4), 363–376 (1968)
13. Y. Rabotnov, *Creep of Construction Elements* (Nauka, Moscow, 1966). (in Russian)
14. Y. Rabotnov, *Elements of Hereditary Mechanics of Solid Bodies* (Nauka, Moscow, 1977). (in Russian)
15. V.D. Vantorin, Motion along a plane with anisotropic friction, pp. 81–120 (1962). (in Russian)

# Stability of a Tensioned Axially Moving Plate Subjected to Cross-Direction Potential Flow

**Juha Jeronen, Tytti Saksa and Tero Tuovinen**

**Abstract** We analyze the stability of an axially moving Kirchhoff plate, subjected to an axial potential flow perpendicular to the direction of motion. The dimensionality of the problem is reduced by considering a cross-directional cross-section of the plate, approximating the axial response with the solution of the corresponding problem of a moving plate in vacuum. The flow component is handled via a Green's function solution. The stability of the cross-section is investigated via the classical Euler type static linear stability analysis method. The resulting eigenvalue problem is solved numerically using Hermite type finite elements. As a result, the critical velocity and the corresponding eigenfunction are determined. It is seen that even at very low free-stream fluid velocities, the buckling shape may become antisymmetric in the cross direction.

**Keywords** Axially moving · Kirchhoff plate · Stability · Eigenvalue problem

**Mathematical Subject Classification:** 74F10 · 74B05 · 76B99 · 65N25

## 1 Introduction

Models of out-of-plane vibrations of axially moving materials are commonly considered in the context of industrial production processes, such as paper making. Typical models include axially moving strings, beams, panels (plates with cylindrical deformation), membranes and plates. Research into the field began at the end of the 19th

J. Jeronen (✉) · T. Saksa · T. Tuovinen
Department of Mathematical Information Technology, University of Jyväskylä,
P.O. Box 35 (Agora), FI-40014 Jyväskylä, Finland
e-mail: Juha.Jeronen@jyu.fi

T. Saksa
e-mail: Tytti.Saksa@jyu.fi

T. Tuovinen
e-mail: Tero.Tuovinen@jyu.fi



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_7

century [15]. Other important classical studies include, e.g., Sack [13], Archibald and Emslie [1], Swope and Ames [16], and Simpson [14]. The field has remained active to this day; stability problems of axially moving materials have been considered, e.g., by Parker [11], Kong and Parker [9], and Wang et al. [17].

Problems of out-of-plane behaviour of axially moving materials share some of their mathematical formulation with those of axially compressed stationary materials and those of gyroscopic systems, leading to questions of stability. The problem parameter of interest is the axial velocity of the material.

In the case of lightweight materials, such as paper, the fluid–structure interaction between the travelling material and the surrounding air must be accounted for, because the inertial contribution of the surrounding air is significant. The surrounding air is known to change both the frequencies of natural vibration and the critical velocity of the travelling material (see, e.g., [7, 10, 12]).

The present study concentrates on the stability analysis of an axially moving Kirchhoff plate in an open draw, subjected to an axial potential flow perpendicular to the direction of motion. The dimensionality of the problem is reduced by considering a cross-directional cross-section of the plate, approximating the axial response with the solution of the corresponding problem of a moving plate in vacuum. The flow component is handled via a Green's function solution [6, 8], leading to a one-dimensional integrodifferential model. The stability of the cross-section is investigated via the classical Euler type static linear stability analysis method. The eigenvalue problem is solved numerically using Hermite type finite elements.

## 2 Problem Setup

Consider a travelling, rectangular, isotropic Kirchhoff plate in the plane region

$$\Omega \equiv \{(x, y) \mid 0 < x < \ell,\ -b < y < b\}, \tag{1}$$

simply supported on the edges $x = 0, \ell$ and free of tractions on the edges $y = \pm b$, see Fig. 1.

The dynamic equation of small vibrations of an isotropic, axially moving Kirchhoff plate, travelling at constant velocity $V_0$ in the $x$ direction, subjected to a constant axial tension $T_0$, applied at the rollers, and an aerodynamic reaction loading $q_\mathrm{f}(w)$, is

$$mw_{,tt} + 2mV_0w_{,xt} + (mV_0^2 - T_0)w_{,xx} + D(w_{,xxxx} + 2w_{,xxyy} + w_{,yyyy}) = q_\mathrm{f}(w). \tag{2}$$

Subscripts after a comma denote partial differentiation. Here $w$ is the transverse displacement, $m$ is the mass per unit area of the middle surface of the plate, and the bending rigidity $D$ is given by

$$D = \frac{Eh^3}{12(1 - \nu^2)}, \tag{3}$$

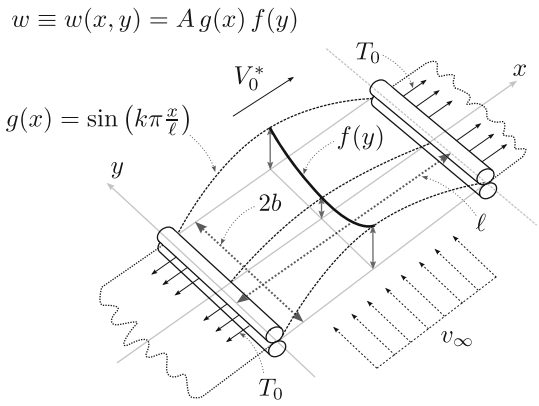



**Fig. 1** Axially travelling plate subjected to cross-direction flow. Steady state

where $E$ is the Young's modulus of the material, $\nu$ its Poisson ratio, and $h$ the thickness of the plate.

The boundary conditions are set as SFSF:

$$w = w_{,xx} = 0 \quad \text{at } x = 0, \ell, \tag{4}$$

$$w_{,yy} + \nu w_{,xx} = 0 \quad \text{at } y = \pm b, \tag{5}$$

$$w_{,yyy} + (2 - \nu) w_{,xxy} = 0 \quad \text{at } y = \pm b. \tag{6}$$

Static stability analysis, applied to Eq. (2), is concerned with determining non-trivial steady-state solutions and the corresponding critical velocities as eigenfunction-eigenvalue pairs $(w, V_0)$. In the steady state, (2) reduces to

$$(m V_0^2 - T_0) w_{,xx} + D(w_{,xxxx} + 2w_{,xxyy} + w_{,yyyy}) = q_{\mathrm{f}}(w). \tag{7}$$

It can be observed [5, 8] that near the middle point of an open draw, the buckling shape is not much altered by the introduction of an aerodynamic load, when compared to the vacuum case. Based on this observation, let us introduce the following approximation.

The steady-state solution in the vacuum case is of the form [4]

$$w(x, y) = A g(x) f(y) = A \sin\left(k\pi \frac{x}{\ell}\right) f(y), \tag{8}$$

where $A$ is an arbitrary constant and $k = 1, 2, 3, \ldots$. We approximate the flow as two-dimensional in the $yz$ plane for each fixed $x$, and approximate these slices as being independent of each other. Observe that (8) fulfills the boundary condition (4) identically.

By differentiating (8) twice with respect to $x$, we obtain

$$w_{,xx} = -\left(\frac{k\pi}{\ell}\right)^2 w \equiv \beta w, \tag{9}$$

where the constant $\beta$ is defined by the obvious identification. Using (8) as a trial function and inserting (9) to (7), and taking into account that $q_\mathrm{f}(w)$, describing the aerodynamic reaction of a potential flow, is linear in $w$, we obtain

$$(mV_0^2 - T_0)\beta f + D(\beta^2 f + 2\beta f'' + f^{(4)}) = q_\mathrm{f}(f), \tag{10}$$

which is an approximate equation for the steady-state solution near the midpoint of a long open draw. The $x$ dependence has been eliminated; $f = f(y)$.

## 3 Solution Approach

Let us define

$$\alpha = (mV_0^2 - T_0)\beta + D\beta^2, \tag{11}$$

and transform (10) into dimensionless coordinates in the usual manner:

$$\hat{y} \equiv y/b, \quad \partial(\cdot)/\partial y \to (1/b)\partial(\cdot)/\partial\hat{y}, \quad \hat{f}(\hat{y}) \equiv f(b\hat{y}). \tag{12}$$

Applying the transformation and collecting terms, we have

$$\alpha\hat{f} + D\left(\frac{2\beta}{b^2}\hat{f}'' + \frac{1}{b^4}\hat{f}^{(4)}\right) = q_\mathrm{f}(\hat{f}), \quad -1 < \hat{y} < 1. \tag{13}$$

Equation (13) can be considered as an eigenvalue problem for the pair $(\alpha, \hat{f})$. The eigenvalue $\alpha$ gives the critical velocity $V_0^*$ via Eq. (11), while the eigenfunction $\hat{f}$ describes the slice of the buckling form in the $yz$ plane (refer to Fig. 1).

If $\alpha > 0$, we observe that the axial tension is seen by the cross-directional cross-section as a linear elastic foundation with stiffness $\alpha$. Considering that $\beta < 0$, we find that $\alpha > 0$ at least until $V_0 = \sqrt{T/m}$, and slightly further if $D > 0$.

Now the aerodynamic reaction $q_\mathrm{f}(f)$ can be written explicitly in terms of $f(y)$ via a Green's function solution for the Neumann problem of the Laplace equation for a plane with a slit [5, 6, 8]. In the general dynamic case, we have

$$q_\mathrm{f}(\hat{f}, t) = -\frac{b\rho_\mathrm{f}}{\tau^2}\left(\frac{\partial}{\partial t} + \frac{\tau}{b}v_\infty\frac{\partial}{\partial\hat{y}}\right)\int_{-1}^{1} N(\eta, \hat{y})\left(\frac{\partial}{\partial t} + \frac{\tau}{b}v_\infty\frac{\partial}{\partial\eta}\right)\hat{f}(\eta, t)\,\mathrm{d}\eta. \tag{14}$$

The parameter $\tau$ is an arbitrary characteristic time used for nondimensionalizing the time coordinate, $v_\infty$ is the free-stream velocity of the potential flow, and $N(\eta, y)$ is the aerodynamic kernel (Green's function):

$$N(\eta,\hat{y}) \equiv \frac{1}{\pi}\ln\left|\frac{1+\Lambda(\eta,\hat{y})}{1-\Lambda(\eta,\hat{y})}\right|, \quad \text{where } \Lambda(\eta,\hat{y}) \equiv \left[\frac{(1-\hat{y})(1+\eta)}{(1-\eta)(1+\hat{y})}\right]^{1/2}. \tag{15}$$

For details, see the references.

Specializing to steady state, as $\partial/\partial t \to 0$, the characteristic time $\tau$ cancels, and we obtain

$$q_{\mathrm{f}}(\hat{f}) = -\frac{\rho_{\mathrm{f}} v_\infty^2}{b}\frac{\partial}{\partial \hat{y}}\left(\int_{-1}^{1} N\left(\eta,\hat{y}\right)\frac{\partial}{\partial \eta}\hat{f}(\eta)\,\mathrm{d}\eta\right). \tag{16}$$

This results in a one-dimensional integro-differential model in terms of $f(y)$. In this approach, a flow solver is not needed; the pressure difference is obtained directly by solving the strongly coupled model.

After inserting (16) into (13), multiplying both sides of the equation by a test function $\psi$, integrating over the domain of $\hat{y}$, performing the appropriate integrations by parts, moving the aerodynamic reaction term to the left-hand side, and finally multiplying both sides of the equation by $b^4$, we have the weak form

$$\begin{aligned}
&\alpha b^4\int_{-1}^{1}\hat{f}\psi\,\mathrm{d}\hat{y} - 2\beta b^2 D\int_{-1}^{1}\hat{f}'\psi'\,\mathrm{d}\hat{y} + D\int_{-1}^{1}\hat{f}''\psi''\,\mathrm{d}\hat{y}\\
&\qquad + 2\beta b^2 D\left[\hat{f}'\psi\right]_{\hat{y}=-1}^{1} + D\left[\hat{f}^{(3)}\psi\right]_{\hat{y}=-1}^{1} - D\left[\hat{f}''\psi'\right]_{\hat{y}=-1}^{1}\\
&\qquad - \rho_{\mathrm{f}} b^3 v_\infty^2\int_{-1}^{1}\left(\int_{-1}^{1} N(\eta,\hat{y})\frac{\partial}{\partial\eta}\hat{f}(\eta)\,\mathrm{d}\eta\right)\psi'(\hat{y})\,\mathrm{d}\hat{y}\\
&\qquad + \rho_{\mathrm{f}} b^3 v_\infty^2\left[\left(\int_{-1}^{1} N(\eta,\hat{y})\frac{\partial}{\partial\eta}\hat{f}(\eta)\,\mathrm{d}\eta\right)\psi(\hat{y})\right]_{\hat{y}=-1}^{1}\\
&\qquad = 0, \quad -1 < \hat{y} < 1.
\end{aligned} \tag{17}$$

Note that with the F boundary conditions, all boundary terms give a nonzero contribution.

To implement the free edge boundary conditions, we insert the trial function (8) into the boundary conditions (5) and (6), and solve for $\hat{f}''$ and $\hat{f}^{(3)}$, respectively. We obtain that on the free edges,

$$\frac{1}{b^2}\hat{f}'' + \nu\beta\hat{f} = 0 \quad \Rightarrow \quad \hat{f}'' = -\nu\beta b^2\hat{f}, \tag{18}$$

$$\frac{1}{b^3}\hat{f}^{(3)} + \frac{1}{b}(2-\nu)\beta\hat{f}' = 0 \quad \Rightarrow \quad \hat{f}^{(3)} = (\nu-2)\beta b^2\hat{f}'. \tag{19}$$

Expressions (18) and (19) are then inserted into the boundary terms in (17).

Finally, the weak form is discretized using finite elements. We use the standard Galerkin representation

$$\hat{f}(\hat{y}) = \sum_{n=1}^{\infty} f_n\varphi_n(\hat{y}), \tag{20}$$

where $f_n$ are the unknown coefficients, and $\varphi_n$ are basis functions for each global degree of freedom. In the numerical implementation, the series is truncated at a finite upper limit $N$.

Continuity of the basis functions across element boundaries must be $C^1$ due to the bending term, which involves second derivatives in the weak form. In this study, this was achieved by using elements of the Hermite type.

We proceed by the standard Galerkin method, where the test functions $\varphi_j$ are taken as the same functions as the basis. Inserting (20) to (17), and moving the first term to the right-hand side, this leads to a generalized linear eigenvalue problem in the following form:

$$\begin{aligned}\Big\{2\beta b^2 D(\mathbf{C}_{jn} + \boldsymbol{\Gamma}^{\mathrm{C}}_{jn}) + D\left[\mathbf{D}_{jn} + (\nu-2)\beta b^2 \boldsymbol{\Gamma}^{1}_{jn} - \nu\beta b^2 \boldsymbol{\Gamma}^{2}_{jn}\right] \\ + (\rho_{\mathrm{f}} b^3 v_\infty^2)(\mathbf{c}_{jn} + \boldsymbol{\Gamma}^{\mathrm{c}}_{jn})\Big\} f_n = -\alpha b^4 \mathbf{A}_{jn} f_n, \quad \text{for all } j = 1, \ldots, N,\end{aligned} \tag{21}$$

where summation over $n = 1, \ldots, N$ is implied. In this problem for $\hat{f}(\hat{y})$, there are no Dirichlet boundary conditions, so there is no need to eliminate any of the degrees of freedom.

The matrices in (21) are given by

$$\begin{aligned}\mathbf{A}_{jn} &= \int_{-1}^{1} \varphi_n \varphi_j \,\mathrm{d}\hat{y}, \qquad \mathbf{C}_{jn} = -\int_{-1}^{1} \varphi_n' \varphi_j' \,\mathrm{d}\hat{y}, \qquad \boldsymbol{\Gamma}^{\mathrm{C}}_{jn} = \left[\varphi_n' \varphi_j\right]_{\hat{y}=-1}^{1}, \\ \mathbf{D}_{jn} &= \int_{-1}^{1} \varphi_n'' \varphi_j'' \,\mathrm{d}\hat{y}, \qquad \boldsymbol{\Gamma}^{1}_{jn} = \left[\varphi_n' \varphi_j\right]_{\hat{y}=-1}^{1}, \qquad \boldsymbol{\Gamma}^{2}_{jn} = -\left[\varphi_n \varphi_j'\right]_{\hat{y}=-1}^{1}, \\ \mathbf{c}_{jn} &= -\int_{-1}^{1} \left( \int_{-1}^{1} N(\eta, \hat{y}) \frac{\partial}{\partial \eta} \varphi_n(\eta) \,\mathrm{d}\eta \right) \varphi_j'(\hat{y}) \,\mathrm{d}\hat{y}, \\ \boldsymbol{\Gamma}^{\mathrm{c}}_{jn} &= \left[ \left( \int_{-1}^{1} N(\eta, \hat{y}) \frac{\partial}{\partial \eta} \varphi_n(\eta) \,\mathrm{d}\eta \right) \varphi_j(\hat{y}) \right]_{\hat{y}=-1}^{1}.\end{aligned} \tag{22}$$

These matrices are assembled by summing local elementwise contributions in the standard manner. Due to the support of each global basis function spanning only two elements (one element at the ends of the domain), most matrices will be sparse, with the exception of $\mathbf{c}_{jn}$ and $\boldsymbol{\Gamma}^{\mathrm{c}}_{jn}$, which will be full due to the inner integral. The full matrices, however, present no practical issue, because the problem is one-dimensional, and thus the number of degrees of freedom will be relatively small.

Note that Eq. (17) is always valid, whereas in (21) we have applied the boundary conditions (18) and (19). The matrices $\boldsymbol{\Gamma}^{C}_{jn}$ and $\boldsymbol{\Gamma}^{1}_{jn}$ in (22) are the same; this is a coincidence due to the boundary condition (19).

Equation (21) is solved by applying a standard solver for generalized linear eigenvalue problems. The critical velocity $V_0^*$ is computed from the eigenvalue $\alpha$ via

Eq. (11), and the cross-directional slice of the buckling shape, $\hat{f}(\hat{y})$, is assembled via Eq. (20) (using a finite upper limit $N$). Finally, the full buckling shape is obtained from (8).

## 4 Numerical Examples

In the following are some numerical examples. The number of elements used was 40. The problem parameters were set as follows, considered typical for paper production applications:

$$T_0 = 500\,\mathrm{N/m}, \quad m = 0.08\,\mathrm{kg/m^2}, \quad E = 10^9\,\mathrm{N/m^2},$$
$$b = 0.5\,\mathrm{m}, \quad h = 10^{-4}\,\mathrm{m}, \quad \nu = 0.3, \quad \rho_\mathrm{f} = 1.225\,\mathrm{kg/m^3}. \tag{23}$$

The bending rigidity $D$ was calculated from Eq. (3). The constants $\alpha$ and $\beta$ were determined by (11) and (9), respectively.

The aspect ratio of the plate and the fluid free-stream velocity were varied across the examples. To change the aspect ratio, the span length $\ell$ was varied, keeping the span half-width $b$ as constant.

Figure 2 shows the behaviour of the lowest critical velocity $V_0^*$ as a function of the plate shape parameter $\ell/b$ and fluid free-stream velocity $v_\infty$. The behaviour is qualitatively similar to that of a travelling panel subjected to axial flow, when the fluid flow parameters are changed (see [8, Fig. 40]).

Figures 3, 4, 5, 6, 7, 8, 9, 10 and 11 show the critical buckling mode for various aspect ratios and fluid velocities. In each figure, the left subfigure shows the shape

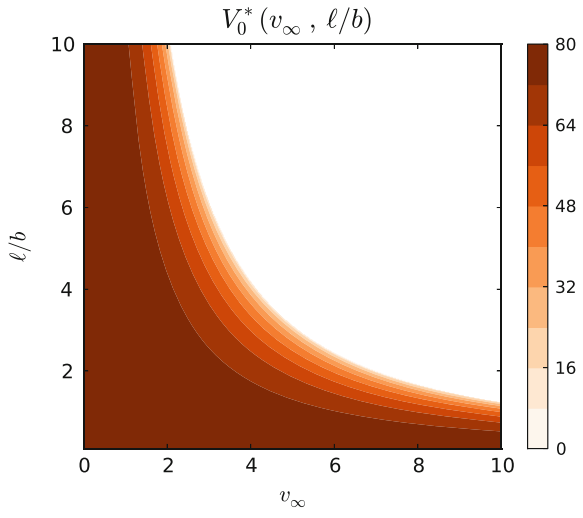



**Fig. 2** Behaviour of the lowest critical velocity $V_0^*$ as a function of the plate shape parameter $\ell/b$ and fluid free-stream velocity $v_\infty$. In the *blank area*, there is no physically meaningful solution

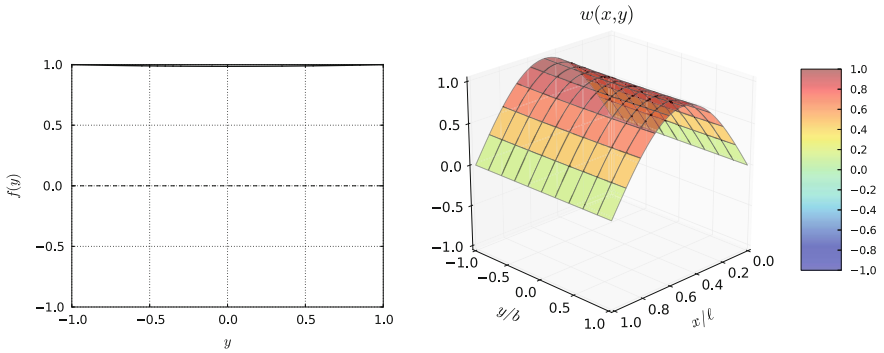



**Fig. 3** Critical buckling mode for $\ell/b = 10$, $\nu_\infty = 0$

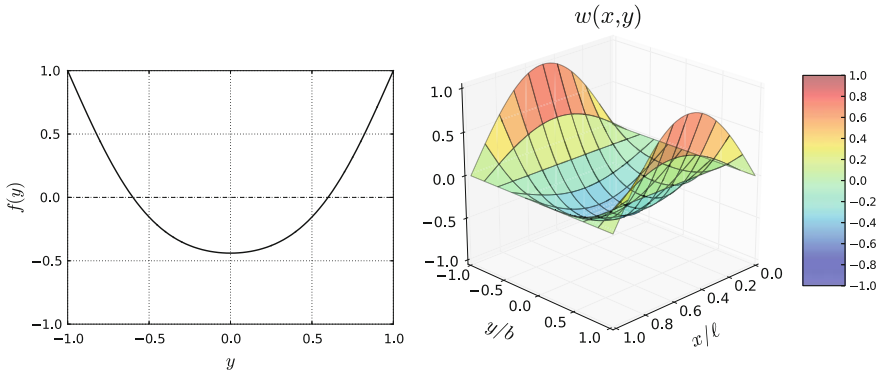



**Fig. 4** Critical buckling mode for $\ell/b = 10$, $\nu_\infty = 0.2$

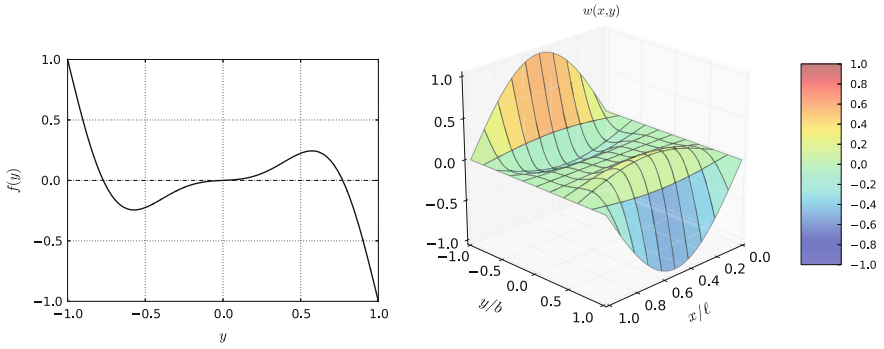



**Fig. 5** Critical buckling mode for $\ell/b = 10$, $\nu_\infty = 0.5$

of the slice, $\hat{f}(\hat{y})$, and the right subfigure shows the complete buckling mode $w$, composed according to Eq. (8). The free constant $A$ was chosen to normalize the maximum of $w$ to 1.

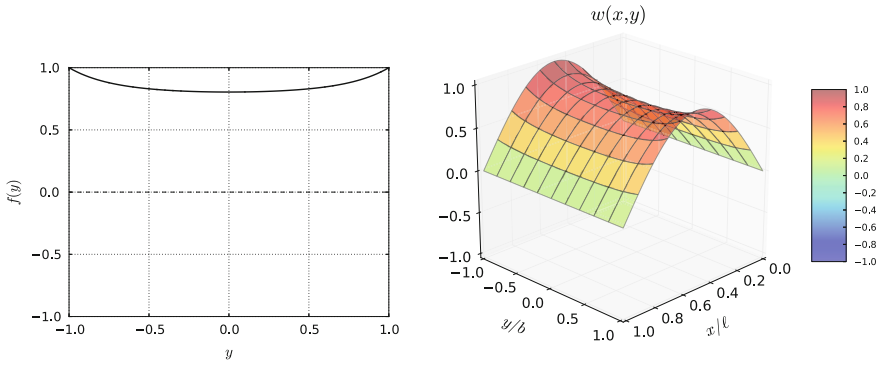



**Fig. 6** Critical buckling mode for $\ell/b = 1$, $v_\infty = 0$

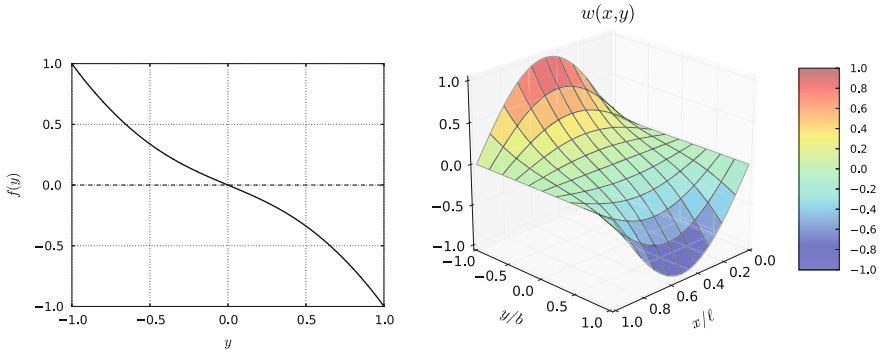



**Fig. 7** Critical buckling mode for $\ell/b = 1$, $v_\infty = 0.2$

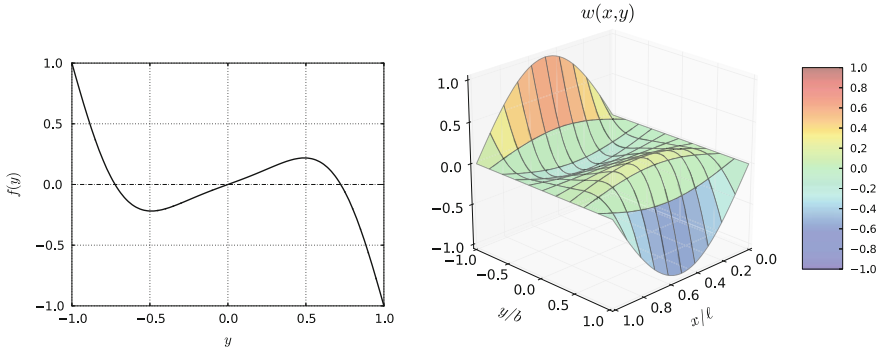



**Fig. 8** Critical buckling mode for $\ell/b = 1$, $v_\infty = 0.5$

Figures 3, 4 and 5 demonstrate a long and narrow span ($\ell/b = 10$), Figs. 6, 7 and 8 a span that is twice as wide as long ($\ell/b = 1$, leading to aspect ratio $\ell/(2b) = 1/2$), and Figs. 9, 10 and 11 a span that is short and wide. The effect of the aspect ratio

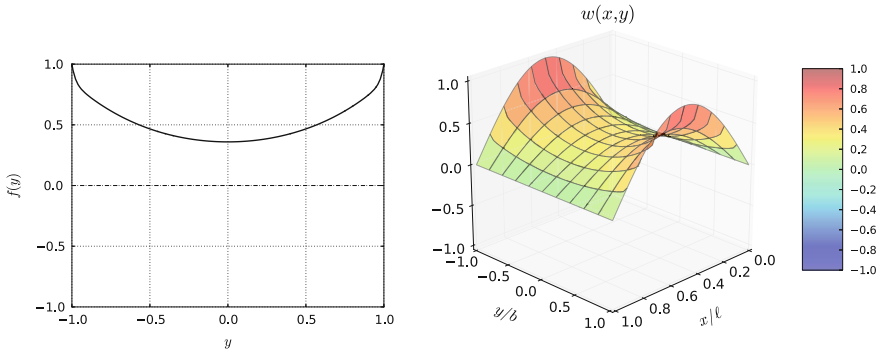



**Fig. 9** Critical buckling mode for $\ell/b = 0.1$, $v_\infty = 0$

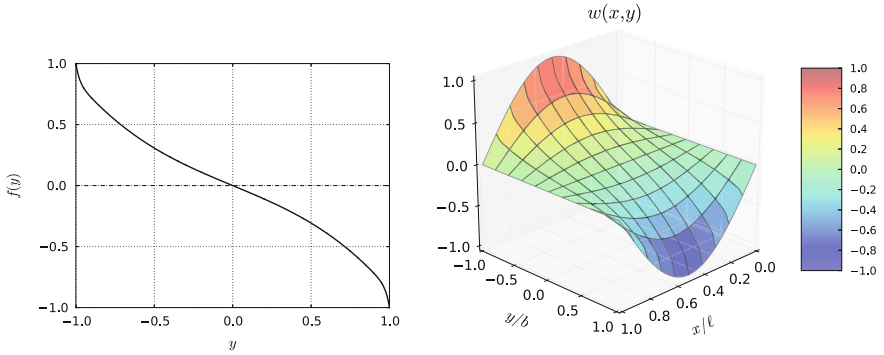



**Fig. 10** Critical buckling mode for $\ell/b = 0.1$, $v_\infty = 1.0$

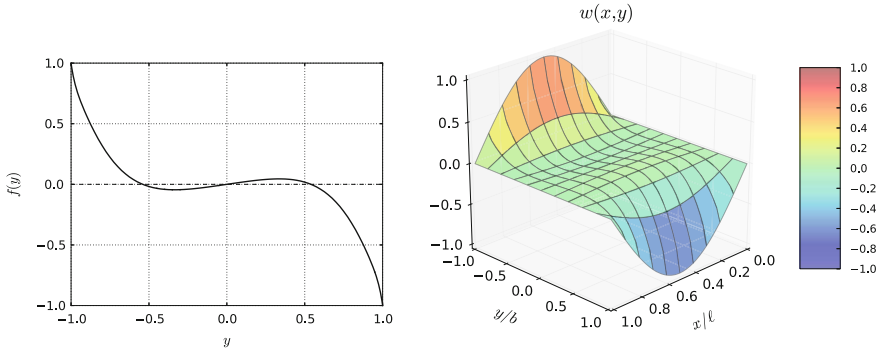



**Fig. 11** Critical buckling mode for $\ell/b = 0.1$, $v_\infty = 2.5$

is qualitatively similar to the corresponding vacuum case; for short and wide spans, the displacement in the buckling mode becomes localized near the free edges of the travelling plate (see [4]).

From the figures, we observe that first, the buckling mode is very sensitive to the parameter values, while the critical velocity is not. Secondly, for some values of the problem parameters, the buckling mode becomes antisymmetric with respect to $\hat{y} = y/b$. In the corresponding vacuum case, the buckling mode that corresponds to the lowest critical velocity is always symmetric [4].

## 5 Conclusion

The present study concerned the stability analysis of an axially moving Kirchhoff plate in an open draw, subjected to an axial potential flow perpendicular to the direction of the travelling motion. The problem was solved in a two-dimensional approximation, which allowed semi-analytical solution of the fluid flow subproblem. The approximate problem was reduced into an eigenvalue problem by applying classical Euler stability analysis. The lowest critical plate velocity and the buckling shape were determined numerically using Hermite type finite elements.

It was seen that the parameters affecting the plate behaviour are the plate aspect ratio, $\ell/(2b)$, and the fluid effect coefficient, which is controlled by the free-stream fluid velocity $v_\infty$. The aspect ratio of the plate was seen to have the same effect as in the vacuum case (see [4]). For short and wide spans, the displacement in the buckling mode becomes localized near the free edges.

When either of the two parameters was increased from zero, the critical velocity decreased from its vacuum value. At fixed $\ell/b > 0$ (respectively $v_\infty$), there was a limit value of $v_\infty$ (resp. $\ell/b$), where the buckling occurs already at $V_0 = 0$. This behaviour of the lowest critical velocity as a function of the problem parameters is typical for this kind of equations.

It was observed that the buckling mode is very sensitive to the parameter values, while the critical velocity is not. The same observation holds for several related models, see, e.g., [2–4].

For some values of the problem parameters, the buckling mode was observed to become antisymmetric with respect to the plate width coordinate $y$. In the corresponding vacuum case, the buckling mode that corresponds to the lowest critical velocity is always symmetric.

**Acknowledgments** This research was supported by the Finnish Cultural Foundation. The authors wish to congratulate professor Banichuk on the occasion of his 70th birthday, and to extend their thanks to him for many interesting and fruitful technical discussions over the years, hoping for many more in the years to come.

## References


1. F.R. Archibald, A.G. Emslie, The vibration of a string having a uniform motion along its length. J. Appl. Mech. **25**, 347–348 (1958)
2. N. Banichuk, J. Jeronen, M. Kurki, P. Neittaanmäki, T. Saksa, T. Tuovinen, On the limit velocity and buckling phenomena of axially moving orthotropic membranes and plates. Int. J. Solids Struct. **48**(13), 2015–2025 (2011). doi:10.1016/j.ijsolstr.2011.03.010
3. N. Banichuk, J. Jeronen, P. Neittaanmäki, T. Saksa, T. Tuovinen, Theoretical study on travelling web dynamics and instability under non-homogeneous tension. Int. J. Mech. Sci. **66**, 132–140 (2013). doi:10.1016/j.ijmecsci.2012.10.014
4. N. Banichuk, J. Jeronen, P. Neittaanmäki, T. Tuovinen, On the instability of an axially moving elastic plate. Int. J. Solids Struct. **47**(1), 91–99 (2010). doi:10.1016/j.ijsolstr.2009.09.020
5. N. Banichuk, J. Jeronen, P. Neittaanmäki, T. Tuovinen, Static instability analysis for travelling membranes and plates interacting with axially moving ideal fluid. J. Fluids Struct. **26**(2), 274–291 (2010). doi:10.1016/j.jfluidstructs.2009.09.006
6. N. Banichuk, J. Jeronen, P. Neittaanmäki, T. Tuovinen, Dynamic behaviour of an axially moving plate undergoing small cylindrical deformation submerged in axially flowing ideal fluid. J. Fluids Struct. **27**(7), 986–1005 (2011). doi:10.1016/j.jfluidstructs.2011.07.004
7. T. Frondelius, H. Koivurova, A. Pramila, Interaction of an axially moving band and surrounding fluid by boundary layer theory. J. Fluids Struct. **22**(8), 1047–1056 (2006)
8. J. Jeronen, On the mechanical stability and out-of-plane dynamics of a travelling panel submerged in axially flowing ideal fluid: a study into paper production in mathematical terms. Ph.D. thesis, University of Jyväskylä (2011)
9. L. Kong, R.G. Parker, Approximate eigensolutions of axially moving beams with small flexural stiffness. J. Sound Vibr. **276**(1–2), 459–469 (2004)
10. A. Kulachenko, P. Gradin, H. Koivurova, Modelling the dynamical behaviour of a paper web. Part II. Comput. Struct. **85**(3–4), 148–157 (2007)
11. R.G. Parker, On the eigenvalues and critical speed stability of gyroscopic continua. J. Appl. Mech. **65**(1), 134–140 (1998)
12. A. Pramila, Sheet flutter and the interaction between sheet and air. TAPPI J. **69**(7), 70–74 (1986)
13. R.A. Sack, Transverse oscillations in travelling strings. Br. J. Appl. Phys. **5**(6), 224–226 (1954)
14. A. Simpson, Transverse modes and frequencies of beams translating between fixed end supports. J. Mech. Eng. Sci. **15**(3), 159–164 (1973)
15. R. Skutch, Uber die Bewegung eines gespannten Fadens, welcher gezwungen ist durch zwei feste Punkte, mit einer constanten Geschwindigkeit zu gehen, und zwischen denselben in Transversalschwingungen von geringer Amplitude versetzt wird. Ann. Phys. Chem. **61**, 190–195 (1897)
16. R.D. Swope, W.F. Ames, Vibrations of a moving threadline. J. Frankl. Inst. **275**(1), 36–55 (1963)
17. Y. Wang, L. Huang, X. Liu, Eigenvalue and stability analysis for transverse vibrations of axially moving strings based on Hamiltonian dynamics. Acta Mech. Sin. **21**(5), 485–494 (2005)

# Multiaxial Fatigue Criteria and Durability of Titanium Compressor Disks in Low- and Very-high-cycle Fatigue Modes

**Nikolay Burago and Ilia Nikitin**

**Abstract** Life duration for titanium disks of low temperature part of compressor aero-engine D30-Ku is investigated. Several criteria and models are tested under conditions of low-cycle fatigue (LCF) and very-high-cycle fatigue (VHCF). Parameters of the criteria and models are determined from uniaxial fatigue tests for titanium alloy VT3-1. Stress-strain state of disks and blades is calculated taking into account cyclic centrifugal, aerodynamic, contact loads and blade vibrations. Calculated stresses and strains are used as input data for multiaxial models of LCF and VHCF regimes. Location and scales of fracture as well as time to fracture are calculated. The results of calculations are in good agreement with observations during engine exploitation and correspond to data of fractographic investigations of damaged disks.

**Keywords** Fracture and damage · Fatigue criteria

**Mathematical Subject Classification:** 74R20

## 1 Introduction

In this paper we consider the problem of determining the duration of safe operation of structures. In experiments [1] it is shown that under the action of cyclic loads after several millions or billions of cycles the material may be damaged even if during this time only small elastic deformations were observed and in the material there were no signs of macroscopic defects. To date, several phenomenological models of fatigue failure have been developed [2, 5–9, 11–13], generalizing the experimental

N. Burago (✉)
Ishlinski Institute for Problems in Mechanics of RAS, Vernadskogo 101, Block 1, 119526 Moscow, Russia
e-mail: burago@ipmnet.ru

I. Nikitin
Institute for Computer Aided Design of RAS, 19/18 2nd Brestskaya Street, 123056 Moscow, Russia
e-mail: i_nikitin@list.ru



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_8

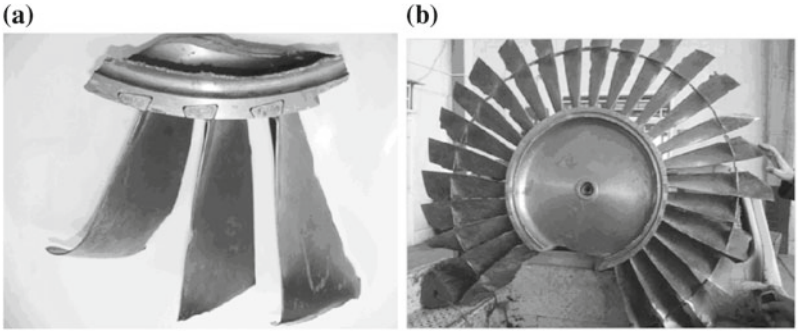



**Fig. 1** Fracture in disks for compressor of aero engines D30-KU-154

data for the case of a multiaxial stress-strain state. To determine the safe operation life of structures for such models it is sufficient calculate the stress-strain state using the linear theory of elasticity. Available models of fatigue failure are divided into three groups, the first of which is based on the criteria of stress state [5, 7, 11], the second group is based on criteria for strain state [2, 6, 12], and third group is based on the calculation of the kinetics of damage [8]. The physical nature of damage in the material structure under the action of cyclic loads, which is investigated in e.g. [1, 10], is still an active topic of study.

Fatigue fracture of disks of gas turbine engines (GTE) is a well-known phenomenon [10]. The gas turbine engines are subjected to various cyclic loads. The cycles of "take-off-flight-landing" correspond to low cycle fatigue. The presence of small vibrations with $R = \sigma_{\min}/\sigma_{\max} > 0.8$ corresponds to very high cycle fatigue.

It is demonstrated that such additional loads can essentially alter the picture of damage accumulation in service. There were several unpredicted cases of damage to titanium rotor disks in the low pressure compressor stage in the D30-KU-154 engine (Fig. 1). The fatigue failures may take place earlier than in accordance with the LCF criteria, and that is why a new numerical approach is developed here in order to estimate and to compare life duration predictions not only for the LCF but also for the VHCF regimes.

The finite element model for the disk with blades under consideration is developed in [3] and the 3D stress-strain state is analyzed there taking into account centrifugal and aerodynamic loads, contact and vibration loads. Aero-elastic effects due to mutual influence of aerodynamic loads and structural shape changes are also taken into account.

It is assumed that during the flight cycle the maximum values of stresses and strains correspond to the aircraft flight velocity of 200 m/s and the disk rotation frequency of 3000 rpm. It is assumed that during many years of safe exploitation the disks are subjected only to elastic deformations and do not contain cracks. Our first goal is to calculate disk life duration as the limiting number of cycles to failure and to detect the location of failure using fatigue criteria [2, 4–9, 11–13] for LCF regime of cyclic loading. The results are compared to available in-flight data.

Our second goal is to study the VHCF regime of cyclic loading due to additional action of high frequency axial vibrations of blade shroud ring. The maximum vibration amplitude is assumed to be at a disk rotation frequency of 3000 rpm. Evaluations of life duration are presented in terms of the number of vibration cycles according to various VHCF criteria.

Until now, in the literature there are no experimental data and theoretical multiaxial models applicable to the considered material (titanium alloy VT3-1) for VHCF regime. Therefore the known multiaxial LCF criteria are generalized here and are applied to study the VHCF regime. The generalization is performed using similarity of the left and right branches of bimodal fatigue curves. The values of parameters for generalized criteria are determined using the few available experimental data found for VHCF regime.

There are no other publications on the life duration estimates for three-dimensional structure in VHCF regime in scientific literature yet. The calculated results for the low-cycle and very-high-cycle fatigue are compared. It is found that in the life duration estimates are close to each other. That is why the VHCF mechanism should be taken into account in the resource estimates of GTE.

## 2 Computational Model of Contact Structure "Disk and Blades"

The application of finite element method for contact structure disk and blades is described in [3]. The three-dimensional stress-strain state of the contact system of the compressor disk and blades (Fig. 2) is numerically analyzed using finite-element method. The distributed aerodynamic loads are approximated using analytical methods based on modification of classical solutions to the problem of flow about a lattice of plates at arbitrary angle of attack. The solution of aerodynamic problem is obtained using theory of complex variable methods and the isolated profile hypothesis with the blade deformable shape changes taken into account [3]. The combined action of centrifugal, aerodynamic and contact loads is taken into account. First, stress-strain state is calculated for the full computational model "disk with 22 blades" (Fig. 2a) using a rough grid with a number of elements of about $10^5$. Then, the solution obtained from the calculation of the full model is used to move the boundary conditions onto the sides of the disk sector with a single blade (Fig. 2b), which is calculated using the refined grid with the same number of finite elements of about $10^5$, which is quite acceptable for calculations on a personal computer. Calculated stresses and strains are used in LCF and VHCF models for life duration estimations.

Extended numerical model is used for VHCF analysis taking into account low-amplitude axial vibrations of shroud ring. The vibrations cause axial displacements of the shroud ring. The vibrations along the ring take the form of 12–16 half waves. For the disc-blade sector calculation it was assumed that the displacement of right side of shroud ring is equal to zero and that the displacement of its left side varies

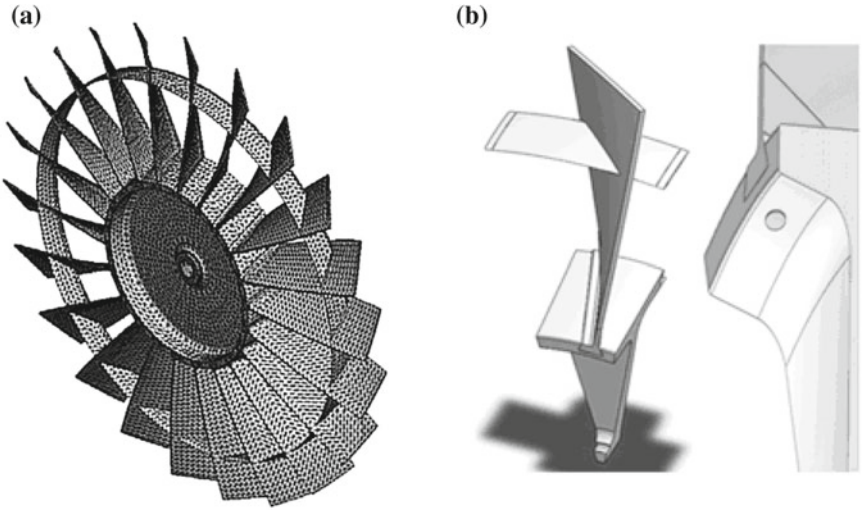



**Fig. 2** The contact system of the compressor disk and blades: **a** disk-blades contact structure, **b** disk sector with blade and part of shroud ring

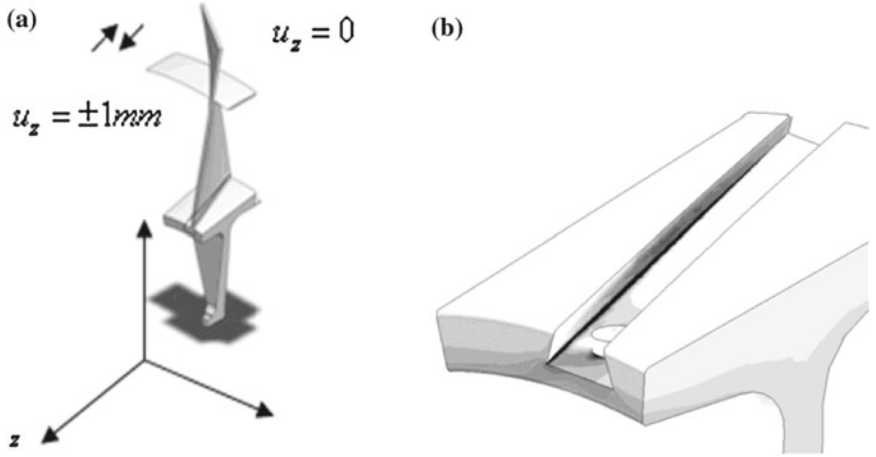



**Fig. 3** The contact system of the compressor disk and blades: **a** schematic for vibration analysis, **b** the slot of dovetail-type connection

in the range the maximal vibration amplitude of $\pm 1$ mm (Fig. 3a) for a frequency of 3000 rpm. Vibration stresses are imposed on the basic stresses and then are used in VHCF models for life duration estimations. The most dangerous area is shown in Fig. 3b.

# 3 Low Cycle Fatigue Models

## *3.1 LCF Models Based on the Stress State Estimation*

The coefficients in criteria of fatigue fracture are determined from the experiments for uniaxial cyclic loading for different values of stress ratio $R = \sigma_{\min}/\sigma_{\max}$, where $\sigma_{\max}$ and $\sigma_{\min}$ are the maximum and minimum stresses during the cycle. These values are used to define the stress amplitude $\sigma_a = (\sigma_{\max} - \sigma_{\min})/2$. In the case of uniaxial deformation, the test data are described using Wohler curves, which can be analytically written by using the Basquin formula [4]:

$$\sigma = \sigma_u + \sigma_c N^{\beta} \tag{1}$$

Here $\sigma_u$ is the fatigue limit, $\sigma_c$ is the fatigue strength factor, $\beta$ is the fatigue strength exponent, and $N$ is the number of cycles to fracture. A typical amplitude fatigue curve is depicted in Fig. 4. The curve consists of two branches corresponding to two fatigue regimes, the low cycle fatigue regime with fatigue limit $\sigma_u$ and very high cycle fatique regime with fatigue limit $\tilde{\sigma}_u$. The regime of interest is located in the left branch of the curve for life duration $N < 10^7$ cycles. The problem of fatigue fracture is that the spatial function of the life duration distribution $N$ must be determined from equations in the form (1) generalized to multi-axis stress state and containing the calculated stresses for the structure under study. Below the basic methods of generalizing the results of uniaxial tests to multi-axis stress state are considered [4].

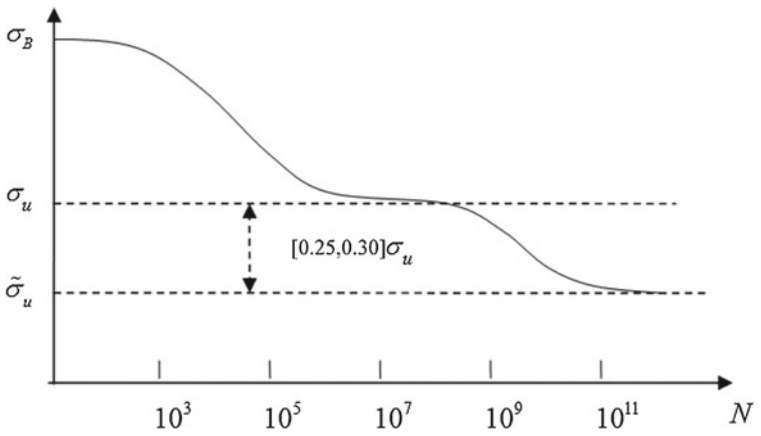



**Fig. 4** Typical fatigue Wohler's curve for metals. Here $\sigma_B$ is the strength limit, $\sigma_u$ and $\tilde{\sigma}_u$ are fatigue limits for LCF and VHCF regimes respectively

### 3.1.1 Sines Model

According to Sines [11], the uniaxial fatigue curve (1) may be generalized to multi-axis stress state as

$$\Delta\tau/2 + \alpha_s \sigma_{\text{mean}} = S_0 + AN^{\beta} \tag{2}$$

where

$$\sigma_{\text{mean}} = (\sigma_1 + \sigma_2 + \sigma_3)_{\text{mean}}$$
$$\Delta\tau = \sqrt{(\Delta\sigma_1 - \Delta\sigma_2)^2 + (\Delta\sigma_1 - \Delta\sigma_3)^2 + (\Delta\sigma_2 - \Delta\sigma_3)^2}/3$$

Here the parameter $\sigma_{\text{mean}}$ is the mean stress over a loading cycle. The parameter $\Delta\tau$ is the change in the octahedral tangent stress per cycle. The parameter $\Delta\tau/2$ is the octahedral tangent stress amplitude. Parameters $\alpha_s$, $S_0$, $A$ and $\beta$ are experimental. The model parameters for uniaxial fatigue curves are determined in [4]:

$$S_0 = \sqrt{2}\sigma_u/3, \quad A = 10^{-3\beta}\sqrt{2}(\sigma_B - \sigma_u)/3$$
$$\alpha_s = \sqrt{2}(2k_{-1} - 1)/3, \quad k_{-1} = \sigma_u/(2\sigma_{u0})$$

Here $\sigma_u$ and $\sigma_{u0}$ are the fatigue limits according to fatigue curves for $R = -1$ and $R = 0$, $N \approx 10^7 - 10^8$ cycles. It is assumed that the decrease of strength limit $\sigma_B$ is negligible up to the values $N \approx 10^3$ (Fig. 4).

### 3.1.2 Crossland Model

According to Crossland [5], the uniaxial fatigue curve (1) may be generalized to multi-axis stress state as

$$\Delta\tau/2 + \alpha_c(\overline{\sigma}_{\max} - \Delta\tau/2) = S_0 + AN^{\beta}, \quad \overline{\sigma}_{\max} = (\sigma_1 + \sigma_2 + \sigma_3)_{\max} \tag{3}$$

Here $\overline{\sigma}_{\max}$ is the maximum sum of principal stresses in a loading cycle, and $\alpha_c$, $S_0$, $A$ and $\beta$ parameters determined from experimental data. The final expressions for model parameters of the multiaxial model are determined in [4] for $R = -1$ and $R = 0$ as

$$S_0 = \sigma_u\left[\sqrt{2}/3 + (1 - \sqrt{2}/3)\alpha_c\right]$$
$$A = 10^{-3\beta}\left[\sqrt{2}/3 + (1 - \sqrt{2}/3)\alpha_c\right](\sigma_B - \sigma_u)$$
$$\alpha_c = (k_{-1}\sqrt{2}/3 - \sqrt{2}/6)\Big/\left[(1 - \sqrt{2}/6) - k_{-1}(1 - \sqrt{2}/3)\right]$$
$$k_{-1} = \sigma_u/(2\sigma_{u0})$$

### 3.1.3 Findley Model

The form of this model for the multi-axis stress state is proposed by Findley [7]

$$(\Delta\tau_s/2 + \alpha_F\sigma_n)_{\max} = S_0 + AN^{\beta} \tag{4}$$

Here $\tau_s$ and $\sigma_n$ are the absolute magnitudes of tangent stress and normal stress for the plane with normal vector $n_i$. For this plane, the combination $\Delta\tau_s/2+\alpha_F\sigma_n$ takes a maximum value. The model parameters are

$$S_0 = \sigma_u\left(\sqrt{1+\alpha_F{}^2}+\alpha_F\right)/2,$$
$$A = 10^{-3\beta}(\sqrt{1+\alpha_F{}^2}+\alpha_F)(\sigma_B-\sigma_u)/2$$
$$\alpha_F = \left[\sqrt{5k_{-1}{}^2-2k_{-1}}/2-k_{-1}(1-k_{-1})\right]\left[k_{-1}(2-k_{-1})\right]^{-1}$$

Approximate parameter values for the titanium alloy Ti-6Al-4V [4] (which are used in the computational example considered below) are limit strength of $\sigma_B = 1100$ MPa, fatigue limits based on $\sigma_a(N)$ amplitude curves for $R = -1$ and $R = 0$ of $\sigma_u = 450$ MPa and $\sigma_{u0} = 350$ MPa, exponent in the power-law dependence on the number of cycles of $\beta = -0.45$, Young's modulus of $E = 116$ GPa, shear modulus of $G = 44$ GPa, and Poisson's ratio of $\nu = 0.32$.

## *3.2 LCF Models Based on the Strain State Estimation*

Classical Coffin-Manson relation [10] describing uniaxial fatigue fracture on the basis of deformations is

$$\Delta\varepsilon/2 = (2N)^b\sigma_c/E + \varepsilon_c(2N)^c$$

Here $\sigma_c$ is the (axial) fatigue strength coefficient, $\varepsilon_c$ is the (axial) fatigue plasticity coefficient, $b$ and $c$ are the fatigue strength and fatigue plasticity exponents. Models generalizing the Coffin-Manson relation to the case of multi-axis fatigue fracture are briefly outlined below.

### 3.2.1 Brown-Miller Model

This model is proposed by Brown and Miller [2]; it takes into account the influence of tensile strains at the plane of maximum shear strain:

$$\frac{\Delta\gamma_{\max}}{2} + \alpha_{bm}\Delta\varepsilon_{\perp} = \beta_1\frac{\sigma_c - 2\sigma_{\perp\text{mean}}}{E}(2N)^b + \beta_2\varepsilon_c(2N)^c \tag{5}$$

Here $\gamma_{ij} = 2\varepsilon_{ij}$, $\varepsilon_{ij}$ are the strain tensor components; $\Delta\gamma_{\max}/2$ is the range of the maximum shear strains attained on the plane; $\Delta\varepsilon_{\perp}$ is the range of the tensile strains on this plane, and $\sigma_{\perp\text{mean}}$ is the cycle-average tensile stress on this plane. Approximate values for the coefficients are provided in [**?** ]: $\alpha_{bm} = 0.3$, $\beta_1 = (1+\nu)+(1-\nu)\alpha_{bm}$, $\beta_2 = 1.5 + 0.5\alpha_{bm}$.

### 3.2.2 Fatemi-Socie Model

This model is proposed by Fatemi and Socie [6]; it takes into account the influence of the normal stresses at the plane of maximum shear strains:

$$\frac{\Delta\gamma_{\max}}{2}(1 + k\frac{\sigma_{\perp\,\max}}{\sigma_y}) = \frac{\tau_c}{G}(2N)^{b_0} + \gamma_c(2N)^{c_0} \tag{6}$$

Here $\sigma_{\perp\,\max}$ is the cycle-maximum normal stress on the plane where $\gamma_{\max}$ is attained, $\sigma_y$ is the material yield strength, $\tau_c$ is the fatigue (shear) strength coefficient, $\gamma_c$ is the fatigue (shear) plasticity coefficient, $b_0$ and $c_0$ are the fatigue strength and fatigue plasticity exponents. The coefficient $k$ is approximately equal to $k = 0.5$ [4].

### 3.2.3 Smith-Watson-Topper Model

This model is described in [12] and accounts for the influence of the normal stress at the plane of maximum tensile strain:

$$\frac{\Delta\varepsilon_1}{2}\sigma_{\perp 1\,\max} = \frac{{\sigma_c}^2}{E}(2N)^{2b} + \sigma_c\varepsilon_c(2N)^{b+c} \tag{7}$$

Here $\Delta\varepsilon_1$ is the change in the maximum principal strain per cycle and $\sigma_{\perp 1\,\max}$ is the maximum normal stress at the plane of maximum tensile strain. The fatigue parameters for titanium alloys for this class of models are selected based on experimental data [4]: $\sigma_c = 1445$ MPa, $\varepsilon_c = 0.35$, $b = -0.095$, $c = -0.69$, $\tau_c = 835$ MPa, $\gamma_c = 0.20$, $b_0 = -0.095$, $c_0 = -0.69$, $\sigma_y = 910$ MPa.

## *3.3 LCF Models Based on Damage Estimation*

### 3.3.1 Lemaitre-Chaboche Model

The differential equation for damage D accumulated under multi-axis cyclic loading is proposed in [8] and after integration may be written as

$$N = \frac{1}{(1+\beta)a_M}\left[\frac{(1-3b_2\overline{\sigma})}{A_{IIa}}\right]^{\beta}\left\langle\frac{(\sigma_u - \sigma_{VM})}{(A_{IIa} - A^*)}\right\rangle \tag{8}$$

Here the notation from [8] is used

$$A_{IIa} = 0.5\sqrt{1.5\left(S_{ij,\max} - S_{ij,\min}\right)\left(S_{ij,\max} - S_{ij,\min}\right)}$$
$$\sigma_{VM} = \sqrt{0.5 S_{ij,\max} S_{ij,\max}}$$
$$\overline{\sigma} = (\sigma_1 + \sigma_2 + \sigma_3)_{\text{mean}}/3$$
$$A^* = \sigma_{10}\left(1 - 3b_1\overline{\sigma}\right)$$
$$a_M = a/M_0^{\beta}$$

The parameters $S_{ij,\max}$ and $S_{ij,\min}$ are maximum and minimum values of stress deviator during loading cycle; the angle brackets are defined as: $\langle X\rangle = 0$ for $X < 0$ and $\langle X\rangle = X$ for $X \geq 0$. The model parameters for a titanium alloy are given in [8]: $\beta = 7.689$, $b_1 = 0.0012$, $b_2 = 0.00085$ 1/MPa, $a_M = 4.1 \times 10^{-28}$, $\sigma_{10} = 395$ MPa, and $\sigma_u = 1085$ MPa.

### 3.3.2 The Liege University (LU) Model

This model is proposed and validated in [6]. In this case, the integrated differential equation for the damage is

$$N = \frac{\gamma + 1}{C}\left\langle \frac{\sigma_u - \theta \cdot \sigma_{VM}}{A_{IIa} - A^*} \right\rangle f_{\text{cr}}^{-(\gamma+1)} \tag{9}$$

Here the notation from [8] is used

$$f_{\text{cr}} = \frac{1}{b}\left(A_{IIa} + a\sigma_H - b\right), \quad f_{\text{cr}} > 0$$
$$A^* = \sigma_{-1}\left(1 - 3s\sigma_H\right)$$
$$\sigma_H = (\sigma_1 + \sigma_2 + \sigma_3)_{\max}/3$$

The model parameters are taken from [8]: $a = 0.467$, $b = 220$ MPa, $\gamma = 0.572$, $C = 7.12 \times 10^{-5}$, $\theta = 0.75$, $s = 0.00105$ 1/MPa, $\sigma_{-1} = 350$ MPa, $\sigma_u = 1199$ MPa.

## *3.4 Results of LCF Calculations*

As an example, we consider the problem of fatigue fracture of GTE compressor disks under low-cycle-fatigue (LCF) conditions. For each FLC (flight loading cycle) it is assumed that maximal loads and rotation correspond to cruising speed of aircraft. The problem is to calculate the life duration of the disk (N—the number of FLCs before fracture) from relations (2)–(9). To this end, it is necessary to calculate the stress state of the contact system of the compressor disk and blades under the combined action of centrifugal, aerodynamic and contact loads.

The input parameters include the angular velocity of rotation $\omega = 314$ rad/s (3000 rpm), the dynamic pressure at infinity $\rho v_\infty^2/2 = 26000$ N/m$^2$, corresponding to a flow velocity of 200 m/s and the air density of 1.3 kg/m$^3$. The total number of finite elements does not exceed 1,00,000 making it possible to solve the problem using usual personal computer. The material properties are $E = 116$ GPa, $\nu = 0.32$, and $\rho = 4370$ kg/m$^3$ for the disk (titanium alloy), $E = 69$ GPa, $\nu = 0.33$ and $\rho = 2700$ kg/m$^3$ for the blades and blade shroud ring (aluminum alloy), and $E = 207$ GPa, $\nu = 0.27$ and $\rho = 7860$ kg/m$^3$ for the fixing pins (steel).

The computations [3] indicate that the most dangerous areas are situated in the contact area of dovetail-type between the disk and the blades. The computations [3] also indicate that the best correspondence of computational and experimental data for stress concentration is provided when the detachment and slip of contacting elements (disk and blades) are taken into account. At the fixing pins boundary (Fig. 2b) the conditions of complete adhesion are used according to technological considerations. The zone of maximum tensile stress concentration is situated in the left (rounded) corner of the contact area of dovetail-type (Fig. 3b). The stress concentration is increased from the front to the rear portion of the groove according to observable nucleation of fatigue failure in the rear portion of the disk [10].

### *3.5 Estimate of Service Life for Structure Elements According to LCF Criteria*

In Fig. 5a–h, the computed number of flight cycles before fracture N for the chosen criteria and multi-axis models of fatigue fracture are displayed for the left corner of disk-blade contact joint of dovetail-type (in the zones of maximum stress concentration). The boundary of contact zone near the left corner of the groove is depicted by solid line (Fig. 6).

In Fig. 5a–h, the horizontal axis represents the dimensionless coordinate of the rounding of the groove's left corner; the vertical axis represents the dimensionless coordinate across the groove depth. The Sines, Lemaitre-Chaboche, Brown-Miller, and Smith-Watson-Topper criteria provide estimates for the service life of gas turbine engine disks of approximately 20,000–50,000 cycles. The Crossland and LU criteria predict the possibility of fatigue fracture after fewer than 20,000 flight cycles. On the whole, all of the criteria predict similar locations for the fatigue fracture regions. The Fatemi-Socie criterion gives a service life prediction of approximately 1,00,000 cycles. The deviation of the Fatemi-Socie estimate from the results based on the other criteria suggests that the shear mechanism of multi-axis fatigue fracture, which is reflected in this criterion, is not purely realized in flight loading. Remark that 25,000 flight cycles correspond to an in-service lifetime of 50,000 h for two hour flights.

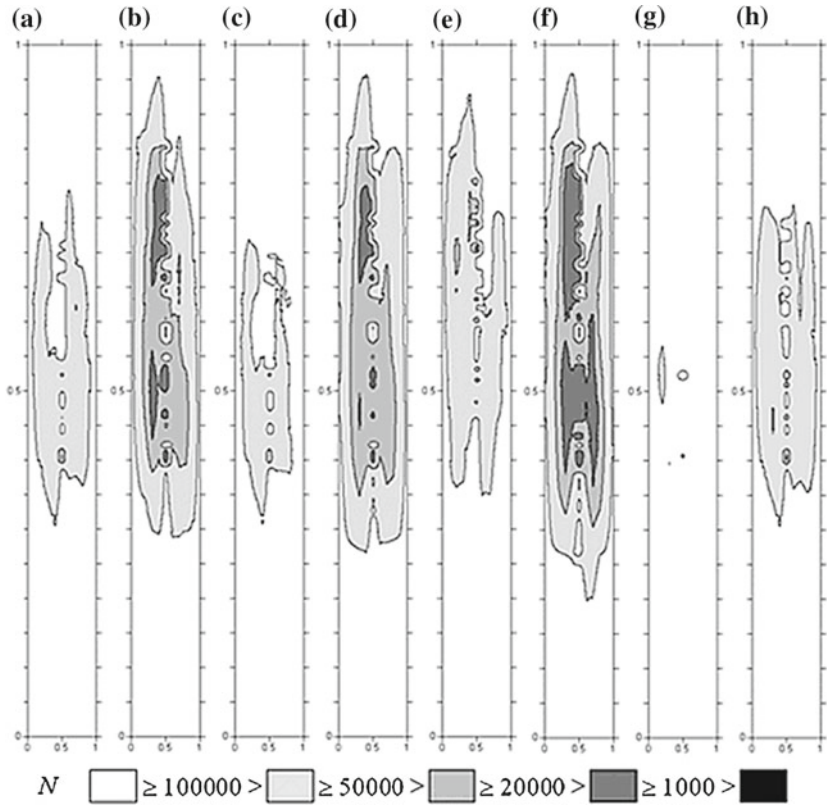
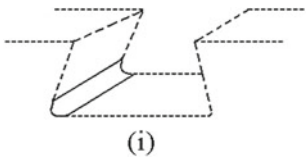



**Fig. 5** Life duration estimates in the area of failure initiation for LCF models: **a** Sines, **b** Crossland, **c** Findley, **d** Brown-Miller, **e** Fatemi-Socie, **f** Smith-Watson-Topper, **g** Lemaitre-Chaboche, **h** Liege University

**Fig. 6** The area of failure initiation; it is indicated by *solid lines* and situated in the slot of dovetail-type connection




## 4 Very High Cycle Fatigue Models

An alternative fatigue mechanism may also be examined for high frequency axial vibrations of the shroud ring. The amplitude of vibrations and stress state disturbances near stress concentrators are relatively small, but the number of high frequency vibrations can be as high as $10^9-10^{10}$, and evaluation of the very-high-cycle fatigue (VHCF) regime is necessary because fatigue may take place even if stress levels are below classical fatigue limits [7]. At present there is no experimentally verified

multi-axis VHCF theory for titanium alloy. In order to obtain life duration estimates the known multi-axis LCF models (2), (4), and (6) are used, taking into account general assumptions about VHCF curves. A typical fatigue curve is presented in Fig. 4, and in case of VHCF the right portion of the curve for $N > 10^8$ is of interest.

### 4.1 Generalization of Sines Model

The VHCF parameters are determined using one-dimensional fatigue curves in a manner similar to the LCF case. The similarity between the left and right halves of the fatigue curve is taken into account through the substitution $\sigma_B \to \sigma_u$, $\sigma_u \to \tilde{\sigma}_u$ and $\sigma_{u0} \to \tilde{\sigma}_{u0}$, where $\tilde{\sigma}_u$ and $\tilde{\sigma}_{u0}$ are new fatigue limits for right half of the fatigue curve for asymmetry factors $R = -1$ and $R = 0$, $N > 10^{11}$ cycles (Fig. 4). The VHCF parameter values for the generalized Sines model (2) are

$$S_0 = \sqrt{2}\tilde{\sigma}_u/3, \quad A = 10^{-8\beta}\sqrt{2}(\sigma_u - \tilde{\sigma}_u)/3,$$
$$\alpha_s = \sqrt{2}(2k_{-1} - 1)/3, \quad k_{-1} = \tilde{\sigma}_u/\tilde{\sigma}_{u0}/2$$

### 4.2 Generalization of Crossland Model

By analogy the VHCF parameters for the generalized Crossland model (4) are

$$S_0 = \tilde{\sigma}_u \left[\sqrt{2}/3 + (1 - \sqrt{2}/3)\alpha_c\right], \quad A = 10^{-8\beta}(\sigma_u - \tilde{\sigma}_u)\left[\sqrt{2}/3 + (1 - \sqrt{2}/3)\alpha_c\right]$$

### 4.3 Generalization of Findley Model

The VHCF parameters for the generalized Findley model (6) are

$$S_0 = \tilde{\sigma}_u \left(\sqrt{1 + \alpha_F{}^2} + \alpha_F\right)/2, \quad A = 10^{-8\beta}(\sigma_u - \tilde{\sigma}_u)(\sqrt{1 + \alpha_F{}^2} + \alpha_F)/2$$

For titanium alloy the following parameter values are used $\sigma_u = 450$ MPa, $\tilde{\sigma}_u = 250$ MPa, $\tilde{\sigma}_{u0} = 200$ MPa. $\beta = -0.3$

### 4.4 Results of VHCF Calculations

The maximum stress concentration occurs near the rounding of the groove's left corner. The calculated limits of $N$ (number of safe vibration cycles) for a selected

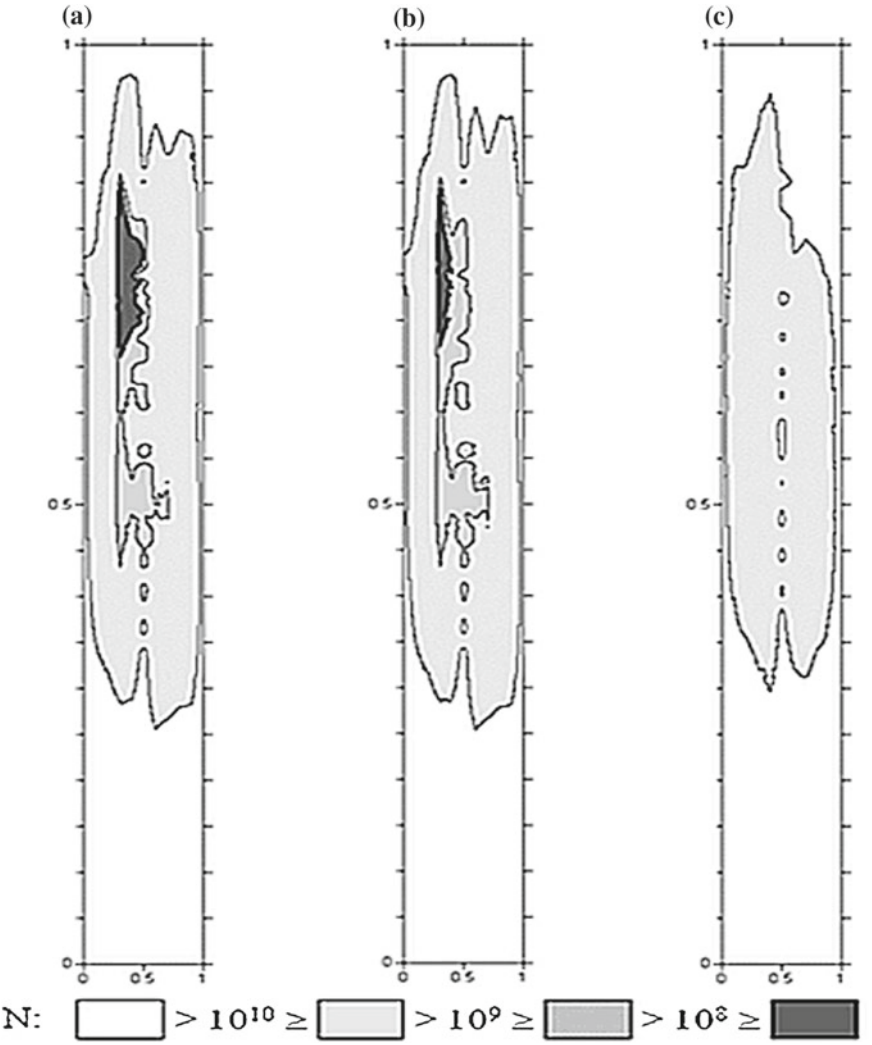



**Fig. 7** Life duration estimates in the area of failure initiation for VHCF models: **a** Sines, **b** Crossland, **c** Findley

area of the left corner are depicted in Fig. 7. The results were obtained using the VHCF generalized criteria of Sines, Crossland, and Findley. Despite the rather low additional vibration stress amplitude level in this case, zones of fatigue failure are also appeared. The fatigue failure zones are situated near the rear portion of the groove's left corner (in the same location as in the LCF case). The safe vibration loading cycle number is approximately equal to $10^9 - 10^{10}$, corresponding to an in-service lifetime of 50,000 h.

## 5 Conclusions

A numerical model is developed to estimate the service life of structural elements for both LCF (flight cycles) and VHCF (vibrations) regimes. Comparative estimates of life duration for GTE compressor disk-blade contact structures were obtained using Sines, Crossland, and Findley fatigue models. The life duration estimates obtained for LCF and VHCF mechanisms coincided closely with the observed service life of titanium compressor disks in the D30-KU-154 GTE. So, original and generalized fatigue failure criteria may be used for estimating in-service life duration of titanium disks.

Although presented life duration estimates are rather approximate, they highlight the possibility of fatigue fracture development in structural elements for both LCF (flight cycles) and VHCF (high frequency low amplitude vibrations) regimes. The most serious hazard may happen due to mutual action of the both mechanisms because they may cause the fatigue failure developed almost simultaneously and in the same location.

**Acknowledgments** The research was supported by the Russian Foundation for Basic Research under projects 15-08-02392-a.

## References

1. C. Bathias, P.C. Paris, *Gigacycle fatigue in mechanical practice* (Marcel Dekker Verlag, 2004)
2. M.W. Brown, K.J. Miller, A theory for fatigue failure under multiaxial stress-strain conditions. Proc. Inst. Mech. Eng. **187**(1), 745–755 (1973)
3. N.G. Burago, A.B. Zhuravlev, I.S. Nikitin, Analysis of stress state of the contact system "disc-blad". Comput. Contin. Mech. **4**(2), 5–16 (2011)
4. N.G. Burago, A.B. Zhuravlev, I.S. Nikitin, Models of multiaxial fatigue fracture and service life estimation of structural elements. Mech. Solids **46**(6) (2011)
5. B. Crossland, Effect of large hydrostatic pressures on the torsional fatigue strength of an alloy steel. In *Proceedings of the International Conference on Fatigue of Metals* (Institution of Mechanical Engineers, London, 1956), pp. 138–149
6. A. Fatemi, D.F. Socie, A critical plane approach to multiaxial fatigue damage including out-of-phase loading. Fatigue Fract. Eng. Mater. Struct. **11**(3), 149–165 (1988)
7. W.N. Findley, A theory for the effect of mean stress on fatigue of metals under combined torsion and axial load or bending. J. Eng. Ind. **81**(4), 301–306 (1959)
8. A.K. Marmi, A.M. Habraken, L. Duchene, Multiaxial fatigue damage modelling at macro scale of Ti-6Al-4V alloy. Int. J. Fatigue **31**(11–12), 2031–2040 (2009)
9. I.V. Papadopoulos, P. Davoli, C. Gorla, M. Filippini, A. Bernasconi, A comparative study of multiaxial high-cycle fatigue criteria for metals. Int. J. Fatigue **19**(3), 219–235 (1997)
10. A.A. Shanyavskiy, *Modeling of Metal Fatigue Fracture* (Monografiya, Ufa, 2007) (in Russian)
11. G. Sines, Behavior of metals under complex static and alternating stresses, in *Metal Fatigue*, ed. by G. Sines, J.L. Waisman (McGraw-Hill, New York, 1959), pp. 145–169
12. R.N. Smith, P. Watson, T.H. Topper, A stress-strain parameter for the fatigue of metals. J. Mater. **5**(4), 767–778 (1970)
13. Y.-Y. Wang, W.-X. Yao, Evaluation and comparison of several multiaxial fatigue criteria. Int. J. Fatigue **26**(1), 17–25 (2004)

# Dynamic Analysis for Axially Moving Viscoelastic Poynting–Thomson Beams

**Tytti Saksa and Juha Jeronen**

**Abstract** This paper is concerned with dynamic characteristics of axially moving beams with the standard linear solid type material viscoelasticity. We consider the Poynting–Thomson version of the standard linear solid model and present the dynamic equations for the axially moving viscoelastic beam assuming that out-of-plane displacements are small. Characteristic behaviour of the beam is investigated by a classical dynamic analysis, i.e., we find the eigenvalues with respect to the beam velocity. With the help of this analysis, we determine the type of instability and detect how the behaviour of the beam changes from stable to unstable.

**Keywords** Beams · Viscoelasticity · Stability

**Mathematical Subject Classification:** 35Q74 · 65N25 · 74D05 · 74G55 · 74G60 · 74H55 · 74K10 · 74S20

## 1 Introduction

Stability of axially moving beams has been studied for a long time beginning in the 1970 s when Simpson pointed out that the behaviour of translating beams differs from that of stationary beams [14]. Simpson studied the natural frequencies of the translating beam and found out that the beam undergoes divergence instability at a sufficiently high translation velocity. Stability of axially moving elastic beams has been further studied, e.g., by Wickert and Mote, who presented the equations of motion in a canonical form and the expressions for the critical transport velocities

T. Saksa (✉) · J. Jeronen
Department of Mathematical Information Technology, University of Jyväskylä,
P.O. Box 35 (Agora), FI-40014 Jyväskylä, Finland
e-mail: Tytti.Saksa@jyu.fi

J. Jeronen
e-mail: Juha.Jeronen@jyu.fi



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_9

explicitly [20]. Kong and Parker derived an analytical expression for the natural frequencies of the translating elastic beam having small bending stiffness [4].

Eigenvalues (related to eigenfrequencies or natural frequencies), stability, and critical velocities for axially moving viscoelastic beams were studied by Oh et al. [10] and Lee and Oh [6]. They used the (two-parameter) Kelvin–Voigt model for viscoelasticity and the partial time derivative in the constitutive relations. Mockensturm and Guo [9] suggested that for axially moving materials, one should use the material time derivative in the viscoelastic constitutive relations. The material time derivative has been used in the recent studies for moving viscoelastic materials (see, e.g., [3, 16]). For example, Saksa et al. [11] studied the stability of axially moving viscoelastic Kelvin–Voigt beams and panels with the help of eigenvalues and using the material time derivative. They also introduced a fifth boundary condition for the dynamic equation, which involves spatial derivatives up to the fifth order.

