Christian Constanda
Andreas Kirsch
Editors

# Integral Methods in Science and Engineering

## Theoretical and Computational Advances

Birkhäuser

Christian Constanda • Andreas Kirsch
Editors

# Integral Methods in Science and Engineering

## Theoretical and Computational Advances

Birkhäuser

*Editors*
Christian Constanda
Department of Mathematics
The University of Tulsa
Tulsa, OK, USA

Andreas Kirsch
Department of Mathematics
Karlsruhe Institute of Technology
Karlsruhe, Germany

ISBN 978-3-319-16726-8 ISBN 978-3-319-16727-5 (eBook)
DOI 10.1007/978-3-319-16727-5

Library of Congress Control Number: 2015949822

Mathematics Subject Classification (2010): 00B25, 35-06, 41-06, 44-06, 45-06, 65-06, 76-06, 86-06, 86A10

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)

# Preface

Since 1985, the international conferences on Integral Methods in Science and Engineering (IMSE) have brought together researchers in various theoretical and applied areas, whose work makes use, in one form or another, of integration techniques. This type of mathematical procedures are efficient, elegant, and powerful in their diversity, offering a common ground to, and serving as a linchpin between, many areas of academic endeavor.

The first 12 IMSE conferences took place in a variety of venues all over the world:

1985, 1990: University of Texas–Arlington, USA;
1993: Tohoku University, Sendai, Japan;
1996: University of Oulu, Finland;
1998: Michigan Technological University, Houghton, MI, USA;
2000: Banff, AB, Canada (organized by the University of Alberta, Edmonton);
2002: University of Saint-Étienne, France;
2004: University of Central Florida, Orlando, FL, USA;
2006: Niagara Falls, ON, Canada (organized by the University of Waterloo);
2008: University of Cantabria, Santander, Spain;
2010: University of Brighton, UK;
2012: Bento Gonçalves, Brazil (organized by the Federal University of Rio Grande do Sul).

The 2014 meeting, hosted by Karlsruhe Institute of Technology, July 21–25, and attended by participants from 11 countries on 4 continents, continued and strengthened the well-established IMSE reputation as an important international forum where scientists and engineers from all over the world have a fruitful and stimulating exchange of novel research ideas and projects.

IMSE 2014 was, as expected, organized to a very high standard, and the participants wish to thank Deutsche Forschungsgemeinschaft and the Department of

Mathematics at Karlsruhe Institute of Technology for their financial support. Special thanks are due to the members of the Local Organizing Committee:

Andreas Kirsch (Karlsruhe Institute of Technology), Chairman,
Tilo Arens (Karlsruhe Institute of Technology),
Frank Hettlich (Karlsruhe Institute of Technology).

A new feature of IMSE 2014 was the inclusion of three minisymposia—on Asymptotic Analysis: Homogenization and Thin Structures, Wave Phenomena, and Inverse Problems.

The next IMSE conference will be hosted by the University of Padova, Italy, in July 2016. Further details will be posted in due course on the conference web site.

The peer-reviewed chapters of this volume, arranged alphabetically by first author's name, are based on 58 papers from among those presented in Karlsruhe. The editors would like to thank the staff at Birkhäuser for their courteous and professional handling of the publication process.

Tulsa, OK, USA Christian Constanda
January 2015

# Contents

# Contributors

**Y. Alipouri** University of Science and Technology, Narmak, Tehran, Iran

**L.J. Alvarez-Vázquez** University of Vigo, Vigo, Spain

**A.C.M. Alvim** Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil

**A. Amosov** National Research University, Moscow Power Engineering Institute, Moscow, Russia

**C.E. Athanasiadis** National and Kapodistrian University of Athens, Athens, Greece

**C. Avila** Chevron Energy Technology Company, Houston, TX, USA

**T. Banaee** Mashhad University of Medical Sciences, Mashhad, Iran

**Andrea Barbarino** Politecnico di Torino, Torino, Italy

**J.B. Bardaji** Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**R.C. Barros** State University of Rio de Janeiro, Nova Friburgo, RJ, Brazil

**B.E.J. Bodmann** Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**C. Bozkaya** Middle East Technical University, Ankara, Turkey

**J.R.G. Braga** National Institute for Space Research, São José dos Campos, SP, Brazil

**D. Buoso** Politecnico di Torino, Torino, Italy

**D. Buske** Federal University of Pelotas, Pelotas, RS, Brazil

**K. Chen** University of Liverpool, Liverpool, UK

**W.A. Coberly** The University of Tulsa, Tulsa, OK, USA

**C. Constanda** The University of Tulsa, Tulsa, OK, USA

**J.C. da Costa** Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**M. Costabel** Université de Rennes 1, Rennes, France

**M. Dalla Riva** Department of Mathematics, The University of Tulsa, Tulsa, OK, USA

**G.A. Degrazia** Federal University of Santa Maria, Santa Maria, RS, Brazil

**D.R. Doty** The University of Tulsa, Tulsa, OK, USA

**T.T. Dufera** Addis Ababa University, Addis Ababa, Ethiopia

**S. Dulla** Politecnico di Torino, Torino, Italy

**M. Er** The University of Tulsa, Tulsa, OK, USA

**C. Etchegaray** Université Paris-Sud, Orsay, France

**F.D. Farimani** Ferdowsi University of Mashhad, Mashhad, Iran

**J.C.L. Fernandes** Federal University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

**C.P. Ferreira** Federal University of Santa Maria, Santa Maria, RS, Brazil

**I.C. Furtado** Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**N. García-Chan** University of Guadalajara, Guadalajara, Mexico

**M. Ghaniee Zarch** University of Science and Technology, Narmak, Tehran, Iran

**D.L. Gisch** Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**V.C. Gomes** Department of Science and Aerospace Technology (DCTA/IEAv), São José dos Campos, SP, Brazil

**D. Gómez** University of Cantabria, Santander, Spain

**S.J. Gotts** Laboratory of Brain and Cognition National Institute of Mental Health Intramural Research Program, Bethesda, MD, USA

**B. Grec** Université Paris Descartes, Paris, France

**S. Greggio** Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**P.J. Harris** University of Brighton, Brighton, UK

**E.B. Hauser** Pontifical Catholic University of Rio Grande do Sul, Rio Grande do Sul, Porto Alegre, RS, Brazil

**R. Hernández Torres** National Institute for Space Research, São José dos Campos, SP, Brazil

**H. Hirayama** Kanagawa Institute of Technology, Atsugi-Shi, Kanagawa-Ken, Japan

**E. Imani** Ferdowsi University of Mashhad, Mashhad, Iran

**O. Khoshdel Borj** Islamic Azad University, Mashhad, Iran

**G. Kouba** Energy Technology Company, Houston, TX, USA

**C.A. Ladeia** Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**M. Lanza de Cristoforis** Dipartimento di Matematica, Università degli Studi di Padova, Via Trieste 63, 35121 Padova, Italy

**T.V. Lisbôa** Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**M. Lobo** University of Cantabria, Santander, Spain

**J.F. Loeck** Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**E.F.P. Luz** National Institute for Space Research, São José dos Campos, SP, Brazil

**S. Maldaner** Federal University of Santa Maria, Santa Maria, RS, Brazil

**R.J. Marczak** Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**A. Martínez** University of Vigo, Vigo, Spain

**H.R. Mashhadi** Ferdowsi University of Mashhad, Mashhad, Iran

**B. Maury** Université Paris-Sud, Orsay, France

**T.A. Mel'nyk** Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

**N. Meunier** Université Paris Descartes, Paris, France

**S.E. Mikhailov** Brunel University London, Uxbridge, UK

**D. Mitrea** University of Missouri, Columbia, MO, USA

**I. Mitrea** Temple University, Philadelphia, PA, USA

**M. Mitrea** University of Missouri, Columbia, MO, USA

**R. Mohan** The University of Tulsa, Tulsa, OK, USA

**V.S. Moreira** Federal University of Pampa, Itaqui, RS, Brazil

**P. Musolino** Department of Mathematics, University of Padova, Padova, Italy

**D. Natroshvili** Georgian Technical University, Tbilisi, Georgia

**L. Navoret** Université de Strasbourg, Strasbourg, France

**F. Oliveira** Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**M.E. Pérez** University of Cantabria, Santander, Spain

**K. Orav-Puurand** Institute of Mathematics of the University of Tartu, Tartu, Estonia

**J. Orlik** Fraunhofer ITWM, Kaiserslautern, Germany

**A. Pedas** University of Tartu, Tartu, Estonia

**E. Pereyra** The University of Tulsa, Tulsa, OK, USA

**C.Z. Petersen** Federal University of Pelotas, Pelotas, RS, Brazil

**A. Plaza** University of Extremadura, Cáceres, Spain

**J. Plaza** University of Extremadura, Cáceres, Spain

**S. Pomeranz** The University of Tulsa, Tulsa, OK, USA

**A. Popov** Taras Shevchenko National University Kyiv, Kyiv, Ukraine

**C.F. Portillo** Brunel University London, Uxbridge, UK

**J. Poshtan** University of Science and Technology, Narmak, Tehran, Iran

**S. Potapenko** University of Waterloo, Waterloo, ON, Canada

**H.R. Pourreza** Ferdowsi University of Mashhad, Azadi Square, Mashhad, Iran

**L. Provenzano** University of Padova, Padova, Italy

**M. Puhl** The University of Tulsa, Tulsa, OK, USA

**R.S. Quadros** Federal University of Pelotas, Pelotas, RS, Brazil

**N. Ramezani** Islamic Azad University, Mashhad, Iran

**F.M. Ramos** National Institute for Space Research, São José dos Campos, SP, Brazil

**P. Ravetto** Politecnico di Torino, Torino, Italy

**U. Rizza** Institute of Atmospheric Sciences and Climate, Lecce, Italy

**H.M. Ruivo** National Institute for Space Research, São José dos Campos, SP, Brazil

**G. Sampaio** National Institute for Space Research, São José dos Campos, SP, Brazil

**B. Schmutzler** University of Missouri, Columbia, MO, USA

**M. Schramm** Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**C.F. Segatto** Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**V. Sevroglou** University of Piraeus, Piraeus, Greece

**T.A. Shaposhnikova** Moscow State University, Moscow, Russia

**E.H. Shiguemori** Department of Science and Aerospace Technology (DCTA/ IEAv), São José dos Campos, SP, Brazil

**V. Shiryaev** Fraunhofer ITWM, Kaiserslautern, Germany

**O. Shoham** The University of Tulsa, Tulsa, OK, USA

**J.J.A. Silva** Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil

**V.C. Silveira** Federal University of Santa Maria, Santa Maria, RS, Brazil

**W.K. Simmons** Laureate Institute for Brain Research, Tulsa, OK, USA
University of Tulsa, Tulsa, OK, USA

**I.G. Stratis** National and Kapodistrian University of Athens, Athens, Greece

**Tadie** Universitet Copenhagen, Universitet Sparken 5, Copenhagen, Denmark

**E. Tamme** University of Tartu, Tartu, Estonia

**T. Tirabassi** Institute of Atmospheric Sciences and Climate of National Council of Research, Bologna, Italy

**F.K. Tomaschewski** Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**G. Vainikko** Institute of Mathematics of the University of Tartu, Tartu, Estonia

**V.B. Vasilyev** Lipetsk State Technical University, Lipetsk, Russia

**R. Vasques** Federal University of Rio Grande do Sul, Av. Osvaldo Aranha 99/4, Porto Alegre 90046-900, Rio Grande do Sul, RS, Brazil

**M.E. Vázquez-Méndez** University of Santiago de Compostela, Escola Politécnica Superior, Lugo, Spain

**H.F.C. Velho** National Institute for Space Research, São José dos Campos, SP, Brazil

**G.T. Venturin** Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**M. Vikerpuur** University of Tartu, Tartu, Estonia

**M.T.M.B. Vilhena** Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**Y. Wang** University of Oklahoma, Norman, OK, USA

**M. Wollmann da Silva** Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**K. Yunt** General Control Design, Zürich, Switzerland

**B. Zubik-Kowal** Boise State University, Boise, ID, USA

**M.N. Zubova** Moscow State University, Moscow, Russia

# Chapter 1
# Solvability of a Nonstationary Problem of Radiative–Conductive Heat Transfer in a System of Semi-transparent Bodies

**A. Amosov**

## 1.1 Introduction

The problems of radiative–conductive heat transfer, where the radiative heat transfer and the conductive heat transfer should be taken into account simultaneously, appear in various fields of science and engineering. There is a huge amount of books and articles discussing such problems from the physical point of view. But the corresponding mathematical theory is not yet satisfactory. The main part of mathematical results received to date is devoted to the problems of complex heat transfer in materials which are nontransparent for radiation (opaque). Some of the recent results in this area and the corresponding bibliography can be found in [Am10a, Am10b].

The questions of solvability of the problems of radiation–conductive heat transfer in semi-transparent materials were studied in a relatively few number of articles [Am79a, Am79b, Ke96, LaTi98, LaTi01, La02, KoEtAl14, KoCh14].

## 1.2 Physical Statement of the Problem

We consider the problem of nonstationary radiative–conductive heat transfer in a system $G = \bigcup_{j=1}^{m} G_j$ of semi-transparent bodies $G_j$, separated by a vacuum. Each body $G_j$ is a bounded domain in $\mathbb{R}^3$ with a boundary $\partial G_j \in C^1$. We assume that domains $G_i$ and $G_j$ are pairwise disjoint whereas their boundaries can intersect for

A. Amosov (✉)
National Research University "Moscow Power Engineering Institute",
Krasnokazarmennaya Street 14, 111250 Moscow, Russia
e-mail: AmosovAA@mpei.ru



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_1

some $i \neq j$. Assume that each body $G_j$ is occupied by a semi-transparent medium with absorption coefficient $\varkappa_{j,\nu} > 0$, scattering coefficient $s_{j,\nu} \geq 0$ and refraction exponent $k_{j,\nu} > 1$ depending on radiation frequency $\nu$. We set $\varkappa_\nu(x) = \varkappa_{j,\nu}$, $s_\nu(x) = s_{j,\nu}$ and $k_\nu(x) = k_{j,\nu}$ for $x \in G_j$, $1 \leq j \leq m$.

The sought functions $u(x,t)$ and $I_\nu(\omega,x,t)$ are interpreted as the absolute temperature and the intensity of the radiation with frequency $\nu$ at a point $x \in G$ when the radiation propagates along the direction $\omega \in \Omega$. Here $\Omega = \{\omega \in \mathbb{R}^3 \mid |\omega| = 1\}$ is the unit sphere in $\mathbb{R}^3$ (the sphere of directions).

The emission and absorption of energy occur at frequencies $\nu \in \mathfrak{N} = \mathcal{N} \cup \{\nu_\ell\} \subset \mathbb{R}^+ = (0, +\infty)$. Here $\nu_\ell$ are the frequencies corresponding to spectral lines with widths $\Delta\nu_\ell > 0$. The set $\mathcal{N}$ is measurable and the set $\{\nu_\ell\}$ may be countable, finite or empty (the last in the event when spectral lines absent).

To describe the process of radiative–conductive heat transfer we use a system, consisting of the heat equation

$$c_p \frac{\partial u}{\partial t} - \operatorname{div}(\lambda(x,u)\nabla u) + H(x,u) = \int\limits_{\mathcal{N}} \varkappa_\nu \int\limits_{\Omega} I_\nu \, d\omega \, d\nu + \sum_{\ell} \varkappa_{\nu_\ell} \int\limits_{\Omega} I_{\nu_\ell} \, d\omega \, \Delta\nu_\ell + f,$$
$$(x,t) \in Q_T = G \times (0,T) \qquad (1.1)$$

and the radiative transfer equation

$$\omega \cdot \nabla I_\nu + (\varkappa_\nu + s_\nu) I_\nu = s_\nu \mathscr{S}_\nu(I_\nu) + \varkappa_\nu k_\nu^2 h_\nu(u),$$
$$(\omega,x,t) \in D \times (0,T), \quad \nu \in \mathfrak{N}, \qquad (1.2)$$

where $D = \Omega \times G = \bigcup\limits_{j=1}^{m} D_j$, $\quad D_j = \Omega \times G_j$, $\quad 1 \leq j \leq m$.

Here $c_p(x)$ is the coefficient of heat capacity, $\lambda(x,u)$ is the coefficient of thermal conductivity and $f(x,t)$ is the density of heat sources. The function

$$H(x,u) = 4\pi \int\limits_{\mathcal{N}} \varkappa_\nu(x) k_\nu^2(x) h_\nu(u) \, d\nu + 4\pi \sum_{\ell} \varkappa_{\nu_\ell}(x) k_{\nu_\ell}^2(x) h_{\nu_\ell}(u) \Delta\nu_\ell$$

corresponds to the density of the radiative energy, and the first two terms in the right-hand side of the equation (1.1) correspond to the density of an absorbed energy. The function $h_\nu$ corresponds to the Planck spectral distribution

$$h_\nu(u) = \frac{2\nu^2}{c_0^2} \frac{\hbar\nu}{\exp\left(\hbar\nu/(\widehat{k}u)\right) - 1}$$

where $\hbar > 0$ is the Planck constant, $\widehat{k} > 0$ is the Boltzmann constant and $c_0$ is the light speed in a vacuum. According to the physical interpretation, this function is

defined only for positive values of $u$. We extend it for $u \le 0$ by setting $h_\nu(u) = -h_\nu(|u|)$ if $u < 0$ and $h_\nu(0) = 0$.

In Equation (1.2) the term $\omega \cdot \nabla I_\nu = \sum_{i=1}^{3} \omega_i \frac{\partial}{\partial x_i} I_\nu$ means the derivative of a function $I_\nu$ along the direction $\omega$ and $\mathscr{S}_\nu$ denotes the scattering operator

$$\mathscr{S}_\nu(\varphi)(\omega,x) = \frac{1}{4\pi} \int_\Omega \theta_{j,\nu}(\omega' \cdot \omega)\varphi(\omega',x)\, d\omega', \quad (\omega,x) \in D_j, \ \ 1 \le j \le m,$$

where the scattering indicatrix $\theta_{j,\nu}$ satisfies

$$\theta_{j,\nu} \in L^1(-1,1), \quad \theta_{j,\nu} \ge 0, \quad \frac{1}{2} \int_{-1}^{1} \theta_{j,\nu}(\mu)\, d\mu = 1$$

for all $\nu \in \mathfrak{N}$ and $1 \le j \le m$.

Complete the system (1.1), (1.2) by the boundary value conditions

$$\lambda(x,u) \frac{\partial u}{\partial n_j} = 0, \quad (x,t) \in \partial G_j \times (0,T), \quad 1 \le j \le m, \tag{1.3}$$

$$I_\nu|_{\Gamma^-} = \mathfrak{B}_\nu(I_\nu|_{\Gamma^+}) + \mathfrak{C}(J_{*\nu}), \quad (\omega,x,t) \in \Gamma^- \times (0,T), \ \ \nu \in \mathfrak{N} \tag{1.4}$$

and the initial condition

$$u|_{t=0} = u^0, \quad x \in G. \tag{1.5}$$

Hereinafter $n_j$ is the outward normal to the boundary $\partial G_j$ of the domain $G_j$. The condition (1.3) means that the conductive heat flow on the boundary is missing. (Recall that the bodies $G_j$ are separated by vacuum).

The condition (1.4) describes reflection and refraction of the radiation on the boundaries of $G_i$. Here $I_\nu|_{\Gamma^-}$ and $I_\nu|_{\Gamma^+}$ are the intensities of incoming and outgoing radiations, $J_{*\nu}$ is the intensity of incoming radiation originating from the outside. The operators $\mathfrak{B}_\nu$ and $\mathfrak{C}$ will be defined in the next section; the sets $\Gamma^-$, $\Gamma^+$ are defined by (1.6), (1.7).

## 1.3 Boundary Value Problem for the Radiative Transfer Equation with Reflection and Refraction Conditions

In this section we briefly discuss the boundary value problem for the radiative transfer equation with reflection and refraction conditions. Detailed explanation can be found in [Am13a, Am13b, Am14].

### 1.3.1 Some Notations and Function Spaces

Let $x \cdot y = \sum\limits_{i=1}^{3} x_i y_i$ be the inner product in $\mathbb{R}^3$.

Hereinafter we will use the notation $\mu_j = \omega \cdot n_j(x)$ for $(\omega, x) \in \Omega \times \partial G_j$, $1 \le j \le m$. We introduce the sets

$$\Gamma = \Omega \times \partial G = \bigcup_{j=1}^{m} \Gamma_j, \quad \Gamma_j = \Omega \times \partial G_j, \quad 1 \le j \le m,$$

$$\Gamma^- = \bigcup_{j=1}^{m} \Gamma_j^-, \quad \Gamma_j^- = \{(\omega, x) \in \Gamma_j \mid \mu_j < 0\}, \quad 1 \le j \le m, \tag{1.6}$$

$$\Gamma^+ = \bigcup_{j=1}^{m} \Gamma_j^+, \quad \Gamma_j^+ = \{(\omega, x) \in \Gamma_j \mid \mu_j > 0\}, \quad 1 \le j \le m, \tag{1.7}$$

$$\Gamma^0 = \bigcup_{j=1}^{m} \Gamma_j^0, \quad \Gamma_j^0 = \{(\omega, x) \in \Gamma_j \mid \mu_j = 0\}, \quad 1 \le j \le m.$$

Assume that the measure $d\Gamma(\omega, x) = d\omega d\sigma(x)$ is introduced on $\Gamma$. Here $d\omega$ and $d\sigma(x)$ are the measures induced by the Lebesgue measure in $\mathbb{R}^3$ on $\Omega$ and $\partial G$, respectively. On $\Gamma^+$ and $\Gamma^-$ we introduce the measures

$$\widehat{d}\Gamma^+(\omega, x) = \omega \cdot n_j(x)\, d\omega d\sigma(x), \ \ (\omega, x) \in \Gamma_j^+, \ \ 1 \le j \le m;$$

$$\widehat{d}\Gamma^-(\omega, x) = |\omega \cdot n_j(x)|\, d\omega d\sigma(x), \ \ (\omega, x) \in \Gamma_j^-, \ \ 1 \le j \le m.$$

Let $1 \le p \le \infty$ and $E^\pm$ be a subset of $\Gamma^\pm$, measurable with respect to the measure $d\Gamma$. Denote by $\mathfrak{M}(E^\pm)$ the set of functions on $E^\pm$ that are measurable with respect to the measure $d\Gamma$ and introduce the Banach spaces $\widehat{L}^p(E^\pm)$ of functions $g \in \mathfrak{M}(E^\pm)$ which possess the finite norms

$$\|g\|_{\widehat{L}^p(E^\pm)} = \begin{cases} \left( \int\limits_{E^\pm} |g(\omega, x)|^p \, \widehat{d}\Gamma^\pm(\omega, x) \right)^{1/p}, & 1 \le p < \infty, \\ \operatorname*{ess\,sup}\limits_{(\omega, x) \in E^\pm} |g(\omega, x)|, & p = \infty. \end{cases}$$

Let $L^p_{loc}(\Gamma^\pm)$ be the space of functions $g \in \mathfrak{M}(\Gamma^\pm)$ such that $g \in L^p(K)$ for any compact subset $K \subset \Gamma^\pm$.

Denote by $L^p(D)$ the Banach space of functions $f$ defined on $D$ and measurable with respect to the measure $d\omega dx$ with the finite norm

$$\|f\|_{L^p(D)} = \begin{cases} \left( \int\limits_D |f(\omega, x)|^p d\omega dx \right)^{1/p}, & 1 \le p < \infty, \\ \operatorname*{ess\,sup}\limits_{(\omega, x) \in D} |f(\omega, x)|, & p = \infty. \end{cases}$$

Denote by $\mathscr{W}^p(D)$ the Banach space of functions $f \in L^p(D)$ possessing the weak derivative $\omega \cdot \nabla f \in L^p(D)$ and equipped with the norm

$$\|f\|_{\mathscr{W}^p(D)} = \begin{cases} \left( \|f\|^p_{L^p(D)} + \| \omega \cdot \nabla f\|^p_{L^p(D)} \right)^{1/p}, & 1 \le p < \infty, \\ \max\{\|f\|_{L^\infty(D)}, \| \omega \cdot \nabla f\|_{L^\infty(D)}\}, & p = \infty. \end{cases}$$

We will denote by $f|_{\Gamma^\pm}$ and $f|_{\Gamma_j^\pm}$ the traces of function $f \in \mathscr{W}^p(D)$ on $\Gamma^\pm$ and $\Gamma_j^\pm$, respectively. It is known that the traces $f|_{\Gamma^\pm}$ of a function $f \in \mathscr{W}^p(D)$ with $1 \le p < \infty$ do not necessarily belong to $\widehat{L}^p(\Gamma^\pm)$, but $f|_{\Gamma^\pm} \in L^p_{loc}(\Gamma^\pm)$.

Let $\nu \in \mathfrak{N}$ and $\widehat{\mu}_{j,\lim,\nu} = \sqrt{1 - 1/k^2_{j,\nu}}$. Introduce the sets

$$\widehat{\Gamma}^-_\nu = \bigcup_{j=1}^m \widehat{\Gamma}^-_{j,\nu}, \quad \widehat{\Gamma}^-_{j,\nu} = \{(\omega,x) \in \Gamma_j^- \mid -\widehat{\mu}_{j,\lim,\nu} < \mu_j < 0\},$$

$$\widehat{\Gamma}^+_\nu = \bigcup_{j=1}^m \widehat{\Gamma}^+_{j,\nu}, \quad \widehat{\Gamma}^+_{j,\nu} = \{(\omega,x) \in \Gamma_j^+ \mid 0 < \mu_j < \widehat{\mu}_{j,\lim,\nu}\},$$

$$\overset{\vee}{\Gamma}{}^-_\nu = \Gamma^- \setminus \widehat{\Gamma}^-_\nu = \bigcup_{j=1}^m \overset{\vee}{\Gamma}{}^-_{j,\nu}, \quad \overset{\vee}{\Gamma}{}^-_{j,\nu} = \{(\omega,x) \in \Gamma_j^- \mid -1 \le \mu_j \le -\widehat{\mu}_{j,\lim,\nu}\},$$

$$\overset{\vee}{\Gamma}{}^+_\nu = \Gamma^+ \setminus \widehat{\Gamma}^+_\nu = \bigcup_{j=1}^m \overset{\vee}{\Gamma}{}^+_{j,\nu}, \quad \overset{\vee}{\Gamma}{}^+_{j,\nu} = \{(\omega,x) \in \Gamma_j^+ \mid \widehat{\mu}_{j,\lim,\nu} \le \mu_j \le 1\}.$$

### *1.3.2 Boundary Operators*

Introduce the operators $\mathscr{R}^-_\nu$ and $\mathscr{R}^+_\nu$ of outer and inner reflections by the formulas

$$\mathscr{R}^-_\nu(\varphi)(\omega,x) = r^-_{j,\nu}(\mu_j)\varphi(\omega - 2\mu_j n_j(x),x), \quad (\omega,x) \in \Gamma_j^-, \quad 1 \le j \le m;$$
$$\mathscr{R}^+_\nu(\psi)(\omega,x) = r^+_{j,\nu}(\mu_j)\psi(\omega - 2\mu_j n_j(x),x), \quad (\omega,x) \in \Gamma_j^+, \quad 1 \le j \le m.$$

Here $r^-_{j,\nu}$ and $r^+_{j,\nu}$ are the coefficients of the outer and inner reflections, defined by Fresnel formulas

$$r^-_{j,\nu}(\mu_j) = \frac{1}{2}\left[ \left( \frac{\eta^+_{j,\nu}(\mu_j) + k_{j,\nu}\mu_j}{\eta^+_{j,\nu}(\mu_j) - k_{j,\nu}\mu_j} \right)^2 + \left( \frac{k_{j,\nu}\eta^+_j(\mu_j) + \mu_j}{k_{j,\nu}\eta^+_{j,\nu}(\mu_j) - \mu_j} \right)^2 \right], \ (\omega,x) \in \overset{\vee}{\Gamma}{}^-_{j,\nu},$$

$$r^-_{j,\nu}(\mu_j) = 1, \quad (\omega,x) \in \widehat{\Gamma}^-_{j,\nu},$$

$$r^+_{j,\nu}(\mu_j) = \frac{1}{2}\left[ \left( \frac{\mu_j - k_{j,\nu}\eta^-_{j,\nu}(\mu_j)}{\mu_j + k_{j,\nu}\eta^-_{j,\nu}(\mu_j)} \right)^2 + \left( \frac{k_{j,\nu}\mu_j - \eta^-_{j,\nu}(\mu_j)}{k_{j,\nu}\mu_j + \eta^-_{j,\nu}(\mu_j)} \right)^2 \right], \ (\omega,x) \in \Gamma_j^+,$$

where

$$\eta^+_{j,\nu}(\mu_j)=\sqrt{1-k^2_{j,\nu}(1-\mu^2_j)},\quad \eta^-_{j,\nu}(\mu_j)=\sqrt{1-\frac{1}{k^2_{j,\nu}}(1-\mu^2_j)}.$$

Note that the effect of complete inner reflection holds for $(\omega,x)\in\widehat{\Gamma}^-_{j,\nu}$:

$$\mathscr{R}^-_\nu(\varphi)(\omega,x)=\varphi(\omega-2\mu_j n_j(x),x),\quad (\omega,x)\in\widehat{\Gamma}^-_{j,\nu},\quad 1\le j\le m.$$

Define the operators $\mathscr{P}^-_\nu$ and $\mathscr{P}^+_\nu$ of refraction inside and outside $G$ by the formulas

$$\mathscr{P}^-_\nu(\psi)(\omega,x)=\begin{cases}(1-r^-_{j,\nu}(\mu_j))\,k^2_{j,\nu}\,\psi(\omega_{\mathscr{P}^-_{j,\nu}}(\omega,x),x),\ (\omega,x)\in\overset{\vee}{\Gamma}{}^-_{j,\nu}, & 1\le j\le m,\\ 0,\qquad (\omega,x)\in\widehat{\Gamma}^-_{j,\nu},\end{cases}$$

$$\mathscr{P}^+_\nu(\varphi)(\omega,x)=(1-r^+_{j,\nu}(\mu_j))\frac{1}{k^2_{j,\nu}}\,\varphi(\omega_{\mathscr{P}^+_{j,\nu}}(\omega,x),x),\ (\omega,x)\in\Gamma^+_j,\quad 1\le j\le m,$$

where

$$\omega_{\mathscr{P}^-_{j,\nu}}(\omega,x)=-\eta^+_{j,\nu}(\mu_j)\,n_j(x)+k_{j,\nu}(\omega-\mu_j n_j(x)),$$

$$\omega_{\mathscr{P}^+_{j,\nu}}(\omega,x)=\eta^-_{j,\nu}(\mu_j)\,n_j(x)+\frac{1}{k_{j,\nu}}(\omega-\mu_j n_j(x)).$$

It is proved that

$$\mathscr{R}^-_\nu\colon L^p_{loc}(\Gamma^+)\to L^p_{loc}(\Gamma^-),\ \mathscr{R}^-_\nu\colon\widehat{L}^p(\widehat{\Gamma}^+_\nu)\to\widehat{L}^p(\widehat{\Gamma}^-_\nu),\ \mathscr{R}^+_\nu\colon\widehat{L}^p(\Gamma^-)\to\widehat{L}^p(\Gamma^+),$$

$$\mathscr{P}^-_\nu\colon\widehat{L}^p_{1-r^+_\nu}(\Gamma^-)\to\widehat{L}^p(\Gamma^-),\ \ \mathscr{P}^+_\nu\colon\widehat{L}^p_{1-r^-_\nu}(\overset{\vee}{\Gamma}{}^+_\nu)\to\widehat{L}^p(\Gamma^+),\quad 1\le p\le\infty.$$

Here

$$L^p_{1-r^+_\nu}(\Gamma^-)=\{g\in\mathfrak{M}(\Gamma^-)\mid\sum_{j=1}^n\int\limits_{\Gamma^-_j}|g(\omega,x)|^p[1-r^+_j(\mu_j)]\widehat{d}\Gamma^-(\omega,x)<\infty\},$$

$$L^p_{1-r^-_\nu}(\overset{\vee}{\Gamma}{}^+_\nu)=\{g\in\mathfrak{M}(\overset{\vee}{\Gamma}{}^+_\nu)\mid\sum_{j=1}^n\int\limits_{\overset{\vee}{\Gamma}{}^+_{j,\nu}}|g(\omega,x)|^p[1-r^-_j(\mu_j)]\widehat{d}\Gamma^+(\omega,x)<\infty\}$$

for $1\le p<\infty$ and $L^\infty_{1-r^+_\nu}(\Gamma^-)=L^\infty(\Gamma^-)$, $L^\infty_{1-r^-_\nu}(\overset{\vee}{\Gamma}{}^+_\nu)=L^\infty(\overset{\vee}{\Gamma}{}^+_\nu)$.

Let $\partial G_j \cap \partial G_j \neq \emptyset$ for some $i \neq j$ and

$$\Gamma_{ij}^{-+} = \Gamma_i^- \cap \Gamma_j^+, \quad \Gamma_{ij}^{+-} = \Gamma_i^+ \cap \Gamma_j^-, \quad \overset{\vee}{\Gamma}{}_{ij,\nu}^{-+} = \overset{\vee}{\Gamma}{}_{i,\nu}^- \cap \Gamma_j^+,$$
$$\widehat{\Gamma}_{ij,\nu}^{-+} = \widehat{\Gamma}_{i,\nu}^- \cap \Gamma_j^+, \quad \overset{\vee}{\Gamma}{}_{ij,\nu}^{+-} = \overset{\vee}{\Gamma}{}_{i,\nu}^+ \cap \Gamma_j^-, \quad \widehat{\Gamma}_{ij,\nu}^{+-} = \widehat{\Gamma}_{i,\nu}^+ \cap \Gamma_j^-.$$

Introduce the operators $\mathscr{R}_{ij,\nu}^-$ and $\mathscr{P}_{ij,\nu}^-$ by formulas

$$\mathscr{R}_{ij,\nu}^-(\varphi)(\omega,x) = \begin{cases} r_{ij,\nu}^-(\mu_i)\varphi(\omega - 2\mu_i n_i(x),x), & (\omega,x) \in \overset{\vee}{\Gamma}{}_{ij,\nu}^{-+}, \\ \varphi(\omega - 2\mu_i n_i(x),x), & (\omega,x) \in \widehat{\Gamma}_{ij,\nu}^{-+}, \end{cases}$$

$$\mathscr{P}_{ij,\nu}^-(\psi)(\omega,x) = \begin{cases} (1 - r_{ij,\nu}^-(\mu_i))\dfrac{k_{i,\nu}^2}{k_{j,\nu}^2}\psi(\omega_{\mathscr{P}_{ij,\nu}^-}(\omega,x),x), & (\omega,x) \in \overset{\vee}{\Gamma}{}_{ij,\nu}^{-+}, \\ 0, & (\omega,x) \in \widehat{\Gamma}_{ij,\nu}^{-+}. \end{cases}$$

Here

$$\omega_{\mathscr{P}_{ij,\nu}^-}(\omega,x) = -\eta_{ij,\nu}^+(\mu_i)n_i(x) + \frac{k_{i,\nu}}{k_{j,\nu}}(\omega - \mu_i n_i(x)), \quad \mu_i = \omega \cdot n_i(x),$$
$$\eta_{ij,\nu}^+(\mu_i) = \sqrt{1 - \frac{k_{i,\nu}^2}{k_{j,\nu}^2}(1-\mu_i^2)},$$
$$r_{ij,\nu}^-(\mu_i) = \frac{r_{i,\nu}^-(\mu_i) + r_{j,\nu}^+(\eta_{i,\nu}^+(\mu_i)) - 2r_{i,\nu}^-(\mu_i)r_{j,\nu}^+(\eta_{i,\nu}^+(\mu_i))}{1 - r_{i,\nu}^-(\mu_i)r_{j,\nu}^+(\eta_{i,\nu}^+(\mu_i))}.$$

It is proved that

$$\mathscr{R}_{ij,\nu}^- : L_{loc}^p(\Gamma_{ij}^{+-}) \to L_{loc}^p(\Gamma_{ij}^{-+}), \quad \mathscr{R}_{ij,\nu}^- : \widehat{L}^p(\overset{\vee}{\Gamma}{}_{ij,\nu}^{+-}) \to \widehat{L}^p(\overset{\vee}{\Gamma}{}_{ij,\nu}^{-+}),$$
$$\mathscr{P}_{ij,\nu}^- : \widehat{L}^p(\overset{\vee}{\Gamma}{}_{ji,\nu}^{+-}) \to \widehat{L}^p(\overset{\vee}{\Gamma}{}_{ij,\nu}^{-+}), \quad 1 \leq p \leq \infty.$$

Introduce the sets

$$S_j^- = \{(\omega,x) \in \Gamma_j^- \mid x \in \partial G_j \setminus \bigcup_{i \neq j} \partial G_i\}, \quad S^- = \bigcup_{j=1}^m S_j^-,$$
$$\overset{*}{S}{}^- = \{(\omega,x) \in S^- \mid \{x - t\omega \mid t > 0\} \cap \overline{G} = \emptyset\}.$$

Let $(\omega,x) \in S^- \setminus \overset{*}{S}{}^- = \{(\omega,x) \in S^- \mid \{x - t\omega \mid t > 0\} \cap \overline{G} \neq \emptyset\}$. We set

$$\tau^-(\omega,x) = \inf\{t > 0 \mid x - t\omega \in \overline{G}\} \quad \text{and} \quad X^-(\omega,x) = x - \tau^-(\omega,x)\omega.$$

Note that $X^-(\omega,x) \in \partial G$ and $(\omega, X^-(\omega,x)) \in \Gamma^+ \cup \Gamma^0$.

Introduce the set

$$\widetilde{S}^- = \{(\omega,x) \in S^- \setminus \overset{*}{S}^- \mid (\omega, X^-(\omega,x)) \in \Gamma^+\}$$

and define translation operator $T$ by formula

$$T\varphi(\omega,x) = \varphi(\omega, X^-(\omega,x)), \quad (\omega,x) \in \widetilde{S}^-.$$

### 1.3.3 Statement of the Reflection and Refraction Conditions

Remember that $I_\nu|_{\Gamma^-}$ and $I_\nu|_{\Gamma^+}$ are the intensity of entering $G$ and outgoing from $G$ radiation. Let $J_\nu$ be the intensity of the radiation propagating in the vacuum and falling to $\partial G$.

For $(\omega,x) \in \widetilde{S}^-$ the radiation $J_\nu$ falling from the vacuum to $\partial G$ comes from the points $X^-(\omega,x) \in \partial G$. This radiation is composed of the reflected and refracted radiations at the point $X^-(\omega,x)$:

$$J_\nu = T\mathscr{R}_\nu^+(J_\nu) + T\mathscr{P}_\nu^+(I_\nu|_{\Gamma^+}), \quad (\omega,x) \in \widetilde{S}^-.$$

For $(\omega,x) \in \overset{*}{S}^-$ the radiation $J_\nu$ falling from the vacuum to $\partial G$ goes outside and we can assume that it is prescribed:

$$J_\nu = J_{*\nu}, \quad (\omega,x) \in \overset{*}{S}^-.$$

For $(\omega,x) \in S^-$ entering $G$ radiation is composed of the reflected and refracted radiations:

$$I_\nu|_{\Gamma^-} = \mathscr{R}_\nu^-(I|_{\Gamma^+}) + \mathscr{P}_\nu^-(J_\nu), \quad (\omega,x) \in S^-.$$

For $(\omega,x) \in \Gamma_{ij}^{-+}$ entering $G_i$ radiation is composed of the reflected and refracted radiations also:

$$I_\nu|_{\Gamma_i^-} = \mathscr{R}_{ij,\nu}^-(I_\nu|_{\Gamma_i^+}) + \mathscr{P}_{ij,\nu}^-(I_\nu|_{\Gamma_j^+}), \quad (\omega,x) \in \Gamma_{ij}^{-+}.$$

Here $I_\nu|_{\Gamma_i^-}$ and $I_\nu|_{\Gamma_i^+}$ are the values of the intensities of entering $G_i$ and outgoing from $G_i$ radiations.

### 1.3.4 Boundary Value Problem for Radiative Transfer Equation with Reflection and Refraction Conditions

Consider the problem

$$\omega\cdot\nabla I_\nu+(\varkappa_\nu+s_\nu)I_\nu=s_\nu\mathscr{S}_\nu(I_\nu)+\varkappa_\nu k_\nu^2F_\nu,\quad(\omega,x)\in D,\tag{1.8}$$

$$I_\nu|_{\Gamma^-}=\mathscr{R}_\nu^-(I_\nu|_{\Gamma^+})+\mathscr{P}_\nu^-(J_\nu),\quad(\omega,x)\in S^-,\tag{1.9}$$

$$I_\nu|_{\Gamma_i^-}=\mathscr{R}_{ij,\nu}^-(I|_{\Gamma_i^+})+\mathscr{P}_{ij,\nu}^-(I_\nu|_{\Gamma_j^+}),\quad(\omega,x)\in\Gamma_{ij}^{-+},\quad i\neq j,\tag{1.10}$$

$$J_\nu=T\mathscr{R}_\nu^+(J_\nu)+T\mathscr{P}_\nu^+(I_\nu|_{\Gamma^+}),\quad(\omega,x)\in\widetilde{S}^-,\tag{1.11}$$

$$J_\nu=J_{*\nu},\quad(\omega,x)\in\overset{*}{S}{}^-.\tag{1.12}$$

describing the radiative transfer in the system of semi-transparent bodies with taking into account reflection and refraction on their boundaries. The functions $F_\nu\in L^1(D)$ and $J_{*\nu}\in\widehat{L}^1(\overset{*}{S}{}^-)$ are prescribed.

It is proved [Am10a] that a solution $J_\nu\in\widehat{L}^1_{1-r_\nu^+}(S^-)$ of subproblem (1.11), (1.12) exists, is unique and may be represented in the form $J_\nu=\mathscr{B}_\nu(I_\nu|_{\Gamma^+})+\mathscr{C}_\nu(J_{*\nu})$, where $\mathscr{B}_\nu$ and $\mathscr{C}_\nu$ are linear bounded operators from $\widehat{L}^1(\overset{\vee}{\Gamma}{}_\nu^-)$ to $\widehat{L}^1_{1-r_\nu^+}(S^-)$ and from $\widehat{L}^1(\overset{*}{S}{}^-)$ to $\widehat{L}^1_{1-r_\nu^+}(S^-)$, respectively.

So the unknown function $J_\nu$ can be excluded and the problem (1.8)–(1.12) may be reduced to the following problem

$$\omega\cdot\nabla I_\nu+(\varkappa_\nu+s_\nu)I_\nu=s_\nu\mathscr{S}_\nu(I_\nu)+\varkappa_\nu k_\nu^2F_\nu,\quad(\omega,x)\in D,\tag{1.13}$$

$$I_\nu|_{\Gamma^-}=\mathfrak{B}_\nu(I_\nu|_{\Gamma^+})+\mathfrak{C}(J_{*\nu}),\quad(\omega,x)\in\Gamma^-\tag{1.14}$$

with one unknown function $I_\nu$. Here

$$\mathfrak{B}_\nu(I_\nu|_{\Gamma^+})(\omega,x)=\begin{cases}[\mathscr{R}_\nu^-(I_\nu|_{\Gamma^+})+\mathscr{P}_\nu^-\mathscr{B}_\nu^-(I_\nu|_{\Gamma^+})](\omega,x),\ (\omega,x)\in S^-,\\ [\mathscr{R}_{ij,\nu}^-(I_\nu|_{\Gamma_i^+})+\mathscr{P}_{ij,\nu}^-(I_\nu|_{\Gamma_j^+})](\omega,x),\ (\omega,x)\in\Gamma_{ij}^-,\ i\neq j,\end{cases}$$

$$\mathfrak{C}_\nu(J_{*\nu})(\omega,x)=\begin{cases}\mathscr{P}_\nu^-\mathscr{C}_\nu^-(J_{*\nu})(\omega,x),\ (\omega,x)\in S^-,\\ 0,\quad(\omega,x)\in\Gamma_{ij}^-,\ i\neq j.\end{cases}$$

By a solution to the problem (1.13), (1.14) we mean a function $I_\nu\in\mathscr{W}^1(D)$, that satisfies equation (5.1) almost everywhere on $D$ and condition (1.14) almost everywhere on $\Gamma^-$.

The following theorem holds.

**Theorem 1.** *Let $F_\nu \in L^1(D)$, $J_\nu \in \widehat{L}^1(\overset{*}{S}{}^-)$. Then the solution $I_\nu \in \mathscr{W}^1(D)$ to the problem (1.13), (1.14) exists and is unique. If additionally $F_\nu \in L^p(D)$, $J_{*\nu} \in \widehat{L}^p(\overset{*}{S}{}^-)$ with some $p \in (1,\infty]$, then $I_\nu \in \mathscr{W}^p(D)$.*

*The following estimates hold:*

$$\|\varkappa_\nu^{1/p} k_\nu^{2/p-2} I_\nu\|_{L^p(D)} \le \left(\|\varkappa_\nu^{1/p} k_\nu^{2/p} F_\nu\|^p_{L^p(D)} + \|J_{*\nu}\|^p_{\widehat{L}^p(\overset{*}{S}{}^-)}\right)^{1/p},$$

$$\|\varkappa_\nu^{1/p-1} k_\nu^{2/p-2}\, \omega \cdot \nabla I_\nu\|_{L^p(D)} \le$$

$$\le \frac{2}{1-\varpi_{\max,\nu}} \left(\|\varkappa_\nu^{1/p} k_\nu^{2/p} F_\nu\|^p_{L^p(D)} + \|J_{*\nu}\|^p_{\widehat{L}^p(\overset{*}{S}{}^-)}\right)^{1/p}$$

*for $1 \le p < \infty$ and the estimates*

$$\|k_\nu^{-2} I_\nu\|_{L^\infty(D)} \le \max\{\|F_\nu\|_{L^\infty(D)}, \|J_{*\nu}\|_{L^\infty(\overset{*}{S}{}^-)}\},$$

$$\|\varkappa_\nu^{-1} k_\nu^{-2}\, \omega \cdot \nabla I_\nu\|_{L^\infty(D)} \le \frac{2}{1-\varpi_{\max,\nu}} \max\{\|F_\nu\|_{L^\infty(D)}, \|J_{*\nu}\|_{L^\infty(\overset{*}{S}{}^-)}\}$$

*for $p = \infty$.*

*Here* $\varpi_{\max,\nu} = \max\limits_{1\le j\le m} \dfrac{s_{j,\nu}}{\varkappa_{j,\nu} + s_{j,\nu}} < 1.$

Thus, the solution to the problem (1.13), (1.14) can be represented in the form

$$I_\nu = \mathscr{A}_\nu(F_\nu) + \mathscr{D}_\nu(J_{*\nu}), \tag{1.15}$$

where $\mathscr{A}_\nu : L^p(D) \to W^p(D)$ and $\mathscr{D}_\nu : \widehat{L}^p(\overset{*}{S}{}^-)$ are the linear bounded operators. The operator $\mathscr{A}_\nu$ maps the function $F_\nu \in L^p(D)$ into the solution of the problem (1.13), (1.14) with $J_{*\nu} = 0$, and the operator $\mathscr{A}_\nu$ maps the function $J_{*\nu} \in \widehat{L}^p(\overset{*}{S}{}^-)$ into the solution of the problem (1.13), (1.14) with $F_\nu = 0$.

## 1.4 Mathematical Statement of the Problem and Main Results

Let us return to the problem (1.1)–(1.5). Using formula (1.15) with $F_\nu = h_\nu(u)$ we can exclude the unknown function $I_\nu$ from the problem and rewrite it in the following form:

$$c_p\frac{\partial u}{\partial t}-\operatorname{div}(\lambda(x,u)\nabla u)+H(x,u)=\mathscr{H}[u]+F,\quad (x,t)\in Q_T, \tag{1.16}$$

$$\lambda(x,u)\frac{\partial u}{\partial n_j}=0,\quad (x,t)\in \partial G_j\times(0,T),\quad 1\le j\le m, \tag{1.17}$$

$$u(x,0)=u^0(x),\quad x\in G, \tag{1.18}$$

where

$$\mathscr{H}[u]=\int\limits_{\mathscr{N}}\varkappa_\nu\int\limits_{\Omega}\mathscr{A}_\nu(h_\nu(u))\,d\omega\,d\nu+\sum_\ell\varkappa_{\nu_\ell}\int\limits_{\Omega}\mathscr{A}_{\nu_\ell}(h_{\nu_\ell}(u))\,d\omega\,\Delta\nu_\ell,$$
$$F=f+\int\limits_{\mathscr{N}}\varkappa_\nu\int\limits_{\Omega}\mathscr{D}_\nu(J_{*,\nu})\,d\omega\,d\nu+\sum_\ell\varkappa_{\nu_\ell}\int\limits_{\Omega}\mathscr{D}_{\nu_\ell}(J_{*\nu_\ell})\,d\omega\,\Delta\nu_\ell.$$

Remember that

$$H(x,u)=4\pi\int\limits_{\mathscr{N}}\varkappa_\nu(x)k_\nu^2(x)h_\nu(u)\,d\nu+4\pi\sum_\ell\varkappa_{\nu_\ell}(x)k_{\nu_\ell}^2(x)h_{\nu_\ell}(u)\Delta\nu_\ell.$$

We assume that the following assumptions on data hold:

$(A_1)$ $c_p\in L^\infty(G)$ and there are positive constants $\underline{c}_p$, $\overline{c}_p$ such that

$$0<\underline{c}_p\le c_p(x)\le\overline{c}_p\quad\forall x\in G.$$

$(A_2)$ The function $\lambda(x,u)$ is defined on $G\times\mathbb{R}$ and for all $u\in\mathbb{R}$ is measurable on $x$. Besides, there are positive constants $\underline{\lambda}\le\overline{\lambda}$ and $A$ such that

$$0<\underline{\lambda}\le\lambda(x,u)\le\overline{\lambda}\quad\forall(x,u)\in G\times\mathbb{R}$$

and the following Hölder condition holds:

$$|\lambda(x,u+v)-\lambda(x,u)|\le A|v|^{1/2}\quad\forall(x,u)\in G\times\mathbb{R},\ \forall v\in[-1,1].$$

$(A_3)$ The function $H(x,u)$ is defined on $G\times\mathbb{R}$ and the following inequality holds:

$$|H(x,u)|\le c_H(|u|^s+1)\quad\forall(x,u)\in G\times\mathbb{R},$$

where $s>0$, $c_H>0$ are constants.

$(A_4)$ $u^0\in L^p(G)$, where $p\in[\max\{2,3s/5\},\infty]$;

$(A_5)$ $F\in L^r(0,T;L^q(G))$, where $r,q\in[1,\infty]$, $2/r+3/q\le 2+3/p$.

$(A_6)$ $\varkappa_{\nu,j}>0$, $s_{\nu,j}\ge 0$, $k_{\nu,j}>1$ for all $\nu\in\mathfrak{N}$ and $1\le j\le m$. Besides, coefficients $\varkappa_{\nu,j}$, $s_{\nu,j}$, $k_{\nu,j}$ are measurable functions of $\nu\in\mathscr{N}$.

$(A_7)$ The scattering indicatrix $\theta_{j,\nu}$ is such that:

$$\theta_{j,\nu} \in L^1(-1,1), \quad \theta_{j,\nu} \geq 0, \quad \frac{1}{2}\int\limits_{-1}^{1} \theta_{j,\nu}(\mu)\, d\mu = 1 \quad \text{for all } \nu \in \mathfrak{N} \text{ and } 1 \leq j \leq m.$$

Besides, the function $\nu \to \theta_{j,\nu}$, considering as a mapping from $\mathcal{N}$ to $L^1(-1,1)$, is strongly measurable.
We introduce the Banach space $V_2(Q_T) = L^2(0,T;W_2^1(G)) \cap L^\infty(0,T;L^2(G))$, equipped with the norm

$$\|u\|_{V_2(Q_T)} = \|\nabla u\|_{L^2(0,T;L^2(G))} + \|u\|_{L^\infty(0,T;L^2(G))}.$$

By a solution to the problem (1.16)–(1.18) we mean a function $u \in V_2(Q_T) \cap L^s(Q_T) \cap C([0,T];L^1(G))$ such that it satisfies the integral identity

$$-\int\limits_{Q_T} c_p u \frac{\partial \varphi}{\partial t}\, dxdt + \int\limits_{Q_T} \lambda(x,u)\nabla u \cdot \nabla \varphi\, dxdt + \int\limits_{Q_T} H(x,u)\varphi\, dxdt =$$
$$= \int\limits_{G} c_p u^0 \varphi|_{t=0}\, dx + \int\limits_{Q_T} \mathscr{H}[u]\varphi\, dxdt + \int\limits_{Q_T} F\varphi\, dxdt$$

for all $\varphi(x,t) = v(x)\eta(t)$, where $v \in W_2^1(G) \cap L^\infty(G)$ and $\eta \in C^\infty[0,T]$, $\eta(T) = 0$.

**Theorem 2.** *A solution to the problem (1.16)–(1.18) exists, is unique and satisfies the following estimate:*

$$\|u\|_{V_2(Q_T)} \leq C_1(\|u^0\|_{L^p(G)} + \|F\|_{L^r(0,T;L^q(G))}). \tag{1.19}$$

*If $p > 2$, then $|u|^{\gamma-1}u \in V_2(Q)$ for all $\gamma \in (1, p/2]$ and the following estimate holds:*

$$\||u|^{\gamma-1}u\|_{V_2(Q_T)} \leq C_2(\|u^0\|_{L^p(G)} + \|F\|_{L^r(0,T;L^q(G))}). \tag{1.20}$$

*If $p = \infty$ and exponents $q, r$ are such that $2/r + 3/q < 2$, then $u \in L^\infty(Q_T)$ and the following estimate holds:*

$$\|u\|_{L^\infty(Q)} \leq C_3(\|u^0\|_{L^\infty(G)} + \|F\|_{L^r(0,T;L^q(G))}). \tag{1.21}$$

*In estimates (1.19)–(1.21) $C_1$, $C_2$, $C_3$ are positive constants depending on $G$, $T$, $\underline{c}_p$, $\overline{c}_p$, $\underline{\lambda}$, $r$ and $q$.*

The following comparison theorem holds.

**Theorem 3.** *Let $u^1, u^2$ be solutions to problem (1.16)–(1.18), corresponding to the data $u^{0,1}, u^{0,2} \in L^p(G)$ and $F^1, F^2 \in L^r(0,T;L^q(G))$.*
*If $u^{0,1} \leq u^{0,2}$ and $F^1 \leq F^2$, then $u^1 \leq u^2$.*

**Corollary 1.** *If* $u^0 \geq 0$ *and* $F \geq 0$*, then* $u \geq 0$*.*

**Acknowledgements** This work was supported by the Russian Scientific Foundation (agreement 14-11-00306) and Board grants of the President of Russia (grant NSh-2081.2014.1).

## References

[Am10a] Amosov, A.A.: Stationary nonlinear nonlocal problem of radiative-conductive heat transfer in a system of opaque bodies with properties depending on radiation frequency. J. Math. Sci. **164**, issue 3, 309–344, (2010).

[Am10b] Amosov, A.A.: Nonstationary nonlinear nonlocal problem of radiative-conductive heat transfer in a system of opaque bodies with properties depending on the radiation frequency. J. Math. Sci. **165**, issue 1, 1–41 (2010).

[Am79a] Amosov, A.A.: The solvability of a problem of radiation heat transfer. Sov. Phys., Dokl. **24**, No. 4, 261–262 (1979).

[Am79b] Amosov, A.A.: The limit connection between two problems of radiation heat transfer. Sov. Phys., Dokl. **24**, No. 6, 439–441 (1979).

[Ke96] Kelley, C.T.: Existence and uniqueness of solutions of nonlinear systems of conductive-radiative heat transfer equations. Transport Theory and Statistical Physics. **25:2**, 249–260 (1996).

[LaTi98] Laitinen, M.T., Tiihonen, T.: Integro-differential equation modelling heat transfer in conducting, radiating and semitransparent materials. Mathematical Methods in Applied Sciences. **21**, 375–392 (1998).

[LaTi01] Laitinen, M.T., Tiihonen, T.: Conductive-radiative heat transfer in grey materials. Quart. Appl. Math. **59**, 737–768 (2001).

[La02] Laitinen, M.: Asymptotic analysis of conductive-radiative heat transfer. Asympt. Anal. **29**, No. 3–4, 323–342 (2002).

[KoEtAl14] Kovtanyuk, A.E., Chebotarev, A.Yu., Botkin, N.D., Hoffmann, K.-H.: The unique solvability of a complex 3D heat transfer problem. J. Math. Anal. Appl. **409**, 808–815 (2014).

[KoCh14] Kovtanyuk, A.E., Chebotarev, A.Yu.: Steady-state problem of complex heat transfer. Comp. Math. Math. Phys. **54**, No. 4, 711–719 (2014).

[Am13a] Amosov, A.A.: Boundary value problem for the radiation transfer equation with reflection and refraction conditions. J. Math. Sci., **191**, issue 2, 101–149 (2013).

[Am13b] Amosov, A.A.: The radiation transfer equation with reflection and refraction conditions. Continuous dependence of solutions on the data and limit passage to the problem with "shooting conditions. J. Math. Sci., **195**, issue 5, 569–608 (2013).

[Am14] Amosov, A.A.: The conjugate boundary value problem for radiation transfer equation with reflection and refraction conditions. J. Math. Sci., **202**, Issue 2, 113–129 (2014).

# Chapter 2
# The Nonstationary Radiative–Conductive Heat Transfer Problem in a Periodic System of Grey Heat Shields. Semidiscrete and Asymptotic Approximations

**A. Amosov**

## 2.1 Introduction

In applications, it is of great importance to study the heat transfer process in periodic media containing vacuum interlayers or cavities through which the heat transfer is realized by radiation. A direct numerical solving of such problems requires considerable computational efforts and becomes, in fact, impossible for a large number of heat transferring elements, especially in the case of two-dimensional and three-dimensional structures. Therefore, it is important to construct effective approximation methods which, in particular, could be based on constructing special homogenizations of the problem.

Formal homogenized equations for such problems were constructed in [Ba82, BaPa89]. But neither rigorous mathematical justification of the homogenized equations nor homogenization of the boundary and initial conditions were provided there. We also refer to [AlHa13a, AlHa13b] for theoretical and numerical analysis of the radiative–conductive heat transfer problem in a periodic perforated domain.

One of the simplest problems of such a type is the problem describing the heat transfer in a system of $n$ grey parallel plate heat shields of width $\varepsilon = 1/n$, separated by vacuum interlayers. The heat shields are layers of a homogeneous heat-conductive material with constant specific heat $c > 0$, heat conductivity coefficient $\lambda > 0$, density $\rho > 0$, and emittance $0 < \theta \leq 1$. The solvability of this problem was studied in [Am10]. Special semidiscrete and asymptotic approximations were proposed in [Am07, AmGu08], and justified in [Am11]. In this article we give a brief and somewhat simplified summary of results [Am10, Am11].

A. Amosov (✉)
National Research University "Moscow Power Engineering Institute",
Krasnokazarmennaya Street 14, 111250 Moscow, Russia
e-mail: AmosovAA@mpei.ru



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_2

## 2.2 Statement and Some Properties of the Radiative–Conductive Heat Transfer Problem in a Periodic System of Grey Shields

### *2.2.1 Physical Statement of the Problem*

With the $i$th heat shield we associate the interval $\Omega_i = (x_{i-1}, x_i)$, $1 \le i \le n$, where $x_i = \varepsilon i$, $\varepsilon = 1/n$. We set

$$\Omega = \bigcup_{i=1}^{n} \Omega_i, \quad Q_{i,T} = \Omega_i \times (0,T), \quad Q_T = \Omega \times (0,T) = \bigcup_{i=1}^{n} Q_{i,T}.$$

For the sake of brevity, we denote the partial derivatives in the following way:

$$D_t = \frac{\partial}{\partial t}, \quad D = \frac{\partial}{\partial x}, \quad D^2 = \frac{\partial^2}{\partial x^2}.$$

The heat transfer inside each shield $\Omega_i$ is described by the heat equation

$$c\rho D_t u = \lambda D^2 u, \quad (x,t) \in Q_{i,T}, \ \ 1 \le i \le n,$$

where $u(x,t)$ is the absolute temperature. We may rewrite this equation in the following form:

$$c\rho D_t u = Dw, \quad w = \lambda Du, \quad (x,t) \in Q_{i,T}, \ \ 1 \le i \le n,$$

where $w = \lambda Du$ coincides up to sign with the heat flux density.

On the interface of heat shields, there is the radiative heat transfer. We recall that the radiation energy flux density on the surface of a grey body is equal to $\theta h(u)$. The function $h(u) = \sigma_0 |u|^3 u$ for $u > 0$ corresponds to the Stefan–Boltzmann law, $0 < \sigma_0$ is the Stefan–Boltzmann constant. The coefficient $\theta$, called the *emittance*, indicates which part of the radiation energy coming from outside is absorbed by the surface. The remaining unabsorbed radiation energy is reflected by the surface. In the case of a grey body, $0 < \theta \le 1$. In the case of an absolutely black body, $\theta = 1$.

Let $x_i$ be a point corresponding to the interface of heat shields $\Omega_i$ and $\Omega_{i+1}$. The heat energy flux coming to the shield surface by heat conductivity is equal to the difference between the energies absorbed and radiated by the surface:

$$\lambda Du\big|_{x=x_i-0} = \lambda Du\big|_{x=x_i+0} = w_i = \varkappa[h(u_i^+) - h(u_i^-)], \quad 0 < i < n,$$

where $\varkappa = \theta/(2-\theta)$ is the apparent emittance. Hereinafter,

$$u_i^+ = u|_{x=x_i+0}, \quad u_i^- = u|_{x=x_i-0}.$$

We emphasize that the temperature $u$ is discontinuous at the points $x_i$, $0 < i < n$. At the same time, the function $w = \lambda Du$ possesses the same limit values $w_i$ at these points.

We assume that the system of heat shields is located in an ambient medium with the temperature $u_\ell(t)$ from the left of the shields and the temperature $u_r(t)$ from the right of the shields. On the left and right boundaries, the radiation heat exchange with the ambient medium is expressed as

$$\lambda Du\big|_{x=x_0+0} = w_0 = \varkappa_\ell[h(u_0^+) - h(u_\ell)],$$

$$\lambda Du\big|_{x=x_n-0} = w_n = \varkappa_r[h(u_r) - h(u_n^-)],$$

where $\varkappa_\ell$ and $\varkappa_r$ are constants, $0 < \varkappa_\ell \le \theta \le 1$, $0 < \varkappa_r \le \theta \le 1$. At the initial time $t = 0$, the absolute temperature is assumed to be given:

$$u\big|_{t=0} = u^0(x), \quad x \in \Omega.$$

*Remark 1.* Note that the data of the problem and its solution essentially depend on the values of small parameter $\varepsilon$. Therefore, one should use the notation $u^0_\varepsilon$, $u_{\ell,\varepsilon}$, $u_{r,\varepsilon}$, $u_\varepsilon$, and $\mathbf{w}_\varepsilon$. However, we omit the superscript $\varepsilon$ for the sake of simplicity.

*Remark 2.* Note that the length $T$ of the time-interval is also an important large parameter since the study of the heat transfer processes on large time-intervals $T \sim 1/\varepsilon$ is of great importance in applications.

### 2.2.2 Well-Known Asymptotic Approximations

In [Ba82] the problem was studied in an infinite system of heat shields under the assumption that the shield width $\varepsilon \to 0$, and formal homogenized equations with arbitrary order in $\varepsilon$ for an asymptotic approximation $v$ to $u$ were constructed. Eliminating terms containing $\varepsilon^m$ with $m \ge 2$ from these homogenized equations, we obtain *an approximation of the radiative heat conductivity*

$$c\rho D_t v = D\left(\lambda_\varepsilon(v)Dv\right), \quad \text{where } \lambda_\varepsilon(v) = \varepsilon\varkappa h'(v). \tag{2.1}$$

In this equation, the role of the heat conductivity coefficient is played by the radiative heat conductivity coefficient $\lambda_\varepsilon(u) = \varepsilon\varkappa h'(v)$. It is obvious that the description of the heat transfer process by means of (2.1) is not perfect since the value of the heat conductivity coefficient $\lambda$ is ignored.

Eliminating terms containing $\varepsilon^m$, $m \ge 3$, in the expansions [Ba82], we find

$$c\rho D_t v = D\left(\lambda_\varepsilon(v)Dv\right), \quad \text{where } \lambda_\varepsilon(v) = \varepsilon\varkappa h'(v)\left(1 - \varepsilon\varkappa\frac{h'(v)}{\lambda}\right). \tag{2.2}$$

Here, the contribution of the heat conductivity to the heat transfer process is taken into account, but for $\varepsilon\varkappa\dfrac{h'(v)}{\lambda} > 1$ the modified radiative heat conductivity coefficient $\lambda_\varepsilon(v)$ becomes negative and Equation (2.2) is inapplicable.

Experts in the field of thermal protection use the following equation instead of Equation (2.2):

$$c\rho D_t v = D\left(\lambda_\varepsilon(v) Dv\right), \quad \text{where } \lambda_\varepsilon(v) = \frac{\varepsilon\varkappa h'(v)}{1+\varepsilon\varkappa\dfrac{h'(v)}{\lambda}} \tag{2.3}$$

Formally, this equation is different from (2.2) by only terms of order $O(\varepsilon^3)$. However, it can be used for any values of parameters.

### 2.2.3 Mathematical Statement of the Original Problem

We formulate the original problem as a problem of finding a function $u$ and a vector-valued function $\mathbf{w} = (w_0, w_1, \dots, w_n)$ such that

1) for all $i = 1, 2, \dots, n$ the function $u$ is such that: $u \in W^{1,2}(Q_{i,T})$, $D^2u \in L^2(Q_{i,T})$ and $u$ is a solution to the parabolic problem

$$c\rho D_t u = \lambda D^2 u, \quad (x,t) \in Q_{i,T}, \tag{2.4}$$

$$\lambda Du\big|_{x=x_{i-1}} = w_{i-1}, \quad \lambda Du\big|_{x=x_i} = w_i, \quad t \in (0,T), \tag{2.5}$$

$$u\big|_{t=0} = u^0, \quad x \in \Omega_i; \tag{2.6}$$

2) $\mathbf{w} \in W^{1,1}(0,T; L^2(\bar\omega^\varepsilon))$ and the following conditions are satisfied:

$$w_i = \varkappa[h(u_i^+) - h(u_i^-)], \quad 0 < i < n, \tag{2.7}$$

$$w_0 = \varkappa_\ell[h(u_0^+) - h(u_\ell)], \quad w_n = \varkappa_r[h(u_r) - h(u_n^-)]. \tag{2.8}$$

Here $\bar\omega^\varepsilon = \{x_i = \varepsilon i,\ 0 \le i \le n\}$ is a grid and $L^2(\bar\omega^\varepsilon)$ is the space of grid functions $Y : \bar\omega^\varepsilon \to \mathbb{R}$ with the norm $\|Y\|_{L^2(\bar\omega^\varepsilon)} = \left(\sum\limits_{i=0}^{n} Y(x_i)^2\varepsilon\right)^{1/2}$.

The functions $u^0$, $u_\ell$, and $u_r$ are given and possess the following properties:

$$u^0 \in W^{1,2}(\Omega_i), \quad 1 \le i \le n; \quad u_{\min} \le u^0(x) \le u_{\max} \quad \forall x \in \Omega,$$

$$u_\ell, u_r \in L^\infty(\mathbb{R}^+), \quad D_t u_\ell, D_t u_r \in L^1(\mathbb{R}^+),$$

$$u_{\min} \le u_\ell(t) \le u_{\max}, \quad u_{\min} \le u_r(t) \le u_{\max} \quad \forall t \in \mathbb{R}^+,$$

where $0 < u_{\min} < u_{\max}$ are some fixed constants. Assume also that

$$\|Du^0\|_{L^2(\Omega)} + \left( |u_0^{0,+} - u_\ell(0)|^2 + \sum_{i=1}^{n-1} (u_i^{0,+} - u_i^{0,-})^2 + |u_r(0) - u_n^{0,-}|^2 \right)^{1/2} \le N,$$

where $N$ is independent of $u^0, u_\ell, u_r$, and $\varepsilon$.

By assumption $D_t u_\ell, D_t u_r \in L^1(\mathbb{R}^+)$, the following limits exist:

$$u_\ell^s = \lim_{t\to\infty} u_\ell(t), \qquad u_r^s = \lim_{t\to\infty} u_r(t).$$

We also assume that $h(u_\ell) - h(u_\ell^s),\ h(u_r) - h(u_r^s) \in L^2(\mathbb{R}^+)$; moreover,

$$\|h(u_\ell) - h(u_\ell^s)\|_{L^2(\mathbb{R}^+)} + \|h(u_r) - h(u_r^s)\|_{L^2(\mathbb{R}^+)} \le N,$$

$$\|D_t h(u_\ell)\|_{L^1(\mathbb{R}^+)} + \|D_t h(u_r)\|_{L^1(\mathbb{R}^+)} \le N,$$

where $N$ is independent of $u_\ell$, $u_r$, and $\varepsilon$.

**Theorem 1.** *The original problem has a unique solution $(u, \mathbf{w})$ and $u$ satisfies the two-sided estimates*

$$u_{\min} \le u(x,t) \le u_{\max}, \quad (x,t) \in Q_T.$$

*For $T \le A/\varepsilon$ the following estimates hold:*

$$\|Du\|_{L^2(Q_T)} + \|\mathbf{w}\|_{L^2(0,T;L^2(\bar\omega^\varepsilon))} \le C(A)\sqrt{\varepsilon},$$

$$\|D_t u\|_{L^2(Q_T)} + \|Du\|_{C([0,T];L^2(\Omega))} + \|D^2 u\|_{L^2(Q_T)} \le C(A).$$

*If $u_\ell^s = u_r^s$, then the following $T$-uniform estimates hold*:

$$\|Du\|_{L^2(Q_T)} + \|\mathbf{w}\|_{L^2(0,T;L^2(\bar\omega^\varepsilon))} \le C\sqrt{\varepsilon},$$

$$\|D_t u\|_{L^2(Q_T)} + \|Du\|_{C([0,T];L^2(\Omega))} + \|D^2 u\|_{L^2(Q_T)} \le C.$$

Hereinafter $C$ and $C(A)$ are some positive constants, which may depend on $N$, $c\rho$, $\lambda$, $\varkappa$, $\varkappa_\ell$, $\varkappa_r$, but do not depend on $\varepsilon$ and $T$.

## 2.3 Semidiscrete Approximations

### *2.3.1 Grids, Grid Functions, and Grid operators*

Introduce the grids

$$\omega^\varepsilon = \{x_i = \varepsilon i,\ 0 < i < n\}, \quad \omega^\varepsilon_{1/2} = \{x_{i-1/2} = \varepsilon(i-1/2),\ 1 \le i \le n\}.$$

and denote by $L^2(\omega^\varepsilon)$ and $L^2(\omega^\varepsilon_{1/2})$ the spaces of grid functions on $\omega^\varepsilon$ and $\omega^\varepsilon_{1/2}$, respectively. We equip these spaces with the norms

$$\|Y\|_{L^2(\omega^\varepsilon)} = \left(\sum_{0<i<n} Y_i^2 \varepsilon\right)^{1/2}, \quad \|Y\|_{L^2(\omega^\varepsilon_{1/2})} = \left(\sum_{1\le i\le n} Y_{i-1/2}^2 \varepsilon\right)^{1/2}.$$

Hereinafter $Y_i$ is the value $Y(x_i)$ of the grid function $Y$ at the grid point, where $i$ is an integer or a half-integer subscript.

Introduce the difference operator $\delta$ by the formula $\delta Y_{i-1/2} = \dfrac{Y_i - Y_{i-1}}{\varepsilon}$, where $i$ is an integer or a half-integer subscript.

### 2.3.2 The Basic Semidiscrete Problem

We begin with a formal derivation of semidiscrete analogs of the original problem (2.4)–(2.8) whose solutions are regarded as approximations to the solution to the original problem.

Averaging Equation (2.4) over $\Omega_i$, we have

$$c\rho D_t[u]_{i-1/2} = \delta w_{i-1/2}, \quad 1 \le i \le n,$$

where $[u]_{i-1/2}(t) = \dfrac{1}{\varepsilon}\int\limits_{\Omega_i} u(x,t)\,dx$.

Note that $u_i^+ \approx [u]_{i+1/2} - \dfrac{\varepsilon}{2\lambda} w_i, \quad u_i^- \approx [u]_{i-1/2} + \dfrac{\varepsilon}{2\lambda} w_i$.

Thus, the conditions (2.7), (2.8) can be approximated as follows:

$$w_i \approx \varkappa\left[h\left([u]_{i+1/2} - \frac{\varepsilon}{2\lambda} w_i\right) - h\left([u]_{i-1/2} + \frac{\varepsilon}{2\lambda} w_i\right)\right], \quad 0 < i < n,$$

$$w_0 \approx \varkappa_\ell\left[h\left([u]_{1/2} - \frac{\varepsilon}{2\lambda} w_0\right) - h(u_\ell)\right], \; w_n \approx \left[h(u_r) - h\left([u]_{n-1/2} + \frac{\varepsilon}{2\lambda} w_n\right)\right].$$

Introducing the functions $U_{i-1/2}(t)$ and $W_i(t)$ that approximate $[u]_{i-1/2}(t)$ and $w_i(t)$, we obtain *the basic semidiscrete problem*

$$c\rho D_t U_{i-1/2} = \delta W_{i-1/2}, \quad 1 \le i \le n, \tag{2.9}$$

$$W_i = \varkappa\left[h\left(U_{i+1/2} - \frac{\varepsilon}{2\lambda} W_i\right) - h\left(U_{i-1/2} + \frac{\varepsilon}{2\lambda} W_i\right)\right], \; 0 < i < n, \tag{2.10}$$

$$W_0 = \varkappa_\ell\left[h\left(U_{1/2} - \frac{\varepsilon}{2\lambda} W_0\right) - h(u_\ell)\right], \tag{2.11}$$

$$W_n = \varkappa_r \left[ h(u_r) - h\left( U_{n-1/2} + \frac{\varepsilon}{2\lambda} W_n \right) \right], \tag{2.12}$$

$$U_{i-1/2}(0) = U^0_{i-1/2}. \quad 1 \le i \le n. \tag{2.13}$$

Hereinafter, $U^0_{i-1/2} = [u^0]_{i-1/2}$.

Equation (2.10) is a nonlinear equation with respect to $W_i$ and defines $W_i$ as a function of $U_{i-1/2}$, $U_{i+1/2}$. Similarly, Equations (2.11) and (2.12) define $W_0$ as a function of $u_\ell$, $U_{1/2}$ and $W_n$ as a function of $U_{n-1/2}$, $u_r$. In whole, we have a Cauchy problem for the system of nonlinear differential equations with respect to the unknowns $U_{i-1/2}(t)$, $1 \le i \le n$.

This problem is of independent interest. As numerical experiments show, its solution $U$ is a good approximation to the solution $u$ to the original problem. Below, the basic semidiscrete problem is used as a basis for constructing other semidiscrete problems which, in turn, can be considered as approximations of initial-boundary value problems approximating the original problem.

### 2.3.3 The First Semidiscrete Problem

We simplify Equations (2.10)–(2.12) by eliminating terms of the form $\frac{\varepsilon}{2\lambda} W_i$ from their right-hand sides. Then we arrive at *the first semidiscrete problem*

$$c\rho D_t U_{i-1/2} = \delta W_{i-1/2}, \quad 1 \le i \le n, \tag{2.14}$$

$$W_i = \varkappa \left[ h(U_{i+1/2}) - h(U_{i-1/2}) \right], \quad 0 < i < n, \tag{2.15}$$

$$W_0 = \varkappa_\ell \left[ h(U_{1/2}) - h(u_\ell) \right], \quad W_n = \varkappa_r \left[ h(u_r) - h(U_{n-1/2}) \right], \tag{2.16}$$

$$U_{i-1/2}(0) = U^0_{i-1/2}, \quad 1 \le i \le n. \tag{2.17}$$

### 2.3.4 The Second Semidiscrete Problem

We return to the basic semidiscrete problem and note that the following approximations hold:

$$W_i = \varkappa \left[ h\left( U_{i+1/2} - \frac{\varepsilon}{2\lambda} W_i \right) - h\left( U_{i-1/2} + \frac{\varepsilon}{2\lambda} W_i \right) \right] \approx$$

$$\approx \varkappa \left[ h(U_{i+1/2}) - h'(U_{i+1/2}) \frac{\varepsilon}{2\lambda} W_i - h(U_{i-1/2}) - h'(U_{i-1/2}) \frac{\varepsilon}{2\lambda} W_i \right],$$

$0 < i < n$, which implies

$$W_i \approx \frac{\varkappa[h(U_{i+1/2}) - h(U_{i-1/2})]}{1 + \dfrac{\varepsilon\varkappa}{2\lambda}\left(h'(U_{i+1/2}) + h'(U_{i-1/2})\right)} \approx \varkappa \int\limits_{U_{i-1/2}}^{U_{i+1/2}} \frac{h'(s)}{1 + \dfrac{\varepsilon\varkappa}{\lambda}h'(s)}\, ds =$$
$$= \varkappa[H(U_{i+1/2}) - H(U_{i-1/2})],$$

where $H(u) = \int\limits_0^u \dfrac{h'(s)}{1 + \dfrac{\varepsilon\varkappa}{\lambda}h'(s)}\, ds$. In the same manner

$$W_0 = \varkappa_\ell \left[h\left(U_{1/2} - \frac{\varepsilon}{2\lambda}W_0\right) - h(u_\ell)\right] \approx \varkappa_\ell \left[h\left(U_{1/2}\right) - h'(U_{1/2})\frac{\varepsilon}{2\lambda}W_0 - h(u_\ell)\right],$$

which implies

$$W_0 \approx \frac{\varkappa_\ell[h(U_{1/2}) - h(u_\ell)]}{1 + \dfrac{\varepsilon\varkappa_\ell}{2\lambda}h'(U_{1/2})} \approx \varkappa_\ell \int\limits_{u_\ell}^{U_{1/2}} \frac{h'(s)}{1 + \dfrac{\varepsilon\varkappa_\ell}{2\lambda}h'(s)}\, ds = \varkappa_\ell[H_\ell(U_{1/2}) - H_\ell(u_\ell)],$$

where $H_\ell(u) = \int\limits_0^u \dfrac{h'(s)}{1 + \dfrac{\varepsilon\varkappa_\ell}{2\lambda}h'(s)}\, ds$. Similarly,

$$W_n \approx \frac{\varkappa_r[h(u_r) - h(U_{n-1/2})]}{1 + \dfrac{\varepsilon\varkappa_r}{2\lambda}h'(U_{n-1/2})} \approx \varkappa_r[H_r(u_r) - H_r(U_{n-1/2})],$$

where $H_r(u) = \int\limits_0^u \dfrac{h'(s)}{1 + \dfrac{\varepsilon\varkappa_r}{2\lambda}h'(s)}\, ds$.

These arguments lead to the *second semidiscrete problem*

$$c\rho D_t U_{i-1/2} = \delta W_{i-1/2}, \quad 1 \le i \le n, \tag{2.18}$$
$$W_i = \varkappa[H(U_{i+1/2}) - H(U_{i-1/2})], \quad 0 < i < n, \tag{2.19}$$
$$W_0 = \varkappa_\ell[H_\ell(U_{1/2}) - H_\ell(u_\ell)], \quad W_n = \varkappa_r[H_r(u_r) - H_r(U_{n-1/2})], \tag{2.20}$$
$$U_{i-1/2}(0) = U^0_{i-1/2}, \quad 1 \le i \le n. \tag{2.21}$$

## 2.4 Asymptotic Approximations

### 2.4.1 The First Homogenized Problem

As a differential analog of the first semidiscrete problem and an asymptotic approximation of the original problem we consider the following initial-boundary value problem (called *the first homogenized problem*)

$$c\rho D_t v = \varepsilon \varkappa D^2 h(v), \quad (x,t) \in Q_T^\varepsilon = \Omega^\varepsilon \times (0,T), \tag{2.22}$$

$$-\varepsilon \varkappa Dh(v)\big|_{x=\varepsilon/2} + c\rho \frac{\varepsilon}{2} D_t v_\ell + \varkappa_\ell h(v_\ell) = \varkappa_\ell h(u_\ell), \quad t \in (0,T), \tag{2.23}$$

$$\varepsilon \varkappa Dh(v)\big|_{x=1-\varepsilon/2} + c\rho \frac{\varepsilon}{2} D_t v_r + \varkappa_r h(v_r) = \varkappa_r h(u_r), \quad t \in (0,T), \tag{2.24}$$

$$v\big|_{t=0} = v^0, \quad x \in \Omega^\varepsilon = (\varepsilon/2, X - \varepsilon/2). \tag{2.25}$$

Here $v_\ell = v\big|_{x=\varepsilon/2}$, $v_r = v\big|_{x=1-\varepsilon/2}$. The function $v^0$ is piecewise linear: $v^0(x) = \sum_{i=1}^{n} [u^0]_{i-1/2} \widehat{e}(x - x_{i-1/2})$. where $\widehat{e}(x) = 1 - |x|/\varepsilon$ for $x \in [-\varepsilon, \varepsilon]$, $\widehat{e}(x) = 0$ for $x \notin [-\varepsilon, \varepsilon]$.

The values $v_{i-1/2}(t) = v(x_{i-1/2}, t)$ are considered as approximations to the values $[u]_{i-1/2}(t)$ and $u(x_{i-1/2}, t)$, $1 \le i \le n$.

*Remark 3.* Note that Equation (2.22) coincides with an approximation of the radiative heat conductivity (2.1).

*Remark 4.* The solution to the original problem depends on the heat conductivity coefficient $\lambda$. At the same time, there is no information about the value of $\lambda$ in the first semidiscrete and in the first homogenized problems. Hence one can expect that the solutions to these problems are rather rough approximations to the solution to the original problem.

### 2.4.2 The Second Homogenized Problem

The second semidiscrete problem can be regarded as an approximation of the following initial-boundary value problem (*the second homogenized problem*):

$$c\rho D_t v = \varepsilon \varkappa D^2 H(v), \quad (x,t) \in Q_T^\varepsilon, \tag{2.26}$$

$$-\varepsilon \varkappa DH(v)\big|_{x=\varepsilon/2} + c\rho \frac{\varepsilon}{2} D_t v_\ell + \varkappa_\ell H_\ell(v_\ell) = \varkappa_\ell H_\ell(u_\ell), \quad t \in (0,T), \tag{2.27}$$

$$\varepsilon \varkappa DH(v)\big|_{x=1-\varepsilon/2} + c\rho \frac{\varepsilon}{2} D_t v_r + \varkappa_r H_r(v_r) = \varkappa_r H_r(u_r), \quad t \in (0,T), \tag{2.28}$$

$$v\big|_{t=0} = v^0, \quad x \in \Omega^\varepsilon. \tag{2.29}$$

*Remark 5.* Since $\varepsilon \varkappa DH(v) = \lambda_\varepsilon(v) Dv$, where $\lambda_\varepsilon(v) = \dfrac{\varepsilon \varkappa h'(v)}{1 + \varepsilon \dfrac{\varkappa h'(v)}{\lambda}}$, Equations (2.26) and (2.3) coincide.

*Remark 6.* Since the right-hand sides of Equations (2.22) and (2.26) contain the factor $\varepsilon$, the homogenized problems become singularly perturbed. It should be taken into account that the boundary conditions (2.23), (2.24), (2.27), and (2.28) have a nonstandard form.

*Remark 7.* Semidiscrete problems are an important intermediate stage between the original problem and its homogenizations. In particular, such problems are essentially used for estimating an error of the asymptotic approximations.

## 2.5 Semidiscrete Problems. Existence and Uniqueness of a Solution. A Priori Estimates for Solutions

By a solution to the first (second) semidiscrete problem we mean a function $U \in C^1([0,T]; L^2(\omega^\varepsilon_{1/2}))$ such that $U$ satisfies Equation (2.14) (Equation (2.18)), where $W$ is defined by (2.15) and (2.16) ($W$ is defined by (2.19) and (2.20)), and $U$ satisfies the initial condition $U|_{t=0} = U^0$.

**Theorem 2.** *A solution to the first (second) semidiscrete problem exists and is unique. The following the $T$-uniform estimates hold:*

$$u_{min} \le U \le u_{max}, \quad (x,t) \in \omega^\varepsilon_{1/2} \times [0,T],$$

$$\|D_t U\|_{L^2(0,T;L^2(\omega^\varepsilon_{1/2}))} + \sqrt{\varepsilon} \|\delta U\|^2_{C([0,T];L^2(\bar\omega^\varepsilon))} \le C.$$

## 2.6 Error Estimates for Solutions to Semidiscrete Problems

Let $u$ be a solution to the original problem. We denote by $\mathbf{u}$ a function on $\bar\omega^\varepsilon_{1/2} \times [0,T]$ with the values $\mathbf{u}_{i-1/2}(t) = u(x_{i-1/2}, t)$, $1 \le i \le n$.

**Theorem 3.** *Let $U$ be a solution to the first semidiscrete problem. Then for $T \le A/\varepsilon$ the following estimate holds*:

$$\|U - \mathbf{u}\|_{L^2(0,T;L^2(\omega^\varepsilon_{1/2}))} \le C(A)\sqrt{\varepsilon}.$$

*If $u_\ell^s = u_r^s$, then the following $T$-uniform estimate holds*:

$$\|U - \mathbf{u}\|_{L^2(0,T;L^2(\omega^\varepsilon_{1/2}))} \le C\sqrt{\varepsilon},$$

**Theorem 4.** *Let $U$ be a solution to the second semidiscrete problem. Then for $T \le A/\varepsilon$ the following estimate holds*:

$$\|U - \mathbf{u}\|_{L^2(0,T;L^2(\omega^\varepsilon_{1/2}))} \le C(A)\varepsilon.$$

*Remark 8.* The same accuracy orders in $\varepsilon$ as established in this section were observed for the first and second semidiscrete methods in computational experiments. These experiments also show that the basic semidiscrete method possesses the first-order accuracy with respect to $\varepsilon$ and has an error less than the second semidiscrete method has. Unfortunately, a rigorous mathematical analysis of this method has not been provided yet.

## 2.7 Homogenized Problems. Existence and Uniqueness of a Solution. A Priori Estimates and Comparison Theorem

By a solution to the first (second) homogenized problem we mean a function $v \in W^{1,2}(Q_T^\varepsilon)$ such that:

1) $Dv \in L^\infty(0,T;L^2(\Omega^\varepsilon))$, $D_t v \in L^2(Q_T^\varepsilon)$, $D^2H(v) \in L^2(Q_T^\varepsilon)$, $v_\ell, v_r \in W^{1,2}(0,T)$;
2) $v$ satisfies Equation (2.22) (Equation (2.26)) in $L^2(Q_T^\varepsilon)$, the boundary conditions (2.23) and (2.24) (the boundary conditions (2.27) and (2.28)) in $L^2(0,T)$, and the initial condition $u|_{t=0} = u^0$ in the classical sense.

**Theorem 5.** *A solution to the first (second) homogenized problem exists and is unique. The following $T$-uniform estimates hold*:

$$u_{\min} \le v \le u_{\max} \quad \forall\ (x,t) \in Q_T^\varepsilon,$$
$$\|D_t v\|_{L^2(Q_T^\varepsilon)} + \sqrt{\varepsilon}\|Dv\|_{L^\infty(0,T;L^2(\Omega^\varepsilon))} \le C,$$
$$\sqrt{\varepsilon}\|D_t v_\ell\|_{L^2(0,T)} + \sqrt{\varepsilon}\|D_t v_r\|_{L^2(0,T)} \le C.$$

*For $T \le A/\varepsilon$ the following estimate holds*:

$$\sqrt{\varepsilon}\|Dv\|_{L^2(Q_T^\varepsilon)} + \varepsilon^{3/4}\|Dv\|_{L^2(0,T;L^\infty(\Omega^\varepsilon))} + \varepsilon^{5/4}\|D^2 v\|_{L^2(Q_T^\varepsilon)} \le C(A).$$

*If $u_\ell^s = u_r^s$, then the following $T$-uniform estimate holds*:

$$\sqrt{\varepsilon}\|Dv\|_{L^2(Q_T^\varepsilon)} + \varepsilon^{3/4}\|Dv\|_{L^2(0,T;L^\infty(\Omega^\varepsilon))} + \varepsilon^{5/4}\|D^2 v\|_{L^2(Q_T^\varepsilon)} \le C.$$

We proved the following comparison theorem.

**Theorem 6.** *Let $v^{(1)}$ and $v^{(2)}$ be two solutions to the first (second) homogenized problem corresponding to the data $v^{0,(1)}$, $u_\ell^{(1)}$, $u_r^{(1)}$ and $v^{0,(2)}$, $u_\ell^{(2)}$, $u_r^{(2)}$, respectively. Suppose that $v^{0,(1)} \le v^{0,(2)}$, $u_\ell^{(1)} \le u_\ell^{(2)}$, and $u_r^{(1)} \le u_r^{(2)}$. Then $v^{(1)} \le v^{(2)}$.*

## 2.8 Error Estimates for Solutions to the Homogenized Problems

We denote by $\mathbf{v}$ a function on $\bar{\omega}_{1/2}^\varepsilon \times [0,T]$ with the values $\mathbf{v}_{i-1/2}(t) = v(x_{i-1/2},t)$, $1 \le i \le n$.

**Theorem 7.** *Let $v$ be a solution to the first homogenized problem. Then for $T \le A/\varepsilon$ the following estimate holds*:

$$\|\mathbf{v} - \mathbf{u}\|_{L^2(0,T;L^2(\omega_{1/2}^\varepsilon))} \le C(A)\sqrt{\varepsilon}.$$

*If $u_\ell^s = u_r^s$, then the following $T$-uniform estimate holds*:

$$\|\mathbf{v} - \mathbf{u}\|_{L^2(0,T;L^2(\omega_{1/2}^\varepsilon))} \le C\sqrt{\varepsilon}.$$

**Theorem 8.** *Let $v$ be a solution to the second homogenized problem. Then for $T \le A/\varepsilon$ the following estimate holds*:

$$\|\mathbf{v} - \mathbf{u}\|_{L^2(0,T;L^2(\omega_{1/2}^\varepsilon))} \le C(A)\varepsilon^{3/4}.$$

*Remark 9.* In computational experiments, the convergence of order $O(\sqrt{\varepsilon})$ and $O(\varepsilon)$ has been observed for the first and second homogenized problems. In [Am11], for the second homogenized problem we were able to establish only the convergence of order $O(\varepsilon^{3/4})$. At present, it is not clear whether this fact is caused by the nature of the case or by the lack of tools for investigation.

**Acknowledgements** The work was financially supported by the Russian Foundation for Basic Research (grant 13-01-00201a), the Ministry of Education and Science of the Russian Federation (assignment No- 1.756.2014/K) and Board grants of the President of Russia (grant NSh-2081.2014.1).

## References

[Ba82] Bakhvalov, N. S.: Averaging of the heat-transfer process in periodic media with radiation. Differ. Equations, **17**, 1094–1100 (1982).

[BaPa89] Bakhvalov, N.S. and Panasenko, G.P.: Averaging processes in periodic media. Kluwer, Dordrecht etc. (1989).

[AlHa13a] Allaire, G. and Habibi, Z.: Second order corrector in the homogenization of a conductive-radiative heat transfer problem. Discrete and Continuous Dynamical Systems - Series B, **18**, Issue 1, 1–36 (2013).

[AlHa13b] Allaire, G. and Habibi, Z.: Homogenization of a conductive, convective, and radiative heat transfer problem in a heterogeneous domain. SIAM J. Math. Anal. **45**, Issue 3, 1136–1178 (2013).

[Am10] Amosov, A.A.: Nonstationary radiative-conductive heat transfer problem in a periodic system of grey heat shields. J. Math. Sci., **169**, No. 1, 1-45 (2010).

[Am07] Amosov, A.A.: Semidiscrete and asymptotic approximations to a solution to the heat transfer problem in a system of heat shields under radiation [in Russian]. In: Modern Problems of Mathematical Simulating, 21-36, Rostov-na-Donu (2007).

[AmGu08] Amosov, A.A., Gulin, V.V.: Semidiscrete and asymptotic approximations in the heat transfer problem in a system of heat shields under radiation [in Russian]. MPEI Bulletin, No. 6, 5–15 (2008).

[Am11] Amosov, A. A.: Semidiscrete and asymptotic approximations for the nonstationary radiative-conductive heat transfer problem in a periodic system of grey heat shields. J. Math. Sci., **176**, No. 3, 361–408 (2011).

# Chapter 3
# A Mixed Impedance Scattering Problem for Partially Coated Obstacles in Two-Dimensional Linear Elasticity

**C.E. Athanasiadis, D. Natroshvili, V. Sevroglou, and I.G. Stratis**

## 3.1 Introduction

In this chapter, we consider the scattering problem of time-harmonic plane elastic waves by a non-penetrable partially coated obstacle embedded in a $(N+1)$-layered background medium. Without loss of generality in this work we only consider the case $N=1$ which means that our obstacle is buried in a two-layered background medium. From the point of view of applications, a medium of this type which is consisted by a finite number of homogeneous layers (having a nested body) appears in remote sensing, nondestructive testing, radars, etc. [AtNa10].

Let $D$ denote the piecewise homogeneous medium which is a bounded and closed subset of $\mathbb{R}^2$ with a boundary $S_0$. We also denote by $D_0$ the exterior of $D$, i.e., $D_0 = \mathbb{R}^2 \setminus \overline{D}$. The interior of $D$ is divided into two disjoint layers $D_1$ and $D_2$, with $D_2$ being the non-penetrable obstacle, which is an open bounded domain having a $C^2$-boundary $\Gamma$. The layer $D_1$ is the domain between $S_0$ and $\Gamma$ and clearly $\partial D_1 = \Gamma \cup S_0$. Furthermore, we assume that the domains $D_j$, $j=0,1$ are occupied by dissimilar homogeneous isotropic elastic media with densities $\rho_j$ and Lam*é* constants $\lambda_j$ and $\mu_j, j=0,1$, respectively, while $D_2$ is occupied by a rigid body which is treated as a non-penetrable obstacle. The boundary $\Gamma$ of $D_2$ is divided into two parts, a Dirichlet (rigid) one $\Gamma_D$ and a Robin (impedance) one $\Gamma_I$. These two parts

C.E. Athanasiadis • I.G. Stratis
National and Kapodistrian University of Athens, Panepistimiopolis, 15784 Athens, Greece
e-mail: cathan@math.uoa.gr; istratis@math.uoa.gr

D. Natroshvili
Georgian Technical University, Tbilisi, Georgia
e-mail: natrosh@hotmail.com

V. Sevroglou (✉)
University of Piraeus, Piraeus, Greece
e-mail: bsevro@unipi.gr



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_3

are simply connected disjoint sub-manifolds and the boundary $\Gamma$ has a dissection $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_I$. The latter part is due to a coating on the Robin part of the boundary with a material of constant surface impedance $c$. We note here that the case $\Gamma_I = \emptyset$ corresponds to a rigid body, whereas the case $\Gamma_D = \emptyset$, $c = 0$ yields a Neumann boundary condition which corresponds to a cavity.

Our paper is organized as follows: In Section 3.2 we present the direct scattering problem in a dyadic form by describing it with a mixed impedance transmission boundary value problem. Issues of existence, uniqueness and stability are also briefly discussed. In Section 3.3 we establish the unique determination of both the non-penetrable obstacle $D_2$ with its boundary condition and the interface $S_0$ from a knowledge of the far-field pattern for incident elastic plane-waves. Our uniqueness will be based on a generalization of a mixed reciprocity relation; that is for the uniqueness of the partially coated obstacle $D_2$ and its boundary conditions, whereas the uniqueness of the penetrable interface $S_0$ between the two-layered media will be based on a uniqueness result of a specific mixed impedance Robin boundary value problem.

The results of the present work concerning the direct scattering problem have been studied extensively in [AtNa14, AtNa11, NaTe01] and they are very useful for the results obtained in Section 3.3 concerning inverse problems which deal with partially coated obstacles in the case of elastic layered background media. A more analytical study for the latter is in progress and will be communicated separately.

## 3.2 The Direct Scattering Problem

We consider the direct scattering problem of a given harmonic elastic wave by a partially coated obstacle $D_2$ in $\mathbb{R}^2$ buried in a two-layered piecewise homogeneous medium. In what follows we consider the problem in a dyadic formulation due to the dyadic nature of the fundamental Green's function. We mention here that, as Twersky [Tw67] pointed out for electromagnetic waves, the dyadic scattering problem – because of its higher symmetry – is easier than the corresponding vector scattering problem. The reason for that is the fact that the propagation vector alone suffices to specify the incident field. For dyadic formulation of various scattering problems in two-dimensional elasticity, we refer to [AtSe06, PeSe03], whereas properties of dyadics can be found in the excellent source of the book [Ta94].

We continue our analysis by introducing the free-space Green's dyadic of the Navier equation in $\mathbb{R}^2$, given by [Se05]

$$\widetilde{\Gamma}(r,r') = \frac{i}{4}\{\frac{1}{\mu}\widetilde{I}H_0^{(1)}(k_s|r-r'|) - \frac{1}{\omega^2}\nabla_r \otimes \nabla_r \left[H_0^{(1)}(k_p|r-r'|) - H_0^{(1)}(k_s|r-r'|)\right]\} \tag{3.1}$$

for $r, r' \in \mathbb{R}^2$, $r \neq r'$, where $H_0^{(1)}(z)$ is the cylindrical Hankel function of the first kind and zero order, $\tilde{I}$ is the $2 \times 2$ identity matrix and the "$\sim$" in this paper will denote dyadic fields. Furthermore, "$\nabla_r$" denotes the action of the gradient operator with respect to the variable $r$, and "$\otimes$" is the juxtaposition between two vectors (this gives a dyadic). The fundamental solution $\widetilde{\Gamma}(r, r')$ satisfies the following equation

$$\Delta^* \widetilde{\Gamma}(r, r') + \rho\, \omega^2\, \widetilde{\Gamma}(r, r') = -\widetilde{I}\, \delta(r - r'), \quad r, r' \in \mathbb{R}^2$$

where $\delta(r - r')$ represents the Dirac measure concentrated at the point $r$ and $\Delta^*$ stands for the Lam*é* operator (see (3.11)).

Our scatterer $D_2$ is irradiated either by incident elastic plane-waves, or by elastic point-sources. Elastic plane-waves and point-sources are of special interest and will play an important role in the study of the inverse problem (see later, Section 3.3).

For the case of plane-waves we have the following analysis: An incident plane P-wave (pressure wave) is given by

$$\widetilde{u}_p^{inc}(r, \hat{d}) := \hat{d} \otimes \hat{d}\, e^{ik_{0,p}\, r \cdot \hat{d}} \tag{3.2}$$

where $\hat{d}$ is the incident direction of propagation, i.e., $\hat{d} \in \Omega := \{ r \in \mathbb{R}^2 : |r| = 1 \}$, whereas a plane S-wave (shear wave) is of the form

$$\widetilde{u}_s^{inc}(r, \hat{d}) := (\widetilde{I} - \hat{d} \otimes \hat{d})\, e^{ik_{0,s}\, r \cdot \hat{d}} \tag{3.3}$$

where $\widetilde{I} - \hat{d} \otimes \hat{d} = \hat{d}^{\perp} \otimes \hat{d}^{\perp}$ with $\hat{d}^{\perp}$ being the polarization vector. Due to the incident plane-wave the corresponding scattered field is denoted by $\widetilde{u}^{sct}(r, \hat{d})$, given by

$$\widetilde{u}^{sct}(r, \hat{d}) := \widetilde{u}_p^{sct}(r, \hat{d}) + \widetilde{u}_s^{sct}(r, \hat{d}), \tag{3.4}$$

where the dyadics $\widetilde{u}_p^{sct}$, $\widetilde{u}_s^{sct}$ are referred to as the longitudinal (pressure) and transverse (shear) parts of $\widetilde{u}^{sct}$, respectively. Then, the total elastic field $\widetilde{u}^t$ is the superposition of the incident field and the corresponding scattered field, i.e.,

$$\widetilde{u}^t(r, \hat{d}) = \widetilde{u}^{inc}(r, \hat{d}) + \widetilde{u}^{sct}(r, \hat{d}) \tag{3.5}$$

*Remark*: The incident plane-wave field $\widetilde{u}^{inc}$ could also be considered as the superposition of P and S-wave, in the form

$$\widetilde{u}^{inc}(r, \hat{d}) = \hat{d} \otimes \hat{d}\, e^{ik_{0,p}\, r \cdot \hat{d}} + (\widetilde{I} - \hat{d} \otimes \hat{d})\, e^{ik_{0,s}\, r \cdot \hat{d}} \tag{3.6}$$

It is worth mentioning that, due to elasticity theory, independently of what form the incident field is, (3.2), (3.3), or (3.6) the corresponding scattered field always will be consisted of the scattered P-wave and S-wave as well (see (3.4)).

Next, we present the notation for elastic point-sources. We denote by $\widetilde{u}_a^{inc}$ an incident point-source due to a source located at a point $a \in \mathbb{R}^2$ with corresponding scattered field $\widetilde{u}_a^{sct}$ and total field given by

$$\widetilde{u}_a^t(r) = \widetilde{u}_a^{inc}(r) + \widetilde{u}_a^{sct}(r). \tag{3.7}$$

From the mathematical point of view our direct scattering problem for a partially coated obstacle embedded in a two-layered homogenous medium is described by the following mixed impedance transmission boundary value problem:

Given $\widetilde{h} \in H_2^{1/2}(\Gamma_D)$, $\widetilde{g} \in H_2^{-1/2}(\Gamma_I)$, $\widetilde{p} \in H_2^{-1/2}(S_0)$ and $\widetilde{f} \in H_2^{1/2}(S_0)$, find functions $\widetilde{u} \in H^1_{2,loc}(D_0)$ and $\widetilde{\upsilon} \in H^1_{2,loc}(D_1)$ satisfying the differential equations

$$\Delta^* \widetilde{u} + \rho_0\, \omega^2\, \widetilde{u} = \widetilde{0} \quad \text{in } D_0,$$
$$\Delta^* \widetilde{\upsilon} + \rho_1\, \omega^2\, \widetilde{\upsilon} = \widetilde{0} \quad \text{in } D_1,$$

the mixed impedance boundary conditions

$$\widetilde{\upsilon} = \widetilde{h} \quad \text{on } \Gamma_D, \tag{3.8}$$
$$T\,\widetilde{\upsilon} + i\,\omega\, c\,\widetilde{\upsilon} = \widetilde{g} \quad \text{on } \Gamma_I, \tag{3.9}$$

the transmission boundary conditions

$$\widetilde{u} - \widetilde{\upsilon} = \widetilde{f}, \quad T\widetilde{u} - T\widetilde{\upsilon} = \widetilde{p} \quad \text{on } S_0$$

and the Sommerfeld–Kupradze radiation conditions

$$\lim_{|r|\to\infty} \sqrt{r} \left( \frac{\partial \widetilde{u}_\beta(r)}{\partial r} - ik_\beta\, \widetilde{u}_\beta(r) \right) = \widetilde{0}, \ \ \beta = p, s, \tag{3.10}$$

where the explicit expression for $\Delta^* \widetilde{u}$ is given by

$$\Delta^* \widetilde{u}(r) := \mu(r)\, \Delta \widetilde{u}(r) + (\,\lambda(r) + \mu(r)\,)\, \operatorname{grad} \operatorname{div} \widetilde{u}(r) \tag{3.11}$$

while the stress operator $T$ acting on $\widetilde{u}$, with outward normal unit vector $n = (n_1, n_2)$ at the point $r \in \Gamma$ or $S_0$, is defined as:

$$T\widetilde{u} := (2\mu\, n \cdot \operatorname{grad} + \lambda\, n \operatorname{div} + \mu\, n \times \operatorname{curl})\, \widetilde{u}, \tag{3.12}$$

Throughout this paper $c$ (the surface impedance for the boundary $\Gamma_I$) will be considered a positive constant, while the above piecewise constant functions, $\lambda_j$ and $\mu_j$ are the Lamé constants, $\rho_j$ are the densities of the elastic layers, satisfying the relations

$$\mu_j > 0, \ \ \lambda_j + 2\mu_j > 0, \ \ \rho_j > 0, \ \ j = 0, 1$$

and $\omega \in \mathbb{R}$ is the so-called frequency parameter. The Sommerfeld-Kupradze radiation conditions (3.10) are assumed to hold uniformly in all directions $\hat{r} = r/|r|$. Further, $k_{j,p}, k_{j,s}, j = 0,1$ are the wave numbers for the longitudinal and the transverse waves, respectively, given by

$$k_{j,p} = \omega\sqrt{\frac{\rho_j}{\lambda_j + 2\mu_j}},\ k_{j,s} = \omega\sqrt{\frac{\rho_j}{\mu_j}},\ j = 0,1 \tag{3.13}$$

Note that in (3.10) $k_\beta = k_{0,\beta}$, $\beta = p,s$ (see also relations (3.2), (3.3) and (3.6)).

It is very important to mention that, the incident field for our scattering problem (3.8), (3.10) could be a plane-wave or a point-source as well. In the latter case let $\widetilde{u}_a^{inc}(r) := \widetilde{\Gamma}(r,a)$, $r \neq a$ be a source point, with

$$\widetilde{\Gamma}(r,a) = -\frac{i}{4\rho_0\omega^2}\nabla_r\nabla_r H_0^{(1)}(k_{0,p}|r-a|)$$

$$+\frac{i}{4\rho_0\omega^2}(\nabla_r\nabla_r + k_{0,s}\widetilde{I})\,H_0^{(1)}(k_{0,s}|r-a|)\,\}, \tag{3.14}$$

located at a point with position vector $a \in \mathbb{R}^2$. This is actually similar to the fundamental solution (3.1).

The solvability of the direct scattering problem (3.8)–(3.10) for partially piecewise homogeneous and inhomogeneous layered obstacles has been proved via the potential method in [AtNa14]. In particular, we reduced the mixed impedance transmission problem to an equivalent system of pseudodifferential equations and proved that the corresponding boundary operators are invertible in appropriate Bessel potential and Besov spaces. We also established existence of solution of the mixed impedance problem as well as regularity results. Furthermore, the case of Lipschitz surfaces was also studied and treated separately. For a detailed analysis, we refer to [AtNa14] and [AtNa11].

## 3.3 The Inverse Scattering Problem

In this section we establish uniqueness results for the inverse elastic scattering problem. To the best of the author's knowledge, there are no results concerning the unique determination of both the non-penetrable partially coated obstacle embedded in a two-layered piecewise homogeneous medium and the interface between the layered media. Therefore, in this article we prove that both the penetrable interface $S_0$ and the non-penetrable obstacle $D_2$ with its physical property (3.8), (3.9) can be suniquely determined by a knowledge of the far-field pattern of the scattered field. The obtained results also hold for a partially coated obstacle buried in multi-layered media; they are valid as well for the three-dimensional case.

In what follows we need the following Sobolev space setting:
Let $H_2^1(D), H_2^1(\mathbb{R}^2 \setminus D)$ be the classical Sobolev spaces with $H_2^{1/2}(\Gamma)$ being their trace space. In order to study mixed impedance boundary value problems with boundary conditions such as in relations (3.8) and (3.9) we need Sobolev spaces on an open part of the boundary. Hence, for a proper subset $\Gamma_0 \subset \Gamma$, let

$$H_2^{1/2}(\Gamma_0) := \{\widetilde{u}|_{\Gamma_0} : \widetilde{u} \in H_2^{1/2}(\Gamma)\}$$

$$\left(H_2^{1/2}(\Gamma_0)\right)^* := \{\widetilde{u} \in H_2^{1/2}(\Gamma) : \text{supp } \widetilde{u} \subseteq \overline{\Gamma}_0\}$$

$$H_2^{-1/2}(\Gamma_0) := \left(\left(H_2^{1/2}(\Gamma_0)\right)^*\right)' \text{ (i.e., the dual space of } \left(H_2^{1/2}(\Gamma_0)\right)^*)$$

$$\left(H_2^{-1/2}(\Gamma_0)\right)^* := \left(H_2^{1/2}(\Gamma_0)\right)' \text{ (i.e., the dual space of } H_2^{1/2}(\Gamma_0))$$

We will also need, corresponding to (3.8)–(3.10), a specific *interior mixed impedance boundary value problem* in $D_1$, which is stated as follows: Find $\widetilde{u} \in H_2^1(D_1)$ such that

$$\Delta^* \widetilde{u} + \rho_1 \omega^2 \widetilde{u} = \widetilde{F} \quad \text{in } D_1, \tag{3.15}$$

$$\widetilde{u} = \widetilde{h} \quad \text{on } \Gamma_D, \tag{3.16}$$

$$T\widetilde{u} + i\omega c \widetilde{u} = \widetilde{g} \quad \text{on } \Gamma_I, \tag{3.17}$$

$$T\widetilde{u} + i\omega \nu \widetilde{u} = \widetilde{f} \quad \text{on } S_0, \tag{3.18}$$

where $\widetilde{F} \in L_2(D_1)$, $\widetilde{h} \in H_2^{1/2}(\Gamma_D)$, $\widetilde{g} \in H_2^{-1/2}(\Gamma_I)$, $\widetilde{f} \in H_2^{-1/2}(S_0)$ and $\nu : S_0 \to \mathbb{R}$ with $\nu \leq 0$ a continuously differentiable function ($\nu$ does not vanish identically). A study for the problem (3.15)–(3.18) has been considered, and in particular integral representations of solutions have been derived. Basic uniqueness, existence and regularity results for these mixed impedance problems have been established as well. We mention here that the boundary value problem (3.15)–(3.18) also possesses a unique solution if we assume (without loss of generality) that $\widetilde{h} = \widetilde{0}$ on $\Gamma_D$. This follows from the setting of the following space

$$H_2^1(D_1; \Gamma_D) := \{\widetilde{u} \in H_2^1(D_1) : \widetilde{u}|_{\Gamma_D} = \widetilde{0}\} \tag{3.19}$$

by easily noting that $H_2^1(D_1; \Gamma_D)$ is a closed subspace of $H_2^1(D_1)$.

The elastic inverse scattering problem by a mixed partially coated obstacle embedded in a two-layered background medium is described by the following mixed impedance transmission boundary value problem: *Determine uniquely the partially coated obstacle $D_2$ and its physical properties as well as the penetrable interface (boundary) $S_0$ of the piecewise homogeneous medium $D$*, if the following conditions hold:

$$\begin{aligned}
&\Delta^* \widetilde{u} + \rho_0\, \omega^2\, \widetilde{u} = \widetilde{0} \quad \text{in } D_0,\\
&\Delta^* \widetilde{u} + \rho_1\, \omega^2\, \widetilde{u} = \widetilde{0} \quad \text{in } D_1,\\
&\widetilde{u} = \widetilde{0} \quad \text{on } \Gamma_D,\\
&T\widetilde{u} + i\,\omega\, c\, \widetilde{u} = \widetilde{0} \quad \text{on } \Gamma_I,\\
&\widetilde{u}^{ext} = \widetilde{u}^{int}, \quad T_e\, \widetilde{u}^{ext} = T_i\, \widetilde{u}^{int} \quad \text{on } S_0\\
&\lim_{|r|\to\infty} \sqrt{r} \left( \frac{\partial \widetilde{u}_\beta^{sct}}{\partial r} - ik\beta\, \widetilde{u}_\beta^{sct} \right) = \widetilde{0}, \quad \beta = p, s,
\end{aligned} \tag{3.20}$$

where $\widetilde{u} = \widetilde{u}^{inc} + \widetilde{u}^{sct}$ in $D_0 \cup D_1$, $\widetilde{u}^{int}$, $\widetilde{u}^{ext}$ denote the interior and exterior one sides limits (traces) on the interface $S_0$, respectively, and the notations $T_i\, \widetilde{u}^{int}$, $T_e\, \widetilde{u}^{ext}$ are given by relation (3.12), if we replace the Lam*é* constants $\lambda$, $\mu$ with the appropriate values $\lambda_j$, $\mu_j$, $j = 0$ or $1$.

In order now to establish uniqueness, let us first prove an essential mixed reciprocity relation. We need to consider incident plane-waves as well as point-sources. In what follows, and for the reader's convenience we recall the following notation: For incident plane fields: $\widetilde{u}^t(r;\hat{d}) = \widetilde{u}^{inc}(r;\hat{d}) + \widetilde{u}^{sct}(r;\hat{d})$, whereas for point-sources $\widetilde{u}_a^t(r) = \widetilde{u}_a^{inc}(r) + \widetilde{u}_a^{sct}(r)$. Our incident point-source field is given by $\widetilde{u}_a^{inc}(r) := \widetilde{\Gamma}(r,a)$, $r \neq a$, where $\widetilde{\Gamma}(r,a)$ is given by (3.14). We also denote by $\widetilde{u}_\infty(\hat{r};\hat{d})$ and $\widetilde{u}_{\infty,a}(\hat{r})$ the far-field patterns of $\widetilde{u}^{sct}(r;\hat{d})$ and $\widetilde{u}_a^{sct}(r)$, respectively.

In the sequel we refer that the following mixed scattering principle is based on a specific functional, the so-called *Reciprocity Gap Functional* defined as:

$$[\widetilde{u}, \widetilde{v}]_{S_0} := \int_{S_0} \left[ (T\widetilde{v})^\top \cdot \widetilde{u} - \widetilde{v}^\top \cdot (T\widetilde{u}) \right] ds \tag{3.21}$$

which is also employed in the study of various acoustic or elastic inverse problems (see, e.g., [AtNa10, CaCo05, XiBo10] and the references therein).

We are now ready to state and prove the following theorem.

**Theorem 1.** *Let $\widetilde{u}_a^{inc}(r) := \widetilde{\Gamma}(r,a)$ be an incident point-source wave field, and let $\widetilde{u}^{inc}(r;-\hat{b})$ be an incident plane-wave propagating in the direction $-\hat{b}$. Then the following relations hold:*

*(i) For $a \in D_0$*

$$\widetilde{u}_{\infty,a}(\hat{b}) = \left( \widetilde{u}^{sct}(a;-\hat{b}) \right)^\top \tag{3.22}$$

*(ii) For* $a \in D_1$

$$\widetilde{u}_{\infty,a}(\hat{b}) = (\widetilde{\gamma} - \widetilde{I}) \left(\widetilde{u}^{inc}(a; -\hat{b})\right)^{\top} + \widetilde{\gamma} \left(\widetilde{u}^{sct}(a; -\hat{b})\right)^{\top} \tag{3.23}$$

*with*

$$\widetilde{\gamma} = -(\hat{a} \otimes \hat{a}) - \frac{\lambda + \mu}{\mu} \left(\widetilde{I} - \hat{a} \otimes \hat{a}\right) \tag{3.24}$$

*and*

$$\widetilde{u}_{\infty,a}(\hat{b}) :\equiv \widetilde{u}^{p}_{\infty,a}(\hat{b}) + \widetilde{u}^{s}_{\infty,a}(\hat{b}) \tag{3.25}$$

*Proof*: Using relation (3.21) and its bilinearity we can arrive at

$$\begin{aligned} [\widetilde{u}^{t}_{a}(r), \widetilde{u}^{t}(r; -\hat{b})]_{S_0} &= [\widetilde{u}^{inc}_{a}(r), \widetilde{u}^{t}(r; -\hat{b})]_{S_0} \\ &\quad + [\widetilde{u}^{sct}_{a}(r), \widetilde{u}^{t}(r; -\hat{b})]_{S_0} \end{aligned} \tag{3.26}$$

(i) Let us first consider the case when $a \in D_0$. We have to calculate the surface integrals $[\widetilde{u}^{sct}_{a}(r), \widetilde{u}^{t}(r; -\hat{b})]_{S_0}$ and $[\widetilde{u}^{inc}_{a}(r), \widetilde{u}^{t}(r; -\hat{b})]_{S_0}$. Via Betti's formulas, the Sommerfeld–Kupradze radiation conditions and the integral representation of the far-field pattern $\widetilde{u}_{\infty,a}(\hat{b})$, we have that

$$\widetilde{u}_{\infty,a}(\hat{b}) = [\widetilde{u}^{sct}_{a}(r), \widetilde{u}^{t}(r; -\hat{b})]_{S_0}, \tag{3.27}$$

or, equivalently

$$\widetilde{u}_{\infty,a}(\hat{b}) = \int_{S_0} \left[ \left(T\widetilde{u}^{t}(r; -\hat{b})\right)^{\top} \cdot \widetilde{u}^{sct}_{a}(r) - \left(\widetilde{u}^{t}(r; -\hat{b})\right)^{\top} \cdot T\widetilde{u}^{sct}(r) \right] ds$$

With the aid of (3.7) and the boundary conditions (3.20) on $S_0$, the latter can be written as

$$\begin{aligned} \widetilde{u}_{\infty,a}(\hat{b}) &= \int_{S_0} \left[ \left(T\widetilde{u}^{t}(r; -\hat{b})\right)^{\top} \cdot \widetilde{u}^{t}_{a}(r) - \left(\widetilde{u}^{t}(r; -\hat{b})\right)^{\top} \cdot T\widetilde{u}^{t}_{a}(r) \right] ds \\ &\quad - \int_{S_0} \left[ \left(T\widetilde{u}^{t}(r; -\hat{b})\right)^{\top} \cdot \widetilde{u}^{inc}_{a}(r) - \left(\widetilde{u}^{t}(r; -\hat{b})\right)^{\top} \cdot T\widetilde{u}^{inc}_{a}(r) \right] ds \end{aligned} \tag{3.28}$$

Taking now into account second Betti's formula, relation (3.28) is written as

$$\begin{aligned}\widetilde{u}_{\infty,a}(\hat{b}) = &\int_{\Gamma}\left[\left(T\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{t}(r)-\left(\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot T\widetilde{u}_{a}^{t}(r)\right]ds\\ &+\int_{D_1}\left[\left(\Delta^{*}\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{t}(r)-\left(\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\Delta^{*}\widetilde{u}_{a}^{t}(r)\right]d\upsilon\\ &-\int_{\Gamma}\left[\left(T\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{inc}(r)-\left(\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot T\widetilde{u}^{inc}(r)\right]ds\\ &-\int_{D_1}\left[\left(\Delta^{*}\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{inc}(r)-\left(\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\Delta^{*}\widetilde{u}_{a}^{inc}(r)\right]d\upsilon\end{aligned} \tag{3.29}$$

and the fact that $\widetilde{u}^{t}(r;-\hat{b})$, $\widetilde{u}_{a}^{t}(r)$ are regular solutions of the Navier equation in $D_1$, equation (3.29) yields

$$\begin{aligned}\widetilde{u}_{\infty,a}(\hat{b}) = &-\int_{\Gamma}\left[\left(T\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{inc}(r)-\left(\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot T\widetilde{u}_{a}^{inc}(r)\right]ds\\ &+\int_{D_1}\left(\rho_1-\rho_0\right)\omega^{2}\left(\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{inc}(r)\,d\upsilon\end{aligned} \tag{3.30}$$

Concerning the other surface integral $[\widetilde{u}_{a}^{inc}(r),\widetilde{u}^{t}(r;-\hat{b})]_{S_0}$, we have to do the following manipulation

$$\begin{aligned}&[\widetilde{u}_{a}^{inc}(r),\widetilde{u}^{t}(r;-\hat{b})]_{S_0}=[\widetilde{u}_{a}^{in}(r),\widetilde{u}^{inc}(r;-\hat{b})]_{S_0}\\ &+[\widetilde{u}_{a}^{inc}(r),\widetilde{u}^{sct}(r;-\hat{b})]_{S_0}\\ =&\int_{S_0}\left[\left(T\widetilde{u}^{inc}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{inc}(r)-\left(\widetilde{u}^{inc}(r;-\hat{b})\right)^{\top}\cdot T\widetilde{u}_{a}^{inc}(r)\right]ds\\ &+\int_{S_0}\left[\left(T\widetilde{u}^{sct}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{inc}(r)-\left(\widetilde{u}^{sct}(r;-\hat{b})\right)^{\top}\cdot T\widetilde{u}_{a}^{inc}(r)\right]ds\end{aligned}$$

Using now the exterior integral representation (recall $a\in D_0$) for the scattered field $\widetilde{u}^{sct}(a;-\hat{b})$, the latter arrives to

$$\begin{aligned}&-\left(\widetilde{u}^{sct}(a;-\hat{b})\right)^{\top}\\ =&\int_{S_0}\left[\left(T\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{inc}(r)-\left(\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot T\widetilde{u}_{a}^{inc}(r)\right]ds\end{aligned} \tag{3.31}$$

With the aid now of the boundary conditions (3.20) and the second Betti's formula, relation (3.31) takes the form

$$\begin{aligned}
&\left(\widetilde{u}^{sct}(a;-\hat{b})\right)^{\top} \\
&= -\int_{\Gamma}\left[\left(T\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{inc}(r)-\left(\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot T\widetilde{u}_{a}^{inc}(r)\right]ds \\
&\quad -\int_{D_1}\left[\left(\Delta^{*}\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{inc}(r)-\left(\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\Delta^{*}\widetilde{u}_{a}^{inc}(r)\right]d\upsilon
\end{aligned} \tag{3.32}$$

Since now $\widetilde{u}^{t}(r;-\hat{b})$ and $\widetilde{u}_{a}^{inc}(r)$ are regular solutions of the Navier equation in $D_1$ and $D_0$, respectively, relation (3.32) can be written as follows:

$$\begin{aligned}
&\left(\widetilde{u}^{sct}(a;-\hat{b})\right)^{\top} \\
&= -\int_{\Gamma}\left[\left(T\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{inc}(r)-\left(\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot T\widetilde{u}_{a}^{inc}(r)\right]ds \\
&\quad +\int_{D_1}\left(\rho_1-\rho_0\right)\omega^{2}\left(\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{inc}(r)\,d\upsilon
\end{aligned} \tag{3.33}$$

Combining relations (3.26), (3.30) and (3.33) the first assertion of the theorem easily follows.

(ii) We deal now with the case $a\in D_1$. We calculate the two surface integrals of the right-hand side of relation (3.26). Taking into account Betti's formulas, Sommerfeld-Kupradze radiation conditions and the integral representation of the far-field pattern $\widetilde{u}_{\infty,a}(\hat{b})$, after lengthy calculations, we arrive at

$$\begin{aligned}
&[\widetilde{u}_{a}^{inc}(r),\widetilde{u}^{t}(r;-\hat{b})]_{S_0} \\
&=\int_{\Gamma}\left[\left(T\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{inc}(r)-\left(\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot T\widetilde{u}_{a}^{inc}(r)\right]ds \\
&\quad +\int_{D_1\setminus\Omega(a;\varepsilon)}\left(\rho_0-\rho_1\right)\omega^{2}\left(\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{inc}(r)\,d\upsilon \\
&\quad +\int_{\Omega(a;\varepsilon)}\left[\left(T\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot\widetilde{u}_{a}^{inc}(r)-\left(\widetilde{u}^{t}(r;-\hat{b})\right)^{\top}\cdot T\widetilde{u}_{a}^{inc}(r)\right]ds
\end{aligned}$$

or, equivalently in the more conventional for

$$\begin{aligned}
&[\widetilde{u}_a^{inc}(r), \widetilde{u}^t(r;-\hat{b})]_{S_0} - [\widetilde{u}_a^{inc}(r), \widetilde{u}^t(r;-\hat{b})]_{\Gamma} \\
&= \int_{D_1\setminus\Omega(a;\varepsilon)} (\rho_1-\rho_0)\,\omega^2 \left(\widetilde{u}^t(r;-\hat{b})\right)^{\top}\cdot \widetilde{u}_a^{inc}(r)\,d\upsilon \\
&\quad + [\widetilde{u}_a^{inc}(r), \widetilde{u}^t(r;-\hat{b})]_{\Omega(a;\varepsilon)}
\end{aligned} \tag{3.34}$$

where $\Omega(a;\varepsilon) := \{r \in \mathbb{R}^2 : |r-a| = \varepsilon\}$ and $D_1 \setminus \Omega(a;\varepsilon) := \{r \in D_1 : |r-a| > \varepsilon\}$. Using the mean value theorem for the line integral $[\widetilde{u}_a^{inc}(r), \widetilde{u}^t(r;-\hat{b})]_{\Omega(a;\varepsilon)}$ of (3.34) and letting $\varepsilon \to 0$, we get

$$\begin{aligned}
&[\widetilde{u}_a^{inc}(r), \widetilde{u}^t(r;-\hat{b})]_{S_0} - [\widetilde{u}_a^{inc}(r), \widetilde{u}^t(r;-\hat{b})]_{\Gamma} \\
&= \int_{D_1\setminus\Omega(a;\varepsilon)} (\rho_1-\rho_0)\,\omega^2 \left(\widetilde{u}^t(r;-\hat{b})\right)^{\top}\cdot \widetilde{u}_a^{inc}(r)\,d\upsilon \\
&\quad + \widetilde{\gamma}\left(\widetilde{u}^t(a;-\hat{b})\right)^{\top}.
\end{aligned} \tag{3.35}$$

Similar arguments for the second line integral $[\widetilde{u}_a^{sct}(r), \widetilde{u}^t(r;-\hat{b})]_{S_0}$ yield

$$\begin{aligned}
&[\widetilde{u}_a^{sct}(r), \widetilde{u}^t(r;-\hat{b})]_{S_0} - [\widetilde{u}_a^{inc}(r), \widetilde{u}^t(r;-\hat{b})]_{\Gamma} \\
&= \int_{D_1\setminus\Omega(a;\varepsilon)} (\rho_1-\rho_0)\,\omega^2 \left(\widetilde{u}^t(r;-\hat{b})\right)^{\top}\cdot \widetilde{u}_a^{inc}(r)\,d\upsilon \\
&\quad + [\widetilde{u}_a^{sct}(r), \widetilde{u}^t(r;-\hat{b})]_{\Omega(a;\varepsilon)}
\end{aligned} \tag{3.36}$$

Using again the mean value theorem, and letting $\varepsilon \to 0$, the last term of the right-hand side of (3.36) vanishes. Combining relations (3.35), (3.36) and taking into account the representation (3.27) (note that (3.27) also holds for $a \in D_1$), i.e.,

$$\widetilde{u}_{\infty,a}(\hat{b}) = [\widetilde{u}_a^{sct}(r), \widetilde{u}^t(r;-\hat{b})]_{S_0},\ a \in D_1, \tag{3.37}$$

we arrive at

$$\widetilde{u}_{\infty,a}(\hat{b}) - \widetilde{\gamma}\left(\widetilde{u}^t(a;-\hat{b})\right)^{\top} = -[\widetilde{u}_a^{inc}(r), \widetilde{u}^t(r;-\hat{b})]_{S_0},$$

and by the superposition (3.5) the latter takes the form

$$\begin{aligned}
\widetilde{u}_{\infty,a}(\hat{b}) - \widetilde{\gamma}\left(\widetilde{u}^t(a;-\hat{b})\right)^{\top} &= -[\widetilde{u}_a^{inc}(r), \widetilde{u}^{inc}(r;-\hat{b})]_{S_0} \\
&\quad -[\widetilde{u}_a^{inc}(r), \widetilde{u}^{sct}(r;-\hat{b})]_{S_0}
\end{aligned} \tag{3.38}$$

The second integral on the right-hand side of (3.38) vanishes for $a \in D_1$, whereas the first integral, due to integral representation theorem for the elastic incident field [PeSe03], equals to $-\left(\widetilde{u}^{inc}(a;-\hat{b})\right)^{\top}$, therefore

$$\widetilde{u}_{\infty,a}(\hat{b}) = \widetilde{\gamma}\left(\widetilde{u}^{t}(a,-\hat{b})\right)^{\top} - \left(\widetilde{u}^{inc}(a,-\hat{b})\right)^{\top}$$

and hence, the second assertion of the theorem easily follows.

We remark that the above mixed reciprocity principle can be extended in three or multi-layered background medium.

We now proceed with the following useful result.

**Theorem 2.** *Assume that $D_2$ and $\breve{D}_2$ are two subsets of $D$ and $\mathfrak{G}$ be the unbounded component of $\mathbb{R}^2 \setminus \left(\overline{D}_2 \cup \overline{\breve{D}}_2\right)$. Furthermore, let $\widetilde{\breve{u}^{sct}}(r,\hat{d})$ being the scattered field due to obstacle $\breve{D}_2$ with corresponding far-field pattern $\widetilde{\breve{u}}_\infty(\hat{r},\hat{d})$ for all $\hat{r}$, $\hat{d} \in \Omega$. If $\widetilde{u}^{sct} = \widetilde{u}^{sct}(r;a)$ is the unique solution of the mixed impedance transmission boundary value problem*

$$\Delta^* \widetilde{u}^{sct} + \rho_0\,\omega^2\,\widetilde{u}^{sct} = \widetilde{0} \quad \text{in } D_0 \tag{3.39}$$

$$\Delta^* \widetilde{u}^{sct} + \rho_1\,\omega^2\,\widetilde{u}^{sct} = (\rho_0-\rho_1)\,\omega^2\,\widetilde{\Gamma}(r,a) \quad \text{in } D_1 \tag{3.40}$$

$$\widetilde{u}_e^{sct}(r) = \widetilde{u}_i^{sct}(r) \;\; T_e\widetilde{u}_e^{sct}(r) = T_i\widetilde{u}_i^{ext}(r) \quad \text{on } S_0 \tag{3.41}$$

$$\widetilde{u}^{sct} = -\widetilde{\Gamma}(r,a) \quad \text{on } \Gamma_D \tag{3.42}$$

$$T\widetilde{u}^{sct} + i\,\omega c\,\widetilde{u}^{sct} = -T\widetilde{\Gamma}(r,a) - i\,\omega c\,\widetilde{\Gamma}(r,a) \quad \text{on } \Gamma_I \tag{3.43}$$

$$\lim_{|r|\to\infty} \sqrt{r}\left(\frac{\partial \widetilde{u}_\beta^{sct}(r)}{\partial r} - ik_\beta\,\widetilde{u}_\beta^{sct}(r)\right) = \widetilde{0}, \;\; \beta = p,s, \tag{3.44}$$

*for $r \neq a \in D \cap \mathfrak{G}$, and further we assume that $\widetilde{\breve{u}^{sct}} := \widetilde{\breve{u}^{sct}}(r;a)$ is the unique solution of the mixed impedance transmission problem (3.39)–(3.44), but this time replacing $D_1$ and $D_2$ by $\breve{D}_1 := D \setminus \overline{\breve{D}}_2$ and $\breve{D}_2$, respectively, then*

$$\widetilde{u}^{sct}(r,a) = \widetilde{\breve{u}^{sct}}(r,a), \;\; r \in \overline{\mathfrak{G}}$$

Recall here the boundary value problem (3.8)–(3.10) has a unique solution. Taking into account the well-posedness of the interior mixed impedance boundary value problem (3.15)–(3.18), the mixed reciprocity principle of Theorem 1 and Theorem 2, we have the following main result.

**Theorem 3.** *Assume that $D_2$ and $\breve{D}_2$ are two scattering non-penetrable partially coated obstacles embedded in the same elastic piecewise-constant background medium in $\mathbb{R}^2$ with $c > 0$, $\breve{c} > 0$ the corresponding surface impedance constants.*

*If $S_0$, $\breve{S}_0$ are two penetrable interfaces, and the far-field patterns of the scattered fields for the same incident plane-wave coincide at a fixed frequency, for all incident direction $\hat{d} \in \Omega$, and observation direction $\hat{r} \in \Omega$, then*

$$D_2 = \breve{D}_2$$
$$\Gamma_D = \breve{\Gamma}_D, \quad \Gamma_I = \breve{\Gamma}_I, \quad c = \breve{c}$$
$$S_0 = \breve{S}_0$$

## References


[AtNa14] Athanasiadis, C. E., Natroshvili, D., Sevroglou, V. Stratis, I. G.: Mixed impedance transmission problems for vibrating layered elastic problems. Math. Methods Appl. Sc. (accepted 2014)

[AtNa11] Athanasiadis, C. E., Natroshvili, D., Sevroglou V., Stratis, I. G.: A boundary integral equations approach for direct mixed impedance problems in elasticity. J. Integral Eqns. Appl. **23**, 183–222 (2011)

[AtNa10] Athanasiadis, C. E., Natroshvili, D., Sevroglou V., Stratis, I. G.: An application of the reciprocity gap functional to inverse mixed impedance problems in elasticity. Inverse Problems. **26**, 085011 19pp (2010)

[AtSe06] Athanasiadis, C. E., Sevroglou V., Stratis, I. G.: Scattering relations for point-generated dyadic fields in two-dimensional linear elasticity. Quart. Appl. Math. **4**, 695–710 (2006)

[CaCo05] Cakoni, F., Colton, D.: Qualitative Methods in Inverse Electromagnetic Scattering Theory. Springer-Verlag (2005)

[XiBo10] Xiaodong L., Bo, Z.: Direct and inverse obstacle scattering problems in a piecewise homogeneous medium. SIAM J. Appl. Math. **70**, No 8 (2010)

[NaTe01] Natrosvilli, D., Z. Tediashvili, Z.: Mixed type direct and inverse scattering problems. In: Elschner, J., Gohberg, I., Silbermann, B. (eds.) Operator Theory: Advances and Applications, **121**, 366–389 Birkhäuser, Basel (2001)

[PeSe03] Pelekanos G., Sevroglou, V.: Inverse scattering by penetrable objects in two-dimensional elastodynamics. J. Comp. Appl. Math. **151**, 129–140 (2003)

[Se05] Sevroglou, V.: The far-field operator for penetrable and absorbing obstacles in 2D inverse elastic scattering. Inverse Problems. **17**, 717–738 (2005)

[Ta94] Tai, C. T.: Dyadic Greens Functions in Electromagnetic Theory. IEEE, New York (1994)

[Tw67] Twersky, V.: Multiple scattering of electromagnetic waves by arbitrary configurations. J. Math. Phys. **8**, 589–610 (1967)

# Chapter 4
# Half-Life Distribution Shift of Fission Products by Coupled Fission–Fusion Processes

**J.B. Bardaji, B.E.J. Bodmann, M.T. Vilhena, and A.C.M. Alvim**

## 4.1 Introduction

Nuclear reactions are by far the most efficient processes of energy release that may be used for energy production. Besides the nuclear decay, that is responsible for maintaining the interior of the earth heated up, fission reactions are nowadays exploited in nuclear power reactors, whereas fusion processes are considered a potential nuclear power perspective of the future. Following a traditional paradigm based on binding energy per nucleon considerations, fission is considered feasible by fragmenting heavy nuclei with neutron emission while fusion shall be attained by melting together light nuclei into heavier ones following several conceptions such as Tokamaks, for instance, or involving laser based techniques among others. In the present contribution we outline a different reasoning considering a combined fission–fusion scenario where strongly negative iso-spin projection of unstable fragments from fission may open a pathway for a fusion process when running into a light nucleus. More specifically, negative iso-spin excess of the fission fragments numerically around 3–4, which is a measure for instability of the nucleus, indicates that there might exist possibilities to merge a fission fragment with a light nucleus (${}^{1}_{1}H$, ${}^{2}_{1}H$, ${}^{7}_{3}Li$, ${}^{9}_{4}Be$ or others) and thus places the produced heavier nucleus closer to the stability line (with iso-spin excess close to zero). Here iso-spin excess refers to the difference in the Bethe–Weizsäcker asymmetry term of the

J.B. Bardaji (✉) • B.E.J. Bodmann • M.T. Vilhena
Federal University of Rio Grande do Sul, Av. Osvaldo Aranha 99/4,
Porto Alegre 90046-900, RS, Brazil
e-mail: bialkowskiknight@gmail.com; bardo.bodmann@ufrgs.br; mtmbvilhena@gmail.com

A.C.M. Alvim
Universidade Federal do Rio de Janeiro, Avenida Horácio Macedo, 2030, Cidade Universitária,
21941-972, Rio de Janeiro, RJ, Brazil
e-mail: aalvim@gmail.com



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_4

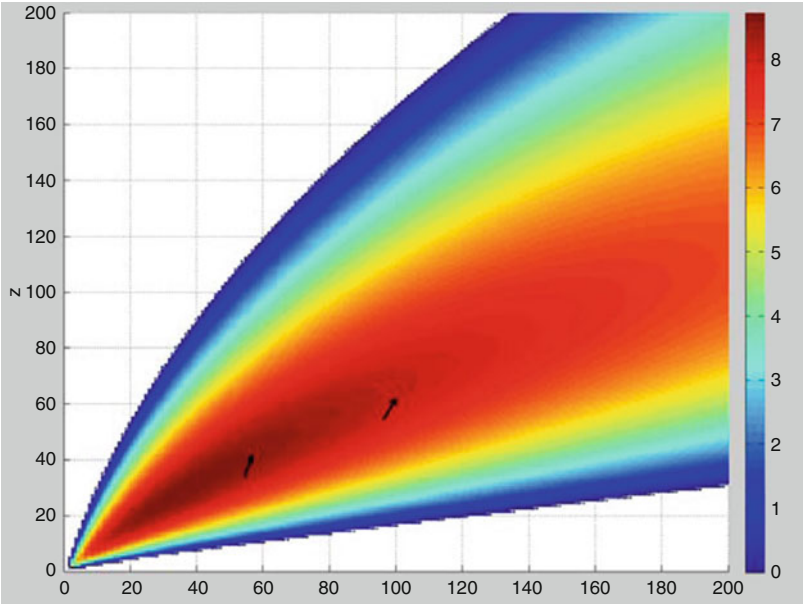


**Fig. 4.1** Binding energy per nucleon as a function of neutron and proton number. The marked transitions indicate possible fusion reactions of the fragments with a light nucleus following a fission process.

unstable nuclei and its corresponding stable counterpart. An example is indicated in figure 4.1, where the binding energy per nucleon is presented depending on the number of protons and neutrons, respectively. The dark red region refers to largest binding energy per nucleon ($\sim 7$–$8\,MeV$) and shades from light red to blue indicate decreasing stability ($6\,MeV \rightarrow 0\,MeV$) so that the indicated possible processes could release an energy amount comparable to decay processes. In the present work we analyze less the energy balance aspect of the combined fission–fusion process but focus on another issue related to stability, i.e. the distribution of half-life times of nuclei after fission and fusion following fission, respectively. The discussion that follows presents first properties from propagation of fission products through nuclear fuel material with addition of light elements and shows the kinetic energies of the fission fragments that overcome the Coulomb repulsion of the respective fusion partners. In a second step, the probability for fusion is estimated using the uncertainty principle, due to the fact that there does not exist a model that covers the variety of possible fusion reactions and thus would allow to calculate those reactions by an approach like Fermi's golden rule. Combining the two previous steps allows to evaluate a comparison of the half-life distribution considering fission only to the distribution from combined fission–fusion.

## 4.2 The Coulomb Barrier

Considering fission of Uranium-235, a variety of fission modes are possible which release mostly two or three neutrons in the fragmentation process besides the two fragments that appear with a mass ratio of approximately $\sim \frac{2}{3}$. The resulting distribution of the fission products from Uranium-235 depending on the proton ($Z$) and neutron number ($N$) is shown in figure 4.2. Since the fission process is a many body process with typically four or five constituents the most probable kinetic energy $\langle E(Z,A)\rangle$ for each fragment is given in equation (4.1) [ViKwWa85].

$$\langle E(Z,A)\rangle = 0.1166 MeV \frac{Z^2}{A^{1/3}} + 9.0 MeV. \tag{4.1}$$

The kinetic energy distribution for the heavy fission products generated by a Monte Carlo simulation for $10^5$ fissions of $^{235}U$ is shown in figure 4.3.