Three-parameter models for linear viscoelasticity have been also applied in models of axially moving beams. The standard linear solid (SLS) model, consisting of two springs and one dashpot, has two variants, both of which are often referred as the SLS model. One of the variants is also known as the Poynting–Thomson model, consisting of a Kelvin–Voigt body and a spring connected in series. The other one is known as the Zener model, consisting of a Maxwell body and a spring connected in parallel. Marynowski and Kapitaniak [8] used the Zener model for modelling viscoelasticity in an axially moving beam with time-dependent tension. They concentrated mainly on bifurcation phenomena of a non-linear model but considered also the stability of the linearized system. They found out that the instability occurs at some critical velocity in a form of flutter and that the critical velocity increases if the damping coefficient characterizing the viscoelasticity is increased. Seddighi and Eipakchi [12] computed natural frequencies and critical speeds for axially moving Euler-Bernoulli and Timoshenko beams using the Zener version of the standard linear solid model for viscoelasticity. In their study, the critical speeds (divergence velocities) were determined by solving the steady-state equations. However, they did not perform dynamic analysis to find out if the divergence instability is the first instability. They reported that viscoelasticity had no effect on the critical speed. In all the above studies with the standard linear solid model, the material time derivative was used in the viscoelastic constitutive relations.

The Poynting–Thomson model has been used for axially moving beams by Wang [18], Wang and Chen [19]. They concentrated on asymptotic stability analysis and steady-state response determination.

Here, we study the stability of axially moving viscoelastic beams using the Poynting–Thomson model and classical dynamic analysis. The eigenvalues are determined with respect to the beam velocity to characterize the behaviour and the possible types of stability. The derivation of the dynamic equations for an axially moving SLS beam has been given in [8, 12, 18, 19]. The derivation method presented in [8] differs from the derivation of the others in the definition of bending moment and, thus, results in different equations. We will follow quite closely the lines of [12, 18, 19]. Since the dynamic analysis has not been performed for this form of equations, we will focus on that. In addition, we will use five boundary conditions for the resulting

dynamic equation with up to fifth order spatial derivatives, whereas in the previous studies only four boundary conditions were used. The equations will be discretized using the finite difference method and numerical results will be presented.

## 2 Axially Moving Viscoelastic Poynting–Thomson Beam

We consider an axially moving viscoelastic beam, travelling at a constant velocity $V_0$ in the positive $x$ direction (Fig. 1). The beam is supported at $x = 0$ and $x = \ell$. The function describing the transverse displacement of the beam is denoted by $w = w(x, t)$. For the standard linear solid, viscoelasticity is characterised by the following stress–strain relation:

$$\Gamma\sigma = \Xi\varepsilon,$$

where

$$\varepsilon = -z\frac{\partial^2 w}{\partial x^2},$$

and

$$\Gamma(\cdot) = a_0(\cdot) + a_1\frac{\mathrm{d}}{\mathrm{d}t}(\cdot), \quad \Xi(\cdot) = b_0(\cdot) + b_1\frac{\mathrm{d}}{\mathrm{d}t}(\cdot), \quad \frac{\mathrm{d}}{\mathrm{d}t}(\cdot) = \frac{\partial}{\partial t}(\cdot) + V_0\frac{\partial}{\partial x}, \quad (1)$$

$\sigma$ is the normal stress due to bending, and $\varepsilon$ is the axial bending strain. In (1), the constants $a_0$, $a_1$, $b_0$ and $b_1$ describe the rheological properties of the standard linear solid. Table 1 shows the parameters $a_i$ and $b_i$ ($i = 0, 1$) in the case of Poynting–Thomson and Zener models.

In this study, we will concentrate on the Poynting–Thomson version of the standard linear solid. The dashpot–spring model for the Poynting–Thomson body is shown in Fig. 2. In the limit $E_1 \to \infty$, we obtain the Kelvin–Voigt body. If we

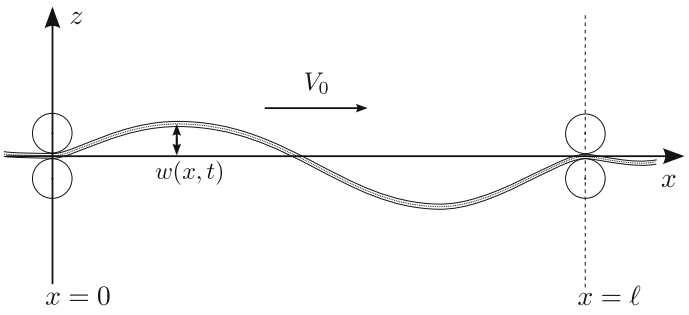



**Fig. 1** A travelling beam

**Table 1** Rheological parameters for Poynting–Thomson and Zener models

| | $a_0$ | $a_1$ | $b_0$ | $b_1$ |
|---|---|---|---|---|
| Poynting–Thomson | $(E_1 + E_2)$ | $\eta$ | $E_1 E_2$ | $E_1 \eta$ |
| Zener | $E_2$ | $\eta$ | $E_1 E_2$ | $(E_1 + E_2)\eta$ |

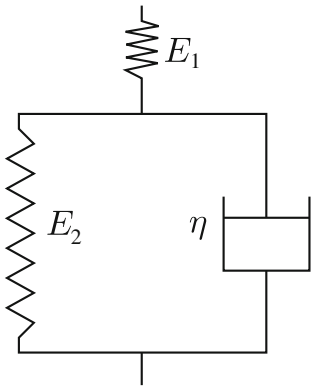



**Fig. 2** Dashpot–spring model for the Poynting–Thomson body

remove the spring described by $E_2$, i.e. we set $E_2 = 0$, then we will have a Maxwell model of viscoelasticity. For the viscosity coefficient $\eta$, we have

$$\eta = t_{\mathrm{R}} E_2, \tag{2}$$

where $t_{\mathrm{R}}$ is the retardation/creep time.

The dynamic equation for the axially moving beam is expressed as (see, e.g., [7, 8, 19])

$$m \frac{\mathrm{d}^2 w}{\mathrm{d}t^2} = \frac{\partial^2 M}{\partial x^2} + T_0 \frac{\partial^2 w}{\partial x^2}, \tag{3}$$

where $m$ is the mass of the beam per unit length, $M$ is the bending moment, and $T_0$ is constant tension applied at the ends. We denote by $\Gamma M$ the equivalent bending moment (see, e.g., [15]) and by $J$ the moment of inertia:

$$\Gamma M = -J \Xi \frac{\partial^2 w}{\partial x^2}, \qquad J = \int_A z^2 \mathrm{d}\, A, \tag{4}$$

where $A$ is the cross-sectional area of the beam. We operate by $\Gamma(\cdot)$ on both sides of (3) and insert (4) assuming sufficient continuity for $M$ to obtain

$$m \Gamma \left( \frac{\mathrm{d}^2 w}{\mathrm{d}t^2} \right) = -J \frac{\partial^2}{\partial x^2} \left( \Xi \frac{\partial^2 w}{\partial x^2} \right) + T_0 \Gamma \left( \frac{\partial^2 w}{\partial x^2} \right). \tag{5}$$

Inserting (1) into (5), we finally have

$$\frac{a_1}{a_0}\left[\frac{\partial^3 w}{\partial t^3}+3V_0\frac{\partial^3 w}{\partial x\partial t^2}+\left(3V_0^2-\frac{T_0}{m}\right)\frac{\partial^3 w}{\partial x^2\partial t}+V_0\left(V_0^2-\frac{T_0}{m}\right)\frac{\partial^3 w}{\partial x^3}\right]+\frac{\partial^2 w}{\partial t^2}$$
$$+2V_0\frac{\partial^2 w}{\partial x\partial t}+\left(V_0^2-\frac{T_0}{m}\right)\frac{\partial^2 w}{\partial x^2}+\frac{Jb_0}{ma_0}\frac{\partial^4 w}{\partial x^4}+\frac{Jb_1}{ma_0}\left(\frac{\partial^5 w}{\partial x^4\partial t}+V_0\frac{\partial^5 w}{\partial x^5}\right)=0. \tag{6}$$

The boundary conditions read

$$w(0,t)=\frac{\partial w}{\partial x}(0,t)=\frac{\partial^2 w}{\partial x^2}(0,t)=0,\qquad w(\ell,t)=\frac{\partial w}{\partial x}(\ell,t)=0. \tag{7}$$

In the derivation of (7), we assume continuity of the equivalent bending moment $\Gamma M$ instead of the actual bending moment. The derivation of the boundary conditions (7) in the case of a Kelvin–Voigt beam (panel) with assumption of the continuity of the bending moment $M$ is given in [11].

The characteristic behaviour of the beam will be studied by inserting the time-harmonic trial function

$$w(x,t)=\exp(st)W(x) \tag{8}$$

into the dynamic equation (6) and the boundary conditions (7). Here, $s=\mathrm{i}\omega$ and $\omega$ is the angular frequency of small transverse vibrations. The complex eigenvalues $s$ characterize the behavior of the system. If real part of $s$ is positive, unstable behavior is encountered, and otherwise the behavior is considered stable.

## 3 Dimensionless Form and Numerical Considerations

For the numerical considerations, we transform the dynamic equation (6) into a dimensionless form. We choose the dimensionless parameters (characterizing the velocity, bending stiffness, relaxation time, and retardation/creep time respectively) as follows:

$$c=\frac{V_0}{\sqrt{T_0/m}},\qquad \alpha=\frac{Jb_0}{a_0T_0\ell^2},\qquad \beta=\frac{a_1\sqrt{T_0}}{a_0\ell\sqrt{m}},\qquad \gamma=\frac{b_1\sqrt{T_0}}{b_0\ell\sqrt{m}}, \tag{9}$$

and the characteristic time is chosen to be

$$\tau=\ell\sqrt{\frac{m}{T_0}}. \tag{10}$$

The $x$-axis is scaled by $\ell$ and the $t$-axis by $\tau$.

With the chosen dimensionless parameters given in (9)–(10) and inserting the standard time-harmonic trial function (8) into (6), we obtain the dimensionless form of the equation

$$s^3\beta W + s^2\left(3c\beta\frac{\partial W}{\partial x} + W\right) + s\left[2c\frac{\partial W}{\partial x} + \beta(3c^2-1)\frac{\partial^2 W}{\partial x^2} + \alpha\gamma\frac{\partial^4 W}{\partial x^4}\right] + \left[(c^2-1)\frac{\partial^2 W}{\partial x^2} + c(c^2-1)\beta\frac{\partial^3 W}{\partial x^3} + \alpha\frac{\partial^4 W}{\partial x^4} + \alpha\gamma c\frac{\partial^5 W}{\partial x^5}\right] = 0. \tag{11}$$

This Eq. (11) generates an eigenvalue problem such that $s$ is the eigenvalue, and the dimensionless axial speed $c$ will be varied. The boundary conditions for the dimensionless function $W$ are

$$W(0) = \frac{\partial W}{\partial x}(0) = \frac{\partial^2 W}{\partial x^2}(0) = 0, \qquad W(1) = \frac{\partial W}{\partial x}(1) = 0.$$

The eigenvalue problem, (11), is reduced to a first order problem:

$$\begin{bmatrix} -M_2 & -M_1 & -M_0 \\ \beta I & 0 & 0 \\ 0 & \beta I & 0 \end{bmatrix} \begin{bmatrix} s^2 W \\ sW \\ W \end{bmatrix} = s\beta \begin{bmatrix} s^2 W \\ sW \\ W \end{bmatrix} \tag{12}$$

in which $s\beta$ are the eigenvalues to be found and the operators $M_i$ are defined by the relations

$$M_0(\cdot) = (c^2-1)\frac{\partial^2}{\partial x^2}(\cdot) + c(c^2-1)\beta\frac{\partial^3}{\partial x^3}(\cdot) + \alpha\frac{\partial^4}{\partial x^4}(\cdot) + \alpha\gamma c\frac{\partial^5}{\partial x^5}(\cdot),$$
$$M_1(\cdot) = 2c\frac{\partial}{\partial x}(\cdot) + \beta(3c^2-1)\frac{\partial^2}{\partial x^2}(\cdot) + \alpha\gamma\frac{\partial^4}{\partial x^4}(\cdot),$$
$$M_2(\cdot) = 3c\beta\frac{\partial}{\partial x}(\cdot) + I(\cdot),$$

where $I$ is the identity operator. Notice that the matrix equation (12) cannot be reduced to the elastic case by setting $\beta = \gamma = 0$ (since the order of the elastic system is lower). However, Eq. (11) is reduced to the elastic case by setting $\beta = \gamma = 0$ and to Kelvin–Voigt by setting $\beta = 0$.

The matrix equation, (12), was discretized via the central finite differences of the second order asymptotic accuracy. The reader will find a systematic discussion of the discretization in [11].

## 4 Numerical Solution

In the numerical examples, we study the dynamic characteristics of the axially moving Poynting–Thomson beam with two different types of materials, both of which are often studied in the context of axially moving materials. For the first material example, we use parameters representing steel [6–8], and for the second one, parameters representing paper [1, 5, 11, 17].

Geometric parameters for the span length $\ell = 1\,\text{m}$ and width $b = 0.2\,\text{m}$ were chosen to be equal for the both studied material examples. Parameters depending on the studied material are given in Table 2. If we refer to a beam made of steel, then we mean that the problem parameters in Table 2 assigned for 'steel' have been used in the computations. Similarly, we refer to a beam made of paper. The creep time, (2), was given the values $t_{\text{R}} = 1 \times 10^{-4}\,\text{s},\ 1 \times 10^{-3}\,\text{s}$. In computations, the parameter $E_1$ was varied. In the finite difference method, we chose the number of computation points on the $x$-axis to be $n = 200$.

The parameters (in Table 2) representing steel were chosen similarly to the numerical studies in [6–8]. In those studies, the maximal value used for the creep time in the computations was $6.8 \times 10^{-4}$ s (lying between our choices $1 \times 10^{-4}\,\text{s}$ and $1 \times 10^{-3}$ s). For paper, we use the same material parameter values as was used, e.g., in [2, 11] except for the elastic Young modulus, for which we use (a more realistic value of) $2 \times 10^9\,\text{N/m}^2$ instead of $10^9\,\text{N/m}^2$. The chosen parameter values represent approximately a paper material with a low basis weight [13, 21].

To compare the dynamic characteristics predicted by the Kelvin–Voigt and Poynting–Thomson models, we present in Fig. 3 the three first eigenvalue pairs for the moving Kelvin–Voigt beam made of steel. In Figs. 4, 5, 6 and 7, the three first eigenvalue pairs are presented for the Poynting–Thomson beam with three different values for $E_1$. Recall that having $E_1 = \infty$, we obtain the Kelvin–Voigt beam case. Notice also that in the discretization for the Kelvin–Voigt beam case, the matrix equation (12) is not applicable. A systematic discussion on the solution process for the Kelvin–Voigt beam case is described in [11].

Comparing Figs. 3, 4, 5 and 6, we see that the bigger the value of $E_1$ is, the closer the results given by the Poynting–Thomson model are qualitatively to the results given by the Kelvin–Voigt model, as expected. In addition, if $E_1 < E_2$, increasing the creep time $t_{\text{R}}$ changes the type of the first instability from divergence to flutter as seen in Figs. 6 and 7.

**Table 2** Material parameters representing a beam made of steel or paper

| | $h$ (m) | $\rho$ (kg/m$^3$) | $A = bh$ (m$^2$) | $m = \rho A$ (kg/m) | $E_2$ (N/m$^2$) | $T_0$ (N) |
|---|---|---|---|---|---|---|
| Steel | 0.0015 | 7800 | $3 \times 10^{-4}$ | 2.34 | $2 \times 10^{11}$ | 2500 |
| Paper | 0.0001 | 800 | $2 \times 10^{-5}$ | 0.016 | $2 \times 10^{9}$ | 100 |

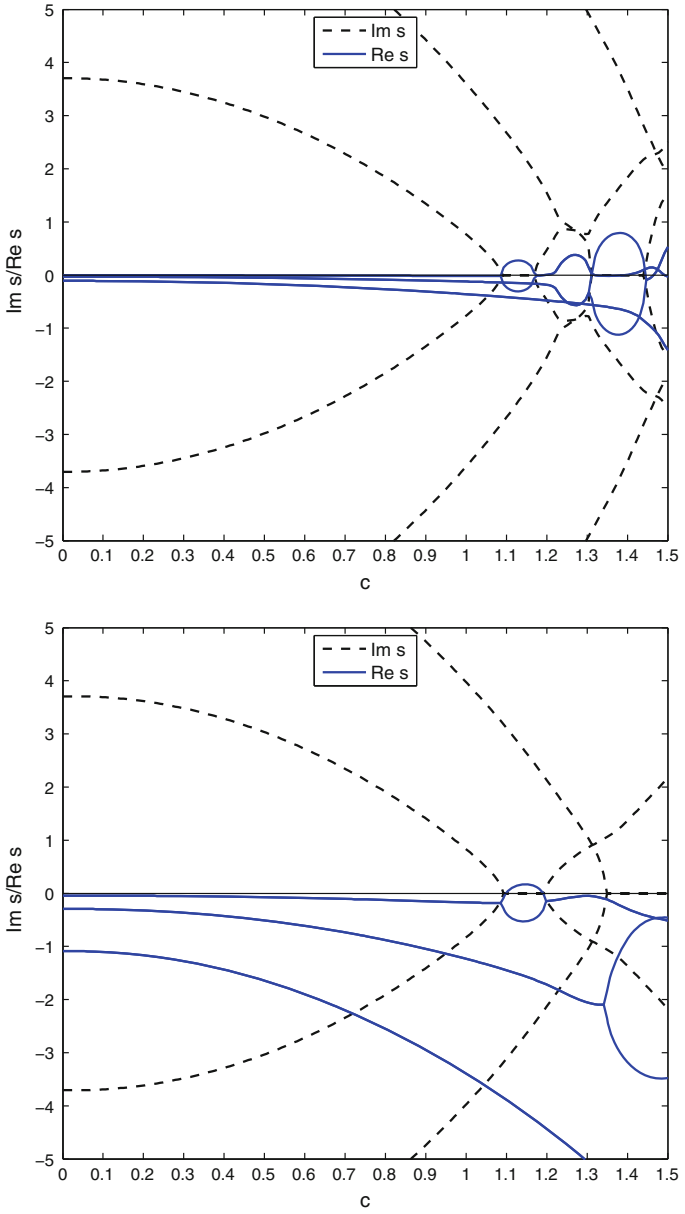



**Fig. 3** Eigenvalues (first three pairs) for an axially moving Kelvin–Voigt beam made of steel. Different numerical solver is applied for this case with "$E_1 = \infty$" in the Poynting–Thomson parameters. The eigenvalues are plotted with respect to the dimensionless beam velocity $c$. *Top* We present the results for the case $t_R = 1 \times 10^{-4}$ s. *Bottom* We present the results for the case $t_R = 1 \times 10^{-3}$ s

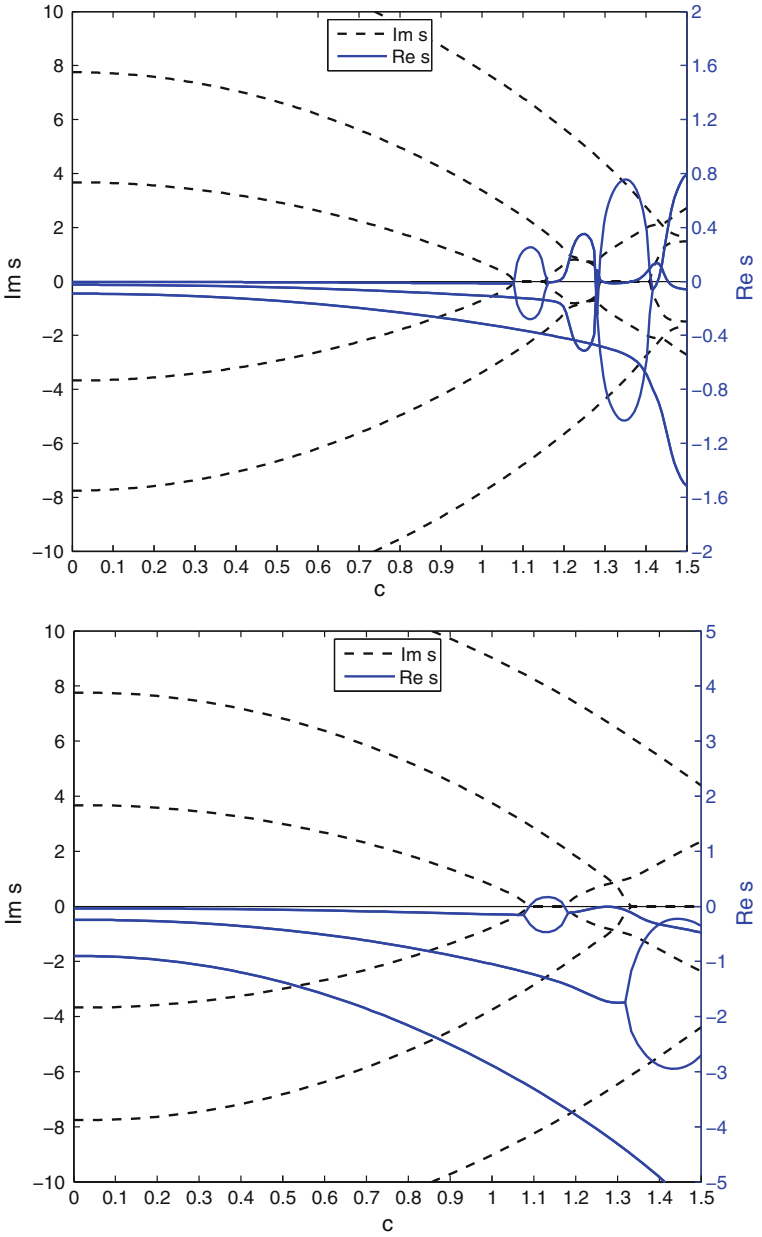



**Fig. 4** Eigenvalues (first three pairs) for an axially moving Poynting–Thomson beam made of steel. The eigenvalues are plotted with respect to the dimensionless beam velocity $c$. Results are presented for the case $E_1 = 10 \times E_2$. Notice that the scaling of the $y$ axis is different for the real and the imaginary parts of $s$. *Top* We present the results for the case $t_R = 1 \times 10^{-4}$ s. *Bottom* We present the results for the case $t_R = 1 \times 10^{-3}$ s

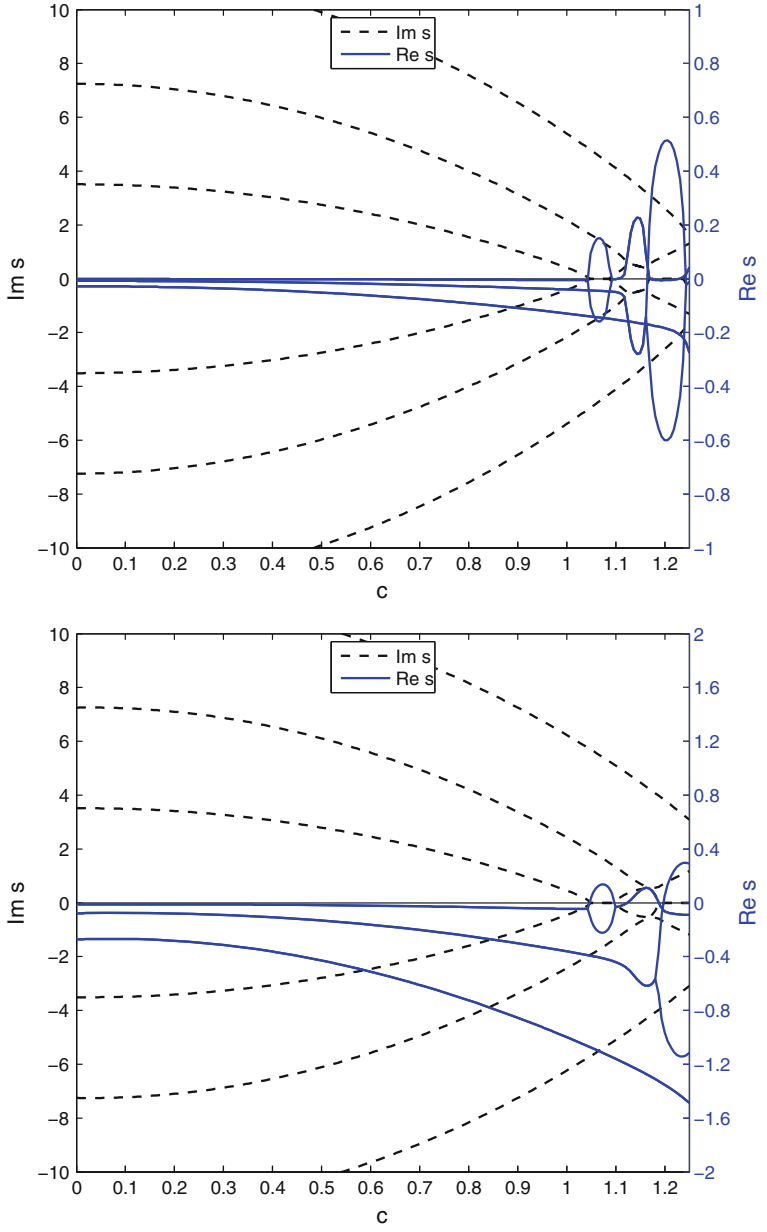



**Fig. 5** Eigenvalues (first three pairs) for an axially moving Poynting–Thomson beam made of steel. The eigenvalues are plotted with respect to the dimensionless beam velocity $c$. Results are presented for the case $E_1 = E_2$. Notice that the scaling of the $y$ axis is different for the real and the imaginary parts of $s$. *Top* We present the results for the case $t_R = 1 \times 10^{-4}$ s. *Bottom* We present the results for the case $t_R = 1 \times 10^{-3}$ s

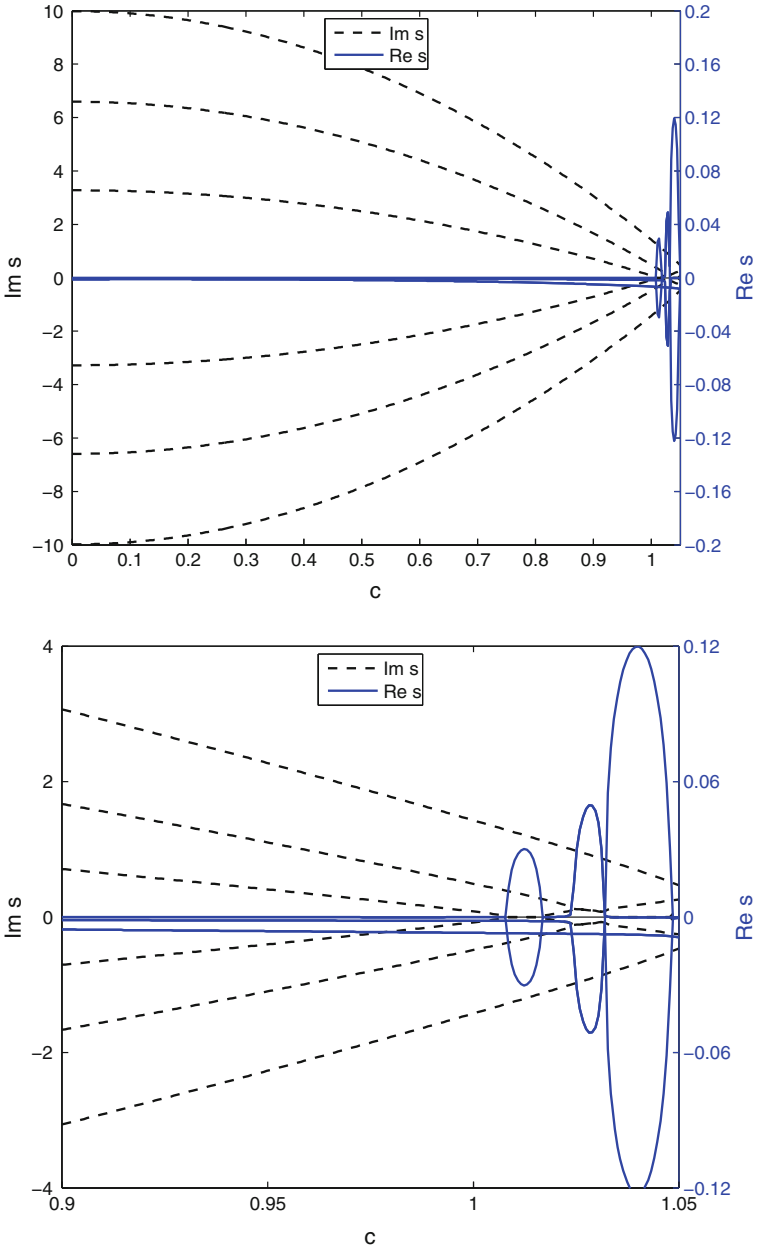



**Fig. 6** Eigenvalues (first three pairs) for an axially moving Poynting–Thomson beam made of steel. The eigenvalues are plotted with respect to the dimensionless beam velocity $c$. Results are presented for the case $E_1 = 0.1 \times E_2$ and $t_R = 1 \times 10^{-4}$ s. Notice that the scaling of the $y$ axis is different for the real and the imaginary parts of $s$. *Top* The value of dimensionless velocity $c$ is between 0 and 1.05. *Bottom* Zoom to the area, where the first instability is detected

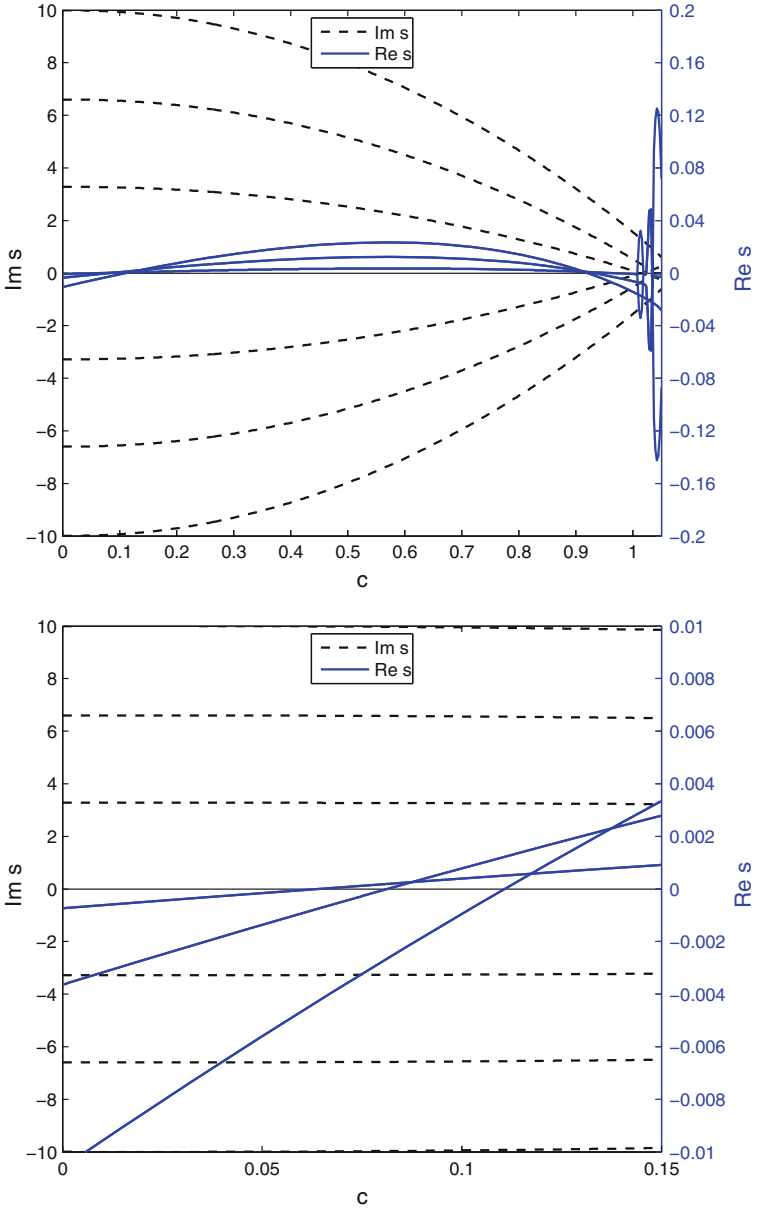



**Fig. 7** Eigenvalues (first three pairs) for an axially moving Poynting–Thomson beam made of steel. The eigenvalues are plotted with respect to the dimensionless beam velocity $c$. Results are presented for the case $E_1 = 0.1 \times E_2$ and $t_R = 1 \times 10^{-3}$ s. Notice that the scaling of the $y$ axis is different for the real and the imaginary parts of $s$. *Top* The value of dimensionless velocity $c$ is between 0 and 1.05. *Bottom* Zoom to the area, where the first instability is detected

The phenomenon that increasing the creep time may change the type of instability from divergence to flutter has been also encountered in previous studies for viscoelastic beam models. Marynowski and Kapitaniak [7] compared Kelvin–Voigt and Bürgers internal damping for models for axially moving viscoelastic beams. For their four-parameter Bürgers model (which is obtained by adding a viscous damper in series to a Poynting–Thomson body), they obtained that with small values of internal damping the dynamic behaviour of the beam was similar to that of a Kelvin–Voigt beam, and for larger values of internal damping, one obtains flutter instability as a first type of instability.

In their later research [8], Marynowski and Kapitaniak continued on Zener internal damping in the context of axially moving beams. As mentioned in the introduction, their definition for the bending moment differed from that of ours and some other authors, so that the dynamic equation was also a bit of different form (also otherwise than the constants that are defined differently for the Zener and Poynting–Thomson models). Their version of the Zener model predicted a flutter type of instability as the first instability.

Comparing the two cases in Figs. 6 and 7, we see that the critical flutter velocity for a large value of creep time ($t_R = 1 \times 10^{-3}$ s, Fig. 7) is drastically lower than the critical divergence velocity for a lower value of creep time ($t_R = 1 \times 10^{-4}$ s, Fig. 6). Also the flutter velocity of Poynting–Thomson beam is lower than the divergence velocity of the Kelvin–Voigt beam. Marynowski and Kapitaniak [8] found also that the critical velocities for the non-linear parametrically excited beam predicted by the Bürgers and Zener models are significantly lower than the critical velocity in the case of a Kelvin–Voigt model.

To compare the dynamic behaviour of different types of materials, the three first eigenvalue pairs were computed also for an axially moving beam made of paper (see Table 2). The eigenvalues for the case of paper material are presented in Figs. 8, 9, 10, 11, 12 and 13.

Compared to steel, paper presents material with a small density and a small bending stiffness. For such material, we see from Figs. 8, 9, 10, 11, 12 and 13 that we encounter flutter already with a relatively large ratio of $E_1/E_2$ compared to steel. In the previously studied case of the Kelvin–Voigt model, viscosity introduced to an elastic model had a stabilizing effect [11], which was pronounced in the case of paper material. On contrary, the Poynting–Thomson model introduces an instability region of flutter, which, with respect to the beam velocity, is encountered a lot earlier than the divergence.

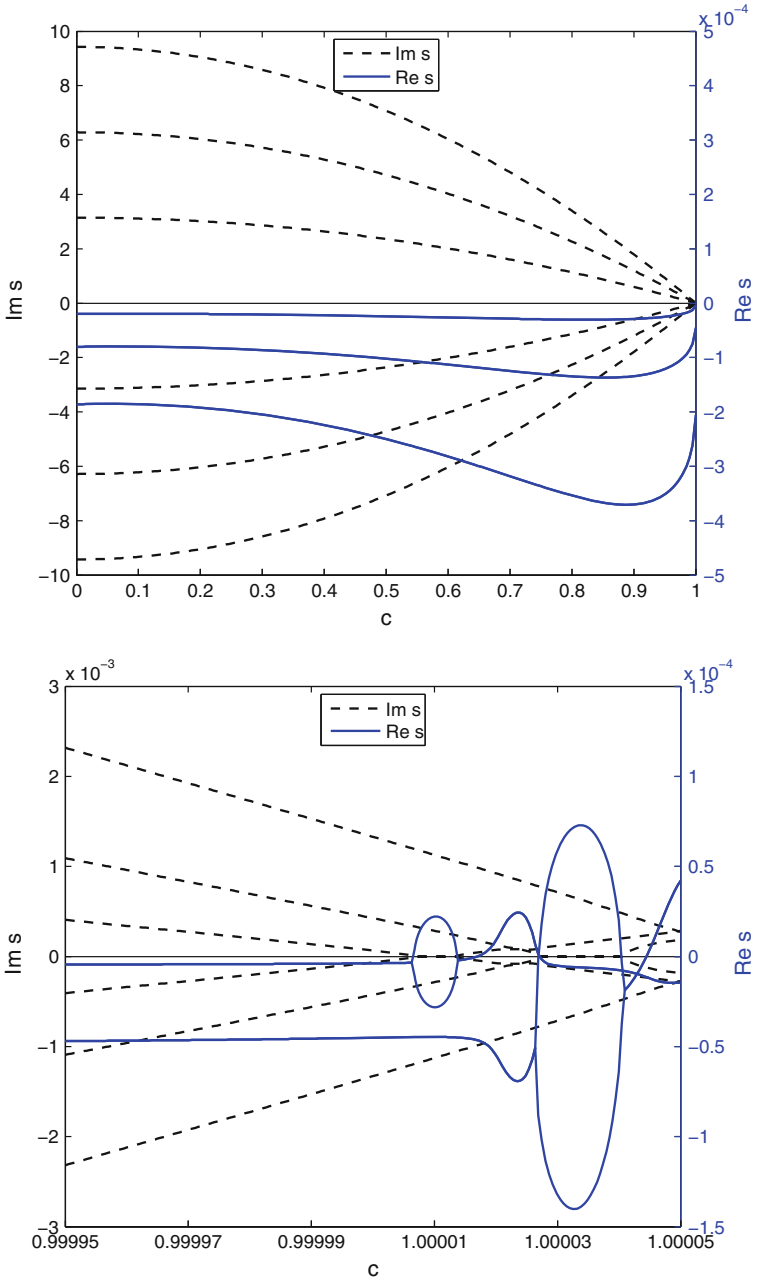



**Fig. 8** Eigenvalues (first three pairs) for an axially moving Poynting–Thomson beam made of paper. The eigenvalues are plotted with respect to the dimensionless beam velocity $c$. Results are presented for the case $E_1 = 100 \times E_2$ and $t_R = 1 \times 10^{-4}$ s. Notice that the scaling of the $y$ axis is different for the real and the imaginary parts of $s$. *Top* The value of dimensionless velocity $c$ is between 0 and 1.0. *Bottom* Zoom to the area, where the first instability is detected

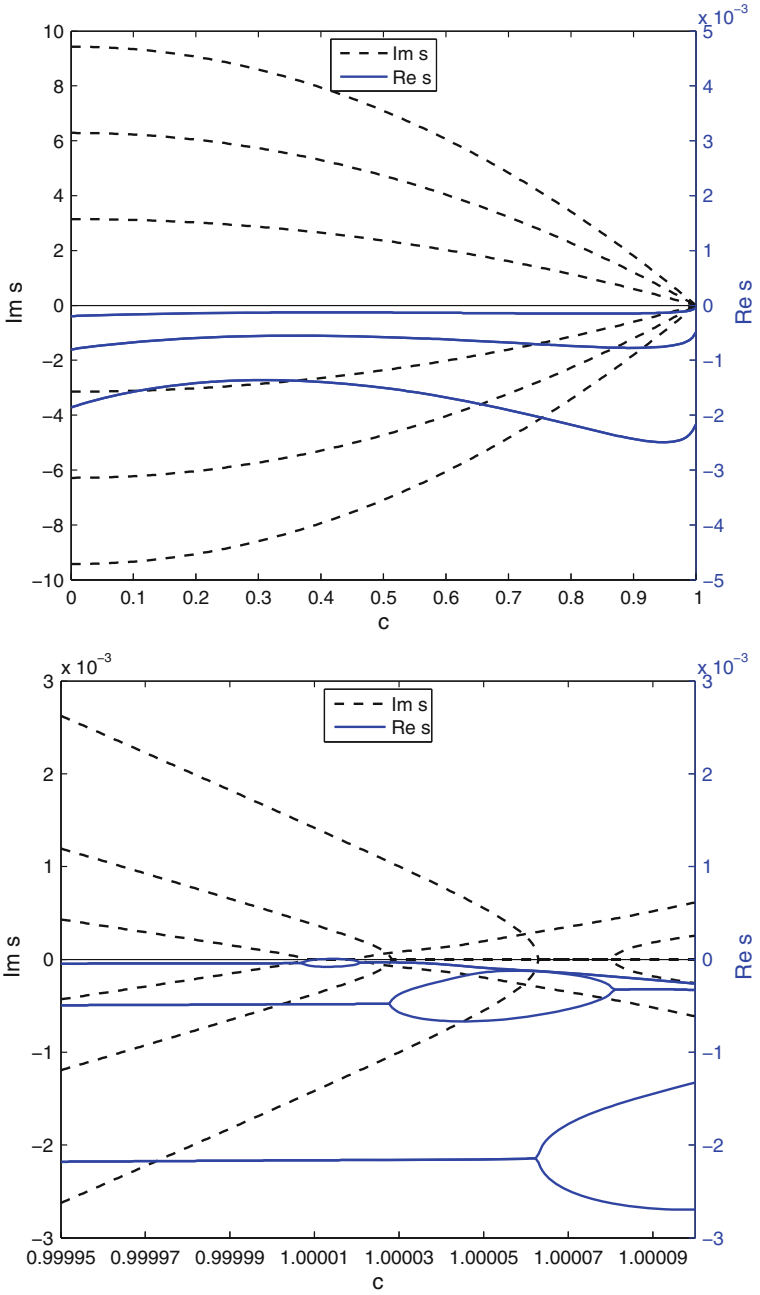



**Fig. 9** Eigenvalues (first three pairs) for an axially moving Poynting–Thomson beam made of paper. The eigenvalues are plotted with respect to the dimensionless beam velocity $c$. Results are presented for the case $E_1 = 100 \times E_2$ and $t_R = 1 \times 10^{-3}$ s. Notice that the scaling of the $y$ axis is different for the real and the imaginary parts of $s$. *Top* The value of dimensionless velocity $c$ is between 0 and 1.0. *Bottom* Zoom to the area, where the first instability is detected

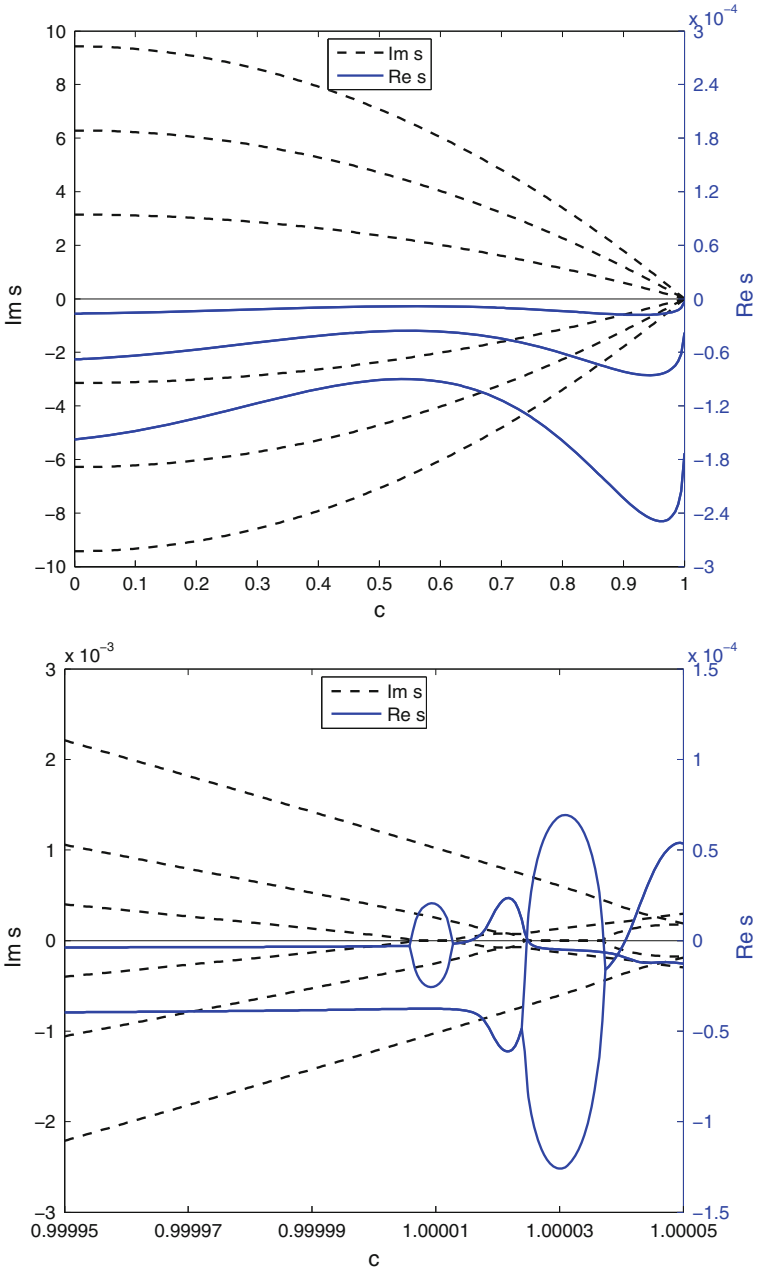



**Fig. 10** Eigenvalues (first three pairs) for an axially moving Poynting–Thomson beam made of paper. The eigenvalues are plotted with respect to the dimensionless beam velocity $c$. Results are presented for the case $E_1 = 10 \times E_2$ and $t_R = 1 \times 10^{-4}$ s. Notice that the scaling of the $y$ axis is different for the real and the imaginary parts of $s$. *Top* The value of dimensionless velocity $c$ is between 0 and 1.0. *Bottom* Zoom to the area, where the first instability is detected

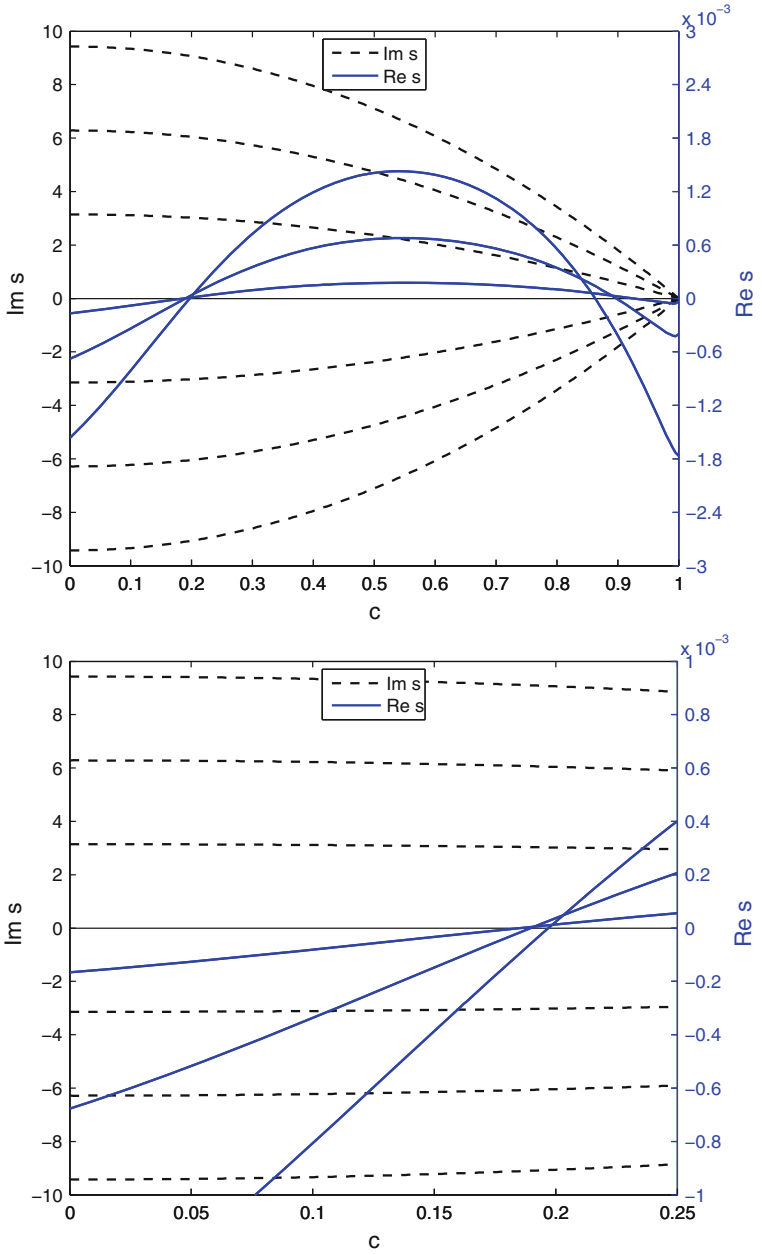



**Fig. 11** Eigenvalues (first three pairs) for an axially moving Poynting–Thomson beam made of paper. The eigenvalues are plotted with respect to the dimensionless beam velocity $c$. Results are presented for the case $E_1 = 10 \times E_2$ and $t_R = 1 \times 10^{-3}$ s. Notice that the scaling of the $y$ axis is different for the real and the imaginary parts of $s$. *Top* The value of dimensionless velocity $c$ is between 0 and 1.0. *Bottom* Zoom to the area, where the first instability is detected

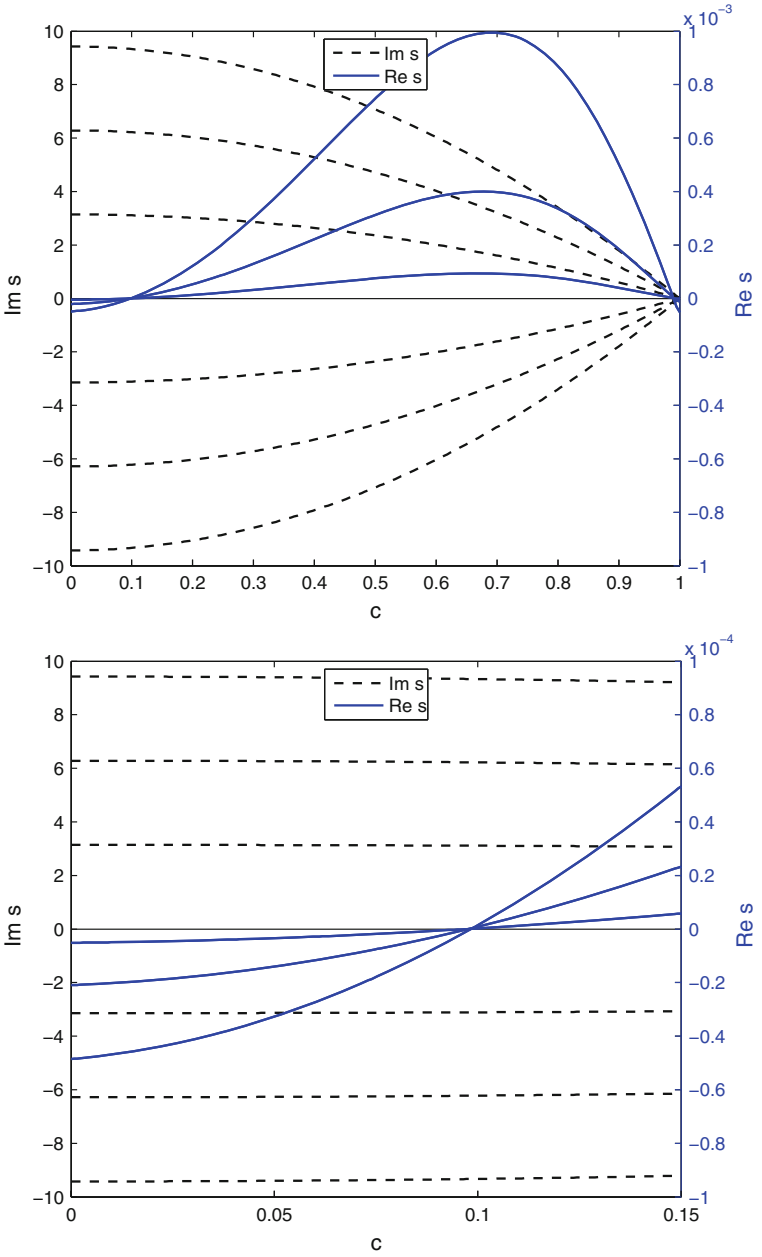



**Fig. 12** Eigenvalues (first three pairs) for an axially moving Poynting–Thomson beam made of paper. The eigenvalues are plotted with respect to the dimensionless beam velocity $c$. Results are presented for the case $E_1 = E_2$ and $t_R = 1 \times 10^{-4}$ s. Notice that the scaling of the $y$ axis is different for the real and the imaginary parts of $s$. *Top* The value of dimensionless velocity $c$ is between 0 and 1.0. *Bottom* Zoom to the area, where the first instability is detected

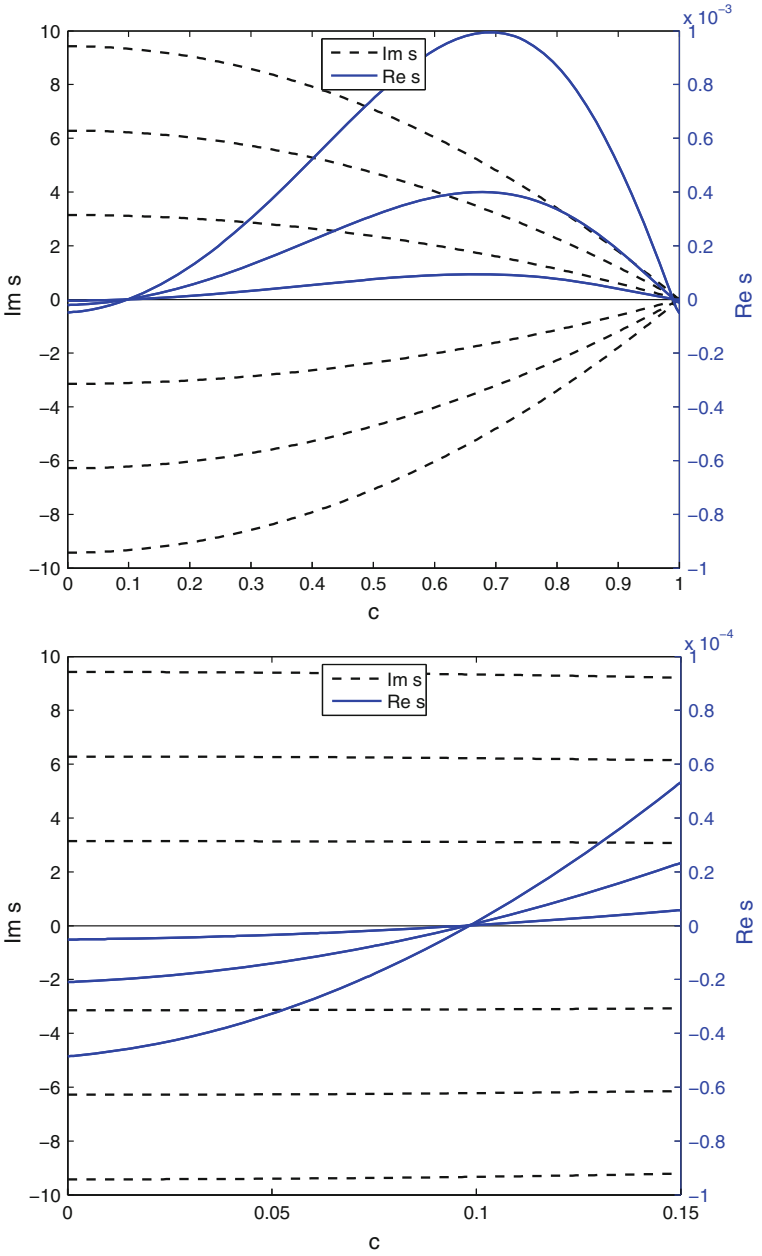



**Fig. 13** Eigenvalues (first three pairs) for an axially moving Poynting–Thomson beam made of paper. The eigenvalues are plotted with respect to the dimensionless beam velocity $c$. Results are presented for the case $E_1 = E_2$ and $t_R = 1 \times 10^{-3}$ s. Notice that the scaling of the $y$ axis is different for the real and the imaginary parts of $s$. *Top* The value of dimensionless velocity $c$ is between 0 and 1.0. *Bottom* Zoom to the area, where the first instability is detected

## 5 Conclusions

In this paper, characteristic behaviour of an axially moving viscoelastic beam was investigated with respect to the beam velocity modelling the viscoelasticity by the Poynting–Thomson version of the standard linear solid. The derivation of the dynamic equation describing the transverse displacement of the beam was outlined briefly. For the resulting equation of the fifth order in space, five boundary conditions were introduced following the derivation in [11].

In the numerical examples, two different types of materials were studied: steel and paper. The former presents material with a relatively high density and a large bending stiffness and the latter is an example of material with a lower density and a very small bending stiffness. In the case of steel-like material, the standard linear solid (Poynting–Thomson version) beam model gave qualitatively similar predictions to the Kelvin–Voigt beam model. We may conclude that, for steel, the Kelvin–Voigt model gave as good predictions of the characteristic behaviour of axially moving viscoelastic beams as the more complicated Poynting–Thomson model.

In the case of paper, the moving beam was found to lose stability in the sense of flutter if the creep time is large enough. Also the ratio of the elastic moduli $E_1$ and $E_2$ was found to significantly affect the predictions of the dynamic behaviour of the beam.

Recall that the Poynting–Thomson body is composed of a Kelvin–Voigt body connected in series with a spring. Thus, we expected the dynamic characteristics of the Poynting–Thomson beam to be qualitatively similar to that of the Kelvin–Voigt beam, if the elastic modulus $E_1$ related to the spring becomes large enough. For both steel and paper, the numerical results of the dynamic analysis were in agreement with this expectancy. Decreasing the ratio $E_1/E_2$ changed the predictions of the type of the first instability from divergence to flutter. The first critical flutter velocity was drastically lower than the first divergence velocity. This effect was more pronounced in the case of material with a relatively low density and a very small bending stiffness.

**Acknowledgments** This research was supported by the Jenny and Antti Wihuri Foundation and the Finnish Cultural Foundation.

## References

1. N. Banichuk, J. Jeronen, M. Kurki, P. Neittaanmäki, T. Saksa, T. Tuovinen, On the limit velocity and buckling phenomena of axially moving orthotropic membranes and plates. Int. J. Solids Struct. **48**(13), 2015–2025 (2011). doi:10.1016/j.ijsolstr.2011.03.010
2. N. Banichuk, J. Jeronen, P. Neittaanmäki, T. Tuovinen, Dynamic behaviour of an axially moving plate undergoing small cylindrical deformation submerged in axially flowing ideal fluid. J. Fluids Struct. **27**(7), 986–1005 (2011). doi:10.1016/j.jfluidstructs.2011.07.004
3. M.H. Ghayesh, M. Amabili, H. Farokhi, Coupled global dynamics of an axially moving viscoelastic beam. Int. J. Non-Linear Mech. **51**, 54–74 (2013). doi:10.1016/j.ijnonlinmec.2012.12.008

4. L. Kong, R.G. Parker, Approximate eigensolutions of axially moving beams with small flexural stiffness. J. Sound Vib. **276**(1–2), 459–469 (2004)
5. A. Kulachenko, P. Gradin, H. Koivurova, Modelling the dynamical behaviour of a paper web. Part I. Comput. Struct. **85**(3–4), 131–147 (2007)
6. U. Lee, H. Oh, Dynamics of an axially moving viscoelastic beam subject to axial tension. Int. J. Solids Struct. **42**(8), 2381–2398 (2005). doi:10.1016/j.ijsolstr.2004.09.026
7. K. Marynowski, T. Kapitaniak, Kelvin-Voigt versus Bürgers internal damping in modeling of axially moving viscoelastic web. Int. J. Non-Linear Mech. **37**(7), 1147–1161 (2002). doi:10.1016/S0020-7462(01)00142-1
8. K. Marynowski, T. Kapitaniak, Zener internal damping in modelling of axially moving viscoelastic beam with time-dependent tension. Int. J. Non-Linear Mech. **42**(1), 118–131 (2007). doi:10.1016/j.ijnonlinmec.2006.09.006
9. E.M. Mockensturm, J. Guo, Nonlinear vibration of parametrically excited, viscoelastic, axially moving strings. J. Appl. Mech. **72**(3), 374–380 (2005). doi:10.1115/1.1827248
10. H. Oh, J. Cho, U. Lee, Spectral element analysis for an axially moving viscoelastic beam. J. Mech. Sci. Tech. **18**(7), 1159–1168 (2004). doi:10.1007/BF02983290
11. T. Saksa, N. Banichuk, J. Jeronen, M. Kurki, T. Tuovinen, Dynamic analysis for axially moving viscoelastic panels. Int. J. Solids Struct. **49**(23–24), 3355–3366 (2012). doi:10.1016/j.ijsolstr.2012.07.017
12. H. Seddighi, H. Eipakchi, Natural frequency and critical speed determination of an axially moving viscoelastic beam. Mech. Time-Depend. Mater. **17**(4), 529–541 (2013). doi:10.1007/s11043-012-9201-1
13. Y.B. Seo, Determination of in-plane shear properties by an off-axis tension method and laser speckle photography. J. Pulp Paper Sci. **25**(9), 321–325 (1999)
14. A. Simpson, Transverse modes and frequencies of beams translating between fixed end supports. J. Mech. Eng. Sci. **15**(3), 159–164 (1973)
15. Z. Sobotka, *Rheology of Materials and Engineering Structures* (Elsevier, Amsterdam, 1984)
16. Y.-Q. Tang, L.-Q. Chen, Stability analysis and numerical confirmation in parametric resonance of axially moving viscoelastic plates with time-dependent speed. Eur. J. Mech. A Solids **37**, 106–121 (2013). doi:10.1016/j.euromechsol.2012.05.010
17. M. Vaughan, A. Raman, Aeroelastic stability of axially moving webs coupled to incompressible flows. J. Appl. Mech. **77**(2), 021001–021001–17 (2010). doi:10.1115/1.2910902
18. B. Wang, Asymptotic analysis on weakly forced vibration of axially moving viscoelastic beam constituted by standard linear solid model. Appl. Math. Mech. **33**(6), 817–828 (2012). doi:10.1007/s10483-012-1588-8
19. B. Wang, L.-Q. Chen, Asymptotic stability analysis with numerical confirmation of an axially accelerating beam constituted by the standard linear solid model. J. Sound Vib. **328**(4–5), 456–466 (2009). doi:10.1016/j.jsv.2009.08.016
20. J.A. Wickert, C.D. Mote, Classical vibration analysis of axially moving continua. J. Appl. Mech. **57**(3), 738–744 (1990). doi:10.1115/1.2897085
21. T. Yokoyama, K. Nakai, Evaluation of in-plane orthotropic elastic constants of paper and paperboard, in *SEM Annual Conference and Exposition on Experimental and Applied Mechanics 2007 (Springfield, MA)*, pp. 1505–1511, Bethel, CT. Society for Experimental Mechanics (2007)

# A Projection Approach to Analysis of Natural Vibrations for Beams with Non-symmetric Cross Sections

**Vasily Saurin and Georgy Kostin**

**Abstract** A projection approach based on the method of integro-differential relations and semi-discretization technique is applied to analyze natural variations of rectilinear elastic beams with non-symmetric cross sections. A numerical algorithm is proposed to compose compatible approximating systems of ordinary differential equations. It is shown that the beam vibrations cannot be separated into four independent types of longitudinal, bending, and torsional motions if a non-symmetric cross section is considered. In this case, all these motions can interact with one another. Nevertheless, only one type of displacement and stress fields makes the largest contribution in the amplitudes of the corresponding vibrations. Several eigenfrequencies and eigenforms of a beam with the isosceles cross section are presented and analyzed.

**Keywords** Natural beam vibrations · Eigenvalue problem · Theory of elasticity · Projection approach · Semi-discretization

**Mathematical Subject Classification:** 65M60 · 74K10 · 74B05 · 74H45

## 1 Introduction

In applications, the conventional classes of beam-type flexible bodies are prismatic rods, shafts, pipes, etc. To construct a reliable beam model in the frame of the linear theory of elasticity, it is important to consider the spatial distribution of displacement and stress fields. Initial-boundary or eigenvalue problems of beam motions may possess symmetry properties which allow us to reduce the dimension of the original system and effectively apply novel numerical approaches.

V. Saurin (✉) · G. Kostin
Institute for Problems in Mechanics RAS, Vernadskogo 101-1, 119526 Moscow, Russia
e-mail: saurin@ipmnet.ru

G. Kostin
e-mail: kostin@ipmnet.ru



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_10

Among simplified models proposed for approximate solution of spatial dynamic problems in elasticity, a special place takes the theory of Euler–Bernoulli beams based on hypotheses put forward by Bernoulli [1]. Despite the fact that this theory is applicable for a wide class of problems, it does not consider such effects of stress-strain state in some elastic beam as shear displacements, warping, deformation of the cross section, interaction of longitudinal and lateral motions depending on Poisson's ratio, etc.

Advanced formulae for longitudinal and bending motions of elastic beams were proposed allowing to take into account compression–tension in the cross section (Rayleigh's correction [7]) as well as its rotations and shears (Timoshenko beam [8]). In the classical model of beam torsion discussed, for example, in [9], the cross-sectional warping is taken into account, which is found as a solution of Poisson's plane problem. In the model proposed by Reissner [6], a variational approach combined with a pre-defined displacement distribution in the lateral direction is used to derive the equations describing the elastic bending of a thin plate (beam). Variational formulations are also utilized to obtain compatible beam equations of higher order by taking into account in different ways the spatial distribution of displacements in a lengthy elastic body [5].

The presented paper is devoted to an approach in which the original problem in partial differential equations (PDEs) is approximated by a system of ordinary differential equation (ODEs) based on the method of integro-differential relations (MIDR) [4]. The projection technique developed in [2] was applied to 3D linear elasticity problems with semi-discrete approximations for the displacement vector and the stress tensor. The approximations include a polynomial expansion of finite dimension over some coordinate components and unknown functions over one remaining component [3].

As compared to the classical Galerkin method, the variational approach described in other studies of the authors [4] possesses some advantages: it guarantees optimality properties of the numerical solution for some approximations, provides explicit error estimates, and supposes the exact implementation of the initial, boundary, and momentum balance conditions. Some drawbacks of the variational method are doubling the dimension of the resulting ODE system and worsening the numerical stability of the results. The projection approach based on the MIDR allows us to reduce the ODE dimension and to improve the numerical stability in such a way that all the advantages mentioned about the variational technique still hold.

In this study, free natural vibrations are analyzed for an elastic body (beam) shaped like a right prism with an isosceles triangular cross section. The choice of the object under study is stipulated by the following circumstances. First of all, such an elastic prism contains specific features which are typical for linear elasticity problems associated with various types of boundary conditions and with the presence of corner points. Secondly, it allows to use polynomial representations of stresses and displacements to approximate an elastic state of the prism. Another reason is that such a beam element can be used to compose structures with more complicated cross-sectional shapes (e.g., thin-walled beams) and applied to specific FEM algorithms.

The paper is structured as follows: In Sect. 2, a 3D eigenvalue problem of free vibrations is considered for some elastic body occupying a prismatic domain with the triangular cross section. A semi-discretization procedure is described in Sect. 3 to approximate displacements and stresses by means of polynomial trial functions with respect to the cross-sectional coordinates. In Sect. 4, a projection algorithm is presented to compose approximating test systems of ODEs. The results of numerical analysis for an elastic beam with the isosceles cross section are discussed in Sect. 5. Finally, brief conclusions are given.

## 2 Modelling of Elastic Beam Dynamics

Consider a long rectilinear prismatic body (beam) with a triangular cross section as in Fig. 1. The origin of the Cartesian coordinate system is placed at the barycenter of one prism base. The axis $x$ is directed to the other base along the beam length, and, hence, the axes $y$ and $z$ are parallel to the body cross sections. It is assumed that the beam is made of homogeneous isotropic material with the volume density $\rho$, Young's modulus $E$, and Poisson's ratio $\nu$.

The elastic vibrations of the beam are described by the following PDE system

$$\varepsilon - \mathsf{C}^{-1} : \sigma = 0 \,, \quad \nabla \cdot \sigma + \rho\omega^2 \mathbf{u} = \mathbf{0} \,. \tag{1}$$

Here, $\mathsf{C}$ is the elastic moduli tensor, $\sigma$ is the stress tensor with the components $\sigma_x$, $\sigma_y$, $\sigma_z$, $\tau_{xy}$, $\tau_{xz}$, and $\tau_{yz}$. The components of the displacement vector $\mathbf{u}$ denote as $u$, $v$, $w$; $\omega$ is the frequency of natural vibrations. The Cauchy strain tensor $\varepsilon$ has the components

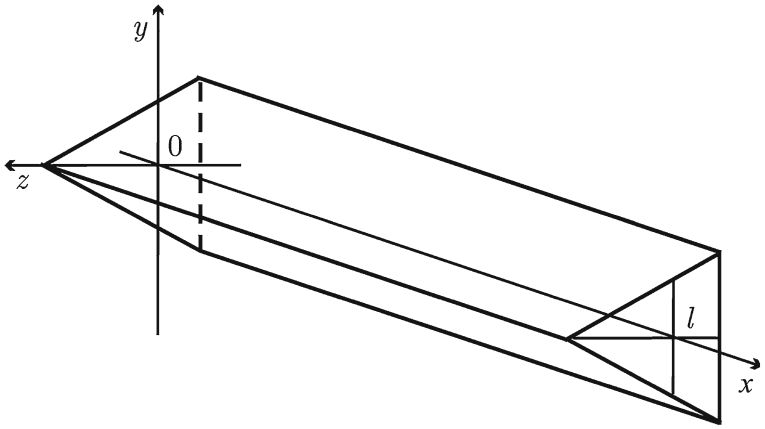



**Fig. 1** Prismatic beam with a triangular cross section

$$\begin{aligned} \varepsilon_x &= \frac{\partial u}{\partial x}\,, \quad \varepsilon_y = \frac{\partial v}{\partial y}\,, \quad \varepsilon_z = \frac{\partial w}{\partial z}\,, \quad \varepsilon_{xy} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right), \\ \varepsilon_{xz} &= \frac{1}{2}\left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}\right), \quad \varepsilon_{yz} = \frac{1}{2}\left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y}\right). \end{aligned} \tag{2}$$

The boundary constraints can be divided into two groups. The conditions on the lateral sides of the beam can be attributed to the first part. In frame of the projection approach discussed in this paper, these relations have to be satisfied before constituting an approximating ODE system. In contrast, the second group consisting of the conditions on the prism bases is implemented together with the system of ODEs.

Let us represent first the boundary conditions which are referred to as the first group. To be more particular, only a free beam is studied here. This means that no displacements are determined on the prism faces. The boundary conditions in the stresses defined on the lateral sides of the beam can be divided in turn into two subgroups due to the fact that these sides are parallel to the $x$-axis. The equations which relate to the shear stresses $\tau_{xy}$ and $\tau_{xz}$ have the form

$$\tau_{xy} n_y^{(i)} + \tau_{xz} n_z^{(i)} = 0\,. \tag{3}$$

The other relations combine the components $\sigma_y$, $\sigma_z$, and $\tau_{yz}$:

$$\sigma_y n_y^{(i)} + \tau_{yz} n_z^{(i)} = 0 \quad \text{and} \quad \tau_{yz} n_y^{(i)} + \sigma_z n_z^{(i)} = 0 \tag{4}$$

for the prism faces $\Gamma_i$ with $i = 1, 2, 3$. Here $n_x^{(i)} = 0$, $n_y^{(i)}$, $n_z^{(i)}$ are the components of the normal vectors $\mathbf{n}^{(i)}$ to $\Gamma_i$.

The following boundary conditions are given on the bases of the prism at $x = 0$ and $x = l$ (the second group):

$$\begin{aligned} \sigma_x(0, y, z) &= \tau_{xy}(0, y, z) = \tau_{xz}(0, y, z) = 0\,, \\ \sigma_x(l, y, z) &= \tau_{xy}(l, y, z) = \tau_{xz}(l, y, z) = 0\,, \end{aligned} \tag{5}$$

where $l$ is the length of the beam.

## 3 Semi-discretization of Displacement and Stress Fields

In accordance with the semi-discretization method, the unknown trial functions $\mathbf{u}(x, y, z)$ and $\sigma(x, y, z)$ are found as complete polynomials with respect to the beam lateral coordinates $y$ and $z$. Approximations of the displacement and stress fields are taken in the form

$$
\begin{aligned}
u &= \sum_{i+j+k=N_1} u^{(ijk)}(x)\,p_{ijk}\,, \quad v = \sum_{i+j+k=N_2} v^{(ijk)}(x)\,p_{ijk}\,,\\
w &= \sum_{i+j+k=N_3} w^{(ijk)}(x)\,p_{ijk}\,, \quad \sigma_x = \sum_{i+j+k=N_4} \sigma_x^{(ijk)}(x)\,p_{ijk}\,,\\
\tau_{xy} &= \sum_{i+j+k=N_5} \tau_{xy}^{(ijk)}(x)\,p_{ijk}\,, \quad \tau_{xz} = \sum_{i+j+k=N_6} \tau_{xz}^{(ijk)}(x)\,p_{ijk}\,,\\
\sigma_y &= \sum_{i+j+k=N_7} \sigma_y^{(ijk)}(x)\,p_{ijk}\,, \quad \sigma_z = \sum_{i+j+k=N_8} \sigma_z^{(ijk)}(x)\,p_{ijk}\,,\\
\tau_{yz} &= \sum_{i+j+k=N_9} \tau_{yz}^{(ijk)}(x)\,p_{ijk}\,, \quad p_{ijk} = [g_1(y,z)]^i\,[g_2(y,z)]^j\,[g_3(y,z)]^k\,.
\end{aligned}
\tag{6}
$$

The choice of the numbers $N_i$, $i = 1, \ldots, 9$, in Eq. (6) is discussed in Sect. 4. The following system of linear basis functions defined by the coordinates $y$ and $z$ of the triangular cross section is introduced in Eq. (6) as

$$
\begin{aligned}
g_1 &= \frac{z_2 - z_3}{2S}(y - y_2) - \frac{y_2 - y_3}{2S}(z - z_2)\,,\\
g_2 &= \frac{z_3 - z_1}{2S}(y - y_3) - \frac{y_3 - y_1}{2S}(z - z_3)\,,\\
g_3 &= \frac{z_1 - z_2}{2S}(y - y_1) - \frac{y_1 - y_2}{2S}(z - z_1)\,.
\end{aligned}
\tag{7}
$$

Here, $y_i$ and $z_i$ for $i = 1, 2, 3$ are the coordinates of the triangle vertices and

$$
2S = y_1 z_2 + y_2 z_3 + y_3 z_1 - y_2 z_1 - y_3 z_2 - y_1 z_3 \tag{8}
$$

is the doubled area of the cross section. In literature [10], the functions $g_i, i = 1, 2, 3$, given in Eq. (7) are referred to as barycentric coordinates or $L$ -coordinates, but it is worth noting that they are linear functions ($g$-functions in what follows), which have a simple geometric interpretation. The function $g_1(x)$ is equal to zero at the triangle edge defined by the coordinates $(y_2, z_2)$, $(y_3, z_3)$ and reaches its maximum value in the triangle at the vertex with the coordinates $(y_1, z_1)$. The same properties are inherent to the other functions $g_2$ and $g_3$ with the only difference in the permutation of indices: $1 \to 2$, $2 \to 3$, $3 \to 1$. Additionally, the sum of all these functions is equal to unity:

$$
g_1(y, z) + g_2(y, z) + g_3(y, z) \equiv 1\,.
$$

To integrate expressions depending on the $g$-functions over the triangular element, it is useful to know the following analytical formula

$$
\int_S (g_1)^i (g_2)^j (g_3)^k\, dS = \frac{i!\,j!\,k!}{(i + j + k + 2)!} 2S\,, \tag{9}
$$

where the section area $S$ is defined by Eq. (8).

The basic idea in solving the problem (1)–(5) is to apply the approximations (6) and the projection approach discussed in [3]. These approximations must satisfy exactly the boundary conditions (3)–(5). After that, an ODE system with respect to the unknown coefficients introduced in Eq. (6) is composed through the components

$$\begin{aligned} r_x &= \frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{xz}}{\partial z} + \rho\omega^2 u \,, \\ r_y &= \frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_y}{\partial y} + \frac{\partial \tau_{yz}}{\partial z} + \rho\omega^2 v \,, \\ r_z &= \frac{\partial \tau_{xz}}{\partial x} + \frac{\partial \tau_{yz}}{\partial y} + \frac{\partial \sigma_z}{\partial z} + \rho\omega^2 w \end{aligned} \tag{10}$$

of the equilibrium vector $\mathbf{r}$ and the components

$$\begin{aligned} \xi_x &= \frac{\partial u}{\partial x} - \frac{\sigma_x}{E} + \frac{\nu}{E}\left(\sigma_y + \sigma_z\right) \,, \quad & \xi_{yz} &= \frac{1}{2}\left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y}\right) - \frac{\tau_{yz}}{2G} \,, \\ \xi_y &= \frac{\partial v}{\partial y} - \frac{\sigma_y}{E} + \frac{\nu}{E}\left(\sigma_x + \sigma_z\right) \,, \quad & \xi_z &= \frac{\partial w}{\partial z} - \frac{\sigma_z}{E} + \frac{\nu}{E}\left(\sigma_x + \sigma_y\right) \,, \\ \xi_{xy} &= \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) - \frac{\tau_{xy}}{2G} \,, \quad & \xi_{xz} &= \frac{1}{2}\left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}\right) - \frac{\tau_{xz}}{2G} \end{aligned} \tag{11}$$

of Hooke's tensor $\xi$, where the shear modulus $G = E/(2+2\nu)$ is introduced. By using the notation introduced in Eqs. (10) and (11), the constitutive relations (1) can be rewritten in the compact form

$$\xi = 0 \,, \quad \mathbf{r} = \mathbf{0} \,. \tag{12}$$

Here, the tensor $\xi$ reflects Hooke's law, i.e. the linear dependence of the Cauchy strain tensor on the stress one. In turn, the equilibrium vector $\mathbf{r}$ characterizes the relation between the stresses and the momentum density. The second equation in (12) describes the momentum balance. All the components of $\xi$ and $\mathbf{r}$ become zero (either strongly or weakly) on the exact solution. If the ansatz functions (6) are used then the nonzero values of these components define the quality of the corresponding approximate solution.

In the proposed approach, the polynomial approximations (6) of degrees $N_m$ for $m = 1, \ldots, 9$ have to possess the following properties. Firstly, it is necessary that such approximations are able to satisfy exactly the boundary conditions (3) and (4). Secondly, it is important to select correctly a specific space of test functions (13) corresponding to each component of the constitutive vector-function

$$\left\{r_x,\ r_y,\ r_z,\ \xi_x,\ \xi_y,\ \xi_z,\ \xi_{xy},\ \xi_{xz},\ \xi_{yz}\right\} \,,$$

which is composed of the relations (10) and (11). This choice must guarantee that the system of differential-algebraic equations (DAEs), which results from the corresponding projections of the components on these polynomial spaces, is consistent.

Note that the choice of polynomial approximations (6) designates the way to determine a subspace of test functions for which the integral projections should be calculated. The space of complete polynomials of degree $K_m > 0$ ($m = 1, \ldots, 9$)

$$P_{K_m} = \left\{ \sum_{i+j+k=K_m} c_{ijkm}\, p_{ijk}(y, z)\ , \quad c_{ijkm} \in \mathbb{R} \right\} \tag{13}$$

with the monomials $p_{ijk}$ defined in Eq. (6) is suitable for this purpose.

It is also desirable in a numerical algorithm that the structure of these DAE systems does not change with the approximation order. The rather complicated form of the constitutive and equilibrium relations (1) as well as the boundary conditions (3), (4) implies that, in general case, the parameters $N_i$ and $K_i$ may differ from one another.

After implementation of the boundary conditions (3) and (4), the equilibrium equations of Eq. (1) contain $\tilde{N}_d$ independent stresses $\{\sigma_x^{(ijk)}(x),\ \tau_{xy}^{(ijk)}(x),\ \tau_{xz}^{(ijk)}(x)\}$ and, in accordance with Eq. (10), their derivatives with respect to the spatial coordinate $x$. Similarly, $\tilde{N}_u$ independent derivatives of $\{u^{(ijk)}(x), v^{(ijk)}(x), w^{(ijk)}(x)\}$ are included in the components of Hooke's tensor (11). It is possible to show that the total differential order of the DAEs is equal to $\tilde{N}_d + \tilde{N}_u$ and the following inequality $\tilde{N}_d \neq \tilde{N}_u$ is valid. It is worth noting that only the above-mentioned stress functions can be used to satisfy the boundary constraints (5) ($2\tilde{N}_d$ conditions altogether). Therefore, the DAE system would be consistent only if $\tilde{N}_d = \tilde{N}_u$. To improve the system a certain number of the stress or displacement functions should be eliminate. The maximal differential order $N_d$ of a desirable system is chosen according to

$$N_d = 2 \min \left\{ \tilde{N}_d, \tilde{N}_u \right\} .$$

This condition brings some complexity in the composing of such a system. Nevertheless these difficulties can be eliminated, as it is shown below, by choosing appropriate displacements and stresses as well as corresponding projections of the constitutive relations.

## 4 Integral Projections in the Eigenvalue Problem

The choice of trial and test spaces is no unique procedure as shown, for example, in [4]. Let us constrain ourselves to the case when the test spaces $P_m$ defined in Eq. (13) are identical to each other, that is $K_m = N_0$ for all $m = 1, \ldots, 9$. Here $N_0$ is a positive integer. This number is simultaneously the degree of approximations of $\sigma_x$ in Eq. (6) ($N_4 = N_0$). Due to the fact that the components of displacements are not subject to any boundary conditions on the prism sides, it is suitable to define the related integers as

$$N_1 = N_2 = N_3 = N_0 .$$

Note that the projection of the components $\xi_x, \xi_{xy}, \xi_{xz}$ of Eq. (11) (Hooke's tensor) on the space $P_{N_0}$ gives the following ODEs of the first order

$$\begin{cases} \int_S \xi_x \, p_{ijk} dS = 0 \\ \int_S \xi_{xy} \, p_{ijk} dS = 0 \quad \text{for} \quad i+j+k = N_0 \\ \int_S \xi_{xz} \, p_{ijk} dS = 0 \end{cases} \tag{14}$$

with respect to all the displacement functions $u^{(ijk)}(x)$, $v^{(ijk)}(x)$, $w^{(ijk)}(x)$. The dimension of the system (14) is

$$\frac{N_d}{2} = \frac{3}{2}(N_0+1)(N_0+2) \, .$$

Thus, the final number of independent stress functions selected from the whole set $\{\sigma_x^{(ijk)}(x), \ \tau_{xy}^{(ijk)}(x), \ \tau_{xz}^{(ijk)}(x)\}$ in Eq. (6) has to be equal to $N_d/2$ as well.

The other projections of Hooke's law

$$\begin{cases} \int_S \xi_y \, p_{ijk} dS = 0 \\ \int_S \xi_z \, p_{ijk} dS = 0 \quad \text{for} \quad i+j+k = N_0 \\ \int_S \xi_{yz} \, p_{ijk} dS = 0 \end{cases} \tag{15}$$

define $N_d/2$ algebraic equations with respect to stress and displacement functions.

In the second step, the boundary conditions (3) on the lateral faces of the beam are satisfied. To make it and build the necessary number of differential equations, let us define the following integers

$$N_5 = N_6 = N_7 = N_8 = N_9 = N_0 + 2 \, .$$

This means that it is necessary to fulfill $N_0 + 3$ boundary conditions on each beam side (linear relations at each monomial), or $3(N_0 + 3)$ as a whole, with respect to the stress functions $\tau_{xy}^{(ijk)}(x)$, $\tau_{xz}^{(ijk)}(x)$. Consequently, the approximations of $\tau_{xy}$ and $\tau_{xz}$ contain only

$$\tilde{N}_\tau = (N_0+3)(N_0+4) - 3(N_0+3) \tag{16}$$

independent coefficients $\tau_{xy}^{(ijk)}(x)$, $\tau_{xz}^{(ijk)}(x)$ after implementation of these equations.

By introducing a new notation $\tilde{\tau}_m(x)$ with $m = 1, \ldots, \tilde{N}_\tau$ for the remaining coefficients $\tau_{xy}^{(ijk)}(x)$ and $\tau_{xz}^{(ijk)}(x)$, the approximation of the shear stresses satisfying the boundary conditions (3) can be presented as

$$\left\{ \begin{array}{c} \tau_{xy} \\ \tau_{xz} \end{array} \right\} = \sum_{m=0}^{\tilde{N}_\tau} \tilde{\tau}_m(x) \left\{ \begin{array}{c} \tilde{\vartheta}_{xy}^{(m)}(y,z) \\ \tilde{\vartheta}_{xz}^{(m)}(y,z) \end{array} \right\}. \tag{17}$$

Here, $\tilde{\vartheta}_{xy}^{(m)}(y,z)$ and $\tilde{\vartheta}_{xz}^{(m)}(y,z)$ are basis functions obtained in agreement with these boundary conditions (3).

After resolving the relations in Eq. (4), the approximations of $\sigma_y$, $\sigma_z$, and $\tau_{yz}$ contain

$$N_\sigma = \frac{3}{2}(N_0+3)(N_0+4) - 3(2N_0+5) = \frac{N_d}{2}$$

independent coefficients $\sigma_y^{(ijk)}(x)$, $\sigma_z^{(ijk)}(x)$, $\tau_{yz}^{(ijk)}(x)$. Renumbering of these coefficients can provide the relevant stress approximation in the form

$$\left\{ \begin{array}{c} \sigma_y \\ \sigma_z \\ \tau_{yz} \end{array} \right\} = \sum_{m=0}^{N_\sigma} \sigma_m(x) \left\{ \begin{array}{c} \vartheta_y^{(m)}(y,z) \\ \vartheta_z^{(m)}(y,z) \\ \vartheta_{yz}^{(m)}(y,z) \end{array} \right\}. \tag{18}$$

Here $\vartheta_y^{(m)}(y,z)$, $\vartheta_z^{(m)}(y,z)$, and $\vartheta_{yz}^{(m)}(y,z)$ are new basis functions consistent with the boundary conditions (4). So, the number $N_\sigma$ of the functions $\sigma_m(x)$ is enough to solve exactly system (15) with respect to these coefficients.

Approximation of the equilibrium equations implies the vanishing of the following complete projections of the vector components $r_x$, $r_y$, $r_z$ in Eq. (10):

$$\left\{ \begin{array}{l} \int_S r_x p_{ijk} dS = 0 \\ \int_S r_y p_{ijk} dS = 0 \quad \text{for} \quad i+j+k = N_0 \, . \\ \int_S r_z p_{ijk} dS = 0 \end{array} \right. \tag{19}$$

It can be seen that two last relations of Eq. (19) define $N_\tau = N_d/3$ differential equations, which only include the derivatives of the functions $\tau_{xy}$ and $\tau_{xz}$. According to Eq. (16), the number of the available variables $\tilde{N}_\tau$ is bigger than $N_\tau$ and this difference is equal to

$$\tilde{N}_\tau - N_\tau = N_0 + 2 \, .$$

To reduce the number of the variables $\tilde{\tau}_m(x)$, $m = 1, \ldots \tilde{N}_\tau$, the corresponding approximation of Eq. (17) is transformed in the following way. First, the complete projections of the functions $\tau_{xy}$ and $\tau_{xz}$ on the space $P_{N_0}$ are calculated:

$$\int_S \tau_{xy} p_{ijk} dS = 0 \quad \text{and} \quad \int_S \tau_{xz} p_{ijk} dS = 0 \quad \text{for} \quad i+k+l = N_0 \, . \tag{20}$$

After that, system (20) is resolved with respect to $N_\tau$ coefficients $\tilde{\tau}_m(x)$ selected arbitrarily.

In composing a system of consistent ODEs, it is necessary to solve underdetermined systems of algebraic equations with respect to $\tilde{\tau}_m(x)$. If the calculations are performed analytically then the choice of variables for which the Eq. (20) are resolved is not so essential. But in numerical computations, a special approach should be applied, e.g., based on the Gauss elimination method to diminish computational errors. At the beginning of this successive process, the equation is chosen which contains the coefficient of maximum absolute value. After that, the variable at the maximum coefficient is expressed from this equation. The procedure accompanied with an appropriate transformation is repeated $N_\tau - 1$ times.

After solving Eq. (20) and substituting the result into Eq. (17), the following expression is obtained

$$\left\{ \begin{array}{c} \tau_{xy}^O \\ \tau_{xz}^O \end{array} \right\} = \sum_{m=N_\tau+1}^{\tilde{N}_\tau} \tau_m(x) \left\{ \begin{array}{c} \vartheta_{xy}^{(m)}(y,z) \\ \vartheta_{xz}^{(m)}(y,z) \end{array} \right\}. \tag{21}$$

Here, $\tau_m(x)$ for $m = N_\tau + 1, \ldots, \tilde{N}_\tau$ are new coefficients, which are linear combinations of $\tilde{\tau}_n(x)$ for $n = 1, \ldots, \tilde{N}_\tau$, whereas $\vartheta_{xy}^{(m)}(y,z)$ and $\vartheta_{xz}^{(m)}(y,z)$ are new basis functions orthogonal to the polynomial space $P_{N_0}$.

Let us find a representation of $\tau_{xy}$ and $\tau_{xz}$ equivalent to Eq. (20) through a new basis including the obtained functions $\vartheta_{xy}^{(m)}(y,z)$ and $\vartheta_{xz}^{(m)}(y,z)$. For this purpose, compose the following system of equations

$$\int_S \left( \tau_{xy} \vartheta_{xy}^{(m)}(y,z) + \tau_{xz} \vartheta_{xz}^{(m)}(y,z) \right) dydz = 0 \quad \text{for} \quad m = N_\tau + 1, \ldots, \tilde{N}_\tau \tag{22}$$

and resolve it with respect to some coefficients $\tilde{\tau}_i(x)$ by the Gauss elimination method.

After substituting the solution of Eq. (22) into Eq. (17) and collecting similar terms, the following expression is obtained

$$\left\{ \begin{array}{c} \tau_{xy}^P \\ \tau_{xz}^P \end{array} \right\} = \sum_{m=1}^{N_\tau} \tau_m(x) \left\{ \begin{array}{c} \vartheta_{xy}^{(m)}(y,z) \\ \vartheta_{xz}^{(m)}(y,z) \end{array} \right\}.$$

Here similarly to Eq. (21), $\tau_m(x)$ for $m = 1, \ldots, N_\tau$ are new coefficients, whereas $\vartheta_{xy}^{(m)}(y,z)$ and $\vartheta_{xz}^{(m)}(y,z)$ are new basis components, which are orthogonal to $\tau_{xy}^O$ and $\tau_{xz}^O$. It is able to verify that the obtained approximations

$$\left\{ \begin{array}{c} \tau_{xy} \\ \tau_{xz} \end{array} \right\} = \left\{ \begin{array}{c} \tau_{xy}^P \\ \tau_{xz}^P \end{array} \right\} + \left\{ \begin{array}{c} \tau_{xy}^O \\ \tau_{xz}^O \end{array} \right\}$$

satisfy the boundary conditions (3).

Thus, the final consistent DAE system includes Eqs. (14), (15), and (19) with variables $\sigma_x^{(ijk)}$, $u^{(ijk)}$, $v^{(ijk)}$, $w^{(ijk)}$, $\tau_m$, $\sigma_m$. The systems of differential equations (14) and (19) can be resolved with respect to the first derivatives of the corresponding variables $\sigma_x^{(ijk)}$, $u^{(ijk)}$, $v^{(ijk)}$, $w^{(ijk)}$, $\tau_m$. It is necessary to do so by taking into account the solution of the algebraic system (15) with respect to the variables $\sigma_m$.

After that, all the differential variables

$$\sigma_x^{(ijk)}(x)\ ,\quad u^{(ijk)}(x)\ ,\quad v^{(ijk)}(x),\quad w^{(ijk)}(x)\quad \text{for}\quad i+j+k=N_0\quad \text{and}$$
$$\tau_m(x)\quad \text{for}\quad m=1,\ldots,N_\tau$$

can be collected into a vector $a(x) \in \mathbb{R}^{N_d}$ of design parameters with the dimension

$$N_d = 3(N_0+1)(N_0+2)\ .$$

After assembling the differential equations, the resulting ODE system can be rewritten in the vector form

$$\frac{da}{dx} + Ka = 0\ , \tag{23}$$

where $K \in \mathbb{R}^{N_d \times N_d}$ is a square matrix.

In this case, the characteristic equation takes the following form

$$\det\left(K(\omega) + \lambda I\right) = 0 \tag{24}$$

with the unit matrix $I$. Equation (24) does not contain the zero root $\lambda(\omega) = 0$ at $\omega \neq 0$. In other words, the general solution of the eigenvalue problem is a linear combination of only exponentials.

Let us assemble a vector $a_1$ of design parameters $\sigma_x^{(ijk)}(x)$, $v^{(ijk)}(x)$, $w^{(ijk)}(x)$ and an other vector $a_2$ of $u^{(ijk)}(x)$, $\tau_m(x)$. The vector $a$ can be rewritten in such a way that $a = [a_1^{\mathrm{T}}, a_2^{\mathrm{T}}]^{\mathrm{T}}$. The components of the derivative $\frac{da_2(x)}{dx}$ are included in the first relations of Eq. (14) and the two last relations of Eq. (19) that depend on the components of vector $a_1$ in accordance with Eqs. (10), (11), (18). Vice versa, the components of $\frac{da_1(x)}{dx}$ are included in the first relations of Eq. (19) and the two last relations of Eq. (14) that depend on the components of vector $a_2$ (see Eqs. (10) and (11)). This means that the vector equation (23) has the following form

$$\frac{d}{dx}\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} 0 & K_{12} \\ K_{21} & 0 \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = 0\ ,\quad K_{12}, K_{21} \in \mathbb{R}^{\frac{N_d}{2}\times\frac{N_d}{2}}\ . \tag{25}$$

By using the specific structure of the matrix $K$, it is always possible to reduce the ODE system of first order (23) to an equivalent system of $N_d/2$ differential equations including the purely second derivatives of stress and displacement functions. After identical transformations, the ODEs (25) can be presented as

$$\frac{d^2 a_1}{dx^2} - K_{12} K_{21} a_1 = 0 \,. \tag{26}$$

Hence, a new eigenvalue $\mu = \lambda^2$ can be introduced to solve effectively the boundary value problem (5), (26). The characteristic equation for Eq. (26)

$$\det \left( K_{12}(\omega) K_{21}(\omega) + \mu I \right) = 0$$

with respect to $\mu$ has the polynomial order twice less than Eq. (24). By exploiting the symmetry properties of the boundary value problem with respect to the cross section $x = l/2$ and the form of the general solution of Eq. (26) (the eigenvalues $\lambda$ are always paired), the original problem can be decomposed onto two subproblems (symmetric and antisymmetric) with the dimensions twice less than the total one.

The natural frequencies $\omega_i$ for $i = 1, \ldots, N_d$ as well as the corresponding eigenforms of stresses and displacements are found under condition that the determinant of a boundary algebraic system is equal to zero. This system is obtained after substituting the general solution of the system (23) into the boundary conditions (5).

## 5 Natural Vibrations of a Beam with the Isosceles Cross Section

Consider free natural vibrations of the rectilinear beam shown in Fig. 1 with the cross section in the form of an isosceles triangle. Let the base $b$ be parallel to the $y$-axis and the height $h$ be oriented along the $z$-axis. The homogeneous isotropic elastic material is described by three dimensional constants: Young's modulus, Poisson's ratio and the volume density. By using the $\pi$ theorem, it is possible to redefine three dimensional parameters for linear elasticity problems. Here, Young's modulus $E = 1$, the volume density $\rho = 1$, and the base of the isosceles triangle $b = 1$ are defined. The other system parameters are not to choose arbitrary. Poisson's ration $\nu = 0.3$ given in the study is typical for many structural materials. The dimensionless height of the beam cross section $h = 1$ and the beam length $l$ are chosen to underline that the considered elastic body has the shape of a thin beam.

Due to the symmetry of the cross section with respect to the $z$-axis, the governing ODE system can be decomposed into two independent subsystems. At that, one of the subsystems describes coupled *bending-torsional* ($bt$) and *torsional-bending* ($tb$) motions of the beam. This system involves, for example, only even functions $\sigma_x$ with respect to the variable $z$. The other subsystem describes the *bending-longitudinal* ($bl$) and *longitudinal-bending* ($lb$) beam vibrations. Only odd functions $\sigma_x$ of Eq. (6) fit to this subsystem.

The coupling of bending with either tension or torsion is caused by an asymmetry of the beam cross section with respect to the $y$-axis. In this case, natural vibrations cannot be separated into four independent types of longitudinal, bending, and tor-

sional motions as it is supposed for beams with a symmetric cross sections [3]. Nevertheless, only one type of displacement and stress fields makes the largest contribution in the corresponding amplitudes of vibrations. This is a reason to introduce the classification of eigenfrequencies and attendant eigenforms with two letters abbreviating corresponding fields. The first letter denotes the dominant type of motion.

To obtain a reliable numerical solution, a sufficiently high degree $N_0$ of polynomial projections should be exploited. In accordance with the MIDR, the constitutive functionals proposed in [4] can be applied to estimate the quality of the solution. In this paper, the convergence of six real eigenvalues $\mu_n \in \mathbb{R}$, $n = 1, \ldots, 6$, are alternatively analyzed to define a necessary approximation dimension. Only these eigenvalues of the ODE system (26) tend to zero if the frequency $\omega$ vanishes and mainly determine the beam eigenforms for lower frequencies. If $\omega \ll 1$ then the others $\mu_n$, $n = 7, \ldots, N_d/2$, govern only transient processes near the boundary at $x = 0$ and $x = l$ (Saint-Venant's effects).

For example, the first six eigenvalues $\mu_n$ versus the polynomial degree $N_0$ are shown in Fig. 2 for $\omega = 1$. The pair of eigenvalues $\mu_1 < 0$ and $\mu_2 > 0$ presented by the lines with squares determines the bending-longitudinal forms described above. The negative values $\mu_3 < 0$ marked by circles are for the longitudinal-bending motions. The pair of lines passing through the extremal values $\mu_4 < 0$ and $\mu_5 > 0$ marked by triangles corresponds to the bending-torsional vibrations. The coupled torsional-bending motions is related with the last eigenvalue $\mu_6 < 0$ marked by rhombuses. All the values $\mu_n$ converge rather fast. At $N_0 = 4$, the maximum relative error

$$\left| \frac{\mu_n(N_0) - \mu_n(N_0 - 1)}{\mu_n(N_0 - 1)} \right| = 0.02$$

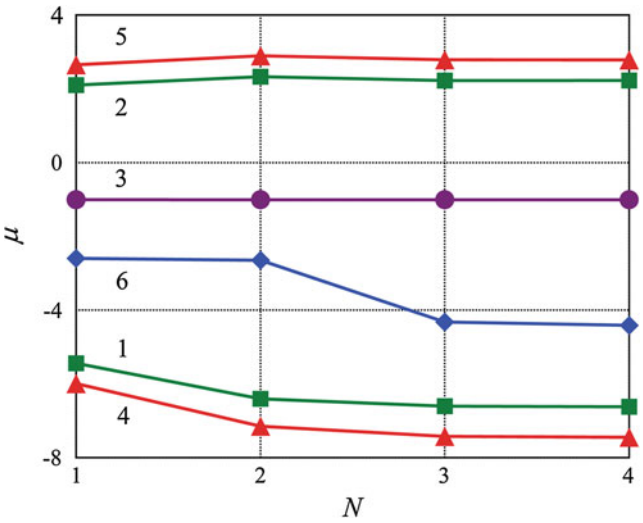



**Fig. 2** Six basic eigenvalues $\mu_n$ versus approximation order $N_0$

is found for the torsional-bending form. This approximation order will be used in what follows to solve the eigenvalue problem of beam vibrations.

A important attribute of elastic structure dynamics is the wave-frequency characteristics of a system. The dependence of the eigenvalues $\mu_n$, $n = 1, \ldots, 6$, on the frequency $\omega$ is represented in Figs. 3 and 4. As it has been mentioned above, the six basic values $\mu_n(\omega)$ depicted by curves 1–3 start at the coordinate origin. When

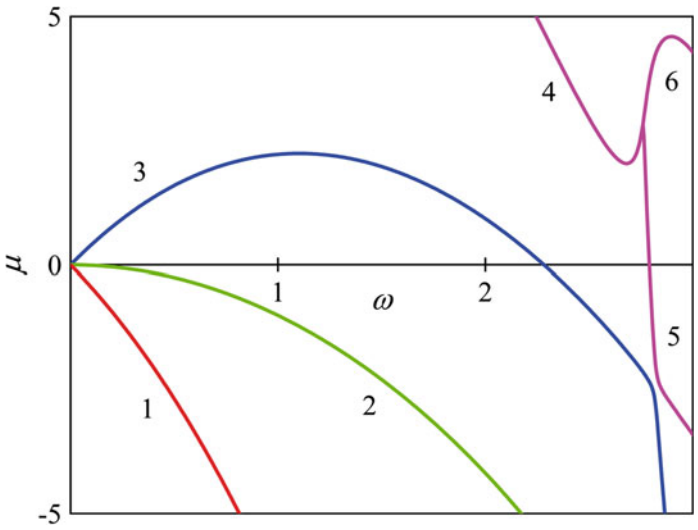



**Fig. 3** Real parts of eigenvalues $\mu$ versus the frequency $\omega$ for the longitudinal-bending vibrations

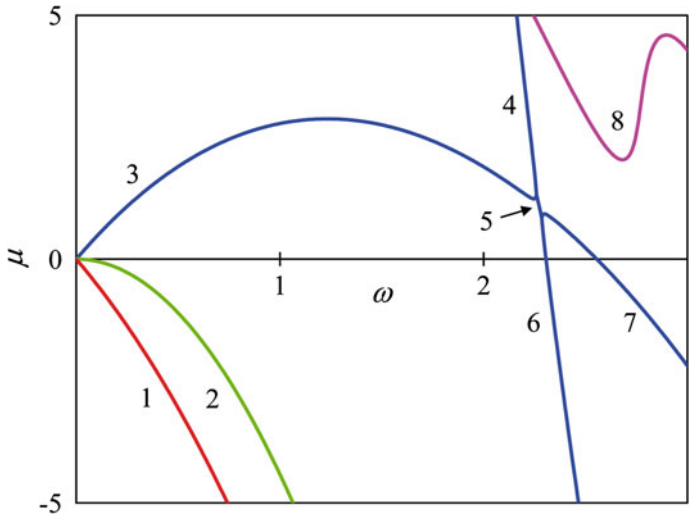



**Fig. 4** Real parts of eigenvalues $\mu$ versus the frequency $\omega$ for the torsional-bending vibrations

the frequency $\omega$ tends to zero, the asymptotic behavior of these curves agrees with the wave-frequency characteristics for longitudinal, torsional, and lateral vibrations derived in the frame of the Euler–Bernoulli beam model. As the quantity $\mu$ corresponds with the square of the wave number $\lambda$, the function $\mu_3(\omega)$ related with the beam compression and tension congruences to the classical one

$$\mu_3 = -\omega^2 + O(\omega^3)$$

(curves 2 in Fig. 3). The similar characteristic for torsion

$$\mu_6 = -4.416\omega^2 + O(\omega^3)$$

(curves 2 in Fig. 4) is also in a good correspondence with the conventional model [4].

For the Euler–Bernoulli beam bending around every principal axis of inertia, two real and two imaginary eigenvalues $\lambda_n$, $n = 1, 2, 3, 4$, are equal to each other in absolute magnitude and depend on the frequency $\omega$ as a square root. This characteristic holds asymptotically for the 3D beam model under study (curves 1 and 3 in Figs. 3 and 4) so that

$$\begin{aligned} &\mu_1 = -\frac{\omega}{\sqrt{J_y}} + O(\omega^2)\,, \quad \mu_2 = \frac{\omega}{\sqrt{J_y}} + O(\omega^2)\,, \\ &\mu_4 = -\frac{\omega}{\sqrt{J_z}} + O(\omega^2)\,, \quad \mu_5 = \frac{\omega}{\sqrt{J_z}} + O(\omega^2)\,, \\ &J_y = \int_S z^2 dS = \frac{1}{36}\,, \quad J_z = \int_S y^2 dS = \frac{1}{48}\,. \end{aligned} \tag{27}$$

Here, $J_y$ and $J_z$ are the moments of inertia respectively about the axes $y$ and $z$.

In contrast to the conventional model, these bending-longitudinal and bending-torsional functions are convex. The negative eigenvalues $\mu_1(\omega)$ and $\mu_4(\omega)$ are strictly decreasing, whereas the positive ones $\mu_2(\omega)$ and $\mu_5(\omega)$ attain their maximal values at $\omega \approx 1.104$ and $\omega \approx 1.232$, respectively.

The function $\mu_2(\omega)$ changes its sign at the critical frequency $\omega_l^{(1)} \approx 1.233$. It was found out [4] that the beam eigenforms change dramatically when passing through such critical points. At the next singular frequency $\omega_l^{(2)} \approx 2.763$, one complex conjugate pair of longitudinal-bending eigenvalues (their real part is shown in Fig. 3 by curve 4) turns into two positive real values, one of which (curve 6) is strictly increasing, whereas the other (curve 5) decreases, changes its sign at the frequency $\omega_l^{(3)} \approx 2.794$, and quickly gets flatter after passing closely over the eigenvalue $\mu_2(\omega)$ (curve 3).

For the bending-torsional vibrations, the interreaction of the corresponding eigenvalues is more sophisticated. A positive eigenvalue (curve 4 in Fig. 4) meets the value $\mu_5(\omega)$ (curve 3) at the singular point $\omega_t^{(1)} \approx 2.255$. The newly formed functions (curve 5) keep complex conjugate for $\omega \in (\omega_t^{(1)}, \omega_t^{(2)})$ with $\omega_t^{(2)} \approx 2.285$ where they turn into real and then diverge from each other for $\omega \geq \omega_t^{(2)}$. One of these

curves (6) crosses the $\omega$-axis at $\omega_t^{(3)} \approx 2.306$; the other (7) do so at $\omega_t^{(3)} \approx 2.555$. One more real eigenvalue (curve 8) appears in the chosen wave-frequency domain.

The next step of the proposed algorithm is in resolving the homogeneous boundary constraints (5) on the basis of the general solution obtained for the system (23). In accordance with the dimension of the stress functions $\sigma_x(x, y, z)$, $\tau_{xy}(x, y, z)$, $\tau_{xz}(x, y, z)$ with respect to the cross-sectional coordinates $y$ and $z$, this solution depends on $N_d$ unknown coefficients which are determined with $N_d/2$ conditions at the beam end $x = 0$ and the same number at the other end $x = l$. The linear algebraic system resulting from these conditions has a nontrivial solution if it is degenerate or, in other words, the determinant of the corresponding system matrix equals to zero.

Another important characteristic can be defined by finding all the beam lengths $l(\omega)$ at which such a degeneration takes place for some fixed frequency. As it can be seen in Fig. 5 for the bending-longitudinal as well as longitudinal-bending vibrations, the function $l(\omega)$ generates two sets of curves (solid and dashed, respectively) on the plane $\{\omega, l\} \in \mathbb{R}^2$. The curves of one set do not intersect with the other; the points of intersection correspond to the case of multiple determinant roots and require special attention (only one such point $\{\omega, l\} = \{0.233, 13.38\}$ presents in Fig. 5). All the curves asymptotically converge to the axes $\omega = 0$ and $l = 0$. Much the same picture appears for the bending-torsional and torsional-bending subsystem (solid and dashed curves in Fig. 6, respectively). In the depicted frequency-space domain, as much as two multiple roots have a place ($\{\omega, l\} = \{0.194, 7.56\}$ and $\{\omega, l\} = \{0.244, 12.26\}$).

By fixing the beam length (as an example, $l = 10$) in the Figs. 5 and 6, the whole frequency spectrum of the elastic beam in the chosen range of $\omega < 0.4$ can be restored. The numerical values of the natural frequencies for corresponding

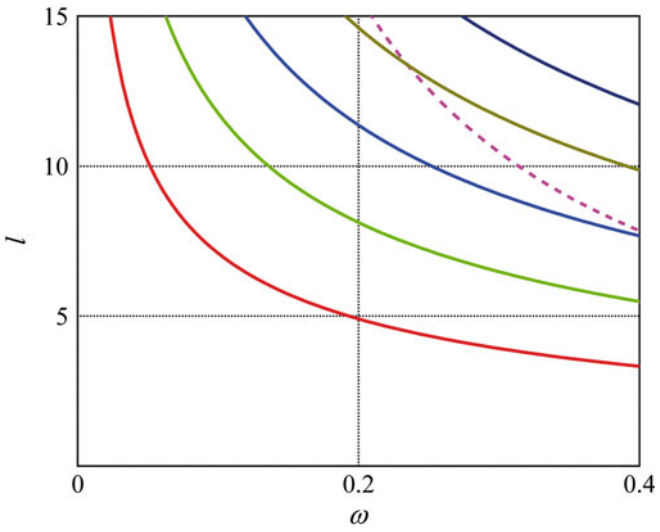



**Fig. 5** Beam length $l$ versus frequency $\omega$ for the longitudinal-bending vibrations

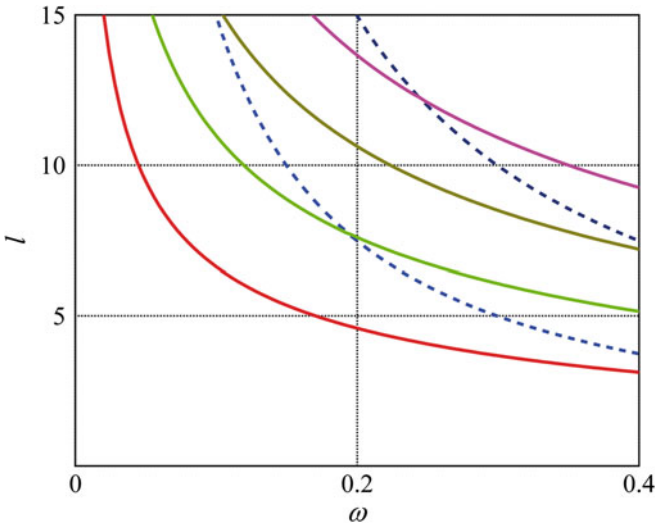



**Fig. 6** Beam length $l$ versus frequency $\omega$ for the torsional-bending vibrations

**Table 1** Eigenfrequencies for the beam with the isosceles cross section for $\omega < 0.4$

| $i$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $\omega_{bt}^{(i)}$ | 0.0448 | 0.1193 | 0.2235 | 0.3497 |
| $\omega_{tb}^{(i)}$ | 0.1496 | 0.2992 | – | – |
| $\omega_{bl}^{(i)}$ | 0.0515 | 0.1359 | 0.2521 | 0.3909 |
| $\omega_{lb}^{(i)}$ | 0.3140 | – | – | – |

eigenmodes obtained in accordance with the proposed projection algorithms at $N_0 = 4$ are given in Table 1.

The first mode of bending-longitudinal vibrations corresponding to the frequency $\omega_{bl}^{(1)}$ is shown in Fig. 7. In contrast to the form of bending vibrations according to the Euler–Bernoulli model, the bending-longitudinal motions are characterized not only by transverse displacement $w_0(x)$ but also the longitudinal displacements $u_0(x)$. Here $w_0$ and $u_0$ are the following integral characteristics

$$w_0(x) = \frac{1}{S}\int_S w(x, y, z)dS \quad \text{and} \quad u_0(x) = \frac{1}{S}\int_S u(x, y, z)dS\,.$$

The amplitudes $w_0(x)$ and $u_0(x)$ calculated at $N_0 = 4$ can be compared by the following ratio

$$\beta_1 = \frac{\max_{x\in[0,l]}|w_0(x)|}{\max_{x\in[0,l]}|u_0(x)|} = 68350\,. \tag{28}$$

At that, the deflections $w_0$ are strongly dominant over $u_0$.

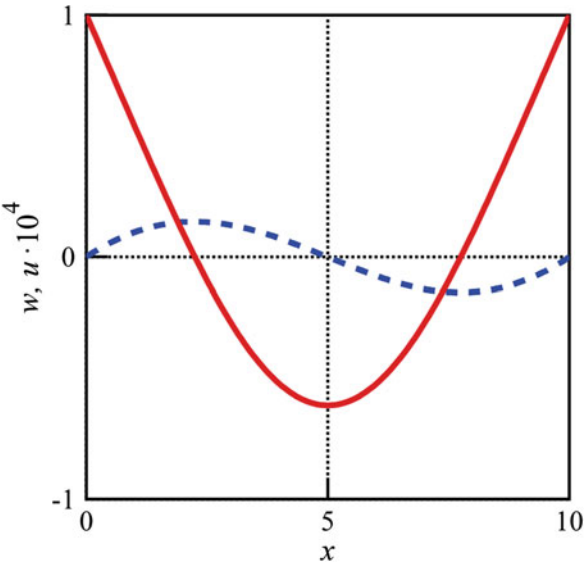



**Fig. 7** Bending-longitudinal eigenform: lateral ($w_0(x)$, *solid curve*) and longitudinal ($u_0(x)$, *dashed curve*) displacements for the frequency $\omega_{bl}^{(1)}$

The first longitudinal-bending mode of natural vibrations with the frequency $\omega_{lb}^{(1)}$ is shown in Fig. 8. This form includes not only the component of the longitudinal displacements $u_0$, as follows from the classical concept, but also the lateral $w_0$. At that, the displacements $u_0$ are dominant over $w_0$. The inverse amplitude ratio

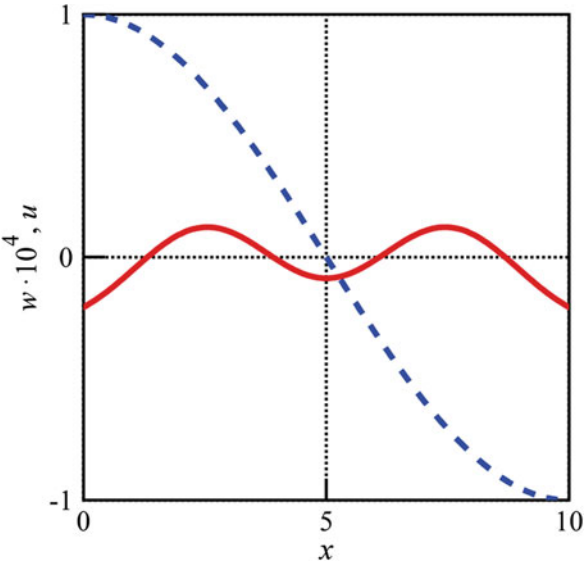



**Fig. 8** Longitudinal-bending eigenform: lateral ($w_0(x)$, *solid curve*) and longitudinal ($u_0(x)$, *dashed curve*) displacements for the frequency $\omega_{lb}^{(1)}$

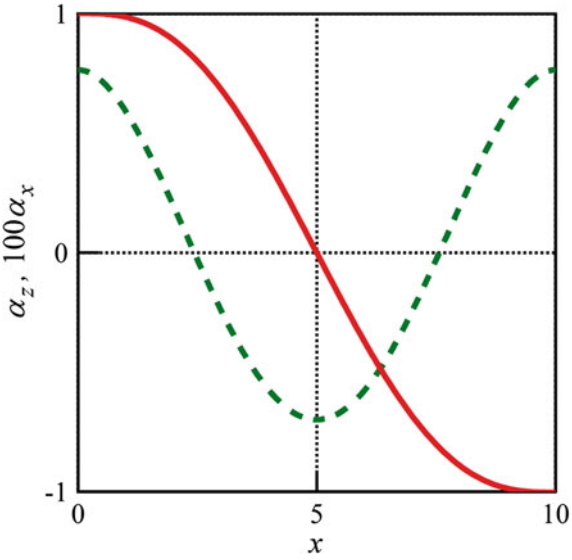



**Fig. 9** Bending-torsional eigenform: bending ($\alpha_z(x)$, *solid curve*) and torsional ($\alpha_z(x)$, *dashed curve*) angles for the frequency $\omega_{bt}^{(1)}$

$$\beta_2 = \frac{\max_{x\in[0,l]} |u_0(x)|}{\max_{x\in[0,l]} |w_0(x)|} = 48540\ . \tag{29}$$

at $\omega_{bl}^{(1)}$ is quite large conversely to the longitudinal-bending vibrations at $\omega_{lb}^{(1)}$. By taking into account Eq. (29), the influence of the longitudinal displacements can be neglected in the most cases. The values $\beta_1$ and $\beta_2$ show that the relationship between longitudinal and lateral vibrations is quite weak and can be ignored under certain assumptions for both eigenforms.

The first bending-torsional and torsional-bending forms of natural beam vibrations are shown in Figs. 9 and 10, respectively. Amplitude relations change appreciably for the bending-torsional and torsional-bending modes. To compare the bending and torsion, the functions

$$\alpha_z(x) = \frac{1}{J_z}\int_S u(x, y, z) y dS$$

and

$$\alpha_x(x) = \frac{1}{2J_x}\int_S (v(x, y, z)z - w(x, y, z)y) dS \quad \text{with} \quad J_x = J_y + J_z$$

(solid and dashed curves in Figs. 9 and 10, respectively) are introduced. Here, $\alpha_z(x)$ is the integral rotation of the beam cross section with respect to the $z$-axis, $\alpha_x(x)$ is the average angle of cross-sectional rotation around the $x$-axis. The principal moments of inertia $J_y$ and $J_z$ for the isosceles cross section are introduced in Eq. (27).

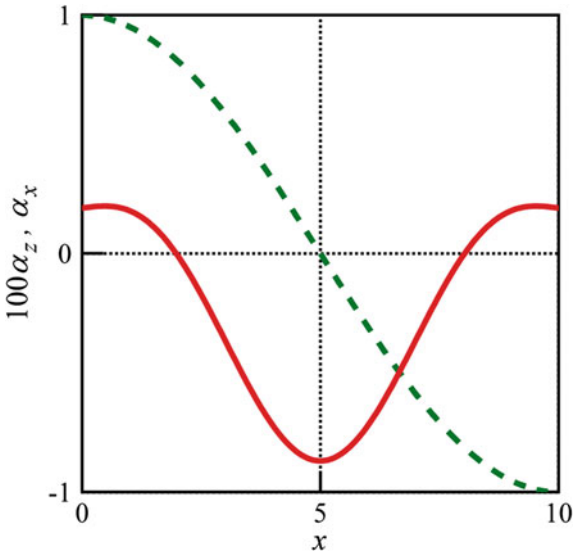



**Fig. 10** Torsional-bending eigenform: bending ($\alpha_z(x)$, *solid curve*) and torsional ($\alpha_x(x)$, *dashed curve*) angles for the frequency $\omega_{tb}^{(1)}$

The function $\alpha_z(x)$ is dominant for the bending-torsion type of natural beam motions. However, the characteristic amplitude ratio for the frequency $\omega_{bt}^{(1)}$

$$\beta_3 = \frac{\max_{x\in[0,l]} |\alpha_z(x)|}{\max_{x\in[0,l]} |\alpha_x(x)|} \approx 131$$

is not so large as in the previous two examples (see Eqs. (28) and (29)).

The inverse amplitude ratio

$$\beta_4 = \frac{\max_{x\in[0,l]} |\alpha_x(x)|}{\max_{x\in[0,l]} |\alpha_z(x)|} \approx 115$$

characterizes the relation between torsion and bending beam vibrations for the eigenfrequency $\omega_{bt}^{(1)}$.

## 6 Conclusions

The influence of cross-sectional asymmetry on the natural frequencies and forms of elastic beam vibrations has been discussed in the paper. It is found out that the natural motions of such beams cannot be divided into purely longitudinal, bending, or torsional. Due to asymmetry, such motions are coupled to each other. The frequency-wave analysis of free elastic vibrations is performed for the right triangular prism. The general features of eigenfrequencies and eigenforms are illustrated on the example

of combined torsional, longitudinal, and lateral vibrations for the beams with an isosceles cross section. Spectrum characteristics of the beams and their specific resonance properties caused by the lack of symmetry are discussed.

**Acknowledgments** This work was supported by the Russian Foundation for Basic Research, project nos. 12-01-00789, 13-01-00108, 14-01-00282, the Leading Scientific Schools Grants NSh-2710.2014.1, NSh-2954.2014.1.

## References

1. L.H. Donnell, *Beams, Plates and Shells* (McGraw-Hill, New York, 1976)
2. G.V. Kostin, V.V. Saurin, Asymptotic approach to free beam vibration analysis. J. Aerosp. Eng. **22**(4), 456–459 (2009)
3. G.V. Kostin, V.V. Saurin, Modelling and analysis of the natural oscillations of a prismatic elastic beam based on a projection approach. J. Appl. Math. Mech. **75**(6), 700–710 (2011)
4. G.V. Kostin, V.V. Saurin, *Integrodifferential Relations in Linear Elasticity* (De Gruyter, Berlin, 2012)
5. M. Levinson, On Bickford's consistent higher order beam theory. Mech. Res. Commun. **12**(1), 1–9 (1985)
6. E. Reissner, The effect of transverse shear deformation on the bending of elastic plates. J. Appl. Mech. **12**, A68–A77 (1945)
7. J.W. Strutt Rayleigh, *Theory of Sound*, vol. I, 2nd edn. (MacMillan, London, 1926)
8. S. Timoshenko, *Strength of Materials. Part I: Elementary Theory and Problems*, 3rd edn. (Van Nostrand, New York, 1955)
9. S.P. Timoshenko, J.N. Goodier, *Theory of Elasticity*, 3rd edn. (McGraw-Hill, New York, 1970)
10. O.C. Zienkiewicz, *The Finite Element in Engineering Science*, 2nd enlarged edition (McGraw-Hill, London, 1971)

# On Bifurcation Analysis of Implicitly Given Functionals in the Theory of Elastic Stability

**Nikolay Banichuk, Alexander Barsuk, Juha Jeronen, Pekka Neittaanmäki and Tero Tuovinen**

**Abstract** In this paper, we analyze the stability and bifurcation of elastic systems using a general scheme developed for problems with implicitly given functionals. An asymptotic property for the behaviour of the natural frequency curves in the small vicinity of each bifurcation point is obtained for the considered class of systems. Two examples are given. First is the stability analysis of an axially moving elastic panel, with no external applied tension, performing transverse vibrations. The second is the free vibration problem of a stationary compressed panel. The approach is applicable to a class of problems in mechanics, for example in elasticity, aeroelasticity and axially moving materials (such as paper making or band saw blades).

**Keywords** Axially moving · Elastic panel · Elastic beam · Stability · Bifurcation · Eigenvalue problem

**Mathematical Subject Classification:** 74B05 · 74H45 · 74H55 · 74H60 · 74K10

N. Banichuk (✉)
Institute for Problems in Mechanics RAS, Prospect Vernadskogo 101, Bld. 1, 119526 Moscow, Russian Federation
e-mail: banichuk@ipmnet.ru

A. Barsuk
Moldova State University, A. Mateevici 60, MD-2009 Kishinev, Moldova
e-mail: a.a.barsuk@mail.ru

J. Jeronen · P. Neittaanmäki · T. Tuovinen
Department of Mathematical Information Technology, University of Jyväskylä, P.O. Box 35 (Agora), 40014 Jyväskylä, Finland
e-mail: juha.jeronen@jyu.fi

P. Neittaanmäki
e-mail: pekka.neittaanmäki@jyu.fi

T. Tuovinen
e-mail: tero.tuovinen@jyu.fi



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_11

## 1 Introduction

Elastic stability analysis comes with a long tradition. The present form of static stability analysis was originally developed by [1], for a differential equation describing the bending of a beam or column. Dynamic stability analysis for linear elastic systems, extending Euler's method, is due to [2] following the pioneering work by Lyapunov. According to [3], the stability behaviour of some axially moving materials is mathematically analogous to the buckling of a compressed column, enabling the use of these techniques also in the context of axially moving materials.

Previously (see, e.g., [4–7]), we have considered many approaches for modelling of moving materials. The conclusions drawn can be applied, for example, to the processing of paper or steel, fabric, rubber or some other continuous material, and looping systems such as band saws and timing belts.

The most often used models for an axially moving material have been travelling flexible strings, membranes, beams, and plates. The research field of axially moving materials can be traced back to [8]. Among the first English-language papers on moving materials were [9, 10]. All these studies considered axially moving ideal strings. The analytical solution describing the free vibrations of the axially moving ideal string was derived by [11]. Dynamics and stability considerations were first reviewed in the article by [12].

The effects of axial motion of the web on its frequency spectrum and eigenfunctions were investigated in the classic papers by [13, 14]. It was shown that the natural frequency of each mode decreases when the transport speed increases, and that the travelling string and beam both experience divergence instability at a sufficiently high speed. However, in the case of the string, this result was recently contrasted by [15], who showed using Hamiltonian mechanics that the ideal string remains stable at any speed.

The loss of stability was studied with an application of dynamic and static approaches in the article by [16]. It was shown by means of numerical analysis that in the all cases instability occurs when the frequency is zero and the critical velocity coincides with the corresponding velocity obtained from static analysis. Similar results were obtained for travelling plates by [17].

The dynamical properties of moving plates have been studied by [18, 19], and the properties of a moving paper web have been studied in the two-part article by [20, 21]. Critical regimes and other problems of stability analysis have been studied, e.g., by [22, 23]. Moreover, in the articles [24–26] the author discusses widely dynamical aspects of the axially moving web. In [27] the authors considered transverse vibrations of the axially accelerating viscoelastic beam and in [28] dynamic behavior of a simply supported beam subjected to an axial transport of mass is studied. An extensive literature review related to areas presented in this paper can be found, for example, in [29]. Some approaches to bifurcation problems and estimation of critical parameters were also presented by [30, 31].

In this article, we analyze the stability and bifurcation of elastic systems using a general scheme developed for problems with implicitly given functionals.

As examples, the approach is first applied to the stability problem of a vibrating, axially moving elastic panel, with no external applied tension, and finally a stationary compressed panel is considered.

## 2 Bifurcation Method of Stability Analysis

In this section, we present the general formal scheme, which explains in general terms the idea used in the subsequent sections. For any particular case, the scheme requires proper clarification and imposing additional conditions that guarantee the mathematical correctness of the considered problem. Particularly important is that the scheme, as presented here, requires that the spectral problem depends continuously on the problem data. Often this is reasonable, but not always.

Consider the generalized spectral boundary value problem (for a sheaf of operators) described by the equation

$$\mathscr{L}(u(x), \lambda, \gamma) = \sum_{k=0}^{m} \sum_{\ell=0}^{n} \lambda^k \gamma^\ell \mathscr{L}_{k\ell}(u(x)) = 0 , \tag{1}$$

where $\gamma$ is a real-valued loading parameter, $\lambda$ is a spectral parameter, and $\mathscr{L}_{k\ell}(u(x))$ are given differential operators applied to the behaviour function $u(x)$, defined in the domain $\Omega$ ($x \in \Omega$). Boundary conditions are considered as included in the differential operator $\mathscr{L}(u(x))$.

The particular choice of the form (1) is motivated by applications in mechanics, especially elastic systems. Although time does not appear explicitly as a variable, the form lends itself to investigation of time-harmonic solutions, because these can be recast as complex-valued pseudo-steady-state problems (typically quadratic eigenvalue problems). Many models describing e.g. free vibrations and stability of elastic systems interacting with external media (liquid or gas) can be reduced to the considered form. The form is also encountered in problems related to axially moving materials, as demonstrated in the examples further below.

In mechanics applications, the spectral parameter $\lambda$ is often a complex-valued frequency of free vibration, and the load parameter $\gamma$ can be e.g. an external force, or the drive velocity of an axially moving material.

Let the function $v(x)$ be the eigenfunction (corresponding to an eigenvalue $\lambda$) of the spectral problem

$$\mathscr{L}^*(v(x), \lambda, \gamma) = 0 , \tag{2}$$

which is adjoint to the problem (1). (In the special case of a self-adjoint problem, $v(x)$ coincides with $u(x)$. In this general treatment, we do not assume self-adjointness.)

For simplicity, we leave aside the case of eigenvalues $\lambda$ with multiplicity higher than one, only noting that then $u(x)$ and $v(x)$ should be chosen so that they correspond to the same mode (which then cannot be identified by the eigenvalue alone).

Multiple eigenvalues arise at bifurcation points, and sometimes also due to symmetries in the problem. In the latter case, the multiplicity is not confined to a single point on the problem parameter axis. One particular example are the free vibrations of a homogeneous, isotropic square plate with simply supported boundary conditions on all edges; e.g. modes (1, 2) and (2, 1) then share the same eigenfrequency for the obvious geometric reason.

We multiply equation (1) by $v(x)$ and integrate over the domain, obtaining

$$\Phi\left(\lambda, J_{00}, \ldots, J_{mn}, \gamma\right) := \sum_{k=0}^{m} \sum_{\ell=0}^{n} \lambda^k \gamma^\ell J_{k\ell} = 0 \,, \tag{3}$$

where the functionals $J_{k\ell}$, $k = 1, 2, \ldots m$; $\ell = 1, 2, \ldots, n$ are defined as

$$J_{k\ell}(v, u) := (v, \mathscr{L}_{k\ell} u) = \int_\Omega v(x) \mathscr{L}_{k\ell} u(x) \, \mathrm{d}\Omega \,. \tag{4}$$

For the remainder of this section, let us formally consider $\gamma$ as an independent variable, and $\lambda$ as a dependent variable.

In this view, the relation (3) can be considered as an implicit expression for $\lambda$. The function $\Phi\left(\lambda, J_{00}, \ldots, J_{mn}, \gamma\right)$ is a polynomial of the degree $m$ with respect to $\lambda$, thus having exactly $m$ complex-valued roots (due to the fundamental theorem of algebra), which are the eigenvalues

$$\lambda_1 = \varphi_1\left(J_{00}, \ldots, J_{mn}, \gamma\right), \ldots, \lambda_m = \varphi_m\left(J_{00}, \ldots, J_{mn}, \gamma\right) \,. \tag{5}$$

The functions $\varphi_1, \ldots, \varphi_m$ represent (3) as formally solved for $\lambda$ (once for each of the $m$ roots), with $\gamma$ and $J_{k\ell}$ (for all $k, \ell$) considered as parameters. The functions input to the functionals $J_{k\ell}(v, u)$, here $v(x)$ and $u(x)$, also play the role of (function-valued) parameters inside the right-hand sides of (5).

With this in mind, consider the expression $J_{k\ell}(g, f)$ for arbitrary admissible functions $g = g(x)$ and $f = f(x)$, which are taken not to depend on $k$, $\ell$. It can be shown that at any fixed value of the problem parameter $\gamma$, if $\partial\Phi/\partial\lambda \neq 0$ at the solution points $(u, v, \lambda)$ (subscript omitted), then the functions $\varphi_1, \ldots, \varphi_m$ have zero variation at the point $(g, f) = (v, u)$, i.e. when the input functions are chosen as the eigenfunctions $u(x)$ and $v(x)$ of the direct and adjoint spectral problems (1), (2). At this point, the functions $\varphi_1, \ldots, \varphi_m$ obtain the values $\lambda_1, \ldots, \lambda_m$.

Let us show that the claimed properties hold by examining the behaviour of $\Phi$ around a solution point $(u, v, \lambda)$, where $u(x)$ and $v(x)$ are the eigenfunctions of the problems (1) and (2), respectively, and $\lambda$ is the corresponding eigenvalue (subscript omitted). Because this point is a solution of (3), if we (formally) write (5) at this point, the left-hand side will be the eigenvalue. Thus we actually only need to show the zero variation property.

Consider arbitrary small variations of the functions

$$u(x) \to u(x) + \delta u(x), \quad v(x) \to v(x) + \delta v(x) \,, \tag{6}$$

giving rise to an unknown small variation in the output of (5),

$$\lambda \to \lambda + \delta\lambda \,. \tag{7}$$

At any fixed $\gamma$, the perturbed value of $\Phi$ is then given by

$$\begin{aligned} \widetilde{\Phi} &= \Phi(\lambda + \delta\lambda,\, J_{00} + \delta J_{00},\, \ldots,\, J_{mn} + \delta J_{mn},\, \gamma) \\ &= \sum_{k=0}^{m} \sum_{\ell=0}^{n} (\lambda + \delta\lambda)^k \gamma^\ell (J_{k\ell} + \delta J_{k\ell}) \,. \end{aligned} \tag{8}$$

Using Eq. (1) for $u(x)$ and adjoint equation (2) for $v(x)$, noting definition (4) for $J_{k\ell}$, and performing elementary operations, we obtain (up to first order in the perturbations)

$$\begin{aligned} \widetilde{\Phi} &= \Phi(\lambda,\, J_{00},\, \ldots,\, J_{mn},\, \gamma) + \frac{\partial \Phi}{\partial \lambda}\delta\lambda + \sum_{k=0}^{m} \sum_{\ell=0}^{n} \lambda^k \gamma^\ell \delta J_{k\ell} \\ &= \Phi(\lambda,\, J_{00},\, \ldots,\, J_{mn},\, \gamma) + \frac{\partial \Phi}{\partial \lambda}\delta\lambda + (\delta v,\, \mathscr{L}u) + (v,\, \delta \mathscr{L} u) \\ &= \Phi(\lambda,\, J_{00},\, \ldots,\, J_{mn},\, \gamma) + \frac{\partial \Phi}{\partial \lambda}\delta\lambda + (\delta v,\, \mathscr{L}u) + \left(\mathscr{L}^* v,\, \delta u\right) = 0 \,, \end{aligned} \tag{9}$$

where on the last line we have required that the perturbed point is also a solution point of (3).

The first (unperturbed) term is zero because of (3). The third and fourth terms are also equal to zero because $u(x)$ and $v(x)$ satisfy, respectively, the equations $\mathscr{L}u = 0$, $\mathscr{L}^* v = 0$. Thus, it follows from (9) that in order to stay on the set of solutions, the perturbation must satisfy

$$\frac{\partial \Phi}{\partial \lambda}\delta\lambda = 0 \,. \tag{10}$$

Recall that $\delta\lambda$ is an unknown to be determined. If $\partial\Phi/\partial\lambda \neq 0$ at the unperturbed value $\lambda$ (which was assumed), then we must have $\delta\lambda = 0$, as claimed. Thus, provided that the condition $\partial\Phi/\partial\lambda \neq 0$ holds, the point $(g,\, f) = (v,\, u)$ is either an extremum or an inflection point for each of the functions $\varphi_1, \ldots, \varphi_m$.

Next, let us study the dependence of $\lambda_k$, $k = 1, 2, \ldots, m$, on the parameter $\gamma$. It can be shown that the functionals $J_{00}, \ldots, J_{mn}$ can be considered as constant when the function $\Phi(\lambda,\, J_{00},\, \ldots,\, J_{mn},\, \gamma)$ is differentiated with respect to $\gamma$.

This property follows simply by writing the total derivative

$$\frac{\mathrm{d}\Phi}{\mathrm{d}\gamma} = \frac{\partial\Phi}{\partial\lambda}\frac{\mathrm{d}\lambda}{\mathrm{d}\gamma} + \frac{\partial\Phi}{\partial\gamma} + \sum_{k=0}^{m}\sum_{\ell=0}^{n}\frac{\partial\Phi}{\partial J_{k\ell}}\frac{\mathrm{d}J_{k\ell}}{\mathrm{d}\gamma}\,, \tag{11}$$

and evaluating the double sum:

$$\begin{aligned}\sum_{k=0}^{m}\sum_{\ell=0}^{n}\frac{\partial\Phi}{\partial J_{k\ell}}\frac{\mathrm{d}J_{k\ell}}{\mathrm{d}\gamma} &= \sum_{k=0}^{m}\sum_{\ell=0}^{n}\lambda^k\gamma^\ell\frac{\mathrm{d}J_{k\ell}}{\mathrm{d}\gamma}\\ &= \sum_{k=0}^{m}\sum_{\ell=0}^{n}\lambda^k\gamma^\ell\left[\left(\frac{\mathrm{d}v}{\mathrm{d}\gamma},\,\mathscr{L}_{k\ell}u\right) + \left(v,\,\mathscr{L}_{k\ell}\frac{\mathrm{d}u}{\mathrm{d}\gamma}\right)\right]\\ &= \left[\left(\frac{\mathrm{d}v}{\mathrm{d}\gamma},\,\mathscr{L}u\right) + \left(v,\,\mathscr{L}\frac{\mathrm{d}u}{\mathrm{d}\gamma}\right)\right]\\ &= \left[\left(\frac{\mathrm{d}v}{\mathrm{d}\gamma},\,\mathscr{L}u\right) + \left(\mathscr{L}^*v,\,\frac{\mathrm{d}u}{\mathrm{d}\gamma}\right)\right] = 0\,,\end{aligned} \tag{12}$$

where we have used the linearity of $\Phi$ with respect to each $J_{k\ell}$, and the equalities $\mathscr{L}u = 0$ and $\mathscr{L}^*v = 0$. Thus we are left with

$$\frac{\mathrm{d}\Phi}{\mathrm{d}\gamma} = \frac{\partial\Phi}{\partial\lambda}\frac{\mathrm{d}\lambda}{\mathrm{d}\gamma} + \frac{\partial\Phi}{\partial\gamma}\,. \tag{13}$$

From the perspective of parametric studies in mechanics, where the stability of the system is investigated as a function of the problem parameter $\gamma$, the function $\Phi = \Phi(\lambda, J_{00}, \ldots, J_{mn}, \gamma)$ can be considered as a function of only two variables $\lambda$ and $\gamma$, and denoted as $F(\lambda, \gamma)$, i.e.

$$F(\lambda,\gamma) := \sum_{k=0}^{m}\sum_{\ell=0}^{n}\lambda^k\gamma^n J_{k\ell} = 0\,. \tag{14}$$

This equation can be taken as an implicit relation for $\lambda = \lambda(\gamma)$, determining a set of functions $\lambda_1(\gamma), \ldots, \lambda_m(\gamma)$.

In correspondence with the fundamental theorem on implicit functions (see, e.g., [32]), a unique solution of (14) exists in a small vicinity of the fixed values $\lambda = \widetilde{\lambda}$, $\gamma = \widetilde{\gamma}$, if $\partial F/\partial\lambda \neq 0$ at the point $(\lambda^*, \gamma^*)$.

Thus nonuniqueness of the solution of (14), or in other words, a bifurcation of the considered system, can occur for some values $\lambda = \lambda^*$, $\gamma = \gamma^*$ when the condition of the theorem on implicit functions is violated. Hence the bifurcation values $\lambda^*$ and $\gamma^*$ are found with the help of the equations

$$F(\lambda^*,\gamma^*) = 0, \quad \frac{\partial F(\lambda^*,\gamma^*)}{\partial\lambda} = 0\,. \tag{15}$$

Let us denote by $(\lambda_1^*, \gamma_1^*)$, $(\lambda_2^*, \gamma_2^*)$, ... the solutions of the nonlinear system of equations (15), representing points on the $(\lambda, \gamma)$ plane, and investigate the behaviour of the functions $\lambda_i = \lambda_i(\gamma)$ in a small vicinity of the bifurcation points $(\lambda_k^*, \gamma_k^*)$. For brevity, the subscript indices of the considered functions and points will be omitted.

Let us represent the function $F(\lambda, \gamma)$ in a small vicinity of the point $(\lambda^*, \gamma^*)$ as a series expansion,

$$\begin{aligned} F(\lambda,\gamma) = F(\lambda^*,\gamma^*) &+ \frac{\partial F(\lambda^*,\gamma^*)}{\partial\lambda}\left[\lambda-\lambda^*\right] + \frac{\partial F(\lambda^*,\gamma^*)}{\partial\gamma}\left[\gamma-\gamma^*\right] \\ &+ \frac{1}{2}\frac{\partial^2 F(\lambda^*,\gamma^*)}{\partial\lambda^2}\left[\lambda-\lambda^*\right]^2 + \frac{\partial^2 F(\lambda^*,\gamma^*)}{\partial\lambda\partial\gamma}\left[\lambda-\lambda^*\right]\left[\gamma-\gamma^*\right] \\ &+ \frac{1}{2}\frac{\partial^2 F(\lambda^*,\gamma^*)}{\partial\gamma^2}\left[\gamma-\gamma^*\right]^2 + \cdots \end{aligned} \tag{16}$$

At a bifurcation point, we have the relation (15), i.e. $F(\lambda^*, \gamma^*) = 0$ and $\partial F/\partial\lambda = 0$. This eliminates the first two terms on the right-hand side. Retaining only the lowest-order nonzero terms, we are left with

$$\begin{aligned} F(\lambda,\gamma) = &\frac{\partial F(\lambda^*,\gamma^*)}{\partial\gamma}\left[\gamma-\gamma^*\right] + \frac{\partial^2 F(\lambda^*,\gamma^*)}{\partial\lambda\partial\gamma}\left[\lambda-\lambda^*\right]\left[\gamma-\gamma^*\right] \\ &+ \frac{1}{2}\frac{\partial^2 F(\lambda^*,\gamma^*)}{\partial\lambda^2}\left[\lambda-\lambda^*\right]^2 + \cdots \end{aligned} \tag{17}$$

Let us now represent the behaviour of the function $\lambda = \lambda(\gamma)$ in the vicinity of the bifurcation point $(\lambda^*, \gamma^*)$ as

$$\lambda(\gamma) = \lambda^* + \alpha\left[\gamma-\gamma^*\right]^{\varepsilon} + \cdots , \tag{18}$$

where $\alpha$ and $\varepsilon$ are determined with the help of the condition $F(\lambda, \gamma) = 0$. By substituting (18) into (17), Eq. (17) is transformed into

$$\widetilde{F} = \widetilde{F}(\gamma-\gamma^*) \equiv 0 ,$$

which must be satisfied identically. Here $\widetilde{F}(\gamma-\gamma^*)$ is a series expansion with respect to $\gamma - \gamma^*$. As a result we have

$$\begin{aligned} F(\lambda,\gamma) = &\frac{\partial F(\lambda^*,\gamma^*)}{\partial\gamma}\left[\gamma-\gamma^*\right] + \alpha\frac{\partial^2 F(\lambda^*,\gamma^*)}{\partial\lambda\partial\gamma}\left[\gamma-\gamma^*\right]^{1+\varepsilon} \\ &+ \frac{\alpha^2}{2}\frac{\partial^2 F(\lambda^*,\gamma^*)}{\partial\lambda^2}\left[\gamma-\gamma^*\right]^{2\varepsilon} + \cdots \equiv 0 . \end{aligned} \tag{19}$$

To find an approximation in the lowest nonzero order, we pick the value of $\varepsilon$ to match the orders of different terms in (19). Once a value is chosen, we omit any remaining higher-order terms and analyze the result.

There are three possibilities. First, $\varepsilon = 0$ matches the orders of the first two terms, but eliminates them in favor of the third term, which becomes a constant. If this constant is nonzero, this is not a solution of (19). The second possibility is matching the orders of the last two terms with $\varepsilon = 1$, eliminating them and leaving only the first term. If the coefficient $\partial F/\partial\gamma \neq 0$, this is not a solution of (19).

The final possibility is to match the orders of the first and third terms with $2\varepsilon = 1$, eliminating the second term. This is the typical general case. It is valid when

$$\frac{\partial F(\lambda^*, \gamma^*)}{\partial \gamma} \neq 0, \quad \frac{\partial^2 F(\lambda^*, \gamma^*)}{\partial \lambda^2} \neq 0 \,. \tag{20}$$

If either or both of these terms vanish, the analysis must be repeated retaining the lowest-order nonzero terms for that particular case.

Inserting $\varepsilon = 1/2$ into (19) and dropping the higher-order term obtains

$$F(\lambda, \gamma) = \frac{\partial F(\lambda^*, \gamma^*)}{\partial \gamma}\left[\gamma - \gamma^*\right] + \frac{\alpha^2}{2}\frac{\partial^2 F(\lambda^*, \gamma^*)}{\partial \lambda^2}\left[\gamma - \gamma^*\right] + \cdots \equiv 0 \,, \tag{21}$$

which is satisfied identically in the lowest nonzero order by

$$\alpha^2 = -2\left(\frac{\partial F(\lambda^*, \gamma^*)}{\partial \gamma}\right)\left(\frac{\partial^2 F(\lambda^*, \gamma^*)}{\partial \lambda^2}\right)^{-1} \,. \tag{22}$$

From (18) we have the asymptotic representation

$$\lambda(\gamma) = \lambda^* + \alpha\sqrt{\gamma - \gamma^*}, \quad \left|\gamma - \gamma^*\right| \ll 1 \,, \tag{23}$$

provided that the inequalities (20) are satisfied.

As is seen from (14) and (22), the value $\alpha$ is expressed in terms of (derivatives of) the functional $F$, and does not require the analytical solution of the behavioural equation in an explicit manner.

Most importantly, the asymptotic result (23) is general for systems of the form (1) (with the appropriate additional assumptions for each particular case). The square root property in the dependence $\lambda(\gamma)$ around each bifurcation point holds for any member of the class of systems to which the general scheme applies.

## 3 Stability of an Axially Moving Panel

As an example, let us consider the stability problem of an axially moving elastic panel (plate undergoing cylindrical deformation), with no external applied tension, performing transverse vibrations. In the fixed (laboratory, Euler) coordinate system the equation of small transverse vibrations and the corresponding boundary

conditions can be written as

$$\begin{aligned} &\frac{\partial^2 w}{\partial t^2} + 2V_0 \frac{\partial^2 w}{\partial x \partial t} + V_0^2 \frac{\partial^2 w}{\partial x^2} + \frac{D}{\rho S} \frac{\partial^4 w}{\partial x^4} = 0 \,, \\ &w(0,t) = w(\ell,t) = 0, \quad D \frac{\partial^2 w(0,t)}{\partial x^2} = D \frac{\partial^2 w(\ell,t)}{\partial x^2} = 0 \,, \end{aligned} \tag{24}$$

where $w = w(x,t)$ describes the transverse displacement, $\rho$ is the density of the material, $S$ the cross-sectional area of the panel, $t$ time and $x \in [0, \ell]$.

Time-harmonic transverse vibrations of the panel are represented as

$$w(x,t) = e^{i\omega t} u(x) \,, \tag{25}$$

and the dimensionless variables

$$x = \ell \widetilde{x}, \quad \widetilde{\omega}^2 = \frac{\rho S \omega^2 \ell^4}{D}, \quad \widetilde{V_0}^2 = \frac{\rho S \ell^2}{D} V_0^2 \tag{26}$$

will be used. The tilde will be omitted.

We obtain

$$\begin{aligned} &\omega^2 u - 2i\omega V_0 \frac{\mathrm{d}u}{\mathrm{d}x} - V_0^2 \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} - \frac{\mathrm{d}^4 u}{\mathrm{d}x^4} = 0 \,, \\ &u(0) = u(1) = 0, \quad \left( \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} \right)_{x=0} = \left( \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} \right)_{x=1} = 0 \,. \end{aligned} \tag{27}$$

In (25)–(27) $\omega$ is a (complex-valued) frequency (playing the role of the eigenvalue $\lambda$), $u = u(x)$ is the amplitude function, and $i$ the imaginary unit. The axial drive velocity $V_0$ plays the role of the loading parameter $\gamma$.

After multiplication of the Eq. (27) by the complex conjugate amplitude function $u^*(x)$ and performing integration, taking into account the boundary conditions (27), we obtain

$$\Phi = a\omega^2 + 2bV_0\omega + V_0^2 c - d = 0 \,, \tag{28}$$

where

$$\begin{aligned} a &= \int_0^1 u u^* \,\mathrm{d}x > 0 \,, \\ ib &= \int_0^1 u^* \frac{\mathrm{d}u}{\mathrm{d}x} \,\mathrm{d}x = -\int_0^1 u \frac{\mathrm{d}u^*}{\mathrm{d}x} \,\mathrm{d}x > 0, \quad (b \text{ real}) \\ c &= -\int_0^1 u^* \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} \,\mathrm{d}x = \int_0^1 \frac{\mathrm{d}u}{\mathrm{d}x} \frac{\mathrm{d}u^*}{\mathrm{d}x} \,\mathrm{d}x > 0 \,, \\ d &= \int_0^1 u^* \frac{\mathrm{d}^4 u}{\mathrm{d}x^4} \,\mathrm{d}x = \int_0^1 \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} \frac{\mathrm{d}^2 u^*}{\mathrm{d}x^2} \,\mathrm{d}x > 0 \,. \end{aligned} \tag{29}$$

The quantities $a$, $b$ and $d$ are obviously real-valued, because each integrand (considered pointwise) is of the form $zz^* \equiv \|z\|$ for some complex number $z$. As for $b$, the product $z_1 z_2^*$, where $z_1 = x_1 + y_1 i$ and $z_2 = x_2 + y_2 i$ are arbitrary complex numbers, is real-valued only if $x_1 y_2 + x_2 y_1 = 0$, which does not hold in general.

Instead, consider the middle two forms in the definition of $b$ in (29). Their equality incorporates additional information from the boundary conditions, namely $u(0) = u(1) = 0$ (and correspondingly for $u^*$); this can be interpreted as resulting from an integration by parts. By summing the two forms of the definition, we have

$$2ib = \int_0^1 \left[ u^* \frac{\mathrm{d}u}{\mathrm{d}x} - u \frac{\mathrm{d}u^*}{\mathrm{d}x} \right] \mathrm{d}x = \int_0^1 \left[ u^* \frac{\mathrm{d}u}{\mathrm{d}x} - \left( u^* \frac{\mathrm{d}u}{\mathrm{d}x} \right)^* \right] \mathrm{d}x \,. \tag{30}$$

The integrand (considered pointwise) is of the form $z - z^*$, and hence the real part cancels. Thus $2ib$ must be pure imaginary, and $b$ is real-valued.

Using the notation $a$, $b$, $c$, $d$ for the considered functionals, determined by the expressions (29), we find the coefficient $\alpha$ in the asymptotic representation of the function $\lambda(\gamma)$. We have ($\Phi = F$)

$$\frac{\partial F}{\partial V_0} = 2\,(b\omega + cV_0)\,, \quad \frac{\partial^2 F}{\partial \omega^2} = 2a\,, \tag{31}$$

and consequently,

$$\alpha^2 = -2\frac{b\omega + cV_0}{a}\,. \tag{32}$$

Thus we have the following asymptotic representation for the dependence $\omega(V_0)$ in the vicinity of the bifurcation point $(\omega_k^*, V_0^*)$:

$$\begin{aligned} \omega(V_0) &\approx \omega^* \pm \sqrt{-2\frac{b\omega^* + cV_0^*}{a}}\sqrt{V_0 - V_0^*} \\ &= \omega^* \pm \sqrt{2\frac{b^2 - ac}{a^2}V_0^*}\sqrt{V_0 - V_0^*}\,. \end{aligned} \tag{33}$$

Note that for the considered problem, the equation $\Phi(\omega, a, b, c, d, V_0)$ can be solved with respect to the variable $\omega$. As a result, we have

$$\omega_{1,2}(V_0) = \frac{-bV_0 \pm \sqrt{(b^2 - ac)V_0^2 + ad}}{a}\,. \tag{34}$$

It is possible now to analyze the dependence $\omega(V_0)$, determined by expression (34) in the small vicinity of the bifurcation point $(\omega^*, V_0^*)$. Taking into account the representations for the bifurcation values of harmonic vibration frequency and velocity of axial motion,

$$\omega^* = -\frac{b}{a} V_0^*, \quad (b^2 - ac)(V_0^*)^2 = -ad \ , \tag{35}$$

and the asymptotic expression

$$V_0^2 \approx (V_0^*)^2 + 2V_0^* \left[V - V_0^*\right], \quad \left|V_0 - V_0^*\right| \ll 1 \ , \tag{36}$$

we obtain the asymptotic result

$$\omega_{1,2} \approx \omega^* \pm \sqrt{2\frac{b^2 - ac}{a^2} V_0^*} \sqrt{V_0 - V_0^*}, \quad \left|V_0 - V_0^*\right| \ll 1 \ , \tag{37}$$

which coincides with the asymptotic representation (33), as expected.

## 4 Harmonic Vibrations of a Compressed Panel

As a second example, we consider the problem of harmonic vibrations of a (stationary, not axially moving) panel compressed by the force $\gamma$ ($\gamma > 0$). The following relations will be used for the amplitude functions $u(x)$ ($x \in [0, 1]$):

$$\begin{aligned} &\frac{\mathrm{d}^4 u}{\mathrm{d}x^4} + \gamma \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} - \omega^2 u = 0, \\ &u(0) = u(1) = 0, \quad \left(\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}\right)_{x=0} = \left(\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}\right)_{x=1} = 0 \ . \end{aligned} \tag{38}$$

Let us investigate the asymptotic behaviour of the frequency $\omega$ as a function of the loading parameter $\gamma$, i.e. $\omega = \omega(\gamma)$, using the discussed perturbation method. To do this, we multiply the Eq. (38) by the function $u(x)$, which coincides in the considered case with $u^*(x)$ (because the problem (38) is self-adjoint), and perform integration.

As a result, we will find the following expression for $\Phi$ as a function of the functionals $a$, $c$ and $d$ (as defined in (29)). We have

$$\Phi(\omega, a, c, d, \gamma) = -a\omega^2 - \gamma c + d = 0 \ . \tag{39}$$

The functionals $a$, $c$ and $d$ can be expressed with the help of eigenmodes of vibrations

$$u_k(x) = B_k \sin(k\pi x) \ .$$

We find

$$
\begin{aligned}
a_k &= \int_0^1 (u_k(x))^2 \, \mathrm{d}x = \frac{B_k^2}{2} \, , \\
c_k &= \int_0^1 \left( \frac{\mathrm{d}u_k}{\mathrm{d}x} \right)^2 \mathrm{d}x = \frac{k^2\pi^2}{2} B_k^2 \, , \\
d_k &= \int_0^1 \left( \frac{\mathrm{d}^2u_k}{\mathrm{d}x^2} \right)^2 \mathrm{d}x = \frac{k^4\pi^4}{2} B_k^2 \, .
\end{aligned}
\tag{40}
$$

In correspondence with the general formulas (22)–(23), the asymptotic behaviour of the frequencies in the vicinity of the bifurcation points

$$
\omega_k^* = 0 \, , \quad \gamma_k^* = k^2\pi^2 \tag{41}
$$

will be described by the expressions

$$
\omega_k = \omega_k(\gamma) = \pm\alpha\sqrt{\gamma - k^2\pi^2}, \quad \left| \gamma - k^2\pi^2 \right| \ll 1 \, , \tag{42}
$$

and the value of the coefficient $\alpha$ will be given by

$$
\alpha^2 = -2 \left( \frac{\partial F(\omega_k^*, \gamma_k^*)}{\partial \gamma} \right) \left( \frac{\partial^2 F(\omega_k^*, \gamma_k^*)}{\partial \omega^2} \right)^{-1} = -k^2\pi^2 \, . \tag{43}
$$

## 5 Conclusions

In this paper, we studied the stability and bifurcation of elastic systems using a general scheme developed for problems with implicitly given functionals. The most important observation gained from the analysis is, that in the small vicinity of each bifurcation point, a square root type of dependence takes place for the eigenvalue (complex eigenfrequency) as a function of the problem parameter. This is a general property, which holds for any system of the class considered.

After a general exposition, the approach was applied to two examples: the transverse elastic stability of an axially moving elastic panel with no external applied tension, and the free vibrations of a compressed panel.

The presented approach has applications, for example, in elasticity, aeroelasticity and axially moving materials.