According to [Be88] fusion dynamics at low energies (i.e., of the order of magnitude of classical potential energies) is governed by theoretical quantum tunnelling through the Coulomb barrier (a semi-classical approach). The phenomenon occurs already for kinetic energies from $\gtrsim 10^1\, keV$ onwards. Note that there do exist some

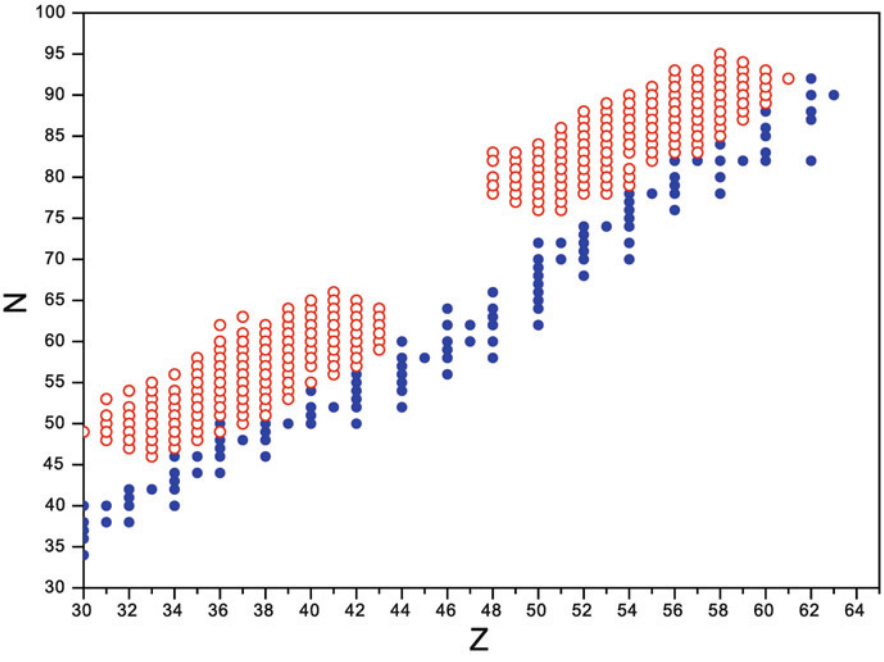


**Fig. 4.2** Fission product distribution from Uranium-235 depending on proton and neutron number (red circles), for comparison stable nuclei (blue circles).

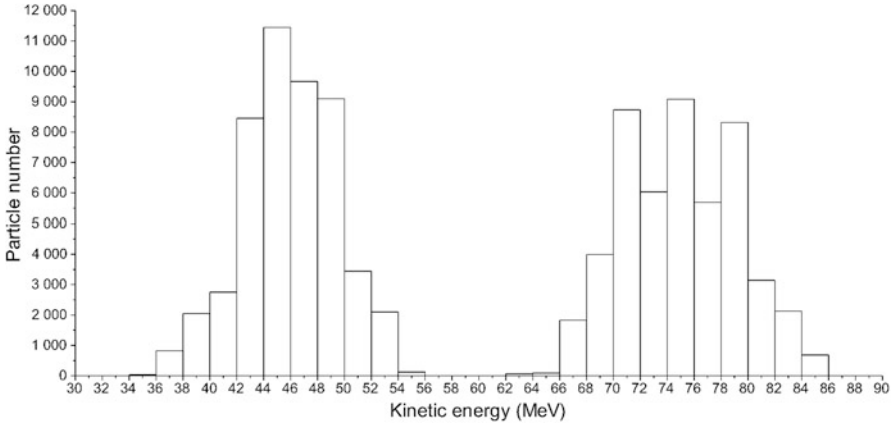


**Fig. 4.3** Kinetic energy distribution of fission products from Uranium-235.

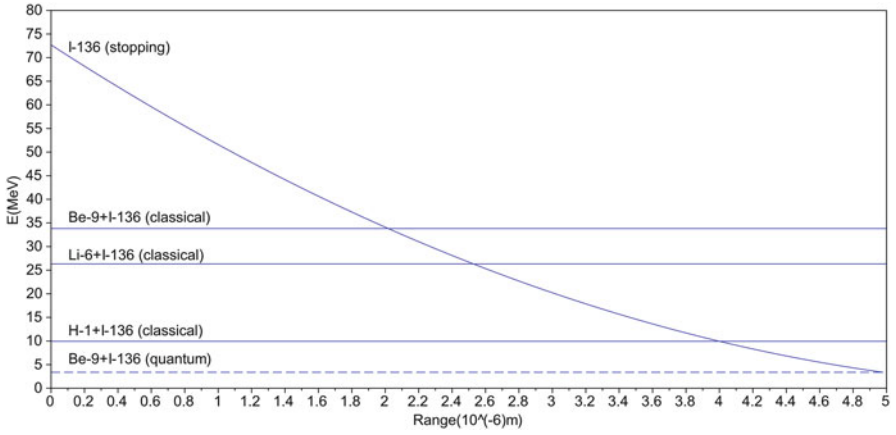


**Fig. 4.4** Kinetic energy of the fission fragment $^{136}I$ and classical Coulomb repulsion barriers for fusion with $^{1}H$, $^{2}H$, $^{6}Li$ and $^{9}Be$.

models based on the Schrödinger equation [NiDaLa89, HaRoKr99, ZaSa04], that allow to calculate fusion cross sections, however with acceptable results for energies well beyond $10^{1}\,MeV$, only. Other approaches that consider fusion cross sections [NaEtAl04, VoEtAl09, SiEtAl10] do not focus on a general description for fusion reaction but are designed to describe specific reactions from heavy ion collision experiments.

A crude estimate by comparing classical barriers and the kinetic energy along the propagation depth shows that for a considerable path length the kinetic energy of the fission fragment is above the repulsive potential energy. Even for a simplified treatment, where the fission fragment has an associated wave function, whereas the repulsive interaction is represented by a classical potential, the effective limit

is lowered down by at least one order in magnitude, due to the so-called tunnel effect. Thus from the kinetic energy and Coulomb repulsion comparison fusion is in principle possible, as illustrated in figure 4.4.

## 4.3 Particle Stopping in Nuclear Fuel

Fission fragments with its initial kinetic energies lose their energies through collisions with the surrounding material. The collisions may occur with the electrons or the nucleus of the atoms $\frac{dE}{dx} = \frac{dE}{dx}\big|_n + \frac{dE}{dx}\big|_e$, respectively. In these processes energy is transferred which for energy loss with electrons may be described to a reasonable accuracy by the Bethe–Bloch formula. Here the energy loss is given in units of $MeV/(mg/cm^2)$.

Collisions with the nucleus are more likely to happen for kinetic energies below the Coulomb barrier and may involve significant energy transfers with associated larger deflections in the ion trajectory, whereas ion electron collisions typically show small energy losses and almost no deflection in the particle trajectory and are dominant in the energy range of interest, i.e. where fusion can occur.

The stopping by the nucleus is described in detail in reference [ZiZiBi10] and can be calculated by

$$\left.\frac{dE}{dx}\right|_n = -\frac{N\pi a_{TF}^2 T_M}{\varepsilon^2}\int_0^{T_M} f\left(t^{1/2}\right)\, dt\,.$$

Here $N$ (in $cm^{-3}$) is the atomic density, $a_{TF} = \frac{1}{2}\left(\frac{3\pi}{4}\right)^{3/2}\frac{\hbar^2}{m_e e^2 Z^{1/3}}$ is the Thomas–Fermi screening length with the elementary charge $e$ and the electron mass $m_e$, $T_M = \frac{4A_1A_2}{(A_1+A_2)^2}E$ is the maximum energy transfer, $f(t^{1/2})$ is the Lindhard scaling function and the dimensionless collision parameter $t = \varepsilon^2\frac{T}{T_M}$ with the dimensionless energy $\varepsilon = \frac{A_1}{A_1+A_2}\frac{a_{TF}}{Z_1Z_2e^2}E$.

For the calculation of the electronic stopping (in units of $MeVcm^2/mg$) of an ion with speed $v$, we used the model proposed by the authors [SrMu76] presented below, due to the fact that its results showed acceptable agreement with experimental data. According to the findings in [MuSr74] equation (4.2) may be applied for particles that are partially or totally ionized. However, it is noteworthy that for the case $\chi \gg 1$ or $\chi \ll 1$ they are not valid rigorously i.e., for the example of the fission products.

$$\frac{dE}{dx} = \frac{2\pi z^2 e^4 N}{m_e v^2}(J_1 + J_2 + J_3) \tag{4.2}$$

$$J_1 \equiv \sum_{U_s=0}^{U_s'} \ln(\eta_s^2\chi^{-2}) \qquad J_2 \equiv \sum_{U_s=0}^{U_s=2v\chi^{-1}} \ln(\eta_s^2) \qquad J_3 \equiv \sum_{U_s=2v\chi^{-1}}^{U_s''} \ln(\eta_s^3\chi^{-1})$$

Here $\eta_s = 2v/v_0$, $\chi = 2zv_0/v$, $v_0 = e^2/\hbar$ and the upper limit of the sum $U_s$ is the electron speed in the $S$ orbit, that turns the logarithmic terms equal zero and guarantees that energy transfer is semi-positive definite. In the following we present the parametrizations for the calculation of the range, that depends basically on $v$, $\chi$, the effective charge of the projectile $z$, and the charge of the target particle $Z_2$. The effective charge of a particle with speed $v$ is calculated following reference [No60] $z = Z_1\left(1 - 2.03\exp\left(\frac{-2vf(Z_1)}{Z_1 v_0}\right)\right)^{1/2}$ with the scaling function $f(Z_1) = 0.28Z_1^{2/3}$ for $Z_1 \leq 45.5$ and $f(Z_1) = Z_1^{1/3}$ for $Z_1 \geq 45.5$. Further, in the formalism below the mean excitation energy ($\bar{I}$) for the target given by [SrMu76] was replaced by the values of [Ah80] which provide the better results.

1. $\chi > 1$ and $v \geq \frac{1}{2}Z_2 v_0 \chi$:
   For $\chi > 1$, $J_1$ all values for $U_s$ are included together with the electron's speed in the $K$ shell, which are given by $Z_2 v_0$, for $v \geq \frac{1}{2}Z_2 v_0 \chi$. Using $\eta_s^2 = (2v/U_s)^2 = 2mv^2/I_s$, where the ionization potential of the $S$ shell ($I_s$) is given by $I_s = mU_s^2/2$, one may express $J_1$

$$J_1 = \sum_{s=1}^{Z_2} \ln(\eta_s^2 \chi^{-2}) = Z_2 \ln\left(\frac{2mv^2}{\bar{I}\chi^2}\right)$$

   where $\bar{I} = \frac{1}{Z_2}\sum_{s=1}^{Z_2} \ln I_s$ is the mean ionization potential. For $v = \frac{1}{2}Z_2 v_0 \chi$ the superior limit of the sum for $J_2$, $U_s = 2v\chi^{-1}$, is equal to $U_s = Z_2 v_0$.

$$J_2 = \sum_{U_s=0}^{Z_2 v_0} \ln\left(\frac{2v}{U_s}\right)^2 = \sum_{s=1}^{Z_2} \ln\left(\frac{2mv^2}{I_s}\right) = Z_2 \ln\left(\frac{2mv^2}{\bar{I}}\right)$$

   In the sum $J_3$ the inferior limit is the maximum value, so that there is no contribution. Thus the expression for the electronic stopping according to the established conditions is

$$\frac{dE}{dx} = \frac{63.65\frac{MeVcm^3}{gs^2} z^2 Z_2}{A_2 v^2} \log_{10}\left(\frac{11.39\frac{keVs^2}{cm^2}v^2}{I\chi}\right).$$

2. $\chi > 1$ and $\frac{1}{2}Z_2 v_0 \chi^{1/3} \leq v < \frac{1}{2}Z_2 v_0 \chi$:
   For $v < \frac{1}{2}Z_2 v_0 \chi$ the value of $U_s'$ for $J_1$ is $2v\chi^{-1}$ and $J_1$ is negative for $U_s > 2v\chi^{-1}$. According to [Bo48], the sum may be replaced by an integral

$$J_1 = \sum_{U_s=0}^{2v\chi^{-1}} \ln\left(\frac{2v\chi^{-1}}{U_s}\right)^2 = \int_0^{n(2v\chi^{-1})} \ln\left(\frac{2v\chi^{-1}}{U_s}\right)^2 dn(U_s)$$

   where $n(U_s)$ is the number of electronic orbitals with speeds inferior $U_s$. The author of reference [MuSr74] showed that for a medium with atomic number

$Z_2$, $n(U_s)$ is given by $n(U_s) = \frac{f(Z_2)U_s}{v_0}$ with $f(Z_2) = 0.28Z_2^{2/3}$ for $Z_2 \leq 45.5$ and $f(Z_2) = Z_2^{1/3}$ for $Z_2 \geq 45.5$. In order to reproduce the electron states with their respective speeds so that $\int_{U_s=0}^{2v_0Z_2} dn(U_s) = Z_2$, [Mu75] applied the expression for $n(U_s)$ to all electrons, except for the two electrons of the $K$ shell.

$$J_1 = \int_{U_s=0}^{2v\chi^{-1}} \ln\left(\frac{2v\chi^{-1}}{U_s}\right)^2 dn(U_s) = \frac{4f(Z_2)\chi^{-1}v}{v_0}$$

The term for $J_2$ was obtained in a similar fashion.

$$J_2 = \int_{U_s=0}^{2v\chi^{-1}} \ln\left(\frac{2v}{U_s}\right)^2 dn(U_s) = \frac{4f(Z_2)\chi^{-1}(1+\ln\chi^{-1})v}{v_0}$$

For $J_3$ the upper limit corresponds to $U_s'' = 2v\chi^{-1/3}$ for $v \geq Z_2v_0\chi^{1/3}/2$. Recalling that the expression $n(U_s)$ does not include the $K$ shell electrons, these have to be added separately.

$$J_3 = \frac{f(Z_2)}{v_0}\int_{U_s=2v\chi^{-1}}^{(Z_2-2)v_0/f(Z_2)} \ln\left(\frac{2v}{U_s\chi^{1/3}}\right)^3 dU_s + 2\ln\left(\frac{2v}{Z_2v_0\chi^{1/3}}\right)$$

Upon solving the integral and summing the expressions $\sum_{i=1}^{3} J_i$ yields

$$\frac{dE}{dx} = \frac{13.79\frac{MeVcm^3}{gs^2}z^2}{A_2v^2}\left(3(Z_2-2)\left(1+\ln\frac{2f(Z_2)v}{(Z_2-2)v_0}\right)+6\ln\frac{2v}{Z_2v_0} + \frac{2f(Z_2)v}{v_0\chi} - Z_2\ln\chi\right)$$

3. $\chi > 1$ and $v < \frac{1}{2}Z_2v_0\chi^{1/3}Z_2v_0\chi$:
This case was elaborated in detail by [Bo48] and a general expression was given by [MuSr74].

$$\frac{dE}{dx} = \frac{12.68\frac{MeVcm^2}{gs}f(Z_2)z^2}{A_2v}\left(3\chi^{-1/3}+\chi^{-1}\right)$$

4. $\chi < 1$ and $v \geq \frac{1}{2}Z_2v_0$:
For $v \geq \frac{1}{2}Z_2v_0$ equation (4.2) simplifies to the form frequently found in the literature [Ah80]. Now, if $\chi < 1$, the particle is capable of ionizing the inner most electrons of the target. The maximum electron speed after energy and momentum transfer by a projectile with speed $v$ is $2v$ and if the speed of an electron in the $K$ shell is $Z_2v_0$, then the second condition is valid for $v \geq \frac{1}{2}Z_2v_0$. Thus, $J_1$ and $J_2$ are equal and $J_3$ is redundant, so that the stopping power is applicable for relativistic particles.

**Table 4.1** Comparison of calculated ($R_{cal}$), simulated ($R_{SRIM}$), and observed ranges ($R_{exp}$).

| Ion | Target | E ($MeV$) | $R_{cal}(\mu m)$ | $R_{exp}(\mu m)$ | $R_{SRIM}(\mu m)$ | Reference |
|---|---|---|---|---|---|---|
| $^{12}C$ | $^{27}Al$ | 124.8 | 239.49 | 231.68 | 230.10 | [Br62] |
| $^{20}Ne$ | $^{27}Al$ | 208.0 | 162.10 | 156.92 | 151.98 | [Br62] |
| $^{16}O$ | $^{63}Cu$ | 75.2 | 21.02 | 21.08 | 23.96 | [Og59] |
| $^{12}C$ | $^{12}C$ | 24.0 | 24.32 | 23.80 | 18.83 | [TaBiBa97] |
| $^{86}Se$ | $^{238}U$ | 39.54 | 4.25 | – | 3.53 | – |
| $^{100}Zr$ | $^{238}U$ | 49.19 | 4.34 | – | 4.02 | – |
| $^{153}Pm$ | $^{238}U$ | 90.12 | 5.53 | – | 5.47 | – |

$$\frac{dE}{dx} = \frac{63.65 \frac{MeVcm^3}{gs^2} z^2 Z_2}{A_2 v^2} \left( \log_{10} \frac{11.39 \frac{keVs^2}{cm^2} v^2}{I(1-\beta^2)} - \frac{\beta^2}{2.303} \right)$$

5. $\chi < 1$ and $v < \frac{1}{2} Z_2 v_0$:
   In this case the appropriate limit for $U_s'$ in $J_1$ and $J_2$ shall be used, because not all electrons are capable of participating in the electronic stopping process. The logarithmic term in $J_1$ and $J_2$ are zero for $U_s' = 2v$, thus, according to [Bo48, MuSr74] and [BeAs53]

$$\frac{dE}{dx} = \frac{50.6 \frac{MeVcm^2}{gs} f(Z_2) z^2}{A_2 v}$$

The energy loss results may be integrated from the initial energy to zero to yield the range of the particle $R = \rho^{-1} \int_{E_0}^{0} (dE/dx)^{-1}\, dE$. Here $E_0$ is the initial kinetic energy (in $MeV$) of the ion from fission. Note that the range does not represent the total length of the particle trajectory but the effective penetration depth. Further, one may also calculate the stopping time $t_S = \rho^{-1} \int_{E_0}^{0} v^{-1} (dE/dx)^{-1}\, dE$. Table 4.1 shows some results for kinetic energies, ranges determined from the present approach, experimental data and from a benchmark simulation [ZiZiBi10]. The stopping time in all cases was of the $10^0$ $ns$ order of magnitude and thus orders of magnitudes larger than the time scale for possible fusion processes.

## 4.4 Fusion Following Fission

The procedure to analyze combined fission–fusion and compare them to pure fission is based on the steps, definition of the chemical composition of the nuclear fuel plus light substances, generating the fission products and evaluation of a possible nuclear fusion reaction. To this end a Monte Carlo simulation platform was developed, where the used databases for the half-lives and binding energies were taken from [IAEA13] and [AuWa93], respectively.

In a Monte Carlo simulation $10^5$ fissions of $^{235}U$ by thermal neutrons were simulated according to the probabilities in [EnRi94] thus covering 98.7% of the nuclide spectrum expected in a real situation. For each fission fragment the kinetic energy was calculated using the formula for the most likely energy. The generated distribution for the kinetic energies is shown in figure 4.3. The propagation of the fission fragment was calculated using the parametrization from section 4.3 and fusion reactions were generated while the kinetic energy was sufficiently high to overcome the Coulomb repulsion. As potential fusion partners we considered hydrogen, deuterium, lithium-6, and beryllium-9. Although data were generated for all targets and some mixtures, they qualitatively produce the same effect, namely the shift in the half-life distribution towards larger values, therefore, only one example ($^9Be$) will be presented.

It is noteworthy that in spite of an increase in heavy ion collision research, no model exists that would allow to calculate fusion cross sections or transition rates. Hence, we circumvent this problem making use of mass defect measurements together with the uncertainty principle $\Delta E \Delta t \geq \frac{\hbar}{2}$ in the spirit of Fermi's Golden Rule (see, for instance, [Sc04]). While $\Delta E$ is determined by the mass defect, $\Delta t$ is related to the inverse transition rate, that in theoretical calculations may be determined from the quantum transition matrix $\langle \beta | V | \alpha \rangle$ and the density of states $\rho_{\alpha\beta}$. The transition rate $\Gamma_{\alpha\beta}$ from initial state $\alpha$ to final state $\beta$ is thus given by

$$\frac{1}{\Delta t} \approx \Gamma_{\alpha\beta} = \frac{1}{h} \rho_{\alpha\beta} |\langle \beta | V | \alpha \rangle|^2 . \tag{4.3}$$

The probability for a specific nuclear reaction is given by the ratio of the transition rate for the specific reaction divided by all possible reactions.

$$P_{(Z,A)} = \frac{N_i \tau_i \frac{4\pi}{h} \Delta E_j}{\sum_{k=1}^{n} \sum_{l=0}^{m} N_k \tau_k \frac{4\pi}{h} \Delta E_l}$$

Here, $N_i \tau_i$ is a measure for the probability of a fission fragment to hit a target nucleus ($H$, $Li$ or $Be$) for fusion or $U$ and $O$, respectively, with no fusion. The term $4\pi / h \Delta E_j$ stems from Fermi's golden rule and estimates the transition rate for a specific reaction, that in case of fusion could be a simple fusion (no neutron emission) or with emission of $j$ neutrons, unless the energy balance $E_{reactants} \leq E_{products}$ holds. Table 4.2 exemplifies this probabilities for a fusion reaction of germanium with lithium. The obtained collision probabilities for the target nuclei $H$, $Li$, or $Be$ are 0.277, 0.377, and 0.523, considering a particle density admixture of 30%.

For each target type three simulations were performed and the following variables recorded:

- The average energy released by the fusion reaction, $\overline{\triangle B}$, in $MeV$, which is related to the variation of the binding energy;
- the average number of neutrons $N$ emitted in the fusion reactions;
- the proportion $P_r$ of all nuclei that increased their half-life $t_{1/2}^{fus} > t_{1/2}^{fis}$ and

**Table 4.2** Probability for a fusion reaction $^{80}Ge + {}^{6}Li$ with $E_{reactants} = 722112.567\ keV$.

| Reaction | $E_{product}(keV)$ | $t\ (y)$ | Probability |
|---|---|---|---|
| $^{80}Ge + {}^{6}Li \rightarrow {}^{86}Br$ | 742386.902 | 16.23 | 0.5629 |
| $^{80}Ge + {}^{6}Li \rightarrow {}^{85}Br + 1n$ | 737287.894 | 21.69 | 0.4214 |
| $^{80}Ge + {}^{6}Li \rightarrow {}^{84}Br + 2n$ | 728384.015 | 58.07 | 0.0157 |
| $^{80}Ge + {}^{6}Li \rightarrow {}^{83}Br + 3n$ | 721545.735 | – | $0\ (\Delta E < 0)$ |

**Table 4.3** Results for a target with (30%) particle density admixture by $^{9}Be$.

| Simulation | no. fusions | $\overline{\Delta B}(MeV)$ | no. neutrons | $\%\ P_r\ t_{1/2}^{fus} > t_{1/2}^{fis}$ | $\%\ P_t\ t_{1/2} > 1y$ |
|---|---|---|---|---|---|
| 1 | 52159 | 10.06 | 0.317 | 60.17 | 1.990 |
| 2 | 52169 | 10.07 | 0.317 | 60.19 | 2.049 |
| 3 | 52325 | 10.23 | 0.318 | 60.12 | 2.041 |

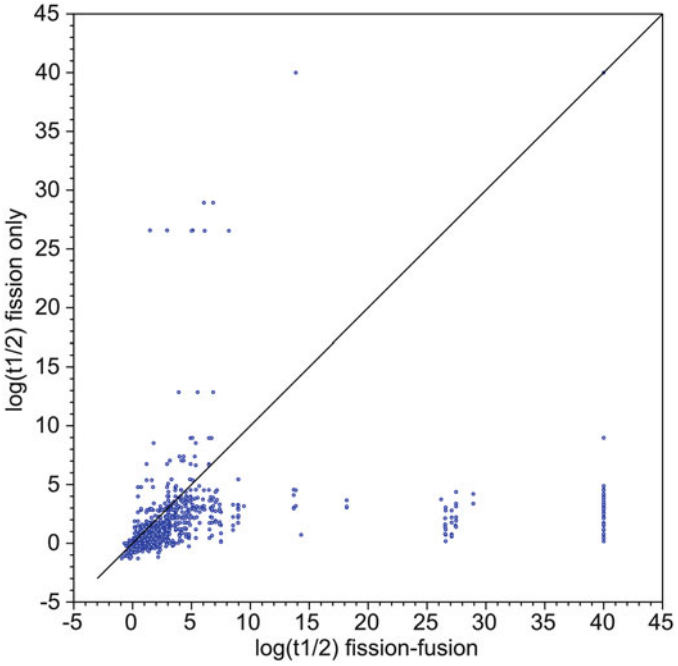


**Fig. 4.5** Shift in the half lives of fission products that underwent a subsequent fusion reaction with 30% admixture of $^{9}Be$ in the nuclear fuel.

- The final state portion of all particles from the $10^5$ fissions with half-life larger than a year ($P_t\ t_{1/2} > 1$) compared to the value for fission only 0,209%.

Table 4.3 shows the results for simulations with a target composition containing $^{9}Be$. For this target the mean energy and mean neutron number released in this reaction are $10.10\,MeV$ and 0,317, respectively. The proportion of half-lives larger than a year increased by an order of magnitude ( a factor 9,70).

Figure 4.5 shows the shift in the half-lives of fission products that underwent a subsequent fusion reaction with $^{9}Be$. The set of points below the bisector clearly show the tendency for increasing the half-life after fusion reaction, only a small proportion reduced the half-life.

## 4.5 Conclusions

The present discussion is an initial study on the possibility to explore a coupled fission–fusion scenario in a nuclear reactor core. In the outlined conception, fission products could suffer a subsequent fusion reaction if some light nuclei was mixed with the nuclear fuel. The discussion showed that one significant consequence of such a scenario is manifest in the change of the half-life distribution of the produced nuclei after fission–fusion when compared to the case where only fission occurs. In the present simulations the coupled fission–fusion enhanced at an average the half-lives by more than one order in magnitude.

The simulations with target nuclei $^{1}H$ and $^{6}Li$ yielded largest differences in comparison with pure fission. Of all particles only 0.209% presented half-lives larger than one year, that after inclusion of fusion shifted to 1.585% and 1.617% for $^{1}H$ and $^{6}Li$, respectively. The targets $^{2}H$ and $^{9}Be$ obtained values of 1.005% and 1.087%, respectively. From the energetic point of view a gain is insignificant, representing 5.8% and 7.8% of the released energy by nuclear fission, an expected fact due to the small atomic number of the targets. This energy quantity is of a similar magnitude as the energy release by the decay of nuclei that would occur in a pure fission scenario.

Evidently, changes in the composition of materials present in the fuel and moderator assembly have several consequences that we have not mentioned in our discussion, such as the important question of criticality. Aspects that should be analyzed are the relation moderator to fuel together with geometrical arrangements, that have crucial consequences in such a new conception for nuclear reactors. Furthermore, kinetics is expected to suffer from significant changes due to the reduction of decay-chains relevant for the delayed neutron precursor production together with the fusion reactions that produce neutrons. The average number of neutrons that may be supplied is strongly related to the target in question. For a case of thousand fusion reactions an average of 297, 145, 549, and 317 neutrons are added for the targets of $^{1}H$, $^{2}H$, $^{6}Li$, and $^{9}Be$, respectively.

Moreover transport properties of neutrons should also suffer modifications, since the various cross sections also changed. These aspects will be considered in future studies, where point kinetics will be adapted to the cases considered and shall give insight in the criticality aspect of this new conception. Also studies that analyze changes in geometries are in order that may be addressed in approaches like diffusion equation with heterogeneous domains and their associated parameter sets.

Finally, the change in the half-life distribution has its impact on safety, reactor dynamics, and nuclear fuel handling, especially the used fuel. For instance, decays

of the remaining fission products provide approximately 8% of the total energy in reactor operation. After shut-down though this energy is still released according to the decay sequences and makes necessary monitoring and cooling for some additional time. The reported increase in the half-life distribution has the effect of reducing activity of the nuclear waste and consequently its heat production. More aspects could be mentioned but by virtue the presented discussion has still some speculative character and answers to more fundamental questions are required. Hence, this work shall be considered as a first step into a new direction, where pathways may be opened that consider energy production by a fission–fusion conception with its new scientific and technological challenges and its possible benefits.

## References


[Ah80] Ahlen, S.P.: Theoretical and experimental aspects of the energy loss of relativistic heavily ionizing particles. Reviews of Modern Physics **52**, 121–173 (1980).

[AuWa93] Audi, G., Wapstra, A.H.: The 1993 atomic mass evaluation: (I) Atomic mass table. Nuclear Physics A **565**, 1–65 (1993).

[Be88] Beckerman, M.: Sub-barrier fusion of two nuclei. Reports on Progress in Physics **51**, 1047–1103 (1988).

[BeAs53] Bethe, H.A., Ashkin, J.: Experimental Nuclear Physics. Wiley (1953).

[Bo48] Bohr, N.: The penetration of atomic particles through matter. Vidensk. Selsk Mat. Fys. Medd. **18** 1–144 (1948).

[Br62] Brustad, T.: Biological effects of neutron irradiation. Adv. Biol. Med. Phys. **8**, 161–220 (1962)

[EnRi94] England, T.R., Rider, B.F.: Evaluation and Compilation of Fission Product Yields. Los Alamos National Laboratory (1994).

[HaRoKr99] Hagino, H., Rowley, N., Kruppa, A.T.: A program for coupled-channel calculations with all order couplings for heavy-ion fusion reactions. Computer Physics Communications **123**, 143–152 (1999).

[IAEA13] IAEA: http://www-nds.iaea.org/relnsd/vcharthtml/VChartHTML March (2013).

[MuSr74] Mukherji, S., Srivasta, B.K.: Universal range-velocity and stopping-power equations for fission fragments and partially stripped heavy ions in solid media. Physical Review B **9**, 3708–3719 (1974).

[Mu75] Mukherji, S.: Calculation of the mean ionization potentials of the elements for stopping-power computations. Physical Review B **12**, 3530–3532 (1975).

[NaEtAl04] Navin, A. Tripathi, V., Blumenfeld, Y., Nanal, V., Simenel, C.: Direct and compound reactions induced by unstable helium beams near the Coulomb barrier. Physical Review C **70**, 044601 (2004).

[NiDaLa89] Niello, J.F., Dasso, C.H., Landowne S.: CCDEF - A simplified coupled-channel code for fusion cross sections. Computer Physics Communications **54**, 409–412 (1989).

[No60] Northcliffe, L.C.: Energy Loss and Effective Charge of Heavy Ions in Aluminum. Physical Review **120**, 1744–1757 (1960).

[Og59] Oganesyan, Y.T.: Zh. Eksp. Teor. Fiz. **36**, (1959).

[Sc04] Schwabl, F.: Quatum Mechanics. Springer Verlag Publishing (2004).

[SiEtAl10] Sinha, M., Majumdar, H., Basu, P., Roy, S., Biswas, M., Pradhan, M.K.: Sub- and above-barrier fusion of loosely bound $^{6}Li$ with $^{28}Si$. The European Physical Journal A **44**, 403–410 (2010).

[SrMu76] Srivasta, B.K., Mukherji, S.: Range and stopping-power equations for heavy ions. Physical Review A **14**, 718–725 (1976).

[TaBiBa97] Tai, H., Bichsel, H., Badavi, F.F.: Comparison of Stopping Power and Range Databases for Radiation Transport Study. NASA (1997).

[VoEtAl09] Vinodkumar, A.M., Lovel, W., Sprunger, P.H., Prisbey, L., Trinczek, M., Dombsky, M., Machule, P., Kolata, J.J., Roberts, A.: Fusion of $^{9}Li$ with $^{208}Pb$. Physical Review C **80**, 054609 (2009).

[ViKwWa85] Viola, V.E., Kwiatkowski, K., Walker, M.: Systematics of fission fragment total kinetics energy release. Physical Review C **31**, 1550–1552 (1985).

[ZaSa04] Zagrebaev, V.I., Samarin, V.V.: Near-Barrier Fusion of Heavy Nuclei: Coupling of Channels. Physics of Atomic Nuclei **67**, 1462–1477 (2004).

[ZiZiBi10] Ziegler, J.F., Ziegler, M.D., Biersack, J.P.: SRIM - The stopping and range of ions in matter. Nuclear Instruments and Methods in Physics Research Section B **268**, 1818–1823 (2010).

# Chapter 5
# DRBEM Simulation on Mixed Convection with Hydromagnetic Effect

**C. Bozkaya**

The steady and laminar mixed convection flow of a viscous, incompressible, and electrically conducting fluid under the effect of an inclined magnetic field is numerically investigated. Specifically, the two-dimensional flow in a lid-driven cavity with a linearly heated wall is considered. The dual reciprocity boundary element method is used for solving the coupled nonlinear differential equations in terms of stream function, vorticity, and temperature. The study focuses on the effects of the physical parameters, such as Richardson and Hartmann numbers, on the flow field and the temperature distribution at different inclinations of the applied magnetic field. The streamlines and isotherms are used for the visualization of the flow and temperature fields. The code validations in terms of average Nusselt numbers show good agreement with the results given in the literature.

## 5.1 Introduction

A combined forced and free convection flow of an electrically conducting fluid and heat transfer in the lid-driven cavities in the presence of a magnetic field is of great interest due to its many industrial applications such as material processing, dynamics of lakes, geothermal reservoirs, cooling of nuclear reactors, thermal insulation, crystal growing, metal casting, and so on. In mixed convection, the temperature differences across the cavity cause a buoyancy driven whereas the movement of a wall generates a forced convection. The effect of an externally applied magnetic field on the system of a mixed convection flow in enclosures has been investigated numerically in some recent studies. Hossain et al [HoHa05] worked on the buoyancy

C. Bozkaya (✉)
Middle East Technical University, Dumlupinar Bulvari No. 1, 06800 Ankara, Turkey
e-mail: bcanan@metu.edu.tr



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_5

and thermocapillary driven convection flow of an electrically conducting fluid in an enclosure with heat generation subject to a uniform magnetic field by using a finite difference method. They concluded that the applied magnetic field resists the flow and retards the velocity field. Another finite difference solution to the magnetohydrodynamic (MHD) mixed convection at high Hartmann numbers was studied by Kalapurakal et al [KaCh13]. It was observed that the heat transfer was more pronounced only with increased Richardson number. On the other hand, Chatterjee [Ch13] analyzed the magnetoconvective flow and heat transfer in a vertical lid-driven square enclosure with two different types of heat sources by a finite volume approach. He showed that the heat transfer rate and bulk fluid temperature both had increasing function of mixed convective strength. Al-Salem et al [AlOz12] investigated the effects of moving lid direction on MHD mixed convection in a linearly heated cavity by using a finite volume approach. It was found that direction of lid was more effective on heat transfer and fluid flow in the case of mixed convection than it was the case in forced convection, and the heat transfer was decreased with increasing magnetic field. The works of Sivasankaran [SiMa11] and Oztop [OzAl11] are also the finite volume solutions of MHD mixed convection in a lid-driven cavity with the walls of different types of heating.

Rahman et al [RaOz11, RaOz12] investigated the effects of Joule heating and the heater position, respectively, on the flow field and heat transfer by discretizing the governing equation with a finite element approach. It was found in the latter work that the highest heat transfer was obtained when the isothermal heater was located at the right vertical wall. Kefayati et al [KeGo12] used the Lattice Boltzmann simulation of mixed convection in a lid-driven cavity with linearly heated wall under the effect of magnetic field.

In the present study, we focus on the dual reciprocity boundary element method (DRBEM) solution of the MHD mixed convection flow problem, introduced in the work [KeGo12], in a lid-driven square cavity subject to an inclined magnetic field. The effects of controlling parameters including Hartmann and Richardson numbers, and the influence of various inclination angles of the magnetic field on the flow field and temperature distribution are investigated. In this paper, the stream function–vorticity–temperature formulation of mixed convection flow under the effect of an external magnetic field is followed. The DRBEM, which is a boundary only nature technique, is used to treat the terms except the Laplace operator as the inhomogeneity. These three equations are solved iteratively with the given boundary conditions for stream function and the temperature. However, the vorticity boundary conditions are obtained by using radial basis functions in the stream function equation, which is an advantage of the DRBEM.

## 5.2 Problem Formulation and Governing Equations

Figure 5.1 displays the schematic of the considered model. It is a two-dimensional cavity of which top wall moves horizontally at a constant velocity $U_0$ and is considered to be adiabatic. The vertical left wall is linearly heated whereas the right

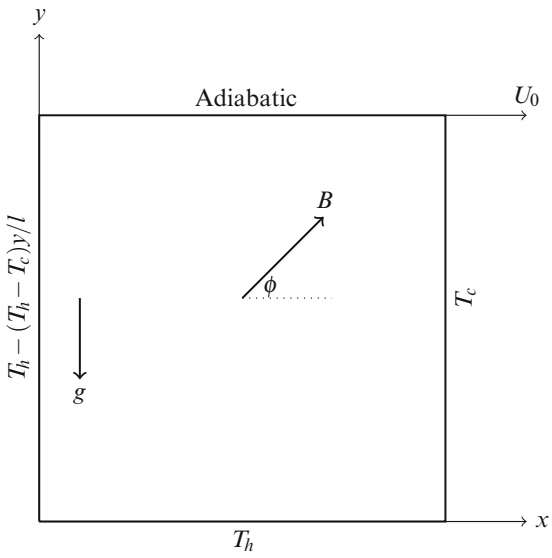


**Fig. 5.1** Geometry of the problem

wall is kept at a constant cold temperature $T_c$. The bottom wall of the cavity is maintained at a constant hot temperature $T_h$. Air is selected as the working fluid at a Prandtl number ($Pr$) of 0.71. The gravity acts downwards and the uniform magnetic field of a constant strength $B_0$ is imposed with an inclination angle $\phi$. The viscous dissipation and Joule heating effects are taken as negligible. In addition, the magnetic Reynolds number is assumed to be small so that the induced magnetic field is neglected. All fluid physical properties are assumed to be constant except the density variations according to the Boussinesq approximation. The governing equations for the problem under consideration are based on the conservation laws of mass, momentum, and thermal energy in two dimensions. Thus, following the aforementioned assumptions, these equations of the steady, laminar flow of a viscous and electrically conducting fluid subjected to a uniform inclined magnetic field can be written in the non-dimensional form [KeGo12] as

$$\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}=0 \tag{5.1}$$

$$\begin{aligned}\frac{1}{Re}\nabla^2 u &= \frac{\partial p}{\partial x}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}-\frac{Ha^2}{Re}(v\sin\phi\cos\phi-u\sin^2\phi)\\ \frac{1}{Re}\nabla^2 v &= \frac{\partial p}{\partial y}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}-RiT-\frac{Ha^2}{Re}(u\sin\phi\cos\phi-v\cos^2\phi)\end{aligned} \tag{5.2}$$

$$\frac{1}{RePr}\nabla^2 T=u\frac{\partial T}{\partial x}+v\frac{\partial T}{\partial y} \tag{5.3}$$

where $u$, $v$, $p$, $T$ are the $x$- and $y$-velocity components, pressure, and the temperature of the fluid, respectively. In these equations, the nondimensional parameters are defined as: $Re = U_0 l/\nu$, the Reynolds number, $Pr = \nu/\alpha$, the Prandtl number, $Gr = g\beta\Delta T l^3/\nu^2$, the Grashof number, $Ha = B_0 l\sqrt{\sigma/\mu}$, the Hartmann number and $Ri = Gr/Re^2$, Richardson number where this ratio is used to indicate the relative strengths of the two modes of convection in a mixed convection. Here, $\nu$, $\alpha$, $\beta$, $\sigma$, and $\mu$ are the kinematic viscosity, the thermal diffusivity, the thermal expansion coefficient, electrical conductivity, and the viscosity coefficients of the fluid, respectively. The temperature difference is $\Delta T = T_h - T_c$, and $U_0$, $l$ are the reference velocity and length, respectively.

The nondimensional boundary conditions corresponding to the considered problem are

$$\begin{array}{ll}
\text{On the sliding top lid:} & u = 1,\ v = 0,\ \partial T/\partial n = 0 \\
\text{On the bottom wall:} & u = v = 0, \quad T = 1 \\
\text{On the left vertical wall:} & u = v = 0, \quad T = 1 - y \\
\text{On the right vertical wall:} & u = v = 0, \quad T = 0\,.
\end{array}$$

In order to eliminate the pressure and to satisfy the continuity equation automatically, we introduce the stream function $\psi(x,y)$ and the vorticity $w(x,y)$ with

$$\frac{\partial \psi}{\partial y} = u, \quad \frac{\partial \psi}{\partial x} = -v, \quad w = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\,.$$

Then, the equations (5.1)–(5.3) are transformed into

$$\nabla^2 \psi = -w \tag{5.4}$$

$$\begin{aligned}
\nabla^2 w = Re\Big(\frac{\partial w}{\partial x}\frac{\partial \psi}{\partial y} - \frac{\partial w}{\partial y}\frac{\partial \psi}{\partial x}\Big) - ReRi\frac{\partial T}{\partial x} \\
-Ha^2\left(\frac{\partial^2 \psi}{\partial x \partial y}\sin 2\phi + \frac{\partial^2 \psi}{\partial x^2}\cos^2\phi + \frac{\partial^2 \psi}{\partial y^2}\sin^2\phi\right)
\end{aligned} \tag{5.5}$$

$$\nabla^2 T = PrRe\Big(\frac{\partial T}{\partial x}\frac{\partial \psi}{\partial y} - \frac{\partial T}{\partial y}\frac{\partial \psi}{\partial x}\Big) \tag{5.6}$$

with the corresponding boundary conditions

$$\begin{array}{ll}
\text{On the sliding top lid:} & \psi = 0,\ \psi_x = 0,\ \psi_y = 1, \quad \partial T/\partial n = 0 \\
\text{On the bottom wall:} & \psi = 0,\ \psi_x = \psi_y = 0, \quad T = 1 \\
\text{On the left vertical wall:} & \psi = 0,\ \psi_x = \psi_y = 0, \quad T = 1 - y \\
\text{On the right vertical wall:} & \psi = 0,\ \psi_x = \psi_y = 0, \quad T = 0\,.
\end{array} \tag{5.7}$$

On the other hand, the unknown boundary values for vorticity will be obtained from the stream function equation (5.4) by using a radial basis function approximation during the application of DRBEM.

## 5.3 Method of Solution

The governing equations (5.4)–(5.6) along with boundary conditions (5.7) are discretized using the dual reciprocity boundary element method. Since the MHD mixed convection equations are nonlinear and coupled in terms of stream function, vorticity, and temperature, they are solved in an iterative manner.

The aim of the DRBEM is to transform the governing equations of the problem into boundary integral equations. In the application, the terms except the Laplacian will be treated as inhomogeneity, [BrPa92], and the equations (5.4)–(5.6) are weighted with the two-dimensional fundamental solution of Laplace equation, $u^* = 1/2\pi \ln(1/r)$. Following the application of the Green's second identity, the equations (5.4)–(5.6) become

$$c_i \psi_i + \int_\Gamma (q^* \psi - u^* \frac{\partial \psi}{\partial n}) d\Gamma = - \int_\Omega (-w) u^* d\Omega \tag{5.8}$$

$$\begin{aligned} c_i w_i + \int_\Gamma (q^* w - u^* \frac{\partial w}{\partial n}) d\Gamma = - \int_\Omega \Bigg( & Re(\frac{\partial w}{\partial x}\frac{\partial \psi}{\partial y} - \frac{\partial w}{\partial y}\frac{\partial \psi}{\partial x}) - ReRi\frac{\partial T}{\partial x} \\ & -Ha^2(\frac{\partial^2 \psi}{\partial x \partial y}\sin 2\phi + \frac{\partial^2 \psi}{\partial x^2}\cos^2\phi + \frac{\partial^2 \psi}{\partial y^2}\sin^2\phi) \Bigg) u^* d\Omega \end{aligned} \tag{5.9}$$

$$c_i T_i + \int_\Gamma (q^* T - u^* \frac{\partial T}{\partial n}) d\Gamma = - \int_\Omega PrRe(\frac{\partial T}{\partial x}\frac{\partial \psi}{\partial y} - \frac{\partial T}{\partial y}\frac{\partial \psi}{\partial x}) u^* d\Omega \tag{5.10}$$

where $q^* = \partial u^*/\partial n$, $\Gamma$ is the boundary of the domain $\Omega$ and the subscript $i$ denotes the source point. The constant $c_i$ is given by $c_i = \theta_i/2\pi$ with the internal angle $\theta_i$ at the source point.

The integrands of the domain integrals on the right-hand side of Equations (5.8)–(5.10) are treated as inhomogeneity. Thus, they are approximated by a set of radial basis functions $f_j(x,y)$ linked with the particular solutions $\hat{u}_j$ to the equation $\nabla^2 \hat{u}_j = f_j$. The approximations for these integrands are given by $\sum_{j=1}^{N+L} \alpha_j f_j(x,y)$, $\sum_{j=1}^{N+L} \beta_j f_j(x,y)$, and $\sum_{j=1}^{N+L} \gamma_j f_j(x,y)$, respectively, for Equations (5.8), (5.9), and (5.10). The coefficients $\alpha_j$, $\beta_j$, and $\gamma_j$ are undetermined constants. The numbers of the boundary and the internal nodes are denoted by $N$ and $L$, respectively. Now, the right-hand sides of Equations (5.8)–(5.10) also involve the multiplication of the Laplace operator with the fundamental solution $u^*$, which can be treated in a similar manner by the use of DRBEM, [BrPa92], to obtain the following boundary only integrals,

$$c_i\psi_i + \int_\Gamma (q^*\psi - u^*\frac{\partial \psi}{\partial n})d\Gamma = \sum_{j=1}^{N+L} \alpha_j \left[ c_i\hat{u}_{ji} + \int_\Gamma (q^*\hat{u}_j - u^*\hat{q}_j)d\Gamma \right] \tag{5.11}$$

$$c_i w_i + \int_\Gamma (q^* w - u^*\frac{\partial w}{\partial n})d\Gamma = \sum_{j=1}^{N+L} \beta_j \left[ c_i\hat{u}_{ji} + \int_\Gamma (q^*\hat{u}_j - u^*\hat{q}_j)d\Gamma \right] \tag{5.12}$$

$$c_i T_i + \int_\Gamma (q^* T - u^*\frac{\partial T}{\partial n})d\Gamma = \sum_{j=1}^{N+L} \gamma_j \left[ c_i\hat{u}_{ji} + \int_\Gamma (q^*\hat{u}_j - u^*\hat{q}_j)d\Gamma \right] \tag{5.13}$$

where $\hat{q} = \partial\hat{u}_j/\partial n$. The use of constant boundary elements for the discretization of the boundary leads to the corresponding matrix-vector form of Equations (5.11)–(5.13)

$$H\psi - G\frac{\partial \psi}{\partial n} = (H\hat{U} - G\hat{Q})F^{-1}\{-w\}\,, \tag{5.14}$$

$$\begin{aligned}(Hw - G\frac{\partial w}{\partial n}) = (H\hat{U} - G\hat{Q})F^{-1}\Big\{ & Re(\frac{\partial w}{\partial x}\frac{\partial \psi}{\partial y} - \frac{\partial w}{\partial y}\frac{\partial \psi}{\partial x}) - ReRi\frac{\partial T}{\partial x} \\ & - Ha^2(\frac{\partial^2 \psi}{\partial x \partial y}\sin 2\phi + \frac{\partial^2 \psi}{\partial x^2}\cos^2\phi + \frac{\partial^2 \psi}{\partial y^2}\sin^2\phi)\Big\}\end{aligned} \tag{5.15}$$

$$(HT - G\frac{\partial T}{\partial n}) = (H\hat{U} - G\hat{Q})F^{-1}\left\{ PrRe(\frac{\partial T}{\partial x}\frac{\partial \psi}{\partial y} - \frac{\partial T}{\partial y}\frac{\partial \psi}{\partial x}) \right\} \tag{5.16}$$

where the matrices $\hat{U}$ and $\hat{Q}$ are constructed by taking each of the vectors $\hat{u}_j$ and $\hat{q}_j$ as columns, respectively. The $(N+L)\times(N+L)$ matrix $F$ contains the coordinate functions $f_j$ as columns. The components of the matrices $H$ and $G$ are

$$H_{ij} = c_i\delta_{ij} + \frac{1}{2\pi}\int_{\Gamma_j} \frac{\partial}{\partial n}\left(\ln(\frac{1}{r})\right) d\Gamma_j, \quad H_{ii} = -\sum_{j=1, j\neq i}^{N} H_{ij}$$

$$G_{ij} = \frac{1}{2\pi}\int_{\Gamma_j} \ln(\frac{1}{r})\, d\Gamma_j, \quad G_{ii} = \frac{A}{2\pi}(\ln(2/A) + 1)$$

where $r$ is the distance from node $i$ to element $j$, $A$ is the length of the element, and $\delta_{ij}$ is the Kronecker delta function. In order to solve the resulting DRBEM equations, which are nonlinear and coupled, we need to use an iterative process with initial estimates of vorticity and temperature. First, the stream function equation (5.14) is solved by giving an initial estimate for vorticity. Thus, we obtain both the interior and boundary values of stream function, which will be used to calculate the $x$- and $y$-derivatives of itself by means of polynomial type radial basis functions.

The insertion of these derivative values in the vorticity equation (5.15) and the use of an initial estimate for the temperature lead to the linearization of the vorticity equation. Once the vorticity values are obtained at all points in the domain, a similar procedure is employed for the solution of the energy equation (5.16). In each iteration, the required space derivatives of the unknowns $\psi$, $w$, and $T$, and also the unknown vorticity boundary conditions are obtained by using the coordinate matrix $F$ as

$$\frac{\partial R}{\partial x} = \frac{\partial F}{\partial x}F^{-1}R, \quad \frac{\partial R}{\partial y} = \frac{\partial F}{\partial y}F^{-1}R, \quad w = -(\frac{\partial^2 F}{\partial x^2}F^{-1}\psi + \frac{\partial^2 F}{\partial y^2}F^{-1}\psi)$$

where $R$ is one of the unknowns $\psi$, $w$, or $T$. This can be regarded as one of the advantages of DRBEM. The iterative procedure will stop when a preassigned tolerance is reached between two successive iterations.

## 5.4 Numerical Results and Discussion

The DRBEM analysis for the two-dimensional MHD mixed convection flow under consideration is performed at the Reynolds number of $Re = 100$ to investigate the effects of the Hartmann and Richardson numbers on the flow and temperature fields. The boundaries of the cavity are discretized by using an adequate number of constant boundary elements according to different combination of $Ha$ and $Ri$ values. For example, maximum $N = 240$ constant boundary elements are used for the case when $Ha = 50$ and $Ri = 100$.

First, to determine the accuracy of the present numerical algorithm, the average Nusselt number is calculated for a lid-driven cavity whose top wall is moving to the right with a constant velocity in the absence of a magnetic field ($Ha = 0$) when $Gr = 100$ and $Pr = 0.71$. The agreement of the present results with the ones available in the literature is presented in Table 5.1 for several values of Reynolds and Richardson numbers.

Figure 5.2 shows the effect of the Richardson number in the absence of the magnetic field, $Ha = 0$. As $Ri$ increases, the core vortex of the streamlines concentrated at the top of the cavity in the direction of the lid-driven velocity moves downwards

**Table 5.1** Comparison of average Nusselt number with available results in literature.

| $Re$ | $Ri$ | Present | [Wa09] | [KeGo12] | [KhCh99] | [IwHy93] |
|---|---|---|---|---|---|---|
| 1 | 100 | 0.99258 | 1.00033 | 1.0094 | – | – |
| 100 | 0.01 | 2.0883 | 2.03116 | 2.09 | 2.01 | 1.94 |
| 400 | 0.00062 | 4.2914 | 4.0246 | 4.08082 | 3.91 | 3.84 |
| 1000 | 0.0001 | 6.5134 | 6.48423 | 6.54687 | 6.33 | 6.33 |

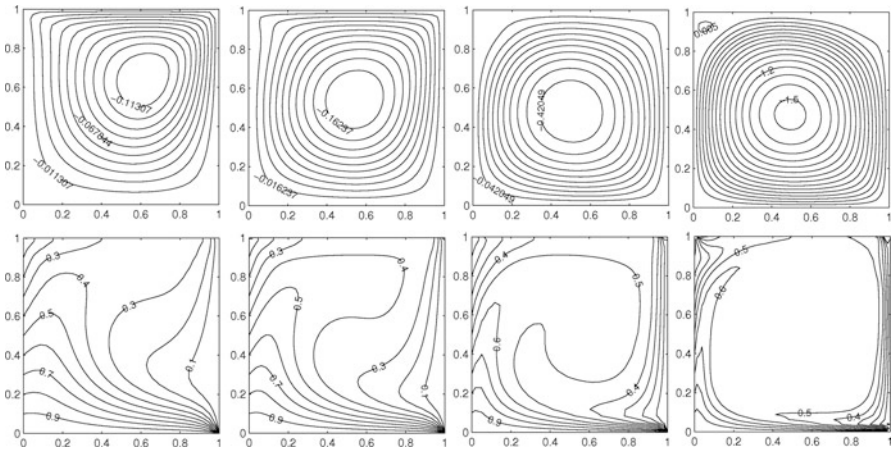

**Fig. 5.2** The effect of Richardson number on streamlines (top) and isotherms (bottom) for $Ha = 0$: (a) $Ri = 0.01$, (b) $Ri = 1$, (c) $Ri = 10$, (d) $Ri = 100$.

through the center of the cavity. In addition, the stream function values increase in magnitude as the flow regime is transformed from forced convection ($Ri = 0.01$) to mixed convection ($Ri = 1$), and finally to natural convection ($Ri = 10$, 100). When $Ri = 100$, a small secondary vortex is formed at the top left corner of the cavity. On the other hand, the gradient of the temperature at heated and linearly heated walls increases as $Ri$ increases. Moreover, the isotherms move towards to the walls of the cavity with an increase in $Ri$.

The flow behavior and the temperature distribution for the cases of the forced convection $Ri = 0.01$, the mixed convection $Ri = 1$, and the natural convection $Ri = 100$ (from top to bottom) when (a) $Ha = 25$, (b) $Ha = 50$ are visualized in Figure 5.3 and Figure 5.4 for the inclination angles $\phi = 0$ and $\phi = \pi/2$, respectively. The formation of new circulations occurs in the flow field with the application of the magnetic field horizontally ($\phi = 0$) at $Ri = 0.01$. That is, as Hartman number increases from $Ha = 0$ to $Ha = 25$ (see top rows of Figure 5.2 and Figure 5.3), a secondary vortex develops at the bottom of the cavity; and the value of the stream function for the main flow decreases in magnitude. When Hartmann number increases to $Ha = 50$, the values of the $\psi$ continue to decline in magnitude, and a tertiary weak circulation which agrees with the core vortex develops at the bottom of the cavity. Further, the secondary circulation, which is counterclockwise, improves and the main flow at the top of the cavity weakens for the higher value of Hartmann number $Ha = 50$. On the other hand, when the magnetic field is applied vertically ($\phi = \pi/2$) at $Ri = 0.01$ (see top row of Figure 5.4), the core of the main flow moves towards the top wall and extends horizontally for both $Ha = 25$ and $Ha = 50$. Moreover, the values of $\psi$ drop steadily in magnitude by an increase in $Ha$ similar to the case when $\phi = 0$, $Ri = 0.01$.

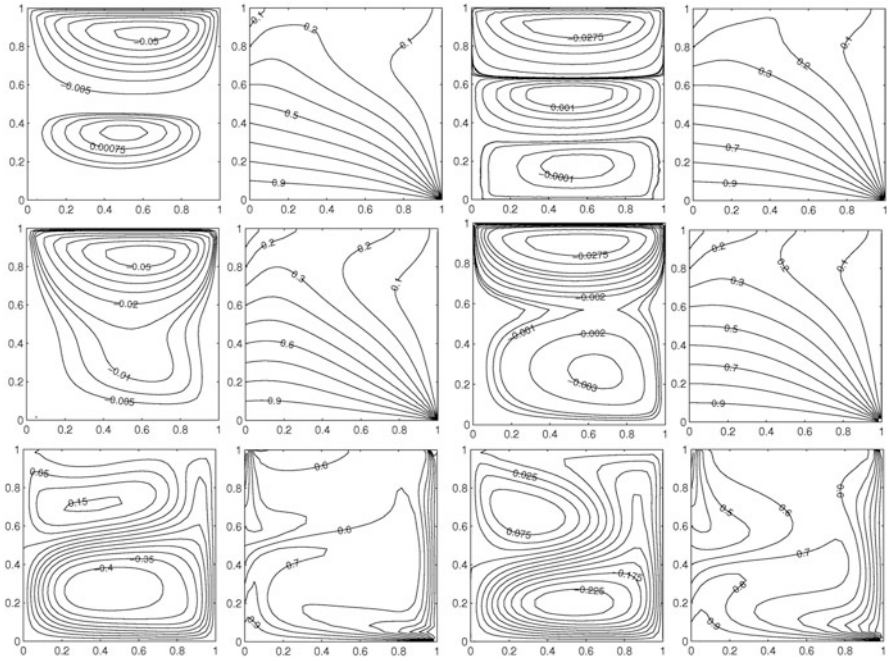

**Fig. 5.3** Streamlines and isotherms for $Ri = 0.01$ (top), $Ri = 1$ (middle), $Ri = 100$ (bottom), $\phi = 0$: (a) $Ha = 25$, (b) $Ha = 50$.

At $Ri = 1$ when the magnetic field is applied horizontally (see middle row in Figure 5.3), the core vortex of the flow moves towards the top right corner of the cavity with an increase in $Ha$ from 0 to 25, and the main vortex is separated into two pieces with the application of a higher strength of magnetic field of $Ha = 50$. The isotherms have similar profiles with $Ri = 0.01$ for both Hartmann numbers. On the other hand, for the case of vertically applied magnetic field $\phi = \pi/2$ (see middle row in Figure 5.4), the core circulation of the flow moves to the right and up and extends in $y$-direction as $Ha$ increases. The behavior of the isotherms is also similar to the ones at $Ri = 0.01$. However, at $Ha = 25$ the isotherms where $T = 0.3$ show more convection and rise towards the top wall. For this case, the streamlines have also a downward inclination towards the cold wall and the values of $\psi$ increase in magnitude when compared to the case $Ri = 0.01$. It is also observed that the inclusion of the magnetic field with inclination angles $\phi = 0,\ \pi/2$ reduces the effect of an increase in Richardson number from $Ri = 0.01$ to $Ri = 1$ on isotherms, especially when $T = 0.3$, 0.4 in the case of absence of magnetic field displayed in Figure 5.2(b).

At the highest Richardson number $Ri = 100$ when $\phi = 0$, the isotherms rise towards the top wall and a depletion of the temperature gradient on the hot bottom wall is seen as $Ha$ increases. At $Ha = 25$, the core of the main flow is suppressed down following the formation of a secondary circulation close to the upper side of

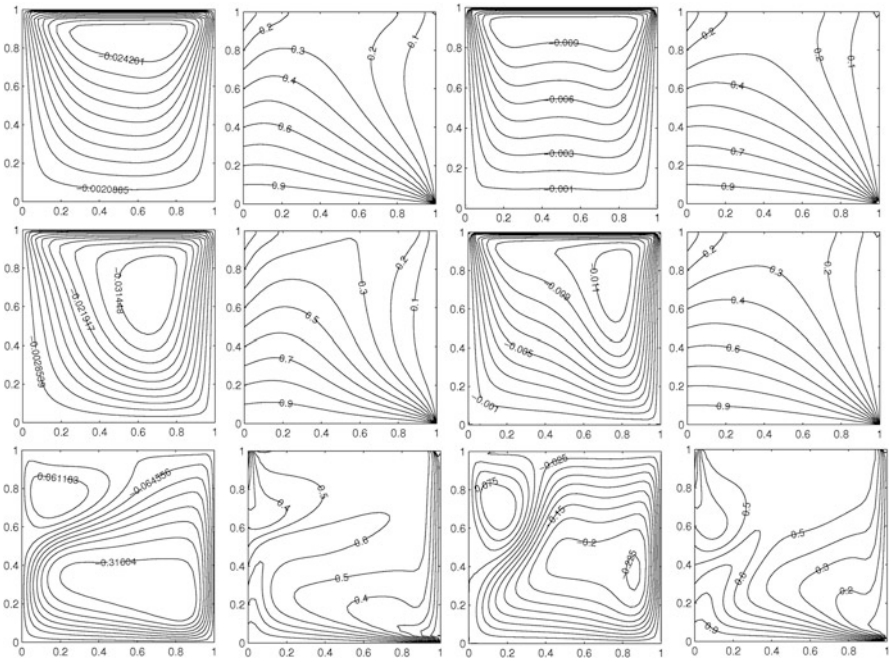

**Fig. 5.4** Streamlines and isotherms for $Ri = 0.01$ (top), $Ri = 1$ (middle), $Ri = 100$ (bottom), $\phi = \pi/2$ : (a) $Ha = 25$, (b) $Ha = 50$.

linearly heated wall. This secondary circulation weakens and a movement at the top of the cavity for $\psi$ is observed as $Ha$ increases to 50. On the other hand, when the magnetic field is applied vertically $\phi = \pi/2$, similar to the case when $\phi = 0$, a secondary vortex formation occurs at the left top corner of the cavity and it becomes intense as $Ha$ increases to 50. Moreover, the main core of the flow gains an elliptical form when compared to the case $Ri = 1$. The isotherms are noticeably affected at $Ri = 100$ and the isotherms close to the cold wall moves up, that is, the heat transfer increases.

Finally, the effect of the direction of the applied magnetic field on the flow behavior and temperature distribution is analyzed by taking three inclination angles $\phi = \pi/6, \pi/4, \pi/3$ at fixed $Ha = 25$ and $Ri = 1$. The streamlines and the isotherms are displayed in Figure 5.5. Although the isotherms show almost a similar behavior in each $\phi$, the isotherms where $T = 0.3$ rise slightly towards the top wall in the direction of applied magnetic field as $\phi$ increases. On the other hand, the core vortex of the streamlines is affected significantly with a change in the direction of the magnetic field. That is, the main circulation concentrated at the top right corner of the cavity when $\phi = \pi/6$ extends in the direction of the applied magnetic field and its magnitude decreases slightly as the inclination angle increases to $\pi/3$.

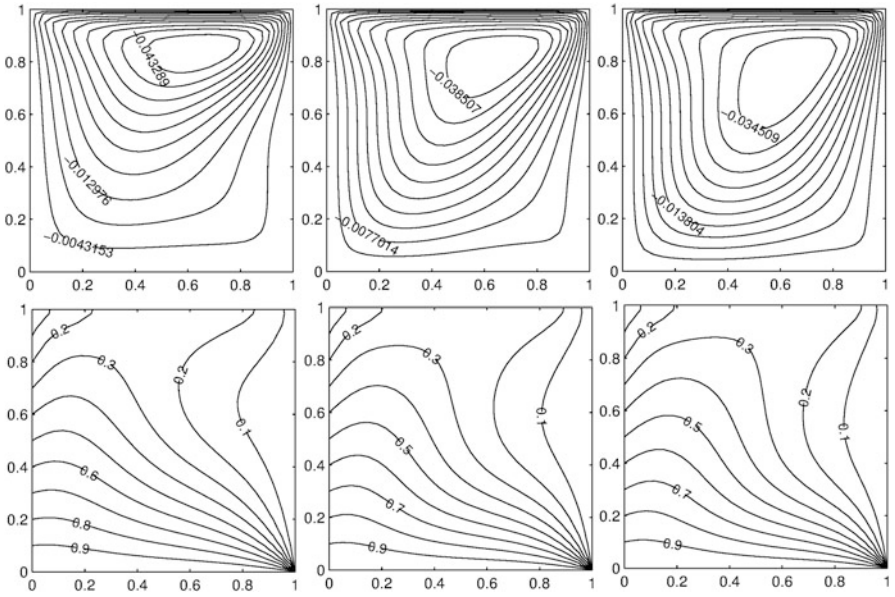


**Fig. 5.5** The effect of the inclination angle $\phi$ on streamlines (top) and isotherms (bottom) for $Ha = 25$, $Ri = 1$: (a) $\phi = \pi/6$, (b) $\phi = \pi/4$, (c) $\phi = \pi/3$.

## 5.5 Conclusions

The mixed convection flow in a lid-driven cavity with a linearly heated wall is analyzed by using the dual reciprocity BEM with constant elements. Numerical simulations are carried out for pertinent parameters in ranges: $Ri = 0.01 - 100$, $Ha =$ 0–50, and for a fixed Reynolds number $Re = 100$ with several inclination angles. The obtained results show that the flow behavior and the heat transfer characteristics are significantly influenced by the use of different combination of Richardson and Hartmann numbers. Formation of additional circulations is observed in the flow field by a transition from the forced and mixed convection flow regimes to the dominating natural convection flow regime as the Hartmann number increases. The application of the magnetic field reduces the effect of the Richardson number on the temperature distribution during the transition between forced, mixed, and natural convection regimes. Thus, an external magnetic field in different directions can be used to control the behavior of the flow and heat transfer in a cavity. All the present results are in good agreement with the previously published results given in [KeGo12].

## References

[HoHa05] Hossain, M.A., Hafiz, M.Z., Rees, D.A.S.: Buoyancy and thermocapillary driven convection flow of an electrically conducting fluid in an enclosure with heat generation, Int. J. Thermal Sciences, **44**: 676–684 (2005)

[KaCh13] Kalapurakal, D., Chandy, A.J.: Accurate and efficient numerical simulations of magnetohydrodynamic (MHD) mixed convection at high Hartmann numbers, Numeric. Heat Transfer, Part A, **64**: 527–550 (2013)

[Ch13] Chatterjee, D.: MHD mixed convection in a lid-driven cavity including a heated source, Numeric. Heat Transfer, Part A, **64**: 235–254 (2013)

[AlOz12] Al-Salem, K., Oztop, H.F., Pop, I., Varol, Y.: Effects of moving lid direction on MHD mixed convection in a linearly heated cavity, Int. J. Heat Mass Transfer **55**: 1103–1112 (2012)

[SiMa11] Sivasankaran, S., Malleswaran, A., Lee, J., Sundar, P.: Hydro-magnetic combined convection in a lid-driven cavity with sinusoidal boundary conditions on both sidewalls, Int. J. Heat Mass Transfer **54**: 512–525 (2011)

[OzAl11] Oztop, H.F., Al-Salem, K., Pop, I.: MHD mixed convection in a lid-driven cavity with corner heater, Int. J. Heat Mass Transfer **54**: 3494–3504 (2011)

[RaOz12] Rahman, M.M., Oztop, H.F., Rahim, N.A., Saidur, R., Al-Salem, K., Amin, N.: Computational analysis of mixed convection in a channel with a cavity heated from different sides, Int. Communications Heat Mass Transfer **39**: 78–84 (2012)

[RaOz11] Rahman, M.M., Oztop, H.F., Rahim, N.A., Saidur, R., Al-Salem, K.: MHD mixed convection with Joule heating effect in a lid-driven cavity with a heated semicircular source using the finite element technique, Numeric. Heat Transfer, Part A, **60**: 543–560 (2011)

[KeGo12] Kefayati, GH.R., Gorji-Bandpy, M., Sajjadi, H., Ganji, D.D.: Lattice Boltzmann simulation of MHD mixed convection in a lid-driven square cavity with linearly heated wall, Scientia Iranica B **19(4)**: 1053–1065 (2012)

[BrPa92] Brebbia, C.A., Partridge, P.W., Wrobel, L.C.: The Dual Reciprocity Boundary Element Method. Computational Mechanics Publications, Southampton, Boston (1992)

[KhCh99] Khanafer, K., Chamkha A.J.: Mixed convection flow in a lid-driven enclosure filled with a fluid-satured porous medium, Int. J. Heat Mass Transfer **42**: 2465–2481 (1999)

[IwHy93] Iwatsu, R., Hyun, J.M., Kuwahara, K.: Mixed convection in a driven cavity with a stable vertical temperature gradient, Int. J. Heat Mass Transfer, **36**: 1601–1608 (1993)

[Wa09] Waheed, M.A.: Mixed convective heat transfer in rectangular enclosures driven by a continuously moving horizontal plate, Int. J. Heat Mass Transfer, **52**: 5055–5063 (2009)

# Chapter 6
# Nonlinear Method of Reduction of Dimensionality Based on Artificial Neural Network and Hardware Implementation

**J.R.G. Braga, V.C. Gomes, E.H. Shiguemori, H.F.C. Velho, A. Plaza, and J. Plaza**

## 6.1 Introduction

The technological development of imaging sensors of high spectral resolution, called multi- or hyper-spectral sensors, enables the acquisition of information on dozens up to thousand of spectral bands. Due to the large amount of available information, the reduction of the dimensions for the provided data, without loss of information, is a challenge [Ch13].

There are several schemes for reducing the dimensionality of data, one of them is the Principal Component Analysis (PCA) [An09]. Such analysis deals with linear transformation, and this limitation can influence the data classification for hyper-spectral sensors [LiEtAl12]. This is a motivation to study of nonlinear techniques for data reduction. One of such techniques is the Nonlinear Principal Component Analysis (NL-PCA), based on artificial neural networks (ANN) [LiEtAl12, DeLiDu09].

In this chapter, a multi-layer perceptron ANN [SiEtAl13] classifier, with back propagation for training, is employed. The general procedure to configure an ANN is an empirical one, where the ANN architecture is defined by an expert. Here, a self-configuring strategy is applied, where the optimal NN architecture is obtained

J.R.G. Braga (✉) • H.F.C. Velho
National Institute for Space Research, Av. dos Astronautas 1758,
São José dos Campos, SP, Brazil
e-mail: jgarciabraga@gmail.com; haroldo@lac.inpe.br

V.C. Gomes • E.H. Shiguemori
Department of Science and Aerospace Technology, São José dos Campos, SP, Brazil
e-mail: vcconrado@gmail.com; elcio@ieav.cta.br

A. Plaza • J. Plaza
University of Extremadura, Av. de la Universidad s/n, 10003 Cáceres, Spain
e-mail: aplaza@unex.es; jplaza@unex.es



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_6

by solving an optimization problem. A new metaheuristic, named Multi-particle Collision Algorithm (MPCA) [LuBeVe08], is used to compute the minimum value for the objective function.

Finally, all optimal MLP-NNs are implemented on a hardware component: Field Programmable Gate Arrays. The use of hardware divide allows a fast parallel image processing with low energy demand.

## 6.2 Methodology

Figure 6.1 exhibits the methodology followed in this study.

### *6.2.1 Principal Component Analysis*

The Principal Component Analysis (PCA) can be used for data reduction by eliminating less representative information [GoWo00]. The PCA reduction is based on selecting a smaller data set, but with almost the same variance from the original data. An algorithm for finding the principal components from a data set is expressed below:

1. Given a data set with $n$ vectors with dimension $m$;

$$x_1 = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \ldots x_n = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}.$$

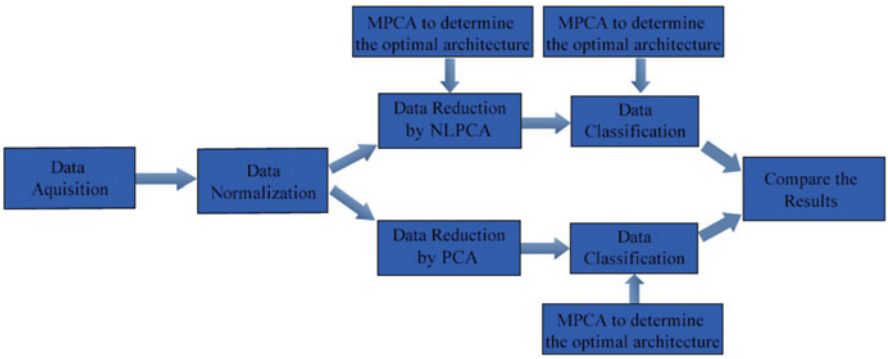


**Fig. 6.1** Flowchart of the methodology used.

2. From these vectors, calculate the average $\mu_x$.
3. Compute a new vector from the data set: $v_i = x_i - \mu_x$ $(i = 1, 2, 3, \ldots, n)$.
4. Multiplying the vector $v_i$ by its transpose: $A_i = v_i \times v_i)^T$.
5. Covariance matrix: perform the sum of matrices above and divide by $n$:

$$M_{\text{cov}} = \frac{1}{n} \sum_{i=1}^{n} A_i \ . \tag{6.1}$$

6. Compute all the eigenvalues and eigenvectors of the covariance matrix (the QR method, or even the deflation technique [An09]). The eigenvector set consists of the principal components from the data set.
7. For generating the data set with reduced size, determine a reference eigenvalue. After that, consider only the reduced matrix containing the maximum eigenvalue (in module) up to the reference eigenvalue.

### *6.2.2 Artificial Neural Network*

The Artificial Neural Network (ANN) is a machine designed to emulate the human brain [Ha01], where:

(a) knowledge is acquired through a learning process;
(b) the basic unit of operation is the artificial neuron;
(c) connections among neurons, called synapses, store the acquired knowledge.

The ANN is usually implemented using electronic components, it can be simulated by programming in a digital computer. The output of an ANN is given by

$$y_k = \varphi(v_k) \quad (k = 1, 2, \ldots, m) \tag{6.2}$$

where $\varphi(.)$ is the activation function, and $v_k$ is a linear combination of all inputs $x_j$ $(j = 1, 2, ..., n)$ multiplied by their respective synaptic weight $w_{kj}$. The activation function is the nonlinear component for this mapping. Heaviside, sigmoid, hyperbolic tangent functions are usually used as an activation function in an artificial neuron.

Figure 6.2 displays a representation for an artificial neuron. More than one hidden layer can be employed to define an ANN. A very popular topology for ANN is the multi-layer perceptron (MLP). Figure 6.2 shows a schematic of artificial neuron.

The most popular algorithm to determine the connection weights (learning phase) is the error back-propagation algorithm [RuHiWi86]. The latter algorithm is an example of supervised learning by error correction [Ha01]. Such learning algorithm can be divided into two distinct steps:

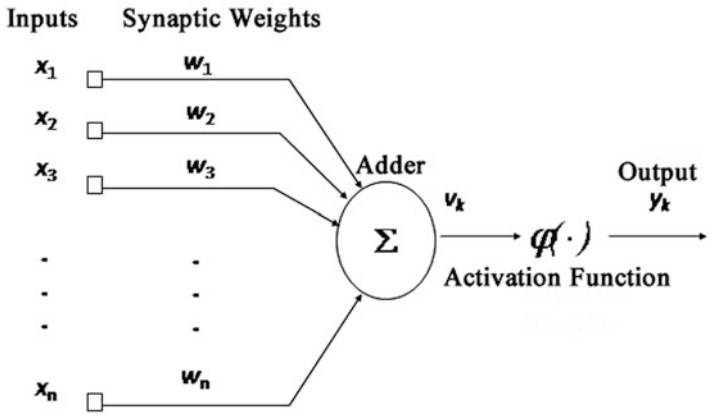


**Fig. 6.2** Representation for an artificial neuron.

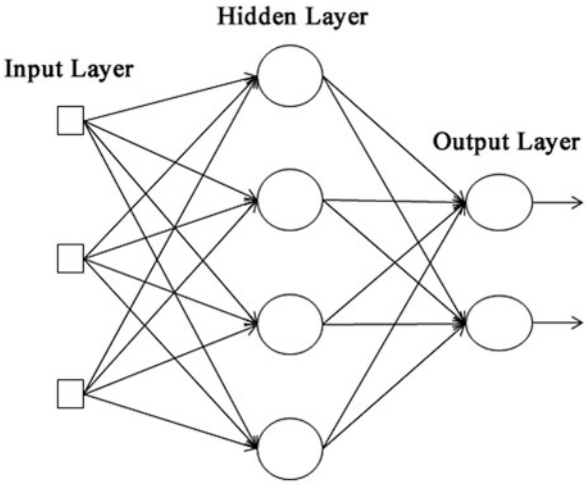


**Fig. 6.3** Multi-layer perceptron ANN architecture.

(i) Each input pattern produces a response (output), and a value of error is obtained by comparison with the target set.
(ii) The weights are updated from the calculated error.

### *6.2.3 Self-Associative Artificial Neural Network*

Consider a fully connected MLP neural network with three hidden layers—see Figure 6.4. The purpose of such NN is to produce an output identical to the input data [DeLiDu09, LiEtAl12].

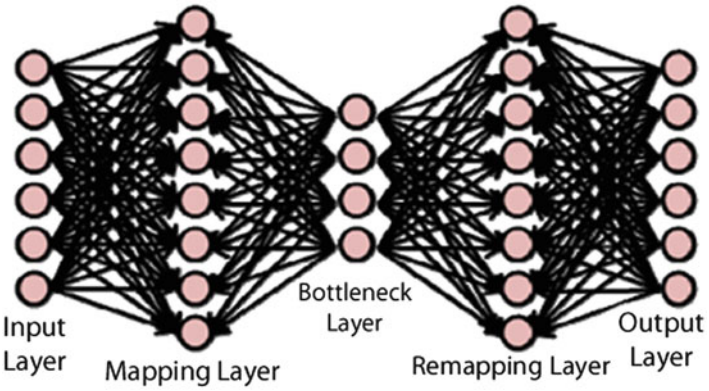


**Fig. 6.4** Outline for a self-associative ANN.

The ANN with the architecture showed in Figure 6.4 is the operator of Nonlinear Principal Component Analysis (NL-PCA). The numbers of neurons in the input and output layers are the same. The mapping layer and re-mapping layer have a sufficient amount of neurons (equal). The output of this ANN is an approximation of the input data. The bottleneck layer has a (much) smaller amount of neurons than the other hidden layers. This is a nonlinear representation of the input data with dimension reduction. For practical purposes, we will not be dealing with raw input data, but with data emerging from the bottleneck layer [DeLiDu09].

### *6.2.4 Multi-Particle Collision Algorithm*

Artificial neural networks have huge success in many applications. However, a tedious job that requires participation of an expert is the configuration of a neural network. Here, the problem of finding an optimal configuration for the neural network is formulated as an optimization problem, where the objective function is expressed as:

$$J(z) = penalty \times \left( \frac{\rho_1 \times E_{\text{train}} + \rho_2 \times E_{\text{gen}}}{\rho_1 + \rho_2} \right) \tag{6.3}$$

where $\rho_1 = 1$ and $\rho_2 = 0.1$ are the same values proposed by [CaRaCh11], which are adjustment factors that magnify the relevance attributed to the training error $E_{\text{train}}$ (see Eq. 6.4), and generalization error $E_{\text{gen}}$ (see Eq. 6.5), respectively [CaRaCh11].

$$E_{\text{train}} = \frac{1}{N} \sum_{k=1}^{N} (y_k - \hat{y}_k)^2 \tag{6.4}$$

$$E_{\text{gen}} = \frac{1}{M-(N+1)} \sum_{k=N+1}^{M} (y_k - \hat{y}_k)^2 \tag{6.5}$$

with $y_k$ and $\hat{y}_k$ being the ANN output and the target value, respectively. The unknown vector $z$ has 5 entries: # hidden layers (max = 3), # neurons for each hidden layer (max = 32), the learning ratio, momentum parameter – both used during the training phase, and type of activation function (only three: logarithmic, sigmoid, and hyperbolic tangent).

The *penalty* term is used to look for a simpler ANN, with the smallest number of neurons and the fastest convergence for calculating the connection weights. However, the *penalty* term will not be used in our applications.

The minimum for the objective function $J(z)$ (Eq. 6.3) is computed by the Multi-Particle Collision Algorithm (MPCA) [LuBeVe08], based on Particle Collision Algorithm [SaOl86]. The MPCA was modified by [AnEtAl14] to find the best value for objective function 6.3.

## 6.3 Results

The data set used was obtained from the Institute of Advanced Studies (IEAv, Brazil) and covers an area from São José dos Campos (SP), Brazil. Images were acquired by air-transported Hyperspectral Scanner Sensor (HSS), with 37 bands from the electromagnetic spectrum into range $[0.44\ \mu m,\ 4\ \mu m]$. The image spatial resolution for the HSS sensor ranging between 2 and 9 meters [Ca03]. There is a ground truth of 4 regions of interest used to evaluate the land classification. Figure 6.5 shows a region on São José dos Campos area, with the 4 regions of interest and the respective ground truth.

An MLP-NN is used as an image classifier. For self-configuring the NN, 105 images were selected, where each region is represented by the average of the pixels in a $3 \times 3$ matrix for each band. The data was split into three sets: training set, validation set, and testing set – see Table 6.1.

Before the data reduction by PCA or NL-PCA, and classification by MLP-NN, the pixels (radiance) were normalized:

$$p_N = \frac{p - p_{\text{Min}}}{p_{\text{Max}} - p_{\text{Min}}} \tag{6.6}$$

where $p$ is the raw (pixel) data, $p_N$ is the normalized pixel value, and $p_{\text{Min}}$ and $p_{\text{Max}}$ are the lowest the largest pixel values found in the data set.

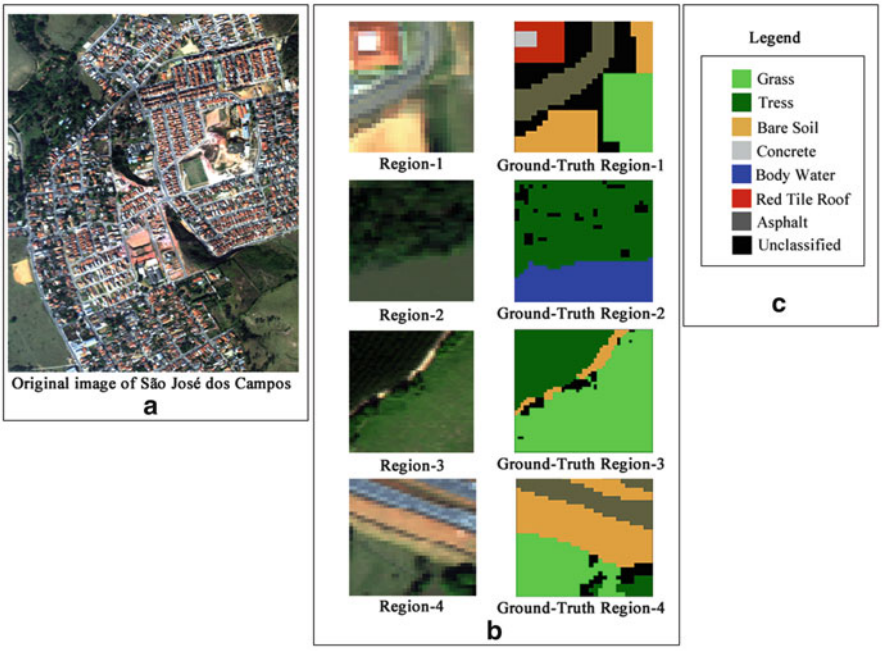


**Fig. 6.5** (A) Original image obtained by HSS sensor, (B1) left: 4 regions of interest, right: the ground-truth for 4 cited regions, and (C) the legend for the classes.

**Table 6.1** Data organization for training and testing the MLP-NN.

| | |
|---|---|
| Numbers of patterns to training | 55 |
| Numbers of patterns to validation | 15 |
| Numbers of patterns to test | 35 |

After the data normalization, the PCA method was applied. Only 6 principal components represent 99% the variability of the data. The MPCA was employed to find the optimal configuration of the MLP-NN to promote the data reduction by NL-PCA method. The MPCA was also used to determine the optimal architecture for the neural classifier.

Three ANNs were designed: (A) for performing the NL-PCA, (B) neural classifier with input data from standard PCA, and (C) neural classifier with input data from NL-PCA. The optimal configuration obtained with MPCA meta-heuristic is shown in Table 6.2.

To evaluate the data classification, the $\kappa$-index was used. The $\kappa$-index is a measure to quantify the deviation (classified data) from the exact values. The evaluation results for classification of $\kappa$-index average for 4 regions, overall accuracy, and average accuracy are shown in Table 6.3.

**Table 6.2** Optimal architectures of the MLP found by MPCA.

| | Configuration A | Configuration B | Configuration C |
|---|---|---|---|
| Input Layer | 37 | 6 | 7 |
| First Hidden Layer | 25 | 25 | 20 |
| Second Hidden Layer | 7 | — | — |
| Third Hidden Layer | 25 | — | — |
| Output Layer | 37 | 3 | 3 |
| Activation Function | Hyperbolic Tangent | Hyperbolic Tangent | Hyperbolic Tangent |
| Learning Rate | 0.05 | 0.25 | 0.22 |
| Momemtum | 0.9 | 0.83 | 0.87 |

**Table 6.3** Evaluation results obtained from classification by PCA of 4 regions using HSS sensor.

| | Total Accuracy | Average Accuracy | $\kappa$-Index |
|---|---|---|---|
| NL-PCA + Classification | 68.58% | 61.61% | 0.55 |
| PCA + Classification | 67.86% | 65.55% | 0.59 |

### *6.3.1 Execution of NLPCA in Hardware*

The data reduction by the NL-PCA method was also implemented on the hardware device Xilinx Virtex II Pro FPGA. The VHDL[1] was used to configure the FPGA. The Cray XD1 hybrid computer system has 6 interconnected processing nodes, where each node has 2 AMD processors (CPU) and one FPGA.

If an FPGA is configured as an artificial neural network, the device can be identified as a *neuro-computer*. The implementation of the MLP-NN on FPGA has four different modules: (a) the MAC (**M**ultiplier **A**nd **A**ccumulator): designed to do the product between inputs and weights (or bias); (b) artificial neuron: using the MAC and control structures; (c) combination of neurons: the inputs are connected by a single bus; (d)LUT (**L**ook**U**p **T**able) unit: the neurons can receive data, and the results (outputs) are flowing to the LUT unit – defined to address 524,288 values of activation function. Finally, the layers can be concatenated in series forming the MLP-NN. Figure 6.6 shows all implemented components of our neuro-computer.

The activation phase of the optimal architecture for both MLP-NNs, one used as NL-PCA operator and another one to perform the classification, was implemented in the FPGA. A comparison between the results produced by the implementation of software and hardware is performed, the 4 regions used to evaluate the classification in software were used to evaluate the classification in hardware. Table 6.4 shows the average of result of $\kappa$-index, overall accuracy, and average accuracy of 4 regions performed in FPGA.

[1] VHDL: **V**HSIC (Very High Speed Integrated Circuits) **H**ardware **D**escription **L**anguage.

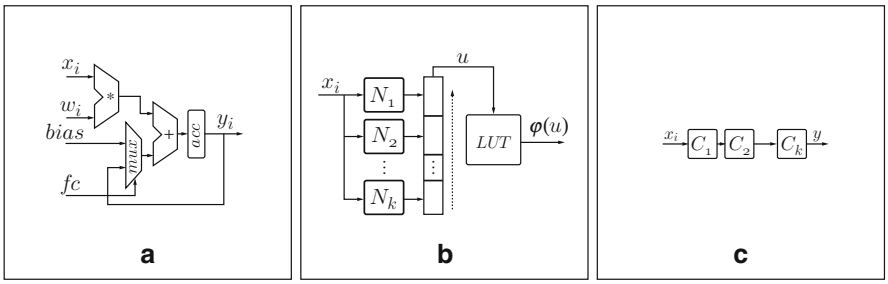


**Fig. 6.6** Implementation of MLP-NN on FPGA: (A) MAC unit, (B) pipeline for LUT unit, (C) the MLP implemented in FPGA.

**Table 6.4** Results obtained for classification by means of PCA with 4 HSS sensor regions.

| | Total Accuracy | Average Accuracy | $\kappa$-Index |
|---|---|---|---|
| NLPCA + Classification in Hardware | 68.58% | 61.61% | 0.55 |

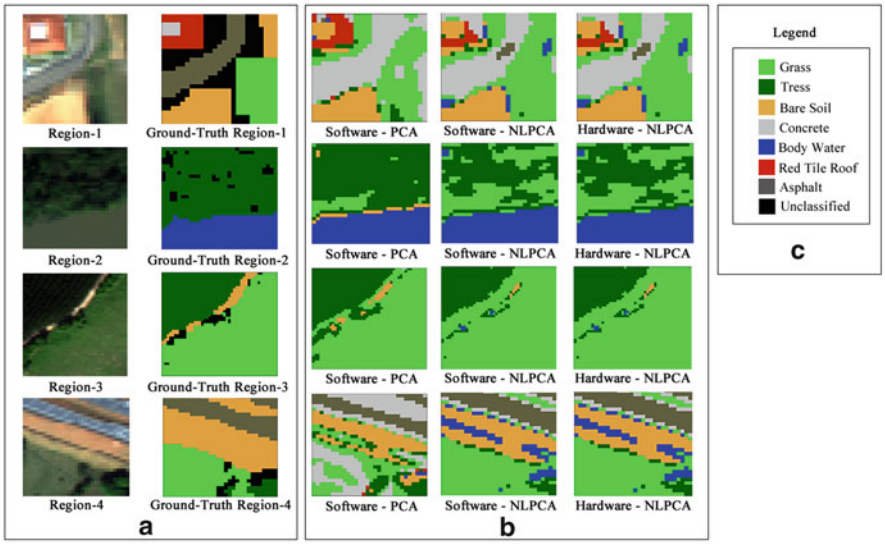


**Fig. 6.7** Neural classifier for 4 regions with data reduction: PCA and NL-PCA on software, and NL-PCA on FPGA.

Figure 6.7 displays the classification results by using data reduction considering two strategies: PCA + NN-classifier (software), and NL-PCA + NN-classifier (software), and NL-PCA + NN-classifier (hardware = FPGA). The results express a good performance of NL-PCA for data reduction. The MLP-NN implementation on FPGA produced very good results in comparison with the software implementation.

## 6.4 Conclusions

A case study was presented to evaluate the NL-PCA method, a nonlinear scheme for reducing data dimensionality. The NL-PCA utilizes a self-associative MLP-NN as the reduction operator. The standard procedure by using PCA was also used for comparison. The data reduction was employed to deal with image processing with data from multi- or hyper-spectral sensors. In our context, image processing means image classification. An artificial neural network was designed as an image classifier.

A self-configuring scheme, formulated as an optimization problem, was applied to define the best configuration for all ANNs employed. The optimization problem was solved by the MPCA meta-heuristic. Such procedure does not need an expert to define a workable neural network.

Finally, the ANN implemented on FPGA produced good results in comparison with software implementation. This is an important result, because the system can be embedded in aircrafts or satellites, allowing a HPC (High Performance Computing) environment working in parallel (data acquisition, pre-processing (data reduction), and image processing (image classification)) with low energy consumption.

## References

[AnEtAl14] Anochi, J. A., Velho, H. F. C., Furtado, H. C. M., Luz E. F. P.: Self-configuring two types neural networks by MPCA. 2nd International Symposium on Uncertainty Quantification and Stochastic Modeling (2014)

[An09] Andrecut, M.: Parallel GPU implementation of iterative PCA algorithms. Journal of Computational Biology, 1593–1599, 16 (2009)

[CaRaCh11] Carvalho, A., Ramos, F.M., and Chaves, A.C.: Metaheuristics for the feedforward artificial neural network (ANN) architecture optimization problem. Neural Computing and Applications, 1273–1284, 20 (2011)

[Ca03] Castro, A.P.A.: Edge Detection and Autonomous Navigation using Artificial Neural Networks. M.Sc. Thesis on Applied Computing, Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, SP, Brazil, 2003 (Portuguese).

[Ch13] Chein-I, C.: Hyperspectral Data Processing: algorithm and analysis. Wiley, Hoboken, NJ (2013)

[DeLiDu09] Del Frate, F., Licciardi, G., and Duca, R.: Autoassociative Neural Networks For Features Reduction of Hyperspectral Data. IEEE Geoscience and Remote Sensing Letters, 447–451, 9 (2009)

[GoWo00] Gonzalez, R.C. and Woods, R.E.: Digital Image Processing. Blucher, Saõ Paulo, SP (2000).

[Ha01] Haykin, S.: Artificial Neural Networks. Bookman, Porto Alegre, RS (2001)

[LiEtAl12] Licciardi, G., Reddy, P.M., Chanussot, J., and Benediktsson, J.A.: Linear Versus Nonlinear PCA for the Classification of Hyperspectral Data Based on the Extended Morphological Profiles. IEEE Geoscience and Remote Sensing Letters, 447–451, 9 (2012)

[LuBeVe08] Luz, E.F.P., Becceneri, J.C., Velho, H.F.C.: A new multi-particle collision algorithm for otimization in a high-performance environment. Journal of Computational Interdisciplinary Sciences, 1–7, 1 (2008)

[RuHiWi86] Rumelhart, D.E., Hinton, G.E., and Willians, R.J.: Learning representations by back-propagating errors. Nature, 533–536, 323 (1986)

[SaOl86] Sacco, W. F. and Oliveira, C.R.E.A.: A new stochastic optimization algorithm based on a particle collision metaheuristic. 6th World Congress of Structural and Multidisciplinary Optimization, Rio de Janeiro (1986)

[SiEtAl13] Silva, W. and Habermann, M. and Shiguemori, E. H. and Andrade, L. L. and Castro, R. M.: Multispectral Image Classification using Multilayer Perceptron and Principal Components Analysis. 1st BRICS Countries Congress on Computational Intelligence, Porto de Galinhas (2013)

# Chapter 7
# On the Eigenvalues of a Biharmonic Steklov Problem

**D. Buoso and L. Provenzano**

## 7.1 Introduction

Let $\Omega$ be a bounded domain (i.e., a bounded connected open set) of class $C^2$ in $\mathbb{R}^N$, $N \geq 2$ and $\tau > 0$. We consider the following Steklov eigenvalue problem for the biharmonic operator

$$\begin{cases} \Delta^2 u - \tau \Delta u = 0, & \text{in } \Omega, \\ \frac{\partial^2 u}{\partial \nu^2} = 0, & \text{on } \partial\Omega, \\ \tau \frac{\partial u}{\partial \nu} - \operatorname{div}_{\partial\Omega}\left(D^2 u.\nu\right) - \frac{\partial \Delta u}{\partial \nu} = \lambda u, & \text{on } \partial\Omega, \end{cases} \tag{7.1}$$

in the unknowns $\lambda$ (the eigenvalue) and $u$ (the eigenfunction). Here $\nu$ denotes the unit outer normal to $\partial\Omega$, $\operatorname{div}_{\partial\Omega}$ the tangential divergence operator, and $D^2 u$ the Hessian matrix of $u$. The spectrum consists of a diverging sequence of eigenvalues of finite multiplicity

$$0 = \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_j \leq \cdots,$$

where the eigenvalues are repeated according to their multiplicity.

When $N = 2$, problem (7.1) arises in the study of the vibration modes of a free elastic plate subject to lateral tension (represented by the parameter $\tau$) whose total mass is concentrated at the boundary. We can describe this concentration phenomenon as follows.

D. Buoso
Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
e-mail: davide.buoso@polito.it

L. Provenzano (✉)
University of Padova, Via Trieste 63, 35121 Padova, Italy
e-mail: proz@math.unipd.it



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_7

For any $\varepsilon$ sufficiently small we consider the $\varepsilon$-neighborhood of $\partial\Omega$, namely $\omega_\varepsilon = \{x \in \Omega : 0 < d(x,\partial\Omega) < \varepsilon\}$. We fix $M > 0$ and define the function $\rho_\varepsilon$ on $\Omega$ as follows:

$$\rho_\varepsilon = \begin{cases} \varepsilon, & \text{in } \Omega \setminus \overline{\omega}_\varepsilon\,, \\ \frac{M-\varepsilon|\Omega\setminus\overline{\omega}_\varepsilon|}{|\omega_\varepsilon|}, & \text{in } \omega_\varepsilon\,. \end{cases}$$

For any $x \in \Omega$ we have $\rho_\varepsilon(x) \to 0$ as $\varepsilon \to 0$. Moreover, $\int_\Omega \rho_\varepsilon = M$ for all $\varepsilon > 0$. Then we consider the following eigenvalue problem for the biharmonic operator subject to Neumann boundary conditions

$$\begin{cases} \Delta^2 u - \tau\Delta u = \lambda(\varepsilon)\rho_\varepsilon u, & \text{in } \Omega\,, \\ \frac{\partial^2 u}{\partial \nu^2} = 0, & \text{on } \partial\Omega\,, \\ \tau\frac{\partial u}{\partial \nu} - \operatorname{div}_{\partial\Omega}\left(D^2 u.\nu\right) - \frac{\partial \Delta u}{\partial \nu} = 0, & \text{on } \partial\Omega\,. \end{cases} \tag{7.2}$$

The spectrum consists of a diverging sequence of eigenvalues of finite multiplicity

$$0 = \lambda_1(\varepsilon) < \lambda_2(\varepsilon) \le \cdots \le \lambda_j(\varepsilon) \le \cdots,$$

where the eigenvalues are repeated according to their multiplicity. Here we emphasize the dependence of the eigenvalues on the parameter $\varepsilon$.

We remark that for $N = 2$ problem (7.2) provides the fundamental modes of vibration of a free elastic plate with mass density $\rho_\varepsilon$ and total mass $M$, as discussed in [Ch11, Chasman]. We refer to [Ch11] for the derivation and the physical interpretation of problem (7.2).

It is possible to prove that the eigenvalues and the eigenfunctions of (7.2) converge to the eigenvalues and eigenfunctions of (7.1) as $\varepsilon$ goes to zero (see, e.g., [ArJiRo08, BuPr14, LaPr14]).

The aim of this paper is to study a few properties concerning the dependence of the eigenvalues of (7.1) upon perturbations of the domain $\Omega$ which preserve the measure.

First, we study the asymptotic behavior of the eigenvalues of (7.2) as $\varepsilon \to 0$ providing an interpretation of (7.1) as the model of a free vibrating plate with all the mass concentrated at the boundary (see Theorem 1). This fact suggests that (7.1) is the natural fourth order generalization of the classical Steklov eigenvalue problem for the Laplace operator, see [St02] and the recent [La14] for related problems.

Second, we consider the problem of the optimal shape of $\Omega$ for the eigenvalues of (7.1) under the constraint that the measure of $\Omega$ is fixed. This problem has been largely investigated for the Laplace operator subject to different homogeneous boundary conditions. We refer to [He06, Henrot] for a collection of results on the subject. See also [Ba80, Bandle]. As far as the biharmonic operator is concerned, only a few results exist in literature. It has been proved in [Na95, Nadirashvili] for $N = 2$ and soon generalized in [AsBe95, Ashbaugh, Benguria] for $N = 3$ that the ball is a minimizer for the first eigenvalue of the biharmonic operator subject to Dirichlet boundary conditions. In the recent paper [Ch11], it has been proved that the first positive eigenvalue of problem (7.2) with constant mass density $\rho \equiv 1$ is maximized by the ball among those sets with a fixed measure.

As for Steklov boundary conditions, we refer to [BuFe09, Bucur, Ferrero, Gazzola] and the references therein. The authors consider the following eigenvalue problem

$$\begin{cases} \Delta^2 u = 0, & \text{in } \Omega\,, \\ u = 0, & \text{on } \partial\Omega\,, \\ \Delta u = \lambda \frac{\partial u}{\partial \nu}, & \text{on } \partial\Omega\,. \end{cases} \tag{7.3}$$

Problem (7.3) should not be confused with problem (7.1) and reveals a rather different nature. (We note that one may refer to Steklov-type boundary conditions for those problems where the spectral parameter enters the boundary conditions.)

By following the approach developed in [BuLa13, BuLa14] we prove that simple eigenvalues and the symmetric functions of multiple eigenvalues of (7.1) depend real analytically upon transformations of the domain $\Omega$ (see Theorem 2) and we characterize those critical transformations which preserve the measure (see Corollary 1). See also [LaLa04, LaLa06, LaLa07]. Then we show that the ball is a critical point for all simple eigenvalues and all symmetric functions of the eigenvalues under measure constraint in the sense of Theorem 3.

Finally, we prove the following isoperimetric inequality: "*The ball is a maximizer for the first positive eigenvalue of problem (7.1) among those bounded domains with a fixed measure*" (see Theorem 4). To do so, we follow the approach of [Ch11] and in particular we study problem (7.1) when $\Omega$ is the unit ball in $\mathbb{R}^N$, identifying the first positive eigenvalue and the corresponding eigenfunctions.

Detailed proofs of the results announced in this paper can be found in [BuPr14].

## 7.2 Asymptotic Behavior of Neumann Eigenvalues

Let $\Omega$ be a bounded domain in $\mathbb{R}^N$ of class $C^2$. Let $\lambda_j$ and $\lambda_j(\varepsilon)$, $j \in \mathbb{N}\setminus\{0\}$, be the eigenvalues of (7.1) and (7.2), respectively. For the sake of simplicity and without any loss of generality, we assume that $M = |\partial\Omega|$. We recall that $\lambda_1 = \lambda_1(\varepsilon) = 0$, while $\lambda_2, \lambda_2(\varepsilon) > 0$ for all $\varepsilon > 0$.

We have the following result concerning the spectral convergence of problem (7.2) to problem (7.1).

**Theorem 1.** *Let $\Omega$ be a bounded domain in $\mathbb{R}^N$ of class $C^2$. Then $\lambda_j(\varepsilon) \to \lambda_j$ for all $j \in \mathbb{N}\setminus\{0\}$. Moreover the projections on the eigenspaces associated with the eigenvalues converge in norm.*

This theorem can be proved by using the notion of compact convergence for the resolvent operators which implies, in the case of self-adjoint operators, convergence in norm. It is well known that if a family of self-adjoint operators $A_\varepsilon$ converges in norm to a self-adjoint operator $A$, then isolated eigenvalues of $A$ are exactly the limits of eigenvalues of $A_\varepsilon$ counting multiplicity. Moreover, eigenprojections converge in norm. We refer to [BuPr14, LaPr14] for more details. We also refer to the recent paper [ArLa13] for a general approach to the shape sensitivity analysis of higher-order operators.

Theorem 1 justifies our interpretation of problem (7.1) as the equations of a free vibrating plate whose mass is concentrated at the boundary. However, we can also directly obtain (7.1) by deriving the equations of motions of a free vibrating plate with constant surface density. To do so, we follow the approach of [We52, ch.10-8] in the case $N = 2$. We represent the displacement at rest of the plate by means of a domain $\Omega \subset \mathbb{R}^2$ and we describe the vertical deviation from the equilibrium during the vibration of each point $(x,y) \in \Omega$ at time $t$ by means of a function $v(x,y,t) \in C^2(\Omega \times [t_1,t_2])$. Then we write the Hamilton's integral $\mathscr{H}$ of the system

$$\mathscr{H} = \frac{1}{2}\int_{t_1}^{t_2}\int_{\partial\Omega} \dot{v}^2 d\sigma dt - \frac{1}{2}\int_{t_1}^{t_2}\int_{\Omega} \left(v_{xx}^2 + v_{yy}^2 + 2v_{xy}^2\right) + \tau\left(v_x^2 + v_y^2\right) dxdydt. \tag{7.4}$$

According to Hamilton's Variational Principle, the actual motion of the system minimizes such integral. Let $v \in C^2(\Omega \times [t_1,t_2])$ be a minimizer for $\mathscr{H}$. Then by differentiating (7.4) it follows that $v$ satisfies

$$\begin{aligned} &-\int_{t_1}^{t_2}\int_{\partial\Omega} \eta \ddot{v} d\sigma dt - \int_{t_1}^{t_2}\int_{\Omega} \eta\left(\Delta^2 v - \tau\Delta v\right) dxdydt \\ &-\int_{t_1}^{t_2}\int_{\partial\Omega} \frac{\partial\eta}{\partial\nu}\frac{\partial^2 v}{\partial\nu^2} - \eta\left(\tau\frac{\partial v}{\partial\nu} - \mathrm{div}_{\partial\Omega}\left(D^2 v.\nu\right)_{\partial\Omega} - \frac{\partial\Delta v}{\partial\nu}\right) d\sigma dt = 0, \end{aligned}$$

for all $\eta \in C^2(\Omega \times [t_1,t_2])$. We refer [Ch11] for the details. By the arbitrary choice of $\eta$ we obtain

$$\begin{cases} \Delta^2 v - \tau\Delta v = 0, & \text{in } \Omega, \\ \frac{\partial^2 v}{\partial\nu^2} = 0, & \text{on } \partial\Omega, \\ \ddot{v} + \tau\frac{\partial v}{\partial\nu} - \mathrm{div}_{\partial\Omega}\left(D^2 v.\nu\right) - \frac{\partial\Delta v}{\partial\nu} = 0, & \text{on } \partial\Omega, \end{cases}$$

for all $t \in \mathbb{R}$. As is customary, by looking for solution of the form $v(x,y,t) = u(x,y)\psi(t)$. We find that the temporal component $\psi(t)$ solves the ordinary differential equation $-\ddot{\psi} = \lambda\psi$ for all $t \in [t_1,t_2]$, while the spatial component $u$ solves problem (7.1).

## 7.3 Isovolumetric Perturbations

Given a bounded domain in $\mathbb{R}^N$ of class $C^2$, we set

$$\Phi(\Omega) = \left\{\phi \in \left(C^2\left(\overline{\Omega}\right)\right)^N : \phi \text{ injective and } \inf_{\Omega}|\det D\phi| > 0\right\}.$$

We observe that if $\Omega$ is of class $C^2$ and $\phi \in \Phi(\Omega)$, it makes sense to study problem (7.1) on $\phi(\Omega)$. For any $\phi \in \Phi(\Omega)$ we denote by $\lambda_j(\phi), j \in \mathbb{N} \setminus \{0\}$, the eigenvalues of (7.1) on $\phi(\Omega)$.

We plan to study the dependence of the eigenvalues upon the function $\phi$. In general, one cannot expect differentiability of the eigenvalues with respect to $\phi$. This is due, for example, to well-known bifurcation phenomena that occur when multiple eigenvalues split from a simple eigenvalue. However, as is pointed out in [BuLa13, BuLa14], in the case of multiple eigenvalues it is possible to prove analyticity for the symmetric functions of the eigenvalues. Namely, given a finite set of indexes $F \subset \mathbb{N} \setminus \{0\}$, one can consider the symmetric functions of the eigenvalues with indexes in $F$

$$\Lambda_{F,s}(\phi) = \sum_{j_1 < \cdots < j_s \in F} \lambda_{j_1}(\phi) \cdots \lambda_{j_s}(\phi),$$

and prove that such functions are real analytic on the set

$$\mathscr{A}_\Omega[F] = \left\{ \phi \in \Phi(\Omega) : \lambda_l(\phi) \notin \left\{ \lambda_j(\phi) : j \in F \right\} \forall l \in \mathbb{N} \setminus (F \cup \{0\}) \right\}.$$

Then it is possible to find formulas for the Fréchet derivatives of the symmetric functions of the eigenvalues. It is convenient to set

$$\Theta_\Omega[F] = \left\{ \phi \in \mathscr{A}_\Omega[F] : \lambda_{j_1}(\phi) = \lambda_{j_2}(\phi), \forall j_1, j_2 \in F \right\}.$$

**Theorem 2.** *Let $\Omega$ be a bounded domain of $\mathbb{R}^N$ of class $C^2$. Let $F$ be a finite nonempty subset of $\mathbb{N} \setminus \{0\}$. Then $\mathscr{A}_\Omega$ is open in $\Phi(\Omega)$ and $\Lambda_{F,s}$ are real analytic in $\mathscr{A}_\Omega$. Moreover, let $\tilde{\phi} \in \Theta_\Omega[F]$ be such that $\partial\tilde{\phi}(\Omega) \in C^4$. Let $v_1, \ldots, v_{|F|}$ be a orthonormal basis of the eigenspace associated with the eigenvalue $\lambda_F(\tilde{\phi})$. Then*

$$d|_{\phi=\tilde{\phi}}(\Lambda_{F,s})(\psi) = -\lambda_F^s(\tilde{\phi}) \binom{|F|-1}{s-1} \sum_{l=1}^{|F|} \int_{\partial\tilde{\phi}(\Omega)} \Big( \lambda_F K v_l^2 + \lambda_F \frac{\partial(v_l^2)}{\partial \nu} - \tau |\nabla v_l|^2 - |D^2 v_l|^2 \Big) \mu \cdot \nu d\sigma, \tag{7.5}$$

*for all $\psi \in (C^2(\Omega))^N$, where $\mu = \psi \circ \phi^{(-1)}$, and $K$ denotes the mean curvature on $\partial\tilde{\phi}(\Omega)$.*

The proof follows the lines of the corresponding results provided in [BuLa13] and [BuLa14] for general polyharmonic operators subject to Dirichlet boundary conditions and for the biharmonic operator subject to hinged boundary conditions.

We consider now the problem of finding critical points $\phi \in \Phi(\Omega)$ for the symmetric functions of the eigenvalues under the condition that $\phi$ preserves the measure. We set $\mathscr{V}(\phi) = \int_{\phi(\Omega)} dy = \int_\Omega |\mathrm{detD}\phi| dx$. We fix $\mathscr{V}_0 \in ]0, +\infty[$ and consider

the set $V(\mathscr{V}_0) = \{\phi \in \Phi(\Omega) : \mathscr{V}(\phi) = \mathscr{V}_0\}$. Given $\Omega$ such that $|\Omega| = \mathscr{V}_0$, $V(\mathscr{V}_0)$ is the subset of $\Phi(\Omega)$ of those functions $\phi$ preserving the measure. By formula (7.5) and by the Lagrange Multipliers Theorem we can characterize the critical points.

**Corollary 1.** *Let all the assumptions of Theorem 2 hold. Then $\tilde{\phi} \in \Phi(\Omega)$ is a critical point for $\Lambda_{F,s}$ on $V(\mathscr{V}_0)$ if and only if there exists a constant $c \in \mathbb{R}$ such that*

$$\sum_{l=1}^{|F|} \left( \lambda_F(\tilde{\phi}) \left( K v_l^2 + \frac{\partial v_l^2}{\partial \nu} \right) - \tau |\nabla v_l|^2 - |D^2 v_l|^2 \right) = c, \text{ a.e. on } \partial\tilde{\phi}(\Omega),$$

*where $K$ denotes the mean curvature on $\partial\tilde{\phi}(\Omega)$.*

Thanks to Corollary 1 we can prove that balls are critical points for the symmetric functions of the eigenvalues under measure constraint, in the sense of the following

**Theorem 3.** *Let $\tilde{\phi} \in \Phi(\Omega)$ be such that $\tilde{\phi}(\Omega)$ is a ball. Let $\tilde{\lambda}$ be an eigenvalue of the problem in $\tilde{\phi}(\Omega)$, and let $F$ be the set of all $j \in \mathbb{N} \setminus \{0\}$ such that $\lambda_j(\tilde{\phi}) = \tilde{\lambda}$. Then $\Lambda_{F,s}$ has a critical point at $\tilde{\phi}$ on $V(\mathscr{V}_0)$, for all $s = 1, \dots, |F|$.*

The proof can be carried out as in [BuLa13, BuLa14]. Namely, given $\lambda$ an eigenvalue of problem (7.1) on the unit ball $B$ in $\mathbb{R}^N$, consider the subset $F$ of $\mathbb{N} \setminus \{0\}$ of those indexes $j$ such that the $j$-th eigenvalue of problem (7.1) in $B$ coincides with $\lambda$. Consider then $v_1, \dots, v_{|F|}$ an orthonormal basis of the eigenspace associated with the eigenvalue $\lambda$, where the orthonormality is taken with respect to the scalar product in $L^2(\partial B)$. Then it is possible to show that the quantities $\sum_{j=1}^{|F|} v_j^2$, $\sum_{j=1}^{|F|} |\nabla v_j|^2$ and $\sum_{j=1}^{|F|} |D^2 v_j|^2$ are radial functions. This fact and the fact that the mean curvature $K$ is constant on the ball allow to conclude.

## 7.4 The Isoperimetric Inequality

Let us consider problem (7.1) when $\Omega = B$ is the unit ball in $\mathbb{R}^N$. It is convenient to use spherical coordinates $(r, \theta)$ in $\mathbb{R}^N$, where $\theta = (\theta_1, \dots, \theta_{N-1})$, with $r \in [0, 1[$, $\theta_1, \dots, \theta_{N-2} \in [0, \pi]$, $\theta_{N-1} \in [0, 2\pi]$. In this case the boundary conditions can be written in the following form

$$\begin{cases} \frac{\partial^2 u}{\partial r^2}_{|_{r=1}} = 0, \\ \tau \frac{\partial u}{\partial r} - \frac{1}{r^2} \Delta_S \left( \frac{\partial u}{\partial r} - \frac{u}{r} \right) - \frac{\partial \Delta u}{\partial r}_{|_{r=1}} = \lambda u_{|_{r=1}}, \end{cases} \tag{7.6}$$

where $\Delta_S$ is the angular part of the Laplacian (see [Ch11] for details). Then, the eigenfunctions of problem (7.1) on the ball can be described explicitly as in the following lemma. We refer to [AbSt64, ch.9] for well-known definitions and properties of Bessel functions.

**Lemma 1.** *Let $B$ be the unit ball in $\mathbb{R}^N$. An eigenfunction $u$ of (7.1) is of the form $u(r,\theta) = R_l(r)Y_l(\theta)$, where $Y_l(\theta)$ is a spherical harmonic of some order $l \in \mathbb{N}$,*

$$R_l(r) = A_l r^l + B_l i_l(\sqrt{\tau}r) \tag{7.7}$$

*and $A_l$ and $B_l$ are suitable constants such that*

$$B_l = \frac{l(1-l)}{\tau i_l''(\sqrt{\tau})} A_l. \tag{7.8}$$

*Here $i_l$ denotes the ultraspherical modified Bessel function of the first kind, which is defined by*

$$i_l(z) = z^{1-\frac{N}{2}} I_{\frac{N}{2}-1+l}(z),$$

*where $I_l(z)$ denotes the modified Bessel function of the first kind.*

We note that equality (7.8) is obtained by imposing the boundary conditions (7.6) to the function (7.7).

Lemma 1 allows to find explicit formulas for the eigenvalues. In the sequel we will denote by $\lambda_{(l)}$ the eigenvalue corresponding to the eigenfunction $u_l$ defined in Lemma 1.

**Lemma 2.** *The eigenvalues $\lambda_{(l)}$ of problem (7.1) on $B$ are delivered by the formula*

$$\begin{aligned}\lambda_{(l)} = l\Big((1-l)li_l(\sqrt{\tau}) + \tau i_l''(\sqrt{\tau})\Big)^{-1} \Big[&3(l-1)l(l+N-2)i_l(\sqrt{\tau}) \\ &-(l-1)\sqrt{\tau}\big(N-1+2Nl+2l(l-2)l+\tau\big)i_l'(\sqrt{\tau}) \\ &+\tau\big((l-1)(l+2N-3)+\tau\big)i_l''(\sqrt{\tau}) \\ &+(l-1)\tau\sqrt{\tau}i_l'''(\sqrt{\tau})\Big],\end{aligned}$$

*with $l \in \mathbb{N}$.*

Now we need to identify the index $l$ satisfying $\lambda_{(l)} = \lambda_2$, that is the first positive eigenvalue of (7.1). This is done by means of the following

**Lemma 3.** *The first positive eigenvalue of problem (7.1) on $B$ is $\lambda_2 = \lambda_{(1)} = \tau$. The corresponding eigenspace is generated by the coordinate functions $\{x_1, \ldots, x_N\}$.*

The proof of Lemma 3 consists in two steps. In the first step we observe that $0 = \lambda_{(0)} < \lambda_{(1)} = \tau$. Moreover, by using well-known recurrence relations for modified ultraspherical Bessel functions of the first kind and their derivatives we are able to prove that $\lambda_{(1)} < \lambda_{(2)}$. In the second step we show that for any smooth radial function $R(r)$, the Rayleigh quotient

$$\mathscr{Q}(R(r)Y_l(\theta)) = \frac{\int_\Omega |D^2(R(r)Y_l(\theta))|^2 + \tau|\nabla(R(r)Y_l(\theta))|^2 dx}{\int_{\partial\Omega} R(r)^2 Y_l(\theta)^2 d\sigma}$$

is an increasing function of $l$ for $l \geq 2$. This, combined with the variational characterization of the eigenvalues, allows us to conclude that $\lambda_{(l)}$ is an increasing function of $l$ for $l \geq 2$.

We are ready to state the isoperimetric inequality.

**Theorem 4.** *Among all bounded domains of class $C^2$ with fixed measure, the ball maximizes the first nonnegative eigenvalue, that is $\lambda_2(\Omega) \leq \lambda_2(\Omega^*)$, where $\Omega^*$ is a ball with the same measure as $\Omega$.*

The proof can be carried out as in [He06, par.7.3]. Namely, we use the following variational characterization of the sum of inverse of eigenvalues.

$$\sum_{l=2}^{N+1} \frac{1}{\lambda_l(\Omega)} = \max\left\{ \sum_{l=2}^{N+1} \int_{\partial\Omega} v_l^2 d\sigma \right\}, \tag{7.9}$$

where $\{v_l\}_{l=2}^{N+1}$ is a family in $H^2(\Omega)$ satisfying $\int_\Omega D^2 v_i : D^2 v_j + \tau \nabla v_i \cdot \nabla v_j dx = \delta_{ij}$ and $\int_{\partial\Omega} v_l d\sigma = 0$ for all $l = 2, ..., N+1$. We plug the functions $v_l = (\tau|\Omega|)^{-\frac{1}{2}} x_l$, with $l = 1, ..., N$, into (7.9) and we use the inequality

$$\int_{\partial\Omega} f(|x|) d\sigma \geq \int_{\partial\Omega^*} f(|x|) d\sigma, \tag{7.10}$$

where $\Omega^*$ is the ball with the same measure of $\Omega$ and $f$ is a continuous, non-negative, non-decreasing function defined on $[0, +\infty)$ and moreover is such that the map $t \mapsto \big(f(t^{1/N}) - f(0)\big) t^{1-(1/N)}$ is convex. Then the isoperimetric inequality easily follows. We refer to [HiXu93] for the proof of (7.9) and to [BeBr99] for the proof of (7.10).

**Acknowledgements** The authors are deeply thankful to Prof. Pier Domenico Lamberti who suggested the problem, and also for many useful discussions. The authors acknowledge financial support from the research project 'Singular perturbation problems for differential operators,' Progetto di Ateneo of the University of Padova. The authors are members of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

## References

[AbSt64] Abramowitz, M., Stegun, I.A.: *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* National Bureau of Standards Applied Mathematics Series, U.S. Government Printing Office, Washington, D.C., (1964)

[ArJiRo08] Arrieta, J.M., Jiménez-Casas, A., Rodríguez-Bernal, A.: *Flux terms and Robin boundary conditions as limit of reactions and potentials concentrating in the boundary.* Rev. Mat. Iberoam. **24**, no. 1, 183–211 (2008)

[ArLa13] Arrieta, J.M., Lamberti, P.D.: *Spectral stability results for higher-order operators under perturbations of the domain.* C. R. Math. Acad. Sci. Paris **351**, no. 19–20, pp 725–730 (2013)

[AsBe95] Ashbaugh, M.S., Benguria, R.D.: *On Rayleigh's conjecture for the clamped plate and its generalization to three dimensions.* Duke Math. J. **78**, no. 1, 17 (1995)

[Ba80] Bandle, C.: *Isoperimetric inequalities and applications.* Pitman advanced publishing program, monographs and studies in mathematics, **7** (1980)

[BeBr99] Betta, F., Brock, F., Mercaldo, A., Posteraro, M.R.: *A weighted isoperimetric inequality and applications to symmetrization.* J. Inequal. Appl., **4**, no. 3, pp 215–240 (1999)

[BuFe09] Bucur, D., Ferrero, A., Gazzola, F.: *On the first eigenvalue of a fourth order Steklov problem.* Calculus of Variations and Partial Differential Equations, **35**, pp 103–131 (2009)

[BuLa13] Buoso, D., Lamberti, P.D.: *Eigenvalues of polyharmonic operators on variable domains.* ESAIM: COCV, **19**, pp 1225–1235 (2013)

[BuLa14] Buoso, D., Lamberti, P.D.: *Shape deformation for vibrating hinged plates.* Mathematical Methods in the Applied Sciences, **37**, pp 237–244 (2014)

[BuPr14] Buoso, D., Provenzano, L.: *A few shape optimization results for a Biharmonic Steklov problem.* J. Differential Equations, (2015). http://dx.doi.org/10.1016/j.jde.2015.03.013.

[Ch11] Chasman, L.M.: *An isoperimetric inequality for fundamental tones of free plates.* Comm. Math. Phys., **303**, no. 2, pp 421–449 (2011)

[He06] Henrot, A.: *Extremum problems for eigenvalues of elliptic operators.* Frontiers in Mathematics, Birkhäuser Verlag, Basel, (2006)

[HiXu93] Hile, G.N., Xu, Z.Y.: *Inequalities for sums of reciprocals of eigenvalues.* J. Math. Anal. Appl., **180** no. 2, pp 412–430 (1993)

[La14] Lamberti, P.D.: *Steklov-type eigenvalues associated with best Sobolev trace constants: domain perturbation and overdetermined systems.* Complex Var. Elliptic Equ. **59** no. 3, pp 309–323 (2014)

[LaLa04] Lamberti, P.D., Lanza de Cristoforis, M.: *A real analyticity result for symmetric functions of the eigenvalues of a domain dependent Dirichlet problem for the Laplace operator.* J. Nonlinear Convex Anal, **5**, no. 1, pp 19–42 (2004)

[LaLa06] Lamberti, P.D., Lanza de Cristoforis, M.: *Critical points of the symmetric functions of the eigenvalues of the Laplace operator and overdetermined problems.* J. Math. Soc. Japan, **58**, no. 1, pp 231–245 (2006)

[LaLa07] Lamberti, P.D., Lanza de Cristoforis, M.: *A real analyticity result for symmetric functions of the eigenvalues of a domain-dependent Neumann problem for the Laplace operator.* Mediterr. J. Math., **4**, no. 4, pp 435–449 (2007)

[LaPr14] Lamberti, P.D., Provenzano, L.: *Viewing the Steklov eigenvalues of the Laplace operator as critical Neumann eigenvalues.* Current Trends in Analysis and its Applications, Proceedings of the 9th ISAAC Congress, Kraków 2013 (2015)

[Na95] Nadirashvili, N.S.: *Rayleigh's conjecture on the principal frequency of the clamped plate.* Arch. Rational Mech. Anal., **129**, no. 1, pp 1–10 (1995)

[St02] Stekloff, W.: *Sur les problémes fondamentaux de la physique mathèmatique (suite et fin).* Ann. Sci. École Norm. Sup., **3**, 19, pp 455–490 (1902)

[We52] Weinstock, R.: *Calculus of variations with applications to physics and engineering.* McGraw-Hill Book Company Inc., New York-Toronto-London (1952)

# Chapter 8
# Shape Differentiability of the Eigenvalues of Elliptic Systems

**D. Buoso**

Let $\Omega$ be a bounded open set in $\mathbb{R}^N$ of class $C^1$, $m \in \mathbb{N}$. By $H^1(\Omega)$ we denote the Sobolev space of functions in $L^2(\Omega)$ with derivatives in $L^2(\Omega)$, and by $H_0^1(\Omega)$ we denote the closure in $H^1(\Omega)$ of the space of $C^\infty$-functions with compact support in $\Omega$.

We consider the following eigenvalue problem in the weak form

$$\int_\Omega \sum_{\alpha,\beta=1}^{N} \sum_{i,j=1}^{m} a_{\alpha\beta}^{ij} \frac{\partial u_i}{\partial x_\alpha} \frac{\partial \varphi_j}{\partial x_\beta} dx = \lambda \int_\Omega u \cdot \varphi dx, \tag{8.1}$$

for any $\varphi \in V(\Omega)^m$, in the unknowns $u \in V(\Omega)^m$ (the eigenfunction), $\lambda \in \mathbb{R}$ (the eigenvalue), where $V(\Omega)$ denotes either $H_0^1(\Omega)$ (for Dirichlet boundary conditions) or $H^1(\Omega)$ (for Neumann boundary conditions).

Note that the classical formulation of the Dirichlet problem reads

$$\begin{cases} -\sum_{\alpha,\beta=1}^{N} \sum_{i=1}^{m} a_{\alpha\beta}^{ij} \frac{\partial^2 u_i}{\partial x_\alpha \partial x_\beta} = \lambda u_j, j = 1,\dots,m, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{8.2}$$

while the classical formulation of the Neumann problem reads

$$\begin{cases} -\sum_{\alpha,\beta=1}^{N} \sum_{i=1}^{m} a_{\alpha\beta}^{ij} \frac{\partial^2 u_i}{\partial x_\alpha \partial x_\beta} = \lambda u_j, j = 1,\dots,m, & \text{in } \Omega, \\ \sum_{\alpha,\beta=1}^{N} \sum_{i=1}^{m} a_{\alpha\beta}^{ij} \nu_\beta \frac{\partial u_i}{\partial x_\alpha} = 0, j = 1,\dots,m, & \text{on } \partial\Omega, \end{cases} \tag{8.3}$$

where $\nu$ denotes the outer unit normal to $\partial\Omega$.

D. Buoso (✉)
Politecnico di Torino, corso Duca degli Abruzzi 24, 10129 Torino, Italy
e-mail: davide.buoso@polito.it



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_8

Here and in the sequel, $a^{ij}_{\alpha\beta} \in \mathbb{R}$ are constant coefficients satisfying $a^{ij}_{\alpha\beta} = a^{ji}_{\beta\alpha}$ and the Legendre–Hadamard condition, i.e.,

$$\sum_{\alpha,\beta=1}^{N} \sum_{i,j=1}^{m} a^{ij}_{\alpha\beta} \xi_i \xi_j \eta_\alpha \eta_\beta \geq \theta |\xi|^2 |\eta|^2, \quad \forall \xi \in \mathbb{R}^m, \forall \eta \in \mathbb{R}^N, \tag{8.4}$$

for some $\theta > 0$.

We consider in $H^1(\Omega)^m$ the bilinear form

$$< u, v >= \int_\Omega \sum_{\alpha,\beta=1}^{N} \sum_{i,j=1}^{m} a^{ij}_{\alpha\beta} \frac{\partial u_i}{\partial x_\alpha} \frac{\partial v_j}{\partial x_\beta} dx, \tag{8.5}$$

for any $u, v \in H^1(\Omega)^m$. Note that, for instance, it is possible to prove that the bilinear form (8.5) defines on $H^1_0(\Omega)^m$ a scalar product whose induced norm is equivalent to the standard one.

Note that problem (8.1) includes some important problems in linear elasticity. For instance, the choice $a^{ij}_{\alpha\beta} = \delta_{ij}\delta_{\alpha\beta} + \mu\delta_{i\alpha}\delta_{j\beta}$, where $\delta_{ij}$ is the Kronecker delta and $\mu \geq 0$ a constant, leads to the Lamé eigenvalue problem

$$\begin{cases} -\Delta u - \mu \nabla \mathrm{div} u = \lambda u, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega. \end{cases} \tag{8.6}$$

Problem (8.6) is very similar to the Reissner–Mindlin system

$$\begin{cases} -\frac{\mu}{12}\Delta\beta - \frac{\mu+\lambda}{12}\nabla \mathrm{div}\beta - \frac{\mu k}{t^2}(\nabla w - \beta) = \frac{\gamma t^2}{12}\beta, & \text{in } \Omega, \\ -\frac{\mu k}{t^2}(\Delta w - \mathrm{div}\beta) = \gamma w, & \text{in } \Omega, \\ \beta = 0, \ \ w = 0, & \text{on } \partial\Omega, \end{cases} \tag{8.7}$$

which arises in the study of the vibrations of a clamped plate. Here $\mu, \lambda, k$, and $t$ are physical constants, $\gamma$ is the eigenvalue, and $(\beta, w)$ is the eigenvector. Note that the current discussion does not comprehend problem (8.7), since it presents lower-order terms. However, the arguments we use can be easily adapted in order to treat problem (8.7) as well (see [BuLa]).

Thanks to condition (8.4), it is possible to show that the eigenvalues of problem (8.1) are nonnegative, have finite multiplicity, and can be represented as a non-decreasing divergent sequence $\lambda_k[\Omega]$, $k \in \mathbb{N}$ where each eigenvalue is repeated according to its multiplicity. In particular,

$$\lambda_k[\Omega] = \min_{\substack{E \subset V(\Omega)^m \\ \dim E = k}} \max_{\substack{u \in E \\ u \neq 0}} R[u],$$

for all $k \in \mathbb{N}$, where $V(\Omega)$ denotes either $H_0^1(\Omega)$ (for problem (8.2)) or $H^1(\Omega)$ (for problem (8.3)), and $R[u]$ is the Rayleigh quotient defined by

$$R[u] = \frac{\int_\Omega \sum_{\alpha,\beta=1}^N \sum_{i,j=1}^m a_{\alpha\beta}^{ij} \frac{\partial u_i}{\partial x_\alpha} \frac{\partial u_j}{\partial x_\beta} dx}{\int_\Omega |u|^2 dx}.$$

For the sake of brevity, in the sequel we shall use Einstein notation, hence summation symbols will be dropped.

In Section 8.1 we examine the problem of shape differentiability of the eigenvalues of problem (8.1). We consider problem (8.1) in $\phi(\Omega)$ and pull it back to $\Omega$, where $\phi$ belongs to a suitable class of diffeomorphisms. This analysis was exploited in [LaLa04, LaLa07] for the Laplace operator, in [BuLa13, BuLa14, BuPr14] for polyharmonic operators and in [BuLa] for the Reissner–Mindlin system (8.7). In particular, we derive Hadamard–type formulas for the symmetric functions of the eigenvalues of problem (8.1).

In Section 8.2 we consider the problem of finding critical points for the symmetric functions of the eigenvalues of problem (8.1), under volume constraint. This is strictly related to the problem of shape optimization of the eigenvalue (see [He06] for a detailed discussion on the topic). Similarly to what was done in [BuLa13, BuLa14, BuLa, BuPr14], and [LaLa06], we provide a characterization for the critical domains, and show that, for a particular class of coefficients $a_{\alpha\beta}^{ij}$, balls are critical domains for all the symmetric functions of the eigenvalues.

## 8.1 Analyticity Results

Let $\Omega$ be a bounded open set in $\mathbb{R}^N$ of class $C^1$. We shall consider problem (8.1) in a family of open sets parameterized by suitable diffeomorphisms $\phi$ defined on $\Omega$. Namely, we set

$$\mathscr{A}_\Omega = \left\{ \phi \in C^1(\overline{\Omega}; \mathbb{R}^N) : \inf_{\substack{x_1, x_2 \in \overline{\Omega} \\ x_1 \neq x_2}} \frac{|\phi(x_1) - \phi(x_2)|}{|x_1 - x_2|} > 0 \right\},$$

where $C^1(\overline{\Omega}; \mathbb{R}^N)$ denotes the space of all functions from $\overline{\Omega}$ to $\mathbb{R}^N$ of class $C^1$. Note that if $\phi \in \mathscr{A}_\Omega$ then $\phi$ is injective, Lipschitz continuous, and $\inf_{\overline{\Omega}} |\det \nabla \phi| > 0$. Moreover, $\phi(\Omega)$ is a bounded open set of class $C^1$ and the inverse map $\phi^{(-1)}$ belongs to $\mathscr{A}_{\phi(\Omega)}$. Thus it is natural to consider problem (8.1) on $\phi(\Omega)$ and study the dependence of $\lambda_k[\phi(\Omega)]$ on $\phi \in \mathscr{A}_\Omega$. To do so, we endow the space $C^1(\overline{\Omega}; \mathbb{R}^N)$ with its usual norm. Note that $\mathscr{A}_\Omega$ is an open set in $C^1(\overline{\Omega}; \mathbb{R}^N)$, see [LaLa04, Lemma 3.11]. Thus, it makes sense to study differentiability and analyticity properties of the maps $\phi \mapsto \lambda_k[\phi(\Omega)]$ defined for $\phi \in \mathscr{A}_\Omega$. For simplicity, we write $\lambda_k[\phi]$ instead of $\lambda_k[\phi(\Omega)]$. We fix a finite set of indexes $F \subset \mathbb{N}$ and we

consider those maps $\phi \in \mathscr{A}_\Omega$ for which the eigenvalues with indexes in $F$ do not coincide with eigenvalues with indexes not in $F$; namely, we set

$$\mathscr{A}_{F,\Omega} = \{\phi \in \mathscr{A}_\Omega : \lambda_k[\phi] \neq \lambda_l[\phi], \forall\, k \in F, l \in \mathbb{N} \setminus F\}.$$

It is also convenient to consider those maps $\phi \in \mathscr{A}_{F,\Omega}$ such that all the eigenvalues with index in $F$ coincide and set

$$\Theta_{F,\Omega} = \left\{\phi \in \mathscr{A}_{F,\Omega} : \lambda_{k_1}[\phi] = \lambda_{k_2}[\phi], \ \forall\, k_1, k_2 \in F\right\}.$$

For $\phi \in \mathscr{A}_{F,\Omega}$, the elementary symmetric functions of the eigenvalues with index in $F$ are defined by

$$\Lambda_{F,s}[\phi] = \sum_{\substack{k_1,\dots,k_s \in F \\ k_1 < \cdots < k_s}} \lambda_{k_1}[\phi] \cdots \lambda_{k_s}[\phi], \quad s = 1, \dots, |F|. \tag{8.8}$$

We have the following

**Theorem 1.** *Let $\Omega$ be a bounded open set in $\mathbb{R}^N$ of class $C^1$ and $F$ be a finite set in $\mathbb{N}$. The set $\mathscr{A}_{F,\Omega}$ is open in $\mathscr{A}_\Omega$, and the real-valued maps $\Lambda_{F,s}$ are real-analytic on $\mathscr{A}_{F,\Omega}$, for all $s = 1, \dots, |F|$. Moreover, if $\tilde{\phi} \in \Theta_{F,\Omega}$ is such that the eigenvalues $\lambda_k[\tilde{\phi}]$ assume the common value $\lambda_F[\tilde{\phi}]$ for all $k \in F$, and $\tilde{\phi}(\Omega)$ is of class $C^2$, then the Frechét differential of the map $\Lambda_{F,s}$ at the point $\tilde{\phi}$ is delivered by the formula*

$$d|_{\phi=\tilde{\phi}}(\Lambda_{F,s})[\psi] = -\lambda_F^s[\tilde{\phi}] \binom{|F|-1}{s-1} \sum_{l=1}^{|F|} \int_{\partial\tilde{\phi}(\Omega)} a^{ij}_{\alpha\beta} \frac{\partial v_i^{(l)}}{\partial y_\alpha} \frac{\partial v_j^{(l)}}{\partial y_\beta} \zeta \cdot \nu d\sigma, \tag{8.9}$$

*for problem (8.2), or*

$$d|_{\phi=\tilde{\phi}}(\Lambda_{F,s})[\psi] = -\lambda_F^s[\tilde{\phi}] \binom{|F|-1}{s-1} \sum_{l=1}^{|F|} \int_{\partial\tilde{\phi}(\Omega)} \left( \lambda_F |v^{(l)}|^2 - a^{ij}_{\alpha\beta} \frac{\partial v_i^{(l)}}{\partial y_\alpha} \frac{\partial v_j^{(l)}}{\partial y_\beta} \right) \zeta \cdot \nu d\sigma, \tag{8.10}$$

*for problem (8.3), for all $\psi \in C^1(\overline{\Omega}; \mathbb{R}^N)$, where $\zeta = \psi \circ \tilde{\phi}^{(-1)}$ and $\{v^{(l)}\}_{l \in F}$ is an orthonormal basis in $V(\tilde{\phi}(\Omega))^m$ (with respect to the scalar product (8.5)) of the eigenspace associated with $\lambda_F[\tilde{\phi}]$.*

The proof of Theorem 1 can be easily done adapting that of [LaLa04, Theorem 3.38] for the Dirichlet problem (8.2), and that of [LaLa07, Theorem 2.5] for the Neumann problem (8.3), and it can be found in [BuTh].

## 8.2 Isovolumetric Perturbations

We consider the following extremum problems for the symmetric functions of the eigenvalues

$$\min_{V[\phi]=\text{const}} \Lambda_{F,s}[\phi] \quad \text{or} \quad \max_{V[\phi]=\text{const}} \Lambda_{F,s}[\phi], \tag{8.11}$$

where $V[\phi]$ denotes the $N$-dimensional Lebesgue measure of $\phi(\Omega)$. Note that if $\tilde{\phi} \in \mathscr{A}_\Omega$ is a minimizer or maximizer in (8.11) then $\tilde{\phi}$ is a critical domain transformation for the map $\phi \mapsto \Lambda_{F,s}[\phi]$ subject to volume constraint, i.e.,

$$\operatorname{Ker} \mathrm{d}|_{\phi=\tilde{\phi}} V \subset \operatorname{Ker} \mathrm{d}|_{\phi=\tilde{\phi}} \Lambda_{F,s},$$

where $V$ is the real-valued function defined on $\mathscr{A}_\Omega$ which takes $\phi \in \mathscr{A}_\Omega$ to $V[\phi]$.

The following theorem provides a characterization of all critical domain transformations $\phi$ (see also [BuLa13, BuLa14, BuLa, BuPr14], and [LaLa06]).

**Theorem 2.** *Let $\Omega$ be a bounded open set in $\mathbb{R}^N$ of class $C^1$, and $F$ be a finite subset of $\mathbb{N}$. Assume that $\tilde{\phi} \in \Theta_{F,\Omega}$ is such that $\tilde{\phi}(\Omega)$ is of class $C^2$ and that the eigenvalues $\lambda_j[\tilde{\phi}]$ have the common value $\lambda_F[\tilde{\phi}]$ for all $j \in F$. Let $\{v^{(l)}\}_{l\in F}$ be an orthonormal basis in $V(\tilde{\phi}(\Omega))^m$ (with respect to the scalar product (8.5)) of the eigenspace corresponding to $\lambda_F[\tilde{\phi}]$. Then $\tilde{\phi}$ is a critical domain transformation for any of the functions $\Lambda_{F,s}$, $s = 1, \ldots, |F|$, with volume constraint if and only if there exists $c \in \mathbb{R}$ such that*

$$\sum_{l=1}^{|F|} a^{ij}_{\alpha\beta} \frac{\partial v_i^{(l)}}{\partial y_\alpha} \frac{\partial v_j^{(l)}}{\partial y_\beta} = c, \quad \text{on } \partial\tilde{\phi}(\Omega), \tag{8.12}$$

*for problem (8.2), or*

$$\sum_{l=1}^{|F|} \left( \lambda_F |v^{(l)}|^2 - a^{ij}_{\alpha\beta} \frac{\partial v_i^{(l)}}{\partial y_\alpha} \frac{\partial v_j^{(l)}}{\partial y_\beta} \right) = c, \quad \text{on } \partial\tilde{\phi}(\Omega), \tag{8.13}$$

*for problem (8.3).*

*Proof.* The proof is a straightforward application of Lagrange Multipliers Theorem combined with formulas (8.9) and (8.10).

Now we introduce the following definition, which is a generalization of the notion of rotation invariance for scalar operators to the case of vectorial operators.

**Definition 3.** The operator $\mathscr{L}$ defined by

$$\mathscr{L}(u)_j = -a^{ij}_{\alpha\beta} \frac{\partial^2 u_i}{\partial x_\alpha \partial x_\beta}$$

is said to be rotation invariant if there exists a group homomorphism

$$S : O_N(\mathbb{R}) \to O_m(\mathbb{R}),$$

(i.e., $S(AB) = S(A)S(B)$ for all $A, B \in O_N(\mathbb{R})$) such that

$$\mathscr{L}\left(S(R)^t u \circ R\right) = S(R)^t \mathscr{L}(u) \circ R,$$

for any $R \in O_N(\mathbb{R})$, and for any $u \in H^2_{loc}(\mathbb{R}^N)^m$.

**Theorem 4.** *Suppose that the operator associated with problem (8.1) is rotation invariant. Let $B$ be the unit ball in $\mathbb{R}^N$ centered at zero, and let $\lambda$ be an eigenvalue of problem (8.1) in $B$. Let $F$ be the subset of $\mathbb{N}$ of all $k$ such that the $k$-th eigenvalue of problem (8.1) in $B$ coincides with $\lambda$. Let $v^{(1)}, \dots, v^{(|F|)}$ be an orthonormal basis of the eigenspace associated with the eigenvalue $\lambda$ in $V(B)^m$. Then there exists $c \in \mathbb{R}$ such that condition (8.12) (condition (8.13) respectively) holds.*

*Proof.* First of all, note that by standard regularity theory (cf. [AgDoNi64, §10.3]), the functions $v^{(l)} \in C^\infty(\overline{B})$ for all $l \in F$.