## References


1. L. Euler, De motu vibratorio tympanorum. Novi Commentarii academiae scientiarum imperialis Petropolitanae **10**, 243–260 (1766)
2. V.V. Bolotin, *Nonconservative Problems of the Theory of Elastic Stability* (Pergamon Press, New York, 1963)
3. C.D. Mote Jr, J.A. Wickert, Response and discretization methods for axially moving materials. Appl. Mech. Rev. **44**(11), S279–S284 (1991)
4. N. Banichuk, J. Jeronen, M. Kurki, P. Neittaanmäki, T. Saksa, T. Tuovinen, On the limit velocity and buckling phenomena of axially moving orthotropic membranes and plates. Int. J. Solids Struct. **48**(13), 2015–2025 (2011)
5. N. Banichuk, J. Jeronen, P. Neittaanmäki, T. Saksa, T. Tuovinen, Theoretical study on travelling web dynamics and instability under non-homogeneous tension. Int. J. Mech. Sci. **66C**, 132–140 (2013)
6. N. Banichuk, J. Jeronen, P. Neittaanmäki, T. Tuovinen, Dynamic behaviour of an axially moving plate undergoing small cylindrical deformation submerged in axially flowing ideal fluid. J. Fluids Struct. **27**(7), 986–1005 (2011)
7. N. Banichuk, M. Kurki, P. Neittaanmäki, T. Saksa, M. Tirronen, T. Tuovinen, Optimization and analysis of processes with moving materials subjected to fatigue fracture and instability. Mech. Based Des. Struct. Mach. Int. J. **41**(2), 146–167 (2013)
8. R. Skutch, Uber die Bewegung eines gespannten Fadens, weicher gezwungen ist durch zwei feste Punkte, mit einer constanten Geschwindigkeit zu gehen, und zwischen denselben in Transversal-schwingungen von gerlinger Amplitude versetzt wird. Annalen der Physik und Chemie **61**, 190–195 (1897)
9. W.L. Miranker, The wave equation in a medium in motion. IBM J. Res. Dev. **4**, 36–42 (1960)
10. R.A. Sack, Transverse oscillations in traveling strings. Br. J. Appl. Phys. **5**, 224–226 (1954)
11. R.D. Swope, W.F. Ames, Vibrations of a moving threadline. J. Franklin Inst. **275**, 36–55 (1963)
12. C.D. Mote, Dynamic stability of axially moving materials. Shock Vibr. Digest **4**(4), 2–11 (1972)
13. F.R. Archibald, A.G. Emslie, The vibration of a string having a uniform motion along its length. ASME J. Appl. Mech. **25**, 347–348 (1958)
14. A. Simpson, Transverse modes and frequencies of beams translating between fixed end supports. J. Mech. Eng. Sci. **15**, 159–164 (1973)
15. Y. Wang, L. Huang, X. Liu, Eigenvalue and stability analysis for transverse vibrations of axially moving strings based on Hamiltonian dynamics. Acta. Mech. Sin. **21**, 485–494 (2005)
16. J.A. Wickert, Non-linear vibration of a traveling tensioned beam. Int. J. Non-Linear Mech. **27**(3), 503–517 (1992)
17. C.C. Lin, Stability and vibration characteristics of axially moving plates. Int. J. Solids Struct. **34**(24), 3179–3190 (1997)
18. J.Y. Shen, L. Sharpe, W.M. McGinley, Identification of dynamic properties of plate-like structures by using a continuum model. Mech. Res. Commun. **22**(1), 67–78 (1995)
19. C. Shin, J. Chung, W. Kim, Dynamic characteristics of the out-of-plane vibration for an axially moving membrane. J. Sound Vib. **286**(4–5), 1019–1031 (2005)
20. A. Kulachenko, P. Gradin, H. Koivurova, Modelling the dynamical behaviour of a paper web. Part I. Comput. Struct. **85**, 131–147 (2007)
21. A. Kulachenko, P. Gradin, H. Koivurova, Modelling the dynamical behaviour of a paper web. Part II. Comput. Struct. **85**, 148–157 (2007)
22. R. Sygulski, Stability of membrane in low subsonic flow. Int. J. Non-Linear Mech. **42**(1), 196–202 (2007)
23. X. Wang, Instability analysis of some fluid-structure interaction problems. Comput. Fluids **32**(1), 121–138 (2003)
24. K. Marynowski, Non-linear dynamic analysis of an axially moving viscoelastic beam. J. Theor. Appl. Mech. **2**(40), 465–482 (2002)

25. K. Marynowski, Non-linear vibrations of an axially moving viscoelastic web with time-dependent tension. Chaos Solitons Fractals **21**(2), 481–490 (2004). doi:10.1016/j.chaos.2003.12.020
26. K. Marynowski, Non-linear vibrations of the axially moving paper web. J. Theor. Appl. Mech. **46**(3), 565–580 (2008)
27. X.-D. Yang, L.-Q. Chen, Bifurcation and chaos of an axially accelerating viscoelastic beam. Chaos Solitons Fractals **23**(1), 249–258 (2005)
28. F. Pellicano, F. Vestroni, Nonlinear dynamics and bifurcations of an axially moving beam. J. Vib. Acoust. **122**(1), 21–30 (2000)
29. M.H. Ghayesh, M. Amabili, M.P. Païdoussis, Nonlinear dynamics of axially moving plates. J. Sound Vib. **332**(2), 391–406 (2013)
30. P. Neittaanmäki, K. Ruotsalainen, On the numerical solution of the bifurcation problem for the sine-Gordon equation. Arab J. Math. **6**(1 and 2) (1985)
31. J. Nečas, A. Lehtonen, P. Neittaanmäki, On the construction of Lusternik-Schnirelmann critical values with application to bifurcation problems. Appl. Anal. **25**(4), 253–268 (1987)
32. K. Rektorys, *Survey of Applicable Mathematics* (Iliffe Books London Ltd, 1969)

# Part III
# Optimization

# Proximal Bundle Method for Nonsmooth and Nonconvex Multiobjective Optimization

**Marko M. Mäkelä, Napsu Karmitsa and Outi Wilppu**

**Abstract** We present a proximal bundle method for finding weakly Pareto optimal solutions to constrained nonsmooth programming problems with multiple objectives. The method is a generalization of proximal bundle approach for single objective optimization. The multiple objective functions are treated individually without employing any scalarization. The method is globally convergent and capable of handling several nonconvex locally Lipschitz continuous objective functions subject to nonlinear (possibly nondifferentiable) constraints. Under some generalized convexity assumptions, we prove that the method finds globally weakly Pareto optimal solutions. Concluding, some numerical examples illustrate the properties and applicability of the method.

**Keywords** Multiobjective optimization · Nonsmooth optimization · Bundle methods

**Mathematical Subject Classification:** 90C26 · 90C29 · 65K05 · 46N10

## 1 Introduction

Nonsmooth (nondifferentiable) optimization problems arise in very many fields of applications, for example, in optimal shape design (see, e.g., [2, 5, 13]), economics [21] and mechanics [19]. On the other hand, instead of one criterion the applications typically have several, often conflicting objectives. During the last three decades the rapid development has been characteristic to the areas of nonsmooth (see, e.g.,

M.M. Mäkelä (✉) · N. Karmitsa · O. Wilppu
Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland
e-mail: makela@utu.fi

N. Karmitsa
e-mail: napsu@karmitsa.fi

O. Wilppu
e-mail: omwilp@utu.fi



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_12

[1, 4, 8, 10–12, 18, 22]) and multiobjective optimization (see, e.g., [16, 17, 20, 23]), separately. Conversely the consideration of both of these approaches in the same framework, i.e. nonsmooth multiobjective optimization, is much less frequent. Thus there exists an increasing demand to be able to solve efficiently optimization problems with several, possible nonsmooth, objective functions.

In this paper we present a proximal bundle based method for constrained nonconvex nonsmooth programming problems with multiple objectives. The method generalizes the proximal bundle approach for single objective optimization [9] by employing the ideas presented in [7, 17, 24]. We can prove, that under some generalized convexity assumptions [15] the method can find globally weakly Pareto optimal solutions. Unlike the most multicriteria optimization methods the multiple objective functions are treated individually without employing any scalarization.

The method is readily implementable and descent, i.e., the value of each objective function is expected to get an improvement at each iteration. Thus the starting point is projected to the weakly Pareto optimal set through the negative orthant of the decision space. This means that the user can control via choosing the starting point which kind of optimal solution is generated. Hence the method can be used either as a part of interactive multiobjective optimization methods producing the efficiently (weakly) Pareto optimal counterparts of nonoptimal solutions or, by choosing several starting points, a good spread of (weakly) Pareto optimal solutions.

The paper is organized as follows. Section 2 contains some preliminary concepts and results of nonsmooth and multiobjective optimization theory. The algorithm of the multicriteria proximal bundle (MPB) method is described in Sect. 3. Some convergence results are presented in Sect. 4. Finally, Sect. 5 is devoted to some numerical examples illustrating the properties and applicability of the method.

## 2 Preliminaries

Let us consider a nonsmooth multiobjective optimization problem of the form

$$\begin{cases} \text{minimize} & \{f_1(x), \ldots, f_k(x)\} \\ \text{subject to} & x \in S, \end{cases} \tag{1}$$

where

$$S = \{\mathrm{x} \in \mathbb{R}^n \mid g_j(x) \leq 0, \ \ j = 1, \ldots, m\}.$$

The objective functions $f_i : \mathbb{R}^n \to \mathbb{R}$ and the constraint functions $g_j : \mathbb{R}^n \to \mathbb{R}$ are supposed to be *locally Lipschitz continuous* (not necessarily smooth nor convex). For a locally Lipschitz continuous function $f : \mathbb{R}^n \to \mathbb{R}$ the *Clarke generalized directional derivative* [3] at x in the direction $\mathrm{d} \in \mathbb{R}^n$ is defined by

$$f^\circ(\mathrm{x};\mathrm{d}) = \limsup_{\substack{\mathrm{y}\to\mathrm{x}\\ t\downarrow 0}} \frac{f(\mathrm{y}+t\mathrm{d}) - f(\mathrm{y})}{t}$$

and the *Clarke subdifferential* [3] of $f$ at x by

$$\partial f(\mathrm{x}) = \{\xi \in \mathbb{R}^n \mid f^\circ(\mathrm{x};\mathrm{d}) \geq \xi^T \mathrm{d} \text{ for all } \mathrm{d} \in \mathbb{R}^n\},$$

which is a nonempty, convex and compact subset of $\mathbb{R}^n$. Note, that if a locally Lipschitz continuous function attains its local minimum at $\mathrm{x}^*$, then

$$0 \in \partial f(\mathrm{x}^*). \tag{2}$$

For a finite maximum of locally Lipschitz continuous functions we have the following subderivation rule.

**Theorem 1** *([3]) Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous at* x *for all $i = 1, \ldots, m$. Then the function*

$$f(\mathrm{x}) = \max_{i=1,\ldots,m} [f_i(\mathrm{x})]$$

*is locally Lipschitz continuous at* x *and*

$$\partial f(\mathrm{x}) \subseteq \mathrm{conv}\{\partial f_i(\mathrm{x}) \mid (f_{\mathrm{J}i}(\mathrm{x}) = f(\mathrm{x}),\ i = 1, \ldots, m\}, \tag{3}$$

*where* conv *denotes the convex hull of a set.*

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *weakly semismooth* if the classical directional derivative

$$f'(\mathrm{x},\mathrm{d}) = \lim_{t\downarrow 0} \frac{f(\mathrm{x}+t\mathrm{d}) - f(\mathrm{x})}{t}$$

exists for all x and d, and

$$f'(\mathrm{x},\mathrm{d}) = \lim_{t\downarrow 0} \xi(\mathrm{x}+t\mathrm{d})^T \mathrm{d},$$

where $\xi(\mathrm{x}+t\mathrm{d}) \in \partial f(\mathrm{x}+t\mathrm{d})$.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is $f^\circ$*-pseudoconvex* [6], if it is locally Lipschitz continuous and for all $\mathrm{x}, \mathrm{y} \in \mathbb{R}^n$

$$f(\mathrm{y}) < f(\mathrm{x}) \quad \text{implies} \quad f^\circ(\mathrm{x};\mathrm{y}-\mathrm{x}) < 0$$

and $f^\circ$*-quasiconvex*, if

$$f(\mathrm{y}) \leq f(\mathrm{x}) \quad \text{implies} \quad f^\circ(\mathrm{x};\mathrm{y}-\mathrm{x}) \leq 0.$$

Note, that a convex function is always $f^\circ$-pseudoconvex, which again is $f^\circ$-quasiconvex (see, e.g., [15]). Next we present two important properties of $f^\circ$-pseudoconvex functions.

**Theorem 2** *([6]) An $f^\circ$-pseudoconvex function $f$ attains its global minimum at $\mathrm{x}^*$, if and only if*

$$0 \in \partial f(\mathrm{x}^*).$$

The proof of the following useful result can be found, for example, in [1].

**Theorem 3** *Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be $f^\circ$-pseudoconvex for all $i = 1, \ldots, m$. Then the function*

$$f(\mathrm{x}) = \max_{i=1,\ldots,m} [f_i(\mathrm{x})]$$

*is also $f^\circ$-pseudoconvex.*

Note, that for an $f^\circ$-quasiconvex function $f$ the level set $\mathrm{lev}_\alpha f := \{\mathrm{x} \in \mathbb{R}^n \mid f(\mathrm{x}) \leq \alpha\}$ is a convex set for all $\alpha \in \mathbb{R}$ (see, e.g., [15]).

A vector $\mathrm{x}^*$ is said to be a *global Pareto optimum* of (1), if there does not exist $\mathrm{x} \in S$ such, that

$$f_i(\mathrm{x}) \leq f_i(\mathrm{x}^*) \text{ for all } i = 1, \ldots, k \quad \text{and} \quad f_j(\mathrm{x}) < f_j(\mathrm{x}^*) \text{ for some } j.$$

Vector $\mathrm{x}^*$ is said to be a *global weak Pareto optimum* of (1), if there does not exist $\mathrm{x} \in S$ such, that

$$f_i(\mathrm{x}) < f_i(\mathrm{x}^*) \quad \text{for all } i = 1, \ldots, k.$$

Vector $\mathrm{x}^*$ is a *local (weak) Pareto optimum* of (1), if there exists $\delta > 0$ such, that $\mathrm{x}^*$ is a global (weak) Pareto optimum on $B(\mathrm{x}^*; \delta) \cap S$. Trivially every Pareto optimal point is weakly Pareto optimal.

The *contingent cone* and *polar cone* of set $S \subseteq \mathbb{R}^n$ at point x are defined respectively as

$$K_S(\mathrm{x}) = \left\{\mathrm{d} \in \mathbb{R}^n \mid \text{there exist } t_i \downarrow 0 \text{ and } \mathrm{d}_i \to \mathrm{d} \text{ with } \mathrm{x} + t_i \mathrm{d}_i \in S\right\}$$
$$S^{\leq} = \{\mathrm{d} \in \mathbb{R}^n \mid \mathrm{s}^T \mathrm{d} \leq 0, \text{ for all } s \in S\}.$$

The closure of a set $S$ is denoted by $\mathrm{cl}\, S$. A set $C \subset \mathbb{R}^n$ is a *cone* if $\lambda \mathrm{x} \in C$ for all $\lambda \geq 0$ and $\mathrm{x} \in C$. We also denote

$$\mathrm{ray}\, S = \{\lambda x \mid \lambda \geq 0,\ x \in S\} \quad \text{and} \quad \mathrm{cone}\ S = \mathrm{ray}\, \mathrm{conv}\ S.$$

In other words ray $S$ is the smallest cone containing $S$ and the conic hull cone $S$ the smallest convex cone containing $S$. Furthermore, let

$$F(\mathrm{x}) = \bigcup_{i=1}^{k} \partial f_i(\mathrm{x})$$

and

$$G(\mathrm{x}) = \bigcup_{j \in J(\mathrm{x})} \partial g_j(\mathrm{x}), \quad \text{where } J(\mathrm{x}) = \{ j \mid g_j(\mathrm{x}) = 0 \}.$$

For the optimality condition we pose the following *constraint qualification*

$$G^{\leq}(\mathrm{x}) \subseteq K_S(\mathrm{x}). \tag{4}$$

Now we can present the following generalized KKT optimality conditions.

**Theorem 4** *([15]) If* $\mathrm{x}^*$ *is a local weak Pareto optimum of* (1) *and the constraint qualification* (4) *is valid, then*

$$0 \in \operatorname{conv} F(\mathrm{x}^*) + \operatorname{cl} \operatorname{cone} G(\mathrm{x}^*). \tag{5}$$

*Moreover, if* $f_i$ *are* $f^\circ$*-pseudoconvex for all* $i = 1, \ldots, k$ *and* $g_j$ *are* $f^\circ$*-quasiconvex for all* $j = 1, \ldots, m$*, then the condition* (5) *is sufficient for* $\mathrm{x}^*$ *to be a global weak Pareto optimum of* (1).

A feasible point $x^* \in S$ is called a *substationary point* for problem (1), if it satisfies the necessary optimality condition (5).

## 3 Multiobjective Proximal Bundle Method

In this section we develop the MPB (Multiobjective Proximal Bundle) method. The original proximal bundle method of [9] for nonsmooth convex and unconstrained single objective optimization was generalized to handle nonconvex and constrained problems in [13]. The MPB method is a further extension into a multiobjective case. The strategy of handling several objective functions is based on the ideas presented in [7, 17, 24]. The idea, in brief, is to move into a direction where the values of all the objective functions improve simultaneously.

### *3.1 Direction Finding*

The MPB method is not directly based on employing any scalarizing function. Some kind of scalarization is, however, needed in deriving the minimization method for all the objective functions. Theoretically, we utilize the *improvement function* $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$H(x, y) = \max_{\substack{i=1,\ldots,k \\ j=1,\ldots,m}} [f_i(x) - f_i(y), g_j(x)].$$

Now we obtain the following connection between the improvement function and the problem (1).

**Theorem 5** *A necessary condition for $x^* \in \mathbb{R}^n$ to be a global weak Pareto optimum of (1) is that*

$$x^* = \underset{x \in \mathbb{R}^n}{arg\,min}\, H(x, x^*). \tag{6}$$

*Moreover, if $f_i$ are $f^\circ$-pseudoconvex for all $i = 1, \ldots, k$, $g_j$ are $f^\circ$-quasiconvex for all $j = 1, \ldots, m$ and the constraint qualification (4) is valid, then the condition (6) is sufficient for* x* *to be a global weak Pareto optimum of (1).*

*Proof* Suppose first, that $x^* \in \mathbb{R}^n$ is a global weak Pareto optimum of (1). Since $x^* \in S$ we have $g_j(x^*) \leq 0$ for all $j = 1, \ldots, m$, thus $H(x^*, x^*) = 0$. If $x^*$ would not be a global minimizer of $H(\cdot, x^*)$, there would exist $y^* \in \mathbb{R}^n$ such that

$$H(y^*, x^*) < H(x^*, x^*) = 0.$$

Then we have $g_j(y^*) < 0$ for all $j = 1, \ldots, m$, in other words, $y^* \in S$. Furthermore, $f_i(y^*) < f_i(x^*)$ for all $i = 1, \ldots, k$, which contradicts the global weak Pareto optimality of $x^*$.

Suppose next that (6) holds true. Suppose also that $f_i$ are $f^\circ$-pseudoconvex for all $i = 1, \ldots, k$, $g_j$ are $f^\circ$-quasiconvex for all $j = 1, \ldots, m$ and (4) is valid. Since $x^*$ is a global minimizer of $H(\cdot, x^*)$ by (2), Theorem 3 and Lemma 2.10 of [14] we have

$$\begin{aligned} 0 \in \partial H(x^*, x^*) &= \operatorname{conv} \left\{ \bigcup_{i=1}^{k} \partial f_i(x^*) \cup \bigcup_{j \in J(x^*)} \partial g_j(x^*) \right\} \\ &= \operatorname{conv} \left\{ F(x^*) \cup G(x^*) \right\} \\ &\subseteq \operatorname{conv} \left\{ \operatorname{conv} F(x^*) \cup \operatorname{conv} G(x^*) \right\} \\ &= \left\{ \lambda \operatorname{conv} F(x^*) + (1 - \lambda) \operatorname{conv} G(x^*) \mid \lambda \in [0, 1] \right\}. \end{aligned}$$

Then there exists $\lambda \in (0, 1]$ such that

$$\begin{aligned} 0 &\in \operatorname{conv} F(x^*) + \frac{1 - \lambda}{\lambda} \operatorname{conv} G(x^*) \\ &\subseteq \operatorname{conv} F(x^*) + \operatorname{ray} \operatorname{conv} G(x^*) \\ &= \operatorname{conv} F(x^*) + \operatorname{cone} G(x^*) \\ &\subseteq \operatorname{conv} F(x^*) + \operatorname{cl} \operatorname{cone} G(x^*). \end{aligned}$$

Thus, Theorem 4 implies that x* is a global weak Pareto optimum of (1). □

Let $x^h$ be the current approximation to the solution of (1) at the iteration $h$. Then, by Theorem 5, we seek for the search direction $d^h$ as a solution of

$$\begin{cases} \text{minimize} & H(x^h + d, x^h) \\ \text{subject to} & d \in \mathbb{R}^n. \end{cases} \tag{7}$$

Since (7) still is a nonsmooth problem, we must approximate it somehow. Let us assume for a moment that the problem (1) is convex. We suppose that, at the iteration $h$ besides the current iteration point $x^h$, we have some auxiliary points $y^j \in \mathbb{R}^n$ from the past iterations and subgradients $\xi_{f_i}^j \in \partial f_i(y^j)$ for $j \in J^h = \{1, \ldots, h\}$, $i = 1, \ldots, k$, and $\xi_{g_l}^j \in \partial g_l(y^j)$ for $j \in J^h, l = 1, \ldots, m$. We linearize the objective and the constraint functions at the point $y^j$ by

$$\bar{f}_{i,j}(x) = f_i(y^j) + (\xi_{f_i}^j)^T(x - y^j) \quad \text{for all } i = 1, \ldots, k,\ j \in J^h, \quad \text{and}$$
$$\bar{g}_{l,j}(x) = g_l(y^j) + (\xi_{g_l}^j)^T(x - y^j) \quad \text{for all } l = 1, \ldots, m,\ j \in J^h.$$

Now we can define a convex piecewise linear approximation to the improvement function by

$$\hat{H}^h(x) = \max_{\substack{i=1,\ldots,k \\ l=1,\ldots,m \\ j \in J^h}} [\bar{f}_{i,j}(x) - f_i(x^h), \bar{g}_{l,j}(x)]$$

and we get an approximation to (7) by

$$\begin{cases} \text{minimize} & \hat{H}^h(x^h + d) + \frac{1}{2}u^h\|d\|^2 \\ \text{subject to} & d \in \mathbb{R}^n, \end{cases} \tag{8}$$

where $u^h > 0$ is some weighting parameter. The penalty term $\frac{1}{2}u^h\|d\|^2$ is added to guarantee the existence and uniqueness of a solution to (8) and also to keep the approximation local enough. Notice that (8) still is a nonsmooth problem, but due to its minmax-nature it is equivalent to the following (smooth) quadratic problem

$$\begin{cases} \text{minimize} & v + \frac{1}{2}u^h\|d\|^2 \\ \text{subject to} & -\alpha_{f_i,j}^h + (\xi_{f_i}^j)^T d \le v, \quad i = 1, \ldots, k,\ j \in J^h \\ & -\alpha_{g_l,j}^h + (\xi_{g_l}^j)^T d \le v, \quad l = 1, \ldots, m,\ j \in J^h, \end{cases} \tag{9}$$

where

$$\alpha_{f_i,j}^h := f_i(\mathrm{x}^h) - \bar{f}_{i,j}(\mathrm{x}^h), \quad i = 1, \ldots, k,\ j \in J^h, \quad \text{and}$$
$$\alpha_{g_l,j}^h := -\bar{g}_{l,j}(\mathrm{x}^h), \quad l = 1, \ldots, m,\ j \in J^h,$$

are so-called *linearization errors*.

In the nonconvex case, we replace the linearization errors by *subgradient locality measures*

$$\beta^h_{f_i,j} := \max[|\alpha^h_{f_i,j}|, \gamma_{f_i}\|\mathrm{x}^h - \mathrm{y}^j\|^2],$$
$$\beta^h_{g_l,j} := \max[|\alpha^h_{g_l,j}|, \gamma_{g_l}\|\mathrm{x}^h - \mathrm{y}^j\|^2],$$

where $\gamma_{f_i} \geq 0$ for $i = 1, \ldots, k$ and $\gamma_{g_l} \geq 0$ for $l = 1, \ldots, m$, ($\gamma_{f_i} = 0$ if $f_i$ is convex and $\gamma_{g_l} = 0$ if $g_l$ is convex).

## 3.2 Line Search

Let $(d^h, v^h)$ be a solution of (9). We perform the following two-point line search strategy, which will detect discontinuities in the gradients of the objective functions. We assume that $m_L \in (0, \frac{1}{2})$, $m_R \in (m_L, 1)$ and $\bar{t} \in (0, 1]$ are some fixed line search parameters. First, we search for the largest number $t^h_L \in [0, 1]$ such that

$$\max_{i=1,\ldots,k}[f_i(x^h + t^h_L d^h) - f_i(x^h)] \leq m_L t^h_L v^h, \quad \text{and}$$
$$\max_{l=1,\ldots,m}[g_l(x^h + t^h_L d^h)] \leq 0.$$

If $t^h_L \geq \bar{t}$, we take a *long serious step*:

$$x^{h+1} = x^h + t^h_L d^h \quad \text{and} \quad y^{h+1} = x^{h+1},$$

if $0 < t^h_L < \bar{t}$, then we take a *short serious step*:

$$x^{h+1} = x^h + t^h_L d^h \quad \text{and} \quad y^{h+1} = x^h + t^h_R d^h$$

and if $t^h_L = 0$, we take a *null step*:

$$x^{h+1} = x^h \quad \text{and} \quad y^{h+1} = x^h + t^h_R d^h,$$

where $t^h_R > t^h_L$ is such that

$$-\beta^{h+1}_{f_i,h+1} + (\xi^{h+1}_{f_i})^T d^h \geq m_R v^h.$$

The iteration is terminated when

$$-\tfrac{1}{2}v^h < \varepsilon_s,$$

where $\varepsilon_s > 0$ is an accuracy parameter supplied by the user.

### 3.3 Algorithm

Next we aggregate the previous subsections and present the algorithm of the multi-objective proximal bundle method.

**Algorithm 1. MPB**

Step 1. (*Initialization*) Select a feasible starting point $x^1 \in S$, a final accuracy tolerance $\varepsilon_s > 0$, an initial weight $u^1 > 0$, line search parameters $m_L \in (0, \frac{1}{2})$, $m_R \in (m_L, 1)$ and $\bar{t} \in (0, 1]$. Choose the distance measure parameters $\gamma_{f_i} \geq 0$ for $i = 1, \ldots, k$ and $\gamma_{g_l} \geq 0$ for $l = 1, \ldots, m$, ($\gamma_{f_i} = 0$ if $f_i$ is convex and $\gamma_{g_l} = 0$ if $g_l$ is convex). Set $h := 1$, $y^1 := x^1$ and calculate $\xi_{f_i}^1 \in \partial f_i(y^1)$ for $i = 1, \ldots, k$ and $\xi_{g_l}^1 \in \partial g_l(y^1)$ for $l = 1, \ldots, m$.

Step 2. (*Direction finding*) Solve the problem (9) in order to get the solution $(d^h, v^h)$.

Step 3. (*Stopping criterion*) If $-\frac{1}{2}v^h < \varepsilon_s$, then STOP.

Step 4. (*Line search*) Find the step sizes $t_L^h \in [0, 1]$ and $t_R^h \in [t_L^h, 1]$. Set

$$x^{h+1} = x^h + t_L^h d^h \quad \text{and} \quad y^{h+1} = x^h + t_R^h d^h.$$

Step 5. (*Updating*) Set $h := h + 1$, calculate $\xi_{f_i}^h \in \partial f_i(y^h)$ for $i = 1, \ldots, k$ and $\xi_{g_l}^h \in \partial g_l(y^h)$ for $l = 1, \ldots, m$. Choose $J^h \subseteq \{1, \ldots, h\}$ and update the weight $u^h$. Go to Step 2.

The subgradient aggregation strategy due to [8] is used to bound the storage requirements (i.e., the size of the index set $J^h$). We use the line search algorithm of [13] to produce the step-sizes $t_L^h$ and $t_R^h$ in Step 4, and a modification of the weight updating algorithm of [9] is used to update the weight $u^h$ in Step 5.

## 4 Convergence Analysis

Next we give two important convergence results. First we prove, that for $f^\circ$-pseudoconvex functions Algorithm 1 produces a global weak Pareto optimum of the problem (1), while in more general case it ends up with a substationary point.

**Theorem 6** *Let $f_i$ and $g_j$ be $f^\circ$-pseudoconvex and weakly semismooth functions for all $i = 1, \ldots, k$ and $j = 1, \ldots, m$, and the constraint qualification (4) be valid. If Algorithm 1 stops with a finite number of iterations, then the solution is a global weak Pareto optimum of (1). On the other hand, any accumulation point of an infinite sequence of solutions generated by Algorithm 1 is global weak Pareto optimum of (1).*

*Proof* Due to Theorem 3 the improvement function $H$ is $f^\circ$-pseudoconvex. The formulation of Algorithm 1 implies, that it is equivalent to the proximal bundle algorithm applied to unconstraint single objective optimization of $H$. According

to the convergence analysis of the standard proximal bundle algorithm (see, e.g., [9, 22]) if it stops with a finite number of iterations, then the solution $x^h$ is a substationary point of a weakly semismooth $H$, in other words $0 \in \partial H(x^h, x^h)$. Then by Theorem 2 function $H$ attains its global minimum at $x^h$. Since every $f^\circ$-pseudoconvex function is also $f^\circ$-quasiconvex, the first assertion follows from Theorem 5. The proof of the case, when Algorithm 1 generates an infinite sequence of solutions, goes similarly. □

Note, that in order to guarantee the $f^\circ$-pseudoconvexity of the improvement function $H$ also the constraint functions $g_j$ are supposed to be $f^\circ$-pseudoconvex in Theorem 6 although only the $f^\circ$-quasiconvexity was required in Theorem 5.

Finally we show, that in more general case the algorithm produces substationary points of the problem (1).

**Theorem 7** *Let the functions of* (1) *be weakly semismooth. If Algorithm 1 stops with a finite number of iterations, then the solution is a substationary point (i.e. satisfies the necessary optimality condition* (5)*). On the other hand, any accumulation point of an infinite sequence of solutions generated by Algorithm 1 is a substationary point.*

*Proof* The proof is analogous to that of Theorem 6. □

## 5 Numerical Experiments

The efficiency and the reliability of the method is shown by some numerical experiments. Algorithm 1 was implemented in Fortran 77. The test runs have been performed on an Intel® Core™ 2 Duo CPU E8400 (3.00 GHz, 2.99 GHz) PC computer.

### *5.1 General Tests*

We wanted to test the method in different functions classes. First we formulated several $f^\circ$-pseudoconvex objective functions. Next we combined $f^\circ$-pseudoconvex functions with classical convex test examples from [1]. Finally some nonconvex test examples [1] being not $f^\circ$-pseudoconvex nor $f^\circ$-quasiconvex were solved. Furthermore, some $f^\circ$-quasiconvex constraint functions were used in all the test examples. Thus the used function classes were

1. $f^\circ$-pseudoconvex objective functions
2. $f^\circ$-pseudoconvex + convex objective functions
3. Non(generalized)convex objective functions.

In all the test cases the number of variables $n$ varied from 2 to 4, the number of objective functions $k$ from 2 to 4, and the number of constraint functions $m$ from 0 to 2. The numerical results are presented in Table 1, where the first column refers

**Table 1** Computational results

| Test class | # Problems | Iterations | Func. calls |
|---|---|---|---|
| 1 | 36 | 5.1 | 6.7 |
| 2 | 70 | 10.4 | 15.4 |
| 3 | 6 | 8.7 | 13.2 |
| All | 112 | 8.6 | 12.5 |

to the above mentioned test classes, the second column tells the number of the test problems in the class. Finally, the last two columns are devoted to the average of the used iterations and function evaluations, respectively. The last line summarizes the overall average numbers. The parameters of MPB were tuned as follows: $\varepsilon_s = 10^{-5}$, $m_L = 0.01$, $m_R = 0.5$, $\bar{t} = 0.01$, $\gamma_{f_i} = 0.5$ (0 for convex objectives) for $i = 1, \ldots, k$ and $\gamma_{g_l} = 0.5$ for $l = 1, \ldots, m$. The initial weight was chosen by

$$u^1 = \frac{1}{k} \sum_{i=1}^{k} \|\xi_{f_i}^1\|.$$

In order to summarize the numerical results reported in Table 1 we can state that MPB method seems to be reliable and efficient in all the test classes. The reason why it needed more resources in class 2 with convex problems is the complexity of some single test problems.

## *5.2 Numerical Example*

In order to illustrate the functioning of MPB in more detail we consider the following problem:

$$\begin{cases} \text{minimize} & f_1(\mathrm{x}) = \sqrt{\|\mathrm{x}\| + 2} \\ & f_2(\mathrm{x}) = \max\left\{-x_1 - x_2, -x_1 - x_2 + x_1^2 + x_2^2 - 1\right\} \\ \text{subject to} & g(\mathrm{x}) = \max\left\{x_1^2 + x_2^2 - 10, 3x_1 + x_2 + 1.5\right\} \leq 0, \end{cases}$$

**Table 2** Results of the numerical example

| $h$ | $x^h$ | $(f_1(x), f_2(x))$ | Accuracy |
|---|---|---|---|
| 0 | $(-0.5000000, -0.5000000)$ | $(1.645329, 1.000000)$ | $0.5181928$ |
| 1 | $(-0.4153649, -0.3124033)$ | $(1.587367, 0.7277682)$ | $0.9826704 \times 10^{-2}$ |
| 2 | $(-0.4360219, -0.2067399)$ | $(1.575612, 0.6427618)$ | $0.3053751 \times 10^{-2}$ |
| 3 | $(-0.4641460, -0.1123331)$ | $(1.574022, 0.5764790)$ | $0.4805499 \times 10^{-3}$ |
| 4 | $(-0.4622420, -0.1137555)$ | $(1.573542, 0.5759975)$ | $0.4842027 \times 10^{-4}$ |
| 5 | $(-0.4620497, -0.1138994)$ | $(1.573493, 0.5759491)$ | $0.4867036 \times 10^{-5}$ |

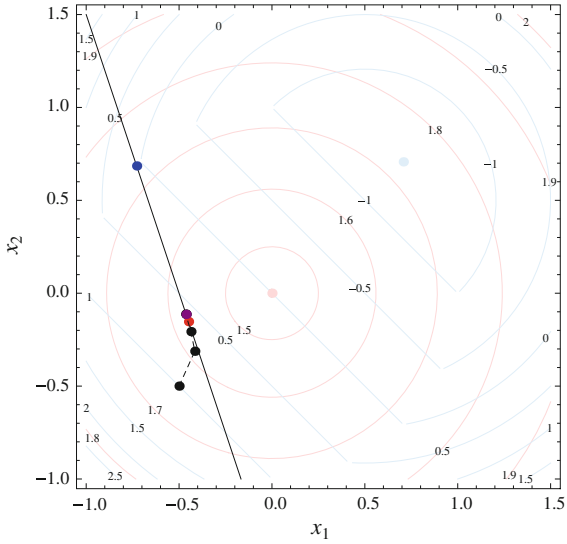



**Fig. 1** Iteration points of the numerical example

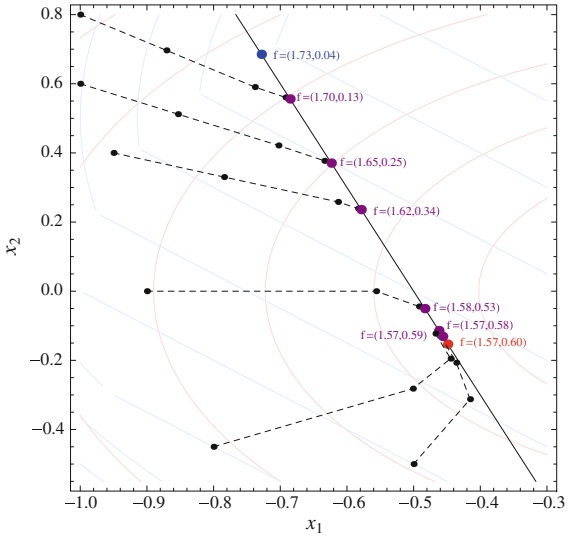



**Fig. 2** Results with several starting points

where $f_1$ is clearly $f^\circ$-pseudoconvex (see [1]), $f_2$ is convex and $g$ convex and thus $f^\circ$-quasiconvex. We used first the starting point $\mathrm{x}^1 = (-0.5, -0.5)$ and the solution iteration by iteration is reported in Table 2.

The numerical results are depicted in Fig. 1, where red and blue colors refer to $f_1$ and $f_2$, respectively. Together with the iteration points (black, final solution violet) there can be seen also the unconstrained and constrained optima and contour lines of the objectives. Note, that the Pareto optimal solutions lie on the line segment between red and blue points.

In Fig. 2 we illustrate the functioning of MPB by starting the optimization from several starting points. According to the character of the method, MPB projects those points to the Pareto optimal set by using the Chebyshev metric.

## 6 Conclusions

We have derived a multiobjective version of the proximal bundle method for nonsmooth and nonconvex optimization. The objective functions are treated individually without employing any scalarization. The method is globally convergent and descent, and under some generalized convexity assumptions it can be proved to find globally weakly Pareto optimal solutions. This kind of method is needed in many application areas. Especially, it can be used as a part of interactive multiobjective optimization methods producing efficiently (weakly) Pareto optimal counterparts of nonoptimal solutions [16, 17, 20] or, by choosing several starting points, generating a good spread of (weakly) Pareto optimal solutions.

**Acknowledgments** This work is financially supported by the University of Turku and the Vilho, Yrjö and Kalle Väisälä foundation.

## References

1. A. Bagirov, N. Karmitsa, M.M. Mäkelä, *Introduction to Nonsmooth Optimization: Theory, Practice and Software* (Springer, Heidelberg, 2014). doi:10.1007/978-3-319-08114-4
2. N.V. Banichuk, P. Neittaanmäki, *Structural Optimization with Uncertainties*, vol. 162, Solid Mechanics and Its Applications (Springer, Berlin, 2010)
3. F.H. Clarke, *Optimization and Nonsmooth Analysis* (Wiley, New York, 1983)
4. M. Gaudioso, M.F. Monaco, Quadratic approximations in convex nondifferentiable optimization. SIAM J. Control Optim. **29**(1), 58–70 (1991)
5. J. Haslinger, P. Neittaanmäki, *Finite Element Approximation for Optimal Shape, Material and Topology Design* (Wiley, Chichester, 1996)
6. J.-B. Hiriart-Urruty, New concepts in nondifferentiable programming. Bull. Soc. Math. France Mém **60**, 57–85 (1979)
7. K.C. Kiwiel, A descent method for nonsmooth convex multiobjective minimization. Large Scale Syst. **8**(2), 119–129 (1985)
8. K.C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, vol. 1133, Lecture Notes in Mathematics (Springer, Berlin, 1985)
9. K.C. Kiwiel, Proximity control in bundle methods for convex nondifferentiable minimization. Math. Program. **46**(1), 105–122 (1990)
10. C. Lemaréchal. *Nondifferentiable Optimization*, eds. by G.L. Nemhauser, A.H.G.Rinnooy Kan, M.J. Todd. Optimization (North-Holland, Amsterdam, 1989), pp. 529–572
11. L. Lukšan, J. Vlček, Globally convergent variable metric method for convex nonsmooth unconstrained minimization. J. Optim. Theory Appl. **102**(3), 593–613 (1999)
12. M.M. Mäkelä, Survey of bundle methods for nonsmooth optimization. Optim. Methods Softw. **17**(1), 1–29 (2002)
13. M.M. Mäkelä, P. Neittaanmäki, *Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control* (World Scientific, Singapore, 1992)

14. M.M. Mäkelä, V.-P. Eronen, N. Karmitsa, On nonsmooth optimality conditions with generalized convexities. TUCS Technical Reports 1056, Turku Centre for Computer Science, Turku, 2012
15. M.M. Mäkelä, V.-P. Eronen, N. Karmitsa, On nonsmooth multiobjective optimality conditions with generalized convexities, in *Optimization in Science and Engineering*, ed. by ThM Rassias, C.A. Floudas, S. Butenko (Springer, New York, 2014), pp. 333–357
16. K. Miettinen, *Nonlinear Multiobjective Optimization* (Kluwer, Boston, 1999)
17. K. Miettinen, M.M. Mäkelä, Interactive bundle-based method for nondifferentiable multiobjective optimization: NIMBUS. Optimization **34**(3), 231–246 (1995)
18. R. Mifflin, A modification and an extension of Lemarechal's algorithm for nonsmooth minimization. Math. Program. Stud. **17**, 77–90 (1982)
19. J.J. Moreau, P.D. Panagiotopoulos, G. Strang (eds.), *Topics in Nonsmooth Mechanics* (Birkhäuser, Basel, 1988)
20. H. Mukai, Algorithms for multicriterion optimization. IEEE Trans. Autom. Control **25**(2), 177–186 (1980)
21. J. Outrata, M. Kočvara, J. Zowe, Nonsmooth approach to optimization problems with equilibrium constraints. *Theory, Applications and Numerical Results* (Kluwer, Dordrecht, 1998)
22. H. Schramm, J. Zowe, A version of the bundle idea for minimizing a nonsmooth functions: conceptual idea, convergence analysis, numerical results. SIAM J. Optim. **2**, 121–152 (1992)
23. R.E. Steuer, *Multiple Criteria Optimization: Theory, Computation, and Application* (Wiley, New York, 1986)
24. S. Wang, Algorithms for multiobjective and nonsmooth optimization. In *Methods of Operations Research 58* (Athenäum, Frankfurt am Main, 1989), pp. 131–142

# Efficient Parallel Nash Genetic Algorithm for Solving Inverse Problems in Structural Engineering

**Jacques Périaux and David Greiner**

**Abstract** A parallel implementation of a game-theory based Nash Genetic Algorithm (Nash-GAs) is presented in this paper for solving reconstruction inverse problems in structural engineering. We compare it with the standard panmictic genetic algorithm in a HPC environment with up to eight processors. The procedure performance is evaluated on a fifty-five bar sized test case of discrete real cross-section types structural frame. Numerical results obtained on this application show a significant achieved increase of performance using the parallel Nash-GAs approach compared to the standard GAs or Parallel GAs.

**Keywords** Parallel genetic algorithms · Nash games · Structural optimization · Finite element analysis · Inverse problems · Bar structures

**Mathematical Subject Classification:** 74P99 · 68Y20 · 49M27

## 1 Introduction

Hybridization of game theory based methods with evolutionary optimization algorithms has been used to enhance the performance of optimizers in hilly search spaces, both in single-objective and multi-objective optimization problems. Particularly, in the case of aeronautical engineering problems the use of Nash based

J. Périaux (✉)
Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland
e-mail: jperiaux@gmail.com

J. Périaux
International Center for Numerical Methods in Engineering (CIMNE), Universidad Politécnica de Cataluña (UPC), Barcelona, Spain

D. Greiner
Institute of Intelligent Systems and Numerical Applications in Engineering (SIANI), Universidad de Las Palmas de Gran Canaria (ULPGC), 35017 Las Palmas, Spain
e-mail: david.greiner@ulpgc.es



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_13

algorithms and its parallelization have improved the efficiency of these methods, where the computation of the fitness function (objective function or cost function in terms of evolutionary computation) is associated frequently with high computational CPU costs in detailed design (see, e.g., Périaux et al. [15, 16]). A summary paper of parallel evolutionary algorithms in CFD applications can be found in [1], where it is explained that excepted the master-slave model (only acting as a hardware accelerator), the parallelization can change significantly the algorithm behaviour and convergence.

In this paper, we introduce and show the advantages of the use of a parallel implementation of the Nash evolutionary algorithm [10–12, 17] to speed up solving inverse problems in structural engineering. In Sect. 2 we briefly describe the Nash evolutionary algorithm, while Sect. 3 introduces the structural engineering problem to be optimized. In Sect. 4, the structural test case used in this work is implemented on a HPC environment. Section 5 covers the results and their analysis, and finally in Sect. 6 conclusions of the research are outlined.

## 2 Nash Evolutionary Algorithms

Nash evolutionary algorithms were introduced by Sefrioui and Periaux [17] for solving Computational Fluid Dynamics (CFD) optimization problems. They are based on hybridizing the mathematical concepts of Nash equilibrium [13, 14] with the nature inspired search procedure.

Nash equilibrium is a symmetric competitive game where players maximize their payoffs while taking into account the strategies of their competitors. Therefore, a set of sub-populations co-evolve simultaneously each of which deals only with a partition of the design variables. These subpopulations interact to evolve towards the Nash equilibrium; in the case of a single objective problem, a virtual Nash game approach can also be applied in inverse shape optimization CFD problems as a speed up technique versus the standard panmictic evolutionary algorithm [10, 11].

In a Nash Evolutionary Algorithm, the solution candidates (chromosomes) of each subpopulation, only include as genotypic information used in the mating and mutation processes, those belonging to its partition (player in terms of game theory vocabulary); the rest of the genotypic information is introduced mandatorily by the best solution obtained by the other subpopulations in the previous generation, as shown in Fig. 1. Therefore a search space partitioning in the sense of evolutionary optimization is performed, which results in an accelerated convergence (as commonly accepted in the evolutionary computation community: the shorter the chromosome length used, the faster the convergence curve achieved).

This approach has been successfully applied in the case of inverse problems where the fitness function objective is a sum of separable terms (such as the case of many

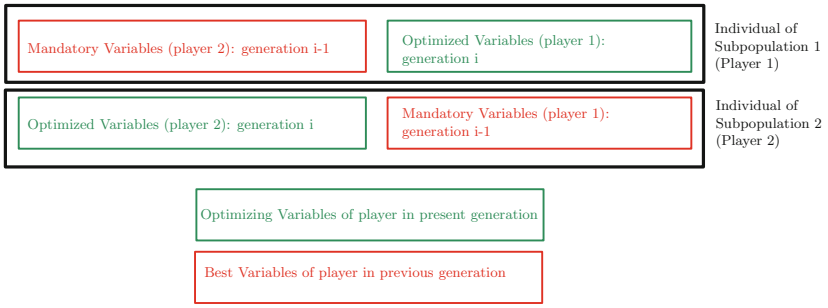



**Fig. 1** Nash genetic algorithm population scheme; 2 players case

shape optimization problems). The use of these Nash evolutionary algorithms in the case of structural engineering problems has been introduced successfully by Greiner et al. [5, 6].

In this paper, the above innovative strategies are applied to capture the solution of a structural engineering problem with a parallel implementation of the Nash evolutionary algorithm and evaluate the particular gain due to its parallelization.

## 3 Structural Problem

Structural inverse problems are handled in this research. The objective is to obtain the structure which most fits the maximum stresses of reference. The optimum structural bar design is defined as a design in which some location of every bar member in the structure has a maximum stress value as accurately equal as the maximum stress of reference for that bar. The fitness function is

$$\text{fitness function} = \min \sqrt{\sum_{i=1}^{Nbars} (\sigma_{\max,i} - \sigma_{\max,R_i})^2}, \tag{1}$$

where $\sigma_{\max,i}$ is the maximum calculated stress and $\sigma_{\max,R_i}$ the maximum stress of reference, both corresponding to bar $i$, and *Nbars* is the number of bars. Variable search space is constrained and described in Sect. 4 (see Table 8 for the particular test case handled in this paper); no additional constraints are present in this optimization problem. Stresses of the structural bars are calculated using a direct stiffness matrix calculation program (with Hermite approximation functions), which implies the resolution of its associated linear equation system (see, e.g., [9]).

A value of zero means a perfect fit in maximum stresses between the searched solution and the reference solution. In the special case of defining as maximum stresses of reference, the set of a previously known structural design, then the prob-

**Table 1** Overview of geometry parameters

| Geometry parameters | Value (m) |
|---|---|
| Column length (Height)—$H$ | 2.8 |
| Beam length (Width)—$W$ | 5.6 |

lem is a reconstruction inverse (RI) problem. Results corresponding to this type of problem are shown and discussed in Sect. 5.

## 4 Test Case Definition

The purpose of this benchmark is to set a fifty-five bars sized frame structure test case in the context of inverse and/or optimization problems, defined by Greiner et al. [4]. This test case has already been used considering the problem of weight minimization, and following the template introduced in the Finnish design test case database [18].

We consider the case of discrete cross-section type variables. In this structural problem, the variables of the chromosome are the type of cross-section of each bar, which can be selected from a fixed database of normalized cross-sections; in this test case, the standard European IPE (e.g., IPE-100, IPE-120, IPE-200, etc.) and HEB (e.g., HEB-200, HEB-300, etc.) types are selected. As these variables are not continuous, but discrete, they are coded in a discrete manner using 1 s and 0 s (binary codification type, particularly binary gray codification); a set size of 16 cross-sections is assigned for each bar. Table 8 shows the detailed search space used.

The main objective of this benchmark is to test and compare different optimization approaches for structural design. We refer to Greiner et al. [3, 7].

The computational domain, boundary conditions, loadings and design variable numbering are shown in Fig. 2. Boundary conditions consist in embedded connections at the bottom joining the columns of the structure to the ground (horizontal and vertical displacements and in-plane rotations are not allowed). Calculation of the bar stresses are evaluated through a stiffness matrix calculation software (bar frames).

Geometry parameters, height $H$ and width $W$, are given in Table 1, and material properties: density, Young modulus, and maximum stresses of columns and beams, which correspond to standard construction steel, are shown in Table 2. Each bar variable has associated an independent optimization variable whose selection set (or search space) is described in Table 8 (see Appendix).

Description with respect to modeling and physical properties is presented as follows: frame bar structure stiffness matrix calculation (rigid nodes: resisting moment capabilities); elastic behaviour of steel is assumed; no buckling effect is included.

One loading case is included, with nodal loads as in Fig. 2 (values in tons) and with vertical uniform load at every beam $D = 39{,}945$ N/m. The own weight of steel bars has also to be considered into account.

**Table 2** Material properties (standard steel)

| Parameter | Value |
|---|---|
| Density | 7850 kg/m$^3$ |
| Young modulus | $2.1 \times 10^5$ MPa |
| Maximum stress (columns) | 200 MPa |
| Maximum stress (beams) | 220 MPa |

Stresses constraints (MPa) are:

$$\sigma_{\max,i} \leq 200, \qquad \forall i, \text{ with } i = 1, \ldots, N \textit{ columns},$$
$$\sigma_{\max,i} \leq 220, \qquad \forall i, \text{ with } i = 1, \ldots, N \textit{ beams},$$

The quantities of interest are:

- Values of the fitness function as described in Sect. 3,
- Cross-section type sizing for each bar, and,
- Maximum stress of each bar.

With respect to the inverse reconstruction problem, the cross-section types corresponding to the design shown in Table 7 (see Appendix) have been taken as reference, where also the correspondent maximum value of each bar stress is shown in Table 9 (see Appendix). The stresses of reference are those belonging to the real design corresponding to IPE330 in all beams (bars 1 to 25) and HEB450 in all columns (bars 26 to 55).

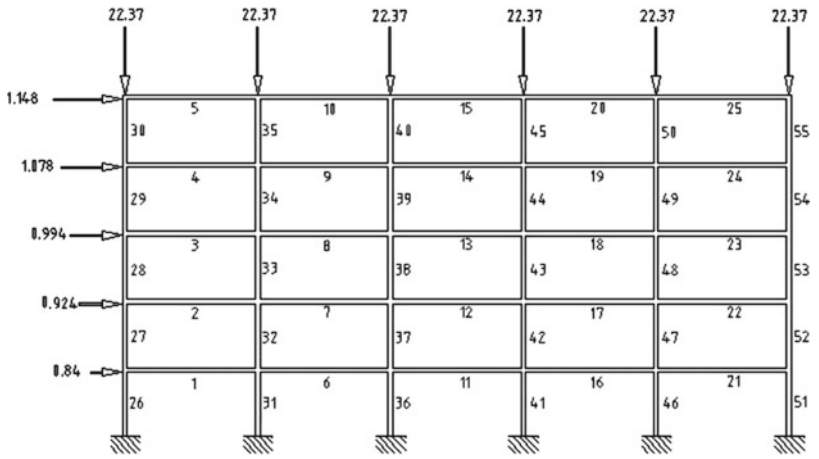



**Fig. 2** Computational domain, boundary conditions and loadings available in [4]

# 5 Results and Discussion

## 5.1 Nash Variable Territory Splitting

Two different split of territories used by Nash players have been compared in the experimental results:

1. A two players Nash partition (two subpopulations) where the split territory is set among beams and columns (Nash 2pl) is used as shown in Fig. 3. Therefore, beams are optimized by player—subpopulation P1 and columns are optimized by player—subpopulation P2. This territory splitting does not include boundary conditions used by each player.
2. A three players Nash partition (three subpopulations) where the split territory is set among left-center-right (Nash 3pl) has been used as shown in Fig. 4. Therefore, the left part is optimized by player—subpopulation P1, the center part is optimized by player—subpopulation P2 and the right part is optimized by player—subpopulation P3. This territory splitting includes boundary conditions used by each player.

These algorithm configurations are also compared with the standard panmictic Evolutionary Algorithm, where no splitting of territory and a whole single population are considered for numerical experiments.

## 5.2 Experiment Definition: Nash Genetic Algorithm

Statistics obtained from one hundred independent executions for each case will be considered to measure the relative increased performance of the game theory based evolutionary algorithm approach versus the standard approach.

Numerical results corresponding to a population size of 100 individuals, uniform crossover, a mutation rate of 0.4 %, an elitist generational replacement strategy keeping the two best individuals, and a stopping criterion of 100,000 function evaluations are considered here. Gray coding is used, in accordance with its good behaviour in structural frame optimum design (see, e.g., [2, 8]).

The performance of the panmictic GA strategy versus the Nash GA strategies (as described in Sect. 5.1) has been compared. In each case, also results from a standard panmictic population are shown for comparison, each of them executed 100 independent times to perform the following statistics and figures.

A Parallel Nash-GA has been implemented with MPI-C/C++ and the three algorithms are tested: Standard Panmictic GA, Nash GA 2 players (beam-column partitioning), Nash GA 3 players (left-center-right partitioning). For each, eight configurations of central processing unit (CPU) are taken into account, in a master-slave parallel evolutionary algorithm configuration, as follows:

- Sequential time (only 1 CPU): Seq.
- 1 Master + 1 Slave (2 CPU): 1M + 1S

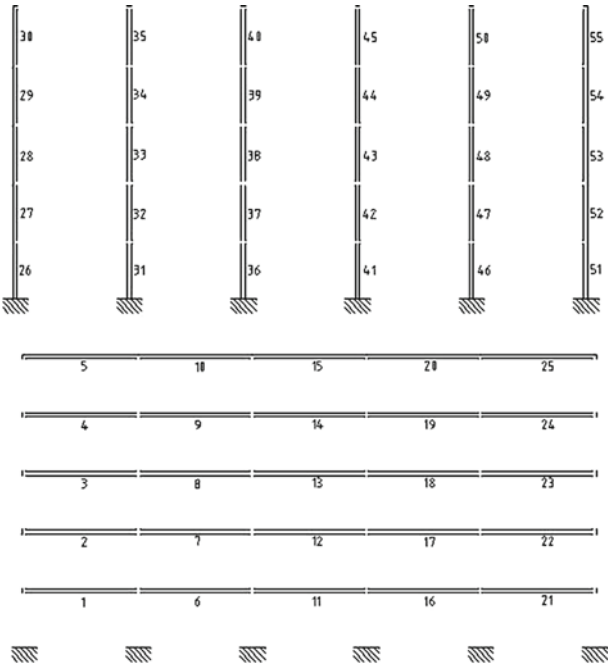



**Fig. 3** A beam-column territory domain decomposition of Nash GA (Nash 2pl)

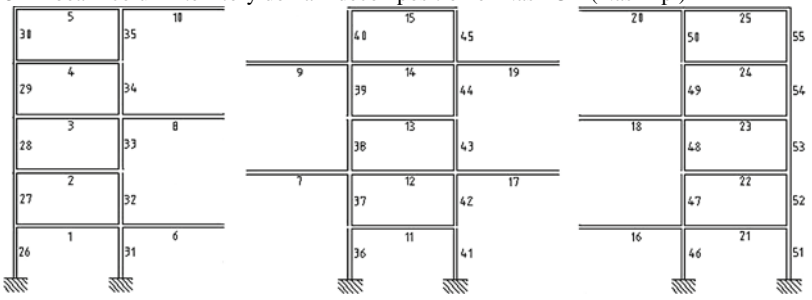



**Fig. 4** A *left-center-right* domain territory decomposition of three partitioning Nash GA (Nash 3pl)

- 1 Master + 2 Slaves (3 CPU): 1M + 2S
- 1 Master + 3 Slaves (4 CPU): 1M + 3S
- 1 Master + 4 Slaves (5 CPU): 1M + 4S
- 1 Master + 5 Slaves (6 CPU): 1M + 5S
- 1 Master + 6 Slaves (7 CPU): 1M + 6S
- 1 Master + 7 Slaves (8 CPU): 1M + 7S

An Intel Core i7-3770-3.40 GHz processor (four cores and eight threads with enabled hyper-threading technology) is used as a hardware platform.

**Table 3** Comparison of CPU time required to achieve the best solution (null value)

| Time (seconds) | Panmictic GA | | Nash GA 2 players | | Nash GA 3 players | |
|---|---|---|---|---|---|---|
| | Average | Std Dev | Average | Std Dev | Average | Std Dev |
| Seq. | 5.654 | 1.158 | 3.348 | 0.373 | 3.044 | 0.384 |
| 1M+1S | 5.456 | 1.102 | 3.465 | 0.383 | 3.062 | 0.361 |
| 1M+2S | 2.871 | 0.583 | 1.804 | 0.249 | 1.538 | 0.175 |
| 1M+3S | 2.006 | 0.362 | 1.220 | 0.130 | 1.091 | 0.139 |
| 1M+4S | 2.213 | 0.485 | 1.403 | 0.182 | 1.206 | 0.164 |
| 1M+5S | 2.032 | 0.503 | 1.154 | 0.137 | 1.046 | 0.131 |
| 1M+6S | 1.684 | 0.368 | 0.991 | 0.123 | 0.881 | 0.107 |
| 1M+7S | 1.386 | 0.249 | 0.837 | 0.106 | 0.744 | 0.085 |

**Table 4** Comparison of number of whole total fitness function evaluations required to achieve the best solution (null value)

| Number of evaluations | Panmictic GA | | Nash GA 2 players | | Nash GA 3 players | |
|---|---|---|---|---|---|---|
| | Average | Std Dev | Average | Std Dev | Average | Std Dev |
| Seq. | 52039.0 | 10655.3 | 30793.6 | 3460.4 | 28312.3 | 3590.1 |
| 1M+1S | 50321.1 | 10136.0 | 31824.6 | 3517.9 | 28271.2 | 3361.5 |
| 1M+2S | 51425.5 | 10425.9 | 31315.0 | 4127.1 | 27468.0 | 3110.2 |
| 1M+3S | 51628.4 | 9407.0 | 31260.1 | 3382.3 | 27962.5 | 3529.3 |
| 1M+4S | 50770.9 | 10266.9 | 32142.1 | 3563.5 | 27392.5 | 3382.3 |
| 1M+5S | 54724.2 | 13518.8 | 31195.4 | 3721.7 | 28053.6 | 3487.2 |
| 1M+6S | 52998.4 | 11633.6 | 31211.1 | 3887.2 | 27545.0 | 3333.1 |
| 1M+7S | 51292.3 | 9354.8 | 31442.4 | 4029.2 | 27724.3 | 3247.4 |

## *5.3 Results: Nash Genetic Algorithm*

Associated to the reconstruction problem, Tables 3, 4, 5 and 6 include the results corresponding to 100 independent executions of the different algorithm strategies. Table 3 shows the average and standard deviation of the time in seconds required to the achievement of the best solution (zero value); Table 4 shows the average and standard deviation of the whole total number of fitness function evaluations required to the achievement of the best solution (zero value); Table 5 shows the average of the number of fitness function evaluations per subpopulation (Nash player) required to the achievement of the best solution (zero value); and Table 6 shows the average of speed-up obtained by the Nash GA algorithms in terms of gain expressed in number of fitness evaluations per subpopulation.

From Table 3, a maximum difference of computing time is obtained when comparing the sequential hardware resources in the standard panmictic GA elapsing an

**Table 5** Average number of fitness evaluations per subpopulation comparison

| Number of evaluations | Panmictic GA | Nash GA 2 players | Nash GA 3 players |
|---|---|---|---|
| Seq. | 52039.02 | 15396.8 | 9437.4 |
| 1M+1S | 50321.08 | 15912.3 | 9423.7 |
| 1M+2S | 51425.54 | 15657.5 | 9156.0 |
| 1M+3S | 51628.4 | 15630.1 | 9320.8 |
| 1M+4S | 50770.9 | 16071.1 | 9130.8 |
| 1M+5S | 54724.22 | 15597.7 | 9351.2 |
| 1M+6S | 52998.44 | 15605.6 | 9181.7 |
| 1M+7S | 51292.26 | 15721.2 | 9241.4 |

average standard deviation time of $5.654 \pm 1.158$ s, versus the case of one master and seven slaves in the Nash GA with 3 players elapsing $0.744 \pm 0.085$ s, implying a time reduction factor of 7.6. When comparing the gain obtained only due to the increased hardware resources (dividing the time of the case of sequential algorithm by the time of the case of one master and seven slaves), panmictic GA, Nash GA 2 players and Nash GA 3 players achieve a speed-up of 4.08, 4.00 and 4.09, respectively.

From Table 4, we observe that the range of whole total fitness function evaluations required by each algorithm type is independent of the available hardware resources, and the variations shown in the table are only due to the stochastic nature of each set of 100 executions per case. Therefore, the maximum and minimum average $\pm$ standard deviation number of fitness functions evaluations of panmictic GA, Nash GA 2 players and Nash GA 3 players, are (50321.1; 54724.2) $\pm$ (9354.8; 13518.8), (30793.6; 32142.1) $\pm$ (3382.3; 4127.1) and (27392.5; 28312.3)$\pm$(3110.2; 3590.1), respectively.

Table 5 includes the number of fitness evaluations considering the number of players used by each algorithm. Therefore, when considering more than one player, (more than one population/subpopulation), the number of evaluations required by each subpopulation decreases in the cases of the Nash GAs, while the case of panmictic GA remains unchanged. The maximum and minimum average number of fitness functions evaluations per subpopulation of Nash GA with 2 players and Nash GA with 3 players, are (15396.8; 16071.1) and (9130.8; 9437.4), respectively.

**Table 6** Average number of fitness evaluations per subpopulation speed-up

| Number of evaluations | Panmictic GA | Nash GA 2 players | Nash GA 3 players |
|---|---|---|---|
| Seq. | 1 | 3.38 | 5.51 |
| 1M+1S | 1 | 3.16 | 5.34 |
| 1M+2S | 1 | 3.28 | 5.62 |
| 1M+3S | 1 | 3.30 | 5.54 |
| 1M+4S | 1 | 3.16 | 5.56 |
| 1M+5S | 1 | 3.51 | 5.85 |
| 1M+6S | 1 | 3.40 | 5.77 |
| 1M+7S | 1 | 3.26 | 5.55 |

Table 6 expresses the previous values of Table 5 in terms of net speed-up with respect to the panmictic GA. The maximum and minimum average gain per subpopulation of Nash GA with 2 players and Nash GA with 3 players, are (3.16; 3.51) and (5.34; 5.85), respectively.

Figures 5, 6 and 7, show for each algorithm type, panmictic GA, Nash GA with 2 players and Nash GA with 3 players, respectively, the relation among the CPU time required to achieve the best solution (null value) and the respective total number of evaluations required. Each dot in these figures corresponds to one of each 100 independent executions which was able to obtain that best solution. The different hardware resources are distinguished inside each figure. Results when comparing the performance of each algorithm with constant hardware resources are shown in Figs. 10, 11, 12, 13, 14, 15, 16 and 17.

Figures 8 and 9 include, respectively, the average and standard deviation out of 100 independent executions, of the whole total number of fitness evaluations versus the total elapsed MPI-time required to achieve the best solution (null value), classified by algorithm type.

## *5.4 Discussion: Nash Genetic Algorithm*

The speed up of the Nash GAs in terms of reduction of number of fitness function evaluations per subpopulation, according to shown results, particularly Tables 4, 5, 6 and Figs. 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, is inherent to the algorithm subpopulation structure and procedure, and independent of the hardware platform of execution. A super-linear speed-up is observed both in the Nash GA with two players (two subpopulations), where a gain between 3.16 and 3.51 is achieved, and in the Nash GA with three players (three subpopulations), where a gain between 5.34 and 5.85 is achieved.

Results of speed up in terms of CPU time, according to shown results, particularly Table 3 and Figs. 5, 6, 7, 10, 11, 12, 13, 14, 15, 16, and 17, should be interpreted

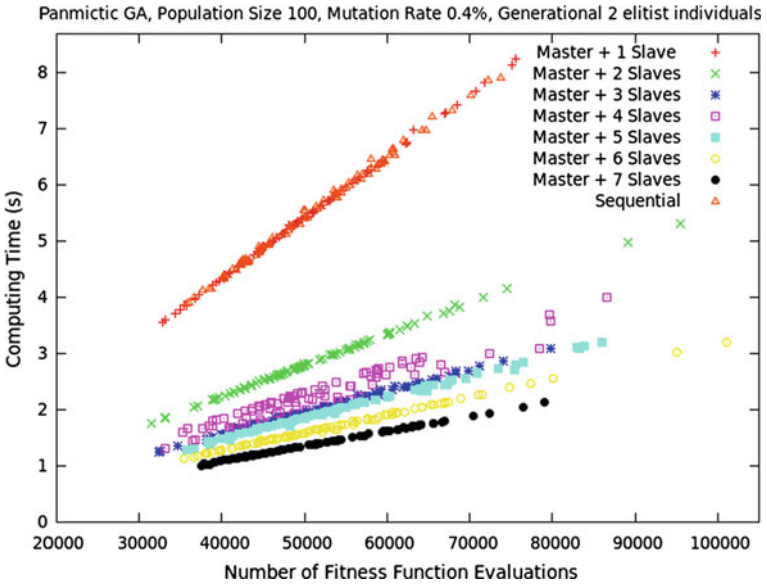



**Fig. 5** Panmictic standard genetic algorithm. Comparing parallel implementations with different hardware resources and the sequential version. 100 independent executions

conditioned by the particular hardware used in this work, an Intel Core i7-3770-3.40 GHz processor (4 cores with 8 threads using hyperthreading technology). Some hardware slowdown can be clearly observed in Fig. 14 (as well as in Figs. 5, 6 and 7), where the fifth thread is activated when increasing the CPU demand from 4 to 5. Nevertheless, as Table 3 gain per algorithm type indicates, finally a calculation time gain of 4 is achieved in all cases (panmictic GA, Nash GA 2 players, Nash GA 3 players) when running the case of one master and seven slaves.

## 6 Conclusions

The performance of a master-slave parallel implementation of Nash genetic algorithms in inverse problems, particularly the reconstruction problem, in structural engineering, has been tested in a fifty-five bar sized frame test case. It has shown an important increased speed-up, even achieving super-linear gains in terms of the fitness function evaluations per player/subpopulation when compared with the standard panmictic genetic algorithm.

In addition, it has been considered the benefits coming from enhanced parallel implementation capability of this type of algorithms, where significant CPU time reduction has been attained.

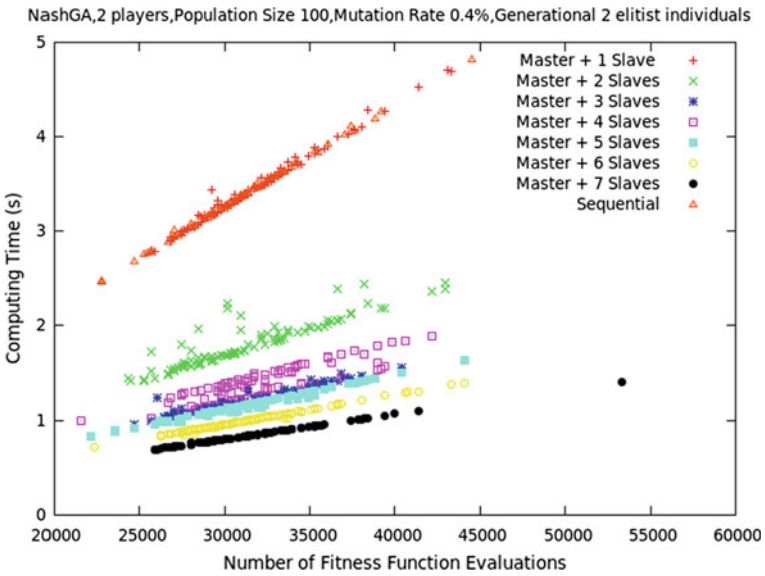



**Fig. 6** Nash genetic algorithm with two players. Comparing parallel implementations with different hardware resources and the sequential version. One hundred independent executions

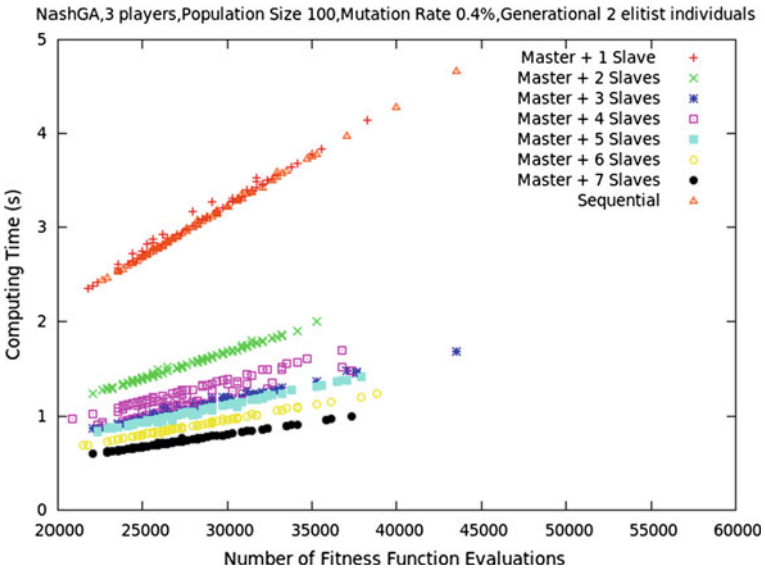



**Fig. 7** Nash genetic algorithm, three players. Comparing parallel implementations with different hardware resources and the sequential version. One hundred independent executions

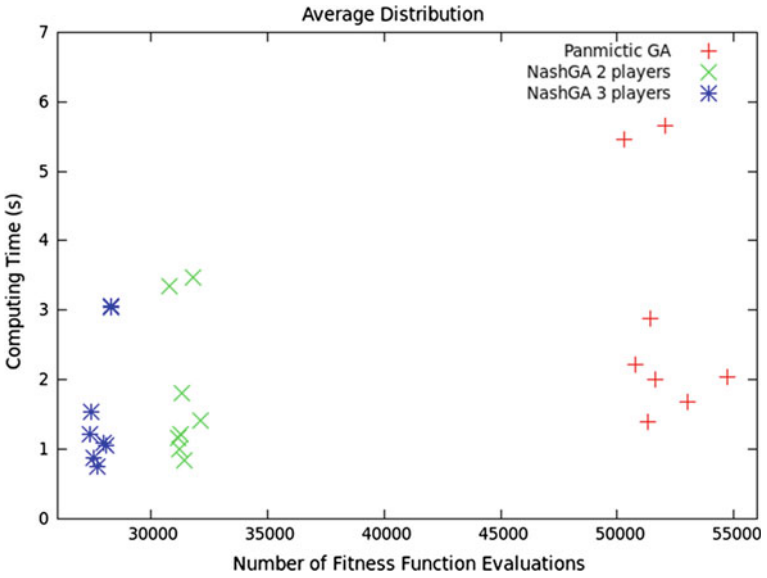



**Fig. 8** Number of fitness function evaluations average required to achieve the best solution (null value), classified by type of algorithm, from one hundred independent executions

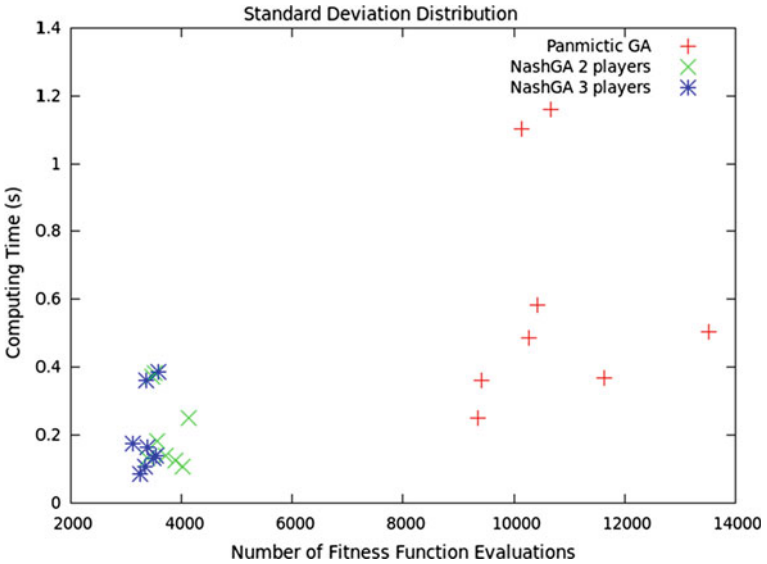



**Fig. 9** Number of fitness function evaluations standard deviation required to achieve the best solution (null value), classified by type of algorithm, from one hundred independent executions

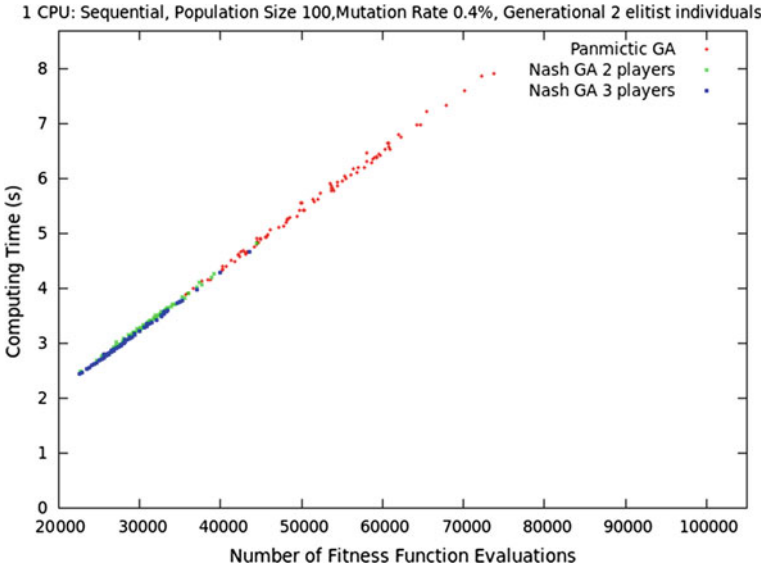



**Fig. 10** Comparing the three algorithms when using only one CPU (sequential computation)

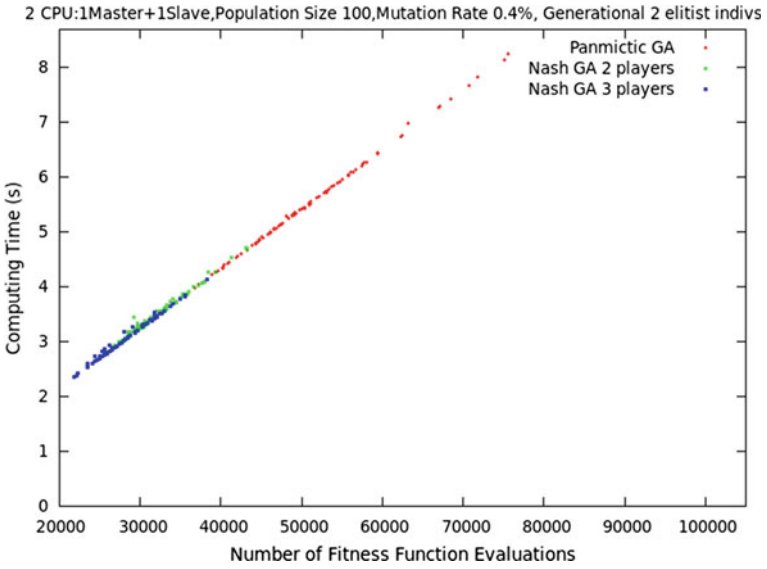



**Fig. 11** Comparing the three algorithms when using two CPUs (master and one slave computation)

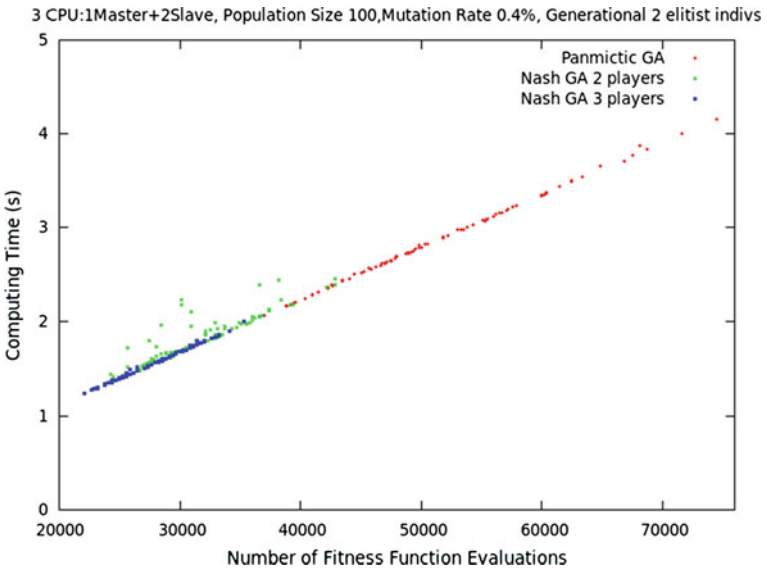



**Fig. 12** Comparing the three algorithms when using three CPU (master and two slaves computation)

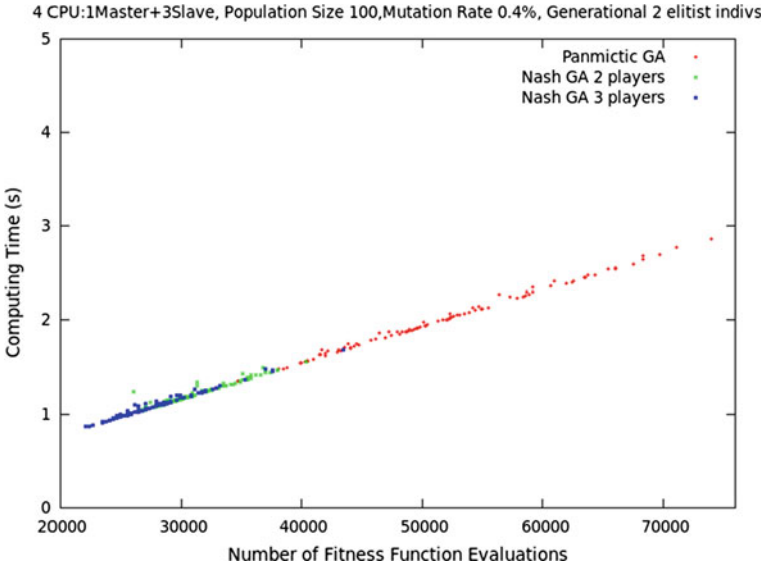



**Fig. 13** Comparing the three algorithms when using four CPUs (master and three slaves computation)

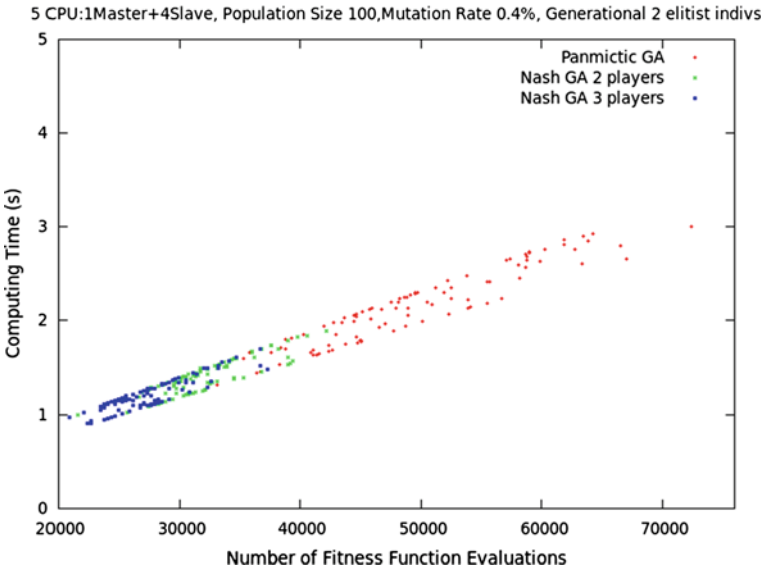



**Fig. 14** Comparing the three algorithms when using five CPUs (master and four slaves computation)

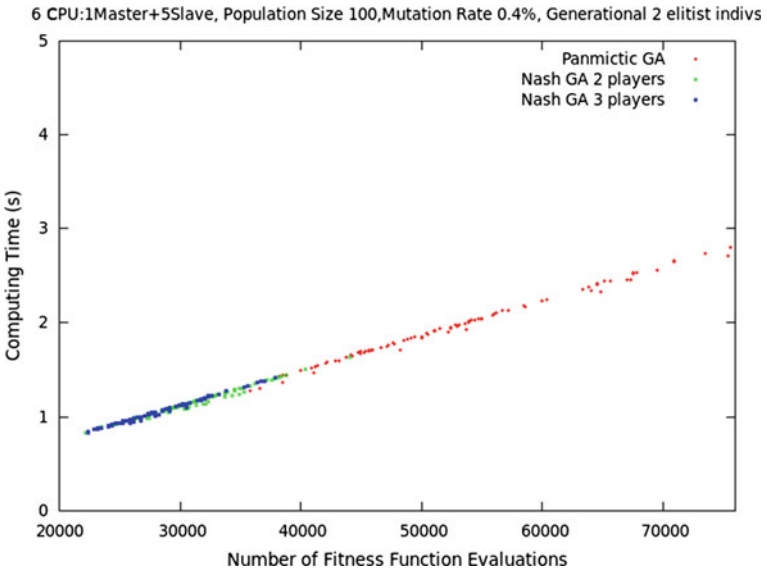



**Fig. 15** Comparing the three algorithms when using six CPUs (master and five slaves computation)

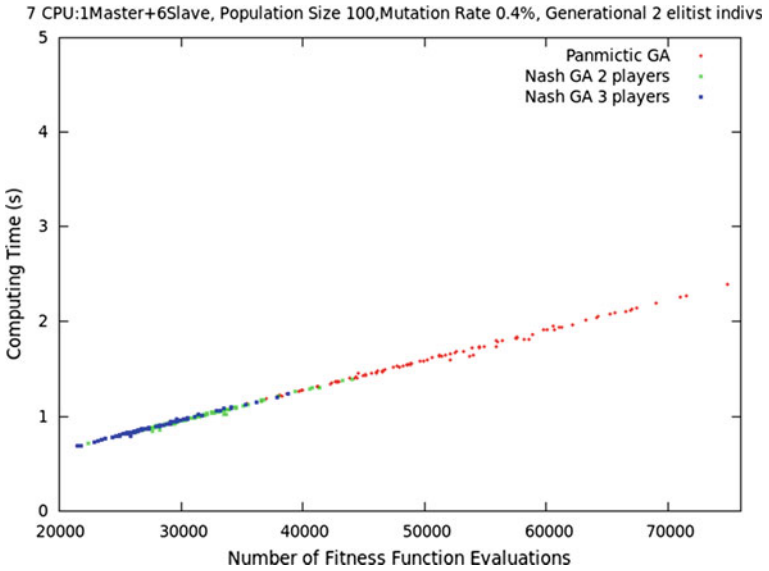



**Fig. 16** Comparing the three algorithms when using seven CPUs (master and seven slaves computation)

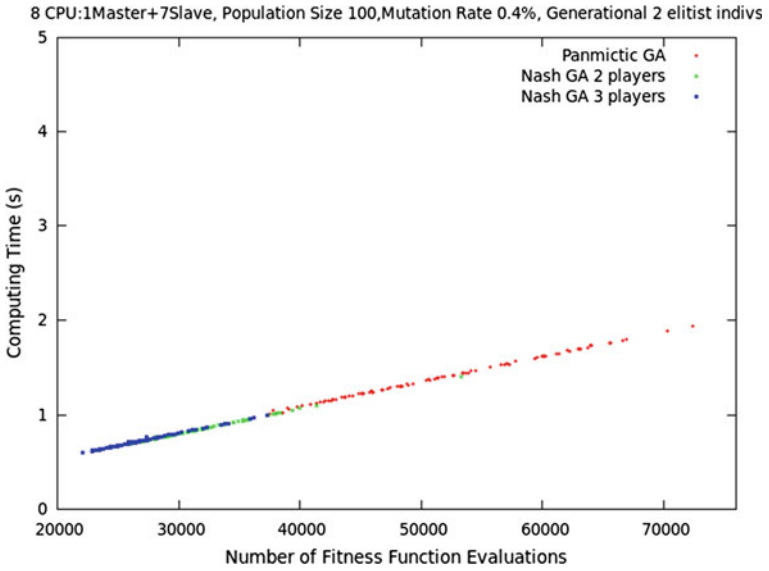



**Fig. 17** Comparing the three algorithms when using eight CPUs (master and eight slaves computation)

**Acknowledgments** This research work is funded through contract CAS12/00400 by Ministerio de Educación, Cultura y Deporte of the Government of Spain, through the Programa Nacional de Movilidad de Recursos Humanos del Plan Nacional de I+D+I 2008-2011 "José Castillejo", extended by agreement of Consejo de Ministros of 7th October 2011. The second author gratefully acknowledges support at the Mathematical Information Technology Department, University of Jyväskylä (Finland) given by, in particular, Prof. Pekka Neittaanmäki.