Thanks to the rotation invariance, $\{(S(R)^t v_l) \circ R : l = 1, \dots, |F|\}$ is another orthonormal basis for the eigenspace associated with $\lambda$, whenever $R \in O_n(\mathbb{R})$, where $S(R)$ is defined as in Definition 3. Since both $\{v^{(l)} : l = 1, \dots, |F|\}$ and $\{(S(R)^t v^{(l)}) \circ R : l = 1, \dots, |F|\}$ are orthonormal bases, then there exists $A[R] \in O_N(\mathbb{R})$ with matrix $(A_{rh}[R])_{r,h=1,\dots,|F|}$ such that

$$(S(R)^t v^{(r)}) \circ R = \sum_{l=1}^{|F|} A_{rl}[R] v^{(l)}. \tag{8.14}$$

Using (8.14) we get

$$\begin{aligned}\sum_{k=1}^{|F|} |v^{(k)}|^2 \circ R &= \sum_{k=1}^{|F|} |(S(R)^t v^{(k)}) \circ R|^2 \\ &= \sum_{k=1}^{|F|} \left( \sum_{l=1}^{|F|} A_{kl}[R] v^{(l)} \right) \cdot \left( \sum_{h=1}^{|F|} A_{kh}[R] v^{(h)} \right) \\ &= \sum_{k=1}^{|F|} \sum_{l,h=1}^{|F|} A_{lk}[R] A_{kh}[R] (v^{(l)} \cdot v^{(h)}) = \sum_{l=1}^{|F|} |v^{(l)}|^2,\end{aligned}$$

and similarly,

$$\sum_{k=1}^{|F|} \left( a^{ij}_{\alpha\beta} \frac{\partial v_i^{(k)}}{\partial y_\alpha} \frac{\partial v_j^{(k)}}{\partial y_\beta} \right) \circ R = \sum_{l=1}^{|F|} \left( a^{ij}_{\alpha\beta} \frac{\partial v_i^{(l)}}{\partial y_\alpha} \frac{\partial v_j^{(l)}}{\partial y_\beta} \right).$$

This concludes the proof.

Thus we get the following

**Corollary 5.** *Let $\Omega$ be a domain in $\mathbb{R}^N$ of class $C^1$. Suppose that the operator associated with problem (8.1) is rotation invariant. Let $\tilde{\phi} \in \mathscr{A}_\Omega$ be such that $\tilde{\phi}(\Omega)$ is a ball. Let $\tilde{\lambda}$ be an eigenvalue of problem (8.1) in $\tilde{\phi}(\Omega)$, and let $F$ be the set of $j \in \mathbb{N}$ such that $\lambda_j[\tilde{\phi}] = \tilde{\lambda}$. Then $\tilde{\phi}$ is a critical point $\Lambda_{F,s}$ under volume constraint, for all $s = 1, \ldots, |F|$.*

**Acknowledgements** The author wishes to thank Prof. Pier Domenico Lamberti for his useful comments and remarks. The author acknowledges financial support from the research project 'Singular perturbation problems for differential operators' Progetto di Ateneo of the University of Padova. The author is a member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

## References

[AgDoNi64] S. Agmon, A. Douglis, L. Nirenberg, Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. II, Comm. Pure Appl. Math., 17, 35–92, 1964.

[BuTh] D. Buoso, Shape sensitivity analysis of the eigenvalues of polyharmonic operators and elliptic systems, Ph.D. Thesis, Università degli Studi di Padova, Padova, Italy, (2015)

[BuLa13] D. Buoso, P.D. Lamberti, Eigenvalues of polyharmonic operators on variable domains, ESAIM: COCV, 19 (2013), 1225–1235.

[BuLa14] D. Buoso, P.D. Lamberti, Shape deformation for vibrating hinged plates, Mathematical Methods in the Applied Sciences, 37 (2014), 237–244.

[BuLa] D. Buoso, P.D. Lamberti, Shape sensitivity analysis of the eigenvalues of the Reissner-Mindlin system, SIAM J. Math. Anal., 47 (2015), 407–426.

[BuPr14] Buoso, D., Provenzano, L.: A few shape optimization results for a biharmonic Steklov problem, J. Differential Equations (2015), http://dx.doi.org/10.1016/j.jde.2015.03.013.

[He06] A. Henrot, Extremum problems for eigenvalues of elliptic operators, Frontiers in Mathematics, Birkhäuser Verlag, Basel, 2006.

[LaLa04] P.D. Lamberti, M. Lanza de Cristoforis, A real analyticity result for symmetric functions of the eigenvalues of a domain dependent Dirichlet problem for the Laplace operator, J. Nonlinear Convex Anal 5 (2004), no. 1, 19–42.

[LaLa07] P.D. Lamberti, M.Lanza de Cristoforis, A real analyticity result for symmetric functions of the eigenvalues of a domain-dependent Neumann problem for the Laplace operator, Mediterr. J. Math. 4 (2007), no. 4, 435–449.

[LaLa06] P.D. Lamberti, M. Lanza de Cristoforis, Critical points of the symmetric functions of the eigenvalues of the Laplace operator and overdetermined problems, J. Math. Soc. Japan 58 (2006), no. 1, 231–245.

# Chapter 9
# Pollutant Dispersion in the Atmosphere: A Solution Considering Nonlocal Closure of Turbulent Diffusion

**D. Buske, M.T.B. Vilhena, B.E.J. Bodmann, R.S. Quadros, and T. Tirabassi**

## 9.1 Introduction

Increasing problems that involve pollution in the atmospheric boundary layer call for countermeasures, and simulation of pollutant dispersion is one of them. Hence, in the last years, analytical solutions for the advection–diffusion equation received attention in order to describe the pollutant dispersion in the boundary layer. So far there do exist analytical solutions in the literature, however, for specific and particular problems, see, for instance, the works [Ro55] [De78] [NiHa81] [Ti89] [ShSiYa96] [Ti03]. In fact, all these solutions are valid for very specialized practical situations with restrictions on wind and vertical profiles of eddy diffusivities. Costa et al. [CoEtAl06] presented a semi-analytical solution of the multidimensional advection–diffusion equation for more realistic physical scenario using an integral formulation. The solution is valid for a limited atmospheric boundary layer and general wind and vertical eddy diffusivity profiles, that are approximated by a stepwise function [MoEtAl06a] [CoEtAl11].

Finally a general two-dimensional solution without any restriction in the spatial function of wind and eddy diffusion coefficients was presented by [Wo05] [MoEtAl06b] [BuEtAl10]. The solving methodology was the Generalized Integral Laplace Transform Technique (GILTT) that is an analytical series solution including

D. Buske (✉) • R.S. Quadros
Federal University of Pelotas, Rua Tiradentes 2515, Pelotas 96010-900, RS, Brazil
e-mail: daniela.buske@ufpel.edu.br; regis.quadros@ufpel.edu.br

M.T.B. Vilhena • B.E.J. Bodmann
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: vilhena@pq.cnpq.br; bejbodmann@gmail.com

T. Tirabassi
Institute of Atmospheric Sciences and Climate of National Council of Research, Bologna, Italy
e-mail: t.tirabassi@isac.cnr.it



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_9

the solution of an associated Sturm–Liouville problem, expansion of the pollutant concentration in a series in terms of the attained eigenfunctions, replacement of this expansion in the advection–diffusion equation and, finally, taking moments. This procedure leads to a set of ordinary differential equations that are solved analytically by Laplace transform technique. A complete review of the GILTT method is given in [MoEtAl09]. More recently, the three-dimensional GILTT solution (3D-GILTT) considering local closure of turbulence was presented by Buske et al. [BuEtAl11]. The solution procedure makes use of the integral transform in the y-direction and then the resultant two-dimensional problem is solved following the previous works [Wo05] [MoEtAl06b] [MoEtAl09] [BuEtAl10] [BuEtAl11]. Note that no approximation is made along the solution derivation so that an exact solution is obtained except for round-off errors.

In this work we consider a three-dimensional problem with nonlocal closure of generic turbulent diffusion. The counter-gradient term in the turbulence closure made additional terms to appear in the advection–diffusion equation and these terms are related to the asymmetrical transport in the convective boundary layer. This new equation is solved by the 3D-GILTT method. Numerical results and statistical comparisons with experimental data are presented.

## 9.2 The Advection-Diffusion Equation and the 3D-GILTT Method

The stationary advection–diffusion equation of air pollution in the atmosphere is essentially a statement of conservation of the suspended material and it can be written as

$$\overline{u}\frac{\partial \overline{c}}{\partial x}+\overline{v}\frac{\partial \overline{c}}{\partial y}+\overline{w}\frac{\partial \overline{c}}{\partial z}=-\frac{\partial \overline{u'c'}}{\partial x}-\frac{\partial \overline{v'c'}}{\partial y}-\frac{\partial \overline{w'c'}}{\partial z}+S, \tag{9.1}$$

in which $\overline{c}$ denotes the average concentration of a passive contaminant $(g/m^3)$, $\overline{u}$, $\overline{v}$, and $\overline{w}$ in units of $(m/s)$ are the mean wind components along the axes $x$, $y$, and $z$, respectively, and $S$ is a source term. The terms $\overline{u'c'}$, $\overline{v'c'}$, $\overline{w'c'}$ represent, respectively, the turbulent fluxes of contaminants $(g/sm^2)$ in the longitudinal, crosswind, and vertical directions.

Observe that eqn. (9.1) has four unknown variables (the concentration $\overline{c}$ and turbulent fluxes) which lead us to the known turbulence closure problem. One of the most widely used closures for eqn. (9.1) is based on the gradient hypothesis (or K-theory) which, in analogy with Fick's law of molecular diffusion, assumes that turbulence causes a net movement of material following the gradient of material concentration at a rate which is proportional to the magnitude of the gradient [SePa98]:

$$\overline{u'c'} = -K_x \frac{\partial \overline{c}}{\partial x}; \overline{v'c'} = -K_y \frac{\partial \overline{c}}{\partial y}; \overline{w'c'} = -K_z \frac{\partial \overline{c}}{\partial z}, \tag{9.2}$$

Here $K_x$, $K_y$ and $K_z$ are the Cartesian components of the turbulent diffusion ($m^2/s$) in the $x$, $y$, and $z$ directions, respectively. In first-order closure all the information of the turbulence complexity is contained in the eddy diffusivities.

Eqn. (9.2), combined with the mass continuity equation, leads to the Cartesian advection–diffusion eqn. [Bl97]:

$$\overline{u}\frac{\partial \overline{c}}{\partial x} + \overline{v}\frac{\partial \overline{c}}{\partial y} + \overline{w}\frac{\partial \overline{c}}{\partial z} = \frac{\partial}{\partial x}\left(K_x \frac{\partial \overline{c}}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial \overline{c}}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial \overline{c}}{\partial z}\right) + S \tag{9.3}$$

The simplicity of the K-theory of turbulent diffusion has led to widespread use of this theory as mathematical basis for simulating pollutant dispersion (open country, urban, photochemical pollution, etc.). However, K-closure has its own limitations: in contrast to molecular diffusion, turbulent diffusion is scale-dependent. This means that the rate of diffusion of a cloud of material generally depends on the cloud dimension and the intensity of turbulence. As the cloud grows, larger eddies are incorporated in the expansion process, so that a progressively larger fraction of turbulent kinetic energy is available for the cloud expansion.

Another problem is that the down-gradient transport hypothesis is inconsistent with observed features of turbulent diffusion in the upper portion of the mixed layer for convective cases, where counter-gradient material fluxes are known to occur [DeWi75]. Because counter-gradient fluxes are thought to be indicative of boundary layer scale eddies, as opposed to small scales, such fluxes are often called nonlocal fluxes. Local K-theory is a method that parametrizes the effects of turbulent mixing based on how small eddies mix quantities along a local gradient of the transported quantity.

Already some decades ago it was noted that in the upper part of convectively driven boundary layers, the flux of scalars are counter to the gradient of the mean scalar profile [De66]. The mean potential temperature gradient and the flux change sign at different levels introducing a certain region in the convective boundary layer, where they have the same sign. This was in contrast to the common view in first order turbulent closure, that turbulent diffusion is directed along the down-gradient. In order to describe diffusion also in these regions, Ertel [Er42] and Deardoff [De66, De72] proposed to modify the usual applied flux-gradient relationship in K-theory approach according to

$$\overline{w'c'} = -K_z \left(\frac{\partial \overline{c}}{\partial z} - \gamma\right) \tag{9.4}$$

where $\gamma$ represents the counter-gradient term.

Many schemes and parametrizations for counter-gradient terms have been developed in the literature. Here, we use the parametrization proposed by van Dop

and Verver [Va01], $\gamma = -\frac{\beta}{K_z}\frac{\partial \overline{w'c'}}{\partial z}$ where $\beta = \frac{S_k \sigma_w T_{l_w}}{2}$, which is based on the work of Wyngaard and Weil [WyWe91].

$$\left[1 + \left(\frac{S_k \sigma_w T_{l_w}}{2}\right)\frac{\partial}{\partial z}\right]\overline{w'c'} = -K_z \frac{\partial \overline{c}}{\partial z} \tag{9.5}$$

Here $S_k = \overline{w'^3}/\overline{w'^2}^{3/2}$ is the skewness of the vertical turbulent velocity ($w'$), $\sigma_w$ is the vertical turbulent velocity standard deviation ($m/s$), and $T_{l_w}$ is the Lagrangian time scale ($s$). The second term in the operator (in the brackets) represents the nonlocal counter-gradient term.

Using eqns. (9.4) and (9.5), together with $\overline{w'c'} = \beta\,\overline{u}\frac{\partial \overline{c}}{\partial x} - K_z\frac{\partial \overline{c}}{\partial z} + \beta\,K_y\frac{\partial^2 \overline{c}}{\partial y^2}$, the turbulence closure problem is solved using a non-Fickian closure (also known as nonlocal closure). This approach models a more consistent kinetic eddy energy spectrum in different heights and the effect of the asymmetric transport of pollutant concentration by turbulent dispersion.

Applying the above eqns. in eqn. (9.1), in the Eulerian framework in which the $x$ direction coincides with that of the average wind field, yields

$$\overline{u}\frac{\partial \overline{c}}{\partial x} = K_y\frac{\partial^2 \overline{c}}{\partial y^2} + \frac{\partial}{\partial z}\left(K_z\frac{\partial \overline{c}}{\partial z}\right) - \frac{\partial}{\partial z}\left(\beta\overline{u}\frac{\partial \overline{c}}{\partial x}\right) + \frac{\partial}{\partial z}\left(\beta K_y\frac{\partial^2 \overline{c}}{\partial y^2}\right) \tag{9.6}$$

for $0 < z < h$, $0 < y < L_y$, and $x > 0$. In this work we neglect the diffusion component $K_x$ because we assume that the advection is dominant in the $x$-direction and also consider that $K_y$ depends only on the $z$-direction. Equation (9.6) is subject to the boundary conditions

$$K_z\frac{\partial \overline{c}}{\partial z} = 0 \qquad \text{at} \qquad z = 0, h \tag{9.7}$$

$$K_y\frac{\partial \overline{c}}{\partial y} = 0 \qquad \text{at} \qquad y = 0, L_y \tag{9.8}$$

and to the source condition,

$$\overline{u}c(0,y,z) = Q\delta(y - y_0)\delta(z - H_s)\,, \tag{9.9}$$

where $h$ is the boundary layer height ($m$), $L_y$ is a domain limit far from the source ($m$), $H_s$ is the height of the source ($m$), $Q$ is the emission rate ($g/s$), and $\delta$ is the Dirac delta functional.

To solve problem (9.6) by the GILTT method (see [BuEtAl07, MoEtAl09, BuEtAl11]), we initially apply the integral transform technique in the $y$ variable. To this end, we expand the pollutant concentration,

$$\overline{c}(x,y,z) = \sum_{m=0}^{M} \overline{c}_m(x,z)\, Y_m(y)\ , \tag{9.10}$$

where $Y_m(y) = cos(\lambda_m y)$ are orthogonal eigenfunctions with eigenvalues $\lambda_m = m\pi/L_y$ $(m = 0, 1, 2, ...)$.

To determine the unknown coefficients $\overline{c}_m(x,z)$, we substitute eqn. (9.10) in eqn. (9.6) and then apply the operator $\int_0^{L_y}(\cdot)Y_n(y)dy$. This procedure leads to the set with $M+1$ two-dimensional diffusion equations

$$\overline{u}\frac{\partial \overline{c}_m}{\partial x} = \frac{\partial}{\partial z}\left(K_z \frac{\partial \overline{c}_m}{\partial z}\right) - \frac{\partial}{\partial z}\left(\beta \overline{u}\frac{\partial \overline{c}_m}{\partial x}\right) - \lambda_m^2 K_y \overline{c}_m - \lambda_m^2 K_y \frac{\partial}{\partial z}(\beta \overline{c}_m) \tag{9.11}$$

The problem (9.11) is then solved analytically by the GILTT method following the works [BuEtAl07, MoEtAl09, BuEtAl10], where the solution of problem (9.11) is given by

$$\overline{c}_m(x,z) = \sum_{l=0}^{L} \overline{c}_{m,l}(x)\, \zeta_l(z)\ . \tag{9.12}$$

Here $\zeta_l(z) = \cos(\eta_l z)$ are a set of orthogonal eigenfunctions, with eigenvalues $\eta_l = l\pi/h$ (l=0,1,2,...).

Replacing eqn. (9.12) in eqn. (9.11) and taking moments, we get the first-order matrix differential equation

$$\frac{dP_m}{dx}(x) + G.P_m(x) = 0, \tag{9.13}$$

for $m = 0, \ldots, M$, where $P_m(x)$ is the column vector whose components are $\{\overline{c}_{m,l}\}$ for $l = 0, \ldots, L$. The matrix $G$ is composed by $G = B_1^{-1}B_2$, with the entries of matrices $B_1$ and $B_2$.

$$(B_1)_{l,j} = -\int_0^h \overline{u}\, \zeta_l(z)\, \zeta_j(z)\, dz$$

and

$$(B_2)_{l,j} = \int_0^h K_z' \zeta_l'(z)\, \zeta_j(z)\, dz - \lambda_l^2 \int_0^h K_z\, \zeta_l(z) \zeta_j(z)\, dz - \eta_l^2 \int_0^h K_y\, \zeta_l(z)\, \zeta_j(z)\, dz$$

A similar procedure leads to the boundary condition of problem (9.13):

$$P_m(0) = \overline{c}_{m,l}(0) = QA^{-1}\, \zeta_j(H_s)\, Y(y_0)\, dz \tag{9.14}$$

where $A^{-1}$ is the inverse of matrix $A$ with the entries: $A_{l,j} = \int_0^h \overline{u}\,\zeta_l(z)\,\zeta_j(z)\,dz$. In a fashion already shown in [MoEtAl09], we solve the problem (9.13) applying Laplace transform and diagonalization that leads to

$$\overline{P_m}(s) = X(sI+D)^{-1}\xi \tag{9.15}$$

where $\xi = X^{-1}P_m(0)$ is found from the equation $X\xi = P_m(0)$, and the values are calculated by $LU$-decomposition, whose computational effort is smaller than that of a matrix inversion. The elements of the diagonal matrix $(sI+D)$ are of the form $s+d_i$ where $d_i$ are the eigenvalues of the matrix $G$ and the elements of $(sI+D)^{-1}$ are $\frac{1}{s+d_i}$ whose Laplace transformed inverse is $e^{(-d_i x)}$. Let be $E(x)$ the diagonal matrix whose elements are $e^{(-d_i x)}$ the final solution is then given by

$$P_m(x) = XE(x)\xi\ . \tag{9.16}$$

Finally, using formula (9.12), we obtain the solution of the 2D problem, where $\zeta_l(z) = \cos(\zeta_l z)$ and $\overline{c}_{m,l}(x)$ is the solution of the transformed problem given by eqn. (9.13). Once $\overline{c}_m(x,z)$ are known, the final three-dimensional solution of problem (9.6) is given by expression (9.10), henceforth called 3D-GILTT (three-dimensional GILTT solution). It is noteworthy that the advection–diffusion equation with Fickian closure [BuEtAl11] is recovered in the limit $\beta \to 0$.

## 9.3 Turbulent Parameterization

In the literature one finds a considerable variety for calculating the vertical turbulent diffusion coefficient [DeMoVi01]. In order to validate the solution against experimental data, we use the vertical and lateral diffusion parametrization as suggested by Degrazia et al. [DeCaCa97] for convective conditions:

$$K_z = 0.22 w_* h \left(\frac{z}{h}\right)^{\frac{1}{3}} \left(1-\frac{z}{h}\right)^{\frac{1}{3}} \left(1 - e^{\frac{4z}{h}} - 0.0003 e^{\frac{8z}{h}}\right) \tag{9.17}$$

$$K_y = \frac{\sqrt{\pi}\sigma_v}{16(f_m)_v q_v} \tag{9.18}$$

with

$$\sigma_v^2 = \frac{0.98 c_v}{(f_m)_v^{\frac{2}{3}}} \left(\frac{\psi_\varepsilon}{q_v}\right)^{\frac{2}{3}} \left(\frac{z}{h}\right)^{\frac{2}{3}} w_*^2 \tag{9.19}$$

$$\psi_{\varepsilon}^{\frac{1}{3}} = \left( \left(1 - \frac{z}{h}\right)^2 \left(-\frac{z}{L}\right)^{-\frac{2}{3}} + 0.75 \right)^{\frac{1}{2}} \tag{9.20}$$

$(f_m)_v = 0.16$ and $q_v = 4.16\frac{z}{h}$. Here, $k$ is the von Karman constant ($k = 0.4$), $w_*$ is the convective velocity scale, $\sigma_v$ Eulerian standard deviation of the longitudinal turbulent velocity, $q_v$ is the stability function, $\psi_\varepsilon$ is the non-dimensional molecular dissipation rate function, and $(f_m)_v$ is the peak wavelength of the turbulent velocity spectrum.

In order to evaluate the vertical wind velocity variance $\sigma_w$ and Lagrangian time scale $T_{l_w}$ in $\beta = \frac{S_k \sigma_w T_{l_w}}{2}$ the following expressions were used [DeMoVi01, KaEtAl76, Ca82].

$$T_{L_w} = \frac{0.55}{4} \frac{1}{\sigma_w} \frac{z}{(f_m^*)_w}, \tag{9.21}$$

where $(f_m^*)_w = \frac{z}{(\lambda_m)_w}$ is the reduced frequency of the convective spectral peak and

$$(\lambda_m)_w = 1.8h \left[ 1 - exp\left(-\frac{4z}{h}\right) - 0.0003exp\left(\frac{8z}{h}\right) \right] \tag{9.22}$$

is the peak wavelength of the turbulent velocity spectrum. For the vertical wind velocity variance we use

$$\sigma_w^2 = 1.06 c_w \frac{\psi^{2/3}}{(f_m^*)_w^{2/3}} \left(\frac{z}{h}\right)^{2/3} w_*^2, \tag{9.23}$$

where $\psi = 1.5 - 1.2\left(\frac{z}{h}\right)^{1/3}$ is the turbulent molecular dissipation [DrEtAl83].

The wind speed profile can be described by a power law, according to [PaDu88]:

$$\frac{\overline{u_z}}{\overline{u_1}} = \left(\frac{z}{z_1}\right)^n, \tag{9.24}$$

where $\overline{u_z}$ and $\overline{u_1}$ are the horizontal mean wind speeds at heights $z$ and $z_1$ and $n$ is an exponent that is related to the intensity of turbulence [Ir89]. All components are determined so that the model together with the parametrization may be applied to an experiment, where the ground-level concentrations of emissions released from an elevated continuous source point in an unstable boundary layer are presented.

## 9.4 Application to a Meteorological Scenario

The obtained solution together with the eddy diffusivity parametrization for convective regimes is validated using data of the Copenhagen experiment [GrLy84]. The Copenhagen campaign [GrLy84] took place in the suburbs of Copenhagen, where an $SF_6$ tracer was released without buoyancy from a tower at a height of $115m$ and collected at ground level on arcs located at distances of 2000, 4000, and 6000 meters from the release point. The site is mainly residential with a roughness length of $0.6m$. The meteorological conditions during the dispersion experiments ranged from moderately unstable to a convective regime. Tracer releases typically started one hour before tracer sampling and stopped after a sampling period of two hours.

Due to the fact that there does not exist a consensus as to what numerical value should be attributed to the skewness parameter typical for a specific turbulence regime, in the present analysis we compare findings for $S_k = 0.6$, as proposed in Wyngaard and Weil [WyWe91] to those for $S_k = 1.0$ suggested by van Dop and Verver [Va01]. Results for both parametrizations are shown in Figure 9.1, where observed concentrations are plotted against the predictions by the advection–diffusion model with local and non-local closure for turbulence normalized by the emission source rate ($c/Q$).

In the sequel, we use standard statistical tools established by Hanna [Ha89] to compare the quality of the new approach. Table 9.1 presents the model results for the experiment mentioned above, and characterized in the following. The reduced mean

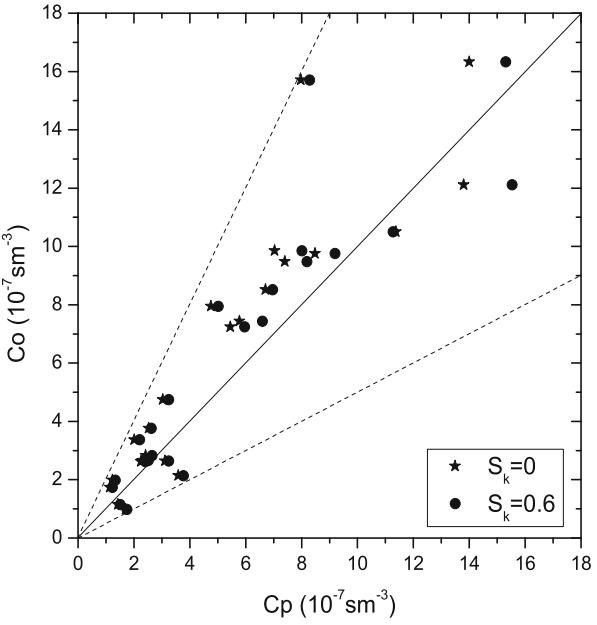


**Fig. 9.1** Scatter diagram of the observed versus predicted maximum ground level concentrations normalized by the emission rate.

**Table 9.1** Statistical results obtained with the 3D-GILTT method compared with the Copenhagen experiment.

| 3D-GILTT | NMSE | COR | FB | FS |
|---|---|---|---|---|
| $S_k = 0.0$ | 0.15 | 0.91 | 0.20 | 0.17 |
| $S_k = 0.6$ | 0.12 | 0.91 | 0.12 | 0.09 |
| $S_k = 1.0$ | 0.14 | 0.88 | 0.08 | −0.01 |

square error (NMSE = $\overline{(C_o - C_p)^2}/\overline{C_p}\,\overline{C_o}$ ) represents the model value dispersion with respect to data dispersion, where the subscripts $o$ and $p$ refer to observed and predicted quantities, respectively, and the overbar indicates an averaged value. The correlation coefficient (COR = $\overline{(C_o - \overline{C_o})(C_p - \overline{C_p})}/\sigma_o\sigma_p$ ) indicates the agreement between the mean values determined by the deterministic model against experimentally sampled values of an unknown distribution. It is noteworthy that differently than in parametric inference applications one does not expect a perfect correlation due to the stochastic nature of the observed phenomenon. The fractional bias (FB = $\overline{C_o} - \overline{C_p}/0.5(\overline{C_o} + \overline{C_p})$ ) indicates an asymmetry of the distribution of data points above and below the bisector, whereas the fractional standard deviations (FS = $(\sigma_o - \sigma_p)/0.5(\sigma_o + \sigma_p)$ ) indicate an asymmetry in the spread of data points.

Note that the predicted concentrations are of deterministic origin, whereas the data spread of stochastic experimental findings naturally are located in the larger intervals. Moreover, the statistical analysis indicates the nonlocal model with $S_k = 0.6$ as the more appropriate model for the Copenhagen scenario, since the model with $S_k = 1.0$ has a smaller standard deviation for the stochastic data in comparison with the standard deviation of the deterministic model. As a general statement we emphasize that the statistical analysis does not have the same meaning as the same concepts in statistical inference applications, due to the above-mentioned difference in the mean value character of the deterministic predictions, whereas experimental data are by nature of stochastic origin. Furthermore, only one sample of an unknown distribution is acquired, once repetitions are in general not feasible because of the variability of meteorological regimes. A reinterpretation of the statistical analysis is beyond the scope of the present contribution and will be discussed in a future work.

## 9.5 Conclusions

In the present contribution we presented a general solution of the three-dimensional steady state advection–diffusion equation considering nonlocal turbulence closure, which can be applied in operative models for simulation of turbulent dispersion of many scalar quantities, such as air pollution, radioactive material, among others. As model validation we chose the Copenhagen experiment with its predominantly convective regime. The theoretical model supplied mean concentrations that were compared to simple samples of a stochastic phenomenon with in general

non-negligible fluctuations and higher statistical moments, respectively. The analysis of the results showed acceptable agreement between computed values against experimental findings.

The already established three-dimensional steady state advection–diffusion model was generalized admitting also non-Fickian closure for turbulence. For this model a solution was determined in analytical representation. One of the emerging features is the coupling of the vertical and crosswind degrees of freedom, which was attained introducing a counter term in the Fickian closure according to the reasoning in references [Er42, De66, De72, DeWi75]. It is worth mentioning that the considered model, once solved provides mean concentrations as a consequence of the derivation of the model, i.e. the reduction of a deterministic-stochastic to purely deterministic model by the closure hypothesis.

From the numerical findings the model for $S_k = 0.6$ seems to be the better model, at least for comparable scenarios to the Copenhagen experiment. Note that for $S_k = 0.0$, which represents the advection-diffusion model with Fickian closure no updraft down-draft asymmetry is contemplated. This is different for presented models with $S_k = 0.6$ and $S_k = 1.0$. However, other validations shall be performed to verify as to which specific $0 \leq S_k \leq 1.0$ shall be associated with specific scenarios. Such a knowledge will open pathways for further generalizations of the discussed model and consequently will allow to implement a broader class of simulations where air quality and control is the principal issue.

**Acknowledgements** The authors thank CNPq and FAPERGS for partial financial support of this work.

## References

[Bl97] Blackadar, A.K.: *Turbulence and diffusion in the atmosphere: lectures in Environmental Sciences.* Springer-Verlag, 185pp. (1997).

[BuEtAl07] Buske, D., Vilhena, M.T., Moreira, D.M., and Tirabassi, T.: An analytical solution of the advection-diffusion equation considering non-local turbulence closure. Environ. Fluid Mechanics 7, 43–54 (2007).

[BuEtAl10] Buske, D., Vilhena, M.T., Moreira, D.M. and Tirabassi, T.: An Analytical Solution for the Transient Two-Dimensional Advection-Diffusion Equation with Non-Fickian Closure in Cartesian Geometry by Integral Transform Technique. *Integral Methods in Science and Engineering: Computational methods*, Boston: Birkhauser, 33–40 (2010).

[BuEtAl11] Buske, D., Vilhena, M.T., Segatto, C.F. and Quadros, R.S.: A General Analytical Solution of the Advection-Diffusion Equation for Fickian Closure. *Integral Methods in Science and Engineering: Computational and Analytic Aspects*, Boston: Birkhauser, 25–34 (2011).

[Ca82] Caughey, S.J.: Observed characteristics of the atmospheric boundary layer. In: Atmospheric turbulence and air pollution modeling, Boston, 1982.

[CoEtAl06] Costa, C.P., Vilhena, M.T., Moreira, D.M. and Tirabassi, T.: Semi-analytical solution of the steady three-dimensional advection-diffusion equation in the planetary boundary layer. Atmos. Environ. **40**, n. 29, 5659–5669 (2006).

[CoEtAl11] Costa, C.P.; Tirabassi, T.; Vilhena, M.T. & Moreira, D.M. (2011). A general formulation for pollutant dispersion in the atmosphere. *J. Eng. Math.*, Published online. Doi 10.1007/s10665-011-9495-z.

[De66] Deardorff, J.W.: The countergradient heat flux in the lower atmosphere and in the laboratory. J. Atmo. Sci. **23**, 503–506 (1966).

[De72] Deardorff, J.W.: Numerical investigation of neutral and unstable planetary boundary layers. J. Atmo. Sci. **29**, 91–115 (1972).

[DeWi75] Deardorff, J.W. and Willis, G.E.: A parameterization of diffusion into the mixed layer. J. Appl. Meteor. **14**, 1451–1458 (1975).

[DeCaCa97] Degrazia, G.A., Campos Velho, H.F., Carvalho, J.C.: Nonlocal exchange coefficients for the convective boundary layer derived from spectral properties. Cont. Atm. Phys., 57–64 (1997).

[De78] Demuth, C.: A contribution to the analytical steady solution of the diffusion equation for line sources. Atm. Env. **12**, 1255–1258 (1978).

[DeMoVi01] Degrazia, G.A., Moreira, D.M. and Vilhena, M.T.: Derivation of an eddy diffusivity depending on source distance for vertically inhomogeneous turbulence in a convective boundary layer. J. Appl. Meteor. **40**, 1233–1240 (2001).

[DrEtAl83] Druilhet, A., Frangi, J.P, Guedalia, D. and Fontan, J.: Experimental studies of the turbulence structure parameters of the convective boundary layer. J. Clim. Appl. Meteorol. **22**, 594–608 (1983).

[Er42] Ertel, H.: Der vertikale turbulenz-wärmestrom in der atmosphäre. Meteor. Z. **59**, 250–253 (1942).

[GrLy84] Gryning, S.E. and Lyck, E.: Atmospheric dispersion from elevated source in an urban area: comparison between tracer experiments and model calculations. J. Appl. Meteor. **23**, 651–654 (1984).

[Ha89] Hanna, S.R.: Confidence limit for air quality models as estimated by bootstrap and jacknife resampling methods. Atm. Env. **23**, 1385–1395 (1989).

[Ir89] Irwin, J.S.: A theoretical variation of the wind profile power-low exponent as a function of surface roughness and stability. Atm. Env. **13**, 191–194(1979).

[KaEtAl76] Kaimal, J.C., Wyngaard, J.C., Haugen, D.A., Coté, O.R., Izumi, Y., Caughey, S.J. and Readings, C.J.: Turbulence structure in the convective boundary layer. J. Atmos. Sci. **33**, 2152–2169 (1976).

[MoEtAl06a] Moreira, D.M., Vilhena, M.T., Tirabassi, T., Costa, C. and Bodmann, B.: Simulation of pollutant dispersion in atmosphere by the Laplace transform: the ADMM approach. Water, Air and Soil Pollution **177**, 411–439 (2006a).

[MoEtAl06b] Moreira, D.M., Vilhena, M.T., Buske, D. and Tirabassi, T.: The GILTT solution of the advection-diffusion equation for an inhomogeneous and nonstationary PBL. Atm. Env. **40**, 3186–3194 (2006b).

[MoEtAl09] Moreira, D. M., Vilhena, M. T., Buske, D. and Tirabassi, T.: The state-of-art of the GILTT method to simulate pollutant dispersion in the atmosphere. Atm. Research **92**, 1–17 (2009).

[NiHa81] Nieuwstadt F.T.M. and de Haan B.J.: An analytical solution of one-dimensional diffusion equation in a nonstationary boundary layer with an application to inversion rise fumigation. Atmos. Environ. **15**, 845–851 (1981).

[PaDu88] Panofsky, A.H., Dutton, J.A.: *Atmospheric Turbulence.* John Wiley & Sons, New York (1988).

[Ro55] Rounds, W.: Solutions of the two-dimensional diffusion equation. Trans. Am. Geophys. Union **36**, 395–405 (1955).

[SePa98] Seinfeld, J.H. and Pandis, S.N. (1998). *Atmospheric chemistry and physics*. John Wiley & Sons, New York, 1326 pp.

[ShSiYa96] Sharan, M., Singh, M.P. and Yadav, A.K.: A mathematical model for the atmospheric dispersion in low winds with eddy diffusivities as linear functions of downwind distance. Atmos. Environ. **30**, n.7, 1137–1145 (1996).

[Ti89] Tirabassi, T.: Analytical air pollution and diffusion models. Water, Air and Soil Pollution **47**, 19–24 (1989).

[Ti03] Tirabassi T.: Operational advanced air pollution modeling. PAGEOPH **160**, n. 1-2, 05–16 (2003).

[Va01] van Dop, H., Verver, G.S.: Countergradient transport revisited. J. Atm. Sci. **58**, 2240–2247 (2001).

[WyWe91] Wyngaard, J.C. and Weil, J.C.: Transport asymmetry in skewed turbulence. Phys. Fluids A **3**, 155–162 (1991).

[Wo05] Wortmann, S., Vilhena, M.T., Moreira, D.M. and Buske, D.: A new analytical approach to simulate the pollutant dispersion in the PBL. Atm. Env. **39**, 2171–2178 (2005).

# Chapter 10
# The Characteristic Matrix of Nonuniqueness for First-Kind Equations

**C. Constanda and D.R. Doty**

## 10.1 Introduction

Let $S$ be a finite domain in $\mathbb{R}^2$, bounded by a simple, closed, $C^2$ curve $\partial S$. We denote by $x$ and $y$ generic points in $S \cup \partial S$ and by $|x-y|$ the distance between $x$ and $y$ in the Cartesian metric. Also, let $C^{0,\alpha}(\partial S)$ and $C^{1,\alpha}(\partial S)$, $\alpha \in (0,1)$, be, respectively, the spaces of Hölder continuous and Hölder continuously differentiable functions on $\partial S$. In what follows, Greek and Latin indices take the values $1, 2$ and $1, 2, 3$, respectively, and a superscript T denotes matrix transposition.

For any function $f$ continuous on $\partial S$, we define the 'calibration' functional $p$ by

$$pf = \int_{\partial S} f \, ds.$$

Using the fundamental solution for the two-dimensional Laplacian

$$g(x,y) = -\frac{1}{2\pi} \ln |x-y|,$$

we define the single-layer harmonic potential of density $\varphi$ by

$$(V\varphi)(x) = \int_{\partial S} g(x,y)\varphi(y) \, ds(y).$$

C. Constanda (✉) • D.R. Doty
The University of Tulsa, 800 S. Tucker Drive, Tulsa, OK 74104, USA
e-mail: christian-constanda@utulsa.edu; dale-doty@utulsa.edu



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_10

The proof of the following assertion can be found, for example, in [Co94] or [Co00].

**Theorem 1.** *For any $\alpha \in (0,1)$, there are a unique nonzero function $\Phi \in C^{0,\alpha}(\partial S)$ and a unique number $\omega$ such that*

$$V\Phi = \omega \quad \text{on } \partial S, \quad p\Phi = 1.$$

It is easy to see that $\Phi$ and $\omega$ depend on $g$ and $\partial S$.

The numbers $2\pi\omega$ and $e^{-2\pi\omega}$ are called *Robin's constant* and the *logarithmic capacity* of $\partial S$.

For a circle with the center at the origin and radius $R$, both $\Phi$ and $\omega$ can be determined explicitly:

$$\Phi = \frac{1}{2\pi R}, \quad \omega = -\frac{1}{2\pi} \ln R.$$

For other boundary curves, $\Phi$ and $\omega$ are practically impossible to determine analytically and must be computed by numerical methods.

If the solution of the Dirichlet problem in $S$ with data function $f$ on $\partial S$ is sought as $u = V\varphi$, then $\varphi$ is a solution of the (weakly singular) first-kind boundary integral equation

$$V\varphi = f \quad \text{on } \partial S.$$

This is a well-posed problem if and only if $\omega \neq 0$. If $\omega = 0$, the above equation has infinitely many solutions, which are expressed in terms of $\Phi$.

## 10.2 Plane Elastic Strain

Consider a plate made of a homogeneous and isotropic material with Lamé constants $\lambda$ and $\mu$, which undergoes deformations in the $(x_1, x_2)$-plane. If the body forces are negligible, then its (static) displacement vector $u = (u_1, u_2)^{\mathrm{T}}$ satisfies the equilibrium system of equations [Co00]

$$Au = 0 \quad \text{in } S,$$

where

$$A(\partial_1, \partial_2) = \begin{pmatrix} \mu\Delta + (\lambda+\mu)\partial_1^2 & (\lambda+\mu)\partial_1\partial_2 \\ (\lambda+\mu)\partial_1\partial_2 & \mu\Delta + (\lambda+\mu)\partial_2^2 \end{pmatrix}.$$

It is not difficult to show [Co00] that the columns $F^{(i)}$ of the matrix

$$F = \begin{pmatrix} 1 & 0 & x_2 \\ 0 & 1 & -x_1 \end{pmatrix}$$

form a basis for the space of rigid displacements.

The 'calibrating' vector-valued functional $p$ is defined for continuous $2 \times 1$ vector functions $f$ by

$$pf = \int_{\partial S} F^{\mathrm{T}} f \, ds.$$

A matrix of fundamental solutions for $A$ is [Co00]

$$D(x,y) = -\frac{1}{4\pi\mu(\gamma+1)} \times \begin{pmatrix} 2\gamma \ln|x-y| + 2\gamma + 1 - \dfrac{2(x_1-y_1)^2}{|x-y|^2} & -\dfrac{2(x_1-y_1)(x_2-y_2)}{|x-y|^2} \\ -\dfrac{2(x_1-y_1)(x_2-y_2)}{|x-y|^2} & 2\gamma \ln|x-y| + 2\gamma + 1 - \dfrac{2(x_2-y_2)^2}{|x-y|^2} \end{pmatrix},$$

$$\gamma = \frac{\lambda + 3\mu}{\lambda + \mu}.$$

The single-layer potential of density $\varphi$ is defined by

$$(V\varphi)(x) = \int_{\partial S} D(x,y)\varphi(y) \, ds(y).$$

The proof of the following assertion can be found in [Co00].

**Theorem 2.** *There is a unique $2 \times 3$ matrix function $\Phi \in C^{0,\alpha}(\partial S)$ and a unique $3 \times 3$ constant symmetric matrix $\mathscr{C}$ such that the columns $\Phi^{(i)}$ of $\Phi$ are linearly independent and*

$$V\Phi = F\mathscr{C} \quad \text{on } \partial S, \quad p\Phi = I,$$

*where $I$ is the identity matrix.*

Clearly, $\Phi$ and $\mathscr{C}$ depend on $A$, $D$, and $\partial S$.

In the so-called alternative indirect method [Co00], the solution of the Dirichlet problem in $S$ with data function $f$ on $\partial S$ is sought in the form

$$u = V\varphi. \tag{10.1}$$

Then the problem reduces to the (weakly singular) boundary integral equation

$$V\varphi = f \quad \text{on } \partial S. \tag{10.2}$$

**Theorem 3.** *Equation* (10.2) *has a unique solution* $\varphi \in C^{0,\alpha}(\partial S)$, $\alpha \in (0,1)$, *for any* $f \in C^{1,\alpha}(\partial S)$ *if and only if* $\det \mathscr{C} \neq 0$. *In this case,* (10.1) *is the unique solution of the Dirichlet problem.*

If $\det \mathscr{C} = 0$, then the unique solution of the Dirichlet problem is obtained by solving an ill-posed modified boundary integral equation whose infinitely many solutions are constructed with $\Phi$ and $\mathscr{C}$.

In the so-called refined indirect method [Co00], the solution of the Dirichlet problem is sought as a pair $\{\varphi, c\}$ such that

$$u = V\varphi - Fc \quad \text{in } S, \quad p\varphi = s,$$

where $s$ a constant $3 \times 1$ vector chosen (arbitrarily) a priori and $c$ is a constant $3 \times 1$ vector. This leads to the system of boundary integral equations

$$V\varphi - Fc = f \quad \text{on } \partial S, \quad p\varphi = s. \tag{10.3}$$

**Theorem 4.** *System* (10.3) *has a unique solution* $\{\varphi, c\}$ *with* $\varphi \in C^{0,\alpha}(\partial S)$ *for any* $f \in C^{1,\alpha}(\partial S)$, $\alpha \in (0,1)$, *and any* $s$.

It is important to evaluate the arbitrariness in the representation of the solution with respect to the prescribed 'calibration' $s$.

**Theorem 5.** *If* $\{\varphi^{(1)}, c^{(1)}\}$, $\{\varphi^{(2)}, c^{(2)}\}$ *are two solutions of* (10.3) *constructed with* $s^{(1)}$ *and* $s^{(2)}$, *respectively, then*

$$\varphi^{(2)} = \varphi^{(1)} + \Phi(s^{(2)} - s^{(1)}),$$
$$c^{(2)} = c^{(1)} + \mathscr{C}(s^{(2)} - s^{(1)}).$$

It is not easy to compute $\Phi$ and $\mathscr{C}$ analytically, or even numerically, in arbitrary domains $S$, but this can be accomplished if $S$ is a circular disk. Let $\partial S$ be the circle with center at the origin and radius $R$. In this case, $\Phi$ and $\mathscr{C}$ can be determined analytically as

$$\Phi = \frac{1}{2\pi R} \begin{pmatrix} 1 & 0 & R^{-2}x_2 \\ 0 & 1 & -R^{-2}x_1 \end{pmatrix},$$

$$\mathscr{C} = -\frac{1}{4\pi\mu(\lambda+2\mu)R^2} \times \begin{pmatrix} (\lambda+3\mu)R^2(\ln R+1) & 0 & 0 \\ 0 & (\lambda+3\mu)R^2(\ln R+1) & 0 \\ 0 & 0 & -(\lambda+\mu) \end{pmatrix}.$$

Clearly, $\det \mathscr{C} = 0$ if and only if $R = e^{-1}$.

Analytic computation of $\Phi$ and $\mathscr{C}$ is practically impossible for non-circular domains, and must be performed numerically.

We choose four $3 \times 1$ constant vectors $s^{(0)}, s^{(i)}$ such that the set $\{s^{(i)} - s^{(0)}\}$ is linearly independent, and form the $3 \times 3$ matrix $\Sigma$ with columns $s^{(i)} - s^{(0)}$. Also, we choose an arbitrary function $f$. Next, we compute the solutions $\{\varphi^{(0)}, c^{(0)}\}$, $\{\varphi^{(i)}, c^{(i)}\}$ of (10.3) corresponding to $s^{(0)}$, $s^{(i)}$, respectively, and $f$, by the refined indirect method, then form the $2 \times 3$ matrix function $\Psi$ with columns $\varphi^{(i)} - \varphi^{(0)}$ and the constant $3 \times 3$ matrix $\Gamma$ with columns $c^{(i)} - c^{(0)}$.

From Theorem 4 it follows that

$$\varphi^{(i)} - \varphi^{(0)} = \Phi(s^{(i)} - s^{(0)}),$$
$$c^{(i)} - c^{(0)} = \mathscr{C}(s^{(i)} - s^{(0)}),$$

or, what is the same,

$$\Phi\Sigma = \Psi, \quad \mathscr{C}\Sigma = \Gamma;$$

hence,

$$\Phi = \Psi\Sigma^{-1}, \quad \mathscr{C} = \Gamma\Sigma^{-1}.$$

A similar analysis can be performed for other two-dimensional linear elliptic systems with constant coefficients—for example, the system modeling bending of elastic plates with transverse shear deformation [Co00]. No apparent connection exists between the matrix $\mathscr{C}$ and the characteristic constant $\omega$ of $\partial S$.

## 10.3 Numerical Examples

Consider a steel plate with scaled Lamé coefficients

$$\lambda = 11.5, \quad \mu = 7.69,$$

and let $\partial S$ (see Figure 10.1) be the curve of parametric equations

$$x_1(t) = 2\cos(\pi t) - \tfrac{4}{3}\cos(2\pi t) + \tfrac{10}{3},$$
$$x_2(t) = 2\sin(\pi t) + 2, \quad 0 \le t \le 2.$$

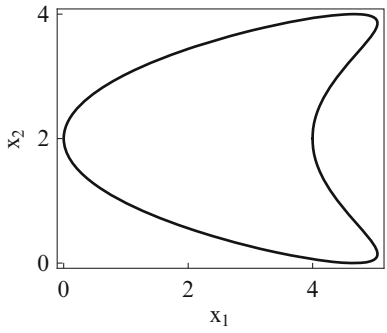


**Fig. 10.1** The boundary curve $\partial S$.

We choose the vectors

$$s^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad s^{(1)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad s^{(2)} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad s^{(3)} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

$$f(x) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The approximating functions for computing $\varphi^{(0)}(t)$ and $\varphi^{(i)}(t)$ are piecewise cubic Hermite splines on 12 knots; that is, the interval $0 \le t \le 2$ is divided into 12 equal subintervals. Then the characteristic matrix (with entries rounded off to 5 decimal places) is

$$\mathscr{C} = \begin{pmatrix} -0.01627 & -0.01083 & -0.00370 \\ -0.01083 & -0.00892 & 0.00542 \\ -0.00370 & 0.00542 & 0.00185 \end{pmatrix}.$$

Here,

$$\det \mathscr{C} = 1.08273 \times 10^{-6}.$$

The graphs of the components $\Phi_{\alpha i}$ of $\Phi$ are shown in Figure 10.2.

As a second example, consider the 'expanding' ellipse $\partial S$ of parametric equations

$$x_1(t) = 2k\cos(\pi t),$$
$$x_2(t) = k\sin(\pi t), \quad 0 \le t \le 2.$$

The graph of $\det \mathscr{C}$ as a function of $k$ is shown in Figure 10.3.

Here, $\det \mathscr{C} = 0$ for $k = 0.22546$ and $k = 0.26934$.

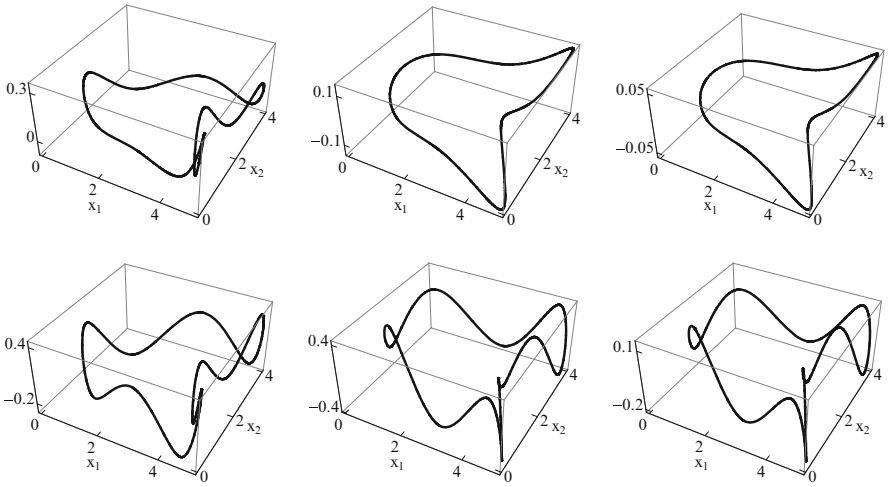


**Fig. 10.2** Graphs of the $\Phi_{\alpha i}$.

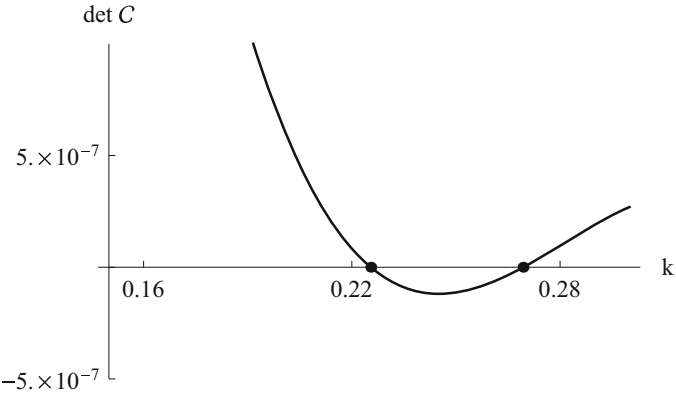


**Fig. 10.3** Graph of $\det \mathscr{C}$.

## References

[Co94] Constanda, C.: On integral solutions of the equations of thin plates. *Proc. Roy. Soc. London Ser. A*, **444**, 261–268 (1994).

[Co00] Constanda, C.: *Direct and Indirect Boundary Integral Equation Methods*, Chapman & Hall/CRC, Boca Raton, FL (2000).

# Chapter 11
# On the Spectrum of Volume Integral Operators in Acoustic Scattering

**M. Costabel**

## 11.1 Volume Integral Equations in Acoustic Scattering

Volume integral equations have been used as a theoretical tool in scattering theory for a long time. A classical application is an existence proof for the scattering problem based on the theory of Fredholm integral equations. This approach is described for acoustic and electromagnetic scattering in the books by Colton and Kress [CoKr83, CoKr98] where volume integral equations appear under the name Lippmann–Schwinger equations.

In electromagnetic scattering by penetrable objects, the volume integral equation (VIE) method has also been used for numerical computations. In particular the class of discretization methods known as 'discrete dipole approximation' [PuPe73, DrFl94] has become a standard tool in computational optics applied to atmospheric sciences, astrophysics and recently to nano-science under the keyword 'optical tweezers' (see the survey article [YuHo07] and the literature quoted there). In sharp contrast to the abundance of articles by physicists describing and analyzing applications of the VIE method, the mathematical literature on the subject consists only of a few articles. An early spectral analysis of a VIE for magnetic problems was given in [FrPa84], and more recently [Ki07, KiLe09] have found sufficient conditions for well-posedness of the VIE in electromagnetic and acoustic scattering with variable coefficients. In [CoDK10, CoDS12], we investigated the essential spectrum of the VIE in electromagnetic scattering under general conditions on the complex-valued coefficients, finding necessary and sufficient conditions for well-posedness in the sense of Fredholm in the physically relevant energy spaces. A detailed presentation of these results can be found in the thesis [Sa14]. Publications based on the thesis are in preparation. Curiously, whereas the study of VIE in electromagnetic scattering

M. Costabel (✉)
IRMAR, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes, France
e-mail: martin.costabel@univ-rennes1.fr



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_11

has thus been completed as far as questions of Fredholm properties are concerned, the simpler case of acoustic scattering does not seem to have been covered in the same depth. It is the purpose of the present chapter to close this gap.

The basic idea of the VIE method in scattering by a penetrable object is to consider the effect of the scatterer as a perturbation of a whole-space constant coefficient problem and to solve the latter by convolution with the whole-space fundamental solution. In the acoustic case, we consider the scalar linear elliptic equation

$$\operatorname{div} a(x)\nabla u + k(x)^2 u = f \qquad \text{in } \mathbb{R}^d \tag{11.1}$$

where we suppose that the (in general, complex-valued) coefficients $a$ and $k$ are constant outside of a compact set:

$$a(x) \equiv 1, \quad k(x) \equiv k \in \mathbb{C} \quad \text{outside of the bounded domain } \Omega.$$

and $f$ has compact support. We further assume that $u$ satisfies the outgoing Sommerfeld radiation condition. It is well known that under very mild conditions on the regularity of the coefficients $a$ and $k$, there is at most one solution of this problem.

We then rewrite (11.1) as a perturbed Helmholtz equation:

$$(\Delta + k^2)u = f - \operatorname{div} \alpha \nabla u - \beta u \tag{11.2}$$

with

$$\alpha(x) = a(x) - 1, \; \beta(x) = k(x)^2 - k^2 .$$

Let now $G_k$ be the outgoing full-space fundamental solution of the Helmholtz equation, i.e. the unique distribution in $\mathbb{R}^d$ satisfying $(\Delta + k^2)G_k = -\delta$ and the Sommerfeld radiation condition. In dimension $d = 3$, we have

$$G_k(x) = \frac{e^{ik|x|}}{4\pi|x|} .$$

We obtain the VIE from the following well-known lemma.

**Lemma 1.** *Let $u$ be a distribution in $\mathbb{R}^d$ satisfying-$(\Delta + k^2)u = v$, where $v$ has compact support, and the Sommerfeld radiation condition. Then $u = G_k * v$, and if $v$ is an integrable function, the convolution can be written as an integral:*

$$u(x) = \int G_k(x-y)\, v(y)\, dy .$$

Applying this lemma to (11.2), we obtain the equation

$$u = -G_k * f + \operatorname{div} G_k * (\alpha \nabla u) + G_k * (\beta u),$$

valid in the distributional sense on $\mathbb{R}^d$. This can be written as a VIE

$$u(x) - \operatorname{div} \int_\Omega G_k(x-y)\alpha(y)\nabla u(y)\,dy - \int_\Omega G_k(x-y)\beta(y)u(y)\,dy = u^{\text{inc}}(x) \tag{11.3}$$

where we use the notation

$$u^{\text{inc}}(x) := -\int G_k(x-y)f(y)\,dy\,.$$

The fact that the coefficients $\alpha$ and $\beta$ vanish outside of $\Omega$ permits to consider the integral equation (11.3) on any domain $\widehat{\Omega}$ satisfying $\Omega \subset \widehat{\Omega} \subset \mathbb{R}^d$. Once $u$ solves (11.3) on $\widehat{\Omega}$, one can use the same formula (11.3) to extend $u$ outside of $\widehat{\Omega}$. It is clear that the resulting function $u$ will not depend on $\widehat{\Omega}$ and will be a solution of the original scattering problem (11.1). In the following we will make the minimal choice $\widehat{\Omega} = \Omega$ and therefore consider (11.3) as an integral equation on $\Omega$. We shall abbreviate this integral equation as

$$u - Au = u^{\text{inc}} \tag{11.4}$$

with

$$Au(x) = \operatorname{div} \int_\Omega G_k(x-y)\alpha(y)\nabla u(y)\,dy + \int_\Omega G_k(x-y)\beta(y)u(y)\,dy\,. \tag{11.5}$$

Assuming that $\Omega$ is a bounded Lipschitz domain, one can consider the VIE (11.4) in the standard Sobolev spaces $H^s(\Omega)$. The natural energy space associated with the second-order PDE (11.1) is $H^1(\Omega)$, but other values of $s$ can be interesting, too, in particular $s = 0$, i.e. the space $L^2(\Omega)$, which seems naturally associated with the apparent structure of (11.4) as a second kind integral equation and may be useful for analyzing certain numerical algorithms for its solution.

The convolution with $G_k$ is a pseudo-differential operator of order $-2$, mapping distributions with compact support and Sobolev regularity $s$ to $H^{s+2}_{\text{loc}}(\mathbb{R}^d)$ for any $s \in \mathbb{R}$, which implies immediately boundedness of the operator $A$ in low-order Sobolev spaces:

**Proposition 1.** *Let $\alpha, \beta \in L^\infty(\Omega)$. Then*

$$A : H^1(\Omega) \to H^1(\Omega) \ \textit{is bounded}\,.$$

*If, in addition, $\nabla\alpha \in L^\infty(\Omega)$, then $A$ is a bounded operator in $L^2(\Omega)$.*

Another immediate observation is that the second integral operator in (11.5) maps $L^2$ to $H^2$, and is therefore compact as an operator in $L^2$ and in $H^1$. This is relevant if $a(x)$ is constant everywhere, since then $\alpha \equiv 0$ and the first integral operator in (11.5), which is not compact, in general, is absent.

**Theorem 1.** *Let $a(x) = 1$ in $\mathbb{R}^d$ and $k \in L^\infty(\mathbb{R}^d)$. Then the VIE (11.3) is a second kind Fredholm integral equation with a weakly singular kernel and the Fredholm alternative holds: The operator $\mathbb{I} - A$ is a Fredholm operator of index zero in $L^2(\Omega)$ and in $H^1(\Omega)$.*

## 11.2 Smooth Coefficients

Besides the case of the Laplace operator addressed in Theorem 1, another situation is well known and is studied, for example, in the book [CoKr83]. This is the case of a coefficient $a(x)$ that is smooth on all of $\mathbb{R}^d$. In this case, $\alpha = 0$ on the boundary $\Gamma = \partial\Omega$, and the first integral operator in (11.5) can be transformed by integration by parts:

$$\begin{aligned}
\operatorname{div} G_k * (\alpha \nabla u)(x) &= -\operatorname{div} \int_\Omega \nabla_y \big( G_k(x-y)\alpha(y) \big) u(y)\, dy \\
&= \Delta \int_\Omega G_k(x-y)\alpha(y)u(y)\, dy - \operatorname{div} \int_\Omega G_k(x-y)(\nabla\alpha)(y)\, u(y)\, dy \\
&= -\alpha(x)u(x) - k^2 \int_\Omega G_k(x-y)\alpha(y)u(y)\, dy - \operatorname{div} \int_\Omega G_k(x-y)(\nabla\alpha)(y)\, u(y)\, dy.
\end{aligned}$$

This allows us to write the VIE (11.3) in an equivalent form that shows its nature as a Fredholm integral equation of the second kind with a weakly singular kernel:

$$\begin{aligned}
&a(x)u(x) - \int_\Omega G_k(x-y)(\beta(y) - k^2\alpha(y))u(y)\, dy \\
&\quad + \operatorname{div} \int_\Omega G_k(x-y)(\nabla\alpha)(y)\, u(y)\, dy - \int_\Omega G_k(x-y)\beta(y)u(y)\, dy = u^{\mathrm{inc}}(x)
\end{aligned} \tag{11.6}$$

**Theorem 2.** *Let $a \in C^1(\mathbb{R}^d)$ and $k \in L^\infty(\mathbb{R}^d)$. Then the operator $\mathbb{I} - A$ is a Fredholm operator of index zero in $L^2(\Omega)$ and in $H^1(\Omega)$ if and only if $a(x) \neq 0$ for all $x \in \Omega$.*

## 11.3 Piecewise Smooth Coefficients

In obstacle scattering, the case of a globally smooth coefficient $a(x)$ is not natural. There one expects rather a sharp interface where the material properties change discontinuously. We thus assume that the coefficient $a$ is piecewise $C^1$, which means that $\alpha \in C^1(\overline{\Omega})$.

One can then still carry out the partial integration as in the previous section, but there will appear an additional term on the boundary $\Gamma = \partial\Omega$:

$$\begin{aligned}
&\operatorname{div} G_k * (\alpha\nabla u)(x) \\
&\quad = -\operatorname{div} \int_\Omega \nabla_y \big(G_k(x-y)\alpha(y)\big)\, u(y)\, dy + \operatorname{div} \int_\Gamma n(y) G_k(x-y)\alpha(y)u(y)\, ds(y) \\
&\quad = -\alpha(x)u(x) - k^2 \int_\Omega G_k(x-y)\alpha(y)u(y)\, dy - \operatorname{div} \int_\Omega G_k(x-y)(\nabla\alpha)(y)\, u(y)\, dy \\
&\qquad\qquad - \int_\Gamma \partial_{n(y)} G_k(x-y)\alpha(y)u(y)\, ds(y)\,.
\end{aligned}$$

The additional term is just the Helmholtz double-layer potential with density $\alpha u$, which we can abbreviate as $\mathscr{D}\gamma(\alpha u)$. Here $\gamma : H^1(\Omega) \to H^{\frac{1}{2}}(\Gamma)$ is the trace operator. We obtain our volume integral operator in the form

$$(\mathbb{I} - A)u(x) = a(x)u(x) + A_1 u(x) + \mathscr{D}\gamma(\alpha u)(x) \tag{11.7}$$

with

$$\begin{aligned}
A_1 u(x) = -k^2 &\int_\Omega G_k(x-y)\alpha(y)u(y)\, dy \\
&+ \operatorname{div} \int_\Omega G_k(x-y)(\nabla\alpha)(y)\, u(y)\, dy - \int_\Omega G_k(x-y)\beta(y)u(y)\, dy\,.
\end{aligned}$$

The operator $A_1$ is bounded from $L^2(\Omega)$ to $H^1(\Omega)$, hence compact as an operator in $H^1(\Omega)$.

The operator $u \mapsto \mathscr{D}\gamma(\alpha u)$ is bounded in $H^1(\Omega)$ but not compact, in general. It is also not continuous with respect to the $L^2(\Omega)$-norm of $u$. This implies that the operator $\mathbb{I} - A$, despite being generated from a pseudo-differential operator of order zero, does not have a continuous extension to $L^2(\Omega)$ from the dense subspace $H^1(\Omega)$. It does have a continuous extension to $L^2(\Omega)$ from the subspace $H_0^1(\Omega)$, but this is a different operator, where the last term in (11.7) is missing.

### *11.3.1 Extension to a Boundary–Domain System*

From the VIE (11.4) with the integral operator written in the form (11.7), we can get an equation on the boundary by taking the trace on $\Gamma$:

$$a\gamma u + \gamma A_1 u + \gamma\mathscr{D}\gamma(\alpha u) = \gamma u^{\mathrm{inc}}\,. \tag{11.8}$$

We now treat the trace $\gamma u$ as if it was an additional unknown, denoted by $\phi$, and consider the two equations (11.4) and (11.8) as a coupled boundary–domain integral equation system.

Taking into account the jump relation for the double-layer potential

$$\gamma \mathscr{D}\phi = -\tfrac{1}{2}\phi + K\phi,$$

where $K$ is the Helmholtz double layer-potential operator evaluated on $\Gamma$, as well as the fact that the commutator $[K,\alpha]$ between $K$ and the multiplication by $\alpha$ is compact in the trace space $H^{\frac{1}{2}}(\Gamma)$, we can write this coupled system in the matrix form

$$\begin{pmatrix} a\mathbb{I}+A_1 & \mathscr{D}(\gamma\alpha\cdot) \\ \gamma A_1 & \frac{1}{2}(1+a)\mathbb{I}+\alpha K+[K,\alpha] \end{pmatrix} \begin{pmatrix} u \\ \phi \end{pmatrix} = \begin{pmatrix} u^{\mathrm{inc}} \\ \psi \end{pmatrix} \tag{11.9}$$

It is easy to see that this system is equivalent to the original VIE in the following sense.

**Proposition 2.** *Let $\Omega$ be a bounded Lipschitz domain with boundary $\Gamma$. Let $\alpha \in C^1(\overline{\Omega})$ and $\beta \in L^\infty(\Omega)$, and let $u^{\mathrm{inc}} \in H^1(\Omega)$ be given.*

*If $u \in H^1(\Omega)$ is a solution of the VIE (11.4), then $\begin{pmatrix} u \\ \phi \end{pmatrix} = \begin{pmatrix} u \\ \gamma u \end{pmatrix}$ solves the coupled system (11.9) with $\psi = \gamma u^{\mathrm{inc}}$.*

*Conversely, let $\psi \in H^{\frac{1}{2}}(\Gamma)$ be given and $\begin{pmatrix} u \\ \phi \end{pmatrix} \in H^1(\Omega) \times H^{\frac{1}{2}}(\Gamma)$ be a solution of the coupled system (11.9). If $\psi = \gamma u^{\mathrm{inc}}$, and if $\gamma a \neq 0$ a.e. on $\Gamma$, then $\phi = \gamma u$, and $u$ is a solution of the VIE (11.4).*

*Proof.* The construction of the coupled system shows that it is satisfied by any solution of the VIE and its trace on the boundary. To show the converse, one subtracts the trace of the first equation in (11.9) from the second and finds

$$\gamma a\,(\gamma u - \phi) = 0.$$

Since we assume that $\gamma a$ does not vanish on a set of positive measure, $\phi = \gamma u$ follows.

### *11.3.2 Lipschitz Boundary*

The system (11.9) is easier to analyze than the original VIE (11.4). This is due to the fact that now the main difficulty is pushed to the boundary integral operator $K$, which is a well-studied classical boundary integral operator [Co88]. Indeed, splitting off the operators that we already have identified as compact operators, and taking

into account that the coupling operator $\phi \mapsto \mathscr{D}(\gamma\alpha\phi)$ is bounded from $H^{\frac{1}{2}}(\Gamma)$ to $H^1(\Gamma)$ [Co88], we see that the Fredholm alternative holds for the system (11.9) (and therefore for the VIE (11.4)) if and only if the operator

$$\widehat{A} = \begin{pmatrix} a\mathbb{I} & \mathscr{D}(\gamma\alpha\cdot) \\ 0 & \frac{1}{2}(1+a)\mathbb{I} + \alpha K \end{pmatrix}$$

is a Fredholm operator of index zero in the space $H^1(\Omega) \times H^{\frac{1}{2}}(\Gamma)$. This, in turn, is the case if and only if both

$$a\mathbb{I} : H^1(\Omega) \to H^1(\Omega) \quad \text{and} \quad \frac{1}{2}(1+a)\mathbb{I} + \alpha K : H^{\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma)$$

are Fredholm of index zero. We have shown the following result.

**Theorem 3.** *Let $\Omega$ be a bounded Lipschitz domain with boundary $\Gamma$. Let $\alpha \in C^1(\overline{\Omega})$ and $\beta \in L^\infty(\Omega)$. Then for the VIE (11.3) the Fredholm alternative holds in $H^1(\Omega)$ if and only if*

(i) *$a(x) \neq 0$ in $\overline{\Omega}$ and*
(ii) *$\frac{1}{2}(1+a)\mathbb{I} + \alpha K$ is Fredholm of index zero in $H^{\frac{1}{2}}(\Gamma)$.*

Condition (ii) can be made more precise by using information about the essential spectrum of the operator $\frac{1}{2}\mathbb{I} + K$. This operator differs by a compact operator from the corresponding operator for $k = 0$, i.e. the trace of the harmonic double layer potential operator. The latter is known to be a positive self-adjoint contraction in $H^{\frac{1}{2}}(\Gamma)$ if this space is equipped with a suitable scalar product, see [Co07].

Therefore its essential spectrum, which is also the essential spectrum of the operator $\frac{1}{2}\mathbb{I} + K$, is a compact subset $\Sigma$ of the open interval $(0,1)$. It is known that for any Lipschitz boundary $\frac{1}{2} \in \Sigma$, that for smooth boundaries $\Sigma = \{\frac{1}{2}\}$, and that for polygons in $\mathbb{R}^2$, $\Sigma$ is an interval depending on the corner angles.

If the coefficient function $a$ is piecewise constant, so that $\alpha = a - 1$ is a constant on $\Gamma$, the operator $\frac{1}{2}(1+a)\mathbb{I} + \alpha K$ is either the identity if $\alpha = 0$ or a multiple of the operator $\sigma\mathbb{I} - (\frac{1}{2}\mathbb{I} + K)$ with

$$\frac{1+a}{2(1-a)} = \sigma - \frac{1}{2} \quad \Longleftrightarrow \quad a = \frac{\sigma - 1}{\sigma}. \tag{11.10}$$

It follows that the operator $\frac{1}{2}(1+a)\mathbb{I} + \alpha K$ is Fredholm of index zero if and only if $\sigma \neq \Sigma$.

If the function $\alpha$ is not constant on $\Gamma$, one can use the fact that the operator $K$ commutes modulo compact operators with multiplications by $C^1$ functions and apply standard localization procedures. The result is that if for each point $x \in \Gamma$, the number $\sigma$ from (11.10) does not belong to the essential spectrum $\Sigma$, then the operator $\frac{1}{2}(1+a)\mathbb{I} + \alpha K$ is Fredholm. This condition

$$\forall x \in \Gamma : \quad \frac{1}{1-a(x)} \notin \Sigma \tag{11.11}$$

is, in general, only a sufficient condition. In order to obtain a necessary condition, one would need a 'localized' version $\Sigma_x$ of $\Sigma$, which is only known in some cases, namely when $\Gamma$ has a suitable tangent cone at $x$.

We summarize this discussion.

**Theorem 4.** *Assume the hypotheses of Theorem 3. Let $\Sigma \subset (0,1)$ be the essential spectrum of the operator $\frac{1}{2}\mathbb{I} + K$ in $H^{\frac{1}{2}}(\Gamma)$. If the coefficient $a \in C^1(\overline{\Omega})$ is constant on $\Gamma$, then the volume integral operator $\mathbb{I} - A$ is Fredholm of index zero in $H^1(\Omega)$ if and only if*

(i) $a(x) \neq 0$ *in* $\overline{\Omega}$ *and*
(ii) $a(x) \neq \frac{\sigma-1}{\sigma}$ *for* $x \in \Gamma$, $\sigma \in \Sigma$.

*If $a$ is not constant on $\Gamma$, then the conditions* (i) *and* (ii) *imply that the volume integral operator is Fredholm in $H^1(\Omega)$.*

### *11.3.3 Smooth Boundary*

If $\Gamma$ is smooth ($C^{1+\varepsilon}$ with $\varepsilon > 0$), then the boundary integral operator $K$ has a weakly singular kernel and is compact in $H^{\frac{1}{2}}(\Gamma)$. This implies that $\Sigma = \{\frac{1}{2}\}$ in Theorem 4. But it also implies directly that the operator $\frac{1}{2}(1+a)\mathbb{I} + \alpha K$ is Fredholm of index zero if and only if $1+a$ does not vanish. We obtain immediately as a corollary of Theorem 3 the following result.

**Theorem 5.** *Let $\Omega$ be a bounded smooth (Lyapunov) domain. Let $\alpha \in C^1(\overline{\Omega})$ and $\beta \in L^\infty(\Omega)$. Then for the VIE (11.3) the Fredholm alternative holds in $H^1(\Omega)$ if and only if*

(i) $a(x) \neq 0$ *in* $\overline{\Omega}$ *and*
(ii) $a(x) \neq -1$ *on* $\Gamma$.

The conditions on the coefficient $a(x)$ obtained in Theorem 5 have been known for a long time as conditions for Fredholm properties of the scattering problem (11.1). In [CoSt85], the case of piecewise constant coefficients was treated. Using the method of boundary integral equations, the case of smooth boundaries in any dimension and the case of polygons in dimension two were studied. In the thesis [Ch12] and the paper [BBCC12], variational methods for the interface problem were used to obtain the same conditions in the case of smooth domains and also necessary and sufficient conditions for some non-smooth domains.

## References

[BBCC12] Bonnet-Ben Dhia, A.-S., Chesnel, L., Ciarlet, P. Jr.: $T$-coercivity for scalar interface problems between dielectrics and metamaterials. ESAIM Math. Model. Numer. Anal. **46**(6), 1363–1387 (2012)

[Ch12] Chesnel, L.: Investigation of some transmission problems with sign changing coefficients, Application to metamaterials. PhD thesis, École Polytechnique (2012)

[CoKr83] Colton, D., Kress, R.: Integral equation methods in scattering theory. Pure and Applied Mathematics (New York). John Wiley & Sons Inc., New York (1983)

[CoKr98] Colton, D., Kress, R.: Inverse acoustic and electromagnetic scattering theory, volume 93 of Applied Mathematical Sciences. Springer-Verlag, Berlin, second edition (1998)

[Co88] Costabel, M.: Boundary integral operators on Lipschitz domains, elementary results. SIAM J. Math. Anal. **19**(3), 613–626 (1988)

[Co07] Costabel, M.: Some historical remarks on the positivity of boundary integral operators. In: Boundary element analysis, volume 29 of Lect. Notes Appl. Comput. Mech., pp. 1–27. Springer, Berlin (2007)

[CoDK10] Costabel, M., Darrigrand, E., Koné, E.-H.: Volume and surface integral equations for electromagnetic scattering by a dielectric body. J. Comput. Appl. Math. **234**(6), 1817–1825 (2010)

[CoDS12] Costabel, M., Darrigrand, E., Sakly, H.: The essential spectrum of the volume integral operator in electromagnetic scattering by a homogeneous body. Comptes Rendus Mathématique, 350, 193–197 (2012)

[CoSt85] Costabel, M., Stephan, E.: A direct boundary integral equation method for transmission problems. J. Math. Anal. Appl. **106**(2), 367–413 (1985)

[DrFl94] Draine, B. T., Flatau, P. J.: Discrete-dipole approximation for scattering calculations. J. Opt. Soc. Am. A **11**(4), 1491–1499 (1994)

[FrPa84] Friedman, M. J., Pasciak, J. E.: Spectral properties for the magnetization integral operator. Math. Comp. **43**(168), 447–453 (1984)

[Ki07] Kirsch, A.: An integral equation approach and the interior transmission problem for Maxwell's equations. Inverse Probl. Imaging **1**(1), 159–179 (2007)

[KiLe09] Kirsch, A., Lechleiter, A.: The operator equations of Lippmann-Schwinger type for acoustic and electromagnetic scattering problems in $L^2$. Appl. Anal. **88**(6), 807–830 (2009)

[PuPe73] Purcell, E. M., Pennypacker, C. R.: Scattering and adsorption of light by nonspherical dielectric grains. Astrophys. J. **186**, 705–714 (1973)

[Sa14] Sakly, H.: Opérateur intégral volumique en théorie de diffraction électromagnétique. PhD thesis, Université de Rennes 1 (2014)

[YuHo07] Yurkin, M. A., Hoekstra, A. G.: The discrete dipole approximation, an overview and recent developments. J. Quant. Spectrosc. Radiat. Transf. **106**(1), 558–589 (2007)

# Chapter 12
# Modeling and Implementation of Demand Dispatch Approach in a Smart Micro-Grid

**F.D. Farimani and H.R. Mashhadi**

## 12.1 Introduction

### *12.1.1 Motivation*

Since today power distribution systems have experienced fundamental changes in recent decades, management mode of the system should also be impressed by the changes. Development of control and communication systems along with new concept of smart grid enables demand side assets to participate in dispatching process. This paper incentive is to precisely model DD, define the DD Aggregator problem, and finally implement the model on a real case study.

### *12.1.2 Literature Review*

Demand dispatch was firstly introduced in [BrEtAl10], as a new way of thinking about demand response (DR) due to the development of communication and control (C&C) technologies in demand side. Many loads are now equipped with C&C and could receive the aggregator dispatch/control command. The article presents the required infrastructure of DD, especially communication part and necessary characteristics of the internet for a successful implementation of DD. Some of the differences between DD and DR are also explained. Moreover, the main required characteristics of the loads, which could be dispatched and remotely controlled (DLs) are presented. The aggregated loads are used for ancillary services and

F.D. Farimani • H.R. Mashhadi (✉)
Ferdowsi University of Mashhad, P.O. Box 91775-1111, Mashhad, Iran
e-mail: fateme.daburi@stu-mail.um.ac.ir; h_mashhadi@um.ac.ir



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_12

the benefits of load-based ancillary services are pointed. Finally, smart charging of PEVs is simulated as an example of demand dispatch. Botterud [BoEtAl13] employed DD combined with a powerful probabilistic wind power forecasting method (WPF). Combination of DD and WPF could increase integration of wind power into the power grid. A UC model considering WPF was developed by Wang [WaEtAl11] and its formulation was extended in [BoEtAl13], to include DD. The results show the ability of DD to handle uncertainty of wind power. DD not only results reserve improvement and less load curtailment but also provides lower wind power curtailment. Daburi [DaRa13] employed DD on an autonomous hybrid PV-wind-battery-diesel system and concluded that DD could reduce the capacity of required backup of battery energy storage and diesel units due to their lower dispatch commands. Berardino [BeNwMi11] presented a method for economic dispatch of some buildings along with DR purposes. In this paper, a generic formulation for DD problem from the perspective of end user is presented and thus topology and constraints of distribution system are not considered. Using DD for smart charging of PEVs is performed in [WuEtAl12]. Stochastic charging of PEVs is an important issue in distribution systems, which causes negative impacts on the grid. This issue is addressed by [WuEtAl12] using 3 smart charging patterns. Daburi [DaRa13] presented a priority list algorithm for the aggregator to implement DD on 900 DLs with the aim of wind generation following by the DLs. Unfortunately, correlation of wind power production and residential power consumption is usually low due to higher wind speed and wind–load correlation and reduces the total operation cost. DOE/NETL [Do11] prepared a comprehensive report of DD approach covering DD definition and characteristics, comparison of DD with SD, its benefits and implementation barriers. Current state of DD is also presented. One of the practical projects on DD is the project in New Brunswick, Canada, which was reported by Power Shift Atlantic. The goal of the project is to balance the variable energy of wind turbines with the residential loads. It is the first project in the world, which uses aggregated loads for integration of wind power into the grid. In this project more than 1000 homes were monitored. One of the benefits of this approach is that replaces the need of supporting wind power with costly generation systems. Actually the existing assets, here dispatchable loads are used in spite of expensive conventional power plants.

### *12.1.3 Chapter Content*

The rest of this chapter is organized as follows. Section 12.2 generally explains DD modeling. Firstly, a definition for DD is provided, and then a structure for dispatching process is suggested to define the DDA identity and his relations with the structure. After that, DDA and end user (EU) roles are modeled. This section finally presents a structure for market relations between DDA, EUs, and MO. In Section 12.3, the problem of aggregator is formulated. DDA problem is first described in part A. The objective function and constraints of the optimization

problem are presented in the next parts. Input data of the problem, DDA commands to the DLs and the results will be analyzed. Asset optimization, as a benefit of DD, is also proofed in the simulations. Finally, the paper is summarized and concluded in Section 12.4. Recommendations for future works are also included in this section.

## 12.2 Demand Dispatch Modeling

### *12.2.1 DD Definition*

Demand dispatch refers to remotely, dynamically and real time control of demand (especially DLs) by the grid operator through whole day to help balance generation and load. Dispatchable loads provide a more flexible system operation. About 33 percent of all loads are estimated to be dispatchable [BrEtAl10]. Section 12.3 will explain this flexibility in detail. DD is a complement approach to SD for more effective operation optimization. The conventional economic dispatch, actually supply dispatch (SD), matches with the conventional centralized power plants while DD is more compatible with decentralized generations and especially variable renewable distributed generations since it is based on the strategy of generation following (GF) as opposed to SD, which uses load following (LF) paradigm. A perfect comparison between DD and SD is presented in [DaRa13]. Some of the main differences between DD and DR are mentioned in [BrEtAl10] and [DaRa13].

### *12.2.2 Dispatching Structure*

Dispatching process is considered to have some levels from the first level dispatcher, which is the independent system operator, to the fourth level one, shown in Figure 12.1. Number of dispatchers is normally more than presented in the structure. Aggregators at the end of the structure are assumed to send their optimum control commands to the related level-4 dispatcher, which is responsible for remote controlling of DLs. Every Level-4 dispatcher receives the control commands of his own aggregators as illustrated in Figure 12.2. The grid scale within a level 4 dispatcher scope is about 3 percent of a city electric network. Level-3 ones, which are distribution dispatchers, receive level-4 dispatchers information and will transmit the information to their own level-2 dispatcher. Level-2 dispatchers are the regional dispatchers under level 1 dispatcher. Therefore, there is a down-up data transfer direction in DD, since the load follows the generation pattern. We call this approach generation following (GF). In the conventional dispatching approach,

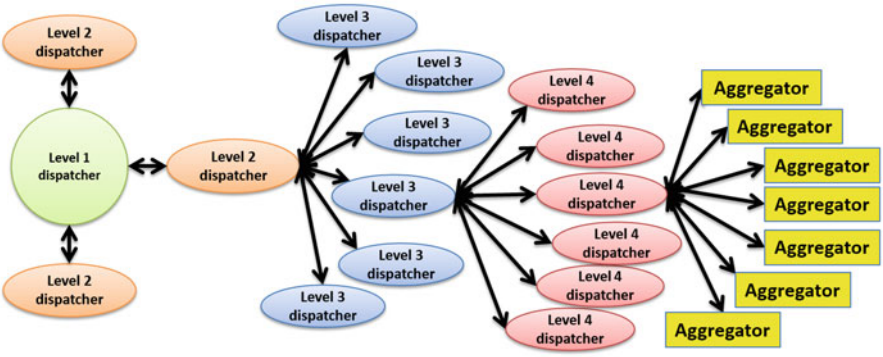


**Fig. 12.1** Dispatching structure with the aggregators at the end.

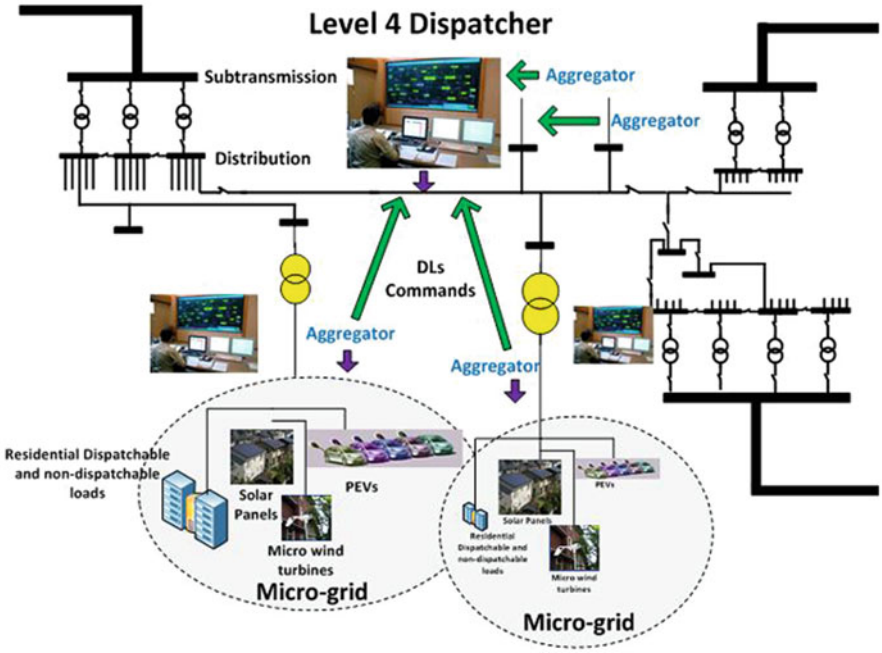


**Fig. 12.2** Aggregators and level 4 dispatchers relations.

which we call it load following (LF), generation follows the load pattern. In LF, the load profile is first forecasted and the generation units dynamically follow it through unit commitment and economic dispatch methods.

### 12.2.3 Demand Dispatch Aggregator Modeling

The DD aggregator as mentioned before is a small system operator that is the closest dispatcher to the residential customers. He receives the information of DLs from the end users, and solves an optimization problem to determine the best commands to the DLs. As depicted in Figure 12.2, every DDA reports the information of commands to the level 4 dispatcher, which is responsible for connecting the DLs in the right time. It could be possible for level-4 dispatcher to solve a new optimization problem based on the commands received from the aggregators. In this way, every micro-grid could play the role of a virtual power plant (VPP) to be considered in the higher level dispatcher optimization.

### 12.2.4 End-User Modeling

End-users play an important role in DD implementation. They are expected to set some information on their appliances after plugging it. We speak about dispatchable loads, which are flexible in time of performance like dishwashers, EVs, washing machines, clothes dryers, pool pumps, and so on. When it is possible for the customer to wait more than the required time for his/her appliance to accomplish its work (for example, a dishwasher needs 30 minutes but it is possible for the user to wait 8 hours for clean dishes), he/she could simply participate in DD. Assuming that the appliances are equipped with communication technology needed, the user will be able to set some required information on the appliance. The user determines his/her waiting period by setting start time (is) and the end time (ie); see Table 12.1. In order to provide a more flexible time scheduling, the day is divided into 144 time intervals. So every time interval lasts 10 minutes. The user might plug the appliance at 7:00 for example, and selects the waiting period between 10:00 to 18:30 (equals with 42 to 111). The end user should also specify the number of intervals needed for the appliance, which is shown in column 4 of Table 12.1. It is assumed that the appliance energy consumption characteristics are known by the aggregator. Since the grid is smart, the aggregator is able to assign an IP for every DL connected to the smart plugs and easily obtain the energy usage of the appliance. The end users expect the aggregator to supply their loads exactly in the determined waiting period. Assume 4 types of DLs like clothes dryer, dishwasher, washing machine, and electric vehicle to participate in DD with energy characteristics of Table 12.2. We will discuss on the number of each DL type participated in DD for the scale of micro-grid in the next section.

**Table 12.1** Dispatchable loads information.

| DL type | $i_s$ (1-144) | $i_e$ (1-144) | Time intervals | Interval energy (kWh) | No. of DLs |
|---|---|---|---|---|---|
| `Type1` | $i_{s,1}$ | $i_{e,1}$ | 6 | 0.650 | 10 |
| `Type2` | $i_{s,2}$ | $i_{e,2}$ | 3 | 0.200 | 25 |
| `Type3` | $i_{s,3}$ | $i_{e,3}$ | 4 | 0.470 | 15 |
| `Type4` | $i_{s,4}$ | $i_{e,4}$ | 18 | 0.544 | 10 |

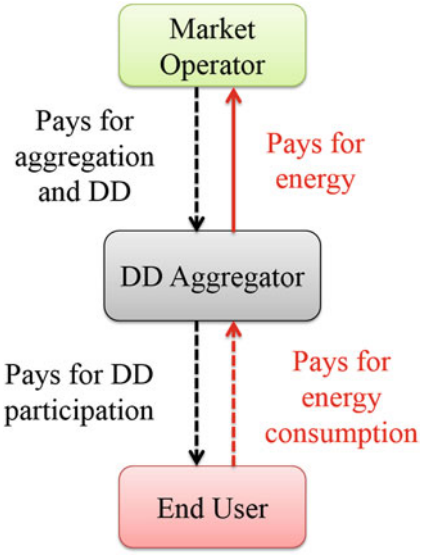


**Fig. 12.3** DD Aggregator, end user, and market operator relations.

### *12.2.5 DDA, EU, and Market Operator*

According to Figure 12.3, end-users' contract with DDA for their energy supply and demand dispatch participation. They pay for their energy consumption to DDA and receive a reward from DDA due to DD participation. It could be also assumed they are rewarded by lower energy payments. DDA should pay for energy to the market operator. MO pays DDA for his DD implementation and load aggregation. DDA tries to minimize his/her cost, which consists of pay for energy to MO and pay for DD participation to EUs.

## 12.3 Problem Formulation

### *12.3.1 What is the DDA Problem?*

The aggregator is responsible for aggregation of candidate DLs to minimize the operation cost. As mentioned before, from the electricity market perspective, DDA buys energy from the MO and supplies the EUs (see Figure 12.3). Final goal is to daily schedule the DLs to minimize operation cost. The micro-grid consists of 40 houses. For the considered 40 houses, about 10 dryers, 25 dishwashers, 15 washing machines, and 5 to 20 EVs could be candidate DLs for an especial day.

### *12.3.2 Objective Function*

DDA's goal is to minimize the operation cost for a period of 24 hours. The mathematical formulation could be described as Equation 12.1.

The first term refers to the energy cost paid by the DDA to the MO. The second term describes the cost of not supplying the load due to overloading of the transformer. The power flow is illustrated in Figure 12.4.

$$Minimize \quad \sum_{i=1}^{144} E_{grid,i}\rho_i + E_{OL,i}\rho_{NS,i} \tag{12.1}$$

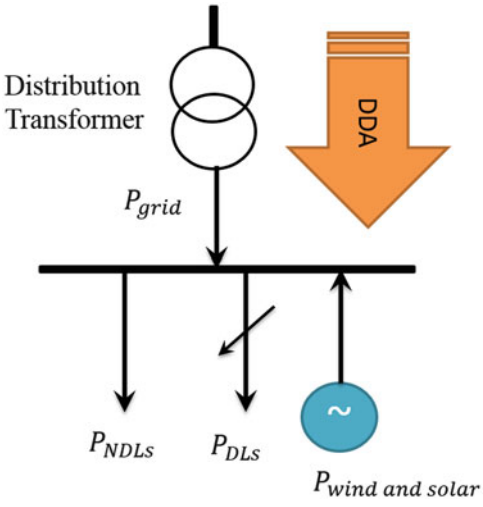


**Fig. 12.4** Power flows of the micro-grid.

### 12.3.3 Constraints

DDA problem constraints include network limitations and DL constraints.

The transformer loading limitation is considered by Equation 12.2. Generation and load balance is formulated as Equation 12.3.

$$P_{grid,i} \leq P_{max} \tag{12.2}$$

$$P_{grid,i} - P_{NDL,i} + P_{Wind,i} + P_{PV,i} - (\sum_{dl=1}^{dl_{max}} E_{dl} S_{dl,i})/T) + E_{OL,i}/T = 0 \tag{12.3}$$

$$\sum_{i_s}^{i_e} S_{dl,i} = N_{dl,i} \qquad \forall dl \tag{12.4}$$

The operation time of the dispatchable loads should lie within the allowable period of that load and equal with the up time needed as formulated in Equation 12.4.

$S_{dl,i}$ denotes the status of the DLs, which is the result of DD. It is a vector containing 144 binary elements. As an example, assume that 30 minutes is needed for a dishwasher to accomplish its work. Assume that the user period is from 37 to 131 (from 07:00 to 21:50). There must be 3 ones ($3 * 10 = 30 min$) in the status vector:

$$S_{dl,i} = [0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0 \quad ,..., \quad 0 \quad 0 \quad 0]$$

For continuous operation of the DLs, the operation period should be uninterrupted. So, one-valued elements should appear sequential as in Equation 12.5. Equation 12.6 represents the relationship between the DLs status and the indicators of startup and shut down of the loads. In order to avoid the appliance to simultaneously receive on and off commands, inequality Equation 12.7 is considered in the constraints. Since the loads are assumed to participate in DD one time a day, they should receive on and off commands, just one time as formulated in Equation 12.8 and Equation 12.9.

$$\sum_{I=i}^{i+N_{dl}-1} S_{dl,I} \geq N_{dl} M dl, i \qquad \forall dl, \forall i \leq 144 - N_{r,dl} + 1 \tag{12.5}$$

$$M_{dl,i} - K_{dl,i} = S_{dl,i} - S_{dl,i-1} \qquad \forall dl, \forall i \tag{12.6}$$

$$M_{dl,i} + K_{dl,i} \leq 1 \qquad \forall dl, \forall i \tag{12.7}$$

$$\sum_{i=1}^{144} M_{dl,i} = 1 \qquad \forall dl \tag{12.8}$$

$$\sum_{i=1}^{144} K_{dl,i} = 1 \qquad \forall dl \tag{12.9}$$

For a better comparison between the conventional approach and a smart grid system using DD, the micro-grid system is also simulated without using DD, when all the loads are non-dispatchable in three scenarios for different time of use of the DLs. Assume the micro-grid shown in Figure 12.5. The PEVs consume energy of 9.8 kWh in 3 hours (18 time intervals) to get full-charged. The energy characteristics of the other three dispatchable loads are given in Table 12.1. In the input DLs information matrix, all of the DLs are assumed to have the same energy characteristics, but with different waiting periods. A 50 Kw wind turbine with the characteristics given in Table 12.2 is assumed for the micro-grid. Data of wind speed in the first of October 2008 in Khaf were converted to wind power using Equation 12.10.

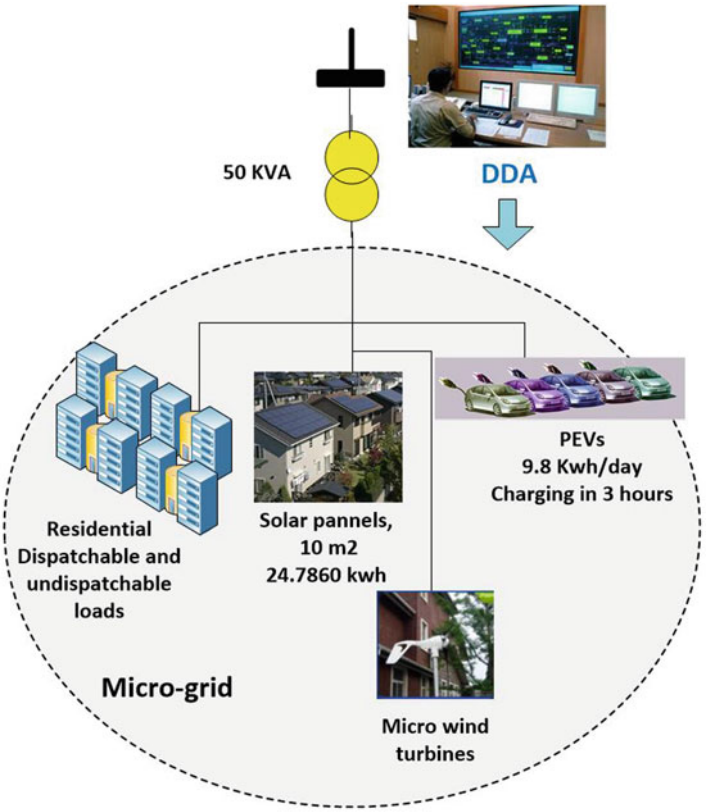


**Fig. 12.5** The micro-grid case studies.

**Table 12.2** Wind Turbine Characteristics.

| | |
|---|---|
| $ins_{cap}$ | 50 kW |
| $ci_s$ | 3.5 m/s |
| $co_s$ | 25 m/s |
| $r_s$ | 14 m/s |

$$P_W(K) = \begin{cases} 0 & \text{if } s_w(k) \geq co_s \quad or \quad s_w(k) \leq ci_s \\ ins_{cap} & \text{if } s_w(k) \geq r_s \quad or \quad s_w(k) \leq co_s \\ \frac{ins_{cap}}{r_s - ci_s} S(k) co_s & \text{if } s_w(k) \geq ci_s \quad or \quad s_w(k) \leq r_s \end{cases} \tag{12.10}$$

### *12.3.4 Case 1: Micro-Grid Operation Without DD*

- Scenario 1: the customers plug their DLs based on their habits without considering the energy price.
- Scenario 2: the customers use their dispatchable appliances in low price intervals of daytime when they are awake.
- Scenario 3: the customers plug in their DLs in low-price intervals even in the midnight.

### *12.3.5 Case 2: Micro-Grid Operation Using DD*

The aggregator should determine the start time of the DLs based on the information received from the EUs. DLs status is shown in Figure 12.6. The first 10 loads are clothes dryers with 60 minutes required time. The next 25 DLs are the dishwashers with 30 minutes duration required. The next 15 DLs are the washing machines with 50 minutes required time and the last 10 loads are the EVs with 18 time intervals for charging. According to the energy price curve in simulation results, the DLs are charged in low cost intervals except the ones which the related waiting periods of them were settled in higher cost intervals by the customer. As shown in Table 12.3, operation cost of the DDA reduces to 276.65 dollars/day when using DD and the dispatchable loads are dispatched by the aggregator. Moreover, the transformer overloading is reduced to zero. One other important result of using DD is Wind-load correlation increment. In the first 6 hours of the day, we have a high amount of wind power. Without using DD, the surplus wind energy is injected to the main grid due to the lack of loads while it is locally consumed by the micro-grid DLs when using DD. In this way, the energy is locally used by the loads and results less

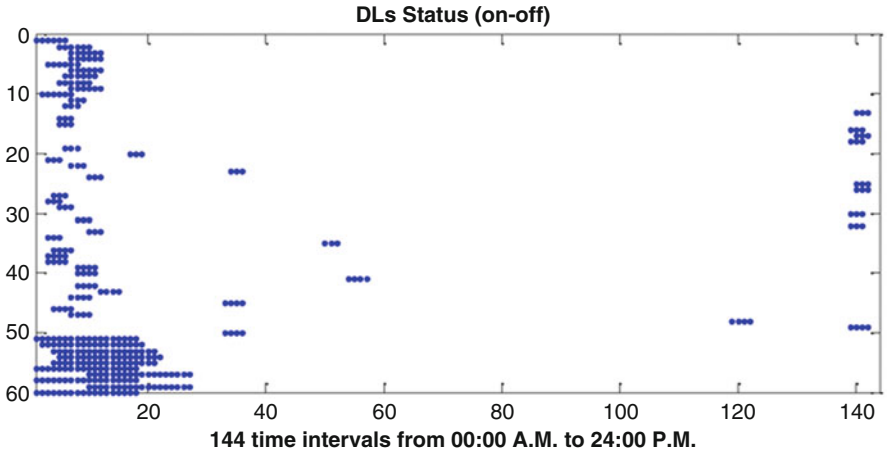


**Fig. 12.6** Dispatchable Loads charging period over the time.

**Table 12.3** Comparison between conventional generation and DD.

| | Micro-grid-DDA operation cost (dollars/day) | | Operation cost reduction (as percentage) |
|---|---|---|---|
| NDL Scenarios | Without DD | Using DD | |
| Scenario 1 | 368.86 | 276.65 | 24.99 |
| Scenario 2 | 338.22 | | 18.20 |
| Scenario 3 | 289.29 | | 4.37 |

power loss in comparison with transferring energy out of the micro-grid. DD could reduce the installation capacity of the transformer. To see how DD helps in asset optimization, especially in distribution transformer loading, operation cost versus different capacities of the transformer for both case studies of no DD and using DD is depicted in Figure 12.7. The horizontal axis represents the nominal capacity of the distribution transformer in kVA from 10 kVA to 70 kVA. The vertical axis is the DDA operation cost. As illustrated in Figure 12.7, minimum operation cost without DD approach occurs with a 55 kVA transformer, while a 40 kVA transformer is enough when using DD. In order to obtain a valid installation capacity of the transformer and speak on planning issues, operation cost of one year load profile, at least, should be drawn versus different transformer capacities.

## 12.4 Conclusions

In this paper, DD approach was modeled as an optimization problem, from the perspective of the aggregator of a micro-grid connected to a distribution transformer. Simulation results obtained from the case studies revealed that Demand Dispatch would lessen the operation cost, optimize asset utilization, reduce overloading of transformers and increase wind–load correlation.

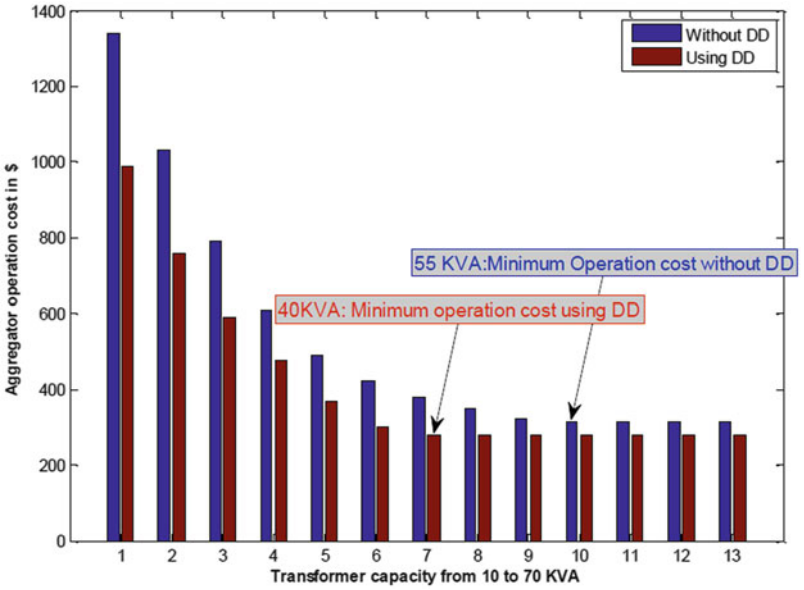


**Fig. 12.7** Operation cost for different transformer capacities from 10 to 70 kVA.

## Nomenclature

| | |
|---|---|
| $i$ | Time interval indicator from 1 to 144 |
| $dl$ | The indicator of DLs from 1 to dlmax (60 for the case study) |
| $t$ | Time interval duration equals to 1/6 hour |
| $E_{grid,i}$ | energy exchanged between the main grid and the micro-grid |
| $\rho_i$ | The energy price in ith time interval (cent/kWh) |
| $E_{OL}$,i | The over-loaded energy of the transformer during ith time interval |
| $\rho_{NS,i}$ | Energy not supplied cost in dollars |
| $P_{max}$ | Maximum power limitation of transformer |
| $P_{NDL,i}$ | Non-dispatchable loads power during ith time interval |
| $P_{wind,i}$ | Produced wind power during ith time interval |
| $P_{pv,i}$ | Produced photovoltaic power during ith time interval |
| $E_{dl}$ | Required energy of a dispatchable load for a 10-minute period |
| $S_{dl,i}$ | binary state of the DLs in every time interval |
| $N_{dl}$ | The up time needed for the dlth DL |
| $M_{dl,i}$ | A binary indicator for the startup period of dlth DL |
| $K_{dl,i}$ | A binary indicator for the shutdown period of dlth DL |

## References

[BrEtAl10] Brooks, A., Lu, E., Reicher,D., Spirakis, C., and Weihl, B.: *Demand Dispatch.* IEEE Power EnergyMag, 8, no. 3, 20–29, 2010.

[BoEtAl13] Botterud, A., Zhi, Z., Jianhui, W., Sumaili, J., Keko, H, Mendes, J., Bessa, R.J., and Miranda, V.,: *Demand Dispatch and Probabilistic Wind Power Forecasting in Unit Commitment and Economic Dispatch: A Case Study of Illinois.* IEEE Transactions on Sustainable Energy, 4, no. 1, 250–261, 2013.

[WaEtAl11] Wang, J., Botterud, A., Bessa, R., Keko, H., Carvalho, L, Issicaba, D., Sumaili, J., and Miranda, V.: *Wind power forecasting uncertainty and unit commitment.* Appl. Energy, 88, 4014–4023, 2011.

[DaRa13] Daburi Farimani, F. and Rajabi Mashhadi, H.: *Effects of demand dispatch on operation of smart hybrid energy systems.* Power System conference (PSC), Tehran, Iran, 2013.

[BeNwMi11] Berardino, J., Nwankpa, C., and Miu, K.: *Economic Demand Dispatch of Controllable Building Electric Loads for Demand Response.* Proceedings of the IEEE PowerTech 2011 Conference, Trondheim, Norway, 2011.

[WuEtAl12] Wu, T., Wu, G., Bao, Z., Yang, Q., Yan, W., and Pen, N.: *Demand Dispatch of Smart Charging for Plug-in Electric Vehicles.* IEEE International Conference on Control Engineering and Communication Technology, 803–806, 2012.

[DaRa13] Daburi Farimani, F. and Rajabi Mashhadi, H.: *Wind Generation Following Using Demand Dispatch Via Smart Grid Platform.* Smart Grid Conference (SGC), Tehran, 2013.

[Do11] Dodrill, K.: *Demand Dispatch-Intelligent Demand for a More Efficient Grid.* U.S. Department of Energy (DOE), National Energy Technology Laboratory (NETL), 2011.

# Chapter 13
# Harmonic Functions in a Domain with a Small Hole: A Functional Analytic Approach

**M. Dalla Riva and P. Musolino**

## 13.1 Introduction

In this survey, we present some recent results obtained by the authors on the asymptotic behavior of harmonic functions in a bounded domain with a small hole. Particular attention will be paid to the case of the solutions of a Dirichlet problem for the Laplace operator in a perforated domain. We fix once for all

$$n \in \mathbb{N} \setminus \{0,1\}, \qquad \alpha \in ]0,1[.$$

Then we take two open sets $\Omega^i$ and $\Omega^o$ in the $n$-dimensional Euclidean space $\mathbb{R}^n$. We assume that $\Omega^i$ and $\Omega^o$ satisfy the following condition.

$$\begin{aligned} &\Omega^i \text{ and } \Omega^o \text{ are open bounded connected subsets of } \mathbb{R}^n \text{ of} \\ &\text{class } C^{1,\alpha} \text{ such that } \mathbb{R}^n \setminus \mathrm{cl}\Omega^i \text{ and } \mathbb{R}^n \setminus \mathrm{cl}\Omega^o \text{ are connected,} \\ &\text{and such that the origin 0 of } \mathbb{R}^n \text{ belongs both to } \Omega^i \text{ and } \Omega^o. \end{aligned} \tag{13.1}$$

Here cl denotes the closure of a set. For the definition of functions and sets of the usual Schauder classes $C^{0,\alpha}$ and $C^{1,\alpha}$, we refer, for example, to Gilbarg and Trudinger [GiTr01, §6.2]. The set $\Omega^o$ will represent the unperturbed domain where

M. Dalla Riva
Department of Mathematics, The University of Tulsa, Tulsa, OK, USA
e-mail: matteo-dallariva@utulsa.edu

P. Musolino (✉)
Department of Mathematics, University of Padova, Padova, Italy
e-mail: musolinopaolo@gmail.com



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_13

we make a hole. On the other hand, the set $\Omega^i$ will play the role of the shape of the perforation. Here, the letter '$i$' stands for 'inner domain' and the letter '$o$' stands for 'outer domain.'

We note that condition (13.1) implies that $\Omega^i$ and $\Omega^o$ have no holes and that there exists a real number $\varepsilon_0$ such that

$$\varepsilon_0 > 0 \text{ and } \varepsilon \mathrm{cl}\Omega^i \subseteq \Omega^o \text{ for all } \varepsilon \in ]-\varepsilon_0, \varepsilon_0[.$$

We are now in the position to define the perforated domain $\Omega(\varepsilon)$:

$$\Omega(\varepsilon) \equiv \Omega^o \setminus \varepsilon \mathrm{cl}\Omega^i \qquad \forall \varepsilon \in ]-\varepsilon_0, \varepsilon_0[\,.$$

In other words, the set $\Omega(\varepsilon)$ is obtained by removing from $\Omega^o$ the closure of the set $\varepsilon\Omega^i$, which can be seen as a hole.

If $\varepsilon \in ]-\varepsilon_0, \varepsilon_0[\setminus\{0\}$, then by a simple topological argument one can see that $\Omega(\varepsilon)$ is an open bounded connected subset of $\mathbb{R}^n$ of class $C^{1,\alpha}$. For each $\varepsilon \in ]-\varepsilon_0, \varepsilon_0[$, the boundary $\partial\Omega(\varepsilon)$ of $\Omega(\varepsilon)$ consists of the two connected components $\partial\Omega^o$ and $\varepsilon\partial\Omega^i$. In particular, $\partial\Omega^o$ is the 'outer boundary' of $\Omega(\varepsilon)$ and $\varepsilon\partial\Omega^i$ is the 'inner boundary.' We also note that $\Omega(0) = \Omega^o \setminus \{0\}$.

For each $\varepsilon \in ]-\varepsilon_0, \varepsilon_0[\setminus\{0\}$ we want to consider a Dirichlet problem for the Laplace operator in the perforated domain $\Omega(\varepsilon)$. In order to do so, we fix two functions $f^i \in C^{1,\alpha}(\partial\Omega^i)$ and $f^o \in C^{1,\alpha}(\partial\Omega^o)$, and we define the Dirichlet datum $f_\varepsilon \in C^{1,\alpha}(\partial\Omega(\varepsilon))$ as follows:

$$f_\varepsilon(x) \equiv \begin{cases} f^i(x/\varepsilon) & \text{if } x \in \varepsilon\partial\Omega^i\,, \\ f^o(x) & \text{if } x \in \partial\Omega^o\,. \end{cases}$$

Then for each $\varepsilon \in ]-\varepsilon_0, \varepsilon_0[\setminus\{0\}$ we consider the following boundary value problem

$$\begin{cases} \Delta u = 0 & \text{in } \Omega(\varepsilon)\,, \\ u(x) = f_\varepsilon(x) & \text{for } x \in \partial\Omega(\varepsilon)\,. \end{cases} \tag{13.2}$$

As is well known, the problem in (13.2) has a unique solution in $C^{1,\alpha}(\mathrm{cl}\Omega(\varepsilon))$, and we denote such solution by $u_\varepsilon$. Our aim is to investigate the behavior of the solution $u_\varepsilon$ of (13.2) as $\varepsilon$ tends to 0. We observe that problem (13.2) is clearly singular when $\varepsilon = 0$. Indeed, the domain $\Omega(\varepsilon)$ is degenerate for $\varepsilon = 0$ and also the Dirichlet datum on $\varepsilon\partial\Omega^i$ does not make sense for $\varepsilon = 0$.

Therefore, in order to study the behavior of $u_\varepsilon$, we can fix a point which belongs to $\Omega(\varepsilon)$ for all $\varepsilon$ that are close to 0, and see what happens to the value of the solution $u_\varepsilon$ at this fixed point as $\varepsilon$ approaches 0. Also, we can choose to approach to the degenerate value $\varepsilon = 0$, for example, from positive values of $\varepsilon$. So we assume that

$$p \in \Omega^o \setminus \{0\}, \tag{13.3}$$

and that

$$\varepsilon_p \in ]0, \varepsilon_0[ \text{ is such that } p \in \Omega(\varepsilon) \text{ for all } \varepsilon \in ]0, \varepsilon_p[. \tag{13.4}$$

We note that (13.4) implies that the point $p$ belongs to the domain of the function $u_\varepsilon$ for all $\varepsilon \in ]0, \varepsilon_p[$, and therefore it makes sense to consider the value $u_\varepsilon(p)$ for $\varepsilon \in ]0, \varepsilon_p[$. Thus we can ask the following question.

$$\begin{aligned} &\text{What can be said of the map from } ]0, \varepsilon_p[ \text{ to } \mathbb{R} \\ &\text{which takes } \varepsilon \text{ to } u_\varepsilon(p) \text{ when } \varepsilon \text{ is close to } 0? \end{aligned} \tag{13.5}$$

The behavior of the solutions of boundary value problems in domains with small holes has been investigated, for example, with methods of asymptotic analysis. With such approach, one would try to answer to the question in (13.5) by producing an asymptotic expansion of $u_\varepsilon(p)$ for $\varepsilon$ close to 0. It is impossible to provide a complete list of all the contributions with this method. As an example, we mention the works by Bonnaillie–Noël and Dambrine [BoDa13], Il'in [Il92], Maz'ya, Movchan, and Nieves [MaMoNi13], Maz'ya, Nazarov, and Plamenevskij [MaNaPl00a, MaNaPl00b]. Moreover, the study of problems of this type has revealed to be a powerful tool in the frame of shape optimization (cf. Novotny and Sokołowsky [NoSo13]). Applications of these investigations, for example, to inverse problems are widely illustrated in Ammari and Kang [AmKa07] and Ammari, Kang, and Lee [AmKaLe09].

Here instead we wish to characterize the behavior of $u_\varepsilon$ at $\varepsilon = 0$ by a different approach. For example, we would try to represent $u_\varepsilon(p)$ for $\varepsilon > 0$ in terms of real analytic functions of the variable $\varepsilon$ defined on a whole neighborhood of 0, and of possibly singular at $\varepsilon = 0$ but explicitly known functions of $\varepsilon$ (such as $\log \varepsilon$, $\varepsilon^{-1}$, *etc*.). Then, if we knew, for example, that $u_\varepsilon(p)$ equals for positive values of $\varepsilon$ a real analytic function of the variable $\varepsilon$ defined on a whole neighborhood of 0, we would be able to deduce the existence of $\varepsilon' \in ]0, \varepsilon_p[$ and of a sequence $\{c_j\}_{j=0}^{\infty}$ of real numbers such that

$$u_\varepsilon(p) = \sum_{j=0}^{\infty} c_j \varepsilon^j \qquad \forall \varepsilon \in ]0, \varepsilon'[,$$

where the series in the right-hand side converges absolutely on $]-\varepsilon', \varepsilon'[$. As we shall see, this is indeed the case when $n \geq 3$ (cf. Theorem 1 below).

This method has been applied to investigate perturbation problems for the conformal representation and for boundary value problems for the Laplace operator in a bounded domain with a small hole (cf., *e.g.*, Lanza de Cristoforis [La02, La08]). Later on, the approach has been extended to nonlinear traction problems in elastostatics (cf., *e.g.*, [DaLa11]), to the Stokes flow (cf., *e.g.*, [Da13]), and to the case of an infinite periodically perforated domain (cf., *e.g.*, [LaMu14]). Moreover, the authors have analyzed the effective properties of dilute composite materials by

this technique (see [DaMu13]). Finally, also (regular) domain perturbation problems in spectral theory have been analyzed with this approach (cf., *e.g.*, Buoso and Lamberti [BuLa13], Lamberti and Provenzano [LaPr13]).

## 13.2 What Happens When $\varepsilon$ is Positive and Close to 0?

In the following theorem, we answer question (13.5) on the behavior of $u_\varepsilon(p)$ as $\varepsilon \to 0^+$, by exploiting the functional analytic approach proposed by Lanza de Cristoforis. We find convenient to denote by $u_0$ the unique function in $C^{1,\alpha}(\mathrm{cl}\Omega^o)$ such that

$$\begin{cases} \Delta u_0 = 0 & \text{in } \Omega^o, \\ u_0(x) = f^o(x) & \text{for } x \in \partial\Omega^o. \end{cases} \tag{13.6}$$

**Theorem 1 (Lanza de Cristoforis [La08]).** *Let $p$ be as in (13.3).*

(i) *If $n = 2$, then there exist $\varepsilon_p$ as in (13.4), $\varepsilon_p < 1$, and a real analytic function $U_p^\#$ from $]-\varepsilon_p, \varepsilon_p[\setminus]1/\log\varepsilon_p, -1/\log\varepsilon_p[$ to $\mathbb{R}$ such that*

$$u_\varepsilon(p) = U_p^\#[\varepsilon, 1/\log\varepsilon] \quad \forall \varepsilon \in ]0, \varepsilon_p[,$$

*and that $u_0(p) = U_p^\#[0,0]$.*

(ii) *If $n \geq 3$, then there exist $\varepsilon_p$ as in (13.4) and a real analytic function $U_p$ from $]-\varepsilon_p, \varepsilon_p[$ to $\mathbb{R}$ such that*

$$u_\varepsilon(p) = U_p[\varepsilon] \quad \forall \varepsilon \in ]0, \varepsilon_p[,$$

*and that $u_0(p) = U_p[0]$*

Now, instead of considering the behavior of the value of $u_\varepsilon$ at a fixed point, as done in Theorem 1, we could consider the restriction of $u_\varepsilon$ to the closure of a suitable open subset of $\Omega^o \setminus \{0\}$. More precisely, we note that if

$$\tilde{\Omega} \text{ is a bounded open subset of } \Omega^o \text{ such that } 0 \notin \mathrm{cl}\tilde{\Omega}, \tag{13.7}$$

then there exists $\varepsilon_{\tilde{\Omega}}$ such that

$$\varepsilon_{\tilde{\Omega}} \in ]0, \varepsilon_0[ \text{ and } \mathrm{cl}\tilde{\Omega} \cap \varepsilon \mathrm{cl}\Omega^i = \emptyset \text{ for all } \varepsilon \in ]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[. \tag{13.8}$$

In particular, $\mathrm{cl}\tilde{\Omega} \subseteq \mathrm{cl}\Omega(\varepsilon)$ for all $\varepsilon \in ]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[$. As consequence, if $\varepsilon \in ]0, \varepsilon_{\tilde{\Omega}}[$, then it makes sense to consider the restriction of $u_\varepsilon$ to $\mathrm{cl}\tilde{\Omega}$. Then we describe the behavior of $u_{\varepsilon|\mathrm{cl}\tilde{\Omega}}$ in the following.

**Theorem 2 (Lanza de Cristoforis [La08]).** *Let $\tilde{\Omega}$ be as in (13.7).*

(i) *If $n = 2$, then there exist $\varepsilon_{\tilde{\Omega}}$ as in (13.8), $\varepsilon_{\tilde{\Omega}} < 1$, and a real analytic map $U^{\#}_{\tilde{\Omega}}$ from $]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[\times]1/\log\varepsilon_{\tilde{\Omega}}, -1/\log\varepsilon_{\tilde{\Omega}}[$ to $C^{1,\alpha}(\mathrm{cl}\tilde{\Omega})$ such that*

$$u_\varepsilon(x) = U^{\#}_{\tilde{\Omega}}[\varepsilon, 1/\log\varepsilon](x) \qquad \forall x \in \mathrm{cl}\tilde{\Omega}\,, \forall \varepsilon \in ]0, \varepsilon_{\tilde{\Omega}}[\,, \tag{13.9}$$

*and that $u_{0|\mathrm{cl}\tilde{\Omega}} = U^{\#}_{\tilde{\Omega}}[0,0]$.*

(ii) *If $n \geq 3$, then there exist $\varepsilon_{\tilde{\Omega}}$ as in (13.8) and a real analytic map $U_{\tilde{\Omega}}$ from $]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[$ to $C^{1,\alpha}(\mathrm{cl}\tilde{\Omega})$ such that*

$$u_\varepsilon(x) = U_{\tilde{\Omega}}[\varepsilon](x) \qquad \forall x \in \mathrm{cl}\tilde{\Omega}\,, \forall \varepsilon \in ]0, \varepsilon_{\tilde{\Omega}}[\,, \tag{13.10}$$

*and that $u_{0|\mathrm{cl}\tilde{\Omega}} = U_{\tilde{\Omega}}[0]$.*

We note that in Theorem 2 the real analytic maps $U^{\#}_{\tilde{\Omega}}$ and $U_{\tilde{\Omega}}$ have values in the Banach space $C^{1,\alpha}(\mathrm{cl}\tilde{\Omega})$. Here we just recall that if $\mathscr{X}$, $\mathscr{Y}$ are (real) Banach spaces and if $F$ is a map from an open subset $\mathscr{W}$ of $\mathscr{X}$ to $\mathscr{Y}$, then $F$ is real analytic in $\mathscr{W}$ if for every $x_0 \in \mathscr{W}$ there exist $r > 0$ and continuous symmetric $j$-linear operators $A_j$ from $\mathscr{X}^j$ to $\mathscr{Y}$ such that $\sum_{j\geq 1} \|A_j\| r^j < \infty$ and $F(x_0 + h) = F(x_0) + \sum_{j\geq 1} A_j(h, \ldots, h)$ for all $h \in \mathscr{X}$ with $\|h\|_{\mathscr{X}} \leq r$ (cf., *e.g.*, Deimling [De85, p. 150]).

Theorem 2 has been proved in Lanza de Cristoforis [La08, Theorem 5.3], where also real analyticity properties of the solution upon perturbations of $\Omega^o$ and $\Omega^i$ are considered. Furthermore, Theorem 2 could also be deduced by some more recent results concerning real analytic families of harmonic functions (cf. [DaMu12, Proposition 4.1] and [DaMu15, Theorem 3.1]).

Moreover, we observe that if $p \in \mathrm{cl}\tilde{\Omega}$, then the map which takes a function $u \in C^{1,\alpha}(\mathrm{cl}\tilde{\Omega})$ to $u(p)$ is linear and continuous (and thus real analytic). Since the composition of real analytic maps is real analytic, by Theorem 2 we deduce the validity of Theorem 1.

## 13.3 What Happens for $\varepsilon$ Negative?

Now we would like to investigate the validity of equalities (13.9) and (13.10) when $\varepsilon$ is negative. As we have seen, the behavior of $u_\varepsilon$ for $\varepsilon$ close to 0 in case $n = 2$ and in case $n \geq 3$ are different. As a consequence, we need to analyze separately these two cases.

### *13.3.1 Case of Dimension $n \geq 3$*

We now observe that both $u_\varepsilon$ and $U_{\tilde{\Omega}}[\varepsilon]$ in equality (13.10) are defined also for negative values of $\varepsilon$. However, by Theorem 2, we just know that the equality in (13.10) holds when $\varepsilon$ is small and positive. As a consequence, it is natural to ask the following question.

Does the equality in (13.10) hold also for $\varepsilon$ negative? (13.11)

In [DaMu12], it has been shown that the answer to the question in (13.11) depends on the parity of the dimension $n$.

The following theorem says that if the dimension $n$ is even and bigger than 3 (*i.e.*, $n = 4, 6, 8, \dots$), then the equality in (13.10) holds also for $\varepsilon < 0$ (cf. [DaMu12, Theorem 3.1 and Proposition 4.1]). Moreover, if $u_0$ is the solution of problem (13.6), equality (13.10) holds in a whole neighborhood of 0, and in particular also for $\varepsilon = 0$.

**Theorem 3.** *Let $n$ be even and $n \geq 3$. Let $\tilde{\Omega}$, $\varepsilon_{\tilde{\Omega}}$ be as in (13.7), (13.8), respectively. Then there exists a real analytic map $U_{\tilde{\Omega}}$ from $]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[$ to $C^{1,\alpha}(\mathrm{cl}\tilde{\Omega})$ such that*

$$u_\varepsilon(x) = U_{\tilde{\Omega}}[\varepsilon](x) \qquad \forall x \in \mathrm{cl}\tilde{\Omega}\,, \forall \varepsilon \in ]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[\,.$$

We now turn to consider case $n$ odd. As we shall see, if $n$ is odd (*i.e.*, $n = 3, 5, 7, \dots$), then the validity of the equality in (13.10) also for $\varepsilon < 0$ has to be considered as a very exceptional situation. Indeed, we have the following theorem (cf. [DaMu12, Proposition 4.3]).

**Theorem 4.** *Let $n$ be odd and $n \geq 3$. Then the following statements are equivalent.*

(i) *There exist $\tilde{\Omega}$, $\varepsilon_{\tilde{\Omega}}$ as in (13.7), (13.8), respectively, and a real analytic map $U_{\tilde{\Omega}}$ from $]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[$ to $C^{1,\alpha}(\mathrm{cl}\tilde{\Omega})$ such that*

$$u_\varepsilon(x) = U_{\tilde{\Omega}}[\varepsilon](x) \qquad \forall x \in \mathrm{cl}\tilde{\Omega}\,, \forall \varepsilon \in ]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[\setminus\{0\}\,.$$

(ii) *There exists $c \in \mathbb{R}$ such that*

$$f^i(x) = c \quad \forall x \in \partial\Omega^i\,, \qquad f^o(x) = c \quad \forall x \in \partial\Omega^o$$

*(and thus $u_\varepsilon(x) = c$ for all $x \in \mathrm{cl}\Omega(\varepsilon)$ and $\varepsilon \in ]-\varepsilon_0, \varepsilon_0[\setminus\{0\}$).*

Clearly, if statement (ii) of Theorem 4 holds and $\tilde{\Omega}$, $\varepsilon_{\tilde{\Omega}}$ are as in (13.7), (13.8), respectively, then the map $U_{\tilde{\Omega}}$ from $]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[$ to $C^{1,\alpha}(\mathrm{cl}\tilde{\Omega})$ defined by

$$U_{\tilde{\Omega}}[\varepsilon](x) = c \qquad \forall x \in \mathrm{cl}\tilde{\Omega}\,, \quad \forall \varepsilon \in ]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[\,,$$

is such that the equality in (13.10) holds for $\varepsilon \in ]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[\setminus\{0\}$, and therefore we deduce the validity of statement (i). On the other hand, Theorem 4 says in particular that if there exists at least one open subset $\tilde{\Omega}$ as in (13.7) for which we can find a small positive number $\varepsilon_{\tilde{\Omega}}$ as in (13.8) and a real analytic map $U_{\tilde{\Omega}}$ from $]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[$ to $C^{1,\alpha}(\mathrm{cl}\tilde{\Omega})$ such that the equality in (13.10) holds for $\varepsilon \in ]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[\setminus\{0\}$, then we are in the very exceptional situation that $f^i$ and $f^o$ are both equal to the same constant $c \in \mathbb{R}$ (and that accordingly $u_\varepsilon = c$ on $\mathrm{cl}\Omega(\varepsilon)$ for all $\varepsilon \in ]-\varepsilon_0, \varepsilon_0[\setminus\{0\}$). Hence, if $n$ is odd, the validity of equality (13.10) also for $\varepsilon$ negative has to be considered as a very special situation which happens only in the trivial case in which the functions $u_\varepsilon$ for $\varepsilon \in ]-\varepsilon_0, \varepsilon_0[\setminus\{0\}$ are all equal to the same constant.

### 13.3.2 Case of Dimension $n = 2$

We now turn to consider the case of dimension $n = 2$. Also in this case, we would like to say something about the validity of equality (13.9) for $\varepsilon < 0$. In particular we would like to replace the pair $(\varepsilon, 1/\log\varepsilon)$, where the map $U^{\#}_{\tilde{\Omega}}$ is evaluated when $\varepsilon > 0$, by a convenient pair which makes sense also for $\varepsilon < 0$, in a way to preserve the validity of equality (13.9) for $\varepsilon$ in a whole neighborhood of 0. We do so in the following theorem (cf. [DaMu15, Theorem 3.1]).

**Theorem 5.** *Let $n = 2$. Let $\tilde{\Omega}$, $\varepsilon_{\tilde{\Omega}}$ be as in (13.7), (13.8), respectively, with $\varepsilon_{\tilde{\Omega}} < 1$. Then there exist an open neighborhood $\mathscr{U}_{\tilde{\Omega}}$ of $\{(\varepsilon, 1/\log|\varepsilon|) : \varepsilon \in ]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[\setminus\{0\}\}$ $\cup\{(0,0)\}$ in $\mathbb{R}^2$ and a real analytic map $U^{\#}_{\tilde{\Omega}}$ from $\mathscr{U}_{\tilde{\Omega}}$ to $C^{1,\alpha}(\mathrm{cl}\tilde{\Omega})$ such that*

$$u_\varepsilon(x) = U^{\#}_{\tilde{\Omega}}[\varepsilon, 1/\log|\varepsilon|](x) \qquad \forall x \in \mathrm{cl}\tilde{\Omega}\,, \forall \varepsilon \in ]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[\setminus\{0\}\,.$$

Now we would like to consider boundary data $f^o$ and $f^i$ in such a way to get rid of the logarithmic behavior of $u_\varepsilon$ for $\varepsilon$ small. In other words, we would like that the following condition (a) holds.

(a) For all $\tilde{\Omega}$, $\varepsilon_{\tilde{\Omega}}$ as in (13.7), (13.8), respectively, there exists a real analytic map $V_{\tilde{\Omega}}$ from $]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[$ to $C^{1,\alpha}(\mathrm{cl}\tilde{\Omega})$ such that

$$u_\varepsilon(x) = V_{\tilde{\Omega}}[\varepsilon] \qquad \forall x \in \mathrm{cl}\tilde{\Omega}\,, \forall \varepsilon \in ]-\varepsilon_{\tilde{\Omega}}, \varepsilon_{\tilde{\Omega}}[\,.$$

In [DaMu15, Theorem 3.6], we show that condition (a) is equivalent to the following condition (b).

(b) There exist $p$, $\varepsilon_p$ as in (13.3), (13.4), respectively, and a real analytic map $V_p$ from $]-\varepsilon_p, \varepsilon_p[$ to $\mathbb{R}$ such that $p \in \Omega(\varepsilon)$ for all $\varepsilon \in ]-\varepsilon_p, \varepsilon_p[$ and

$$u_\varepsilon(p) = V_p[\varepsilon] \qquad \forall \varepsilon \in ]0, \varepsilon_p[\,.$$

This means that either $u_\varepsilon(p)$ displays a logarithmic behavior for every point $p \in \Omega^o \setminus \{0\}$, or $u_\varepsilon(p)$ does not display a logarithmic behavior for any point $p \in \Omega^o \setminus \{0\}$. Also, by [DaMu15, Theorem 3.6] there exists a pair of functions $(\rho^o[\varepsilon], \rho^i[\varepsilon]) \in C^{0,\alpha}(\partial\Omega^o) \times C^{0,\alpha}(\partial\Omega^i)$ which depends only on $\varepsilon$, $\partial\Omega^o$, and $\partial\Omega^i$, such that (a) and (b) are equivalent to the following condition (c).

(c) It holds $\int_{\partial\Omega^o} f^o \rho^o[\varepsilon]\, d\sigma + \int_{\partial\Omega^i} f^i \rho^i[\varepsilon]\, d\sigma = 0$ for all $\varepsilon \in ]-\varepsilon_0, \varepsilon_0[$.

The advantage of condition (c) with respect to (a) and (b) is that (c) can be verified on the boundary data $f^o$ and $f^i$ and does not require the knowledge of the solution $u_\varepsilon$ of (13.2). In [DaMu15, §3], we also observe that under some convenient assumptions, condition (c) can become very explicit. For example, if $f^o$ and $f^i$ are both constant functions, then condition (c) is equivalent to the fact that $f^o$ and $f^i$ are identically equal to the same real number. If instead both $\Omega^o$ and $\Omega^i$ coincide with the unit ball $\mathbb{B}_2$ of $\mathbb{R}^2$, then condition (c) is equivalent to $\int_{\partial\mathbb{B}_2} f^o\, d\sigma = \int_{\partial\mathbb{B}_2} f^i\, d\sigma$.

## 13.4 Asymptotic Expansion of the Solution of a Dirichlet Problem in a Perforated Domain ($n = 2$)

As already mentioned, the functional approach of the authors can be used to compute asymptotic expansions for the solutions of boundary value problems in perforated domains. In particular, the results of [DaMu15] have been exploited in [DaMuRo] to prove the following expansions for the solution $u_\varepsilon$ of the Dirichlet problem (13.2) for the Laplace operator in a bounded planar domain with a small hole.

**Theorem 6.** *Let $n = 2$. Then there exist a family $\{\lambda_{M,(j,l)}\}_{(j,l)\in\mathbb{N}^2,\, l\leq j+1}$ of functions from $(\mathrm{cl}\Omega^o)\setminus\{0\}$ to $\mathbb{R}$, and a family $\{\lambda_{m,(j,l)}\}_{(j,l)\in\mathbb{N}^2,\, l\leq j+1}$ of functions from $\mathbb{R}^2\setminus\Omega^i$ to $\mathbb{R}$, and $r_0\in\mathbb{R}$ such that the following statements hold.*

(i) *Let $\Omega_M\subseteq\Omega^o$ be open and such that $0\notin\mathrm{cl}\Omega_M$. Then there exists $\varepsilon'_M\in\,]0,\varepsilon_0]\cap]0,1[$ such that $\mathrm{cl}\Omega_M\cap\varepsilon\mathrm{cl}\Omega^i=\emptyset$ for all $\varepsilon\in\,]-\varepsilon'_M,\varepsilon'_M[$ and such that*

$$u_{\varepsilon|\mathrm{cl}\Omega_M}=\sum_{j=0}^{\infty}\varepsilon^j\sum_{l=0}^{j+1}\frac{\lambda_{M,(j,l)|\mathrm{cl}\Omega_M}}{(r_0+(2\pi)^{-1}\log|\varepsilon|)^l}$$

*for all $\varepsilon\in\,]-\varepsilon'_M,\varepsilon'_M[\setminus\{0\}$. Moreover, the series*

$$\sum_{j=0}^{\infty}\varepsilon^j\sum_{l=0}^{j+1}\frac{\lambda_{M,(j,l)|\mathrm{cl}\Omega_M}\eta^l}{(r_0\eta+(2\pi)^{-1})^l}$$

*converges in $C^{1,\alpha}(\mathrm{cl}\Omega_M)$ uniformly for $(\varepsilon,\eta)$ belonging to the product of intervals $]-\varepsilon'_M,\varepsilon'_M[\times]1/\log\varepsilon'_M,-1/\log\varepsilon'_M[$.*

(ii) *Let $\Omega_m\subseteq\mathbb{R}^2\setminus\mathrm{cl}\Omega^i$ be open and bounded. Then there exists $\varepsilon'_m\in]0,\varepsilon_0]\cap]0,1[$ such that $\varepsilon\mathrm{cl}\Omega_m\subseteq\Omega^o$ for all $\varepsilon\in\,]-\varepsilon'_m,\varepsilon'_m[$ and such that*

$$u_\varepsilon(\varepsilon\cdot)_{|\mathrm{cl}\Omega_m}=\sum_{j=0}^{\infty}\varepsilon^j\sum_{l=0}^{j+1}\frac{\lambda_{m,(j,l)|\mathrm{cl}\Omega_m}}{(r_0+(2\pi)^{-1}\log|\varepsilon|)^l}$$

*for all $\varepsilon\in\,]-\varepsilon'_m,\varepsilon'_m[\setminus\{0\}$. Moreover, the series*

$$\sum_{j=0}^{\infty}\varepsilon^j\sum_{l=0}^{j+1}\frac{\lambda_{m,(j,l)|\mathrm{cl}\Omega_m}\eta^l}{(r_0\eta+(2\pi)^{-1})^l}$$

*converges in $C^{1,\alpha}(\mathrm{cl}\Omega_m)$ uniformly for $(\varepsilon,\eta)$ belonging to the product of intervals $]-\varepsilon'_m,\varepsilon'_m[\times]1/\log\varepsilon'_m,-1/\log\varepsilon'_m[$.*

Here above, we denote by $u_\varepsilon(\varepsilon\cdot)$ the rescaled function which takes $x \in \varepsilon^{-1}\mathrm{cl}\Omega(\varepsilon)$ to $u_\varepsilon(\varepsilon x)$, for all $\varepsilon \in ]-\varepsilon_0, \varepsilon_0[\setminus\{0\}$. Moreover, the letter '$M$' stands for 'macroscopic,' while the letter '$m$' stands for 'microscopic.' Indeed, in Theorem 6 (i) we analyze the behavior of $u_\varepsilon$ far from the hole and in Theorem 6 (ii) we consider the solution in proximity of the perforation.

Finally, we note that in [DaMuRo] explicit formulae for the families of functions $\{\lambda_{M,(j,l)}\}_{(j,l)\in\mathbb{N}^2,\, l\leq j+1}$ and $\{\lambda_{m,(j,l)}\}_{(j,l)\in\mathbb{N}^2,\, l\leq j+1}$ are provided.

## 13.5 Real Analytic Families of Harmonic Functions in a Bounded Domain with a Small Hole

The results of Subsections 13.3.1 and 13.3.2 can be deduced from those in the papers [DaMu12] and [DaMu15], where we introduce and study *real analytic families of harmonic functions*, which are not required to be the solutions of any particular boundary value problem. Therefore, in the present section, the functions $u_\varepsilon$ are not necessarily solutions of problem (13.2).

**Definition 1.** Let $\varepsilon_1 \in ]0, \varepsilon_0]$. We say that $\{u_\varepsilon\}_{\varepsilon\in]0,\varepsilon_1[}$ is a *right real analytic family of harmonic functions on* $\Omega(\varepsilon)$ if it satisfies the following conditions (a0)–(a2).

(a0) $u_\varepsilon \in C^{1,\alpha}(\mathrm{cl}\Omega(\varepsilon))$ and $\Delta u_\varepsilon = 0$ in $\Omega(\varepsilon)$ for all $\varepsilon \in ]0, \varepsilon_1[$.

(a1) Let $\Omega_M \subseteq \Omega^o$ be open and such that $0 \notin \mathrm{cl}\Omega_M$. Let $\varepsilon_M \in ]0, \varepsilon_1]$ be such that $\mathrm{cl}\Omega_M \cap \varepsilon\mathrm{cl}\Omega^i = \emptyset$ for all $\varepsilon \in ]-\varepsilon_M, \varepsilon_M[$. Then there exists a real analytic map $U_M$ from $]-\varepsilon_M, \varepsilon_M[$ to $C^{1,\alpha}(\mathrm{cl}\Omega_M)$ such that

$$u_{\varepsilon|\mathrm{cl}\Omega_M} = U_M[\varepsilon] \qquad \forall \varepsilon \in ]0, \varepsilon_M[\,.$$

(a2) Let $\Omega_m \subseteq \mathbb{R}^n \setminus \mathrm{cl}\Omega^i$ be open and bounded. Let $\varepsilon_m \in ]0, \varepsilon_1]$ be such that $\varepsilon\mathrm{cl}\Omega_m \subseteq \Omega^o$ for all $\varepsilon \in ]-\varepsilon_m, \varepsilon_m[$. Then there exists a real analytic map $U_m$ from $]-\varepsilon_m, \varepsilon_m[$ to $C^{1,\alpha}(\mathrm{cl}\Omega_m)$ such that

$$u_\varepsilon(\varepsilon\cdot)_{|\mathrm{cl}\Omega_m} = U_m[\varepsilon] \qquad \forall \varepsilon \in ]0, \varepsilon_m[\,.$$

**Definition 2.** Let $\varepsilon_1 \in ]0, \varepsilon_0]$. We say that $\{v_\varepsilon\}_{\varepsilon\in]-\varepsilon_1,\varepsilon_1[}$ is a *real analytic family of harmonic functions on* $\Omega(\varepsilon)$ if it satisfies the following conditions (b0)–(b2):

(b0) $v_0 \in C^{1,\alpha}(\mathrm{cl}\Omega^o)$ and $\Delta v_0 = 0$ in $\Omega^o$, $v_\varepsilon \in C^{1,\alpha}(\mathrm{cl}\Omega(\varepsilon))$ and $\Delta v_\varepsilon = 0$ in $\Omega(\varepsilon)$ for all $\varepsilon \in ]-\varepsilon_1, \varepsilon_1[\setminus\{0\}$.

(b1) Let $\Omega_M \subseteq \Omega^o$ be open and such that $0 \notin \mathrm{cl}\Omega_M$. Let $\varepsilon_M \in ]0, \varepsilon_1]$ be such that $\mathrm{cl}\Omega_M \cap \varepsilon\mathrm{cl}\Omega^i = \emptyset$ for all $\varepsilon \in ]-\varepsilon_M, \varepsilon_M[$. Then there exists a real analytic map $V_M$ from $]-\varepsilon_M, \varepsilon_M[$ to $C^{1,\alpha}(\mathrm{cl}\Omega_M)$ such that

$$v_{\varepsilon|\mathrm{cl}\Omega_M} = V_M[\varepsilon] \qquad \forall \varepsilon \in ]-\varepsilon_M, \varepsilon_M[\,.$$

(b2) Let $\Omega_m \subseteq \mathbb{R}^n \setminus \mathrm{cl}\Omega^i$ be an open and bounded subset. Let $\varepsilon_m \in ]0, \varepsilon_1]$ be such that $\varepsilon \mathrm{cl}\Omega_m \subseteq \Omega^o$ for all $\varepsilon \in ]-\varepsilon_m, \varepsilon_m[$. Then there exists a real analytic map $V_m$ from $]-\varepsilon_m, \varepsilon_m[$ to $C^{1,\alpha}(\mathrm{cl}\Omega_m)$ such that

$$v_\varepsilon(\varepsilon\,\cdot)_{|\mathrm{cl}\Omega_m} = V_m[\varepsilon] \qquad \forall \varepsilon \in ]-\varepsilon_m, \varepsilon_m[\setminus\{0\}.$$

**Definition 3.** Let $\varepsilon_1 \in ]0, \varepsilon_0]$. We say that $\{w_\varepsilon\}_{\varepsilon \in ]-\varepsilon_1, \varepsilon_1[}$ is a *real analytic family of harmonic functions on* $\Omega^o$ if it satisfies the following conditions (c0), (c1).

(c0) $w_\varepsilon \in C^{1,\alpha}(\mathrm{cl}\Omega^o)$ and $\Delta w_\varepsilon = 0$ in $\Omega^o$ for all $\varepsilon \in ]-\varepsilon_1, \varepsilon_1[$.
(c1) The map from $]-\varepsilon_1, \varepsilon_1[$ to $C^{1,\alpha}(\mathrm{cl}\Omega^o)$ which takes $\varepsilon$ to $w_\varepsilon$ is real analytic.

Then the following assertion holds (see [DaMu12] and [DaMu15]).

**Theorem 7.** (i) *If the dimension $n$ is even and $\{u_\varepsilon\}_{\varepsilon \in ]0, \varepsilon_1[}$ is a* right real analytic family of harmonic functions on $\Omega(\varepsilon)$*, then there exists a* real analytic family of harmonic functions $\{v_\varepsilon\}_{\varepsilon \in ]-\varepsilon_1, \varepsilon_1[}$ *on $\Omega(\varepsilon)$ such that $u_\varepsilon = v_\varepsilon$ for all $\varepsilon \in ]0, \varepsilon_1[$.*
(ii) *If the dimension $n$ is odd and $\{v_\varepsilon\}_{\varepsilon \in ]-\varepsilon_1, \varepsilon_1[}$ is a* real analytic family of harmonic functions on $\Omega(\varepsilon)$*, then there exists a* real analytic family of harmonic functions $\{w_\varepsilon\}_{\varepsilon \in ]-\varepsilon_1, \varepsilon_1[}$ on $\Omega^o$ *such that $v_\varepsilon = w_{\varepsilon|\mathrm{cl}\Omega(\varepsilon)}$ for all $\varepsilon \in ]-\varepsilon_1, \varepsilon_1[$.*

In particular, we note that for $n$ odd, Theorem 7 (ii) implies that for each value of $\varepsilon \in ]-\varepsilon_1, \varepsilon_1[$ the function $v_\varepsilon$ can be extended inside the hole $\varepsilon\Omega^i$ to a harmonic function defined on the whole of $\Omega^o$.

**Acknowledgements** The research of M. Dalla Riva was supported by Portuguese funds through the CIDMA—Center for Research and Development in Mathematics and Applications, and the Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia"), within project UID/MAT/04106/2013. The research of M. Dalla Riva was also supported by the Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia") with the research grant SFRH/BPD/ 64437/2009. P. Musolino is member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). The work of M. Dalla Riva and P. Musolino is also supported by "Progetto di Ateneo: Singular perturbation problems for differential operators – CPDA120171/12" of the University of Padova.

## References

[AmKa07] Ammari, H., Kang, H.: Polarization and moment tensors. With applications to inverse problems and effective medium theory. Applied Mathematical Sciences, **162**. Springer, New York (2007)
[AmKaLe09] Ammari, H., Kang, H., Lee, H.: Layer potential techniques in spectral analysis. Mathematical Surveys and Monographs, **153**. American Mathematical Society, Providence, RI (2009)
[BoDa13] Bonnaillie-Noël, V., Dambrine, M.: Interactions between moderately close circular inclusions: the Dirichlet-Laplace equation in the plane. Asymptot. Anal. **84**, 197–227 (2013)

[BuLa13] Buoso, D., Lamberti, P.D.: Eigenvalues of polyharmonic operators on variable domains. ESAIM Control Optim. Calc. Var. **19**, 1225–1235 (2013)

[De85] Deimling, K.: Nonlinear functional analysis, Springer Verlag, Berlin (1985)

[Da13] Dalla Riva, M.: Stokes flow in a singularly perturbed exterior domain. Complex Var. Elliptic Equ. **58**, 231–257 (2013)

[DaLa11] Dalla Riva, M., Lanza de Cristoforis, M.: Weakly singular and microscopically hypersingular load perturbation for a nonlinear traction boundary value problem: a functional analytic approach. Complex Anal. Oper. Theory **5**, 811–833 (2011)

[DaMuRo] Dalla Riva, M., Musolino, P., Rogosin, S.V.: Series expansions for the solution of the Dirichlet problem in a planar domain with a small hole. Asymptot. Anal. **92**, 339–361 (2015)

[DaMu12] Dalla Riva, M., Musolino, P.: Real analytic families of harmonic functions in a domain with a small hole. J. Differential Equations **252**, 6337–6355 (2012)

[DaMu13] Dalla Riva, M., Musolino, P.: A singularly perturbed nonideal transmission problem and application to the effective conductivity of a periodic composite. SIAM J. Appl. Math. **73**, 24–46 (2013)

[DaMu15] Dalla Riva, M., Musolino, P.: Real analytic families of harmonic functions in a planar domain with a small hole. J. Math. Anal. Appl. **442**, 37–55 (2015)

[GiTr01] Gilbarg, D., Trudinger, N.S.: Elliptic partial differential equations of second order. Classics in Mathematics, Springer-Verlag, Berlin (2001)

[Il92] Il'in, A.M.: Matching of asymptotic expansions of solutions of boundary value problems. Translations of Mathematical Monographs **102**, American Mathematical Society, Providence (1992)

[LaPr13] Lamberti, P.D., Provenzano, L.: Eigenvalues of polyharmonic operators on variable domains. Eurasian Math. J. **4**, 70–83 (2013)

[La02] Lanza de Cristoforis, M.: Asymptotic behaviour of the conformal representation of a Jordan domain with a small hole in Schauder spaces. Comput. Methods Funct. Theory **2**, 1–27 (2003)

[La08] Lanza de Cristoforis, M.: Asymptotic behaviour of the solutions of the Dirichlet problem for the Laplace operator in a domain with a small hole. A functional analytic approach. Analysis (Munich) **28**, 63–93 (2008)

[LaMu14] Lanza de Cristoforis, M., Musolino, P.: A quasi-linear heat transmission problem in a periodic two-phase dilute composite. A functional analytic approach. Commun. Pure Appl. Anal. **13**, 2509–2542 (2014)

[MaMoNi13] Maz'ya, V., Movchan, A., Nieves, M.: Green's kernels and meso-scale approximations in perforated domains. Lecture Notes in Mathematics **2077**, Springer, Berlin (2013)

[MaNaPl00a] Maz'ya, V., Nazarov, S., Plamenevskij, B.: Asymptotic theory of elliptic boundary value problems in singularly perturbed domains. Vol. I. Volume **111** of Operator Theory: Advances and Applications, Birkhäuser Verlag, Basel (2000)

[MaNaPl00b] Maz'ya, V., Nazarov, S., Plamenevskij, B.: Asymptotic theory of elliptic boundary value problems in singularly perturbed domains. Vol. II. Volume **112** of Operator Theory: Advances and Applications, Birkhäuser Verlag, Basel (2000)

[NoSo13] Novotny, A.A.; Sokołowski, J.: Topological derivatives in shape optimization. Interaction of Mechanics and Mathematics. Springer, Heidelberg (2013)

# Chapter 14
# Employing Eddy Diffusivities to Simulate the Contaminants Dispersion for a Shear Dominated-Stable Boundary Layer

**G.A. Degrazia, S. Maldaner, C.P. Ferreira, V.C. Silveira, U. Rizza, V.S. Moreira, and D. Buske**

## 14.1 Introduction

For the convective planetary boundary layer there is a large number of mathematical models to describe the transport and the dispersion of contaminants. Generally, the turbulent parameterizations that are utilized in such models are well known and statistical quantities as eddy diffusivities, dispersion parameters, velocity variances, and time scales are represented by a convective similarity theory originated from a physical system in a state of quasi-equilibrium. Differently, in comparison with the convective boundary layer, the number of turbulent parameterizations employed in a dispersion model for a shear dominated stable boundary layer (SBL) is quite

G.A. Degrazia • C.P. Ferreira • V.C. Silveira
Federal University of Santa Maria, Av. Roraima 1000, Santa Maria 97105-900, RS, Brazil
e-mail: gevasiodegrazia@gmail.com; cecilia.perobelliferreira@gmail.com; viliamcardoso@gmail.com

S. Maldaner (✉)
Federal University of Santa Maria, Coordenadoria Acadêmica – campus Cachoeira do Sul, Cachoeira do Sul, RS, Brazil
e-mail: silvana.maldaner@ufsm.br

U. Rizza
Institute of Atmospheric Sciences and Climate (ISAC) of the Italian National Research Council (CNR), Bologna, Italy
e-mail: u.rizza@isac.cnr.it

V.S. Moreira
Federal University of Pampa, Itaqui, RS, Brazil
e-mail: virneimoreira@gmail.com

D. Buske
Federal University of Pelotas, Pelotas, Brazil
e-mail: daniela.buske@ufpel.edu.br



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_14

reduced. One of the major problems concerning to the shear dominated SBL is the determination of its height. This particular vertical depth is a relevant quantity to describe the processes that govern the SBL development. It is important to note that the SBL height has a significant influence on the mixing properties. Furthermore, the inhomogeneous character associated with the turbulence in the SBL becomes difficult the derivation of eddy diffusivities and dispersion parameters. Nonetheless, the local similarity theory (LST) allied to the spectral Taylor statistical diffusion theory allows to construct local expressions for the turbulence parameters in a shear dominated SBL. Therefore, in the present study we employ the LST and the turbulent velocity spectra, in the Taylor statistical diffusion theory to derive eddy diffusivities for a shear dominated SBL. This new formulation is used in a bidimensional Eulerian dispersion model to simulate the observed contaminant concentrations in the classical Hanford experiment [DoHo85].

## 14.2 Derivation of Eddy Diffusivities

An expression for the eddy diffusivities in the planetary boundary layer can be written as ([Ba49, PaSi83, DeMoVi01])

$$K_\alpha = \frac{\sigma_i^2 \beta_i}{2\pi} \int_0^\infty F_i^E(n) \frac{sin(2\pi t/\beta_i)}{n} dn \tag{14.1}$$

with $\alpha = x, y, z$ and $i = u, v, w$, where $F_i^E(n)$ is the Eulerian spectrum of energy normalized by the velocity variance, $n$ is the frequency, $\beta_i$ is the ratio of the Lagrangian to the Eulerian integral time scales, $\sigma_i^2$ is the turbulent velocity variance, and $t$ is the travel time. According to [WaKo62], $\beta_i$ can be described by

$$\beta_i = \left(\frac{\pi U^2}{16\sigma_i^2}\right)^{1/2} \tag{14.2}$$

where $U$ is the mean wind speed.

The velocity spectra for a shear dominated stable boundary layer can expressed as [DeEtAl00]

$$\frac{S_i(n)}{U_*^2} = \frac{1.5 c_i z/U \phi^{2/3}}{(fm)_i^{5/3}\left(1+1.5\frac{f^{5/3}}{(fm)_i^{5/3}}\right)} \tag{14.3}$$

where $c_i = \alpha_i (0.5 \pm 0.05)(2\pi k)^{-2/3}$ with $\alpha_i = 1, 4/3$ and $4/3$ for $u, v$, and $w$ components, respectively, $U_*$ is the local friction velocity ($U_* = (1 - z/h)^{3/4} u_*$), $h$ is the turbulent SBL height, $\phi_\varepsilon$ is the dissipation rate, $f$ is the reduced frequency ($f = nz/U$), $z$ is the height above the surface, $k = 0.4$ is the von Karman constant, and $(fm)_i$ is the normalized frequency of the spectral peak described by

$$(fm)_i = (fm)_{0i}(1 + 3.7\frac{z}{\Lambda}) \tag{14.4}$$

where $(fm)_{0i}$ is the frequency of the spectral peak in the surface for neutral conditions and $\Lambda$ is the local Monin–Obukhov length:

$$\Lambda = L(1 - z/h)^{5/4}$$

where $L$ is the Monin–Obukhov length. By integrating $S_i(n)$ (equation 14.3) over the whole frequency range, one obtains the following variance [DeEtAl00]

$$\sigma_i^2 = \frac{2.32 c_i \phi^{2/3} U_*^2}{(fm)_i^{2/3}} \tag{14.5}$$

Eq. (14.1) together with the equations (14.3), (14.2), and (14.4) leads to the following parameterization for the eddy diffusivities in a shear dominated SBL:

$$\frac{K_\alpha}{u_* h} = \frac{0.07\sqrt{c_i}(1-z/h)^{3/4} z/h}{(fm)_i^{4/3}} \int_0^\infty \frac{sin[(18.24(1-z/h)^{3/4} X')(fm)_i^{2/3} \frac{h}{z} n']}{(1+n'^{5/3})n'} dn' \tag{14.6}$$

where $n = \frac{1.5z}{(fm)_i U} n'$, $X' = \frac{x u_*}{hU}$. For $\alpha = z$ in equation (14.6) result:

$$\frac{K_z}{u_* h} = \frac{0.04(1-z/h)^{3/4} z/h}{(fm)_w^{4/3}} \int_0^\infty \frac{sin[(18.24(1-z/h)^{3/4} X')(fm)_w^{2/3} \frac{h}{z} n']}{(1+n'^{5/3})n'} dn \tag{14.7}$$

## 14.3 Test of the Proposed Parameterization Employing the Hanford Observed Concentration Data

The eddy diffusivities are used in a Eulerian dispersion model to simulate the observed contaminant concentrations in the classical Hanford experiment. The study of transport and dispersion of contaminants in the planetary boundary layer is described by the advection–diffusion equation, which is obtained parameterizing the turbulent fluxes in the continuity equation utilizing the gradient transport model or K-theory. For a cartesian coordinate system in which the $x$-direction coincides with that one of the wind speed magnitude, the steady state advection–diffusion equation is written as

$$U\frac{\partial \bar{c}}{\partial x} = \frac{\partial}{\partial x}(K_x \frac{\partial \bar{c}}{\partial x}) + \frac{\partial}{\partial y}(K_y \frac{\partial \bar{c}}{\partial y}) + \frac{\partial}{\partial z}(K_z \frac{\partial \bar{c}}{\partial z}) \tag{14.8}$$

where $\bar{c}$ is the average concentration of a contaminant, $U$ is the wind speed magnitude in $x$ direction, and $K_x$, $K_y$, and $K_z$ are the eddy diffusivities. Neglecting the longitudinal diffusion in comparison to wind advection, the integration of the equation (14.8) leads to

$$U\frac{\partial \bar{c_y}}{\partial x} = \frac{\partial}{\partial z}(K_z \frac{\partial \bar{c_y}}{\partial x}) \tag{14.9}$$

where $\bar{c_y}$ is the average crosswind integrated concentration in the vertical region $0 < z < z_i$ and for $X > 0$, considering the following boundary conditions and emission rate $Q$:

$$K_z \frac{\partial \bar{c_y}}{\partial z} = 0 \tag{14.10}$$

at $z = 0, h$ (zero concentration flux at the surface and CBL top) and

$$\bar{u}\bar{c_y}(0,z) = Q(z - H_s) \tag{14.11}$$

(emission rate at source height $H_s$).

In the present study, the solution for the problem defined by the equations (14.9), (14.10), and (14.11) is obtained by the GILTT method (see [BuEtAl11] and [MoEtAl09]). This general method to simulate pollutants dispersion in a planetary boundary layer is described in a detailed form in [BuEtAl11, MoEtAl09]. The vertical eddy diffusivity as given by equation (14.7) is introduced in equation (14.9) and solved with the GILTT method with the aim of evaluating the performance of this new parameterization in reproducing the observed ground level concentrations. To accomplish this task, observed concentration data from the Hanford dispersion experiment were simulated. The Hanford diffusion experiments were accomplished in Washington region [DoHo85]. The tracer SF6 was released at a height of $2m$ and sampled in arcs of $100, 200, 800, 1600$, and $3200m$ from the source. The $u_*$ and $L$ were determined by measurements obtained from sonic anemometer [DoHo85]. The height of the stable layer was determined by the following formulation [Ni84]:

$$h = 0.4(\frac{u_* L}{f})^{1/2}$$

The wind speed profile employed in the simulations is expressed by a power law provided by the following relation [AlEtAl12]

$$\frac{\overline{u_z}}{u_1} = (\frac{z}{z_1})^n$$

where $\overline{u_1}$ and $\overline{u_z}$ are the mean horizontal wind speeds at heights $z$ and $z_1$, while $n$ is an exponent that is related to the intensity of turbulence. For shear forcing conditions $n = 0.1$ [Ir79]. Table 14.1 shows a summary of the meteorological conditions during the Hanford stable experiments [DoHo85].

The performance of the GILTT method employing the vertical eddy diffusivity as given by equation (14.7) is shown in Table 14.2 and Figure 14.1.

Table 14.2 exhibits the statistical analysis that allows to compare observed and simulated magnitudes of the ground level crosswind integrated concentration $\overline{c_y}/Q$. The statistical indices to evaluate the performance of the new vertical eddy

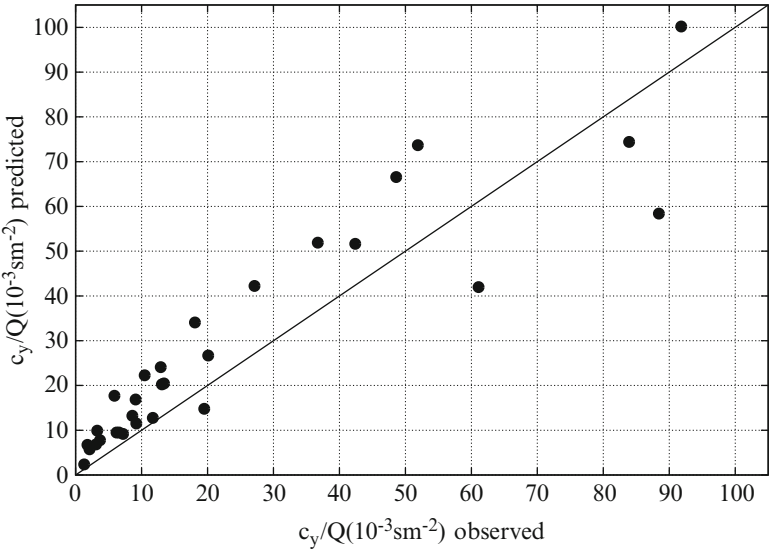


**Fig. 14.1** Scatter diagram between observed and predicted $c_y/Q$.

diffusivity were proposed by [Ha89]. The *NMSE* is the normalized mean square error, *FA*2 is Factor 2, *COR* is the correlation coefficient, *FB* is the fractional bias and *FS* is the fractional standard deviations. The meaning of these indices is discussed and explained in a detailed form in [MoEtAl11] and [MaEtAl13]. The statistical indices *NMSE*, *FB*, and *FS* represent good results when they approach zero, whereas *COR* and *FA*2 are optimized at the value 1. Therefore, analyzing the magnitude of the statistical indices in Table 14.2 and observing the scatter diagram in Figure 14.1, it is possible to conclude that the advection–diffusion equation (14.9), solved by GILTT method, employing the vertical eddy diffusivity for a shear dominated SBL obtained from Taylor statistical diffusion and local similarity theory, reproduces very well the observed crosswind integrated concentration from the Hanford stable experiment.

**Table 14.1** Meteorological parameters measured during the Hanford stable experiment.

| Run | Data | $h(m)$ | $L(m)$ | $u_*$ |
|---|---|---|---|---|
| 11 | May 18, 1983 | 325 | 166 | 0.40 |
| 12 | May 26, 1983 | 135 | 44 | 0.26 |
| 13 | June 5, 1983 | 182 | 77 | 0.27 |
| 14 | June 12, 1983 | 104 | 34 | 0.20 |
| 15 | June 24, 1983 | 157 | 59 | 0.26 |
| 16 | June 27, 1983 | 185 | 71 | 0.30 |

**Table 14.2** Statistical indices of the model performance for the Hanford stable diffusion experiments.

| | |
|---|---|
| Normalized Mean Square Error (*NMSE*) | 0.18 |
| Correlation coefficient (*COR*) | 0.92 |
| Fractional bias (*FB*) | 0.06 |
| Fractional Standard deviation (*FS*) | −0.18 |
| Factor 2 (*FA*2) | 0.77 |

## 14.4 Conclusion

Eddy diffusivities for a shear dominated stable planetary boundary layer turbulence are derived. The model is based upon Taylor statistical diffusion and local similarity theory. These eddy diffusivities can be applied to parameterize turbulent dispersion in the near, the intermediate and far field of an elevated continuous point source. The present development allows to construct a vertical eddy diffusivity expressed in terms of the source distance for an inhomogeneous turbulent field in a stable PBL. In this aspect $K_z$ is dependent on the nondimensional distance $X$ , on the stability parameter $z/\Lambda$ and of the nondimensional height $z/h$. To evaluate the new vertical eddy diffusivity (equation 14.7) in a Eulerian dispersion model, we employ equation (14.7) in the advection–diffusion equation (14.9) to simulate the Hanford observed ground level crosswind integrated concentrations. The results show that there is a good agreement between simulated and observed concentrations. Therefore, the new eddy diffusivity applied to a shear dominated SBL, expressed by the equation (14.7), depending on source distance and describing an inhomogeneous turbulence, can be applied in regulatory air pollution modeling.

**Acknowledgements** The authors thank CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and FAPERGS (Fundação de Apoio a Pesquisa do Estado do Rio Grande do Sul) for partial financial support of this work.

## References

[AlEtAl12] Alves, I.P., Degrazia, G.A., Buske, D., Vilhena, M.T., Moraes, O.L.L., Acevedo, O.C.: Derivation of an eddy diffusivity coefficient depending on source distance for a shear dominated planetary boundary layer. Physica A: Statistical Mechanics and its Applications, 6577–6586, 391, (2012).

[Ba49] Batchelor, G.K.: Diffusion in a field of homogeneous turbulence, Eulerian analysis. Australian Journal of Scientific Research 437–450, 2, (1949).

[BuEtAl11] Buske, D., Vilhena, M.T., Segatto, C.F., Quadros, R.S.: A general analytical solution of the advection-diffusion equation for Fickian closure, in: Integral Methods in Science and Engineering: Computational and Analytic Aspects, Birkhauser, Boston, 25–34, (2011).

[DeMoVi01] Degrazia, G. A., Moreira, D. M., Vilhena, M. T. Derivation of an eddy diffusivity depending on source distance for vertically inhomogeneous turbulence in a convective boundary layer. Journal of Applied Meteorology 1233–1240, 40, (2001).

[DeEtAl00] Degrazia, G. A., Anfossi, D., Carvalho, J. C., Mangia, C., Tirabassi, T., Campos Velho, H. F.: Turbulence parameterisation for PBL dispersion models in all stability conditions. Atmospheric Environment 3575–3583, 34, (2000).

[DoHo85] Doran, J.C. and Horst, T.W.: An evaluation of Gaussian plume-depletion models with dual-tracer field measurements, Atmospheric Environment, 939–951, 19, (1985).

[Ha89] Hanna, S R.: Confidence limits for air quality model evaluations, as estimated by bootstrap and jackknife resampling methods. Atmospheric Environment, 1385–1398, 23, (1989).

[Ir79] Irvin, J.S.: A theoretical variation of the wind profile power-low exponent as a function of surface roughness and stability. Atmospheric Environment 191-184, 13, (1979).

[MaEtAl13] Maldaner, S., Degrazia, G.A., Rizza, U., Moreira, V.S., Puhales, F.S., Acevedo, O.C., da Costa Carvalho, J.: Derivation of third-order vertical velocity turbulence moment in the convective boundary layer from large eddy simulation data: an application to the dispersion modeling. Atmospheric Pollution Research, 191–198, 4, (2013).

[MoEtAl11] Moreira, V.S., Degrazia, G.A., Roberti, D.R., Timm, A.U., da Costa Carvalho, J.: Employing a Lagrangian stochastic dispersion model and classical diffusion experiments to evaluate two turbulence parameterization schemes. Atmospheric Pollution Research, 384–393, 2, (2011).

[MoEtAl09] Moreira, D.M., Vilhena M.T., Buske, D., Tirabassi T.: The state-of-art of the GILTT method to simulate pollutant dispersion in the atmosphere, Atmospheric Research, 1–17, 92, (2009).

[Ni84] Nieuwstadt, F.T.M.: The turbulent structure of the stable, nocturnal boundary layer. Journal of the Atmospheric Sciences, 2202–2216, 41, (1984).

[PaSi83] Pasquill, F. and Smith, F. B. Atmospheric diffusion: Study of the dispersion of windborne material from industrial and other sources., John Wiley & Sons, New York, 94–100 (1983).

[WaKo62] Wandel, C.F. and Kofoed-Hansen, O.: On the Eulerian-Lagrangian transform in the statistical theory of turbulence. Journal of Geophysical Research, 3089–3093, 67, (1962).

# Chapter 15
# Analysis of Boundary–Domain Integral Equations for Variable-Coefficient Dirichlet BVP in 2D

**T.T. Dufera and S.E. Mikhailov**

## 15.1 Preliminaries

Let $\Omega$ be a domain in $\mathbb{R}^2$ bounded by simple closed infinitely differentiable curve $\partial\Omega$, the set of all infinitely differentiable function on $\Omega$ with compact support is denoted by $\mathscr{D}(\Omega)$. The function space $\mathscr{D}'(\Omega)$ consists of all continuous linear functionals over $\mathscr{D}(\Omega)$. For $s \in \mathbb{R}$, we denote by $H^s(\mathbb{R}^2)$ the Bessel potential space. Note that the space $H^1(\mathbb{R}^2)$ coincides with the Sobolev space $W_2^1(\mathbb{R}^2)$ with equivalent norm and $H^{-s}(\mathbb{R}^2)$ is the dual space to $H^s(\mathbb{R}^2)$. For any nonempty open set $\Omega \in \mathbb{R}^n$ we define $H^s(\Omega) = \{u \in \mathscr{D}'(\Omega) : u = U|_\Omega \text{ for some } U \in H^s(\mathbb{R}^n)\}$. The space $\widetilde{H}^s(\Omega)$ is defined to be the closure of $\mathscr{D}(\Omega)$ with respect to the norm of $H^s(\mathbb{R}^n)$. For $s \in (-\frac{1}{2}, \frac{1}{2})$, $H^s(\Omega)$ can be identified with $\widetilde{H}^s(\Omega)$, see e.g. [Mc00, HsWe08].

We shall consider the scalar elliptic differential equation

$$Au(x) = \sum_{i=1}^{2} \frac{\partial}{\partial x_i}\left[a(x)\frac{\partial u(x)}{\partial x_i}\right] = f(x) \qquad \text{in } \Omega \tag{15.1}$$

with $a(x) \in C^\infty(\mathbb{R}^2)$, $a(x) > 0$.

T.T. Dufera (✉)
Addis Ababa University, P.O. Box 1176, Addis Ababa, Ethiopia
e-mail: tamirat.temesgen@astu.edu.et

S.E. Mikhailov
Brunel University London, Uxbridge, UK
e-mail: sergey.mikhailov@brunel.ac.uk



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_15

For given functions $\varphi_0 \in H^{\frac{1}{2}}(\partial\Omega)$ and $f \in L_2(\Omega)$, we will consider the Dirichlet boundary value problem for function $u \in H^1(\Omega)$,

$$Au = f \qquad \text{in } \Omega, \tag{15.2}$$

$$\gamma^+ u = \varphi_0 \qquad \text{on } \partial\Omega. \tag{15.3}$$

Here equation (15.2) is understood in the distributional sense and (15.3) in the trace sense.

In applications, the BVP (15.2)–(15.3) may describe a stationary heat transfer boundary value problem in isotropic inhomogeneous two-dimensional body $\Omega$, where $u(x)$ is an unknown temperature, $a(x)$ is a known variable thermal conductivity coefficient, $f(x)$ is a known distributed heat source, $\varphi_0(x)$ is known temperature on the boundary.

We define as in [Gr85, Co88, Mi11], the subspace

$$H^{1,0}(\Omega;A) := \{g \in H^1(\Omega) : Ag \in L_2(\Omega)\}$$

endowed with the norm $\|g\|^2_{H^{1,0}(\Omega;A)} := \|g\|^2_{H^1(\Omega)} + \|Ag\|^2_{L_2(\Omega)}$.

For $u \in H^{1,0}(\Omega;A)$ we can define the (canonical) conormal derivative $T^+u \in H^{-\frac{1}{2}}(\partial\Omega)$ in the weak form (see, e.g., [Co88, Mi11] and the references therein),

$$\langle T^+u, w\rangle := \int_\Omega [(\gamma^+_{-1}w)Au + E(u, \gamma^+_{-1}w)]dx \qquad \forall w \in H^{\frac{1}{2}}(\partial\Omega), \tag{15.4}$$

where $\gamma^+_{-1} : H^{\frac{1}{2}}(\partial\Omega) \to H^1(\Omega)$ is a continuous right inverse of the continuous interior trace operator $\gamma^+ : H^1(\Omega) \to H^{\frac{1}{2}}(\partial\Omega)$, while $E(u,v) := a(x)\nabla u(x) \cdot \nabla v(x)$ is the symmetric bilinear form.

For $u \in H^s(\Omega)$, $s > 3/2$, the canonical conormal derivative defined by (15.4) coincides with the classical one, defined in the trace sense, i.e.,

$$T^+u = an \cdot \gamma^+ \nabla u, \tag{15.5}$$

where $n(x)$ is the exterior unit normal vector.

*Remark 1.* The first Green identity holds for any $u \in H^{1,0}(\Omega;A)$ and $v \in H^1(\Omega)$ ([Co88, Mi11]), i.e.,

$$\int_\Omega E(u,v)dx = -\langle Au, v\rangle_\Omega + \langle T^+u, \gamma^+v\rangle_{\partial\Omega}$$

and the second Green identity holds for any $u, v \in H^{1,0}(\Omega;A)$,

$$\int_\Omega (vAu - uAv)dx = \langle T^+u, \gamma^+v\rangle_{\partial\Omega} - \langle T^+v, \gamma^+u\rangle_{\partial\Omega}.$$

## 15.2 Parametrix-Based Potential Operators

A function $P(x,y)$ is a parametrix (Levi function) for the operator $A$ if

$$A_x P(x,y) = \delta(x-y) + R(x,y),$$

where $\delta$ is the Dirac-delta distribution, while $R(x,y)$ is a remainder possessing at most a weak singularity at $x = y$.

In particular, see, e.g., [Mi02], the function

$$P(x,y) = \frac{1}{2\pi a(y)} \log|x-y|, \quad x,y \in \mathbb{R}^2$$

is a parametrix for the operator $A$ and the corresponding remainder is

$$R(x,y) = \sum_{i=1}^{2} \frac{x_i - y_i}{2\pi a(y)|x-y|^2} \frac{\partial a(x)}{\partial x_i}, \quad x,y \in \mathbb{R}^2.$$

If $a(x) = 1$, then $A$ becomes the Laplace operator, $\Delta$, and the parametrix $P(x,y)$ becomes its fundamental solution, $P_\Delta(x,y)$.

If $u \in H^{1,0}(\Omega;A)$, then from the second Green identity, we have the following parametrix-based third Green identity for $y \in \Omega$, [Mi02],

$$\begin{aligned} u(y) = & \int_{\partial\Omega} [\gamma^+ u(x) T_x^+ P(x,y) - P(x,y) T^+ u(x)] dx \\ & - \int_\Omega R(x,y) u(x) dx + \int_\Omega P(x,y) f(x) dx, \quad y \in \Omega. \end{aligned} \tag{15.6}$$

Note that the direct substitution of $v(x)$ by $P(x,y)$ in the second Green identity is not possible as it has singularity at $x = y$. This difficulty is avoided by replacing $\Omega$ by $\Omega \setminus B(y,\varepsilon)$, where $B(y,\varepsilon)$ is a disc of radius $\varepsilon$ centered at $y$; taking the limit $\varepsilon \to 0$, we then arrive at (15.6), cf. e.g. [Mi70].

The parametrix-based logarithmic and remainder potential operators are defined, similar to [ChMiNa09a] in the 3D case, as

$$\mathscr{P}g(y) := \int_\Omega P(x,y) g(x) dx, \qquad \mathscr{R}g(y) := \int_\Omega R(x,y) g(x) dx.$$

The single-layer and double-layer potential operators, corresponding to the parametrix $P(x,y)$, are defined for $y \notin \partial\Omega$ as

$$Vg(y) := -\int_{\partial\Omega} P(x,y) g(x) ds_x, \qquad Wg(y) := -\int_{\partial\Omega} T_x^+ P(x,y) g(x) ds_x.$$

The following boundary integral (pseudo-differential) operators are also defined for $y \in \partial\Omega$,

$$\mathscr{V}g(y) := -\int_{\partial\Omega} P(x,y)g(x)ds_x, \qquad \mathscr{W}g(y) := -\int_{\partial\Omega} T_x^+ P(x,y)g(x)ds_x,$$

$$\mathscr{W}'g(y) := -\int_{\partial\Omega} T_y^+ P(x,y)g(x)ds_x, \qquad \mathscr{L}^+g(y) := T_y^+ Wg(y).$$

Let $\mathscr{P}_\Delta, V_\Delta, W_\Delta, \mathscr{V}_\Delta, \mathscr{W}_\Delta, \mathscr{L}_\Delta^+$ denote the potentials corresponding to the operator $A = \Delta$. Then the following relations hold (cf. [ChMiNa09a] for 3D case),

$$\mathscr{P}g = \frac{1}{a}\mathscr{P}_\Delta g, \qquad \mathscr{R}g = \frac{-1}{a(y)}\sum_{i=1}^{2}\partial_i \mathscr{P}_\Delta[g(\partial_i a)], \tag{15.7}$$

$$Vg = \frac{1}{a}V_\Delta g, \qquad Wg = \frac{1}{a}W_\Delta(ag) \tag{15.8}$$

$$\mathscr{V}g = \frac{1}{a}\mathscr{V}_\Delta g, \qquad \mathscr{W}g = \frac{1}{a}\mathscr{W}_\Delta(ag), \tag{15.9}$$

$$\mathscr{W}'g = \mathscr{W}'_\Delta g + \left[a\frac{\partial}{\partial n}\left(\frac{1}{a}\right)\right]\mathscr{V}_\Delta g, \tag{15.10}$$

$$\mathscr{L}^+g = \mathscr{L}_\Delta^+(ag) + \left[a\frac{\partial}{\partial n}\left(\frac{1}{a}\right)\right]W_\Delta^+(ag). \tag{15.11}$$

**Theorem 1.** *For $s \in \mathbb{R}$, the following operators are continuous,*

$$V : H^s(\partial\Omega) \to H^{s+\frac{3}{2}}(\Omega),$$

$$W : H^s(\partial\Omega) \to H^{s+\frac{1}{2}}(\Omega),$$

$$\mathscr{V} : H^s(\partial\Omega) \to H^{s+1}(\partial\Omega),$$

$$\mathscr{W}, \mathscr{W}' : H^s(\partial\Omega) \to H^{s+1}(\partial\Omega),$$

$$\mathscr{L}^+ : H^s(\partial\Omega) \to H^{s-1}(\partial\Omega).$$

*Proof.* We have the corresponding mappings for the corresponding constant-coefficient operators. Then (15.8)–(15.11) imply the theorem claim. □

**Theorem 2.** *Let $u \in H^{-\frac{1}{2}}(\partial\Omega)$ and $v \in H^{\frac{1}{2}}(\partial\Omega)$. Then the following jump relation holds on $\partial\Omega$*

$$\gamma^+ Vu(y) = \mathscr{V}u(y), \tag{15.12}$$

$$\gamma^+ W v(y) = -\frac{1}{2} v(y) + \mathscr{W} v(y), \tag{15.13}$$

$$T^+ V u(y) = \frac{1}{2} u(y) + \mathscr{W}' u(y). \tag{15.14}$$

*Proof.* For the constant coefficient case, this theorem is well known. Then taking into account the relations (15.8)–(15.10), we can prove the theorem for the variable positive coefficient $a \in C^\infty(\mathbb{R}^2)$ as well.

**Theorem 3.** *Let $\Omega$ be a bounded open domain in $\mathbb{R}^2$ with closed, infinitely smooth boundary $\partial\Omega$. The following operators are continuous.*

$$\mathscr{P} : \widetilde{H}^s(\Omega) \to H^{s+2}(\Omega), \qquad s \in \mathbb{R}; \tag{15.15}$$

$$: H^s(\Omega) \to H^{s+2}(\Omega), \qquad s > -\frac{1}{2}; \tag{15.16}$$

$$\mathscr{R} : \widetilde{H}^s(\Omega) \to H^{s+1}(\Omega), \qquad s \in \mathbb{R}; \tag{15.17}$$

$$: H^s(\Omega) \to H^{s+1}(\Omega), \qquad s > -\frac{1}{2}; \tag{15.18}$$

$$\gamma^+ \mathscr{P} : \widetilde{H}^s(\Omega) \to H^{s+\frac{3}{2}}(\partial\Omega), \qquad s > -\frac{3}{2}; \tag{15.19}$$

$$: H^s(\Omega) \to H^{s+\frac{3}{2}}(\partial\Omega), \qquad s > -\frac{1}{2}; \tag{15.20}$$

$$\gamma^+ \mathscr{R} : \widetilde{H}^s(\Omega) \to H^{s+\frac{1}{2}}(\partial\Omega), \qquad s > -\frac{1}{2}; \tag{15.21}$$

$$: H^s(\Omega) \to H^{s+\frac{1}{2}}(\partial\Omega), \qquad s > -\frac{1}{2}; \tag{15.22}$$

$$T^+ \mathscr{P} : \widetilde{H}^s(\Omega) \to H^{s+\frac{1}{2}}(\partial\Omega), \qquad s > -\frac{1}{2}; \tag{15.23}$$

$$: H^s(\Omega) \to H^{s+\frac{1}{2}}(\partial\Omega), \qquad s > -\frac{1}{2}; \tag{15.24}$$

$$T^+ \mathscr{R} : \widetilde{H}^s(\Omega) \to H^{s-\frac{1}{2}}(\partial\Omega), \qquad s > \frac{1}{2}; \tag{15.25}$$

$$: H^s(\Omega) \to H^{s-\frac{1}{2}}(\partial\Omega), \qquad s > \frac{1}{2}. \tag{15.26}$$

*Proof.* The operator $\mathscr{P}_\Delta$ is a homogeneous pseudo-differential operator of order $-2$ on $\mathbb{R}^2$, mapping $\mathscr{P}_\Delta : H^s_{\text{comp}}(\mathbb{R}^2) \to H^{s+2}_{\text{loc}}(\mathbb{R}^2)$ continuously for any $s \in \mathbb{R}$. Hence the application of trace theorem along with the relations (15.7), the operators (15.15), (15.17),(15.19),(15.21), (15.23), and (15.25) are continuous. For $s \in (-\frac{1}{2}, \frac{1}{2})$, $\widetilde{H}^s(\Omega)$ is identified with $H^s(\Omega)$, and (15.16) directly follows from (15.15). To prove the case $s \in (\frac{1}{2}, \frac{3}{2})$, we implement the Gauss divergence theorem and the fact that $\dfrac{\partial}{\partial x_j} \log|x-y| = -\dfrac{\partial}{\partial y_j} \log|x-y|$ and obtain

$$\frac{\partial}{\partial y_j}(\mathscr{P}_\Delta g)(y) = -\frac{1}{2\pi}\int_\Omega g(x)\frac{\partial}{\partial x_j}\log|x-y|dx$$

$$= \frac{1}{2\pi}\int_\Omega \log|x-y|\frac{\partial}{\partial x_j}g(x)dx - \frac{1}{2\pi}\int_{\partial\Omega}\log|x-y|n_j\gamma^+ g(x)ds_x$$

$$= \mathscr{P}_\Delta(\partial_j g)(y) + V_\Delta(n_j\gamma^+ g)(y). \tag{15.27}$$

Now for $s \in (\frac{1}{2}, \frac{3}{2})$, since $\partial_j : H^s(\Omega) \to H^{s-1}(\Omega)$ is continuous, we have $\mathscr{P}_\Delta\partial_j : H^s(\Omega) \to H^{s+1}(\Omega)$ is continuous, and from trace theorem $\gamma^+ g \in H^{s-\frac{1}{2}}(\partial\Omega)$ and the properties of the single-layer potential, we conclude that $\nabla\mathscr{P}_\Delta : H^s(\Omega) \to H^{s+1}(\Omega)$ is continuous. This implies that $\mathscr{P}_\Delta : H^s(\Omega) \to H^{s+2}(\Omega)$ is continuous, which along with the relation $\mathscr{P}g = \frac{1}{a}\mathscr{P}_\Delta$ leads to the continuity of operator (15.16) for $s \in (\frac{1}{2}, \frac{3}{2})$.

Further, with the help of these results and the relation (15.27), we can verify by induction that the operator (15.16) is continuous for $s \in (k-\frac{1}{2}, k+\frac{1}{2})$, where $k$ is an arbitrary nonnegative integer. For the values $s = k+\frac{1}{2}$ the continuity of the operator (15.16) then follows due to the complex interpolation properties of Bessel potential spaces.

The trace theorem will give the continuity proof for the operators (15.19) and (15.20). We can follow the same procedure to prove the claim of the theorem concerning the operator $\mathscr{R}$. The continuity of the operators (15.23)–(15.26) follows if we remark that for the chosen $s$ the conormal derivative can be understood in the classical sense (15.5). □

By the Rellich compact embedding theorem (see, e.g., [Mc00, Theorem 3.27]), Theorems 1 and 3 imply the following two assertions.

**Corollary 1.** *Let $s \in \mathbb{R}$. The following operators are compact,*

$$\mathscr{V} : H^s(\partial\Omega) \to H^s(\partial\Omega) \tag{15.28}$$

$$\mathscr{W} : H^s(\partial\Omega) \to H^s(\partial\Omega) \tag{15.29}$$

$$\mathscr{W}' : H^s(\partial\Omega) \to H^s(\partial\Omega) \tag{15.30}$$

**Corollary 2.** *The following operators are compact for any $s > \frac{1}{2}$,*

$$\mathscr{R} : H^s(\Omega) \to H^s(\Omega),$$

$$\gamma^+\mathscr{R} : H^s(\Omega) \to H^{s-\frac{1}{2}}(\partial\Omega),$$

$$T^+\mathscr{R} : H^s(\Omega) \to H^{s-\frac{3}{2}}(\partial\Omega).$$

## 15.3 Invertibility of the Single-Layer Potential Operator

It is well known (see, e.g., [Co00, Remark 1.42(ii)], [St08, proof of Theorem 6.22]) that for some 2D domains the kernel of the operator $\mathscr{V}_\Delta$ is non-zero, which by (15.9) also implies that $\ker \mathscr{V} \neq \{0\}$ for the same domains.

In order to have invertibility for the single-layer potential operator in 2D, we define the following subspace of the space $H^{-\frac{1}{2}}(\partial\Omega)$, see, e.g., [St08, Eq. (6.30)],

$$H_*^{-\frac{1}{2}}(\partial\Omega) := \{\phi \in H^{-\frac{1}{2}}(\partial\Omega) : \langle \phi, 1\rangle_{\partial\Omega} = 0\},$$

where the norm in $H_*^{-\frac{1}{2}}(\partial\Omega)$ is the induced by the norm in $H^{-\frac{1}{2}}(\partial\Omega)$.

**Theorem 4.** *If $\psi \in H_*^{-\frac{1}{2}}(\partial\Omega)$ satisfies $\mathscr{V}\psi = 0$ on $\partial\Omega$, then $\psi = 0$.*

*Proof.* The theorem holds for the operator $\mathscr{V}_\Delta$ (see, e.g., [Mc00, Corollary 8.11(ii)]), which by (15.9) implies it for the operator $\mathscr{V}$ as well. □

**Theorem 5.** *Let $\Omega \subset \mathbb{R}^2$ have the diameter* $\operatorname{diam}(\Omega) < 1$. *Then the single layer potential $\mathscr{V} : H^{-\frac{1}{2}}(\partial\Omega) \to H^{\frac{1}{2}}(\partial\Omega)$ is invertible.*

*Proof.* By [St08, Theorem 6.23], for $diam(\Omega) < 1$ the operator $\mathscr{V}_\Delta : H^{-\frac{1}{2}}(\partial\Omega) \to H^{\frac{1}{2}}(\partial\Omega)$ is $H^{-\frac{1}{2}}(\partial\Omega)-$elliptic and since it is also bounded, c.f. Theorem 1 for $s = -1/2$, the Lax–Milgram theorem implies its invertibility. Then by the first relation in (15.10) the invertibility of the operator $\mathscr{V} : H^{-\frac{1}{2}}(\partial\Omega) \to H^{\frac{1}{2}}(\partial\Omega)$ also follows. □

## 15.4 The Third Green Identity

For $u \in H^{1,0}(A;\Omega)$, let us write the third Green identity (15.6) using the surface and volume potential operator notations,

$$u + \mathscr{R}u - V T^+ u + W\gamma^+ u = \mathscr{P}Au \qquad \text{in } \Omega. \tag{15.31}$$

Applying the *trace operator* to equation (15.31) and using the jump relations from Theorem 2, we have

$$\frac{1}{2}\gamma^+ u + \gamma^+ \mathscr{R}u - \mathscr{V} T^+ u + \mathscr{W}\gamma^+ u = \gamma^+ \mathscr{P}Au \qquad \text{on } \partial\Omega. \tag{15.32}$$

Similarly, applying the *conormal derivative operator* to equation (15.31), and using again the jump relation, we obtain

$$\frac{1}{2}T^+u + T^+\mathscr{R}u - \mathscr{W}'T^+u + \mathscr{L}^+\gamma^+u = T^+\mathscr{P}Au \qquad \text{on } \partial\Omega. \tag{15.33}$$

For some functions $f$, $\Psi$ and $\Phi$ let us consider a more general indirect integral relation associated with equation (15.31).

$$u + \mathscr{R}u - V\Psi + W\Phi = \mathscr{P}f \qquad \text{in } \Omega. \tag{15.34}$$

**Lemma 1.** *Let $u \in H^1(\Omega)$, $f \in L_2(\Omega)$, $\Psi \in H^{-\frac{1}{2}}(\partial\Omega)$, and $\Phi \in H^{\frac{1}{2}}(\partial\Omega)$ satisfy equation (15.34). Then $u$ belongs to $H^{1,0}(\Omega;A)$ and is a solution of PDE $Au = f$ in $\Omega$ and*

$$V(\Psi - T^+u)(y) - W(\Phi - \gamma^+u)(y) = 0, \qquad y \in \Omega$$

*Proof.* The proof follows word for word the corresponding proof in 3D case in [ChMiNa09a, Theorem 4.1]. □

**Lemma 2.** *(i) Let either $\Psi^* \in H^{-\frac{1}{2}}(\partial\Omega)$ and $\operatorname{diam}(\Omega) < 1$, or $\Psi^* \in H_*^{-\frac{1}{2}}(\partial\Omega)$. If $V\Psi^* = 0$ in $\Omega$, then $\Psi^* = 0$ on $\partial\Omega$.*

*(ii) Let $\Phi^* \in H^{\frac{1}{2}}(\partial\Omega)$. If $W\Phi^* = 0$ in $\Omega$, then $\Phi^* = 0$ on $\partial\Omega$.*

*Proof.* (i) Taking the trace of equation in Lemma 2(i) on $\partial\Omega$, by the jump relation (15.13) we have $\mathscr{V}\Psi^*(y) = 0$ on $\partial\Omega$. If $\Psi^* \in H^{-\frac{1}{2}}(\partial\Omega)$ and $\operatorname{diam}(\Omega) < 1$, then the result follows from the invertibility of the single-layer potential given by Theorem 5. On the other hand, if $\Psi^* \in H_*^{-\frac{1}{2}}(\partial\Omega)$, then the result is implied by Theorem 4.

(ii) Let us take the trace of equation in Lemma 2(ii) on $\partial\Omega$, and use the jump relation (15.14) to obtain,

$$-\frac{1}{2}\Phi^* + \mathscr{W}\Phi^* = 0 \ \text{ on } \partial\Omega.$$

Multiplying this equation by $a(y)$, denoting $\hat{\Phi}^* = a\Phi^*$ and using the second relation in (15.9), we obtain equation

$$-\frac{1}{2}\hat{\Phi}^* + \mathscr{W}_\Delta\hat{\Phi}^* = 0 \ \text{ on } \partial\Omega.$$

It is well known that this equation has only the trivial solution. It is particularly due to the contraction property of the operator $\frac{1}{2}I + \mathscr{W}_\Delta$, see [StWe01, Theorem 3.1]. Since $a(y) \neq 0$, the result follows. □

## 15.5 Boundary–Domain Integral Equations (BDIEs)

To reduce the variable-coefficient Dirichlet BVP (15.2)–(15.3) to a *segregated* boundary-domain integral equation system, let us denote the unknown conormal derivative as $\psi := T^+ u \in H^{-\frac{1}{2}}(\partial\Omega)$ and will further consider $\psi$ as formally independent on $u$.

Assuming that the function $u$ satisfies PDE $Au = f$, by substituting the Dirichlet condition into the third Green identity (15.31) and either into its trace (15.32) or into its conormal derivative (15.33) on $\partial\Omega$, we can reduce the BVP (15.2)–(15.3) to two different systems of Boundary–Domain-Integral Equations for the unknown functions $u \in H^{1,0}(\Omega;A)$ and $\psi := T^+ u \in H^{-\frac{1}{2}}(\partial\Omega)$.

**BDIE system (D1)** obtained from equations (15.31) and (15.32) is

$$\begin{aligned} u + \mathscr{R} u - V\psi &= F_0 \qquad \text{in } \Omega, \\ \gamma^+ \mathscr{R} u - \mathscr{V}\psi &= \gamma^+ F_0 - \varphi_0 \qquad \text{on } \partial\Omega, \end{aligned}$$

where

$$F_0 := \mathscr{P} f - W\varphi_0 \ \text{ in } \Omega. \tag{15.35}$$

The system can be written in matrix form as $\mathscr{A}^1 \mathscr{U} = \mathscr{F}^1$, where $\mathscr{U} := [u, \psi]^\top \in H^{1,0}(\Omega;A) \times H^{-\frac{1}{2}}(\partial\Omega)$ and

$$\mathscr{A}^1 := \begin{bmatrix} I + \mathscr{R} & -V \\ \gamma^+ \mathscr{R} & -\mathscr{V} \end{bmatrix}, \qquad \mathscr{F}^1 = \begin{bmatrix} F_0 \\ \gamma^+ F_0 - \varphi_0 \end{bmatrix}.$$

From the mapping properties of $W$ in Theorem 1 and $\mathscr{P}$ in Theorem 3, we get the inclusion $F_0 \in H^{1,0}(\Omega;A)$, and the trace theorem implies $\gamma^+ F_0 \in H^{\frac{1}{2}}(\partial\Omega)$. Therefore, $\mathscr{F}^1 \in H^1(\Omega) \times H^{\frac{1}{2}}(\partial\Omega)$. Due to the mapping properties of the operators involved in $\mathscr{A}^1$, the operator $\mathscr{A}^1 : H^{1,0}(\Omega;A) \times H^{-\frac{1}{2}}(\partial\Omega) \to H^1(\Omega) \times H^{\frac{1}{2}}(\partial\Omega)$ is bounded.

**BDIE system (D2)** obtained from equations (15.31) and (15.33) is

$$\begin{aligned} u + \mathscr{R} u - V\psi &= F_0 \qquad \text{in } \Omega, \\ \frac{1}{2}\psi + T^+ \mathscr{R} u - \mathscr{W}'\psi &= T^+ F_0 \qquad \text{on } \partial\Omega, \end{aligned}$$

where $F_0$ is given by (15.35). In matrix form it can be written as $\mathscr{A}^2 \mathscr{U} = \mathscr{F}^2$, where

$$\mathscr{A}^2 = \begin{bmatrix} I + \mathscr{R} & -V \\ T^+ \mathscr{R} & \frac{1}{2} I - \mathscr{W}' \end{bmatrix}, \qquad \mathscr{F}^2 = \begin{bmatrix} F_0 \\ T^+ F_0 \end{bmatrix}$$

Note that the operator $\mathscr{A}^2 : H^{1,0}(\Omega;A) \times H^{-\frac{1}{2}}(\partial\Omega) \to H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega)$ is bounded.

## 15.6 Equivalence and Invertibility Theorems

In the following theorem we shall see the equivalence of the original Direchlet boundary value problem to the boundary–domain integral equation systems.

**Theorem 6.** *Let* $\varphi_0 \in H^{\frac{1}{2}}(\partial\Omega)$ *and* $f \in L_2(\Omega)$.

*(i) If some* $u \in H^1(\Omega)$ *solves the BVP(15.2)–(15.3), then the pair* $(u, \psi)$*,where*

$$\psi = T^+ u \in H^{-\frac{1}{2}}(\partial\Omega), \tag{15.36}$$

*solves BDIE systems (D1) and (D2).*

*(ii) If a pair* $(u, \psi) \in H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega)$ *solves BDIE system (D1), and* $\operatorname{diam}(\Omega) < 1$*, then u solves BDIE system (D2) and BVP(15.2)–(15.3), this solution is unique, and* $\psi$ *satisfies (15.36).*

*(iii) If a pair* $(u, \psi) \in H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega)$ *solves BDIE system (D2), then u solves BDIE system (D1) and BVP(15.2)–(15.3), this solution is unique, and* $\psi$ *satisfies (15.36).*

*Proof.* (i) Let $u \in H^1(\Omega)$ be solution of the BVP(15.2)–(15.3). Since $f \in L_2(\Omega)$, we have that $u \in H^{1,0}(\Omega;A)$. Setting $\psi$ by (15.36) and recalling how BDIE systems (D1) and (D2) were constructed, we obtain that $(u, \psi)$ solve them.

Let now a pair $(u, \psi) \in H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega)$ solve system (D1) or (D2). Due to the first equations in the BDIE systems, the hypotheses of Lemma (1) are satisfied implying that $u$ belongs to $H^{1,0}(\Omega;A)$ and solves PDE (15.2) in $\Omega$, while the following equation also holds,

$$V(\psi - T^+ u)(y) - W(\varphi_0 - \gamma^+ u)(y) = 0, \qquad y \in \Omega. \tag{15.37}$$

(ii) Let $(u, \psi) \in H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega)$ solve system (D1). Taking the trace of the first equation in (D1) and subtracting the second equation from it, we get $\gamma^+ u = \varphi_0$ on $\partial\Omega$. Thus, the Dirichlet boundary condition is satisfied, and using it in (15.37), we have $V(\psi - T^+ u)(y) = 0$, $y \in \Omega$. Lemma 2(i) then implies $\psi = T^+ u$.

(iii) Let now $(u, \psi) \in H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega)$ solve system (D2). Taking the conormal derivative of the first equation in (D2) and subtracting the second equation from it, we get $\psi = T^+ u$ on $\partial\Omega$. Then inserting this in (15.37) gives $W(\varphi_0 - \gamma^+ u)(y) = 0$, $y \in \Omega$ and Lemma 2(ii) implies $\varphi_0 = \gamma^+ u$ on $\partial\Omega$.

The uniqueness of the BDIE system solutions follows from the fact that the corresponding homogeneous BDIE systems can be associated with the homogeneous Dirichlet problem, which has only the trivial solution. Then paragraphs (ii) and (iii) above imply that the homogeneous BDIE systems also have only the trivial solutions. □

**Theorem 7.** *If* $\operatorname{diam}(\Omega) < 1$, *then the following operators are invertible,*

$$\mathscr{A}^1 : H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega) \to H^1(\Omega) \times H^{\frac{1}{2}}(\partial\Omega), \tag{15.38}$$

$$\mathscr{A}^1 : H^{1,0}(\Omega;A) \times H^{-\frac{1}{2}}(\partial\Omega) \to H^{1,0}(\Omega;A) \times H^{\frac{1}{2}}(\partial\Omega). \tag{15.39}$$

*Proof.* Theorem 6(ii) implies that operators (15.38) and (15.39) are injective.

Let us denote $\mathscr{A}^1_0 := \begin{bmatrix} I & -V \\ 0 & -\mathscr{V} \end{bmatrix}$. Then $\mathscr{A}^1_0 : H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega) \to H^1(\Omega) \times H^{\frac{1}{2}}(\partial\Omega)$ is bounded. It is invertible due to its triangular structure and invertibility of its diagonal operators $I : H^1(\Omega) \to H^1(\Omega)$ and $-\mathscr{V} : H^{-\frac{1}{2}}(\partial\Omega) \to H^{\frac{1}{2}}(\partial\Omega)$ (see Theorem 5).

By Corollary 2 the operator

$$\mathscr{A}^1 - \mathscr{A}^1_0 = \begin{bmatrix} R & 0 \\ \gamma^+ R & 0 \end{bmatrix} : H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega) \to H^1(\Omega) \times H^{\frac{1}{2}}(\partial\Omega)$$

is compact, implying that operator (15.38) is a Fredholm operator with zero index, see, e.g., [Mc00, Theorem 2.26]. Then the injectivity of operator (15.38) implies its invertibility, see e.g. [Mc00, Theorem 2.27].

To prove invertibility of operator (15.39), we remark that for any element $\mathscr{F}^1 \in H^{1,0}(\Omega;A) \times H^{\frac{1}{2}}(\partial\Omega)$, a solution of the equation $\mathscr{A}^1\mathscr{U} = \mathscr{F}^1$ can be written as $\mathscr{U} = (\mathscr{A}^1)^{-1}\mathscr{F}^1$, where $(\mathscr{A}^1)^{-1} : H^1(\Omega) \times H^{\frac{1}{2}}(\partial\Omega) \to H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega)$ is the continuous inverse to operator (15.38). But due to Lemma 1 the first equation of system (D1) implies that $\mathscr{U} = (\mathscr{A}^1)^{-1}\mathscr{F}^1 \in H^{1,0}(\Omega;A) \times H^{-\frac{1}{2}}(\partial\Omega)$ and moreover, the operator $(\mathscr{A}^1)^{-1} : H^{1,0}(\Omega;A) \times H^{\frac{1}{2}}(\partial\Omega) \to H^{1,0}(\Omega;A) \times H^{-\frac{1}{2}}(\partial\Omega)$ is continuous, which implies invertibility of operator (15.39). □

The following similar assertion for the operator $\mathscr{A}^2$ holds without the limitation on the diameter of $\Omega$.

**Theorem 8.** *The following operators are invertible.*

$$\mathscr{A}^2 : H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega) \to H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega), \tag{15.40}$$

$$\mathscr{A}^2 : H^{1,0}(\Omega;A) \times H^{-\frac{1}{2}}(\partial\Omega) \to H^{1,0}(\Omega;A) \times H^{-\frac{1}{2}}(\partial\Omega). \tag{15.41}$$

*Proof.* Theorem 6(iii) implies that operators (15.40) and (15.41) are injective.

Let us denote $\mathscr{A}^2_0 = \begin{bmatrix} I & -V \\ 0 & \frac{1}{2}I \end{bmatrix}$. Then $\mathscr{A}^2_0 : H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega) \to H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega)$ is bounded. It is invertible due to its triangular structure and invertibility of its diagonal operators $I : H^1(\Omega) \to H^1(\Omega)$ and $I : H^{-\frac{1}{2}}(\partial\Omega) \to H^{-\frac{1}{2}}(\partial\Omega)$. By Corollaries 1 and 2 the operator

$$\mathscr{A}^2 - \mathscr{A}_0^2 = \begin{bmatrix} R & 0 \\ T^+R & -\mathscr{W}' \end{bmatrix} : H^{1,0}(\Omega;A) \times H^{-\frac{1}{2}}(\partial\Omega) \to H^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega)$$

is compact. This implies that operator (15.40) is a Fredholm operator with zero index and then the injectivity of operator (15.40) implies its invertibility.

The invertibility of operator (15.41) is then proved similar to the last paragraph of the proof of Theorem 7. □

## 15.7 Conclusions

In this paper, we have considered the interior Dirichlet problem for variable coefficient PDE in a two-dimensional domain, where the right-hand side function is from $L_2(\Omega)$ and the Dirichlet data from the space $H^{\frac{1}{2}}(\partial\Omega)$. The BVP was reduced to two systems of Boundary–Domain Integral Equations and their equivalence to the original BVP was shown. The invertibility of the associated operators in the corresponding Sobolev spaces was also proved.

In a similar way one can consider also the 2D versions of the BDIEs for the Neumann problem, mixed problem in interior and exterior domains, united BDIEs as well as the localized BDIEs, which were analyzed for 3D case in [ChMiNa09a, ChMiNa13, Mi06, ChMiNa09b].

**Acknowledgements** The first author work on this paper is a part of his PhD project supported by DAAD. He would like also to thank his PhD adviser Dr. Tsegaye Gedif Ayele for discussing the results.

## References

[ChMiNa09a] Chkadua, O., Mikhailov, S.E., and Natroshvili, D.: Analysis of direct boundary-domain integral equations for a mixed BVP with variable coefficient. I: Equivalence and invertibility. *J. Integral Equations Appl.* **21**, 499–543 (2009).

[ChMiNa09b] Chkadua, O., Mikhailov, S.E., and Natroshvili, D.: Analysis of some localized boundary-domain integral equations. *J. Integral Equations Appl.* **21**, 405–445 (2009).

[ChMiNa13] Chkadua, O., Mikhailov, S.E., and Natroshvili, D.: Analysis of direct segregated boundary-domain integral equations for variable-coefficient mixed BVPs in exterior domains. *Analysis and Appl.* **11**, 1350006(1–33) (2013).

[Co00] Constanda, C.: *Direct and Indirect Boundary Integral Equation Methods*. Chapman & Hall/CRC (2000).

[Co88] Costabel, M.: Boundary integral operators on Lipschitz domains: elementary results. *SIAM J. Mathematical Anal.* **19**, 613–626 (1988).

[Gr85] Grisvard, P.: *Elliptic Problems in Nonsmooth Domains*. Pitman, Boston–London–Melbourne (1985).

[HsWe08] Hsiao, G.C. and Wendland, W.L.: *Boundary Integral Equations*. Springer, Berlin (2008).

[Mc00] McLean, W.: *Strongly Elliptic Systems and Boundary Integral Equations.* Cambrige University Press, Cambrige (2000).

[Mi02] Mikhailov, S.E.: Localized boundary-domain integral formulations for problems with variable coefficients. *Int. J. Engineering Analysis with Boundary Elements*, **26**, 681–690 (2002).

[Mi06] Mikhailov, S.E.: Analysis of united boundary-domain integro-differential and integral equations for a mixed BVP with variable coefficient. *Math. Methods Appl. Sci.* **29**, 715–739 (2006).

[Mi11] Mikhailov, S.E.: Traces, extensions and co-normal derivatives for elliptic systems on Lipschitz domains. *Math. Anal. Appl.* **378**, 324–342 (2011).

[Mi70] Miranda, C.: *Partial Differential Equations of Elliptic Type.* Springer, Berlin-Heidelberg-New York (1970).

[St08] Steinbach, O.: *Numerical Approximation Methods for Elliptic Boundary Value Problems: Finite and Boundary Elements.* Springer (2007).

[StWe01] Steinbach, O. and Wendland, W.L.: On C. Neumann's method for second-order elliptic systems in domains with non-smooth boundaries. *Math. Anal. Appl.* **262**, 733–748 (2001).

# Chapter 16
# Onset of Separated Water-Layer in Three-Phase Stratified Flow

**M. Er, R. Mohan, E. Pereyra, O. Shoham, G. Kouba, and C. Avila**

## 16.1 Introduction

Three-phase gas-oil-water stratified flow schematic is shown in Figure 16.1. As shown in the figure, two possible flow configurations of the liquid-phase may occur. The first flow configuration is a separated liquid-phase, namely, the oil and the water flow separately as layers. The second possible flow configuration is a mixed liquid phase, whereby one of the phases is dispersed into the other. The flow configuration of the liquid-phase in a three-phase stratified flow in pipelines can affect the operation of the line. When the liquid-phase is separated, water can accumulate in low locations along the pipeline. The water accumulation may increase the pressure upstream, and eventually the accumulated water will be pushed forward by the gas in the form of a water slug. The water slug may cause operational problems in downstream separation facilities, which might require shutdown of the system. Additionally, water accumulation may lead to Under Deposit Corrosion. Thus, it is desirable to operate a three-phase stratified flow pipeline under dispersed liquid-phase conditions, avoiding accumulation, corrosion, and slugging of water in the pipeline. There are numerous publications on stratified three-phase flow; however, most of them focus on the interaction between the gas and liquid phases. As of the writing of this chapter, no studies have been carried out on the interaction between the oil and water phases under three-phase stratified flow conditions.

M. Er • R. Mohan • E. Pereyra • O. Shoham (✉)
The University of Tulsa, 800 S. Tucker Drive, Tulsa, OK 74104, USA
e-mail: morkuner@tpao.gov.tr; ram-mohan@utulsa.edu; ep@utulsa.edu; ovadia-shoham@utulsa.edu

G. Kouba • C. Avila
Chevron Energy Technology Company, Houston, TX, USA
e-mail: genekouba@chevron.com; c.avila@chevron.com



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_16

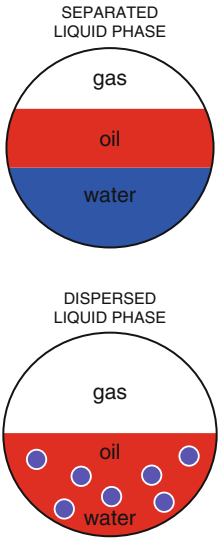


**Fig. 16.1** Liquid-phase behaviors in three-phase stratified flow.

The objective of this study is to acquire data and to develop a model for the prediction of the transition boundary between the separated liquid-phase and dispersed liquid-phase regimes, namely, the onset of a separated water-layer, in a horizontal three-phase stratified flow. This represents a novel study, since no studies have been conducted on this topic before.

## 16.2 Experimental Program

This section provides details of the three-phase flow experimental facility used to investigate the flow behavior of the liquid-phase in three-phase stratified flow. The test matrix, fluid physical properties, and testing procedure are also presented, as well as the acquired data on the liquid-phase flow behavior. Refer to Er [Er10] for additional details.

### *16.2.1 Experimental Facility*

The three-phase oil-water-gas flow loop, shown schematically in Figure 16.2, is a fully instrumented state-of-the-art facility. The three-phase flow loop consists of two major sections, namely, the storage and metering section and the test section, which are described briefly next.

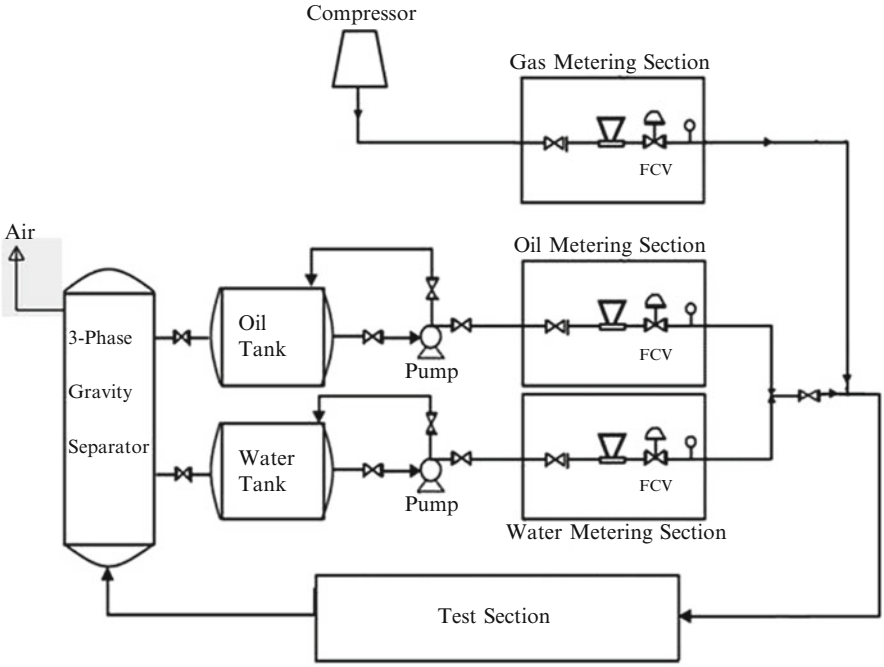


**Fig. 16.2** Schematic of three-phase flow loop.

#### 16.2.1.1 Storage and Metering Section

Two separate tanks are installed for oil and water storage, with a capacity of 400 gallons each. The oil and water flow from the three-phase separator into the respective storage tanks.

Two 3656 model pumps are connected to each of the tanks in order to deliver oil and water to the test section. One of the pumps' size is $1 \times 2$–8 with a 10 HP motor, delivering 25 gpm rotating at 3600 rpm. The size of the second pump is $1.5 \times 2$–10 with a 25 HP motor, delivering 110 gpm rotating at 3600 rpm. Gas is provided by a compressor, which delivers 240 scfm at 100 psig.

The fluids pass through the metering section before reaching the test section. Oil, water, and gas densities and flow rates are measured utilizing Micromotion® Coriolis meters, and the flow rates are controlled by control valves. Pressure and temperature transducers, and check valves are also installed in the metering section.

The oil and water are mixed in an impacting tee, which is located upstream of a second impacting tee that combines the gas with the oil and water mixture to obtain gas-oil-water flow.

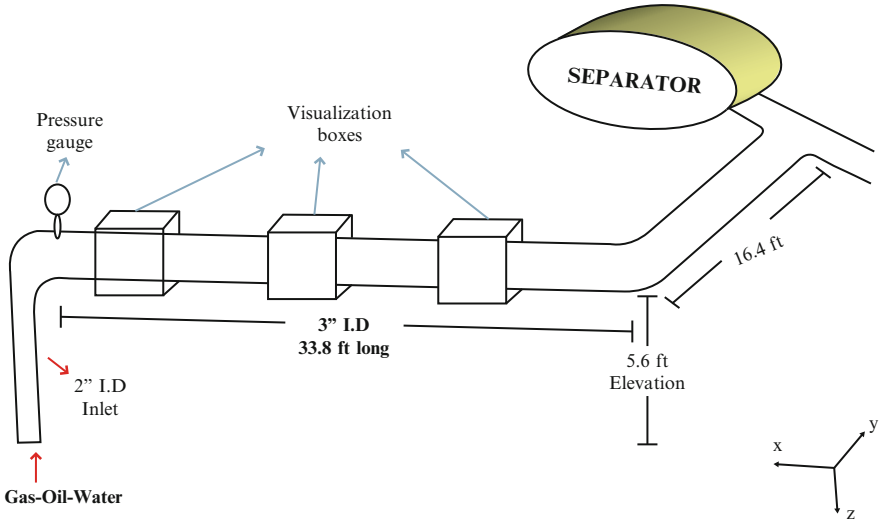


**Fig. 16.3** Schematic of first test section.

### 16.2.1.2 Test Section

The horizontal test section, shown in Figure 16.3, is 33.8 ft (10.3 m) long, constructed of a 3-in.-ID PVC pipe. The elevation of the test section is 5.6 ft (around eye level), facilitating visual observations. A three-phase separator is located downstream of the test section, operating at 7 psig, where the phases are separated. The air is discharged to the atmosphere, and the separated oil and water flow back into their respective storage tanks.

The inlet section is a vertical 2-in.-ID, 2-ft long PVC pipe. A static mixer is installed at the bottom of the inlet, to ensure well-mixed gas-oil-water flow. The gas-oil-water mixture flows through the vertical inlet section into the horizontal test section. Three visualization boxes were installed along the test section. These boxes are filled with Glycerin to prevent light reflection, making observations and measurements more accurate. The visualization boxes are located at 1.5, 4.5, and 7.5 m from the inlet. The first visualization box is used to verify that all fluids are well mixed. The two others are used to observe the liquid-phase flow behavior and to measure the heights of the oil and water layers. A pressure gauge is installed at the inlet of the test section in order to obtain the average pressure in the test section and adjust the gas flow rate accordingly.

The measured oil, water, and gas flow rates and densities are transferred to a computer through LabView software. The mass flow rates are controlled by using the front panel of the program. Volumetric flow rates, superficial velocities, densities, and system pressure and temperature are also depicted on the front panel. The acquired data can be saved in an Excel file for further analysis.

**Table 16.1** Physical Properties of Tap Water and Tulco Tech Oil.

| Tap Water | | Tulco Tech Oil | |
|---|---|---|---|
| Density ($\rho$) @70°F | 1.0 g/cm$^3$ | Specific gravity ($\gamma$) | 0.857 |
| Viscosity ($\mu$) @70°F | 1.25 cp | Viscosity ($\mu$) @100°F | 13.6 cp |
| Surface Tension @77°F | 71.97 dyne/cm | Surface Tension @77°F | 29.14 dyne/cm |

## *16.2.2 Test Matrix*

The physical properties of the test fluids, detailed information on the test matrix, and test procedure are presented next.

### 16.2.2.1 Test Fluids

The working fluids used in this study are air, tap water, and Tulco Tech 80 oil. The Tulco Tech 80 oil was selected because of its fast separability and stability. For all the experimental runs, the temperature was between 67 and 70°F and the average pressure was around 21.4 psia. The physical properties of tap water at atmospheric conditions and Tulco Tech 80 oil are summarized in Table 16.1.

### 16.2.2.2 Test Conditions

The experimental test matrix is shown in Figure 16.4. The horizontal and vertical axes represent, respectively, $v_{SW}$ and $v_{SO}$, namely, the water and oil superficial velocities. Three different liquid superficial velocities, $v_{SL}\,(v_{SL} = v_{SW} + v_{SO})$, are used, namely 0.01, 0.02, and 0.03 m/s. The liquid superficial velocities are chosen to ensure that stratified gas-liquid flow occurs. Water cut values of 5, 10, 20, 30, and 40% were used for each superficial liquid velocity. Thus, a total of $5 \times 3 = 15$ data points were acquired for each superficial gas velocity. Five different superficial gas velocities, $v_{SG}$, i.e.: 0.3, 1.5, 3.0, 4.6 and 6.1 m/s were run, resulting in a total of $5 \times 15 = 75$ data points.

## *16.2.3 Experimental Results*

The experimental results include the observed flow configuration of the liquid-phase for all the runs given in the test matrix, namely, separated or dispersed oil and water flow. Also, the heights of the oil and water layers (for separated liquid-phase) or the liquid-phase height (for dispersed liquid-phase) are presented.

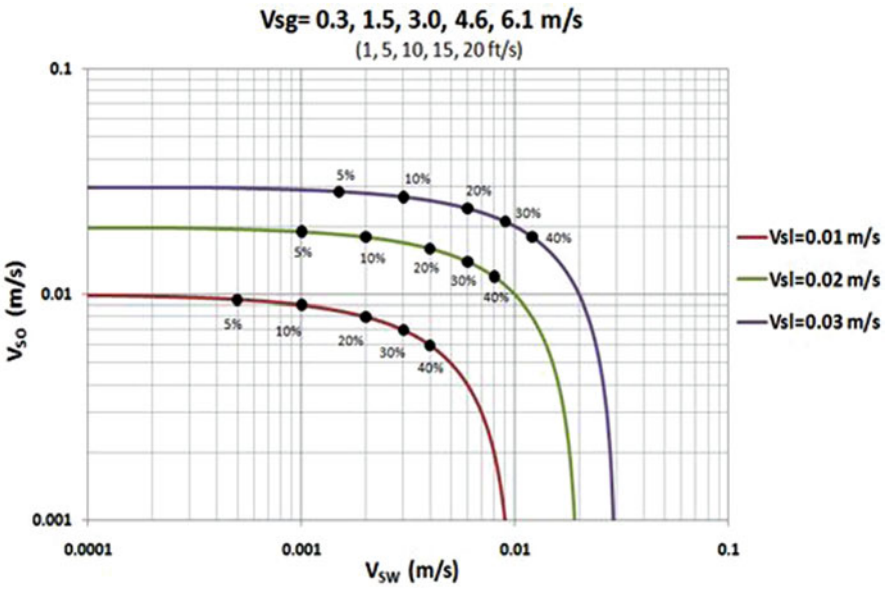


**Fig. 16.4** Experimental test matrix map.

#### 16.2.3.1 Flow Patterns

In this study, the flow patterns for three-phase stratified flow are defined according to the gas-liquid and oil-water interactions, as shown in Figure 16.5. The oil-water interaction has been classified into two cases, namely, separated or dispersed liquid-phase. The separated liquid-phase represents the condition where a water layer flows at the bottom of pipe and an oil layer flows on top of the water layer. On the other hand, the oil and water are completely mixed for the dispersed liquid-phase configuration. The gas-liquid interface is also considered in the flow pattern classification. For each of the liquid-phase cases, depending on the configuration of the gas-liquid interface, either stratified smooth or stratified wavy may occur. Thus, a total of four flow patterns are possible as shown in Figure 16.5.

#### 16.2.3.2 Experimental Results

The experimental results are presented in Figures 16.6(a) through 16.7(c), each of which is for a fixed superficial gas velocity. The flow patterns classified in the previous section are represented with different symbols and colors. The gas-liquid interaction is depicted as follows: diamonds represent stratified smooth and triangles represent stratified wavy gas-liquid interface. Colors are used to define the oil-water interaction. Red and black represent the separated and the dispersed liquid-phase, respectively. For instance, a data point represented by a red diamond

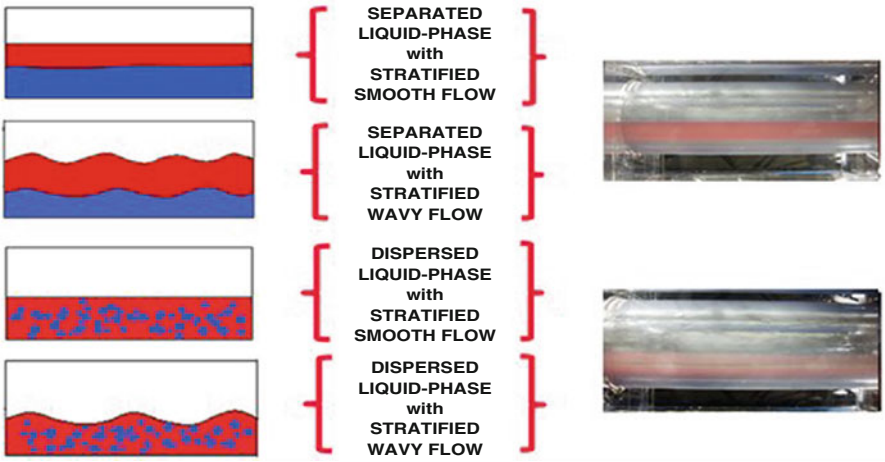


**Fig. 16.5** Schematic of three-phase stratified flow patterns: gas (white), oil (red), water (blue).

indicates that the liquid-phase is separated and the gas-liquid interface is smooth. As another example, black triangle stands for dispersed liquid-phase and wavy gas-liquid interface. The cross marker (x) represents inlet perturbation, which is defined later.

Figures 16.6(a) and 16.6(b) present the results for the low superficial gas velocities of 0.3 and 1.5 m/s, respectively. For the 0.3 m/s case, as shown in Figure 16.6(a), the oil and water are separated and the gas-liquid interface is smooth, namely, Separated-Liquid-Phase Stratified-Smooth flow occurs. Also, flow perturbations are observed for 30% and 40% water cuts with 0.02 m/s and 0.03 m/s superficial liquid velocities.

The inlet perturbations occur due to the vertical inlet section. When operating at low superficial gas velocities, liquid accumulates in the vertical section. Periodically the gas pushes the accumulated liquid from the vertical inlet section into the test section. This inlet perturbation creates a disturbance wave in the test section for high water cut values. For the of 0.3 and 1.5 m/s superficial gas velocity cases, just before the disturbance occurs, Separated-Liquid-Phase Stratified-Smooth flow is observed for these four perturbation data points. Therefore, it is expected that the inlet perturbed data are also separated liquid-phase, as are all the other data points for this case.

Similarly, the experimental results for 1.5 m/s superficial gas velocity are shown in Figure 16.6(b). The flow behavior for 1.5 m/s superficial gas velocity is similar to the behavior of the 0.3 m/s superficial gas velocity. For all flow conditions Separated-Liquid-Phase Stratified-Smooth flow occurs. Similarly, four perturbation runs occur at the same superficial liquid velocities and water cuts.

For the 3 m/s superficial gas velocity results, dispersed liquid-phase occurs at some flow conditions, as shown in Figure 16.7(a). The liquid-phase is dispersed for the lowest water cut, namely, 5% for 0.01, 0.02, and 0.03 m/s superficial

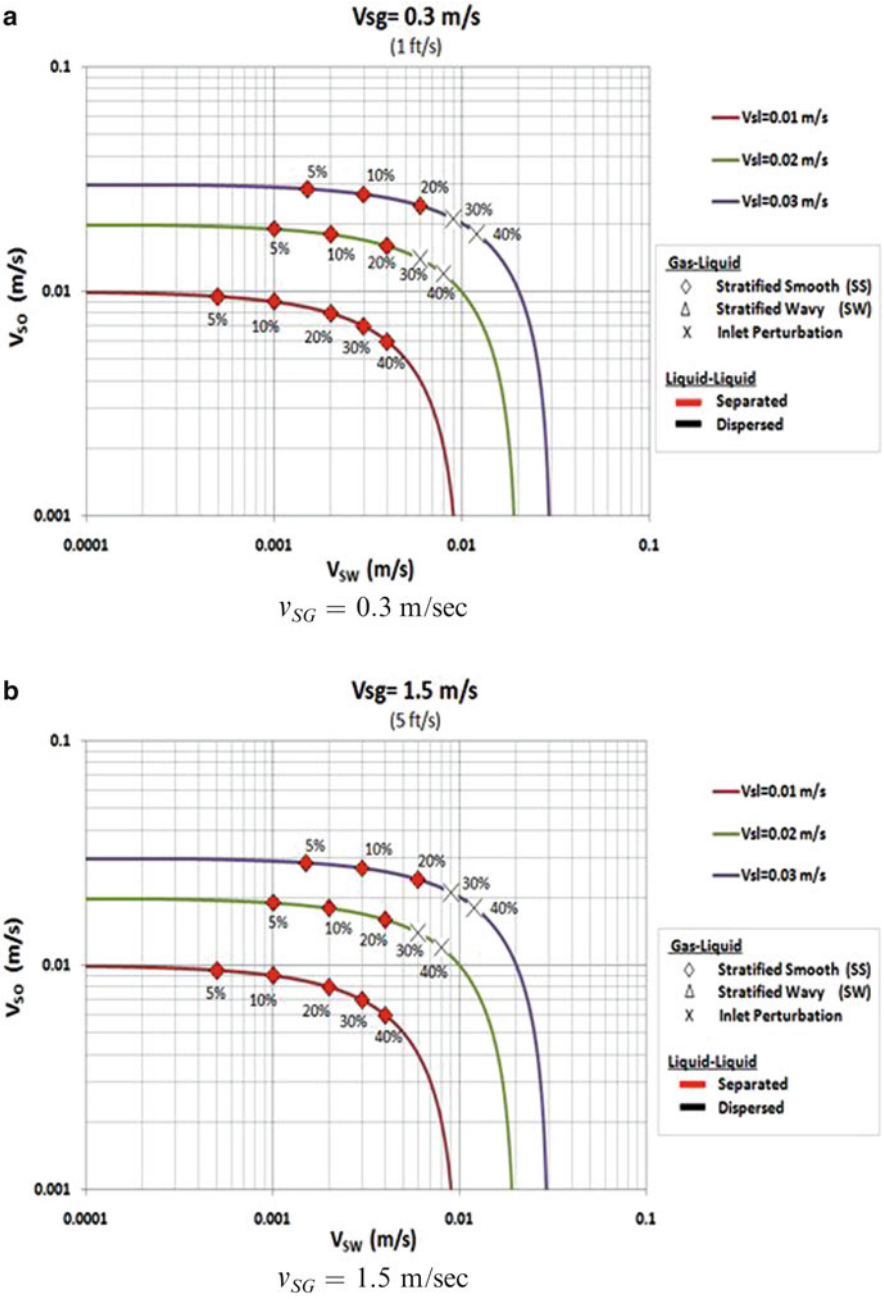


**Fig. 16.6** Observed three-phase flow patterns.

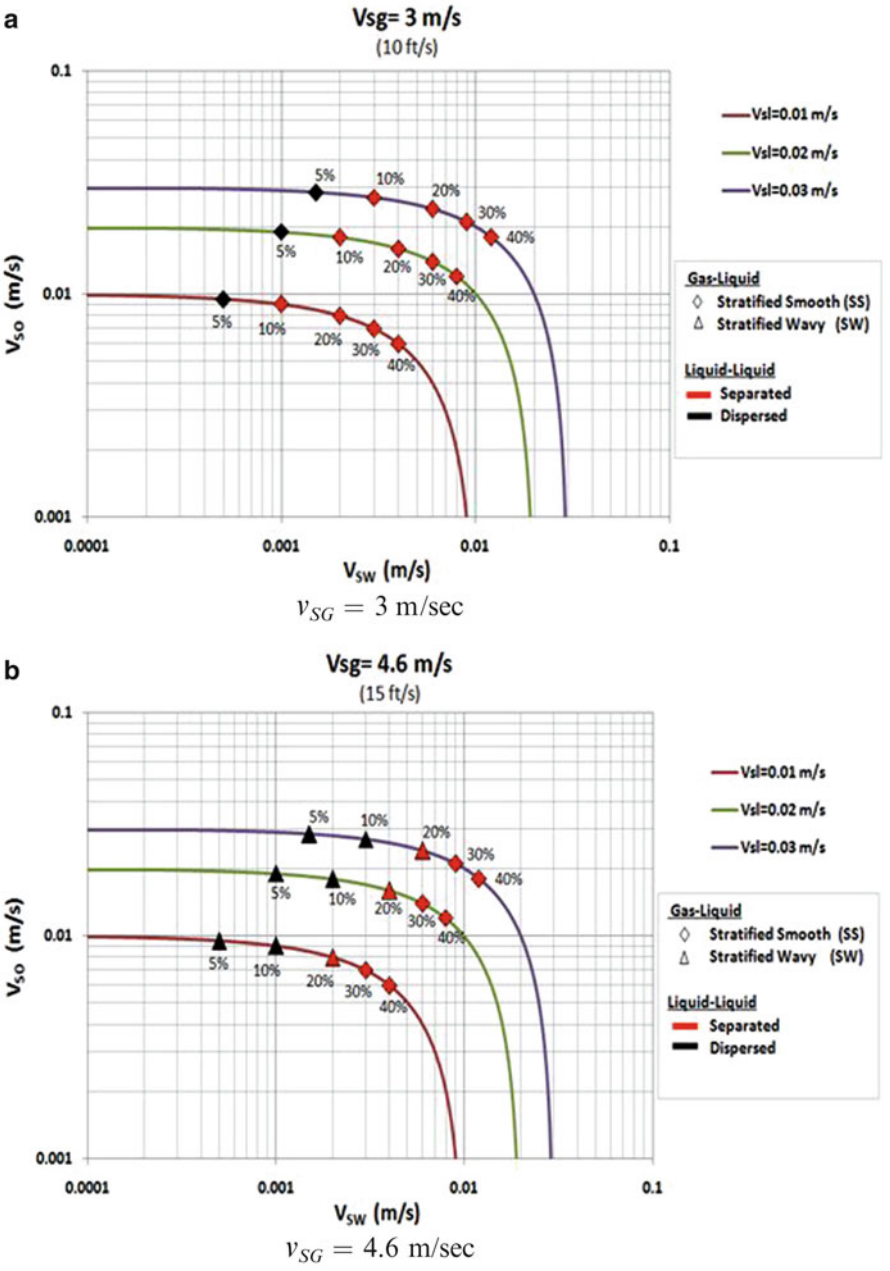


**Fig. 16.7** Observed three-phase flow patterns.

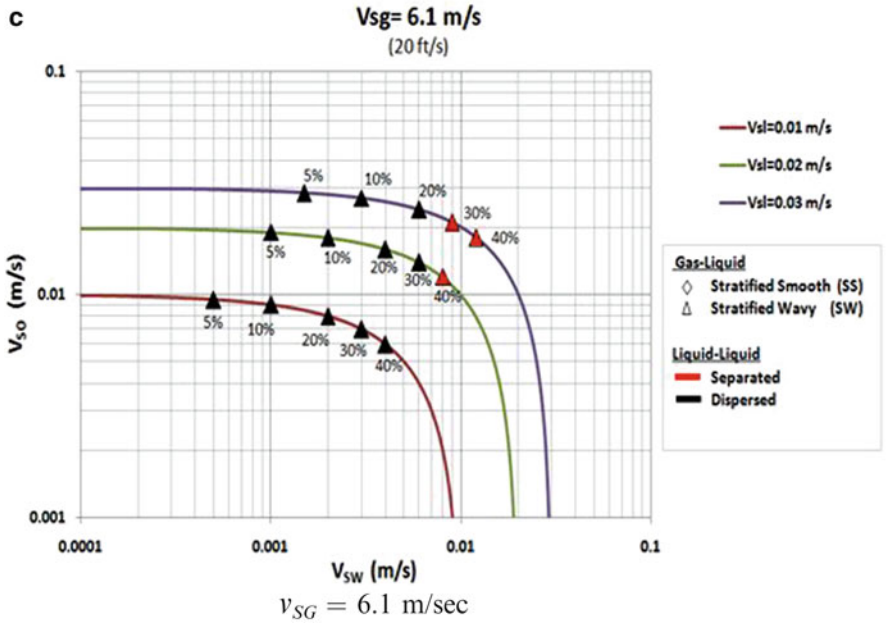


**Fig. 16.7** (continued)

liquid velocities. For higher water cuts, oil and water are separated from each other. The gas-liquid interface is still smooth for all the data points of this case. Moreover, no inlet perturbations are observed for the 3 m/s and higher superficial gas velocities. The results for 4.6 m/s superficial gas velocity can be seen in Figure 16.7(b). For 5% and 10% water cuts, the liquid-phase is dispersed for all three superficial liquid velocities. With increase in water cut, transition from dispersed liquid-phase to separated liquid-phase occurs. Another effect of increasing the water cut is observed at the gas-liquid interface. With 20% water cut the interface becomes wavy. However, when the water cut reaches 30%, the gas-liquid interface becomes smooth.

Figure 16.7(c) presents the experimental results for the highest superficial gas velocity of this study, namely, 6.1 m/s. For this case, the dispersed liquid phase region expands. The oil and water phases are separated only for 30% and 40% water cuts with 0.03 m/s superficial liquid velocity, and 40% water cut for 0.02 m/s liquid superficial velocity. For all other conditions the liquid phase is dispersed. The gas-liquid interface is wavy for all data points, and the wave frequency is higher, as compared to the lower gas velocity runs.

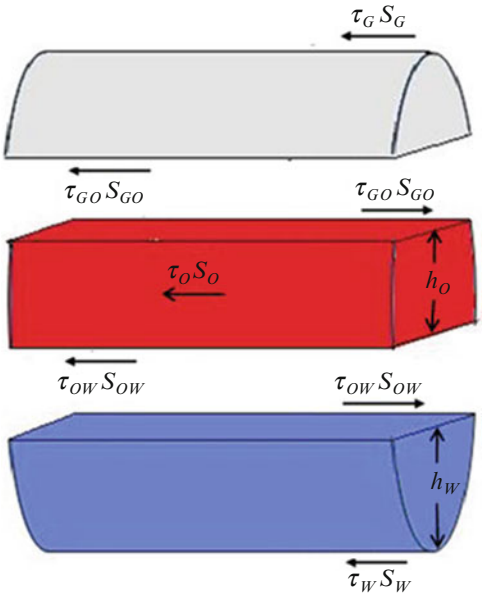


**Fig. 16.8** Schematic of forces acting on gas, oil, and water phases.

## 16.3 Model Development

This section presents the developed mechanistic model for predicting the transition between separated and dispersed liquid-phase under horizontal three-phase stratified flow conditions, namely, the onset to water layer. The model consists of two parts. The first part consists of the three-phase stratified flow model developed by Taitel et al. [TaBa94]. The second part of the model utilizes the results of the first part to develop a criterion for the transition between separated and dispersed liquid-phase conditions.

### *16.3.1 Three-Phase Stratified Flow Model*

The proposed model requires as input the three-phase stratified flow variables, which are determined based on the Taitel et al. [TaBa94] model. The Taitel et al. [TaBa94] model for separated three-phase stratified flow was developed by applying momentum balance equations for the gas, oil, and water phases. Figure 16.8 shows the acting forces on the three phases. Neglecting the rate of change of momentum (steady state), the momentum balance equation reduces to a force balance equation. The momentum (force) balance equations for inclined flow for the gas, oil, and water are given, respectively, by

$$-A_G\left(\frac{dp}{dL}\right)_G - \tau_G S_G - \tau_{GO} S_{GO} - \rho_G A_G g \sin\beta = 0, \tag{16.1}$$

$$-A_O\left(\frac{dp}{dL}\right)_O - \tau_O S_O - \tau_{OW} S_{OW} + \tau_{GO} S_{GO} - \rho_O A_O g \sin\beta = 0 \tag{16.2}$$

and

$$-A_W\left(\frac{dp}{dL}\right)_W - \tau_W S_W + \tau_{OW} S_{OW} - \rho_W A_W g \sin\beta = 0. \tag{16.3}$$

The momentum balance equation for the total liquid-phase (oil and water) can be obtained by summing the oil and water momentum equations. Adding Eqs. 16.2 and 16.3 yields

$$-A_L\left(\frac{dp}{dL}\right)_L - \tau_L S_L + \tau_{GO} S_{GO} - \rho_L A_L g \sin\beta = 0, \tag{16.4}$$

where $A_L = A_W + A_O$, $\rho_L = \frac{\rho_W A_W + \rho_O A_0}{A_L}$ and $\tau_L S_L = \tau_W S_W + \tau_O S_O$.

In Eq. 16.4, $A$ is cross sectional area, $\frac{dp}{dL}$ is pressure gradient, $\tau$ is the shear stress, $S$ is the perimeter, $\rho$ is the density, $g$ is the acceleration of gravity, and $\beta$ is the inclination angle. Gas, oil, water, and total liquid-phase are represented by subscripts $G$, $O$, $W$, and $L$, respectively. The subscripts $GO$ and $OW$ represent, respectively, the gas-oil and oil-water interfaces.

The cross-sectional areas and perimeters are calculated utilizing geometrical relationships based on the pipe diameter and the heights of the water layer, $h_W$, and oil layer, $h_O$ as shown in Figure 16.8. Refer to Shoham [Sh06] for these geometrical relationships. On the other hand, determination of the wall and interfacial shear stresses is more complex, which can be obtained by different correlations. The shear stresses between each phase and the pipe wall are determined as follows:

$$\tau_G = f_G \frac{\rho_G v_G^2}{2}, \qquad \tau_O = f_O \frac{\rho_O v_O^2}{2}, \quad \text{and} \quad \tau_W = f_W \frac{\rho_W v_W^2}{2}$$

The interfacial shear stresses are calculated using

$$\tau_{OW} = f_{OW} \frac{\rho_O (v_O - v_W) \cdot |v_O - v_W|}{2} \quad \text{and} \quad \tau_{GO} = f_{GO} \frac{\rho_G (v_G - v_O) \cdot |v_G - v_O|}{2}$$

The friction factors between the pipe wall and the gas, oil and water phases are calculated by the Blasius correlation (for smooth pipes), namely,

$$f = C \cdot \mathrm{Re}^{-n}$$

where Re is the Reynolds number and $C$ and $n$ are constants: $C = 0.046$ and $n = 0.2$ for turbulent flow and $C = 16$ and $n = 1$ for laminar flow.

The Reynolds numbers of the gas, oil, and water phases are

$$\mathrm{Re}_G = \frac{4 \cdot v_G A_G \rho_G}{(S_G + S_{GO})\,\mu_G} \qquad \mathrm{Re}_O = \frac{4 \cdot v_O A_O \rho_O}{S_O \mu_O} \quad \text{and} \quad \mathrm{Re}_W = \frac{4 \cdot v_W A_W \rho_W}{S_W \mu_W}$$

There are several correlations for the interfacial shear stress friction factor. Taitel et al. [TaBa94] followed the Cohen and Hanratty [CoHa68] correlation, as follows:

If $f_G < 0.014$, then $f_{GO} = 0.014$; otherwise, $f_{GO} = f_G$, and if $f_W < 0.014$, then $f_{OW} = 0.014$, otherwise $f_{OW} = f_W$, where $f_{GO}$ and $f_{OW}$ are the friction factors of gas-oil and oil-water interface and $f_G$ and $f_W$ are the gas-wall and water-wall friction factors.

Equating the pressure gradient terms in the gas and liquid momentum balance equations, namely, Equations 16.1 and 16.4, yields the combined momentum balance equation of the gas and liquid phases given by

$$-\frac{\tau_L S_L}{A_L} + \frac{\tau_G S_G}{A_G} + \tau_{GO} S_{GO} \left( \frac{1}{A_L} + \frac{1}{A_G} \right) - (\rho_L - \rho_G)\, g \sin\beta = 0. \tag{16.5}$$

Similarly, equating the pressure gradient terms in the oil and water momentum equations, which are given in Equations 16.2 and 16.3, results in a second combined momentum balance equation of the oil and water phases, namely,

$$-\frac{\tau_W S_W}{A_W} + \frac{\tau_O S_O}{A_O} - \frac{\tau_{GO} S_{GO}}{A_O} + \tau_{OW} S_{OW} \left( \frac{1}{A_W} + \frac{1}{A_O} \right) - (\rho_W - \rho_O)\, g \sin\beta. \tag{16.6}$$

The two combined momentum equations are implicit equations for the heights of the oil and water layers, $h_O$ and $h_W$, (see [Sh06]). Equations 16.5 and 16.6 must be solved simultaneously in order to obtain $h_O$ and $h_W$. Note that the height of the gas can be determined based on the pipe diameter.

### *16.3.2 Transition Between Separated and Dispersed Liquid-Phase*

The three-phase stratified flow model presented in the previous section is used to find the oil and water layer heights under a given set of flow conditions. However, these heights represent the equilibrium heights of the oil and water layers. Thus, the three-phase flow model does not address the liquid-phase flow behavior, which is the main objective of the current study. A simple mechanistic model is developed in this study for determining the transition between separated and dispersed liquid-phase under three-phase stratified flow.

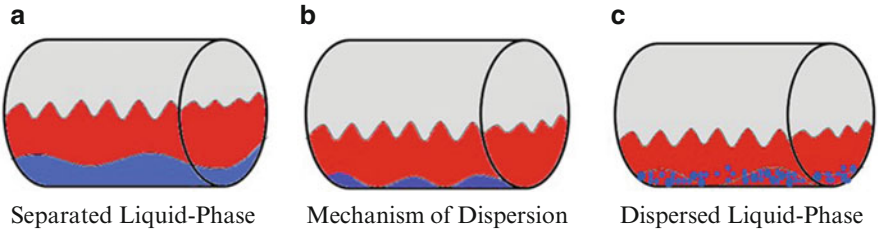


**Fig. 16.9** Liquid-phase dispersion transition mechanism.

#### 16.3.2.1 Transition Mechanism

The experimental results reveal that dispersion of the oil and water phases in three-phase stratified flow depends on the gas velocity and water cut. Increasing $v_G$ results in the occurrence of waves at the gas-liquid and oil-water interfaces. If the oil-water interfacial waves bridge the bottom of the pipe, they sweep the water layer and disperse it. Illustration of the dispersion mechanism is given in Figure 16.9. As shown in Figure 16.9(a), waves at oil-water interface may not be sufficiently large to bridge the pipe bottom. However, as shown in Figure 16.9(b), for low water cuts the waves reach the bottom of pipe, and sweep the liquid-phase, generating a dispersion. Figure 16.9(c) shows the dispersed liquid-phase with Stratified Wavy Flow. This observation is the basis for modeling the transition from separated liquid-phase to dispersed liquid-phase.

#### 16.3.2.2 Transition Criterion

Based on the physical phenomena presented in the previous section, a simple transition criterion for the liquid-phase is proposed, based on the Froude number. The Froude number has previously been used by several authors for different applications. Taitel and Dukler [TaDu76] utilized the Froude number for characterization of transition boundary between stratified to non-stratified flow in gas-liquid flow. The Froude number was also used by Petalas and Aziz [PeAz00] to determine the occurrence of waves in two-phase stratified flow. Hong et al. [Hong01] found that corrosion inhibitor films are washed away from the pipe surface under high Froude number condition. The Froude number is defined as the ratio of inertial forces to the gravitational forces, as given by

$$Fr^2 = \frac{\rho_G}{(\rho_W - \rho_G)} \cdot \frac{v_G^2}{g \cdot h_W \cos\beta} \tag{16.7}$$

where $v_G$ is actual gas velocity, $h_W$ is water height, $g$ is the acceleration due to gravity, $\beta$ is the inclination angle, and $\rho_G$ and $\rho_W$ are the gas and water densities, respectively. Note that $v_G$ is a function of the oil and the water heights (which are

outputs of the Three-Phase Stratified Flow Model solution) and $\rho_G$ is a function of pressure. Therefore, Eq. 16.7 is dependent on the liquid layer thickness and pressure.

The Froude number has been predicted by the proposed model for each of the experimental runs. It was found that for all cases where the liquid-phase was separated, the Froude number was less than $1.28 \pm 0.145$. On the other hand, for all cases where the liquid-phase was dispersed, the Froude number is equal or greater than $1.28 \pm 0.145$.

Thus, the criterion for the onset of water layer (separated liquid-phase) is given by

$$Fr^2 < 1.28 \pm 0.145 \tag{16.8}$$

The developed transition criterion was also calculated with the measured values of the variables in the Froude number. The height of the water layer, $h_W$, was measured directly and the gas velocity, $v_G$, was determined based on the measured gas-phase height, $h_G$. The calculated Froude number for this approach resulted in the same criterion as given in Eq. 16.8.

## 16.4 Results and Discussion

This section provides a comparison between the acquired data and model predictions for the transition boundaries between the separated and the dispersed liquid-phases. The transition boundary between separated liquid-phase and dispersed liquid-phase in three-phase stratified flow, or onset to liquid layer, is predicted based on the Froude number criterion approach. Figures 16.10(a), 16.10(b), and 16.10(c) show the transition boundaries, which are represented by red dashed lines, and the experimental data for 6.1 m/s, 4.6 m/s and 3.1 m/s superficial gas velocities, respectively. Because the predicted Froude number is less than 1.28 for the .3 m/s and 1.5 m/s superficial gas velocity cases, the predicted liquid-phase configuration is always separated. Therefore, for these cases, the transition line does not exist, as confirmed by the experimental results. As shown in Figure 16.10(a), the predicted transition boundary for the 6.1 m/s superficial gas velocity case accurately separates the dispersed liquid-phase data points from the separated liquid-phase data points. All the data points on the left of transition boundary are indeed separated liquid-phase, while the points on the right of transition line are dispersed liquid-phase. A comparison between model predictions and experimental data for the 4.6 m/s superficial gas velocity case is shown in Figure 16.10(b). The predicted transition boundary between separated and dispersed liquid-phase passes around 12% water cut, showing a good agreement with the experimental results. Figure 16.10(c) shows a similar comparison for the 3 m/s superficial gas velocity runs. Although the line does not pass between the separated liquid-phase and dispersed liquid-phase data points, it passes very close to the separated liquid-phase boundary. The transition

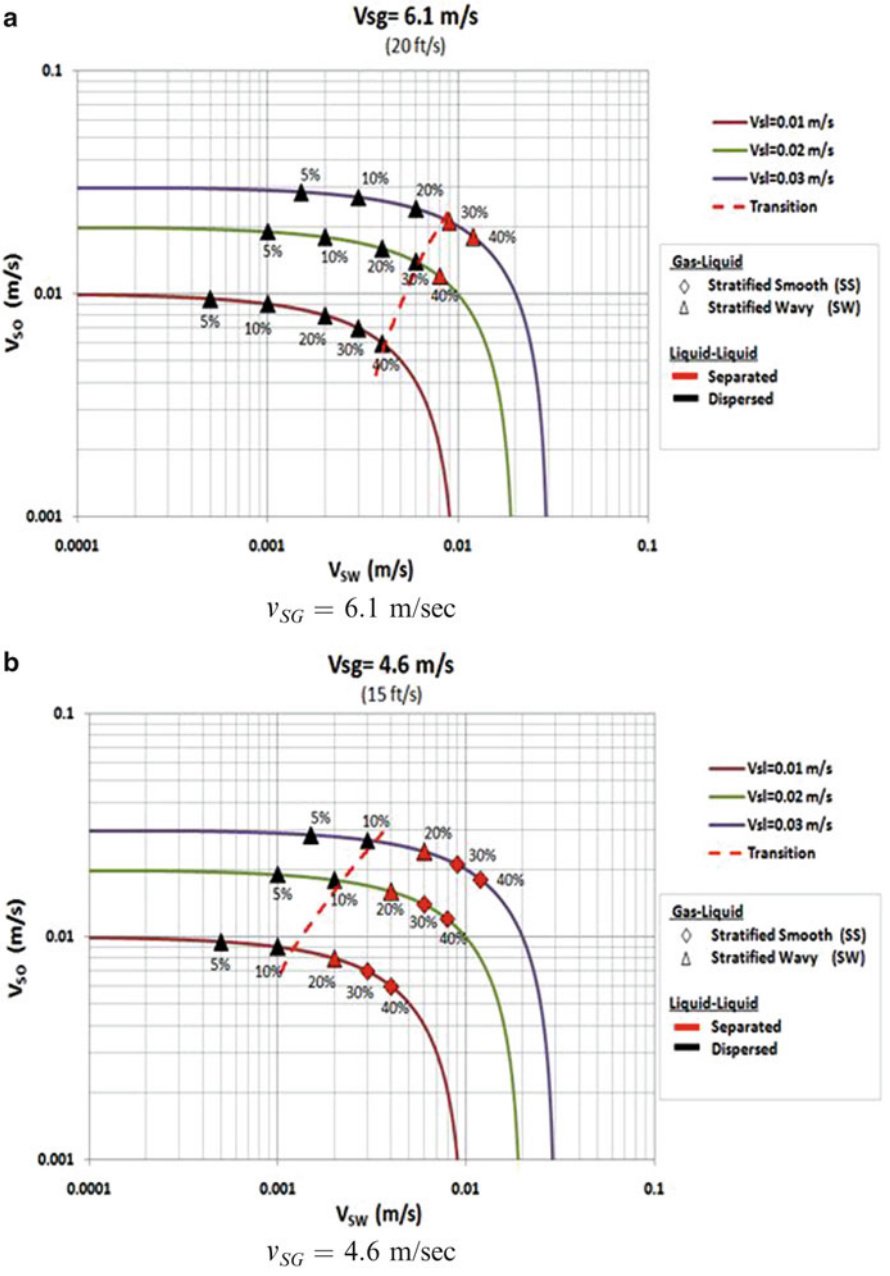


**Fig. 16.10** Comparison between model prediction and experimental data.

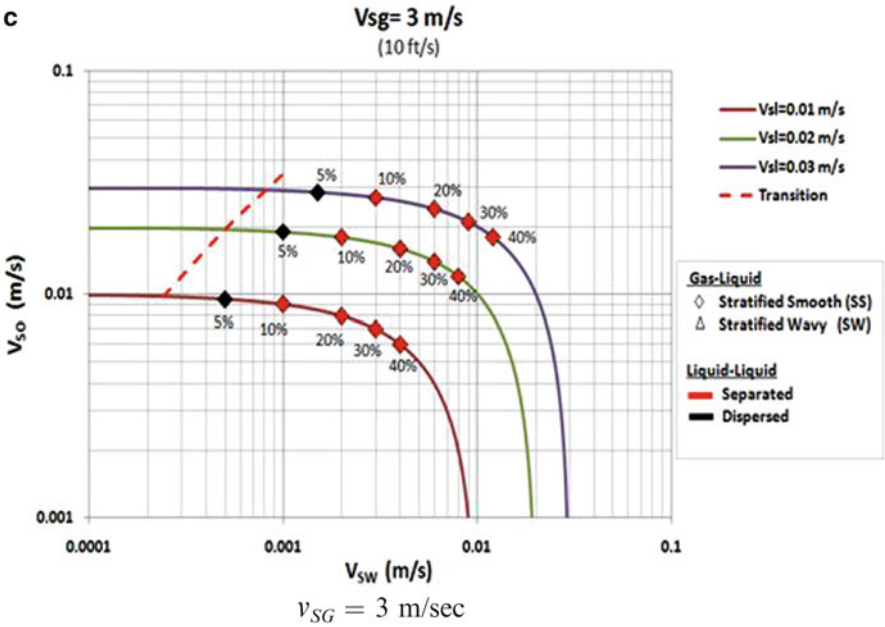


**Fig. 16.10** (continued)

line occurs around 3% water cut while the observed transition occurs at 5% water cut values, which constitutes a fair agreement. This slight difference is due to the uncertainty of the water height at low water cuts.

## 16.5 Conclusions

A total of 75 experimental runs under horizontal gas-oil-liquid three-phase stratified flow were conducted varying the water cut between 5% and 40%. The experimental results provide the transition boundary between the separated liquid-phase and the dispersed liquid-phase flow configurations, namely, the onset of water layer. A mechanistic model was developed for the prediction of the transition boundary between the separated and dispersed liquid-phase under three-phase stratified flow. The model predictions of the transition boundary between the separated and dispersed liquid-phase flow show good agreement with the acquired experimental data.

## References

[CoHa68] Cohen, S. L., Hanratty, T. J.: Effects of Waves at a Gas-Liquid Interface on a Turbulent Air Flow. Journal of Fluid Mechanics, **31**, 467–469 (1968)

[Er10] Er, M. O.: Onset to Separated Water-Layer in Three-Phase Stratified Flow. M.S Thesis, The University of Tulsa, (2010)

[Hong01] Hong, T., Sun, Y. H., Jepson, W. P.: Study on Corrosion Inhibitor in Large Pipelines under Multiphase Flow using EIS. NSF I/UCRC Corrosion in Multiphase System Center, Institute for Corrosion and Multiphase Technology, Ohio University, Ohio, USA, (2001)

[PeAz00] Petalas, N., Aziz, K.: A Mechanistic Model for Multiphase Flow in Pipes. Journal of Canadian Petroleum Technology, **39**, 43–55 (2000)

[Sh06] Shoham, O.: Mechanistic Modeling of Gas-Liquid Two-Phase Flow in Pipes. SPE Books, (2006)

[TaBa94] Taitel, Y., Barnea, D., Brill, J. P.: Stratified Three-Phase Flow in Pipes. International Journal of Multiphase Flow, **21**, 53–60 (1994)

[TaDu76] Taitel, Y., Dukler, A. E.: A model for predicting flow regime transitions in horizontal and near horizontal gas-liquid flow. AIChE Journal, **22**, 47–55. (1976)

# Chapter 17
# An Integro-Differential Equation for 1D Cell Migration

**C. Etchegaray, B. Grec, B. Maury, N. Meunier, and L. Navoret**

## 17.1 Introduction

Cell migration is a fundamental biological phenomenon involved, for example, in development, wound healing, cancer, and immune response. Understanding its key features is therefore a burning issue.

Some cells can move on an adherent substrate by a *crawling* process, where motion comes from the formation of finger-like extensions named *filopodia* that adhere to the substrate for some time (see Fig. 17.1). When the cell contracts, non-adherent filopodia retract, whereas adherent ones exert forces that induce motion. We refer to [AnEh07] for a complete description of cell crawling.

When a cell is set on a flat homogenous substrate, it performs a random-like motion. However, it can also become polarized and move in a preferential direction. How this direction is chosen is a question that is still driving many experimental and modeling efforts. In [CaVoRi14], a nonhomogenous substrate imposes geometrical constraints that are sufficient to direct 1D cell motion. This paper focuses on the 1D motion brought forth by the filopodial activity.

We introduce a simplified model of 1D cell migration (see Fig. 17.1) relying on the filopodial activity. In what follows, the substrate is supposed to be flat and homogenous, but more complex settings could also be described. Let us consider

C. Etchegaray (✉) • B. Maury
Université Paris-Sud, 15 Rue Georges Clémenceau, 91400 Orsay, France
e-mail: christele.etchegaray@math.u-psud.fr; bertrand.maury@math.u-psud.fr

B. Grec • N. Meunier
Université Paris Descartes, Paris, France
e-mail: berenice.grec@parisdescartes.fr; nicolas.meunier@parisdescartes.fr

L. Navoret
Université de Strasbourg, Strasbourg, France
e-mail: laurent.navoret@math.unistra.fr



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_17

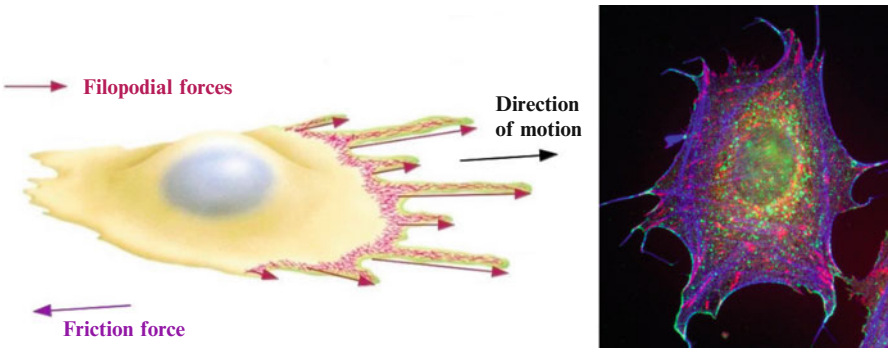


**Fig. 17.1** Illustration of a moving cell [LoEtAl02], and picture of a fibroblast [MaEtAl08].

the center of mass of the cell, whose position at time $t$ is denoted $x(t) \in \mathbb{R}$. Force equilibrium leads to

$$C\frac{dx}{dt}(t) = -F_\ell(t) + F_r(t), \tag{17.1}$$

where $F_r \geq 0$ (resp. $F_\ell \geq 0$) is the force exerted by filopodia located on the right (resp. on the left) of the cell and $C$ is the friction parameter, that can be set equal to $1\,nN.h.\mu m^{-1}$ [WoTa11].

We can now focus on the forces $F_{r,\ell}$. Following biological knowledge, we assume that the forces exerted by filopodia on the cell at time $t$ depend on

- densities of filopodia sent to the right and left, denoted by $\psi_{r,\ell} > 0$,
- their existence time, fixed by the lifetime function $\mathscr{P} : \mathbb{R}^+ \to \mathbb{R}_+$,
- the force $f_{r,\ell}$ exerted by one filopodium on the cell, related to its orientation. Moreover, we assume that $f_{r,\ell} = f_{r,\ell}(x(t'), x(t))$ depends on the positions of both the tip of the filopodium, related to the cell position at creation time $t'$, and the actual cell position.

Consequently, equation (17.1) rewrites as an integro-differential equation

$$\begin{aligned} &\frac{dx}{dt}(t) = \int_0^t \mathscr{P}(a)\Big(\psi_r f_r(x(t-a), x(t)) - \psi_\ell f_\ell(x(t-a), x(t))\Big)da, \\ &x(0) = x_0, \end{aligned} \tag{17.2}$$

where $\psi_{r,\ell}$ are positive constants, $x : \mathbb{R}_+ \to \mathbb{R}$, and $f_{r,\ell} : \mathbb{R}^2 \to \mathbb{R}$. Let us assume for simplicity that $x_0 = 0$.

Problem (17.2) can be treated more or less easily depending on the force functions $f_{r,\ell}$. In this work, we shall first investigate one case of nonlinear elastic force, where only existence and uniqueness of a solution can be proved. Then, we shall consider a simplified case of linear force functions, where a linear Volterra

equation can be obtained. We shall see how this formalism allows us to get more information on the sign, boundedness, and asymptotic behavior of the solution in general, as well as explicit solutions for some special cases.

## 17.2 Nonlinear Force Functions

Let us start with the force functions

$$f_r(y,x) = k\,[\ell - (x-y)]_+, \text{ and } f_\ell(y,x) = k\,[\ell - (y-x)]_+,$$

where $[.]_+$ denotes the positive part function and $k, \ell \in \mathbb{R}_+$ are two constants. Taking $f_r(x(t-a),x(t))$ and $f_\ell(x(t-a),x(t))$, it corresponds to the hypothesis of filopodia having a constant size $\ell$, and exerting elastic forces as long as the cell at position $x(t)$ has not reached their tips $x(t-a) \pm \ell$. Equation (17.2) now writes

$$\begin{aligned}\frac{dx}{dt}(t) = k\int_0^t \mathscr{P}(a)\Big(\psi_r[\ell + x(t-a) - x(t)]_+ \\ - \psi_\ell\,[\ell + x(t) - x(t-a)]_+\Big)da.\end{aligned} \tag{17.3}$$

### *17.2.1 Existence and Uniqueness*

We prove the following result :

**Theorem 1.** *For $\mathscr{P} \in L^1(\mathbb{R}_+)$, there exists a unique solution $x \in \mathscr{C}^1(\mathbb{R}_+,\mathbb{R})$ of (17.3).*

*Proof.* After integration, equation (17.3) writes

$$\begin{aligned}x(t) = k\int_0^t\int_0^s \mathscr{P}(a)\Big(\psi_r\big[x(s-a) + \ell - x(s)\big]_+ \\ - \psi_\ell\big[x(s) - x(s-a) + \ell\big]_+\Big)dads =: \Phi(x)(t)\end{aligned}$$

with

$$\begin{aligned}\Phi : \big(\mathscr{C}([0,T],\mathbb{R}), \|\ \|_\infty\big) &\longrightarrow \big(\mathscr{C}([0,T],\mathbb{R}), \|\ \|_\infty\big) \\ x &\longmapsto \Phi(x) = (t \mapsto \Phi(x)(t)),\end{aligned}$$

for some $T \geq 0$. We are looking for existence and uniqueness of a fixed point for $\Phi$. Let us construct a sequence $(x^n)_{n\geq 0}$ in $\mathscr{C}([0,T],\mathbb{R})$ such that

$$x^0 \equiv x_0, \quad x^{n+1} = \Phi(x^n)\ \forall n \geq 0.$$

As $[0,T]$ is compact, $\big(\mathscr{C}([0,T],\mathbb{R}), \| \ \|_\infty\big)$ is a Banach space and we can use the Banach fixed-point theorem. All we need to show now is that $\Phi$ is a contraction mapping. Considering $(y,z) \in \big(\mathscr{C}([0,T],\mathbb{R}), \| \ \|_\infty\big)^2$ and denoting

$$g_{s,a}(y) = y(s-a) + \ell - y(s), \ \text{ and } \ h_{s,a}(y) = y(s) - y(s-a) + \ell,$$

we have

$$\begin{aligned}
\|\Phi(y) - \Phi(z)\|_\infty = \sup_{t\in[0,T]} \Bigg| k \int_0^t \int_0^s \mathscr{P}(a) \Big( \psi_r\big([g_{s,a}(y)]_+ - [g_{s,a}(z)]_+\big) \\
- \psi_\ell\big([h_{s,a}(y)]_+ - [h_{s,a}(z)]_+\big)\Big) da ds \Bigg| \\
\leq kT \sup_{s\in[0,T]} \int_0^s |\mathscr{P}(a)| \times \Big( \psi_r \big|[g_{s,a}(y)]_+ - [g_{s,a}(z)]_+\big| \\
+ \psi_\ell \big|[h_{s,a}(y)]_+ - [h_{s,a}(z)]_+\big|\Big) da,
\end{aligned}$$

since $\psi_{r,\ell} \geq 0$. Denote $\psi := \psi_r + \psi_\ell$. Now, for $(A,B) \in \mathbb{R}^2$, the inequality $\big|[A]_+ - [B]_+\big| \leq \big|A - B\big|$ holds, leading to

$$\begin{aligned}
\| \Phi(y) - \Phi(z) \|_\infty &\leq kT \sup_{s\in[0,T]} \int_0^s \psi\, |\mathscr{P}(a)|\, \big|(y-z)(s-a) - (y-z)(s)\big| da, \\
&\leq 2kT\psi \, \| \mathscr{P} \|_{L^1(\mathbb{R}_+)} \| \, y - z \, \|_\infty .
\end{aligned}$$

For $T$ small enough such that $2k\psi \, \| \mathscr{P} \|_{L^1(\mathbb{R}_+)} \, T < 1$, we deduce that $\Phi$ is a contraction mapping. As a consequence of the Banach fixed-point theorem, there exists a unique $x \in \mathscr{C}([0,T],\mathbb{R})$ which is solution of (17.3).

Iterating the same reasoning on time intervals of size $T$, one can extend this result to prove that (17.3) admits a unique solution $x \in \mathscr{C}(\mathbb{R}_+,\mathbb{R})$. Finally, using (17.3), it is clear that $x \in \mathscr{C}^1(\mathbb{R}_+,\mathbb{R})$, and this concludes the proof.

### *17.2.2 Numerical Simulations*

We consider the lifetime function $\mathscr{P} : a \mapsto e^{-a}$. This means that a density of filopodia will exponentially decrease with time, as more and more filopodia will

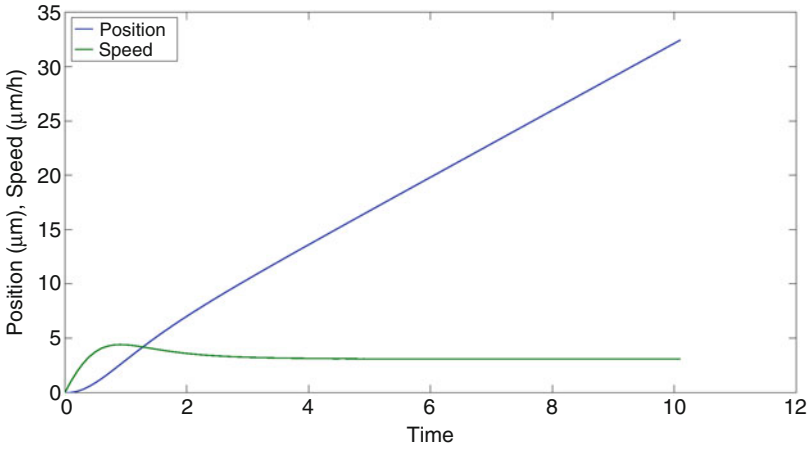


**Fig. 17.2** Numerical simulation of a particle speed and trajectory during $T = 10h$, for $dt = 10^{-2}h$, $C = 1\,nN.h.\mu m^{-1}$, $k = 1\,nN.\mu m^{-1}$, $\ell = 20.5\mu m$, and $(\psi_r, \psi_\ell) = (1.5, 1)$.

have disappeared. For the filopodia's length, we use an experimental value from [CaVoRi14]. Moreover, we impose a bias on the densities of filopodia ($\psi_r > \psi_\ell$).

Figure 17.2 represents the cell position and velocity over time computed with an explicit Euler time discretization and a rectangle integration method of equation (17.3). What can be observed is that the bias in the produced forces seems to lead to a nonzero asymptotic velocity. Further simulations with different parameter values and/or lifetime functions confirm this tendency.

## 17.3 Linear Forces

The presence of the positive part function in the previous case prevents getting analytical properties of the solution. In this section, we will take the following linearized forces functions:

$$f_r(x(t-a), x(t)) = k\,(x(t-a) + \ell - x(t))\,,$$
$$f_\ell(x(t-a), x(t)) = k\,(x(t) - x(t-a) + \ell)\,,$$

with $k, \ell \in \mathbb{R}_+$. This assumption is less relevant from the modeling point of view, since if the cell overtakes the tip of a filopodium, then it will experience a force in the opposite direction. However, for $k$ small enough, and an appropriate lifetime function, we can assume that the cell is slow enough so that it does not reach any existing filopodium tip.

Equation (17.2) can be written as a linear Volterra equation, which will lead to more analytical results.

### 17.3.1 Linear Volterra Equation Formalism

Let us rewrite equation (17.2) as

$$\begin{aligned} v(t) &= k\int_0^t \mathscr{P}(a)\Big(\psi_r(x(t-a)+l-x(t)) - \psi_\ell(x(t)-x(t-a)+\ell)\Big)da \\ &= kl(\psi_r-\psi_\ell)\int_0^t \mathscr{P}(a)da + k\psi\int_0^t \mathscr{P}(a)\left(x(t-a)-x(t)\right)da. \end{aligned} \tag{17.4}$$

Denoting $Q : t \mapsto \displaystyle\int_0^t \mathscr{P}(a)da$ and integrating by parts, we obtain

$$v(t) = k\ell(\psi_r-\psi_\ell)Q(t) + k\psi\left(\int_0^t Q(a)v(t-a)da - Q(t)x(t)\right),$$

since $Q(0)=0$ and $x(0)=0$. After the change of variable $s=t-a$, we get

$$v(t) = f(t) - k\psi\int_0^t (Q(t)-Q(t-s))v(s)ds, \tag{17.5}$$

$$\text{with} \quad f(t) = k\ell(\psi_r-\psi_\ell)Q(t). \tag{17.6}$$

which is a linear Volterra integro-differential equation for $v$.

### 17.3.2 Existence and Uniqueness of a Solution

With similar arguments to Theorem 1, we can prove the following property:

**Theorem 2.** *Equation (17.5) admits a unique solution $v \in \mathscr{C}(\mathbb{R}_+,\mathbb{R})$ for any $\mathscr{P} \in L^1(\mathbb{R}_+)$.*

*Remark 1 (The resolvent formalism).* Let us define the operator

$$h \star v : t \mapsto \int_0^{+\infty} h(t,s)v(s)ds.$$

Equation (17.5) can then be written as a convolution-like equation:

$$\begin{aligned} &v(t) + (h\star v)(t) = f(t), \text{ with} \\ &h(t,s) = k\psi(Q(t)-Q(t-s))\mathbf{1}_{[0,t]}(s). \end{aligned}$$

Existence and uniqueness of a solution can be proved by showing that $h$ is a Volterra kernel of $L^\infty$ type. For more details, we refer to [GrLoSt90].

### 17.3.3 Sign and Boundedness Property

We now prove a result showing how important the function $f$ is in controlling the migration. Indeed, it captures no less than the range of forces exerted with $k$, aging, and the potential asymmetry $\psi_r - \psi_\ell$ in the formation of filopodia.

**Theorem 3.** *If $\mathscr{P}$ is positive and decreasing, then the solution to equation (17.5) satisfies*

$$\begin{aligned} \psi_r \geq \psi_\ell \quad &\Rightarrow \quad \forall t \geq 0,\ v(t) \in [0, f(t)], \\ \psi_r \leq \psi_\ell \quad &\Rightarrow \quad \forall t \geq 0,\ v(t) \in [f(t), 0]. \end{aligned}$$

*Proof.* First, suppose that $\psi_r \geq \psi_\ell$. Consequently, $\forall t \geq 0$, $f(t) \geq 0$ and $f'(t) \geq 0$. By derivation of (17.5), we obtain

$$v'(t) = f'(t) - k\psi(Q(t) - Q(0))v(t) - k\psi \int_0^t (\mathscr{P}(t) - \mathscr{P}(t-s))v(s)ds,$$

with $f'(t) = k\ell(\psi_r - \psi_\ell)\mathscr{P}(t)$. Suppose there exists $t^*$ such that $\forall t < t^*$, $v(t) > 0$ and $v(t^*) = 0$, then

$$v'(t^*) = f'(t^*) - k\psi Q(t^*)v(t^*) - k\psi \int_0^{t^*} (\mathscr{P}(t^*) - \mathscr{P}(t^* - s))v(s)ds$$

is positive (since all the terms are positive). Consequently, $\forall t \geq 0$, $v(t) \geq 0$. This implies that $x(t-a) - x(t) \leq 0$, $\forall t \geq a \geq 0$. Going back to the equivalent equation (17.4), this shows that $\forall t \geq 0$, $v(t) \leq f(t)$. Now, let us assume that $\psi_\ell \geq \psi_r$, which means that $\forall t \geq 0$, $f(t) \leq 0$ and $f'(t) \leq 0$. In a similar way, if there exists $t^*$ such that $\forall t < t^*$, $v(t) < 0$ and $v(t^*) = 0$, then $v'(t^*)$ is negative. And considering again (17.4), we easily show that $\forall t \geq 0$, $v(t) \geq f(t)$, which concludes the proof.

### 17.3.4 Asymptotic Velocity

We now give an expression for the asymptotic velocity of the cell. Here again, $f$ has a crucial importance. The proof of the following result is similar to the one done in a forthcoming work [GrEtAl], and we do not repeat it here.

**Theorem 4.** *Let $v$ be the solution of (17.5), and denote $\gamma = \lim_{t\to+\infty} f(t)$. Assume that $v$ is uniformly continuous on $\mathbb{R}_+$. Then,*

$$v(t) \underset{t\to+\infty}{\longrightarrow} v_\infty = \begin{cases} \frac{\gamma}{1 + k\psi \int_0^{+\infty} a\mathscr{P}(a)da} & \text{if } a \mapsto a\mathscr{P}(a) \in L^1(\mathbb{R}_+), \\ 0 & \text{if } a \mapsto a\mathscr{P}(a) \notin L^1(\mathbb{R}_+). \end{cases}$$

Having two different cases can be interpreted as follows: if the mean lifetime of filopodia is finite, then the cell is permanently escaping from the action of older forces. As a consequence, if $\psi_r - \psi_\ell \neq 0$, it can get off its position all the time. However, if the mean lifetime of filopodia is infinite, all of them exert elastic forces on the cell, which will be stabilized in finite time.

### *17.3.5 Particular Cases*

Some choices of function $\mathscr{P}$ can give more explicit information on the solution.

#### 17.3.5.1 Infinite Existence Time of Forces ($\mathscr{P} \equiv 1$).

Taking $\mathscr{P} \equiv 1$, we are considering elastic forces that never disappear. Here, $\mathscr{P}$ does not fulfill the hypothesis of Theorem 2, but Theorem 4 applies. Equation (17.5) writes

$$v(t) = k\ell(\psi_r - \psi_\ell)t - k\psi \int_0^t sv(s)ds,$$

and can be solved after derivation with the variation of constants method, to give

$$v(t) = v(0)e^{-k\psi t^2/2} + \ell\sqrt{\frac{k\pi}{2}}(\psi_r - \psi_\ell)\sqrt{e^{-\frac{k\psi t^2}{2}}\left(1 - e^{-\frac{k\psi t^2}{2}}\right)}. \tag{17.7}$$

Figure 17.3 represents the solution $v$ as well as the corresponding forcing function $f$. As expected, the cell is stabilized in finite time. Moreover, we can

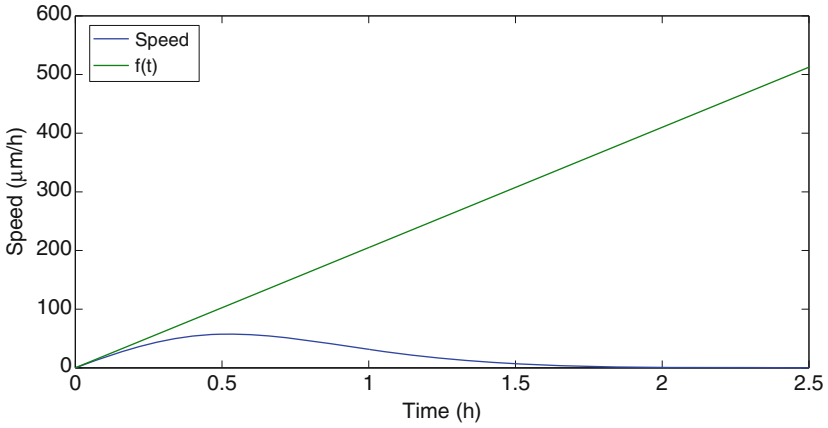


**Fig. 17.3** Graph of $t \mapsto v(t)$ and $t \mapsto f(t)$ for $\mathscr{P} \equiv 1$, with $C = 1nN.h.\mu m^{-1}$, $\ell = 20.5\mu m$, $k = 5nN.\mu m^{-1}$, $(\psi_r, \psi_\ell) = (6, 4)$ and $v(0) = 0$.

observe numerically the sign and boundedness property (given in Theorem 3), where in this case the $f(t)$ bound is optimal.

It is easy to check analytically the convergence of $v$ to $v_\infty = 0$ since we have the equivalence

$$v(t) \underset{t\to+\infty}{\sim} \ell\sqrt{\frac{k\pi}{2}}(\psi_r - \psi_\ell)e^{-k\psi t^2/4}.$$

#### 17.3.5.2 Exponential Decay ($\mathscr{P}(a) = e^{-a}$)

We now assume that $\mathscr{P}(a) = e^{-a}$, which was the function chosen in Section 17.2.2. All the results demonstrated before apply. However, we can actually directly solve the equation. Noting that $Q(t) = 1 - e^{-t}$, equation (17.5) becomes

$$v(t) = k\ell(\psi_r - \psi_\ell)(1 - e^{-t}) - k(\psi_r + \psi_\ell)e^{-t}A(t), \tag{17.8}$$
$$\text{with} \quad A(t) = \int_0^t (e^s - 1)v(s)ds.$$

**Proposition 1.** *The solution to* (17.8) *is given by*

$$v(t) = k\ell(\psi_r - \psi_\ell)\left(1 - e^{-t}\right) - k^2 l(\psi_r - \psi_\ell)\psi e^{-(k\psi+1)t + k\psi - k\psi e^{-t}}J(t), \tag{17.9}$$
$$\textit{with} \quad J(t) = \int_0^t (e^s + e^{-s} - 2)e^{k\psi(s-1+e^{-s})}ds.$$

*Proof.* Deriving $A$ with respect to time leads to

$$A'(t) = k\ell(\psi_r - \psi_\ell)(e^t + e^{-t} - 2) - k\psi(1 - e^{-t})A(t).$$

Now, by the variation of parameters method, we find that

$$A(t) = k\ell(\psi_r - \psi_\ell)e^{-k\psi(t-1+e^{-t})}\int_0^t (e^s + e^{-s} - 2)e^{k\psi(s-1+e^{-s})}ds,$$

leading to expression (17.9).

As a consequence, an asymptotic equivalent of the solution can be given.

**Theorem 5.** *The following equivalence holds:*

$$v(t) \underset{t\to+\infty}{\sim} k\ell(\psi_r - \psi_\ell)\left(1 - \frac{k\psi}{k\psi+1} - \frac{k\psi}{k\psi-1}e^{-2t} + 2e^{-t}\right), \tag{17.10}$$

*and v converges to the asymptotic velocity*

$$v_\infty := k\ell(\psi_r - \psi_\ell)\left(1 - \frac{k\psi}{k\psi+1}\right).$$

*Proof.* Using the following equivalence

$$\int_0^t e^{\alpha s} ds \underset{t\to+\infty}{\sim} \frac{e^{\alpha t}}{\alpha},$$

we easily obtain

$$J(t) \underset{t\to+\infty}{\sim} e^{-k\psi}\left(\frac{e^{(k\psi+1)t}}{k\psi+1} + \frac{e^{(k\psi-1)t}}{k\psi-1} - 2\frac{e^{k\psi t}}{k\psi}\right).$$

Considering expression (17.9), we have

$$v(t) \underset{t\to+\infty}{\sim} kl(\psi_r - \psi_\ell)\left[1 - k\psi e^{-(k\psi+1)t}\left(\frac{e^{(k\psi+1)t}}{k\psi+1} + \frac{e^{(k\psi-1)t}}{k\psi-1} - 2\frac{e^{k\psi t}}{k\psi}\right)\right],$$

which leads to the result.

We can then deduce that the asymptotic behavior of the cell depends on the range of filopodial forces $k$, on the global filopodial activity $\psi$, and on the asymmetry $\psi_r - \psi_\ell$. Moreover, the bigger $k$ and $\psi$ are, the faster the cell velocity reaches equilibrium. The non-trivial equilibrium is a consequence of the lifetime function that lets newer forces lead motion, whereas the older ones are 'silenced'. The initial asymmetry is then maintained over time. In Figure 17.4, a numerical simulation illustrates this behavior.

#### 17.3.5.3 Constant Existence Time ($\mathscr{P}(a) = 1_{[0,\tau]}(a)$)

Now, let us look at $\mathscr{P}(a) = 1_{[0,\tau]}(a)$ with $\tau > 0$, meaning that all filopodia exert forces during the same finite amount of time. For $t \geq 0$, we have

$$Q(t) = \begin{cases} t & \text{if } t < \tau, \\ \tau & \text{if } t \geq \tau. \end{cases}$$

Existence and uniqueness of a continuous solution to (17.5) comes from Theorem 2. Moreover, we can find an explicit solution for $t \leq \tau$, and bounds for the solution for $t \geq \tau$. Equation (17.5) then writes

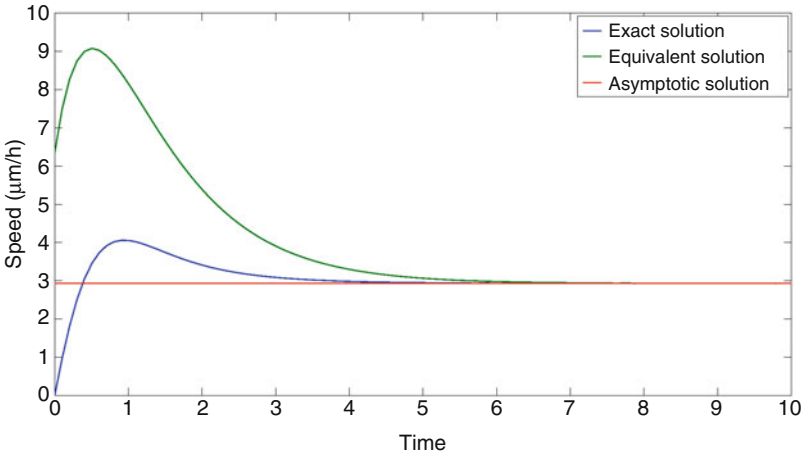


**Fig. 17.4** Numerical simulation of the exact solution, its equivalent at infinity and the asymptotic velocity for $C = 1\,nN.h.\mu m^{-1}$, $k = 1\,nN.\mu m^{-1}$, $\ell = 20.5\mu m$ and $(\psi_r, \psi_\ell) = (1.5, 1)$.

$$v(t) = k\ell(\psi_r - \psi_\ell)t - k\psi \int_0^t sv(s)ds, \qquad \text{for } t \leq \tau, \tag{17.11}$$

$$v(t) = k\ell\tau(\psi_r - \psi_\ell) - k\psi \int_{t-\tau}^t v(s)(\tau - (t-s))ds, \qquad \text{for } t \geq \tau. \tag{17.12}$$

**Theorem 6.** *The unique solution to equations (17.11)–(17.12) satisfies*

$$v(t) = \ell(\psi_r - \psi_\ell)\sqrt{\frac{k\pi}{2\psi}}\sqrt{(1 - e^{\frac{-k\psi}{2}t^2})} \qquad \textit{for } t \leq \tau, \tag{17.13}$$

$$v(\tau)\exp\left(-k\psi\frac{(t^2-\tau^2)}{2}\right) \leq v(t) \leq k\ell(\psi_r - \psi_\ell) \qquad \textit{for } t \geq \tau. \tag{17.14}$$

*Proof.* Let us first study the case where $t \leq \tau$. By derivation of (17.11), we have

$$v'(t) = k\ell(\psi_r - \psi_\ell) - k\psi tv(t),$$

and the variation of parameters method leads to expression (17.11).