## Appendix

See Tables 7, 8 and 9.

**Table 7** Design of reference in inverse problem design (cross-section type detail)

| Bar number | Bar 1 | Bar 2 | Bar 3 | Bar 4 | Bar 5 | Bar 6 | Bar 7 | Bar 8 | Bar 9 | Bar 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Cross section | IPE330 | IPE330 | IPE330 | IPE330 | IPE330 | IPE330 | IPE330 | IPE330 | IPE330 | IPE330 |
| Bar number | Bar 11 | Bar 12 | Bar 13 | Bar 14 | Bar 15 | Bar 16 | Bar 17 | Bar 18 | Bar 19 | Bar 20 |
| Cross section | IPE330 | IPE330 | IPE330 | IPE330 | IPE330 | IPE330 | IPE330 | IPE330 | IPE330 | IPE330 |
| Bar number | Bar 21 | Bar 22 | Bar 23 | Bar 24 | Bar 25 | Bar 26 | Bar 27 | Bar 28 | Bar 29 | Bar 30 |
| Cross section | IPE330 | IPE330 | IPE330 | IPE330 | IPE330 | HEB450 | HEB450 | HEB450 | HEB450 | HEB450 |
| Bar number | Bar 31 | Bar 32 | Bar 33 | Bar 34 | Bar 35 | Bar 36 | Bar 37 | Bar 38 | Bar 39 | Bar 40 |
| Cross section | HEB450 | HEB450 | HEB450 | HEB450 | HEB450 | HEB450 | HEB450 | HEB450 | HEB450 | HEB450 |
| Bar number | Bar 41 | Bar 42 | Bar 43 | Bar 44 | Bar 45 | Bar 46 | Bar 47 | Bar 48 | Bar 49 | Bar 50 |
| Cross section | HEB450 | HEB450 | HEB450 | HEB450 | HEB450 | HEB450 | HEB450 | HEB450 | HEB450 | HEB450 |
| Bar number | Bar 51 | Bar 52 | Bar 53 | Bar 54 | Bar 55 | | | | | |
| Cross section | HEB450 | HEB450 | HEB450 | HEB450 | HEB450 | | | | | |

**Table 8** Search space of variables (beams 1–25 and columns 26–55)

| Bar number | Bar variable | Cross-section type set | Bar number | Bar variable | Cross-section type set |
|---|---|---|---|---|---|
| 1 | v1 | From IPE080 to IPE500 | 26 | v26 | From HEB100 to HEB450 |
| 2 | v2 | From IPE080 to IPE500 | 27 | v27 | From HEB100 to HEB450 |
| 3 | v3 | From IPE080 to IPE500 | 28 | v28 | From HEB100 to HEB450 |
| 4 | v4 | From IPE080 to IPE500 | 29 | v29 | From HEB100 to HEB450 |
| 5 | v5 | From IPE080 to IPE500 | 30 | v30 | From HEB100 to HEB450 |
| 6 | v6 | From IPE080 to IPE500 | 31 | v31 | From HEB100 to HEB450 |
| 7 | v7 | From IPE080 to IPE500 | 32 | v32 | From HEB100 to HEB450 |
| 8 | v8 | From IPE080 to IPE500 | 33 | v33 | From HEB100 to HEB450 |
| 9 | v9 | From IPE080 to IPE500 | 34 | v34 | From HEB100 to HEB450 |
| 10 | v10 | From IPE080 to IPE500 | 35 | v35 | From HEB100 to HEB450 |
| 11 | v11 | From IPE080 to IPE500 | 36 | v36 | From HEB100 to HEB450 |
| 12 | v12 | From IPE080 to IPE500 | 37 | v37 | From HEB100 to HEB450 |
| 13 | v13 | From IPE080 to IPE500 | 38 | v38 | From HEB100 to HEB450 |
| 14 | v14 | From IPE080 to IPE500 | 39 | v39 | From HEB100 to HEB450 |
| 15 | v15 | From IPE080 to IPE500 | 40 | v40 | From HEB100 to HEB450 |
| 16 | v16 | From IPE080 to IPE500 | 41 | v41 | From HEB100 to HEB450 |

(continued)

**Table 8** (continued)

| Bar number | Bar variable | Cross-section type set | Bar number | Bar variable | Cross-section type set |
|---|---|---|---|---|---|
| 17 | v17 | From IPE080 to IPE500 | 42 | v42 | From HEB100 to HEB450 |
| 18 | v18 | From IPE080 to IPE500 | 43 | v43 | From HEB100 to HEB450 |
| 19 | v19 | From IPE080 to IPE500 | 44 | v44 | From HEB100 to HEB450 |
| 20 | v20 | From IPE080 to IPE500 | 45 | v45 | From HEB100 to HEB450 |
| 21 | v21 | From IPE080 to IPE500 | 46 | v46 | From HEB100 to HEB450 |
| 22 | v22 | From IPE080 to IPE500 | 47 | v47 | From HEB100 to HEB450 |
| 23 | v23 | From IPE080 to IPE500 | 48 | v48 | From HEB100 to HEB450 |
| 24 | v24 | From IPE080 to IPE500 | 49 | v49 | From HEB100 to HEB450 |
| 25 | v25 | From IPE080 to IPE500 | 50 | v50 | From HEB100 to HEB450 |
| | | | 51 | v51 | From HEB100 to HEB450 |
| | | | 52 | v52 | From HEB100 to HEB450 |
| | | | 53 | v53 | From HEB100 to HEB450 |
| | | | 54 | v54 | From HEB100 to HEB450 |
| | | | 55 | v55 | From HEB100 to HEB450 |

**Table 9** Stresses (MPa $\times 10^{-1}$) of reference in inverse problem design

| Bar number | Bar 1 | Bar 2 | Bar 3 | Bar 4 | Bar 5 | Bar 6 | Bar 7 | Bar 8 | Bar 9 | Bar 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Maximum stress | 1658.58 | 1659.14 | 1633.58 | 1621.73 | 173668 | 1653.40 | 1669.11 | 1656.48 | 1656.57 | 1711.20 |
| Bar number | Bar 11 | Bar 12 | Bar 13 | Bar 14 | Bar 15 | Bar 16 | Bar 17 | Bar 18 | Bar 19 | Bar 20 |
| Maximum stress | 1653.24 | 1665.31 | 1657.52 | 1645.52 | 1684.86 | 1655.73 | 1664.14 | 1660.26 | 1627.26 | 1662.95 |
| Bar number | Bar 21 | Bar 22 | Bar 23 | Bar 24 | Bar 25 | Bar 26 | Bar 27 | Bar 28 | Bar 29 | Bar 30 |
| Maximum stress | 1646.07 | 1670.28 | 1677.24 | 1660.08 | 1644.60 | 524.75 | 450.94 | 414.77 | 368.59 | 427.37 |
| Bar number | Bar 31 | Bar 32 | Bar 33 | Bar 34 | Bar 35 | Bar 36 | Bar 37 | Bar 38 | Bar 39 | Bar 40 |
| Maximum stress | 709.85 | 580.16 | 451.51 | 320.76 | 242.27 | 714.46 | 576.51 | 454.40 | 338.05 | 233.66 |
| Bar number | Bar 41 | Bar 42 | Bar 43 | Bar 44 | Bar 45 | Bar 46 | Bar 47 | Bar 48 | Bar 49 | Bar 50 |
| Maximum stress | 717.98 | 572.95 | 455.64 | 351.48 | 223.30 | 722.61 | 567.58 | 457.35 | 373.64 | 238.59 |
| Bar number | Bar 51 | Bar 52 | Bar 53 | Bar 54 | Bar 55 | | | | | |
| Maximum stress | 533.56 | 511.80 | 440.31 | 404.61 | 449.23 | | | | | |

## References


1. B. Galván, D. Greiner, J. Périaux, M. Sefrioui, G. Winter, Parallel evolutionary computation for solving complex CFD optimization problems: a review and some nozzle applications, in *Parallel Computational Fluid Dynamics: New Frontiers and Multi-Disciplinary Applications*, ed. by K. Matsuno, A. Ecer, J. Périaux, N. Satofuka, P. Fox (Elsevier, Amsterdam, 2003), pp. 573–602
2. D. Greiner, N. Diaz, J.M. Emperador, B. Galván, G. Winter, A comparative study of the influence of codification on discrete optimum design of frame structures, in *Proceedings of the Third International Conference on Soft Computing Technology in Civil, Structural and Environmental Engineering*, Stirlingshire, 2013. Civil-Comp Press. Paper 6
3. D. Greiner, J.M. Emperador, G. Winter, Multiobjective optimization of bar structures by Pareto-GA, in *Proceedings of the European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS) (Barcelona, 2000)*. CD-ROM, 17 pp
4. D. Greiner, J.M. Emperador, G. Winter, Single and multiobjective frame optimization by evolutionary algorithms and the auto-adaptive rebirth operator. Comput. Methods Appl. Mech. Eng. **193**(33–35), 3711–3743 (2004)
5. D. Greiner, J. Périaux, J.M. Emperador, B. Galván, G. Winter, A hybrid Nash genetic algorithm for reconstruction inverse problems in structural engineering, in *Reports of the Department of Mathematical Information Technology*, Series B, Scientific Computing B 5/2013, University of Jyväskylä, Jyväskylä, 2013
6. D. Greiner, J. Périaux, J.M. Emperador, B. Galván, G. Winter, A study of Nash-evolutionary algorithms for reconstruction inverse problems in structural engineering, in *Advances in Evolutionary and Deterministic Methods for Design, Optimization and Control in Engineering and Sciences*, ed. by D. Greiner, B. Galván, J. Périaux, N. Gauger, K. Giannakoglou, G. Winter, Computational Methods in Applied Sciences, vol. 36 (Springer, Berlin, 2015), pp. 321–333
7. D. Greiner, G. Winter, J.M. Emperador, Optimising frame structures by different strategies of genetic algorithms. Finite Elem. Anal. Des. **37**(5), 381–402 (2001)
8. D. Greiner, G. Winter, J.M. Emperador, B. Galván, Gray coding in evolutionary multicriteria optimization: application in frame structural optimum design, in *Evolutionary Multi-Criterion Optimization (EMO 2005)*, Lecture Notes in Computer Science, vol. 3410 (Springer, Berlin, 2005), pp. 576–591
9. A. Kassimali, *Matrix Analysis of Structures*, 2nd edn. (Cengage Learning, 2011)
10. D.S. Lee, J. Périaux, L.F. Gonzalez, K. Srinivas, E. Onate, Active flow control bump design using hybrid Nash-Game coupled to evolutionary algorithms, in *Proceedings of the Fifth European Conference on Computational Fluid Dynamics ECCOMAS CFD 2010*, ed. by J.C.F. Pereira, A. Sequeira, J.M.C. Pereira (2010) CD-ROM, 14 pp
11. J. Leskinen, J. Périaux, Distributed evolutionary optimization using Nash games and GPUs—applications to CFD design problems. Comput. Fluids **80**, 190–201 (2013)
12. J. Leskinen, H. Wang, J. Périaux, Increasing parallelism of evolutionary algorithms by Nash games in design inverse flow problems. Eng. Comput. **30**(4), 581–600 (2013)
13. J.F. Nash, Equilibrium points in $n$-person games. Proc. Nat. Acad. Sci. USA **36**, 48–49 (1950)
14. J.F. Nash, Non-cooperative games. Ann. Math. **2**(54), 286–295 (1951)
15. J. Périaux, F. Gonzalez, D.S. Lee, D. Greiner, Multi hybridization techniques for advanced parallel evolutionary design in aerospace and structure engineering, in *OPT-i International Conference on Engineering and Applied Sciences Optimization*, 4–6 June 2014, Kos Island, 2014. Plenary lecture
16. J. Périaux, F. González, D.S.C. Lee, Evolutionary optimization and game strategies for advanced multi-disciplinary design: applications to aeronautics, in *Intelligent Systems, Control and Automation: Science and Engineering*, vol. 75 (Springer, Berlin, 2015)

17. M. Sefrioui, J. Périaux, Nash genetic algorithms: examples and applications, in *Proceedings of the 2000 Congress on Evolutionary Computation CEC00* (IEEE, 2000), pp. 509–516
18. T. Varis, T. Tuovinen, Open benchmark database for multidisciplinary optimization problems, in *Proceedings of the International Conference on Modeling and Applied Simulation* (2012), pp. 23–30

# Efficient Variational Design Sensitivity Analysis

**Franz-Joseph Barthold, Nikolai Gerzen, Wojciech Kijanski and Daniel Materna**

**Abstract** The authors' variant of variational design sensitivity analysis in structural optimisation is highlighted in detail. A rigorous separation of physical quantities into geometry and displacement mappings based on an intrinsic presentation of continuum mechanics build up the first step. The variations with respect to design and displacements are easily available in a second step. The subsequent discrete matrix expressions are used to formulate the finite element equations in a third step. The fourth step elaborates the derived Matlab implementation while the fifth step shows the computational behaviour for an academic example. Both, the general case of nonlinear structural behaviour and the linearised approximation are outlined. The advocated scheme is compared with the well-known analytical differentiation approach of the discrete finite element equations.

## 1 Introduction

Several approaches to sensitivity analysis such as an overall finite difference scheme, a semi-analytical approach, a discrete analytical method, an automatic differentiation technique and some different flavours of the variational approach are well-known in structural optimisation. All mentioned techniques finally yield the correct gradient

F.-J. Barthold (✉) · N. Gerzen · W. Kijanski
Numerische Methoden und Informationsverarbeitung, TU Dortmund, August-Schmidt-Str. 8, D-44227 Dortmund, Germany
e-mail: franz-joseph.barthold@tu-dortmund.de

N. Gerzen
e-mail: nikolai.gerzen@tu-dortmund.de

W. Kijanski
e-mail: wojciech.kijanski@tu-dortmund.de

D. Materna
Department of Civil Engineering, Ostwestfalen-Lippe University of Applied Sciences, Emilienstrasse 45, D-32756 Detmold, Germany
e-mail: daniel.materna@tu-dortmund.de; daniel.materna@hs-owl.de



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_14

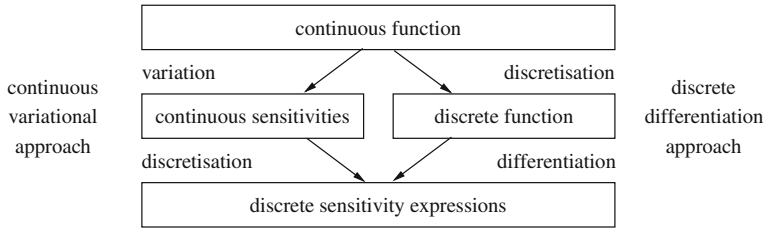



**Fig. 1** Different order of steps within variational and analytical sensitivity analysis

value up to some manageable approximation errors. Nevertheless, the methods differ significantly if not only the correctness of the resulting gradient value is considered.

This paper focuses on the authors' variant of *variational design sensitivity analysis* which is based on an *intrinsic presentation* of continuum mechanics. The advantages of this viewpoint are described in comparison with the *discrete analytical differentiation* of the finite element equations. The later method is well-known and widely used in structural optimisation, see e.g. [11, 21] for details.

The central differences of both methods, i.e. an initial variation followed by the discretisation step in case of the variational approach and the discretisation of the continuous equations accompanied by the analytical differentiation of the resulting discrete equations, respectively, are highlighted in Fig. 1.

The overall scheme for variational design sensitivity analysis is outlined in Sect. 2 and the intrinsic presentation of continuum mechanics is summarised in Sect. 3. In Sect. 4, the discretisation of variational sensitivities is performed. Some remarks on its implementation follow in Sect. 5. The alternative differentiation approach is explained in Sect. 6. In Sect. 7, the computational performance of both approaches is compared. The most significant results are summarised in Sect. 8.

## 2 Outline of Variational Design Sensitivity Analysis

The layout of sensitivity analysis is described and its discrete formulation is derived.

### *2.1 Continuous Formulation of Sensitivity Analysis*

In structural analysis, the displacements $\mathbf{u} \in \mathcal{V}$ are computed for any given design $\mathbf{X} \in \mathcal{S}$ by solving the weak form of equilibrium $R(\mathbf{u}, \mathbf{X}; \mathbf{v}) = 0$ for any test function $\mathbf{v} \in \mathcal{V}$. Here, $\mathcal{V}$ and $\mathcal{S}$ denote the spaces of admissible displacements and designs, which are parametrized by time $t$ and design $s$, respectively, see Sect. 3. Thus, the partial variations of any function $(\cdot)$ are indicated by $\delta_u(\cdot)$ and $\delta_X(\cdot)$ with variations $\delta\mathbf{u} \in \mathcal{V}$ and $\delta\mathbf{X} \in \mathcal{S}$, respectively. But, no variation must violate equilibrium, i.e.

$$\delta R = \delta_X R + \delta_u R = p(\mathbf{u}, \mathbf{X}; \mathbf{v}, \delta\mathbf{X}) + k(\mathbf{u}, \mathbf{X}; \mathbf{v}, \delta\mathbf{u}) = 0. \tag{1}$$

Here, $\delta_u R = k(\mathbf{u}, \mathbf{X}; \mathbf{v}, \delta\mathbf{u})$ is the *tangent stiffness operator* and $\delta_X R = p(\mathbf{u}, \mathbf{X}; \mathbf{v}, \delta\mathbf{X})$ denotes the *tangent pseudo load operator*. We can solve the sensitivity equation for a given design variation $\delta\mathbf{X}$ to obtain the displacement variation $\delta\mathbf{u}$.

$$k(\mathbf{u}, \mathbf{X}; \mathbf{v}, \delta\mathbf{u}) = -Q(\mathbf{u}, \mathbf{X}; \mathbf{v}) \qquad \forall\, \mathbf{v} \in \mathcal{V} \tag{2}$$

Herein, $Q(\mathbf{u}, \mathbf{X}; \mathbf{v}) := p(\mathbf{u}, \mathbf{X}; \mathbf{v}, \delta\mathbf{X})$ is the *pseudo load* which maps the (material) design variation $\delta\mathbf{X}$ to a (physical) pseudo load $Q$. This observation motivates the term *pseudo load operator* for the operator $p$. Furthermore, a sensitivity equation for the *material* or *inverse motion problem* as well as for the *dual* or *adjoint problem* can be derived in the same manner, see [24] for details.

As a consequence, the existence of a (linear) *sensitivity operator* $S$, evaluated at the current equilibrium point $(\mathbf{u}, \mathbf{X})$, can be deduced from Eq. (1) with

$$\delta\mathbf{u} = S(\mathbf{u}, \mathbf{X}; \delta\mathbf{X}). \tag{3}$$

The optimisation model consists of objective and constraint functions $f(\mathbf{u}, \mathbf{X})$. Therefore, $\beta(\mathbf{u}, \mathbf{X}; \delta\mathbf{X})$ and $\gamma(\mathbf{u}, \mathbf{X}; \delta\mathbf{u})$ denote linear forms obtained by varying $f$ with respect to design or displacements. Thus, the total variation of the function $f$ with respect to any design variation $\delta\mathbf{X}$ yields the linear operator $\alpha = \beta + \gamma \circ S$

$$\delta f = \delta_X f + \delta_u f = \beta(\delta\mathbf{X}) + \gamma(\delta\mathbf{u}) = (\beta + \gamma \circ S)\,(\delta\mathbf{X}) = \alpha(\delta\mathbf{X}). \tag{4}$$

For a specific function $f_i$ and any design variation $\delta\mathbf{X}_j$, Eq. (4) transforms to

$$\delta f_{ij} = \alpha_i(\mathbf{u}, \mathbf{X}; \delta\mathbf{X}_j). \tag{5}$$

*Remark 1 (Linear operators in sensitivity analysis)* All variations of residuals, objectives and constraints can be obtained by a straightforward analysis on the continuum mechanical level. At any equilibrium point $(\mathbf{u}, \mathbf{X})$, the resulting linear operator $S : \mathcal{S} \to \mathcal{V}$ and the linear form $\alpha : \mathcal{S} \to \mathbb{R}$ describe the reaction of the mechanical systems in case of a design perturbation. The central sensitivity operator $S$ is only implicitly known and further progress can only be achieved on the discrete level.

Special emphasis is given to a combined presentation of the tangent stiffness operator and the tangent pseudo load operator permitting a minimal overlay for performing the variational design sensitivity analysis, see Sect. 3.

## 2.2 Fundamentals of Discrete Approximations

The general idea of discretisation is outlined without referring to a specific method.

We introduce the discrete approximations for the state $\mathbf{u}_h$ and the design $\mathbf{X}_h$ to obtain a matrix description of the derived residual and tangent forms. These

approximations depend on the displacement parameters $\hat{\boldsymbol{U}} \in \mathbb{R}^{nu}$ and the design parameters $\hat{\boldsymbol{X}} \in \mathbb{R}^{nx}$. Here, *nu* and *nx* are the dimensions of the introduced approximation spaces, i.e. *nu* denotes the number of the discrete state variables in $\mathcal{V}_h \subset \mathcal{V}$ and *nx* the number of the discrete design variables in $\mathcal{S}_h \subset \mathcal{S}$. We introduce in the same manner the discrete approximations for the corresponding variations, i.e. $\delta\hat{\boldsymbol{U}} \in \mathbb{R}^{nu}$ and $\delta\hat{\boldsymbol{X}} \in \mathbb{R}^{nx}$. Furthermore, the test function $\mathbf{v}_h$ is given by $\hat{\boldsymbol{V}} \in \mathbb{R}^{nu}$.

The continuous forms can be evaluated for any discrete design $\mathbf{X}_h$ and the associated discrete displacements $\mathbf{u}_h$, i.e. we obtain the functionals and bilinear forms

$$R(\mathbf{u}_h, \mathbf{X}_h; \mathbf{v}_h) = \hat{\boldsymbol{V}}^{\mathrm{T}} \hat{\boldsymbol{R}} \quad \text{with residual vector } \hat{\boldsymbol{R}} \in \mathbb{R}^{nu}, \tag{6}$$

$$k(\mathbf{u}_h, \mathbf{X}_h; \mathbf{v}_h, \delta\mathbf{u}_h) = \hat{\boldsymbol{V}}^{\mathrm{T}} \hat{\boldsymbol{K}}\, \delta\hat{\boldsymbol{U}} \quad \text{with stiffness matrix } \hat{\boldsymbol{K}} \in \mathbb{R}^{nu \times nu}, \tag{7}$$

$$p(\mathbf{u}_h, \mathbf{X}_h; \mathbf{v}_h, \delta\mathbf{X}_h) = \hat{\boldsymbol{V}}^{\mathrm{T}} \hat{\boldsymbol{P}}\, \delta\hat{\boldsymbol{X}} \quad \text{with pseudo load matrix } \hat{\boldsymbol{P}} \in \mathbb{R}^{nu \times nx}. \tag{8}$$

The details in case of the finite element method are explained in Sect. 4.3.

## 2.3 Properties of Discrete Sensitivity Relations

The discrete version of Eq. (1) evaluated at $(\mathbf{u}_h, \mathbf{X}_h)$ reads

$$\delta R = \hat{\boldsymbol{V}}^{\mathrm{T}} \delta\hat{\boldsymbol{R}} = \hat{\boldsymbol{V}}^{\mathrm{T}} \left[ \hat{\boldsymbol{K}}\, \delta\hat{\boldsymbol{U}} + \hat{\boldsymbol{P}}\, \delta\hat{\boldsymbol{X}} \right] = 0 \tag{9}$$

and we obtain the well-known discrete condition

$$\delta\hat{\boldsymbol{R}} = \hat{\boldsymbol{K}}\, \delta\hat{\boldsymbol{U}} + \hat{\boldsymbol{P}}\, \delta\hat{\boldsymbol{X}} = \mathbf{0}. \tag{10}$$

Additionally, the discrete version of Eq. (2) with $\hat{\boldsymbol{Q}} \in \mathbb{R}^{nu}$ being the pseudo load vector of the physical residual problem associated to the functional $Q(\mathbf{u}_h, \mathbf{X}_h; \cdot)$ is

$$\hat{\boldsymbol{K}}\, \delta\hat{\boldsymbol{U}} = -\hat{\boldsymbol{Q}} \qquad \text{with} \qquad \hat{\boldsymbol{Q}} := \hat{\boldsymbol{P}}\, \delta\hat{\boldsymbol{X}}. \tag{11}$$

Furthermore, the discrete version of Eq. (3) takes the form

$$\delta\hat{\boldsymbol{U}} = \hat{\boldsymbol{S}}\, \delta\hat{\boldsymbol{X}} \qquad \text{with} \qquad \hat{\boldsymbol{S}} := -\hat{\boldsymbol{K}}^{-1} \hat{\boldsymbol{P}}, \tag{12}$$

where $\hat{\boldsymbol{S}} \in \mathbb{R}^{nu \times nx}$ denotes the *sensitivity operator matrix*, i.e. we can evaluate the sensitivity equation for arbitrary admissible variations $\delta\hat{\boldsymbol{X}}$ in the material space.

Finally, the sensitivity of the function $f$ with respect to design variations can be deduced from Eq. (4) to yield

$$\delta f = \hat{\boldsymbol{b}}^{\mathrm{T}} \delta\hat{\boldsymbol{X}} + \hat{\boldsymbol{c}}^{\mathrm{T}} \delta\hat{\boldsymbol{U}}, \tag{13}$$

where $\delta\hat{\boldsymbol{U}}$ and $\delta\hat{\boldsymbol{X}}$ denote the variation of the discrete displacement and nodal coordinate vectors, respectively. The discretised variation of the displacements given in Eq. (12) can be inserted. Considering several objectives $f_i$ with $i = 1, 2, \ldots, nf$ and several variations of design $\delta\hat{\boldsymbol{X}}_j$ with $j = 1, 2, \ldots, ndv$ yields

$$\delta f_{ij} = \left[ \hat{\boldsymbol{b}}_i^{\mathrm{T}} - \hat{\boldsymbol{c}}_i^{\mathrm{T}} \, \hat{\boldsymbol{K}}^{-1} \, \hat{\boldsymbol{P}} \right] \delta\hat{\boldsymbol{X}}_j = \left[ \hat{\boldsymbol{b}}_i^{\mathrm{T}} + \hat{\boldsymbol{c}}_i^{\mathrm{T}} \, \hat{\boldsymbol{S}} \right] \delta\hat{\boldsymbol{X}}_j = \hat{\boldsymbol{a}}_i^{\mathrm{T}} \, \delta\hat{\boldsymbol{X}}_j \, . \tag{14}$$

*Remark 2 (Discrete form of sensitivity analysis).* The continuous linear mapping $\delta\mathbf{u} = S(\delta\mathbf{X})$ transfers to the computable discrete equation $\delta\hat{\boldsymbol{U}} = \hat{\boldsymbol{S}} \, \delta\hat{\boldsymbol{X}}$ and the linear form $\delta f = \alpha(\delta\mathbf{X})$ transfers to the computable discrete equation $\delta f = \hat{\boldsymbol{a}}^{\mathrm{T}} \, \delta\hat{\boldsymbol{X}}$.

## 3 Basics on the Intrinsic Viewpoint of Continuum Mechanics

The authors' viewpoint on variational design sensitivity analysis within a general continuum mechanical framework relies on a rigorous separation of physical quantities into geometry and displacement mappings, see [4–6] for further details. This viewpoint is based on *intrinsic coordinates* which have been introduced to mechanics in [30] advancing the traditional presentation [34]. The intrinsic coordinates are also named *convected*, *curvilinear*, *local* or *natural* coordinates depending on which feature should be highlighted in the specific setting. Unfortunately, the mathematical background, i.e. the body can be described as differentiable manifold, is not always present. However, all available knowledge should be explored to obtain general theoretical results and efficient numerical formulations, see Sect. 3.5.

The advocated integrated approach for deriving the variations within direct analysis and variational design sensitivity analysis based on intrinsic coordinates differs conceptually from the *material derivative approach*, see e.g. [2], and the *domain parametrization approach*, see e.g. [33], respectively. Furthermore, the relationship to *configurational mechanics* has been outlined in [24, 27]. These different approaches are briefly sketched in Sect. 3.6.

This section summarises the theoretical framework used for variational design sensitivity analysis. The results can be used to apply a subsequent discretisation step.

### *3.1 Separation of Point and Tangent Mappings*

The separation approach yields a decomposition of the design ($s$) and time ($t$) dependent deformation mapping $\mathbf{x} = \boldsymbol{\varphi}_{\mathrm{X}}(\mathbf{X}(s), t)$ into two independent mappings, i.e. a design dependent geometry mapping $\mathbf{X} = \boldsymbol{\kappa}_{\Theta}(\boldsymbol{\Theta}, s)$ and a time dependent motion mapping $\mathbf{x} = \boldsymbol{\mu}_{\Theta}(\boldsymbol{\Theta}, t)$. Furthermore, an intrinsic displacement mapping $\mathbf{u} = \boldsymbol{v}_{\Theta}(\boldsymbol{\Theta}, s, t) = \boldsymbol{\mu}_{\Theta}(\boldsymbol{\Theta}, t) - \boldsymbol{\kappa}_{\Theta}(\boldsymbol{\Theta}, s)$ can be introduced. All intrinsic quantities are indicated using a subscript $\Theta$.

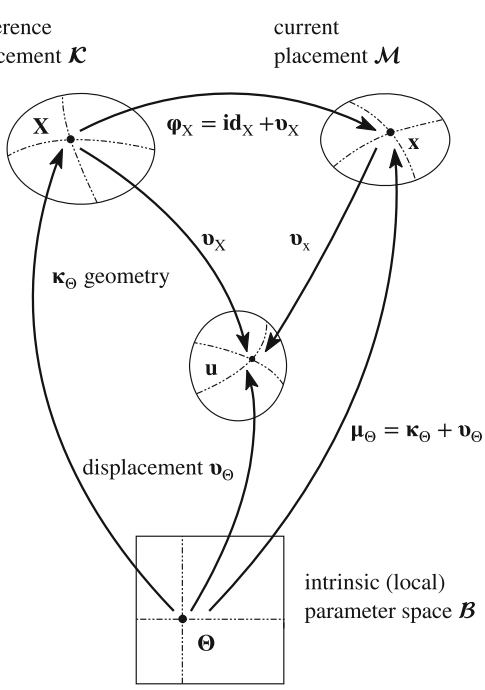



**Fig. 2** Configurations and mappings

This viewpoint using intrinsic coordinates $\boldsymbol{\Theta} \in \mathcal{B}$ can be enhanced by the common Lagrangian viewpoint based on the reference coordinates $\mathbf{X} \in \mathcal{K}$ and indicated by a subscript X. Furthermore, the Eulerian viewpoint is based on the current coordinates $\mathbf{x} \in \mathcal{M}$ and is indicated by a subscript x. Consequently, the referential deformation mapping is given by the composition $\boldsymbol{\varphi}_X = \boldsymbol{\mu}_\Theta \circ \boldsymbol{\kappa}_\Theta^{-1}$, see Fig. 2.

*Remark 3 (Design as part of mechanics)* The most significant enhancement of the advocated approach over the traditional presentation of continuum mechanics is the introduction of the geometry mapping $\boldsymbol{\kappa}_\Theta$. Thus, continuum mechanics is understood as a theory of two fundamental mappings, i.e. geometry $\boldsymbol{\kappa}_\Theta : \mathcal{B} \to \mathcal{K}$ and motion $\boldsymbol{\mu}_\Theta : \mathcal{B} \to \mathcal{M}$, which decompose the deformation mapping $\boldsymbol{\varphi}_X = \boldsymbol{\mu}_\Theta \circ \boldsymbol{\kappa}_\Theta^{-1}$. Therefore, the design of structures can be rigorously founded, see [4–6].

## *3.2 Gradients, Strains and Stresses*

Different gradient operators grad, Grad and GRAD corresponding to the variables $\mathbf{x}$, $\mathbf{X}$ and $\boldsymbol{\Theta}$ of the considered domains $\mathcal{M}$, $\mathcal{K}$ and $\mathcal{B}$, respectively, can be defined. The intrinsic *motion gradient* $\mathbf{M}_\Theta = \mathrm{GRAD}\, \boldsymbol{\mu}_\Theta$ and the intrinsic *geometry gradient* $\mathbf{K}_\Theta = \mathrm{GRAD}\, \boldsymbol{\kappa}_\Theta$ are used to decompose the referential *deformation gradient*

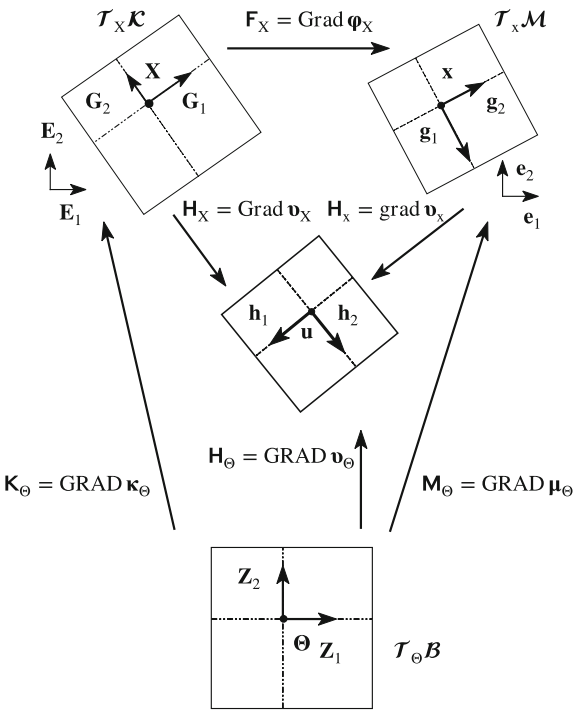



**Fig. 3** Tangent mappings between tangent spaces

$$\mathbf{F}_X = \mathrm{Grad}\,\boldsymbol{\varphi}_X = \mathrm{GRAD}\,\boldsymbol{\mu}_\Theta\ [\mathrm{GRAD}\,\boldsymbol{\kappa}_\Theta]^{-1} = \mathbf{M}_\Theta\,\mathbf{K}_\Theta^{-1}. \tag{15}$$

The pull-back and push-forward transformation between the tangent spaces at the current placement $\mathcal{T}_x\mathcal{M}$, at the reference placement $\mathcal{T}_X\mathcal{K}$ and at the intrinsic parameter space $\mathcal{T}_\Theta\mathcal{B}$, see Fig. 3, are based on $\mathbf{K}_\Theta$, $\mathbf{M}_\Theta$, $\mathbf{F}_X$ and on their determinants $J_{K_\Theta} = \det\mathbf{K}_\Theta$, $J_{M_\Theta} = \det\mathbf{M}_\Theta$ and $J_{F_X} = \det\mathbf{F}_X$.

Similarly, the referential *displacement gradient* $\mathbf{H}_X = \mathrm{Grad}\,\boldsymbol{v}_X$ can be split into the intrinsic displacement gradient and the inverse intrinsic geometry gradient

$$\mathbf{H}_X = \mathrm{Grad}\,\boldsymbol{v}_X = \mathrm{GRAD}\,\boldsymbol{v}_\Theta\ [\mathrm{GRAD}\,\boldsymbol{\kappa}_\Theta]^{-1} = \mathbf{H}_\Theta\,\mathbf{K}_\Theta^{-1}. \tag{16}$$

*Remark 4 (Decomposition of referential deformation and displacement gradients)* The decomposition of referential gradients into two independent intrinsic gradients, see e.g. Eqs. (15) and (16), is a central prerequisite for efficiently deriving variations, see Sect. 3.3. Furthermore, it serves as a master copy for discrete computational methods, e.g. the technique to compute Cartesian derivatives in the finite element method, see Sect. 4.2.

The (referential) Green-Lagrange strain tensor $\mathbf{E}$ and the linear strain tensor $\boldsymbol{\varepsilon}$ can be considered as functions of the (referential) displacement gradient

$$\mathbf{E} = \frac{1}{2}\left(\mathbf{H}_{\mathrm{X}} + \mathbf{H}_{\mathrm{X}}^{\mathsf{T}} + \mathbf{H}_{\mathrm{X}}^{\mathsf{T}}\,\mathbf{H}_{\mathrm{X}}\right) \quad \text{and} \quad \boldsymbol{\varepsilon} = \frac{1}{2}\left(\mathbf{H}_{\mathrm{X}} + \mathbf{H}_{\mathrm{X}}^{\mathsf{T}}\right). \tag{17}$$

The Cauchy stress tensor $\mathbf{T}$, the Kirchhoff stress tensor $\boldsymbol{\tau} = \boldsymbol{J}_{\mathbf{F}_{\mathrm{X}}}\,\mathbf{T}$, the (1) Piola-Kirchhoff stress tensor $\mathbf{P} = \boldsymbol{\tau}\,\mathbf{F}_{\mathrm{X}}^{-\mathsf{T}}$ and the (2) Piola-Kirchhoff stress tensor $\mathbf{S} = \mathbf{F}_{\mathrm{X}}^{-1}\,\mathbf{P}$ collapse to the linear stress tensor $\boldsymbol{\sigma}$ in case of linearised elasticity.

## *3.3 Variations of Gradients, Strains and Stresses*

The total variation of the deformation gradient $\mathbf{F}_{\mathrm{X}}$ can be derived using the multiplicative decomposition $\mathbf{F}_{\mathrm{X}} = \mathbf{M}_\Theta\,\mathbf{K}_\Theta^{-1}$ and the variations of the tangent mappings $\delta\mathbf{K}_\Theta = \mathrm{GRAD}\,\delta\mathbf{X}$ and $\delta\mathbf{M}_\Theta = \mathrm{GRAD}\,\delta\mathbf{x}$, i.e.

$$\delta\mathbf{F}_{\mathrm{X}} = \delta\mathbf{M}_\Theta\,\mathbf{K}_\Theta^{-1} + \mathbf{M}_\Theta\,\delta\mathbf{K}_\Theta^{-1} = \delta\mathbf{M}_\Theta\,\mathbf{K}_\Theta^{-1} - \mathbf{M}_\Theta\,\mathbf{K}_\Theta^{-1}\,\delta\mathbf{K}_\Theta\,\mathbf{K}_\Theta^{-1}, \tag{18}$$

where $\delta[\mathbf{K}_\Theta^{-1}] = -\mathbf{K}_\Theta^{-1}\,\delta\mathbf{K}_\Theta\,\mathbf{K}_\Theta^{-1}$ has been used. The total variation $\delta\mathbf{F}_{\mathrm{X}}$ can be split into partial variations $\delta_u\mathbf{F}_{\mathrm{X}}$ and $\delta_X\mathbf{F}_{\mathrm{X}}$, i.e. w.r.t. displacements $\delta\mathbf{u}$ or geometry $\delta\mathbf{X}$

$$\delta\mathbf{F}_{\mathrm{X}} = \delta_u\mathbf{F}_{\mathrm{X}} + \delta_X\mathbf{F}_{\mathrm{X}} = \left[\mathbf{F}_{\mathrm{X}}\right]'_u + \left[\mathbf{F}_{\mathrm{X}}\right]'_X = \mathrm{Grad}\,\delta\mathbf{u} - \mathrm{Grad}\,\mathbf{u}\;\mathrm{Grad}\,\delta\mathbf{X} \tag{19}$$

using the notation $\left[\mathbf{F}_{\mathrm{X}}\right]'_u(\mathbf{u}, \delta\mathbf{u}) = \mathrm{Grad}\,\delta\mathbf{u}$ and $\left[\mathbf{F}_{\mathrm{X}}\right]'_X(\mathbf{u}, \delta\mathbf{X}) = -\mathrm{Grad}\,\mathbf{u}\,\mathrm{Grad}\,\delta\mathbf{X}$. Furthermore, the variations of the Jacobians $\boldsymbol{J}_{\mathbf{K}_\Theta}$, $\boldsymbol{J}_{\mathbf{M}_\Theta}$ and $\boldsymbol{J}_{\mathbf{F}_{\mathrm{X}}}$ can be performed based on $\delta(\det\mathbf{B}) = \det\mathbf{B}\;\mathbf{B}^{-\mathsf{T}} : \delta\mathbf{B}$, where $\mathbf{B}$ is either $\mathbf{K}_\Theta$, $\mathbf{M}_\Theta$ or $\mathbf{F}_{\mathrm{X}}$.

The first partial variations of the Green-Lagrange strain tensor $\mathbf{E}(\mathbf{u})$ yield

$$\begin{aligned}
\mathbf{E}'_u(\mathbf{u}, \delta\mathbf{u}) &= \mathrm{sym}\left(\mathbf{A}_u^{\mathsf{T}}\;\mathrm{Grad}\,\delta\mathbf{u}\right) = \mathrm{sym}\left(\mathbf{F}_{\mathrm{X}}^{\mathsf{T}}\;\mathrm{Grad}\,\delta\mathbf{u}\right)\\
\mathbf{E}'_X(\mathbf{u}, \delta\mathbf{X}) &= \mathrm{sym}\left(\mathbf{A}_X^{\mathsf{T}}\;\mathrm{Grad}\,\delta\mathbf{X}\right) = -\mathrm{sym}\left(\mathbf{F}_{\mathrm{X}}^{\mathsf{T}}\;\mathrm{Grad}\,\mathbf{u}\;\mathrm{Grad}\,\delta\mathbf{X}\right),
\end{aligned} \tag{20}$$

where the abbreviations $\mathbf{A}_u = \mathbf{F}_{\mathrm{X}}$ and $\mathbf{A}_X = -\mathrm{Grad}^{\mathsf{T}}\mathbf{u}\,\mathbf{F}_{\mathrm{X}}$ have been used. The second partial variations with additional variation of the displacements $\Delta\mathbf{u}$ read

$$\begin{aligned}
\mathbf{E}''_{uX}(\mathbf{u}, \delta\mathbf{u}, \delta\mathbf{X}) &= -\mathrm{sym}(\mathrm{Grad}^{\mathsf{T}}\delta\mathbf{X}\,\mathbf{H}_{\mathrm{X}}^{\mathsf{T}}\;\mathrm{Grad}\,\delta\mathbf{u} + \mathbf{F}_{\mathrm{X}}^{\mathsf{T}}\;\mathrm{Grad}\,\delta\mathbf{u}\;\mathrm{Grad}\,\delta\mathbf{X})\\
\mathbf{E}''_{uu}(\mathbf{u}, \delta\mathbf{u}, \Delta\mathbf{u}) &= \mathrm{sym}(\mathrm{Grad}^{\mathsf{T}}\delta\mathbf{u}\;\mathrm{Grad}\,\Delta\mathbf{u}).
\end{aligned} \tag{21}$$

The variations of the linearised strain tensor $\boldsymbol{\varepsilon}$ are less complex, i.e.

$$\boldsymbol{\varepsilon}'_u(\delta\mathbf{u}) = \mathrm{sym}(\mathrm{Grad}\,\delta\mathbf{u}) \quad \text{and} \quad \boldsymbol{\varepsilon}'_X(\mathbf{u}, \delta\mathbf{X}) = -\mathrm{sym}(\mathbf{H}_{\mathrm{X}}\;\mathrm{Grad}\,\delta\mathbf{X}), \tag{22}$$

and the second partial variations are

$$\boldsymbol{\varepsilon}''_{uu} = \mathbf{0} \quad \text{and} \quad \boldsymbol{\varepsilon}''_{uX}(\delta\mathbf{u}, \delta\mathbf{X}) = -\mathrm{sym}(\mathrm{Grad}\,\delta\mathbf{u}\ \mathrm{Grad}\,\delta\mathbf{X}). \tag{23}$$

The variation of stresses in case of hyperelastic material behaviour deliver the referential and linear elasticity tensors, i.e. $\mathbb{C}$ and $\mathbb{E}$, respectively, and the expressions

$$\delta\mathbf{S} = \frac{\partial \mathbf{S}}{\partial \mathbf{E}} : \delta\mathbf{E} = \mathbb{C} : \delta\mathbf{E} \quad \text{and} \quad \delta\boldsymbol{\sigma} = \frac{\partial \boldsymbol{\sigma}}{\partial \boldsymbol{\varepsilon}} : \delta\boldsymbol{\varepsilon} = \mathbb{E} : \delta\boldsymbol{\varepsilon}. \tag{24}$$

### *3.4 Weak Form of Equilibrium and Its Variations*

The weak form of equilibrium $R = R_{\mathrm{int}} - R_{\mathrm{ext}} = 0$ is a linear form of the test function $\mathbf{v}$ evaluated for current geometry and displacement mappings, i.e.

$$R(\mathbf{u}, \mathbf{X}; \mathbf{v}) = \int_{\mathcal{K}} \mathbf{S} : \mathbf{E}'_u(\mathbf{u}, \mathbf{v})\, \mathrm{d}V_{\mathrm{X}} - F(\mathbf{X}; \mathbf{v}) \tag{25}$$

in case of a general nonlinear theory. We consider for physical body forces $\mathbf{b}_{\mathrm{X}}$

$$R_{\mathrm{ext}} = F(\mathbf{X}; \mathbf{v}) = \int_{\mathcal{K}} \mathbf{b}_{\mathrm{X}} \cdot \mathbf{v}\, \mathrm{d}V_{\mathrm{X}} = \int_{\mathcal{B}} \mathbf{b}_{\mathrm{X}} \cdot \mathbf{v} \boldsymbol{J}_{\mathrm{K}_\Theta}\, \mathrm{d}V_\Theta, \tag{26}$$

i.e. the external virtual work is deformation independent for notational simplicity.

An outline of the variational design sensitivity analysis is given in Sect. 2, and Eq. (1) needs two partial variations of the weak form. In the general nonlinear case, the *tangent stiffness operator* $k$ is defined by $k = \delta_u R$ with

$$k(\mathbf{u}, \mathbf{X}; \mathbf{v}, \delta\mathbf{u}) = \int_{\mathcal{K}} \left\{ \mathbf{E}'_u(\mathbf{u}, \mathbf{v}) : \mathbb{C} : \mathbf{E}'_u(\mathbf{u}, \delta\mathbf{u}) + \mathbf{S} : \mathbf{E}''_{uu}(\mathbf{v}, \delta\mathbf{u}) \right\} \mathrm{d}V_{\mathrm{X}} \tag{27}$$

with variations of the Green-Lagrange strain tensor $\mathbf{E}$ given in Eqs. (20) and (21).

The corresponding variation $\delta_X R$ leads to the *pseudo load operator* $p$,

$$\begin{aligned} p(\mathbf{u}, \mathbf{X}; \mathbf{v}, \delta\mathbf{X}) = \int_{\mathcal{K}} \big\{ & \mathbf{S} : \mathbf{E}''_{uX}(\mathbf{u}, \mathbf{v}, \delta\mathbf{X}) + \mathbf{E}'_u(\mathbf{u}, \mathbf{v}) : \mathbb{C} : \mathbf{E}'_X(\mathbf{u}, \delta\mathbf{X}) \\ & + \mathbf{S} : \mathbf{E}'_u(\mathbf{u}, \mathbf{v})\, \mathrm{Div}\,\delta\mathbf{X} \big\}\, \mathrm{d}V_{\mathrm{X}} - F'_X(\mathbf{X}; \mathbf{v}, \delta\mathbf{X}). \end{aligned} \tag{28}$$

The design variation of Eq. (26), to be used in Eqs. (28) and (32), leads to

$$F'_X(\mathbf{X}; \mathbf{v}, \delta\mathbf{X}) = \int_{\mathcal{K}} \mathbf{b}_{\mathrm{X}} \cdot \mathbf{v}\ \mathrm{Div}\,\delta\mathbf{X}\, \mathrm{d}V_{\mathrm{X}}. \tag{29}$$

The linearised version on the linear domain $\Omega$ simplifies to the expression

$$R_{\text{lin}}(\mathbf{u}, \mathbf{X}; \mathbf{v}) = \int_{\Omega} \text{Grad}\,\mathbf{v} : \boldsymbol{\sigma}\, \text{d}\Omega - F(\mathbf{X}; \mathbf{v}). \tag{30}$$

Consequently, the *linear stiffness operator* of the linear theory is constant

$$k_{\text{lin}}(\mathbf{X}; \mathbf{v}, \delta\mathbf{u}) = \int_{\Omega} \text{Grad}\,\mathbf{v} : \mathbb{E} : \text{sym}\{\text{Grad}\,\delta\mathbf{u}\}\, \text{d}\Omega, \tag{31}$$

and the *linear pseudo load operator* depends linearly on the displacements

$$\begin{aligned} p_{\text{lin}}(\mathbf{u}, \mathbf{X}; \mathbf{v}, \delta\mathbf{X}) = &\int_{\Omega} \text{Grad}\,\mathbf{v} : \mathbb{E} : \text{sym}\{-\text{Grad}\,\mathbf{u}\ \text{Grad}\,\delta\mathbf{X}\}\, \text{d}\Omega \\ &- \int_{\Omega} \{\boldsymbol{\sigma} :\ \text{Grad}\,\mathbf{v}\ \text{Grad}\,\delta\mathbf{X} - \boldsymbol{\sigma} :\ \text{Grad}\,\mathbf{u}\ \text{Div}\,\delta\mathbf{X}\}\, \text{d}\Omega \\ &- F'_X(\mathbf{X}; \mathbf{v}, \delta\mathbf{X}). \end{aligned} \tag{32}$$

Further equivalent formulations can be derived and the overall expression can be transformed onto the intrinsic parameter space $\mathcal{B}$ using $\text{d}V_{\text{X}} = \boldsymbol{J}_{\text{K}_\Theta}\, \text{d}V_\Theta$.

To summarise, all continuous expressions needed for structural and sensitivity analysis have been derived on the continuous level using the intrinsic viewpoint.

*Remark 5 (Efficiency of theoretical development)* A significant advantage of the variational approach is the similarity of all variations in the *physical space*, i.e. with respect to displacements, with those in the *material space* or *design space*, i.e. with respect to geometry. The advocated intrinsic viewpoint offers the possibility to derive both variations in parallel (at the same time and done by the same researcher) minimising the overall effort.

## 3.5 Impact of Differential Geometry on Computations

Continuum mechanics is based on differential geometry and the obtainable results can be presented with different mathematical rigour due to the readers background. We report the most important features and refer to literature for further reading.

*Manifolds* can be characterised as sets which can be *covered* by a finite number of *charts* consisting of a *subset* and *coordinate mapping*. Thus, for any element of the set, there is at least one chart which describes the body in an open environment of the chosen element using a coordinate system. All those charts together build up an *atlas*. Importantly, the special choice of the atlas does not effect the properties of the set. The manifold is termed *differentiable* if the transformation between different coordinate descriptions is sufficiently smooth.

The manifold properties can be summarised as follows: *The material bodies under consideration are described by numerous local coordinate systems.* This property is valid on all levels, i.e. in theory as well as in computations. The consequences for the finite element method are outlined in Sect. 4.1.

*Remark 6 (Efficiency of interaction between continuous and discrete formulations)* The interaction between theoretical investigations on the continuous level and the computational strategies on the discrete level are organised consistently using the intrinsic viewpoint. This minimises the number of concepts used in theory and computations and increases the efficiency of the development. Thus, all theoretical results have a discrete representation. And conversely, any discrete technique has a continuous origin.

## 3.6 Comparison with Other Variational Approaches

Several continuous formulations of sensitivity analysis have been available since decades, see e.g. [3, 12, 18, 19, 31] among others for further details and references. Two concepts, namely the *material derivative approach* (MDA), see e.g. [2, 13, 29], and the *domain parametrization approach* (DPA), see e.g. [17, 33], respectively, which are based on the traditional viewpoint published in [34] attracted wide attention in literature. In MDA, the effect of a design variation is understood by using the analogy to a *physical time derivative* consisting of a local and a convective part. In DPA, an additional *master reference placement* is introduced to which the different designs are related. Both approaches are valuable and important results have been derived based on them, see e.g. [28].

Another concept of mechanics to which the presented approach can be compared with is *Eshelbian mechanics* or *configurational mechanics*, see [14, 16, 20]. The differences occur in the description of modifications in the material space, i.e. an inverse motion is considered from an artificially fixed current placement. Again, the traditional concept from [34] is used to model phenomena which are actual beyond the theoretical framework, see [24, 27] for further hints.

To summarise, none of the mentioned concepts use the manifold structure of continuum mechanics outlined in [30] in order to benefit from the decompositions outlined in Figs. 2 and 3. These elements are an significant improvement which eases the modelling of inverse geometry problems.

Besides these differences, all variational variants have the same strategy for sensitivity in common, see Fig. 1. Thus, most remarks highlighting the advantages of the variational approach over the discrete approach are valid for all variants. The computational methods are generated from a continuous theory, which requires an infinite dimensional space of admissible geometry mappings $\mathcal{S}$ and an infinite dimensional space of admissible displacement mappings $\mathcal{V}$, by a discretisation step. Hereby, the infinite dimensional function spaces $\mathcal{S}$ and $\mathcal{V}$ are substituted by finite dimensional subspaces, say $\mathcal{S}_h \subset \mathcal{S}$ and $\mathcal{V}_h \subset \mathcal{V}$.

*Remark 7* Most properties of the discrete solution are already available and known for the original continuous problem. Thus, a thorough theoretical knowledge eases the interpretation of computational techniques as well as computational results.

## 4 FEM Discretisation of Continuous Sensitivity Expressions

In computations, the geometry $\boldsymbol{\kappa}_\Theta$ is realised by *computer aided geometric design* (CAGD) using *shape functions* such as non-uniform rational B-splines (NURBS). The displacement $\boldsymbol{v}_\Theta$ is approximated by the *finite element method* (FEM) using *Ansatz functions*. In case of an *isoparametric finite element method*, both shape and Ansatz functions are low order polynomials, say (bi- or tri-) linear functions. The details for sensitivity expressions are outlined in this section.

Three steps must be performed to derive the finite element equations from the continuous setting. Firstly, the considered domain must be subdivided into elements and an efficient computational scheme must be set up, see Sect. 4.1. Secondly, the tensorial notation should be transformed into a matrix notation, see Sect. 4.2. Thirdly, the chosen approximations for geometry and displacements must be inserted into the matrix formulation, see Sect. 4.3.

### *4.1 Adaptation of the Manifold Properties to the FE Method*

The material bodies are differentiable manifolds, see Sect. 3.5, i.e. any discretisation should not destroy this property. Therefore, all results quoted in Sect. 3 are independent from the choice of the intrinsic coordinates $\{\Theta^1, \Theta^2, \Theta^3\}$ and the intrinsic parameter domain $\mathcal{B}$, i.e. both can be adopted to the special needs.

Within the finite element framework and without loss of generality, a suitable atlas is introduced by partitioning the placements $\mathcal{K}$ and $\mathcal{M}$ as well as the intrinsic parameter domain $\mathcal{B}$ into a finite number of sub-domains $\mathcal{K}_e$, $\mathcal{M}_e$ and $\mathcal{B}_e$. These sub-domains are linked to the *local parameter domain* $\mathcal{R}_e \equiv \mathcal{R} = [-1, 1]^m$, where $m$ refers to the dimension of the problem. The local parameter domain $\mathcal{R}$ is unique and constant for all sub-domains and its *local coordinates* are labelled $\boldsymbol{\zeta}$. Furthermore, the point mappings $\boldsymbol{\kappa}_\zeta : \mathcal{R} \to \mathcal{K}$ and $\boldsymbol{\mu}_\zeta : \mathcal{R} \to \mathcal{M}$ as well as the tangent mappings $\mathbf{K}_\zeta : \mathcal{T}_\zeta \mathcal{R} \to \mathcal{T}_X \mathcal{K}$ and $\mathbf{M}_\zeta : \mathcal{T}_\zeta \mathcal{R} \to \mathcal{T}_x \mathcal{M}$, respectively, replace the mappings introduced in Sect. 3. We omit the index $e$ in most cases to shorten the notation.

*Remark 8 (Finite element mesh is a special atlas of the body)* The above description characterises the finite element mesh and the mappings from the local parameter space onto each element. Thus, every modification so far is a direct consequence of the intrinsic presentation of continuum mechanics. The finite element mesh, i.e. the subdivision of the body and its placements in a finite number of subdomains, is a special atlas suitable for efficient computations. Thus, this adaptation to the finite element method is exact because no approximation has been introduced so far.

## 4.2 Matrix Representation of Gradients, Strains and Stresses

Using the introduced Cartesian basis systems $\{\mathbf{Z}_i\}$, $\{\mathbf{E}_i\}$, $\{\mathbf{e}_i\}$, see Fig. 3, the displacement and reference placement vectors read

$$\mathbf{u} = \boldsymbol{v}_{\mathrm{X}}(\mathbf{X}) = u^i\,\mathbf{E}_i \qquad \text{and} \qquad \mathbf{X} = \boldsymbol{\kappa}_\Theta(\boldsymbol{\Theta}) = X^i\,\mathbf{E}_i\,. \tag{33}$$

Here, the coefficients $u^i$ can be considered either as functions of the referential coordinates $X^i$ or of the intrinsic coordinates $\Theta^i$, i.e.

$$u^i = v^i_{\mathrm{X}}(X^1, X^2, X^3) = v_\Theta(\Theta^1, \Theta^2, \Theta^3) \quad \text{and} \quad X^i = \kappa_\Theta(\Theta^1, \Theta^2, \Theta^3), \tag{34}$$

respectively. The referential and intrinsic gradients of the displacement mapping as well as the intrinsic gradient of the reference mapping and its inverse are given by

$$\begin{aligned} \mathbf{H}_{\mathrm{X}} &= \mathrm{Grad}\,\boldsymbol{v}_{\mathrm{X}} = \tfrac{\partial v^i_{\mathrm{X}}}{\partial X^j}\,\mathbf{E}_i \otimes \mathbf{E}^j\,, \qquad \mathbf{H}_\Theta = \mathrm{GRAD}\,\boldsymbol{v}_\Theta = \tfrac{\partial v^i_\Theta}{\partial \Theta^j}\,\mathbf{E}_i \otimes \mathbf{Z}^j \\ \mathbf{K}_\Theta &= \mathrm{GRAD}\,\boldsymbol{\kappa}_\Theta = \tfrac{\partial \kappa^i_\Theta}{\partial \Theta^j}\,\mathbf{E}_i \otimes \mathbf{Z}^j\,, \qquad \mathbf{K}^{-1}_\Theta = \mathrm{Grad}\,\boldsymbol{\kappa}^{-1}_\Theta = \tfrac{\partial \Theta^i}{\partial X^j}\,\mathbf{Z}_i \otimes \mathbf{E}^j\,. \end{aligned} \tag{35}$$

All quantities are depicted using bold letters in upright shape, i.e. serif type for vectors (e.g. $\boldsymbol{\Theta}$, $\mathbf{X}$, $\mathbf{x}$, $\mathbf{u}$) and sans serif type for tensors of 2nd order (e.g. $\mathbf{K}_\Theta$, $\mathbf{M}_\Theta$, $\mathbf{F}_{\mathrm{X}}$).

The relationship between the referential gradient $\mathrm{Grad}\,\mathbf{H}_{\mathrm{X}}$ and the intrinsic gradients $\mathrm{GRAD}\,\mathbf{H}_\Theta$ and $\mathrm{GRAD}\,\mathbf{K}_\Theta$ has been established in Sect. 3, see Eq. (16). Consequently, the matrix and coordinate versions can be derived yielding

$$\boldsymbol{H}_{\mathrm{X}} = \boldsymbol{H}_\Theta\,\boldsymbol{K}^{-1}_\Theta \quad \text{with} \quad \frac{\partial v^i_{\mathrm{X}}}{\partial X^j} = \frac{\partial v^i_\Theta}{\partial \Theta^k}\,\frac{\partial \Theta^k}{\partial X^j}. \tag{36}$$

Herein, the *Jacobian matrix* of the coordinate transformation is

$$\boldsymbol{K}_\Theta \equiv \boldsymbol{J}_{K_\Theta} = \left[\frac{\partial \kappa^i_\Theta}{\partial \Theta^j}\right] \quad \text{with} \quad \boldsymbol{K}^{-1}_\Theta \equiv \boldsymbol{J}^{-1}_{K_\Theta} = \left[\frac{\partial \Theta^i}{\partial X^j}\right] = \left[\frac{\partial \kappa^i_\Theta}{\partial \Theta^j}\right]^{-1}. \tag{37}$$

The coefficient matrix $\boldsymbol{K}_\Theta$ of the tensorial geometry gradient $\mathbf{K}_\Theta$ is labelled as *Jacobian matrix* $\boldsymbol{J}_{K_\Theta}$ of the coordinate transformation between intrinsic and referential coordinates. Here, $\boldsymbol{J}_{K_\Theta}$ is preferred over $\boldsymbol{K}_\Theta$ to distinguish it properly from the stiffness matrix $\hat{\boldsymbol{K}}$. Thus, Eq. (36) and a similar expression for local coordinates read

$$\boldsymbol{H}_{\mathrm{X}} = \boldsymbol{H}_\Theta\,\boldsymbol{K}^{-1}_\Theta = \boldsymbol{H}_\Theta\,\boldsymbol{J}^{-1}_{K_\Theta} \quad \text{and} \quad \boldsymbol{H}_{\mathrm{X}} = \boldsymbol{H}_\zeta\,\boldsymbol{K}^{-1}_\zeta = \boldsymbol{H}_\zeta\,\boldsymbol{J}^{-1}_{K_\zeta}\,. \tag{38}$$

Similar results are available for the deformation gradient

$$\boldsymbol{F}_{\mathrm{X}} = \boldsymbol{M}_{\Theta}\, \boldsymbol{K}_{\Theta}^{-1} = \boldsymbol{M}_{\Theta}\, \boldsymbol{J}_{K_{\Theta}}^{-1} \quad \text{and} \quad \boldsymbol{F}_{\mathrm{X}} = \boldsymbol{M}_{\zeta}\, \boldsymbol{K}_{\zeta}^{-1} = \boldsymbol{M}_{\zeta}\, \boldsymbol{J}_{K_{\zeta}}^{-1}\,. \tag{39}$$

The coefficient matrices $\boldsymbol{F}_{\mathrm{X}} = [x_{i,j}]$ and $\boldsymbol{H}_{\mathrm{X}} = [u_{i,j}]$ are outlined above, where $x_{i,j}$ and $u_{i,j}$ denote derivatives of $x_i$ and $u_i$ with respect to $X_j$, respectively. The unity matrix is $\boldsymbol{I} = [\delta_{ij}]$. The coefficient matrix of the 2. Piola-Kirchhoff stress tensor $\boldsymbol{S} = [S_{ij}]$ can be written in *Voigt notation* $\bar{\boldsymbol{S}} = [S_{11}, S_{22}, S_{12}]^{\mathrm{T}}$. Similarly, $\boldsymbol{E} = [E_{ij}]$ and $\bar{\boldsymbol{E}} = [E_{11}, E_{22}, 2\,E_{12}]^{\mathrm{T}}$ are used for the Green-Lagrange strain tensor. Finally, $\boldsymbol{C} = [C_{ij}]$ is the corresponding Voigt matrix of a fourth-order elasticity tensor, i.e. either $\mathbb{C}$ in the nonlinear case or $\mathbb{E}$ in the linear case, respectively.

*Remark 9 (From tensorial to matrix notation)* The reduction of the general tensorial to a matrix notation with respect to Cartesian base vectors is a necessary step to obtain a numerical method. But, it is nearly impossible to generalise results obtained on the matrix level back to the more general tensorial theory.

## *4.3 Fundamentals of Finite Element Approximation*

The finite element approximation is based on shape and Ansatz functions. Following the isoparametric concept, the state $\boldsymbol{u}$ and the geometry $\boldsymbol{X}$ are approximated by the same functions $N_I(\boldsymbol{\zeta})$ defined on a fixed (local) parameter space with (local) coordinates $(\zeta^1, \zeta^2, \zeta^3)$. According to the classical Bubnov-Galerkin technique also the test functions are interpolated using the shape functions $N_I(\zeta)$.

The state function $\boldsymbol{u}$, the test function $\boldsymbol{v}$ and the geometry $\boldsymbol{X}$ are approximated in every element $\mathcal{K}_e$ in the form

$$\boldsymbol{u}_h^e = \sum_{I=1}^{n} N_I\, \boldsymbol{u}_I\,, \quad \boldsymbol{v}_h^e = \sum_{I=1}^{n} N_I\, \boldsymbol{v}_I \quad \text{and} \quad \boldsymbol{X}_h^e = \sum_{I=1}^{n} N_I\, \boldsymbol{X}_I\,. \tag{40}$$

The corresponding displacement and design variations are

$$\delta\boldsymbol{u}_h^e = \sum_{I=1}^{n} N_I\, \delta\boldsymbol{u}_I \qquad \text{and} \qquad \delta\boldsymbol{X}_h^e = \sum_{I=1}^{n} N_I\, \delta\boldsymbol{X}_I\,, \tag{41}$$

where $n$ denotes the number of nodes per element. The vectors for nodal values read

$$\boldsymbol{u}_I = \begin{bmatrix} u_I^1 \\ u_I^2 \end{bmatrix}, \ \boldsymbol{v}_I = \begin{bmatrix} v_I^1 \\ v_I^2 \end{bmatrix}, \ \boldsymbol{X}_I = \begin{bmatrix} X_I^1 \\ X_I^2 \end{bmatrix}, \ \delta\boldsymbol{u}_I = \begin{bmatrix} \delta u_I^1 \\ \delta u_I^2 \end{bmatrix}, \ \delta\boldsymbol{X}_I = \begin{bmatrix} \delta X_I^1 \\ \delta X_I^2 \end{bmatrix}. \tag{42}$$

For notational simplicity, the explicit forms of all derived matrices are given for the two-dimensional case only.

The discrete residual and the tangent forms, see Sect. 2.2, can now be specified. The discrete element contributions $R_h^e$, $k_h^e$, $p_h^e$ are collected over all elements and the element matrices $\boldsymbol{r}^e$, $\boldsymbol{k}^e$ and $\boldsymbol{p}^e$ consists of all nodal contributions, i.e. overall

$$R(\boldsymbol{u}_h, \boldsymbol{X}_h; \boldsymbol{v}_h) = \sum_{e=1}^{nel} \left[ \sum_{I=1}^{n} \boldsymbol{v}_I^{\mathrm{T}} \boldsymbol{r}_I^e \right] = \hat{\boldsymbol{V}}^{\mathrm{T}} \hat{\boldsymbol{R}}, \tag{43}$$

$$k(\hat{\boldsymbol{u}}_h, \hat{\boldsymbol{X}}_h; \boldsymbol{v}_h, \delta\boldsymbol{u}_h) = \sum_{e=1}^{nel} \left[ \sum_{I=1}^{n} \sum_{J=1}^{n} \boldsymbol{v}_I^{\mathrm{T}} \boldsymbol{k}_{IJ}^e \, \delta\boldsymbol{u}_J \right] = \hat{\boldsymbol{V}}^{\mathrm{T}} \hat{\boldsymbol{K}} \, \delta\hat{\boldsymbol{U}}, \tag{44}$$

$$p(\hat{\boldsymbol{u}}_h, \hat{\boldsymbol{X}}_h; \boldsymbol{v}_h, \delta\boldsymbol{X}_h) = \sum_{e=1}^{nel} \left[ \sum_{I=1}^{n} \sum_{J=1}^{n} \boldsymbol{v}_I^{\mathrm{T}} \boldsymbol{p}_{IJ}^e \, \delta\boldsymbol{X}_J \right] = \hat{\boldsymbol{V}}^{\mathrm{T}} \hat{\boldsymbol{P}} \, \delta\hat{\boldsymbol{X}}. \tag{45}$$

The techniques to assemble the global quantities $\hat{\boldsymbol{U}}$, $\hat{\boldsymbol{V}}$, $\hat{\boldsymbol{X}}$, $\delta\hat{\boldsymbol{U}}$, $\delta\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{R}}$, $\hat{\boldsymbol{K}}$, $\hat{\boldsymbol{P}}$ are well-known, see Sect. 5 for hints on the implementation. Thus, the nodal contributions $\boldsymbol{r}_I^e$, $\boldsymbol{k}_{IJ}^e$ and $\boldsymbol{p}_{IJ}^e$ still need to be specified.

Further details as well as the implementation of sensitivity relations of the *material or inverse motion problem* and for the *dual or adjoint problem* is given in [24].

## 4.4 FEM Approximations of Gradient and Divergence

The gradient and the divergence of any quantity $\boldsymbol{a}_h^e$, which is either $\boldsymbol{u}$, $\boldsymbol{v}$, $\delta\boldsymbol{u}$ or $\delta\boldsymbol{X}$, with respect to the approximation $\boldsymbol{X}_h^e$ of the referential geometry, defined on each element in form of Eq. (40), is given for

$$\boldsymbol{a}_h^e = \sum_{I=1}^{n} N_I \, \boldsymbol{a}_I \quad \text{by} \quad \operatorname{Grad} \boldsymbol{a}_h^e = \sum_{I=1}^{n} \boldsymbol{a}_I \, \boldsymbol{L}_I^{\mathrm{T}} \quad \text{and} \quad \operatorname{Div} \boldsymbol{a}_h^e = \sum_{I=1}^{n} \boldsymbol{L}_I^{\mathrm{T}} \, \boldsymbol{a}_I, \tag{46}$$

where $\boldsymbol{L}_I$ denotes the gradient of the shape function $N_I$, i.e.

$$\boldsymbol{L}_I := \operatorname{Grad} N_I = \left[ N_{I,1} \; N_{I,2} \right]^{\mathrm{T}} = \left[ N_{I,X} \; N_{I,Y} \right]^{\mathrm{T}}, \tag{47}$$

where the notation $(X, Y)$ can be used instead of $(X^1, X^2)$.