Let us now consider the case where $t \geq \tau$. After a change of variable, equation (17.12) becomes

$$v(t) = k\ell\tau(\psi_r - \psi_\ell) - k\psi \int_0^\tau (\tau - s)v(t-s)ds, \tag{17.15}$$

$$= k\ell\tau(\psi_r - \psi_\ell) - k\psi \int_0^t (t-s)v(t-s)ds + h(t), \tag{17.16}$$

with

$$h(t) = k\psi \int_{\tau}^{t} (t-s)v(t-s)ds + k\psi \int_{0}^{\tau} (t-\tau)v(t-s)ds.$$

Since $\mathscr{P}$ is positive and decreasing, we deduce from Theorem 3 that $\forall t \geq 0$, $v(t) \geq 0$. Hence, we know that $h \geq 0$ on $[\tau, +\infty[$. Moreover, differentiating $h$ with respect to $t$, we obtain

$$\begin{aligned} h'(t) &= k\psi \left( \int_{\tau}^{t} \frac{d}{dt}((t-s)v(t-s))ds + \int_{0}^{\tau} \frac{d}{dt}((t-\tau)v(t-s))ds \right), \\ &= k\psi \left( (t-\tau)v(t) + (x(t) - x(t-\tau)) \right) \geq 0, \end{aligned}$$

as $v \geq 0$. Then, differentiating equation (17.16) leads to

$$v'(t) \geq -k\psi t v(t),$$

from which we deduce that

$$v(t) \geq v(\tau) \exp\left( -k\psi \frac{(t^2 - \tau^2)}{2} \right),$$

leading to the first inequality. Moreover, as $v \geq 0$, the second one is obtained from equation (17.15), and this concludes the proof.

## 17.4 Conclusions and Perspectives

In this chapter, we have introduced a simple deterministic model of 1D cell migration, based on the filopodial activity of the cell. It describes the formation of antagonist elastic forces by filopodia on each side of the cell.

This model is not able to describe realistic trajectories, as the filopodial activity is taken constant, but it relates explicitly filopodial statistics to the cell velocity and asymptotic behavior, and hence represents a first step in the global description of cell trajectories from filopodial activity.

In this work, we have studied a realistic case where filopodia stop exerting a force as soon as the cell overtook their tips. In this case, the highly nonlinear forces prevent us from getting more than an existence and uniqueness result.

The case of linear elastic forces is richer, as it gave more information about the sign, boundedness, and asymptotic behavior of the solution. It is important to keep in mind that the linear model is realistic only in a particular setting: if the cell is slow enough and filopodia's lifetime short enough, then they won't be reached by the cell. Typical velocity and filopodium lifetime are closely related to the cell type and experimental setting. Indeed, considering the force $k\ell$ exerted by a

filopodium of length $\ell$ on the substrate, it is known that $\ell$ is variable among cell types. Moreover, $k$ highly depends on the rigidity of the substrate: the more it is rigid, the larger the forces are [LoEtAl00]. Another key player in the filopodial forces is the adhesiveness of the substrate, that scales how strong it is coupled to filopodia, hence how large forces will be. However, a very adherent substrate is also less likely to let go of filopodia during the contraction of the cell, leading to a longer lifetime for them. This results in a bell-shaped curve relating velocity and adhesiveness, as described in [PaLoHo97]. As a consequence, it is likely that for a substrate of low (or very large) adhesiveness and low rigidity, cells velocity would be low enough so that the linear model fits with experimental conditions. This first-step model describing filopodial activity and trajectories is simple enough to give analytical information about the cell velocity, but still rich enough to be compared to different kinds of experimental 1D migration assays. In future works, it will be crucial to consider nonconstant densities of filopodia, to take into account the effect of motion itself on the filopodial activity. This would probably lead to much more realistic trajectories, where changes of direction would be described.

## References

[LoEtAl02] Lodish, Harvey et al.: Molecular Cell Biology, 4th ed., W.H. Freeman, 2002.

[GrLoSt90] Gripenberg, G., Londen, S. O., Staffans, O.: Volterra Integral and Functional Equations. 1st ed. Cambridge: Cambridge University Press, 1990.

[AnEh07] Ananthakrishnan, R., Ehrlicher, A.: The Forces Behind Cell Movement. Int J Biol Sci 2007; 3**5**:303–317.

[MaEtAl08] Machesky, L.,Simon, A., Bramble, J., Yeung, C., Mende, P.: Arp2/3 complex activity in filopodia of spreading cells. BMC Cell Biology 2008, **9**:65.

[CaVoRi14] Caballero, D., Voituriez, R., Riveline, D.: Protrusion Fluctuations Direct Cell Motion. Biophys J., **107**(1), 34–42 (2014)

[GrEtAl] Grec, B., Maury, B., Meunier, N., Navoret, L.: The role of ligands binding in shear induced leukocyte rolling. In preparation.

[LoEtAl00] Lo, C-M., Wang, H-B., Dembo, M., Wang, Y-L.: Cell Movement Is Guided by the Rigidity of the Substrate. Biophys J., **79**(1), 144–152 (2000)

[WoTa11] Wong, H.C., Tang, W.C.: Finite element analysis of the effects of focal adhesion mechanical properties and substrate stiffness on cell migration. J. Biomech., **44**, 1046–1050 (2011)

[PaLoHo97] Palecek, S.P., Loftus, J.C., Horwitz, A.F.: Integrin-ligand binding properties govern cell migration speed through cell-substratum adhesiveness. Nature, **385**, 537–540 (1997)

# Chapter 18
# The Multi-Group Neutron Diffusion Equation in General Geometries Using the Parseval Identity

**J.C.L. Fernandes, F. Oliveira, B.E.J. Bodmann, and M.T.B. Vilhena**

## 18.1 Multi-Group Steady State Diffusion in General Geometry

Our starting point is the steady state multi-energy group neutron diffusion equation, with the usual diffusion, removal, out-scattering, fission and in-scattering terms as presented in [Bd10] and solved in cylindrical geometry in [Fe12]. Here $D_g$ is the diffusion coefficient for energy group $g$, for general geometry we have

$$-D_g \Delta_\gamma \phi_g + \left( \Sigma_{ag} + \sum_{\substack{g'=1 \\ g' \neq g}}^{G} \Sigma^s_{g \to g'} \right) \phi_g = \chi_g \sum_{g'=1}^{G} \nu \Sigma_{fg'} \phi_{g'} + \sum_{g'=1}^{G} \Sigma_{g' \to g} \phi_{g'}$$

where $\Delta_\gamma = x^{-\gamma} \partial_x \left( x^\gamma \partial_x \right)$. This operator represents for each $\gamma$ the suitable geometry of a multi-group neutron diffusion problem. Here, $\Sigma_{Rg} = \Sigma_{ag} + \sum_{g'=1}^{G} \Sigma^s_{g \to g'}$ (for $g' \neq g$) are the respective removal cross section, $\Sigma_{g \to g'}$, $\Sigma_{g' \to g}$ the out- and in-scattering cross sections represented by $\Sigma_{gg'}$ and $\Sigma_{gg'}$, respectively. Further, $\nu \Sigma_{fg}$ is the fission cross section times the average neutron yield per fission, $\chi_g$ the spectral weight of energy group $g \in [1, G]$ and we add a generic source term $S_g$ per energy group.

J.C.L. Fernandes (✉) • F. Oliveira • B.E.J. Bodmann • M.T.B. Vilhena
Federal University of Rio Grande do Sul, Rua Sarmento Leite,
425, 90050-170 Porto Alegre, RS, Brazil
e-mail: julio.lombaldo@ufrgs.br; fernando.rodrigues@ufrgs.br; bardo.bodmann@ufrgs.br;
vilhena@math.ufrgs.br



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_18

The nonhomogeneous system is

$$-D_g\Delta_\gamma\phi_g+\Sigma_{Rg}\phi_g=\chi_g\nu\Sigma_{fg'}\phi_{g'}+\Sigma_{gg'}\phi_{g'}+S_g \tag{18.1}$$

Let $x_\gamma$ be the variable in the coordinate system generated by $\gamma$ and $X_\gamma$ the space such that $x_\gamma\in X_\gamma$. To solve the initial system, we define the general integral transform that takes $X_\gamma$ into $\overline{X}_\gamma$:

$$\overline{f}(\bar{x}_\gamma)=\mathscr{I}_\gamma[f(x_\gamma);x_\gamma\to\bar{x}_\gamma]=\int_{X_\gamma}f(x_\gamma)K_\gamma(x_\gamma,\bar{x}_\gamma)\,dx_\gamma$$

where $K_\gamma$ is the respective kernel of the integral transform with two remarkable properties, namely

i) $I_\gamma[\kappa(f(x_\gamma)+g(x_\gamma))]:=\kappa\overline{f}(\bar{x}_\gamma)+\kappa\overline{g}(\bar{x}_\gamma), f,g\in X_\gamma$ and $\kappa\in\mathbf{R}$;
ii) $I_\gamma[\Delta_\gamma f(x_\gamma)]:=h(\bar{x}_\gamma)\overline{f}(\bar{x}_\gamma)$.

where $h(\bar{x}_\gamma)$ is an algebraic function that depends on $\bar{x}_\gamma$. Now, applying $I_\gamma$ in the system (18.1), we get

$$-D_gh(\bar{x}_\gamma)\overline{\phi}_g(\bar{x}_\gamma)+\Sigma_{Rg}\overline{\phi}_g(\bar{x}_\gamma)=\chi_g\nu\Sigma_{fg'}\overline{\phi}_{g'}(\bar{x}_\gamma)+\Sigma_{gg'}\overline{\phi}_{g'}(\bar{x}_\gamma)+\overline{S}_g \tag{18.2}$$

Without loss of generality, we consider the case with two energy groups. After application of $\mathscr{I}_\gamma$ in (18.2) we obtain

$$\begin{bmatrix}-D_1h(\bar{x}_\gamma)+\Sigma_{R1} & -(\chi_1\nu\Sigma_{f2}+\Sigma_{12})\\ -(\chi_2\nu\Sigma_{f1}+\Sigma_{21}) & -D_2h(\bar{x}_\gamma)+\Sigma_{R2}\end{bmatrix}\begin{bmatrix}\overline{\phi}_1\\ \overline{\phi}_2\end{bmatrix}=\begin{bmatrix}\overline{S}_1\\ \overline{S}_2\end{bmatrix} \tag{18.3}$$

For convenience, we define the constant $\mu_{gg'}=\chi_g\nu\Sigma_{fg'}+\Sigma_{gg'}$ and the function $A_g(\bar{x}_\gamma)\in\overline{X}_\gamma$, with $A_g(\bar{x}_\gamma)=-D_gh(\bar{x}_\gamma)+\Sigma_{Rg}$. This way, the system (18.3), is of the form

$$\begin{bmatrix}A_1(\bar{x}_\gamma) & -\mu_{12}\\ -\mu_{21} & A_2(\bar{x}_\gamma)\end{bmatrix}\begin{bmatrix}\overline{\phi}_1\\ \overline{\phi}_2\end{bmatrix}=\begin{bmatrix}\overline{S}_1\\ \overline{S}_2\end{bmatrix}$$

The solution in $\overline{X}_\gamma$, in matrix notation is $M(\bar{x}_\gamma)\overline{\Phi}=\overline{S}$, where $\overline{\Phi}=[\overline{\phi}_1,\overline{\phi}_2]^T$ and $\overline{S}=[\overline{S}_1,\overline{S}_2]^T$.

$$Det(M(\bar{x}_\gamma))=\left|M(\bar{x}_\gamma)\right|=A_1(\bar{x}_\gamma)A_2(\bar{x}_\gamma)-\mu_{12}\mu_{21}\neq 0$$

The general solution is given by

$$\overline{\phi}_g=\frac{1}{\left|M(\bar{x}_\gamma)\right|}A_{g'}(\bar{x}_\gamma)\overline{S}_g+\frac{1}{\left|M(\bar{x}_\gamma)\right|}\mu_{gg'}\overline{S}_{g'}\,.$$

To find the solution in $X_\gamma$ we apply the inverse operator $\mathscr{I}_\gamma^{-1}[\overline{f}(\bar{x}_\gamma);\bar{x}_\gamma \to x_\gamma] := \int_{\overline{X}_\gamma} \overline{f}(\bar{x}_\gamma) K_\gamma^{-1}(x_\gamma,\bar{x}_\gamma)\, d\bar{x}_\gamma$. Then

$$\phi_g(x_\gamma) = \underbrace{\int_{\overline{X}_\gamma} \frac{1}{\left|M(\bar{x}_\gamma)\right|} A_{g'}(\bar{x}_\gamma)\overline{S}_g(\bar{x}_\gamma) K_\gamma^{-1}(x_\gamma,\bar{x}_\gamma)\, d\bar{x}_\gamma}_{\phi_g^{(1)}} + \underbrace{\int_{\overline{X}_\gamma} \frac{1}{\left|M(\bar{x}_\gamma)\right|} \mu_{gg'}\overline{S}_{g'}(\bar{x}_1) K_\gamma^{-1}(x_\gamma,\bar{x}_\gamma)\, d\bar{x}_\gamma}_{\phi_g^{(2)}} \tag{18.4}$$

To evaluate each integral in (18.4) we need define $\mathscr{I}_\gamma$ and the respective kernel $K_\gamma(x_\gamma,\bar{x}_\gamma)$ for each $\gamma$, i.e., for the considered

$$K_\gamma(x_\gamma,\bar{x}_\gamma) := \left[\frac{1}{\sqrt{2\pi}} e^{-i x_\gamma \bar{x}_\gamma}\right]^{\delta_{0,\gamma}} \times \left[x_\gamma J_0(x_\gamma \bar{x}_\gamma)\right]^{\delta_{1,\gamma}}$$

$$K_\gamma^{-1}(x_\gamma,\bar{x}_\gamma) := \left[\frac{1}{\sqrt{2\pi}} e^{i x_\gamma \bar{x}_\gamma}\right]^{\delta_{0,\gamma}} \times \left[\bar{x}_\gamma J_0(x_\gamma \bar{x}_\gamma)\right]^{\delta_{1,\gamma}},$$

where $\delta_{i,j}$ is the Kronecker delta. To solve the equation for $\phi^{(1)}(x_\gamma)$ and $\phi^{(2)}(x_\gamma)$ we will use the Parseval relation given by the following Theorem.

**Theorem 1 (Parsenval's relation in general geometries [Sn72]).** *If the functions $f(x'_\gamma)$ and $g(x'_\gamma)$ are piecewise continuous and absolutely integrable on the positive real line and if $\overline{f}(\bar{x}_\gamma)$ and $\overline{f}(\bar{x}_\gamma)$ denote the respective integral transforms by $\mathscr{I}_\gamma$, then*

$$\int_{X'_\gamma} \omega(x'_\gamma) f(x'_\gamma) g(x'_\gamma)\, dx'_\gamma = \int_{\overline{X}_\gamma} \omega(\bar{x}_\gamma)\overline{f}(\bar{x}_\gamma)\overline{g}(\bar{x}_\gamma)\, d\bar{x}_\gamma \tag{18.5}$$

*where $\omega(x_\gamma)$ is a weight function that depends on $\gamma$.*

Using the Parseval relation and the kernels previously mentioned for $\phi_g^{(1)}(x_\gamma)$, we get

$$\int_{\overline{X}_\gamma} \omega(\bar{x}_\gamma) \left\{\frac{A_{g'}(\bar{x}_\gamma) K_\gamma^{-1}(x_\gamma,\bar{x}_\gamma)}{\left|M(\bar{x}_\gamma)\right|}\right\} \overline{S}_g(\bar{x}_\gamma)\, d\bar{x}_\gamma = \int_{X'_\gamma} \omega(x'_\gamma) I_\gamma^{-1} \left\{\frac{A_{g'}(\bar{x}_\gamma) K_\gamma^{-1}(x_\gamma,\bar{x}_\gamma)}{\left|M(\bar{x}_\gamma)\right|}\right\} S_g(x'_\gamma)\, dx'_\gamma$$

or

$$I_\gamma^{-1}\left\{\frac{A_{g'}(\bar{x}_\gamma)K_\gamma^{-1}(x_\gamma,\bar{x}_\gamma)}{\left|M(\bar{x}_\gamma)\right|}\right\}=\int_{\bar{X}_\gamma}\left\{\frac{A_{g'}(\bar{x}_\gamma)K_\gamma^{-1}(x_\gamma,\bar{x}_\gamma)}{\left|M(\bar{x}_\gamma)\right|}\right\}K(x'_\gamma,\bar{x}_\gamma)\,d\bar{x}_\gamma\,.$$

So that, the final expression for $\phi_g^{(1)}(x_\gamma)$ is

$$\phi_g^{(1)}(x_\gamma)=\int_{X'_\gamma}\int_{\bar{X}_\gamma}\omega(x'_\gamma)\frac{A_{g'}(\bar{x}_\gamma)}{\left|M(\bar{x}_\gamma)\right|}K^{-1}(x_\gamma,\bar{x}_\gamma)K(x'_\gamma,\bar{x}_\gamma)S_g(x'_\gamma)\,d\bar{x}_\gamma\,dx'_\gamma \quad (18.6)$$

By a similar procedure, we can obtain the general solution for $\phi_g^{(2)}(x_\gamma)$ again using the Parseval relation:

$$\phi_g^{(2)}(x_\gamma)=\int_{X'_\gamma}\int_{\bar{X}_\gamma}\frac{\omega(x'_\gamma)\mu_{gg'}}{\left|M(\bar{x}_\gamma)\right|}K^{-1}(x_\gamma,\bar{x}_\gamma)K(x'_\gamma,\bar{x}_\gamma)S_{g'}(x'_\gamma)\,d\bar{x}_\gamma\,dx'_\gamma \quad (18.7)$$

### *18.1.1 Homogeneous Associated Solution*

In the sequel, we determine the associated homogeneous solution of the system. The homogeneous equation of this system for two energy groups is

$$\Delta_\gamma\phi_g-\alpha_g\phi_g+\frac{\mu_{gg'}}{D_g}\phi_{g'}=0\,,$$

and in matrix representation

$$\Delta_\gamma\boldsymbol{\Phi}-P\boldsymbol{\Phi}=0 \qquad \text{with} \qquad P=\begin{bmatrix}\alpha_1 & -\frac{\mu_{12}}{D_1}\\ -\frac{\mu_{21}}{D_2} & \alpha_2\end{bmatrix}$$

Upon diagonalizing the matrix $P=UDU^{-1}$ where $D$ is a diagonal matrix, using $W=U^{-1}\boldsymbol{\Phi}$, the last matrix equation turns into $\Delta_\gamma W-DW=0$. At this point, the solution is given by $\boldsymbol{\Phi}=UW$ after solving the equation system for $W$.

## 18.2 Solution for Cylindrical Geometry

The Laplace operator in cylinder coordinates is $\Delta_1=\frac{1}{r}\frac{\partial}{\partial r}+\frac{\partial^2}{\partial r^2}$, assuming translational symmetry of the neutron flux $\phi_g$ along the cylinder axis ($\partial_{zz}\phi=0$). The diffusion problem is subject zero current density boundary conditions at the center

of the cylinder $\frac{\partial \phi_g}{\partial x_1}(0) = 0$ and zero flux at the border $\phi_g(R) = 0$. Let $X_1$ be the domain, with $X_1 = [0, \infty)$ and $\Delta_1$ a linear operator as described in [Fe11]. By the definition of $\mathscr{I}_\gamma$,

$$\begin{aligned} \mathscr{I}_1[f(x_1); x_1 \to \bar{x}_1] &:= \int_{X_1} f(x_1) K(x_1, \bar{x}_1)\, dx_1 \\ K_1(x_1, \bar{x}_1) &:= x_1 J_0(x_1 \bar{x}_1) \\ K_1^{-1}(x_1, \bar{x}_1) &:= \bar{x}_1 J_0(x_1 \bar{x}_1) \end{aligned} \tag{18.8}$$

Using property $(ii)$ of $\mathscr{I}_\gamma$

$$\mathscr{I}_1[\Delta_1 f(x_1)] := -\bar{x}_1^2 \bar{f}(\bar{x}_1)$$

where $h(\bar{x}_1) = -\bar{x}_1^2$. This way, we can use the same general methodology shown before and after application of the transform $\mathscr{I}_1$ to system (18.1), one obtains the general solution given by equation (18.6).

Using the Parseval relation (18.5) where $\gamma = 1$ and $\omega_1(x_1) = x_1$ and considering that $\phi_g(\bar{x}_1) := \phi_g^{(1)}(\bar{x}_1) + \phi_g^{(2)}(\bar{x}_1)$ we obtain

$$\begin{aligned} \phi_g^{(1)}(\bar{x}_1) &= \int_{\bar{X}_1} \bar{x}_1 \frac{1}{|M(\bar{x}_1)|} A_{g'} \bar{S}_g(\bar{x}_1) J_0(x_1 \bar{x}_1)\, d\bar{x}_1 \\ &= \int_{\bar{X}_1} \bar{x}_1 \frac{A_{g'}(\bar{x}_1)}{A_g(\bar{x}_1) A_{g'}(\bar{x}_1) - \mu_{gg'} \mu_{g'g}} \bar{S}_g(\bar{x}_1) J_0(x_1 \bar{x}_1)\, d\bar{x}_1 \\ \phi_g^{(2)}(\bar{x}_1) &= \int_{\bar{X}_1} \bar{x}_1 \frac{1}{|M(\bar{x}_1)|} \mu_{gg'} \bar{S}_{g'}(\bar{x}_1) J_0(x_1 \bar{x}_1)\, d\bar{x}_1 \\ &= \int_{\bar{X}_1} \bar{x}_1 \frac{\mu_{gg'} \bar{S}_{g'}(\bar{x}_1) J_0(x_1 \bar{x}_1)}{A_g(\bar{x}_1) A_{g'}(\bar{x}_1) - \mu_{gg'} \mu_{g'g}}\, d\bar{x}_1 \end{aligned} \tag{18.9}$$

The final solution for this couple of equations can be expressed using (18.6) and (18.7), and has the form

$$\begin{aligned} \phi_g^{(1)}(x_1) &= \int_{X_1'} \int_{\bar{X}_1} \frac{A_{g'}(\bar{x}_1)}{|M(\bar{x}_1)|} K^{-1}(x_1, \bar{x}_1) K(x_1', \bar{x}_1) S_g(x_1')\, d\bar{x}_1\, dx_1' \\ \phi_g^{(2)}(x_1) &= \int_{X_1'} \int_{\bar{X}_1} \frac{\mu_{gg'}}{|M(\bar{x}_1)|} K^{-1}(x_1, \bar{x}_1) K(x_1', \bar{x}_1) S_{g'}(x_1')\, d\bar{x}_1\, dx_1' \end{aligned} \tag{18.10}$$

Solving first $\phi_g^{(1)}(x_1)$, using (18.8) and (18.10) one gets

$$\phi_g^{(1)}(x_1) = \int_{X_1'} \int_{\bar{X}_1} \frac{A_{g'}(\bar{x}_1) \bar{x}_1 J_0(\bar{x}_1 x_1) x_1' J_0(x_1' \bar{x}_1) S_g(x_1')}{A_g(\bar{x}_1) A_{g'}(\bar{x}_1) - \mu_{gg'} \mu_{g'g}}\, d\bar{x}_1\, dx_1'$$

To solve the inner integral, we use the Parseval relation and make use of the generally valid relation

$$\frac{\mu_{gg'}\mu_{g'g}}{A_g(\bar{x}_1)A_{g'}(\bar{x}_1)} \ll 1$$

From the identity

$$\begin{aligned}\frac{A_{g'}(\bar{x}_1)}{A_g(\bar{x}_1)A_{g'}(\bar{x}_1)-\mu_{gg'}\mu_{g'g}} &= \frac{A_{g'}(\bar{x}_1)}{A_g(\bar{x}_1)A_{g'}(\bar{x}_1)}\frac{1}{1-\frac{\mu_{gg'}\mu_{g'g}}{A_g(\bar{x}_1)A_{g'}(\bar{x}_1)}} \\ &= \frac{1}{A_g(\bar{x}_1)}\sum_{n=0}^{\infty}\left(\frac{\mu_{gg'}\mu_{g'g}}{A_g(\bar{x}_1)A_{g'}(\bar{x}_1)}\right)^n\end{aligned}$$

We consider only the dominant term of this series,

$$\phi_g^{(1)}(x_1) = \int_{X_1'}\int_{\bar{X}_1}\frac{1}{A_g(\bar{x}_1)}\bar{x}_1 J_0(\bar{x}_1 x_1)x_1' J_0(x_1'\bar{x}_1)S_g(x_1')\,d\bar{x}_1\,dx_1'$$

That allows us to solve the integral analytically:

$$\begin{aligned}&\int_{\bar{X}_1}\bar{x}_1\frac{J_0(x_1\bar{x}_1)}{A_g(\bar{x}_1)}J_0(x_1'\bar{x}_1)\,d\bar{x}_1 = \\ &\begin{cases}\frac{1}{D_g}I_0(\sqrt{\alpha_g}x_1')K_0(\sqrt{\alpha_g}x_1) & \text{for} \quad 0<x_1'<x_1 \\ \frac{1}{D_g}I_0(\sqrt{\alpha_g}x_1)K_0(\sqrt{\alpha_g}x_1') & \text{for} \quad x_1<x_1'<\infty\end{cases},\end{aligned}$$

where $\alpha_g = \frac{\Sigma_{Rg}}{D_g}$. Here, $I_0$ and $K_0$ are the modified Bessel functions. To complement this part of the solution we use the fact that there does not exist any source outside the cylinder, i.e., $S_g = 0$ for $x_1 > R$. We can express the solution as

$$\phi_g^{(1)} = \mathscr{T}_g^{(1)}[S_g](x_1)$$

where

$$\begin{aligned}\mathscr{T}_g^{(1)}[\cdot] := &\frac{K_0(\sqrt{\alpha_g}x_1)}{D_g}\int_0^{x_1}x_1' I_0(\sqrt{\alpha_g}x_1')[\cdot]\,dx_1' \\ &+ \frac{I_0(\sqrt{\alpha_g}x_1)}{D_g}\int_{x_1}^{R}x_1' K_0(\sqrt{\alpha_g}x_1')[\cdot]\,dx_1'\end{aligned} \tag{18.11}$$

Now, we determine the solution of $\phi_g^{(2)}$. Again, we can use the above-mentioned theorems to obtain

$$\int_{\bar{X}_1}\bar{x}_1\frac{J_0(x_1\bar{x}_1)}{|M(\bar{x}_1)|}\bar{S}_{g'}\,d\bar{x}_1 = \int_{X_1}x_1'\mathscr{I}_0^{-1}\left\{\frac{J_0(x_1\bar{x}_1)}{|M(\bar{x}_1)|}\right\}S_{g'}(x_1')\,dx_1'$$

and, by definition,

$$\mathscr{I}_0^{-1}\left\{\frac{J_0(x_1\bar{x}_1)}{|M(\bar{x}_1)|}\right\}=\int_{\bar{X}_1}\bar{x}_1\frac{J_0(x_1\bar{x}_1)}{|M(\bar{x}_1)|}J_0(x_1'\bar{x}_1)\,d\bar{x}_1$$

Using analogue arguments as for the fast flux, we arrive at

$$\begin{aligned}\mathscr{I}_0^{-1}\left\{\frac{J_0(x_1\bar{x}_1)}{|M(\bar{x}_1)|}\right\}&=\frac{1}{(\Sigma_{Rg}D_{g'}-\Sigma_{Rg'}D_g)}\int_{X_1}\bar{x}_1\frac{J_0(x_1\bar{x}_1)J_0(x_1'\bar{x}_1)}{\bar{x}_1^2+(\sqrt{\alpha_{g'}})^2}\,d\bar{x}_1\\&\quad-\frac{1}{(\Sigma_{Rg}D_{g'}-\Sigma_{Rg'}D_g)}\int_{\bar{X}_1}\bar{x}_1\frac{J_0(x_1\bar{x}_1)J_0(x_1'\bar{x}_1)}{\bar{x}_1^2+(\sqrt{\alpha_g})^2}\,d\bar{x}_1\\&=\begin{cases}\frac{I_0(\sqrt{\alpha_{g'}}x_1')K_0(\sqrt{\alpha_{g'}}x_1)-I_0(\sqrt{\alpha_g}x_1')K_0(\sqrt{\alpha_g}x_1)}{(\Sigma_{Rg}D_{g'}-\Sigma_{Rg'}D_g)} & \text{for}\quad 0<x_1'<x_1\\ \frac{I_0(\sqrt{\alpha_{g'}}x_1)K_0(\sqrt{\alpha_{g'}}x_1')-I_0(\sqrt{\alpha_g}x_1)K_0(\sqrt{\alpha_g}x_1')}{(\Sigma_{Rg}D_{g'}-\Sigma_{Rg'}D_g)} & \text{for}\quad x_1<x_1'<\infty\end{cases}.\end{aligned}$$

This way,

$$\phi_g^{(2)}=c_{gg'}(D_{g'}\mathscr{T}_{g'}^{(1)}[S_{g'}](x_1)-D_g\mathscr{T}_g^{(1)}[S_{g'}](x_1)),$$

where $c_{gg'}=\frac{\mu_{gg'}}{(\Sigma_{Rg}D_{g'}-\Sigma_{Rg'}D_g)}$ and by the similarity of the solutions of the integral expressions may be used to formulate the complete solution for the group $g$ using (18.11)

$$\phi_g(x_1)=\mathscr{T}_g^{(1)}[S_g](x_1)+c_{gg'}(D_{g'}\mathscr{T}_{g'}^{(1)}[S_{g'}](x_1)-D_g\mathscr{T}_g^{(1)}[S_{g'}](x_1)) \quad (18.12)$$

The solution for the group $g'$ is obtained upon changing $g'$ with $g$ in (18.12). This last equation together with associated homogeneous solutions for $\gamma=1$ represents the complete profile of neutron flux in cylindrical geometry.

## 18.3 Solution for Cartesian Geometry

In this section, we solve (18.1) for $\gamma=0$, i.e., in Cartesian geometry. The multi-group diffusion equation (18.1) is subject to the boundary conditions $\phi_g(L/2)=\phi_g(-L/2)=0$ and by the definition of $\mathscr{I}_\gamma$ we have

$$\mathscr{I}_0[f(x_0);x_0\to\bar{x}_0]:=\int_{X_0}f(x_0)K(x_0,\bar{x}_0)\,dx_0$$

$$K_0(x_0,\bar{x}_0):=\frac{1}{\sqrt{2\pi}}e^{-ix_0\bar{x}_0}\text{ and }K_0^{-1}(x_0,\bar{x}_0):=\frac{1}{\sqrt{2\pi}}e^{ix_0\bar{x}_0} \quad (18.13)$$

Using the property $(ii)$ of $\mathscr{I}_\gamma$ and for this case $h(\bar{x}_0) = -\bar{x}_0^2$, one obtains

$$\mathscr{I}_0[\Delta_0 f(x_0)] := -\bar{x}_0^2 \bar{f}(\bar{x}_0)$$

This way, we can express the general solution using (18.13) as

$$\phi_g(x_0) = \underbrace{\frac{1}{\sqrt{2\pi}} \int_{\bar{X}_0} \frac{A_{g'}(x_0)}{|M(\bar{x}_0)|} e^{-ix_0\bar{x}_0} S_g(\bar{x}_0)\, d\bar{x}_0}_{\phi_g^{(1)}(x_0)} + \underbrace{\frac{1}{\sqrt{2\pi}} \int_{\bar{X}_0} \frac{\mu_{gg'}}{|M(\bar{x}_0)|} e^{-ix_0\bar{x}_0} S_{g'}(\bar{x}_0)\, d\bar{x}_0}_{\phi_g^{(2)}(x_0)},$$

where $X_0 = \bar{X}_0 = (-\infty, \infty)$. Firstly, solving for $\phi_g^{(1)}$ we define

$$G_1^*(\bar{x}_0) = \frac{A_{g'}(x_0)}{|M(\bar{x}_0)|} e^{ix_0\bar{x}_0} \text{ and } F_1(\bar{x}_0) = S_g(\bar{x}_0)$$

and using the Parseval theorem again with $w(x_0) = 1/\sqrt{2\pi}$. Then

$$\phi_1^{(1)}(x_0) = \frac{1}{\sqrt{2\pi}} \int_{\bar{X}_0} G_1^{(1)*}(\bar{x}_0) F_1^{(1)}(\bar{x}_0)\, d\bar{x}_0 = \frac{1}{\sqrt{2\pi}} \int_{X_0'} g_1^{(1)*}(x_0') S_1(x_0')\, dx_0'$$

where

$$g_1^{(1)*}(x_0') = \mathscr{F}^{-1}\left\{G_1^{(1)*}(\bar{x}_0)\right\} = \frac{1}{\sqrt{2\pi}} \int_{\bar{X}_0} e^{i\bar{x}_0(x_0'-x_0)} \frac{A_2(\bar{x}_0)}{|M(\bar{x}_0)|}\, d\bar{x}_0. \tag{18.14}$$

We define the following constant $C_{gg'} = \alpha_g \alpha_{g'} - \frac{\mu_{gg'}\mu_{g'g}}{D_g D_{g'}}$. Thus, according to definition (18.14):

$$g_1^{(1)*}(x_0') = \frac{1}{D_g\sqrt{2\pi}} \int_{\bar{X}_0} \frac{\bar{x}_0^2 + \alpha_{g'}}{\bar{x}_0^4 + (\alpha_g + \alpha_{g'})\bar{x}_0^2 + C_{gg'}} e^{i\bar{x}_0(x_0'-x_0)}\, d\bar{x}_0$$

To obtain an expression for $g_1^{(1)*}(x_0')$, we make use of the residue theorem of Cauchy. To this end we consider the integral

$$\int_\Gamma \frac{z^2 + \alpha_{g'}}{z^4 + (\alpha_g + \alpha_{g'})z^2 + C_{gg'}} e^{iz(t-x)}\, dz$$

for $x_0' - x_0 > 0$, where $\Gamma$ is a closed Jordan curve in the sector and oriented such that all poles are inside the curve oriented counterclockwise. The poles at the integrand are $y = z^2$ and observing that $(\alpha_g + \alpha_{g'})^2 - 4C_{gg'} > 0$, further, from that fact that the nuclear parameters are all positive, we conclude that there are two distinct real

roots for $y$ and consequently four different roots for $z$. Observe that the signal of $y_1$ can be positive or negative while $y_2 < 0$, $y_1 < 0$ with $0 < \mu_{gg'}\mu_{g'g} < \Sigma_{Rg}\Sigma_{Rg'}$. All poles of the integrand are located on the imaginary axis.

$$z_{\pm\mp} = \pm i\frac{\sqrt{2}}{2}\left((\alpha_g + \alpha_{g'}) \mp \sqrt{(\alpha_g - \alpha_{g'})^2 + \frac{4C_{gg'}}{D_g D_{g'}}}\right)^{\frac{1}{2}}$$

Knowing the poles, we can choose $\Gamma = \Gamma_+ \cup \Gamma_R$, where $\Gamma_+ = \{z \in \mathbf{C} : z = Re^{i\theta}, 0 \leq \theta \leq \pi\}$ and $\Gamma_R = \{z \in \mathbf{R} : 0 < |z| < R\}$. Then, by the residue theorem of Cauchy we get

$$\int_\Gamma u_1^{(1)}(z)\,dz = 2\pi i\left(\sum_{n=1}^{2} Res_{z=z_n} u_1^{(1)}(z)\right) = \int_{-R}^{R} u_1^{(1)}(z)\,dz + \int_{\Gamma_+} u_1^{(1)}(z)\,dz$$

with $u_1^{(1)}(z) = \frac{z^2+\alpha_{g'}}{z^4+(\alpha_g+\alpha_{g'})z^2+C_{gg'}} e^{iz(x_0'-x_0)}$. In the limit of $R$ going to infinity and the integral along $\Gamma_+$ is null, so that

$$\begin{aligned}\lim_{R\to\infty}\int_{-R}^{R} u_1^{(1)}(z)\,dz &= \int_{-\infty}^{\infty} \frac{z^2+\alpha_{g'}}{z^4+\left(\alpha_g+\alpha_{g'}\right)z^2+C_{gg'}} e^{iz\left(x_0'-x_0\right)}\,dz \\ &= 2\pi i\left[Res_{z=z_1} u_1^{(1)}(z) + Res_{z=z_2} u_1^{(1)}(z)\right]\end{aligned}$$

where after some manipulations

$$Res_{z=z_1} u_1^{(1)}(z) = \frac{(\alpha_g - \alpha_{g'} + \sqrt{\beta})e^{-\frac{\sqrt{2}}{2}(x_0'-x_0)(\alpha_g+\alpha_{g'}+\sqrt{\beta})^{1/2}}}{2\sqrt{2}i\sqrt{\beta}(\alpha_g+\alpha_{g'}+\sqrt{\beta})^{1/2}}$$

$$Res_{z=z_2} u_1^{(1)}(z) = \frac{(\alpha_g - \alpha_{g'} - \sqrt{\beta})e^{-\frac{\sqrt{2}}{2}(x_0'-x_0)(\alpha_g+\alpha_{g'}-\sqrt{\beta})^{1/2}}}{2\sqrt{2}i\sqrt{\beta}(\alpha_g+\alpha_{g'}-\sqrt{\beta})^{1/2}}$$

where $\beta = (\alpha_g - \alpha_{g'})^2 + \frac{4\mu_{gg'}\mu_{g'g}}{D_g D_{g'}}$. This way, we obtain

$$\begin{aligned} g_1^{(1)*}(x_0') = & \\ &+ \pi\frac{(\alpha_g+\alpha_{g'}-\sqrt{\beta})^{\frac{1}{2}}(\alpha_g-\alpha_{g'}+\sqrt{\beta})e^{-\frac{\sqrt{2}}{2}(x_0'-x_0)(\alpha_g+\alpha_{g'}+\sqrt{\beta})^{\frac{1}{2}}}}{4D_g\sqrt{\pi\beta C_{gg'}}} \\ &- \pi\frac{(\alpha_g+\alpha_{g'}+\sqrt{\beta})^{\frac{1}{2}}(\alpha_g-\alpha_{g'}-\sqrt{\beta})e^{-\frac{\sqrt{2}}{2}(x_0'-x_0)(\alpha_g+\alpha_{g'}-\sqrt{\beta})^{\frac{1}{2}}}}{4D_g\sqrt{\pi\beta C_{gg'}}}\end{aligned}$$

We express $\phi_g^{(1)}(x_0)$, bearing in mind that we have no source outside the slab, i.e., $S_g = 0$ for $|x_0| > L/2$,

$$\phi_g^{(1)}(x_0) = \mathscr{T}_-^{(1)}[S_g](x_0) - \mathscr{T}_+^{(1)}[S_g](x_0)\,,$$

where

$$\begin{aligned}\mathscr{T}_\mp^{(1)}[\cdot] = {} & \frac{(\alpha_g + \alpha_{g'} \pm \sqrt{\beta})^{\frac{1}{2}}(\alpha_g - \alpha_{g'} \mp \sqrt{\beta})}{4D_g\sqrt{2\beta C_{gg'}}} \\ & \times \left[\int_{-\frac{L}{2}}^{x_0} e^{\frac{\sqrt{2}}{2}(x_0'-x_0)(\alpha_g+\alpha_{g'}\mp\sqrt{\beta})^{\frac{1}{2}}}[\cdot]\,dx_0'\right. \\ & + \left.\int_{x_0}^{\frac{L}{2}} e^{-\frac{\sqrt{2}}{2}(x_0'-x_0)(\alpha_g+\alpha_{g'}\mp\sqrt{\beta})^{\frac{1}{2}}}[\cdot]\,dx_0'\right]\end{aligned}$$

To obtain the solution for $\phi_g^{(2)}(x_0)$ we may use a similar procedure to end up with the final expression

$$\phi_g^{(2)} = \mathscr{T}_-^{(2)}[S_{g'}](x_0) - \mathscr{T}_+^{(2)}[S_{g'}](x_0)$$

where

$$\begin{aligned}\mathscr{T}_\mp^{(2)}[\cdot] = {} & \frac{\mu_{gg'}\left(\alpha_g + \alpha_{g'} \pm \sqrt{\beta}\right)^{\frac{1}{2}}}{2\sqrt{2\beta C_{gg'}}} \\ & \times \left[\int_{-\frac{L}{2}}^{x_0} e^{\frac{\sqrt{2}}{2}(x_0'-x_0)\left(\alpha_g+\alpha_{g'}\mp\sqrt{\beta}\right)^{\frac{1}{2}}}[\cdot]\,dx_0'\right. \\ & + \left.\int_{x_0}^{\frac{L}{2}} e^{-\frac{\sqrt{2}}{2}(x_0'-x_0)\left(\alpha_g+\alpha_{g'}\mp\sqrt{\beta}\right)^{\frac{1}{2}}}[\cdot]\,dx_0'\right]\end{aligned}$$

Thus, the full representation for the solution in Cartesian coordinates is described by

$$\phi_g(x_0) = \mathscr{T}_-^{(1)}[S_g](x_0) - \mathscr{T}_+^{(1)}[S_g](x_0) + \mathscr{T}_-^{(2)}[S_{g'}](x_0) - \mathscr{T}_+^{(2)}[S_{g'}](x_0)$$

This last expression together with the associated homogeneous solutions for $\gamma = 0$ represents the complete profile of neutron flux in slab geometry.

**Table 18.1** Sets 1-2 of Nuclear Parameters used in cylinder simulations ($^a$ is $7.3760 \times 10^{-1}$ ).

| Set 1 | $g=1$ | $g=2$ | Set 2 | $g=1$ | $g=2$ |
|---|---|---|---|---|---|
| $D_g$ | 2.4449 | 1.2272 | $D_g$ | 2.4449 | 1.2272 |
| $\chi_g$ | $7.3760(-1)^a$ | $2.6220(-1)$ | $\chi_g$ | $7.3760(-1)$ | $2.6220(-1)$ |
| $\Sigma_{Rg}$ | $5.8938(-2)$ | $6.7201(-2)$ | $\Sigma_{Rg}$ | $5.8938(-2)$ | $6.7201(-2)$ |
| $\Sigma_{gg'}$ | $1.0000(-2)$ | $1.0000(-4)$ | $\Sigma_{gg'}$ | $1.0000(-1)$ | $1.0000(-4)$ |
| $\nu\Sigma_{fg}$ | $9.6350(-4)$ | $1.1530(-3)$ | $\nu\Sigma_{fg}$ | $1.46025(-3)$ | $1.7295(-3)$ |
| $S_g$ | 1.0000 | $1.0000(-1)$ | $S_g$ | 1.0000 | $1.0000(-1)$ |

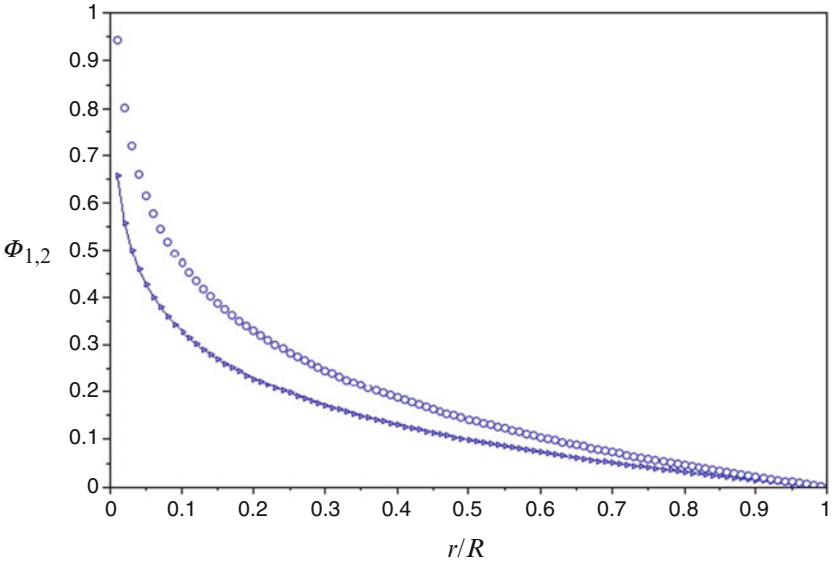


**Fig. 18.1** Profile of $\phi_1$ ($\triangleright$) and $\phi_2$ ($\circ$), $R=1$ and using set 1.

## 18.4 Results

For simulations we consider a set of parameters (see table 18.1) for two energy groups in a homogeneous medium. For different cases we simulated the flux for both groups (see figures 18.1 and 18.2) and present normalized results in the domain $X_\gamma$ (see tables 18.2 and 18.3).

## 18.5 Conclusions

In this work, a multi-group neutron diffusion problem in different geometries is discussed. The analytical expressions found represent an accurate solution for the multi-group steady state diffusion equation in a slab and cylinder coordinates.

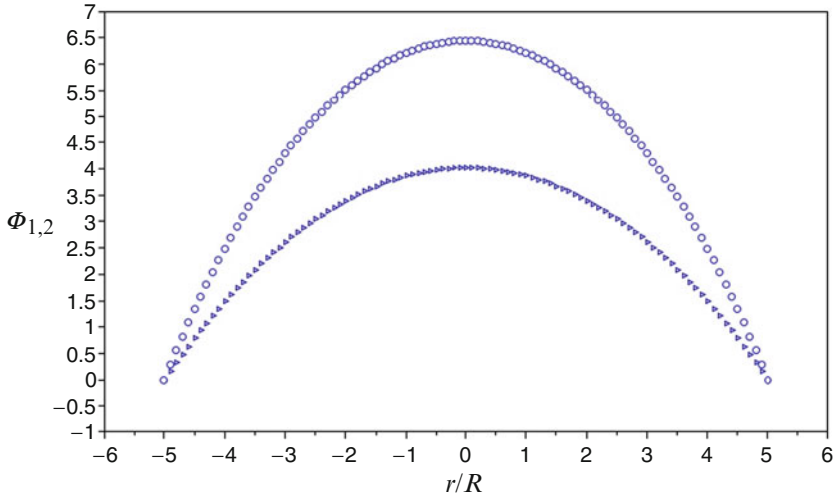


**Fig. 18.2** Profile of $\phi_1$ (▷) and $\phi_2$ (○), $L = 5$ and using set 1.

**Table 18.2** Normalized profile for two energy groups by two different parameters in cylindrical geometry with $R = 5$.

| Set 1 | | | Set 2 | | |
|---|---|---|---|---|---|
| $x_1/R$ | $\phi_1$ | $\phi_2$ | $x_1/R$ | $\phi_1$ | $\phi_2$ |
| 0.0 | 1.0000000 | 1.0000000 | 0.0 | 1.0000000 | 1.0000000 |
| 0.1 | 0.3284663 | 0.2729200 | 0.1 | 0.4715268 | 0.3282991 |
| 0.2 | 0.2295664 | 0.1897752 | 0.2 | 0.3295612 | 0.2294560 |
| 0.3 | 0.1717303 | 0.1412999 | 0.3 | 0.2465389 | 0.1716520 |
| 0.4 | 0.1307049 | 0.1070547 | 0.4 | 0.1876467 | 0.1306485 |
| 0.5 | 0.0988880 | 0.0806290 | 0.5 | 0.1419720 | 0.0988476 |
| 0.6 | 0.0728922 | 0.0591645 | 0.6 | 0.1046525 | 0.0728640 |
| 0.7 | 0.0509098 | 0.0411342 | 0.7 | 0.0730935 | 0.0508911 |
| 0.8 | 0.0318613 | 0.0256255 | 0.8 | 0.0457457 | 0.0318503 |
| 0.9 | 0.0150500 | 0.0120486 | 0.9 | 0.0216089 | 0.0150451 |
| 1.0 | 0.0000000 | 0.0000000 | 1.0 | 0.0000000 | 0.0000000 |

An immediate conclusion that may be drawn from this work is that for neutron diffusion scenarios, using the Parseval Identity is a considerably efficient technique for solving this type of problem. As can be seen from the formulation, the present method provides the correct final expression without making use of approximations or simplifications. In procedures where an analytical solution was obtained by a spectral theory approach, the solution had been expressed as an expansion of orthogonal functions with a predefined base. It is noteworthy that Parseval's identity indicates the base that "naturally" should be used. Concluding, this method in these geometries can be considered a reliable tool for solving more general problems in

**Table 18.3** Normalized profile for two energy groups by two different parameters in slab geometry with $L = 5$.

| Set 1 | | | Set 2 | | |
|---|---|---|---|---|---|
| $x_0/(L/2)$ | $\phi_1$ | $\phi_2$ | $x_0/(L/2)$ | $\phi_1$ | $\phi_2$ |
| 0.0 | 1.0000000 | 1.0000000 | 0.0 | 1.0000000 | 1.0000000 |
| 0.1 | 0.9904005 | 0.9910545 | 0.1 | 0.9903972 | 0.9910538 |
| 0.2 | 0.9615538 | 0.9640957 | 0.2 | 0.9615411 | 0.9640928 |
| 0.3 | 0.9133158 | 0.9187549 | 0.3 | 0.9132886 | 0.9187487 |
| 0.4 | 0.8454454 | 0.8544121 | 0.4 | 0.8454005 | 0.8544018 |
| 0.5 | 0.7576040 | 0.7701872 | 0.5 | 0.7575410 | 0.7701728 |
| 0.6 | 0.6493545 | 0.6649283 | 0.6 | 0.6492767 | 0.6649105 |
| 0.7 | 0.5201600 | 0.5371958 | 0.7 | 0.5200749 | 0.5371763 |
| 0.8 | 0.3693819 | 0.3852425 | 0.8 | 0.3693026 | 0.3852244 |
| 0.9 | 0.1962778 | 0.2069900 | 0.9 | 0.1962243 | 0.2069778 |
| 1.0 | 0.0000000 | 0.0000000 | 1.0 | 0.0000000 | 0.0000000 |

neutron diffusion, for example, with more than two energy groups. We further plan to investigate results for more realistic problems in nuclear reactor physics where the application of this methodology will be generalized to heterogeneous problems.

## References

[Bd10] Bodmann, B.E.J., Vilhena, M.T., Ferreira, L.S., Bardaji, J.B.: An analytical solver for the multi-group two dimensional neutron-diffusion equation by integral transform techniques. *Il Nuovo Cimento della Societ Italiana di Fisica,* **C 33**, 1–10 (2010).

[Fe12] Fernandes, J.C.L., Bodmann, B.E.J., Vilhena, M.T.: A Novel to Approach to The Hankel Transform Inversion of the Neutron Diffusion Problem Using Parseval Identity *Integral Methods in Science and Engineering*, Birkhauser, 105–114, (2012).

[Sn72] Sneddon, I.A.: *The use of integral transforms.* McGraw-Hill, New York (1972).

[Fe11] Fernandes, J.C.L.: *Solução Analítica da equação de difusão de nêutrons multi-grupo em cilindro infinito homogêneo através da transformada de Hankel*, PhD Tesis, UFRGS, Porto Alegre, Brazil (2011).

# Chapter 19
# Multi-Group Neutron Propagation in Transport Theory by Space Asymptotic Methods

**J.C.L. Fernandes, S. Dulla, P. Ravetto, and M.T.B. Vilhena**

## 19.1 Introduction

The study of the neutronic response to a localized pulsed source requires accurate models, due to the strong spatial effects coming into play associated with the wavefront propagation. Furthermore, usually neutron sources inject particles at high energy and, consequently, important spectral effects may come into play. In previous works, the propagation phenomenon has been investigated for idealized configurations in exact transport [DuGaRa06], restricting the analysis to one-group problems. Furthermore, the numerical effects associated with the application of discrete ordinate and spherical harmonics methods have been analyzed [DuRa04, DuRa08], illustrating the appearance of time-dependent ray effects.

The present work aims at the characterization of the response of a system to localized pulsed source also taking into account spectral phenomena, which may have an important role in the interpretation of pulsed experiments [DuEtAl13]. To this aim, the multi-group transport equation is considered in one-dimensional plane geometry, extending what was done for the one-velocity model. The contribution of delayed neutrons is not considered, since in these short-term transients the role of delayed emissions is negligible. The interest of the present work is mainly focused on gaining insight into basic physics. Therefore, simple configurations are analyzed and the treatment is analytical as much as possible to clearly evidence physical effects. Furthermore, the analytical approach yields reference solutions free from

J.C.L. Fernandes (✉) • M.T.B. Vilhena
Federal University of Rio Grande do Sul, Rua Sarmento Leite, 425, 90050-170 Porto Alegre, RS, Brazil
e-mail: julio.lombaldo@ufrgs.br; vilhena@math.ufrgs.br

S. Dulla • P. Ravetto
Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
e-mail: sandra.dulla@polito.it; piero.ravetto@polito.it



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_19

numerical discretization drawbacks and the results can be proposed as benchmarks for numerical methods and codes [Ga83].

## 19.2 Problem Formulation

The multi-group linear transport model considered for this study concerns a 1D slab geometry, assuming a homogeneous medium, isotropic emissions, and vanishing initial conditions for the angular fluxes $\varphi_g$:

$$\frac{1}{\mathrm{v}_g}\frac{\partial \varphi_g(x,\mu,t)}{\partial t}+\mu\frac{\partial \varphi_g(x,\mu,t)}{\partial x}+\sigma_g\varphi_g(x,\mu,t)=\frac{1}{2}\sum_{g'=1}^{G}\sigma_{g'}c_{gg'}\int_{-1}^{1}\varphi_{g'}(x,\mu',t)d\mu'$$
$$+\frac{1}{2}S_g(x,t)=\frac{1}{2}\sum_{g'=1}^{G}\sigma_{g'}c_{gg'}\Phi_{g'}(x,t)+\frac{1}{2}S_g(x,t),\quad g=1,\ldots,G, \tag{19.1}$$

where the system spans in the interval $x\in[-h/2:h/2]$ and is surrounded by vacuum; a total number of $G$ energy groups is assumed.

The application of Laplace ($t\rightarrow s$) and Fourier ($x\rightarrow B$) transforms in time and space (see [DuGaRa06]) allows to derive the following algebraic expression for the transformed fluxes:

$$\frac{s}{\mathrm{v}_g}\tilde{\tilde{\varphi}}_g(B,\mu,s)-iB\mu\,\tilde{\tilde{\varphi}}_g(B,\mu,s)+\sigma_g\tilde{\tilde{\varphi}}_g(B,\mu,s)$$
$$=\frac{1}{2}\sum_{g'=1}^{G}\sigma_{g'}c_{gg'}\tilde{\tilde{\Phi}}_{g'}(B,s)+\frac{1}{2}\tilde{\tilde{S}}_g(B,s),\quad g=1,\ldots,G. \tag{19.2}$$

The $g$-group angular flux can be made explicit from the left-hand side of Eq. (19.2):

$$\varphi_g(B,\mu,s)=\frac{1}{\sigma_g+\frac{s}{\mathrm{v}_g}-iB\mu}\left[\frac{1}{2}\sum_{g'=1}^{G}\sigma_{g'}c_{gg'}\Phi_{g'}(B,s)+\frac{1}{2}S_g(B,s)\right] \tag{19.3}$$

and both sides of Eq. (19.3) can be integrated with respect to $\mu$, to obtain an expression involving the scalar flux and the external source only:

$$\int_{-1}^{1}\varphi_g(B,\mu,s)d\mu=\sum_{g'=1}^{G}\sigma_{g'}c_{gg'}\Phi_{g'}(B,s)\frac{1}{2}\int_{-1}^{1}\frac{1}{(\sigma_g+\frac{s}{\mathrm{v}_g})-iB\mu}d\mu$$
$$+S_g(B,s)\frac{1}{2}\int_{-1}^{1}\frac{1}{(\sigma_g+\frac{s}{\mathrm{v}_g})-iB\mu}d\mu. \tag{19.4}$$

If we define the integral expression

$$A^g_{l,k}(B,s) = \frac{1}{2}\int_{-1}^{1} \frac{P_l(\mu)P_k(\mu)}{(\sigma_g + \dfrac{s}{\mathrm{v}_g}) - iB\mu}\, d\mu, \tag{19.5}$$

where $P_l(\mu)$ is the Legendre polynomial of order $l$, we can re-write Eq. (19.4) in a more compact form as

$$\Phi_g(B,s) = A^g_{0,0}(B,s)\sum_{g'=1}^{G} \sigma_{g'} c_{gg'} \Phi_{g'}(B,s) + A^g_{0,0}(B,s) S_g(B,s), \quad g = 1,\ldots,G. \tag{19.6}$$

To solve the system of equations (19.6) we assume an expansion for the flux in each energy group in the form:

$$\Phi_g(B,s) := \sum_{i=0}^{\infty} \Phi^{(i)}_g(B,s), \tag{19.7}$$

where the first term $\Phi^{(0)}_g$ represents the solution obtained in the one-group case and the other terms allow to take into account the coupling among energy groups. The solution procedure starts from the solution of the monokinetic problem, as solved in [DuGaRa06], and then evaluate the following terms of the expansion through this equation:

$$\begin{aligned}\Phi^{(i)}_g(B,s) &= A^g_{0,0}(B,s)\sigma_g c_{gg}\Phi^{(i)}_g(B,s) + A^g_{0,0}(B,s)\sum_{g'\neq g}^{G} \sigma_{g'} c_{gg'} \Phi^{(i-1)}_{g'}(B,s) \\ &+ A^g_{0,0} S_g(B,s), \quad i > 0,\ g = 1,\ldots,G,\end{aligned} \tag{19.8}$$

where the $i-$th term of (19.8) can be made explicit as:

$$\begin{aligned}\Phi^{(i)}_g(B,s) &= \frac{A^g_{0,0}(B,s)}{1 - \sigma_g c_{gg} A^g_{0,0}(B,s)} \sum_{g'\neq g}^{G} \sigma_{g'} c_{gg'} \Phi^{(i-1)}_{g'}(B,s) \\ &+ \frac{A^g_{0,0}(B,s)}{1 - \sigma_g c_{gg} A^g_{0,0}(B,s)} S_g(B,s), \quad i > 0,\ g = 1,\ldots,G.\end{aligned} \tag{19.9}$$

The transfer functions in the transformed space appearing in (19.9) can be defined in a general form as

$$\Gamma_{gg'}(B,s) := \frac{A^g_{0,0}(B,s)}{1 - \sigma_{g'} c_{gg'} A^g_{0,0}(B,s)} \tag{19.10}$$

and the solution for the $i-$th term takes the form

$$\Phi_g^{(i)}(B,s) = \Gamma_{gg}(B,s)\left\{\sum_{g'\neq g}^{G}\sigma_{g'}c_{gg'}\Phi_{g'}^{(i-1)}(B,s) + S_g(B,s)\right\}. \tag{19.11}$$

The computation of the scalar flux as defined in the expansion (19.7) requires the definition of a truncation order $N$. Having introduced the scalar flux obtained with such truncation $\Phi_g^N$, we can write its explicit form as

$$\begin{aligned}
&\Phi_g^N(B,s) = \Phi_g^{(0)}(B,s) + \Phi_g^{(1)}(B,s) + \ldots + \Phi_g^{(N)}(B,s) \qquad \text{for } N \geq 1 \\
&= \Gamma_{gg}(B,s)S_g(B,s) + \sum_{i=1}^{N}\Phi_g^{(i)}(B,s) \\
&= \underbrace{(N+1)\Gamma_{gg}(B,s)S_g(B,s)}_{\text{one-group solution}=\hat{\Phi}_g(B,s)} + \underbrace{\Gamma_{gg}(B,s)\sum_{i=1}^{N}\sum_{g'\neq g}^{G}\sigma_{g'}c_{gg'}\Phi_{g'}^{(i-1)}(B,s)}_{\text{energy group interaction}=\check{\Phi}_g^N(B,s)} \\
&:= \hat{\Phi}_g(B,s) + \check{\Phi}_g^N(B,s),
\end{aligned} \tag{19.12}$$

allowing its evaluation up to the desired order $N$.

In order to obtain the solution in the physical space-time domain, it is then required to perform the inverse Fourier and Laplace transform on the solution (19.12). The inverse Fourier transform is tackled first; the general form of the integral to be performed is

$$\begin{aligned}
&I_{m_1,m_2,\ldots,m_G}^{S_g}(x,s) = \int_{-\infty}^{\infty}(\Gamma_{11})^{m_1}(\Gamma_{22})^{m_2}\ldots(\Gamma_{GG})^{m_G}(B,s)S_g(B,s)e^{-iBx}dB \\
&= \int_{-\infty}^{\infty}\prod_{\gamma=1}^{G}(\Gamma_{\gamma\gamma})^{m_\gamma}(B,s)S_g(B,s)e^{-iBx}dB.
\end{aligned} \tag{19.13}$$

Expression (19.13) is obtained starting from the term of Eq. (19.12) providing the coupling among groups, $\check{\Phi}_g^N$, and making explicit all the flux terms appearing in such expression, so that they can all be referred directly to the external sources $S_g$. This process results in successive application of the transport kernel $\Gamma_g g$, as made explicit in (19.13) by the exponentials $m_1 \ldots m_G$.

At this point the Fourier transform can be worked out easily if we assume that a generic symmetric source in the domain $[-h/2;h/2]$ can be expanded in Helmholtz eigenfunctions as

$$S_g(B,s) = \sqrt{\frac{2}{h}}\sum_{n=1}^{\infty}s_n^g(s)\left[\frac{\delta(B-B_n)+\delta(B+B_n)}{2}\right], \tag{19.14}$$

where $B_n = (2n-1)\pi/h$. This assumption implies that also the solution will satisfy the same boundary conditions as the harmonics adopted, i.e. it vanishes at the system boundary. This condition is not correct when applied to the scalar flux, but is still

physically significant in the analysis of source pulses for short times, when the neutrons still have not reached the system boundary. Once we introduce (19.14) into (19.13) the inverse Fourier transform is easily obtained, and we can proceed to the evaluation of the inverse Laplace transform:

$$I^{S_g}_{m_1,m_2,\ldots,m_G}(x,t) = \sqrt{\frac{2}{h}}\sum_{n=1}^{\infty}\left\{\int_0^t\left[\frac{1}{2\pi i}\int_{-\sigma_g-i\infty}^{-\sigma_g+i\infty}\prod_{\gamma=1}^{G}\Gamma_{\gamma\gamma}^{m_\gamma}(B,s)e^{s(t-t')}ds\right]s_n^g(t')dt'\right\}\cos(B_n x) \tag{19.15}$$

Making use of the integral definitions above, we can provide as example the solution in the transformed space of the flux when expanded to progressively higher orders $N$. For $N=0$ we have

$$\begin{cases}\Phi_1^0(B,s)=\Gamma_{11}S_1(B,s)\\ \Phi_2^0(B,s)=\Gamma_{22}S_2(B,s)\\ \vdots\\ \Phi_G^0(B,s)=\Gamma_{GG}S_G(B,s)\end{cases}\Rightarrow\begin{cases}\Phi_1^0(x,t)=I_1^{S_1}(x,t)\\ \Phi_2^0(x,t)=I_{0,1}^{S_2}(x,t)\\ \vdots\\ \Phi_G^0(x,t)=I^{S_G}_{\underbrace{0,0,\ldots,1}_{\times G}}(x,t).\end{cases} \tag{19.16}$$

When expanding to higher orders, we obtain, as expected, expressions of increasing complexity. For the first group, the case $N=1$ reads as

$$\begin{aligned}\Phi_1^1(B,s) &= 2\Gamma_{11}S_1(B,s)+\Gamma_{11}\sum_{i=1}^{1}\sum_{g'\neq g}^{G}\sigma_{g'}c_{1g'}\Phi_{g'}^{(i-1)}(B,s) \\ &= 2\Gamma_{11}S_1(B,s)+\Gamma_{11}\left(\sigma_2 c_{12}\Phi_2^{(0)}(B,s)+\cdots+\sigma_G c_{1G}\Phi_G^{(0)}(B,s)\right) \\ &= 2\Gamma_{11}S_1(B,s)+\sigma_2 c_{12}\Gamma_{11}\Gamma_{22}S_2(B,s)+\cdots+\sigma_G c_{1G}\Gamma_{11}\Gamma_{GG}S_G(B,s)\end{aligned} \tag{19.17}$$

and for the generic group $g$ in the transformed space we have

$$\begin{aligned}\Phi_g^1(B,s) &= 2\Gamma_{gg}S_g(B,s)+\Gamma_{gg}\sum_{g'\neq g}^{G}\sigma_{g'}c_{gg'}\Phi_{g'}^{(0)}(B,s) \\ &= 2\Gamma_{gg}S_g(B,s)+\sum_{g'\neq g}^{G}\sigma_{g'}c_{gg'}\Gamma_{gg}\Gamma_{g'g'}S_{g'}(B,s).\end{aligned} \tag{19.18}$$

The solution obtained when expanding up to order $N=2$ writes as

$$\begin{aligned}\Phi_g^2(B,s) &= 3\Gamma_{gg}S_g(B,s)+\Gamma_{gg}\sum_{i=1}^{2}\sum_{g'\neq g}^{G}\sigma_{g'}c_{gg'}\Phi_{g'}^{(i-1)}(B,s) \\ &= 3\Gamma_{gg}S_g(B,s)+\Gamma_{gg}\left(\sum_{g'\neq g}^{G}\sigma_{g'}c_{gg'}\Phi_{g'}^{(0)}(B,s)+\sum_{g'\neq g}^{G}\sigma_{g'}c_{gg'}\Phi_{g'}^{(1)}(B,s)\right)\end{aligned}$$

$$= 3\Gamma_{gg}S_g(B,s) + 2\sum_{g'\neq g}^{G}\sigma_{g'}c_{gg'}\Gamma_{gg}\Gamma_{g'g'}S_{g'}(B,s)$$
$$+ \sum_{g'\neq g}^{G}\sum_{g''\neq g'}^{G}\sigma_{g'}c_{gg'}\sigma_{g''}c_{g'g''}\Gamma_{gg}\Gamma_{g'g'}\Gamma_{g''g''}S_{g''}(B,s). \tag{19.19}$$

The compact form at the end of expression (19.19) provides a guideline for the determination of the higher order solutions; the case $N = 3$ is

$$\begin{aligned}\Phi_g^3(B,s) &= 4\Gamma_{gg}S_g(B,s) + 3\sum_{g'\neq g}^{G}\sigma_{g'}c_{gg'}\Gamma_{gg}\Gamma_{g'g'}S_{g'}(B,s)\\ &+ 2\sum_{g'\neq g}\sum_{g''\neq g'}\sigma_{g'}c_{gg'}\sigma_{g''}c_{g'g''}\Gamma_{gg}\Gamma_{g'g'}\Gamma_{g''g''}S_{g''}(B,s)\\ &+ \sum_{g'\neq g}\sum_{g''\neq g'}\sum_{g'''\neq g''}\sigma_{g'}c_{gg'}\sigma_{g''}c_{g'g''}\sigma_{g'''}c_{g''g'''}\Gamma_{gg}\Gamma_{g'g'}\Gamma_{g''g''}\Gamma_{g'''g'''}S_{g'''}(B,s).\end{aligned} \tag{19.20}$$

The general case of expansion up to order $N$ can be written as

$$\begin{aligned}\Phi_g^N(B,s) &= (N+1)\Gamma_{gg}S_g(B,s)\\ &+ N\sum_{g_1\neq g}\sigma_{g_1}c_{gg_1}\Gamma_{gg}\Gamma_{g_1g_1}S_{g_1}(B,s)\\ &+ (N-1)\sum_{g_1\neq g}\sum_{g_2\neq g_1}\sigma_{g_1}c_{gg_1}\sigma_{g_2}c_{g_1g_2}\Gamma_{gg}\Gamma_{g_1g_1}\Gamma_{g_2g_2}S_{g_2}(B,s)\\ &\vdots\\ &+ \underbrace{\sum_{g_1\neq g}\sum_{g_2\neq g_1}\cdots\sum_{g_N\neq g_{N-1}}}_{\times N}\sigma_{g_1}c_{gg_1}\ldots\sigma_{g_N}c_{g_{N-1}g_N}\Gamma_{gg}\Gamma_{g_1g_1}\ldots\Gamma_{g_Ng_N}S_{g_N},\end{aligned} \tag{19.21}$$

where the group indexes in the sums have been generalized as $g_j, j = 1,\ldots,N$.

In order to obtain a compact and elegant expression for the group fluxes in the transformed and direct space, we introduce the constants $\Xi_{g,\ell_{max}}$ and the operators $T_{g,j_{max}}(\cdot)$ by

$$\Xi_{g,\ell_{max}} = \left\{\prod_{\ell=1}^{\ell_{max}}\sigma_{g_{\ell+1}}c_{g_\ell g_{\ell+1}}\right\}\sigma_{g_1}c_{gg_1}, \qquad \ell_{max}\in\{0,\ldots,N-1\}$$
$$T_{g,j_{max}}(\cdot) = \left\{\prod_{j=1}^{j_{max}}\Gamma_{k_jk_j}\right\}\Gamma_{gg}(\cdot), \qquad j_{max}\in\{0,\ldots,N-1\}. \tag{19.22}$$

where $\Xi_{g,0} = \sigma_{g_1} c_{gg_1}$ and $T_{g,0}(\cdot) = \Gamma_{gg}(\cdot)$. As a consequence, we can rewrite the expression (19.21) as

$$\begin{aligned}\Phi_g^N(B,s) &= (N+1)T_{g,0}(S_g(B,s)) \\ &+ N\sum_{g_1\neq g}\Xi_{g,0}T_{g,1}(S_{g_1}(B,s)) \\ &+ (N-1)\sum_{g_1\neq g}\sum_{g_2\neq g_1}\Xi_{g,1}T_{g,2}(S_{g_2}(B,s)) \\ &\vdots \\ &+ \underbrace{\sum_{g_1\neq g}\sum_{g_2\neq g_1}\cdots\sum_{g_N\neq g_{N-1}}}_{\times N}\Xi_{g,N-1}T_{g,N}(S_{g_N}(B,s)),\end{aligned} \tag{19.23}$$

or

$$\Phi_g^N(B,s) = (N+1)T_{g,0}(S_g(B,s)) + \sum_{\omega=1}^{N}(N-\omega+1)\left\{\underbrace{\sum\cdots\sum}_{\times\omega}\Xi_{g,\omega-1}T_{g,\omega}(S_{g_\omega})\right\}. \tag{19.24}$$

Expression (19.24) can be inverted in order to obtain the group fluxes $\Phi_g^N(x,t)$ noticing that the operator $T_{g,j_{max}}$, when the inverse Fourier and Laplace transforms are applied, satisfies the following relation:

$$\mathscr{L}^{-1}\left[\mathscr{F}^{-1}\left[T_{g,j_{max}}(S_g(B,s))\right]\right] = I^{S_g}_{m_1,\ldots,m_{g_{j_{max}}}}(x,t). \tag{19.25}$$

## 19.3 The Singularities of the Laplace Transform

The Laplace inverse transform is carried out by the use of the residue theorem, and, thus, it is required to study the distribution of the singularities of the transforms of the group fluxes: to this purpose, the behavior of the function $\Gamma(B,s)$ must be analyzed. This function depends on the integral term $A_{00}(B,s)$:

$$A_{00}(B,s) = \int_{-1}^{1}\frac{1}{\left(\sigma+\frac{s}{\mathrm{v}}\right)-iB\mu}d\mu = \frac{1}{B}\arctan\left(\frac{B}{\sigma+\frac{s}{\mathrm{v}}}\right), \tag{19.26}$$

that can be written in terms of the logarithm in the complex plane as

$$\arctan\left(\frac{B}{\sigma+\frac{s}{\mathrm{v}}}\right) = \frac{1}{2i}\log\left[\frac{\sigma+\frac{s}{\mathrm{v}}+iB}{\sigma+\frac{s}{\mathrm{v}}-iB}\right]. \tag{19.27}$$

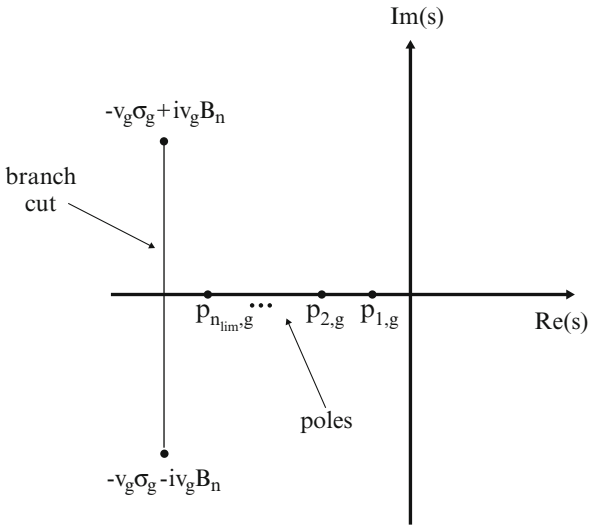


**Fig. 19.1** Poles and branch cut in complex plane.

The presence of a logarithmic terms introduces as a consequence a vertical branch cut is introduced in the complex plane for each group between the points $(-\mathrm{v}_g\sigma_g - i\mathrm{v}_g B)$ and $(-\mathrm{v}_g\sigma_g + i\mathrm{v}_g B)$. Being $B_n$ an increasing sequence, the span of the cut in the complex plane is increasing with $n$ (see Figure 19.1).

The characteristic equation to be solved in order to determine the location of the poles is

$$1 - c\frac{\sigma}{B_n}\arctan\left(\frac{B_n}{\sigma + \frac{s}{\mathrm{v}}}\right) = 0, \tag{19.28}$$

where $c$ is the number of secondaries per collision within each energy group. Letting $\tau_n = B_n/\sigma$, it is immediate to verify that a solution to Eq. (19.28) exists only if $\tau_n/c < \pi/2$. Therefore, the following chain of consequences holds:

$$\frac{B_n}{c\sigma} < \frac{\pi}{2} \quad \Rightarrow \quad \frac{(2n-1)\pi}{hc\sigma} < \frac{\pi}{2} \quad \Rightarrow \quad n < \frac{hc\sigma + 2}{4}. \tag{19.29}$$

Therefore, a polar singularity exists provided the following inequality is satisfied:

$$n \leq n_{lim} := \frac{hc\sigma + 2}{4}. \tag{19.30}$$

This evaluation holds for the transport kernel associated with each energy group, $\Gamma gg$: the existence of poles to be considered in the inverse Laplace transform is given by the condition:

$$n \leq n_{lim,g} := \frac{hc_{gg}\sigma_g + 2}{4}, \tag{19.31}$$

as can be seen graphically in Figure 19.1.

In conclusion, the inverse transform shall be constituted by the sum of an integral part along the branch cut and a time exponential term introduced by the residue at the polar singularity, whenever it exists.

## 19.4 Numerical Solution for Three Energy Groups

The analytical model for the solution of the linear transport equation in the multi-group case in the presence of a pulsed source is now applied to the case of three energy groups. The system is supposed to experience a pulsed source, distributed in all energy groups, located symmetrically in the center of the system, to preserve the symmetry of the solution we adopted as hypothesis. Therefore, the source in each energy group $g$ is constant, unitary and spans on the interval $[-x_{0,g}; x_{0,g}]$. The basic material and geometrical data adopted to evaluate the numerical solution are reported in Table 19.1, while the scattering matrix containing the energy transfer probabilities is assumed as

$$M_{scat} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} = \begin{bmatrix} 0.40 & 0.05 & 0.40 \\ 0.30 & 0.50 & 0.20 \\ 0.25 & 0.35 & 1.00 \end{bmatrix}. \tag{19.32}$$

As a first step, the evaluation of the poles and branch cuts of the Laplace-transformed transport kernels for each energy group is performed, as a function of the number of Helmholtz harmonics adopted for the inverse Fourier transform. Some example of the results obtained is given in Figure 19.2, where the physical dimension of the system $h$ has been changed to highlight how this parameter affects both the extension of the branch cut and the appearance of polar singularities.

The time-dependent solution to the transport problem in response to the pulsed source described above has been evaluated for the three energy groups, adopting 10 terms in the expansion (19.7) and using 300 Helmholtz eigenfunctions for the spatial representation of the flux. The behavior at different time instants is given in Figure 19.3, showing the initial distribution associated with the source shape and the consequent propagation at finite speed $\mathrm{v}_g$ characterizing the fluxes behavior at subsequent times. Moreover, the effects associated with the truncated series of spatial harmonic adopted for the representation are visible, especially for the third energy group where the flux should be null.

**Table 19.1** Material and geometrical data assumed in the transient evaluation. Dimensionless quantities are used.

| $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\mathrm{v}_1$ | $\mathrm{v}_2$ | $\mathrm{v}_3$ | $h$ | $x_{0,1}$ | $x_{0,2}$ | $x_{0,3}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1.00 | 1.00 | 2.00 | 1.00 | 0.50 | 0.10 | 10.00 | 0.50 | 0.25 | 0.25 |

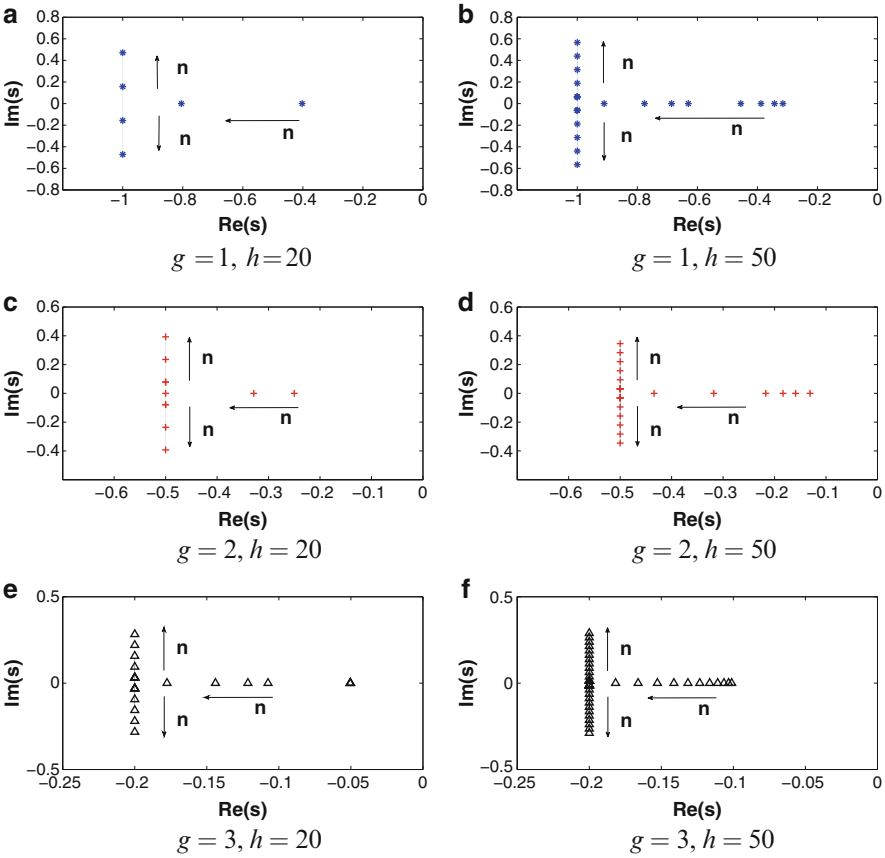


**Fig. 19.2** Localization of poles and extension of branch cut obtained in the Laplace transform inversion of the transport kernel for each energy group $g$ and different system dimensions $h$.

## 19.5 Conclusions

An analytical method to study the propagation of neutron pulses in the frame of multi-group neutron transport is presented. The method is based on the use of the double Fourier-Laplace transform, to deal with space and time, respectively. The technique yields exact results for times smaller than the transit time of the pulse through the spatial domain, to reach the boundary of the system. Use is made of the solution for the one-group problem, in combination with a multiple collision approach. Some results are presented and discussed for the three-group case. The convergence with respect to the number of harmonics used in the spatial series and of number of collision is investigated. Future developments will include the comparison with discrete-ordinate and spherical harmonics models. The series representation of the solution could also be accelerated by efficient novel techniques [Ga13].

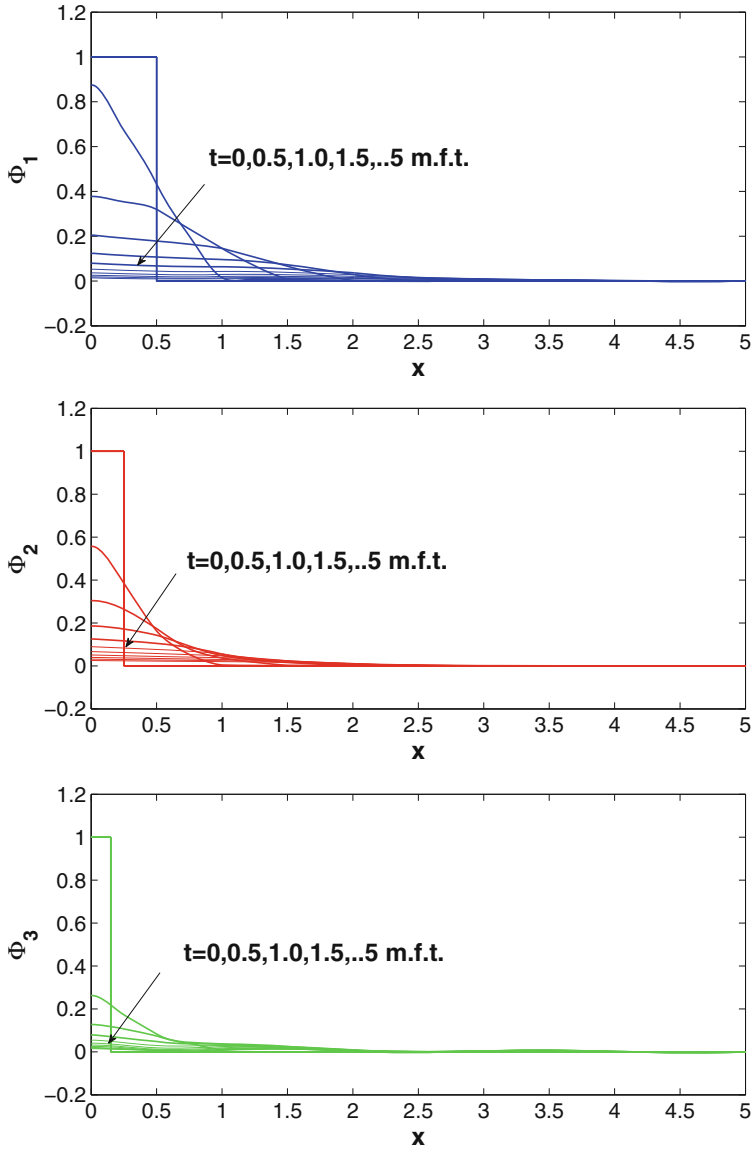


**Fig. 19.3** Spatial distribution of group fluxes $\Phi_g$ for a transient initiated by a unitary pulsed source in all groups. Flux expansion up to order $N = 10$; spatial representation with 300 Helmholtz eigenfunctions. Times are expressed as mean free times.

## References

[DuGaRa06] Dulla, S., Ganapol, B. D., Ravetto, P.: Space asymptotic methods for the study of neutron propagation. Annals of Nuclear Energy, **C 33**, 932–940 (2006).

[DuEtAl13] Dulla, S., Nervo, M., Ravetto, P., Carta, M.: Spatial and spectral effects in sub-critical system pulsed experiments. Proceedings of the International Conference on Mathematics and Computational Methods Applied to Nuclear Science and Engineering, M&C 2013, 1721–1735 (2013).

[DuRa04] Dulla, S., Ravetto, P.: Analytical solutions to discrete ordinate time dependent transport problems. Transactions of the American Nuclear Society, **C 90**, 278–280 (2004).

[DuRa08] Dulla, S., Ravetto, P.: Numerical aspects in the study of neutron propagation. Annals of Nuclear Energy, **C 35**, 656–664 (2008).

[Ga83] Ganapol, B. D.: Analytical benchmarks in time-dependent transport theory via the method of multiple collision. Transactions of the American Nuclear Society, **44**, 283 (1983).

[Ga13] Ganapol, B. D.: What is convergence acceleration anyway?. Integral Methods in Science and Engineering, Birkhauser, 115–135 (2013).

# Chapter 20
# Infiltration in Porous Media: On the Construction of a Functional Solution Method for the Richards Equation

**I.C. Furtado, B.E.J. Bodmann, and M.T.B. Vilhena**

## 20.1 Introduction

In engineering, knowledge about infiltration and water movement in soil emerges as a preventive measure, both to control the destructive action of water on foundations, dams, and pavements and to predict the behavior of flow and transport of pollutants. Mathematical modeling of these infiltration processes in porous media is substantiated by the equations of Richards, or Fokker–Planck. Both equations are highly nonlinear, so that analytical solutions to the equations are extremely difficult to find. In order to turn prognostics in applications more efficient, it is essential to consider field observations, because they are necessary for identification of constitutive relations that govern the phenomenon and may be used in theoretical formulations. The best-known models that relate soil parameters are the models found in refs. [BrCo64, Ge80] and [Ga58]. The Van Genuchten model provides more satisfactory results than others when compared with experimental data, but due to its functional form proposed solutions have limited applicability. On the other hand, the other two models result in simplified equations, leading to cases of linearized equations and their associated solutions, as, for instance, in [Ba99, ChTaCh01, Ba02]. However, most of these solutions are limited to cases with uniform initial conditions and in an infinite domain.

In this contribution, we analyze a problem of transient flow of water in unsaturated media, modeled by the Richards equation. To this end, the constitutive relations of Van Genuchten are employed and a hybrid method of Padé approximations and Adomian decomposition [Ad94] is applied. Although Adomian states in his work that this method should be applicable to any nonlinear problem, the present

I.C. Furtado (✉) • B.E.J. Bodmann • M.T.B. Vilhena
Federal University of Rio Grande do Sul, Av. Osvaldo Aranha, 99/4, Porto Alegre 90046-900, RS, Brazil
e-mail: igorjara@gmail.com; bardo.bodmann@ufrgs.br; vilhena@math.ufrgs.br



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_20

problem is a counterexample, where the method in its original form fails. In order to circumvent this shortcoming, we propose a construction of a functional solution method for the Richards equation, that shall replace the recursion initialization of the decomposition method. The found solution is then optimized and its accuracy evaluated by the nonlinear Richards equation and the profile of the potential matrix is also compared to numerical findings [WePi10]. It is remarkable that for the present parameter set the recursion initialization is already considerably close to the true solution, so that in the present case no further recursion steps are necessary.

## 20.2 The Model

The governing equation describing infiltration in porous media is established by the Darcy–Buckingham and the continuity equation.

$$\vec{q} = -K(\theta)\vec{\nabla}\Phi \quad \text{and} \qquad \frac{\partial\theta}{\partial t} = -\vec{\nabla}\vec{q} \tag{20.1}$$

Here $\vec{q}$ in $(m/s)$ is the specific flow, $K(\theta)$ in $(m/s)$ is the hydraulic conductivity depending on soil moisture $\theta$, and $\Phi$ signifies the hydraulic potential in units of $(m)$. Equation (20.1) has a considerable mathematical complexity due to the nonlinearity present in the hydraulic conductivity. Moreover, this equation is established for steady state condition or dynamical equilibrium. Though, most situations in nature are transients, and to describe such scenarios time dependence is introduced by the continuity equation.

For convenience one may split $\Phi = \psi + z$ into the matrix potential that contains the essential effects attributed to porosity, and the gravitational potential represented by the soil depth. Nevertheless equation system (20.1) needs an additional relation so that the system can be solved with one unique solution for $\Psi$. To this end, one may use the model of Van Genuchten [Ge80], which is capable of characterizing the zone of capillary rise and is applicable from zero to saturation condition. The relationship between volumetric water content $(\theta)$ and the matrix potential $(\psi)$ is parametrized as $\psi = \alpha^{-1}\left(S(\theta)^{-1/m} - 1\right)^{1/q}$, where $m = 1 - 1/q$ is related to the effective saturation, $\alpha, q$ are parameters dependent on soil properties and $S(\theta) = (\theta - \theta_r)/(\theta_s - \theta_r)$, with $\theta_s$ and $\theta_r$ the saturated and residual soil water content. The relationship between $K$ and $\theta$ is given by

$$K(\theta) = K_s S(\theta)^{1/2}\left(1 - \left(1 - S(\theta)^{1/m}\right)^m\right)^2,$$

where $K_s$ is the saturated hydraulic conductivity. Thus, one may write $K$ as a direct function of $\psi$

$$K(\psi) = K_s\left((\alpha\psi)^q + 1\right)^{-m/2}\left(1 - \left(1 - \left((\alpha\psi)^q + 1\right)^{-1}\right)^m\right)^2.$$

Combining the Darcy–Buckingham and continuity equation together with the van Genuchten relation allows to cast the problem in the form

$$C(\psi)\frac{\partial \psi}{\partial t} = \vec{\nabla}[K(\psi)\nabla\psi] + \frac{dK(\psi)}{d\psi}\vec{\nabla}\psi\,, \tag{20.2}$$

where $C(\psi) = \frac{d\theta}{d\psi}$ is called hydraulic capacity. Equation (20.2) is known as the Richards equation. This equation governs the movement of water in unsaturated soil and can be applied in the whole domain even for distinct saturated and unsaturated areas [WePi10]. The specific water capacity of the soil is explicitly given by

$$C(\psi) = \frac{mq\alpha^q(\theta_r - \theta_s)\psi^{q-1}}{(1+(\alpha\psi)^q)^{m+1}}\,.$$

For all numerical calculations that follow we use the parameter of [CeBoZa90], $\alpha = 3.35m^{-1}$, $q = 2$, $K_s = 9.92 \times 10^{-5}ms^{-1}$,$\theta_s = 0.368m^3m^{-3}$ and $\theta_r = 0.102m^3m^{-3}$. The considered depth range in soil is $[0, L = 1m]$ and initial and boundary conditions are $\psi(z,0) = -10m$, $-L \le z \le 0$; $\psi(0,t) = -0.75m$ and $\psi(-L,t) = -10m$ for $t > 0$.

Note that the model of Van Genuchten is widely used in numerical simulations, but due to its complexity in functional form algebraic manipulations are rather complicated that seem to make analytic solutions unattractive. To circumvent some of these difficulties, functional Padé approximants were used to substitute the expressions for $C$ and $K$. Note that these parameters are phenomenological relations, so that the representation by approximants can be used without loss of generality. The simplest Padé representation for $K$ and $C$ in the region of interest is $[1/3]$ with expansion point $\psi = -0.2$ for $C$ and $\psi = -0.1$ for $K$, respectively. A comparison of the original expressions and the Padé approximations for $K$ and $C$ are shown in figure 20.1.

## 20.3 Construction of a Parametrized Solution

In the sequel, we consider the $1 \oplus 1$ dimensional space-time version of the Richards equation with the aforementioned initial and boundary conditions. The initial idea to employ the Adomian decomposition method [Ad94] was abandoned after several attempts to implement the method even in different ways, because the latter did not attain convergent results. In order to recover at least a part of the procedure that that was proven to be useful in other applications, we first construct a solution that shall be a reasonable initial solution for the recursive scheme so that all remaining corrections are sufficiently small and the scheme converges. The equation to be solved for the matrix potential is the Richards equation.

$$C(\psi)\frac{\partial \psi}{\partial t} = \frac{\partial}{\partial z}\left[K(\psi)\left(\frac{\partial \psi}{\partial z} + 1\right)\right]$$

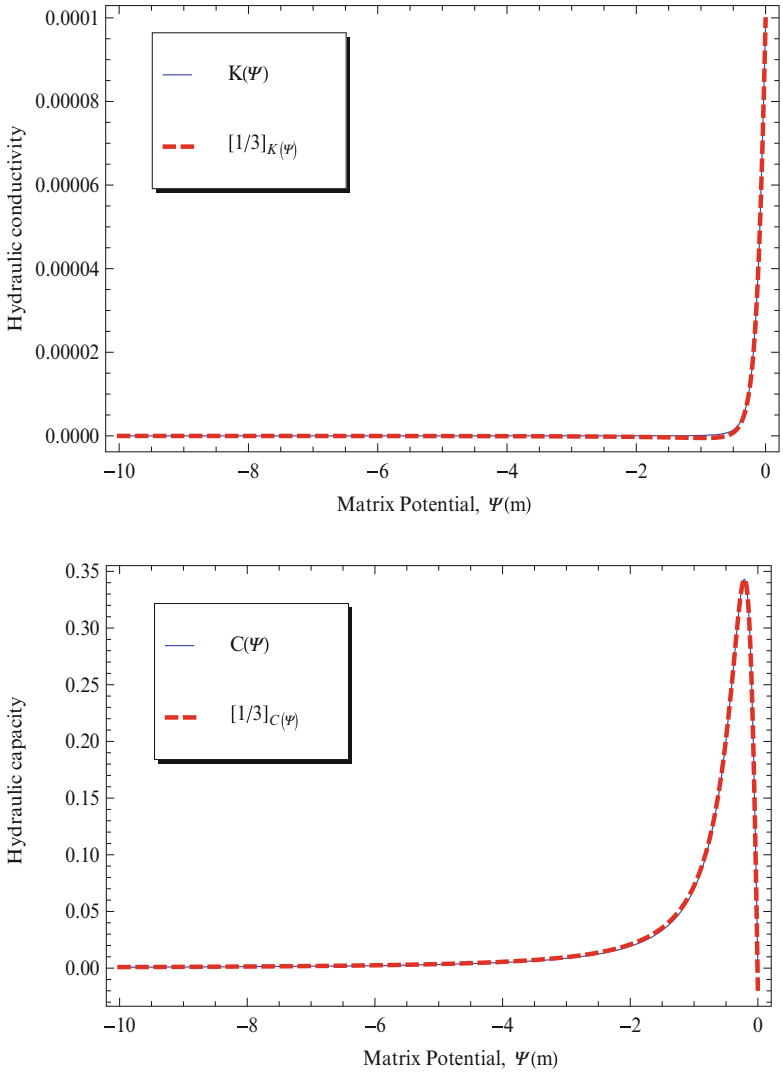


**Fig. 20.1** Comparison of [1/3] Padé approximants for $K(\psi)$ and $C(\psi)$ with the original expressions.

### 20.3.1 Transient and Steady State Regimes

From comparison to experimental findings, one expects the matrix potential to assume negative values in the range of $[-10,0]$. From inspection one observes that for a restriction of $\psi \in [-10,-2]$ the hydraulic conductivity may be approximated by a constant $K(\psi) \approx K$ and the hydraulic capacity may be approximated by a polynomial $C(\psi) \approx a(\psi+10)^6$. For convenience we introduce the substitution $\psi + 10 \to \phi$ and solve the resulting equation

$$a\phi^6 \frac{\partial \phi}{\partial t} = K \frac{\partial^2 \phi}{\partial z^2}.$$

This equation has an implicit travelling wave solution as shown in [PoZa03],

$$\lambda^2 \int \frac{d\phi}{F(\phi)+C_1} = t + \lambda z + C_2\,, \qquad F(\phi) = \int \frac{a\phi^2}{K}\, d\phi\,,$$

where $\lambda, C_1$ and $C_2$ are constants which are determined from the initial and boundary conditions of the problem.

The found solution already allows us to analyze some properties of the dynamics of the system, namely the transient and stationary regime of the matrix potential. Figure 20.2 shows the plot of contours with constant matrix potential $\phi$ as a function of time and depth. One nicely observes that contours with $\phi \geq 3$ (ore equivalently $\psi \geq -7$) effectively reduce the dimension of the problem by one so that either

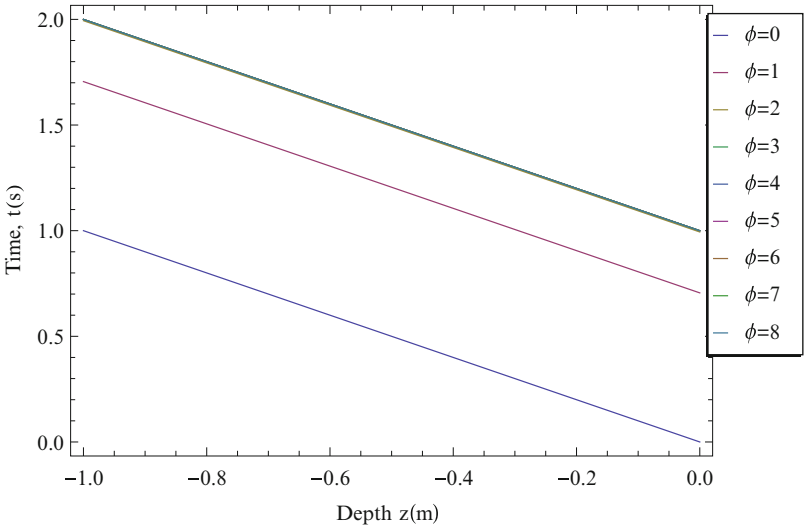


**Fig. 20.2** Transient to steady state evolution of constant matrix potential contours $\phi$.

the time or the space variable may be substituted in the problem and hence can be attributed to stationarity.

### 20.3.2 The Stationary Solution

From the finding that a stationary regime exists for the expected values of $\psi$ and thus $\phi$, we solve the time independent problem but this time using a polynomial expression for the logarithmic hydraulic conductivity $\ln K(\psi) = a(\psi - b)^2 - c$.

$$0 = \frac{\partial K(\psi)}{\partial z}\left[\frac{\partial \psi}{\partial z} + 1\right] + K(\psi)\frac{\partial^2 \psi}{\partial z^2}$$

$$0 = \frac{\partial \ln K(\psi)}{\partial z}\left[\frac{\partial \psi}{\partial z} + 1\right] + \frac{\partial^2 \psi}{\partial z^2}$$

$$0 = a\frac{\partial}{\partial z}(\psi - b)^2\left(\frac{\partial \psi}{\partial z} + 1\right) + \frac{\partial^2 \psi}{\partial z^2}$$

Here, $a$, $b$, and $c$ are parameters determined such as to minimize the difference between the polynomial function to the original expression. Upon substitution of $\psi \rightarrow \phi = \psi - b$ we solve the resulting ordinary differential equation using a decomposition method, where $\phi = \sum_{i=0}^{\infty} \phi_i$.

$$0 = \underbrace{\frac{\partial^2 \phi}{\partial z^2} + a\frac{\partial \phi^2}{\partial z}}_{A\ (Initialisation)} + \underbrace{a\frac{\partial \phi^2}{\partial z}\frac{\partial \phi}{\partial z}}_{B\ (Correction)} \tag{20.3}$$

Equation (20.3) can be solved using a recursive method, where the terms $A$ are used to determine the initialization and terms of $B$ are considered as a correction for the subsequent recursions. Therefore, if we assume that $\phi_0$ is the first term of the recursion, then $\phi_0$ is solution of the following equation

$$\frac{\partial^2 \phi_0}{\partial z^2} + a\frac{\partial \phi_0^2}{\partial z} = 0$$

with known solution

$$\phi_0(z,t) = -\sqrt{\frac{c_1}{a}}\tanh\left(\sqrt{ac_1}(-z + c_2)\right)$$

To determine $\phi_1$ in the second recursion step, the term $B$ is now considered as a source using the previously determined solution $\phi_0$.

$$\frac{\partial^2 \phi_1}{\partial z^2} + a\frac{\partial \phi_1^2}{\partial z} = -a\frac{\partial \phi_0^2}{\partial z}\frac{\partial \phi_0}{\partial z} \tag{20.4}$$

Since $\phi_0$ is the homogeneous solution of eq. (20.4) the method of variation of parameters [ON11] leads to the particular solution. Thus, the particular solution $\phi_p = v(z)\phi_0$ with $v(z)$ is determined from

$$v'(z)\phi_0(z) = -a\frac{\partial \phi_0^2}{\partial z}\frac{\partial \phi_0}{\partial z}$$

and

$$\phi_1(z) = \phi_0 \underbrace{c_3}_{=0} + \phi_0(z)\int_0^z \frac{1}{\phi_0(z')}\left(-a\frac{\partial \phi_0^2}{\partial z'}\frac{\partial \phi_0}{\partial z'}\right) dz' ,$$

in this case $c_3 = 0$ because of the zero boundary conditions. Note that the boundary conditions of the problem were already absorbed in the determination of the solution $\phi_0$ so that all remaining extra boundary conditions are zero, which is a peculiarity of the decomposition method. For all the terms $\phi_i$ $(i > 1)$ the procedure is repeated in an analogue fashion. By inspection one finds that the first term of the form $\psi_0(z,t) = a_1 \tanh(a_3 z + a_4) + a_2$ is the dominant one and the result of the first recursion only a correction. The constants may be found using the Richards equation and minimizing the error.

### *20.3.3 The Time-Dependent Solution*

Phenomenological arguments allow now to extend the stationary solution including a time dependence as follows. With increasing infiltration the surface region approaches local saturation so that the scenario characterized by the initial condition is shifted towards increasing depth. Saturation is already present in the asymptotic behavior of the hyperbolic tangent function and because of the initial condition $(\psi(z,0) = -10, -L \leq z < 0)$ the argument of tanh shall be singular. The simplest way to introduce a shift is adding a term $a_4/t$ to $z$ in the argument of the hyperbolic tangent function. Last, we apply some "cosmetics" to our solution by observing that there exists an asymmetry between the convex and concave parts of the profile, i.e. the edge towards the saturated region is sharper than the one at the edge where the matrix potential assumes a numerical value of approximately $-10\ m$. This may be achieved by multiplying the hyperbolic tangent's argument by a factor $1 + \exp\left(a_3 z + a_4 + \frac{a_5}{t}\right)$. Thus we arrive at a solution in parametrized form, that we evaluate using the original Richards equation.

$$\psi(z,t) = -a_1 \tanh\left(\left(1 + e^{a_3 z + a_4 + \frac{a_5}{t}}\right)\left(a_3 z + a_4 + \frac{a_5}{t}\right)\right) + a_2 \tag{20.5}$$

Now, the matrix potential $\psi$ is given as a parametrized function $\psi = \psi(z,t;\{a_i\})$ with parameter $a_i$ $(i = 1,2,3,4,5)$, where the unknown parameter has to be determined.

### 20.3.4 Optimization

To adjust the parameter set, we insert the parametrized solution ($\psi_P$) given in equation (20.5) into the governing equation, which for convenience we write in a form where all terms are on the left-hand side and consequently the right-hand side is zero. Let $\Omega_R$ be the differential operator that represents the Richards equation with all terms to the left, then for the true solution $\Omega_R[\psi_T] = 0$ holds. Since our solution is an approximate solution the right hand side differs from zero by a residual term $\Omega_R[\psi_P] = R(z,t)$. Thus, the solution presented in eq. (20.5) is optimized minimizing $R(z,t)$ using the method of nonlinear least squares optimization and refined by Newton's method, so that this procedure constitutes a self-consistency test.

Some constants can be determined *a priori* the optimization. We can fix the constants $a_1$ and $a_2$ directly using the boundary conditions where $a_1 = (\psi(0,t) - \psi(L,t))/2$ and $a_2 = \psi(0,t) - a_1$. The remaining parameter are determined using the afore mentioned minimization of $R$. The objective function that is to be minimized is

$$\sum_{i=0}^{M}\sum_{j=0}^{N}\left[C(\psi)\frac{\partial\psi}{\partial t} - \left(\frac{\partial}{\partial z}[K(\psi)\frac{\partial}{\partial z}\psi] + \frac{\partial K(\psi)}{\partial\psi}\frac{\partial\psi}{\partial z}\right)\right]^2\Bigg|_{(z_i,t_j)} \rightarrow \min\ .$$

Since the asymptotics of the solution was fixed using the boundary conditions we use a discrete set of points in the range that contains maximum curvature and the inflection point to optimize $\{a_3,\ldots,a_5\}$. The optimization may then be simplified using an expansion of the hyperbolic tangent function around the inflection point (at $z_0$), i.e. where the argument of the function is zero $a_3 z_0 + a_4 + \frac{a_4}{t} = 0$ allows to solve the minimization problem in a straightforward fashion.

## 20.4 Results

The parameter set [CeBoZa90] that was used for an application refers to a situation that considers the infiltration of water in a column of initially dry homogeneous soil. Figure 20.3 shows the computed matrix potential for a sequence of times and again one observes the already analyzed transition from the transient to the steady state regime. We further show in figure 20.4 the self-consistency test along the vertical coordinate and present some numerical values in table 20.1 for the largest deviations from the parametrized to the true solution.

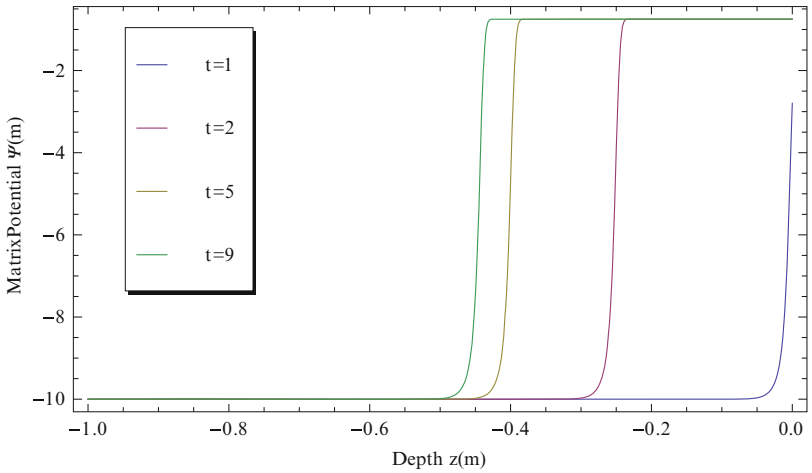


**Fig. 20.3** Matrix potential profile with depth for a selection of times.

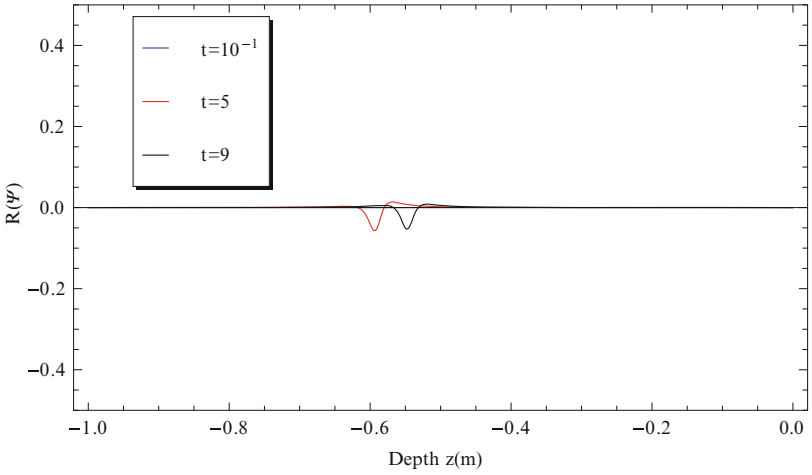


**Fig. 20.4** Self-consistency test profile with depth for a selection of times.

We further compare the parametrized solution to the calculated matrix potential profile from a numerical approach [WePi10], shown in figure 20.5. It is noteworthy that the difference between the two solutions is in a range where the error of the parametrized solution is negligible. From table 20.1 one may conclude that the bumps in the curves that appear in figure 20.4 are small indeed and so is the error of the solution, so that one can say that the parametrized solution is close to the true solution.

**Table 20.1** Self-consistency test along the vertical coordinate.

| Largest deviation $R[\psi]/\psi$ | | |
|---|---|---|
| $t = 10^{-1}$ | $t = 5$ | $t = 9$ |
| 0.00025 | 0.005 | 0.0045 |

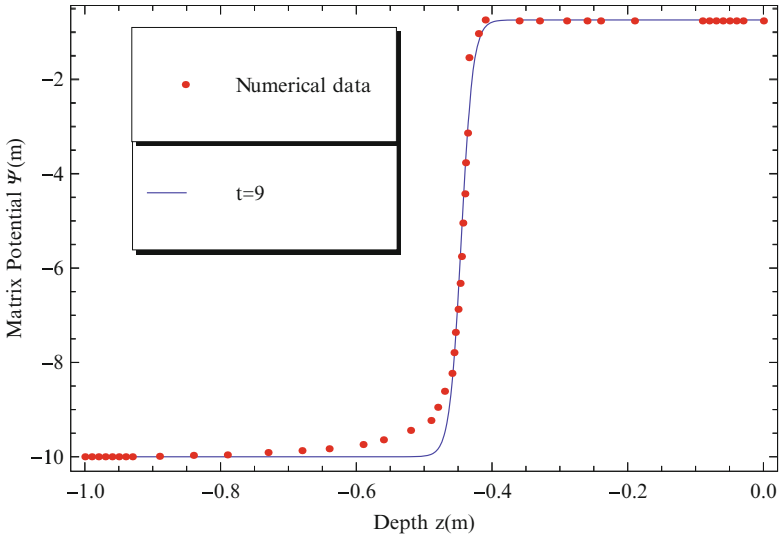


**Fig. 20.5** Parametrized solution compared to a result from a numerical approach [WePi10].

## 20.5 Conclusions

In the present work we discussed a methodology to construct a parametrized solution for the Richards equation. It is remarkable that already the initial solution for a recursive scheme given in a relatively compact formula is close to the true solution, so that one may efficiently simulate one dimensional flow of water in unsaturated and saturated porous media. The main difficulties of the problem were the non-linearity and the initial condition. Although Adomian's decomposition method has been used successfully in a variety of applications, in the present case, this scheme strongly diverges and does not solve the nonlinear Richards equation. Moreover, one could argue that using Padé approximations should simplify the decomposition, in an analytical sense, however the numerically increasing source term with increasing recursion depth clearly turns this procedure a divergent sequence. In this sense the Richards equation is a counterexample to the statement, that Adomians prescription should in principle work for any nonlinearity and lead to a convergent recursion procedure. Various implementations of Adomian's idea suffered from the same problem, the increasing source term contributions to the corrections of the solution per recursion step.

The parametrized solution, which was presented in eq. 20.5, when optimized by the method of least squares and nonlinear Newton's method, gave fairly good results for the matrix potential profile. A self-consistency test accused only small

differences between the true and the parametrized solution. A similar conclusion may be drawn from the comparison to a numerical solution from the literature [WePi10]. For the soil specification used to perform the numerical calculations we found that no further recursion was of need, however, for other soil compositions and their associated parameter sets one cannot expect that the hyperbolic function formula is as good an approximation as for the case discussed in this contribution. Nevertheless, it is quite plausible that using the hyperbolic tangent expression as recursion initialization will necessitate only a few recursion steps to obtain an acceptable solution within a predefined accuracy.

## References


[Ad94] Adomian, G.: Solving Frontier Problems of Physics: The Decomposition Method. Kluwer Academic Publishers, The Netherlands (1994).

[Ba99] Basha, H.A.: Multidimensional linearized nonsteady infiltration with prescribed boundary conditions ate the soil surface. Water Resources Research, **35**(1), 75–83 (1999).

[Ba02] Basha, H.A.: Burgers equation: A general nonlinear solution of infiltration and redistribution. Water Resources Research, **38**(11), 29.1–29.9 (2002).

[BrCo64] Brooks, R.H. and Corey, A.T.: Hydraulic Properties of Porous Media. Hydrol. paper 3., Colorado State University (1964).

[CeBoZa90] Celia, M.A., Bouloutas, E.T., and Zarba, R.L.: A general mass conservative numerical solution for the unsaturated flow equation. Water Resources Research, **26**(30), 1483–1496 (1990).

[ChTaCh01] Chen, J.M., Tan, Y.C., Chen, C.H., and Parlange, J.Y.: Analytical solutions for linearized Richards equation with arbitrary time-dependent surface fluxes. Water Resources Research, **37**(4), 1091–2001 (2001).

[Ga58] Gardner, W.R.: Some steady state solution of unsaturated moisture flow equations with application evaporation from a water table. Soil Science, **85**, 228–232 (1958).

[Ge80] Genuchten, M.T.: A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Science Society of America Journal, **44**, 892–898 (1980).

[ON11] O'Neil, P.V.: Advanced engineering mathematics. International Student Edition, University of Alabama at Birmingham, Ed. **7**, 82–85 (2011).

[PoZa03] Polyanin, A.D. and Zaitsev, V.F.: Handbook of Nonlinear Partial Differential Equations. Chapman and Hall/CRC, (2003).

[WePi10] Wendland, E. and Pizarro, M.L.P.: Modelagem computacional do fluxo unidimensional de Água em meio não saturado do solo. Engenharia Agrícola, Jabotical. **30**(3), 424–434 (2010).

# Chapter 21
# A Soft-Sensor Approach to Probability Density Function Estimation

**M. Ghaniee Zarch, Y. Alipouri, and J. Poshtan**

## 21.1 Introduction

In probability and statistics, density estimation is the construction of an estimate, based on observed data, of an unobservable underlying Probability Density Function (PDL8). There are two main approaches to estimate the PDF. In the first approach, a special sensor is designed to measure the PDF of a signal. In recent years, such sensors are becoming available. For example, one can now use optical sensors and digital cameras to pick profile images and then transfer these imagines into a mathematical representation such as probability density functions. In the second approach, a mathematical tool is utilized to approximate the PDF. Generally, since a PDF is a nonlinear and positive function with an integral constraint, determining the output PDF requires some complicated mathematical techniques such as partial differential equations. Both introduced approaches to estimate the PDF have some drawbacks. The first approach needs physical equipment that may be expensive, and a special sensor must be designed for each specific application. The second approach is usually time-consuming and not well developed.

One approach to fill this gap is using soft-sensor methods. Soft-sensor methods are mixture of both approaches which have advantages of both, without need to design any physical equipment.

One approach to density estimation is *non-parametric*. A variety of approaches to non-parametric density estimation have been proposed, such as Histograms [Or13], Naive estimator [We72], Kernel estimator [MaSc14], Nearest neighbor method [WeTiWa14], Orthogonal series estimators [Sc67], Maximum penalized likelihood estimators [TaGoRe14], General weight function estimators [Si86], Parzen estimators [WaJo95, ScSz01, ZhKw06], Expectation maximization (EM)

M. Ghaniee Zarch (✉) • Y. Alipouri • J. Poshtan
Iran University of Science and Technology, Narmak, Tehran 16846, Iran
e-mail: majidghaniee@iust.ac.ir; yalipouri@iust.ac.ir; jposhtan@iust.ac.ir



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_21

algorithm [McKr97, FiJa02, ZiVa04], Variational estimation [CoBi01, McTi07]. The most basic form of density estimation is a rescaled histogram. It is the oldest and most widely used density estimator, however the discontinuity of histograms causes extreme difficulty if derivatives of the estimates are required [Si86].

In statistics, kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample.

A range of kernel functions are commonly used: uniform, triangular, biweight, triweight, Epanechnikov, normal, and others. The Epanechnikov kernel is optimal in a minimum variance sense [Ep69], though the loss of efficiency is small for the kernels listed previously [WaJo95], and due to its convenient mathematical properties, the normal kernel is often used.

In particular, we focus on Gaussian Mixture Models (GMM), which are known to be a powerful tool in approximating distributions even when their form is far from Gaussian [WaJo95]. A GMM is a probability density function represented as a weighted sum of Gaussian component densities [PrSa14].

There are several techniques available for estimating the parameters of a GMM [Mc88]. By far the most popular and well-established methods are Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation. However, their extension to online estimation of mixture models is nontrivial, since they assume all the data is available in advance (batch learning) [KrSkLe10].

The process of online learning should create, update, and modify models of the perceived data in a continuous manner, while still keeping the representations compact and efficient. Various models and methods for their extraction have been proposed in different contexts and tasks [ArEtAl92, Ar04, KiWeKo05, SoWa05, ArCi05]. To deep review on online methods to estimate GMM refer to [KrSkLe10]. Based on authors' knowledge, there is not any method which is used fuzzy tools to online estimation of kernel density function. Most of the proposed methods have conflict with requirement for online applications and their complexity increase with time. We propose a method using fuzzy approach to fulfill these requirements without losing accuracy.

The remainder of the chapter is outlined as follows: In Section 21.2, the method of online kernel density estimation is proposed. Section 21.3 presents simulation study to clarify the effectiveness of the proposed method. Conclusions are drawn in Section 21.4.

## 21.2 Online Kernel Density Estimation

Throughout this chapter, we will refer to a class of kernel density estimates based on Gaussian kernels, which are commonly known as the Gaussian mixture models. A one-dimensional M-component Gaussian mixture model is a weighted sum of M

component Gaussian densities as given by Equation 21.1

$$\hat{f_\sigma}(x) = \sum_{j=1}^{M} w_j g\left(x|\mu_j, \sigma_j^2\right) \tag{21.1}$$

where $w_j$ is the weight of the $j$th component and $g\left(x|\mu_j, \sigma_j^2\right)$ is a Gaussian-kernel

$$g\left(x|\mu_j, \sigma_j^2\right) = \left(2\pi\sigma_j^2\right)^{\frac{1}{2}} \exp\left(-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right)$$

centered at mean $\mu_j$ with standard deviation $\sigma_j^2$; note that $\sigma_j^2$ is also known as the bandwidth of the Gaussian-kernel. The mixture weights satisfy the constraint that

$$\sum_{j=1}^{M} w_j = 1$$

Suppose that we have observed a set of $n_t$ samples $\{x_i\}_{i=1:nt}$ up to some time-step $t$. The problem of modeling samples by a probability density function can be posed as a problem of kernel density estimation [WaJo95]. Here, a fuzzy approach is proposed for estimating the mixture weights. The fuzzy logic model is empirically based, relying on an operator's experience rather than their technical understanding of the system. It consists of three main operators: fuzzifier, rule inference, and defuzzifier. From input output point of view, the constructed fuzzy model can be written as

$$f(x) = \frac{\sum_{l=1}^{M} \alpha_l w_l \bar{y}_l}{\sum_{l=1}^{M} \alpha_l w_l} \tag{21.2}$$

where $M$ denotes the number of rules and is fixed, $\alpha_l$ denote rules weights, $\bar{y}_l$ is center of output memberships, and $w_l$ are defined as:

$$w_l = \prod_{i=1}^{n} \mu_{A_i^l}(x)$$

where $\mu_{A_i^l}$ are memberships values. The common defuzzifier method is centroid. In centroid defuzzifier method $\bar{y}_l$ is the center of the activated output membership. In this study, the goal is estimating a Gaussian mixture model. Hence, considering Eq. 21.1, $\bar{y}_l$ must be a Gaussian function. Therefore, we change the defuzzifier method such that $\bar{y}_l$ is the activated output membership function (not its centroid value). In the proposed structure, in each step, just one of the output membership

functions which has higher membership value is activated. Therefore, Eq. 21.2 can be rewritten as

$$\hat{f}_{\sigma}(x) = \frac{\sum_{j=1}^{M} w_j(x) \times \alpha_j \times g_j\left(x|\mu_j, \sigma_j^2\right)}{\sum_{j=1}^{M} w_j(x) \times \alpha_j} \tag{21.3}$$

Notice that all membership functions are selected Gaussian.

### 21.2.1 Tuning the Model Parameters

Parameters of model (21.3) which must be adapted are $\left\{\alpha_j, \mu_j, \sigma_j^2\right\}$. To implement this, the adaptation process is performed in two stages. In first stage, one of the output membership parameters $\left(\mu_j, \sigma_j^2\right)$ is updated using observed data at each new samples, then the rules weights are adapted by minimizing the mean square error cost function.

**Stage 1:** As stated above, just one of the output membership functions is activated for each new sample. The parameters of the activated output membership function is updated by the new sample data $x_t$ as follows.

$$\mu_{t+1} = \mu_t + \frac{1}{t+1}(x_t - \mu_t) \tag{21.4}$$

$$\sigma_{t+1}^2 = \frac{t}{t+1}\sigma_t^2 + \frac{t}{(t+1)^2}(x_t - \mu_t)^2 \tag{21.5}$$

Above equations are intuitively derived from

$$\overline{\mu}_T = \frac{1}{T}\sum_{t=1}^{T-1} x_t$$

$$\overline{\mu}_{T+1} = \frac{1}{T+1}\sum_{t=1}^{T} x_t = \frac{1}{T+1}(T\mu_T + x_T) = \mu_T + \frac{1}{T+1}(x_T - \mu_T)$$

and similarly,

$$\sigma_{T+1}^2 = \frac{1}{T+1}\left(\sum_{t=1}^{T} x_t^2\right) - \mu_T^2 = \frac{1}{T+1}\left(T\sigma_T^2 + \left(x_T - \mu_T^2\right)\right) - \mu_T^2$$

$$= \frac{T}{T+1}\sigma_T^2 + \frac{T}{(T+1)^2}(x_T - \mu_T)^2$$

**Stage 2:** The value of the rules weights are updated by minimizing expected $L_2$ risk function, also termed the mean integrated squared error

$$MISE(h) = E\int \left(\hat{f}_h(x) - f(x)\right)^2 dx$$

The above criteria cannot be calculated by the data as the true value $f(x)$ is not previously known. Suppose $X$ is a random variable and that all of the moments exist. Further, suppose the probability distribution of $X$ is completely determined by its moments, i.e., there is no other probability distribution with the same sequence of moments. If

$$\lim_{N\to\infty} E\left(x_N^k\right) = E\left(x^k\right)$$

for all values of $k$, then the sequence $\{X_n\}$ converges to $X$ in distribution. In other words, if two distributions have the same moments (i.e., same moment-generating function), then they are identical at all points [GrSn97]. Therefore, we redefine the cost function as

$$MSE = \sum_{h=1}^{p} c_h \left(\frac{1}{N_1}\sum_{t=1}^{N_1} \hat{x}_t^h - \frac{1}{N_2}\sum_{t=1}^{N_2} x_t^h\right)^2 \tag{21.6}$$

where $c_h$ is weighting coefficient which is intuitively we select $c_{h-1} > c_h$ . $x_t : t = 1,\ldots,N_2$ are last observed data, $N_2$ is the window width, and $\hat{x}_t$ are data produced by the estimated kernel density function $\hat{f}(x)$. From practical point of view, we can suppose $N_1 \gg N_2$.

To minimize cost function (21.6), the gradient descent method has been utilized. The rule weights are updating as

$$\alpha_j^{t+1} = \alpha_j^t - 2\eta\left[\sum_{h=1}^{p} c_h \left(\frac{1}{N_1}\sum_{t=1}^{N_1} \hat{x}_t^h - \frac{1}{N_2}\sum_{t=1}^{N_2} x_t^h\right)\right]\frac{\partial \hat{f}_h(x)}{\partial \alpha_j^t} \tag{21.7}$$

where $\eta$ is the step length.

$$\frac{\partial \hat{f}_h(x)}{\partial \alpha_i} = \frac{\sum_{j=1}^{M} w_j(x) \times w_i(x) \times \alpha_j \times \left(g\left(x|\mu_i,\sigma_i^2\right) - g\left(x|\mu_j,\sigma_j^2\right)\right)}{\left(\sum_{j=1}^{M} w_j(x) \times \alpha_j\right)^2}$$

Considering above explanations, the algorithm is summarized in following steps:

1. Initialize the membership functions, rules and the rules weights.
2. Calculate the membership value for the new input sample.

3. Decide the rule with maximum inference operator
4. Determine the activated output membership function
5. Tune the activated output membership function by Eqs. 21.4 and 21.5.
6. Tune the rules weights by Eq. 21.7.
7. Calculate the cost function (21.6).
8. If the stop criteria is not met, back to step 2.

## 21.3 Simulation

In this section, the capability of the proposed method in online estimation of kernel density function is tested on Gaussian probability density function. In this part, we assume that input variable $x_t$ for PDF belongs to $[0,1]$. In Step 1, we define three input fuzzy sets in $[0,1]$, and three output fuzzy sets initialized in $[0,1]$, where the input and output membership functions are shown in Figure 21.1. Then, the output function (estimated kernel density function) is

$$\hat{f}(x) = \Big\{ w_1(x) \max_{i=1,2,3} (\alpha_i)\, g_1\left(x|\mu_i, \sigma_i^2\right) + w_2(x) \max_{i=4,5,6} (\alpha_i)\, g_2\left(x|\mu_i, \sigma_i^2\right) + w_3(x) \max_{i=7,8,9} (\alpha_i)\, g_3\left(x|\mu_i, \sigma_i^2\right) \Big\} \Big/ \Big\{ w_1(x) \max_{i=1,2,3} (\alpha_i) + w_2(x) \max_{i=4,5,6} (\alpha_i) + w_3(x) \max_{i=7,8,9} (\alpha_i) \Big\}$$

Parameters of the cost function defined in 21.6 are:

$$p = 4, \quad c_1 = 7, \quad c_2 = 5, \quad c_3 = 3, \quad c_1 = 1, \quad N_1 = 2000, \quad N_2 = 500$$

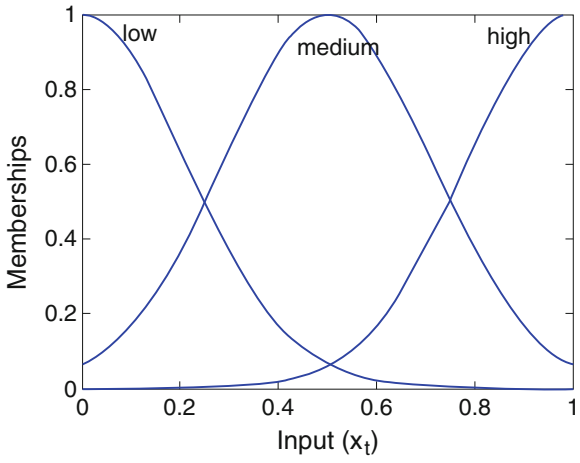


**Fig. 21.1** Membership functions for input and output of fuzzy logic.

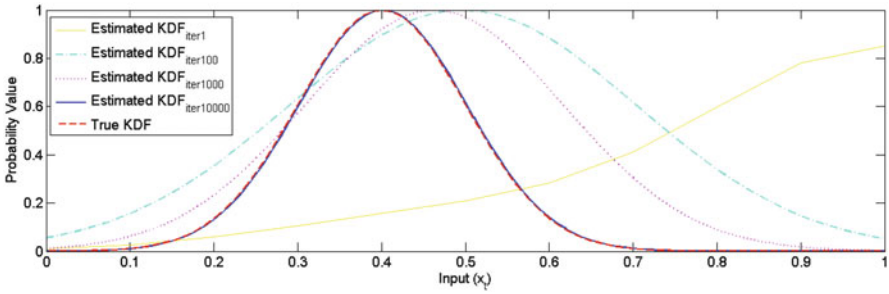


**Fig. 21.2** Estimated KDF at iterations 1, 100, 1000 and 10,000.

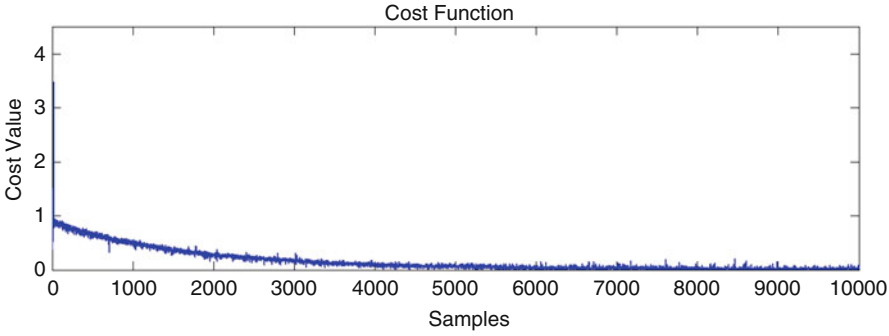


**Fig. 21.3** Value of the cost in each sample.

The step length is 0.01. The singleton fuzzifier block has been selected. The rule weights are initialized randomly. After initialization steps, as introduced above, the proposed method is tested on estimating Gaussian density function with mean *0.4* and variance *0.1*. Figure 21.2 shows the estimated KDF. It can be seen that the algorithm significantly estimates the true KDF. Figure 21.3 shows the cost function (MSE of estimated and true KDF, see Eq. 21.6). It can be seen that the algorithm can find the true model by sample 5000. The results show that the algorithm is successful in estimating the KDF. Note that just three membership functions are used for input and output of fuzzy logic. By increasing the number of membership functions the accuracy will be increased. Besides, the estimation is performing recursively by collecting just 500 last sampled data (not batch learning). By increasing the window of sampled data the accuracy maybe increased. The algorithm in this study is simulated in MATLAB R2009b environment with CPU 2.2 GHz Intel 2 core Duo processor T6600. In this environment, an iteration of running the proposed algorithm requires 0.005 sec. It is suitable for most online real-world applications.

## 21.4 Conclusions

In this study, a fuzzy method has been proposed to estimate kernel density function online. To achieve this goal, Gaussian mixture model has been generated by the fuzzy algorithm. Defuzzifier operator has been modified to make it suitable for this application. Means and variances of the model have been adapted using observed data in each new sample. Then, rules weights have been tuned by minimizing the expected $L_2$ risk function of estimated and true PDFs. In contrast to the existing approaches, our approach does not require fine-tuning parameters for a specific application, we do not assume specific forms of the target distributions and temporal constraints are not assumed on the observed data. The algorithm is simple and easy to use. Simulation results show capability of the proposed algorithm in online and accurate estimation of kernel density function.

## References

[ArCi05] Arandjelovic, O. and Cipolla, R.: Incremental learning of temporally-coherent Gaussian mixture models. British Machine Vision Conference, 759–768(2005)

[ArEtAl92] Ardizzone, E., Chella, A., Frixione, M., and Gaglio, S.: Integrating subsymbolic and symbolic processing in artificial vision. J. Intell. Syst.,**1**(4), 273–308 (1992)

[Ar04] Arsenio, A.M.: Developmental learning on a humanoid robot. IEEE International Joint Conference On Neural Networks, 3167–3172 (2004)

[CoBi01] Corduneanu, A. and Bishop, C.M.: Artificial Intelligence and Statistics. Morgan Kaufmann, 27–34(2001)

[Ep69] Epanechnikov, V.A.: Non-parametric estimation of a multivariate probability density. Theory of Probability and its Applications, **14**, 153–158 (1969), doi:10.1137/1114019

[FiJa02] Figueiredo, M.A.F. and Jain, A.K.: Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Mach. Intell. **24**(3), 381–396 (2002)

[GrSn97] Grinstead, C.M. and Snell, J.L.: Introduction to probability. American Mathematical Society, **2**, 365–380 (1997)

[KiWeKo05] Kirstein, S., Wersing, H., and Körner, E.: Rapid online learning of objects in a biologically motivated recognition architecture. 27th DAGM, 301–308 (2005)

[KrSkLe10] Kristan, M., Skocaj, D., and Leonardis, A.: Online kernel density estimation for interactive learning. Image and Vision Computing, **28**, 1106–1116 (2010)

[MaSc14] Maleca, P. and Schienle, M.: Nonparametric kernel density estimation near the boundary. Computational Statistics & Data Analysis, **72**, 57–76 (2014)

[McTi07] McGrory, C.A. and Titterington, D.M.: Variational approximations in Bayesian model selection for finite mixture distributions. Comput. Stat. Data Analysis, **51**(11), 5352–5367 (2007)

[Mc88] McLachlan, G.: Mixture Models, Marcel Dekker, (1988)

[McKr97] McLachlan, G.J. and Krishan, T.: The EM Algorithm and Extensions. Wiley, (1997)

[Or13] Orlov, Yu.N.: Optimal histogram interval for non-stationary time-series distribution function density estimation. Keldysh Institute preprints, **14**, 1–26 (2013)

[PrSa14] Prabhakar, O.P. and Sahu, N.K.: Performance Improvement of Human Voice Recognition System using Gaussian Mixture Model. International Journal of Advanced Research in Computer and Communication Engineering, **3**1, (2014)

[Sc67] Schwartz, S.C.: Estimation of Probability Density by an Orthogonal Series. The Annals of Mathematical Statistics, **38**(4), 1261–1265(1967)

[ScSz01] Scott, D.W. and Szewczyk, W.F.: From kernels to mixtures. Technometrics, **43**(3), 323–335 (2001)

[Si86] Silverman, W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall/CRC, (1986)

[SoWa05] Song, M. and Wang, H.: Highly efficient incremental estimation of Gaussian mixture models for online data stream clustering. SPIE: Intelligent Computing: Theory and Applications, 174–183 (2005)

[TaGoRe14] Tamuri, U., Goldman, N., and Reis, M.: A Penalized Likelihood Method for Estimating the Distribution of Selection Coefficients from Phylogenetic Data. Genetics, (2014), doi: 10.1534/genetics.114.162263

[WaJo95] Wand, M.P. and Jones, M.C.: Kernel Smoothing. Chapman & Hall/CRC, (1995)

[We72] Wegman, E.J.: Nonparametric Probability Density Estimation: I. A Summary of Available Methods. Technometrics, **14**(3), 533–546 (1972)

[WeTiWa14] Wellsa, J.R., Tinga, K.M., and Washio, T.: LiNearN: A new approach to nearest neighbour density estimator. Pattern Recognition, 2014 (in press), http://dx.doi.org/10.1016/j.patcog.2014.01.013

[ZhKw06] Zhang, K. and Kwok, J.T.: Simplifying mixture models through function approximation. Neural Inf. Proc. Systems, (2006)

[ZiVa04] Zivkovic, Z. and van der Heijden, F.: Recursive unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Mach. Intell., **26**(5), 651–656 (2004)

# Chapter 22
# Two Reasons Why Pollution Dispersion Modeling Needs Sesquilinear Forms

**D.L. Gisch, B.E.J. Bodmann, and M.T.B. Vilhena**

## 22.1 Introduction

Atmospheric dispersion modeling is nowadays a valuable tool that permits to simulate how air pollutants affect the ambient atmosphere. Models are not only used to estimate the downwind concentration of pollutant substances but also allow to reproduce the full three-dimensional pollutant distributions over time, while measurements are typically acquired by a small set of detection locations, only [PeEtAl13, BuEtAl12b, TiEtAl11]. Nowadays, governmental agencies for ambient air quality protection and management employ such models in order to determine whether existing or planned emission sources are in compliance with ambient air quality standards.

The atmospheric boundary layer that extends from the earth's surface to a few kilometers in height is where predominantly emission, transport and dispersion of airborne pollutants take place [PeBoVi11]. The phenomena that characterize the boundary layer dictate the physical content of a model that shall either be incorporated in form of physical laws or if complexity cannot be disentangled into simpler components, parametrizations hide the unknown reality [PuEtAl13, CoAcDe11]. Meteorological conditions such as the wind field and the vertical thermodynamics profile on one side and atmospheric turbulence in diverse stability regimes on the other side [GoEtAl10]. The emission source(s) and locations which may be of point, line, surface or volume type with their characteristic time signatures of discharge

D.L. Gisch (✉) • B.E.J. Bodmann • M.T.B. Vilhena
Federal University of Rio Grande do Sul, Av. Osvaldo Aranha 99/4,
Porto Alegre 90046-900, RS, Brazil
e-mail: debora.gisch@gmail.com; bardo.bodmann@ufrgs.br; vilhena@math.ufrgs.br



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_22

rates, such as instantaneous or steady state releases among many others. Last not least the natural or urban topography shall enter the model so that simulations with a reasonable precision may be performed.

In the sequel, we focus only on one aspect of the rather complex dispersion phenomenon, namely the transport and mixing by the phenomenon of turbulence [PeEtAl13, BuEtAl12b, TiEtAl11]. It is noteworthy that pollutant dispersion for instance smoke from a chimney is being observed along centuries, however insight in what gives life to those filigrane patterns is still some way ahead. The discussion that follows addresses the question as to what is the most efficient way to model the aforementioned patterns without extending the parameter space of the model to an exorbitant dimension.

## 22.2 Modeling

In the literature, one finds two classes of models used to simulate pollutant dispersion in the planetary boundary layer, either deterministic models [PeEtAl13, BuEtAl12b, TiEtAl11] or stochastic models [BoMeVi13, BoViMe10]. While the phenomenon is manifest stochastic, deterministic models provide a solution which describes average pollutant concentrations, so that for each statistical moment, such as variance skewness and bias among others an additional model equation is of need. Only rare cases where the distribution is sufficiently narrow are characterized by the mean values only. In stochastic approaches the pollutant distribution is obtained from a number of realizations that follow a probability density function. These probability density functions are unknown and can only be determined by a validation procedure, after the solution has been determined. In practice this means, that one selects the 'best' solution among the trials that were conducted. In the further we present a novel approach that shall maintain the simplicity of the deterministic models but present some realistic features concerning the stochastic character of the dispersion phenomenon. To this end, we first analyze the traditional way to derive a deterministic model and indicate the minimal modifications that can be introduced that turn the model a stochastic one.

### *22.2.1 A Traditional Deterministic Model*

A convenient starting point is the continuity equation with the time derivative of the concentration, or pollutant density and a current density, i.e. the pollutant flux. The variables are then decomposed into mean values and fluctuation quantities which are eliminated upon applying the Fickian closure. This procedure reduces the originally stochastic model into a deterministic one known as the advection–diffusion equation. Since it is the closure that eliminates the stochastic character, a modification of the closure seems the adequate way to recover at least partially the stochastic nature of the phenomenon. The traditional model is

$$\frac{\partial C}{\partial t} + \mathbf{u}\nabla C = \nabla^{\dagger}\mathbf{K}\nabla C + S\,.$$

Here, $C$ is the mean pollutant concentration, $\mathbf{u}$ represents the wind velocity field, $\mathbf{K} = \mathrm{diag}(K_x, K_y, K_z)$ is the turbulent diffusion coefficient matrix, and $S$ a source term. The domain considered is bounded in the crosswind and vertical direction and semi-open in the wind velocity direction. At the boundary of the domain $y \in [-L_y, L_y] \cup z \in [0, z_{BL}]\}$ with $z_{BL}$ the boundary layer height, the pollutant flux vanishes and as initial condition we assume zero pollutant concentration, except for the point source location.

$$\mathbb{D} = \{\mathbf{x} = (x,y,z) \mid x \in [0,\infty),\ y \in [-L_y, L_y],\ z \in [0, z_{BL}]\}$$
$$\nabla C = 0 \quad \forall\, \mathbf{x} \in \delta\mathbb{D}$$
$$C(x,y,z,0) = 0 \qquad \text{for} \quad t = 0 \quad \text{and} \quad \mathbf{x} \in \mathbb{D}\backslash(0,0,H_s)$$

In this discussion, we consider a continuous source with constant emission rate located at the coordinates $\mathbf{x} = (0,0,H_s)$

$$\mathbf{u}C(0,y,z,t) = \dot{Q}\delta(y)\delta(z - H_s)\,.$$

Due to the choice of the domain, we make use of spectral theory in $\mathbb{D}\backslash\{x \mid x \in [0,\infty)\}$ [BoEtAl12a] together with Laplace transform in $\{x \mid x \in [0,\infty)\} \cup \{t \mid t \in [0,\infty)\}$ to obtain a solution in analytical form. Furthermore, instead of solving the problem for the continuous source, we superimpose solutions from the problem without source term for initial conditions by shifting the initial time to larger values. One could interpret such a procedure as continuous initial condition.

$$C(x,y,z,t) = \textstyle\sum_{n=1}^{\infty}\sum_{l=1}^{\infty} A_{nl}\, c(x,t) \cos\left(\frac{n\pi}{L_y}y\right) \cos\left(\frac{l\pi}{L_z}z\right)$$
$$c(x,t) = \textstyle\int_0^t \frac{1}{2} e^{\frac{u}{2K_x}x}\, e^{\left(\alpha - \frac{u^2}{4K_x}\right)\tau}\, e^{-\left(\frac{x^2}{4K_x\tau}\right)} \left(\frac{x}{\sqrt{\pi K_x \tau^3}} - \frac{5\sqrt{K_x}u}{\sqrt{\pi\tau}}\right) d\tau$$
$$\alpha = -\left(\left(K_z \sin\left(\frac{z\pi}{L_z}\right)\right)\left(\frac{l\pi}{L_z}\right)^2 + K_y\left(\frac{n\pi}{L_y}\right)^2\right)$$

The expansion coefficients are determined by

$$(A_{nl}A_{n'l'}) = \frac{\dot{Q}}{u}\phi_{nln'l'}(y_0,H_s)\left(\int_0^{L_z}\int_{-L_y}^{L_y}\phi^2_{nln'l'}(y,z)\,dy\,dz\right)^{-1}$$
$$= \frac{4\dot{Q}\phi_{nln'l'}(y_0,H_s)}{u(L_yL_z)^2}$$
$$\phi_{nln'l'}(y,z) = \cos\left(\frac{n\pi}{L_y}y\right)\cos\left(\frac{l\pi}{L_z}z\right)\cos\left(\frac{n'\pi}{L_y}y\right)\cos\left(\frac{l'\pi}{L_z}z\right)\,.$$

Although we will change the closure, this solution will still be useful even for our stochastic approach.

### 22.2.2 A New Concept

As the following discussion will show using some axiomatic arguments and reinterpreting the solution will lead us a pollutant density from a deterministic-stochastic approach. Stochastic evolution with turbulent character arises upon interpreting the solution in terms of a probability amplitude. In order to render the concentration compatible with necessary properties for distributions we construct our solution such that the following properties are true. Due to the fact that the pollutant density ($C$) shall be interpreted in terms of probabilities it shall result 'naturally' as a mapping from space-time to a semi-positive definite space $\mathbb{R}^{3\oplus 1} \rightarrow \mathbb{R}^{+}$. The distribution shall be independent on any specific choice of reference frame, i.e. under coordinate transformation the quantity shall have the property of a scalar density $C_A(\mathbf{x}_A, t_A) = J\, C_B(\mathbf{x}_B, t_B)$. One possibility that complies with the aforementioned properties is representing the concentration by a Hermitian form associated quadratic form $\mathbb{C} \times \mathbb{C} \rightarrow \mathbb{R}^{+}$, where semi-positiveness is guaranteed for the Euclidean case [La05, MiHu73, GrWe77].

From the physical point of view, we consider turbulent evolution following the idea of Kolmogorov's eddy spectrum [Su32, Mo83]. Eddies are coherent structures [BoEtAl13] that at least show partially constant space-time correlations and may be implemented by the presence of a phase. The fact that complex functions naturally embody a phase indicates them as a convenient descriptor. Since sesquilinear forms [La05, MiHu73, GrWe77] unite density and evolution aspects they seem an adequate way to describe distributions with structure, where structure means a filigrane appearance as, for instance, the patterns that appear with smoke.

One of the fundamental differences to the traditional model, that maps space-time to a concentration function $C : \mathbb{R}^{3\oplus 1} \rightarrow \mathbb{R}$ using a real diffusion coefficient matrix $\mathbf{K} \in M(3, \mathbb{R})$, is that the quantity that is determined by the complex advection-diffusion equation is not the observable, i.e. the pollutant concentration. The solution of the equation is a complex probability amplitude $\mathscr{C} : \mathbb{R}^{3\oplus 1} \rightarrow \mathbb{C}$ that upon taking the Hermitian form associated quadratic form results in the concentration $C = \mathscr{C}^{\dagger}\mathscr{C}$ which is naturally semi-positive. The complex solution is obtained from the modified Fickian closure, that makes use of a complex diffusion coefficient matrix $\mathbf{K} \in M(3, \mathbb{C})$. For simplicity and to show the effect of a complex contribution, we maintain the coefficients for the wind and cross wind direction real $K_x, K_y \in \mathbb{R}$ and add only to the vertical component an imaginary part $K_z(z) = K_{zR} \sin\left(\frac{z\pi}{L_z}\right) + \imath K_{zI}$. Note that the turbulent character is strongest in the vertical direction due to predominantly vertical heat-flux. The model inherent appearance of structure may be understood from the sesquilinearity conditions, that because of the presence of a phase together with the crossed terms generate structure.

$$C(x,y,z,t) = \left(\sum_i \mathscr{C}_i^\dagger\right)\left(\sum_j \mathscr{C}_j\right)$$
$$= \underbrace{\sum_i \mathscr{C}_i^\dagger \mathscr{C}_i}_{\text{traditional}} + \underbrace{\sum_i \sum_j (1-\delta_{ij})\mathscr{C}_i^\dagger \mathscr{C}_j}_{\text{structure}}$$

Here $\delta_{ij}$ is the Kronecker delta.

## 22.3 Results

For a purely real eddy diffusion coefficient, one obtains a Gaussian solution with its maximum at the center position ($y = 0$), where the source is located. Upon extending the model with a complex valued eddy diffusion coefficient one obtains an intuitively unexpected result, a distribution with two global maxima to the left and right of the center position (see figure 22.1). Due to the probabilistic nature of the obtained distributions this property is a first hint to understand meandering in terms of a deterministic-stochastic model. The reported effect increases with increasing values for the imaginary part of the eddy diffusion coefficient as shown in figures 22.2 and 22.3. Moreover, the larger the imaginary part the more apparent is the wavy character of the distribution. While in figure 22.1 there are no local maxima next the center, the latter exist for $K_{zI} = 0.5$ and $K_{zI} = 0.8$ (shown in figures 22.2 and 22.3), respectively. In order to analyze the spatial distribution of the concentration, i.e. the spatial structure, we compare the concentration depending on the wind direction $x$ and the height $z$ for two cases of wind speeds. Figure 22.4 shows a structure that emerges for low wind speeds $\sim 1\frac{m}{s}$, where along $10^2 m$ of length there are several local maxima of crest like structures, that show a

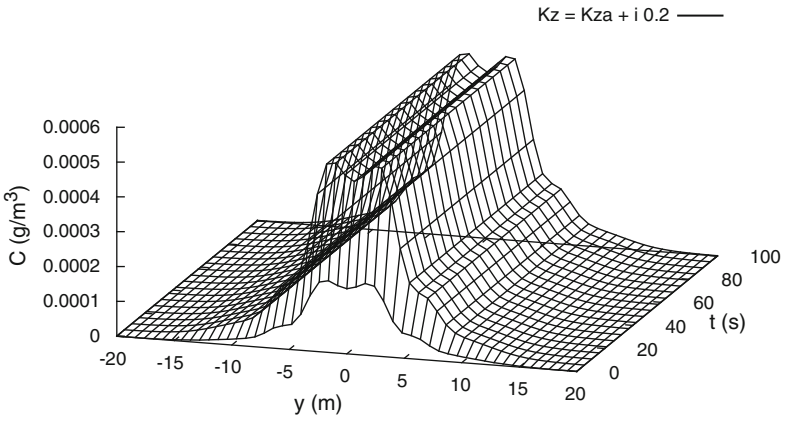


**Fig. 22.1** Concentration evolution in cross wind direction for $K_{zI} = 0.2$.

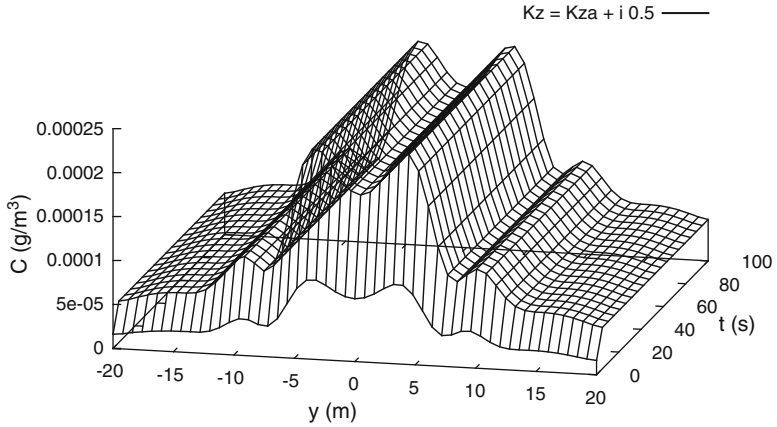


**Fig. 22.2** Concentration evolution in cross wind direction for $K_{zI} = 0.5$.

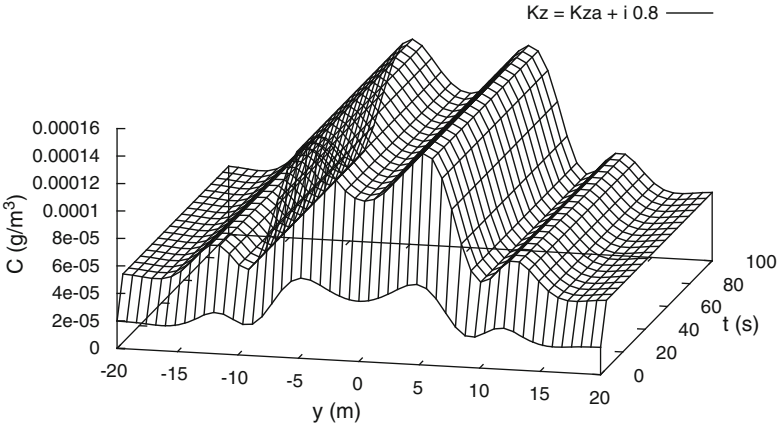


**Fig. 22.3** Concentration evolution in cross wind direction for $K_{zI} = 0.8$.

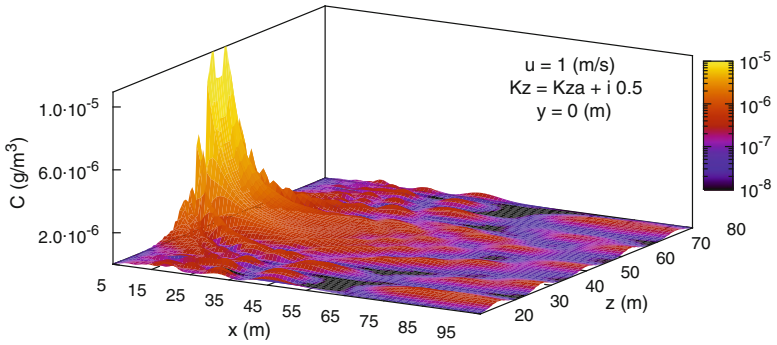


**Fig. 22.4** Concentration in wind and vertical direction for low wind speed.

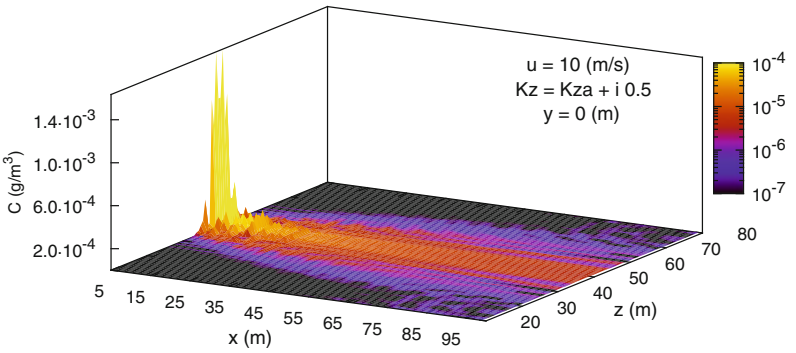


**Fig. 22.5** Concentration in wind and vertical direction for high wind speed.

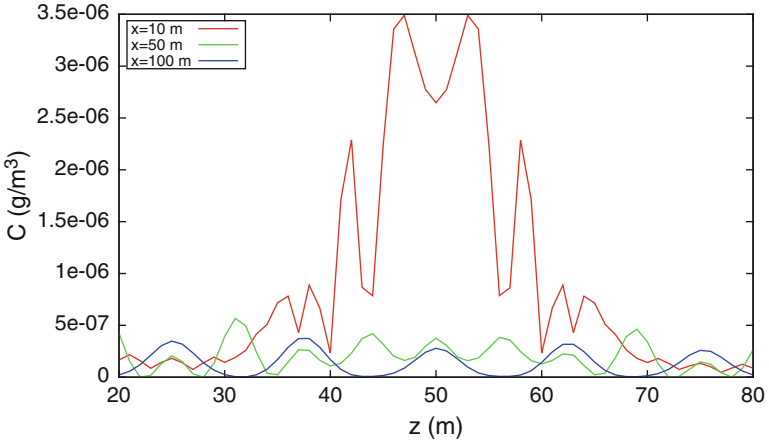


**Fig. 22.6** Concentration in vertical direction for low wind speed.

divergence with increasing distance from the source. In the case for a wind speed of $10\frac{m}{s}$ the plume is stretched with little divergence as the distance to the source increases (see figure 22.5). Furthermore, the already commented meandering related characteristics of the distribution is less pronounced for the large wind speed, an effect also confirmed by observation, i.e. meandering only occurs significantly for low wind speeds. Projections in the vertical direction for horizontal distances $x = 10m$, $50m$ and $100m$ are shown for low and high wind speed in figures 22.6 and 22.7, respectively. For shorter distances the pronounced maxima are visible, whereas for larger distances, dissipative effects leave a remaining weakly wavy distribution. This at least qualitatively corroborates with observation, where turbulence decays and increasingly larger wavelength contributions prevail, an effect of combined dispersion with dissipation.

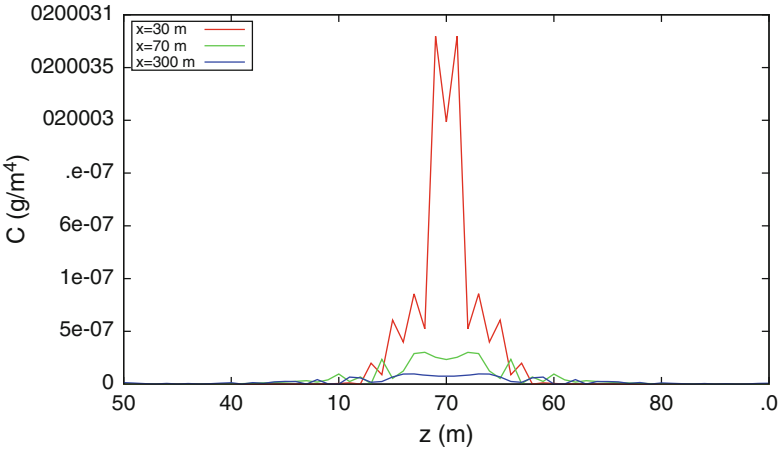


**Fig. 22.7** Concentration in vertical direction for high wind speed.

## 22.4 Conclusion

In this chapter, we presented a novel approach to simulating transport and dispersion of pollutant by the presence of turbulence. In contrast to the traditional approaches either by deterministic or stochastic models, the present model contains a phase responsible for structure patterns in the concentration distribution. Similar patterns may be observed in scenarios where smoke is released or more generally where the pollutant accompanies the wind field with its velocity distribution.

Both methods traditionally used in pollutant dispersion modeling either of deterministic or stochastic character have their shortcomings. In the derivation of deterministic models the closure that should characterize the turbulent character is of deterministic origin. Consequently, essential properties due to turbulence are lost and the dynamics is reduced to an advection-diffusion process. Stochastic dispersion models such as the Langevin equation maintain stochastic character by a probability-driven source term. However, the probability density functions that simulate the turbulent character are not known *a priori* and thus are in general imposed *ad hoc*. One of the consequences is that if the complex structure of the plume is not put in by hand, no such structure emerges from the model.

This is different in the proposed model, where structure in the space time dependent concentration field emerges from the presence of a phase in the model. The phase was included using a complex closure and writing the concentration $C : \mathbb{R}^{3\oplus 1} \to \mathbb{R}^{+}$ as a sesquilinear form associated quadratic form $C = \mathscr{C}^{\dagger}\mathscr{C}$. It is noteworthy that the equation for the amplitude is linear, whereas the observable, i.e. the pollutant concentration is nonlinear, which gives rise to the structure by the presence of interference terms. The model has some remarkable inherent features. At low wind speed the maxima in crosswind and vertical direction are not centered at the line that is oriented along the wind direction and passes through the center of the

sources. This effect disappears with increasing wind speed. Observation agrees with that phenomenon, if the distribution is interpreted as a probability density function then realizations of pollutant flow would either deviate to the left or right and cross over to the adjacent side in the cross wind as well as the vertical plane. This effect is known as meander and there is no model known so far that predicts such a behavior in a model inherent fashion.

**Acknowledgements** The authors wish to thank CAPES, LGSA, PD-ANEEL, and CPPT for financial support.

## References

[BoEtAl12a] Bodmann, B.E.J., Buske, D., Vilhena, M.T., and Tirabassi, T.: Analytical Model for Air Pollution in the Atmospheric Boundary Layer. In: Mukesh Khare. (Org.) Air Pollution – Monitoring, Modelling and Health. InTech, 39–58 (2012).

[BoMeVi13] Bodmann, B.E.J., Mello, K.B., and Vilhena, M.T.: Turbulent Wind Profiles and Tracer Dispersion for Eolic Park Site Evaluation. American Journal of Environmental Engineering, **3**, 147–169 (2013).

[BoViMe10] Bodmann, B.E.J., Vilhena, M.T.M.B., and Mello, K.B.: Stochastic wind profiles determination for radioactive substances released from nuclear power plants. Nuclear Power, 267–292 (2010).

[BoEtAl13] Bodmann, B.E.J., Zabadal, J.R., Schuck, A., Vilhena, M.T., and Quadros, R.: On Coherent Structures from a Diffusion-Type Model. Integral Methods in Science and Engineering: Progress in Numerical and Analytic Techniques. Birkhäuser/Springer, 65–74 (2013).

[BuEtAl12b] Buske, D., Vilhena, M.T., Bodmann, B.E.J., Tirabassi, T.: Air Pollution Steady-State Advection-Diffusion Equation: The General Three-Dimensional Solution. Journal of Environmental Protection, **3**, 1124–1134 (2012).

[CoAcDe11] Costa, F.D., Acevedo, O., Degrazia, G.A.: A Simplified Model for Intermittent Turbulence in the Nocturnal Boundary Layer. Journal of the Atmospheric Sciences, **68**, 1714–1729 (2011).

[GoEtAl10] Goulart, A., Bodmann, B.E.J., Moreira, D.M., Vilhena, M.T.M.B.: On the Time Evolution of the Turbulent Kinetic Energy Spectrum for Decaying Turbulence in the Convective Boundary Layer. Boundary-Layer Meteorology, **1**, 1–15 (2010).

[GrWe77] Gruenberg, K.W., Weir, A.J.: Linear Geometry (Chapter 5.8 Sesquilinear Forms). Springer, 120–124 (1977).

[La05] Lam, T.-Y.: Introduction to Quadratic Forms over Fields. Graduate Studies in Mathematics 67. American Mathematical Society (2005).

[MiHu73] Milnor, J., Husemoller, D.: Symmetric Bilinear Forms. Ergebnisse der Mathematik und ihrer Grenzgebiete **73**. Springer-Verlag (1973).

[Mo83] Moffatt, H.K.: Transport effects associated with turbulence with particular attention to the influence of helicity. Rep. Prog. Phys., **46**, 621–664 (1983).

[PeBoVi11] Pellegrini, C., Bodmann, B.E.J., Vilhena, M.T.M.B.: A Theoretical Study of the Stratified Atmospheric Boundary Layer Through Perturbation Techniques. In: C. Constanda, P. Harris. (eds.). Integral Methods in Science and Engineering. Springer, 273–285 (2011).

[PeEtAl13] Pellegrini, C., Buske, D., Bodmann, B.E.J., Vilhena, M.T.: A First Order Pertubative Analysis of the Advection-Diffusion Equation for Pollutant Dispersion in the Atmospheric Boundary Layer. American Journal of Environmental Engineering, **3**, 48–55 (2013).

[PuEtAl13] Puhales, F.S., Rizza, U., Degrazia, G.A., Acevedo, O.: A simple parameterization for the turbulent kinetic energy transport terms in the convective boundary layer derived from large eddy simulation. Physica A, **392**, 583–595 (2013).

[Su32] Sutton, O.G.: A Theory of Eddy Diffusion in the Atmosphere. Proc. R. Soc. Lond. A , **135**(826), 143–165 (1932).

[TiEtAl11] Tirabassi, T., Tiesi, A., Vilhena, M.T.M.B, Bodmann, B.E.J., Buske, D.: An analytical simple formula for the ground level concentration from a point source. Atmosphere, **2**, 21–35 (2011).

# Chapter 23
# Correcting Terms for Perforated Media by Thin Tubes with Nonlinear Flux and Large Adsorption Parameters

**D. Gómez, M. Lobo, M.E. Pérez, T.A. Shaposhnikova, and M.N. Zubova**

## 23.1 Introduction and Formulation of the Problem

In this chapter, we obtain correcting terms in homogenization problems for the Laplace operator arising e.g., in modeling diffusion of substances in perforated media with large adsorption parameters on the boundary of the perforations (see, e.g., [Go95, CoDi04], for more specific models). These correctors allow us to improve the results on weak convergence obtained in [GoLo14], providing bounds for convergence rates of solutions in $H^1$, for all the possible relations between the parameters arising in the problem, namely, periodicity $\varepsilon$, diameter of tubes $a_\varepsilon$ and adsorption parameters $\beta(\varepsilon)$, $a_\varepsilon$ being $a_\varepsilon \ll \varepsilon$, and $\varepsilon \to 0$. At the same time, we provide proofs for some of the results stated in [GoLo14] for the relations (23.12). The results in this paper complement and improve [GoLo13b] and [GoLo14]. Here, for the sake of simplicity in computations, we assume that the cylinders have a circular transverse section, but we emphasize that the statements of all the theorems can be formulated (with the suitable modifications) for the more general geometrical configuration in [GoLo14] involving both isoperimetric cylinders and a certain non-periodic distribution of the tubes. See also Remark 2 for other extensions of the results in this paper.

We assume that the medium fills a domain $\Omega_\varepsilon$ of $\mathbb{R}^3$ which is obtained by removing thin cylinders $(0,l) \times G_\varepsilon$, *the thin tubes*, of diameter $O(a_\varepsilon)$ from a fixed domain $\Omega$. These cylinders of length $O(1)$ are periodically placed in volume over $\Omega$, parallel to the $x_1$-axis; the basis of the cylinders are circles on the plane $\{x_1 = 0\}$

D. Gómez • M. Lobo • M.E. Pérez (✉)
University of Cantabria, Av. Los Castros s.n., 39005 Santander, Spain
e-mail: gomezdel@unican.es; miguel.lobo@unican.es; meperez@unican.es

T.A. Shaposhnikova • M.N. Zubova
Moscow State University, Moscow, Russia
e-mail: shaposh.tan@mail.ru; zubovnv@mail.ru



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_23

which are periodically distributed at a distance $O(\varepsilon)$ between them. Here, the period $\varepsilon$ is a small parameter which we shall make to go to zero; the radius of the circles $a_\varepsilon$ is such that $a_\varepsilon \ll \varepsilon$. A Dirichlet condition is imposed on $\partial\Omega_\varepsilon \cap \partial\Omega$, while a nonlinear flux is imposed on the rest of the boundary, namely, on the lateral boundary of the thin tubes. This condition involves both a large *adsorption parameter* $\beta(\varepsilon)$ and a nonlinear function $\sigma \in C^1(\overline{\Omega} \times \mathbb{R})$, $\sigma$ being monotonic with respect to the second argument (cf. (23.1) and Remark 1).

An extensive study of these kinds of problems for the case of perforated domains with perforations which are balls has been considered, e.g., in [Go95] and [ZuSh11]. See also [CoDi04, Ti09] and [CaDo12] for the case where the size of the perforations is of the same order of magnitude as the period, and [LoPe11] and [GoPe12] for the case where the perforations are placed along manifolds. Let us mention [ZuSh11, GoPe12] and [GoLo14] for a large list of references on the subject.

Let $\omega$ be a bounded domain of $\mathbb{R}^2$, in the plane $\{x_1 = 0\}$, with a smooth boundary $\partial\omega$. We set $\Omega = (0,l) \times \omega$. Let $\varepsilon$ and $a_\varepsilon$ be small parameters, $0 < \varepsilon \ll 1$ and $0 < a_\varepsilon < \varepsilon$. We set

$$\widetilde{\omega}_\varepsilon = \{x \in \omega : \rho(x, \partial\omega) > 2\varepsilon\},\ \widetilde{\Omega}_\varepsilon = (0,l) \times \widetilde{\omega}_\varepsilon,$$

where $\rho$ denotes the distance to the boundary.

Let $\mathscr{Z}$ be the set of vectors $j = (0, j_2, j_3)$ with integer coordinates. We define

$$G_\varepsilon = \bigcup_{j \in \Upsilon_\varepsilon} (0,l) \times (G_\varepsilon^0 + \varepsilon j) = \bigcup_{j \in \Upsilon_\varepsilon} (0,l) \times G_\varepsilon^j,$$

where $G_\varepsilon^0 = \{\hat{x} \in \mathbb{R}^2_{x_2 x_3}, x_2^2 + x_3^2 < a_\varepsilon^2\}$ and $\Upsilon_\varepsilon = \{j \in \mathscr{Z} : (0,l) \times G_\varepsilon^j \bigcap \widetilde{\Omega}_\varepsilon \neq \emptyset\}$. Note that $|\Upsilon_\varepsilon| \cong d\varepsilon^{-2}$, with some $d > 0$.

We set

$$\Omega_\varepsilon = \Omega \setminus \overline{G_\varepsilon},\ \partial\Omega_\varepsilon = S_\varepsilon \bigcup \Gamma_\varepsilon,$$

where $S_\varepsilon = \bigcup_{j \in \Upsilon_\varepsilon} (0,l) \times \partial G_\varepsilon^j = \bigcup_{j \in \Upsilon_\varepsilon} S_\varepsilon^j$ is the lateral area of cylinders, $S_\varepsilon^j$ the lateral area of $(0,l) \times G_\varepsilon^j$, $\Gamma_\varepsilon = ([0,l] \times \partial\omega) \bigcup \omega_\varepsilon^0 \bigcup \omega_\varepsilon^l$, $\omega_\varepsilon^0 = \omega \setminus \overline{G_\varepsilon}$, $\omega_\varepsilon^l = (\partial\Omega \bigcap \{x_1 = l\}) \setminus G_\varepsilon$ (see Figure 23.1 for the geometrical configuration of $\Omega_\varepsilon$).

Let $\sigma(x,u)$ be a continuously differentiable function of variables $(x,u) \in \overline{\Omega} \times \mathbb{R}$ satisfying $\sigma(x,0) = 0$ and there exist two constants $k_1 > 0$ and $k_2 > 0$ such that

$$k_1 \leq \frac{\partial\sigma}{\partial u}(x,u) \leq k_2, \quad \forall x \in \overline{\Omega}, \quad u \in \mathbb{R}.$$

Note that the above conditions imply that

$$k_1|uv| \leq |\sigma(x,u)v| \leq k_2|uv| \text{ and } (\sigma(x,u) - \sigma(x,v))(u-v) \geq k_1(u-v)^2, \qquad (23.1)$$

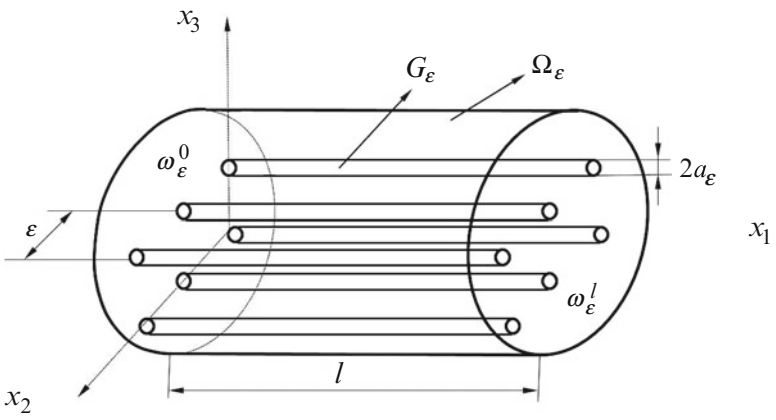


**Fig. 23.1** The domain $\Omega_\varepsilon$.

for any $x \in \overline{\Omega}$ and $u, v \in \mathbb{R}$.

For $f \in L^2(\Omega)$, we consider the boundary value problem

$$\begin{cases} -\Delta u_\varepsilon = f & \text{in } \Omega_\varepsilon, \\ \dfrac{\partial u_\varepsilon}{\partial \nu} + \beta(\varepsilon)\sigma(x, u_\varepsilon) = 0 & \text{for } x \in S_\varepsilon, \\ u_\varepsilon = 0 & \text{on } \Gamma_\varepsilon, \end{cases} \tag{23.2}$$

where $\beta(\varepsilon)$ is a certain strictly positive order function and $\nu$ denotes the unit outward normal vector to $\partial\Omega_\varepsilon$ on $S_\varepsilon$.

The variational formulation of problem (23.2) is: find $u_\varepsilon \in H^1(\Omega_\varepsilon, \Gamma_\varepsilon)$ satisfying

$$\int_{\Omega_\varepsilon} \nabla u_\varepsilon \nabla \psi dx + \beta(\varepsilon) \int_{S_\varepsilon} \sigma(x, u_\varepsilon)\psi ds = \int_{\Omega_\varepsilon} f\psi dx, \quad \forall \psi \in H^1(\Omega_\varepsilon, \Gamma_\varepsilon). \tag{23.3}$$

As usual, we denote by $H^1(\Omega_\varepsilon, \Gamma_\varepsilon)$ the completion with respect to norm $H^1(\Omega)$ of the set of infinitely differentiable function in $\overline{\Omega}_\varepsilon$ vanishing on $\Gamma_\varepsilon$.

On account of the monotonicity of the function $\sigma(x, u)$ with respect to $u$, the following assertion can be proved (see [GoLo13b] and [GoLo14] for the proof):

**Proposition 1.** *Let $\varepsilon > 0$ and $f \in L^2(\Omega)$. For fixed $\varepsilon$, problem (23.3) has a unique solution $u_\varepsilon \in H^1(\Omega_\varepsilon, \Gamma_\varepsilon)$. In addition, for $u_\varepsilon$ the solution of (23.3), there exists $\widetilde{u}_\varepsilon$ an extension of $u_\varepsilon$ to $\Omega$, $\widetilde{u}_\varepsilon \in H_0^1(\Omega)$ with the following properties*

$$\|\widetilde{u}_\varepsilon\|_{H^1(\Omega)} \le K\|u_\varepsilon\|_{H^1(\Omega_\varepsilon)}, \quad \|\nabla\widetilde{u}_\varepsilon\|_{L^2(\Omega)} \le K\|\nabla u_\varepsilon\|_{L^2(\Omega_\varepsilon)},$$

*and*

$$\|\widetilde{u}_\varepsilon\|^2_{H^1(\Omega)} + \beta(\varepsilon)\|u_\varepsilon\|^2_{L^2(S_\varepsilon)} \leq K. \tag{23.4}$$

*In all the estimates above, $K > 0$ denotes a constant independent of $\varepsilon$.*

Thus, from (23.4), we derive that for each sequence $\varepsilon$ we can extract a subsequence (still denoted by $\varepsilon$) such that

$$\widetilde{u}_\varepsilon \rightharpoonup u \text{ in } H^1_0(\Omega) - \text{ weak} \quad \text{and} \quad \widetilde{u}_\varepsilon \to u \text{ in } L^2(\Omega) \text{ as } \varepsilon \to 0, \tag{23.5}$$

for a certain function $u$ which, once identified, provides the convergence (23.5) for the whole sequence of $\varepsilon$.

Under the assumption of *thin cylinders* in the period scale $\varepsilon$ (namely, $a_\varepsilon \ll \varepsilon$), in [GoLo14], we obtain the different averaged problems for $u$, and show the convergence of the solution. As a matter of fact, considering the problem (23.2) and the nine possibilities for the couple of limits

$$\lim_{\varepsilon\to 0} \varepsilon^2 \ln(a_\varepsilon) = -\alpha^2 \quad \text{and} \quad \lim_{\varepsilon\to 0} \beta(\varepsilon) a_\varepsilon \varepsilon^{-2} = C^2, \tag{23.6}$$

where $\alpha^2$ and $C^2$ can be well-determined positive constants, 0 or $\infty$, we obtain five possible different limit behaviors of $u_\varepsilon$. In order to be self-contained, we gather the results in the following theorem (cf. [GoLo14] for details):

**Theorem 1.** *Let $u_\varepsilon$ be the solution of (23.3). Then, the limit function $u$ of the extension of $u_\varepsilon$, which is defined by (23.5), coincides with*

*i). $u = 0$ when $\beta(\varepsilon) a_\varepsilon \varepsilon^{-2} \to \infty$ and $\varepsilon^2 \ln(a_\varepsilon) \to 0$. In addition,*

$$\|\widetilde{u}_\varepsilon\|_{H^1(\Omega)} \leq K \left(\varepsilon^2 a_\varepsilon^{-1} \beta(\varepsilon)^{-1} + \varepsilon^2 |\ln(\varepsilon/2a_\varepsilon)|\right)^{1/4}.$$

*ii). the weak solution of the Dirichlet problem*

$$\begin{cases} -\Delta u = f \text{ in } \Omega, \\ u = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{23.7}$$

*when $\beta(\varepsilon) a_\varepsilon \varepsilon^{-2} \to 0$ or $\varepsilon^2 \ln(a_\varepsilon) \to -\infty$.*

*iii). the weak solution of the problem*

$$\begin{cases} -\Delta u + \dfrac{2\pi}{\alpha^2} u = f \text{ in } \Omega, \\ u = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{23.8}$$

*when $\beta(\varepsilon) a_\varepsilon \varepsilon^{-2} \to \infty$ and $\varepsilon^2 \ln(a_\varepsilon) \to -\alpha^2 < 0$.*

*iv). the weak solution of the problem*

$$\begin{cases} -\Delta u + 2\pi C^2 \sigma(x,u) = f \text{ in } \Omega, \\ u = 0 \qquad\qquad\qquad\quad \text{on } \partial\Omega, \end{cases} \tag{23.9}$$

*when* $\beta(\varepsilon)a_\varepsilon\varepsilon^{-2} \to C^2 > 0$ *and* $\varepsilon^2 \ln(a_\varepsilon) \to 0$.

*v). the weak solution of the problem*

$$\begin{cases} -\Delta u + \dfrac{2\pi}{\alpha^2} H(x,u) = f \text{ in } \Omega, \\ u = 0 \qquad\qquad\qquad\quad \text{on } \partial\Omega, \end{cases} \tag{23.10}$$

*where the function* $H = H(x,\phi)$ *is the unique solution of the functional equation*

$$H = \alpha^2 C^2 \sigma(x, \phi - H), \tag{23.11}$$

*when* $\beta(\varepsilon)a_\varepsilon\varepsilon^{-2} \to C^2 > 0$ *and* $\varepsilon^2 \ln(a_\varepsilon) \to -\alpha^2 < 0$.

In this paper, we construct correctors to improve convergence, and provide estimates for convergence rates of solutions. Section 23.2 contains some preliminary results which are necessary for proofs. Sections 23.3 and 23.4 contain the correctors results for the different possible limits arising in (23.6).

As already mentioned, the closest works in the literature to the problem here considered are [GoLo13b] and [GoLo14]. To be more specific, [GoLo13b] addresses the case of a periodical distribution of the cylinders $(0,l) \times G_\varepsilon^j$ with a circular transverse section, while only the parameters

$$a_\varepsilon = \varepsilon \exp\left(-\frac{\alpha^2}{\varepsilon^2}\right) \quad \text{and} \quad \beta(\varepsilon) = \varepsilon \exp\left(\frac{\alpha^2}{\varepsilon^2}\right) \tag{23.12}$$

have been considered leaving, as open problems, other possible choices for these functions and possible shapes for perforations. [GoLo14] considers the problem for a domain perforated by isoperimetric tiny tubes of arbitrary shape and any choice of $a_\varepsilon$ and $\beta(\varepsilon)$, $a_\varepsilon \ll \varepsilon$. Both papers show convergence of solutions in the weak topology of $H^1$ (see Theorem 1). Also, in [GoLo13b], it is announced without proof a corrector result for the parameters (23.12).

In order to illustrate the limit behaviors, for all the possible relations between parameters, we choose a large set of different orders of magnitude for these parameters, but ranging in the functions $a_\varepsilon = \varepsilon^k e^{-\alpha^2/\varepsilon^\gamma}$ and $\beta(\varepsilon) = \varepsilon^{2-m} e^{\alpha^2/\varepsilon^\gamma}$, for any constants $\gamma > 0$, $k \geq 0$ and $m \geq 0$. However Figure 23.2 shows a general situation for the different relations (23.21), (23.25), (23.33), (23.40), and (23.44).

We note that $\gamma = 2$ provides the *critical size for perforated domains by tubes*. Namely, $\gamma > 2$ implies that the diameter of the cylinders is very small compared with the distance between them. In this case, Figure 23.2 shows that the limit problem is the Dirichlet problem in $\Omega$. That is, asymptotically the solution of (23.2) ignores

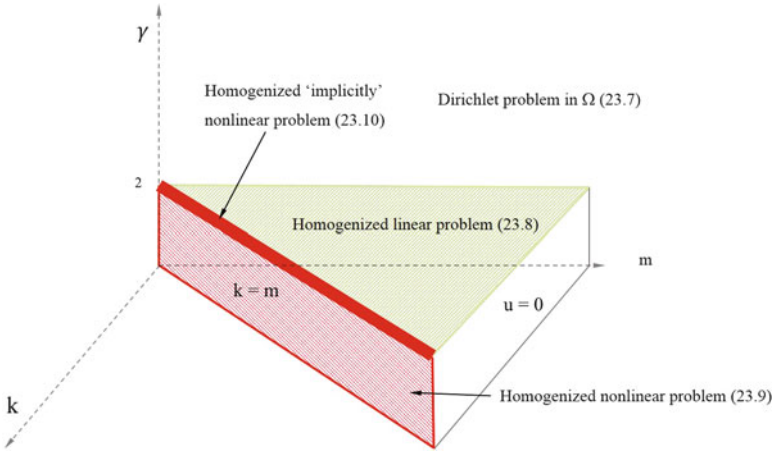


**Fig. 23.2** Sketch of homogenized problems depending on $\gamma, k, m$ for $a_\varepsilon = \varepsilon^k e^{-\alpha^2/\varepsilon^\gamma}$ and $\beta(\varepsilon) = \varepsilon^{2-m} e^{\alpha^2/\varepsilon^\gamma}$.

both the tubes and the boundary conditions for any $\beta(\varepsilon)$. This recalls the case of a Dirichlet condition on the boundary of the tubes when $\beta(\varepsilon) \to \infty$ or a Neumann condition when $\beta(\varepsilon) \to 0$ (cf. [CiMu82] and [LoOl97]). Now, the relations between parameters would satisfy (23.40) or (23.44) and the limiting problem is (23.7). In the case where $\gamma < 2$, the diameter of the tubes is very large compared with the distance between them. Depending on $\beta(\varepsilon)$ we obtain that an extension of the solution of (23.2) converges either towards zero in $H^1(\Omega)$ (case where $k < m$: $\beta(\varepsilon)$ is very large; relations $\beta(\varepsilon) a_\varepsilon \varepsilon^{-2} \to \infty$ and $\varepsilon^2 \ln(a_\varepsilon) \to 0$ hold) or towards the solution of the Dirichlet problem (23.7) (case where $k > m$: $\beta(\varepsilon)$ is very small; relation (23.40) holds).

In the case where $\gamma = 2$, while $k < m$ the adsorption parameter is very large and it seems as though the Dirichlet condition is asymptotically imposed on the boundary of the tubes and, consequently the limiting problem is the classical linear one (23.8) ignoring both the shapes of the tubes and the datum $\sigma$. Note that in this case we fit into the relations (23.33) between parameters (see the open triangle on the plane $\gamma = 2$ in Figure 23.2). Above $\beta(\varepsilon)$ small or large means in comparison with $|\partial G_\varepsilon|^{-1} \equiv O(\varepsilon^2 a_\varepsilon^{-1})$. Note that $\gamma = 2$ is also the classical critical size of the tubes in the homogenization of Dirichlet boundary condition, without adsorption parameters (cf. [CiMu82]).

In the case where $\gamma \in (0,2)$ and $k = m$ we deal with large diameters of tubes and the critical relation between $\beta(\varepsilon)$, $a_\varepsilon$ and $\varepsilon$ given by $(23.25)_1$. This choice of parameters implies that the total area of the tubes multiplied by the adsorption parameter $\beta(\varepsilon)$ is of order $O(1)$, and the homogenized problem contains a nonlinear term $2\pi C^2 \sigma(x,u)$ (cf. problem (23.9)). Here, fixed $\beta(\varepsilon)$ we have a critical relation for $a_\varepsilon$ and vice versa. Finally, the most critical relations are provided by $\gamma = 2$ and $k = m$ (namely, (23.21)) where the nonlinear homogenized problem (23.10) contains

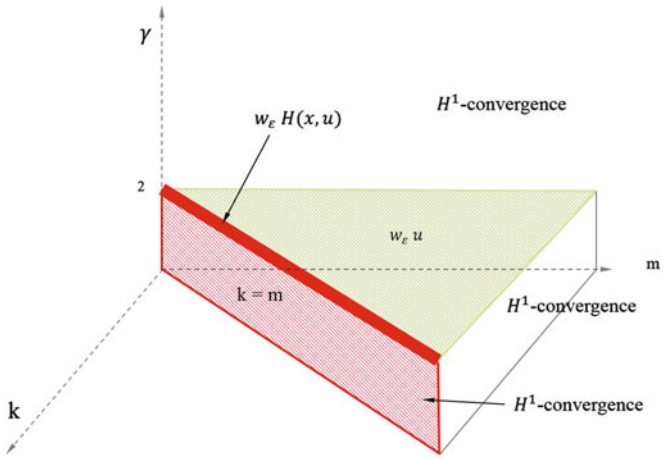


**Fig. 23.3** Sketch of corrector terms depending on $\gamma, k, m$.

what can be considered as the averaged influence of the large adsorption parameter, namely $2\pi\alpha^{-2}H(x,u)$ with $H$ a function of variables $x$ and $u$ defined implicitly by the functional equation (23.11).

Summarizing the above comments, in (23.6), $\alpha^2 > 0$ provides the usual *critical size* of the transversal sections of the cylinders for linear problems with Dirichlet conditions (cf. [CiMu82] and [MaKh05]), but since we deal with large adsorption parameters we also find a *critical relation* for the parameter $\beta(\varepsilon)$ when $\alpha^2 > 0$: relation which implies $|\partial G_\varepsilon|\beta(\varepsilon) = O(1)$, namely $C^2 > 0$, while the limit behavior of the solution changes drastically for $C = 0$ or $C = \infty$. We emphasize that the convergence results hold for a more general geometry of the tubes provided and a non-certain periodically distribution (cf. [GoLo14]).

As regards the correctors that we construct here for a circular transverse section of the cylinders, we use the classical test functions for linear homogenization problems with *forest of cylinders*, namely functions $w_\varepsilon$ defined in (23.14), which we combine suitably with the non-linear function defining the average equation (cf. Theorems 2 and 4). Since, in the case where the size of the tubes is very small compared with the distance between them ($\varepsilon^2 \ln(a_\varepsilon) \to -\infty$), $w_\varepsilon \to 0$ in $H^1(\Omega)$, we need to add the corrector to improve convergence only in the case where $\varepsilon^2 \ln(a_\varepsilon) \to -\alpha^2 < 0$, namely for the critical size of the tubes and limiting problems described in statements *iii*) and *v*) of Theorem 1 (see plane $\gamma = 2$ in Figure 23.3). Note the nonlinear dependence on $u$ of the corrector term in the most critical case. In the extreme cases (see statements *ii*) and *iv*) of Theorem 1) we obtain bounds for convergence rates for the discrepancies $u_\varepsilon - u$ in $H^1$ (cf. Theorems 3, 5 and 6). In the case of big sizes of tubes, namely $\varepsilon^2 \ln(a_\varepsilon) \to 0$, we use also the auxiliary function $\theta_\varepsilon$ (cf. (23.30)) to transform surface integrals into volume integrals and derive estimates for $u_\varepsilon - u$ in $H^1$ (see Theorem 3).

## 23.2 Preliminary Results

In this section, we introduce certain functions which allow us to construct correctors and obtain precise bounds for discrepancies.

Let $P_\varepsilon^j$ be the center of the ball $G_\varepsilon^j$ in $\mathbb{R}^2_{x_2x_3}$ and we denote by $T^j_{\varepsilon/4}$ the ball of radius $\varepsilon/4$ with center $P_\varepsilon^j$. For $j \in \Upsilon_\varepsilon$ let us consider the auxiliary problem

$$\begin{cases} \Delta_{\hat{x}} w_\varepsilon^j = 0 \text{ in } T^j_{\varepsilon/4} \setminus \overline{G_\varepsilon^j}, \\ w_\varepsilon^j = 1 \quad \text{on } \partial G_\varepsilon^j, \\ w_\varepsilon^j = 0 \quad \text{on } \partial T^j_{\varepsilon/4}, \end{cases} \tag{23.13}$$

and we introduce the function defined in $\Omega$ by

$$w_\varepsilon(x) = \begin{cases} w_\varepsilon^j(\hat{x}),\ x \in (0,l) \times (T^j_{\varepsilon/4} \setminus \overline{G_\varepsilon^j}), \ \text{if } j \in \Upsilon_\varepsilon, \\ 1, \qquad x \in (0,l) \times \overline{G_\varepsilon^j}, j \in \Upsilon_\varepsilon, \\ 0, \qquad x \in \Omega \setminus \bigcup_{j\in\Upsilon_\varepsilon} (0,l) \times T^j_{\varepsilon/4}. \end{cases} \tag{23.14}$$

As is well known, the solution of (23.13) can be constructed explicitly:

$$w_\varepsilon^j(\hat{x}) = \frac{\ln(4r/\varepsilon)}{\ln(4a_\varepsilon/\varepsilon)} \quad \text{where} \quad r = |\hat{x} - P_\varepsilon^j|, \tag{23.15}$$

where we assume that $a_\varepsilon < \frac{\varepsilon}{4}$. Thus, we can compute

$$\begin{gathered} \|w_\varepsilon\|_{L^2(\Omega_\varepsilon)} \le K |\ln(4a_\varepsilon/\varepsilon)|^{-1}, \quad \|\nabla w_\varepsilon\|_{L^2(\Omega)} \le K\varepsilon^{-1} |\ln(4a_\varepsilon/\varepsilon)|^{-1/2} \\ \|\nabla w_\varepsilon\|_{L^p(\Omega)} \le K\varepsilon \quad \text{for } p \in (1,2) \end{gathered} \tag{23.16}$$

and, consequently, as $\varepsilon \to 0$,

$$\begin{aligned} &w_\varepsilon \rightharpoonup 0 \text{ in } H^1(\Omega) - \text{ weak} && \text{if } \varepsilon^2 \ln(a_\varepsilon) \to -\alpha^2 < 0, \\ &w_\varepsilon \to 0 \text{ in } H^1(\Omega) && \text{if } \varepsilon^2 \ln(a_\varepsilon) \to -\infty. \end{aligned}$$

Also, for the sake of completeness, we introduce here certain results which prove to be useful for the proofs throughout Sections 23.3 and 23.4. We refer to Proposition 1 in [GoLo14] for the proof of Lemma 1. See Lemma 1.6 in Chapter I of [OlSh92] for the proof of Lemma 2.

In the lemmas below, and in what follows, $K$ denotes a constant independent of $\varepsilon$. Also, in these lemmas, the constant $K$ does not depend on the functions $w, h$ appearing in their statements.

**Lemma 1.** *There exists an operator $\mathscr{P}_\varepsilon$ from $H^1(\Omega_\varepsilon, \Gamma_\varepsilon)$ into $H_0^1(\Omega)$, such that for $w \in H^1(\Omega_\varepsilon, \Gamma_\varepsilon)$ we set $\mathscr{P}_\varepsilon w = \widetilde{w}$ the function which satisfies: $\widetilde{w}(x) = w(x)$ for $x \in \Omega_\varepsilon$, and*

$$\|\widetilde{w}\|_{H^1(\Omega)} \le K\|w\|_{H^1(\Omega_\varepsilon)} \quad \text{and} \quad \|\nabla\widetilde{w}\|_{L^2(\Omega)} \le K\|\nabla w\|_{L^2(\Omega_\varepsilon)}.$$

**Lemma 2.** *Let $g(y)$ be a bounded and measurable function in $y \in \mathbb{R}^3$, 1-periodic and $< g >_Q = 0$ where $Q = (0,1) \times (-1/2,1/2)^2 \subset \mathbb{R}^3$. Then,*

$$\left| \int_\Omega whg(x/\varepsilon)\,dx \right| \le K\varepsilon\|w\|_{H^1(\Omega)}\|h\|_{H^1(\Omega)}, \quad \forall w,h \in H^1(\Omega).$$

**Lemma 3.** *Let $T^j_{\varepsilon/4}$ be the ball of radius $\varepsilon/4$ with center $P^j_\varepsilon$. Then,*

$$\left| \sum_{j\in\Upsilon_\varepsilon} 4\varepsilon \int_{(0,l)\times\partial T^j_{\varepsilon/4}} h\,ds - 2\pi \int_\Omega h\,dx \right| \le K\varepsilon\|h\|_{H^1(\Omega)}, \quad \forall h \in H^1_0(\Omega). \tag{23.17}$$

*(23.17) also holds if $h \equiv h^\varepsilon$ with $\|h^\varepsilon\|_{H^1(\Omega)}$ bounded independently of $\varepsilon$.*

*Proof.* Let $Q = (0,1) \times (-1/2,1/2)^2 \subset \mathbb{R}^3$, $Q_1 = Q \setminus [0,1] \times \overline{T_{1/4}}$, where $T_{1/4}$ is the circle of radius 1/4 with the center in the origin of coordinates. Let $M(y)$ be a solution of the auxiliary problem

$$\begin{cases} \Delta_y M = \mu, & y \in Q_1, \\ \partial_{\nu_y} M = 1, & y \in [0,1] \times \partial T_{1/4}, \\ \partial_{\nu_y} M = 0, & y \in \partial Q_1 \setminus [0,1] \times \partial T_{1/4}, \\ < M >_{Q_1} = 0, \end{cases}$$

where $\mu = \frac{\pi}{2|Q_1|}$. We set $M_\varepsilon(x) = \varepsilon^2 M(\frac{x}{\varepsilon})$. It is easy to see that $M_\varepsilon$ is a solution of the following problem

$$\begin{cases} \Delta_x M_\varepsilon = \mu, & x \in \varepsilon Q_1, \\ \partial_{\nu_x} M_\varepsilon = \varepsilon, & x \in [0,\varepsilon] \times \partial(\varepsilon T_{1/4}), \\ \partial_{\nu_x} M_\varepsilon = 0, & x \in \partial(\varepsilon Q_1) \setminus [0,\varepsilon] \times \partial(\varepsilon T_{1/4}), \\ < M_\varepsilon >_{\varepsilon Q_1} = 0. \end{cases}$$

Then,

$$\int_{\varepsilon Q_1} div(\nabla_x(M_\varepsilon)h)\,dx = \frac{\pi}{2|Q_1|}\int_{\varepsilon Q_1} h\,dx + \int_{\varepsilon Q_1} \nabla_x M_\varepsilon \nabla_x h\,dx = \varepsilon \int_{(0,\varepsilon)\times\partial(\varepsilon T_{1/4})} h\,ds.$$

We can write such an equality for any set $\varepsilon Q^k_j = (\varepsilon Q + \varepsilon(k,j)) \setminus \varepsilon[k,k+1] \times (\overline{\varepsilon T_{1/4}} + \varepsilon j)$, $j \in \Upsilon_\varepsilon$, $k = 0,1,\ldots,K_\varepsilon$, $K_\varepsilon = c\varepsilon^{-1}$ and, taking sums over $j \in \Upsilon_\varepsilon$ and $k = 0,1,\ldots,K_\varepsilon$, we obtain

$$\sum_{j\in\Upsilon_\varepsilon}\varepsilon\int\limits_{(0,l)\times\partial T^j_{\varepsilon/4}} h\,ds=\frac{\pi}{2|Q_1|}\sum_{j\in\Upsilon_\varepsilon}\sum_{k=0}^{K_\varepsilon}\int\limits_{\varepsilon Q^k_j} h\,dx+\sum_{j\in\Upsilon_\varepsilon}\sum_{k=0}^{K_\varepsilon}\int\limits_{\varepsilon Q^k_j}\nabla_x M_\varepsilon\nabla_x h\,dx. \tag{23.18}$$

Now, we consider the function $g$ defined in $Q$ by $g(y)=\frac{1}{|Q_1|}-1$ if $y\in Q_1$ and $g(y)=-1$ if $y\in Q\setminus Q_1$, and extended by periodicity to $\mathbb{R}^3$. Then, applying Lemma 2, we have

$$\frac{1}{|Q_1|}\sum_{j\in\Upsilon_\varepsilon}\sum_{k=0}^{K_\varepsilon}\int\limits_{\varepsilon Q^k_j} h\,dx-\int\limits_{\Omega} h\,dx=\int\limits_{\Omega} hg(x/\varepsilon)\,dx\le K\varepsilon\|h\|_{H^1(\Omega)}. \tag{23.19}$$

In addition, since $M_\varepsilon(x)=\varepsilon^2 M(x/\varepsilon)$, we have

$$\Big|\sum_{j\in\Upsilon_\varepsilon}\sum_{k=0}^{K_\varepsilon}\int\limits_{\varepsilon Q^k_j}\nabla_x M_\varepsilon\nabla_x h\,dx\Big|\le\varepsilon\Big|\sum_{j\in\Upsilon_\varepsilon}\sum_{k=0}^{K_\varepsilon}\int\limits_{\varepsilon Q^k_j}\nabla_y M\nabla_x h\,dx\Big|\le K\varepsilon\|h\|_{H^1(\Omega_\varepsilon)}. \tag{23.20}$$

Finally, gathering (23.18), (23.19), and (23.20) we deduce (23.17) and Lemma 3 holds.

## 23.3 The Case of Nonlinear Homogenized Problems: Corrector and Energy Estimates

In this section we consider the case where the adsorption parameter multiplied by the total area of the tubes is of order $O(1)$; that is, $\lim_{\varepsilon\to 0}\beta(\varepsilon)|S_\varepsilon|\neq 0$. Depending on the size of the tubes, we have two different limiting problems, with an average provided by a nonlinear function which can be $\sigma$ multiplied by some averaged constant, or it can be defined implicitly as the solution of a functional equation (cf. (23.11)). This depends on whether the relation between the diameters of the transverse sections of tubes $O(a_\varepsilon)$ and the cell scale $\varepsilon$ satisfy $\varepsilon^2\ln(a_\varepsilon)=o(1)$ or $\varepsilon^2\ln(a_\varepsilon)$ has a non null finite limit, providing in any case a *critical size* of tubes. The last case is referred to as the most critical case and it is considered in Theorem 2. The case of *large sizes* of transverse sections of tubes is considered in Theorem 3. We recall that $a_\varepsilon\ll\varepsilon$ always holds.

**Theorem 2.** *Let us assume the $\beta(\varepsilon)$ and $a_\varepsilon$ satisfy the following conditions*

$$\beta(\varepsilon)a_\varepsilon\varepsilon^{-2}\to C^2>0\quad\textit{and}\quad\varepsilon^2\ln(a_\varepsilon)\to-\alpha^2<0\quad\textit{as }\varepsilon\to 0. \tag{23.21}$$

*Let $w_\varepsilon$ be the function defined by (23.14). Let $u_\varepsilon$ be the solution of (23.3) and $u \in H_0^1(\Omega)$ the weak solution of problem (23.10), with the additional regularity $u \in C^1(\overline{\Omega})$. Then, we have*

$$\begin{aligned}\|u_\varepsilon - u + w_\varepsilon H(x,u)\|^2_{H^1(\Omega_\varepsilon)} + \beta(\varepsilon)\|u_\varepsilon - u + H(x,u)\|^2_{L^2(S_\varepsilon)} \\ \leq K[\varepsilon + (a_\varepsilon \varepsilon^{-1})^{1/2} + \delta_1^\varepsilon + \delta_2^\varepsilon],\end{aligned} \tag{23.22}$$

*where $\delta_1^\varepsilon = |\beta(\varepsilon)a_\varepsilon\varepsilon^{-2} - C^2|$ and $\delta_2^\varepsilon = |\varepsilon^2 \ln(4a_\varepsilon/\varepsilon) + \alpha^2|$.*

*Proof.* Let us consider (23.3) with $\psi = u_\varepsilon - u + w_\varepsilon H(x,u) \in H^1(\Omega_\varepsilon, \Gamma_\varepsilon)$, with $H$ defined as in Theorem 1, and take in the integral identity for the limit problem (23.10) the test function $v_\varepsilon = \widetilde{u}_\varepsilon - u + w_\varepsilon H(x,u) \in H_0^1(\Omega)$, where $w_\varepsilon$ is defined by (23.14). Subtracting both equalities we obtain

$$\begin{aligned}&\|\nabla(u_\varepsilon - u + w_\varepsilon H(x,u))\|^2_{L^2(\Omega_\varepsilon)} \\ &+ \beta(\varepsilon)\int_{S_\varepsilon} (\sigma(x,u_\varepsilon) - \sigma(x,u - H(x,u)))(u_\varepsilon - u + H(x,u))\,ds \\ &= \int_{G_\varepsilon} \nabla u \nabla(\widetilde{u}_\varepsilon - u + w_\varepsilon H(x,u))\,dx - \int_{G_\varepsilon} f(\widetilde{u}_\varepsilon - u + w_\varepsilon H(x,u))\,dx \\ &+ \frac{2\pi}{\alpha^2}\int_\Omega H(x,u)(\widetilde{u}_\varepsilon - u + w_\varepsilon H(x,u))\,dx \\ &- \beta(\varepsilon)\int_{S_\varepsilon} \sigma(x,u - H(x,u))(u_\varepsilon - u + H(x,u))\,ds \\ &+ \int_{\Omega_\varepsilon} \nabla(w_\varepsilon H(x,u))\nabla(u_\varepsilon - u + w_\varepsilon H(x,u))\,dx.\end{aligned}$$

Taking into account that

$$\begin{aligned}&\int_{\Omega_\varepsilon} \nabla(w_\varepsilon H)\nabla(u_\varepsilon - u + w_\varepsilon H)dx = \int_{\Omega_\varepsilon} \nabla w_\varepsilon \nabla[H(u_\varepsilon - u + w_\varepsilon H)]dx \\ &- \int_{\Omega_\varepsilon} \nabla w_\varepsilon \nabla H(u_\varepsilon - u + w_\varepsilon H)dx + \int_{\Omega_\varepsilon} w_\varepsilon \nabla H \nabla(u_\varepsilon - u + w_\varepsilon H)dx,\end{aligned}$$

we conclude

$$\begin{aligned}
&\|\nabla(u_\varepsilon - u + w_\varepsilon H)\|^2_{L^2(\Omega_\varepsilon)} + \beta(\varepsilon)\int_{S_\varepsilon} [\sigma(x,u_\varepsilon) - \sigma(x,u-H)](u_\varepsilon - u + H)\,ds \\
&= \int_{G_\varepsilon} \nabla u \nabla(\widetilde{u}_\varepsilon - u + w_\varepsilon H)\,dx - \int_{G_\varepsilon} f(\widetilde{u}_\varepsilon - u + w_\varepsilon H)\,dx + R_\varepsilon \\
&+ \int_{\Omega_\varepsilon} w_\varepsilon \nabla H \nabla(u_\varepsilon - u + w_\varepsilon H)dx - \int_{\Omega_\varepsilon} \nabla w_\varepsilon \nabla H(u_\varepsilon - u + w_\varepsilon H)dx
\end{aligned} \tag{23.23}$$

where

$$\begin{aligned}
R_\varepsilon = &\int_{\Omega_\varepsilon} \nabla w_\varepsilon \nabla[H(u_\varepsilon - u + w_\varepsilon H)]dx + \frac{2\pi}{\alpha^2}\int_{\Omega} H(\widetilde{u}_\varepsilon - u + w_\varepsilon H)\,dx \\
&- \beta(\varepsilon)\int_{S_\varepsilon} \sigma(x,u-H)(u_\varepsilon - u + H)\,ds.
\end{aligned}$$

Let us estimate each term on the right-hand side of (23.23). Denoting by $|G_\varepsilon|$ and $|S_\varepsilon|$ the volume of $G_\varepsilon$ and the area of $S_\varepsilon$, respectively, and using the regularity of $u$, (23.4), (23.16), (23.21) and the embeddings of the spaces $L^r(\Omega)$ with $1 \le r \le \infty$ and of $H^1_0(\Omega)$ into $L^6(\Omega)$, we obtain

$$\left|\int_{G_\varepsilon} \nabla u \nabla(\widetilde{u}_\varepsilon - u + w_\varepsilon H)\,dx\right| \le K|G_\varepsilon|^{1/2}\|\nabla(\widetilde{u}_\varepsilon - u + w_\varepsilon H)\|_{L^2(\Omega)} \le K a_\varepsilon \varepsilon^{-1},$$

$$\begin{aligned}
\left|\int_{G_\varepsilon} f(\widetilde{u}_\varepsilon - u + w_\varepsilon H)\,dx\right| &\le \|f\|_{L^{4/3}(G_\varepsilon)}\|\widetilde{u}_\varepsilon - u + w_\varepsilon H\|_{L^4(\Omega)} \\
&\le K|G_\varepsilon|^{1/4}\|f\|_{L^2(G_\varepsilon)}\|\widetilde{u}_\varepsilon - u + w_\varepsilon H\|_{H^1(\Omega)} \le K(a_\varepsilon \varepsilon^{-1})^{1/2},
\end{aligned}$$

$$\begin{aligned}
\left|\int_{\Omega_\varepsilon} w_\varepsilon \nabla H \nabla(u_\varepsilon - u + w_\varepsilon H)dx\right| &\le K\|w_\varepsilon\|_{L^2(\Omega_\varepsilon)}\|u_\varepsilon - u + w_\varepsilon H\|_{H^1(\Omega_\varepsilon)} \\
&\le K|\ln(4a_\varepsilon/\varepsilon)|^{-1} \le K\varepsilon^2,
\end{aligned}$$

$$\Big|\int\limits_{\Omega_\varepsilon}\nabla w_\varepsilon\nabla H(u_\varepsilon-u+w_\varepsilon H)dx\Big|\le K\|\nabla w_\varepsilon\|_{L^{6/5}(\Omega_\varepsilon)}\|\widetilde{u}_\varepsilon-u+w_\varepsilon H\|_{L^6(\Omega)}$$

$$\le K\varepsilon\|\widetilde{u}_\varepsilon-u+w_\varepsilon H\|_{H^1(\Omega)}\le K\varepsilon.$$

Let us estimate $R_\varepsilon$. Using Green's formula and the definitions of $w_\varepsilon$ and $H$ (cf. (23.14) and (23.11)), we obtain

$$R_\varepsilon=\sum_{j\in\Upsilon_\varepsilon}\int\limits_{(0,l)\times(\partial T^j_{\varepsilon/4}\cup\partial G^j_\varepsilon)}\partial_\nu w^j_\varepsilon h_\varepsilon\,ds+\frac{2\pi}{\alpha^2}\int\limits_\Omega h_\varepsilon\,dx-\frac{\beta(\varepsilon)}{\alpha^2C^2}\int\limits_{S_\varepsilon}h_\varepsilon\,ds,$$

where $h_\varepsilon=H(\widetilde{u}_\varepsilon-u+w_\varepsilon H)$. Now, on account of (23.15), Lemma 3, relation $(23.21)_{(2)}$ and the boundedness of $h_\varepsilon$ in $H^1(\Omega)$ (cf. (23.4) and (23.16)), we have

$$\Big|\sum_{j\in\Upsilon_\varepsilon}\int\limits_{(0,l)\times\partial T^j_{\varepsilon/4}}\partial_\nu w^j_\varepsilon h_\varepsilon\,ds+\frac{2\pi}{\alpha^2}\int\limits_\Omega h_\varepsilon\,dx\Big|$$

$$=\Big|\sum_{j\in\Upsilon_\varepsilon}\frac{4}{\varepsilon\ln(4a_\varepsilon/\varepsilon)}\int\limits_{(0,l)\times\partial T^j_{\varepsilon/4}}h_\varepsilon\,ds+\frac{2\pi}{\alpha^2}\int\limits_\Omega h_\varepsilon\,dx\Big|$$

$$\le\Big|\frac{1}{\varepsilon^2\ln(4a_\varepsilon/\varepsilon)}\Big[\sum_{j\in\Upsilon_\varepsilon}4\varepsilon\int\limits_{(0,l)\times\partial T^j_{\varepsilon/4}}h_\varepsilon\,ds-2\pi\int\limits_\Omega h_\varepsilon\,dx\Big]\Big|$$

$$+2\pi\Big|\Big[\frac{1}{\varepsilon^2\ln(4a_\varepsilon/\varepsilon)}+\frac{1}{\alpha^2}\Big]\int\limits_\Omega h_\varepsilon\,dx\Big|$$

$$\le K[\varepsilon+|\varepsilon^2\ln(4a_\varepsilon/\varepsilon)+\alpha^2|]\|h_\varepsilon\|_{H^1(\Omega)}\le K(\varepsilon+\delta_2^\varepsilon).$$

Moreover, by (23.15), the area of $S_\varepsilon$ and (23.21), we get

$$\Big|\sum_{j\in\Upsilon_\varepsilon}\int\limits_{(0,l)\times\partial G^j_\varepsilon}\partial_\nu w^j_\varepsilon h_\varepsilon\,ds-\frac{\beta(\varepsilon)}{\alpha^2C^2}\int\limits_{S_\varepsilon}h_\varepsilon\,ds\Big|$$

$$=\Big|\Big[\frac{1}{a_\varepsilon\ln(4a_\varepsilon/\varepsilon)}+\frac{\beta(\varepsilon)}{\alpha^2C^2}\Big]\int\limits_{S_\varepsilon}H(u_\varepsilon-u+w_\varepsilon H)\,ds\Big|$$

$$\le K\Big|\frac{\varepsilon^2\ln(4a_\varepsilon/\varepsilon)+\alpha^2}{\alpha^2a_\varepsilon\ln(4a_\varepsilon/\varepsilon)}+\frac{\beta(\varepsilon)a_\varepsilon\varepsilon^{-2}-C^2}{\alpha^2C^2a_\varepsilon\varepsilon^{-2}}\Big|[|S_\varepsilon|+\|u_\varepsilon-u+w_\varepsilon H\|^2_{L^2(S_\varepsilon)}]$$

$$\le K[\delta_1^\varepsilon+\delta_2^\varepsilon]+K(\delta_1^\varepsilon+\delta_2^\varepsilon)\beta(\varepsilon)\|u_\varepsilon-u+w_\varepsilon H\|^2_{L^2(S_\varepsilon)}.$$

Thus,

$$|R_\varepsilon| \le K[\varepsilon + \delta_1^\varepsilon + \delta_2^\varepsilon] + K(\delta_1^\varepsilon + \delta_2^\varepsilon)\beta(\varepsilon)\|u_\varepsilon - u + w_\varepsilon H\|^2_{L^2(S_\varepsilon)}.$$

Gathering (23.23) and the above estimates, we conclude that

$$\|\nabla(u_\varepsilon - u + w_\varepsilon H)\|^2_{L^2(\Omega_\varepsilon)} + \beta(\varepsilon)\int_{S_\varepsilon} [\sigma(x,u_\varepsilon) - \sigma(x,u-H)](u_\varepsilon - u + H)\,ds$$
$$\le K[(a_\varepsilon \varepsilon^{-1})^{1/2} + \varepsilon + \delta_1^\varepsilon + \delta_2^\varepsilon] + K(\delta_1^\varepsilon + \delta_2^\varepsilon)\beta(\varepsilon)\|u_\varepsilon - u + w_\varepsilon H\|^2_{L^2(S_\varepsilon)}],$$

and, from (23.1) and (23.21), we derive

$$\|\nabla(u_\varepsilon - u + w_\varepsilon H(x,u))\|^2_{L^2(\Omega_\varepsilon)} + \beta(\varepsilon)\|u_\varepsilon - u + H(x,u)\|^2_{L^2(S_\varepsilon)} \le K[\varepsilon + (a_\varepsilon \varepsilon^{-1})^{1/2} + \delta_1^\varepsilon + \delta_2^\varepsilon]. \tag{23.24}$$

To obtain (23.22) from (23.24), we consider the Poincaré inequality for the $H^1$-extension of $u_\varepsilon - u + w_\varepsilon H(x,u)$ to $\Omega$ in Lemma 1, namely for the function $\mathscr{P}_\varepsilon(u_\varepsilon - u + w_\varepsilon H(x,u)) \in H_0^1(\Omega)$, which satisfies

$$\|\nabla(\mathscr{P}_\varepsilon(u_\varepsilon - u + w_\varepsilon H(x,u)))\|^2_{L^2(\Omega)} \le K\|\nabla(u_\varepsilon - u + w_\varepsilon H(x,u))\|^2_{L^2(\Omega_\varepsilon)},$$

and consequently, we have

$$\|u_\varepsilon - u + w_\varepsilon H(x,u)\|^2_{L^2(\Omega_\varepsilon)} \le K\|\nabla(u_\varepsilon - u + w_\varepsilon H(x,u))\|^2_{L^2(\Omega_\varepsilon)}$$

and (23.22) also holds.

**Theorem 3.** *Let us assume the $\beta(\varepsilon)$ and $a_\varepsilon$ satisfy the following conditions*

$$\beta(\varepsilon)a_\varepsilon\varepsilon^{-2} \to C^2 > 0 \quad \textit{and} \quad \varepsilon^2 \ln(a_\varepsilon) \to 0 \quad \textit{as } \varepsilon \to 0. \tag{23.25}$$

*Let $u_\varepsilon$ be the solution of (23.3) and $u \in H_0^1(\Omega)$ the weak solution of problem (23.9), with the additional regularity $u \in C^1(\overline{\Omega})$. Then, we have*

$$\|u_\varepsilon - u\|^2_{H^1(\Omega_\varepsilon)} + \beta(\varepsilon)\|u_\varepsilon - u\|^2_{L^2(S_\varepsilon)} \le K[\varepsilon + (a_\varepsilon\varepsilon^{-1})^{1/2} + \delta_1^\varepsilon + \tilde{\delta}_2^\varepsilon], \tag{23.26}$$

*where $\delta_1^\varepsilon = |\beta(\varepsilon)a_\varepsilon\varepsilon^{-2} - C^2|$ and $\tilde{\delta}_2^\varepsilon = \varepsilon|\ln(\varepsilon/2a_\varepsilon)|^{1/2}$.*

*Proof.* Let us consider (23.3) with $\psi = u_\varepsilon - u \in H^1(\Omega_\varepsilon, \Gamma_\varepsilon)$ and take in the integral identity for the limit problem (23.9) the test function $v_\varepsilon = \widetilde{u}_\varepsilon - u \in H_0^1(\Omega)$. Subtracting both equalities we obtain

$$\begin{aligned}&\|\nabla(u_\varepsilon-u)\|^2_{L^2(\Omega_\varepsilon)}+\beta(\varepsilon)\int\limits_{S_\varepsilon}(\sigma(x,u_\varepsilon)-\sigma(x,u))(u_\varepsilon-u)\,ds\\&=\int\limits_{G_\varepsilon}\nabla u\nabla(\widetilde{u}_\varepsilon-u)\,dx-\int\limits_{G_\varepsilon}f(\widetilde{u}_\varepsilon-u)\,dx+I_\varepsilon\end{aligned}\tag{23.27}$$

where

$$I_\varepsilon=2\pi C^2\int\limits_{\Omega}\sigma(x,u)(\widetilde{u}_\varepsilon-u)\,dx-\beta(\varepsilon)\int\limits_{S_\varepsilon}\sigma(x,u)(u_\varepsilon-u)\,ds.$$

Computing the volume $G_\varepsilon$ and using (23.5) along with the convergence in Theorem 1, and the embeddings of the spaces $L^r(\Omega)$ with $1\le r\le\infty$ and of $H^1_0(\Omega)$ into $L^6(\Omega)$, we obtain

$$\Big|\int\limits_{G_\varepsilon}\nabla u\nabla(\widetilde{u}_\varepsilon-u)\,dx\Big|\le K|G_\varepsilon|^{1/2}\|\nabla(\widetilde{u}_\varepsilon-u)\|_{L^2(\Omega)}\le Ka_\varepsilon\varepsilon^{-1},\tag{23.28}$$

and

$$\begin{aligned}&\Big|\int\limits_{G_\varepsilon}f(\widetilde{u}_\varepsilon-u)\,dx\Big|\le\|f\|_{L^{4/3}(G_\varepsilon)}\|\widetilde{u}_\varepsilon-u\|_{L^4(\Omega)}\\&\le K|G_\varepsilon|^{1/4}\|f\|_{L^2(G_\varepsilon)}\|\widetilde{u}_\varepsilon-u\|_{H^1(\Omega)}\le K(a_\varepsilon\varepsilon^{-1})^{1/2}.\end{aligned}\tag{23.29}$$

In order to estimate $I_\varepsilon$ we introduce the function $\theta_\varepsilon$ as follows. We define $\theta^0_\varepsilon$ as a solution of the following problem

$$\begin{cases}\Delta_{\hat{x}}\theta^0_\varepsilon=\mu_\varepsilon & \text{in } \varepsilon Y\setminus\overline{G^0_\varepsilon},\\ \partial_\nu\theta^0_\varepsilon=-1 & \text{on } \partial G^0_\varepsilon,\\ \partial_\nu\theta^0_\varepsilon=0 & \text{on } \partial(\varepsilon Y\setminus\overline{G^0_\varepsilon})\setminus\partial G^0_\varepsilon,\end{cases}$$

where $\mu_\varepsilon=\frac{-2\pi a_\varepsilon\varepsilon^{-2}}{1-\pi(a_\varepsilon\varepsilon^{-1})^2}$ and $Y=(-1/2,1/2)^2$. We assume that $\int\limits_{\varepsilon Y\setminus\overline{G^0_\varepsilon}}\theta^0_\varepsilon d\hat{x}=0$. For $j\in\Upsilon_\varepsilon$, we denote by $\theta^j_\varepsilon(x)$ the solution of the problem posed in $Y^j_\varepsilon\setminus\overline{G^j_\varepsilon}=(\varepsilon Y\setminus\overline{G^0_\varepsilon})+\varepsilon j$. We denote by $\widehat{Y}_\varepsilon=\bigcup_{j\in\Upsilon_\varepsilon}(Y^j_\varepsilon\setminus\overline{G^j_\varepsilon})$ and introduce the function $\theta_\varepsilon$ such that $\theta_\varepsilon=\theta^j_\varepsilon$ in $Y^j_\varepsilon\setminus G^j_\varepsilon$. Then, it can be proved that (see [GoLo14] for details)

$$\|\theta_\varepsilon\|_{L^2(\widehat{Y_\varepsilon})}\le Ka_\varepsilon|\ln(\varepsilon/2a_\varepsilon)|^{1/2},\quad\|\nabla\theta_\varepsilon\|_{L^2(\widehat{Y_\varepsilon})}\le Ka_\varepsilon\varepsilon^{-1}|\ln(\varepsilon/2a_\varepsilon)|^{1/2}.\tag{23.30}$$

By means of $\theta_\varepsilon$, the integral on $S_\varepsilon$ of $I_\varepsilon$ can be transformed into a volume integral. Thus, we can write

$$I_\varepsilon = 2\pi C^2 \int_\Omega \sigma(x,u)(\widetilde{u}_\varepsilon - u)\,dx + \beta(\varepsilon) \sum_{j\in\Upsilon_\varepsilon} \mu_\varepsilon \int_{(0,l)\times(Y_\varepsilon^j \setminus G_\varepsilon^j)} \sigma(x,u)(\widetilde{u}_\varepsilon - u)dx$$

$$+ \beta(\varepsilon) \sum_{j\in\Upsilon_\varepsilon} \int_{(0,l)\times(Y_\varepsilon^j \setminus G_\varepsilon^j)} \nabla\theta_\varepsilon^j \nabla(\sigma(x,u)(\widetilde{u}_\varepsilon - u))dx = I_\varepsilon^1 + I_\varepsilon^2 + I_\varepsilon^3$$

where

$$I_\varepsilon^1 = 2\pi\Big[C^2 - \beta(\varepsilon)\frac{a_\varepsilon\varepsilon^{-2}}{1-\pi(a_\varepsilon\varepsilon^{-1})^2}\Big] \int_\Omega \sigma(x,u)(\widetilde{u}_\varepsilon - u)\,dx,$$

$$I_\varepsilon^2 = 2\pi\beta(\varepsilon)\frac{a_\varepsilon\varepsilon^{-2}}{1-\pi(a_\varepsilon\varepsilon^{-1})^2} \int_{\Omega\setminus(0,l)\times\widehat{Y_\varepsilon}} \sigma(x,u)(\widetilde{u}_\varepsilon - u)\,dx,$$

$$I_\varepsilon^3 = \beta(\varepsilon) \int_{(0,l)\times\widehat{Y_\varepsilon}} \nabla\theta_\varepsilon \nabla(\sigma(x,u)(\widetilde{u}_\varepsilon - u))\,dx.$$

Now, from (23.25), the regularity of $u$, (23.5), the convergence in Theorem 1 and (23.30), it follows that

$$|I_\varepsilon^1| \le K[\delta_1^\varepsilon + (a_\varepsilon\varepsilon^{-1})^2]\|\widetilde{u}_\varepsilon - u\|_{L^2(\Omega)} \le K[\delta_1^\varepsilon + (a_\varepsilon\varepsilon^{-1})^2],$$

$$|I_\varepsilon^2| \le K\beta(\varepsilon)a_\varepsilon\varepsilon^{-2}|\Omega\setminus(0,l)\times\widehat{Y}_\varepsilon|^{1/2}\|\widetilde{u}_\varepsilon - u\|_{L^2(\Omega)} \le K[\varepsilon + a_\varepsilon\varepsilon^{-1}],$$

$$|I_\varepsilon^3| \le K\beta(\varepsilon)\|\nabla\theta\|_{L^2(\widehat{Y_\varepsilon})}\|\widetilde{u}_\varepsilon - u\|_{H^1(\Omega)} \le K\beta(\varepsilon)a_\varepsilon\varepsilon^{-1}|\ln(\varepsilon/2a_\varepsilon)|^{1/2} \le K\widetilde{\delta}_2^\varepsilon,$$

and, consequently,

$$|I_\varepsilon| \le K[\delta_1^\varepsilon + a_\varepsilon\varepsilon^{-1} + \varepsilon + \widetilde{\delta}_2^\varepsilon]. \tag{23.31}$$

Gathering (23.27), (23.28), (23.29) and (23.31), we conclude that

$$\|\nabla(u_\varepsilon - u)\|_{L^2(\Omega_\varepsilon)}^2 + \beta(\varepsilon)\int_{S_\varepsilon}(\sigma(x,u_\varepsilon) - \sigma(x,u))(u_\varepsilon - u)\,ds$$

$$\le K[\delta_1^\varepsilon + (a_\varepsilon\varepsilon^{-1})^{1/2} + \varepsilon + \widetilde{\delta}_2^\varepsilon]$$

and, from (23.1),

$$\|\nabla(u_\varepsilon - u)\|^2_{L^2(\Omega_\varepsilon)} + \beta(\varepsilon)\|u_\varepsilon - u\|^2_{L^2(S_\varepsilon)} \leq K[\varepsilon + (a_\varepsilon \varepsilon^{-1})^{1/2} + \delta_1^\varepsilon + \tilde{\delta}_2^\varepsilon]. \tag{23.32}$$

To obtain (23.26) from (23.32), we apply the Poincaré inequality for the extension $\mathscr{P}_\varepsilon(u_\varepsilon - u) \in H_0^1(\Omega)$ as in Theorem 2, and the theorem is proved.

## 23.4 The Case of Linear Homogenized Problems: Corrector and Energy Estimates

In this section, we gather the results for the *extreme relation* between the parameters $\beta(\varepsilon)$ and $a_\varepsilon$; namely, for the cases in which one of the limits (or both limits) $\lim_{\varepsilon\to 0}\beta(\varepsilon)|S_\varepsilon|$ and $\lim_{\varepsilon\to 0}\varepsilon^2\ln(a_\varepsilon)$ is 0 or $\infty$, excepting the case (23.25). Since the Dirichlet problem (without foreign term) arises as the homogenized problem, excepting for the relation (23.33) between parameters, the strong convergence of solutions $u_\varepsilon$ of (23.2) hold and we obtain precise bounds for discrepancies between $u_\varepsilon$ and $u$. In the case where (23.33) we show that the corrector is the same arising in linear problems without any adsorption parameter. We recall that $a_\varepsilon \ll \varepsilon$ always holds.

**Theorem 4.** *Let us assume the $\beta(\varepsilon)$ and $a_\varepsilon$ satisfy the following conditions*

$$\beta(\varepsilon)a_\varepsilon\varepsilon^{-2} \to \infty \quad \textit{and} \quad \varepsilon^2\ln(a_\varepsilon) \to -\alpha^2 < 0 \quad \textit{as } \varepsilon \to 0. \tag{23.33}$$

*Let $w_\varepsilon$ be the function defined by (23.14). Let $u_\varepsilon$ be the solution of (23.3) and $u \in H_0^1(\Omega)$ the weak solution of problem (23.8), with the additional regularity $u \in C^1(\overline{\Omega})$. Then, we have*

$$\|u_\varepsilon - u + w_\varepsilon u\|^2_{H^1(\Omega_\varepsilon)} + \beta(\varepsilon)\|u_\varepsilon\|^2_{L^2(S_\varepsilon)} \leq K[\varepsilon + (a_\varepsilon\varepsilon^{-1})^{1/2} + \tilde{\delta}_1^\varepsilon + \delta_2^\varepsilon], \tag{23.34}$$

*where $\tilde{\delta}_1^\varepsilon = (\beta(\varepsilon)a_\varepsilon\varepsilon^{-2})^{-1/2}$ and $\delta_2^\varepsilon = |\varepsilon^2\ln(4a_\varepsilon/\varepsilon) + \alpha^2|$.*

*Proof.* Let us consider (23.3) with $\psi = u_\varepsilon - u + w_\varepsilon u \in H^1(\Omega_\varepsilon, \Gamma_\varepsilon)$ and take in the integral identity for the limit problem (23.8) the test function $v_\varepsilon = \widetilde{u}_\varepsilon - u + w_\varepsilon u \in H_0^1(\Omega)$, where $w_\varepsilon$ is defined by (23.14). Subtracting both equalities and taking into account that $w_\varepsilon = 1$ in $\overline{G}_\varepsilon$, we have

$$\begin{aligned} &\|\nabla(u_\varepsilon - u + w_\varepsilon u)\|^2_{L^2(\Omega_\varepsilon)} + \beta(\varepsilon)\int_{S_\varepsilon}\sigma(x,u_\varepsilon)u_\varepsilon\,ds \\ &= \int_{G_\varepsilon}\nabla u\nabla\widetilde{u}_\varepsilon\,dx - \int_{G_\varepsilon} f\widetilde{u}_\varepsilon\,dx + J_\varepsilon, \end{aligned} \tag{23.35}$$

where

$$J_\varepsilon = \int_{\Omega_\varepsilon} \nabla(w_\varepsilon u)\nabla(u_\varepsilon - u + w_\varepsilon u)\,dx + \frac{2\pi}{\alpha^2}\int_{\Omega} u(\widetilde{u}_\varepsilon - u + w_\varepsilon u)\,dx.$$

Owing to the volume $G_\varepsilon$ and (23.4), we obtain

$$\Big|\int_{G_\varepsilon} \nabla u \nabla \widetilde{u}_\varepsilon\,dx\Big| \le K|G_\varepsilon|^{1/2}\|\nabla\widetilde{u}_\varepsilon\|_{L^2(\Omega)} \le Ka_\varepsilon\varepsilon^{-1} \tag{23.36}$$

and

$$\begin{aligned}\Big|\int_{G_\varepsilon} f\widetilde{u}_\varepsilon\,dx\Big| &\le \|f\|_{L^{4/3}(G_\varepsilon)}\|\widetilde{u}_\varepsilon\|_{L^4(\Omega)} \le |G_\varepsilon|^{1/4}\|f\|_{L^2(G_\varepsilon)}\|\widetilde{u}_\varepsilon\|_{H^1(\Omega)}\\ &\le K(a_\varepsilon\varepsilon^{-1})^{1/2}.\end{aligned} \tag{23.37}$$

Let us estimate $J_\varepsilon$. Straightforward calculation yields

$$\begin{aligned}\int_{\Omega_\varepsilon} \nabla(w_\varepsilon u)\nabla(u_\varepsilon - u + w_\varepsilon u)\,dx &= \int_{\Omega_\varepsilon} \nabla(u(u_\varepsilon - u + w_\varepsilon u))\nabla w_\varepsilon\,dx\\ &+ \int_{\Omega_\varepsilon} \nabla u\nabla(u_\varepsilon - u + w_\varepsilon u)w_\varepsilon\,dx - \int_{\Omega_\varepsilon} \nabla u\nabla w_\varepsilon(u_\varepsilon - u + w_\varepsilon u)\,dx.\end{aligned}$$

Besides, using the definition of $w_\varepsilon$ and the Green formula, we have that

$$\begin{aligned}&\int_{\Omega_\varepsilon} \nabla(u(u_\varepsilon - u + w_\varepsilon u))\nabla w_\varepsilon\,dx\\ &= \sum_{j\in\Upsilon_\varepsilon}\int_{(0,l)\times\partial T^j_{\varepsilon/4}} \partial_\nu w^j_\varepsilon u(u_\varepsilon - u + w_\varepsilon u)\,ds + \sum_{j\in\Upsilon_\varepsilon}\int_{(0,l)\times\partial G^j_\varepsilon} \partial_\nu w^j_\varepsilon u u_\varepsilon\,ds.\end{aligned}$$

Thus, from (23.15), we can write $J_\varepsilon = J^1_\varepsilon + J^2_\varepsilon + J^3_\varepsilon$ where

$$J^1_\varepsilon = \sum_{j\in\Upsilon_\varepsilon}\frac{4}{\varepsilon\ln(4a_\varepsilon/\varepsilon)}\int_{(0,l)\times\partial T^j_{\varepsilon/4}} u(u_\varepsilon - u + w_\varepsilon u)\,ds + \frac{2\pi}{\alpha^2}\int_{\Omega} u(\widetilde{u}_\varepsilon - u + w_\varepsilon u)\,dx,$$

$$J^2_\varepsilon = -\sum_{j\in\Upsilon_\varepsilon}\frac{1}{a_\varepsilon\ln(4a_\varepsilon/\varepsilon)}\int_{(0,l)\times\partial G^j_\varepsilon} u u_\varepsilon\,ds,$$

$$J_\varepsilon^3 = \int\limits_{\Omega_\varepsilon} \nabla u \nabla (u_\varepsilon - u + w_\varepsilon u) w_\varepsilon \, dx - \int\limits_{\Omega_\varepsilon} \nabla u \nabla w_\varepsilon (u_\varepsilon - u + w_\varepsilon u) \, dx.$$

We denote by $\hat{h}_\varepsilon = \widetilde{u}_\varepsilon - u + w_\varepsilon u$. Using Lemma 3, (23.33), (23.4), (23.16) and the fact that $|S_\varepsilon| \le K a_\varepsilon \varepsilon^{-2}$, we have

$$\begin{aligned}
|J_\varepsilon^1| \le & \Big| \frac{1}{\varepsilon^2 \ln(4a_\varepsilon/\varepsilon)} \Big[ \sum_{j \in \Upsilon_\varepsilon} 4\varepsilon \int\limits_{(0,l)\times \partial T^j_{\varepsilon/4}} u \hat{h}_\varepsilon \, ds - 2\pi \int\limits_\Omega u \hat{h}_\varepsilon \, dx \Big] \Big| \\
& + 2\pi \Big| \Big[ \frac{1}{\varepsilon^2 \ln(4a_\varepsilon/\varepsilon)} + \frac{1}{\alpha^2} \Big] \int\limits_\Omega u \hat{h}_\varepsilon \, dx \Big| \\
\le & K \frac{1}{\varepsilon^2 |\ln(4a_\varepsilon/\varepsilon)|} \varepsilon \|\hat{h}_\varepsilon\|_{H^1(\Omega)} + K \frac{\delta_2^\varepsilon}{\varepsilon^2 |\ln(4a_\varepsilon/\varepsilon)|} \|\hat{h}_\varepsilon\|_{L^2(\Omega)} \le K[\varepsilon + \delta_2^\varepsilon],
\end{aligned}$$

$$|J_\varepsilon^2| \le K \frac{1}{a_\varepsilon |\ln(4a_\varepsilon/\varepsilon)|} |S_\varepsilon|^{1/2} \|u_\varepsilon\|_{L^2(S_\varepsilon)} \le K \widetilde{\delta}_1^\varepsilon \beta(\varepsilon)^{1/2} \|u_\varepsilon\|_{L^2(S_\varepsilon)} \le K \widetilde{\delta}_1^\varepsilon,$$

$$\begin{aligned}
|J_\varepsilon^3| \le & K \|\nabla \hat{h}_\varepsilon\|_{L^2(\Omega)} \|w_\varepsilon\|_{L^2(\Omega)} + K \|\nabla w_\varepsilon\|_{L^{6/5}(\Omega)} \|\hat{h}_\varepsilon\|_{L^6(\Omega)} \\
\le & K |\ln(4a_\varepsilon/\varepsilon)|^{-1} \|\nabla \hat{h}_\varepsilon\|_{L^2(\Omega)} + K\varepsilon \|\hat{h}_\varepsilon\|_{H^1(\Omega)} \le K[\varepsilon^2 + \varepsilon],
\end{aligned}$$

and, hence,

$$|J_\varepsilon| \le K[\varepsilon + \delta_2^\varepsilon + \widetilde{\delta}_1^\varepsilon]. \tag{23.38}$$

Gathering (23.35), (23.36), (23.37), and (23.38), we conclude that

$$\|\nabla(u_\varepsilon - u + w_\varepsilon u)\|^2_{L^2(\Omega_\varepsilon)} + \beta(\varepsilon) \int\limits_{S_\varepsilon} \sigma(x, u_\varepsilon) u_\varepsilon \, ds \le K[\varepsilon + (a_\varepsilon \varepsilon^{-1})^{1/2} + \tilde{\delta}_1^\varepsilon + \delta_2^\varepsilon]$$

and, by (23.1),

$$\|\nabla(u_\varepsilon - u + w_\varepsilon u)\|^2_{L^2(\Omega_\varepsilon)} + \beta(\varepsilon) \|u_\varepsilon\|^2_{L^2(S_\varepsilon)} \le K[\varepsilon + (a_\varepsilon \varepsilon^{-1})^{1/2} + \tilde{\delta}_1^\varepsilon + \delta_2^\varepsilon]. \tag{23.39}$$

Finally, to show (23.34) from (23.39) we rewrite the proof at the end of Theorem 2 with minor modifications.

**Theorem 5.** *Let us assume the $\beta(\varepsilon)$ and $a_\varepsilon$ satisfy the following condition*

$$\beta(\varepsilon) a_\varepsilon \varepsilon^{-2} \to 0 \quad \text{as } \varepsilon \to 0. \tag{23.40}$$

*Let $u_\varepsilon$ be the solution of (23.3) and $u \in H_0^1(\Omega)$ the weak solution of the Dirichlet problem (23.7), with the additional regularity $u \in C^1(\overline{\Omega})$. Then, we have*

$$\|u_\varepsilon - u\|^2_{H^1(\Omega_\varepsilon)} + \beta(\varepsilon)\|u_\varepsilon - u\|^2_{L^2(S_\varepsilon)} \leq K[(a_\varepsilon \varepsilon^{-1})^{1/2} + \beta(\varepsilon) a_\varepsilon \varepsilon^{-2}]. \tag{23.41}$$

*Proof.* Let us consider the variational formulations of (23.2) and (23.7) with $v = u_\varepsilon - u \in H^1(\Omega_\varepsilon, \Gamma_\varepsilon)$ and $v = \widetilde{u}_\varepsilon - u \in H_0^1(\Omega)$ as test functions, respectively. Subtracting both equalities, we have

$$\begin{aligned} &\|\nabla(u_\varepsilon - u)\|^2_{L^2(\Omega_\varepsilon)} + \beta(\varepsilon) \int_{S_\varepsilon} [\sigma(x, u_\varepsilon) - \sigma(x,u)](u_\varepsilon - u)\, ds \\ &= \int_{G_\varepsilon} \nabla u \nabla(\widetilde{u}_\varepsilon - u)\, dx - \int_{G_\varepsilon} f(\widetilde{u}_\varepsilon - u)\, dx - \beta(\varepsilon) \int_{S_\varepsilon} \sigma(x,u)(u_\varepsilon - u)\, ds. \end{aligned} \tag{23.42}$$

Let us estimate the last three terms of (23.42).

Computing $|G_\varepsilon|$ and $|S_\varepsilon|$ and using the regularity of $u$, (23.5) along with the convergence in Theorem 1, we obtain that

$$\left| \int_{G_\varepsilon} \nabla u \nabla(\widetilde{u}_\varepsilon - u)\, dx \right| \leq C|G_\varepsilon|^{1/2} \|\nabla(\widetilde{u}_\varepsilon - u)\|_{L^2(\Omega)} \leq K a_\varepsilon \varepsilon^{-1},$$

$$\begin{aligned} \left| \int_{G_\varepsilon} f(\widetilde{u}_\varepsilon - u)\, dx \right| &\leq \|f\|_{L^{4/3}(G_\varepsilon)} \|\widetilde{u}_\varepsilon - u\|_{L^4(\Omega)} \\ &\leq |G_\varepsilon|^{1/4} \|f\|_{L^2(G_\varepsilon)} \|\widetilde{u}_\varepsilon - u\|_{H^1(\Omega)} \leq K(a_\varepsilon \varepsilon^{-1})^{1/2}, \end{aligned}$$

$$\left| \int_{S_\varepsilon} \sigma(x,u)(u_\varepsilon - u)\, ds \right| \leq \delta \|u_\varepsilon - u\|^2_{L^2(S_\varepsilon)} + K\delta^{-1}|S_\varepsilon| \leq \delta \|u_\varepsilon - u\|^2_{L^2(S_\varepsilon)} + K a_\varepsilon \varepsilon^{-2},$$

with $\delta > 0$ arbitrary. Therefore, from (23.42) and (23.1), it follows that

$$\|\nabla(u_\varepsilon - u)\|^2_{L^2(\Omega_\varepsilon)} + \beta(\varepsilon)(k_1 - \delta)\|u_\varepsilon - u\|^2_{L^2(S_\varepsilon)} \leq K[(a_\varepsilon \varepsilon^{-1})^{1/2} + \beta(\varepsilon) a_\varepsilon \varepsilon^{-2}].$$

Now, choosing $\delta = k_1/2$ in the above expression yields

$$\|\nabla(u_\varepsilon - u)\|^2_{L^2(\Omega_\varepsilon)} + \beta(\varepsilon)\|u_\varepsilon - u\|^2_{L^2(S_\varepsilon)} \leq K[(a_\varepsilon \varepsilon^{-1})^{1/2} + \beta(\varepsilon) a_\varepsilon \varepsilon^{-2}]. \tag{23.43}$$

Since $|S_\varepsilon| \le K a_\varepsilon \varepsilon^{-2}$, by (23.40), we also have

$$\|\nabla(u_\varepsilon - u)\|^2_{L^2(\Omega_\varepsilon)} + \beta(\varepsilon)\|u_\varepsilon\|^2_{L^2(S_\varepsilon)} \le K[(a_\varepsilon\varepsilon^{-1})^{1/2} + \beta(\varepsilon)a_\varepsilon\varepsilon^{-2}].$$

Finally, to show (23.41) from (23.43) we rewrite the proof at the end of Theorem 2 with minor modifications.

**Theorem 6.** *Let us assume the $\beta(\varepsilon)$ and $a_\varepsilon$ satisfy the following condition*

$$\varepsilon^2 \ln(a_\varepsilon) \to -\infty \quad \text{as } \varepsilon \to 0. \tag{23.44}$$

*Let $u_\varepsilon$ be the solution of (23.3) and $u \in H^1_0(\Omega)$ the weak solution of the Dirichlet problem (23.7), with the additional regularity $u \in C^1(\overline{\Omega})$. Then, we have*

$$\|u_\varepsilon - u\|^2_{H^1(\Omega_\varepsilon)} + \beta(\varepsilon)\|u_\varepsilon\|^2_{L^2(S_\varepsilon)} \le K[(a_\varepsilon\varepsilon^{-1})^{1/2} + \varepsilon^{-1}|\ln(4a_\varepsilon/\varepsilon)|^{-1/2}]. \tag{23.45}$$

*Proof.* We consider the variational formulation of (23.2) and (23.7) and take $v = u_\varepsilon - u + w_\varepsilon u \in H^1(\Omega_\varepsilon, \Gamma_\varepsilon)$ and $v = \widetilde{u}_\varepsilon - u + w_\varepsilon u \in H^1_0(\Omega)$ as test functions, respectively, where $w_\varepsilon$ is defined by (23.14). Subtracting both expressions and taking into account that $w_\varepsilon = 1$ in $\overline{G}_\varepsilon$, we have

$$\int_{\Omega_\varepsilon} \nabla(u_\varepsilon - u)\nabla(u_\varepsilon - u + w_\varepsilon u)\,dx + \beta(\varepsilon)\int_{S_\varepsilon} \sigma(x,u_\varepsilon)u_\varepsilon\,ds = \int_{G_\varepsilon} \nabla u\nabla\widetilde{u}_\varepsilon\,dx - \int_{G_\varepsilon} f\widetilde{u}_\varepsilon\,dx$$

and, hence,

$$\|\nabla(u_\varepsilon - u + w_\varepsilon u)\|^2_{L^2(\Omega_\varepsilon)} + \beta(\varepsilon)\int_{S_\varepsilon} \sigma(x,u_\varepsilon)u_\varepsilon ds$$

$$= \int_{G_\varepsilon} \nabla u\nabla\widetilde{u}_\varepsilon dx - \int_{G_\varepsilon} f\widetilde{u}_\varepsilon dx + \int_{\Omega_\varepsilon} \nabla(w_\varepsilon u)\nabla(u_\varepsilon - u + w_\varepsilon u)dx.$$

Owing to (23.4), (23.16), and (23.44), and computing the volume of $G_\varepsilon$, we obtain that

$$\left|\int_{G_\varepsilon} \nabla u\nabla\widetilde{u}_\varepsilon dx\right| \le K|G_\varepsilon|^{1/2}\|\nabla\widetilde{u}_\varepsilon\|_{L^2(\Omega)} \le K a_\varepsilon\varepsilon^{-1},$$

$$\left|\int_{G_\varepsilon} f\widetilde{u}_\varepsilon\,dx\right| \le \|f\|_{L^{4/3}(G_\varepsilon)}\|\widetilde{u}_\varepsilon\|_{L^4(\Omega)} \le |G_\varepsilon|^{1/4}\|f\|_{L^2(G_\varepsilon)}\|\widetilde{u}_\varepsilon\|_{H^1(\Omega)}$$

$$\le K(a_\varepsilon\varepsilon^{-1})^{1/2},$$

and

$$\left| \int_{\Omega_\varepsilon} \nabla(w_\varepsilon u)\nabla(\widetilde{u}_\varepsilon - u + w_\varepsilon u)dx \right| \leq K\|\nabla(w_\varepsilon u)\|_{L^2(\Omega_\varepsilon)} \leq K\varepsilon^{-1}|\ln(4a_\varepsilon/\varepsilon)|^{-1/2}.$$

Thus, we have that

$$\|\nabla(u_\varepsilon - u + w_\varepsilon u)\|^2_{L^2(\Omega_\varepsilon)} + \beta(\varepsilon)\int_{S_\varepsilon} \sigma(x,u_\varepsilon)u_\varepsilon ds$$
$$\leq K[(a_\varepsilon\varepsilon^{-1})^{1/2} + \varepsilon^{-1}|\ln(4a_\varepsilon/\varepsilon)|^{-1/2}],$$

which already shows that $w_\varepsilon u$ is a corrector.

Finally, using again (23.16) and (23.1), it follows

$$\|\nabla(u_\varepsilon - u)\|^2_{L^2(\Omega_\varepsilon)} + \beta(\varepsilon)\|u_\varepsilon\|^2_{L^2(S_\varepsilon)} \leq K[(a_\varepsilon\varepsilon^{-1})^{1/2} + \varepsilon^{-1}|\ln(4a_\varepsilon/\varepsilon)|^{-1/2}]. \tag{23.46}$$

Now to show (23.45) from (23.46) we rewrite the proof at the end of Theorem 2 with minor modifications and the theorem holds.

*Remark 1.* As already noticed in [GoLo14], the hypothesis on $\sigma$ suffices for all the proofs throughout the paper. Nevertheless, depending on the section (namely, depending on the limits (23.6)) this hypothesis can be weakened by prescribing $0 \leq \sigma_u(x,u)$ or $\sigma_u(x,u) \leq k_2(1+|u|^\delta)$ for some $\delta \in [0,2]$.

*Remark 2.* As it is well known, in the case of a linear $\sigma$, estimates for convergence rates of solutions of stationary problems allows us to derive estimates for convergence rates of eigenelements of the associated spectral problems (see Theorems 1.4 and 1.7 in Section III.1 in [OlSh92] for the precise statement). Nevertheless, as outlined in [GoPe12] these estimates should involve the norms of the data $f$ in the natural setting of the spectral problems, and avoiding hypothesis on smoothness for solutions and using the technique in [GoPe13] and [GoLo13a] will likely provide us with weaker bounds in terms of the parameter $\varepsilon$ to be applied to the spectrum.

**Acknowledgements** This work is partially supported by grant MTM2013-44883-P.

## References

[CaDo12] Cabarrubias B., Donato P.: Homogenization of a quasilinear elliptic problem with nonlinear Robin boundary conditions. Appl. Anal. **91**(6), 1111–1127 (2012). DOI: 10.1080/00036811.2011.619982.

[CiMu82] Cioranescu D, Murat F.: Un terme étrange venu d'ailleurs I & II, in: Brezis H, Lions JL (Eds.) Nonlinear Partial Differential Equations and their Applications. Collège de France Séminar, Volume II & III, Research Notes in Mathematics, 60 & 70. Pitman: London, (1982) pp. 98–138 & 154–178.

[CoDi04] Conca C., Díaz J.I., Liñán A., Timofte C.: Homogenization in chemical reactive flows. Electron. J. Differential Equations, **2004**(40) 1–22 (2004).

[Go95] Goncharenko M.: The asymptotic behaviour of the third boundary-value problem solutions in domains with fine-grained boundaries, in: GAKUTO Internat. Ser. Math. Sci. Appl., 9, Homogenization and Applications to Material Sciences. Gakkotosho, Tokyo, (1995) pp. 203–213.

[GoLo13a] Gómez D, Lobo M, Pérez M.E., Shaposhnikova T.A.: On correctors for spectral problems in the homogenization of Robin boundary conditions with very large parameters. Int. J. Appl. Math. **26**, 309–320 (2013). DOI: 10.12732/ijam.v26i3.6.

[GoLo13b] Gómez, D., Lobo, M., Pérez, M.E., Shaposhnikova, T.A., Zubova, M.N.: Homogenization problem in domain perforated by thin tubes with nonlinear Robin type boundary condition. Dokl. Math. **87**, 5–11 (2013).

[GoLo14] Gómez, D., Lobo, M., Pérez, M.E., Shaposhnikova, T.A., Zubova, M.N.: On critical parameters in homogenization of perforated domains by thin tubes with nonlinear flux and related spectral problems. Math. Methods Appl. Sci. (2014). DOI: 10.1002/mma.3246.

[GoPe12] Gómez D., Pérez M.E., Shaposhnikova T.A.: On homogenization of nonlinear Robin type boundary conditions for cavities along manifolds and associated spectral problems. Asymptot. Anal. **80**, 289–322 (2012). DOI: 10.3233/ASY-2012-1116.

[GoPe13] Gómez D, Pérez M.E., Shaposhnikova T.A.: Spectral boundary homogenization problems in perforated domains with Robin boundary conditions and large parameters. in: Integral Methods in Science and Engineering, Progress in Numerical and Analytic Techniques. Birkhäuser/Springer: New York, pp. 155–174 (2013).

[LoOl97] Lobo M, Oleinik OA, Perez ME, Shaposhnikova TA.: On homogenization of solutions of boundary value problem in domains perforated along manifolds. Ann. Scuola Norm. Sup. Pisa Cl. Sci. Ser. 4 **25**, 611–629 (1997).

[LoPe11] Lobo M., Pérez M.E., Sukharev V.V., Shaposhnikova T.A.: Averaging of boundary-value problem in domain perforated along $(n-1)$ - dimensional manifold with nonlinear third type boundary conditions on the boundary of cavities. Dokl. Math. **83**, 34–38 (2011). DOI: 10.1134/S1064562411010108.

[MaKh05] Marchenko VA, Khruslov EYa. Homogenization of Partial Differential Equations. Springer: Boston, (2005).

[OlSh92] Oleinik, O.A., Shamaev, A.S., Yosifian, G.A.: Mathematical Problems in Elasticity and Homogenization. North-Holland, Amsterdam, (1992).

[Ti09] Timofte C.: Homogenization results for enzyme catalyzed reactions through porous media. Acta Math. Sci. Ser. B, **29**, 74–82 (2009). DOI: 10.1016/S0252-9602(09)60008-4.

[ZuSh11] Zubova M.N., Shaposhnikova T.A.: Homogenization of boundary value problems in perforated domains with the third boundary condition and the resulting change in the character of the nonlinearity in the problem. Differ. Equ. **47**, 78–90 (2011). DOI: 10.1134/S0012266111010095.

# Chapter 24
# A Finite Element Method For Deblurring Images

**P.J. Harris and K. Chen**

## 24.1 Introduction

This chapter considers a method for removing the noise and/or blurring from a typical digital image. The fundamental problem is methods for removing either noise or blurring from a digital image is one of the most important tasks in image analysis. The problem has been studied extensively and different formulations and methods have been widely reported in the literature (see [BrCh10, ChCh06, VoOm96, YaChYu12] and the references that they contain). Most recent methods reformulate the problem in terms of nonlinear partial differential equations which have to be solved to give the original, uncontaminated image. One notable feature of nearly all of the previous work on this problem is that the finite difference method has been used to solve the governing equations. Whilst the problem certainly lends itself to the finite difference method as it is essentially dealing with data on a uniform grid, there are circumstances under which this may not be the best approach. For example, if part of the image is masked giving a curved boundary, then it can be complicated to modify the finite difference approach to incorporate this.

An alternative approach is to use a finite element type method. Although the initial formulation is more complicated than for the finite difference method, the finite element method can deal with irregular shaped boundaries with no modifications provided the elements are defined in a sensible manner.

In this paper we will present a finite element method, based on the total variation method, for removing the noise and/or blurring from a digital image. In the next

P.J. Harris (✉)
University of Brighton, Lewes Road, Brighton BN2 4GJ, UK
e-mail: p.j.harris@brighton.ac.uk

K. Chen
University of Liverpool, Liverpool L69 7ZL, UK
e-mail: K.Chen@liverpool.ac.uk



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_24

section we will give a brief description of the finite element method for the general problem. In the third section we will discuss how the integral operator which models the blurring in the image is discretized and finally we will present some results for some typical images.

## 24.2 The Finite Element Formulation of the Problem

Let $u(x,y)$ denote a function which for integer values of $x$ and $y$ gives the intensity level of the pixel located at $(x,y)$ in the required uncontaminated digital image, and let $z(x,y)$ denote a function which gives the corresponding intensity levels when the image is contaminated by noise or blurring. For convenience it is assumed that the intensities are scaled such that $0 \leq u \leq 1$. The actual pixel values can be found by multiplying by an appropriate value (usually 255) and rounding to the nearest integer value.

The total variation method for removing the noise and blurring from an image can be written as [ChPiZa13]

$$\min_u \left( \alpha \int_\Omega |\nabla u|^2 \, dx\, dy + \|(\mathscr{A}u - z)\|^2 \right) \tag{24.1}$$

where $\mathscr{A}$ is an integral operator which introduced the blurring into the image, and $\alpha$ is a regularization parameter. Given the pixel values of $z$ we want to be able to calculate the pixel values of $u$. The integral operator in (24.1) will be discussed in the next section. If we are just removing noise from an image, then this is simply the identity operator.

The Euler–Lagrange equations for the total variation method (24.1) can be written as [ChPiZa13]

$$-\alpha \nabla \cdot \left( \frac{\nabla u}{\|\nabla u\|_\beta} \right) + \mathscr{A}^*(\mathscr{A}u - z) = 0 \tag{24.2}$$

over the domain $\Omega$ of the image with the boundary condition

$$\frac{\partial u}{\partial n} = 0$$

on the boundary $\Gamma$. Here

$$\|u\|_\beta = \sqrt{\nabla u \cdot \nabla u + \beta}$$

where $\beta$ is a small parameter used to avoid problems which can arise if $\nabla u \cdot \nabla u = 0$.

Approximate $u$ by

$$\tilde{u}(x,y) = \sum_{i=1}^{N} u_i \phi_i(x,y) \tag{24.3}$$

where $\{\phi_i(x,y)\}$ is a set of known basis functions and $\{u_i\}$ is a set of constants to be determined. Since (24.3) is not, in general, the exact solution of (24.2) when we substitute (24.3) into (24.2) we get

$$-\alpha\nabla\cdot\left(\frac{\nabla\tilde{u}}{\|\nabla\tilde{u}\|_\beta}\right) + \mathscr{A}^*(\mathscr{A}\tilde{u} - z) = r(x,y) \tag{24.4}$$

where $r$ is a residual function. The Galerkin method now requires that the constants are chosen to make

$$\int_\Omega r\psi_i \, dx\, dy = 0 \tag{24.5}$$

where $\{\psi_1, \psi_2, \cdots, \psi_N\}$ is a set of trial functions. For simplicity, it is usual to take the trial functions to be the same as the basis functions as this often leads to systems where the coefficient matrix has useful properties such as symmetry and positive definiteness.

Letting $\psi_i = \phi_i$, taking the inner product of (24.4) with each trial function and noting the requirement (24.5) leads to

$$-\alpha\int_\Omega \left[\nabla\cdot\left(\frac{\nabla\tilde{u}}{\|\nabla\tilde{u}\|_\beta}\right)\phi_j + \mathscr{A}^*(\mathscr{A}\tilde{u} - z)\phi_j\right] dx\, dy = 0 \tag{24.6}$$

Apply the corollary to the divergence theorem

$$\int_\Omega (\psi\nabla\cdot\mathbf{F} + \mathbf{F}\cdot\nabla\psi)\, dx\, dy = \oint_\Gamma \psi\mathbf{F}\cdot\mathbf{n}\, dC$$

with $\mathbf{F} = \nabla\tilde{u}$ and $\psi = \phi_j$ to (24.6) to get

$$\alpha\int_\Omega \left[\left(\frac{\nabla\tilde{u}}{\|\nabla\tilde{u}\|_\beta}\right)\cdot\nabla\phi_j + \mathscr{A}^*(\mathscr{A}\tilde{u} - z)\phi_j\right] dx\, dy - \oint_\Gamma \left(\frac{\nabla\tilde{u}}{\|\nabla\tilde{u}\|_\beta}\right)\cdot\mathbf{n}\phi_j\, dC = 0.$$

Since $\frac{\partial\tilde{u}}{\partial\mathbf{n}} = 0$ on $\Gamma$ it follows that the integral around the boundary is zero. Hence we get (where, for simplicity, $\tilde{u}$ has been replaced by $u$)

$$\alpha\int_\Omega \left[\left(\frac{\nabla u}{\|\nabla u\|_\beta}\right)\cdot\nabla\phi_j + \mathscr{A}^*(\mathscr{A}u - z)\phi_j\right] dx\, dy = 0. \tag{24.7}$$

Consider the first term on the left of (24.7)

$$\alpha \int_\Omega \left[ \left( \frac{\nabla u}{\|\nabla u\|_\beta} \right) \cdot \nabla \phi_j \right] dx\, dy.$$

This can be written in matrix notation as $K(\mathbf{u})\mathbf{u}$, where $\mathbf{u} = [u_1, u_2, \ldots, u_N]^T$ where

$$K(\mathbf{u}) = \alpha \int_\Omega \left[ \left( \frac{1}{\|\nabla u\|_\beta} \right) \nabla \phi_i \cdot \nabla \phi_j \right] dx\, dy. \tag{24.8}$$

Now consider the second term on the left of (24.7)

$$\int_\Omega \mathscr{A}^*(\mathscr{A}u - z)\phi_i \, dx\, dy.$$

Let $w = \mathscr{A}^*(\mathscr{A}u - z)$ and approximate $w$ as

$$w = \sum_{i=1}^{N} w_i \phi_i$$

in which case the final term on the left of (24.7) becomes

$$\sum_{i=1}^{N} w_i \int_\Omega \phi_i \phi_j \, dx\, dy$$

which can be expressed in matrix notation as $M\mathbf{w}$ where $\mathbf{w}$ is the vector of nodal values of $w$ and

$$M_{ij} = \int_\Omega \phi_i \phi_j \, dx\, dy$$

Hence we can write the discrete version of (24.7) as

$$\alpha K(\mathbf{u})\mathbf{u} + MA^*(A\mathbf{u} - \mathbf{z}) = 0. \tag{24.9}$$

where $A$ denotes the matrix approximation to the integral operator $\mathscr{A}$. We note at this point that if the image contains $N$ pixels, then the size of all of these matrices will be $N \times N$.

We now have to consider the choice of the finite element basis functions $\phi_i(x, y)$. Since an image consists of data on a square grid, at first glance it would appear that bilinear quadrilateral elements are most suitable type of elements to use. However, if such elements are used, then the integrals of the functions which appear in (24.8) have to be evaluated numerically, which adds considerably to the cost of the method as this matrix need to be recalculated at each iteration of the solution process. However, if the square is split into two triangles and linear basis functions used in each, then the resulting integral in (24.8) can be evaluated analytically. Further, using these linear elements means that any given node is only connected to itself and

(at most) the eight which surround it which means any given row of the matrices $K$ and $M$ will contain at most nine non-zero entries, making them both very sparse matrices. For example, for an $n \times n$ image, the full matrix contains $n^4$ entries, but only at most $9n^2$ of these are non-zero. Thus even for moderate values of $n$, $K$ and $M$ are going to be very sparse matrices.

The resulting nonlinear equations can be solved using a simple fixed point iterative scheme obtained by rearranging (24.9) to give

$$(K(\mathbf{u}_j) + MA^*A)\,\mathbf{u}_{j+1} = MA^*\mathbf{z}$$

where the initial estimate of the solution is $\mathbf{u}_1 = \mathbf{z}$, and the solution process stops when

$$\|\mathbf{u}_{j+1} - \mathbf{u}_j\| < \delta$$

for some predetermined tolerance $\delta$. The final image obtained at the end of the solution process will be called the recovered image.

## 24.3 Discretization of the Blurring Operator

The blurring operator $A$ is a first kind Fredholm integral operator which can be expressed in the form

$$\mathscr{A}u = \int_{\Omega} k(|\mathbf{p} - \mathbf{q}|)u(\mathbf{q})\,d\mathbf{q} \tag{24.10}$$

where $k(|\mathbf{p} - \mathbf{q}|)$ is a known kernel function. For Gaussian blurring,

$$k(|\mathbf{p} - \mathbf{q}|) = \frac{1}{2\pi\sigma}\exp\left(-\frac{|\mathbf{p} - \mathbf{q}|^2}{2\sigma^2}\right) \tag{24.11}$$

where $\sigma$ is a constant called the blurring parameter.

In the work presented here, an approximation to the blurring operator is obtained using a piecewise constant approximation to the pixel intensities $u$. Whilst this is not strictly consistent with the finite element approximations used above, it does allow the calculation of the blurred intensities at each pixel which is all that is needed. Applying the collocation method yields a matrix approximation given by

$$A_{ij} = \int_{\Omega_j} k(|\mathbf{p}_i - \mathbf{q}|)\,d\mathbf{q} \tag{24.12}$$

where $\Omega_j$ is the square around the $j^{\text{th}}$ pixel. Depending on the kernel function, it may be possible to evaluate the integrals in (24.12) analytically but in general these integrals will have to be evaluated numerically.

Since the matrix $A$ is an approximation to an integral operator, it will be full and for an $n \times n$ image, $A$ will be an $n^2 \times n^2$ matrix. For all but the smallest of images, it is impractical to form and store the whole matrix. Fortunately, the matrix has some structure which we can exploit. It is easy to show that $A$ can be expressed in block form as

$$\begin{bmatrix} A_1 & A_2 & A_3 & \cdots & A_n \\ A_2 & A_1 & A_2 & \cdots & A_{n-1} \\ A_3 & A_2 & A_1 & \cdots & A_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_n & A_{n-1} & A_{n-2} & \cdots & A_1 \end{bmatrix} \tag{24.13}$$

where each block $A_j$ is an $n \times n$ matrix. Hence by exploiting this structure the storage requirement for the full matrix is reduced from $n^4$ to $n^3$. However, if the matrix is now extended by adding $n-2$ rows and columns as follows

$$\left[\begin{array}{ccccc|ccc} A_1 & A_2 & A_3 & \cdots & A_n & A_{n-1} & \cdots & A_2 \\ A_2 & A_1 & A_2 & \cdots & A_{n-1} & A_n & \cdots & A_3 \\ A_3 & A_2 & A_1 & \cdots & A_{n-2} & A_{n-2} & \cdots & A_4 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ A_n & A_{n-1} & A_{n-2} & \cdots & A_1 & A_2 & \cdots & A_{n-1} \\ \hline A_{n-1} & A_n & A_{n-1} & \cdots & A_2 & A_1 & \cdots & A_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ A_3 & A_4 & A_5 & \cdots & A_{n-2} & A_{n-3} & \cdots & A_2 \\ A_2 & A_3 & A_4 & \cdots & A_{n-1} & A_{n-2} & \cdots & A_1 \end{array}\right] \tag{24.14}$$

the resulting matrix (24.14) is now block circulant. That is, each block column can be formed by shifting the previous block column down one block, and block at the bottom of the previous column becomes the top block in the current column. We also note that this matrix also has block symmetry. Further, each block has the same internal structure as the whole matrix. That is, the internal structure of each $A_j$ is in the form of (24.13) where each entry is a scalar rather than a matrix. If each block is extended in a similar manner by adding $n-2$ rows and columns to each block, then the entire matrix becomes circulant and so only requires $4(n-1)^2$ entries to be stored. In order to find the matrix-vector product $A\mathbf{u}$, the vector $\mathbf{u}$ needs to be extended by inserting extra rows corresponding to the extra columns of $A$. Provided these extra rows contain zero, the correct values of $\mathbf{v} = A\mathbf{u}$ can be obtained from the appropriate rows of $\mathbf{v}$ whilst the values in the extra rows of $\mathbf{v}$ are simply discarded. These matrix-vector products can be found efficiently using a fast Fourier transform method as described in [Da79].

## 24.4 Numerical Examples

In this section, we present the results of applying the finite element method to some typical examples. In order to measure the difference between the original image, the contaminated image and the recovered image, we shall use the root mean square (RMS) error, defined by

$$E(u,z) = \sqrt{\frac{\sum_{i=1}^{N}(u_i - z_i)}{N}}$$

to measure the difference between images with pixel intensities $u$ and $z$, respectively. Here $N$ is the total number of pixels in each image.

Figure 24.1 shows a perfect test image ($u$) on the left, and the contaminated image ($z$) obtained by blurring the perfect image using Gaussian blurring with parameter $\sigma = 3$ is on the right. For this example $E(u,z) = 0.07477$. Figure 24.2 shows the recovered image using $\alpha = 10^{-4}$ and different values of $\beta$. The images on the left is the image recovered using $\beta = 10^{-3}$ and the image on the right in the image recovered using $\beta = 10^{-6}$. The corresponding RMS errors in the recovered images when compared to the original image are 0.03292 and 0.02979, respectively. These results show that the choice of $\beta$ does have an effect on the accuracy of the recovered image, but more work is necessary to investigate this further. Also, the effect of the parameter $\alpha$ has not been studied, so further work is needed here to investigate its effect. Figure 24.3 shows the results for a typical image that arises in medical imaging. The image top left is the original uncontaminated image and the image top right is the one obtained when the image is contaminated with Gaussian blurring with parameter $\sigma = 3$. The RMS error in the contaminated image is 0.05292. The recovered image using $\alpha = 10^{-4}$ and $\beta = 10^{-3}$ is shown bottom left, which the recovered image shown bottom right uses $\alpha = 10^{-4}$ and $\beta = 10^{-6}$. The RMS errors in these recovered images are 0.02819 and 0.02889, respectively.

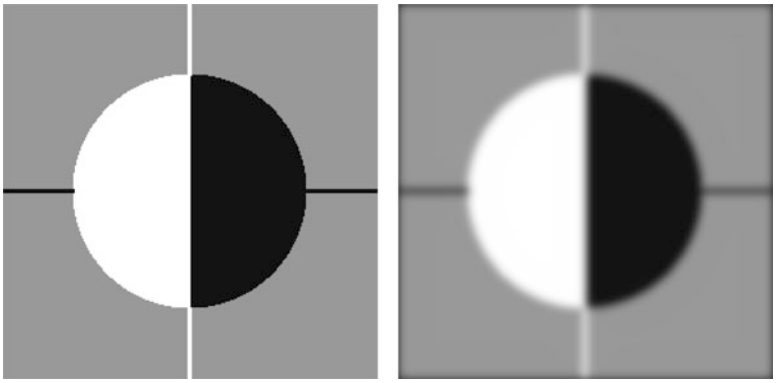

**Fig. 24.1** The original (left) and blurred (right) images.

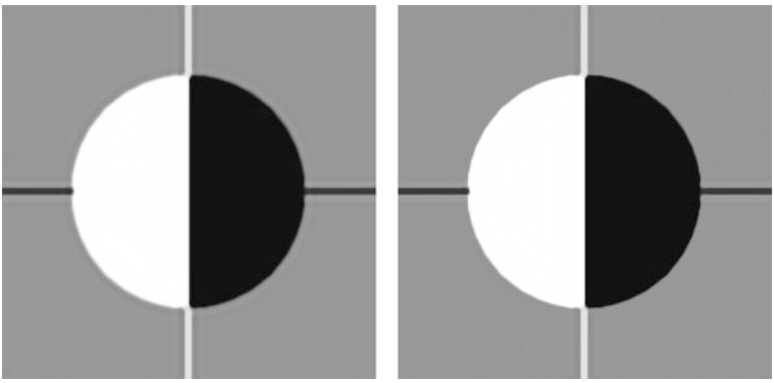

**Fig. 24.2** Recovered image for $\beta = 10^{-3}$ (left) and $\beta = 10^{-6}$ (right).

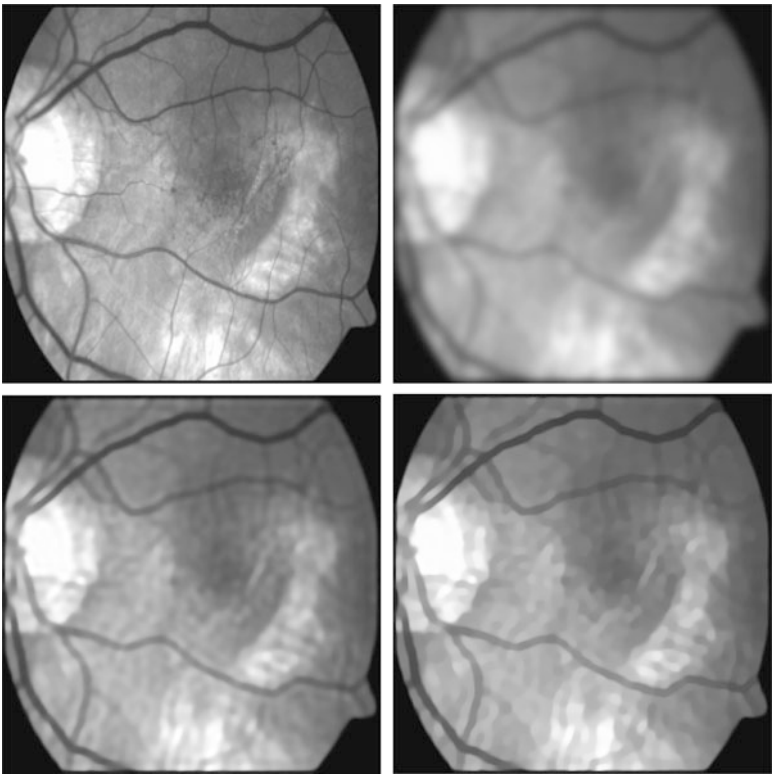

**Fig. 24.3** The original image (top left), blurred image (top right), and recovered images using $\beta = 10^{-3}$ (bottom left) and $\beta = 10^{-6}$ (bottom right) for a typical example used in medical imaging.

## 24.5 Conclusions

This chapter has shown that the finite element method can be used to solve the nonlinear partial differential equation which arises when using the total variation method to remove the noise and/or blurring from a digital image. Although it has not been reported here, the method can be readily adapted to deal with images that are not rectangular, such as when part of the image is masked. Further work needs to be carried out to investigate the influence of the various parameters (such as $\alpha$ and $\beta$) on the accuracy of the final recovered image.

## References

[BrCh10] Brito-Loeza C., Chen K.: Multigrid algorithm for high-order denoising. SIAM J. Im. Sci. **3**(3), 363–389. (2010).

[ChCh06] Chan T.F., Chen K.: On a nonlinear multigrid algorithm with primal relaxation for the image total variation minimisation. Numer Al **41**(4), 387–411. (2006).

[ChPiZa13] Chen, K., Piccolomini E.L. Zama F. An automatic regularization parameter selection algorithm in the total variation model for image deblurring. Numerical Algorithms 1–20 (2013)

[Da79] Davis, P.J. Circulant Matrices. John Wiley & Sons, New York. (1997)

[VoOm96] Vogel C.R., Oman M.E.: Iterative methods for total variational denoising. SIAM J. Sci. Comp. **17**(1), 227–238. (1996)

[YaChYu12] Yang F., Chen K., Yu B.: Homotopy method for a mean curvature based denoising model. App. Numer. Math. **63**, 185–200 (2012).

# Chapter 25
# Mathematical Modeling to Quantify the Pharmacokinetic Process of [18F]2-fluor-2deoxy-D-glucose (FDG)

**E.B. Hauser, G.T. Venturin, S. Greggio, and J.C. da Costa**

## 25.1 Introduction

The main objective of this study is to quantify pharmacokinetic processes - such as absorption, distribution and elimination - of [18F]2-fluor-2deoxy-D-glucose (18F - FDG), by using the Laplace transformation method.

18F - FDG is a glucose analog, labeled with the positron emitter 18F, and is used as a radiopharmaceutical to investigate tissue metabolism in positron emission tomography (PET) studies. PET is a functional imaging technology that allows to study physiological and molecular changes through the administration of radiopharmaceutical tracers into living systems.

When a radiolabeled drug, such as 18F -FDG, is administered intravenously, the absorption is complete, the compound becomes available in the bloodstream to be distributed throughout the whole body in all tissues and fluids, and after that it is eliminated.

Mathematical modeling seeks to describe the processes of distribution and elimination through compartments, where distinct pools of the tracer are assigned to different compartments.

A compartmental model is an important kinetic modeling technique used for quantification in PET imaging. It is described by a system of differential equations, where each equation represents the sum of all the transfer rates to and from a specific compartment. Rate transferring from one compartment to another is proportional to concentration in the compartment of origin. We denote

E.B. Hauser (✉) • G.T. Venturin • S. Greggio • J.C. da Costa
Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: eliete@pucrs.br; gianina.venturin@pucrs.br; samuel.greggio@pucrs.br; jcc@pucrs.br



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_25

$$\frac{d}{dt}C_i(t) = \sum_{j=1, j\neq i}^{N} \left[K_{ij}C_j(t) - K_{ji}C_i(t)\right]$$

where $C_i(t)$ is the concentration of radioactive tracer in compartment $i$, $N$ is the number of sections of the model, and $K_{ij}$ is the rate constant for transfer from compartment $j$ to compartment $i$.

Physiological or biochemical systems are described using models of compartments in which a tracer is distributed between compartments, which represent spatial location or chemical state.

In this chapter, a two-tissue irreversible compartment model is used for kinetic modeling of 18F -FDG uptake by using the Laplace transform method.

The irreversible two-compartment model for 18F -FDG is used for description of this tracer, which is first entering a free compartment, $C1$, and is then metabolized irreversibly in the second compartment $C2$.

In order to determine the parameters of the model, information on the tracer delivery is needed in the form of an input function that represents the time-course of tracer concentration in the arterial blood or plasma [CuJo93, Za06, Kh11].

Quantitative PET studies often require a measure of the input function [La05, VrGe09, KiPi11, Za06, Kuba91].

We estimated the arterial input function in two stages and applied the Levenberg–Marquardt method to solve nonlinear regressions [BaWa88].

The transport of FDG across the arterial blood is very fast in the first ten minutes and then slowly decreases. The main contribution of the present study is that we used the We Heaviside function to represent this compartment modeling for 18F -FDG. We applied the Laplace transform and obtained the analytical solution for the two-tissue irreversible compartment model. The only approach is to determinate the arterial input function.

## 25.2 The Proposed Method for Two-Tissue Irreversible Compartment Model

The glucose metabolism is investigated using the irreversible FDG-model, which was developed for description of the tracer $^{18}F-$2-fluor-2deoxy-D-glucose (see [KiPi11]). The irreversible two-compartment model for FDG is illustrated in Figure 25.1 and is used for description of the tracer, which first enter a free compartment, $C1$, and is then metabolized irreversibly in the second compartment $C2$.

The mathematical model for the problem is expressed by the system of two differential equations

$$\begin{aligned} \frac{d}{dt}C_1(t) &= K1\, C_a(t) - (k2 + k3)\, C_1(t) \\ \frac{d}{dt}C_2(t) &= k3\, C_1(t) \end{aligned} \tag{25.1}$$

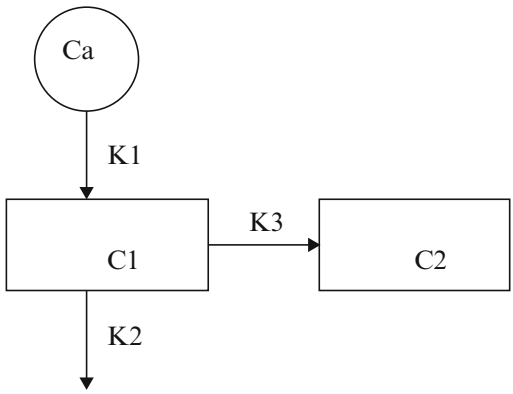


**Fig. 25.1** FDG model.

The tracer concentration in arterial blood $Ca(t)$, the input function depends on the time $t$, is a known quantity.

We apply the Laplace transformation with respect to $t$ in (25.1), denoting

$$\pounds\{C_i(t)\} = \overline{C}_i(s) = \int_0^\infty e^{-st} C_i(t)\, dt$$

and

$$\pounds\left\{\frac{dC_k(t)}{dt}\right\} = s\,\overline{C}_i(s) - C_i(0).$$

We obtain, with $C_1(0) = 0$ and $C_2(0) = 0$ , an algebraic system:

$$\begin{aligned} (s + k2 + k3)\,\overline{C}_1(s) &= K1\,\overline{C}_a(s) \\ -k3\,\overline{C}_1(s) + s\,\overline{C}_2(s) &= 0 \end{aligned} \quad . \tag{25.2}$$

Now we apply the inverse Laplace transformation to equation (25.2):

$$C_i(t) = \pounds^{-1}\left\{\overline{C}_i(s)\right\}.$$

Therefore, we obtain

$$C_1(t) = \pounds^{-1}\left\{\frac{K1\,\overline{C}_a(s)}{(s + k2 + k3)}\right\}, \quad C_2(t) = \pounds^{-1}\left\{\frac{k3\,\overline{C}_1(s)}{s}\right\}. \tag{25.3}$$

Then

$$\begin{aligned} C_1(t) &= K1\,\pounds^{-1}\left\{\frac{1}{(s + k2 + k3)}\right\} * \pounds^{-1}\left\{\overline{C}_a(s)\right\} \\ C_2(t) &= k3 * \pounds^{-1}\left\{\overline{C}_1(s)\right\}, \end{aligned} \tag{25.4}$$

where $*$ denotes the convolution operation.

The representation (25.3) implies that

$$C_1(t) = K1\, e^{-(k2+k3)\,t} * C_a(t) = K1 \int_0^t e^{-(k2+k3)\,(t-u)}\, C_a(u)\, du$$
$$C_2(t) = k3 * C_1(t) = k3 \int_0^t C_1(u)\, du\,. \tag{25.5}$$

The analytic solution of the irreversible two-compartment model for FDG (25.1) is (25.5).

## 25.3 Arterial Input Function

The transport of FDG across the arterial blood is very fast in the first minutes and then slowly decreases. For this reason we have chosen to estimate the arterial input function in two stages.

We defined the arterial input function for the fast stage

$$C_f(t), t \in (a, b)$$

and the arterial input function for the slow phase

$$C_s(t), t \in (b, c)\,.$$

We introduce the Heaviside step function (the unit step function)

$$H(t-a) = \begin{cases} 0, & t < a, \\ 1, & t \geq a, \end{cases}$$

$$H(t-a) - H(t-b) = \begin{cases} 1, & a \leq t < b, \\ 0, & t < a \text{ and } t \geq b, \end{cases}.$$

Our goal is to construct a piecewise input function as

$$C_a(t) = [H(t-a) - H(T-b)]\, C_f(t) + [H(t-b) - H(T-c)]\, C_s(t)\,. \tag{25.6}$$

## 25.4 Illustrative Example

We used the experimental results presented by [KiPi11]. Considering $k1 = 0.4$, $k2 = 0.2$, and $k3 = 0.05$, we obtain the analytical solution for the system (2).

We write the input function as

$$C_a(t) = [H(t) - H(T-10)]\, C_f(t) + [H(t-10) - H(T-60)]\, C_s(t)\,. \tag{25.7}$$

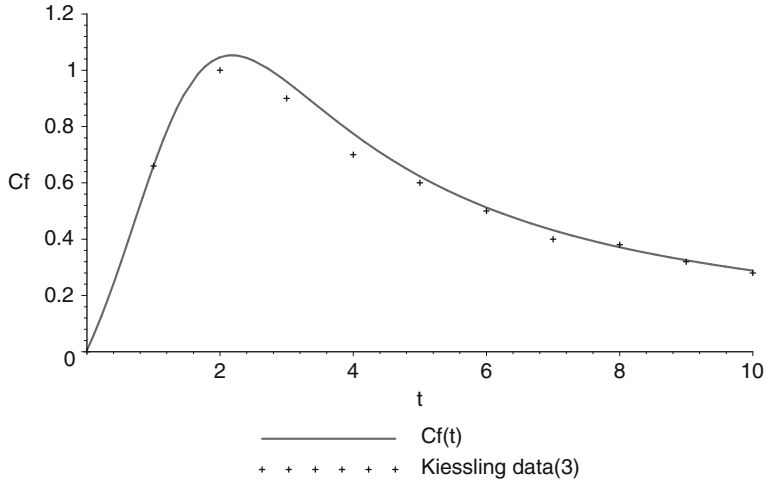


**Fig. 25.2** $C_f(t)$ : Fast phase input function.

After some algebraic manipulations, we get that the arterial input function for the fast stage, $C_f(t)$ , obtained using the Levenberg–Marquardt method, as

$$C_f(t) = \frac{0.51t + 0.005}{0.21t^2 - 0.43t + 1}, \tag{25.8}$$

represented in Fig. 25.2.

And in (25.7), the arterial input function for the slow stage, $C_s(t)$, obtained in a similar fashion is

$$C_S(t) = \frac{(4.96 \times 10^9)\,t + 27.86}{(4.36 \times 10^7)\,t^2 + (4.29 \times 10^{10})\,t + 1}, \tag{25.9}$$

illustrated in Fig. 25.3.

In Fig. 25.4, we present the piecewise input function (25.7).

Then we use the Laplace transform method described in Section 25.2. In Table 25.1 we summarize the properties of the Laplace transform needed to solve the system of two ordinary differential equations of first order (25.1).

In Table 25.1, the special functions are defined by

- Exponential integral:

$$Ei(t) = \int_t^{\infty} \frac{e^{-u}}{u}\, du,$$

- Sine integral:

$$Si(t) = \int_0^t \frac{\sin u}{u}\, du$$

and

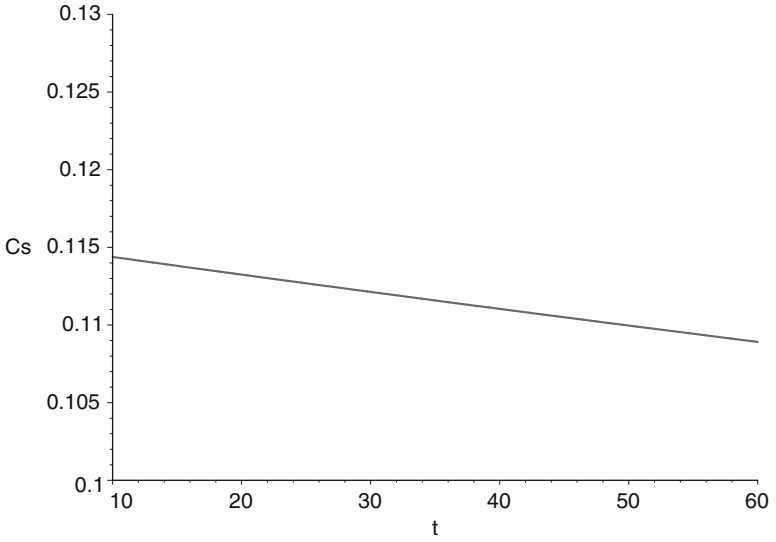


**Fig. 25.3** $C_s(t)$ : Slow phase input function.

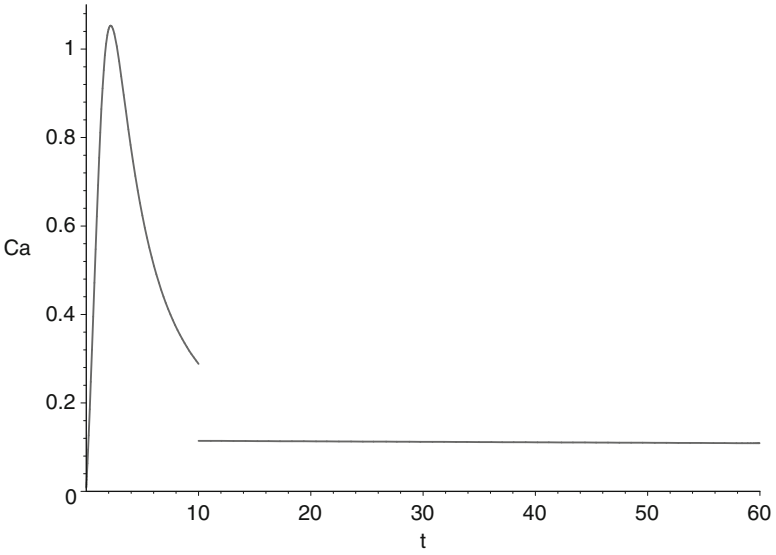


**Fig. 25.4** $C_a(t)$ : Arterial input function.

**Table 25.1** Laplace Transform

| $f(t)$ | $\pounds\{f(t)\} = F(s) = \int_0^\infty e^{-st} f(t)\,dt$ |
|---|---|
| $e^{at} f(t)$ | $F(s-a)$ |
| $f(t-a)\,H(t-a)$ | $e^{-as} F(s)$ |
| $f'(t)$ | $sF(s) - f(0)$ |
| $f(t) * g(t) = \int_0^t f(u)\,g(t-u)\,du$ | $F(s)\,G(s)$ |
| $e^{at}$ | $\frac{1}{s-a}$ |
| $\frac{1}{t+a}$ | $e^{as} Ei(as)$ |
| $\frac{1}{t^2+a^2}$ | $\frac{1}{a}[cos(as)\{\frac{\pi}{2} - Si(as)\} - sen(as)Ci(as)]$ |
| $\frac{t}{t^2+a^2}$ | $sen(as)\{\frac{\pi}{2} - Si(as)\} + cos(as)Ci(as)]$ |

- Cosine integral:

$$Ci(t) = \int_t^\infty \frac{\cos u}{u}\,du.$$

The input function $C_a(t)$ and the response curves, $C_1(t)$ and $C_2(t)$, with transport constants $K1 = 0.4$, $k2 = 0.2$, $k3 = 0.05$, are represented in Figure 25.5. $C_1(t)$ and $C_2(t)$ are the analytical solution for two-tissue irreversible compartment model (25.1). The results are very similar to those obtained in experiment described in [KiPi11]. The only approximation used in this work was the arterial input function $C_a(t)$, expressed by de (25.7), (25.8) and (25.9).

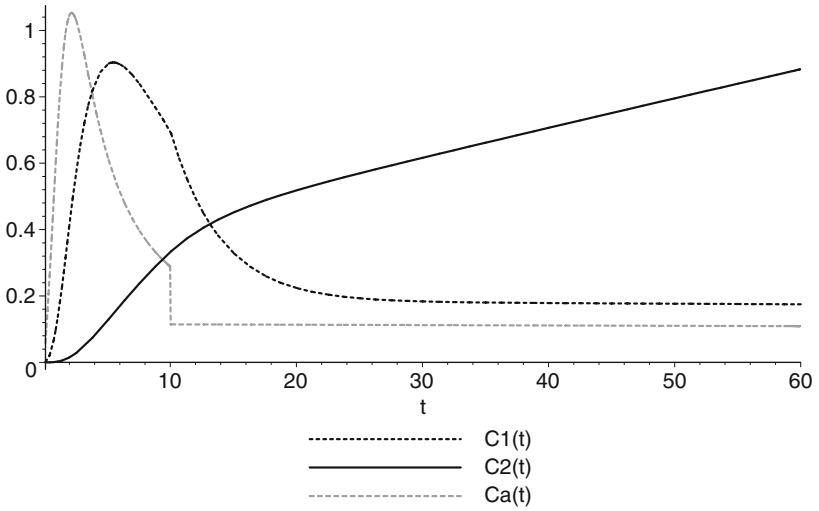


**Fig. 25.5** Solution of the FDG model.

## References

[BaWa88] Bates, D.M., Watts, D.G.: Nonlinear Regression and Its Applications. Wiley, New York (1988)

[CuJo93] Cunningham, V.J., Jones, T.: Spectral analysis of dynamic PET studies. J Cereb Blood Flow Metab. **13**, 15–23(1993).