## 4.5 FEM Approximation of Variations of the Strain Tensor

The variation of the Green-Lagrange strain tensor $\mathbf{E}$ with respect to the state $\mathbf{u}$ is

$$\mathbf{E}'_u(\mathbf{u}, \delta\mathbf{u}) = \operatorname{sym}\left( \mathbf{A}_u^{\mathsf{T}} \operatorname{Grad} \delta\mathbf{u} \right) \quad \text{with} \quad \mathbf{A}_u := \mathbf{F}_{\mathrm{X}},$$

see Eq. (20). The corresponding finite element approximation can be written as

$$\bar{\boldsymbol{E}}'_u(\boldsymbol{u}_h, \delta\boldsymbol{u}_h) = \left[(E'_u)_{11}\ (E'_u)_{22}\ 2\,(E'_u)_{12}\right]^{\mathrm{T}} = \sum_{I=1}^{n} \boldsymbol{B}_{uI}\,\delta\boldsymbol{u}_I \tag{48}$$

with

$$\boldsymbol{B}_{uI} = \begin{bmatrix} A_u^{11}\,N_{I,X} & A_u^{21}\,N_{I,X} \\ A_u^{12}\,N_{I,Y} & A_u^{22}\,N_{I,Y} \\ A_u^{11}\,N_{I,Y} + A_u^{12}\,N_{I,X} & A_u^{21}\,N_{I,Y} + A_u^{22}\,N_{I,X} \end{bmatrix}. \tag{49}$$

Furthermore, the design variation of $\mathbf{E}$ has been introduced in Eq. (20) too

$$\mathbf{E}'_X(\mathbf{u}, \delta\mathbf{X}) = \operatorname{sym}\left\{\mathbf{A}_X^{\mathsf{T}}\operatorname{Grad}\delta\mathbf{X}\right\} \quad \text{with} \quad \mathbf{A}_X := -\operatorname{Grad}^{\mathsf{T}}\mathbf{u}\,\mathbf{F}_{\mathrm{X}}.$$

The finite element approximation can be written in the same manner as above

$$\bar{\boldsymbol{E}}'_X(\mathbf{u}_h, \delta\mathbf{X}_h) = \left[(E'_X)_{11}\ (E'_X)_{22}\ 2\,(E'_X)_{12}\right]^{\mathrm{T}} = \sum_{I=1}^{n} \boldsymbol{B}_{sI}\,\delta\boldsymbol{X}_I \tag{50}$$

with

$$\boldsymbol{B}_{sI} = \begin{bmatrix} A_X^{11}\,N_{I,X} & A_X^{21}\,N_{I,X} \\ A_X^{12}\,N_{I,Y} & A_X^{22}\,N_{I,Y} \\ A_X^{11}\,N_{I,Y} + A_X^{12}\,N_{I,X} & A_X^{21}\,N_{I,Y} + A_X^{22}\,N_{I,X} \end{bmatrix}. \tag{51}$$

### 4.6 Approximation of Residual and Tangent Forms

Using the above introduced finite element approximations, the nodal contributions of the discrete element residual vector at node $I$ is given by

$$\boldsymbol{r}_I^e = \int_{\mathcal{K}^e} \boldsymbol{B}_{uI}^{\mathrm{T}}\,\bar{\boldsymbol{S}}\,\mathrm{d}V_{\mathrm{X}} - \boldsymbol{f}_I^e. \tag{52}$$

The vector $\boldsymbol{f}_I^e$ is the standard nodal vector of the external forces. Furthermore, the nodal contributions of the element matrices $\boldsymbol{k}^e$ and $\boldsymbol{p}^e$ are obtained as

$$\boldsymbol{k}_{IJ}^e = \int_{\mathcal{K}^e} \left\{ \boldsymbol{B}_{uI}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{B}_{uJ} + \boldsymbol{L}_I^{\mathrm{T}}\boldsymbol{S}\boldsymbol{L}_J\boldsymbol{I} \right\} \mathrm{d}V_{\mathrm{X}}, \tag{53}$$

$$\boldsymbol{p}^e_{IJ} = \int_{\boldsymbol{\mathcal{K}}^e} \left\{ \boldsymbol{B}^{\mathrm{T}}_{uI} \boldsymbol{C} \boldsymbol{B}_{sJ} - \boldsymbol{L}^{\mathrm{T}}_I \boldsymbol{S} \boldsymbol{L}_J \boldsymbol{H} - \boldsymbol{F} \boldsymbol{S} \boldsymbol{L}_J \boldsymbol{L}^{\mathrm{T}}_I + \boldsymbol{F} \boldsymbol{S} \boldsymbol{L}_I \boldsymbol{L}^{\mathrm{T}}_J \right\} \mathrm{d}V_{\mathrm{X}} - \int_{\boldsymbol{\mathcal{K}}^e} N_I \, \boldsymbol{b}_{\mathrm{X}} \, \boldsymbol{L}^{\mathrm{T}}_J \, \mathrm{d}V_{\mathrm{X}}. \tag{54}$$

*Remark 10 (Similarity of stiffness and pseudo load matrix)* It is important to observe that the structures of stiffness and pseudo load computations are fully similar. Thus, sensitivity information can be derived, implemented and computed on the element level with a small additional effort compared to 'standard' FEM computations.

### *4.7 Possible Fields of Application*

There is a great number of fields where the presented sensitivity relations can be applied. All quantities are derived with respect to coordinates of FE-nodes. Parameter free shape optimisation can be performed based on these derivatives utilising some additional tools like filters and mesh control techniques. Recent works on this topic are [1, 9, 15, 22, 32]. Gradients for geometry based shape optimisation can be calculated extending the presented sensitivities by the corresponding design velocity fields. R-adaptivity is concerned with improvement of finite element solution on the same mesh. Here, the number of degrees of freedom and mesh topology are fixed. Only the mesh form is changed. A review, much more details and examples on this topic can be found in [26]. Fracture mechanics deals with the propagation of cracks in materials. Here, the strain energy release and the direction of crack growth can be directly derived from the material residuum, see [25] for details and examples. Furthermore, the technique was applied to history dependent problems, see [7].

Last but not least, the pseudo load and sensitivity matrices $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{S}}$, respectively, can be decomposed using a *singular value decomposition* (SVD). The insight gained from the singular value structure and from the interpretation of the corresponding singular vectors can be used for model reduction, see [15].

## 5 Details on Numerical Implementation

In this section, we present a prototype implementation of the quantities and topics concerned with structural optimisation introduced in the previous chapters. Therefore, the well-known structure of the nonlinear finite element method is extended to sensitivity analysis, i.e. the pseudo load matrix $\hat{\boldsymbol{P}}$ from Eq. (54) is added. However, we do not focus on general details of FEM, see the standard literature on finite element analysis, for example [8, 10, 35, 36].

The presented Matlab code, i.e. the element routine, is part of an educational in-house finite element analysis environment for general nonlinear problems. We concentrate on the plane strain two-dimensional case using a quadrilateral four node

element with bilinear shape functions, which is sufficient to explain the necessary steps for sensitivity analysis. The extension to three dimensions and to higher order elements is straightforward.

We slightly differ from the notation used in the previous sections in order to avoid super- and subindices, but the meaning of the variables in the code should be obvious. We abbreviate the Cartesian and local coordinates $(X^1, X^2)$ and $(\zeta^1, \zeta^2)$ with (X, Y) and (a, b), respectively.

We name the function containing the element matrices

```
function[ Rint, Fvol, K, P ] = plane_nl( coorde, mate, be, Ue )
```

for the referred *plane, pure displacement formulation for nonlinear problems*. The overall implementation can be found in the appendix. Its integration into an already existing environment for structural analysis is easy to handle by using the standard Matlab syntax for function calls

```
[Out1, Out2, ...] = FunctionName(In1, In2, ...).
```

The input quantities, namely `coorde`, `mate`, `be`, `Ue`, contain the nodal coordinates of the current element, properties of the chosen material, information about body forces per unit volume acting on the system and the displacements for all degrees of freedom of the element, respectively. These details are specified below.

The discretisation of the domain delivers a matrix with the global `X`- and `Y`-coordinates of all nodes of the generated mesh. On element level, the necessary coordinate matrix for the plane two-dimensional case has the form

$$\texttt{coorde} = \begin{bmatrix} \texttt{X}_1 \ \ldots \ \texttt{X}_\texttt{n} \\ \texttt{Y}_1 \ \ldots \ \texttt{Y}_\texttt{n} \end{bmatrix}^\texttt{T} = \begin{bmatrix} \texttt{X}_1 \ \texttt{X}_2 \ \texttt{X}_3 \ \texttt{X}_4 \\ \texttt{Y}_1 \ \texttt{Y}_2 \ \texttt{Y}_3 \ \texttt{Y}_4 \end{bmatrix}^\texttt{T}, \tag{55}$$

where `n` is the overall number of nodes of the element and in our case is $\texttt{n} = 4$. Furthermore, the introduced matrix `mate` contains the material properties of the current element with the components `mate(1)` for the Young's modulus, `mate(2)` for the Poisson's ration and `mate(3)` for the thickness of the element. The matrix `be` represents the body forces per unit volume in the possible directions `X` and `Y`, i.e. it contains the components $\texttt{be} = \left[\texttt{b}_\texttt{X} \ \texttt{b}_\texttt{Y}\right]^\texttt{T}$. The last input value `Ue` includes the displacements of all nodes in the directions `X` and `Y`. Therefore, its dimension is the number of degrees of freedom of the current element $\texttt{ndof} \times 1$ and in the referred case $8 \times 1$

$$\texttt{Ue} = \left[\texttt{U}_1 \ \ldots \ \texttt{U}_\texttt{ndof}\right]^\texttt{T} = \left[\texttt{U}_1 \ \texttt{U}_2 \ \texttt{U}_3 \ \texttt{U}_4 \ \texttt{U}_5 \ \texttt{U}_6 \ \texttt{U}_7 \ \texttt{U}_8\right]^\texttt{T}. \tag{56}$$

The output quantities `Rint` for the internal residual, `Fvol` for the contribution of body forces to the external residual, `K` as the well known tangent stiffness matrix and the tangent pseudo load matrix `P` were already introduced in the previous sections.

We apply Gauss quadrature with four Gauss points stored in matrix gpoints with the four related weights in weights, which are organised within an external function

```
30  [ gpoints, weights, numgp ] = gaussquad(n);
```

and provide the matrix representations for the chosen $\mathtt{numgp} = 4$ integration points

$$\mathtt{gpoints} = \begin{bmatrix} \mathtt{a}_1 & \mathtt{a}_2 & \mathtt{a}_3 & \mathtt{a}_4 \\ \mathtt{b}_1 & \mathtt{b}_2 & \mathtt{b}_3 & \mathtt{b}_4 \end{bmatrix}^{\mathrm{T}}, \qquad \mathtt{weights} = \begin{bmatrix} \mathtt{w}_1 & \mathtt{w}_2 & \mathtt{w}_3 & \mathtt{w}_4 \end{bmatrix}^{\mathrm{T}}. \tag{57}$$

The bilinear shape functions are delivered by an external function as well

```
38  [ Nmat, N_X, N_Y, detJ ] = shape_plane(a, b, coorde);
```

and depend on the local coordinates a, b of the integration points and the coordinates of the nodes of the element. The output matrices contain the shape functions, summarised in Nmat, the derivatives of the shape functions with respect to the global coordinates X and Y, summarised in N_X, N_Y and the determinant of the Jacobian detJ. The explicit forms of the matrices are as follows

$$\mathtt{Nmat} = \begin{bmatrix} \mathtt{N}_1 & 0 & \mathtt{N}_2 & 0 & \mathtt{N}_3 & 0 & \mathtt{N}_4 & 0 \\ 0 & \mathtt{N}_1 & 0 & \mathtt{N}_2 & 0 & \mathtt{N}_3 & 0 & \mathtt{N}_4 \end{bmatrix} \tag{58}$$

and

$$\mathtt{N\_X} = \begin{bmatrix} \mathtt{N}_{1,\mathtt{X}} & \mathtt{N}_{2,\mathtt{X}} & \mathtt{N}_{3,\mathtt{X}} & \mathtt{N}_{4,\mathtt{X}} \end{bmatrix}^{\mathrm{T}}, \qquad \mathtt{N\_Y} = \begin{bmatrix} \mathtt{N}_{1,\mathtt{Y}} & \mathtt{N}_{2,\mathtt{Y}} & \mathtt{N}_{3,\mathtt{Y}} & \mathtt{N}_{4,\mathtt{Y}} \end{bmatrix}^{\mathrm{T}}. \tag{59}$$

The gradient of the displacements is performed by the function

```
40  Gradu = grad_disp(a, b, coorde, Ue);
```

and delivers a matrix

$$\mathtt{Gradu} = \begin{bmatrix} \mathtt{u}_{1,\mathtt{X}} & \mathtt{u}_{1,\mathtt{Y}} \\ \mathtt{u}_{2,\mathtt{X}} & \mathtt{u}_{2,\mathtt{Y}} \end{bmatrix}. \tag{60}$$

Different constitutive equations, for example St. Venant or Neo-Hooke, can be used. The $(3 \times 3)$ material matrix C and the $(2 \times 2)$ matrix S of the 2. Piola-Kirchhoff stresses are computed in the external function

```
47  [ C, S ] = etensor(mate, Gradu);
```

The number of degrees of freedom dof for the current node are computed using

```
56  dofni = dof*ni-1:dof*ni;
```

Here, dof is the number of overall degrees of freedom per node and ni is the number of the current node of interest. Same holds true for the node nj.

Due to the fact, that the B-operator has a similar structure for each necessary case, we introduce the external function Bmat for its computation

```
85  function B = Bmat(N_X, N_Y, A)
```

Here, the introduced derivatives of the shape functions N_X, N_Y and a matrix A are the required inputs. The matrix A represents the matrix representation $\mathtt{A\_u} = \mathtt{F}$ of the deformation gradient for the matrices for structural analysis or $\mathtt{A\_X} = -\mathtt{Gradu}^{\mathrm{T}}\mathtt{F}$ for the contribution to the pseudo load matrix within sensitivity analysis

```
60  Bui = Bmat(N_X(ni), N_Y(ni), A_u);

70  Buj = Bmat(N_X(nj), N_Y(nj), A_u);
71  Bsj = Bmat(N_X(nj), N_Y(nj), A_X);
```

In the end and with all given hints the element matrices and simultaneously the outputs of the element function can be computed. All output quantities use the number of degrees of freedom for the determination of their dimensions. Therefore, the internal residual Rint as well as the contribution of the body forces to the external residual Fvol have the dimension $\mathtt{ndof} \times 1 = 8 \times 1$ and are organised as follows

$$\mathtt{Rint} = \left[\mathtt{R}_1\ \mathtt{R}_2\ \ldots\ \mathtt{R}_{\mathtt{ndof}}\right]^{\mathrm{T}} = \left[\mathtt{R}_1\ \mathtt{R}_2\ \mathtt{R}_3\ \mathtt{R}_4\ \mathtt{R}_5\ \mathtt{R}_6\ \mathtt{R}_7\ \mathtt{R}_8\right]^{\mathrm{T}} \tag{61}$$

and

$$\mathtt{Fvol} = \left[\mathtt{F}_1\ \mathtt{F}_2\ \ldots\ \mathtt{F}_{\mathtt{ndof}}\right]^{\mathrm{T}} = \left[\mathtt{F}_1\ \mathtt{F}_2\ \mathtt{F}_3\ \mathtt{F}_4\ \mathtt{F}_5\ \mathtt{F}_6\ \mathtt{F}_7\ \mathtt{F}_8\right]^{\mathrm{T}}. \tag{62}$$

The computation of Fvol for the contribution of body loads to external residual vector Rext is realised in the following lines

```
52  Fvol = Fvol + Nmat'*be * dV;
```

where dV results from the integration over the domain.

For the physical residual vector Rint the B-operator Bui as well as the vector representation of the 2. Piola-Kirchhoff stresses

```
48  Svec    = [ S(1,1); S(2,2); S(1,2) ];
```

is necessary and can be updated for each integration point as follows

```
62  Rint(dofni) = Rint(dofni) + Bui'*Svec * dV;
```

The symmetric tangent stiffness matrix K and the tangent pseudo load matrix P, which is not symmetric in general, have the same structure and are both of the dimension $\mathtt{ndof} \times \mathtt{ndof} = 8 \times 8$. Their computation is pretty similar and can be organised even in the same loops over the nodes, due to similar dependencies

```
74  Kij = ( Bui'*C*Buj + Li'*S*Lj*eye(dof,dof) ) * dV;
75  K(dofni,dofnj) = K(dofni,dofnj) + Kij;

77  Pij = ( Bui'*C*Bsj - Li'*S*Lj*Gradu - F*S*Lj*Li' + F*S*Li*Lj' ) * dV;
78  Fsij= Nmat(:,dofni)'*be*Lj' * dV;
79  P(dofni,dofnj) = P(dofni,dofnj) + Pij - Fsij;
```

Section 7 shows a simple application of the presented approach for sensitivity analysis. It is useful to comprehend the mentioned aspects and can be easily realised within tutorials or lectures on structural optimisation using the presented element formulation.

## 6 Analytical Derivatives of Discrete Equations

The *discrete differentiation approach* focuses on the discrete matrix formulation which has been derived for the finite element method, see Fig. 1. In this approach, all derivatives of the discrete functions with respect to the discrete variables are computed based on standard calculus, i.e. using chain and product rules. The method is well-known with extensive discussion in literature, see e.g. [11, 21]. Nevertheless, a few details are presented to highlight essential differences. For simplicity, the design variable $s$ represents any nodal coordinate $X_I^i$ with $i = 1, 2$ and $I = 1, 2, 3, 4$. Furthermore, we abbreviate the Jacobian matrix by $\boldsymbol{J}$ and its determinant by $J = \det \boldsymbol{J}$.

### *6.1 Design Derivatives of Shape Functions and Jacobians*

The isoparametric concept is an important concept in FEM and the computation of the Cartesian derivatives of the shape functions and of the Jacobian determinant play a central role in the discrete differentiation approach.

*Remark 11 (Isoparametric concept is derived from continuous theory)* In teaching finite elements, the (iso-) parametric technique to compute the Cartesian derivatives of the shape functions is often argued to be a novel concept introduced by FEM. This is wrong, because the underlying structure of differential geometry has been ignored.

Thus, the analytical or numerical differentiation of discrete functions belonging to the (iso-) parametric technique re-compute those results which are already available in more general form on the continuous level. Instead of using results from Sect. 3.3, Eq. (16) applied to the displacement approximation (40) is differentiated again

$$\frac{\partial}{\partial s}\left[N_{I,X}\ N_{I,Y}\right] = \frac{\partial}{\partial s}\left(\left[N_{I,a}\ N_{I,b}\right]\boldsymbol{J}^{-1}\right) = \left[N_{I,a}\ N_{I,b}\right]\frac{\partial}{\partial s}\boldsymbol{J}^{-1}. \qquad (63)$$

The Jacobian $\boldsymbol{J}$ as discrete version of $\mathbf{K}_\zeta$ or $\boldsymbol{K}_\zeta \equiv \boldsymbol{J}_{K_\zeta}$, see Eq. (37), is given by

$$\boldsymbol{J} = \begin{bmatrix} X_{,a} & X_{,b} \\ Y_{,a} & Y_{,b} \end{bmatrix} = \begin{bmatrix} X_1 & X_2 & X_3 & X_4 \\ Y_1 & Y_2 & Y_3 & Y_4 \end{bmatrix} \begin{bmatrix} N_{1,a} & N_{1,b} \\ N_{2,a} & N_{2,b} \\ N_{3,a} & N_{3,b} \\ N_{4,a} & N_{4,b} \end{bmatrix}. \tag{64}$$

The derivative of the inverse of the Jacobian with respect to nodal coordinates or design variables can be obtained using the identity $\boldsymbol{I} = \boldsymbol{J}^{-1}\boldsymbol{J}$ which leads to

$$\frac{\partial \boldsymbol{I}}{\partial s} = \frac{\partial \boldsymbol{J}^{-1}}{\partial s}\boldsymbol{J} + \boldsymbol{J}^{-1}\frac{\partial \boldsymbol{J}}{\partial s} = \boldsymbol{0} \quad \text{and therefore to} \quad \frac{\partial \boldsymbol{J}^{-1}}{\partial s} = -\boldsymbol{J}^{-1}\frac{\partial \boldsymbol{J}}{\partial s}\boldsymbol{J}^{-1}. \tag{65}$$

The design variable $s$ is an abbreviation for the nodal coordinates $X_I$, $Y_I$ of all nodes $I = 1, 2, 3, 4$ of the element. Thus, the design derivative of Eq. (64) yields

$$\frac{\partial \boldsymbol{J}}{\partial s} = \begin{cases} \begin{bmatrix} N_{I,a} & N_{I,b} \\ 0 & 0 \end{bmatrix} & \text{for} \quad s = X_I \\ \\ \begin{bmatrix} 0 & 0 \\ N_{I,a} & N_{I,b} \end{bmatrix} & \text{for} \quad s = Y_I \end{cases}. \tag{66}$$

Further necessary quantity is the first derivative of the determinant of the Jacobian $J = \det \boldsymbol{J} = X_{,a}Y_{,b} - Y_{,a}X_{,b}$. It can be obtained by performing the product rule

$$\frac{\partial J}{\partial s} = \frac{\partial X_{,a}}{\partial s}Y_{,b} + X_{,a}\frac{\partial Y_{,b}}{\partial s} - \frac{\partial Y_{,a}}{\partial s}X_{,b} - Y_{,a}\frac{\partial X_{,b}}{\partial s}. \tag{67}$$

The nonlinear B-operator $\boldsymbol{B}_{uI}$ is quoted in Eq. (49). In the framework of the discrete sensitivity analysis its derivative with respect to the design variables has to be provided, i.e. every element of $\boldsymbol{B}_{uI}$ must be differentiated. With the introduced quantity $\boldsymbol{A}_u = \boldsymbol{F}_X = \boldsymbol{I} + \operatorname{Grad}\boldsymbol{u}$, see Sect. 4.5, its derivative corresponds to the derivative of the deformation gradient

$$\frac{\partial \boldsymbol{A}_u}{\partial s} = \frac{\partial \boldsymbol{F}_X}{\partial s} = \frac{\partial}{\partial s}(\boldsymbol{I} + \operatorname{Grad}\boldsymbol{u}) = \frac{\partial \operatorname{Grad}\boldsymbol{u}}{\partial s}. \tag{68}$$

The definition of the gradient in Eq. (46) allows the computation of the derivative with respect to the nodal coordinates or the design variables in the following way

$$\frac{\partial \operatorname{Grad}\boldsymbol{u}}{\partial s} = \frac{\partial}{\partial s}\left(\sum_{I=1}^{n} \boldsymbol{u}_I \boldsymbol{L}_I^{\mathrm{T}}\right) = \sum_{I=1}^{n} \boldsymbol{u}_I \frac{\partial \boldsymbol{L}_I^{\mathrm{T}}}{\partial s} = \sum_{I=1}^{n} \boldsymbol{u}_I \begin{bmatrix} \frac{\partial N_{I,X}}{\partial s} & \frac{\partial N_{I,Y}}{\partial s} \end{bmatrix}. \tag{69}$$

## 6.2 Design Derivatives of the Linear Stiffness Matrix

In the framework of a static and linear finite element analysis, the discrete equilibrium condition is usually presented as $\hat{\boldsymbol{K}}\,\hat{\boldsymbol{U}} = \hat{\boldsymbol{F}}$. Thus, sensitivity analysis reads

$$\frac{\partial \hat{\boldsymbol{K}}}{\partial s}\,\hat{\boldsymbol{U}} + \hat{\boldsymbol{K}}\,\frac{\mathrm{d}\hat{\boldsymbol{U}}}{\mathrm{d}s} = \frac{\partial \hat{\boldsymbol{F}}}{\partial s} \quad \text{and therefore} \quad \frac{\mathrm{d}\,\hat{\boldsymbol{U}}}{\mathrm{d}s} = \hat{\boldsymbol{K}}^{-1}\left[\frac{\partial \hat{\boldsymbol{F}}}{\partial s} - \frac{\partial \hat{\boldsymbol{K}}}{\partial s}\,\hat{\boldsymbol{U}}\right], \tag{70}$$

where $s$ is a scalar valued design variable. This approach suggests to differentiate the linear stiffness matrix by applying the chain rule to all element contributions $\boldsymbol{k}_e$

$$\begin{aligned}\frac{\partial \boldsymbol{k}_e}{\partial s} &= \frac{\partial}{\partial s}\left(\int_{\mathcal{R}^e} \boldsymbol{B}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{B}\,J\,\mathrm{d}V_\zeta\right) = \int_{\mathcal{R}^e}\frac{\partial}{\partial s}\left(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{B}\,J\right)\mathrm{d}V_\zeta \\ &= \int_{\mathcal{R}^e}\left(\frac{\partial \boldsymbol{B}^{\mathrm{T}}}{\partial s}\boldsymbol{C}\boldsymbol{B} + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{C}\frac{\partial \boldsymbol{B}}{\partial s}\right)J\,\mathrm{d}V_\zeta + \int_{\mathcal{R}^e}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{B}\frac{\partial J}{\partial s}\,\mathrm{d}V_\zeta,\end{aligned} \tag{71}$$

where the analytical derivatives of $\boldsymbol{B}$ and $J$ are discussed above.

*Remark 12 (No derivative of the stiffness matrix is necessary)* The variational sensitivity analysis emphasises that the continuous residuum (weak form) must be varied or alternatively, that the discrete residual vector must be differentiated. The special form $\hat{\boldsymbol{K}}\,\hat{\boldsymbol{U}} = \hat{\boldsymbol{F}}$ is irritating and leads to a higher effort than needed, i.e. additional analytical derivatives must be derived and must be implemented. Last but not least, the computational performance is less efficient as outlined in Sect. 7.

## 6.3 Design Derivatives of Nonlinear Residual Vectors

Referring Remark 12, the design sensitivity analysis for the static nonlinear case has to be performed starting with the discrete equilibrium condition for finite element analysis introduced in Eq. (52). In detail, the derivative with respect to the scalar valued design variable $s$ can be evaluated by

$$\frac{\partial \boldsymbol{r}_I^e}{\partial s} = \frac{\partial}{\partial s}\left(\int_{\mathcal{K}^e}\boldsymbol{B}_{uI}^{\mathrm{T}}\bar{\boldsymbol{S}}\,\mathrm{d}V_{\mathrm{X}} - \boldsymbol{f}_I^e\right). \tag{72}$$

For the internal part of the residual the derivative reads

$$\begin{aligned}\frac{\partial}{\partial s}\int_{\mathcal{K}^e}\boldsymbol{B}_{uI}^{\mathrm{T}}\bar{\boldsymbol{S}}\,\mathrm{d}V_{\mathrm{X}} &= \frac{\partial}{\partial s}\int_{\mathcal{R}^e}\boldsymbol{B}_{uI}^{\mathrm{T}}\bar{\boldsymbol{S}}\,J\,\mathrm{d}V_\zeta = \int_{\mathcal{R}^e}\frac{\partial}{\partial s}\left(\boldsymbol{B}_{uI}^{\mathrm{T}}\bar{\boldsymbol{S}}J\right)\mathrm{d}V_\zeta \\ &= \int_{\mathcal{R}^e}\left(\frac{\partial \boldsymbol{B}_{uI}^{\mathrm{T}}}{\partial s}\bar{\boldsymbol{S}} + \boldsymbol{B}_{uI}^{\mathrm{T}}\frac{\partial \bar{\boldsymbol{S}}}{\partial s}\right)J\,\mathrm{d}V_\zeta + \int_{\mathcal{R}^e}\boldsymbol{B}_{uI}^{\mathrm{T}}\bar{\boldsymbol{S}}\frac{\partial J}{\partial s}\,\mathrm{d}V_\zeta\end{aligned} \tag{73}$$

and similar for the external part

$$\frac{\partial \boldsymbol{f}_I^e}{\partial s} = \frac{\partial}{\partial s} \int_{\boldsymbol{\mathcal{K}}^e} N_I \boldsymbol{b}_I \, \mathrm{d}V_{\mathrm{X}} = \frac{\partial}{\partial s} \int_{\boldsymbol{\mathcal{R}}^e} N_I \boldsymbol{b}_I \, J \, \mathrm{d}V_\zeta = \int_{\boldsymbol{\mathcal{R}}^e} N_I \boldsymbol{b}_I \frac{\partial J}{\partial s} \, \mathrm{d}V_\zeta . \quad (74)$$

The essential steps for the derivative of the nonlinear B-operator and of the Jacobian determinant are already detailed in Sect. 6.1. For the derivative of stresses $\boldsymbol{S}$ the derivative of strains $\boldsymbol{E} = \frac{1}{2}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F} - \boldsymbol{I}) = \frac{1}{2}\left(\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{H} + \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\right)$ with $\boldsymbol{H} = \mathrm{Grad}\,\boldsymbol{u}$ has to be discussed. The derivative of $\boldsymbol{H}$ can be found in Eq. (69). A lot of further aspects and details concerning sensitivity analysis of nonlinear systems are discussed and presented in [21] as well.

*Remark 13 (Comparable quantities should not be treated differently)* It has been shown, that the stiffness matrix and pseudo load matrix originate from partial variations of the residual with respect to displacements or design, respectively. It is common and good practice in computational mechanics, to perform all variations with respect to the state variables on the continuous level before applying a subsequent discretisation step. But it is a strong discrepancy, if the differentiation of the discrete residual vector is advocated for deriving the pseudo load matrix.

## 7 Numerical Example

The example illustrates the usage of the variational design sensitivity analysis and should serve as a benchmark problem suitable for comparing different sensitivity analysis techniques. It is equally applicable for linear and nonlinear problems.

### *7.1 Structural Optimisation Problem*

The dimensions of the structure (height and width) are $h = 2$ and $w = 6$. The geometry is defined by a Bézier patch with 8 control points and the corresponding control polygon is pictured in Fig. 4b. The material properties are Young's modulus $E = 21000$ and Poisson's ratio $\nu = 0.3$. The constitutive model is either the classical Hooke's law in the linear case or the Neo-Hooke's law in the nonlinear case. The load $\bar{q} = 3$ is a line load. The applied boundary conditions are pictured in Fig. 4a. The FE-mesh consists of 675 elements and 736 nodes with 1472 degrees of freedom.

Vertical positions ($y$ coordinates) of 4 lower control points are used as design variables. The vertical displacement $u_l$ of the upper right corner is to be minimised taking into account a constant volume constraint $V = V_0$. Here, $V$ and $V_0$ denote the current and initial volumes. The resulting force $Q$ of the line load $\overline{q}$ is kept constant.

Sequential quadratic programming (SQP) is utilised to solve the optimisation problem. The algorithm converges after sixteen iterations. The corresponding

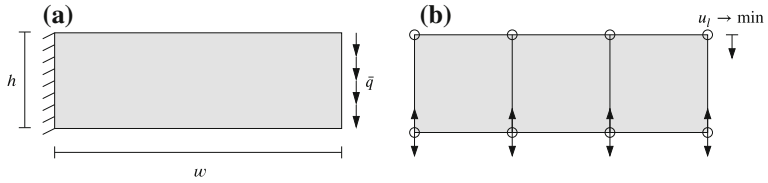



**Fig. 4** Cantilever beam: initial structure. **a** Mechanical system. **b** Optimisation model

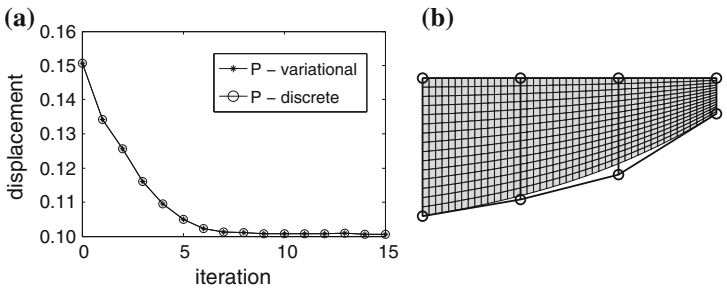



**Fig. 5** Cantilever beam: optimisation results. **a** Displacement history. **b** Optimised design

iteration history (steps 0–15) for the objective function is pictured in Fig. 5a. The displacement $u_l$ is decreased to about two-thirds of its initial value. The optimised design in terms of nonlinear structural analysis is presented in Fig. 5b. The structure is stiffened with respect to bending stress.

## *7.2 Performance Comparison of Different Strategies*

Performance tests have been run for linear and nonlinear structural analysis. In each case, the computation of the pseudo load matrix using either the variational formulation $\hat{\boldsymbol{P}}_{\text{var}}$, see Sect. 4, or the discrete formulation $\hat{\boldsymbol{P}}_{\text{dis}}$, see Sect. 6, have been performed. The absolute as well as the relative computation times for assembling the pseudo load matrices referred to the computation time for assembling the stiffness matrix $\hat{\boldsymbol{K}}$ are presented in Table 1.

**Table 1** Computation times for pseudo load matrices

| | Absolute values in seconds | | | Relative factors compared with $\hat{\boldsymbol{K}}$ | | | Speedup |
|---|---|---|---|---|---|---|---|
| | $\hat{\boldsymbol{K}}$ | $\hat{\boldsymbol{P}}_{\text{var}}$ | $\hat{\boldsymbol{P}}_{\text{dis}}$ | $\hat{\boldsymbol{K}}$ | $\hat{\boldsymbol{P}}_{\text{var}}$ | $\hat{\boldsymbol{P}}_{\text{dis}}$ | |
| Linear | 0.0465 | 0.2210 | 0.5431 | 1 | 4.7 | 11.7 | 2.5 |
| Nonlinear | 0.2758 | 0.3575 | 1.0455 | 1 | 1.3 | 3.8 | 2.9 |

Referring this times, the usage of the variational formulation provides a speed up of 2.5 for the linear and of 2.9 for the nonlinear computations in contrast to the discrete formulation. Beyond that, the times prove that the effort for the provision of the (nonlinear) pseudo load matrix using the variational formulation is in the same range as for the stiffness matrix. Due to the fact, that the linear stiffness matrix is constant and independent of displacements, the factors differ bit more than for the nonlinear case. Regarding the overall computation time for the presented example in the previous section, one ends up at least with a speed up of 1.3 for the nonlinear and of 1.7 for the linear case.

*Remark 14 (Variational technique shows superior computational performance)* The invested effort into a rigorous theoretical development and its careful implementation yields a minimal number of floating-point operations compared to the discrete differentiation approach. This advantage remains for all advanced computational techniques such as *High Performance Computing* (HPC) on all available hardware. Thus, in a long term run, there is no alternative to an investment in theoretical rigour.

## 8 Conclusions

The proposed variational design sensitivity analysis is considered to be the most efficient technique to determine the gradients of objective or constraint functions with respect to design variables in structural optimisation. The effort of a rigorous separation of physical quantities into geometry and displacement mappings based on an intrinsic presentation of continuum mechanics pays off with fundamental insight.

Moreover, the quote *There's nothing more practical than a good theory*, see [23], perfectly characterises the outlined benefits of a thorough theoretical investigation for the subsequent discretisation, implementation and computation of design sensitivity analysis in structural optimisation. Overall fourteen remarks, embedded in the text, substantiate the correctness of this statement.

## Appendix: MATLAB Source Code

The appended Matlab source code contains two functions, i.e. `plane_nl` for the computation of the element matrices and `Bmat` for the B-matrices needed in structural analysis and sensitivity analysis, see Sect. 5 for detailed explanations.

Matlab-source code

```
function[ Rint, Fvol, K, P ] = plane_nl( coorde, mate, be, Ue )
%PLANE_PDF_NL   PLANE NONLINEAR SOLID ELEMENT (PURE DISPLACEMENT FORMULATION)
% ============================================================================
%  Purpose:
%   Element matrices for plane nonlinear solid element
%   with isoparametric shape functions.
%  Input parameters:
%   coorde              global x / y-coordinates
%   mate                material parameters
%   mate(1)             Young's modulus
%   mate(2)             Poisson's ratio
%   mate(3)             thickness
%   be                  body load
%   Ue                  nodal displacements vector
%  Output Parameters:
%   Rint                physical residual vector, internal
%   Fvol                contribution of body loads to external residual vector
%   K                   tangent physical stiffness matrix   for current element
%   P                   tangent physical pseudo load matrix for current element
% ============================================================================
% INIT ELEMENT CONTRIBUTIONS
  n    = size(coorde,1);     % Number of nodes per element
  dof  = size(coorde,2);     % Number of DOFs per node
  Rint = zeros(n*dof,1);     % Physical residual vector,internal
  Fvol = zeros(n*dof,1);     % Contribution of body loads to external R
  K    = zeros(n*dof,n*dof); % Tangent physical stiffness matrix
  P    = zeros(n*dof,n*dof); % Tangent physical pseudo load matrix
  thick= mate(3);            % Thickness of current element
% Gauss quadrature points, weights, number
  [ gpoints, weights, numgp ] = gaussquad(n);
% ============================================================================
% COMPUTATION OF ELEMENT CONTRIBUTIONS
for gpi = 1:numgp
  % Coordinates of integration points
    a = gpoints(gpi,1);
    b = gpoints(gpi,2);
  % Shape functions and derivatives w.r.t. to X and Y
    [ Nmat, N_X, N_Y, detJ ] = shape_plane(a, b, coorde);
  % Displacement gradient
    Gradu = grad_disp(a, b, coorde, Ue);
  % Deformation gradient
    F    = eye(dof) + Gradu;
  % Input for B-operators Bui and Bsi
    A_u   = F;
    A_X   = -Gradu' * F;
  % Elasticity matrix and 2.P.K. stress as tensor and vector
    [ C, S ] = etensor(mate, Gradu);
    Svec     = [ S(1,1); S(2,2); S(1,2) ];
  % Contribution to volume of element
    dV = detJ*weights(gpi)*thick;
  % Contribution of body loads to external residual vector R
    Fvol = Fvol + Nmat'*be * dV;
% LOOP OVER NODES i
  for ni=1:n
    % Current DOF numbers
      dofni = dof*ni-1:dof*ni;
    % Gradient of shape function
      Li = [N_X(ni); N_Y(ni)];
    % Nonlinear B-operator in terms of A_u
      Bui = Bmat(N_X(ni), N_Y(ni), A_u);
    % Physical internal residual vector
      Rint(dofni) = Rint(dofni) + Bui'*Svec * dV;
%.. LOOP OVER NODES j
```

```
64      for nj=1:n
65        % Current DOF numbers
66          dofnj = dof*nj-1:dof*nj;
67        % Gradient of shape function
68          Lj = [N_X(nj); N_Y(nj)];
69        % Nonlinear B-operator in terms of A_u and A_X
70          Buj = Bmat(N_X(nj), N_Y(nj), A_u);
71          Bsj = Bmat(N_X(nj), N_Y(nj), A_X);
72    %.... ELEMENT MATRICES
73        % Tangent physical stiffness matrix
74          Kij = ( Bui'*C*Buj + Li'*S*Lj*eye(dof,dof) ) * dV;
75          K(dofni,dofnj) = K(dofni,dofnj) + Kij;
76        % Tangent physical pseudo load matrix
77          Pij = ( Bui'*C*Bsj - Li'*S*Lj*Gradu - F*S*Lj*Li' + F*S*Li*Lj' ) * dV;
78          Fsij= Nmat(:,dofni)'*be*Lj' * dV;
79          P(dofni,dofnj) = P(dofni,dofnj) + Pij - Fsij;
80      end
81    end
82  end
83  return
84  % ================================================================================
85  function B = Bmat(N_X, N_Y, A)
86  % Computation of B-matrix in terms of A
87    B = zeros(3,2);
88    B(1,1) = A(1,1)*N_X;
89    B(2,1) = A(1,2)*N_Y;
90    B(3,1) = A(1,1)*N_Y + A(1,2)*N_X;
91    B(1,2) = A(2,1)*N_X;
92    B(2,2) = A(2,2)*N_Y;
93    B(3,2) = A(2,1)*N_Y + A(2,2)*N_X;
94  return
```

## References


1. S. Arnout, M. Firl, K.U. Bletzinger, Parameter free shape and thickness optimisation considering stress response. Struct. Multi. Optim. **45**(6), 801–814 (2012)
2. J. Arora, An exposition of the material derivative approach for structural shape sensitivity analysis. Comp. Methods Appl. Mech. Eng. **105**, 41–62 (1993)
3. N. Banichuk, *Problems and Methods of Structural Optimal Design* (Plenum Press, New York, 1983)
4. F.J. Barthold, *Zur Kontinuumsmechanik inverser Geometrieprobleme* (Habilitation, TU Braunschweig, 2001)
5. F.J. Barthold, Remarks on variational shape sensitivity analysis based on local coordinates. Eng. Anal. Bound. Elem. **32**(11), 971–985 (2008)
6. F.J. Barthold, E. Stein, A continuum mechanical based formulation of the variational sensitivity analysis in structural optimization. Part I: analysis. Struct. Multi. Optim. **11**, 29–42 (1996)
7. F.J. Barthold, K. Wiechmann, Variational design sensitivity for inelastic deformations, in *Proceedings of COMPLAS 5*, ed. by D. Owen, E. Onate, E. Hinton (CIMNE, Barcelona, 1997), pp. 792–797
8. K.J. Bathe, *Finite Element Procedures* (Prentice-Hall, 1996)
9. K.U. Bletzinger, M. Firl, J. Linhard, R. Wüchner, Optimal shapes of mechanically motivated surfaces. Comput. Methods Appl. Mech. Eng. **199**(5–8), 324–333 (2010)
10. J. Bonet, R. Wood, *Nonlinear Continuum Mechanics for Finite Element Analysis* (Cambridge University Press, Cambridge, 1997)
11. R. Brockman, Geometric sensitivity analysis with isoparametric finite elements. Commun. Appl. Numer. Methods **3**, 495–499 (1987)

12. K. Choi, N.H. Kim, *Structural Sensitivity Analysis and Optimization*, Mechanical Engineering Series (Springer, Berlin, 2005)
13. K. Dems, Z. Mróz, Variational approach to first- and second-order sensitivity analysis of elastic structures. Int. J. Numer. Methods Eng. **21**, 637–661 (1985)
14. J.D. Eshelby, The force on an elastic singularity. Philos. Trans. R. Soc. Lond. **244**, 87–112 (1951)
15. N. Gerzen, D. Materna, F.J. Barthold, The inner structure of sensitivities in nodal based shape optimisation. Comput. Mech. **49**, 379–396 (2012)
16. M. Gurtin, *Configurational Forces as Basic Concepts of Continuum Physics* (Springer, New York, 2000)
17. R. Haber, A new variational approach to structural shape design sensitivity analysis, in *Computer Aided Optimal Design*, vol. 27, ed. by C. Mota Soares (Springer, New York, 1987), pp. 573–587
18. J. Haslinger, R.A.E. Mäkinen, *Introduction to Shape Optimization* (Society for Industrial and Applied Mathematics, Philadelphia, 2003)
19. E. Haug, K. Choi, V. Komkov, *Design Sensitivity Analysis of Structural Systems* (Academic Press, Orlando, 1986)
20. R. Kienzler, G. Maugin (eds.), *Configurational Mechanics of Materials* (Springer, Wien, 2001)
21. M. Kleiber, H. Antúnez, T. Hien, P. Kowalczyk, *Parameter Sensitivity in Nonlinear Mechanics: Theory and Finite Element Computations* (Wiley, Chichester, 1997)
22. C. Le, T. Bruns, D. Tortorelli, A gradient-based, parameter-free approach to shape optimization. Comput. Methods Appl. Mech. Eng. **200**(9–12), 985–996 (2011)
23. K. Lewin, *Field Theory in Social Science: Selected Theoretical Papers by Kurt Lewin* (Tavistock, London, 1952)
24. D. Materna, *Structural and Sensitivity Analysis for the Primal and Dual Problems in the Physical and Material Spaces* (Shaker Verlag, 2010)
25. D. Materna, F.J. Barthold, Variational design sensitivity analysis in the context of structural optimization and configurational mechanics. Int. J. Fract. **147**(1–4), 133–155 (2007)
26. D. Materna, F.J. Barthold, Goal-oriented r-adaptivity based on variational arguments in the physical and material spaces. Comput. Methods Appl. Mech. Eng. **198**(41–44), 3335–3351 (2009)
27. D. Materna, F.J. Barthold, Theoretical aspects and applications of variational sensitivity analysis in the physical and material space, in *Computational Optimization: New Research Developments*, ed. by R.F. Linton, T.B. Carroll (Nova Science Publishers, 2010), pp. 397–444
28. P. Michaleris, D. Tortorelli, C. Vidal, Tangent operators and design sensitivity formulations for transient non-linear coupled problems with applications to elastoplasticity. Int. J. Numer. Methods Eng. **37**, 2471–2499 (1994)
29. Z. Mróz, *Variational Approach to Sensitivity Analysis and Optimal Design* (Plenum Press, New York, 1986)
30. W. Noll, A new mathematical theory of simple materials. Arch. Ration. Mech. Anal. **102**(1) (1972)
31. O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer Series in Computational Physics (Springer, New York, 1984)
32. M. Scherer, R. Denzer, P. Steinmann, A fictitious energy approach for shape optimization. Int. J. Numer. Methods Eng. **82**, 269–302 (2010)
33. D. Tortorelli, Z. Wang, A systematic approach to shape sensitivity analysis. Int. J. Solids Struct. **30**(9), 1181–1212 (1993)
34. C. Truesdell, W. Noll, The nonlinear field theories of mechanics, in *Handbuch der Physik III/3*, ed. by S. Flügge (Springer, 1965)
35. P. Wriggers, *Nonlinear Finite Element Methods* (Springer, 2008)
36. O. Zienkiewicz, R. Taylor, R. Taylor, *The Finite Element Method* (Butterworth-Heinemann, 2000)

# A Variational Approach to Modelling and Optimization in Elastic Structure Dynamics

**Georgy Kostin and Vasily Saurin**

**Abstract** The paper studies dynamics modelling and control design for elastic systems with distributed parameters. The constitutive laws are specified by an integral equality according with the method of integro-differential relations. The original initial-boundary value problem is reduced to a constrained minimization problem for a nonnegative quadratic functional. A numerical algorithm is developed to solve direct and inverse dynamic problems in linear elasticity based on the Ritz method and finite element technique. The minimized functional is used to define an energy type criteria of solution quality. The efficiency of the approach is demonstrated on the example of a thin rectilinear elastic rod. The control problem is to find motion of a rod from an initial state to a terminal one at a fixed time with the minimal mean energy. The control input is presented by piecewise polynomial displacements on one end of the rod. It is possible to find the exact solution of the problem for a specific relation of the space-time mesh steps. The results of numerical analysis are presented and discussed.

**Keywords** Optimal control · Elastic structure · Dynamic system · Ritz method · Finite elements

**Mathematical Subject Classification:** 49S05 · 49J20 · 65M60 · 74K10

G. Kostin (✉) · V. Saurin
Institute for Problems in Mechanics RAS, Vernadskogo 101-1,
119526 Moscow, Russia
e-mail: kostin@ipmnet.ru

V. Saurin
e-mail: saurin@ipmnet.ru



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_15

## 1 Introduction

The design of control strategies for dynamic systems with distributed parameters has been actively developed in recent decades. Optimization of dynamic models of elastic structures is an important problem arising in a large variety of applications in science and engineering. Theoretical foundations of optimal control problems with linear partial differential equations (PDEs) and convex functionals were established by Lions [17, 18]. Linear hyperbolic equations are studied in [1, 4]. An introduction to the control of vibrations can be found in [13]. Oscillating elastic networks are investigated in [10, 14, 16]. Since accurate modelling of these systems leads to a description in terms of PDEs, control design is usually based on specific approaches to solving direct and inverse problems.

Two different methods to the control design for distributed parameter processes can be emphasized. In the first approach, (the so-called *late lumping*), the control is directly designed for distributed parameter models and then converted to a finite-dimensional approximation. The infinite-dimensional control strategies can rely on specific spectrum analysis of the linear system operator (see, e.g., [3, 7]). The control method considered in [5] enables one to construct a constrained distributed control in closed form and ensures that the system is brought to a given state in a finite time. This method is based on a decomposition of the original system into simple subsystems by the Fourier approach. In [9], a numerical approach for the solution of PDE-constrained optimal control problems is adapted to hyperbolic equations. The method of choice proposed there is either a full discretization method for small size problems or the vertical method of lines for medium size problems.

In applications, the second approach, *early lumping*, is used for numerical control design. In accordance with this approach, the initial-boundary value problem is first discretized and reduced to a system of ordinary differential equations (ODEs), e.g., by means of the Rayleigh-Ritz or Galerkin methods. A family of Galerkin approximations based on solutions of the homogeneous beam equation was constructed and sufficient conditions for stabilizability of such finite-dimensional systems were derived in [19]. Alternatively, the finite-difference or finite-element method (FEM) can be used as it is shown in [2, 6]. The direct discretization approaches are also known in optimal control theory (e.g., see [15]).

One of the disadvantages of the early lumping is that it is rather difficult to relate the discretized system with its original distributed model. However, this connection can be estimated by following the method of integro-differential relations (MIDR) [12]. These estimates allow us to qualify finite-dimensional modelling, refine a coarse solution and make necessary corrections of the control law. The MIDR was extended in [11] to the optimal control design of elastic rod motions. In the paper, this approach is combined with the Ritz method and FEM to minimize the mean energy distributed in an elastic structure during controlled processes.

The paper is structured as follows: In Sect. 2, the PDE system that models the elastic rod dynamics is introduced. A variational formulation of the considered

initial-boundary value problem is proposed in Sect. 3. In Sect. 4, a finite element algorithm is described based on this generalized statement. An optimal control problem for the elastic structure is formulated in Sect. 5. In the next section, the inverse dynamic problem is related with the proposed variational formulation of the direct problem. Section 7 is devoted to the FEM procedure including the successive minimization of the constitutive and control functionals. In Sect. 8, a numerical example of system modelling and optimization is presented and discussed. Finally, conclusions and a brief outlook are given.

## 2 Modelling of Elastic Rod Dynamics

As an example of elastic structure dynamics, longitudinal displacements of a thin rectilinear elastic rod are considered. In the Lagrange coordinate system, one end of the rod at $x = 0$ can move in accordance with some control law $u(t)$, whereas the other end at $x = L$ is free of load [11]. No external distributed forces are supposed.

Small vibrations of the elastic rod can be described by the linear equations

$$\{t, x\} \in \Omega : \quad p = \rho(x) w_t \quad \text{and} \quad s = \kappa(x) w_x, \tag{1}$$

$$\{t, x\} \in \Omega : \quad p_t = s_x \tag{2}$$

with the initial and boundary conditions

$$\begin{gathered} t = 0 : \quad p = p_0(x) \quad \text{and} \quad w = w_0(x), \\ x = 0 : \quad w = w_0(0) + u(t) \quad \text{with} \quad u(0) = 0, \quad x = L : \quad s = 0. \end{gathered} \tag{3}$$

Here, $\Omega = (0, T) \times (0, L)$ is the time-space domain, $T$ is the time instant, $\rho$ denotes the function of rod linear density, and $\kappa$ is its distributed stiffness. The linear momentum density $p(t, x)$, the normal stresses in the cross section $s(t, x)$, and the displacements $w(t, x)$ are unknown functions. Some initial momentum density $p_0(x)$ and displacements $w_0(x)$ are given.

The choice of the example is stipulated by its practical relevance and possible extensions. The equations of elastic rod motions (1)–(3) describe also a wide class of dynamic systems with distributed parameters, starting with the classical spring model, including the elastic shaft torsion, and so on. Although the rod considered in the paper has internal parameters uniformly distributed along its length, the proposed algorithm can be easily generalized onto the non-uniform case. Nevertheless, the hyperbolic system with constant geometrical and mechanical parameters can serve itself a useful application for numerical verification of the control algorithm.

## 3 Variational Statement of the Direct Dynamic Problem

The solution $p^*(t,x)$, $s^*(t,x)$, $w^*(t,x)$ of the initial-boundary value problem (1)–(3) may not exist in the classical sense (this depends on regularity of the functions $\kappa(x)$, $\rho(x)$, $p_0(x)$, $w_0(x)$, $u(t)$). To generalize the problem, we consider an integral statement of the constitutive laws proposed in [11] instead of the local formulation (1).

Let us introduce two auxiliary constitutive functions needed to relate the momentum density and velocities as well as the normal stresses and strains along the elastic rod in accordance with (1):

$$\eta(t,x) = w_t - \frac{p}{\rho(x)}, \quad \xi(t,x) = w_x - \frac{s}{\kappa(x)}. \tag{4}$$

On the solution these functions must be equal to zero.

For the direct problem of elastic rod motions, the generalized statement can be formulated as follows: Find the functions $p^*(t,x)$, $s^*(t,x)$, $w^*(t,x)$ such that the integral equation

$$\Phi[p,s,w] = \int_\Omega \varphi \, d\Omega = 0 \quad \text{with} \quad \varphi = \frac{1}{2}\left(\rho(x)\eta^2 + \kappa(x)\xi^2\right) \tag{5}$$

holds as well as the constraints (2) and (3). Here, $\Phi$ is the constitutive functional in the energy norm with $\varphi$ as the function of quadratic residual for the constitutive equations (1).

It is worth noting that the integrand $\varphi$ defined in (5) has a dimension of linear energy density and nonnegative. This fact directly follows from properties of $\varphi$, which imply that the functional $\Phi$ is also nonnegative. This allows us to reduce the integro-differential problem (2), (3), (5) to a variational one: Find those functions $p^*(t,x)$, $s^*(t,x)$, $w^*(t,x)$ that minimize the functional

$$\Phi[p^*,s^*,w^*] = \min_{p,s,w} \Phi[p,s,w] = 0 \tag{6}$$

subject to the constraints (2) and (3).

Denote the actual and arbitrarily chosen admissible momentum, stress, displacement fields via $p^*$, $s^*$, $w^*$ and $p$, $s$, $w$, respectively. Define

$$p = p^* + \delta p, \quad s = s^* + \delta s, \quad w = w^* + \delta w,$$

where $\delta p$, $\delta s$, $\delta w$ are the respective variations of momentum density, stresses, and displacements. Then,

$$\Phi[p,s,w] = \Phi[p^*,s^*,w^*] + \delta\Phi + \delta^2\Phi.$$

Here, $\Phi[p^*, s^*, w^*] = 0$ in accordance with (5). The first variation of the functional can be presented as the sum $\delta\Phi = \delta_p\Phi + \delta_s\Phi + \delta_w\Phi$ of the variations with respect to the unknowns $p, s, w$. It follows from the quadratic structure of the functional $\Phi$ that the second variation $\delta^2\Phi = \Phi[\delta p, \delta s, \delta w]$ is also nonnegative.

Let us express explicitly the first variation of the functional $\Phi$ and, consequently, the system of Euler–Lagrange equations with the corresponding natural conditions for the variational problem (2), (3), (6). For this purpose, the relation between the momentum function $p$ and the stress function $s$ imposed by the differential equation (2) should be used together with the corresponding relation between their variations

$$\delta p_t = \delta s_x.$$

The necessary condition of stationarity is obtained after integration by parts of the relation for $\delta\Phi$ and taking into account the problem constraints (2) and (3),

$$\delta_p\Phi + \delta_s\Phi + \delta_w\Phi = 0, \tag{7}$$

$$\delta_p\Phi = -\int_\Omega \eta\delta p \, d\Omega, \quad \delta_s\Phi = -\int_\Omega \xi\delta s \, d\Omega,$$

$$\delta_w\Phi = -\int_\Omega (\rho(x)\eta_t + (\kappa(x)\xi)_x)\,\delta w \, d\Omega + \int_0^L [\rho(x)\eta\delta w]_{t=T}\, dx + \int_0^T [\kappa(x)\xi\delta w]_{x=L}\, dt.$$

From (7), we see that $\delta\Phi = 0$ over all admissible variations $\delta p, \delta s, \delta w$ if the equalities (1) hold.

Introduce an auxiliary function

$$\zeta(t, x) = -\int_0^t \eta(\tau, x) d\tau$$

and get the expression for the first variations with respect to $p$ and $s$ after some equivalent transformations as follows

$$\delta_p\Phi + \delta_s\Phi = \int_\Omega \zeta_t\delta p \, d\Omega - \int_\Omega \xi\delta s \, d\Omega = \int_\Omega (\zeta_x - \xi)\,\delta s \, d\Omega + \int_0^L [\zeta\delta p]_{t=T}\, dx + \int_0^T [\zeta\delta s]_{x=0}\, dt. \tag{8}$$

By using (7) and (8), it is possible to derive the Euler-Lagrange system with the corresponding boundary and terminal conditions

$$\begin{gathered}\rho(x)\zeta_{tt} - (\kappa(x)\zeta_x)_x = 0, \quad \xi = \zeta_x; \\ \zeta|_{x=0} = \zeta_x|_{x=L} = \zeta|_{t=T} = \zeta_t|_{t=T} = 0.\end{gathered} \tag{9}$$

This homogeneous system is a terminal-boundary value problem with respect to the variable $\zeta(t, x)$. It can be shown that there is only a trivial solution of this problem and hence $\xi = 0$ and $\eta = 0$. In other words, if the solution $p^*$, $s^*$, $w^*$ of the problem (1)–(3) exists in the classical sense then the system of necessary conditions (9) together with the essential constraints (2) and (3) is equivalent to the original problem of elastic rod motion (1)–(3). This means that the statement (2), (3), (5) is given correctly in terms of the calculus of variations.

## 4 Finite Element Technique Based on the Ritz Method

The system (1)–(3) is solved by variational approach, which is a modification of the Ritz method based on the MIDR discussed in [12]. The law of momentum balance (2) holds automatically if two auxiliary functions (kinematic $\tilde{w}(t, x)$ and dynamic $\tilde{r}(t, x)$) are introduced such that

$$p = \tilde{r}_x(t, x) + p_0(x), \quad s = \tilde{r}_t(t, x), \quad w = \tilde{w}(t, x) + w_0(x). \tag{10}$$

The initial and boundary conditions for the new variables $\tilde{w}$ and $\tilde{r}$ are defined as follows:

$$t = 0: \quad \tilde{w} = 0 \quad \text{and} \quad \tilde{r} = 0, \quad x = 0: \quad \tilde{w} = u(t), \quad x = L: \quad \tilde{r} = 0. \tag{11}$$

Let us restate the initial-boundary value problem (1)–(3) in the variational form. Find the functions $\tilde{w}^*(t, x)$ and $\tilde{r}^*(t, x)$ subject to the constraints (11) and such that

$$\Phi[\tilde{w}^*, \tilde{r}^*] = \min_{\tilde{w}, \tilde{r}} \Phi[\tilde{w}, \tilde{r}], \quad \Phi = \frac{1}{2}\int\limits_{\Omega} \left(\rho(x)\eta^2 + \kappa(x)\xi^2\right) d\Omega, \tag{12}$$

$$\eta = \tilde{w}_t - \frac{\tilde{r}_x + p_0(x)}{\rho(x)}, \quad \xi = \tilde{w}_x + w_0'(x) - \frac{\tilde{r}_t}{\kappa(x)},$$

Here, $\eta$ and $\xi$ are the constitutive functions (4) expressed through the new independent variable $\tilde{w}$ and $\tilde{r}$.

To solve the minimization problem (11)–(12), we use piecewise polynomial approximations with respect to the time and space. For the triangulation of the domain

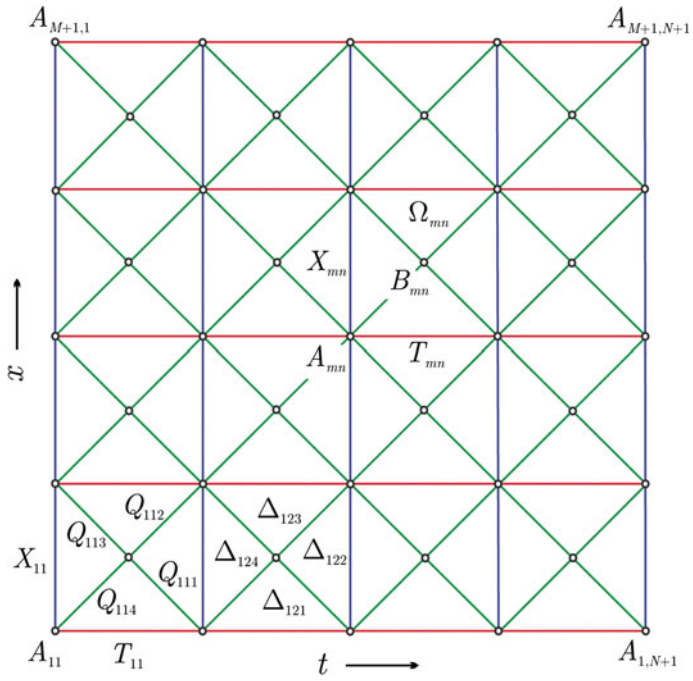



**Fig. 1** Triangulation of the time-space domain $\Omega$

$\Omega$ shown in Fig. 1, these approximations are given by the relations

$$\begin{aligned}
\tilde{w} \in \mathscr{S}_w^h &= \left\{ \begin{array}{ll} \tilde{w}(t,x):\ \tilde{w} = \displaystyle\sum_{k+l=0}^{K} w_{jmn}^{(kl)} t^k x^l, & \{t,x\} \in \Delta_{jmn}, \\ j = 1,\ldots,4, \quad m = 1,\ldots,M, & n = 1,\ldots,N \end{array} \right\} \cap C^0. \\
\tilde{r} \in \mathscr{S}_r^h &= \left\{ \begin{array}{ll} \tilde{r}(t,x):\ \tilde{r} = \displaystyle\sum_{k+l=0}^{K} r_{jmn}^{(kl)} t^k x^l, & \{t,x\} \in \Delta_{jmn}, \\ j = 1,\ldots,4, \quad m = 1,\ldots,M, & n = 1,\ldots,N \end{array} \right\} \cap C^0.
\end{aligned} \tag{13}$$

Here, $\Delta_{jmn}$ denotes the corresponding subdomain of a triangular mesh described in Fig. 1.

The mesh is defined by the nodes on the axes $t$ and $x$ as follows:

$$\begin{aligned}
&x_m > x_{m-1}, \qquad m = 1,\ldots,M+1, \qquad x_1 = 0, \qquad x_{M+1} = 1; \\
&t_n > t_{n-1}, \qquad n = 1,\ldots,N+1, \qquad t_1 = 0, \qquad t_{N+1} = T.
\end{aligned}$$

The domain $\Omega$ is divided by the straight lines $x = x_m$ and $t = t_n$ (see Fig. 1) into $MN$ rectangles $\Omega_{mn} = (t_n, t_{n+1}) \times (x_m, x_{m+1})$ with $m = 1, \ldots, M$ and $n = 1, \ldots, N$.

The rectangle vertices $\{t_n, x_m\}$ are denoted by $A_{mn}$ with the corresponding edges $A_{kl}A_{mn}$ between $A_{kl}$ and $A_{mn}$. For brevity, let $T_{mn} = A_{mn}A_{m,n+1}$ and $L_{mn} = A_{mn}A_{m+1,n}$.

The diagonals of the rectangle $\Omega_{mn}$ (see Fig. 2) divide it in turn into four triangles

$$\begin{aligned} \Delta_{mn,1} &= B_{mn}A_{mn}A_{m,n+1}, & \Delta_{mn,2} &= B_{mn}A_{m,n+1}A_{m+1,n+1}, \\ \Delta_{mn,3} &= B_{mn}A_{m+1,n+1}A_{m+1,n}, & \Delta_{mn,4} &= B_{mn}A_{m+1,n}A_{mn}. \end{aligned} \tag{14}$$

Here, $B_{mn}$ is the intersection point of the diagonals $A_{mn}A_{m+1,n+1}$ and $A_{m,n+1}A_{m+1,n}$. Let us introduce the notation for the inclined edges of the triangle (14). We denote

$$\begin{aligned} Q_{mn,1} &= B_{mn}A_{m,n+1}, & Q_{mn,2} &= B_{mn}A_{m+1,n+1}, \\ Q_{mn,3} &= B_{mn}A_{m+1,n}, & Q_{mn,4} &= B_{mn}A_{mn}. \end{aligned}$$

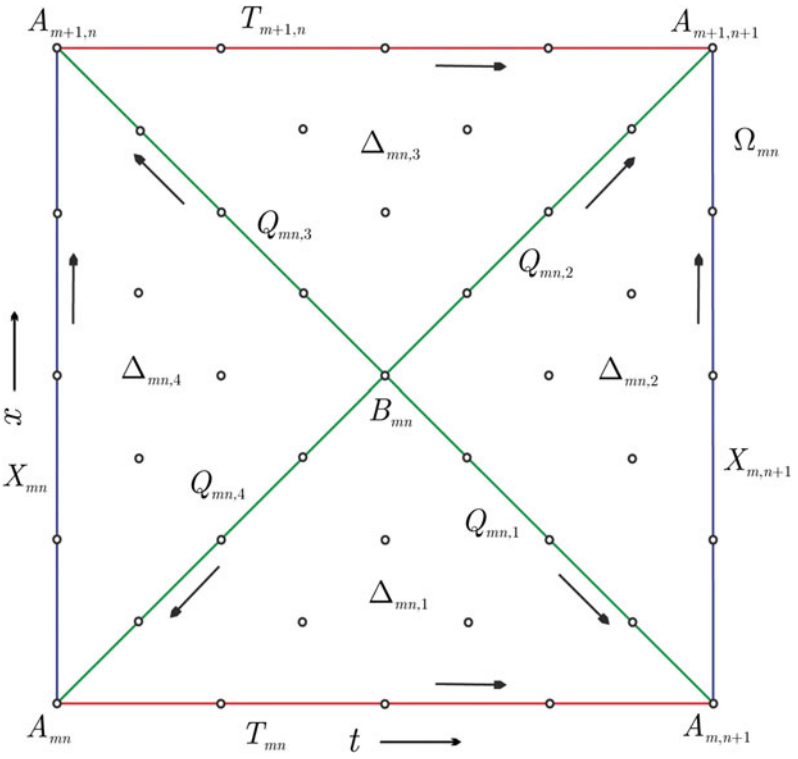



**Fig. 2** Mesh structure on the rectangle $\Omega_{mn}$

The unknown functions $\tilde{w}(t,x)$ and $\tilde{r}(t,x)$ are approximated in each of $4MN$ triangles $\Delta_{nmj}$ by a complete bivariate polynomial of the order $K$ in the Bézier-Bernstein form [8]. In accordance with this form, the variables $\tilde{w}$ and $\tilde{r}$ in any triangle $\Delta \subset \Omega$ with the vertices $A_i = \{t_i, x_i\} \in \overline{\Omega}$, $i = 1,2,3$ (a local vertex indexing is given bypassing the triangle contour counterclockwise) are expressed by the relations

$$w(t,x) = \sum_{k+l=0}^{K} w_{kl} B_{kl}^K(t,x), \quad r(t,x) = \sum_{k+l=0}^{K} r_{kl} B_{kl}^K(t,x),$$
$$B_{kl}^K(t,x) = \frac{K!}{k!l!(K-k-l)!} b_1^k(t,x) b_2^l(t,x) b_3^{K-k-l}(t,x), \tag{15}$$

Here, the linear functions

$$b_1(t,x) = d^{-1}(x_2 - x_3)(t - t_3) - d^{-1}(t_2 - t_3)(x - x_3),$$
$$b_2(t,x) = d^{-1}(x_3 - x_1)(t - t_1) - d^{-1}(t_3 - t_1)(x - x_1),$$
$$b_3(t,x) = d^{-1}(x_1 - x_2)(t - t_2) - d^{-1}(t_1 - t_2)(x - x_2),$$

are introduced and

$$d = \det \mathbf{T}, \quad \mathbf{T} = \begin{bmatrix} t_1 & t_2 & t_3 \\ x_1 & x_2 & x_3 \\ 1 & 1 & 1 \end{bmatrix},$$

where $\mathbf{T}$ is the extended coordinate matrix, which determinant $d$ equals to the doubled area of the triangle $\Delta$. The functions $b_i$, the so-called barycentric coordinates, have the following properties:

$$b_i(t_i, x_i) = 1, \quad b_i(t_j, x_j) = 0, \quad i \neq j, \ i, j = 1,2,3, \quad b_1 + b_2 + b_3 = 1.$$

According to the Eq. (15), for a chosen piecewise polynomial, the total number of parameters $w_{kl}^{(mnj)}$ and $r_{kl}^{(mnj)}$ in the mesh element $\Delta_{mnj}$ is equal to $N_\Delta = (K+1)(K+2)/2$. These degrees of freedom can be symbolically marked by circles as it is shown in Fig. 2 for $K = 4$.

The vector $\hat{\mathbf{z}} = \{\hat{z}_i\} \in \mathbb{R}^{N_l}$ consisting of all such local parameters has the dimension $N_l = 8MNN_\Delta$. The sequence of the vector components $\hat{z}_i$ can be chosen so that

$$\hat{z}_{i_1} = w_{kl}^{(mnj)}, \quad \hat{z}_{i_2} = r_{kl}^{(mnj)}, \quad i_1 = j_0 + k_0, \quad i_2 = N_\Delta + j_0 + k_0,$$
$$j_0 = 4\,(2\,((m-1)N + n - 1) + j - 1)\,N_\Delta, \quad k_0 = \frac{k(2K - k + 3)}{2} + l + 1,$$
$$m = 1, \ldots, M, \quad n = 1, \ldots, N, \quad j = 1,2,3,4,$$
$$k = 0, \ldots, K, \quad l = 0, \ldots, K - k. \tag{16}$$

Here, $j_0$ is the index of the last coefficient for the previous triangle, $k_0$ defines one-dimensional indexing of the Bézier-Bernstein coefficients $w_{kl}^{(mnj)}$ related with $\Delta_{mnj}$.

Let us define the vector of discontinuous basis functions

$$\mathbf{a}(t,x) = \{a_i(t,x)\} \in \mathbb{R}^{N_l},$$

which corresponds to $\hat{\mathbf{z}}$ in accordance with the relation

$$a_{i_1} = a_{i_2} = \begin{cases} B_{kl}^{(mnj)}(t,x), & \{t,x\} \in \Delta_{mnj}, \\ 0, & \{t,x\} \notin \Delta_{mnj}. \end{cases}$$

The Bézier-Bernstein polynomial $B_{kl}^{(mnj)}$ of the order $K$ is introduced in the triangle $\Delta_{mnj}$ by relations similar to (15). The corresponding indices $i_1$ and $i_2$ are given in (16). In this case, the approximations of dynamic and kinematic fields can be presented as follows:

$$\hat{u} = \hat{\mathbf{w}}^{\mathrm{T}}(t,x)\hat{\mathbf{z}}, \quad \hat{r} = \hat{\mathbf{r}}^{\mathrm{T}}(t,x)\hat{\mathbf{z}}, \quad \hat{\mathbf{w}} = \mathbf{E}_w\mathbf{a}(t,x), \quad \hat{\mathbf{r}} = \mathbf{E}_r\mathbf{a}(t,x). \tag{17}$$

Here,

$$\mathbf{E}_w = \begin{bmatrix} \mathbf{E}_w^0 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_w^0 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{E}_w^0 \end{bmatrix} \in \mathbb{R}^{N_l \times N_l}, \qquad \text{where } \mathbf{E}_w^0 = \begin{bmatrix} \mathbf{I}_{N_\Delta} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

$$\mathbf{E}_r = \begin{bmatrix} \mathbf{E}_r^0 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_r^0 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{E}_r^0 \end{bmatrix} \in \mathbb{R}^{N_l \times N_l}, \qquad \text{where } \mathbf{E}_r^0 = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{N_\Delta} \end{bmatrix},$$

and $\mathbf{I}_n$ denotes the identity matrices of the dimension $n \times n$.

## 5 Optimal Control Problem

Now we consider an inverse dynamic problem for the elastic rod model discussed above. In accordance with the variational formulation of the initial-boundary value problem (11)–(12), minimum of the constitutive functional $\Phi[\tilde{r}, \tilde{w}]$ is first sought for any sufficiently smooth function $u(t) \in \mathscr{U}$. The control problem is to find such

a function $u^*(t) \in \mathscr{U}$ that moves the elastic rod at the fixed time instant $T$ to the finale state

$$t = T: \quad p = 0, \quad w = w_T = \text{const}, \quad u = u_T = w_T - w_0(0) \tag{18}$$

and minimizes the mean energy $\bar{E}$ of the rod

$$J[u^*] = \min_{u \in \mathscr{U}} J[u], \quad J = \bar{E}. \tag{19}$$

Here

$$\bar{E} = \frac{\Psi}{T}, \quad \Psi = \int_0^T E\, dt, \quad E = \int_0^L \psi\, dx, \tag{20}$$

$$\psi = \frac{1}{2}\left(\rho^{-1}(x)\,(\tilde{r}_x + p_0(x))^2 + \kappa(x)\left(\tilde{w}_x + w_0'(x)\right)^2\right). \tag{21}$$

Here, $E$ is the total mechanical energy of the moving elastic structure with its linear density $\psi$ expressed through the variables $\tilde{r}(t, x, u)$ and $\tilde{w}(t, x, u)$.

## 6 Discretization and Regularization of the Control Problem

According to (11) and (13), the admissible control function $u(t) = \tilde{w}(t, 0)$ taken in numerical realization has to be piecewise polynomial. Let $\hat{\mathbf{u}} = [u_1, \ldots, u_{KN}]^{\mathrm{T}} \in \mathbb{R}^{KN}$ be the vector of control parameters. In this case, the $KN$ components of the vector $\hat{\mathbf{u}}$ are used to meet $2KM + 2$ terminal conditions in the optimal control problem (18)–(20). Even if the terminal values (18) are admissible for the splines $\tilde{r}(t, x) \in \mathscr{S}_r^h$ and $\tilde{w}(t, x) \in \mathscr{S}_w^h$, the momentum density and displacements resulting from (10) with the approximations (13) for terminal constraints more general than the piecewise polynomial ones cannot be apparently satisfied.

The terminal conditions (18) can be weakened by introducing some tolerance $\varepsilon_1 > 0$. For example, the total energy of the rod at the end of the controlled process can be constrained by some small value

$$E_1 = E(T) \le \varepsilon_1 \ll \bar{E}.$$

As it has been shown in numerical calculations, the accuracy of approximate solutions may dramatically fall down through the optimization of the control input $u(t)$. To regulate the error level and ensure the reliability of modelling, an upper limit of the error functional $\Phi$ should be given

$$E_2 = T^{-1}\Phi \le \varepsilon_2 \ll E_1.$$

Such a tolerance can be guaranteed by two isoperimetric conditions imposed on the energy functionals

$$E_1 = \varepsilon_1 \quad \text{and} \quad E_2 = \varepsilon_2. \tag{22}$$

After parametric optimization of the functional $\Phi$ according to (11)–(13) for a arbitrary vector $\hat{\mathbf{u}}$, the problem (18)–(20) with the integral conditions (22) is equivalent to the following minimization: Find the control vector $\hat{\mathbf{u}}^*$ that moves the rod end at $x = 0$ in the fixed time $T$ to the final position $w_T$ and minimizes the energy functional

$$J(\hat{\mathbf{u}}^*) = \min_{\hat{\mathbf{u}}} J(\hat{\mathbf{u}}), \quad w(T, 0) = w_T; \quad J = \bar{E} + \gamma_1 E_1 + \gamma_2 E_2, \quad \gamma_{1,2} \geq 0. \tag{23}$$

Here, $\bar{E}$ is the mean energy of the rod, $E_1$ is the terminal energy of the system, $E_2$ is the integral error of approximate solution, $\gamma_1$ and $\gamma_2$ are the weighting factors introduced to achieve the given values of $E_1$ and $E_2$ in accordance with (22), $\psi$ is the rod energy density.

The optimal control vector $\hat{\mathbf{u}}^*$ as well as the corresponding function $u^*(t) = u(t, \hat{\mathbf{u}}^*)$, the approximation of displacements $w^*(t, x) = w(t, x, \hat{\mathbf{u}}^*)$, momentum density $p^*(t, x) = p(t, x, \hat{\mathbf{u}}^*)$, and normal stresses $s^*(t, x) = s(t, x, \hat{\mathbf{u}}^*)$ are found in accordance with the algorithm described below.

## 7 Numerical Algorithm of Control Optimization

As it is seen in (13), the unknown functions $\tilde{r}(t, x)$ and $\tilde{w}(t, x)$ on the triangle $\Delta_{mnj}$ of the time-space mesh are defined by the parameters $r_{kl}^{(mnj)}$ and $w_{kl}^{(mnj)}$, which number is equal to $2N_\Delta = (K + 1)(K + 2)$. The local parameters have been collected into a vector $\hat{\mathbf{z}} \in \mathbb{R}^{N_l}$ with the dimension $N_l = 8MNN_\Delta$ in accordance with (16) and the approximations (13) can be presented as in (17).

By satisfying the continuous conditions imposed on the fields $\tilde{r}(t, x)$ and $\tilde{w}(t, x)$, the matrix $\mathbf{Q} \in \mathbb{R}^{N_l \times N_g}$ is derived. It relates the global and local parameter vectors according to the relation $\hat{\mathbf{z}} = \mathbf{Q}\mathbf{z}$. The resulting continuous fields are expressed in the vector form as follows:

$$\tilde{r}(t, x, z) = \mathbf{r}^{\mathrm{T}}(t, x)\mathbf{z}, \quad \tilde{w}(t, x, z) = \mathbf{w}^{\mathrm{T}}(t, x)\mathbf{z}. \tag{24}$$

For the optimal control problem, the vector of global parameters can be presented by the relations

$$\mathbf{z} = \left[\mathbf{y}^{\mathrm{T}}\ \mathbf{u}^{\mathrm{T}}\ \mathbf{q}^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{R}^{N_g}, \quad \mathbf{y} \in \mathbb{R}^{N_y}, \quad \mathbf{u} \in \mathbb{R}^{N_u}, \quad \mathbf{q} \in \mathbb{R}^{N_q},$$
$$N_g = N_y + N_u + N_q, \quad N_y = 4KMN, \quad N_u = KN - 1.$$

Here, $\mathbf{y}$ is the vector of designed parameters, $\mathbf{u}$ denotes the vector of control parameters that remain after satisfying the terminal displacement condition in (23), $\mathbf{q}$ is the vector of system parameters that depends only on the terminal value $u_T$. It is always possible to reduce the problem to the case $u_T = 1$ by scaling and to eliminate the vector $\mathbf{q}$ from consideration.

After substituting the approximation $\tilde{r}(t, x, \mathbf{z})$ and $\tilde{w}(t, x, \mathbf{z})$ from (13) into the functional of (12) and integrating over the domain $\Omega$, we obtain

$$\tilde{\Phi}(\mathbf{z}) = \Phi[\tilde{u}, \tilde{r}] = \frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{F}\mathbf{z} + \mathbf{f}^{\mathrm{T}}\mathbf{z} + f.$$

By taking into account the structure of the vector $\mathbf{z}$ and quadratic form of the functional $\Phi$, the matrix $\mathbf{F} \in \mathbb{R}^{N_g \times N_g}$ and the vector $\mathbf{f} \in \mathbb{R}^{N_g}$ are defined in the form

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_{yy} & \mathbf{F}_{yu} & \mathbf{0} \\ \mathbf{F}_{yu}^{\mathrm{T}} & \mathbf{F}_{uu} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} \mathbf{f}_y \\ \mathbf{f}_u \\ \mathbf{0} \end{bmatrix}, \quad \begin{cases} \mathbf{F}_{yy} = \mathbf{F}_{yy}^{\mathrm{T}} \in \mathbb{R}^{N_y \times N_y} & \mathbf{F}_{uu} = \mathbf{F}_{uu}^{\mathrm{T}} \in \mathbb{R}^{N_u \times N_u}, \\ \mathbf{f}_y \in \mathbb{R}^{N_y}, & \mathbf{f}_u \in \mathbb{R}^{N_u} \quad f \in \mathbb{R}. \end{cases}$$

Minimum of the function $\tilde{\Phi}$ is attained if

$$\mathbf{y} = \tilde{\mathbf{y}} = -\mathbf{F}_{yy}^{-1}\left(\mathbf{F}_{yu}\mathbf{u} + \mathbf{f}_y\right).$$

Similarly, the control functional $\hat{J}(\mathbf{z}) = J[\tilde{u}, \tilde{r}]$ in (23) is quadratic with respect to the vector $\mathbf{z}$ and can be represented in the form

$$\hat{J}(\mathbf{y}, \mathbf{u}) = \frac{1}{2}\begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{J}_{yy} & \mathbf{J}_{yu} \\ \mathbf{J}_{yu}^{\mathrm{T}} & \mathbf{J}_{uu} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix} + \begin{bmatrix} \mathbf{j}_y \\ \mathbf{j}_u \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix} + J_0,$$

where

$$\mathbf{J}_{yy} = \mathbf{J}_{yy}^{\mathrm{T}} \in \mathbb{R}^{N_y \times N_y}, \quad \mathbf{J}_{uu} = \mathbf{J}_{uu}^{\mathrm{T}} \in \mathbb{R}^{N_u \times N_u}, \quad \mathbf{j}_y \in \mathbb{R}^{N_y}, \quad \mathbf{j}_u \in \mathbb{R}^{N_u}, \quad J_0 \in \mathbb{R}.$$

After that, the vector of design parameter $\tilde{\mathbf{y}}$ is substituted into the cost function $\hat{J}(\mathbf{y}, \mathbf{u})$ and we obtain

$$\tilde{J}(\mathbf{u}) = \hat{J}\left(\tilde{\mathbf{y}}(\mathbf{u}), \mathbf{u}\right) = \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{G}\mathbf{u} + \mathbf{g}^{\mathrm{T}}\mathbf{u} + G, \quad \mathbf{G} = \mathbf{G}^{\mathrm{T}}.$$

As a result, the original control problem is reduced to the unconstrained minimization for the function $\tilde{J}(\mathbf{u})$. The optimal control vector is found as $\mathbf{u}^* = -\mathbf{G}^{-1}\mathbf{g}$ and the design parameter vector as $\mathbf{y}^* = \tilde{\mathbf{y}}(\mathbf{u}^*)$. By changing the vector $\mathbf{z}$ for the optimal vector $\mathbf{z}^* = \left[(\mathbf{y}^*)^{\mathrm{T}}\ (\mathbf{u}^*)^{\mathrm{T}}\ \mathbf{q}\right]^{\mathrm{T}}$ in (24) and taking into account (10), approximations of the momentum density, stress and displacement fields are obtained as $\tilde{p}^* = \tilde{p}(t, x, \mathbf{z}^*)$, $\tilde{s}^* = \tilde{s}(t, x, \mathbf{z}^*)$, $\tilde{w}^* = \tilde{w}(t, x, \mathbf{z}^*)$.

The relative energy error $\Delta$ of the approximate solution is given by the relation

$$\Delta = \tilde{\Phi}(\mathbf{z}^*)\tilde{\Psi}^{-1}(\mathbf{z}^*), \quad \tilde{\Psi}(\mathbf{z}) = \Psi[\tilde{r}(t,x,\mathbf{z}), \tilde{w}(t,x,\mathbf{z})],$$

where $\Psi = T\bar{E}$ is the energy integral over the time interval $[0, T]$ defined in (20).

## 8 Simulation and Solution Quality Estimates

We choose the dimensionless parameters of the system $\rho = \kappa = L = 1$, the initial functions $p_0(x) = w_0(x) = 0$, and the control parameters $T = 4, w_T = 1, \gamma_1 = 10^4$, $\gamma_2 = 10^{-4}$. The algebraic order of the approximating system is $N_y = 4MNK^2$. For the test control function

$$u = u_0(t) = 3t^2T^{-2} - 2t^3T^{-3}, \tag{25}$$

the relative integral error $\Delta = E_2\bar{E}^{-1}$ versus the dimension $N_y$ is presented in Fig. 3.

The so-called $h$-convergence is depicted by solid lines for homogeneous meshes ($M = N = 1 \div 7$) and different polynomial orders (from $K = 3$ to $K = 6$). The rate of $p$-convergence when the polynomial degree is varied ($K = 3 \div 7$) is given by a dashed line for the fixed triangulation with $M = N = 1$. We see that the accuracy of numerical solutions grows up fast if the dimension increases.

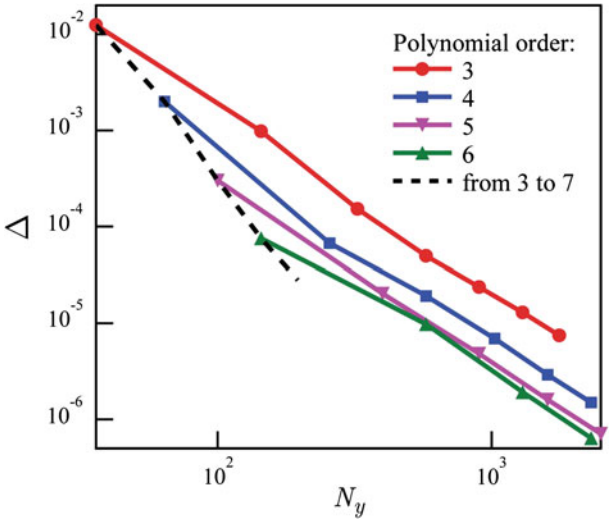



**Fig. 3** Relative error $\Delta$ versus the approximation dimension $N_y$

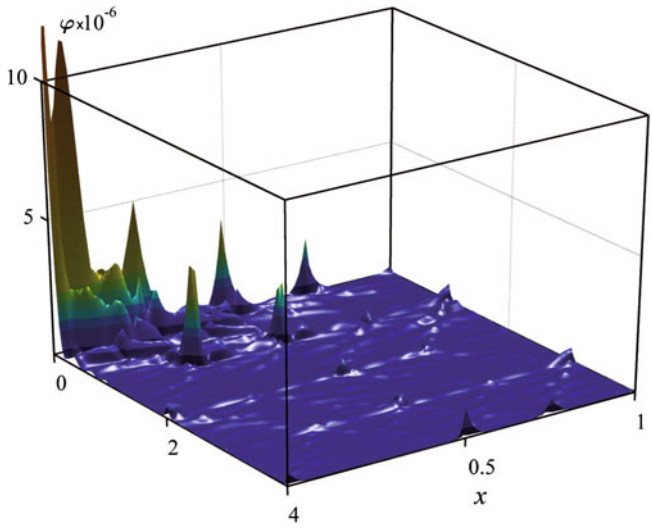



**Fig. 4** Local error distribution $\varphi(t, x)$

Local values of the solution error can be defined by the function $\varphi(t, x)$. The time-space error distribution is depicted in Fig. 4 for the following approximation parameters: $K = 6$ and $M = N = 4$. The relative integral error for the mesh is equal to $\Delta = 6.4 \times 10^{-7}$. The mean mechanical energy over the process equals to $\bar{E} = 0.0984$.

It can be verified that there exists a piecewise polynomial solution for this specific test control parameters. Moreover, the polynomials are defined on those triangular subdomains of the time-space domain $\Omega$ which are bounded by the characteristic lines

$$x - t = 0, \quad x - t = 2, \quad x + t = 2, \quad x + t = 4.$$

The order of the polynomials is equal to 3 and $p(t, x) = s(t, x) = w(t, x) \equiv 0$ if $t \leq x$.

It turns out for the mesh topology under consideration that some of the triangle edges coincide with these characteristic lines if $N = 4M$. In this case, the exact solution can be found (up to the round-off error) by using the finite-element approximations of the unknown functions $\tilde{r}(t, x)$ and $\tilde{w}(t, x)$ with the polynomial order $K \geq 3$. Such a superconvergence property is exploited below to obtained the momentum, stress and displacement fields. The relative displacements of the elastic rod $w(t, x) - u(t)$ with the control input $u = u_0(t)$ are shown in Fig. 5 for $K = 3$, $M = 1$, and $N = 4$.

The distributions of the momentum density $p(t, x)$ and the normal stresses $p(t, x)$ for the same test control $u_0(t)$ are depicted in Figs. 6 and 7, respectively. Here, the nonanaliticity along the characteristic lines can be seen more distinctively. It is certainly difficult to approximate rather accurately place where the function breaks

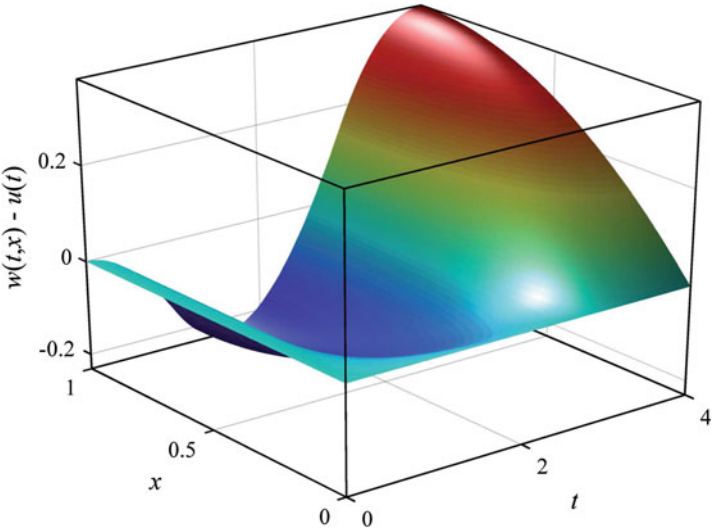



**Fig. 5** Relative rod displacements $w(t, x) - u^0(t)$ for the test motion

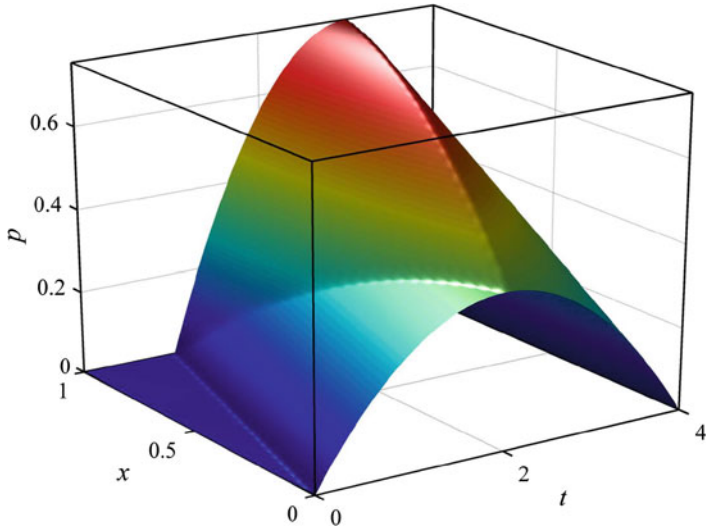



**Fig. 6** Momentum density $p(t, x)$ for the test motion

if such a line is located inside a mesh element. These breaks cause error surges for unappropriate meshes as it can be seen in Fig. 4.

The optimal control as a piecewise polynomial function has been found for the given parameters $K = 3$, $M = 1$, and $N = 4$ ($N_y = 144$, $N_u = 11$). In Fig. 8, the optimal control displacement of the rod end $u^*(t)$ (dash-dot curve) is compared with

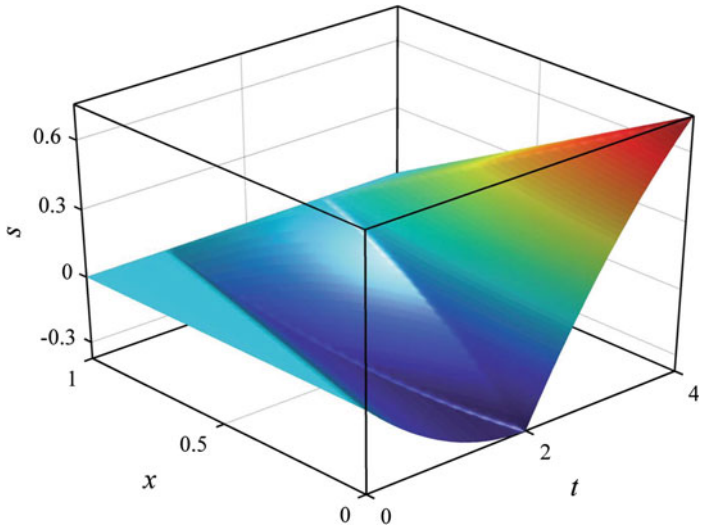



**Fig. 7** Stress distribution $s(t, x)$ for the test motion

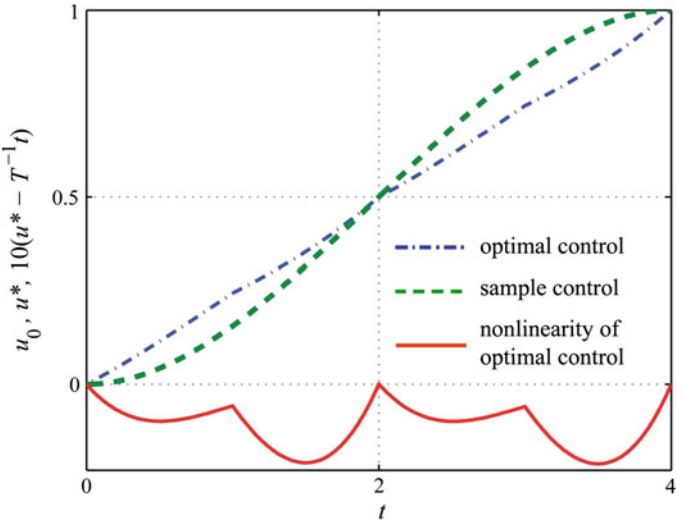



**Fig. 8** Test control $u_0(t)$ versus optimal control $u^*(t)$

the test control $u_0(t)$ considered in (25) (dashed curve). The optimal input is near to linear one, but the moderate deviation from the uniform motion $u^*(t) - T^{-1}t$, which is traced in this figure by solid curve with the scaling factor of 10, influences sufficiently on the whole elastic deformations of the rod.

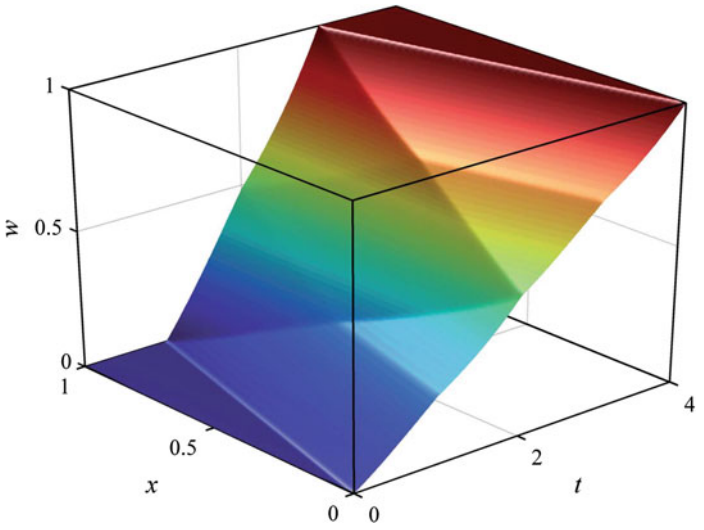



**Fig. 9** Absolute rod displacements $w(t, x)$ for the optimal motion

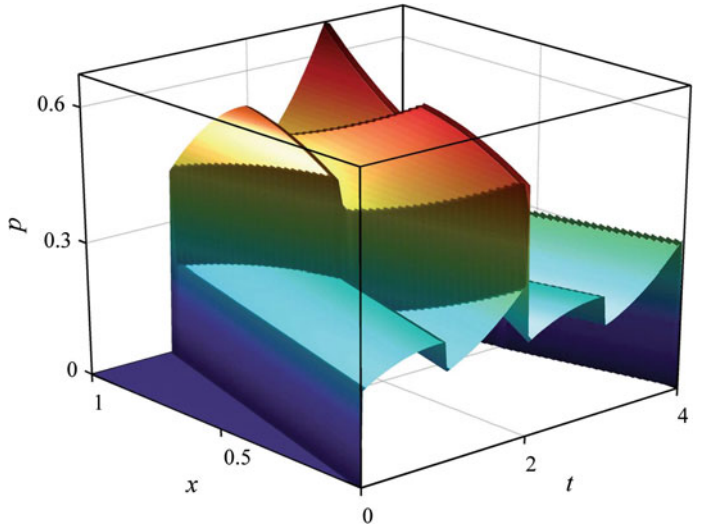



**Fig. 10** Momentum density $p(t, x)$ for the optimal motion

The optimal displacements of the rod points $\tilde{w}^*$ as a function of the time $t$ and coordinate $x$ are shown in Fig. 9. The optimal momentum $\tilde{p}^*(t, x)$ and stresses $\tilde{s}^*(t, x)$ are depicted in Figs. 10 and 11, respectively.

By using the obtained control law, a sufficiently low value of terminal energy $E_1 = 9 \times 10^{-11}$ is attained as compared with the average energy of the elastic rod $\bar{E} = 0.0636$. The relative error achieved for the optimal control does not exceed

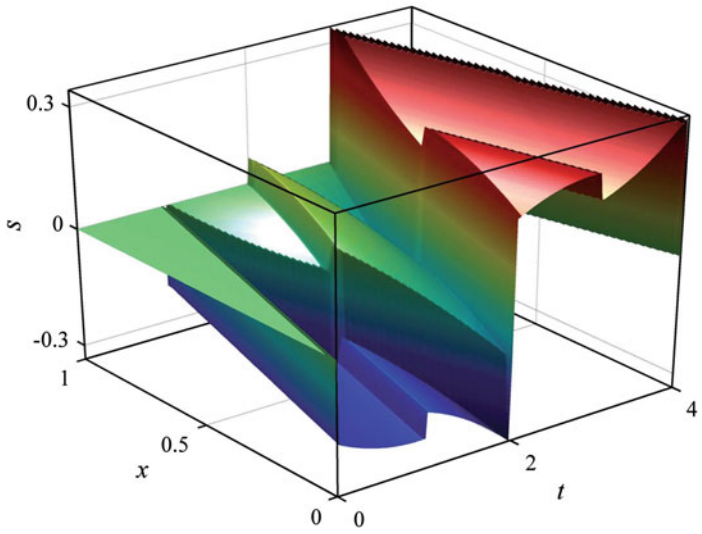



**Fig. 11** Stress distribution $s(t, x)$ for the optimal motion

$\Delta < 10^{-15}$. The weighting coefficients are chosen so that the following inequality holds: $E_2 \ll E_1 \ll \tilde{E}$.

It is worth noting that any significant vibrations of the rod are not excited during the control process. The corresponding changes in the dimension of the spline approximation (13) and, therefore, in control dimension does not cause any significant decreasing of the minimized mean energy for the control process as presented in Table 1.

**Table 1** Optimal energy values versus approximation and control dimensions

| Space intervals | Time intervals | Polynomial order | Control dimension | Mean energy |
|---|---|---|---|---|
| 1 | 4 | 3 | 11 | 0.0636 |
| 1 | 4 | 4 | 15 | 0.0634 |
| 1 | 4 | 5 | 19 | 0.0633 |
| 2 | 8 | 3 | 23 | 0.0634 |
| 2 | 8 | 4 | 31 | 0.0633 |
| 2 | 8 | 5 | 40 | 0.0632 |

## 9 Conclusions and Outlook

In this paper, a control algorithm for energy optimization in structural dynamics has been proposed and discussed. This control strategy is based on the MIDR, variational approach, and on finite element techniques. The verification of optimal control laws has been performed by taking into account the explicit local and integral error estimates.

In a subsequent research, we plan to apply the optimization algorithm proposed in this paper to more complex elastic systems with non-uniformly distributed parameters and to motions of 2D and 3D elastic bodies. Various mesh refinement and mesh adaptation approaches can be applied to increase the solution accuracy. Other dynamical models of solids, e.g., viscoelastic body and structures with geometrical and physical nonlinearity are to be considered from the viewpoint of the calculus of variation. Optimal problems with non-quadratic cost functions and control constraints and other inverse problems such as identification, measurements, etc. can also be considered as a great challenge for the method proposed.

**Acknowledgments** This work was supported by the Russian Foundation for Basic Research, project nos. 12-01-00789, 13-01-00108, 14-01-00282, the Leading Scientific Schools Grants NSh-2710.2014.1, NSh-2954.2014.1.

## References

1. N.U. Ahmed, K.L. Teo, *Optimal Control of Distributed Parameter Systems* (North-Holland, New York, 1981)
2. M.J. Balas, Finite-dimensional control of distributed parameter systems by Galerkin approximation of infinite-dimensional controllers. J. Math. Anal. Appl. **114**(1), 17–36 (1986)
3. S.P. Banks, *State-space and Frequency-domain Methods in the Control of Distributed Parameter Systems* (Peregrinus, London, 1983)
4. A.G. Butkovsky, *Distributed Control Systems* (Elsevier, New York, 1969)
5. F.L. Chernous'ko, I.M. Ananievski, S.A. Reshmin, *Control of Nonlinear Dynamical Systems: Methods and Applications* (Springer, Berlin, 1996)
6. P.D. Christofides, *Nonlinear and Robust Control of PDE Systems: Methods and Applications to Transport-reaction Processes* (Birkhäuser, Boston, 2001)
7. R. Curtain, H. Zwart, *An Introduction to Infinite-dimensional Linear Systems Theory* (Springer, New York, 1995)
8. G. Farin, *Curves and Surfaces for Computer-aided Geometric Design*, 4th edn. (Academic Press, San Diego, 1997)
9. M. Gerdts, G. Greif, H.J. Pesch, Numerical optimal control of the wave equation: optimal boundary control of a string to rest in finite time. Math. Comput. Simul. **79**(4), 1020–1032 (2008)
10. M. Gugat, Optimal nodal control of networked hyperbolic systems: evaluation of derivatives. Adv. Model. Optim. **7**(1), 9–37 (2005)
11. G.V. Kostin, Construction of an optimal control for the motion of elastic bodies using integrodifferential relations. J. Comput. Syst. Sci. Int. **46**(4), 532–542 (2007)
12. G.V. Kostin, V.V. Saurin, *Integrodifferential Relations in Linear Elasticity* (De Gruyter, Berlin, 2012)

13. W. Krabs, *Optimal Control of Undamped Linear Vibrations* (Heldermann, Lemgo, 1995)
14. J.E. Lagnese, G. Leugering, E.J.P.G. Schmidt, *Modeling, Analysis, and Control of Dynamic Elastic Multi-link Structures* (Birkhäuser, Boston, 1984)
15. D. Leineweber, E.I. Bauer, H. Bock, J. Schloeder, An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization. Part 1: theoretical aspects. Comput. Chem. Eng. **27**(2), 157–166 (2003)
16. G. Leugering, Domain decomposition of optimal control problems for dynamic networks of elastic strings. Comput. Optim. Appl. **16**(1), 5–27 (2000)
17. J.L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations* (Springer, New York, 1971)
18. J.L. Lions, Exact controllability, stabilization and perturbations for distributed systems. SIAM Rev. **30**(1), 1–68 (1988)
19. A. Zuyev, O. Sawodny, Stabilization and observability of a rotating Timoshenko beam model. Math. Probl. Eng. (Art. ID 57238), 19 (2007)

# Contact Optimization Problems for Stationary and Sliding Conditions

**István Páczelt, Attila Baksa and Zenon Mróz**

**Abstract** The contact stress distribution is frequently not regular. It may contain singularities reducing the lifetime of machine elements. In order to eliminate such stress singularities, the application of contact pressure control is recommended in the contact conditions. In the paper, several classes of optimization problems are formulated for stationary and sliding contacts. Further, they are illustrated by specific examples. The relation to wear process is made as a natural way to attain the steady state contact profile satisfying the optimality conditions corresponding to minimization of the wear dissipation rate. It is assumed that the displacements and strains are small and the materials of the contacting bodies are elastic.

**Keywords** Contact optimization problems · Steady wear state · Generalized wear dissipation power · Variational principles · $p$-version of the finite element method

**Mathematical Subject Classification:** 74A55 · 74F05 · 74P10 · 74S05

## 1 Introduction

A designer always tends to avoid singularities within the contact regions in order to keep stresses at a low level. This tendency leads to optimal design of contact surface shape and proper material selection, thus generating a class of contact optimization

I. Páczelt (✉) · A. Baksa
Institute of Applied Mechanics, University of Miskolc,
3515 Miskolc-egyetemváros, Hungary
e-mail: mechpacz@uni-miskolc.hu

A. Baksa
e-mail: mechab@uni-miskolc.hu

Z. Mróz
Institute of Fundamental Technological Research, A. Pawińskiego 5B,
02-106 Warsaw, Poland
e-mail: zmroz@ippt.pan.pl



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_16

problems. The design parameters in structural optimization are usually defined as material moduli, structure size and shape, characteristic dimensions, supports, loads, inner links, reinforcement and topology. Here, we refer to the books by Banichuk and Neittaanmäki [1] and Banichuk [2]. The sensitivity analysis for optimization of different kind of structures and loading conditions was developed in Mróz et al. [3–7]. In engineering practice the interaction at connections of machine elements is frequently modeled as unilateral contact problems. Haslinger and Neittaanmäki [8] dealt with the mathematical aspects of contact optimization problems.

Contact pressure distribution is sensitive to friction conditions, geometry, stiffness of contacting bodies and loading conditions. The peak contact pressure distribution and the boundary stress concentration can be reduced by using special mathematical (linear or quadratic) programming techniques and shape sensitivity analysis.

The contact pressure optimization was analyzed for an elastic punch on a rigid substrate assuming the linear elasticity relations, cf. [9–12]. In some earlier works [13–15], the maximum contact pressure was chosen as the objective function, but it was not differentiable. In the articles [8–10, 16], the total potential energy was considered as a cost function and the integral measure of the gap function was used as an isoparametric constraint.

A nearly constant contact pressure distribution was achieved in [14, 15] by an appropriate shape optimization for axially symmetric bodies, assuming that the change in radius has no effect on the stiffness and compliance matrices. Our works [17–20] provide a new type solutions for 2D and 3D problems, in which the contact pressure distribution is partially controlled by minimizing the maximal contact pressure.

Discretization of the domain by $p$-version of the finite element method is advantageous [21], since it results in fast convergence and high order mapping assures accurate geometry for the shape optimization.

An extensive survey of contact pressure optimization problems was presented by Hilding et al. [22]. Optimization problems with frictional contact were investigated in [23–30]. Special methods (level set method, evolutionary approach) were also used for topology optimization [31, 32]. An interesting solution was presented in [33] for multiple load cases and for incomplete external loading data by Banichuk and Ivanova [34]. Mathematical programming technique is used by many authors for shape optimization of structures in frictionless contact cases [35–40]. In [41], a unified shape optimization approach was developed for both minimization of boundary stress concentration and of peak contact pressure. Shape modifications in the iterational process were based on the distribution of the stress field, and the modification step was controlled by the relative deformation.

Numerous papers were devoted to a redesign procedure aimed at the wear reduction for rail and wheel by generation of new profiles of contacting bodies [42–44]. Similarly, the o-ring seal shape optimization is important for infallible work of the o-ring seal construction [45].

In [46–48], several classes of optimization problems have been considered with account of wear process. It was demonstrated that minimization of the generalized wear volume rate, generalized friction dissipation power and generalized wear dissi-

pation power and application of the optimality conditions provides different contact pressure distributions and local wear rates. In general, both singular and regular regimes of the wear rate and pressure distribution may occur [46]. The wear rule was presented by a nonlinear relation of wear rate to friction traction and relative sliding velocity, similar to the Archard rule. It was assumed that the relative sliding velocity between contacting bodies results from translation and rotation of two bodies. In this case for steady wear state the contact pressure is reached by minimization of the wear dissipation power. The specified pressure distribution can next be used for calculation of the wear shape form. Analytical solutions have also been presented for contact wear problems, cf. Goryacheva [49].

Contact optimization problems for kinematical constraints were studied in the paper by Páczelt [50]. Another interesting problem was concerned with optimization of round-off rollers aimed at the maximization of admissible loads of rolling bearings [51]. In the optimization of roller shape, the influence matrix is derived from the solution of the elastic half-space problem [52], and the mirror technique is also applied in this program [20].

Contact analysis problems are non-smooth due to unilateral contact conditions, requiring variational stress or displacement inequalities to be applied for the boundary-value problem solution. In the case of contact friction conditions, typical in treating contact slip or sliding and induced wear growth, the loading-unloading conditions and the active slip rule constitute an additional source of non-smooth response. The solution of contact analysis problems will not be discussed in detail in the paper. The contact optimization problems of min-max or max-min character, related to local state values, are also non-smooth.

## 2 Optimal Design for Controlled Contact Pressure

In our analysis, it is assumed that the bodies are in the conformal contact on the whole subdomain $\Omega_c$ of the contact zone $S_c = \Omega$, (Fig. 1). We introduce the surface coordinates $s$, $t$ and assume that the following pressure distribution is reached due to shape optimization [18]:

$$p_n(\mathbf{x}) = c(\mathbf{x}) p_{n,\max}, \tag{1}$$

where the assumed control function $c(\mathbf{x})$ must satisfy the condition $0 \leq c(\mathbf{x}) \leq 1$, and

$$p_{n,\max} = \max p_n(\mathbf{x}) \quad \mathbf{x} = [s, t]. \tag{2}$$

In the subdomain $\Omega_{nc} (\Omega = \Omega_c \bigcup \Omega_{nc})$, the contact pressure is not controlled and does not exceed the values specified by (1), so that

$$\chi(\mathbf{x}) = c(\mathbf{x}) p_{\max} - p(\mathbf{x}) \geq 0 \qquad \mathbf{x} \in \Omega_{nc}. \tag{3}$$

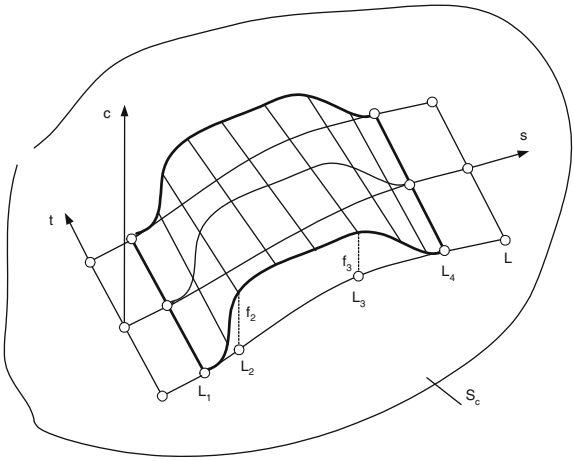



**Fig. 1** Control function $c(\mathbf{x}) = c(s) \cdot \tilde{c}(t)$

Let us introduce the functions depending on the varying position parameter $s$

$$C^*(s) = f_2 + (f_3 - f_2)\frac{s - L_2}{L_3 - L_2}, \quad f_2 \geq 0, \ f_3 \geq 0, \tag{4}$$

and

$$c(s) = \begin{cases} 0, & 0 \leq s \leq L_1, \\ C^*(s)\left\{3\left(\frac{s-L_1}{L_2-L_1}\right)^2 - 2\left(\frac{s-L_1}{L_2-L_1}\right)^3\right\}, & L_1 \leq s \leq L_2, \\ C^*(s), & L_2 \leq s \leq L_3, \\ C^*(s)\left\{1 - 3\left(\frac{s-L_3}{L_4-L_3}\right)^2 + 2\left(\frac{s-L_3}{L_4-L_3}\right)^3\right\}, & L_3 \leq s \leq L_4, \\ 0, & L_4 \leq s \leq L. \end{cases} \tag{5}$$

Here some of the parameters $f_2, f_3, L_i, i = 1, 2, 3, 4$, are fixed while the other are determined in the optimization process. It is assumed that the pressure distribution now is

$$c(\mathbf{x}) = c(s) \cdot \tilde{c}(t), \tag{6}$$

where we set $\tilde{c}(t) = 1$ in view of one parameter variation of contact pressure. Note that for $f_2 = f_3$, $L_1 = L_2 = 0$, $L_3 = L_4 = L$, we obtain uniform distribution pressure over $\Omega_c$.

An extensive study of this type contact optimization problems for 2D and 3D models using the control functions of type (4)–(6) was presented in [19].

# 3 Optimization Problems for Axisymmetric Bodies with Arbitrary Meridian Profile

## *3.1 Specified Punch Displacement*

Assume that the uniform vertical displacement $w_0$ is prescribed on the top punch surface (Fig. 2). The pressure distribution parameters $f_2$, $f_3$, $L_j$, $j = 1, \ldots, 4$, are fixed but the maximum pressure is subject to control. The minimal gap $g_{\min}$ is assumed to be zero. Now $s$ is defined as $s = r - r_i$, where $r$ denotes the radius and $r_i$ is the internal punch radius. We formulate the following optimization problem:

**Problem 1** Minimize the maximal contact pressure $p_{n,\max}$ by determining the initial gap function $g = g(s)$, such that $g(s_*) = g_{\min} = 0$, where $s = 0$ at the internal punch radius $r_i$, i.e.,

$$\min \Big\{ p_{n,\max} \mid p_n \geq 0,\ d = d(s,\ u_n^{(l)}) = g + u_n^{(2)} - u_n^{(1)} = 0,\ l = 1, 2 \\ \chi(s) = c(s)\, p_{n,\max} - p_n(s) = 0,\ \min g = g_{\min} = 0 \Big\}. \tag{7}$$

After determining the optimal gap function $g = g\,(s)$, the resultant contact force can be calculated by the formula

$$F_p^* = 2\pi \int_{r_i}^{r_e} p_n\, \cos\alpha\, r \sqrt{1 + \left(f_m'\right)^2}\, \mathrm{d}r, \tag{8}$$

where $r_e$ denotes the external punch radius, $\alpha$ is the direction of the contact normal, $f_m = f_m(r)$ of the meridian curve, $f_m' = \frac{\mathrm{d}f_m}{\mathrm{d}r}$.

**Problem 2** Assume now that the minimal gap $g_{\min}$ does not vanish but its value is determined in the optimization process. The value of $F_p$ transmitted by the contact area is now specified, so that we have another problem

$$\min \Big\{ p_{n,\max} \mid p_n \geq 0,\ d = d\left(s, u_n^{(l)}\right) = 0, l = 1, 2,\ \chi = 0, \\ F_p = 2\pi \int_{r_i}^{r_e} p_n \cos\alpha\, r \sqrt{1 + \left(f_m'\right)^2}\, \mathrm{d}r \Big\}. \tag{9}$$

**Problem 3** If the constraint on effective stress $\sigma_{eq}$ is introduced, then the value of $F_p$ cannot be selected arbitrarily and its maximum value constitutes an unspecified variable. The problem of maximization of the contact force can be formulated as follows:

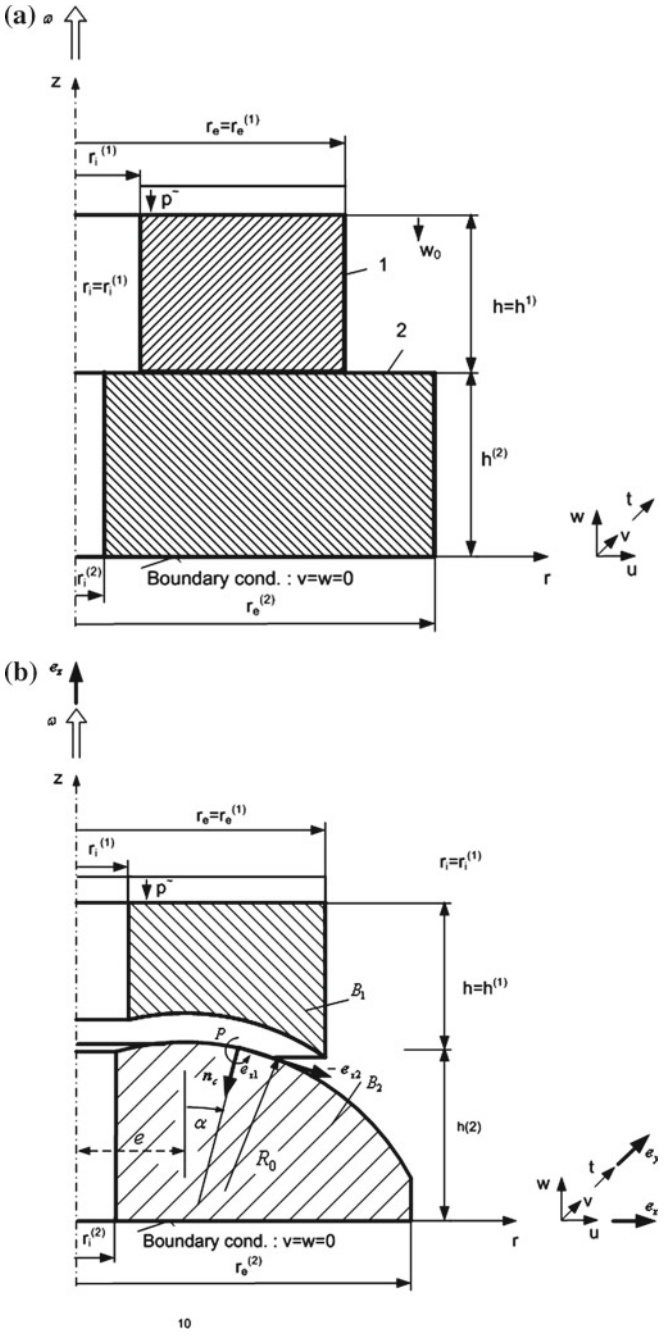



**Fig. 2** Contact of cylindrical bodies: **a** on a plane surface, **b** on a toroidal surface

$$\max\left\{F_p \mid \min\left[p_{n,\max} \mid p_n \geq 0,\ d = 0,\ \chi = 0\right],\ \sigma_{eq} \leq \sigma_u\right\}, \tag{10}$$

where $\sigma_u$ is the ultimate stress.

**Problem 4** An alternative design problem can be considered if the punch displacement $w_0$ is maximized within the imposed stress constraint $\sigma_{eq} \leq \sigma_u$ and the gap constraint $g_{\min} = 0$. Then, the problem is formally defined as follows:

$$\max\left\{w_0 \mid \min\left[p_{n,\max} \mid p_n \geq 0,\ d = 0,\ \chi = 0,\ g_{\min} = 0\right],\ \sigma_{eq} \leq \sigma_u\right\}. \tag{11}$$

This problem belongs to a class of displacement induced contact optimization problems.

## 3.2 Traction Induced Loading

Assume that the uniform axial pressure $\sigma_z = -\tilde{p}$ is applied at the top punch surface (Fig. 2) with the resulting force $F_0 = \pi(r_e^2 - r_i^2)\tilde{p}$. A typical optimization problem is to minimize the maximal contact pressure with specification of initial gap function $g = g(s)$ and proper selection of parameters $L_1, L_2, L_3$, and $L_4$. In the next examples, we set $f_2 = f_3 = 1$. If these parameters are varied and are determined in the optimization process, then $L_1 = L_2 = 0$, $L_3 = L_4 = r_e - r_i$ are the optimal values and the uniform pressure distribution is attained in the contact domain. We have the problem

**Problem 5**

$$\min\left\{p_{\max} \mid p_n \geq 0,\ d = 0,\ \chi = 0,\ g_{\min} = 0\right\}. \tag{12}$$

There are numerous studies of this problem (see, e.g., [12–14]).

## 3.3 Rotating Punch Under Compressive Loading

Assume now that the punch rotates with respect to its axis with the angular velocity $\omega$, (Fig. 2). Denote by $\tau_n$ the shear stress, by $\mu$ the friction coefficient and by $\dot{u}_\tau = r\omega$ the relative velocity. Specify the dissipation power due to frictional sliding at the contact surface. Then,

$$D_F = \int\limits_{S_c} \boldsymbol{\tau}_n \cdot \dot{\mathbf{u}}_\tau \, \mathrm{d}S = \omega\mu \int\limits_{r_i}^{r_e} 2\pi r^2 p_n \cos\alpha \sqrt{1 + (f_m')^2}\, \mathrm{d}r = M_T\omega, \tag{13}$$

where $M_T$ is the torque.

**Problem 6** Assume now that the uniform vertical traction $\sigma_z = -\tilde{p}$ is applied at the top boundary of the punch. Consider the problem of torque maximization assuming the parameters $L_1$ and $L_2$ as unspecified and $L_3$, $L_4$ as fixed. Then, we have

$$\max_{g(s),L_1}\left\{M_T=\mu\int_{r_i}^{r_e}2\pi r^2 p_n\cos\alpha\sqrt{1+(f_m')^2}\,\mathrm{d}r\,\middle|\,p_{n,\max}\le p_0,\ p_n\ge 0,\ d=0,\right.$$
$$\left.F_p-F_0=0,\ \chi=v(s,L_1,L_2(L_1))p_{n,\max}-p_n(s)=0,\ g_{\min}=0\right\},\tag{14}$$

where $p_0$ is a given pressure value. It is obvious that the contact pressure is shifted to the external boundary $r = r_e$.

**Problem 7** A similar solution is obtained when the additional stress constraint is introduced and the value of $p_{n,\max}$ cannot be fixed in advance. The solution is generated by maximizing the value of $L_1$ and the problem formulation is

$$\max_{g(s),L_1}\left\{M_T\mid p_n\ge 0,\ d=0,\ F_p-F_0=0,\ \chi=\chi(s,p_n,L_1)=0,\ \sigma_{eq}\le\sigma_u,\ g_{\min}=0\right\}.\tag{15}$$

**Problem 8** In order to minimize the dissipation power or torque, assume that $L_1 = 0$, $L_2 = 0$ and $L_4 - L_3$ are fixed, however, $L_4$ and $L_3$ may vary. The optimization problem now is formulated as follows:

$$\min_{g(s),L_4}\left\{D_F\mid p_n\ge 0,\ d=0,\ F_p-F_0=0,\ \chi=\chi(s,p_n,L_4)=0,\ g_{\min}=0\right\}.\tag{16}$$

**Problem 9** If the stress constraint $\sigma_{eq} \le \sigma_u$ is imposed, then the dissipation power is minimized with respect to the parameter $L_4$. In this case, we have the problem

$$\min_{g(s),L_4}\left\{D_F\mid p_n\ge 0,\ d=0,\ F_p-F_0=0,\ \chi=\chi(s,p_n,L_4)=0,\ \sigma_{eq}\le\sigma_u,\ g_{\min}=0\right\}.\tag{17}$$

## 4 Optimization Problems for a Steady Wear State

Relative sliding motion of two elastic bodies in contact induces wear process and contact shape evolution. In this case, shape modification is associated with the material removal due to wear and the related boundary motion in the normal contact direction. The velocity of contact shape modification is specified by the wear rate $\dot{w}_{i,n}$ for $i$th body. A modified Archard wear rule [46] specifies the wear rate $\dot{w}_{i,n}$ of the $i$th body in the normal contact direction. Following the previous work [46–48] it is assumed that

$$\dot{w}_{i,n} = \beta_i(\tau_n)^{b_i} \, \|\dot{\mathbf{u}}_\tau\|^{a_i} = \beta_i(\mu p_n)^{b_i} \, \|\dot{\mathbf{u}}_\tau\|^{a_i} = \beta_i(\mu p_n)^{b_i} v_r^{a_i} = \tilde{\beta}_i p_n^{b_i} v_r^{a_i}, \quad i = 1, 2, \tag{18}$$

where $\mu$ is the friction coefficient, $\beta_i, a_i, b_i$ are the wear parameters, $\tilde{\beta}_i = \beta_i \mu^{b_i}$, $v_r$ is the relative velocity between the bodies in the tangential direction of the contact surface. The shear stress at the contact surface is denoted by $\tau_n$ and calculated in terms of the contact pressure $p_n$ by using the Coulomb friction law $\tau_n = \mu p_n$. For analysis of the wear process and contact shape optimization three types of functionals can be taken. First, the generalized wear volume rate [46, 47] can be presented as follows:

$$\dot{W}^{(q)} = \sum_{i=1}^{2} \left( \int_{S_c} \dot{w}_i^q \, \mathrm{d}S \right)^{\frac{1}{q}} = \sum_{i=1}^{2} \left( \int_{S_c} (\tilde{\beta}_i p_n^{b_i} v_r^{a_i})^q \, \mathrm{d}S \right)^{\frac{1}{q}} = \sum_{i=1}^{2} A_i^{\frac{1}{q}}. \tag{19}$$

Next, the generalized friction dissipation power at the surface $S_c$ [46, 47] is

$$D_F^{(q)} = \left( \int_{S_c} (\mu p_n v_r)^q \, \mathrm{d}S \right)^{\frac{1}{q}} = B^{\frac{1}{q}}, \tag{20}$$

and the generalized wear dissipation power [47] in the new modified form equals [53]

$$D_w^{(q)} = \sum_{j=1}^{J} \sum_{i=1}^{2} \left( \int_{S_c^{(j)}} (\mathbf{t}_i^c \cdot \dot{\mathbf{w}}_i)^q \, \mathrm{d}S \right)^{\frac{1}{q}} = \sum_{j=1}^{J} \sum_{i=1}^{2} C_i^{(j)\frac{1}{q}}, \tag{21}$$

where $q$ is called the control parameter. The contact surface is assumed to be separated in $J$ parts (Fig. 3), but the rigid body displacement is assumed to be specified by the same degree of freedom. The introduction of the wear rate vector $\dot{\mathbf{w}}_i$ coaxial with the rigid body wear velocity provides a new concept in the description of the wear process [47].

The transient process tends to a steady state occurring at fixed contact stress and strain distribution. It has been shown in [47, 53] that the steady state conditions for the wear problem are obtained from minimization of the generalized wear dissipation power at $q = 1$

$$D_w = \sum_{j=1}^{J} \sum_{i=1}^{2} \left( \int_{S_c^{(j)}} (\mathbf{t}_i^c \cdot \dot{\mathbf{w}}_i) \, \mathrm{d}S \right) = \sum_{j=1}^{J} \sum_{i=1}^{2} C_i^{(j)}, \tag{22}$$

where $\mathbf{t}_i^c$ is the contact traction, $\dot{\mathbf{w}}_i$ is the wear rate vector of the $i$th body [47, 53].

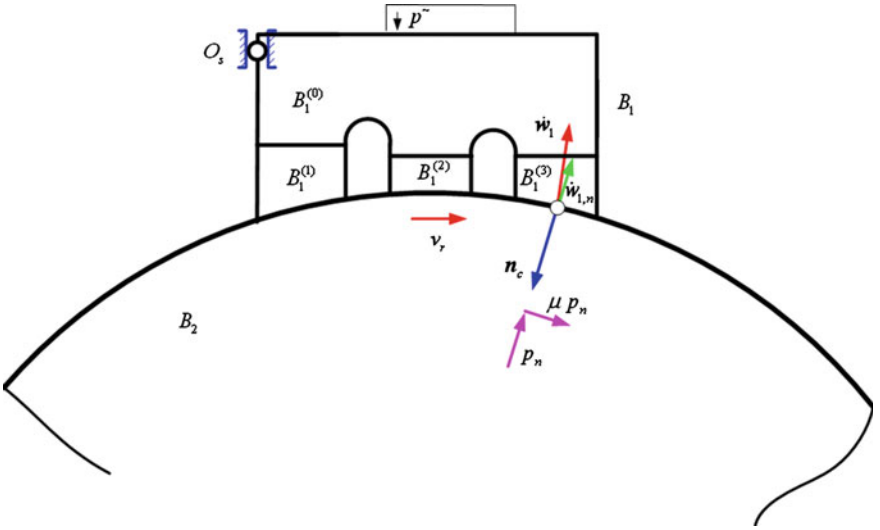



**Fig. 3** Plane structure with three separated contact zones of body $B_1$ can rotate around support $O_s$ and translate vertically. The body $B_2$ (*disk*) is assumed to rotate with constant angular velocity generating the relative sliding velocity $v_r$ on the contact surface

Equations (23), (24) contain three different optimization problems for different objective functions $\dot{W}^{(q)}$, $D_F^{(q)}$, $D_w^{(q)}$. Thus for generalized functionals the optimization problems are stated by requiring minimization of one of the objective functions placed in the column of (23).

**Problem 10**

$$\min \left\{ \begin{matrix} \dot{W}^{(q)} \\ D_F^{(q)} \\ D_w^{(q)} \end{matrix} \;\middle|\; p_n \geq 0,\ d = 0,\ g_{\min} = 0,\ \text{Equilibrium equations for the punch} \right\}. \tag{23}$$

In the alternative class of optimization problems, the local stress constraint is introduced and the max-min formulation for one of the objective functions is applied, thus

**Problem 11**

$$\max_q \left\{ \min \left\{ \begin{matrix} \dot{W}^{(q)} \\ D_F^{(q)} \\ D_w^{(q)} \end{matrix} \;\middle|\; p_n \geq 0,\ d = 0,\ g_{\min} = 0, \right. \right.$$
$$\left. \left. \text{Equilibrium equations for the punch} \right\}, \sigma_{eq} \leq \sigma_u \right\}. \tag{24}$$

The minimum of the wear dissipation power corresponds to the steady wear state. In this case the optimization problem is

**Problem 12**

$$\min \{D_w \mid p_n \geq 0,\ d = 0,\ g_{\min} = 0,\ \text{Equilibrium equations for the punch}\}. \tag{25}$$

We can now formulate a new optimization problem of contact shape by requiring maximization of the loading pressure $\tilde{p}$ subject to the constraint set on the Mises equivalent stress in the steady wear state not exceeding the value $\sigma_u$, that is $\sigma_{eq} \leq \sigma_u$, thus

**Problem 13**

$$\max_{\tilde{p}} \Big\{ \min \{D_w \mid p_n \geq 0,\ d = 0,\ g_{\min} = 0,\ \text{Equilibrium equations for the punch}\}, \\ \sigma_{eq} \leq \sigma_u \Big\}. \tag{26}$$

The set of state relations should be added to all formulations of optimization problems, namely balance of linear momentum, constitutive equation (the Hooke law), strain-displacement relations, and the boundary conditions:

$$\begin{gathered} \boldsymbol{\sigma} \cdot \nabla + \boldsymbol{f} = \boldsymbol{0}, \quad \boldsymbol{\sigma} = \boldsymbol{C} : \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} = \frac{1}{2}\left[\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T\right], \\ \boldsymbol{\sigma} \cdot \boldsymbol{n} = \boldsymbol{t}_0 \quad \text{on } S_t, \quad \boldsymbol{u} = \boldsymbol{u}_0 \quad \text{on } S_u, \end{gathered} \tag{27}$$

where $\boldsymbol{\sigma}$ and $\boldsymbol{\varepsilon}$ are the stress and strain tensors, $\boldsymbol{C}$ is the elasticity 4th order tensor, $\boldsymbol{f}$ denotes the body force per unit volume, $\boldsymbol{t}_0$ and $\boldsymbol{u}_0$ are the tractions and displacements on the boundary portions $S_t$ and $S_u$, $\nabla$ is the gradient differential operator, $\cdot$ denotes the scalar product and : denotes the double scalar product. $^T$ stands for the transpose.

In the contact zone $S_c$, the contact pressure is $p_n = -\boldsymbol{n} \cdot \boldsymbol{\sigma} \cdot \boldsymbol{n}$ and for the Coulomb friction condition, in the slip case, the shear stress is $\tau_n = \mu\, p_n$, where $\mu$ is the coefficient of friction.

For Problems 10–13 formulae for the distribution of the contact pressure can explicitly be defined by the optimality conditions. If the heat generation and thermal distortion are accounted for, then an interesting result is obtained, namely, the contact pressure distribution does not depend on the temperature state, but the corresponding contact surface shape depends on temperature state, that is the shapes are different for state with temperature and for state without temperature [54].

For axisymmetric bodies, the contact pressure distribution in the steady wear state has the following form [46]:

$$\begin{cases} \min \dot{W}^{(q)}, \ p_n = Q_{\dot{W}} r^{-\frac{aq}{bq-1}} (\cos\alpha)^{\frac{1}{bq-1}} & \text{if } bq-1 \neq 0, \\ \min D_F^{(q)}, \ p_n = Q_{D_F} r^{-\frac{q}{q-1}} (\cos\alpha)^{\frac{1}{q-1}} & \text{if } q-1 \neq 0, \\ \min D_w^{(q)}, \ p_n = Q_{D_w} r^{-\frac{aq}{(b+1)q-1}} (\cos\alpha)^{\frac{1}{(b+1)q-1}} & \text{if } (b+1)q-1 \neq 0, \end{cases} \tag{28}$$

where $Q_{\dot{W}}$, $Q_{D_F}$, $Q_{D_w}$ are calculated from the equilibrium condition of the punch. An example demonstrating how the parameter $q$ affects the stress state will be shown in this paper (see part Sect. 6.2.3).

## 5 Iterative Solution of the Contact Shape Optimization Problem

We use the $p$-version of the finite element method for solving the optimization problems. Then, Problems 1–13 are reduced to nonlinear programming problems. This class of problems is solved by a special iteration process. We distinguish iterations of two types. In the *first type iteration process* the optimal shape is determined for fixed values of the control parameters $f_2$, $f_3$, $L_i$, $i = 1, 2, 3, 4$, (in Problems 10–13 using the Eq. (28)) specifying contact pressure distribution. The maximization or minimization of objective function is then performed for unspecified $p_{\max}$ and $g = g(s)$. However, for some problems the selected values of parameters may violate the stress constraint. If the effective Mises stress $\sigma_{eq}$ must be below a prescribed ultimate stress $\sigma_u$, ($\sigma_{eq} \leq \sigma_u$) then the optimization problem includes this additional condition and the solution requires another iteration, labeled as the *second type iteration process*.

### *5.1 First Type Iteration Process*

The first iterative scheme for contact shape optimization was discussed in detail by Páczelt [17]. Here we only outline the consecutive steps.

The iterative process is described by the following scheme:

1. Solution of the original contact problem: specification of the contact pressure

$$p_n^{(0)} = p_n^{(0)}(s), \quad p_{\max}^{(0)}, \quad k = 0.$$

2. $k = k + 1$.
3. Let the controlled pressure distribution be defined in accordance with (3),

$$p_n^{(k)}(s) = c(s)\, p^*.$$

Consider first the displacement induced loading. The value of the parameter $p^*$ is obtained from the contact problem solution at the previous iterative step $k-1$,

$$p^* = \max p_n^{(k-1)}(s).$$

For the case of traction or mixed boundary condition, the solution of contact problem is not required for calculation of the $p^*$. The value of $p^*$ at each step is specified from the load equilibrium condition, thus

$$F_0 = 2\pi \int_{r_i}^{r_e} p^* c(r) \cos\alpha r \sqrt{1 + (f'_m)^2}\, \mathrm{d}r.$$

4. After the previous steps the contact pressure is known. The separated bodies are now loaded by the pressures $p_n^{(k)}(s)$ and $-p_n^{(k)}(s)$ and in the case of frictional contact also by the shear stresses $\mu p_n^{(k)}(s)$ and $-\mu p_n^{(k)}(s)$ in the tangential direction at the contact surface $S_c$. From the finite element solution for separated bodies the normal displacements $u_n^{(1)}$ and $u_n^{(2)}$ are determined ($u_n^{(1)}$ for the punch, $u_n^{(2)}$ for the substrate body).
5. Calculate the discontinuity of normal displacements

$$m(s) = u_n^{(1)} - u_n^{(2)} = [u_n]\,.$$

6. Specify the minimal value $\min m(s) = m(s_*)$.
7. Generate the new initial gap:

$$\begin{cases} \text{in Problems 1, 5–13:} & \text{if } g(s) = m(s) - m(s_*) \Rightarrow g(\mathbf{x}), \\ \text{in Problems 2–4:} & \text{if } g(s) = m(s). \end{cases}$$

8. Solution of the contact problem with the new gap: specification of contact pressure $p_n^{(k)} = p_n^{(k)}(s)$.
9. Repeat the steps 2–8 until the convergence condition is satisfied:

$$g_{tol} = 2\pi \int_{r_i}^{r_e} \left| \frac{g^{(k)} - g^{(k-1)}}{g^{(k)}} \right| r \sqrt{1 + (f'_m)^2}\, \mathrm{d}r \le 10^{-4}.$$

## *5.2 Second Type Iteration Process*

Now, we discuss the second type iteration process, which is coupled with the first type iteration. When the stress constraint $\max \sigma_{eq} \le \sigma_u$ is imposed at any Gaussian integration point, the values of parameters assumed as fixed or specified in the first type iteration should now be updated in order to satisfy the stress constraint. Referring to Problem 4, where the displacement $w_0$ is to be maximized, the value of $(w_0)_{\max}$ assumed in the first type iteration must be reduced in the second type process. Sim-

ilarly, in Problem 6 the value of $L_1$ assumed in the first type iteration should be reduced, when Problem 7 is formulated with stress constraint. In Problem 8 the friction dissipation power is minimized and this leads to minimization of parameter $L_4$. However, in Problem 9 the minimal value of $L_4$ reached in the first type iteration should be increased to meet the stress constraint.

Denote generally by $f$ the parameter which should be updated in the iteration. The loading process is characterized by the variable *istep* specifying the consecutive iteration number, $istep = 1, 2, 3, \ldots$. The value of $f$ is calculated by the following formula: $f = f_0 + \Delta f \cdot (istep - 1)$, where $f_0$ and $\Delta f$ are chosen in advance. For instance, for Problem 9 there is $f_0 = r_e - r_i = L$, $\Delta f = -(r_e - r_i)/10$. The optimization problem is solved by the first type iteration at the fixed $f$. At each *istep* a new shape is determined for the upper body.

The effective stress value $\sigma_{eq}$ is calculated at the Gaussian integral points of the finite elements and at the boundary points as well: $(\xi = -1, \xi_1, \xi_2, \ldots, \xi_{NG}, 1)$ and $(\eta = -1, \eta_1, \eta_2, \ldots, \eta_{NG}, 1)$, where $\xi$, $\eta$ are the local normal co-ordinates, $NG$ denotes the number of integration points. When the equivalent stress constraint is violated at the $istep = istep_{**}$, the value of $f$ is properly updated. Assume that for the value $f = f^*$ at the $istep = istep_{**} - 1$ the effective stress is $\sigma^*_{eq} < \sigma_u$ and the parameter equals $f = f^{**}$ and effective stress exceeds the ultimate value $\sigma^{**}_{eq} \geq \sigma_u$ at the next loading step $istep = istep_{**}$. The optimal value of $f = f^{opt(i)}$ is searched in the interval $f^* < f^{opt(i)} < f^{**}$ by the following linearization process:

$$f^{opt(i)} = f^* + (f^{**(i-1)} - f^*)\frac{\sigma_u - \sigma^*_{eq}}{\sigma^{**(i-1)}_{eq} - \sigma^*_{eq}},$$

where

$$f^{**(0)} = f^{**}, \quad \sigma^{**(0)}_{eq} = \sigma^{**}_{eq}.$$

At each step $i$ of the second type iteration the contact shape is specified in the first iteration process. The second type iteration process will run until

$$\frac{\sigma_u - \sigma^{**(i)}_{eq}}{\sigma_u} \leq 0.005.$$

## 6 Examples

For numerical demonstration of the optimization problems, we select examples of two types. First, the optimal design for maximization of the contact force will be discussed for the interacting bodies depicted in Fig. 2a, and, second, we will present Problems 7 and 9 for cylindrical bodies with toroidal contact surface.

Assume that the body $B_1$ is rotating with the angular velocity $\omega$ around the $-z$ axis, and the second body $B_2$ is fixed. Material parameters of the bodies are as follows:

Young modulus $E = 2 \times 10^5$ MPa, Poisson ratio $\nu_P = 0.3$, coefficient of friction $\mu = 0.25$. The ultimate equivalent stress is assumed at the level $\sigma_u = 250$ MPa.

The results were obtained for $p$-approximations of the order up to $p = 8$. In Problem 7 of the second example, the convergence of calculations is demonstrated with increasing $p$. This is done in order to demonstrate the efficiency of the $p$-version approximations for the considered class of optimal control problems.

## *6.1 First Example: Solution of Problem 4*

We set the geometric parameters of cylinders as follows:

$$r_i^{(1)} = 20\,\text{mm}, \quad r_e^{(1)} = 120\,\text{mm}, \quad r_i^{(2)} = 20\,\text{mm}, \quad r_e^{(2)} = 140\,\text{mm},$$
$$h = h^{(1)} = h^{(2)} = 100\,\text{mm} \quad \text{(see Fig. 4a)}.$$

The control parameters in (5) are $L_1 = 0$, $L_2 = 4$ mm, $L_3 = 96$ mm, $L_4 = 100$ mm, $f_2 = f_3 = 1$. The boundary conditions are: on the cylindrical surfaces $r = r_i$ and $r = r_e^{(l)}$ $(l = 1, 2)$ the traction free boundaries, $\mathbf{t}_0^{(l)} = \mathbf{0}$ $(l = 1, 2)$, the bottom surface $z = 0$ of body $B_2$ is constrained, $v = w = 0$ and the upper surface $z = 2h$ of punch $B_1$ is subjected to the axial displacement $w_0 = 0.15$ mm and to rotational motion $v = r\omega t$ in the circumferential direction (see Fig. 2). In order to achieve an accurate solution of the finite element approximation, we use the quadratic $p$-version finite elements of order 8 in the truncated space [21]. In Fig. 4, the initial discretization mesh is shown and variation of $F_p$, $\max \sigma_{eq}$ and of $D_F/(\mu\omega) = M_T/\mu$ is visualized during the consecutive iteration steps.

The contact condition $d = 0$, $p_n \geq 0$ were checked at the Lobatto integration points $(NL = 9)$. In the initial configuration (flat contact surfaces) there is no initial gap between bodies and the stress singularity occurs at the external edge of contact zone $r = r_e = 120$ mm, $z = h^{(2)} = 100$ mm.

The optimal value of $F_p$ equals $F_p^{opt} = 9223.2$ kN and was computed within the interval $F_p^* = 8 \times 10^6\,\text{N} < F_p^{opt} < F_p^{**} = 10 \times 10^6\,\text{N}$ using the linear interpolation rule according to formula of the $2nd$ type iteration process. In our case, only one step of this iteration was required.

Figure 5a shows the distribution of the radial stress $\sigma_r$, the shearing stress in the circumferential direction, calculated on the contact surface by the formulae $\tau_{tz} = \mu p_n = -\mu\sigma_z$, the normal stress $\sigma_z$ and $\sigma_{eq}$ within the punch in the optimal configuration. Figure 5b shows the fields of the same stresses for the lower body $B_2$. It is seen that $\max \sigma_{eq}$ is reached at the location $r = r_i$, $z = 0$. The control of the contact pressure generates a nonsingular stress state in the bodies, with the stress $\sigma_z$ vanishing at the points $(r = r_i, z = h)$ and $(r = r_e, z = h)$.

Figure 6 illustrates the evolution of shape of the contact surface for different values of the resultant contact force $F_p$, calculated by the formula $F_p = 2 \times 10^6 * (NS - 1)$,

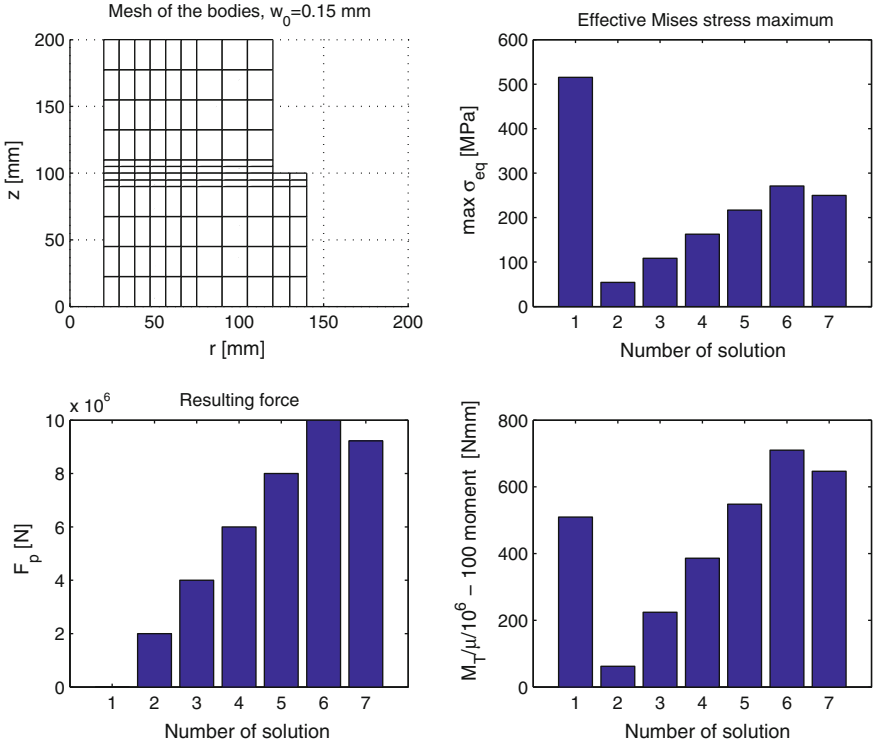



**Fig. 4** The results of the second type iteration steps in the solution of Problem 4. Number of solution ($NS = istep + 1$): 1—initial configuration, 2, 3, …, 6—control function with $F_p = 2 \times 10^6 * (NS - 1)$, 7—optimal solution

where $NS$ is the number of solution, $NS = istep + 1$. The optimal shape is marked by (o).

## *6.2 Second Example*

Consider now two cylindrical bodies interacting on a toroidal contact surface of curvature radius $R_0$ with its centre located at the distance $e$ from the cylinder axis, Fig. 2b. The radius of the contact point P in the meridional plane is $r = e + R_0 \sin\alpha$. The body $B_1$ is uniformly loaded by the axial traction $\sigma_z = -\tilde{p}$ which corresponds to the resultant force $F_0$. The $p$ version finite element mesh is demonstrated in Fig. 7. The number of elements in the contact region in the horizontal direction is 12 and in the vertical direction is 9 for both bodies. The horizontal (curved) and vertical lines correspond to the Lobatto integration points. Integration points are $p + 1 = 9$, where $p$ is the polynomial order of approximated displacement fields. The lower

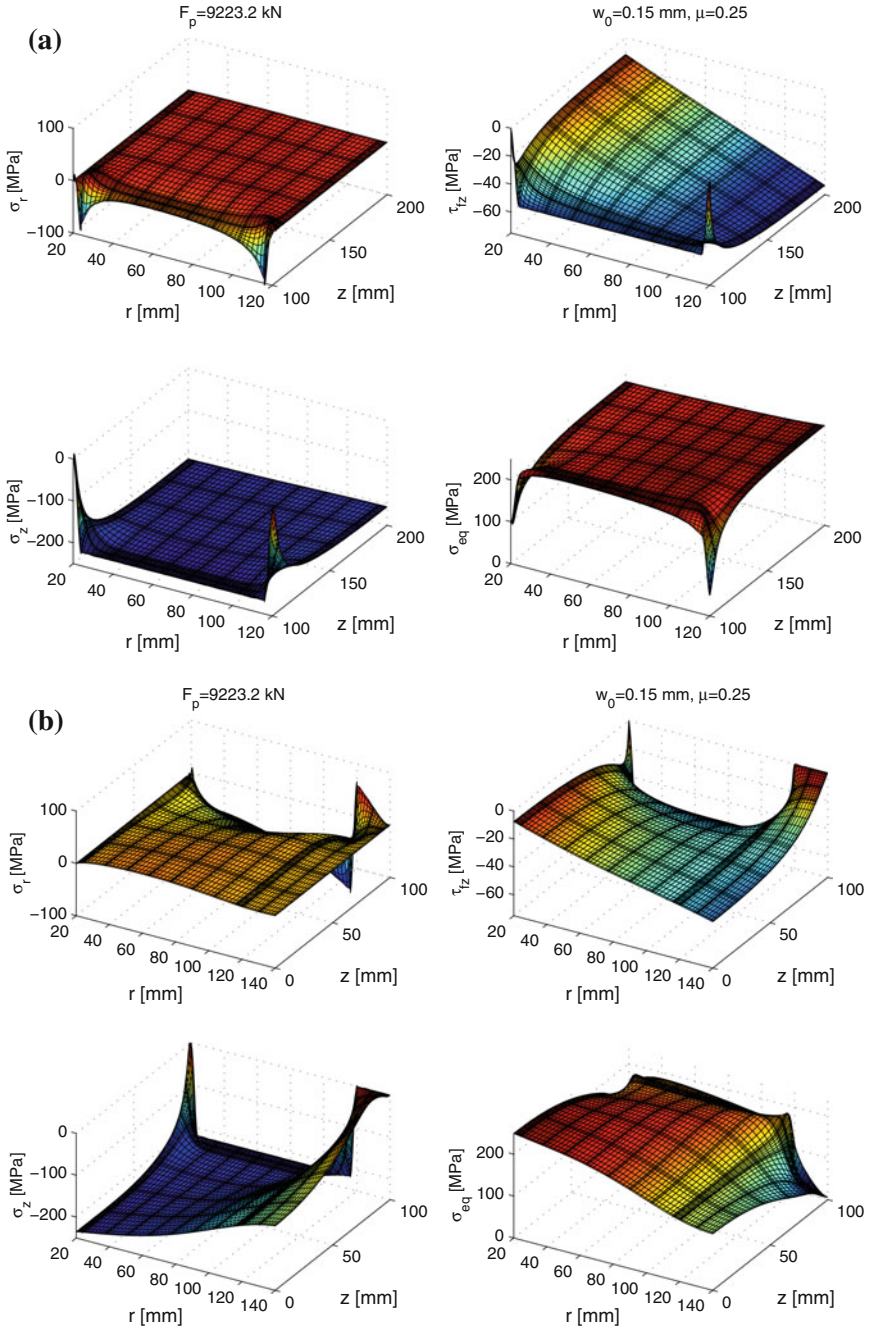



**Fig. 5** Stress distribution for the optimized design of Problem 4. **a** in body $B_1$, **b** in body $B_2$, $\sigma_r$ radial stress, $\sigma_z$ normal stress, $\tau_{fz} \equiv \tau_{\varphi z}$ shear stress and $\sigma_{eq}$ von Mises effective stress

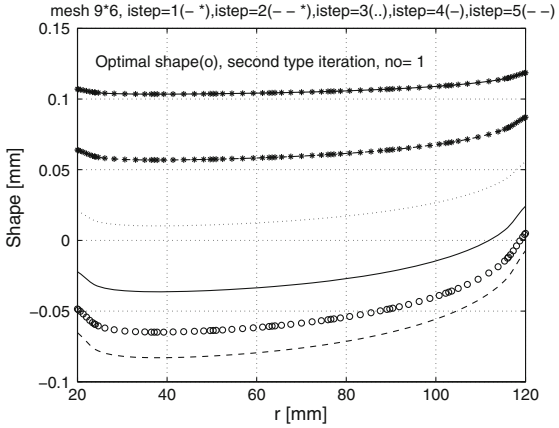



**Fig. 6** Shape evolution in the optimization process for the first example

body in the horizontal interval $r_i \leq r \leq r_e^{(2)}$ is divided by four elements. We select the following geometric and material parameters (the model of debris motion [48] is not applied here):

$$r_i = r_i^{(1)} = r_i^{(2)} = 10\,\text{mm}, \quad r_e = r_e^{(1)} = 100\,\text{mm}, \quad r_e^{(2)} = 120\,\text{mm},$$
$$h = h^{(1)} = h^{(2)} = 80\,\text{mm}, \quad e = 55\,\text{mm}, \quad R_0 = 100\,\text{mm}, \quad E = 2 \times 10^5\,\text{MPa},$$
$$\nu_P = 0.3, \quad \tilde{p} = 80\,\text{MPa}, \quad \mu = 0.25, \quad \mu_d = 0.$$

In calculations, we use the wear parameters $\tilde{\beta}_1 \neq 0$, $\tilde{\beta}_2 = 0$, and $a = b = 1$.

The contact problem was solved by applying the penalty method with the penalty parameter equal $1000\,E$. Contact springs are placed in the normal direction to the contact surface, that is in the direction $\mathbf{n}_c$.

### 6.2.1 Solution of Problem 7

Problem 7 was solved with the following initial values of control parameters are assumed: $L_1 = 0$, $L_2 = 10\,\text{mm}$, $L_3 = 80\,\text{mm}$, $L_4 = 90\,\text{mm}$, $f_2 = f_3 = 1$ (see Fig. 1). The problem is solved by *istep* $= 13$. The final values of control parameters are: $L_1 = 53.798\,\text{mm}$, $L_2 = 63.798\,\text{mm}$, $L_3 = 80\,\text{mm}$, $L_4 = 90\,\text{mm}$. The distribution of contact pressure is presented in Fig. 8a, the gaps at different iterative steps are shown in Fig. 8b. The torque is $M_T = 5.153\text{e} + 07\,\text{Nmm}$.

The distributions of the $\sigma_z$ and $\sigma_{eq}$ are illustrated in Fig. 9, and convergence of the $p$-version solutions is demonstrated in Fig. 10. We selected $3 \leq p \leq 8$ and built the convergence curves for control parameter $L_1$, torque $M_T$, maximal contact pressure $p_{\max}$ and normal gap $g_n$ at the perimeter point of the contact zone $r = 10$. The convergence rate is very high. The number of degrees of freedom (NDF) of the FE

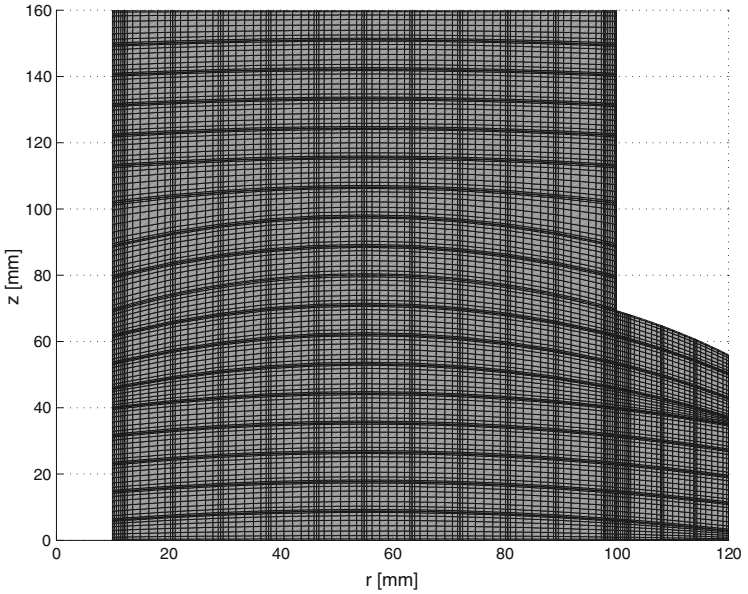



**Fig. 7** Finite element mesh for contacting bodies

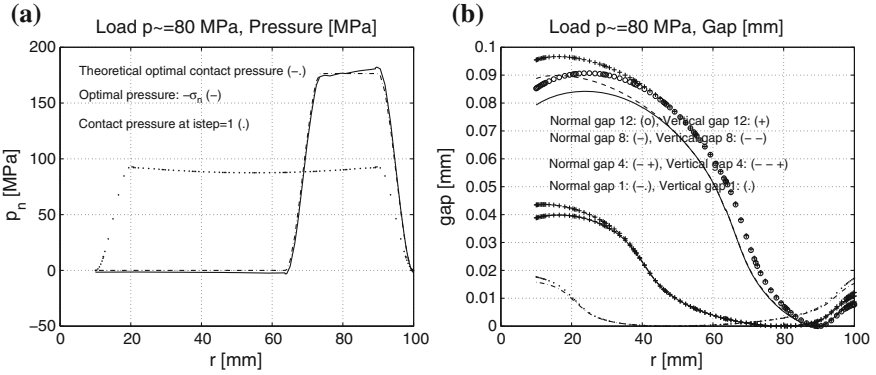



**Fig. 8** Analysis results of Problem 7: **a** contact pressure distribution, **b** contact gap at different iteration steps

model is the following: at $p = 3 \rightarrow NDF = 2714$, $p = 4 \rightarrow NDF = 4290$, $p = 5 \rightarrow NDF = 6370$, $p = 6, \rightarrow NDF = 8954$, $p = 7 \rightarrow NDF = 12042$, $p = 8 \rightarrow NDF = 15634$. Practically the results do not change for $p \geq 6$, thus the applied mesh is correct for the convergence solution with increasing polynomial order of finite elements.