[Kh11] Khalil, M.M.: Basic Sciences of Nuclear Medicine. Springer, Berlin (2011).

[KiPi11] Kiessling, F., Pichler, B.J.: Small Animal Imaging. Springer, Berlin (2011).

[Kuba91] Kuikka, J.T. et al.: Mathematica Modelling in Nuclear Medicine. European Journal of Nuclear Medicine. **18**, 351–362 (1991).

[La05] Laforest, R. et al.: Measurement of input functions in rodents: challenges and solutions. Nuclear Medicine and Biology. **32**, 679–685 (2005)

[Za06] Zaidi, H.: Quantitative Analysis in Nuclear Medicine Imaging, Springer, New York (2006).

[VrGe09] Vriens, D. et al.: A Curve-Fitting Approach to Estimate the Arterial Plasma Input Function for the Assessment Of Glucose Metabolic Rate and Response to Treatment. The Journal of Nuclear Medicine. **50–12**, 1933–1939 (2009)

# Chapter 26
# Multi-Particle Collision Algorithm for Solving an Inverse Radiative Problem

**R. Hernández Torres, E.F.P. Luz, and H.F. Campos Velho**

## 26.1 Introduction

Optimization is the area of the Applied Mathematics that studies the theory and techniques to finding the best available values to optimize (minimize or maximize) some objective function, also called error function or cost function.

Many real problems in science and engineering involves optimization in some way. A class of inverse problems can also be formulated as an optimization problem. The forward problem is characterized for producing a response from input parameters. From measured or desired response, the procedure to identify the input parameters is called inverse problem.

Inverse radiative transfer problem has many relevant applications in science and industry. Some examples are computerized tomography, optical reconstruction in spectroscopy, radiative property estimation, heat conduction, climate modeling, hydrologic optics, and space science [StEtAl10]. These inverse problems can be formulated implicitly and solved as an optimization problem. Inverse problems are typically ill-posed problems. To deal with them, a regularization term is added to the regular objective function.

Stochastic optimization has become an important tool to solve multi-modal cost functions. In addition, sometimes it is hard to compute the gradient of a cost function, or even there is no derivative of such function. Some stochastic methods do not need gradient information or other internal information of the process/system to be applied. Stochastic methods use a random process to generate new solutions, and facilitate the exploration (global search) in the search space, at the same time that exploitation (local search) is made by some methods. The entire search space

R. Hernández Torres (✉) • E.F.P. Luz • H.F. Campos Velho
National Institute for Space Research (INPE), Av. dos Astronautas, 1758 São José dos Campos, SP, Brazil
e-mail: reynier.torres@inpe.br; eduardo.luz@lac.inpe.br; haroldo@lac.inpe.br



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_26

can be visited by generating new randomly candidate solutions, while an intense search is made in neighborhood of this candidate solution – this searching can be applied for some selected candidates.

Meta-heuristic algorithms can be bio-inspired stochastic methods: evolution of species, social behavior of animals (ants, fireflies, bees, etc.), or developed based on physics phenomena (simulated annealing (SA), Particle Collision Algorithm (PCA) [SaOl05]). In particular, the PCA was inspired by the physics of a nuclear particle traveling inside of a nuclear reactor, where scattering and absorption are the main phenomena in the process. PCA is an individual method, where a single particle explores and exploits the search space.

A new version of PCA, named Multi-Particle Collision Algorithm (MPCA) [LuBeVe08], the search space is explored for several particles at the same time. The particles work in a cooperative behavior, and the strategy can easily be implemented in a parallel machine. The MPCA will be used to address an inverse problem on radiative transfer process.

In the next section, the direct problem will be enunciated. The inverse radiative transfer problem is formulated in a later section of the direct problem. The canonical MPCA will be presented in the following section, as well as an MPCA version with pre-regularization. The final sections are results and conclusions.

## 26.2 Forward Problem: Solving the Radiative Transfer Problem

Figure 26.1 represents a participating medium and transparent boundary surfaces [StEtAl10]. The medium is considered as one-dimensional, heterogeneous, gray, with optical thickness $\tau_0$. Radiation generated from external sources with intensities $A_1$ and $A_2$, respectively, arrives on the boundaries at $\tau = 0$ and $\tau = \tau_0$ – the boundary surfaces also diffusely reflect the radiation coming from inside.

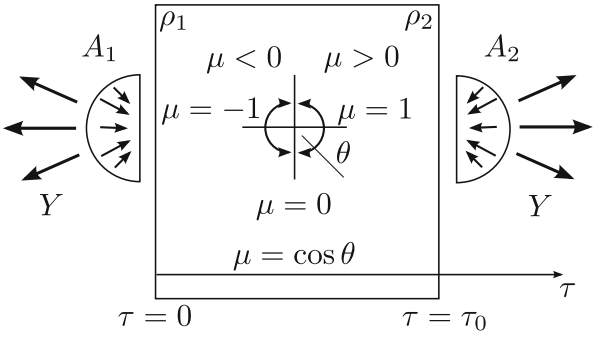


**Fig. 26.1** Plane parallel geometry with external incident radiation.

The mathematical model of the radiative transfer problem, i.e. the radiation interaction with the medium, for constant radiative properties, isotropic scattering, and azimuthal symmetry, is given by the linear version of the Boltzmann equation, and written in the dimensionless form as

$$\mu \frac{\partial I(\tau,\mu)}{\partial \tau} + I(\tau,\mu) = \frac{\omega(\tau)}{2} \int_{-1}^{1} I(\tau,\mu')d\mu', 0 < \tau < \tau_0 \tag{26.1}$$

with the boundary conditions

$$I(0,\mu) = A_1(\mu) + 2\rho_1 \int_0^1 I(0,-\mu')\mu' d\mu', \mu > 0 \tag{26.2}$$

$$I(\tau_0,-\mu) = A_2(\mu) + 2\rho_2 \int_0^1 I(\tau_0,-\mu')\mu' d\mu', \mu < 0 \tag{26.3}$$

where $I$ is the radiation intensity, $\tau \equiv \int_0^{\tau_0} a(x)dx$ the optical variable, $\mu$ cosine of the polar angle, $\rho_1$ and $\rho_2$ diffuse reflectivities at the inner part of the boundary surfaces at $\tau = 0$ and $\tau = \tau_0$, and $\omega(\tau) = b(\tau)/[a(\tau) + b(\tau)]$ is the single albedo ($a(z)$ and $b(z)$ are absorption and scattering coefficients, respectively), expressed in the polynomial form

$$\omega(\tau) = \sum_{k=0}^{K} D_k \tau^k \,. \tag{26.4}$$

The direct problem described in (26.1–26.4) may be solved by using Chandrasekhar's discrete ordinate method, where the scattering angle $\mu$ is taken on discrete directions – see Figure 26.2. The integral term on the right-hand side of Eq. (26.2) is replaced by a Gaussian quadrature. A finite difference approximation is used for the terms on the left-hand side of Eq. (26.2), and by performing forward (from $\tau = 0$ to $\tau = \tau_0$) and backward (from $\tau = \tau_0$ to $\tau = 0$) sweeps, $I(\tau,\mu)$ is determined for all spatial and angular nodes of the discretized computational domain.

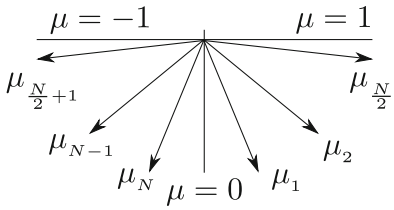


**Fig. 26.2** Discretization of the polar angle domain.

## 26.3 Inversion Formulated as an Optimization Problem

The inverse problem consists of estimating the radiative properties of the medium from the emerging radiation, minimizing the objective function:

$$Q(\vec{Z}) = \sum_{i=1}^{N_d} \left[ I_i^{\text{mod}}(\vec{Z}) - I_i^{\text{exp}} \right]^2 \tag{26.5}$$

where $I_i^{\text{mod}}$ and $I_i^{\text{exp}}$ are, respectively, the calculated and measured values of the radiation intensity.

Half of the data is acquired at the boundary $\tau = 0$ and half at $\tau = \tau_0$ by using external detectors, as represented in Figure 26.3. The space dependent albedo $\vec{Z} = \{\omega_1, \omega_2, \cdots, \omega_{N_u}\}$, with $N_u$ discrete values, is unknown.

Considering that only the left boundary ($\tau = 0$) is subjected to the incidence of isotropic radiation originated at an external source while there is no radiation coming into the medium through the boundary at $\tau = \tau_0$. Also considering the diffuse reflectivities $\rho_1$ and $\rho_2$ as null, the boundary conditions become

$$I(0,\mu) = A_1 \text{ if } \mu > 0, \quad I(\tau_0, -\mu) = 0 \text{ if } \mu < 0.$$

The integro-differential operator associated with the Boltzmann equation is compact. Therefore, the inverse operator does not have a formal inverse. This ill-posedness can be treated by a regularization process. A practical implementation is to use the Tikhonov regularization of order 2. Then, the objective function becomes

$$Q^*(\vec{Z}) = \sum_{i=1}^{N_d} \left[ I_i^{\text{mod}}(\vec{Z}) - I_i^{\text{exp}} \right]^2 + \alpha \Omega(\vec{Z}) \tag{26.6}$$

where $\alpha$ is the regularization parameter and $\Omega(\cdot)$ is the regularization operator. The smoother operator is expressed by the second-order Tikhonov regularization

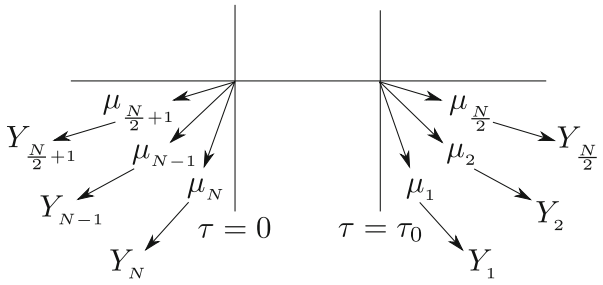


**Fig. 26.3** Schematic representation for experimental data $I_i^{\text{exp}}$ $(i = 1, 2, \cdots, N/2)$ acquired at $\tau = \tau_0$, and $(i = N/2+1, \cdots, N-1, N)$ acquired at $\tau = 0$.

$$\Omega(\vec{Z}) = \sum_{i=1}^{N_u - 1} \left[\omega_{i+1} - 2\omega_i + \omega_{i-1}\right]^2 . \tag{26.7}$$

The parameter $\alpha$ is hard to be determined. Depending on the adopted criterion, many executions may be required for the inverse solver to calculate $\alpha$. A scheme of pre-regularization is implemented, where the regularization parameter is not necessary. The scheme selects the smoothest candidate solutions (albedo, in our case) from a population. The pre-regularization approach was successfully used by solving an inverse hydrological optics Ant Colony Optimization (ACO).

The pre-regularization scheme has some computational advantages: saves extra evaluations to obtain the value of the $\alpha$ parameter, and it is not necessary to solve the forward problem for those no smooth candidates.

## 26.4 Multi-Particle Collision Algorithm (MPCA)

The MPCA [LuBeVe08] is based on the canonical Particle Collision Algorithm (PCA), introduced by Sacco[SaOl05]. The MPCA pseudo-algorithm is represented in Algorithm–1.

The PCA is inspired by the physics of nuclear particle traveling inside of a nuclear reactor, particularly the scattering and the absorption phenomena. In this algorithm, the Perturbation function (see algorithm 2) performs a random variation of the solution within a defined range, allowing the *visit* on different regions in the search space, while the Exploration function (see algorithm 3) performs a local search (applying a small perturbation (see algorithm 4) on the candidate solution). When the new candidate solution has a worse performance (the cost function is enhanced), the Scattering process (see algorithm 5) is activated, in which the particle (the candidate solution) is replaced by a new random solution, according a computed probability from: $[1 - \text{cost\_function}/(\text{best solution})]$ [LuBeVe08, SaOl05].

For the MPCA, more than one particle is applied to explore the search space in a cooperative way. A blackboard strategy is implemented, where the best particle is over-copied for all other particles after some iterations. The process is re-started at every $N_{\text{blackboard}}$ iterations (the textitblackboard cycle), as seen in Algorithm–1.

A parallel version of the MPCA is implemented in FORTRAN 95, using MPI libraries in a multiprocessor architecture with distributed memory machine.

### *26.4.1 MPCA with Pre-regularization*

The use of the pre-regularization in the MPCA implies that a large set of solutions ($T \times N_{particles}$) will be generated and evaluated according the regularization norm – see Eq. 26.7. Therefore, a small subset (the $N_{particles}$ smoothest particles) will be

**Algorithm 1** MPCA (*IL* and *SL* are the lower and upper limits for the local search perturbation intensity; *LB* and *UB* are the minimum and maximum value for each variable; *currentP* is the current particle; *newP* is the new particle; *bestP* is the best particle).

```
Global variables LB, UB, IL, SL
for i ← 1, N_processors do
    for j ← 1, N_particles do
        currentP_{i,j} = RANDOMSOLUTION
iteration = 0
while iteration < N_maxIterations or other stopping criteria not yet met do
    for i ← 1, N_processors do
        if iteration % N_blackboard == 0 then
            bestP_i = UPDATEBLACKBOARD(currentP_{i,−})
        for j ← 1, N_particles do
            newP_{i,j} = PERTURBATION(currentP.Solution_{i,j})
            if newP_{i,j}.Fitness < currentP_{i,j}.Fitness then
                currentP_{i,j} = newP_{i,j}
                currentP_{i,j} = EXPLORATION(currentP_{i,j})
            else
                currentP_{i,j} = SCATTERING(currentP_{i,j}, newP_{i,j}, bestP_i)
            if currentP_{i,j}.Fitness < bestP_i.Fitness then
                bestP_i = currentP_{i,j}
    iteration = iteration + 1
for i ← 1, N_processors do
    bestP_i = UPDATEBLACKBOARD(currentP_{i,−})
return bestP_1
```

**Algorithm 2** Perturbation Function (*P* is the obtained particle, *currentP* is the current particle, *bestP* is the best particle)

```
function PERTURBATION(currentP)
    for d ← 1, N_dimension do
        R = rand(0,1)
        P.Solution_d = currentP.Solution_d + ((UB_d − currentP.Solution_d) * R) − ((P.Solution_d −
LB_d) * (1 − R))
        if P.Solution_d > U then
            P.Solution_d = U
        else if P.Solution_d < L then
            P.Solution_d = L
    P.Fitness = FITNESS(P.Solution)
    return P
```

**Algorithm 3** Exploration Function(*currentP* is the current particle, *newP* is the new particle)

```
function EXPLORATION(currentP)
    for n ← 1, N_maxInternalIterations do
        newP = SMALLPERTURBATION(currentP)
        if newP.Fitness < currentP.Fitness then
            currentP = newP
    return currentP
```

**Algorithm 4** Small Perturbation Function (*P* is the obtained particle, *currentP* is the current particle, *bestP* is the best particle)

```
function SMALLPERTURBATION(currentP)
    for d ← 1, N_dimension do
        U = currentP.Solution_d * rand(1, SL)
        L = currentP.Solution_d * rand(IL, 1)
        R = rand(0, 1)
        if U > UB_d then
            U = UB_d
        if L < LB_d then
            L = LB_d
        P.Solution_d = currentP.Solution_d + ((U − currentP.Solution_d) * R) −
((currentP.Solution_d − L) * (1 − R))
    P.Fitness = FITNESS(P.Solution)
    return P
```

**Algorithm 5** Scattering Function (*P* is the obtained particle, *currentP* is the current particle, *newP* is the new particle, *bestP* is the best particle)

```
function SCATTERING(currentP, newP, bestP)
    p_scattering = 1−(bestP.Fitness/newP.Fitness)
    if p_scattering > rand(0, 1) then
        P = RANDOMSOLUTION
    else
        P = EXPLORATION(currentP)
    return P
```

used for ranking by objective function (Eq. 26.5). This procedure is executed at the moment of generating a new random solution (i.e., creating the initial population after Scattering action).

## 26.5 Experimental Results

For the numerical experiments, machine with 8 processors was used. One of them is the master processor controlling the updating of the blackboard. Each processor works with a single particle ($N_{particles} = 1$), resulting eight particles in the population

for each iteration. The control parameters for the MPCA are $IL = 0.85$ and $SL = 1.15$. In our application, the parameters $LB = 0.0$ and $UB = 1.0$ are assumed for all dimensions. Maximum number of function evaluation ($N_{maxIterations}$) is set to 10000, and the number of evaluation for the internal loop $N_{maxInternalIterations} = 200$. The blackboard updating occurs each 400 function evaluations ($N_{\text{blackboard}} = 400$).

The method is tested using synthetic measurements, where the exit radiation intensities were generated (*in silico*) using the exact values of the radiative properties, and some noise was added (2% and 5%, respectively). Considering the following parameters [StEtAl10]: $A_1 = 1.0$ and $A_2 = 0.0$. The albedo is given as a polynomial with the coefficients $D_0 = 0.2$, $D_1 = 0.2$, and $D_2 = 0.6$, with $N_d = N_u = 10$.

The results were obtained taking an average from 25 realizations, each one with a different random seed. Three cases will be analyzed: noiseless experimental data, and data with 2% and 5% of noise level. The quality of the results is given by the value of the objective function (residue: Eq. 26.5), and the sum of the quadratic error between the exact and the estimated values for the albedo is given by Eq. 26.8. Table 26.1 shows the mean values for all the cases in 25 runs of the algorithm.

$$d^2 = \sum_{i=1}^{N_d} \left( \vec{Z}_i^{\text{exact}} - \vec{Z}_i^{\text{estimated}} \right)^2 \tag{26.8}$$

Each experiment expends approximately 10 seconds evaluating a mean of 10000 times the objective function, each run. The final results have the same quality that those described in [StEtAl10], using the Ant Colony Optimization, but the results obtained with MPCA were much faster: from 10 up to 20 faster, depending on the execution seed. Figures 26.4–26.6 show the average of albedo profiles obtained during 25 runs for all cases.

**Table 26.1** Statistical results for 25 runs of MPCA with regularization.

| | Noise level (%) | Initial Guess | Final Result |
|---|---|---|---|
| Residue | 0 | $3.8559 \times 10^{-1}$ | $3.4563 \times 10^{-6}$ |
| Minimum Residue | | $5.4780 \times 10^{-3}$ | $6.8150 \times 10^{-7}$ |
| Maximum Residue | | $2.1050 \times 10^{0}$ | $5.9920 \times 10^{-6}$ |
| Error | | $9.6275 \times 10^{-2}$ | $4.8789 \times 10^{-3}$ |
| Residue | 2 | $2.2216 \times 10^{-1}$ | $1.7716 \times 10^{-5}$ |
| Minimum Residue | | $1.6470 \times 10^{-2}$ | $1.4170 \times 10^{-5}$ |
| Maximum Residue | | $1.6210 \times 10^{0}$ | $2.1230 \times 10^{-5}$ |
| Error | | $5.2183 \times 10^{-2}$ | $2.3857 \times 10^{-2}$ |
| Mean Residue | 5 | $2.6367 \times 10^{-1}$ | $4.4552 \times 10^{-5}$ |
| Minimum Residue | | $3{,}3150 \times 10^{-3}$ | $4{,}2940 \times 10^{-5}$ |
| Maximum Residue | | $1{,}1540 \times 10^{0}$ | $4{,}7320 \times 10^{-5}$ |
| Error | | $1.0974 \times 10^{-1}$ | $2.2869 \times 10^{-3}$ |

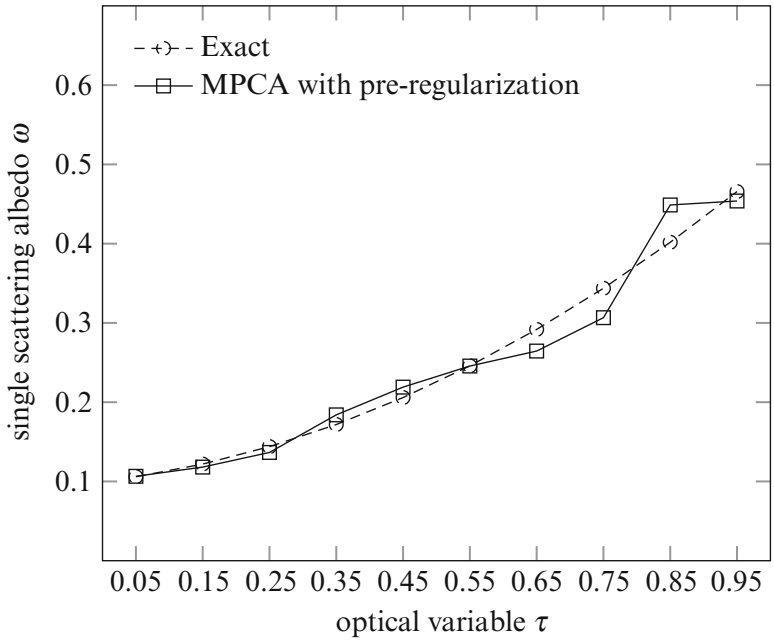


**Fig. 26.4** Comparison of the exact and estimated albedo for mean of the final solution yielded by 25 runs the algorithm using noiseless data.

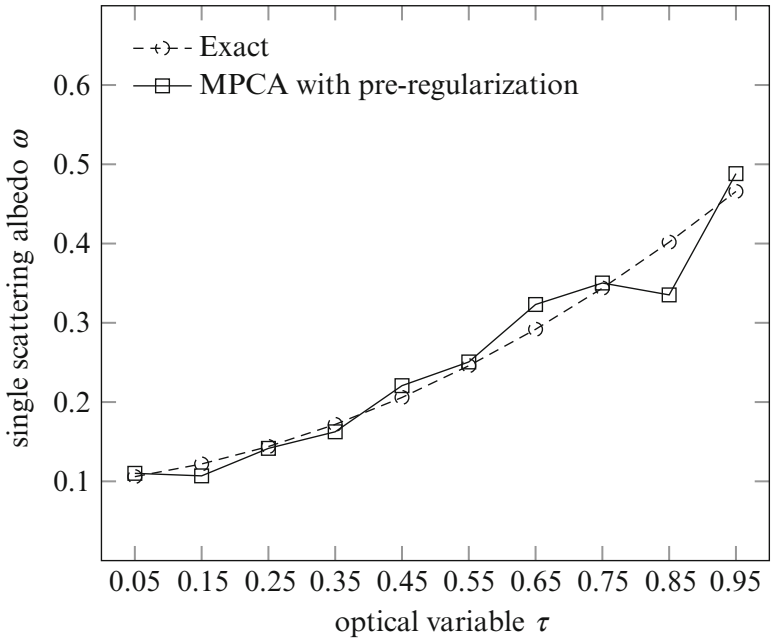


**Fig. 26.5** Comparison of the exact and estimated albedo for mean of the final solution yielded by 25 runs the algorithm using data with 2% of noise.

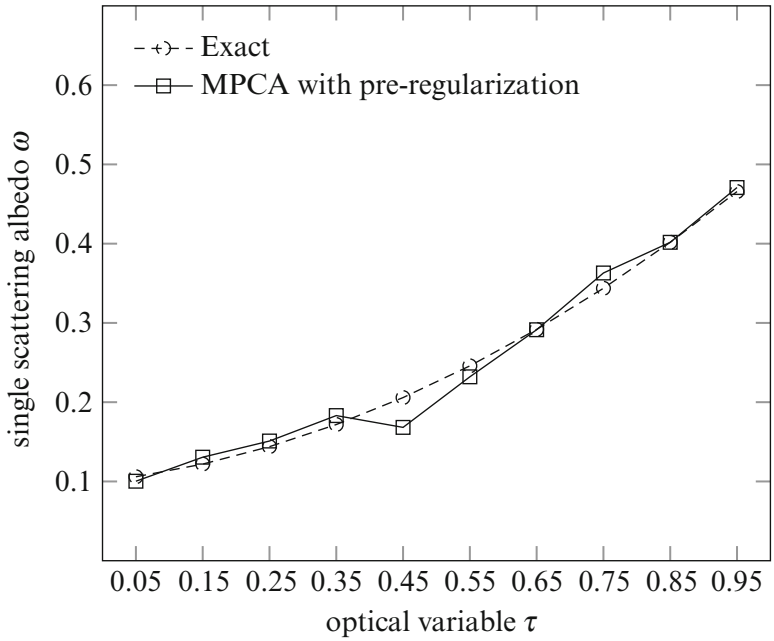


**Fig. 26.6** Comparison of the exact and estimated albedo for mean of the final solution yielded by 25 runs the algorithm using data with 5% of noise.

## 26.6 Conclusions

A pre-regularization scheme used with the MPCA was applied to reconstruction of albedo with spatial dependency, with radiation data acquired by external detectors. This intrinsic regularization scheme saves computational cost and can be applied to any inverse problem, where the extra information (searching for smooth solution) is not embedded in the objective function.

The experiments yielded good estimates for the albedo for all tested cases, noiseless and noisy data. Future works include a hybridization scheme (stochastic optimization associated with a deterministic one, such as LM method) for improving the results.

**Acknowledgements** The authors acknowledge the financial support provided by the Brazilian institution CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico.

## References

[LuBeVe08] Luz, E.F.P., Becceneri, J.C., Campos Velho, H.F.: A new multi-particle collision algorithm for optimization in a high performance environment Journal of Computational Interdisciplinary Sciences **1**, 3–10 (2008)

[SaOl05] Sacco, W.F. and de Oliveira, C.R.: A new stochastic optimization algorithm based on a particle collision metaheuristic Proceedings of 6th World Congress of Structural and Multidisciplinary Optimization, WCSMO Rio de Janeiro (2005)

[StEtAl10] Stephany, S., Becceneri, J. C., Souto, R. P., Campos Velho, H. F., and Silva Neto, A. J.: A pre-regularization scheme for the reconstruction of a spatial dependent scattering albedo using a hybrid ant colony optimization implementation Applied Mathematical Modelling **3**, 34, 561–572 (2010)

# Chapter 27
# Performance of a Higher-Order Numerical Method for Solving Ordinary Differential Equations by Taylor Series

**H. Hirayama**

## 27.1 Prerequisites

We consider the numeric solution of the initial value problem of the following ordinary differential equations.

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}) \qquad \mathbf{y}(x_0) = \mathbf{y}_0$$

As a numerical calculation method of such an ordinary differential equation, the following explicit Runge–Kutta method is often used.

$$\begin{cases} \mathbf{k}_1 &= \mathbf{f}(x_n, \mathbf{y}_n) \\ \mathbf{k}_2 &= \mathbf{f}(x_n + c_2 h, \mathbf{y}_n + a_{21} h \mathbf{k}_1) \\ \mathbf{k}_3 &= \mathbf{f}(x_n + c_3 h, \mathbf{y}_n + a_{31} h \mathbf{k}_1 + a_{32} h \mathbf{k}_2) \\ & \vdots \\ \mathbf{k}_s &= \mathbf{f}(x_n + c_s h, \mathbf{y}_n + a_{s1} h \mathbf{k}_1 + a_{s2} h \mathbf{k}_2 + \cdots + a_{s,s-1} h \mathbf{k}_{s-1}) \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + \sum_{i=1}^{s} b_i \mathbf{k}_i \end{cases} \tag{27.1}$$

When the value $\mathbf{y}_n$ of $\mathbf{y}$ on $x = x_n$ is given, above formula (27.1) gives the value $\mathbf{y}_{n+1}$ of $\mathbf{y}$ on $x_{n+1} = x_n + h$, $a_{ij} (1 \le j < i \le s)$, $b_i (i = 1, \cdots, s)$, $c_i (i = 2, \cdots, s)$ are constants.

H. Hirayama (✉)
Kanagawa Institute of Technology, 1030 Shimo-Ogino, Atsugi-Shi,
Kanagawa-Ken 243-0292, Japan
e-mail: hirayama@sd.kanagawa-it.ac.jp



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_27

It is very difficult to determine these constants as the order increases. In order to make the higher order Runge–Kutta formula, you have to solve a large-scale nonlinear equation. For example, if you want to make a 25 stage 12th order Runge–Kutta formula, you have to solve the nonlinear equation that consists of 7813 equations[On06] to determine the coefficients of the equation. For this reason, the formula of Runge–Kutta which can be used now is to the about 12th order, and we cannot choose more than the 15th order formula.

To solve these disadvantages, there is an implicit Runge–Kutta method (IRK method) as follows.

$$\begin{cases} \mathbf{k}_i &= \mathbf{f}(x_n + c_i h, \mathbf{y}_n + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j) \quad i = 1, \cdots, s \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + h\sum_{i=1}^{s} b_i \mathbf{k}_i \end{cases} \tag{27.2}$$

When $a_{ij} = 0 \quad (j \geq i)$, the above formula (27.2) is called explicit Runge–Kutta, and when other, it is called the implicit Runge–Kutta method.

Implicit Runge–Kutta method, it is possible to select the calculation order freely, even more different from the explicit Runge–Kutta method characterized in that A is stable. To use this formula, it is necessary to solve the simultaneous equations for $\mathbf{k}_i \ (i = 1, \cdots, s)$ in (27.2). Generally this equation is nonlinear simultaneous equations, and it must solve it for every calculation step.

Moreover, the Taylor series solution [HiKoSa02] is known on the textbook of the differential equation. By this calculation method, calculation of arbitrary order is possible.

In this paper, we solve the problem that has been calculated by the IRK [Ko13] in the Taylor series method, discussed its performance and its features.

In the following numerical computation, the computing environment that was used in Taylor series method is Intel i7-3930K, 3.2GHz (6core), Windows 8 (64-bit), MS Visual C ++ 2012, we used a self-made radix $10^8$ multiple-precision program as high precision program. The computing environment that was used in IRK method is Intel i7-3820, 3.6GHz (4core), Scientific Linux 6.3 (64-bit), Intel C ++ 13.0.1, multiple precision program MPFR 3.1.1 / GMP 5.1.1, BNC-pack 0.8.

## 27.2 Numerical Solution of an Ordinary Differential Equation with a High-Order Formula

The problem to deal with is an ordinary differential equation called HIRES. With the equation well quoted as a test problem, it is treated on many books [HaWa93] or Web.

It is a problem which consists of 8 equations as follows:

$$\begin{cases} y_1' = -1.71y_1 + 0.43y_2 + 8.32y_3 + 0.0007 \\ y_2' = 1.71y_1 - 8.75y_2 \\ y_3' = -10.03y_1 + 0.43y_4 + 0.035y_5 \\ y_4' = 8.32y_2 + 1.71y_3 - 1.12y_4 \\ y_5' = -1.745y_5 + 0.43y_6 + 0.43y_7 \\ y_6' = -280y_6y_8 + 0.69y_4 + 1.71y_5 - 0.43y_6 + 0.69y_7 \\ y_7' = 280y_6y_8 - 1.81y_7 \\ y_8' = -280y_6y_8 + 1.81y_7 \end{cases} \tag{27.3}$$

Initial conditions:$y_1(0) = 1, y_2(0) = y_3(0) = y_4(0) = y_5(0) = y_6(0) = y_7(0) = 0$, $y_8(0) = 0.0057$. Integration interval : $0 \leq t \leq 321.8122$.

### *27.2.1 Computer Program*

The computer program can be written easily as follows. Suppose that the coefficient of a Taylor series is expressed with an array in program language. In other words, $n$-th order coefficient of the Taylor series $y_m$ represented by y[m][n]. Here, the formula to calculate the $y_4$ is written only. We can write other formulas similarly. The zero-order constant term is determined using the initial value.

```
y[4][0] = 0 ;
```

To determine the $i+1$th order coefficient of the Taylor series, substituting the Taylor expansions to ordinary differential equations, to compare the coefficient of $t^i$. Because the left side is (i+1)y[4][i+1], the following formula is obtained.

```
y[4][i+1]=(8.32*y[2][i]+1.71*y[3][i]
  -1.12*y[4][i])/(i+1) ;
```

By repeating use of this equation, the coefficient of any order of the Taylor expansion equation is obtained.

$y_6$ and subsequent expressions have a nonlinear term $y_6y_8$. This nonlinear term can be calculated in the following manner. In the following program, using the calculated nonlinear term, we calculate the equation $y_6$.

```
y6y8[i] = 0 ;
for(int j=0 ; j<= i ; j++)y6y8[i]+
  =y[5][j]*y[7][i-j] ;
y[6][i+1] = (-280y6y8[i]+0.69*y[4][i]
  1.71*y[5][i]-0.43*y[6][i]+0.69*y[7][i])/(i+1) ;
```

The nonlinear clause calculated here is applicable also to calculation of $y_7$ and $y_8$. Calculation of this nonlinear clause is also a kind of automatic differentiation[Ra81]. The library of Taylor series [Hi2] was used for actual calculation.

**Table 27.1** Numerical results of HIRES.

| order | comp. time(msec) | No. of steps | max. step size | min. step size |
|---|---|---|---|---|
| 3 | 874.00 | 1244404 | 6.04e-3 | 4.28e-12 |
| 4 | 45.41 | 61444 | 3.73e-2 | 3.06e-8 |
| 5 | 13.06 | 16254 | 1.63e-1 | 2.85e-6 |
| 6 | 9.40 | 10980 | 2.17e-1 | 4.75e-5 |
| 7 | 8.42 | 9179 | 4.15e-1 | 3.03e-4 |
| 8 | 7.69 | 8200 | 4.85e-1 | 8.96e-4 |
| 9 | 7.51 | 7445 | 5.00e-1 | 1.81e-3 |
| 10 | 7.32 | 6870 | 5.33e-1 | 3.48e-3 |
| 11 | 7.08 | 6371 | 5.73e-1 | 6.0ee-3 |
| 12 | 6.90 | 5951 | 6.24e-1 | 9.74e-3 |
| 13 | 6.89 | 5583 | 6.31e-1 | 1.46e-2 |
| 14 | 6.78 | 5261 | 7.12e-1 | 1.97e-2 |
| 15 | 6.77 | 4974 | 7.28e-1 | 2.33e-2 |
| 16 | 6.59 | 4718 | 7.65e-1 | 2.85e-2 |
| 17 | 6.59 | 4487 | 8.14e-1 | 3.46e-2 |
| 18 | 6.59 | 4277 | 8.78e-1 | 3.79e-2 |
| 19 | 6.65 | 4088 | 8.73e-1 | 3.97e-2 |
| 20 | 6.59 | 3914 | 9.00e-1 | 4.15e-2 |
| 25 | 6.59 | 3228 | 1.11e-1 | 5.03e-2 |
| 30 | 6.59 | 2749 | 1.26e-1 | 5.91e-2 |
| 35 | 6.71 | 2395 | 1.22e-1 | 6.79e-2 |

### *27.2.2 Computational Results*

This equation was solved by using Taylor series in the order from 3 to 20 and 25, 30, 35. The results are given in Table 27.1.

Suppose that Taylor series obtained in the calculation assumed the form

$$y(t) = y_0 + y_1(t-t_0) + y_2(t-t_0)^2 + \cdots + y_n(t-t_0)^n \tag{27.4}$$

By use of this formula(27.4), the absolute error is presumed to be $y_n(t-t_0)^n$. If a step size is set to $h$, step size $h$ which fulfills the conditions of the absolute error $\varepsilon_{abs}$ can be written as

$$|y_n h^n| \leq \varepsilon_{abs} \tag{27.5}$$

If $y_0 (\neq 0)$ assumes that it is very large compared to $y_n h^n$, it turns out that step size $h$ fits the formula

$$\left|\frac{y_n}{y_0} h^n\right| \leq \varepsilon_{rel} \tag{27.6}$$

The step size $h$ is calculated for every formula((27.4)and(27.6)) from these conditional expressions. Let the minimum of the obtained calculation results be a step size $h$. It computed using the adapted type numeric solution using this step size. In this computation, it computed as $\varepsilon_{abs} = \varepsilon_{rel}^{[} = 10^{-14}]$.

When the 3rd order formula was used for this problem, only about 12 figures of accuracy were acquired probably for the rounding error. The accuracy of about 13 figures was acquired in 4th order formula. The accuracy of 14 or more figures was acquired in 5th or more order formula.

## 27.3 Comparison with the Implicit Runge–Kutta Method

Here, comparison with the implicit Runge–Kutta method is performed. Although the implicit Runge–Kutta method was known from before that calculation of arbitrary order is possible, since it was necessary to solve a nonlinear equation in the computation, high order calculations were not actually performed. Such calculations are performed recently, we did a comparison between the results and the Taylor series method.

### *27.3.1 Lorenz Model*

We consider Lorenz model as a simple problem. This problem is simple three ordinary differential equations as represented by the following ordinary differential equations.

$$\begin{cases} \dfrac{dy_1}{dx} = \sigma(-y_1 + y_2) \\ \dfrac{dy_2}{dx} = -y_1 y_3 + r y_1 - y_2 \\ \dfrac{dy_3}{dx} = y_1 y_2 - b y_3 \end{cases} \tag{27.7}$$

Initial conditions and constants are

$$y_1(0) = 0,\ y_2(0) = 1,\ y_3(0) = 0,\ \sigma = 10,\ r = \frac{470}{19},\ b = \frac{8}{3}$$

This problem(27.7) is integrated over the interval $[0, 50]$ by the multi-precision floating point number of 200 digits.

This problem is known as one of the nonlinear equations representing a chaotic behavior. In double-precision calculations, for cancelation, it is also known a problem that accurate results cannot be obtained. Here, it is calculated at 200 digits of sufficient accuracy is calculated by decreasing the influence of cancelation.

**Table 27.2** Calculation results of the 200-digit accuracy of the Lorenz model.

| (Tolerance) | Taylor series($10^{-120}$) | | | Implicit Runge–Kutta($10^{-120}$) | | |
|---|---|---|---|---|---|---|
| Order( No. of Stages) | 160 | 200 | 240 | 160(80) | 200(100) | 240(120) |
| CPU time(sec) | 42.5 | 46.1 | 81.7 | 1991.4 | 2317.4 | 2555.0 |
| No. of Steps | 1005 | 706 | 557 | 1661 | 863 | 563 |
| Error | 1.0e-110 | 2.7e-110 | 2.3e-110 | 6.5e-110 | 1.3e-109 | 2.3e-109 |

By using this program, we computed for three kinds of calculation order (100, 120,160) with a required accuracy of $10^{-120}$. The results are shown in Table 27.2. Also shown there are the results of the IRK method according to Kouya [Ko13].

Since the computing environment is different, the comparison is difficult. When compared simply, the Taylor series method can be seen to be about 40 times faster than the IRK method.

### *27.3.2 The P-Dimensional Brusselator Problem*

As a big problem, I can deal with one-dimensional Brusselator problem. This ordinary differential equation is derived from the partial differential equations

$$\begin{cases} \dfrac{\partial u}{\partial t} = 1 + u^2 v - 4 + 0.02 \dfrac{\partial^2 u}{\partial x^2} \\ \dfrac{\partial v}{\partial t} = 3u - u^2 v + 0.02 \dfrac{\partial^2 v}{\partial x^2} \end{cases}$$

If this equation is equally divided into $N$ pieces($N = 500$) in terms of the space variable, $N+1$ ordinary differential equations will be obtained.

Boundary condition : $u(x=0,t)=0, u(x=1,t)=0, v(x=0,t)=3, v(x=0,t)=3$. Initial condition : $u(x,t=0)=1+\sin(2\pi x), v(x,t=0)=3$

$$\begin{cases} \dfrac{du_i}{dt} = 1 + u_i^2 v_i - 4 + 0.02 \dfrac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta x)^2} \\ \dfrac{dv_i}{dt} = 3u_i - u_i^2 v_i - 4 + 0.02 \dfrac{v_{i+1} - 2v_i + v_{i-1}}{(\Delta x)^2} \\ (i = 0, 1, \cdots, N+1) \end{cases}$$

We integrated these ordinary differential equations on $[0,10]$. Computation was done using the multiple-precision number of radix $10^8$ of 10 figures (they contain about 80 digits). Kouya was calculated with an accuracy of 70 digits.

By the IRK method, since the calculation degree became twice a number of stages, the degree calculated using the (40, 60, 80)-th order Taylor series. This calculation result is shown in table 27.3.

**Table 27.3** The calculation result of the 70-digits accuracy of 1D Brusselator problem.

| (Tolerance) | Taylor series($10^{-60}$) | | | Implicit Runge–Kutta($10^{-60}$) | | |
|---|---|---|---|---|---|---|
| Order( No. of Stages) | 40 | 60 | 80 | 40(20) | 60(30) | 80(40) |
| CPU time(sec) | 6518.2 | 9131.3 | 11797.6 | 19712.0 | 11377.8 | 13667.6 |
| No. of Steps | 12313 | 8454 | 6436 | 3249 | 890 | 630 |
| error | 2.9e-63 | 8.6e-65 | 1.2e-65 | 4.8e-53 | 1.1e-43 | 1.7e-44 |

**Table 27.4** The calculation result of the 70-digits accuracy of 1D Brusselator problem at the time of restricting a step size.

| Tolerance($10^{-60}$) | max $h_k \leq 0.005$ | | max $h_k \leq 0.002$ | |
|---|---|---|---|---|
| Order( No. of Stages) | 60(30) | 80(40) | 60(30 ) | 80(40) |
| CPU time(sec) | 21834.0 | 32983.4 | 43723.7 | 74230.3 |
| No. of Steps | 1864 | 1748 | 3856 | 4156 |
| error | 1.1e-49 | 7.4e-49 | 3.4e-53 | 2.9e-53 |

In this problem, although the computation times of the Taylor series method and the IRK method are almost same, when comparing the calculation results by the two methods we see as for the calculation result by the IRK method, we see that the latter are worse after about 10 figures of arithmetic precision.

Kouya has aimed to improve this point, when arithmetic precision restricts a step size. In Table 27.4, restricting a step size $0.002 \leq h \leq 0.005$ shows that arithmetic precision is improvable.

Comparing these improved results with the Taylor series method results, we conclude that the calculation accuracy is almost the same. However, due to limitations of the step size, we see that the computation time of the latter increases considerably. In this case, the IRK method requires from 4.7 to 7.3 times more time than the Taylor expansion method.

## 27.4 Conclusions

When computing with different environments, it is impossible to discuss the relative merits of the two methods. But if the performance of the compiler and the multiple-precision arithmetic program are almost the same, then the Taylor series method is about 40 times faster than the IRK method, which makes the calculation in stiff problems several times faster.

In stiff problems, for many ordinary differential equations using the Taylor series method and high-order calculation methods we can expect accurate computations at high speed.

For the Taylor series method, because the calculation procedure is simple, parallelization is easy. In particular, parallelization in the higher-order calculation is effective. It seems that speed of the higher-order calculation, which requires a longer time, can be improved by parallelization.

## References

[HaWa93] Hairer, E. and Wanner, G., Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems, Springer-Verlag (1993)

[HiKoSa02] Hirayama, H., Komiya, S., and Satou, S., Solving Ordinary Differential Equations by Taylor Series, JSIAM, 12(2002), 1–8 (Japanese)

[Hi2] Hirayama, H., Tateno, H., Asano, N., and Kawaguchi, T., How to use Mathmatical libraly for Taylor series, SENAC, Information Synergy Organization, Tohoku University, 40(2007) 29–68( Japanese)

[Ko13] Kouya, T., Performance Analysis of Parallelized Fully Implicit Runge-Kutta Method in Multiple Precision Computing EnvirromentĄAIPSJ Technical Report, Vol. 2013-HPC-139(2013), No. 18, 1–8

[On06] Ono, H., On the 25 stage 12th-order explicit Runge–Kutta method, JSIAM, 16 (2006), 177–186 (Japanese)

[Ra81] Rall, L.B., Automatic Differentiation-Technique and Applications, Lecture Notes in Computer Science, Vol. 120, Springer-Verlag, Berlin-Heidelberg-New York(1981)

# Chapter 28
# Retinal Image Quality Assessment Using Shearlet Transform

**E. Imani, H.R. Pourreza, and T. Banaee**

## 28.1 Introduction

Eye diseases such as diabetic retinopathy (DR) affect a large number of the population. Retinal fundus photographs are widely used in the diagnosis and treatment of various eye diseases in clinics. It is also one of the main resources for mass screening of diabetic retinopathy. The resulting retinal images must be examined by an expert human grader in a cumbersome and time-consuming diagnosis process. Automated analysis and diagnosis has the potential to reduce the workload and thus increase the cost-effectiveness of such screening initiatives. Nevertheless, there are number of problems that must be solved in order to develop a fully reliable automated retinal images analysis system. Among them, is the need to guarantee that the quality of the retinal images to be graded exceeds a threshold below which the automated analysis procedures may fail [PiOlDa12].

In a DR system, an image is considered poor quality if it is difficult or impossible to make a reliable clinical judgment on the image regarding presence or absence of DR [YuEtAl12]. Performing automated analysis on the image of insufficient quality will produce unreliable results. Images with low quality should be examined by an ophthalmologist and reacquired if necessary [NiAbVa06]. The store and forward teleophthalmology systems involve acquiring images and transmitting them for remote retinopathy detection. This could become problematic when received images do not have enough quality and patient is not accessible. Thus an algorithm

E. Imani (✉) • H.R. Pourreza
Ferdowsi University of Mashhad, Azadi Square, Mashhad, Iran
e-mail: elaheh.imani@gmail.com; hpourreza@um.ac.ir

T. Banaee
Mashhad University of Medical Sciences, Mashhad, Iran
e-mail: banaeet@mums.ac.ir



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_28

with ability of automatically assessing fundus image quality is a necessary tool in preprocessing stage for reliable lesion detection especially in the systems that deliver eye care through telecommunications technology.

Fundus image quality can be affected by a number of factors including patient's head or eye movement, poorly dilated and small pupils, blinking and media opacity. Head or eye movement can result in out-of-focus and incorrectly illuminated images. Poorly dilated pupils may affect image illumination and create dark low-contrast images that can prevent lesion identification. If fundus cameras capture retinal images through cataract, images appear blurred and are often poor quality [HuEtAl11]. In 2006, Zimmer–Galler [ZiZe06] reported that 11% of the images in their study were unreadable. It was estimated that 25% of the poor quality images were caused primarily by poor patient fixation, 25% by poor focus and pupil centering, and 25% were thought to be caused by small pupil size, media opacity, and instrument failure. A specific cause for the unreadable image could not be determined for the remainder. Figure 28.1 shows some instances of good and poor quality retinal images.

Several approaches have been developed to automatically determine the quality of the retinal images. These approaches could be classified into two categories. The first category is based on generic image quality parameters such as sharpness and contrast. These methods make use of simple image measurements to estimate image quality avoiding eye structure segmentation procedures which are usually complex and time-consuming tasks [PiOlDa12]. In 2001, Lalondey [LaGaBo01] proposed a method based on histogram of edge magnitude and local histogram of pixel gray-scale values to evaluate image focus and illumination. In this method, the quality of a given image is determined through the difference between its histogram and the mean histogram of a set of good quality images used as reference. In 2009,

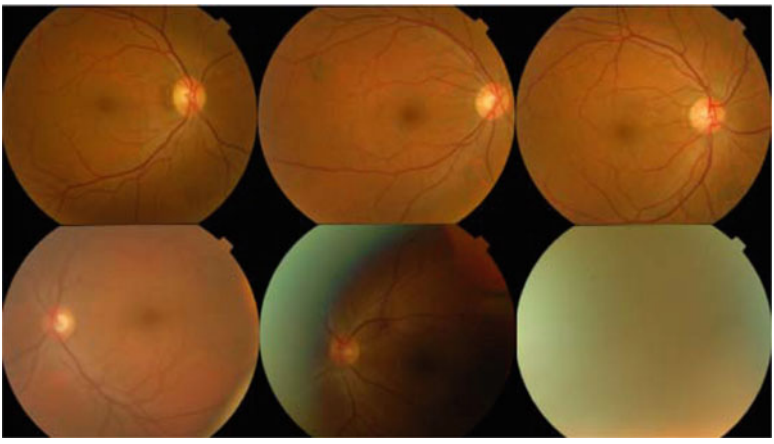

**Fig. 28.1** Examples of good quality and poor quality retinal images: top row are good quality images and bottom row are poor quality images

Davis et al. [DaEtAl09] focused their quality assessment on contrast and luminance features. A method based on sharpness and illumination parameters was proposed by Bartling [BaWaMa09] in 2009. Illumination was measured through evaluation of contrast and brightness and the degree of sharpness was calculated from the spatial frequencies of the image. Image structure clustering, Heralick features, and sharpness measures based on image gradient magnitudes were used by Paulus et al. [PaEtAl10] to classify poor quality retinal images. In 2012, Dias et al. [PiOlDa12] introduced a method based on fusion of generic image quality indicators such as image color, focus, contrast, and illumination.

The advantage of the image quality assessments based on generic image quality measures is their algorithmic simplicity which translates into reduced computational complexity [PiOlDa12].

The second group is based on the structural information of the image which requires segmentation of anatomical landmarks in retinal images. In 2005, Fleming et al. [FlEtAl06] developed a method based on field definition and image clarity. The clarity analysis is based on the vasculature of a circular area around the macula. The authors whether or not a given image has enough quality using presence/absence of small vessels in the selected circular areas. In 2006, Niemeijer [NiAbVa06] proposed a method based on clustering the filter bank response vectors in order to obtain a compact representation of the image structures. In 2008, Giancardo et al. [GiEtAl08] assessed the quality of retinal images based on the eye vasculature. Giancardo concluded used vessel density in local patches as a feature vector for quality assessment. In 2011, Hunter et al. [HuEtAl11] proposed a method based on the clarity of retinal vessels within the macula region and contrast between the fovea region and retina background. The methods based on structural information require anatomical landmarks segmentation which is complex and error prone, especially in the case of poor quality images. This is the major disadvantage of such approaches [PiOlDa12].

The rest of the chapter is organized as follows. In Section 28.2, we give a brief introduction to shearlet transform. In Section 28.3, we propose a retinal image quality assessment based on generic parameters with the usage of shearlet transform. We evaluate the performance of the developed approach in Section 28.5. The results are compared against state-of-the-art retinal image quality assessment methods. In Section 28.6, we finish the paper with some conclusions.

## 28.2 Prerequisites

One of the most useful features of wavelets is their ability to efficiently approximate signals containing pointwise singularities. Consider a one-dimensional signal which is smooth away from point discontinuities. If the signal is approximated using the best M-term wavelet expansion, then the rate of decay of the approximation error, as a function of M, is optimal. In particular, it is significantly better than corresponding Fourier approximation error [Li10]. Since wavelets have isotropic supports, they

fail to capture the geometric regularity along edges. Recently, the novel directional representation system of shearlets [LaEtAl05] proposed to provide efficient tools for analyzing the geometrical structures of a signal using anisotropic window functions. Among directional representation systems, shearlets are the most versatile and successful systems, the reason for this being an extensive list of desirable properties: shearlet systems are generated by one function, they provide precise resolution of wavefront sets, they allow compactly supported analyzing elements, they are associated with fast decomposition algorithms, and they provide a unified treatment of the continuum and the digital realm [KuLeLi12].

### *28.2.1 Brief Introduction to Shearlet Transform*

In many applications in image processing, the important information is often located around edges separating image objects from background. These features correspond to the anisotropic structures in the image. Shearlets are designed to efficiently encode such anisotropic features [KuLeLi12]. For $j \geq 0, k \in Z$, let

$$A_{2^j} = \begin{pmatrix} 2^j & 0 \\ 0 & 2^{\frac{j}{2}} \end{pmatrix} \quad S_k = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix} \quad M_c = \begin{pmatrix} c_1 & 0 \\ 0 & c_2 \end{pmatrix}$$

where $c = (c_1, c_2)$ and $c_1, c_2$ are positive constants. Similarly,

$$\tilde{A_{2^j}} = \begin{pmatrix} 2^j & 0 \\ 0 & 2^{\frac{j}{2}} \end{pmatrix} \quad \tilde{S_k} = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix} \quad \tilde{M_c} = \begin{pmatrix} c_1 & 0 \\ 0 & c_2 \end{pmatrix}$$

We are now ready to define a shearlet transform as follows. Let $c = (c_1, c_2) \in (R_+)^2$. For $\phi, \psi, \tilde{\psi} \in L^2(R^2)$ the cone-adapted discrete shearlet system $SH(\phi, \psi, \tilde{\psi})$ is defined by

$$SH(\phi, \psi, \tilde{\psi}; c) = \Phi(\phi; c_1)\Psi(\psi; c)\tilde{\Psi}(\psi; c)$$

where

$$\Phi(\phi; c) = \{\phi_m = \phi(. - m) : m \in Z^2\}$$
$$\Psi(\psi; c) = \{\psi_{j,k,m} = 2^{3j/4}\psi(S_k A_{2^j}. - m) : j \geq, |k| \leq \lceil 2^{j/2} \rceil, m \in M_c Z^2\}$$
$$\tilde{\Psi}(\tilde{\psi}; c) = \{\tilde{\psi}_{j,k,m} = 2^{3j/4}\tilde{\psi}(\tilde{S}_k \tilde{A}_{2^j}. - m) : j \geq, |k| \leq \lceil 2^{j/2} \rceil, m \in M_c Z^2\}$$

If $SH(\phi, \psi, \tilde{\psi}; c)$ is a frame for $L^2(R^2)$, we refer to $\phi$ as a scaling function and $\psi$ and $\tilde{psi}$ as shearlets. Observe that shearlets are obtained by applying translation, anisotropic scaling matrices $A_{2^j}$ and shear matrices $S_k$ to the fixed generating

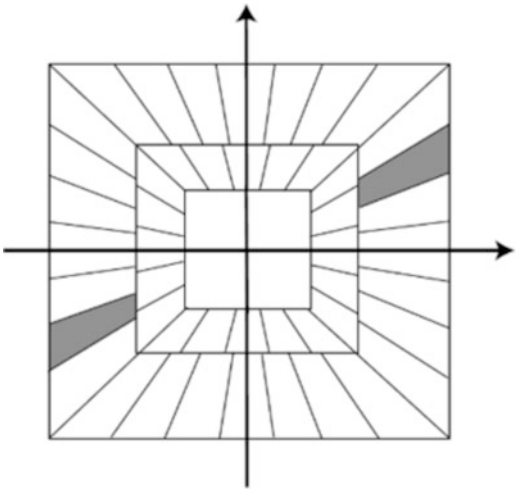

**Fig. 28.2** The tilting of the frequency plane introduce by shearlets in $\Psi$.

functions $\psi$ . The matrices $A_{2^j}$ and $S_k$ lead to windows which can be elongated along arbitrary directions and the geometric structures of singularities in images can be efficiently represented using them [Li10]. Figure 28.2 shows the tilting of the frequency plane using shearlet system $\psi$. It was shown that shearlet $\psi$ can provide nearly optimal approximation for a piecewise smooth function $f$ with $C^2$ smoothness except at points lying on $C^2$ curves [Li10].

## 28.3 Proposed Method

In this work, an automated retinal image quality assessment system is presented. Input to the developed system is a color image of human retina, which is acquired by using a fundus camera, and its output is the quality level of the input image, as shown in Figure 28.3. The proposed method follows a sequences of steps: preprocessing, feature extraction, and classification. In preprocessing step, we remove useless image information in order to decrease the processing time and the green channel of the retinal image selected for further processing. In the second step, we extract generic features with the usage of shearlet transform. Finally by using these features and a supervised classifier, we specify whether the image is of poor or good quality. In this section, the proposed algorithm is described in detail.

### *28.3.1 Preprocessing*

Since green channel of the image provides maximum contrast for retinal landmarks such as vessels among other color image components, this channel is chosen to

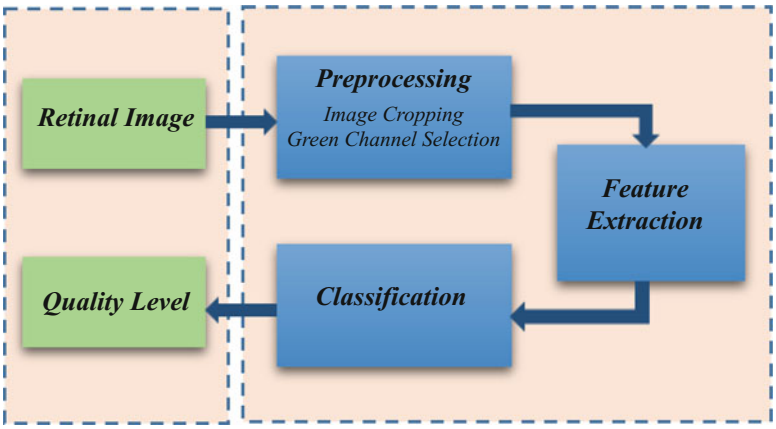


**Fig. 28.3** Block diagram of the proposed method

apply the proposed algorithm. We remove useless information of the retinal image by cropping it in order to include retinal region only. A mask for cropping the retinal image is created using a threshold value and morphological operations. The binary mask is created by applying a threshold value of 3 to green plane of the retinal images. Afterwards, noisy regions on the background and foreground are removed using morphological opening and closing. After creating the retinal mask, we find the bounding box containing the retinal region. Cropping the useful part of retinal image accelerates other processing stages. Finally, the images are resized to $512 \times 512$ pixels.

### *28.3.2 Feature Extraction*

Visual perception is very sensitive to local image structures such as edges. The quality of the image is a function of edge strength. In blurred and low contrast images, the strength of edges is very weak. Thus evaluation of retinal image quality can be made by edge features. In retinal images, these edges arise from vasculature, optic disk, and lesions. The proposed method assesses the quality of retinal images using edge information. The edge features correspond to the anisotropic structures in the data. Since shearlet systems capture such anisotropic features efficiently [KuLeLi12], we use shearlet transform to detect retinal edge features. The degree of image quality could be specified by measuring the alterations in the statistical characteristics of the shearlet coefficients.

We classify retinal images as good quality or poor quality. Some instances of poor quality retinal images are shown in Figure 28.1. As it is shown in this figure, the strength of the edges in poor quality images is lower than good quality ones. Thus, the image quality level can be specified by measuring the changes in statistical

characteristics of the edge information. The retinal image is decomposed using shearlet transform. Each coefficient in shearlet expansion of an image is the result of convolution of the associated shearlet and the image. If a shearlet of a given scale, angle, and location is approximately aligned along a curve, its shearlet coefficient is large, otherwise it is close to zero [KuSa07]. Since changes in image quality level affect the property of curve singularities in the image, the corresponding large shearlet coefficients will be also affected. Hence, the quality of retinal image could be assessed using statistical characteristics of shearlet coefficients.

In order to demonstrate that sub-band statistics are affected by changing in quality levels of the image, Figure 28.5 plots the coefficients distribution of good and poor quality retinal images which were shown in Figure 28.4. As it has been indicated in Figure 28.5, the coefficients distribution of the poor quality retinal

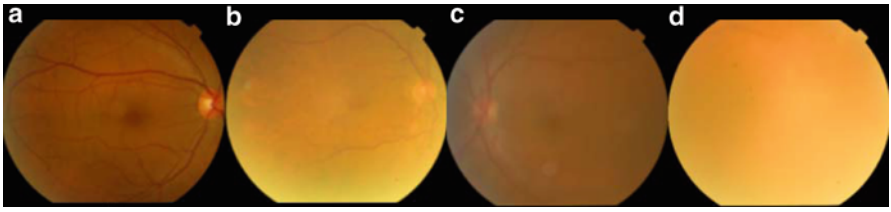


**Fig. 28.4** Some instances of retinal images with different quality level. **(a)**: a good quality retinal image. **(b-d)**: poor quality retinal images.

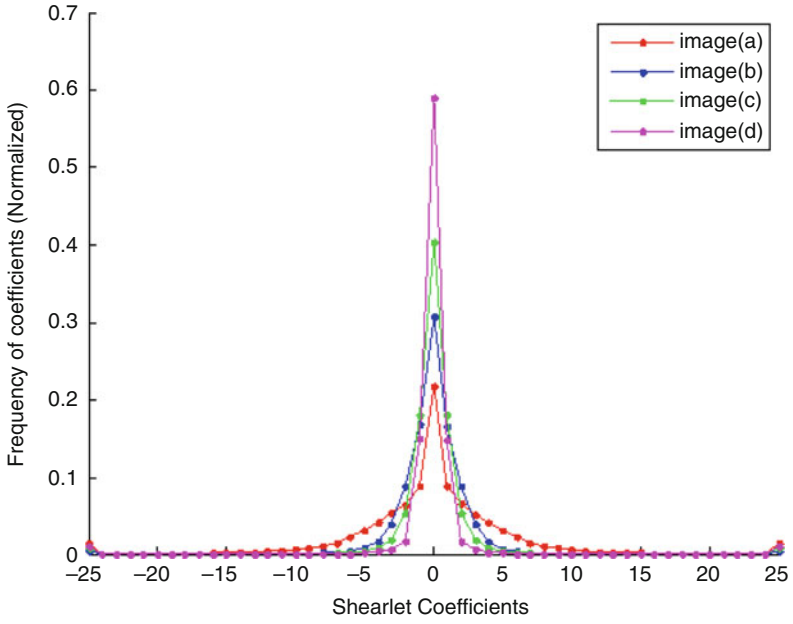


**Fig. 28.5** Shearlet coefficients distribution of retinal images.

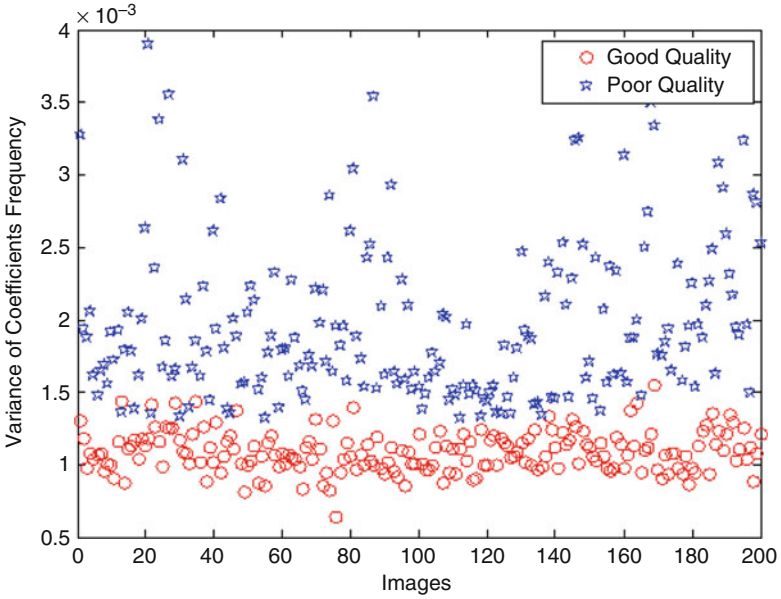


**Fig. 28.6** VSCF values of 400 retinal images with different quality level.

images is more concentrated around zero and falls rapidly. The Variance of the Shearlet Coefficients Frequency (VSCF) is computed to evaluate the quality of the images. By decreasing the quality level of image, the value of VSCF increases. In order to demonstrate the effect of quality level on the VSCF value, Figure 28.6 shows the value of VSCF for 400 images with 200 good quality and 200 poor quality. As it can be seen from Figure 28.6, the VSCF values for good quality retinal images are less than VSCF values for poor quality ones. Thus, VSCF value could be used to classify retinal images as good quality or poor quality.

The images are decomposed into three scales and 8 orientations to form oriented responses. Since the finer scales are more sensitive to noise, the coefficients of the second scale of shearlet transform are used to extract statistical features.

## 28.4 Material

Several retinal image datasets were used to develop and test the retinal image quality assessment. All of the images have been manually graded by ophthalmologists from the Khatam-Al-Anbia eye hospital of Mashhad, Iran, using a software tool provided for image annotation.

### 28.4.1 Messidor Dataset

The images in this dataset were obtained using a color video 3CCD camera on Topcon TRC NW6 non-mydriatic retinograph with a 45 degree field of view. The dataset consists of 1200 eye fundus color images with the size of $1440 \times 960$, $2240 \times 1488$ or $2304 \times 1536$ pixels.

### 28.4.2 Khatam-Al-Anbia Dataset

Khatam-Al-Anbia dataset were obtained in Khatam-Al-Anbia eye hospital of Mashhad, Iran. This dataset includes 1000 retinal images with the resolution of $3872 \times 2592$ pixels.

## 28.5 Results

This section presents the classification results of the image quality assessment algorithm. The retinal images are classified as good quality and poor quality using a supervised classifier and extracted feature vector. A support vector machine (SVM) with different kernels was used as a classifier. Classifier testing was performed by 5-fold cross validation, using 80% of the dataset for training and 20% of the dataset for testing. In order to assess the algorithm performance, three measures were used: sensitivity, specificity, and accuracy. These performance measures are defined as follows:

$$sensitivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{TN + FP}$$

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. The results of the retinal quality assessment using SVM classifier with different kernels on Messidor and Khatam-Al-Anbia datasets are shown in Table 28.1 and Table 28.2. As it is shown in Table 28.1, the best results are obtained using rbf and polynomial kernels on Messidor and Khatam-Al-Anbia datasets. Table 28.3 compares the performance of the proposed method with the method presented in [NiAbVa06], in terms of sensitivity, specificity on Messidor dataset. Results of Niemeijer et al. [NiAbVa06] is provided by the authors. The results show that the performance of the proposed method is higher than this algorithm.

**Table 28.1** Performance achieved by the proposed method on Messidor dataset.

| | sensitivity | specificity | accuracy |
|---|---|---|---|
| linear | 96.00 | 93.59 | 93.58 |
| quadratic | 96.00 | 92.83 | 92.83 |
| polynomial | 92.00 | 93.17 | 93.08 |
| rfb | 96.00 | 93.76 | 93.75 |

**Table 28.2** Performance achieved by the proposed method on Khatam-Al-Anbia dataset.

| | sensitivity | specificity | accuracy |
|---|---|---|---|
| linear | 96.34 | 97.46 | 96.90 |
| quadratic | 96.72 | 97.26 | 97.00 |
| polynomial | 97.15 | 96.69 | 96.90 |
| rfb | 96.34 | 97.46 | 96.90 |

**Table 28.3** Performance achieved by the proposed method and Neimeijter et al method.

| | sensitivity | specificity |
|---|---|---|
| Proposed Method | 96.00 | 93.76 |
| Niemeijer et al. Method | 84.44 | 90.73 |

## 28.6 Conclusions

The proposed method evaluates the retinal image quality with the usage of shearlet transform. Changes in quality level of the retinal image affect the properties of image edges. Therefore, edge information for the images could be used to assess their quality. The edge and curve information of the image are detected using shearlet transform. Image quality levels were specified by measuring the alterations of the statistical characteristics of shearlet coefficients. Experimental results have shown that the proposed method gives comparable results (93.75% for Messidor and 96.90% for Khatam-Al-Anbia) on Messidor and Khatam-Al-Anbia datasets.

## References

[PiOlDa12] Pires Dias, J.M., Oliveira, C.M., and da Silva Cruz, L.A. *Retinal image quality assessment using generic image quality indicators.* Information Fusion (2012).

[YuEtAl12] Yu, H., Agurto, C., Barriga, S., Nemeth, S.C., Soliz, P., and Zamora, G. *Automated image quality evaluation of retinal fundus photographs in diabetic retinopathy screening.* In Image Analysis and Interpretation (SSIAI), 2012 IEEE Southwest Symposium on, pp. 125–128. IEEE, (2012).

[NiAbVa06] Niemeijer, M., Abramoff, M.D., and van Ginneken, B. *Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening.* Medical image analysis 10, no. 6 (2006)

[HuEtAl11] Hunter, A., Lowell, J.A., Habib, M., Ryder, B., Basu, A., and Steel, D. *An automated retinal image quality grading algorithm.* In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, pp. 5955–5958. IEEE, (2011)

[ZiZe06] Zimmer-Galler, I. and Zeimer, R. *Results of implementation of the DigiScope for diabetic retinopathy assessment in the primary care environment.* Telemedicine Journal & e-Health 12, no. 2 (2006)

[LaGaBo01] Lalonde İ, M., Gagnon İ, L., and Boucher, M.C. *Automatic visual quality assessment in optical fundus images.* (2001).

[DaEtAl09] Davis, H., Russell, S., Barriga, E., Abramoff, M., and Soliz, P. *Vision-based, real-time retinal image quality assessment.* In Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on, pp. 1–6. IEEE, (2009).

[BaWaMa09] Bartling, H., Wanger, P., and Martin, L. *Automated quality evaluation of digital fundus photographs.* Acta ophthalmologica 87, no. 6 (2009)

[PaEtAl10] Paulus, J., Meier, J., Bock, R., Hornegger, J., and Michelson, G. *Automated quality assessment of retinal fundus photos.* International journal of computer assisted radiology and surgery 5, no. 6 (2010): 557–564.

[FlEtAl06] Fleming, A.D., Philip, S., Goatman, K.A., Olson, J.A., and Sharp, P.F. *Automated assessment of diabetic retinal image quality based on clarity and field definition.* Investigative ophthalmology & visual science 47, no. 3 (2006)

[GiEtAl08] Giancardo, L., Abramoff, M.D., Chaum, E., Karnowski, T.P., Meriaudeau, F., and Tobin, K.W. *Elliptical local vessel density: a fast and robust quality metric for retinal images.* In Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, pp. 3534–3537. IEEE, (2008)

[Li10] Lim, W.Q. *The discrete shearlet transform: A new directional transform and compactly supported shearlet frames.* Image Processing, IEEE Transactions on 19, no. 5 (2010)

[LaEtAl05] Labate, D., Lim, W.Q., Kutyniok, G., and Weiss, G. *Sparse multidimensional representation using shearlets.* In Optics & Photonics 2005, pp. 59140U-59140U. International Society for Optics and Photonics, (2005)

[KuLeLi12] Kutyniok, G., Lemvig, J., and Lim, W.Q. *Compactly supported shearlets.* In Approximation Theory XIII: San Antonio 2010, pp. 163–186. Springer New York, (2012)

[KuSa07] Kutyniok, G. and Sauer, T. *From Wavelets to Shearlets and back again.* Approximation Theory XII (San Antonio, TX, 2007), CK Chui, M. Neamtu, and L. Schumaker, eds., Nashboro Press, Nashville, TN, to appear (2007).

# Chapter 29
# The Radiative–Conductive Transfer Equation in Cylinder Geometry and Its Application to Rocket Launch Exhaust Phenomena

**C.A. Ladeia, B.E.J. Bodmann, and M.T.B. Vilhena**

## 29.1 Introduction

Evolution of aerospace engineering during the last decades includes among others extensive research on rocket launches [BiLi04]. During launch thrust is produced by burning solid or liquid fuel, where hot combustion products are released into the atmosphere. In particular, we are interested in thermal effects behind the nozzle exit, which is predominantly characterized by radiation and thermal conduction. In this context, we derive a solution for the radiative–conductive transfer problem in a co-moving cylindrical coordinate system. The solution allows to simulate the radiation and temperature field together with conductive and radiative energy transport originating from the exhaust released in the rocket launches. In general, the equation of radiative–conductive transfer in cylinder geometry is difficult to solve without introducing some approximations, such as linearization or discretizing angular terms, that turn the construction of an acceptably precise solution to an approximate problem feasible. Solutions found in the literature are typically determined by numerical means, see, for instance, [Li00, MiKrKi11].

In the sequel, we discuss a semi-analytical approach reducing the original equation, which is continuous in the angular variables, into an equation similar to the Cartesian $S_N$ radiative–conductive transfer problem, but considering cylinder geometry. The solution is constructed using a composite method by Laplace transform and Adomian decomposition method [Ad88]. The Laplace method gives way to use established procedures for linear problems, while the Adomian decomposition method allows to treat the nonlinear contribution as source term of a linear recursive

C.A. Ladeia (✉) • B.E.J. Bodmann • M.T.B. Vilhena
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: cibele_mat_uel@yahoo.com.br; bardo.bodmann@ufrgs.br; vilhena@mat.ufrgs.br



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_29

problem. This recursive scheme opens a pathway to determine a solution, in principle to any prescribed precision. It is noteworthy that this methodology was also successfully applied to attain the solution of the $S_N$ nodal equations in Cartesian geometry [BoViSe11, PaViHa02, PaViHa03].

## 29.2 The Radiative Conductive Transfer Equation in Cylindrical Geometry

We consider the one-dimensional problem in cylindrical geometry and assume that the problem is independent of time $t$. Further, the intensity is integrated over the entire spectrum. This problem of energy transfer is described in [Oz73] by the conductive-radiative transfer equation coupled with the energy equation,

$$\sqrt{1-\xi^2}\left[\gamma\frac{\partial \mathscr{I}(r,\xi,\gamma)}{\partial r}+\frac{1-\gamma^2}{r}\frac{\partial \mathscr{I}(r,\xi,\gamma)}{\partial \gamma}\right]+\mathscr{I}(r,\xi,\gamma)$$
$$=\frac{\omega(r)}{2}\int_0^1\int_{-1}^1 \mathscr{P}(\xi,\xi')\mathscr{I}(r,\xi',\gamma')d\xi'\frac{d\gamma'}{\sqrt{1-\gamma'^2}}+(1-\omega(r))\,\Theta^4(r). \quad (29.1)$$

Here, $\mathscr{I}$ is the radiation intensity, $\omega$ is the single scattering albedo and $\mathscr{P}(\xi)$ signifies the differential scattering coefficient, also called the phase function. The integral on the right-hand side of (29.1) can be written as

$$\int_{-1}^1 \mathscr{P}(\xi,\xi')\mathscr{I}(r,\xi',\gamma')d\xi'=\sum_{l=0}^{\infty}\beta_l\int_{-1}^1 \mathscr{P}_l(\xi)\mathscr{P}_l(\xi')\mathscr{I}(r,\xi',\gamma')d\xi',$$

where the summation index refers to the degree of anisotropy, for details see [BoViSe11]. The energy equation for the temperature that connects the radiative flux to a temperature gradient is

$$r\frac{d^2}{dr^2}\Theta(r)+\frac{d}{dr}\Theta(r)=\frac{1}{4\pi N_c}\frac{d}{dr}[rq_r^*]\,. \quad (29.2)$$

Here, $N_c$ is the conduction-radiation parameter

$$N_c=\frac{k\beta_{ext}}{4\sigma n^2 T_r^3}\,,$$

with $k$ the thermal conductivity, $\beta_{ext}$ the extinction coefficient, $\sigma$ the Stefan-Boltzmann constant and $n$ the refractive index. The dimensionless radiative flux is expressed in terms of the intensity by

$$q_r^*=4\int_0^1\int_{-1}^1 I(r,\xi',\gamma')d\xi'\frac{d\gamma'}{\sqrt{1-\gamma'^2}}\,.$$

The boundary conditions of equation (29.1) are

$$\mathscr{I}(r,\xi,\gamma)|_{r\in\{0,R\}} = \varepsilon(r)\Theta^4(r) + \rho^d(r)\int_0^1\int_{-1}^1 \mathscr{I}(r,\xi',\gamma')d\xi'\frac{d\gamma'}{\sqrt{1-\gamma'^2}}\bigg|_{r\in\{0,R\}},$$

where $\rho^d$ is the diffuse reflectivity, $\varepsilon$ is the emissivity and the boundary conditions of equation (29.2) are

$$\frac{d}{dr}\Theta(r)\bigg|_{r=0} = \Theta_T \quad \text{and} \quad \Theta(r)|_{r=R} = \Theta_B\,.$$

## 29.3 Solution by the Decomposition Method

The equations (29.1) and (29.2) can be simplified using a discrete countable set of angles following the collocation method, which defines the problem of radiative–conductive transfer in cylindrical geometry in the so-called $S_N$ approximation extended by an additional angular variable $\mathscr{I}_{n,m}(r,\xi_n,\gamma_{n,m})$ and represented by the following equations.

$$\gamma_{n,m}\frac{\partial\mathscr{I}_{n,m}}{\partial r} + \left(\frac{1-\gamma_{n,m}^2}{r}\right)\frac{\partial\mathscr{I}_{n,m}}{\partial\gamma} + \frac{1}{\sqrt{1-\xi_n^2}}\mathscr{I}_{n,m} = \tag{29.3}$$

$$= \frac{\omega(r)}{\sqrt{1-\xi_n^2}}\sum_{l=0}^{L}\beta_l\mathscr{P}_l(\xi_n)\sum_{p=1}^{N/2}\sum_{q=1}^{(N/2)-n+1}\varpi_{p,q}\mathscr{P}_l(\xi_p)\mathscr{I}_{p,q} + \frac{(1-\omega(r))}{\sqrt{1-\xi_n^2}}\Theta^4(r)\,,$$

$$\frac{d}{dr}\Theta(r) - \frac{d}{dr}\Theta(r)\bigg|_{r=0} = \frac{1}{N_c}\sum_{p=1}^{N/2}\sum_{q=1}^{(N/2)-n+1}\varpi_{p,q}[\mathscr{I}_{p,q}(r) - \mathscr{I}_{p,q}(0)]\,, \tag{29.4}$$

where $\xi_n$ and $\gamma_{n,m}$ are evaluation points. A sketch of the cylindrical domain together with the definition of the discrete angular variable is shown in figure 29.1.

The integration is carried out over two octants with $1 \le n \le N/2$ and $1 \le m \le N$ and subject to the boundary conditions

$$\mathscr{I}_{n,m}(0) = \varepsilon(0)\Theta^4(0) + \rho^d(0)\sum_{p=1}^{N/2}\sum_{q=1}^{(N/2)-n+1}\varpi_{p,q}\mathscr{I}_{N/2-p+1,q}(0),$$

$$\mathscr{I}_{N/2-n+1,m}(R) = \varepsilon(R)\Theta^4(R) + \rho^d(R)\sum_{p=1}^{N/2}\sum_{q=1}^{(N/2)-n+1}\varpi_{p,q}\mathscr{I}_{p,q}(R),$$

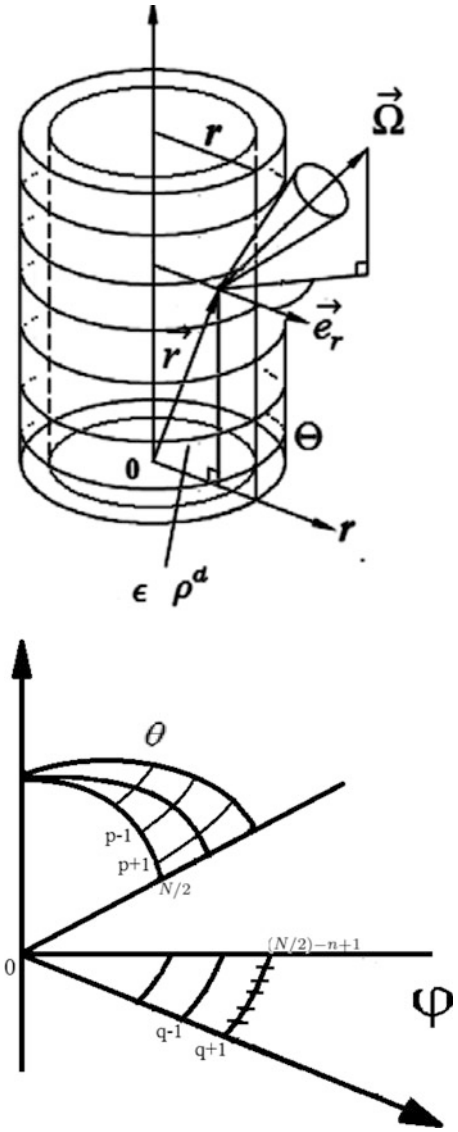


**Fig. 29.1** Representation of the physical domain in cylinder geometry.

Note that the integrals over the angular variables are replaced by a system of Gaussian quadrature with weights $\varpi_p$ using

$$\varpi_{p,q} = \pi \frac{\varpi_p}{N},$$

where weights $\varpi_p$ are normalized to one so that $\varpi_{p,q}$ is normalized to the solid angle of an octant.

$$\sum_{p=1}^{N/2} \varpi_p = 1 \quad \sum_{p=1}^{N/2} \sum_{q=1}^{N} \varpi_p = \pi$$

Here, $p$ indicates a discrete direction of $\xi_p$ and $q$ a discrete direction of $\gamma_{p,q}$, respectively. The equation system (29.3) and (29.4) may be cast in matrix representation

$$A \frac{d}{dr} \mathscr{I}_{p,q} + \left\{ B \frac{d}{d\gamma} \mathscr{I}_{p,q} \right\}_{\gamma=\gamma_p} - C \mathscr{I}_{p,q} = \Psi, \tag{29.5}$$

with $A = \gamma_{n,m}$, $B = (1 - \gamma_{n,m}^2)$ and $C$ is a square matrix,

$$C(i,j) = \begin{cases} \frac{1}{\sqrt{1-\xi_i}} + \frac{\omega_j(r)}{\sqrt{1-\xi_i}} \left[ \sum_{l=0}^{L} \beta_l P_l(\xi_i) P_l(\xi_j) \right] & \text{for } i = j \\ \frac{\omega_j(r)}{\sqrt{1-\xi_i}} \left[ \sum_{l=0}^{L} \beta_l P_l(\xi_i) P_l(\xi_j) \right] & \text{for } i \neq j. \end{cases}$$

The nonlinear terms are

$$\Psi = \left( \frac{(1-\omega(r))}{\sqrt{1-\xi_1^2}} \Theta^4(r), \ldots, \frac{(1-\omega(r))}{\sqrt{1-\xi_N^2}} \Theta^4(r) \right)^T .$$

For each direction $\gamma = \gamma_q$ in equation (29.5), the angular derivative term is discretized by a central difference scheme.

$$\left\{ B \frac{d}{d\gamma} \mathscr{I}_{p,q} \right\}_{\gamma=\gamma_q} \approx \frac{\alpha_{q+1/2} \mathscr{I}_{q+1/2} - \alpha_{q-1/2} \mathscr{I}_{q-1/2}}{\varpi_q}$$

where $\mathscr{I}_{q\pm1/2}$ are the angular intensities in the directions $q \pm 1/2$, and the central difference scheme is adopted to correlate them to the unknown $\mathscr{I}_q$, i.e., $\mathscr{I}_q = \frac{1}{2}(\mathscr{I}_{q+1/2} + \mathscr{I}_{q-1/2})$. The coefficients $\alpha_{q\pm1/2}$ result from azimuthal difference terms. These terms are chosen such so as to establish energy conservation, i.e., the integration of the term $\left( B \frac{d}{d\gamma} \mathscr{I}_{p,q} \right)$ over the whole azimuthal angle shall be equal zero. Details about the selection of $\alpha_{q\pm1/2}$ are explicitly given in ref. [LiOz91]. In shorthand notation equation (29.5) reads now

$$A \frac{d}{dr} \mathscr{I}_{p,q} - E \mathscr{I}_{p,q} = \Psi \,, \tag{29.6}$$

where $E = -[B + C]$ with $B$ a tridiagonal matrix.

According to Adomian's prescription [Ad88] the intensity of radiation is expanded in an infinite series:

$$\mathscr{I}_{p,q} = \sum_{l=0}^{\infty} Y_l \tag{29.7}$$

Equation (29.6) is then

$$\sum_{l=0}^{\infty}\left(A\frac{d}{dr}Y_l - EY_l\right) =$$

$$= \left(\frac{(1-\omega(r))}{\sqrt{1-\xi_1^2}},\ldots,\frac{(1-\omega(r))}{\sqrt{1-\xi_N^2}}\right)\sum_{l=0}^{\infty}\mathscr{A}_{l-1}\left(\{Y_l\}_{l=0}^{\infty}\right). \tag{29.8}$$

In order to solve the equation system (29.8) in a recursive fashion, initialization is chosen to be

$$A\frac{d}{dr}Y_0 - EY_0 = 0$$

together with the boundary conditions and then entering a recursive process of the equations for the remaining components $Y_l$,

$$A\frac{d}{dr}Y_l - EY_l = \left(\frac{(1-\omega(r))}{\sqrt{1-\xi_1^2}},\ldots,\frac{(1-\omega(r))}{\sqrt{1-\xi_N^2}}\right)\mathscr{A}_{l-1}\left(\{Y_l\}_{l=1}^{\infty}\right),$$

with $l = 1,2,\ldots,L$.

Upon applying the Laplace transformation in the radial variable in equation (29.9) together with the boundary conditions, one obtains the solution

$$Y_l(r) = \mathscr{L}^{-1}((sI-U)A^{-1}Y_l(0)) + \mathscr{L}^{-1}((sI-U)\overline{\Psi}(s)),$$

where $\mathscr{L}^{-1}$ denotes the inverse Laplace transformation operator, $s$ is a complex parameter, $U = A^{-1}E$ and the decomposed matrix $U = XDX^{-1}$, $D$ is the diagonal matrix with distinct eigenvalues and $X$ is the eigenvector matrix. Thus, the general solution is given by

$$Y_l(r) = Xe^{D_r}V^l +$$

$$+Xe^{D_r}X^{-1} * \mathscr{A}_{l-1}\left(\{Y_l\}_{l=1}^{\infty}\right)\left(\frac{(1-\omega(r))}{\sqrt{1-\xi_1^2}},\ldots,\frac{(1-\omega(r))}{\sqrt{1-\xi_N^2}}\right),$$

with $l = 0,1,2,\ldots,L$. The nonlinearity is represented by the term $\Theta^4(r)$ and will be represented by Adomian polynomials, given by

$$LY = \sum_{l=0}^{L}\hat{A}_l(r) = \Theta^4(r).$$

Note that one can write the nonlinear term in a generic fashion,

$$LY = \sum_{l=0}^{\infty} \hat{A}_l(r) = \sum_{l=0}^{\infty} \frac{1}{l!} \underbrace{\frac{\partial^l (LY)}{\partial Y^l}\bigg|_{Y=Y_0}}_{f_0^{(l)}} \left( \sum_{\nu=1}^{\infty} Y_\nu \right)^l$$

$$= \lim_{a\to\infty} \sum_{l=0}^{\infty} \frac{1}{l!} f_0^{(l)} \sum_{\substack{b_1,\ldots,b_a \\ \sum b_i = l}} \left( \binom{l}{\{b_i\}_1^a} \prod_{\nu=1}^{a} Y_\nu^{b_\nu} \right)$$

$$= f_0^{(0)} + \sum_{l=1}^{\infty} \left( f_0^{(1)} Y_l + \sum_{j=2}^{l} \frac{1}{j!} f_0^{(j)} \sum_{\substack{b_1,\ldots,b_{l-1} \\ \sum b_i = j}} \left( \binom{j}{\{b_i\}_1^{l-1}} \prod_{\nu=1}^{l-1} Y_\nu^{b_\nu} \right) \right), \quad (29.9)$$

where the notation $f_0^{(l)}$ for the $l$-th derivative at $Y = Y_0$. Already in the last line of equation (29.9) the terms are reorganized so that one identifies the first term $f_0^{(0)}$ and all the subsequent terms of the series that define the Adomian polynomials $\hat{A}_l$.

$$\hat{A}_0(r) = f_0^{(0)} = f(Y_0),$$

$$\hat{A}_1(r) = f_0^{(1)} Y_1 = Y_1 \frac{d}{dY_0} f(Y_0)$$

$$\hat{A}_2(r) = f_0^{(1)} Y_2 + f_0^{(2)} Y_1^2 = Y_2 \frac{d}{dY_0} f(Y_0) + \frac{Y_1^2}{2!} \frac{d}{dY_0} f(Y_0)$$

$$\vdots$$

$$\hat{A}_l(r) = f_0^{(1)} Y_l + \sum_{j=2}^{l} \frac{1}{j!} f_0^{(j)} \sum_{\substack{b_1,\ldots,b_{l-1} \\ \sum b_i = j}} \left( \binom{j}{\{b_i\}_1^{l-1}} \prod_{\nu=1}^{l-1} Y_\nu^{b_\nu} \right)$$

Accordingly, we apply the Adomian decomposition taking the binomial expansion of the term $\Theta^4(r)$

$$\Theta^4(r) = \left( \sum_{l=0}^{L} Y_l(r) \right)^4 = Y_0^4 + 4Y_0^3 \left( \sum_{i=0}^{L} Y_i(r) \right) +$$

$$+ \frac{12}{2!} Y_0^2 \left( \sum_{i=0}^{L} Y_i(r) \right)^2 + \frac{24}{3!} Y_0 \left( \sum_{i=0}^{L} Y_i(r) \right)^3 + \frac{24}{4!} \left( \sum_{i=0}^{L} Y_i(r) \right)^4 .$$

## 29.4 Numerical Results

Below, we implement a fictitious scenario to test consistency of the proposed method. To this end we determine characteristic quantities for the radiative–conductive transfer problem, which are the profiles of the temperature, the radiative, conductive and total heat fluxes, respectively.

$$Q_r(r) = \frac{1}{4\pi N_c} q_r^*, \quad Q_c(r) = -\frac{d}{dr}\Theta(r)$$
$$Q(r) = Q_r(r) + Q_c(r)$$

All the calculations that follow are based on the parameter set given in table 29.1. As already mentioned at the present stage of the work the parameters $\omega$, $\varepsilon$, $\rho$ and $N_c$ are somehow arbitrary but in a subsequent work we will correlate these parameters to the concentration of pollutant in the exhaust that allows to model the near field, the source term for pollutant dispersion during rocket launch. We further use a normalized temperature distribution:

The numerical results for the profile of the temperature, for the conductive heat flux $(Q_c)$, the radiative flux $(Q_r)$ and the total flux $(Q_t)$ along the radial optical depth are shown in Figures 29.2, 29.3, 29.4 and 29.5, respectively, where we consider $r$ in units of $r/R$ that varies between 0 and 1.

**Table 29.1** Parameters of the Problem.

| $\omega(r)$ | $\varepsilon$ | $\rho$ | $N_c$ |
|---|---|---|---|
| 0.92 | 0.8 | 0.2 | 0.05 |

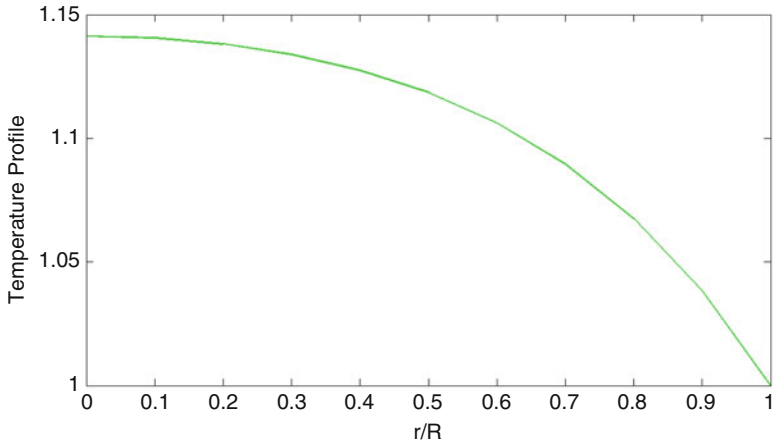


**Fig. 29.2** Temperature profile along the radial optical depth.

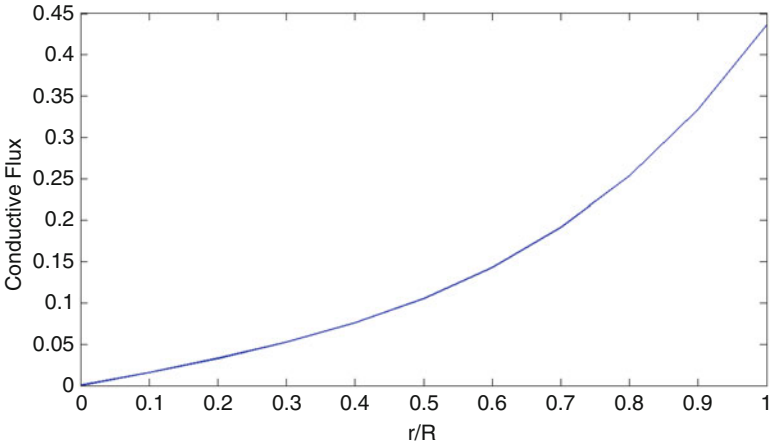


**Fig. 29.3** Conductive heat flux along the radial optical depth.

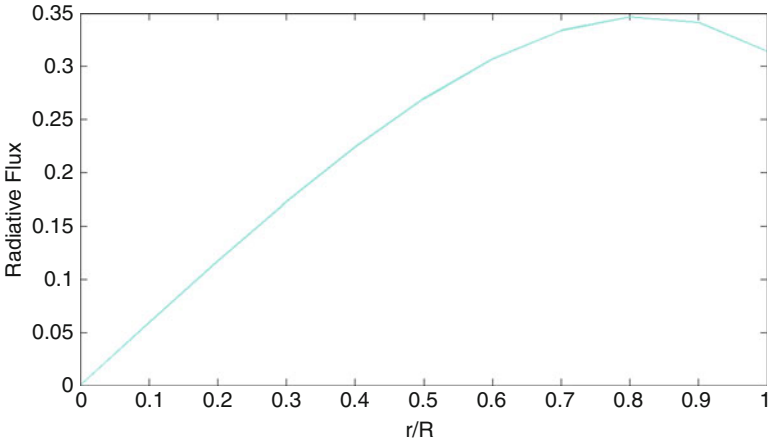


**Fig. 29.4** Radiative heat flux along the radial optical depth.

## 29.5 Conclusions

The present study demonstrates a novel procedure to solve the radiative–conductive transfer equation in cylindrical geometry approximated in form of the doubly discrete ordinate representation analogue to the $S_N$ equation. The original nonlinear problem was decomposed in a recursive scheme of equation systems similar to the decomposition description by Adomian. The initialization of the recursion is a linear equation system with known solution. All the subsequent equation systems to be solved are of linear type, where the nonlinearity appears as source term but containing only terms with the solutions from all previous solutions.

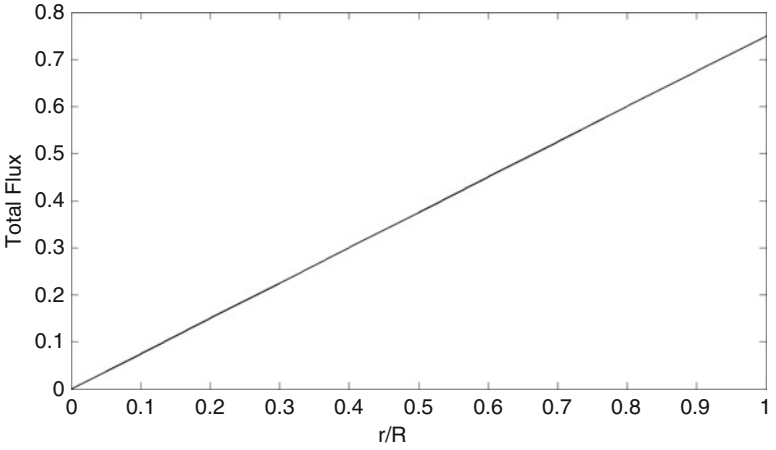


**Fig. 29.5** The total heat flux against the radial optical depth.

The recursive scheme is manifest exact in the infinite recursion depth limit, so that truncation at an adequate finite depth yields an acceptable solution to the approximate problem. We have not discussed the important issue of convergence, however, so far various trials have shown us that for polynomial nonlinearities the recursive scheme converges already for a small recursion depth. A rigorous proof of convergence is in preparation and will be discussed in future work.

From the physical point of view, we have chosen only an arbitrary set of parameter for which we found that the total heat flux increases linearly with increasing radial optical depth as expected. However, future investigations have to show what are the adequate correlations between the physical parameters emissivity, reflectivity and albedo among others, that can be related to the density or pollutant concentration profile in the exhaust released in rocket launch. Such a relation is essential since existing pollutant dispersion models are characterized by the absence of thermal properties of the source, so that an extension of the present study in this direction could open pathways to reduce this shortcoming. In this context the present approach is a first step in this direction.

## References

[Ad88] Adomian, G.: A review of the decomposition method in applied mathematics. J. Math. Anal. Appl. **135**, 501–544 (1988)

[BiLi04] Bille, M., Lishock, E.: The First Space Race: Launching the World's First Satellites Texas A&M University Press. (2004)

[BoViSe11] Bodmann, B., Vilhenna, M.T., Segatto, C.F.: Non-Linear Radiactive-Conductive Heat Transfer in a Heterogeneons Gray Plane-Parallel Participating Medium. In: Amimul Ahsan.(Org.) Heat Transfer: InTech, **1**, 177–196 (2011)

[Li00] Li, H.-Y.: A Two-dimensional Cylindrical Inverse Source Problem in Radiative Transfer. J. Quant. Spectrosc. Radiat. Transfer. **69**, 403–414 (2000)

[LiOz91] Li, H.Y., Ozisik, M.N.: Simultaneous conduction and radiation in a two-dimensional participating cylinder with anisotropic scattering. J. Quant. Spectrosc. Radiative Transfer **46**, 393–404 (1991)

[MiKrKi11] Mishra, S. C., Krishna, Ch. H., Kim, M. Y.: Analysis of Conduction and Radiation Heat Transfer in a 2-D Cylindrical Medium Using the Modified Discrete Ordinate Method and the Lattice Bolztmann Method. Numerical Heat Transfer. **60**, 254–287 (2011)

[PaViHa02] Pazos, R. P., Vilhena, M.T., Hauser E. B.: Solution and study of two-dimensional nodal neutron transport equation. In. Proccedings of $10^{th}$ International Conference of Nuclear Engineering, **1**, Arlington (2002)

[PaViHa03] Pazos, R. P., Vilhena, M.T., Hauser E. B.: Advances in the solution of threedimensional nodal neutron transport equation. In. $11^{th}$ International Conference of Nuclear Engineering, Tokyo (2003)

[Oz73] Ozisik, M.N.: Radiative Transfer and Interaction with Conductions and Convection. John Wiley & Sons Inc., New York (1973)

# Chapter 30
# A Functional Analytic Approach to Homogenization Problems

**M. Lanza de Cristoforis and P. Musolino**

## 30.1 Introduction and Statement of the Problem

We plan to illustrate a functional analytic approach to analyze homogenization problems, which has already been developed for singular perturbation problems in bounded domains with small holes (cf. *e.g.*, [La02, La08, La10, La12].) In the frame of linearized elastostatics and of the Stokes equations, we mention [DaLa10, DaLa11], and [Da13]. Later on, such an approach has been exploited for the analysis of problems in unbounded perforated domains with a fixed periodic structure, for example in [LaMu13, LaMu14, Mu12]. Instead, here we consider the case where also the size of the periodicity cell tends to zero. The results in this chapter are based on the work of [LaMu].

We consider a simple linear model problem, which we now introduce. We fix

$$n \in \mathbb{N} \setminus \{0,1\}\,,$$

and introduce a periodicity cell

$$Q \equiv ]0,1[^n\,.$$

Clearly, $\mathbb{Z}^n$ is the set of vertices of a periodic subdivision of $\mathbb{R}^n$ corresponding to the fundamental cell $Q$.

We plan to perform a periodic set of perforations in $\mathbb{R}^n$. To do so, we fix a point $p \in Q$.

M. Lanza de Cristoforis (✉) • P. Musolino
Dipartimento di Matematica, Università degli Studi di Padova,
Via Trieste 63, 35121 Padova, Italy
e-mail: mldc@math.unipd.it; musolino@math.unipd.it



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_30

Then we fix

$$\alpha \in ]0,1[\,.$$

For the definition of functions of class $C^{0,\alpha}$ or $C^{1,\alpha}$ in the closure or on the boundary of an open set, and for the norm on the corresponding Schauder spaces, we refer to Gilbarg and Trudinger [GiTr83] (see also [Mu12, §2] for the periodic case.)

Next we select a subset $\Omega$ of $\mathbb{R}^n$ satisfying the following assumption.

> *Let $\Omega$ be a bounded open connected subset of $\mathbb{R}^n$ of class $C^{1,\alpha}$.*
>
> *Let $\mathbb{R}^n \setminus \mathrm{cl}\Omega$ be connected. Let $0 \in \Omega$.*

Then there exists $\varepsilon_0 \in ]0,+\infty[$ such that

$$p + \varepsilon \mathrm{cl}\Omega \subseteq Q \qquad \forall \varepsilon \in ]-\varepsilon_0, \varepsilon_0[\,,$$

where cl denotes the closure. To shorten our notation, we set

$$\Omega_{p,\varepsilon} \equiv p + \varepsilon\Omega \qquad \forall \varepsilon \in \mathbb{R}\,.$$

Then we introduce the periodic domains

$$\mathbb{S}[\Omega_{p,\varepsilon}] \equiv \bigcup_{z \in \mathbb{Z}^n} (z + \Omega_{p,\varepsilon})\,,$$

$$\mathbb{S}[\Omega_{p,\varepsilon}]^- \equiv \mathbb{R}^n \setminus \mathrm{cl}\mathbb{S}[\Omega_{p,\varepsilon}]\,,$$

for all $\varepsilon \in ]-\varepsilon_0, \varepsilon_0[$. Then a function $u$ from $\mathrm{cl}\mathbb{S}[\Omega_{p,\varepsilon}]^-$ to $\mathbb{R}$ is $q$-periodic if

$$u(x + e_h) = u(x) \qquad \forall x \in \mathrm{cl}\mathbb{S}[\Omega_{p,\varepsilon}]^-\,,$$

for all $h \in \{1,\dots,n\}$. Here $\{e_1,\dots, e_n\}$ denotes the canonical basis of $\mathbb{R}^n$. Next we introduce a second parameter $\delta \in ]0,+\infty[$ and we consider the rescaled periodic domains

$$\mathbb{S}(\varepsilon,\delta)^- \equiv \delta\mathbb{S}[\Omega_{p,\varepsilon}]^-\,, \qquad \mathbb{S}(\varepsilon,\delta) \equiv \delta\mathbb{S}[\Omega_{p,\varepsilon}]\,,$$

which are associated with the periodicity cell $\delta Q$. We say $\delta q$-periodic the functions which are periodic with respect to the cell $\delta Q$.

We now turn to introduce the data of our problem. Let $f$ be a $q$-periodic real analytic function from $\mathbb{R}^n$ to $\mathbb{R}$ such that

$$\int_Q f\,dx = 0\,.$$

Let

$$g \in C^{1,\alpha}(\partial\Omega)\,.$$

Then we consider the Dirichlet problem

$$\begin{cases} \Delta u(x) = f(\delta^{-1}x) & \forall x \in \mathbb{S}(\varepsilon,\delta)^-\,, \\ u \text{ is } \delta q - \text{periodic in } \mathrm{cl}\mathbb{S}(\varepsilon,\delta)^-\,, & \\ u(x) = g(\delta^{-1}\varepsilon^{-1}(x-\delta p)) & \forall x \in \delta\partial\Omega_{\varepsilon,p}\,, \end{cases} \tag{30.1}$$

for each $(\varepsilon,\delta) \in ]0,\varepsilon_0[\times]0,+\infty[$. As is well known, for each $(\varepsilon,\delta)$ in the set $]0,\varepsilon_0[\times]0,+\infty[$ problem (30.1) has one and only one solution $u(\varepsilon,\delta,\cdot)$ in the space

$$C^{1,\alpha}_{\delta q}(\mathrm{cl}\mathbb{S}(\varepsilon,\delta)^-)$$

of $\delta q$-periodic functions of class $C^{1,\alpha}$ in the domain $\mathrm{cl}\mathbb{S}(\varepsilon,\delta)^-$. We are interested into the behavior of $u(\varepsilon,\delta,\cdot)$ and of its energy integral as $(\varepsilon,\delta)$ degenerates to $(0,0)$. Most of the results in the vast literature on homogenization problems aim at computing the limiting behavior as $(\varepsilon,\delta)$ degenerates to $(0,0)$, or at writing asymptotic expansions.

Here instead, we wish to represent $u(\varepsilon,\delta,\cdot)$ or its energy integral in terms of real analytic maps and in terms of possibly singular at $\varepsilon = 0$, $\delta = 0$, but known functions of $\varepsilon$, $\delta$ in the same spirit of the papers cited at the beginning of the present section.

The chapter is organized as follows. In the next section 30.2, we analyze the behavior of $u(\varepsilon,\delta,\cdot)$. In the following section 30.3, we analyze the behavior of its energy integral.

## 30.2 Analysis of the Solution of Problem (30.1) as $(\varepsilon,\delta)$ Degenerates to $(0,0)$

Let $(\varepsilon,\delta) \in ]0,\varepsilon_0[\times]0,+\infty[$. Then the solution $u(\varepsilon,\delta,\cdot)$ of (30.1) is defined on the domain $\mathrm{cl}\mathbb{S}(\varepsilon,\delta)^-$, which depends upon $(\varepsilon,\delta)$. In order to study the dependence of $u(\varepsilon,\delta,\cdot)$ upon $(\varepsilon,\delta)$, we find convenient to deal with a domain which does not depend upon $(\varepsilon,\delta)$. One way is to extend each function defined on $\mathrm{cl}\mathbb{S}(\varepsilon,\delta)^-$ to be zero outside of $\mathrm{cl}\mathbb{S}(\varepsilon,\delta)^-$. Thus if $v$ is a function from $\mathrm{cl}\mathbb{S}(\varepsilon,\delta)^-$ to $\mathbb{R}$, we denote by $\mathbf{E}_{(\varepsilon,\delta)}[v]$ the function from $\mathbb{R}^n$ to $\mathbb{R}$ defined by

$$\mathbf{E}_{(\varepsilon,\delta)}[v](x) \equiv \begin{cases} v(x) & \forall x \in \mathrm{cl}\mathbb{S}(\varepsilon,\delta)^-\,, \\ 0 & \forall x \in \mathbb{R}^n \setminus \mathrm{cl}\mathbb{S}(\varepsilon,\delta)^-\,. \end{cases}$$

Then we can prove the following well-known 'classical' result.

**Theorem 1.** *There exists a constant $\tilde{c}$ which depends only on $\Omega$ and $g$ such that*

$$\lim_{j\to\infty} \mathbf{E}_{(\varepsilon_j,\delta_j)}[u(\varepsilon_j,\delta_j,\cdot)] = \tilde{c} \qquad \text{in } L^r(V)\,,$$

*for all bounded open subsets $V$ of $\mathbb{R}^n$, and for all $r \in [1,+\infty[$, and for all sequences $\{(\varepsilon_j,\delta_j)\}_{j\in\mathbb{N}}$ in $]0,\varepsilon_0[\times]0,+\infty[$ which converge to $(0,0)$.*

However, in the spirit of this chapter, we want to describe the behavior of the function $\mathbf{E}_{(\varepsilon,\delta)}[u(\varepsilon,\delta,\cdot)]$ when $(\varepsilon,\delta)$ is close to $(0,0)$ by means of analytic maps of $(\varepsilon,\delta)$.

As a first step, we can try to do so in a 'weak form.' More precisely, we can try to describe the behavior of the function

$$(\varepsilon,\delta) \mapsto \int_{\mathbb{R}^n} \mathbf{E}_{(\varepsilon,\delta)}[u(\varepsilon,\delta,\cdot)]\phi\,dx\,,$$

for *all* $\phi \in L^{r'}(\mathbb{R}^n)$ with compact support. Here $(1/r)+(1/r')=1$.

At the moment however, we cannot do so for *all* elements $\phi$ of $L^{r'}(\mathbb{R}^n)$ with compact support, but only for all the elements $\phi$ which belong to a certain dense subspace $\mathcal{T}_q$ of $L^{r'}(\mathbb{R}^n)$ which we now turn to introduce by means of the following.

**Proposition 1.** *The vector subspace $\mathcal{T}_q$ of $L^\infty(\mathbb{R}^n)\cap L^1(\mathbb{R}^n)$ generated by the set of functions*

$$\left\{\chi_{y+sQ} : (y,s) \in \mathbb{R}^n \times (\mathbb{Q}\cap]0,+\infty[)\right\}\,,$$

*is dense in $L^{r'}(\mathbb{R}^n)$ for all $r' \in [1,+\infty[$.*

Then we can prove the following.

**Theorem 2.** *Let $\phi \in \mathcal{T}_q$. Then there exist $\varepsilon'$, $\delta'$, $s \in ]0,+\infty[$ and a real analytic map*

$$H : ]-\varepsilon',\varepsilon'[\times]-\delta',\delta'[\to\mathbb{R}\,,$$

*such that*

$$\int_{\mathbb{R}^n} \mathbf{E}_{(\varepsilon,l^{-1}s)}[u(\varepsilon,l^{-1}s,\cdot)]\phi\,dx = s^n H[\varepsilon,l^{-1}s]$$

*for all $l \in \mathbb{N}\setminus\{0\}$ such that $l > s/\delta'$ and for all $\varepsilon \in ]0,\varepsilon'[$.*

As a consequence, we can expand the integral

$$\int_{\mathbb{R}^n} \mathbf{E}_{(\varepsilon,l^{-1}s)}[u(\varepsilon,l^{-1}s,\cdot)]\phi\,dx$$

into a convergent expansion of powers of $\varepsilon$ and $l^{-1}s$ for $\varepsilon > 0$ small enough and for $l \in \mathbb{N}\setminus\{0\}$ large enough.

## 30.3 Analysis of the Energy Integral of the Solution of Problem (30.1) as $(\varepsilon, \delta)$ Degenerates to $(0,0)$

We now consider the energy integral of the solution $u(\varepsilon, \delta, \cdot)$ in the periodic cell $Q$. Namely, we are interested in the behavior of the integral

$$\mathrm{En}[\varepsilon, \delta] \equiv \int_{Q \cap \mathbb{S}(\varepsilon,\delta)^-} |D_x u(\varepsilon, \delta, x)|^2 \, dx,$$

as $(\varepsilon, \delta)$ approaches $(0,0)$, and we have the following result, which in the spirit of the present chapter describes the behavior of $\mathrm{En}[\varepsilon, \delta]$ in terms of analytic functions of $(\varepsilon, \delta)$ evaluated on a discrete set of values of $\delta$.

**Theorem 3.** *There exist $\varepsilon_e \in ]0, \varepsilon_0[$, $\delta' \in ]0, +\infty[$, and $l_e \in \mathbb{N} \setminus \{0\}$, and real analytic functions*

$$\mathscr{E}^\sharp : ]-\varepsilon_e, \varepsilon_e[ \times ]-\delta', \delta'[ \to \mathbb{R}$$
$$\mathscr{P}^\sharp : ]-\varepsilon_e, \varepsilon_e[ \to \mathbb{R}$$

*such that*

$$\mathrm{En}[\varepsilon, l^{-1}] = l^2 \left\{ \mathscr{E}^\sharp[\varepsilon, l^{-1}] \varepsilon^{n-2} + l^{-4} \mathscr{P}^\sharp[\varepsilon] \right\},$$

*for all $\varepsilon \in ]0, \varepsilon_e[$ and $l \in \mathbb{N} \setminus \{0\}$ such that $l \geq l_e$.*

In particular, we can expand the term in braces into a convergent expansion of powers of $\varepsilon$ and $l^{-1}$ for $\varepsilon > 0$ small enough and for $l \in \mathbb{N} \setminus \{0\}$ large enough.

We note that the coefficient of the term in braces is $l^2$, which diverges as $l$ tends to infinity. However, if $n \geq 3$ and if we choose $\varepsilon = l^{-\frac{2}{n-2}}$, we can prove a convergence result for the energy integral. However, to do so, we need to introduce the following.

**Lemma 1.** *Let $n \geq 3$. Le $\tilde{c}$ be the constant of Theorem 1. Then there exists a unique function $\tilde{u}$ in $C^{1,\alpha}_{\mathrm{loc}}(\mathbb{R}^n \setminus \Omega)$ which solves the 'limiting' exterior Dirichlet problem*

$$\begin{cases} \Delta u(x) = 0 & \forall x \in \mathbb{R}^n \setminus \mathrm{cl}\Omega, \\ u(x) = g(x) & \forall x \in \partial\Omega, \\ \lim_{x \to \infty} u(x) = \tilde{c}. \end{cases} \tag{30.2}$$

Then we can state the following.

**Corollary 1.** *Let $n \geq 3$. There exist $\tilde{\varepsilon} \in ]0, \varepsilon_e[$, and $\tilde{l} \in \mathbb{N} \setminus \{0\}$, and a real analytic function*

$$\mathscr{F} : ]-\tilde{\varepsilon}, \tilde{\varepsilon}[ \to \mathbb{R}$$

*such that*

$$\mathrm{En}[l^{-\frac{2}{n-2}}, l^{-1}] = \mathscr{F}[l^{-\frac{1}{n-2}}]$$

*for all* $l \in \mathbb{N} \setminus \{0\}$ *such that* $l \geq \tilde{l}$. *Moreover,*

$$\mathscr{F}[0] = \int_{\mathbb{R}^n \setminus \mathrm{cl}\Omega} |D\tilde{u}|^2 \, dx,$$

*where* $\tilde{u}$ *is as in Lemma 1.*

As a consequence, we can expand the integral

$$\mathrm{En}[l^{-\frac{2}{n-2}}, l^{-1}]$$

into a convergent expansion of powers of $l^{-\frac{1}{n-2}}$ for $l \in \mathbb{N} \setminus \{0\}$ large enough.

We note that the criticality of the exponent $\frac{2}{n-2}$ has been observed a long time ago by Marčenko and Khruslov [MaKh74] and by Cioranescu and Murat [CiMu82a, CiMu82b] for related problems (see also Maz'ya and Movchan [MaMo10], where the assumption of periodicity of the array of holes is relaxed.) Here, we can deduce by our results, the following corollary, which is in the spirit of those papers, and that no doubt could be proved with those methods.

**Corollary 2.** *Let* $n \geq 3$. *Assume that the boundary datum* $g$ *of problem (30.1) is not a constant function. Let* $h \in ]0, +\infty[$. *Then*

$$\lim_{\delta \to 0} \mathrm{En}[\delta^h, \delta] = \begin{cases} 0 & \text{if } h > \frac{2}{n-2}, \\ \int_{\mathbb{R}^n \setminus \mathrm{cl}\Omega} |D\tilde{u}|^2 \, dx & \text{if } h = \frac{2}{n-2}, \\ +\infty & \text{if } h < \frac{2}{n-2}, \end{cases}$$

*where* $\tilde{u}$ *is the unique solution in* $C^{1,\alpha}_{\mathrm{loc}}(\mathbb{R}^n \setminus \Omega)$ *of the 'limiting' exterior Dirichlet problem (30.2).*

**Acknowledgements** The authors acknowledge the support of "Progetto di Ateneo: Singular perturbation problems for differential operators, CPDA120171/12," University of Padova, and of Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

## References

[CiMu82a] Cioranescu, D., Murat, F.: Un terme étrange venu d'ailleurs. In Nonlinear partial differential equations and their applications. Collège de France Seminar, Vol. II (Paris, 1979/1980), volume **60** of Res. Notes in Math., 98–138, 389–390. Pitman, Boston, Mass. (1982)

[CiMu82b] Cioranescu, D., Murat, F.: Un terme étrange venu d'ailleurs. II. In Nonlinear partial differential equations and their applications. Collège de France Seminar, Vol. III (Paris, 1980/1981), volume **70** of Res. Notes in Math., 154–178, 425–426. Pitman, Boston, Mass. (1982)

[Da13] Dalla Riva, M.: Stokes flow in a singularly perturbed exterior domain. Complex Var. Elliptic Equ. **58**, 231–257 (2013)

[DaLa10] Dalla Riva, M., Lanza de Cristoforis, M.: Microscopically weakly singularly perturbed loads for a nonlinear traction boundary value problem: a functional analytic approach. Complex Var. Elliptic Equ. **55**, 771–794 (2010)

[DaLa11] Dalla Riva, M., Lanza de Cristoforis, M.: Weakly singular and microscopically hypersingular load perturbation for a nonlinear traction boundary value problem: a functional analytic approach. Complex Anal. Oper. Theory **5**, 811–833 (2011)

[GiTr83] Gilbarg, D., Trudinger, N.S.: Elliptic partial differential equations of second order, Springer Verlag, Berlin (1983)

[La02] Lanza de Cristoforis, M.: Asymptotic behaviour of the conformal representation of a Jordan domain with a small hole in Schauder spaces. Comput. Methods Funct. Theory **2**, 1–27 (2002)

[La08] Lanza de Cristoforis, M.: Asymptotic behavior of the solutions of the Dirichlet problem for the Laplace operator in a domain with a small hole. A functional analytic approach. Analysis (Munich) **28**, 63–93 (2008)

[La10] Lanza de Cristoforis, M.: Asymptotic behaviour of the solutions of a non-linear transmission problem for the Laplace operator in a domain with a small hole. A functional analytic approach. Complex Var. Elliptic Equ. **55**, 269–303 (2010)

[La12] Lanza de Cristoforis, M.: Simple Neumann eigenvalues for the Laplace operator in a domain with a small hole. A functional analytic approach. Rev. Mat. Complut. **25**, 369–412 (2012)

[LaMu13] Lanza de Cristoforis, M., Musolino, P.: A singularly perturbed nonlinear Robin problem in a periodically perforated domain: a functional analytic approach. Complex Var. Elliptic Equ. **58**, 511–536 (2013)

[LaMu14] Lanza de Cristoforis, M., Musolino, P.: A quasi-linear heat transmission problem in a periodic two-phase dilute composite. A functional analytic approach. Commun. Pure Appl. Anal. **13**, 2509–2542 (2014)

[LaMu] Lanza de Cristoforis, M., Musolino, P.: Two-parameter anisotropic homogenization for a Dirichlet problem for the Poisson equation in an unbounded periodically perforated domain. A functional analytic approach. Typewritten manuscript (2014)

[MaKh74] Marčenko, V.A., Khruslov, E.Y.: Boundary value problems in domains with a fine-grained boundary. Izdat. "Naukova Dumka", Kiev (1974) (in Russian)

[MaMo10] Maz'ya, V., Movchan, A.: Asymptotic treatment of perforated domains without homogenization. Math. Nachr. **283**, 104–125 (2010)

[Mu12] Musolino, P.: A singularly perturbed Dirichlet problem for the Laplace operator in a periodically perforated domain. A functional analytic approach. Math. Methods Appl. Sci. **35**, 334–349 (2012)

# Chapter 31
# Anisotropic Fundamental Solutions for Linear Elasticity and Heat Conduction Problems Based on a Crystalline Class Hierarchy Governed Decomposition Method

**T.V. Lisboa, R.J. Marczak, B.E.J. Bodmann, and M.T.M.B. Vilhena**

## 31.1 Introduction

Fundamental solutions play an essential role in numerical methods such as the Boundary Elements Method (BEM) and Fundamental Solutions Method (FSM). Important properties and their efficiency come from these solutions, which can be the response of an infinite or semi-infinite domain subject to a point load and submitted to radiation boundary conditions (Sommerfeld Type).

The influence of material on mechanical and thermal responses of structures has been studied for decades. The generalized Hooke's Law describes a stress-strain linear relationship in elastic solids, in which several crystals and metals can be included. Due to the internal symmetries and to the crystalline lattice form, all the elastic materials can be arranged into eight different symmetry groups ([CoMe95, ChVi01]). The Hooke's Law counterpart in heat conduction in solids is known as Fourier's Law, which describes a relationship between the heat flux and the temperature distribution. Elasticity problems can be expressed in a similar way and can be divided into three symmetric groups.

Based on Smith et al. work [SmSm63], Tu [Tu68] has presented an additive decomposition of the fourth-order and second-order flexibility and constitutive tensors, where the symmetries are the criteria for their superposition. Another additive decomposition has been presented by Browaeys & Chevrot [BrCh04]. A constitutive tensor vectorial representation has been used and, with specific matrices, this vector was operated to change its symmetry.

T.V. Lisboa (✉) • R.J. Marczak • B.E.J. Bodmann • M.T.M.B. Vilhena
Federal University of Rio Grande do Sul, Av. Rua Sarmento Leite 425,
Porto Alegre 90050-170, RS, Brazil
e-mail: taleslisboa@daad-alumni.de; rato@mecanica.ufrgs.br; bardo.bodmann@ufrgs.br; vilhena@math.ufrgs.br



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_31

The constitutive symmetries' significance to fundamental solutions lies on the additional complexity, or even the impossibility, of their determination in an analytical closed form. Many researchers have worked on ways to develop these solutions and have presented the use of several analytical and numerical methods to their determination: derivation through the Fourier Transform ([MaDe11]), the Radon Transform ([MaDe11, BuMa14]), the use of auxiliary tensors ([NaTu97, LiSu07]), Stroh's Formalism ([Ti96, ChRe02, BuOr10]), the direct integration of the equations, other complex-variable formalisms [BuMa14], etc.

For three-dimensional elasticity, only two analytical closed form fundamental solutions are known and both are for the simpler materials symmetry: isotropy and transverse homogeneity ([NaTu97, Ti96, Wa97]). Furthermore, internal relations of material properties can originate singularities in anisotropic fundamental solutions, which cause their non-reduction to simpler/higher symmetry cases. This effect is called degeneracy and almost all known anisotropic fundamental solutions can develop it. This degeneracy is related to the differential operator's characteristic equation roots and their associated multiplicity [NaTu97].

Wang [Wa97] has presented three-dimensional fundamental solutions using an integral representation of the Dirac delta distribution together with the residual theorem integration scheme. The derivation process of the fundamental solutions and its first and second derivative has been presented, however not explicitly. Tonon et al. [ToPa01] have used the aforementioned methodology and have presented numerical applications of these solutions in a BEM code. Results were presented for a transversally homogeneous symmetry and have shown good correlations to solutions known in the literature.

Buroni et al. [BuOr10] have presented a methodology using the residual integration theorem scheme to avoid the problem of multiple roots in the integrand's denominator, as done by Wang [Wa97], and have used the second Barnett–Lothe tensor, as done by Távara et al. [TaOr08]. The three-dimensional fundamental solution can be written in a unique formula for all constitutive symmetries, in which the roots multiplicity enters as a parameter for differentiation.

Two anisotropic thin plate fundamental solutions have been introduced by Shi & Bezini [ShBe88]: one for a fully anisotropic material and one for a degenerated case similar to cubic materials. So far, both fundamental solutions cannot be reduced to the isotropic fundamental solution [PaSo02]. As a result, for example, in a BEM code that solves thin plates, three different fundamental solutions are needed.

Two new methods to determine fundamental solutions for anisotropic heat conduction have been proposed by Marczak & Denda [MaDe11]. A new anisotropic fundamental solution has been determined and it has been expressed by a line integral over a unit circle. Buroni et. al [BuMa14] have developed a new complex variable formalism that, along with the Radon Transform, has been produced to analyze heat conduction problems in homogeneous anisotropic solids. For the first time, fundamental solutions for infinite media, half-space, and bi-material system due to heat dipole sources have been developed.

The Adomian decomposition [Ad98] is a recursive methodology to solve, mainly, nonlinear ordinary differential equations. Its three central ideas are: the

differential operator's splitting in a linear, remainder and nonlinear terms; an infinite superposition of the response in which each term is obtained via the previous ones; a nonlinear terms interpolation via polynomial functions known as Adomian polynomials. Removing the nonlinear part, the constitutive tensor's additive decomposition herein presented (Tu [Tu68] and Browaeys & Chevrot [BrCh04]) is exactly what is according to Adomian decomposition prescription with respect to the split of the differential operators.

In this chapter, the constitutive tensor, both for heat conduction and elasticity, is additively decomposed using the hierarchy proposed by Cowin & Mehrabadi [CoMe95] and Chadwick et. al [ChVi01] as a criterion. The decomposition methodology can be the one proposed by Tu [Tu68], Browaeys & Chevrot [BrCh04] or any other additive decomposition. The structural theories of thin plates, two- and three-dimensional elasticity as well as the two- and three-dimensional heat conduction on solids are focused upon in this paper. This methodology, however, can be applied to several other linear operators in Physics and Engineering. Given the differential equations' linearity, the decomposition separates the equations in an identical way to apply the Adomian decomposition for linear operators.

## 31.2 Differential Equations Subject to Decomposition

Any linear Partial Differential Equation (PDE), for a physical problem, can be written as

$$\boldsymbol{L}(\partial_{\boldsymbol{x}})\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x}), \;\; \boldsymbol{L}(\partial_{\boldsymbol{x}}) = \partial^{\mathrm{T}}\mathbf{C}\partial \tag{31.1}$$

in which the vector $\partial$ arranges the partial differentials $\partial_{\boldsymbol{x}}$, $\mathbf{C}$ denotes the constitutive tensor, $\boldsymbol{f}(\boldsymbol{x})$ is the elastic/thermal response, and $\boldsymbol{g}(\boldsymbol{x})$ is the source term. $\boldsymbol{L}(\partial_{\boldsymbol{x}})$ is the differential operator and the material is homogeneous through the domain. The dimensions of all the variables as well as the boundary conditions depend on the particular case: the eq. (31.1) can be related to three- and two-dimensional elasticity, to thin plate theory and three- and two-dimensional heat conduction. The three-dimensional elasticity is used in this paper to present the method's procedure. For the other operators, the modifications are obvious and are discussed throughout the text.

The Fourier transformation is applied in eq. (31.1) and for fundamental solutions, the source term $\boldsymbol{g}(\boldsymbol{x})$ can be replaced by the Dirac delta for the infinitesimal load simulation. Due to the domain homogeneity and the Fourier Transform property that converts partial differentials into algebraic-complex functions, the result can be directly written as

$$\hat{\boldsymbol{L}}(\xi)\hat{\boldsymbol{u}}(\xi) = \boldsymbol{I}\boldsymbol{q}, \;\; \hat{\boldsymbol{L}}(\xi) = \zeta^{\mathrm{T}}\mathbf{C}\zeta \tag{31.2}$$

where $\zeta$ arranges the transformed differentials, $\xi$ maps the spatial into frequency domain, and the hat denotes their definition in the frequency domain. $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{q}$ a unity vector containing the direction of the infinitesimal load, when applicable (for a scalar problem, both are equal to unity), and $\hat{\boldsymbol{u}}(\xi)$ is the fundamental solution in the transformed domain. The PDE (31.1) is elliptic, thus eq. (31.2) is a real-value function and can be manipulated algebraically. Upon application of the inverse Fourier transformation, the fundamental solution $\boldsymbol{u}(\boldsymbol{x})$ can be written as

$$\boldsymbol{u}(\boldsymbol{x}) = \frac{1}{(2\pi)^m} \int_{\Omega_\xi} [\hat{\boldsymbol{L}}(\xi)]^{-1} \boldsymbol{q}\, e^{i\boldsymbol{x}\cdot\xi}\, d\xi \tag{31.3}$$

where $m$ is the dimension of the problem, $\Omega_\xi$ is the integral domain, $d\xi = d\xi_1.d\xi_2.\cdots.d\xi_m$ and $i = \sqrt{-1}$. The transformed operator's inverse - $[\hat{\boldsymbol{L}}(\xi)]^{-1}$ - exists due to the constitutive tensor's positivity. Equation (31.3) is also found in Ting [Ti96] for three-dimensional elasticity. In that case, eq. (31.3) can be modified to represent the fundamental solution as a tensor, written as

$$\hat{\boldsymbol{L}}(\xi)\hat{\boldsymbol{U}}(\xi) = \boldsymbol{I} therefore \boldsymbol{U}(\boldsymbol{x}) = \frac{1}{(2\pi)^m} \int_{\Omega_\xi} [\hat{\boldsymbol{L}}(\xi)]^{-1}\, e^{i\boldsymbol{x}\cdot\xi}\, d\xi,\ \ \boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{U}(\boldsymbol{x})\boldsymbol{q}$$

in which $\boldsymbol{U}(\boldsymbol{x})$ is the tensor fundamental solution.

## 31.3 Constitutive Tensor Decomposition

### *31.3.1 Hooke's Law, Fourier's Law, and Constitutive Tensors*

Hooke's Law describes the mechanical behavior of crystalline solids via a linear stress–strain relation, in which the linearity coefficients are the constitutive properties. These can be written as a second-order tensor:

$$\mathbf{C} = \begin{pmatrix} C_{11} & C_{12} & C_{13} & C_{14} & C_{15} & C_{16} \\ & C_{22} & C_{23} & C_{24} & C_{25} & C_{26} \\ & & C_{33} & C_{34} & C_{35} & C_{36} \\ & & & C_{44} & C_{45} & C_{46} \\ & & & & C_{55} & C_{56} \\ sym & & & & & C_{66} \end{pmatrix} \tag{31.4}$$

For thin plate theory, the third, fourth, and fifth column/line are excluded from the eq. (31.4) and the influences of the transversal properties are inserted into its constitutive tensor as

$$\tilde{\mathbf{C}} = \begin{pmatrix} \tilde{C}_{11} & \tilde{C}_{12} & \tilde{C}_{16} \\ & \tilde{C}_{22} & \tilde{C}_{26} \\ sym & & \tilde{C}_{66} \end{pmatrix}, \quad \tilde{C}_{ij} = C_{ij} - \frac{C_{i3}C_{3j}}{C_{33}} \tag{31.5}$$

in which $i$ and $j$ vary as 1, 2 and 6. The tensor on eq. (31.5) is called reduced constitutive tensor and the tilde over the constitutive properties denotes reduced materials property.

Depending on the internal symmetries of the material, the tensors in eq. (31.4) - (31.5) can change. Cowin & Meharabadi [CoMe95] and Chadwick et. al [ChVi01] described the symmetry classes, and both have shown that there are eight types of constitutive symmetries (Figure 31.1). Symmetry is defined as an invariant spatial transformation on a unitary material cell. It is mathematically defined as an invariant transformation on the constitutive tensor

$$\boldsymbol{R}^{\mathrm{T}}\mathbf{C}'\boldsymbol{R} = \mathbf{C}, \quad \mathbf{C}' = \mathbf{C} \tag{31.6}$$

in which $\boldsymbol{R}$ is a unitary transformation matrix, and its size depends on the constitutive tensor. The transformations matrices have the following properties: closedness, associativity, identity, and inversion [CoMe95].

Figure 31.1 shows the symmetric planes for each crystal class on three-dimensional elasticity. Other structural theories have lesser symmetries due to the geometric simplifications which make some symmetries reduce to another. In thin plates theory, for example, the triclinic and the monoclinic materials generate the same reduced constitutive tensor.

In heat conduction, the tensor that gathers the constitutive parameters can be decomposed into their singular values. Given the nature of the differential operator, this modification just rotates the global coordinate system into the principal direction of the material and it reduces mathematically the number of possible symmetries. The constitutive tensor for heat conduction problems is

$$\mathbf{C} = \begin{pmatrix} k_{11} & 0 & 0 \\ & k_{22} & 0 \\ sym & & k_{33} \end{pmatrix} \tag{31.7}$$

The heat conduction constitutive symmetries are presented in Figure 31.1. For the two-dimensional case, the 3rd line/column is excluded and the constitutive tensor's size is $2 \times 2$. The transformation of the constitutive tensor on heat conduction is exactly the same as presented in eq. (31.6). The tensors in eq. (31.4)–(31.5) and eq. (31.7) are positive and symmetric due to the energy's positivity and the reciprocity's theorem, respectively.

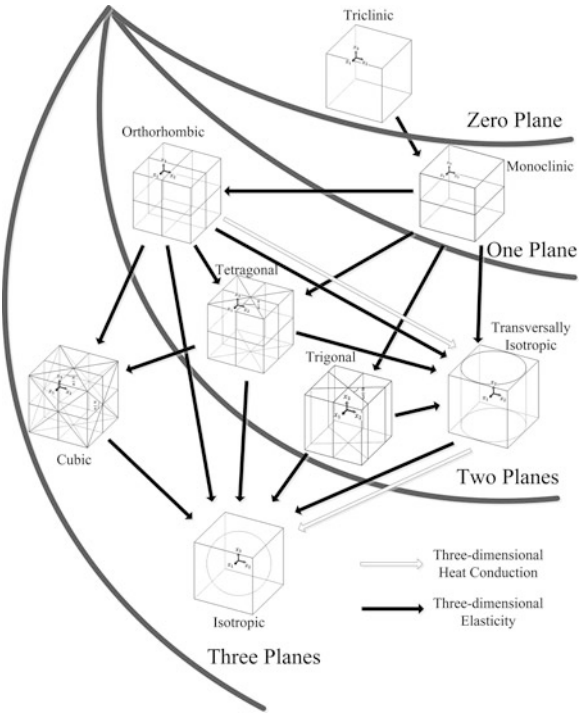


**Fig. 31.1** Symmetric classes and their hierarchy on Tridimensional Elasticity and Heat Conduction.

### *31.3.2 Hierarchy and Decompositions*

The hierarchy of the constitutive symmetries for both elasticity and heat conduction can be condensed into one rule as presented in Figure 31.1. Each cube shows the geometry of the symmetric planes and, when applicable, the angles between them. The four arcs divide the eight classes in terms of necessary planes to develop the symmetry. The arrows address the hierarchy level: the pointed material has all the symmetric planes of the pointing one. In other words, the intersection between the symmetry groups of the two materials is equal to the pointer symmetry group, having consequently a higher hierarchy.

This hierarchy is important in the presented decompositions. The constitutive tensor superposition follows the hierarchy and, hence, the arrows' system. For example: an orthorhombic material can be decomposed into a superposition of an isotropic, a transversally isotropic and an orthorhombic symmetry or, a cubic and an orthorhombic material. It is important to highlight that the decomposed symmetries are fictitious materials that hold only the properties of a specific symmetry. They do not represent any natural or synthetic material.

Tu [Tu68] has developed an additive decomposition of the fourth-order and second-order constitutive and flexibility elasticity tensors. Five types of orthonormal basis have been created and all symmetries but the triclinic can be decomposed.

Browayes & Chevrot [BrCh04] have decomposed the constitutive tensor in a similar way as shown in [Tu68]. However, it has been used a different criterion to categorize the symmetries. Five matrices were used to project the constitutive tensor, described as a 21-D vector form of a given symmetry, to another. Expanding the ideas from both works and knowing that some tensors can be null (which is not possible on the cited papers), a constitutive tensor is decomposed into eight existent symmetries as

$$\mathbf{C} = \mathbf{C}^{iso} + \mathbf{C}^{cub} + \mathbf{C}^{tis} + \mathbf{C}^{tgo} + \mathbf{C}^{tet} + \mathbf{C}^{ort} + \mathbf{C}^{mon} + \mathbf{C}^{tri} \tag{31.8}$$

where the superscripts correspond to the constitutive symmetries in Figure 31.1 and, from the left to the right are: isotropic, cubic, transversally isotropic, trigonal, tetragonal, orthorhombic, monoclinic and triclinic.

Hence, the decompositions can be described by

$$\mathbf{C} = \mathbf{C}^{(1)} + \mathbf{C}^{(2)} \tag{31.9}$$

It is noteworthy that eq. (31.9) does not simplify eq. (31.8). In symmetry (2), all remaining symmetries could be superposed, and could be decomposed again, if necessary. For example,

$$\mathbf{C}^{(1)} = \mathbf{C}^{iso}, \quad \mathbf{C}^{(2)} = \mathbf{C}^{cub} + \mathbf{C}^{tis} + \mathbf{C}^{tgo} + \mathbf{C}^{tet} + \mathbf{C}^{ort} + \mathbf{C}^{mon} + \mathbf{C}^{tri}$$

or could be symmetry (1) adding $\mathbf{C}^{iso}$ and $\mathbf{C}^{tis}$ and the remaining terms are added to symmetry (2).

Both decompositions do not generate positive tensors: only the isotropic tensor has this property. For application of the methodology, the tensor $\mathbf{C}^{(1)}$ needs to be positive. Hence, the isotropic tensor must be in symmetry (1) when the cited decompositions are used.

## 31.4 Recursive Methodology

The recursive methodology used to obtain the fundamental solutions is based on the Adomian decomposition. For nonlinear differential equations, in which this method is normally used, the operator can be divided into three parts: a linear in which the inverse is known, a linear where the inverse is not known (remainder term), and a nonlinear one. The considered PDEs are linear, so are the decomposition terms. The procedure of the recursive decomposition is simple and it is based on five steps:

1. The decomposition of the constitutive tensor into two or more constitutive tensors and the determination of the linear and remainder terms.
2. The remainder part, applied to the response, is placed together with the source term.
3. The fundamental solution is superposed by an infinite sum of terms.

4. Recursively, each of the superposed fundamental solution terms is solved in a manner so that the previous solutions may be used.
5. By using a defined truncation rule, the obtained terms are added to determine the approximate fundamental solution.

The original constitutive tensor is decomposed into two or more symmetries (step 1). Inserting eq. (31.9) into eq. (31.1) yields:

$$\boldsymbol{L}(\partial_x) = \partial^{\mathrm{T}}\mathbf{C}\partial = \partial^{\mathrm{T}}\left[\mathbf{C}^{(1)} + \mathbf{C}^{(2)}\right]\partial =$$

$$= \partial^{\mathrm{T}}\mathbf{C}^{(1)}\partial + \partial^{\mathrm{T}}\mathbf{C}^{(2)}\partial = \boldsymbol{L}^{(1)}(\partial_x) + \boldsymbol{L}^{(2)}(\partial_x)$$

in which the $\boldsymbol{L}^{(1)}(\partial_x)$ and $\boldsymbol{L}^{(2)}(\partial_x)$ are related to symmetry (1) and (2) of eq. (31.9), respectively. The $\boldsymbol{L}^{(1)}(\partial_x)$ fundamental solution is known, therefore, the PDE can be rewritten (step 2) as:

$$\boldsymbol{L}^{(1)}(\partial_x)\boldsymbol{U}(\boldsymbol{x}) = \mathbf{I}\delta(\boldsymbol{x}) - \boldsymbol{L}^{(2)}(\partial_x)\boldsymbol{U}(\boldsymbol{x})$$

The fundamental solution is expanded in an infinite series (step 3):

$$\boldsymbol{U}(\boldsymbol{x}) = \boldsymbol{U}_{(0)}(\boldsymbol{x}) + \boldsymbol{U}_{(1)}(\boldsymbol{x}) + \boldsymbol{U}_{(2)}(\boldsymbol{x}) + \cdots + \boldsymbol{U}_{(n)}(\boldsymbol{x}) + \cdots \tag{31.10}$$

Each term of eq. (31.10) is solved, recursively, (step 4):

$$\boldsymbol{L}^{(1)}(\partial_x)\boldsymbol{U}_{(0)}(\boldsymbol{x}) = \mathbf{I}\delta(\boldsymbol{x}) \tag{31.11}$$

$$\boldsymbol{L}^{(1)}(\partial_x)\boldsymbol{U}_{(n)}(\boldsymbol{x}) = -\boldsymbol{L}^{(2)}(\partial_x)\boldsymbol{U}_{(n-1)}(\boldsymbol{x}), \quad n \geq 1 \tag{31.12}$$

Equation (31.11) is identical to (31.1) when the source term, in the last, is the Dirac delta. The initial condition for the recursive system is the knowledge of the $\boldsymbol{L}^{(1)}(\partial_x)$ fundamental solution, as mentioned before. Using the fundamental solution's identity element property, eqs. (31.11) - (31.12) are defined as

$$\boldsymbol{U}_{(1)}(\boldsymbol{x}) = -\boldsymbol{U}_{(0)}(\boldsymbol{x}) * \boldsymbol{F}_{AN}(\boldsymbol{x}) \tag{31.13}$$

$$\boldsymbol{U}_{(2)}(\boldsymbol{x}) = \boldsymbol{U}_{(0)}(\boldsymbol{x}) * [\boldsymbol{F}_{AN}(\boldsymbol{x}) * \boldsymbol{F}_{AN}(\boldsymbol{x})] \tag{31.14}$$

$$\vdots$$

$$\boldsymbol{U}_{(n)}(\boldsymbol{x}) = (-1)^n \boldsymbol{U}_{(0)}(\boldsymbol{x}) * \left[\underbrace{\boldsymbol{F}_{AN}(\boldsymbol{x}) * \boldsymbol{F}_{AN}(\boldsymbol{x}) * \cdots * \boldsymbol{F}_{AN}(\boldsymbol{x})}_{n \text{ times}}\right] \tag{31.15}$$

where $*$ is the convolution operator and

$$\boldsymbol{F}_{AN}(\boldsymbol{x}) = \boldsymbol{L}^{(2)}(\partial_x)\boldsymbol{U}_{(0)}(\boldsymbol{x}) \tag{31.16}$$

Equation (31.16) denotes the partial insertion of the second symmetry's influence on the first symmetry fundamental solution. This term has an essential importance on the method's convergence. Inserting eq. (31.13) - (31.15) into eq. (31.10), one can find the complete solution, expressed as (step 5)

$$\boldsymbol{U}(\boldsymbol{x}) = \boldsymbol{U}_{(0)}(\boldsymbol{x}) * \{\mathbf{I}\delta(\boldsymbol{x}) - \boldsymbol{F}_{AN}(\boldsymbol{x}) + [\boldsymbol{F}_{AN}(\boldsymbol{x}) * \boldsymbol{F}_{AN}(\boldsymbol{x})] + \\ + \cdots + (-1)^n \left[\underbrace{\boldsymbol{F}_{AN}(\boldsymbol{x}) * \cdots * \boldsymbol{F}_{AN}(\boldsymbol{x})}_{n \text{ times}}\right]\} \tag{31.17}$$

As presented, eq. (31.17) can describe a fundamental solution through a constitutive tensor's decomposition in two different materials. Moreover, the second constitutive tensor does not need to be positive: it does not need to be inverted and hence it may be able to have null eigenvalues. Therefore, the fundamental solution in eq. (31.17) does not develop degeneracy. The only singular behavior, due to the material's symmetry, is that $\boldsymbol{U}_{(0)}(\boldsymbol{x})$ holds. Equation (31.16) shows clearly this statement: if $\boldsymbol{L}^{(2)}(\partial_x)$ possess multiples roots, it will be indifferent to $\boldsymbol{F}_{AN}(\boldsymbol{x})$ and therefore to $\boldsymbol{U}(\boldsymbol{x})$. The singularities of $\boldsymbol{U}_{(0)}(\boldsymbol{x})$, on the other hand, will be carried out through the recursive solution and to the final solution in the eq. (31.17).