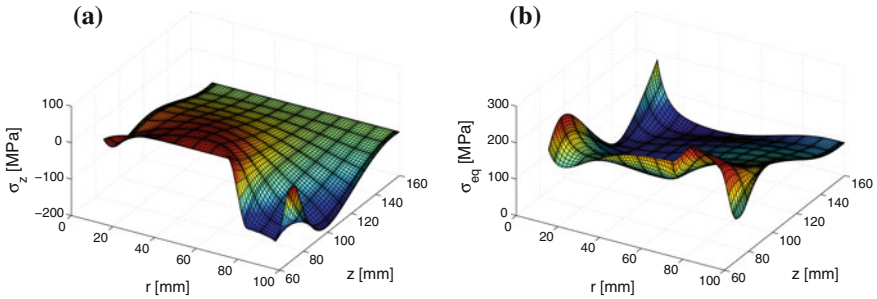



**Fig. 9** Distribution of stress components in body 1 at the maximal moment $M_T$: **a** distribution of $\sigma_z$, **b** distribution of $\sigma_{eq}$

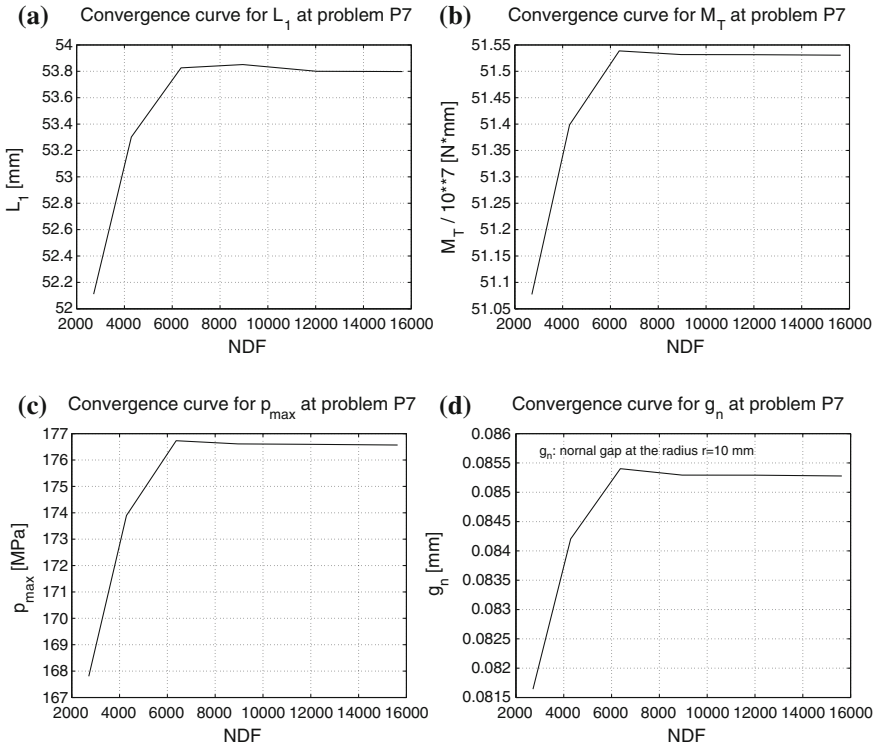



**Fig. 10** Convergence of the finite element solution for elastic system of cylindrical bodies (see Fig. 7), **a** convergence of the control parameter $L_1$, **b** convergence of the torque $M_T$, **c** convergence of the maximal contact pressure $p_{\max}$, **d** convergence of the normal gap at point $r = 10$

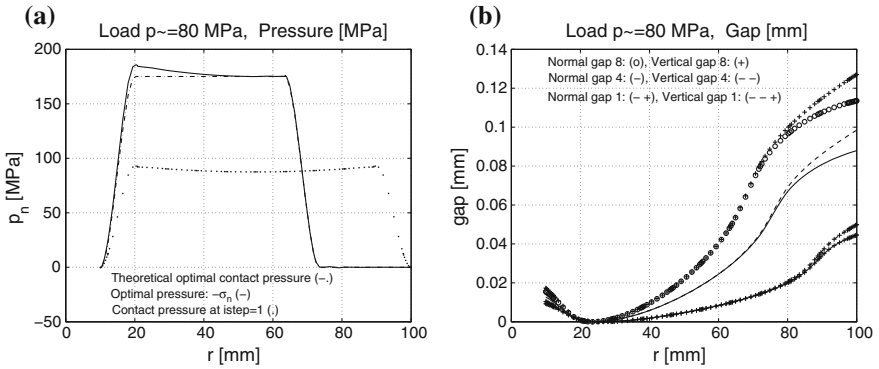



**Fig. 11** Analysis results of Problem 9; **a** contact pressure distribution, **b** gaps at different iteration steps

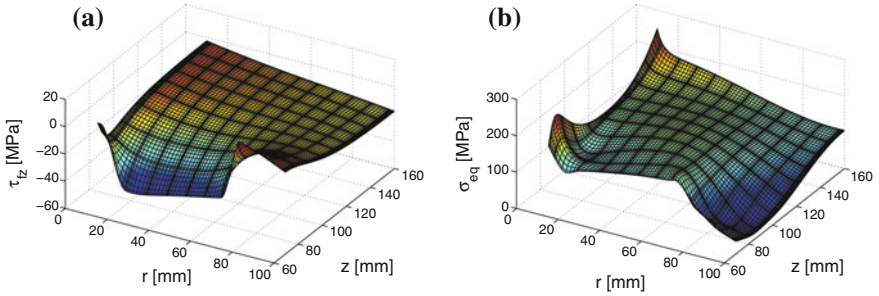



**Fig. 12** Distribution of stress component in body 1 at the minimal moment $M_T$: **a** distribution of $\tau_{fz} \equiv \tau_{\varphi z}$, **b** distribution of $\sigma_{eq}$

### 6.2.2 Solution of Problem 9

Figures 11 and 12 demonstrate results related to Problem 9. The final values of control parameters are $L_1 = 0$, $L_2 = 10\,\text{mm}$, $L_3 = 53.46\,\text{mm}$, $L_4 = 63.46\,\text{mm}$. The torque is $M_T = 2.94\text{e} + 07\,\text{Nmm}$ ($\sigma_{eq} = 250\,\text{MPa}$).

### 6.2.3 The Effect of Value of the Control Parameter $q$, Problems 10 and 11

We have solved Problem 10 defined by (23) for different values of parameter $q$. The calculated torque is presented in Fig. 13 and the contact pressure in Fig. 14. Theoretically the min value of the torque is $M_T^{\min} = r_i \mu F_0 \cos\alpha_0$, the maximum value is $M_T^{\max} = r_e \mu F_0 \cos\alpha_0$. These values are marked by $\circ$ and $\oplus$ in Fig. 13. Here the $\alpha_0$ is the angle of contact normal at radius $r_i$ and $r_e$.

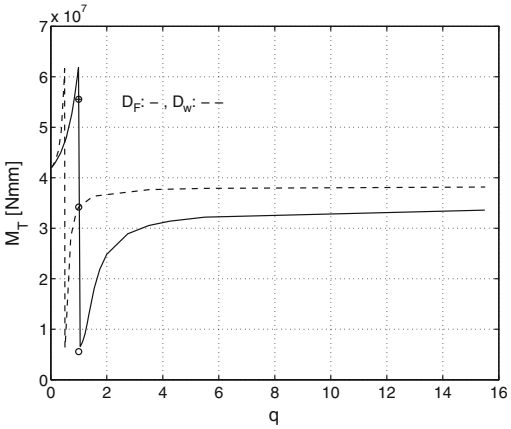



**Fig. 13** Torque variation at different values of parameter $q$ for optimization problems: min $D_F$ and min $D_w$. Theoretically $M_T^{\max} = 5.5549\text{e}+07$ Nmm: $\oplus$, $M_T^{\min} = 0.5549\text{e}+07$ Nmm: o

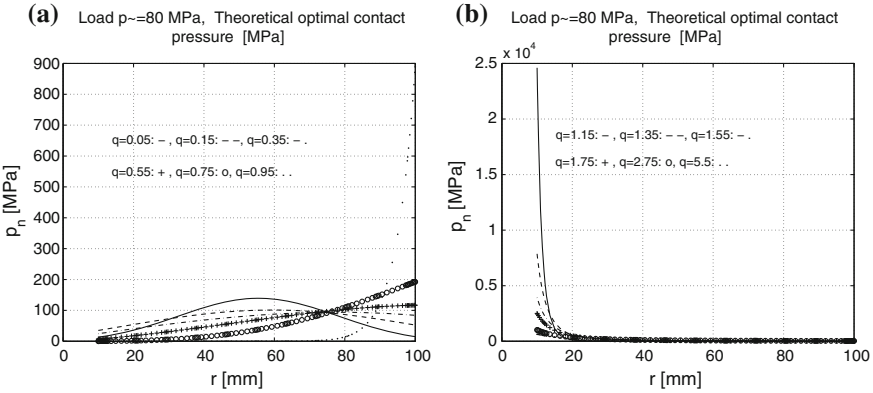



**Fig. 14** Pressure distributions for the problem min $D_F$ at different values of $q$, **a** solutions for values $q < 1$, **b** solutions for values near $q > 1$

If $q$ tends to 0, then the contact pressure has simple forms resulting from the respective minimization problems, namely:

$$\min D_F: \; p_n = Q_{D_F} r^{-\frac{q}{q-1}} \, (\cos\alpha)^{\frac{1}{q-1}} \quad \rightarrow \quad \frac{Q_{D_F}}{\cos\alpha},$$

$$\min D_w: \; p_n = Q_{D_w} r^{-\frac{aq}{(b+1)q-1}} \, (\cos\alpha)^{\frac{1}{(b+1)q-1}} \quad \rightarrow \quad \frac{Q_{D_w}}{\cos\alpha},$$

where from the equilibrium condition of the punch there is

$$Q_{D_F} = Q_{D_w} = \frac{F_0}{S_c} = \frac{2488141.6}{32260.51} = 77.13\,\text{MPa}.$$

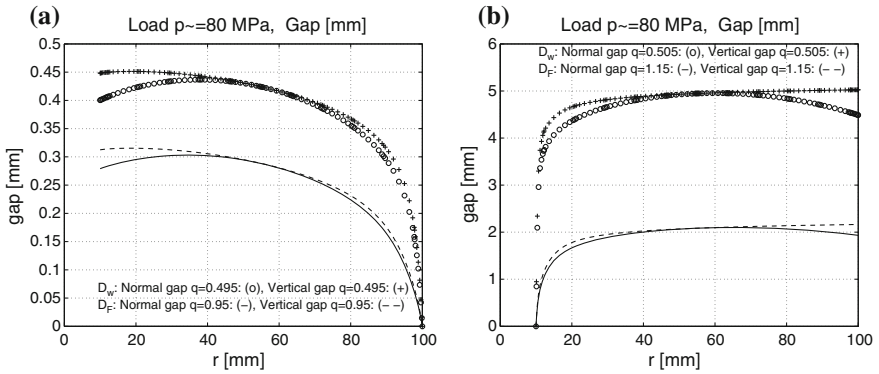



**Fig. 15** Contact gaps for problems min $D_F$ and min $D_w$ at near singular values of $q$, **a** singular solution near $q < 0.5$, **b** singular solution near $q > 0.5$

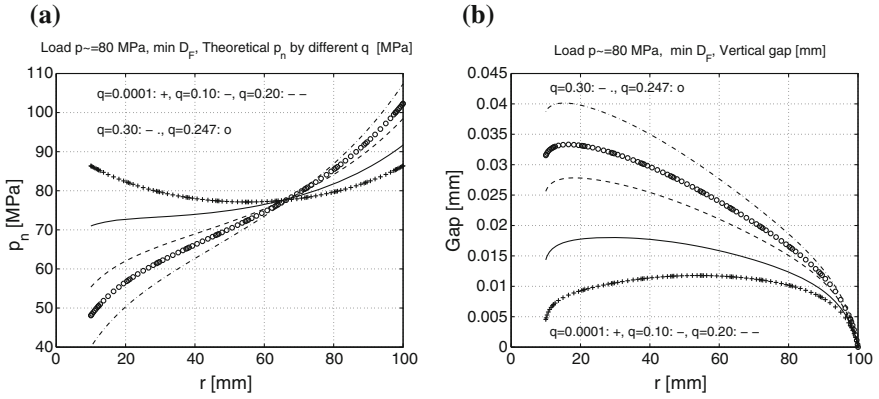



**Fig. 16** Analysis results of Problem 11: **a** contact pressure distribution in the iteration process, **b** contact gap at different values of $q$. Optimal solution is marked by o

If $q$ tends to 1, then the solution becomes singular for the problem min $D_F$, and at $q \to 0.5$ ($a = b = 1$) the solution is singular for the problem min $D_w$. The specified contact shape for $q$ near singular values is demonstrated in Fig. 15. The shape is characterized by sharp punch edges at $r_e$ and $r_i$ where the contact pressure reaches very high values. In this case, the effective stress exceeds the elasticity limit and the problem requires an elastic-plastic analysis. However, the problem can be treated within the linear elasticity relations provided the stress constraint $\sigma_{eq} \leq \sigma_u$ is applied and the admissible maximal value of $q$ is specified, that is Problem 11 is solved. This problem was solved in four iterations. The initial control parameter is $q = 0.01$ and the optimized value is $q^{opt} = 0.247$. The contact pressure and vertical gap forms are shown in Fig. 16. The torque in this case is $M_T^{opt} = 4.4055\text{e} + 07\,\text{Nmm}$.

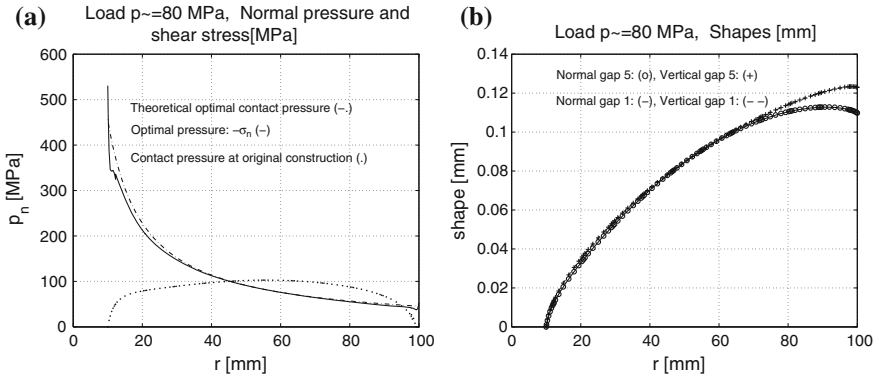



**Fig. 17** Steady wear state: **a** contact pressure distribution for the problem min $D_w^{(q)}$ at $q = 1$, **b** shape (1st type iteration is finished after 5 steps)

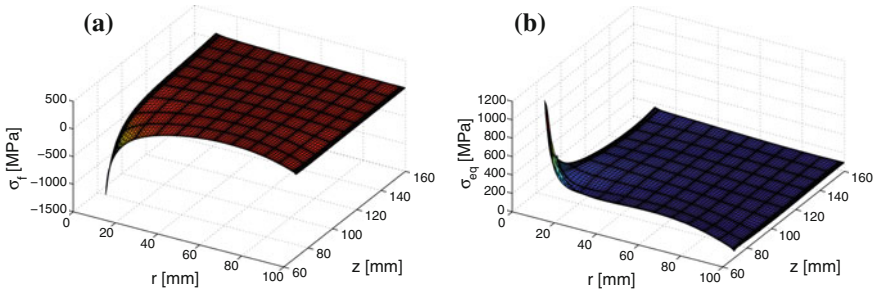



**Fig. 18** Stress distribution in the steady wear state: **a** $\sigma_f \equiv \sigma_\varphi$, **b** $\sigma_{eq}$

### 6.2.4 Solution for the Steady Wear State, Problem 11 and 12

It was stated previously, cf. [46–48] that the steady wear state solution is reached by minimization of the generalized wear dissipation power for the value $q = 1$ of the control parameter. Solving Problem 12, the contact pressure distribution is specified and shown in Fig. 17a, next the optimal gap is determined and presented in Fig. 17b. The distribution of circumferential stress $\sigma_\varphi$ is shown in Fig. 18a and the effective Mises stress distribution is demonstrated in Fig. 18b. It is seen, that much higher values are reached at the lower points of the cylindrical surface $r = r_i$. These values exceed the elasticity limit as $\sigma_u \ll \sigma_{eq}$ with the effective stress much higher than $\sigma_u$. For real description of the wear process, an elastic-plastic analysis should be applied.

The torque is

$$M_T^{\min} < M_T(D_w^{(q=1)}) = 3.42\text{e}+07\,\text{Nmm} < M_T^{\max}.$$

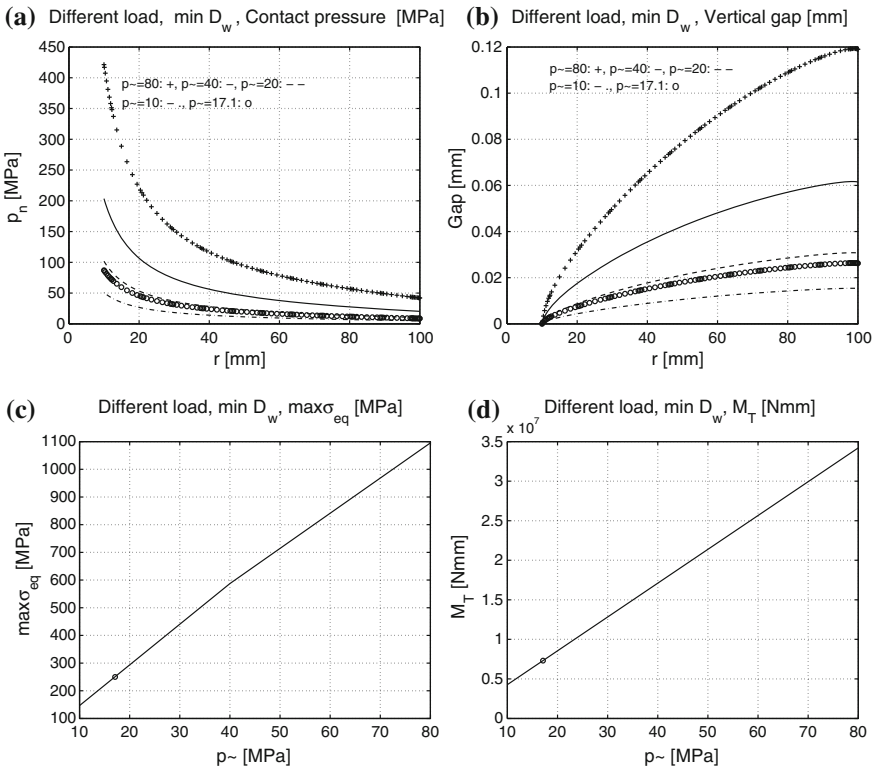



**Fig. 19** Solution for the steady wear state; **a** distribution of the contact pressure, **b** the vertical gap between the toroidal surfaces, **c** maximal values of Mises equivalent stress, **d** torque values for the rotating punch. Optimal solution is marked by o

If Problem 13 is solved, then the specified max $\tilde{p} = 17.1$ MPa. The different value of $\tilde{p}$ provides different contact pressure distribution and wear gaps between the bodies (see Fig. 19a, b). The values of max $\sigma_{eq}$ and of torque are demonstrated in Fig. 19c, d.

## *6.3 Contact Shape Optimization for Sliding Punches with Account for Wear*

In this section, our analysis is referred to the case of monotonically or reciprocally sliding punches with account for friction and wear effect. It will be demonstrated that the wear process, tending to its steady state, generates different contact shapes and pressure distributions for a relative monotonic or reciprocal sliding motion along the contact surface. Also, it is demonstrated that solutions for monotonic sliding motions can be used for approximation of the contact shape for periodic sliding motion.

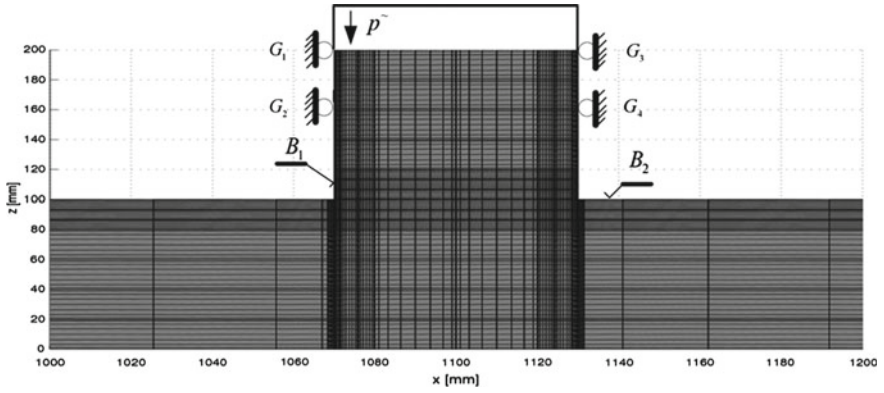



**Fig. 20** Contact between punch and strip

The wear process can be analysed with account for heat generation and thermal strain distortion effect [48]. In this case the thermal distortion does not affect the contact pressure distribution reached in a steady state, but the corresponding contact shape attained in the wear process differs essentially from that attained with neglect of heat generation [54].

### 6.3.1 Strip Monotonically Moving in the Leftward (−) or Rightward (+) Directions

Let us analyse the wear of punch (Body 1) shown in Fig. 20. The following geometric parameters are assumed: the punch height is $h = 100$ mm, its width is $L = 60$ mm, the thickness of punch and strip is $t_{th} = 10$ mm. The punch is loaded by the uniform pressure $\tilde{p} = 16.666$ MPa on the upper boundary $z = 200$ mm. The resultant vertical force is $F_0 = 10$ kN. The wear parameters are: $\tilde{\beta}_1 = 5 \times 10^{-6}$, $\tilde{\beta}_2 = 0$, $a_1 = 1$, $b_1 = 1$. The strip is moving in left (–) or in right (+) direction with the constant relative velocity $v_r = 200$ mm/s. The bodies are in the plane stress state. The coefficient of friction is $\mu = 0.25$. The upper part of punch and strip are assumed to be made of the same materials $E^{(1)} = 2 \times 10^5$ MPa, $\nu_P^{(1)} = 0.3$ and the lower punch portion of height 20 mm is characterized by the material parameters $E^{(2)} = 1.3 \times 10^5$ MPa, $\nu_P^{(2)} = 0.23$. The punch now is allowed to execute a rigid body wear velocity $\dot{\lambda}_F$ [48, 55] parallel to the axis $z$. The optimal pressure for the steady wear state is constant: $p_n^+ = p_n^- = \tilde{p}$.

The calculation of the initial gap that is the wear shape is performed by loading separately each body by the optimal contact pressure and friction stress. In this case the bodies are not allowed for the rigid body motion in the vertical direction. For monotonic sliding the equation requiring the total contact gap to vanish specifies the wear gap $g$, thus the formulae of Sect. 5.1 are used.

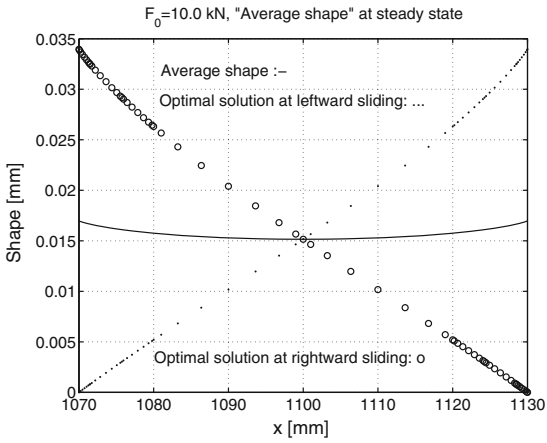



**Fig. 21** Shapes for the steady wear states induced by the strip monotonically translating in *leftward* or *rightward* directions

The steady state contact shapes can be found in Fig. 21. Note that for the leftward motion it is assumed that $g(x = 1070\,\text{mm}) = 0$, and for the rightward sliding $g(x = 1130\,\text{mm}) = 0$. Taking the average value of optimal solutions for monotonic sliding in two directions, a very good approximation of the shape of the periodic sliding of the strip is obtained [55].

### 6.3.2 Strip Sliding Periodically in the Leftward and Rightward Direction

The wear process is induced by the reciprocal strip translation. It is assumed that only the punch undergoes wear (see Fig. 19). The wear parameters are: $\tilde{\beta}_1 = 1.25\pi \times 10^{-8}$, $\tilde{\beta}_2 = 0$, $a_1 = 1$, $b_1 = 1$, the coefficient of friction is $\mu = 0.25$. The displacement of $B_2$ is: $u = -u_0 \cos \omega\tau$, where $u_0 = 1.5\,\text{mm}$, $\omega = 10\,\text{rad/s}$, $\tau$ denotes the time. It is supposed in the steady state wear analysis that the contact pressure distribution is fixed during semi-cycle and varies discontinuously during sliding reversal in the consecutive semi-cycle. The contact pressures are denotes by $p_n^+$, $p_n^-$, respectively.

Using the wear rule (18) the wear dissipation work in one cycle of sliding motion has the following form for $\tilde{\beta}_1 \neq 0$, $\tilde{\beta}_2 = 0$:

$$E_w = \int_0^{T_*} D_w \mathrm{d}\tau = \int_0^{T_*/2} \int_{S_c^{(1)}} p_n^+ \dot{w}_{1,n}^+ \mathrm{d}S\,\mathrm{d}\tau + \int_{T_*/2}^{T_*} \int_{S_c^{(1)}} p_n^- \dot{w}_{1,n}^- \mathrm{d}S\,\mathrm{d}\tau, \qquad (29)$$

and for $a_1 = b_1 = 1$, there is

$$\frac{E_w}{2u_0\tilde{\beta}_1} = \int\limits_{S_c^{(1)}} \left\{ \left(p_n^+\right)^2 + \left(p_n^-\right)^2 \right\} \mathrm{d}S = \frac{E_w^+}{2u_0\tilde{\beta}_1} + \frac{E_w^-}{2u_0\tilde{\beta}_1}. \tag{30}$$

In the steady wear state $E_w$ reaches the minimum value. We note that $p_n^+$ and $p_n^-$ are not uniformly distributed at the contact interface. During the steady periodic response the wear increment accumulated during one cycle should be compatible at each point $x \in S_c$ with the rigid body punch motion. The total accumulated wear increment must be constant along the contact domain. The normal wear increment in one period equals

$$\Delta w_{1,n} = \tilde{\beta}_1 \left[ p_n^{+b} + p_n^{-b} \right] (u_0\omega)^{a_1} \int\limits_0^{T_*/2} |\sin \omega\tau|^{a_1} \, \mathrm{d}\tau,$$

which for $a_1 = b_1 = 1$ is $\Delta w_{1,n} = \tilde{\beta}_1[p_n^+ + p_n^-]2u_0$. The averaged normal wear rate in one period is equal to

$$\dot{\bar{w}}_{1,n} = \frac{\Delta w_{1.n}}{T_*} = \frac{\tilde{\beta}_1[p_n^+ + p_n^-]}{T_*} 2u_0, \tag{31}$$

where $T_*$ is the time period.

To assure the uniform wear increment accumulated during full cycle at each point of the contact zone, the following condition should be satisfied

$$p_n^+ + p_n^- = 2p_m = \text{const}. \tag{32}$$

In our case, $p_m = \tilde{p}$.

According to the Signorini contact condition in the normal direction the contact pressure must be positive in the contact zone and distance after deformation between the bodies is positive:

$$d_n^{\pm} = u_n^{(2)\pm} - u_n^{(1)\pm} + g_n \geq 0, \tag{33}$$

where $u_n^{(i)} = \mathbf{u}^{(i)} \cdot \mathbf{n}_c$ is the normal displacement of the $i$th body, $g_n$ is the initial gap (shape of Body 1 in the steady state). This shape is not specified and must be found in the optimization process. The Signorini conditions for the whole period have the form

$$p_n^{\pm} d_n^{\pm} = 0, \quad p_n^{\pm} \geq 0, \; d_n^{\pm} \geq 0. \tag{34}$$

In this case, the optimization problem is

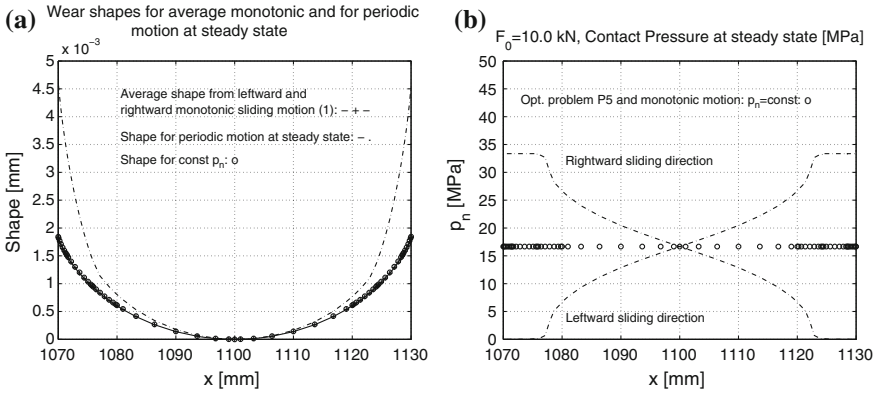



**Fig. 22** Results for periodic sliding; **a** shapes from static solution for the constant pressure distribution without friction (solution of Problem 5): o, the average shape from *leftward* and *rightward* monotonic motion: +, shape from periodic sliding with time integration of the Archard wear rate: –, **b** contact pressure distribution in the steady wear state: for Problem 5: o, for periodic sliding: "–"

**Problem 14**

$$\min_{g_n}\left\{\int_{S_c}\frac{1}{2}(p_n^+ + p_n^- - 2p_m)^2\,\mathrm{d}S \mid p_n^{\pm}\geq 0,\ d_n^{\pm}\geq 0,\ p_n^{\pm}d_n^{\pm}=0,\right.$$

$$\left.\text{Equilibrium equations for the punch}\right\}, \quad (35)$$

where the minimum of (35) provides the contact pressure distribution satisfying (32) and contact conditions (34). Solving first the minimization problem (35), next the initial gap (contact shape of the body 1) will be found (Fig. 22).

## 7 Concluding Remarks

In this paper, we have discussed some optimization problems assuming the prescribed form of contact pressure in the subdomain of contact zone including the optimization problems associated with the wear process. Parameters of the contact pressure distribution were assumed to be fixed, or some of them were determined in the optimization process.

Using the prescribed form of the pressure distribution, a smooth stress state in the contact domain can be specified. A special iteration method was applied for solving optimization problems. In calculations, the $p$-version of finite element method was used. The generalized friction dissipation power and generalized wear dissipation

power provides both regular and singular regimes of the wear and stress states. Also, the characters of singular solutions are discussed.

The steady wear state is reached at minimum of the generalized wear dissipation power for the control parameter $q = 1$. The numerical examples are presented for axisymmetric bodies and plane structures. Problem 4 for a cylindrical punch was discussed, for Problems 7 and 9 the maximal value of torque and minimal value of the friction dissipation on a toroidal contact surface were determined. In these examples, we applied the control function (3)–(6) with fixed parameters (in Problem 4) and with calculated values $L_1$ or $L_4$ for Problems 7 and 9, respectively.

Using the generalized friction dissipation power and wear dissipation power, we can determine also the torque for Problem 11 but its value is smaller than that for Problem 7. Preserving the stress constraint set in Problem 13, the maximum value of loading pressure $\tilde{p}$ can be specified. In this way, the admissible load level of system ensuring the stress constraint $\sigma_{eq} \leq \sigma_u$ in the steady wear state can be determined.

**Acknowledgments** The present research was partially supported by the Hungarian Academy of Sciences, by grants OTKA K67825 and K115701, within the program TÁMOP 4.2.1.B-10/2/KONV-2010-0001.

## References

1. N.V. Banichuk, P.J. Neittaanmäki, *Structural Optimization with Uncertainties*. Solid Mechanics and Its Applications, vol. 162 (Springer, Berlin, 2010)
2. N.V. Banichuk, *Introduction to Optimization of Structures* (Springer, London, 2011). Reprint of the original 1st edn. (1990)
3. Z. Mróz, Optimal design of structures of composite materials. Int. J. Solids Struct. **6**(7), 859–870 (1970)
4. Z. Mróz, Design sensitivity of critical loads and vibration frequencies of nonlinear structures, in *Optimization of Large Structural Systems*, vol. I, of NATO ASI Series, Series E: Applied Sciences, vol. 321, ed. by G.I.N. Rozvany (Kluwer Academic Publishers, Dordrecht, 1993), pp. 455–476
5. Z. Mróz, D. Bojczuk, Shape and topology sensitivity analysis and its application to structural design. Arch. Appl. Mech. **82**(10–11), 1541–1555 (2012)
6. Z. Mróz, R.T. Haftka, Design sensitivity analysis of non-linear structures in regular and critical states. Int. J. Solids Struct. **31**(15), 2071–2098 (1994)
7. Z. Mróz, T. Lekszycki, Optimal support reaction in elastic frame structures. Comput. Struct. **14**(3–4), 179–185 (1981)
8. J. Haslinger, P.P. Neittaanmäki, *Finite Element Approximation for Optimal Shape Design: Theory and Applications* (Wiley, London, 1988)
9. R.L. Benedict, J.E. Taylor, Optimal design for elastic bodies in contact, in *Optimization of Distributed Parameter Structures*, vol. II, ed. by E.J. Haug, J. Cea (Sijthoff and Noordhoff, Alphen aan den Rijn, 1981), pp. 1553–1599
10. N. Kikuchi, J.E. Taylor, Shape optimization for unilateral elastic contact problems, in *Numerical Methods in Coupled Problems* (University College, Swansea, 1981), pp. 430–441
11. A. Klarbring, On the problem of optimizing contact force distributions. J. Optim. Theory Appl. **74**(1), 131–150 (1992)
12. A. Klarbring, J. Haslinger, On almost constant contact stress distributions by shape optimization. Struct. Multidisc. Optim. **5**(4), 213–216 (1993)

13. T.F. Conry, A. Seireg, A mathematical programming method for design of elastic bodies in contact. J. Appl. Mech. **38**(2), 387–392 (1971)
14. J. Oda, J. Sakamoto, K. San, A method for producing a uniform contact stress distribution in composite bodies with interface. Struct. Multidisc. Optim. **3**(1), 23–28 (1991)
15. I. Páczelt, B. Herpai, Some remarks on the solution of contact problems of elastic shells. Archiwum Budowy Maszyn **24**, 197–202 (1977)
16. J. Petersson, Behaviourally constrained contact force optimization. Struct. Multidisc. Optim. **9**(3–4), 189–193 (1995)
17. I. Páczelt, Some new developments in contact pressure optimization. Eng. Trans. **43**(1–2), 297–312 (1995)
18. I. Páczelt, Iterative methods for solution of contact optimization problems. Arch. Mech. **52**(4–5), 685–711 (2000)
19. I. Páczelt, A. Baksa, Examination of contact optimization and wearing problems. J. Comput. Appl. Mech. **3**(1), 61–84 (2002)
20. I. Páczelt, T. Szabó, Optimal shape design for contact problems. Struct. Multidisc. Optim. **7**(1–2), 66–75 (1994)
21. B. Szabó, I. Babuska, *Finite Element Analysis* (Wiley, New York, 1991)
22. D. Hilding, A. Klarbring, J. Petersson, Optimization of structures in unilateral contact. Appl. Mech. Rev. **52**(4), 139–160 (1999)
23. P. Beremlijski, J. Haslinger, M. Kočvara, R. Kučera, J.V. Outrata, Shape optimization in three-dimensional contact problems with coulomb friction. SIAM J. Optim. **20**(1), 416–444 (2009)
24. A. Czekanski, S.A. Meguid, Solution of dynamic frictional contact problems using nondifferentiable optimization. Internat. J. Mech. Sci. **43**(6), 1369–1386 (2001)
25. D. Hilding, The equilibrium state of a structure subject to frictional contact. Eur. J. Mech. A Solids **19**(6), 1029–1040 (2000)
26. D. Hilding, A. Klarbring, Optimization of structures in frictional contact. Comput. Methods Appl. Mech. Eng. **205–208**, 83–90 (2012)
27. Y. Kanno, J.A.C. Martins, Arc-length method for frictional contact problems using mathematical programming with complementarity constraints. J. Optim. Theory Appl. **131**(1), 89–113 (2006)
28. N.H. Kim, K.K. Choi, J.S. Chen, Shape design sensitivity analysis and optimization of elasto-plasticity with frictional contact. AIAA J. **38**(9), 1742–1753 (2000)
29. A. Klarbring, A. Mikelič, M. Shillor, Optimal shape design in contact problems with normal compliance and friction. Appl. Math. Lett. **5**(2), 51–55 (1992)
30. G.E. Stavroulakis, Optimal prestress of structures with frictional unilateral contact interfaces. Arch. Appl. Mech. **66**(1–2), 71–81 (1995)
31. W. Li, Q. Li, G.P. Steven, Y.M. Xie, An evolutionary approach to elastic contact optimization of frame structures. Finite Elem. Anal. Des. **40**(1), 61–81 (2003)
32. A. Myslinski, Level set method for optimization of contact problems. Eng. Anal. Bound. Elem. **32**(11), 986–994 (2008)
33. W. Li, Q. Li, G.P. Steven, Y.M. Xie, An evolutionary shape optimization for elastic contact problems subject to multiple load cases. Comput. Methods Appl. Mech. Eng. **194**(30–33), 3394–3415 (2005)
34. N.V. Banichuk, S.Y. Ivanova, Shape optimization in contact problems of the theory of elasticity with incomplete external loading data. J. Appl. Math. Mech. **73**(6), 696–704 (2009)
35. N.V. Banichuk, Optimization of the contact pressure in the problem of the interaction of a punch and an elastic medium. J. Appl. Math. Mech. **74**(3), 334–340 (2010)
36. B. Desmorat, Structural rigidity optimization with frictionless unilateral contact. Int. J. Solids Struct. **44**(3–4), 1132–1144 (2007)
37. D. Hilding, B. Torstenfelt, A. Klarbring, A computational methodology for shape optimization of structures in frictionless contact. Comput. Methods Appl. Mech. Eng. **190**(31), 4043–4060 (2001)
38. F.F. Mahmoud, A.A. El-Sharkawy, K.M. Hassan, Contour design for contact stress minimization by interior penalty method. Appl. Math. Model. **13**(10), 596–600 (1989)

39. Y. Tada, S. Nishihara, Optimum shape design of contact surface with finite element method. Adv. Eng. Softw. **18**(2), 75–85 (1993)
40. S. Šimunović, S. Saigal, Contact surface optimization using boundary element method. Comput. Struct. **56**(5), 745–750 (1995)
41. H. Ou, B. Lu, Z.S. Cui, C. Lin, A direct shape optimization approach for contact problems with boundary stress concentration. J. Mech. Sci. Tech. **27**(9), 2751–2759 (2013)
42. M. Ignesti, A. Innocenti, L. Marini, E. Meli, A. Rindi, P. Toni, Wheel profile optimization on railway vehicles from the wear viewpoint. Int. J. Non-Linear Mech. **53**, 41–54 (2013)
43. R. Smallwood, J.C. Sinclair, K.J. Sawley, An optimization technique to minimize rail contact stresses. Wear **144**(1–2), 373–384 (1991)
44. S. Zakharov, I. Goryacheva, V. Bogdanov, D. Pogorelov, I. Zharov, V. Yazykov, E. Torskaya, S. Soshenkov, Problems with wheel and rail profiles selection and optimization. Wear **265**(9–10), 1266–1272 (2008)
45. P.C. Sui, S. Anderle, Optimization of contact pressure profile for performance improvement of a rotary elastomeric seal operating in abrasive drilling environment. Wear **271**(9–10), 2466–2470 (2011)
46. I. Páczelt, Z. Mróz, Optimal shapes of contact interfaces due to sliding wear in the steady relative motion. Int. J. Solids Struct. **44**(3–4), 895–925 (2007)
47. I. Páczelt, Z. Mróz, On the analysis of steady-state sliding wear processes. Trib. Int. **42**(2), 275–283 (2009)
48. I. Páczelt, Z. Mróz, Numerical analysis of steady thermo-elastic wear regimes induced by translating and rotating punches. Comput. Struct. **89**(23–24), 2495–2521 (2011)
49. I.G. Goryacheva, *Contact Mechanics in Tribology* (Kluwer Academic Publishers, Dordrecht, 1998)
50. I. Páczelt, Some optimization problems of contact bodies within the linear theory of elasticity, in *Variational methods in the Mechanics of Solids* (Evanston, IL, 1978), ed. by S. Nemat-Nasser (Pergamon Press, Oxford, 1980), pp. 349–356
51. I. Páczelt, A. Baksa, T. Szabó, Product design using a contact-optimization technique. Strojniški Vestnik J. Mech. Eng. **53**(7–8), 442–461 (2007)
52. J.J. Kalker, *Three-dimensional Elastic Bodies in Rolling Contact* (Kluwer Academic Publishers, Dordrecht, 1990)
53. I. Páczelt, Z. Mróz, On steady wear states for monotonic relative sliding of contacting bodies. Key Eng. Mater. **618**, 49–71 (2014)
54. I. Páczelt, Z. Mróz, Analysis of thermo-mechanical wear problems for reciprocal punch sliding. Adv. Eng. Softw. **80**, 139–155 (2015)
55. I. Páczelt, Z. Mróz, Solution of wear problems for monotonic and periodic sliding with p-version of finite element method. Comput. Methods Appl. Mech. Eng. **249–252**, 75–103 (2012)

# Some Problems of Multipurpose Optimization for Deformed Bodies and Structures

**Alexander Sinitsin, Svetlana Ivanova, Evgeniy Makeev and Nikolay Banichuk**

**Abstract** Some problems of multipurpose analysis and optimization of deformed structures and thin-walled structural elements are studied in this paper under some constraints including incomplete data. The first problem is the multipurpose optimization of layered plate made from given set of materials in context of optimization of ballistic limit velocity. Incomplete data concerning the thickness of layers of optimized multilayered shield structure are taken into account. The Pareto-approach and numerical evolutionary method (genetic algorithm) are used for solving of the considered multipurpose problem. The second problem studied in the paper is the shape optimization problem for rigid punch moving on the surface of elastic half-space, which is solved analytically in multipurpose formulation taking into account friction of contacted surfaces, wear of materials and arising pressure distributions. The relative movement is considered in frame of quasi-static formulation. Formulated optimization problem is studied analytically using the developed decomposition approach and exact solutions are obtained for the punch which has a rectangular contact region and moves translationally with a constant velocity.

**Keywords** Shape optimization · Multilayered shield · Penetration · Rigid strikers

**Mathematical Subject Classification:** 65K10

A. Sinitsin (✉) · S. Ivanova · E. Makeev · N. Banichuk
A. Ishlinsky Institute for Problems in Mechanics of the Russian Academy of Sciences, Prospekt Vernadskogo 101, 119526 Moscow, Russia
e-mail: a@sinitsyn.info

S. Ivanova
e-mail: syuivanova@yandex.ru

E. Makeev
e-mail: makeevev@rambler.ru

N. Banichuk
e-mail: banichuk@ipmnet.ru



P. Neittaanmäki et al. (eds.), *Mathematical Modeling and Optimization of Complex Structures*, Computational Methods in Applied Sciences 40,
DOI 10.1007/978-3-319-23564-6_17

## 1 Introduction

Multipurpose optimization in mechanics and optimal design of structures and structural elements plays an important role in modern problems of engineering. There are two basic approaches for solving of multipurpose problems: the Pareto-approach and the Nash-approach [1, 8].

According to Pareto-approach that is used in our considerations the minimization (or maximization) of vector functional

$$J(h) = \{J_1(h), \ldots, J_i(h), \ldots, J_N(h)\}^T \rightarrow \min_{h \in \Lambda_h} \tag{1}$$

must be performed on the set $\Lambda_h$ of admissible design variables $h \in \Lambda_h$. The components of this vector functional are the treated optimization criteria $J_1(h)$, …, $J_i(h)$, …, $J_N(h)$. Minimum in (1) is considered in the Pareto sense, i.e.

$$h^* = \arg \min_{h \in \Lambda_h} J(h) \tag{2}$$

is the solution of the problem (1) and there does not exist any other solution $\tilde{h} \in \Lambda_h$ for which

$$J_i(\tilde{h}) \leq J_i(h^*) \tag{3}$$

and at least for one component $J_s$ the strong inequality

$$J_s(\tilde{h}) < J_s(h^*) \tag{4}$$

is valid. For finding the optimal solution $h^*$ the minimization of an objective weighting functional or a preference functional

$$J_C(h) = \sum_{i=1}^{N} C_i J_i(h) \rightarrow \min_{h \in \Lambda_h} \tag{5}$$

can be realized where $C_i \geq 0$, $i = 1, \ldots, N$ and $\sum_{i=1}^{N} C_i = 1$. For any set of coefficients (factors) $C_i$ there is one optimal solution $h^*$, i.e. one point in the space of functionals $J_1, \ldots, J_N$. All such points create the Pareto-front.

For comparison, the other approach (the so-called Nash-approach) has a game character. According to this approach there are two optimality criteria

$$J_1(h_1, h_2), \quad J_2(h_1, h_2), \quad h_1 \in \Lambda_{h_1}, \quad h_2 \in \Lambda_{h_2} \tag{6}$$

and two "players"—two design variables $h_1$ and $h_2$. The sequence of Nash-minimization may be described as follows:

Step 1: Suppose that the first approximation of optimal solution $h_2^*$ is given.

Step 2: The first criterion is minimized by the first "player" with given $h_2^*$

$$J_1(h_1, h_2^*) \to \min_{h_1 \in \Lambda_{h_1}} \tag{7}$$

and the first approximation of optimal solution $h_1^*$ is found:

$$h_1^* = \arg \min_{h_1 \in \Lambda_{h_1}} J_1(h_1, h_2^*). \tag{8}$$

Step 3: The second criterion is minimized by the second "player" with given $h_1^*$

$$J_2(h_1^*, h_2) \to \min_{h_2 \in \Lambda_{h_2}} \tag{9}$$

and the second approximation of optimal solution $h_2^*$ is found:

$$h_2^* = \arg \min_{h_2 \in \Lambda_{h_2}} J_2(h_1^\star, h_2). \tag{10}$$

Step 4: Return to the Step 2 or stop the process of optimization.

The solution of the Nash-optimization problem defines in the space of functionals some equilibrium point.

Further we consider in the paper the multipurpose optimization problem for multilayered plate structure in context of high-speed striker penetration and the shape multipurpose optimization problem for the rigid punch moving on the surface of elastic half-space.

## 2 Multipurpose Optimization of Multilayered Shield Structure Under Penetration of High-Speed Rigid Strikers

The multilayered structure is considered that is composed of $n$ homogeneous layers made from $r_m$ different materials ($n, r_m$ are given). It is supposed that the thicknesses of these layers are not known apriori but the sum $L$ of them (the total thickness of the shield structure) is given. The material of each layer with number $i$ is characterized by two constant values: dynamical rigidity $A_0^i$ and density $A_2^i$. The quantities of the structure are described by two piecewise functions $A_0(x)$, $A_2(x)$ and for convenience the natural parameterization of the material quantities is applied by using of piecewise function

$$t = t(x), \quad x \in [0, L], \quad t \in \{t_i = i\}, \tag{11}$$

i.e. $A_0^{i+1}, \ A_2^{i+1} \in \{(A_0)_s, (A_2)_s\}, \ \ s = 1, 2, \ldots, r_m$.

For high-speed penetration of rigid axisymmetric strikers the resistance force $D(x)$ of the layered media (shield) is connected with parameters of layer materials

$A_0^i$, $A_2^i$, their thicknesses $\Delta_i$ and the shape $y = y(x)$ of the striker of the length $l$ as [3, 4]

$$D(x) = D_{nose}(x) + D_{lat}(x) = B_0(x) + B_2(x)v^2, \tag{12}$$

$$B_0(x) = \pi r^2 A_0(x) - 2\pi \int_{x_*}^{x_{**}} A_0(\eta) y y_\eta d\eta, \tag{13}$$

$$B_2(x) = \pi r^2 A_2(x) - 2\pi \int_{x_*}^{x_{**}} \frac{A_2(\eta) y y_\eta^3}{1 + y_\eta^2} d\eta, \tag{14}$$

where $D_{nose}(x)$ and $D_{lat}(x)$ are the parts of the resistance force acting on the nose surface and the lateral surface of the striker, $r$ is the radius of truncated nose part of striker, $0 \le x \le L + l$ and $x_*$, $x_{**}$ define the size of the region of interaction of the striker with the shield, $y_\eta = dy/d\eta$. Taking into account the relations (12)–(14) we can describe the penetration process by the ordinarily differential equation

$$M_0 v \frac{dv}{dx} = -D, \quad 0 < x < L \tag{15}$$

with the boundary condition at $x = 0$

$$(v)_{x=0} = v_{imp}. \tag{16}$$

Here $v(0) = v_0 = v_{imp} > 0$ is the striker velocity at the moment of impact, $M_0$ is its mass and $v = v(x)$ is the velocity distribution. The initial (impact) velocity when

$$\begin{aligned} v(x) > 0, \quad x \in [0, L + l), \\ (v)_{x=L+l} = 0 \end{aligned} \tag{17}$$

is called "the ballistic limit velocity", i.e.

$$v(0) = v_{imp} = v_{BLV}. \tag{18}$$

This parameter is treated here as the general characteristic of shield structure. For convenience we introduce the new variable $\xi$ as

$$\xi = L + l - x, \quad d\xi = -dx \tag{19}$$

and transform the problem (15) with the unknown apriori initial condition (16) into the problem of finding of striker velocity distribution $v = v(\xi)$ for given distribution of materials in the layered structure

$$\frac{dv^2}{d\xi} = \beta(\alpha + v^2) \tag{20}$$

with the given initial condition

$$\left(v^2\right)_{\xi=0} = 0, \tag{21}$$

where

$$\alpha = \frac{B_o}{B_2}, \quad \beta = \frac{2B_2}{M_0}. \tag{22}$$

Two optimality criteria are chosen for multipurpose optimization in Pareto-sense of considered multilayered shield structure. The first is maximum of ballistic limit velocity $v_{BLV}$ or minimum of the quantity $(-v_{BLV})$, that is the functional

$$J_v = -v_{BLV}(t, \Delta) \tag{23}$$

is minimized and $t$ is the piecewise function (11), $\Delta = (\Delta_1, \Delta_2, \ldots, \Delta_n)$. The second minimizing functional is the mass of the structure

$$J_m = S_{surf} \int_0^L A_2(t(x))dx, \tag{24}$$

where $S_{surf}$ is the area of the layer surface (for example, the unit of the area). Note that the functional (24) can characterize, for example, the cost (or other mass characteristic) of the structure with the correspond function $A_2(t(x))$. The total thickness of the plate was given, but thicknesses of separate layers were unknown as well as their positions (order) in the structure. Taking into account the scalar functionals (23), (24) we can formulate the multipurpose minimization problem of the vector functional [1, 8]

$$J^* = J(h^*) = \min_{h \in \Lambda_h} \left\{ \begin{matrix} J_v \\ J_m \end{matrix} \right\}, \tag{25}$$

where $h = (t, \Delta) = (t, \Delta_1, \Delta_2, \ldots, \Delta_n)$ and $\Lambda_h$ is the given set of admissible design variables. Note that operation of minimization in (25) is considered in Pareto sense (2)–(4).

For investigation of the multipurpose problem formulated above let us apply the method of minimization of objective weighting functional or preference functional

$$J_C = C_v J_v + C_m J_m \tag{26}$$

under following conditions supposed on the weighting factors $C_v$ and $C_m$:

$$1 \geq C_v \geq 0, \quad 1 \geq C_m \geq 0, \tag{27}$$

$$C_v + C_m = 1. \tag{28}$$

Thus, the problem of minimization of the vector functional (25) is transformed to the problem of minimization of scalar functional (26) with the conditions (27), (28), that is

$$J_C^* = J_C(h^*) = \min_{h \in \Lambda_h} J_C(h). \tag{29}$$

For solving the optimization problem (29) with incomplete data about the layer thicknesses the evolutionary computational method known as genetic algorithm (GA) [5] is applied for different values of problem parameters. It is supposed that the structure is composed of $n$ thin layers, so that

$$\Delta_1 = \Delta_2 = \cdots = \Delta_n = \frac{L}{n} \quad (n = 20), \tag{30}$$

**Table 1** Materials properties

| Materials | s | $(A_0)_s$, N/m$^2$ | $(A_2)_s$, kg/m$^3$ |
|---|---|---|---|
| Aluminum | 1 | $350 \times 10^6$ | 2765 |
| Soft steel | 2 | $1850 \times 10^6$ | 7830 |
| Copper | 3 | $910 \times 10^6$ | 8920 |
| Duraluminum | 4 | $1330 \times 10^6$ | 2765 |

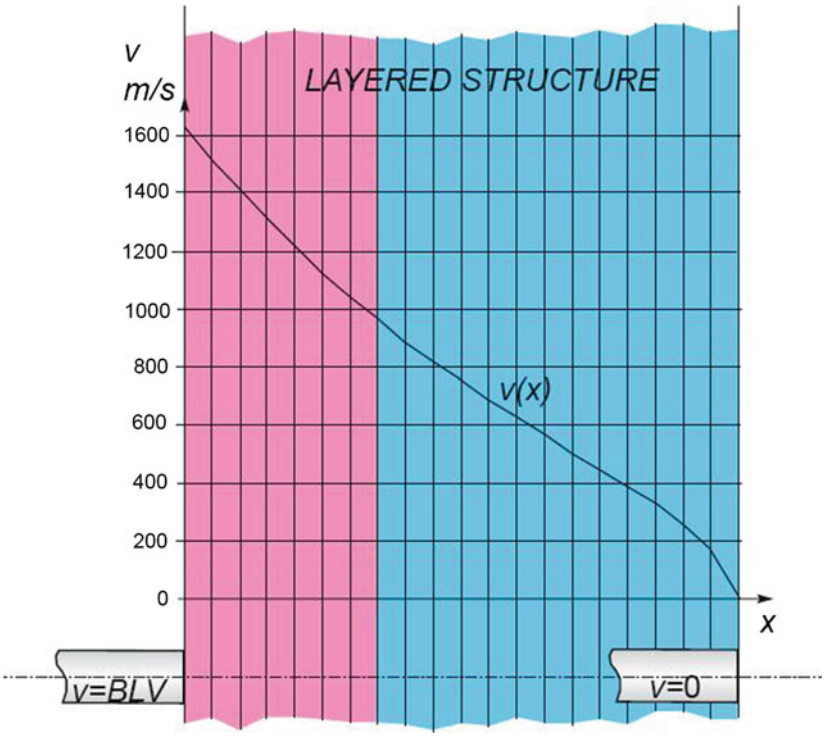



**Fig. 1** Optimal layered plate against penetration of cylindrical striker for $C_m = 0$

each of one can be occupied by one of the possible materials. The minimized functional (26) can be represented as

$$J_C = C_v J_v + C_m J_m = (1 - C_m) J_v + C_m J_m \to \min_{t \in \Lambda_t}, \tag{31}$$

where $0 \leq C_m \leq 1$ and $\Lambda_t$ is set of admissible piecewise $t$. There were 10 individuals (strings or layer distributions) in each generation, the number of generations was equal to 500 for 10 different initializations of the finding process. Four materials are represented in Table 1 as possible for optimized structure [3, 4].

In Fig. 1 the optimal distribution of materials (red—copper, blue—steel) within layered structure ($L = 0.1\,\text{m}$, $S_{surf} = 1\,\text{m}^2$) and the striker velocity decreasing for cylinder striker ($R = 0.005\,\text{m}$, $M_0 = 0.05\,\text{kg}$) are shown for arbitrary mass of the structure ($C_m = 0$).

The material distributions for different values of objective weighing factor $C_m$ of the mass criterion for the case of cylindrical striker are given in Fig. 2. Dark regions in Fig. 2 denote copper layers and gray regions—steel layers. The factor $C_m$ has the following values: (1) 0–0.02; (2) 0.3; (3) 0.4; (4) 0.5; (5) 0.55.

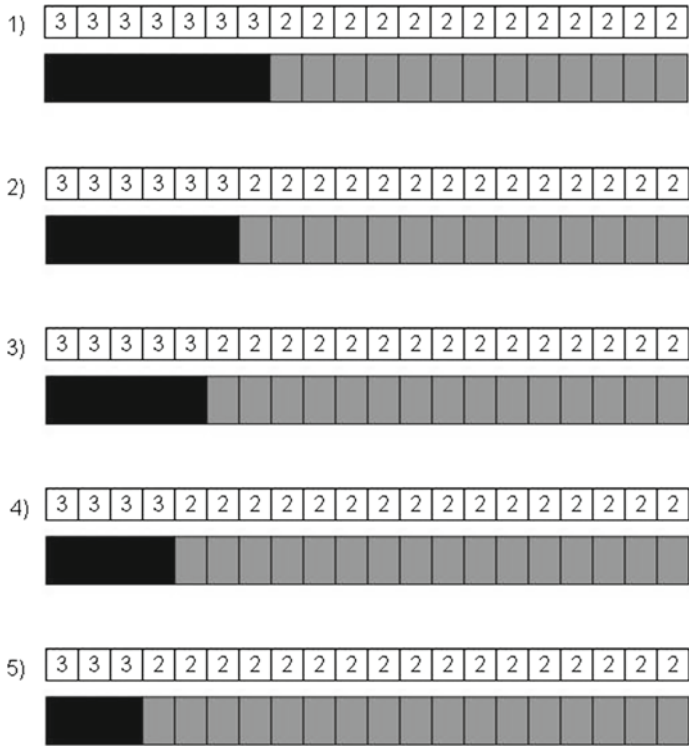



**Fig. 2** Optimal distributions of materials against penetration of cylindrical striker

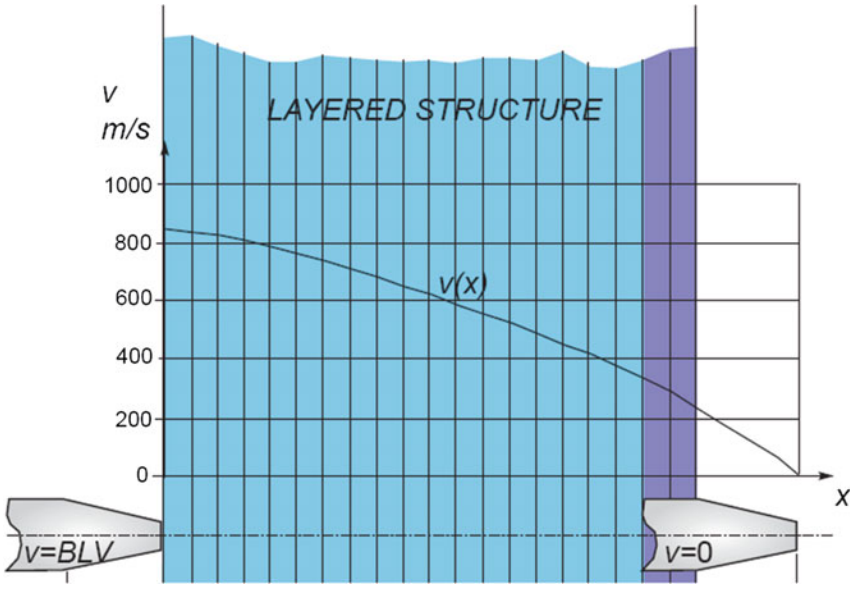



**Fig. 3** Optimal layered plate for truncated conical striker, $C_m = 0.25$

In Fig. 3 the optimal distribution of materials (blue—steel, violet—duraluminum) within layered structure ($L = 0.1$ m) and the striker velocity decreasing for truncated conical striker ($R = 0.005$ m, $M_0 = 0.05$ kg, $r = 0.002$ m, $l = 0.02$ m) are shown for $C_m = 0$.

Figure 4 shows the monotonic dependence of objective weighting functional $J_C$ on the weighting factor $C_m$ for cylindrical (dashed line), truncated conical (thin solid line) and conical (thick solid line) strikers.

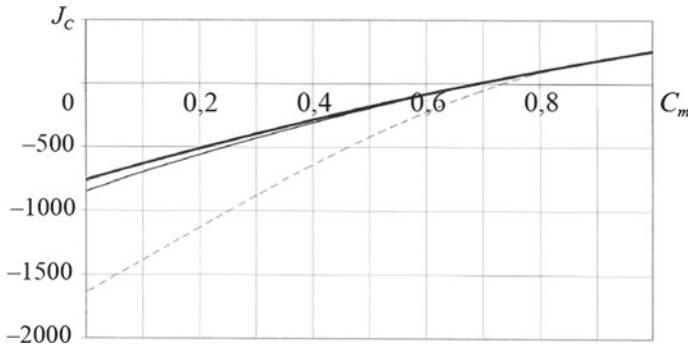



**Fig. 4** Monotonic dependence of objective weighting functional $J_C$ on the factor $C_m$

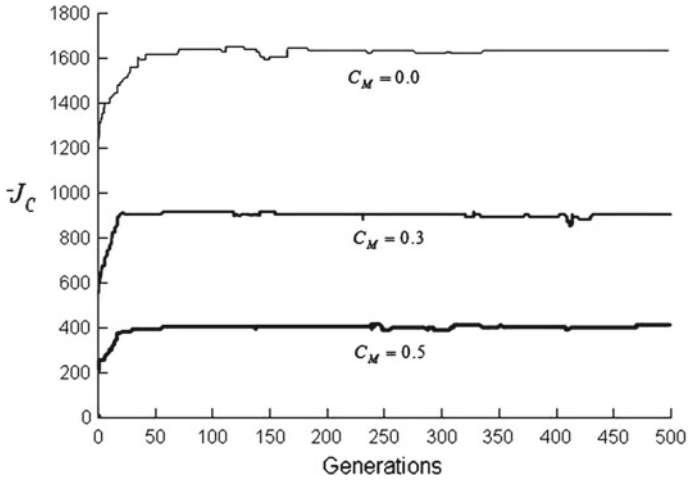



**Fig. 5** Convergence of the genetic algorithm

Convergence of the genetic algorithm is illustrated in Fig. 5 by the dependence of the preference functional value on the number of generation for cylindrical striker and different values of factor $C_m$.

It should be noted that all results are received without taking into account the processes of striker entrance and exit (as well as the layer boundaries influence on this process).

## 3 Multipurpose Optimization of The Rigid Punch Shape

Let us consider the interaction of a rigid punch with the elastic half-space ($z \geq 0$) taking into account friction and wear. The system of coordinates $Oxyz$ is fixed with the punch, which moves in $x$-direction with the constant relative velocity $V$. The relative movement is considered in frame of quasi-static formulation. It is supposed that the rectangular contact region $\Omega_f$ is symmetric with respect to the axes $Ox$, $Oy$: $\Omega_f = [-a \leq x \leq a,\ -b \leq y \leq b]$, where $a$, $b$ are the given constants. The normal stresses $\sigma_{zz}$ and the tangential stresses $\sigma_{xz}$

$$\sigma_{zz}(x, y, 0) = -p(x, y), \quad p(x, y) \geq 0,$$
$$\sigma_{xz} = \tau_0 + \mu p(x, y) \tag{32}$$

are acting in the contact region $\Omega_f$, where $p \geq 0$ is the contact pressure and $\tau_0$, $\mu$ are the friction constants [6]. The boundary conditions for the theory of elasticity

problem in the half-space $z \geq 0$ are written in the form

$$w = f(x, y), \quad \sigma_{xz} = \tau_0 - \mu\sigma_{zz}, \quad \sigma_{yz} = 0 \quad (x, y) \in \Omega_f,$$
$$\sigma_{zz} = 0, \quad \sigma_{xz} = 0, \quad \sigma_{yz} = 0, \quad (x, y) \in \Omega_0. \tag{33}$$

Here $w = w(x, y)$ is the $z$-component of the vector of elastic displacements, $f(x, y)$ denotes the punch shape contacting with the elastic medium and $\Omega_0$—the region of elastic half-space surface free from loads.

For the known distribution $p(x, y)$, $(x, y) \in \Omega_f$ the resulting force $P$ and the total moment $M$ with respect to the $y$-axis which acted on the punch are defined as

$$P = J_P = \int_{\Omega_f} p d\Omega_f, \quad M = J_M = \int_{\Omega_f} xp d\Omega_f. \tag{34}$$

The punch shape $f = f(x, y)$ is accepted as a desirable design variable. The friction dissipation power

$$J_F(f) = \int_{\Omega_f} (\tau_0 + \mu p(f))V d\Omega_f, \tag{35}$$

the wear volume rate

$$J_W(f) = \int_{\Omega_f} K_W p^n(f) V^m d\Omega_f, \tag{36}$$

and the discrepancy functional

$$J_D(f) = \int_{\Omega_f} (p(f) - p_g)^2 d\Omega_f \tag{37}$$

are taken into account as the components of the minimized vector-functional $J$

$$J(f) = \{J_F(f), \ J_W(f), \ J_D(f)\}^T \to \min_f. \tag{38}$$

Here $K_W$ is the wear constant, $n > 1$, $m > 0$ are given parameters of contacting materials [6] and $p_g = p_g(x, y)$ is given objective function, $T$ is the transposition operator.

The optimization problem is to find the optimal punch shape $f(x, y)$ that minimizes the vector-functional (38) under the constraints

$$J_P = P^*, \quad J_M = M^*, \tag{39}$$

where $P^* \geq 0$ and $M^* \geq 0$ are the given values. We suppose that the value of the total force $P^*$ is sufficiently large in comparison with the value of the moment $M^*$, i.e. the condition $p(x, y) \geq 0$, $(x, y) \in \Omega_f$ is satisfied [2].

Formulated optimization problem can be decomposed to two separately solved problems [2]. The first problem consists in finding of the optimal pressure distribution $p_* \in \Omega_f$ that minimizes the vector-functional

$$J(p) = \{J_F(p),\ J_W(p),\ J_D(p)\}^T \to \min_{p \in \Lambda_p}, \tag{40}$$

$$\Lambda_p = \{p \geq 0,\ J_P(p) = P^*,\ J_M(p) = M^*\}. \tag{41}$$

The second problem is to find the normal displacements $w_*(x, y)$ for the optimal pressure distribution $p_*(x, y)$ and corresponding friction forces, i.e. to define the optimal punch shape $f_*(x, y) = w_*(x, y)$.

For solving of the multi-criteria Problems (40)–(41) (Pareto-optimization) the scalar linear weight-functional $J_C$ is constructed and minimized (see, for example, [1, 8]):

$$J_C(p) = C_F J_F(p) + C_W J_W(p) + C_D J_D(p) \to \min_{p \in \Lambda_p}, \tag{42}$$

$$C_F \geq 0,\ C_W \geq 0,\ C_D \geq 0,\ \ C_F + C_W + C_D = 1.$$

As a result the optimal pressure distribution $p_*(x, y)$ depending on the weight coefficients $C_F$, $C_W$, $C_D$ can be defined with application of well known technique of Lagrange multipliers. The necessary condition of optimality of the functional (42) under the constraints (39) can be written as

$$\delta J_C^L(p) = \delta\{J_C(p) - \lambda_P J_P(p) - \lambda_M J_M(p)\}$$
$$= \int_{\Omega_f} \{C_F \mu V + n C_W K_W p^{n-1} V^m + 2C_D(p - p_g) - \lambda_P - \lambda_M x\}\delta p d\Omega_f = 0, \tag{43}$$

where $\lambda_P$ and $\lambda_M$ are the Lagrange multipliers defined from the conditions (39). We assume for simplicity $n = 2$, $m = 1$, $p_g = P^*/S$, where $S$ is the area of contact region $\Omega_f$. If $\mu$ is the constant value and the region $\Omega_f$ is symmetric with respect to the axes $Ox$, $Oy$, the optimal pressure distribution $p_*$ can be defined as

$$p_* = \frac{P^\star}{S} + \frac{M^*}{I_y} x, \tag{44}$$

where $I_y$ is the moment of inertia of $\Omega_f$. Note that in this special case $p_*$ do not depend on the weight coefficients $C_F, C_W, C_D$. If $\mu = \mu(x, y)$ we have more general dependence

$$p_* = \frac{P^\star}{S} + \Delta_1 + \left(\frac{M^*}{I_y} + \Delta_2\right) x, \tag{45}$$

$$\Delta_1 = \frac{C_F V}{2(C_W K_W V + C_D)} \left(\frac{1}{S} \int_{\Omega_f} \mu d\Omega_f - \mu\right),$$

$$\Delta_2 = \frac{C_F V}{2\left(C_W K_W V + C_D\right) I_y} \int_{\Omega_f} \mu x d\Omega_f.$$

In accordance with considered decomposition approach the second problem must be solved to define the optimal shape distribution $f_*(x, y)$ corresponding to the defined optimal pressure $p_*(x, y)$ and acting friction forces. For this purpose we use the exact solution [7]

$$w(x, y) = \kappa_0 \frac{Q_z^0}{\rho} + \kappa_f \frac{(x - x')Q_x^0}{\rho^2}, \quad \rho^2 = (x - x')^2 + (y - y')^2, \qquad (46)$$
$$\kappa_0 = \frac{1 - \nu^2}{\pi E}, \ \kappa_f = \frac{(1 + \nu)(1 - 2\nu)}{2\pi E}$$

of the problem for the pointed force

$$Q = \left\{Q_x, \ Q_y = 0, \ Q_z\right\} = \left\{Q_x^0 \delta(x - x', y - y'), \ 0, \ Q_z^0 \delta(x - x', y - y')\right\},$$
$$Q_x^0 = \int_{\Omega_f} \left(\tau_0 + \mu p_*\left(x', y'\right)\right) dx' dy', \quad Q_z^0 = \int_{\Omega_f} p_*(x', y') dx' dy' \qquad (47)$$

loaded at the point $(x, y)$. Here $\nu$ is a Poisson coefficient, $E$ is Young modulus and $\delta$ is Dirac function. By integration of the right side of the expression (46) on the region $\Omega_f$ we obtain

$$f_*(x, y) = \kappa_0 \int_{\Omega_f} \frac{p_*(x', y') dx' dy'}{\sqrt{(x - x')^2 + (y - y')^2}} + \kappa_f \int_{\Omega_f} \frac{(x - x')(\tau_0 + \mu p_*(x', y')) dx' dy'}{(x - x')^2 + (y - y')^2}. \qquad (48)$$

Let us introduce the new variables $\rho, \theta$. We have

$$x' - x = \rho \cos\theta, \quad y' - y = \rho \sin\theta,$$
$$\rho(x, y, x', y') = \sqrt{(x - x')^2 + (y - y')^2}, \quad d\Omega = \rho d\rho d\theta, \qquad (49)$$

and

$$f_*(x, y) = \kappa_0 \int_0^{2\pi} d\theta \int_0^{R(\theta)} p_*(x + \rho\cos\theta, \ y + \rho\sin\theta) d\rho$$
$$- \mu\kappa_f \int_0^{2\pi} \cos\theta d\theta \int_0^{R(\theta)} p_*(x + \rho\cos\theta, \ y + \rho\sin\theta) d\rho$$
$$- \tau_0 \kappa_f \int_0^{2\pi} R(\theta)\cos\theta d\theta, \qquad (50)$$

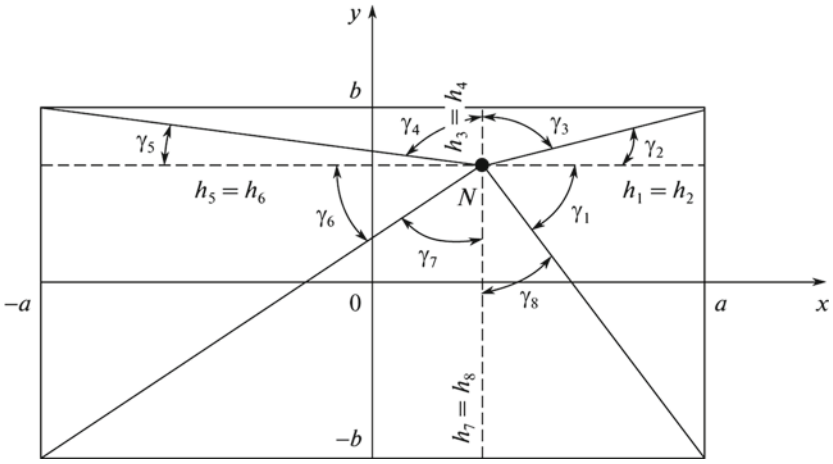



**Fig. 6** Contact region

where $r(\theta)$ is the distance between the fixed point N$(x, y)$ and the boundary of the region $\Omega_f$ (Fig. 6). The expression (44) can be rewritten as

$$p_* = p_*(x) = C + Ax' = C + A(x + \rho\cos\theta), \qquad (51)$$
$$C = \frac{P^*}{S}, \quad A = \frac{M^*}{I_y}.$$

Then we obtain the following relation for the optimal punch shape

$$f_*(x, y) = \kappa_0(C + Ax)D_0(x, y) + \frac{1}{2}\kappa_0 A D_C(x, y)$$
$$- \mu\kappa_f[(C + Ax)D_{0C}(x, y) + \frac{1}{2}A D_{CC}(x, y)] - \tau_0\kappa_f D_{0C}(x, y), \qquad (52)$$

$$D_0 = \int_0^{2\pi} R d\theta, \quad D_C = \int_0^{2\pi} R^2\cos\theta d\theta,$$
$$D_{0C} = \int_0^{2\pi} R\cos\theta d\theta, \quad D_{CC} = \int_0^{2\pi} R^2\cos^2\theta d\theta.$$

For considered rectangular region $\Omega_f$ function $D_0$ can be written in the form

$$D_0 = \frac{1}{2}[(a - x)\varphi_1(x, y) + (b - y)\varphi_2(x, y) + (a + x)\varphi_3(x, y) + (b + y)\varphi_4(x, y)], \qquad (53)$$
$$\varphi_1 = \ln\left(\frac{\rho_1 + b + y}{\rho_1 - b - y}\,\frac{\rho_2 + b - y}{\rho_2 - b + y}\right), \quad \varphi_2 = \ln\left(\frac{\rho_2 + a - x}{\rho_2 - a + x}\,\frac{\rho_3 + a + x}{\rho_3 - a - x}\right),$$

$$\varphi_3 = \ln\left(\frac{\rho_3 + b - y}{\rho_3 - b + y}\,\frac{\rho_4 + b + y}{\rho_4 - b - y}\right),\quad \varphi_4 = \ln\left(\frac{\rho_4 + a + x}{\rho_4 - a - x}\,\frac{\rho_1 + a - x}{\rho_1 - a + x}\right),$$
$$\rho_1^2 = (a-x)^2 + (b+y)^2,\quad \rho_2^2 = (a-x)^2 + (b-y)^2,$$
$$\rho_3^2 = (a+x)^2 + (b-y)^2,\quad \rho_4^2 = (a+x)^2 + (b+y)^2.$$

Performing the analogues transformations we obtain the following expressions for the functions $D_C$, $D_{0C}$, $D_{CC}$:

$$D_C = (a-x)^2\varphi_1/2 - (a+x)^2\varphi_3/2 + (b-y)^2(1/\cos\gamma_3 - 1/\cos\gamma_4) + (b+y)^2(1/\cos\gamma_8 - 1/\cos\gamma_7),$$
$$D_{0C} = (a-x)(\gamma_1+\gamma_2) + (a+x)(\gamma_5+\gamma_6) + (b-y)\ln\frac{\cos\gamma_4}{\cos\gamma_3} - (b+y)\ln\frac{\cos\gamma_8}{\cos\gamma_7},$$
$$D_{CC} = (a-x)^2(\gamma_1+\gamma_2) + (a+x)^2(\gamma_5+\gamma_6) - (b-y)^2(\gamma_3+\gamma_4) - (b+y)^2(\gamma_7+\gamma_8) + (b-y)^2(\cot\gamma_2 + \cot\gamma_5) + (b+y)^2(\cot\gamma_6 + \cot\gamma_1),$$

where

$$\gamma_1 = \arcsin\frac{b+y}{\rho_1},\ \gamma_2 = \arcsin\frac{b-y}{\rho_2},\ \gamma_3 = \arcsin\frac{a-x}{\rho_2},\ \gamma_4 = \arcsin\frac{a+x}{\rho_3},$$
$$\gamma_5 = \arcsin\frac{b-y}{\rho_3},\ \gamma_6 = \arcsin\frac{b+y}{\rho_4},\ \gamma_7 = \arcsin\frac{a+x}{\rho_4},\ \gamma_8 = \arcsin\frac{a-x}{\rho_1}.$$

In Fig. 7 the quantity $z = f_*/\kappa_0$ characterizing the optimal punch shape for $C = 1$, $A = 0.1$, $a = b = 1$, $\nu = 0.25$, $\mu = \tau_0 = 0.3$ is presented.

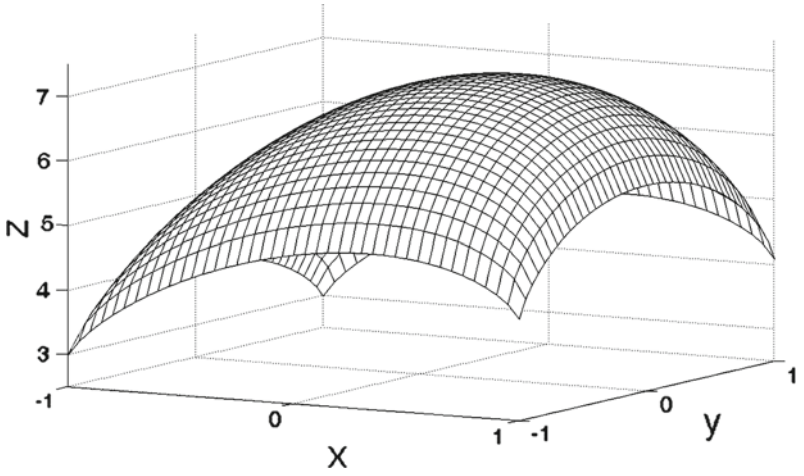



**Fig. 7** Optimal punch shape

## 4 Some Notes and Conclusions

In Sect. 2 of the paper we studied the multipurpose optimization problem for multilayered plate structure dynamically interacting with high-speed axisymmetric strikers. The layers of the structure were made from the given set of materials. Such scalar criterions as ballistic limit velocity, and total mass of the structure are considered as components of the vector-functional and the problem of optimization is treated in the Pareto-sense with application of the method of objective weighting. In contrast to previous investigations of the layered plates and protective coatings we developed the procedure of global optimum finding, based on evolutionary stochastic methods (genetic algorithm). Incomplete data concerning the thickness of layers of optimized multilayered shield structure are taken into account. The optimal structure was constructed numerically, i.e. the sequence and thicknesses of the layers were defined by the genetic algorithm. Note that the number of treated scalar criterions can be increased as well as the number of given materials. It should be also noted that all results are received without taking into account the processes of striker entrance and exit (as well as the layer boundaries influence on this process).

In Sect. 3 of this paper we formulated and investigated the problem of multipurpose optimization of the rigid punch shape that minimizes the wear volume rate, the friction dissipation power and the functional, characterized by the discrepancy between the pressure distribution corresponding to a certain punch shape and the specified (required) pressure distribution. The relative movement of the punch is considered in frame of quasi-static formulation. Effective decomposition method has been developed in frame of multipurpose optimization problem in the Pareto-sense and analytical solution of the shape optimization problem has been found. Note, that the proposed approach can be applied to the case of a system of moving punches. The number of considered and optimized functionals can be increased. It is reasonable to take into account the wear dissipation power in the future investigations.

**Acknowledgments** The research was performed under financial support of RFBR (grant 14-08-00016-a), RAS Program 12, Program of Support of Leading Scientific Schools (grant 2954.2014.1).

## References

1. N.V. Banichuk, *Introduction to Optimization of Structures* (Springer, New York, 1990)
2. N.V. Banichuk, Optimization of contact pressure in the problem of the interaction between a punch and an elastic medium. J. Appl. Math. Mech. **74**(3), 334–340 (2010)
3. N.V. Banichuk, S.Y. Ivanova, E.V. Makeev, Penetration of rigid strikers in layered plates and some problems of global multipurpose structural optimization. Prob. Strength Plast. **74**, 124–133 (2012). (in Russian)

4. G. Ben-Dor, A. Dubinsky, T. Elperin, *Applied High-speed Plate Penetration Dynamics* (Springer, Dordrecht, 2006)
5. D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Wesley)
6. I.G. Goryacheva, *Mechanics of Frictional Interaction* (Nauka, Moscow, 2001)
7. A.E.H. Love, *A Treatise on the Mathematical Theory of Elasticity* (Dover Publications, New York, 1944)
8. W. Stadler, Multicriteria Optimization in Mechanics (a survey)