The hierarchy from Figure 31.1 enters as a parameter when the recursive methodology is used multiple times. That is the reason for describing the symmetries as (1) and (2). The only necessity in the method is the symmetry (1) fundamental solution knowledge. In eq. (31.13)–(31.15), the influence of the symmetry is inserted into the aforementioned solution. At the end, the fundamental solution obtained is with respect to the sum of the two symmetries. Then this solution can be used as initial parameter to a new analysis and to develop a new and less symmetric solution.

## 31.5 Errors, Convergence Criterion, and Its Rate

To describe the proposed method's errors, convergence criterion and its rate, one may employ the Fourier Transform. The error can be calculated in the frequency domain by the Plancherel theorem. First of all, the norm for a tensor and an inner product between two equal tensors are defined as

$$\langle \mathbf{A}(\boldsymbol{y}), \mathbf{A}(\boldsymbol{y}) \rangle = \int_{\Omega_y} \mathrm{tr}\left(\mathbf{A}(\boldsymbol{y})\mathbf{A}^{\mathrm{T}}(\boldsymbol{y})\right) d\boldsymbol{y} = \int_{\Omega_y} |\mathbf{A}(\boldsymbol{y})|^2 d\boldsymbol{y} \tag{31.18}$$

in which $\mathbf{A}$ is a second-order tensor. The absolute error is calculated as the direct difference between the analytical and the approximate fundamental solution. Then

$$\mathbf{E}(\boldsymbol{x}) = \boldsymbol{U}^{ana}(\boldsymbol{x}) - \boldsymbol{U}^{apr}(\boldsymbol{x})$$

where $\boldsymbol{U}^{ana}(\boldsymbol{x})$ and $\boldsymbol{U}^{apr}(\boldsymbol{x})$ are the analytical and approximate fundamental solutions, respectively. By Plancherel theorem, the inner product of the error norm by itself in eq. (31.18) is expressed as

$$\langle \mathbf{E}(\boldsymbol{x}), \mathbf{E}(\boldsymbol{x}) \rangle = \int_{\Omega_x} |\mathbf{E}(\boldsymbol{x})|^2 d\boldsymbol{x} \tag{31.19}$$

and can be manipulated in the frequency domain.

$$\int_{\Omega_x} |\boldsymbol{U}^{ana}(\boldsymbol{x}) - \boldsymbol{U}^{apr}(\boldsymbol{x})|^2 d\boldsymbol{x} = \int_{\Omega_\zeta} |\hat{\boldsymbol{U}}^{ana}(\xi) - \hat{\boldsymbol{U}}^{apr}(\xi)|^2 d\xi$$

The difference between the analytical and the approximate solution, on the frequency domain, is

$$\hat{\boldsymbol{U}}^{ana}(\xi) - \hat{\boldsymbol{U}}^{apr}(\xi) = (-1)^{n+1} \left( \hat{\boldsymbol{L}}^{(1)}(\xi) + \hat{\boldsymbol{L}}^{(2)}(\xi) \right)^{-1} \left( \hat{\boldsymbol{F}}_{AN}(\xi) \right)^{n+1} \tag{31.20}$$

considering that the $\boldsymbol{U}^{apr}(\boldsymbol{x})$ is truncated in some $n$ approximation. Decomposing $\hat{\boldsymbol{F}}_{AN}(\xi)$ into its principal values (eigenvalues), it can be written as

$$\hat{\boldsymbol{F}}_{AN}(\xi) = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}$$

where $\boldsymbol{Q}$ is a square matrix where the columns contain eigenvectors of $\hat{\boldsymbol{F}}_{AN}(\xi)$ and $\boldsymbol{\Lambda}$ is a diagonal matrix whose elements correspond to the eigenvalues of $\hat{\boldsymbol{F}}_{AN}(\xi)$. Therefore, any power of $\hat{\boldsymbol{F}}_{AN}(\xi)$ can be calculated as:

$$\hat{\boldsymbol{F}}^n_{AN}(\xi) = \left[ \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1} \right]^n = \boldsymbol{Q}\boldsymbol{\Lambda}^n\boldsymbol{Q}^{-1} \tag{31.21}$$

Using eq. (31.21) in eq. (31.20) and applying the result in eq. (31.19), the inner product of the error norm is found as:

$$\langle \mathbf{E}(\boldsymbol{x}), \mathbf{E}(\boldsymbol{x}) \rangle = \int_{\Omega_\zeta} \left( \hat{\boldsymbol{L}}^{(1)}(\xi) + \hat{\boldsymbol{L}}^{(2)}(\xi) \right)^{-1} \cdot \cdot \hat{\boldsymbol{F}}^{2(n+1))}_{AN}(\xi) \cdot \left( \hat{\boldsymbol{L}}^{(1)}(\xi) + \hat{\boldsymbol{L}}^{(2)}(\xi) \right)^{-1} d\xi \tag{31.22}$$

Via eq. (31.22), the quadratic error norm can be determined given a recursion truncation $n$. Equations (31.20) and (31.22) impose that:

$$-\mathbf{I} < \boldsymbol{\Lambda} < \mathbf{I} \tag{31.23}$$

for the methodology's absolute convergence. This is also the criterion for all PDE that make use of this recursive method.

For the scalar case, eq. (31.16), transformed into the frequency domain, can be expressed as

$$\hat{F}_{AN}(\xi) = \frac{\hat{L}^{(2)}(\xi)}{\hat{L}^{(1)}(\xi)}$$

And the quadratic norm of the error is:

$$\langle E(\boldsymbol{x}), E(\boldsymbol{x}) \rangle = \int_{\Omega_\xi} \left( \frac{\hat{L}^{(2)}(\xi)}{\hat{L}^{(1)}(\xi)} \right)^{2(n+1)} \cdot \frac{1}{\left( \hat{L}^{(1)}(\xi) + \hat{L}^{(2)}(\xi) \right)^2} d\xi$$

and thus,

$$-1 < \frac{\hat{L}^{(2)}(\xi)}{\hat{L}^{(1)}(\xi)} < 1 \tag{31.24}$$

To corroborate eq. (31.23) and eq. (31.24), we recognize that the recursive methodology, in which the solution is presented in eq. (31.17), generates an alternative series. If $\mathbf{C}^{(2)}$ is positive-definite, eq. (31.23) and eq. (31.24) can be simplified to

$$\boldsymbol{\Lambda} < \mathbf{I}$$

$$\mathbf{C}^{(1)} - \mathbf{C}^{(2)} > \mathbf{0}.$$

The relative error and the convergence rate depend uniquely on eq. (31.16), as well as the convergence criterion. For the relative error, the eq. (31.20) can be pre-multiplied by $\hat{\boldsymbol{U}}^{ana}(\xi)$ resulting in $\left(\hat{F}_{AN}(\xi)\right)^{n+1}$. The convergence rate will depend on the highest eigenvalue of $\hat{F}_{AN}(\xi)$ and how close it is to unity.

## 31.6 Summary and Conclusions

The aim of this chapter was the development of an anisotropic fundamental solution based on a crystalline class hierarchy. An additive decomposition of the constitutive tensor was proposed to simplify the calculus of an anisotropic solution. A known solution is used and recursively, the remainder term is inserted as a source term. The methodology's error and convergence were presented. Even though the solutions obtained through this methodology do not have an analytical close form, they do not degenerate and, due to the solutions' superimposed form, the first and second derivatives can be easily determined. Nevertheless, the solutions reduce themselves

to other more symmetric cases given the fact that the materials' singularities stay in the base solution. The problem in the methodology herein presented are the convolution operations. These have high computational costs, however, they can be solved by semi-analytical or numerical procedures. Future steps are the development of a numerical procedure to determine anisotropic solutions via an isotropic response. Using the results, semi-analytical procedures will pursuit to put the material properties in evidence.

## References

[Ad98] Adomian, G.: Solving Frontier Problems of Physics: The Decomposition Method. Kluwer Academic Publ., Dordrecht (1994)

[BrCh04] Browaeys, J. T., Chevrot, S.: Decomposition of the Elastic Tensor and Geophysical Applications. Geophys. J. Internat., **159**, 667–678 (2004).

[BuMa14] Buroni, F. C., Marczak, R. J., Denda, M., Saez, A.: A Formalism for Anisotropic Heat Transfer Phenomena: Foundations and Green's Functions. Internat. J. Heat Mass Transf., **75**, 399–409 (2014)

[BuOr10] Buroni, F. C., Ortiz, J. E., Saez, A.: Multiple Pole Residue Approach dor 3D BEM Analysis of Mathematical Degenerate and Non-Degenerate Materials. Internat. J. Numer. Methods Engrg., **86**, 1125–1143 (2010)

[ChRe02] Cheng, Z. -Q., Reddy, J. N.: Octet Formalism for Kirchhoff Anisotropic Plates. R. Soc. London Proc. Ser. A, **458**, 1499–1517 (2002)

[ChVi01] Chadwick, P., Vianello, M., Cowin, S. C.: A New Proof that the Number of Linear Elastic Symmetries is Eight. J. Mech. Phys. Solids, **49**, 2471–2492 (2001)

[CoMe95] Cowin, S. C., Mehrabadi, M. M.: Anisotropic Symmetries of Linear Elasticity. Appl. Mech. Rev., **48**, 247–285 (1995)

[LiSu07] Liou, J. Y., Sung, J. C.: On the Generalized Barnett-Lothe Tensors for Monoclinic Piezoeletric Materials. Internat. J. Solids Struct., **44**, 5208–5221 (2007)

[MaDe11] Marczak, R. J., Denda, M.: New Derivations of the Fundamental Solution for Heat Conduction Problems in Three-Dimensional General Anisotropic Media. Internat. J. Heat Mass Transf., **54**, 3605–3612 (2011).

[NaTu97] Nakamura, G., Tanuma, K.: A Formula for the Fundamental Solution of Anisotropic Elasticity. Quart. J. Mech. Appl. Math., **50**, 179–194 (1997)

[PaSo02] Paiva, W. P., Sollero, P., Albuquerque, E. L.: Analysis of the Fundamental Solution for Anisotropic Thin Plates. 15th ASCE Engrg. Mech. Conf.. New York (2002)

[ShBe88] Shi, G., Bezine, G.: A General Boundary Integral Formulation for Anisotropic Plate Bending. J. Compos. Mater., **22**, 694–716 (1988)

[SmSm63] Smith, G. F., Smith, M. M., Rivlin, R. S.: Integrity Bases for a Symmetric Tensor and a Vector - The Crystal Classes. Arch. Ration. Mech. Anal., **12**, 93–133 (1963)

[TaOr08] Távara, L., Ortiz, J. E., Mantič, V., París, F.: Unique Real-Variable Expressions of Displacement and Traction Fundamental Solutions Covering all Transversely Isotropic Elastic Materials for 3D BEM. Internat. J. Numer. Methods Engrg., **74**, 776–798 (2008)

[Ti96] Ting, T. C.: Anisotropic Elasticity: Theory and Applications. Oxford Univ. Press, Inc., New York (1996).

[ToPa01] Tonon, F., Pan, E., Amadei, B.: Green's Functions and Boundary Element Method Formulation for 3D Anisotropic Media. Comput. & Struct., **79**, 469–482 (2001)

[Tu68] Tu, Y. -O.: The Decomposition of an Anisotropic Elastic Tensor. Acta Crystallogr. Sect. A , **24**, 273–282 (1968)

[Wa97] Wang, C. -Y.: Elastic Fields Produced by a Point Source in Solids of General Anisotropy. J. Engrg. Math., **32**, 41–52 (1997)

# Chapter 32
# On a Model for Pollutant Dispersion in the Atmosphere with Partially Reflective Boundary Conditions

**J.F. Loeck, B.E.J. Bodmann, and M.T.B. Vilhena**

## 32.1 Introduction

Air pollutant release of either anthropogenic or natural sources is of increasing relevance because of its possible adverse effects and consequences on the ecosystem including humans. Initiatives related to environmental protocols are one witness to testify the necessity to understand and predict impact of dispersion of substances on environmental health and in case of incidents or accidents evaluate its risks on habitats.

Atmospheric pollution dispersion is commonly modelled by deterministic equations, where the most widely employed approach is based on the advection-diffusion equation. Several works exist in the literature that solve the equation analytically, semi-analytically or numerically (see, for instance, [BuEtAl11, ThMc06, TiVi12] and the references therein). These equations are usually linear equations with a solution that describes the mean value of substance concentrations and some of them are restricted to a compact support others valid on an infinite support, where the present approach is of the second type.

It is noteworthy that the phenomenon of pollutant transport in the atmosphere is nonlinear by virtue of turbulence and second stochastic which is best visible by comparison of experimental findings with theoretical predictions. If a deterministic model was adequate to describe the dispersion process, an improvement of the model should approach unity for the correlation between observed and predicted values. However, publications on the subject suggest that there is an asymptotic limit

J.F. Loeck (✉) • B.E.J. Bodmann • M.T.B. Vilhena
Federal University of Rio Grande do Sul, Av. Osvaldo Aranha 99/4, Porto Alegre 90046-900, RS, Brazil
e-mail: emaildajaque@gmail.com; bardo.bodmann@ufrgs.br; mtmbvilhena@gmail.com



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_32

of approximately $R = 0.9$, which may indicate that besides possible model errors, there are also natural fluctuations present that evidently cannot be reproduced by a purely deterministic model.

The present work is thus an attempt to introduce effects into that model that shall to some extent mimic some stochastic properties. To this end we modify the originally purely geometrical boundary conditions, i.e. the ground level and the boundary layer height, respectively. More specifically, turbulent mixture is believed to take place in various scales, where the largest scale is limited by the boundary layer height, but also smaller scales shall be present. One could think of the boundary layer as a superposition of various boundary layers, however with different ground and upper layer heights. Such a construction could model the escape of pollutant substances across the boundary layer horizon on the one side and the surface boundary on the other side and are modelled by probabilities to quantify the fraction of pollutant that returns into the boundary layer from above and the process of adsorption or deposition on the ground layer. These effects are represented by reflective and distributed boundary conditions that together with advection–diffusion dispersion define the model in consideration. The consequences of the reflections are analyzed using the meteorological conditions and data of the Hanford experiment.

## 32.2 A Locally Gaussian Model

The advection–diffusion equation may be derived in the standard fashion starting from the continuity equation and using the Reynolds decomposition to separate the mean components for the concentration and the velocity fields, respectively. Upon taking averages and substitution of the average fluctuations by Fick's closure, one arrives at the desired equation for mean concentrations and an *a priori* known wind field and with all turbulent characteristics parametrized in a time dependent eddy diffusivity matrix coefficient $\mathbf{K}$. For the present study we further simplify eddy diffusion using locally constant coefficients, which may be justified by the fact that the coefficients vary softly only with changing coordinates and is typical for homogeneous turbulence. For details of the derivation, see, for instance, the textbook by Arya [Ar99].

$$\frac{\partial \bar{c}}{\partial t} + \bar{\mathbf{u}} \nabla \bar{c} = \nabla \mathbf{K} \nabla \bar{c} + \bar{S} \tag{32.1}$$

where $\bar{c}$ represents the mean concentration of a contaminant $(g/m^3)$, $\bar{\mathbf{u}} = (\bar{u}, \bar{v}, \bar{w})$ are the mean wind speeds (in $m/s$) in the longitudinal, vertical and cross wind directions, the nabla symbol $\nabla$ signifies the usual vector differential operator, the eddy diffusivity coefficient is represented by a diagonal matrix $\mathbf{K} = \mathrm{diag}(K_x, K_y, K_z)$ and $\bar{S}$ is a source term.

Considering a point source at height $H_s$ that releases instantaneously a pollutant at a time $t = 0$, and a fixed quantity $Q$, then the source term can be cast in an initial condition and equation (32.1) simplifies to the initial value problem, neglecting further the slowly varying terms $\nabla \mathbf{K} \nabla \bar{c}$.

$$\frac{\partial \bar{c}}{\partial t} + \bar{u}\frac{\partial \bar{c}}{\partial x} + \bar{v}\frac{\partial \bar{c}}{\partial y} + \bar{w}\frac{\partial \bar{c}}{\partial z} = K_x \frac{\partial^2 \bar{c}}{\partial x^2} + K_y \frac{\partial^2 \bar{c}}{\partial y^2} + K_z \frac{\partial^2 \bar{c}}{\partial z^2}$$

$$\bar{c}(x, y, z, 0) = Q\delta(x - x_0)\delta(y - y_0)\delta(z - H_s)$$

This approximate problem can be solved analytically by separation of variables [Oz74] and Fourier transform [SePa06].

$$\bar{c}(x, y, z, t) = \frac{Q}{\sqrt{64\pi^3 K_x K_y K_z t^3}} \exp\left\{ -\frac{(x - x_0 - \bar{u}t)^2}{4K_x t} - \frac{(y - y_0 - \bar{v}t)^2}{4K_y t} - \frac{(z - H_s - \bar{w}t)^2}{4K_z t} \right\} \tag{32.2}$$

However, most dispersion problems are due to continuous emissions, which can be idealized by the superposition of instantaneous emissions. Considering a small time interval $d\tau$ with an instantaneous emission, then the continuous emission is

$$\bar{C}(x, y, z, t) \propto \int_0^t \bar{c}(x, y, z, t - \tau)\, d\tau\,,$$

where $\bar{c}$ is the concentration for the instantaneous and $\bar{C}$ for the continuous emission. Note that such an approach is valid for cases where the concentrations do not influence in the flow characteristics. Considering now an emission rate $\dot{Q}$ instead of the quantity $Q$ and the solution for the instantaneous emission (32.2), then the solution for the continuous emission is

$$\bar{C}(x, y, z, t) = \frac{\dot{Q}}{\sqrt{64\pi^3 K_x K_y K_z}} \int_0^t \frac{1}{\sqrt{(t - \tau)^3}} \exp\left\{ -\frac{[x - x_0 - \bar{u}(t - \tau)]^2}{4K_x(t - \tau)} - \frac{[y - y_0 - \bar{v}(t - \tau)]^2}{4K_y(t - \tau)} - \frac{[z - H_s - \bar{w}(t - \tau)]^2}{4K_z(t - \tau)} \right\} d\tau\,. \tag{32.3}$$

Recalling that the solutions (32.2) and (32.3) were obtained by Fourier transform, they are valid for the infinite ranges $x \in (-\infty, \infty), y \in (-\infty, \infty), z \in (-\infty, \infty)$ in contrast to [BuEtAl11]. However, as a matter of fact the dispersion of contaminants is limited by the ground $(z = 0)$ and the top of the atmospheric boundary layer $(z = z_i)$ so that the infinite range has to be mapped into a finite range.

## 32.3 Reflective Boundary Conditions

To justify the mapping of the infinite range $z \in (-\infty, \infty)$ to the finite $z \in [0, z_i]$ we first consider a cut of the distribution at $z = 0$ and $z = z_i$, respectively. The fact that a non-zero concentration at the boundaries is possible is not a serious problem, but if Fick's hypothesis is understood one expects a flux across these boundaries which contradicts the boundary layer conception. In a second step we copy from observation which suggests that the layer until the height where temperature inversion occurs may be considered at least partially decoupled from the wind flux system beyond. Hence, in an ideally decoupled system the lost contributions should be recovered, which could be obtained by a time dependent normalization (the total amount of pollutant shall be equal to the quantity released until that time) but we adopt another reasoning, namely adopting reflecting boundaries, which intuitively agrees with a simple particle ensemble picture where the pollutant that reaches the ground or the top of the atmospheric boundary layer bounces completely back into the domain. For the distributions that means that even after reflections the Gaussian tails exceeding the allowed domain are mirrored back into the finite range $z \in [0, z_i]$.

Formally, the reflection on the ground and in the atmospheric boundary layer may be viewed as contributions due to a virtual source in some effective heights to both sides below ground and above the boundary layer[Ba01]. The sequences that represent the mirror maxima are

$$\left.\begin{aligned} H_s &\to -H_s - 2nz_i \\ H_s &\to \ \ H_s + 2nz_i \end{aligned}\right\} \ \forall n \in \mathbb{Z}. \tag{32.4}$$

Substituting these two sequences in the solution for the continuous emission (32.3), the solution for continuous emission with complete reflection is obtained

$$\begin{aligned} \bar{C}(x,y,z,t) = \frac{\dot{Q}}{\sqrt{64\pi^3 K_x K_y K_z}} \int_0^t \Bigg[ \frac{1}{\sqrt{(t-\tau)^3}} \exp\left\{ -\frac{[x - x_0 - \bar{u}(t-\tau)]^2}{4K_x(t-\tau)} \right. \\ \left. - \frac{[y - y_0 - \bar{v}(t-\tau)]^2}{4K_y(t-\tau)} \right\} \left( \sum_{n=-\infty}^{\infty} \exp\left\{ -\frac{[z - H_s - 2nz_i - \bar{w}(t-\tau)]^2}{4K_z(t-\tau)} \right\} \right. \\ \left. \left. + \exp\left\{ -\frac{[z + H_s + 2nz_i - \bar{w}(t-\tau)]^2}{4K_z(t-\tau)} \right\} \right) \Bigg] \, d\tau \,, \end{aligned}$$

and is now valid for $x \in (-\infty, \infty), y \in (-\infty, \infty), z \in [0, z_i]$.

So far the model still does not represent any property that might be associated with an effect from a stochastic feature. As already argued before, instead of a boundary layer with rigid limits one could mimic a sample of a distribution with different boundary layer heights upon changing the position of the mirror images that compose the total distributions. To this end we introduce the reduction factor $\omega_b$

and $\omega_g$ in the sequences (32.4). Note that the system still maintains its deterministic character, but a finite sample of boundary layer configurations with different heights and center could be interpreted as a manifestation of its stochastic nature and are used to study the behavior of the new solution:

$$\bar{C}(x,y,z,t) = \frac{\dot{Q}}{\sqrt{64\pi^3 K_x K_y K_z}} \int_0^t \left[ \frac{1}{\sqrt{(t-\tau)^3}} \exp\left\{ -\frac{[x-x_0-\bar{u}(t-\tau)]^2}{4K_x(t-\tau)} \right.\right.$$
$$\left. - \frac{[y-y_0-\bar{v}(t-\tau)]^2}{4K_y(t-\tau)} \right\} \left( \sum_{n=-\infty}^{\infty} \exp\left\{ -\frac{[z-H_s-2n\omega_g z_i-\bar{w}(t-\tau)]^2}{4K_z(t-\tau)} \right\} \right.$$
$$\left.\left. + \exp\left\{ -\frac{[z+H_s+2n\omega_b z_i-\bar{w}(t-\tau)]^2}{4K_z(t-\tau)} \right\} \right) \right] d\tau$$

## 32.4 Turbulent Diffusivity Parametrization

To validate the proposed model, more specifically to analyze the impact of reflections on the results, turbulent diffusivity was parametrized to represent meteorological conditions of the Hanford experiment [DoHo85]. This campaign is a low source experiment (the height of the source $H_s$ was 2 $m$) with stable to quasi-neutral conditions. A non-depositing tracer was released with an average rate of $\dot{Q} = 0.3\ g/s$ and release time interval of 30 minutes, except for experiment run 05, where the release time was 22 minutes. The measurements were performed at distances 100 $m$, 200 $m$, 800 $m$, 1600 $m$, and 3200$m$ from the source. The necessary micro-meteorological data for the parametrization were provided by the experiment and are presented in Table 32.1. The height of the stable boundary layer ($z_{i,s}$) was calculated using the relation $z_{i,s} = 0.4(u_* L/f_c)^{1/2}$, where $f_c = 1.46 \times 10^{-4}\ s^{-1}$ is the Coriolis parameter.

The eddy diffusion coefficient for stable conditions proposed by Degrazia and Moraes [DeMo92] is based on the diffusion theory of Taylor [Ta22] and the turbulent kinetic energy spectrum [PaSm83] and can be computed using the micro-meteorological data set from table 32.1.

**Table 32.1**
Micro-meteorological data for the Hanford experiment.

| Expt | $\bar{u}$ (2 $m$) ($ms^{-1}$) | $u_*$ ($ms^{-1}$) | $L$ ($m$) | $z_{i,s}$ ($m$) |
|---|---|---|---|---|
| 01 | 3.63 | 0.40 | 166 | 269 |
| 02 | 1.42 | 0.26 | 44 | 112 |
| 03 | 2.02 | 0.27 | 77 | 151 |
| 04 | 1.50 | 0.20 | 34 | 86 |
| 05 | 1.41 | 0.26 | 59 | 129 |
| 06 | 1.54 | 0.30 | 71 | 152 |

$$K_z = \frac{0.644 u_* (1 - \frac{z}{z_{i,s}})^{\frac{\alpha_1}{2}} 1.58 z}{8\sqrt{\pi}(f_m)_w} \times \int_0^\infty \frac{\sin\left\{8\sqrt{\pi}1.58(1 - \frac{z}{z_{i,s}})^{\frac{\alpha_1}{2}}(f_m)_w n' X \frac{z_{i,s}}{(1.5)^{\frac{3}{5}}} z\right\}}{(1 + n'^{\frac{5}{3}})n'} dn'$$

Here $u_*$ is the friction velocity, $z$ is the observation height, $z_{i,s}$ is the height of the stable boundary layer, the parameter $\alpha_1 = 1.5$, $(f_m)_w$ is the frequency of the spectral peak in the vertical eddy spectrum, $X$ is the dimensionless distance, and $n'$ is the dimensionless frequency of the turbulent kinetic energy spectrum.

In the further, we introduce a simplification, without imposing restrictions on our numerical findings. We assume that our coordinate system has its $x$-axis aligned with the average wind speed, which is to a good approximation horizontal with respect to the Earth's surface. In order to determine the velocity field $\bar{\mathbf{u}} = U(z)\hat{\mathbf{x}}$ with $\hat{\mathbf{x}}$ a unit vector, we need to fix the vertical wind speed profile. The latter has been parametrized following Obukhov's similarity theory manifest in the so-called OML-model [BeOl86], where close to the surface and because of its roughness, there is a raising profile, whereas sufficiently far from the surface the wind speed remains approximately constant. If $z_b = \min(|L|, 0.1 z_{i,s})$, then

$$U = \frac{u_*}{k}\left(\ln\left[\frac{z}{z_0}\right] - \Psi_m(\frac{z}{L}) + \Psi_m(\frac{z_0}{L})\right), \quad z \le z_b,$$
$$U = \bar{u}(z), \quad z > z_b,$$

where $\Psi_m = -4.7\frac{z}{L}$ is the stability function for stable conditions.

## 32.5 Validation of the Model

To simulate the results, the complete data set of the Hanford experimental data was used, except those for the distances $x = 100m$ and $x = 200m$. The comparison of observed $(C_o)$ against predicted $(C_p)$ concentrations for a variety of reflection parameter $\omega_{b,g}$ and number of reflections is shown in figures 32.1 and 32.3. The corresponding statistical indices [Ha89], i.e. the normalized mean square error ($NMSE = \frac{\overline{(C_o - C_p)^2}}{\bar{C}_o \bar{C}_p}$), the correlation coefficient ($COR = \frac{\overline{(C_o - \bar{C}_o)(C_p - \bar{C}_p)}}{\sigma_o \sigma_p}$), the fractional bias ($FB = \frac{\bar{C}_o - \bar{C}_p}{\frac{1}{2}(\bar{C}_o + \bar{C}_p)}$), and the fractional standard deviation ($FS = \frac{\sigma_0 - \sigma_p}{\frac{1}{2}(\sigma_0 + \sigma_p)}$ are shown in the tables 32.2 and 32.3. Further, the saturation effect for after a certain number of reflections is shown in figures 32.2 and 32.4, where the normalized mean

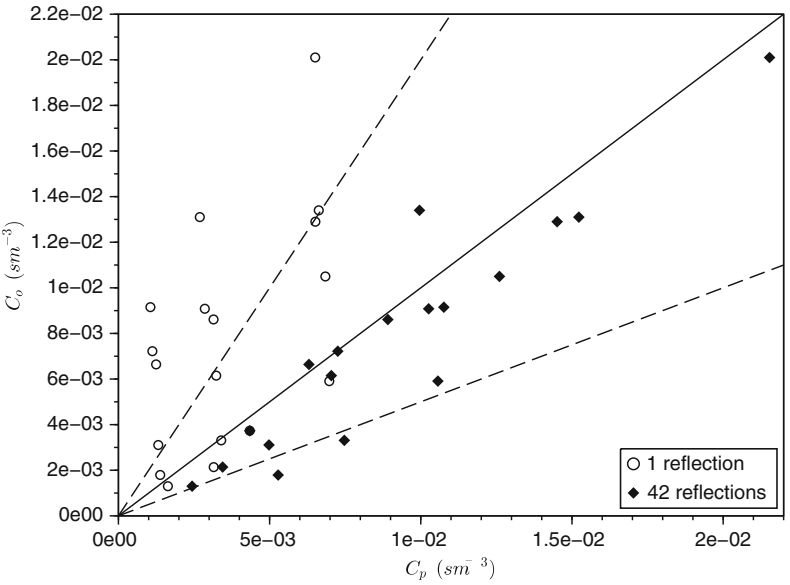


**Fig. 32.1** Scatter plot for observed ($C_o$) and predicted ($C_p$) concentrations for one and 42 reflections with the parameters $\omega_g = 0.1$ and $\omega_b = 0.005$.

**Table 32.2** Statistical evaluation for observed ($C_o$) and predicted ($C_p$) concentrations with $R = 1, \ldots, 8, 14, 20, 30$ and 42 reflections for the parameters $\omega_g = 0.1$ and $\omega_b = 0.005$.

| R | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 14 | 20 | 30 | 42 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NMSE | 1.24 | 0.50 | 0.27 | 0.18 | 0.13 | 0.11 | 0.09 | 0.08 | 0.07 | 0.07 | 0.07 | 0.07 |
| COR | 0.52 | 0.62 | 0.69 | 0.75 | 0.78 | 0.81 | 0.83 | 0.84 | 0.88 | 0.88 | 0.88 | 0.88 |
| FB | 0.73 | 0.39 | 0.20 | 0.09 | 0.02 | −0.02 | −0.06 | −0.08 | −0.14 | −0.16 | −0.16 | −0.16 |
| FS | 0.78 | 0.44 | 0.28 | 0.19 | 0.14 | 0.10 | 0.08 | 0.01 | 0.05 | 0.05 | 0.04 | 0.04 |

**Table 32.3** Statistical evaluation for observed ($C_o$) and predicted ($C_p$) concentrations with $R = 1, \ldots, 4, 10, 20$, and 28 reflections for the parameters $\omega_g = 0.2$ and $\omega_b = 0.01$.

| R | 1 | 2 | 3 | 4 | 10 | 20 | 28 |
|---|---|---|---|---|---|---|---|
| NMSE | 0.39 | 0.16 | 0.10 | 0.08 | 0.06 | 0.06 | 0.06 |
| COR | 0.66 | 0.77 | 0.82 | 0.85 | 0.88 | 0.88 | 0.88 |
| FB | 0.32 | 0.10 | 0.00 | −0.05 | −0.12 | −0.12 | −0.12 |
| FS | 0.37 | 0.19 | 0.12 | 0.10 | 0.08 | 0.07 | 0.07 |

square error (the scale is given by the left vertical axis) as well as the correlation coefficient (the scale is given by the right vertical axis) is plotted against the number of reflections.

A general comment is in order here, although the statistical evaluations mentioned above are similar to those from parametric inference procedures, their

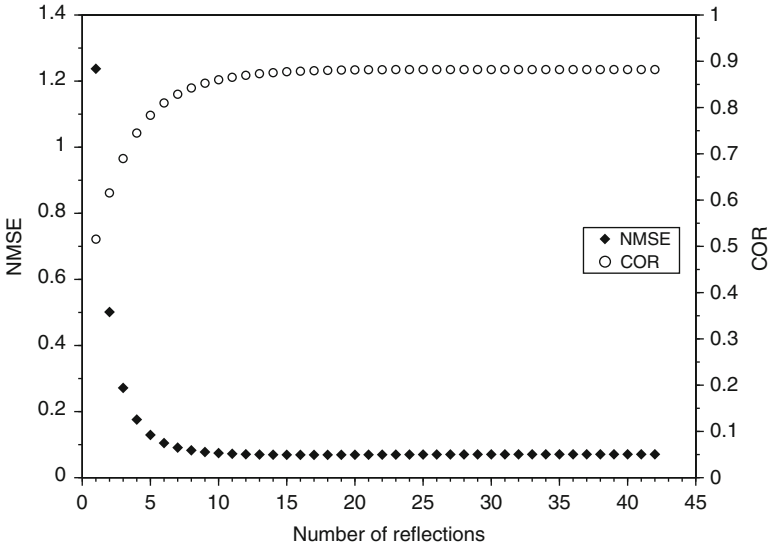


**Fig. 32.2** NMSE (scale left) and COR (scale right) versus number of reflections for $\omega_g = 0.1$ and $\omega_b = 0.005$.

interpretations are different in the present context. In parametric inference the best estimates for parameters were attained for $NMSE \to 0$, but in the present consideration a deterministic model is compared to a relatively small data set from a stochastic phenomenon, so that one does not expect vanishing values for this error. Further, the correlation coefficient does not converge to unity, recalling that the observed data are one sample out of a distribution for a specific situation, that are parametrized using their specific micro-meteorological data. The fractional bias may be interpreted in terms of model fidelity, significant deviations from zero indicate that the model lacks some relevant physical features. Last but not least, figures 32.2, 32.3, 32.4 show the apparent asymptotic limit for the normalized mean square error as well as the correlation coefficient with limit $\sim 0.9$. One observes in the two presented cases that after inclusion of reflections the correlation of experimental and predicted data improves, which may be interpreted as an indication that with our reasoning we have at least made some point, even though we have not solved the important question of how the best values for $\omega_b$ and $\omega_g$ shall be obtained.

## 32.6 Conclusions

This chapter may be considered an attempt to introduce stochastic characteristics in an originally deterministic model, i.e. pollutant dispersion by advection–diffusion dynamics in the planetary boundary layer. One may argue that the effective

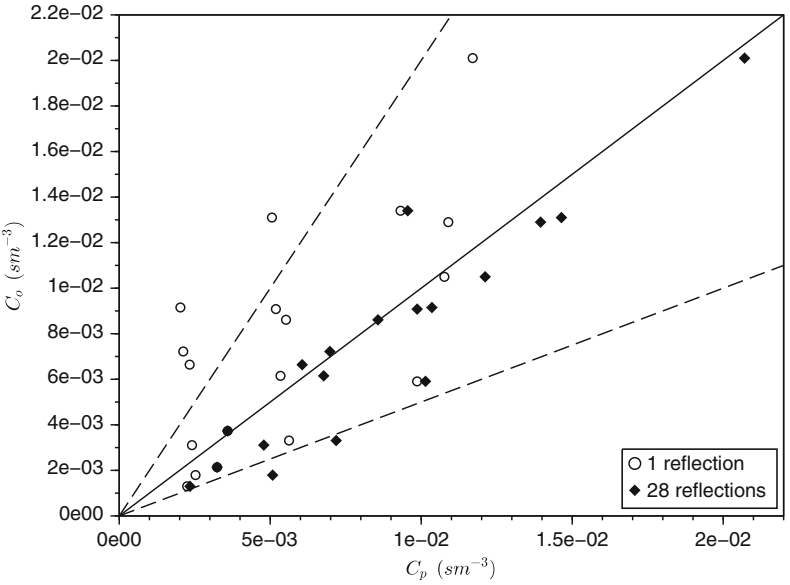


**Fig. 32.3** Scatter plot for observed ($C_o$) and predicted ($C_p$) concentrations for one and 28 reflections with the parameters $\omega_g = 0.2$ and $\omega_b = 0.01$.

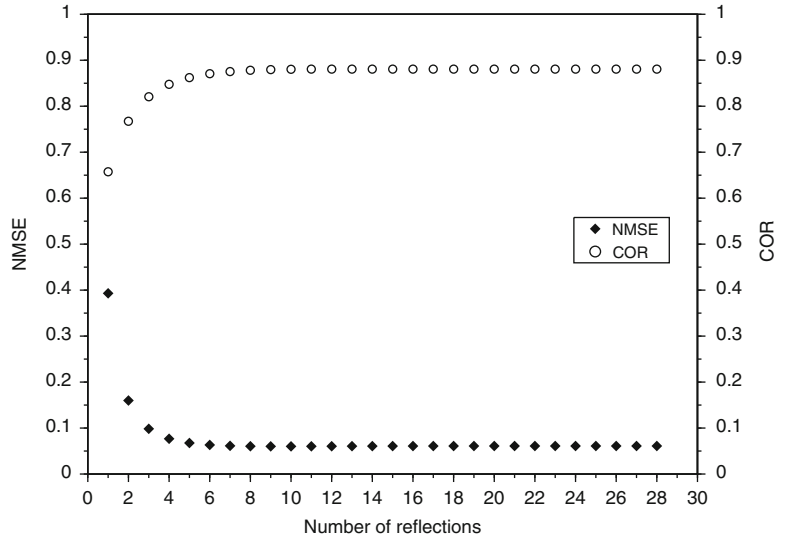


**Fig. 32.4** NMSE versus number of reflections for $\omega_g = 0.2$ and $\omega_b = 0.01$.

boundary layer height is not necessarily a fixed quantity but varies according to the turbulent flow dynamics it incorporates. Thus the layer boundary shall have stochastic character, so that a more realistic flow may be thought of as a superposition of various boundary layer problems but with different effective boundary layer heights. Such a procedure may be interpreted as a discrete set of samples that represent an unknown distribution, which needs an additional model or parametrization hypothesis (for $\omega_b$ and $\omega_g$) and is beyond the scope of the present analysis. So far the goal was to introduce geometrically motivated mirror images as virtual sources and superimpose them such as to mimic a finite size sample of a distribution from different boundary layer height realizations.

At this stage of the work, we are completely aware of the fact that some of the parameters ($\omega_b$, $\omega_g$) need a physically motivated prescription on how to determine them from experimental data. Although one would like to have this kind of information for our model right from the beginning, we think that our reasoning shows that it is well plausible and the boundary layer height distributions exist, however we leave the discussion on this issue for a future work. Nevertheless, a variety of trials have shown us that reflections on the boundary layer horizon and on the ground obtain significant correlations between model and data suggesting that effects on the boundary are essential to model dispersion processes in the atmospheric boundary layer, even though the values for the reflection parameters were established *ad hoc*.

This improvement in the solution can be related to the fact that the deterministic equation predicts only mean values of an unknown distribution and is not capable at all to reproduce stochastic properties, which in our case were modeled by the effects of the considered reflections. Moreover, the model does not consider deposition and adsorption on the ground, but due to the fact that concentration and vertical concentration fluxes are different from zero on the ground level one may reason that the parameter $\omega_g$ is somehow incorporating these properties.

## References

[Ar99] Arya, S.P.: Air pollution meteorology and dispersion. Oxford University Press, New York (1999)

[Ba01] Barratt, R.: Atmospheric Dispersion Modelling: An Introduction to Practical Applications. Earthscan, London, UK (2001)

[BeOl86] Berkowicz, R.R., Olesen, H.R., Torp, U.: The danish gaussian air pollution model (OML): Description, test and sensitivity analysis in view of regulatory applications. In: Air Pollution Modeling and Its Application **10**, Plenum Publishing Corporation, New York, 453–481 (1986).

[BuEtAl11] Buske, D., Vilhena, M.T., Segatto, C.F., Quadros, R.S.: A General Analytical Solution of the Advection-Diffusion Equation for Fickian Closure In: Ch. Constanda, P.J. Harris, Integral Methods in Science and Engineering: Computational and Analytic Aspects, Springer, 25–33 (2011).

[DeMo92] Degrazia, G.A. and Moraes, O.L.L.: A model for eddy diffusivity in a stable boundary layer. Boundary-Layer Meteorology **58**, 205–214 (1992)

[DoHo85] Doran, J.C. and Horst, T.W.: An evaluation of Gaussian plume depletion models with dual-tracer field measurements. Atmospheric Environment **19**, 939–951 (1985)

[Ha89] Hanna, S.R.: Confidence limit for air quality models as estimated by bootstrap and jacknife resampling methods. Atmospheric Environment **23**, 1385–1395 (1989)

[Oz74] Özisik, M.: Heat Conduction. John Wiley & Sons, New York, 2 edition (1974)

[PaSm83] Pasquill, F. and Smith, F.B.: Atmospheric Diffusion. Halsted Press, New York, $3^{rd}$ edition (1983)

[SePa06] Seinfeld, J.H., Pandis, S.N.: Atmospheric chemistry and physics: from air pollution to climate change. John Wiley & Sons, New Jersey, $2^{nd}$ edition (2006)

[Ta22] Taylor, G.I.: Diffusion by Continuous Movements. Proceedings of the London Mathematical Society **20**, 196–212 (1922)

[ThMc06] Thongmoon, M., McKibbin, R.: A comparison of some numerical methods for the advection-diffusion equation. Res. Lett. Inf. Math. Sci. **10**, 49–62 (2006).

[TiVi12] Tirabassi, T. and Vilhena M.: Advection-Diffusion in the Atmosphere: Equations and Solutions. In: R. Grifoni; G. Latini; S. Tascini (eds.) Atmospheric Flow Fields: Theory, Numerical Methods and Software Tools, Bentham Science Publishers, Oak Park, Illinois, 153–173 (2012)

# Chapter 33
# Asymptotic Approximations for Chemical Reactive Flows in Thick Fractal Junctions

**T.A. Mel'nyk**

## 33.1 Introduction

It is known that if some problem under consideration involved a reaction process accompanied by diffusion, then it can be mathematically described with a set of partial differential equations for the unknown quantities of the system. These quantities may be mass concentrations in chemical reaction processes, temperature in heat conduction, population densities in population dynamics, and many others.

To our knowledge, the first works on the study of a reaction-diffusion equation were papers by Kolmogorov, Petrovskii, Piskunov [KoPePi37] and Fisher [Fish37]. As turned out over the years, reaction-diffusion systems are useful models to describe very different phenomena in physics, chemistry, biology, and medicine. At the present time, this field is a well-developed area of the theory of partial differential equations which includes qualitative properties of solutions both for the reaction-diffusion equation and system of equations.

In recent years, materials with complex structure are widely used in engineering devices, biology, and other fields of science. It is known that many properties of materials are controlled by their geometrical structure. Therefore, the study of the influence of the material microstructure can improve its useful properties and reduce undesirable effects. The main methods for this study are asymptotic methods for boundary value problems (BVPs) in domains with complex structure: perforated domains, grid-domains, domains with rapidly oscillating boundaries, thick junctions, etc.

T.A. Mel'nyk (✉)
Taras Shevchenko National University of Kyiv, Volodymyrska St., 64/13, Kyiv 01601, Ukraine
e-mail: melnyk@imath.kiev.ua



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_33

Successful applications of thick-junction constructions in nanotechnologies and microtechnique have stimulated active investigation of BVPs in thick junctions with more complex (see [Mel08, BlGaMe08, DurMel12, CheMel14] and the references therein).

In this chapter, new results for a reaction-diffusion system in a thick junction of a new type, namely *thick fractal junction*, are presented. Many nerve and blood systems, root systems, and industrial systems have structure of thick fractal junctions.

## 33.2 Statement of the Problem

Let $\Omega_0$ be a bounded domain in $\mathbb{R}^2$ with the Lipschitz boundary $\partial\Omega_0$ and $\Omega_0 \subset \{x := (x_1, x_2) \in \mathbb{R}^2 : x_2 > 0\}$. Let $\partial\Omega_0$ contain the segment $I_0 = \{x : x_1 \in [0, a],\ x_2 = 0\}$. We also assume that there exists a positive number $\delta_0$ such that $\Omega_0 \cap \{x : 0 < x_2 < \delta_0\} = \{x : x_1 \in (0, a),\ x_2 \in (0, \delta_0)\}$.

Let $a, l_1, l_2, l_3$ be positive numbers, $h_0, h_{1,1}, h_{1,2}, h_{2,1}, h_{2,2}, h_{2,3}, h_{2,4}$ be fixed numbers from the interval $(0, 1)$ and $h_{1,1} + h_{1,2} < h_0$, $h_{2,1} + h_{2,2} < h_{1,1}$, $h_{2,3} + h_{2,4} < h_{1,2}$. Let us also introduce a small parameter $\varepsilon = \frac{a}{N}$, where $N$ is a large positive integer.

A model thick fractal junction $\Omega_\varepsilon$ (see Figure 33.1) consists of the junction's body $\Omega_0$,

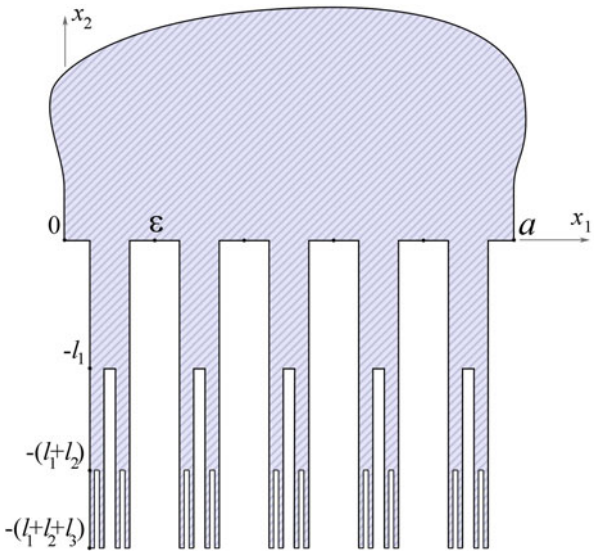


**Fig. 33.1** A model thick fractal junction $\Omega_\varepsilon$.

- a large number of the thin rods $G_\varepsilon^{(0)} = \bigcup_{j=1}^{N-1} G_j^{(0)}(\varepsilon)$,

$$G_j^{(0)}(\varepsilon) = \Big\{ x : \Big| x_1 - \varepsilon(j+\tfrac{1}{2}) \Big| < \frac{\varepsilon h_0}{2}, \quad x_2 \in (-l_1, 0] \Big\},$$

from the zero layer,

- a large number of the thin rods $G_\varepsilon^{(1,m)} = \bigcup_{j=1}^{N-1} G_j^{(1,m)}(\varepsilon)$,

$$G_j^{(1,m)}(\varepsilon) = \Big\{ x : |x_1 - \varepsilon(j+b_{1,m})| < \frac{\varepsilon h_{1,m}}{2}, \quad x_2 \in \big(-l_2 - l_1, -l_1\big] \Big\},$$

from the first branching layer, where $m \in \{1,2\}$ and

$$b_{1,1} = \frac{1 - h_0 + h_{1,1}}{2}, \quad b_{1,2} = \frac{1 + h_0 - h_{1,2}}{2},$$

- and a large number of the thin rods $G_\varepsilon^{(2,m)} = \bigcup_{j=1}^{N-1} G_j^{(2,m)}(\varepsilon)$,

$$G_j^{(2,m)}(\varepsilon) = \Big\{ x : |x_1 - \varepsilon(j+b_{2,m})| < \frac{\varepsilon h_{2,m}}{2}, \\ x_2 \in \big(-l_3 - l_2 - l_1, -l_2 - l_1\big] \Big\},$$

from the second branching layer, where $m \in \{1,2,3,4\}$ and

$$b_{2,1} = \frac{1 - h_0 + h_{2,1}}{2}, \quad b_{2,2} = \frac{1 - h_0 + 2h_{1,1} - h_{2,2}}{2},$$
$$b_{2,3} = \frac{1 + h_0 - 2h_{1,2} + h_{2,3}}{2}, \quad b_{2,4} = \frac{1 + h_0 - h_{2,4}}{2}.$$

Thus, $\Omega_\varepsilon = \Omega_0 \bigcup G_\varepsilon^{(0)} \bigcup G_\varepsilon^{(1)} \bigcup G_\varepsilon^{(2)}$, where $G_\varepsilon^{(1)} = \bigcup_{m=1}^{2} G_\varepsilon^{(1,m)}$, and $G_\varepsilon^{(2)} = \bigcup_{m=1}^{4} G_\varepsilon^{(2,m)}$. The parameter $\varepsilon$ characterizes the distance between neighboring thin branches and also their thickness. Precisely, each branch $G_j^{(i,m)}(\varepsilon)$ has small cross-section of size $\mathcal{O}(\varepsilon)$ and constant height. In addition, at fixed $j \in \{0,1,\dots,N-1\}$ branches $G_j^{(0)}(\varepsilon)$, $\{G_j^{(1,m)}(\varepsilon)\}_{m=1}^2$, $\{G_j^{(2,m)}(\varepsilon)\}_{m=1}^4$ form the tree with two branching levels. These trees are $\varepsilon$-periodically distributed along the segment $I_0$.

In $\Omega_\varepsilon$ we consider the following reaction–diffusion system:

$$\begin{cases} \partial_t \mathbf{u}^\varepsilon - \mathfrak{D}\Delta_x \mathbf{u}^\varepsilon + \mathbf{k}(\mathbf{u}^\varepsilon) = \mathbf{f}(x,t) & \text{in } \Omega_\varepsilon \times (0,T), \\ \partial_\nu \mathbf{u}^\varepsilon + \varepsilon^\alpha \kappa(\mathbf{u}^\varepsilon) = \varepsilon^\beta \mathbf{g}(x,t) & \text{on } \Upsilon_\varepsilon \times (0,T), \\ \partial_\nu \mathbf{u}^\varepsilon = 0 & \text{on } (\partial\Omega_\varepsilon \setminus \Upsilon_\varepsilon) \times (0,T), \\ \mathbf{u}^\varepsilon|_{t=0} = 0 & \text{in } \Omega_\varepsilon, \end{cases} \tag{33.1}$$

where $\mathbf{u}^\varepsilon = (u_1^\varepsilon, \dots, u_N^\varepsilon)$ is an unknown vector-valued function; the reaction terms $\mathbf{k} = (k_1, \dots, k_N)$ and $\kappa = (\kappa_1, \dots, \kappa_N)$ model the chemistry inside of $\Omega_\varepsilon$ and on the vertical boundaries $\Upsilon_\varepsilon$ of the thin rods, respectively; the diagonal matrix $\mathfrak{D} = \operatorname{diag}(D_1, \dots, D_N)$ introduces the diffusion positive constants $D_1, \dots, D_N$; and the parameters $\alpha, \beta \geq 1$.

The main assumptions: given vector-functions

1. $\mathbf{f} \in L^2\big(0,T;L^2(\Omega_\varepsilon;\mathbb{R}^N)\big)$, $\mathbf{g} \in L^2\big(0,T;L^2(D_\ell;\mathbb{R}^N)\big)$ and $\operatorname{supp}(f_j) \subseteq \Omega_0$, $j = 1, \dots, N$, where $D_\ell = (0,a) \times (-l_1 - l_2 - l_3, 0)$;
2. the reaction terms $\mathbf{k} : \mathbb{R}^N \mapsto \mathbb{R}^N$ and $\kappa : \mathbb{R}^N \mapsto \mathbb{R}^N$ are smooth Lipschitz vector-functions such that their Jacobian matrices are positive defined in the following sense: $\exists\, \chi_1, \chi_2 > 0 \;\; \forall \mathbf{p} \in \mathbb{R}^N \;\; \forall \mathbf{q} \in \mathbb{R}^N$ :

$$\chi_1 |\mathbf{q}|^2 \leq \sum_{i,j=1}^{N} \partial_{u_i} k_j(\mathbf{p}) q_i q_j \leq \chi_2 |\mathbf{q}|^2, \tag{33.2}$$

the similar inequalities for $\kappa$.

Our aim is to develop an asymptotic efficient method allowing us to reproduce qualitative properties of the solution to the nonlinear reaction–diffusion system (33.1) in the thick fractal junction $\Omega_\varepsilon$ as $\varepsilon \to 0$, i.e., when the number of the attached thin trees infinitely increases and their thickness vanishes. In particular, we want

- to construct the asymptotic approximation for the solution $\mathbf{u}^\varepsilon$,
- to find the corresponding homogenized (limit) problem for problem (33.1) as $\varepsilon \to 0$,
- to prove the corresponding asymptotic estimate for the difference between the solution $\mathbf{u}^\varepsilon$ and constructed approximation,
- to study the influence of the parameters $\alpha$ and $\beta$ on the asymptotic behavior of the solution.

### *33.2.1 Comments on the Statement*

Standard assumptions for nonlinear terms of BVPs are as follows: $\mathbf{k}$ and $\kappa$ are Lipschitz continuous. This hypothesis in particular implies $|\mathbf{k}(\mathbf{p})| \leq C(1 + |\mathbf{p}|)$ for each $\mathbf{p} \in \mathbb{R}^N$ and some constant $C$. This is enough to state that problem (33.1) has a unique solution. But, if we want to construct some approximation for a solution and to prove the corresponding estimate, we need some kind of a coercivity condition on the nonlinearity. Usually it reads as follows: $\mathbf{k}(\mathbf{p}) \cdot \mathbf{p} \geq C_1 |\mathbf{p}|^2 - C_2$ for all $\mathbf{p} \in \mathbb{R}^N$ and appropriate constants $C_1 > 0$, $C_2 \geq 0$.

Many physical processes, especially in chemistry and medicine, have monotonic nature. Therefore, it is naturally to impose special monotonicity conditions on the nonlinear terms. In our case we propose (33.2). If $N = 1$, then condition (33.2) is equivalent to $\chi_1 \leq k'(p) \leq \chi_2$ for a.e. $p \in \mathbb{R}$. For instance, the following functions

$$k(p)=\lambda p+\cos p \quad (\lambda>1); \qquad k(s)=\frac{\lambda p}{1+\nu p}, \quad p\in\mathbb{R}_+ \quad (\lambda,\nu>0)$$

satisfy this condition. The last one corresponds to the Michaelis–Menten hypothesis in biochemical reactions and to the Langmuir kinetics adsorption models (see [Pao92]).

From condition (33.2) it is easy to deduce the inequalities

$$\begin{aligned}\chi_1|\mathbf{p}-\mathbf{q}|^2\le\big(\mathbf{k}(\mathbf{p})-\mathbf{k}(\mathbf{q})\big)\cdot(\mathbf{p}-\mathbf{q})\le\chi_2|\mathbf{p}-\mathbf{q}|^2,\\ |\mathbf{k}(\mathbf{p})|\le c_1(1+|\mathbf{p}|), \qquad \mathbf{k}(\mathbf{p})\cdot\mathbf{p}\ge c_2|\mathbf{p}|^2-c_3,\end{aligned} \tag{33.3}$$

where $c_1>0,\ c_2>0,\ c_3\ge 0$.

Using these inequalities, we verify that the operator $A_\varepsilon(t)$, which corresponds to problem (33.1) and defined by

$$\langle A_\varepsilon(t)\mathbf{u},\mathbf{v}\rangle_\varepsilon:=\int\limits_{\Omega_\varepsilon}\Big(\sum_{j=1}^N D_j\nabla_x u_j\cdot\nabla_x v_j+\mathbf{k}(\mathbf{u})\cdot\mathbf{v}\Big)dx+\varepsilon^\alpha\int\limits_{\Upsilon_\varepsilon}\kappa(\mathbf{u})\cdot\mathbf{v}\,dx_2$$

for all $\mathbf{u},\mathbf{v}\in H^1(\Omega_\varepsilon;\mathbb{R}^N)$ and a.a. $t\in[0,T]$, is bounded, strictly monotonic, semicontinuous, and coercive. Here the brackets $\langle\cdot,\cdot\rangle_\varepsilon$ denote the pairing of the adjoint $\big(H^1(\Omega_\varepsilon;\mathbb{R}^N)\big)^*$ with $H^1(\Omega_\varepsilon;\mathbb{R}^N)$.

Thus, according to Corollary 4.1 in [Sho97], the problem (33.1) has a unique weak solution for each fixed value of $\varepsilon$.

A function $\mathbf{u}^\varepsilon\in L^2\big(0,T;H^1(\Omega_\varepsilon;\mathbb{R}^N)\big)$, $\partial_t\mathbf{u}^\varepsilon\in L^2\big(0,T;\big(H^1(\Omega_\varepsilon;\mathbb{R}^N)\big)^*\big)$, is a weak solution to the problem (33.1) if

$$\langle\partial_t\mathbf{u}^\varepsilon,\mathbf{v}\rangle_\varepsilon+\langle A_\varepsilon(t)\mathbf{u}^\varepsilon,\mathbf{v}\rangle_\varepsilon=\langle F_\varepsilon(t),\mathbf{v}\rangle_\varepsilon \quad \text{a.e. } t\in(0,T)$$

for each $\mathbf{v}\in H^1(\Omega_\varepsilon;\mathbb{R}^N)$ and $\mathbf{u}^\varepsilon|_{t=0}=0$. Here, $F_\varepsilon(t)\in\big(H^1(\Omega_\varepsilon;\mathbb{R}^N)\big)^*$ is the linear functional defined by

$$\langle F_\varepsilon(t),\mathbf{v}\rangle_\varepsilon:=\int_{\Omega_0}\mathbf{f}\cdot\mathbf{v}\,dx+\varepsilon^\beta\int_{\Upsilon_\varepsilon}\mathbf{g}\cdot\mathbf{v}\,dx_2, \quad \forall\mathbf{v}\in H^1(\Omega_\varepsilon;\mathbb{R}^N),$$

for a.e. $t\in[0,T]$. In addition, it is known that $\mathbf{u}^\varepsilon\in C([0,T];L^2(\Omega_\varepsilon;\mathbb{R}^N))$ and thus the equality $\mathbf{u}^\varepsilon|_{t=0}=0$ makes sense.

It should be noted here that the asymptotic behavior of solutions to the reaction–diffusion equation in different kind of thin domains with the uniform Neumann conditions was studied in [MarRyb01, ACPS11]. The convergence theorems were proved under the following assumptions for the nonlinear function $f$: in [ACPS11] it is a $C^2$-function with bounded derivatives and

$$\limsup_{|p|\to+\infty}\frac{f(p)}{p}<0; \tag{33.4}$$

in [MarRyb01] it is a $C^1$-function, $|f'(p)|\le C(1+|p|^q)$, where $q\in(0,+\infty)$, and the dissipative condition (33.4) is satisfied.

Let us note that the convergence theorem for the solution to our problem (33.1) can be proved under weaker assumptions on the vector-functions $\mathbf{k}$ and $\kappa$, namely $\mathbf{k}(\mathbf{0}) = \mathbf{0}$ and $\kappa(\mathbf{0}) = \mathbf{0}$, all their components are increasing functions from $C^1(\mathbb{R})$, and

$$|\nabla \mathbf{k}(\mathbf{p})| + |\nabla \kappa(\mathbf{p})| \le C\big(1 + |\mathbf{p}|^{q-1}\big),$$

where $q \in [1, +\infty)$.

## 33.3 Formal Asymptotics and Homogenized Problem

In this and next sections, for simplicity and clarity the scalar case $(N = 1)$ is considered. We propose the following asymptotic expansions for the solution:

$$u^\varepsilon(x,t) \approx u_0^+(x,t) + \sum_{n=1}^{+\infty} \varepsilon^n u_n^+(x,t) \tag{33.5}$$

in $\Omega_0 \times (0,T)$; and

$$u^\varepsilon(x,t) \approx u_0^{(i,m)}(x,t) + \sum_{n=1}^{+\infty} \varepsilon^n u_n^{(i,m)}\big(x, \tfrac{x_1}{\varepsilon} - j, t\big) \tag{33.6}$$

in every thin rod $G_j^{(i,m)}(\varepsilon) \times (0,T), j = 0, \dots, N-1$, from each layer $(i = 0,1,2)$. The index $m \in \{1,2\}$ for $i = 1$, $m \in \{1,2,3,4\}$ for $i = 2$, and it is omitted for $i = 0$, i.e., $G_j^{(0,m)}(\varepsilon) = G_j^{(0)}(\varepsilon)$ and $u_n^{(0,m)} = u_n^{(0)}$ in (33.6). The asymptotic expansions (33.5) and (33.6) are usually called *outer expansions*.

To find transmission conditions both in the joint zone $I_0$ and in each of the branching zones $I_1 = \{x : x_1 \in [0,a],\ \ x_2 = -l_1\}$ and $I_2 = \{x : x_1 \in [0,a],\ \ x_2 = -l_1 - l_2\}$, we should construct *inner expansions* for the solution in neighborhoods of these zones.

In a neighborhood of $I_0 \cap \Omega_\varepsilon$, we propose the ansatz

$$u^\varepsilon \approx u_0^+(x_1,0,t) + \varepsilon\Big(Z_1\big(\tfrac{x}{\varepsilon}\big)\partial_{x_1} u_0^+(x_1,0,t) + Z_2\big(\tfrac{x}{\varepsilon}\big)\partial_{x_2} u_0^+(x_1,0,t)\Big) + \dots, \tag{33.7}$$

where $Z_1$ and $Z_2$ are 1-periodic junction-layer solutions to problems

$$\begin{aligned}
-\Delta_\xi\, Z_q(\xi) &= 0, && \xi \in \Pi_0,\\
\partial_{\xi_1}^p Z_q(\xi)|_{\xi_1=0} &= \partial_{\xi_1}^p Z_q(\xi)|_{\xi_1=1}, && \xi \in \partial\Pi^+,\ \ \xi_2 > 0,\ \ p = 0,1.\\
\partial_{\xi_2} Z_q(\xi_1,0) &= 0, && \xi_1 \in (0,1) \setminus \big(\tfrac{1}{2} - \tfrac{h_0}{2}, \tfrac{1}{2} - \tfrac{h_0}{2}\big),\\
\partial_{\xi_1} Z_q(\xi) &= -\delta_{q,1}, && \xi \in \partial\Pi_{h_1}^- \cap \{\xi :\ \xi_2 < 0\},\ \ q = 1,2.
\end{aligned} \tag{33.8}$$

Here $\Pi_0$ is the union of two semi-strips $\Pi^+ := (0,1) \times (0,+\infty)$ and $\Pi_{h_0}^- := \big(\frac{1}{2} - \frac{h_0}{2}, \frac{1}{2} + \frac{h_0}{2}\big) \times (-\infty, 0]$.

The existence and the main asymptotic relations for solutions of problems (33.8) can be obtained from general results about the asymptotic behavior of solutions to elliptic problems in domains with different exits to infinity [KonOle83, NazPla94]. But, thanks to the symmetry of $\Pi_0$ with respect to $\frac{1}{2}$ we can define more exactly the asymptotic relations and detect other properties of junction-layer solutions (see Lemma 4.1 and Corollary 4.1 from [Mel99], see also [MelNaz96]). From those results it follows the following proposition.

**Lemma 1.** *There exist unique solutions $Z_1^{(0)}, Z_2^{(0)} \in H^1_{loc,\xi_2}(\Pi_0)$ to problems (33.8), respectively, which have the following differentiable asymptotics*

$$Z_1^{(0)}(\xi) = \begin{cases} \mathcal{O}(\exp(-2\pi\xi_2)), & \xi_2 \to +\infty, \\ \left(-\xi_1 + \frac{1}{2}\right) + \mathcal{O}(\exp(\pi h_0^{-1}\xi_2)), & \xi_2 \to -\infty, \end{cases}$$

$$Z_2^{(0)}(\xi) = \begin{cases} \xi_2 + \mathcal{O}(\exp(-2\pi\xi_2)), & \xi_2 \to +\infty, \\ \frac{\xi_2}{h_0} + C_2 + \mathcal{O}(\exp(\pi h_0^{-1}\xi_2)), & \xi_2 \to -\infty, \end{cases}$$

*Moreover, function $Z_1^{(0)}$ is odd in $\xi_1$ and function $Z_2^{(0)}$ is even in $\xi_1$ with respect to $\frac{1}{2}$.*

Recall that a function $Z$ belongs to the Sobolev space $H^1_{loc,\xi_2}(\Pi_0)$ if for every $R > 0$ this function $Z \in H^1(\Pi_0 \cap \{\xi : |\xi_2| < R\})$.

In a neighborhood of $I_1 \cap \Omega_\varepsilon$, we propose the ansatz

$$\begin{aligned} u^\varepsilon(x,t) \approx u_0^{(0)}(x_1,-l_1,t) + \varepsilon\Big(Z_1^{(1)}\big(\tfrac{x_1}{\varepsilon}, \tfrac{x_2+l_1}{\varepsilon}\big)\partial_{x_1}u_0^{(0)}(x_1,-l_1,t) \\ + \Big\{\eta_1(x_1,t)\,\Xi_1^{(1)}\big(\tfrac{x_1}{\varepsilon}, \tfrac{x_2+l_1}{\varepsilon}\big) + (1-\eta_1(x_1,t))\,\Xi_2^{(1)}\big(\tfrac{x_1}{\varepsilon}, \tfrac{x_2+l_1}{\varepsilon}\big)\Big\} \\ \times \partial_{x_2}u_0^{(0)}(x_1,-l_1,t)\Big) + \dots \end{aligned} \tag{33.9}$$

where the coefficients $\Xi_1^{(1)}$ and $\Xi_2^{(1)}$ are solutions of the problem

$$\Delta_\xi\, \Xi(\xi) = 0 \quad \text{on } \Pi_1, \qquad \partial_{\nu_\xi}\Xi(\xi) = 0 \quad \text{on } \partial\Pi_1. \tag{33.10}$$

Here $\Pi_1 = \Pi_{h_0}^+ \cup \Pi_{1,1}^- \cup \Pi_{1,2}^-$, $\Pi_{h_0}^+ = \left(\frac{1}{2} - \frac{h_0}{2}, \frac{1}{2} + \frac{h_0}{2}\right) \times (0, +\infty)$, $\Pi_{1,m}^- = \big(b_{1,m} - \frac{h_{1,m}}{2}, b_{1,m} + \frac{h_{1,m}}{2}\big) \times (-\infty, 0]$, $m = 1, 2$.

Again using the approach mentioned before Lemma 1, we prove Lemma 2.

**Lemma 2.** *There exist two solutions $\Xi_1, \Xi_2 \in H^1_{loc,\xi_2}(\Pi_1)$ to problem (33.10), which have the differentiable asymptotics*

$$\Xi_1(\xi)=\begin{cases}\xi_2+\mathcal{O}\big(\exp(-\frac{\pi\xi_2}{h_0})\big), & \xi_2\to+\infty,\ \xi\in\Pi^+_{h_0},\\ \frac{h_0}{h_{1,1}}\,\xi_2+C_1^{(1)}+\mathcal{O}\big(\exp(\frac{\pi\xi_2}{h_{1,1}})\big), & \xi_2\to-\infty,\ \xi\in\Pi^-_{1,1},\\ C_2^{(1)}+\mathcal{O}\big(\exp(\frac{\pi\xi_2}{h_{1,2}})\big), & \xi_2\to-\infty,\ \xi\in\Pi^-_{1,2},\end{cases}\tag{33.11}$$

$$\Xi_2(\xi)=\begin{cases}\xi_2+\mathcal{O}\big(\exp(-\frac{\pi\xi_2}{h_0})\big), & \xi_2\to+\infty,\ \xi\in\Pi^+_{h_0},\\ C_1^{(2)}+\mathcal{O}\big(\exp(\frac{\pi\xi_2}{h_{1,1}})\big), & \xi_2\to-\infty,\ \xi\in\Pi^-_{1,1},\\ \frac{h_0}{h_{1,2}}\,\xi_2+C_2^{(2)}+\mathcal{O}\big(\exp(\frac{\pi\xi_2}{h_{1,2}})\big), & \xi_2\to-\infty,\ \xi\in\Pi^-_{1,2}.\end{cases}\tag{33.12}$$

*Any other solution to the homogeneous problem (33.10), which has polynomial grow at infinity, can be presented as a linear combination* $c_0+c_1\Xi_1+c_2\Xi_2$.

The function $Z_1^{(1)}$ is a solution to the problem

$$\begin{aligned}-\Delta_\xi\, Z(\xi)&=0, && \xi\in\Pi_1,\\ \partial_{\xi_1}Z(\xi)&=-1, && \xi\in\partial_\parallel\Pi_1,\\ \partial_{\xi_2}Z(\xi_1,0)&=0, && (\xi_1,0)\in\partial\Pi_1\setminus\partial_\parallel\Pi_1.\end{aligned}\tag{33.13}$$

**Lemma 3.** *There exists the unique solution* $Z\in H^1_{loc,\xi_2}(\Pi_1)$ *to problems* (33.13), *which has the differentiable asymptotics*

$$Z(\xi)=\begin{cases}-\xi_1+\frac12+\mathcal{O}\big(\exp(-\frac{\pi\xi_2}{h_0})\big), & \xi_2\to+\infty,\ \xi\in\Pi^+_{h_0},\\ -\xi_1+b_{1,1}+C_1+\mathcal{O}\big(\exp(\frac{\pi\xi_2}{h_{1,1}})\big), & \xi_2\to-\infty,\ \xi\in\Pi^-_{1,1},\\ -\xi_1+b_{1,2}+C_2+\mathcal{O}\big(\exp(\frac{\pi\xi_2}{h_{1,2}})\big), & \xi_2\to-\infty,\ \xi\in\Pi^-_{1,2},\end{cases}\tag{33.14}$$

*where* $C_1, C_2$ *are some fixed constants.*

Thus, we set $\Xi_1^{(1)}=\Xi_1$, $\Xi_2^{(1)}=\Xi_2$ and $Z_1^{(1)}=Z$ in (33.9).
In a neighborhood of $I_2\cap\Omega_\varepsilon$, we propose the two ansatzes

$$\begin{aligned}u_\varepsilon(x,t)\approx u_0^{(1,1)}(x_1,0,t)+\varepsilon\Big(&Z_1^{(2,1)}\big(\tfrac{x_1}{\varepsilon},\tfrac{x_2+l_1+l_2}{\varepsilon}\big)\,\partial_{x_1}u_0^{(1,1)}(x_1,0,t)\\ &+\Big\{\eta_{2,1}(x_1,t)\,\Xi_1^{(2,1)}\big(\tfrac{x_1}{\varepsilon},\tfrac{x_2+l_1+l_2}{\varepsilon}\big)\\ &+(1-\eta_{2,1}(x_1,t))\,\Xi_2^{(2,1)}\big(\tfrac{x_1}{\varepsilon},\tfrac{x_2+l_1+l_2}{\varepsilon}\big)\Big\}\partial_{x_2}u_0^{(1,1)}(x_1,0,t)\Big)+\ldots\end{aligned}\tag{33.15}$$

in a neighborhood of $I_2 \cap \Big( G_\varepsilon^{(1,1)} \bigcup \big( \bigcup_{m=1}^{2} G_\varepsilon^{(2,m)} \big) \Big)$, and the second one

$$u_\varepsilon(x,t) \approx u_0^{(1,2)}(x_1,0,t) + \varepsilon \Big( Z_1^{(2,2)} \big( \tfrac{x_1}{\varepsilon}, \tfrac{x_2+l_1+l_2}{\varepsilon} \big)\, \partial_{x_1} u_0^{(1,2)}(x_1,0,t) + \Big\{ \eta_{2,2}(x_1,t)\, \Xi_1^{(2,2)} \big( \tfrac{x_1}{\varepsilon}, \tfrac{x_2+l_1+l_2}{\varepsilon} \big) + (1-\eta_{2,2}(x_1,t))\, \Xi_2^{(2,2)} \big( \tfrac{x_1}{\varepsilon}, \tfrac{x_2+l_1+l_2}{\varepsilon} \big) \Big\} \partial_{x_2} u_0^{(1,2)}(x_1,0,t) \Big) + \dots \tag{33.16}$$

in a neighborhood of $I_2 \cap \Big( G_\varepsilon^{(1,2)} \bigcup \big( \bigcup_{m=3}^{4} G_\varepsilon^{(2,m)} \big) \Big)$.

The coefficients $\Xi_1^{(2,1)}, \Xi_2^{(2,1)}$ and $\Xi_1^{(2,2)}, \Xi_2^{(2,2)}$ are solutions to problem (33.10) but now in $\Pi_2^{(1)}$ and $\Pi_2^{(2)}$, respectively, where $\Pi_2^{(1)} = \Pi_{1,1}^{+} \cup \Pi_{2,1}^{-} \cup \Pi_{2,2}^{-}$ and $\Pi_2^{(2)} = \Pi_{1,2}^{+} \cup \Pi_{2,3}^{-} \cup \Pi_{2,4}^{-}$, $\Pi_{1,m}^{+} = \big( b_{1,m} - \frac{h_{1,m}}{2}, b_{1,m} + \frac{h_{1,m}}{2} \big) \times (0, +\infty)$, $m = 1,2$, $\Pi_{2,m}^{-} = \big( b_{2,m} - \frac{h_{2,m}}{2}, b_{2,m} + \frac{h_{2,m}}{2} \big) \times (-\infty, 0]$, $m = 1,2,3,4$. From Lemma 2 it follows that they have the corresponding differentiable asymptotics (33.11) and (33.12). Functions $\eta_{2,1}$ and $\eta_{2,2}$ are defined from matching conditions.

The coefficients $Z_1^{(2,1)}$ and $Z_1^{(2,2)}$ are solutions to problem (33.13) in $\Pi_2^{(1)}$ and $\Pi_2^{(2)}$, respectively. Applying results of Lemma 3, we can state that there exist the unique solutions with the differentiable asymptotics

$$Z_1^{(2,1)}(\xi) = \begin{cases} -\xi_1 + b_{1,1} + \mathcal{O}\big( \exp(-\frac{\pi \xi_2}{h_{1,1}}) \big), & \xi_2 \to +\infty,\ \xi \in \Pi_{1,1}^{+}, \\ -\xi_1 + b_{2,1} + C_1^{(3)} + \mathcal{O}\big( \exp(\frac{\pi \xi_2}{h_{2,1}}) \big), & \xi_2 \to -\infty,\ \xi \in \Pi_{2,1}^{-}, \\ -\xi_1 + b_{2,2} + C_2^{(3)} + \mathcal{O}\big( \exp(\frac{\pi \xi_2}{h_{2,2}}) \big), & \xi_2 \to -\infty,\ \xi \in \Pi_{2,2}^{-}, \end{cases}$$

$$Z_1^{(2,2)}(\xi) = \begin{cases} -\xi_1 + b_{1,2} + \mathcal{O}\big( \exp(-\frac{\pi \xi_2}{h_{1,2}}) \big), & \xi_2 \to +\infty,\ \xi \in \Pi_{1,2}^{+}, \\ -\xi_1 + b_{2,3} + C_1^{(4)} + \mathcal{O}\big( \exp(\frac{\pi \xi_2}{h_{2,3}}) \big), & \xi_2 \to -\infty,\ \xi \in \Pi_{2,3}^{-}, \\ -\xi_1 + b_{2,4} + C_2^{(4)} + \mathcal{O}\big( \exp(\frac{\pi \xi_2}{h_{2,4}}) \big), & \xi_2 \to -\infty,\ \xi \in \Pi_{2,4}^{-}. \end{cases}$$

If we put (33.5) and (33.6) in problem (33.1) and collect the coefficients of the same power of $\varepsilon$ considering (33.2) and then apply the method of matched asymptotic expansions (the asymptotics of the leading terms of each outer expansions (33.5) and (33.6) as $x_2 \to \pm - l_i$ have to coincide with the corresponding asymptotics of the leading terms of the inner expansions (33.7), (33.9), (33.15), (33.16) as $\xi_2 \to \pm\infty$, respectively), we derive the following homogenized problem (for more detail, see [Mel14]):

$$
\begin{gathered}
\partial_t u_0^+ - \Delta u_0^+ + k(u_0^+) = f \quad \text{in } \Omega_0 \times (0,T),\\
\partial_\nu u_0^+ = 0 \quad \text{on } \big(\partial\Omega_0 \setminus I_0\big) \times (0,T),\\
\partial_t u_0^{(i,m)} - \partial^2_{x_2x_2} u_0^{(i,m)} + k\big(u_0^{(i,m)}\big) + \tfrac{2\delta_{\alpha,1}}{h_{i,m}}\kappa\big(u_0^{(i,m)}\big) = \tfrac{2\delta_{\beta,1}}{h_{i,m}} g \text{ in } D_i \times (0,T),\\
m \in \{1,\dots,2i\}, \quad i = 0,1,2,\\
u_0^+ = u_0^{(0)}, \quad \partial_{x_2} u_0^+ = h_0 \partial_{x_2} u_0^{(0)} \quad \text{on } I_0 \times (0,T),\\
u_0^{(0)} = u_0^{(1,m)} \quad \text{on } I_1 \times (0,T), \quad m = 1,2,\\
h_0 \partial_{x_2} u_0^{(0)} = h_{1,1} \partial_{x_2} u_0^{(1,1)} + h_{1,2} \partial_{x_2} u_0^{(1,2)} \quad \text{on } I_1 \times (0,T),\\
u_0^{(1,1)} = u_0^{(2,m)} \quad \text{on } I_2 \times (0,T), \quad m = 1,2,\\
h_{1,1} \partial_{x_2} u_0^{(1,1)} = h_{2,1} \partial_{x_2} u_0^{(2,1)} + h_{2,2} \partial_{x_2} u_0^{(2,2)} \quad \text{on } I_2 \times (0,T),\\
u_0^{(1,2)} = u_0^{(2,m)} \quad \text{on } I_2 \times (0,T), \quad m = 3,4,\\
h_{1,2} \partial_{x_2} u_0^{(1,2)} = h_{2,3} \partial_{x_2} u_0^{(2,3)} + h_{2,4} \partial_{x_2} u_0^{(2,4)} \quad \text{on } I_2 \times (0,T),\\
\partial_{x_2} u_0^{(2,m)}\big(x_1, -(l_1+l_2+l_3), t\big) = 0, \quad (x_1,t) \in (0,a) \times (0,T), \ \ m = 1,2,3,4,\\
u_0^+|_{t=0} = u_0^{(0)}|_{t=0} = \{u^{(1,m)}\}_{m=1}^2|_{t=0} = \{u^{(2,m)}\}_{m=1}^4|_{t=0} = 0.
\end{gathered}
\tag{33.17}
$$

Recall that the index $m \in \{1,2\}$ for $i=1$, $m \in \{1,2,3,4\}$ for $i=2$, and $m$ is absent if $i=0$.

To give appropriately the definition of a weak solution of the homogenized problem, let us first introduce an anisotropic Sobolev space $\mathbf{H}$ of multi-sheeted functions. A multi-sheeted function

$$
\varphi(x) := \begin{cases} \varphi^+(x), & x \in \Omega_0,\\ \varphi^{(0)}(x), & x \in D_0,\\ \varphi^{(1,m)}(x), & x \in D_1, \ m = 1,2,\\ \varphi^{(2,m)}(x), & x \in D_2, \ m = 1,2,3,4, \end{cases}
$$

belongs to $\mathbf{H}$ if $\varphi^+ \in H^1(\Omega_0)$, $\{\varphi^{(i,m)}\}_{m=1}^{2i} \subset L^2(D_i)$, there exist weak derivatives $\{\partial_{x_2}\varphi^{(i,m)}\}_{m=1}^{2i} \subset L^2(D_i)$, $i=0,1,2$, and

$$
\varphi^+|_{I_0} = \varphi^{(0)}|_{I_0}, \qquad \varphi^{(0)}|_{I_1} = \varphi^{(1,1)}|_{I_1} = \varphi^{(1,2)}|_{I_1},
$$

$$
\varphi^{(1,1)}|_{I_2} = \varphi^{(2,1)}|_{I_2} = \varphi^{(2,2)}|_{I_2}, \qquad \varphi^{(1,2)}|_{I_2} = \varphi^{(2,3)}|_{I_2} = \varphi^{(2,4)}|_{I_2}.
$$

Obviously, the space $\mathbf{H}$ is continuously and densely embedded in the Hilbert space $\mathbf{V}$ of multi-sheeted functions whose components belong to the corresponding $L^2$-spaces. The scalar products in these spaces are defined as follows:

$$(\varphi,\psi)_{\mathbf{V}} := (\varphi^+,\psi^+)_{L^2(\Omega_0)} + \sum_{i=0}^{2}\sum_{m=1}^{2i}(\varphi^{(i,m)},\psi^{(i,m)})_{L^2(D_i)},$$

$$(\varphi,\psi)_{\mathbf{H}} := (\varphi,\psi)_{\mathbf{V}} + (\nabla\varphi^+,\nabla\psi^+)_{L^2(\Omega_0)} + \sum_{i=0}^{2}\sum_{m=1}^{2i}(\partial_{x_2}\varphi^{(i,m)},\partial_{x_2}\psi^{(i,m)})_{L^2(D_i)}.$$

For a.e. $t\in(0,T)$ we introduce the following operator $\mathscr{A}(t):\mathbf{H}\mapsto\mathbf{H}^*$:

$$\langle\mathscr{A}(t)\varphi,\psi\rangle := \int\limits_{\Omega_0}\left(\nabla\varphi^+\cdot\nabla\psi^+ + k(\varphi^+)\psi^+\right)dx + \sum_{i=0}^{2}\sum_{m=1}^{2i}$$

$$\int\limits_{D_i}\left(h_{i,m}\partial_{x_2}\varphi^{(i,m)}\,\partial_{x_2}\psi^{(i,m)} + h_{i,m}k(\varphi^{(i,m)})\psi^{(i,m)} + 2\delta_{\alpha,1}\kappa(\varphi^{(i,m)})\psi^{(i,m)}\right)dx$$

for all $\varphi,\psi\in L^2(0,T;\mathbf{H})$, and a linear functional $\mathbf{F}(t)\in\mathbf{H}^*$

$$\langle\mathbf{F}(t),\psi\rangle := \int_{\Omega_0} f\,\psi^+\,dx + 2\delta_{\beta,1}\sum_{i=0}^{2}\sum_{m=1}^{2i}\int_{D_i} g\,\psi^{(i,m)}\,dx.$$

Here $\langle\cdot,\cdot\rangle$ is the pairing of $\mathbf{H}^*$ and $\mathbf{H}$.

**Definition 1.** A multi-sheeted function $\mathbf{u}\in L^2(0,T;\mathbf{H})$, with $\mathbf{u}'\in L^2(0,T;\mathbf{H}^*)$, is called a weak solution to the homogenized problem (33.17) if

$$\langle\mathbf{u}'(t),\mathbf{v}\rangle + \langle\mathscr{A}(t)\mathbf{u},\mathbf{v}\rangle = \langle\mathbf{F}(t),\mathbf{v}\rangle \quad \forall\,\mathbf{v}\in\mathbf{H} \text{ and for a.e. } t\in(0,T),$$

and $\mathbf{u}|_{t=0}=0$.

With the help of (33.2) we prove that for a.e. $t\in(0,T)$ the operator $\mathscr{A}$ is bounded, strictly monotone, hemicontinuous, and coercive. Thus, due to Corollary 4.1 [Sho97] problem (33.17) has a unique weak solution.

## 33.4 The Main Results

Let $\mathbf{u} = \left(u^+,u^{(0)},\left\{u^{(1,m)}\right\}_{m=1}^{2},\left\{u^{(2,m)}\right\}_{m=1}^{4}\right)\in L^2(0,T;\mathbf{H})$ be a unique weak solution to the homogenized problem (33.17).

An approximating function $R_\varepsilon$ is constructed as the sum of the leading terms of the outer expansions (33.5), (33.6) and the inner expansion (33.7), (33.9), (33.15), (33.16) in neighborhoods of the joint zone $I_0$ and branching zones $I_1, I_2$, respectively, with the subtraction of the identical terms of their asymptotics because they are summed twice (for more detail, see [Mel14]).

**Theorem 1.** *Suppose that in addition to the assumptions made in § 33.2, the following conditions hold: the function* $f \in C^1(\overline{\Omega_0} \times [0,T])$ *and if the parameter* $\beta = 1$, *then the function* $g \in C^1(\overline{D_\ell} \times [0,T])$ *and it and its derivative with respect to* $x_2$ *vanish at* $x_2 = 0$.

*Then for any* $\rho \in (0,1)$ *there exist positive constants* $C_0, \varepsilon_0$ *such that for all values* $\varepsilon \in (0,\varepsilon_0)$ *the difference between the solution* $u^\varepsilon$ *to problem (33.1) and the approximating function* $R_\varepsilon$ *satisfies the following estimate*

$$\max_{0\le t\le T} \|R_\varepsilon(\cdot,t) - v_\varepsilon(\cdot,t)\|_{L^2(\Omega_\varepsilon)} + \|R_\varepsilon - v_\varepsilon\|_{L^2(0,T;H^1(\Omega_\varepsilon))} \le C_0\Big(\varepsilon + \varepsilon^{1-\rho} + \varepsilon^{\delta_{\alpha,1}(2-\alpha)+\alpha-1} + \varepsilon^{\delta_{\beta,1}(2-\beta)+\beta-1}\Big). \tag{33.18}$$

From Theorem 1 it follows directly the Corollary.

**Corollary 1.** *Let assumptions from Theorem 1 hold. Then*

$$\max_{t\in[0,T]} \Big(\|u^\varepsilon(\cdot,t) - u_0^+(\cdot,t)\|_{L^2(\Omega_0)} + \sum_{i=0}^{2}\sum_{m=1}^{2i} \|u^\varepsilon(\cdot,t) - u_0^{(i,m)}(\cdot,t)\|_{L^2(G_\varepsilon^{(i,m)})}\Big) \le C_0\Big(\varepsilon + \varepsilon^{1-\rho} + \varepsilon^{\delta_{\alpha,1}(2-\alpha)+\alpha-1} + \varepsilon^{\delta_{\beta,1}(2-\beta)+\beta-1}\Big),$$

*where* $u^\varepsilon$ *is the solution to problem* (33.1),

$$\mathbf{u}(x) := \begin{cases} u^+(x), & x \in \Omega_0, \\ u^{(0)}(x), & x \in D_0, \\ u^{(1,m)}(x), & x \in D_1, \ m = 1,2, \\ u^{(2,m)}(x), & x \in D_2, \ m = 1,2,3,4, \end{cases}$$

*is the weak multi-sheeted solution to the homogenized problem (33.17).*

**Acknowledgements** This chapter was written during the author's visit at the University of Lübeck in July-August 2014, supported by the European grant EUMLS-FP7-People-2011-IRSES Project number 295164. The author would also like to thank the Alexander von Humboldt Foundation for supporting his participation in the conference IMSE 2014.

## References

[ACPS11] Arrieta, J.M., Carvalho, A.N., Pereira, M.C., Silva, R.P.: Semilinear parabolic problems in thin domains with a highly oscillatory boundary. Nonlinear Analysis, **74**, 5111–5132 (2011)

[BlGaMe08] Blanchard, D., Gaudiello A., Mel'nyk, T.A.: Boundary homogenization and reduction of dimension in a Kirchhoff-Love plate. SIAM J. Math. Anal. **39**, no. 6, 1764–1787 (2008)

[CheMel14] Chechkin, G.A., Mel'nyk, T.A.: Spatial-skin effect for eigenvibrations of a thick cascade junction with "heavy" concentrated masses. Mathematical Models and Methods in Applied Sciences, **37**, 56–74 (2014)

[DurMel12] Durante, T., Mel'nyk, T.A.: Homogenization of quasilinear optimal control problems involving a thick multilevel junction of type 3:2:1. ESAIM: Control, Optimisation and Calculus of Variations, **18** (2), 583–610 (2012)

[Fish37] Fisher R.A.: The wave of advance of advantageous genes. Ann Eugenics, **7**, 355–369 (1937)

[KoPePi37] Kolmogorov, A.N., Petrovskii I., Piskunov N.: A study of the diffusion equation with increase in the amount of substance and its application to a biology problem. Moskow Univ. Bull. Math. A, **1**(6),1–25 (1937)

[KonOle83] Kondrat'ev, V.A., Oleinik, O.A.: Boundary-value problems for partial differential equations in non-smooth domains. Russian Math. Survays, **38**(2) 1–86 (1983)

[MarRyb01] Prizzi, M., Rybakowski, K.P.: The effect of domain squeezing upon the dynamics of reaction-diffusion equations, Journal of Differential Equations, **173**, 271–320 (2001)

[Mel99] Mel'nyk, T.A.: Homogenization of the Poisson equation in a thick periodic junction. Zeitschrift für Analysis und ihre Anwendungen, **18**(4) 953–975 (1999)

[Mel08] Mel'nyk, T.A.: Homogenization of a boundary-value problem with a nonlinear boundary condition in a thick junction of type 3:2:1, Mathematical Models and Methods in Applied Sciences, **31**, 1005–1027 (2008)

[Mel14] Mel'nyk, T.A.: Asymptotic approximation for the solution to a semi-linear parabolic problem in a thick fractal junction. Journal of Mathematical Analysis and Applications (to appear in 2015); see also preprint arXiv: 1408.2717v1 [math.AP] 12 Aug 2014

[MelNaz96] Mel'nyk, T.A., Nazarov, S.A.: Asymptotics of the Neumann spectral problem solution in a domain of "thick comb". Journal of Mathematical Sciences, **85**(6), 2326–2346 (1997).

[Pao92] Pao C.V.: Nonlinear Parabolic and Elliptic Equations, Plenum Press, New York (1992)

[NazPla94] Nazarov, S.A., Plamenevskii, B.A.: Elliptic Problems in Domains with Piecewise Smooth Boundaries. Walter de Gruyter, Berlin (1994).

[Sho97] Showalter, R. E., Monotone Operators in Banach Space and Nonlinear Partial Differential Equations, Mathematical Surveys and Monographs, Vol.49, American Mathematical Society (1997)

# Chapter 34
# BDIE System in the Mixed BVP for the Stokes Equations with Variable Viscosity

**S.E. Mikhailov and C.F. Portillo**

## 34.1 Introduction

The mixed (Dirichlet-Neumann) boundary value problem for the steady-state Stokes system of PDEs for an incompressible viscous fluid with variable viscosity coefficient is reduced to a system of direct segregated Boundary-Domain Integral Equations (BDIEs). Mapping properties of the potential-type integral operators appearing in these equations are presented in appropriate Sobolev spaces. We also prove the equivalence between the original BVP and the corresponding BDIE system.

Let $\Omega = \Omega^+ \subset \mathbb{R}^3$ be a bounded connected domain with boundary $\partial\Omega = S$, which is a closed and simply connected infinitely differentiable manifold of dimension 2, and $\overline{\Omega} = \Omega \cup S$. The exterior of the domain $\Omega$ is denoted as $\Omega^- = \mathbb{R}^3 \setminus \overline{\Omega}$. Moreover, let $S = \overline{S_D} \cup \overline{S_N}$ where both $S_N$ and $S_D$ are nonempty disjointed and simply connected open manifolds of $S$.

Let $v$ be the velocity vector field $p$ the pressure scalar field and $\mu \in \mathscr{C}^\infty(\Omega)$ be the variable kinematic viscosity of the fluid such that $\mu(\boldsymbol{x}) > c > 0$.

For a compressible fluid the stress tensor operator, $\sigma_{ij}$, for an arbitrary couple $(v,p)$ is defined as

$$\sigma_{ji}(v,p)(\boldsymbol{x}) := -\delta_i^j p(\boldsymbol{x}) + \mu(\boldsymbol{x}) \left( \frac{\partial v_i(\boldsymbol{x})}{\partial x_j} + \frac{\partial v_j(\boldsymbol{x})}{\partial x_i} - \frac{2}{3}\delta_{ij} \operatorname{div} v \right),$$

S.E. Mikhailov (✉) • C.F. Portillo
Brunel University London, Uxbridge, UK
e-mail: Sergey.Mikhailov@brunel.ac.uk; Carlos.Fresneda-Portillo@brunel.ac.uk



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_34

and the Stokes operator is defined as

$$\mathscr{A}_j(v,p)(\boldsymbol{x}) := \frac{\partial}{\partial x_i}\sigma_{ji}(v,p)(\boldsymbol{x}) = \frac{\partial}{\partial x_i}\left(\mu(\boldsymbol{x})\left(\frac{\partial v_j}{\partial x_i}+\frac{\partial v_i}{\partial x_j}-\frac{2}{3}\delta_{ij}\mathrm{div} v\right)\right)-\frac{\partial p}{\partial x_j},\ j,i\in\{1,2,3\},$$

where $\delta_i^j$ is Kronecker symbol. Here and henceforth we assume the Einstein summation in repeated indices from 1 to 3. We also denote the Stokes operator as $\mathscr{A}=\{\mathscr{A}_j\}_{j=1}^3$.

For an incompressible fluid $\mathrm{div}\, v=0$, which reduces the stress tensor operator and the Stokes operator, respectively, to

$$\sigma_{ij}(v,p)(\boldsymbol{x}) = -\delta_i^j p(\boldsymbol{x})+\mu(\boldsymbol{x})\left(\frac{\partial v_i(\boldsymbol{x})}{\partial x_j}+\frac{\partial v_j(\boldsymbol{x})}{\partial x_i}\right),$$

$$\mathscr{A}_j(v,p)(\boldsymbol{x}) = \frac{\partial}{\partial x_i}\left(\mu(\boldsymbol{x})\left(\frac{\partial v_j}{\partial x_i}+\frac{\partial v_i}{\partial x_j}\right)\right)-\frac{\partial p}{\partial x_j}.$$

In what follows $H^s(\Omega)=H_2^s(\Omega)$, $H^s(\partial\Omega)=H_2^s(\partial\Omega)$ are the Bessel potential spaces, where $s\in\mathbb{R}$ is an arbitrary real number (see, e.g., [LiMa73, McL00]). We recall that $H^s$ coincide with the Sobolev–Slobodetski spaces $W_2^s$ for any non-negative $s$. We denote by $\widetilde{H}^s(\Omega)$ the subspace of $H^s(\mathbb{R}^3)$, $\widetilde{H}^s(\Omega):=\{g:\ g\in H^s(\mathbb{R}^3),\ \mathrm{supp}\, g\subset\overline{\Omega}\}$; similarly, $\widetilde{H}^s(S_1)=\{g\in H^s(S),\ \mathrm{supp}\ g\subset\overline{S}_1\}$ is the Sobolev space of functions having support in $S_1\subset S=\partial\Omega$. We will also use the notation like $\boldsymbol{H}^s(\Omega)=[H^s(\Omega)]^n$ for the $n$-dimensional counterparts of all the aforementioned spaces. Let $\boldsymbol{H}^s_{\mathrm{div}}(\Omega)=\{v\in\boldsymbol{H}^s(\Omega):\mathrm{div}\, v=0\}$ be the divergence-free Sobolev space.

We will also make use of the following spaces (cf., e.g., [Co88] [CMN09])

$$\mathbb{H}^{1,0}(\Omega;\mathscr{A}) := \{(v,p)\in\boldsymbol{H}^1(\Omega)\times L_2(\Omega):\mathscr{A}(v,p)\in\boldsymbol{L}_2(\Omega)\},$$

$$\mathbb{H}^{1,0}_{\mathrm{div}}(\Omega;\mathscr{A}) := \{(v,p)\in\boldsymbol{H}^1_{\mathrm{div}}(\Omega)\times L_2(\Omega):\mathscr{A}(v,p)\in\boldsymbol{L}_2(\Omega)\},$$

endowed with the same norm, $\|(v,p)\|_{\mathbb{H}^{1,0}_{\mathrm{div}}(\Omega;L)}=\|(v,p)\|_{\mathbb{H}^{1,0}(\Omega;L)}$, where

$$\|(v,p)\|_{\mathbb{H}^{1,0}(\Omega;L)} := \left(\|p\|^2_{L_2(\Omega)}+\|v\|^2_{\boldsymbol{H}^1(\Omega)}+\|\mathscr{A}(v,p)\|^2_{\boldsymbol{L}_2(\Omega)}\right)^{\frac{1}{2}}.$$

For sufficiently smooth functions $v$ and $p$ in $\Omega^{\pm}$, we can write the classical traction operators on the boundary $S$ as

$$T_i^{\pm}(v,p)(\boldsymbol{x}) := \gamma^{\pm}\sigma_{ij}(v,p)(\boldsymbol{x})\,n_j(\boldsymbol{x}), \tag{34.1}$$

where $n_j(\boldsymbol{x})$ denote components of the unit outward normal vector $\boldsymbol{n}(\boldsymbol{x})$ to the boundary $S$ of the domain $\Omega$ and $\gamma^{\pm}$ are the trace operators from inside and outside $\Omega$.

Traction operators (34.1) can be continuously extended to the *canonical* traction operators $\boldsymbol{T}^{\pm}: \mathbb{H}^{1,0}(\Omega^{\pm}, \mathscr{A}) \to \boldsymbol{H}^{-\frac{1}{2}}(S)$ defined in the weak form similar to [Co88, Mi11, CMN09] as

$$\langle \boldsymbol{T}^{\pm}(v,p), \boldsymbol{w}\rangle_S := \pm \int_{\Omega^{\pm}} \left[\mathscr{A}(v,p)\gamma^{-1}\boldsymbol{w} + \mathscr{E}\left((v,p), \gamma^{-1}\boldsymbol{w}\right)\right] d\boldsymbol{x},$$
$$\forall\, (v,p) \in \mathbb{H}^{1,0}(\Omega^{\pm}, \mathscr{A}),\ \forall \boldsymbol{w} \in \boldsymbol{H}^{\frac{1}{2}}(S).$$

Here the operator $\gamma^{-1}: \boldsymbol{H}^{\frac{1}{2}}(S) \to \boldsymbol{H}^1(\mathbb{R}^3)$ denotes a continuous right inverse of the trace operator $\gamma: \boldsymbol{H}^1(\mathbb{R}^3) \to \boldsymbol{H}^{\frac{1}{2}}(S)$, and the bilinear form $\mathscr{E}$ is defined as

$$\mathscr{E}\left((v,p),\boldsymbol{u}\right)(\boldsymbol{x}) := \frac{1}{2}\mu(x)\left(\frac{\partial u_i(\boldsymbol{x})}{\partial x_j} + \frac{\partial u_j(\boldsymbol{x})}{\partial x_i}\right)\left(\frac{\partial v_i(\boldsymbol{x})}{\partial x_j} + \frac{\partial v_j(\boldsymbol{x})}{\partial x_i}\right)$$
$$- \frac{2}{3}\mu(x)\operatorname{div} v(\boldsymbol{x}) \operatorname{div}\boldsymbol{u}(\boldsymbol{x}) - p(\boldsymbol{x})\operatorname{div}\boldsymbol{u}(\boldsymbol{x}).$$

Furthermore, if $(v,p) \in \mathbb{H}^{1,0}(\Omega, \mathscr{A})$ and $\boldsymbol{u} \in \boldsymbol{H}^1(\Omega)$, the following first Green identity holds, cf. [Co88, Mi11, CMN09],

$$\langle \boldsymbol{T}^{+}(v,p), \gamma^{+}\boldsymbol{u}\rangle_S = \int_{\Omega} [\mathscr{A}(v,p)\boldsymbol{u} + \mathscr{E}\left((v,p),\boldsymbol{u}\right)(\boldsymbol{x})] d\boldsymbol{x}. \tag{34.2}$$

For $(v,p) \in \mathbb{H}^{1,0}_{\mathrm{div}}(\Omega^{\pm}, \mathscr{A})$ the *canonical* traction operators can be reduced to $\boldsymbol{T}^{\pm}: \mathbb{H}^{1,0}_{\mathrm{div}}(\Omega^{\pm}, \mathscr{A}) \to \boldsymbol{H}^{-\frac{1}{2}}(S)$ defined as

$$\langle \boldsymbol{T}^{\pm}(v,p), \boldsymbol{w}\rangle_S := \pm \int_{\Omega^{\pm}} \left[\mathscr{A}(v,p)\gamma^{-1}_{\mathrm{div}}\boldsymbol{w} + \mathscr{E}(v, \gamma^{-1}_{\mathrm{div}}\boldsymbol{w})\right] d\boldsymbol{x}$$
$$\forall\, (v,p) \in \mathbb{H}^{1,0}_{\mathrm{div}}(\Omega^{\pm}, \mathscr{A}),\ \forall \boldsymbol{w} \in \boldsymbol{H}^{\frac{1}{2}}(S).$$

Here the operator $\gamma^{-1}_{\mathrm{div}}: \boldsymbol{H}^{\frac{1}{2}}(S) \to \boldsymbol{H}^1_{\mathrm{div}}(\mathbb{R}^3)$ denotes a continuous right inverse of the trace operator $\gamma: \boldsymbol{H}^1_{\mathrm{div}}(\mathbb{R}^3) \to \boldsymbol{H}^{\frac{1}{2}}(S)$, and the bilinear form $\mathscr{E}$ reduces to

$$\mathscr{E}(v,\boldsymbol{u})(\boldsymbol{x}) := \frac{\mu(\boldsymbol{x})}{2}\left(\frac{\partial u_i(\boldsymbol{x})}{\partial x_j} + \frac{\partial u_j(\boldsymbol{x})}{\partial x_i}\right)\left(\frac{\partial v_i(\boldsymbol{x})}{\partial x_j} + \frac{\partial v_j(\boldsymbol{x})}{\partial x_i}\right).$$

For $(v,p) \in \mathbb{H}^{1,0}_{\mathrm{div}}(\Omega, \mathscr{A})$ and $\boldsymbol{u} \in \boldsymbol{H}^1_{\mathrm{div}}(\Omega)$, the first Green identity takes the same form (34.2), where $\mathscr{E}\left((v,p),\boldsymbol{u}\right)(\boldsymbol{x})$ reduces to $\mathscr{E}\left(v,\boldsymbol{u}\right)(\boldsymbol{x})$.

Applying the identity (34.2) to the pairs of elements $(v,p) \in \mathbb{H}^{1,0}_{\rm div}(\Omega,\mathscr{A})$ and $(\boldsymbol{u},q) \in \mathbb{H}^{1,0}_{\rm div}(\Omega,\mathscr{A})$ with exchanged roles and subtracting the one from the other, we arrive at the second Green identity, cf. [McL00, Mi11],

$$\int_\Omega [\mathscr{A}_j(v,p)u_j - \mathscr{A}_j(\boldsymbol{u},q)v_j]\,d\boldsymbol{x} = \int_S [T_j(v,p)u_j - T_j(\boldsymbol{u},q)v_j]\,dS. \tag{34.3}$$

Now we are ready to define the mixed boundary value problem for which we aim to derive equivalent boundary-domain integral equation systems (BDIEs) and investigate the existence and uniqueness of their solutions.

*For* $\boldsymbol{f} \in \boldsymbol{L}_2(\Omega)$, $\varphi_0 \in \boldsymbol{H}^{\frac{1}{2}}(S_D)$ *and* $\psi_0 \in \boldsymbol{H}^{-\frac{1}{2}}(S_N)$, *find* $(v,p) \in \mathbb{H}^{1,0}_{\rm div}(\Omega,\mathscr{A})$ *such that:*

$$\mathscr{A}(v,p)(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega, \tag{34.4a}$$

$$r_{S_D}\gamma^+ v(\boldsymbol{x}) = \varphi_0(\boldsymbol{x}), \quad \boldsymbol{x} \in S_D, \tag{34.4b}$$

$$r_{S_N}\boldsymbol{T}^+(v,p)(\boldsymbol{x}) = \psi_0(\boldsymbol{x}), \quad \boldsymbol{x} \in S_N. \tag{34.4c}$$

The following assertion can be easily proved by the Lax-Milgram lemma.

**Theorem 1.** *Mixed boundary value problem* (34.4) *is uniquely solvable.*

## 34.2 Parametrix and Parametrix-Based Hydrodynamic Potentials

When $\mu(\boldsymbol{x}) = 1$, the operator $\mathscr{A}$ becomes the constant-coefficient Stokes operator $\mathring{\mathscr{A}}$, for which we know an explicit fundamental solution defined by the pair of distributions $(\mathring{\boldsymbol{u}}^k, \mathring{q}^k)$ where $\mathring{u}^k_j$ represent the components of the incompressible velocity fundamental solution and $\mathring{q}^k$ represent the components of the pressure fundamental solution (see, e.g., [La69, KoWe06], [HsWe08]).

$$\mathring{u}^k_j(\boldsymbol{x},\boldsymbol{y}) = -\frac{1}{8\pi}\left\{\frac{\delta^k_j}{|\boldsymbol{x}-\boldsymbol{y}|} + \frac{(x_j-y_j)(x_k-y_k)}{|\boldsymbol{x}-\boldsymbol{y}|^3}\right\},$$

$$\mathring{q}^k(\boldsymbol{x},\boldsymbol{y}) = \frac{x_k-y_k}{4\pi|\boldsymbol{x}-\boldsymbol{y}|^3}, \qquad j,k \in \{1,2,3\}.$$

Therefore $(\mathring{\boldsymbol{u}}^k, \mathring{q}^k)$ satisfy

$$\mathring{\mathscr{A}}_j(\mathring{\boldsymbol{u}}^k, \mathring{q}^k)(\boldsymbol{x}) = \sum_{i=1}^3 \frac{\partial^2 \mathring{u}^k_j}{\partial x_i^2} - \frac{\partial \mathring{q}^k}{\partial x_j} = \delta^k_j \delta(\boldsymbol{x}-\boldsymbol{y})$$

Let us denote $\mathring{\sigma}_{ij}(v,p) := \sigma_{ij}(v,p)|_{\mu=1}$. Then in the particular case, for $\mu = 1$ and the fundamental solution $(\mathring{\boldsymbol{u}}^k, \mathring{q}^k)_{k=1,2,3}$ of the operator $\mathring{\mathscr{A}}$, the stress tensor $\mathring{\sigma}_{ij}(\mathring{\boldsymbol{u}}^k, \mathring{q}^k)(\boldsymbol{x}-\boldsymbol{y})$ reads

$$\mathring{\sigma}_{ij}(\mathring{\boldsymbol{u}}^k, \mathring{q}^k)(\boldsymbol{x}-\boldsymbol{y}) = \frac{3}{4\pi}\frac{(x_i-y_i)(x_j-y_j)(x_k-y_k)}{|\boldsymbol{x}-\boldsymbol{y}|^5},$$

and the boundary traction becomes

$$\begin{aligned}\mathring{T}_i(\boldsymbol{x};\mathring{\boldsymbol{u}}^k, \mathring{q}^k)(\boldsymbol{x},\boldsymbol{y}) :&= \mathring{\sigma}_{ij}(\mathring{\boldsymbol{u}}^k, \mathring{q}^k)(\boldsymbol{x}-\boldsymbol{y})\, n_j(\boldsymbol{x}) \\ &= \frac{3}{4\pi}\frac{(x_i-y_i)(x_j-y_j)(x_k-y_k)}{|\boldsymbol{x}-\boldsymbol{y}|^5} n_j(\boldsymbol{x}).\end{aligned}$$

Let us define a pair of functions $(\boldsymbol{u}^k, q^k)_{k=1,2,3}$ as

$$u_j^k(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{\mu(\boldsymbol{y})}\mathring{u}_j^k(\boldsymbol{x},\boldsymbol{y}) = -\frac{1}{8\pi\mu(\boldsymbol{y})}\left\{\frac{\delta_j^k}{|\boldsymbol{x}-\boldsymbol{y}|} + \frac{(x_j-y_j)(x_k-y_k)}{|\boldsymbol{x}-\boldsymbol{y}|^3}\right\}, \tag{34.5}$$

$$q^k(\boldsymbol{x},\boldsymbol{y}) = \frac{\mu(\boldsymbol{x})}{\mu(\boldsymbol{y})}\mathring{q}^k(\boldsymbol{x},\boldsymbol{y}) = \frac{\mu(\boldsymbol{x})}{\mu(\boldsymbol{y})}\frac{x_k-y_k}{4\pi|\boldsymbol{x}-\boldsymbol{y}|^3}, \quad j,k \in \{1,2,3\}. \tag{34.6}$$

Then

$$\sigma_{ij}(\boldsymbol{x};\boldsymbol{u}^k, q^k)(\boldsymbol{x},\boldsymbol{y}) = \frac{\mu(\boldsymbol{x})}{\mu(\boldsymbol{y})}\mathring{\sigma}_{ij}(\mathring{\boldsymbol{u}}^k, \mathring{q}^k)(\boldsymbol{x}-\boldsymbol{y}),$$

$$T_i(\boldsymbol{x};\boldsymbol{u}^k, q^k)(\boldsymbol{x},\boldsymbol{y}) := \sigma_{ij}(\boldsymbol{x};\boldsymbol{u}^k, q^k)(\boldsymbol{x},\boldsymbol{y})\, n_j(\boldsymbol{x}) = \frac{\mu(\boldsymbol{x})}{\mu(\boldsymbol{y})}\mathring{T}_i(\boldsymbol{x};\mathring{\boldsymbol{u}}^k, \mathring{q}^k)(\boldsymbol{x},\boldsymbol{y}).$$

Substituting (34.5)-(34.6) in the Stokes system gives

$$\mathscr{A}_j(\boldsymbol{x};\boldsymbol{u}^k, q^k)(\boldsymbol{x},\boldsymbol{y}) = \delta_j^k\delta(\boldsymbol{x}-\boldsymbol{y}) + R_{kj}(\boldsymbol{x},\boldsymbol{y}), \tag{34.7}$$

where

$$R_{kj}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{\mu(\boldsymbol{y})}\frac{\partial\mu(\boldsymbol{x})}{\partial x_i}\mathring{\sigma}_{ij}(\mathring{\boldsymbol{u}}^k, \mathring{q}^k)(\boldsymbol{x}-\boldsymbol{y}) = \mathscr{O}(|\boldsymbol{x}-\boldsymbol{y}|)^{-2})$$

is a weakly singular remainder. This implies that $(\boldsymbol{u}^k, q^k)$ is a parametrix of the operator $\mathscr{A}$.

Let us define the parametrix-based Newton-type and remainder vector potentials

$$\mathscr{U}_k\rho(\boldsymbol{y}) = \mathscr{U}_{kj}\rho_j(\boldsymbol{y}) := \int_\Omega u_j^k(\boldsymbol{x},\boldsymbol{y})\rho_j(\boldsymbol{x})\,d\boldsymbol{x},$$

$$\mathscr{R}_k\rho(\boldsymbol{y}) = \mathscr{R}_{kj}\rho_j(\boldsymbol{y}) := \int_\Omega R_{kj}(\boldsymbol{x},\boldsymbol{y})\rho_j(\boldsymbol{x})\,d\boldsymbol{x}, \quad \boldsymbol{x} \in \mathbb{R}^3,$$

for the velocity, and the scalar Newton-type and remainder potentials

$$\mathscr{Q}\rho(\boldsymbol{y}) = \mathscr{Q}_j\rho_j(\boldsymbol{y}) := \int_\Omega \mathring{q}^j(\boldsymbol{x},\boldsymbol{y})\rho_j(\boldsymbol{x})d\boldsymbol{x}, \tag{34.8}$$

$$\mathscr{R}^\bullet\rho(\boldsymbol{y}) = \mathscr{R}^\bullet_j\rho_j(\boldsymbol{y}) := 2\int_\Omega \frac{\partial \mathring{q}^j(\boldsymbol{x},\boldsymbol{y})}{\partial x_i}\frac{\partial \mu(\boldsymbol{x})}{\partial x_i}\rho_j(\boldsymbol{x})d\boldsymbol{x}, \quad \boldsymbol{x}\in\mathbb{R}^3, \tag{34.9}$$

for the pressure. The integral in (34.9) is understood as a 3D strongly singular integral in the Cauchy sense.

For the velocity, let us also define the parametrix-based single-layer potential, double-layer potential and their respective direct values on the boundary, as follows:

$$V_k\rho(\boldsymbol{y}) = V_{kj}\rho_j(\boldsymbol{y}) := -\int_S u^k_j(\boldsymbol{x},\boldsymbol{y})\rho_j(\boldsymbol{x})\,dS_x, \quad \boldsymbol{y}\notin S,$$

$$W_k\rho(\boldsymbol{y}) = W_{kj}\rho_j(\boldsymbol{y}) := -\int_S T_j(\boldsymbol{x};\boldsymbol{u}^k,q^k)(\boldsymbol{x},\boldsymbol{y})\rho_j(\boldsymbol{x})\,dS_x, \quad \boldsymbol{y}\notin S,$$

$$\mathscr{V}_k\rho(\boldsymbol{y}) = \mathscr{V}_{kj}\rho_j(\boldsymbol{y}) := -\int_S u^k_j(\boldsymbol{x},\boldsymbol{y})\rho_j(\boldsymbol{x})\,dS_x, \quad \boldsymbol{y}\in S,$$

$$\mathscr{W}_k\rho(\boldsymbol{y}) = \mathscr{W}_{kj}\rho_j(\boldsymbol{y}) := -\int_S T_j(\boldsymbol{x};\boldsymbol{u}^k,q^k)(\boldsymbol{x},\boldsymbol{y})\rho_j(\boldsymbol{x})\,dS_x, \quad \boldsymbol{y}\in S.$$

Let us also denote

$$\mathscr{W}'_k\rho(\boldsymbol{y}) = \mathscr{W}'_{kj}\rho_j(\boldsymbol{y}) := -\int_S T_j(\boldsymbol{y};\boldsymbol{u}^k,\mathring{q}^k)(\boldsymbol{x},\boldsymbol{y})\rho_j(\boldsymbol{x})\,dS_x, \quad \boldsymbol{y}\in S.$$

For pressure in the variable coefficient Stokes system, we will need the following single-layer and double-layer potentials:

$$\mathscr{P}\rho(\boldsymbol{y}) = \mathscr{P}_j\rho_j(\boldsymbol{y}) := -\int_S \mathring{q}^j(\boldsymbol{x},\boldsymbol{y})\rho_j(\boldsymbol{x})\,dS_x,$$

$$\Pi\rho(\boldsymbol{y}) = \Pi_j\rho_j(\boldsymbol{y}) := -2\int_S \frac{\partial \mathring{q}^j(\boldsymbol{x},\boldsymbol{y})}{\partial n(\boldsymbol{x})}\mu(\boldsymbol{x})\rho_j(\boldsymbol{x})\,dS_x, \quad \boldsymbol{y}\notin S.$$

The parametrix-based integral operators, depending on the variable coefficient $\mu(\boldsymbol{x})$, can be expressed in terms of the corresponding integral operators for the constant coefficient case, $\mu = 1$,

$$\mathscr{U}_k\rho(\boldsymbol{y}) = \frac{1}{\mu(\boldsymbol{y})}\mathring{\mathscr{U}}_k\rho(\boldsymbol{y}), \tag{34.10}$$

$$\mathscr{R}_k\rho(\boldsymbol{y}) = \frac{-1}{\mu(\boldsymbol{y})}\Big[2\frac{\partial}{\partial y_j}\mathring{\mathscr{U}}_{ki}(\rho_j\partial_i\mu)(\boldsymbol{y}) + 2\frac{\partial}{\partial y_i}\mathring{\mathscr{U}}_{kj}(\rho_j\partial_i\mu)(\boldsymbol{y}) + \mathring{\mathscr{Q}}_k(\rho_j\partial_j\mu)(\boldsymbol{y})\Big], \tag{34.11}$$

$$\mathscr{Q}_j\rho_j(\boldsymbol{y}) = \mathring{\mathscr{Q}}_j\rho_j(\boldsymbol{y}), \qquad \mathscr{R}_j^{\bullet}\rho_j(\boldsymbol{y}) = -2\frac{\partial}{\partial y_i}\mathring{\mathscr{Q}}_j(\rho_j\partial_i\mu)(\boldsymbol{y}), \tag{34.12}$$

$$V_k\rho(\boldsymbol{y}) = \frac{1}{\mu(\boldsymbol{y})}\mathring{V}_k\rho(\boldsymbol{y}), \qquad W_k\rho(\boldsymbol{y}) = \frac{1}{\mu(\boldsymbol{y})}\mathring{W}_k(\mu\rho)(\boldsymbol{y}), \tag{34.13}$$

$$\mathscr{V}_k\rho(\boldsymbol{y}) = \frac{1}{\mu(\boldsymbol{y})}\mathring{\mathscr{V}}_k\rho(\boldsymbol{y}), \qquad \mathscr{W}_k\rho(\boldsymbol{y}) = \frac{1}{\mu(\boldsymbol{y})}\mathring{\mathscr{W}}_k(\mu\rho)(\boldsymbol{y}), \tag{34.14}$$

$$\mathscr{P}_j\rho_j(\boldsymbol{y}) = \mathring{\mathscr{P}}_j\rho_j(\boldsymbol{y}), \qquad \Pi_j\rho_j(\boldsymbol{y}) = \mathring{\Pi}_j(\mu\rho_j)(\boldsymbol{y}), \tag{34.15}$$

$$\mathscr{W}_k'\rho = \mathring{\mathscr{W}}_k'\rho - \left(\frac{\partial_i\mu}{\mu}\mathring{\mathscr{V}}_k\rho + \frac{\partial_k\mu}{\mu}\mathring{\mathscr{V}}_i\rho - \frac{2}{3}\delta_i^k\frac{\partial_j\mu}{\mu}\mathring{\mathscr{V}}_j\rho\right)n_i. \tag{34.16}$$

Note that the velocity potentials defined above are *not incompressible for the variable coefficient* $\mu(y)$.

The following assertions of this section are well known for the constant coefficient case, see, e.g., [KoWe06, HsWe08]. Then by relations (34.10)-(34.16) we obtain their counterparts for the variable-coefficient case.

**Theorem 2.** *The following operators are continuous.*

$$\mathscr{U}_{ik} : \widetilde{H}^s(\Omega) \to H^{s+2}(\Omega), \qquad s \in \mathbb{R}, \tag{34.17}$$

$$\mathscr{U}_{ik} : H^s(\Omega) \to H^{s+2}(\Omega), \qquad s > -\frac{1}{2}, \tag{34.18}$$

$$\mathscr{R}_{ik} : \widetilde{H}^s(\Omega) \to H^{s+1}(\Omega), \qquad s \in \mathbb{R}, \tag{34.19}$$

$$\mathscr{R}_{ik} : H^s(\Omega) \to H^{s+1}(\Omega), \qquad s > -\frac{1}{2}, \tag{34.20}$$

$$\mathscr{P}_k : H^{s-\frac{3}{2}}(S) \to H^{s-1}(\Omega), \qquad s \in \mathbb{R}, \tag{34.21}$$

$$\Pi_k : H^{s-\frac{1}{2}}(S) \to H^{s-1}(\Omega), \qquad s \in \mathbb{R}, \tag{34.22}$$

$$\mathscr{Q}_k : \widetilde{H}^{s-2}(\Omega) \to H^{s-1}(\Omega), \qquad s \in \mathbb{R}, \tag{34.23}$$

$$\mathscr{R}_k^{\bullet} : H^s(\Omega) \to H^s(\Omega), \qquad s > -\frac{1}{2}. \tag{34.24}$$

Let us also denote

$$\mathscr{L}_k^{\pm}\rho(\boldsymbol{y}) := T_k^{\pm}(\boldsymbol{W}\rho, \Pi\rho)(\boldsymbol{y}), \quad \boldsymbol{y} \in S,$$

where $T_k^{\pm}$ are the traction operators for the *compressible* fluid.

**Theorem 3.** *Let $s \in \mathbb{R}$. Let $S_1$ and $S_2$ be two non empty manifolds on $S$ with smooth boundary $\partial S_1$ and $\partial S_2$, respectively. Then the following operators are continuous:*

$$V_{ik} : H^s(S) \to H^{s+\frac{3}{2}}(\Omega), \qquad W_{ik} : H^s(S) \to H^{s+\frac{1}{2}}(\Omega),$$
$$\mathscr{V}_{ik} : H^s(S) \to H^{s+1}(S), \qquad \mathscr{W}_{ik} : H^s(S) \to H^{s+1}(S),$$
$$r_{S_2}\mathscr{V}_{ik} : \widetilde{H}^s(S_1) \to H^{s+1}(S_2), \qquad r_{S_2}\mathscr{W}_{ik} : \widetilde{H}^s(S_1) \to H^{s+1}(S_2),$$
$$\mathscr{L}^{\pm}_{ik} : H^s(S) \to H^{s-1}(S), \qquad \mathscr{W}'_{ik} : H^s(S) \to H^{s+1}(S).$$

**Theorem 4.** *If $\tau \in \boldsymbol{H}^{1/2}(S)$, $\rho \in \boldsymbol{H}^{-1/2}(S)$, then the following jump relations hold,*

$$\gamma^{\pm} V_k \rho = \mathscr{V}_k \rho, \qquad \gamma^{\pm} W_k \tau = \mp \frac{1}{2}\tau_k + \mathscr{W}_k \tau$$
$$T^{\pm}_k(\boldsymbol{V}\rho, \mathscr{P}\rho) = \pm\frac{1}{2}\rho_k + \mathscr{W}'_k \rho,$$
$$(\mathscr{L}^{\pm}_k - \hat{\mathscr{L}}_k)\tau = -\gamma^{\pm}\left[(\partial_i \mu)\, W_k(\tau) + (\partial_k \mu)\, W_i(\tau) - \frac{2}{3}\delta^k_i (\partial_j \mu)\, W_j \tau\right] n_i,$$
$$\hat{\mathscr{L}}_k(\tau) = \mathring{\mathscr{L}}_k(\mu\tau).$$

**Proposition 1.** *The following operators are compact,*

$$\mathscr{R}_{ik} : H^s(\Omega) \to H^s(\Omega), \qquad \mathscr{R}^{\bullet}_k : H^s(\Omega) \to H^{s-1}(\Omega), \quad s \in \mathbb{R},$$
$$\gamma^{+}\mathscr{R}_{ik} : H^s(\Omega) \to H^{s-\frac{1}{2}}(S), \qquad T^{\pm}_{ik}(\mathscr{R}, \mathscr{R}^{\bullet}) : H^s(\Omega) \to H^{s-\frac{3}{2}}(S), \quad s > \frac{1}{2}.$$

**Proposition 2.** *Let $s \in \mathbb{R}$ and $S_1$ be a nonempty submanifold of $S$ with smooth boundary. Then the following operators are compact:*

$$(\mathscr{L}^{\pm}_{ik} - \hat{\mathscr{L}}_{ik}) : \widetilde{H}^s(S_1) \to H^{s-1}(S).$$

## 34.3 The Third Green Identities

Let $B(\boldsymbol{y}, \varepsilon) \subset \Omega$ be a ball of a radius $\varepsilon$ around a point $\boldsymbol{y} \in \Omega$. Applying the second Green identity (34.3) in the domain $\Omega \setminus B(\boldsymbol{y}, \varepsilon)$ to any $(v, p) \in \mathbb{H}^{1,0}_{\mathrm{div}}(\Omega; \mathscr{A})$ and to the fundamental solution $(\boldsymbol{u}^k, q^k)$ and taking the limit as $\varepsilon \to 0$, we obtain the following third Green identity

$$v + \mathscr{R}v - \boldsymbol{V}\boldsymbol{T}^{+}(v,p) + \boldsymbol{W}\gamma^{+}v = \mathscr{U}\mathscr{A}(v,p) \quad \text{in } \Omega. \tag{34.25}$$

Similarly, applying the first Green identity (34.2) in the domain $\Omega \setminus B(\boldsymbol{y}, \varepsilon)$ to any $(v, p) \in \mathbb{H}^{1,0}_{\mathrm{div}}(\Omega; \mathscr{A})$ and to the pressure part of the constant-coefficient fundamental solution $\mathring{q}^k$, for $u_k$, and taking the limit as $\varepsilon \to 0$, we obtain the following parametrix-based third Green identity for pressure,

$$p+\mathscr{R}^{\bullet}v-\mathscr{P}\boldsymbol{T}^{+}(v,p)+\Pi\gamma^{+}v=\mathscr{Q}\mathscr{A}(v,p)\quad\text{in }\Omega. \tag{34.26}$$

If the couple $(v,p)\in\mathbb{H}^{1,0}_{\mathrm{div}}(\Omega;\mathscr{A})$ is a solution of the Stokes PDE (34.4a) with variable coefficient, then (34.25) and (34.26) give

$$v+\mathscr{R}v-\boldsymbol{V}\boldsymbol{T}^{+}(v,p)+\boldsymbol{W}\gamma^{+}v=\mathscr{U}\boldsymbol{f}, \tag{34.27}$$

$$p+\mathscr{R}^{\bullet}v-\mathscr{P}\boldsymbol{T}^{+}(v,p)+\Pi\gamma^{+}v=\mathscr{Q}\boldsymbol{f}\quad\text{in }\Omega. \tag{34.28}$$

We will also need the trace and traction of the third Green identities for $(v,p)\in\mathbb{H}^{1,0}_{\mathrm{div}}(\Omega;\mathscr{A})$ on $S$:

$$\frac{1}{2}\gamma^{+}v+\mathscr{R}^{+}v-\mathscr{V}\boldsymbol{T}^{+}(v,p)+\mathscr{W}\gamma^{+}v=\gamma^{+}\mathscr{U}\boldsymbol{f}, \tag{34.29}$$

$$\frac{1}{2}\boldsymbol{T}^{+}(v,p)+\boldsymbol{T}^{+}(\mathscr{R},\mathscr{R}^{\bullet})v-\mathscr{W}'\boldsymbol{T}^{+}(v,p)+\mathscr{L}^{+}\gamma^{+}v=\boldsymbol{T}^{+}(\mathscr{U},\mathscr{Q})\boldsymbol{f}. \tag{34.30}$$

One can prove the following two assertions that are instrumental for proof of equivalence of the BDIEs and the mixed PDE.

**Lemma 1.** *Let $v\in\boldsymbol{H}^{1}_{\mathrm{div}}(\Omega)$, $p\in L_2(\Omega)$, $\boldsymbol{f}\in\boldsymbol{L}_2(\Omega)$, $\Psi\in\boldsymbol{H}^{-\frac{1}{2}}(S)$ and $\Phi\in\boldsymbol{H}^{\frac{1}{2}}(S)$ satisfy the equations*

$$p+\mathscr{R}^{\bullet}v-\mathscr{P}\Psi+\Pi\Phi=\mathscr{Q}\boldsymbol{f}\quad\text{in }\Omega,$$
$$v+\mathscr{R}v-\boldsymbol{V}\Psi+\boldsymbol{W}\Phi=\mathscr{U}\boldsymbol{f}\quad\text{in }\Omega.$$

*Then $(v,p)\in\mathbb{H}^{1,0}_{\mathrm{div}}(\Omega,\mathscr{A})$ and solve the equation $\mathscr{A}(\boldsymbol{y};v,p)=\boldsymbol{f}$. Moreover, the following relations hold true:*

$$\boldsymbol{V}(\Psi-\boldsymbol{T}^{+}(v,p))(\boldsymbol{y})-\boldsymbol{W}(\Phi-\gamma^{+}v)(\boldsymbol{y})=0,\qquad\boldsymbol{y}\in\Omega,$$

$$\mathscr{P}(\Psi-\boldsymbol{T}^{+}(v,p))(\boldsymbol{y})-\Pi(\Phi-\gamma^{+}v)(\boldsymbol{y})=0,\qquad\boldsymbol{y}\in\Omega.$$

**Lemma 2.** *Let $S=\overline{S}_1\cup\overline{S}_2$, where $S_1$ and $S_2$ are open nonempty non-intersecting simply connected submanifolds of $S$ with infinitely smooth boundaries. Let $\Psi^{*}\in\widetilde{\boldsymbol{H}}^{-\frac{1}{2}}(S_1)$, $\Phi^{*}\in\widetilde{\boldsymbol{H}}^{\frac{1}{2}}(S_2)$. If*

$$\boldsymbol{V}\Psi^{*}(\boldsymbol{x})-\boldsymbol{W}\Phi^{*}(\boldsymbol{x})=0\qquad\mathscr{P}(\Psi^{*})-\Pi(\Phi^{*})=0\quad\text{in }\Omega,$$

*then $\Psi^{*}=0$ and $\Phi^{*}=0$ on $S$.*

## 34.4 Boundary–Domain Integral Equation System for the Mixed Problem

We aim to obtain a segregated boundary-domain integral equation system for mixed BVP (34.4). To this end, let the functions $\Phi_0 \in \boldsymbol{H}^{\frac{1}{2}}(S)$ and $\Psi_0 \in \boldsymbol{H}^{-\frac{1}{2}}(S)$ be respective continuations of the boundary functions $\varphi_0 \in \boldsymbol{H}^{\frac{1}{2}}(S_D)$ and $\psi_0 \in \boldsymbol{H}^{-\frac{1}{2}}(S_N)$ from (34.4b) and (34.4c). Let us now represent

$$\gamma^+ v = \Phi_0 + \varphi, \quad \boldsymbol{T}^+(v,p) = \Psi_0 + \psi \text{ on } S, \tag{34.31}$$

where $\varphi \in \widetilde{\boldsymbol{H}}^{\frac{1}{2}}(S_N)$ and $\psi \in \widetilde{\boldsymbol{H}}^{-\frac{1}{2}}(S_D)$ are unknown boundary functions.

Let us now take equations (34.27) and (34.28) in the domain $\Omega$ and restrictions of equations (34.29) and (34.30) to the boundary parts $S_D$ and $S_N$, respectively. Substituting there representations (34.31) and considering further the unknown boundary functions $\varphi$ and $\psi$ as formally independent of (segregated from) the unknown domain functions $v$ and $p$, we obtain the following system of four boundary-domain integral equations for four unknowns, $(v,p) \in \mathbb{H}^{1,0}_{\mathrm{div}}(\Omega, \mathscr{A})$, $\varphi \in \widetilde{\boldsymbol{H}}^{\frac{1}{2}}(S_N)$ and $\psi \in \widetilde{\boldsymbol{H}}^{-\frac{1}{2}}(S_D)$:

$$p + \mathscr{R}^{\bullet} v - \mathscr{P}\psi + \Pi\varphi = F_0 \quad \text{in } \Omega, \tag{34.32a}$$

$$v + \mathscr{R} v - \boldsymbol{V}\psi + \boldsymbol{W}\varphi = \boldsymbol{F} \quad \text{in } \Omega, \tag{34.32b}$$

$$r_{S_D}\gamma^+ \mathscr{R} v - r_{S_D}\mathscr{V}\psi + r_{S_D}\mathscr{W}\varphi = r_{S_D}\gamma^+ \boldsymbol{F} - \varphi_0 \quad \text{on } S_D, \tag{34.32c}$$

$$r_{S_N}\boldsymbol{T}^+(\mathscr{R}, \mathscr{R}^{\bullet}) v - r_{S_N}\mathscr{W}'\psi + r_{S_N}\mathscr{L}^+\varphi = r_{S_N}\boldsymbol{T}^+(\boldsymbol{F}, F_0) - \psi_0 \quad \text{on } S_N, \tag{34.32d}$$

where

$$F_0 = \mathscr{Q}\boldsymbol{f} + \mathscr{P}\Psi_0 - \Pi\Phi_0, \quad \boldsymbol{F} = \mathscr{U}\boldsymbol{f} + \boldsymbol{V}\Psi_0 - \boldsymbol{W}\Phi_0. \tag{34.33}$$

Applying Lemma 1 to (34.33) and taking into account the continuity of operators (34.20) and (34.24), one can prove that $(F_0, \boldsymbol{F}) \in \mathbb{H}^{1,0}(\Omega, \mathscr{A})$.

We denote the right-hand side of BDIE system (34.32) as

$$\mathscr{F}^{11} := [F_0, \boldsymbol{F}, r_{S_D}\gamma^+ \boldsymbol{F} - \varphi_0, r_{S_N}\boldsymbol{T}^+_{\boldsymbol{F},F} - \psi_0]^{\top}, \tag{34.34}$$

which implies $\mathscr{F}^{11} \in \mathbb{H}^{1,0}(\Omega, \mathscr{A}) \times \boldsymbol{H}^{\frac{1}{2}}(S_D) \times \boldsymbol{H}^{-\frac{1}{2}}(S_N)$.

Note that BDIE system (34.32) can be split into the BDIE system of 3 vector equations (34.32b), (34.32c), (34.32d) for 3 vector unknowns, $v$, $\psi$ and $\varphi$, and the separate equation (34.32a) that can be used, after solving the system, to obtain the pressure, $p$. However since the couple $(v,p)$ shares the space $\mathbb{H}^{1,0}_{\mathrm{div}}(\Omega, \mathscr{A})$, equations (34.32b), (34.32c), (34.32d) are not completely separate from equation (34.32a).

**Theorem 5 (Equivalence Theorem).** *Let $\boldsymbol{f} \in \boldsymbol{L}_2(\Omega)$ and let $\Phi_0 \in \boldsymbol{H}^{-\frac{1}{2}}(S)$ and $\Psi_0 \in \boldsymbol{H}^{-\frac{1}{2}}(S)$ be some fixed extensions of $\varphi_0 \in \boldsymbol{H}^{\frac{1}{2}}(S_D)$ and $\psi_0 \in \boldsymbol{H}^{-\frac{1}{2}}(S_N)$, respectively.*

(i) *If $(v,p) \in \mathbb{H}^{1,0}_{\mathrm{div}}(\Omega;\mathscr{A})$ solve (34.4), then*

$$(p,v,\psi,\varphi) \in \mathbb{H}^{1,0}_{\mathrm{div}}(\Omega;\mathscr{A}) \times \widetilde{\boldsymbol{H}}^{-\frac{1}{2}}(S_D) \times \widetilde{\boldsymbol{H}}^{\frac{1}{2}}(S_N),$$

*where*

$$\varphi = \gamma^+ v - \Phi_0, \quad \psi = \boldsymbol{T}^+(v,p) - \Psi_0 \quad \text{on } S, \tag{34.35}$$

*solve BDIE system (34.32).*

(ii) *If $(p,v,\psi,\varphi) \in \mathbb{H}^{1,0}_{\mathrm{div}}(\Omega;\mathscr{A}) \times \widetilde{\boldsymbol{H}}^{-\frac{1}{2}}(S_D) \times \widetilde{\boldsymbol{H}}^{\frac{1}{2}}(S_N)$ solve the BDIE system (34.32), then $(v,p)$ solve mixed BVP (34.4) and the functions $\psi, \varphi$ satisfy (34.35).*

(iii) *System (34.32) is uniquely solvable in $\mathbb{H}^{1,0}_{\mathrm{div}}(\Omega;\mathscr{A}) \times \widetilde{\boldsymbol{H}}^{-\frac{1}{2}}(S_D) \times \widetilde{\boldsymbol{H}}^{\frac{1}{2}}(S_N)$.*

*Proof.* (i) Let $(v,p) \in \mathbb{H}^{1,0}_{\mathrm{div}}(\Omega;\mathscr{A})$ be a solution of the BVP. Let us define the functions $\varphi$ and $\psi$ by (34.35). By the BVP boundary conditions, $\gamma^+ v = \varphi_0 = \Phi_0$ on $S_D$ and $\boldsymbol{T}^+(v,p) = \psi_0 = \Psi_0$ on $S_N$. This implies that $(\psi,\varphi) \in \widetilde{\boldsymbol{H}}^{-\frac{1}{2}}(S_D) \times \widetilde{\boldsymbol{H}}^{\frac{1}{2}}(S_N)$. Taking into account the Green identities (34.26)-(34.30), we immediately obtain that $(p,v,\varphi,\psi)$ solve system (34.32).

(ii) Conversely, let $(p,v,\psi,\varphi) \in \mathbb{H}^{1,0}_{\mathrm{div}}(\Omega;\mathscr{A}) \times \widetilde{\boldsymbol{H}}^{-\frac{1}{2}}(S_D) \times \widetilde{\boldsymbol{H}}^{\frac{1}{2}}(S_N)$ solve BDIE system (34.32). If we take the trace of (34.32b) restricted to $S_D$, use the jump relations for the trace of $W$, see Theorem 5, and subtract it from (34.32c), we arrive at $r_{S_D}\gamma^+ v - \frac{1}{2} r_{S_D}\varphi = \varphi_0$ on $S_D$. As $\varphi$ vanishes on $S_D$, therefore the Dirichlet condition of the BVP is satisfied.

Repeating the same procedure but taking the traction of (34.32a) and (34.32b), restricted to $S_N$, using the jump relations for the traction of $V$ and subtracting it from (34.32d), we arrive at $r_{S_N}\boldsymbol{T}(v,p) - \frac{1}{2} r_{S_N}\psi = \psi_0$ on $S_N$. As $\psi$ vanishes on $S_N$, therefore the Neumann condition of the BVP is satisfied. Since $\varphi_0 = \Phi_0$ on $S_D$ and $\psi_0 = \Psi_0$ on $S_N$, the conditions (34.35) are satisfied, respectively, on $S_D$ and $S_N$.

Also we have that $\Psi \in \boldsymbol{H}^{-\frac{1}{2}}$ and $\Phi \in \boldsymbol{H}^{-\frac{1}{2}}$. We note that if $(v,p) \in L_2(\Omega) \times \boldsymbol{H}^1_{\mathrm{div}}(\Omega)$ then $\mathscr{A}(v,p) = \boldsymbol{f} \in \boldsymbol{L}_2(\Omega)$. Due to relations (34.32a) and (34.32b) the hypotheses of the Lemma 1 are satisfied with $\Psi = \psi + \Psi_0$ and $\Phi = \varphi + \Phi_0$ . As a result we obtain that $(v,p)$ is a solution of $\mathscr{A}(v,p) = \boldsymbol{f}$ satisfying

$$\boldsymbol{V}(\Psi^*) - \boldsymbol{W}(\Phi^*) = 0, \quad \mathscr{P}(\Psi^*) - \Pi(\Phi^*) = 0 \quad \text{in } \Omega, \tag{34.36}$$

where

$$\Psi^* = \psi + \Psi_0 - \boldsymbol{T}^+(v,p) \qquad\qquad \Phi^* = \varphi + \Phi_0 - \gamma^+ v$$

Since $\Psi^* \in \widetilde{\boldsymbol{H}}^{-\frac{1}{2}}(S_D)$ and $\Phi^* \in \widetilde{\boldsymbol{H}}^{\frac{1}{2}}(S_N)$, and (34.36) hold true, applying Lemma 2 for $S_1 = S_D$ and $S_2 = S_N$ we obtain $\Psi^* = \Phi^* = 0$ on $S$. This implies conditions (34.35).

(iii) The uniqueness of the BDIEs (34.32) follows from the uniqueness of the BVP, see Theorem 1, and items (i) and (ii). □

## References


[CMN09] Chkadua, O., Mikhailov, S.E. and Natroshvili, D.: Analysis of direct boundary-domain integral equations for a mixed BVP with variable coefficient, I: Equivalence and invertibility. *J. Integral Equations and Appl.* **21**, 499–543 (2009).

[Co88] Costabel, M.: Boundary integral operators on Lipschitz domains: Elementary results. *SIAM J. Math. Anal.* **19**, 613–626 (1988).

[HsWe08] Hsiao, G.C. and Wendland, W.L.: *Boundary Integral Equations.* Springer, Berlin (2008).

[KoWe06] Kohr, M. and Wendland, W.L.: Variationa boundary integral equations for the Stokes system. *Applicable Anal.* **85**, 1343–1372 (2006).

[La69] Ladyzhenskaya, O.A.: *The Mathematical Theory of Viscous Incompressible Flow.* Gordon & Breach, New York (1969).

[LiMa73] Lions, J.L. and Magenes, E.: *Non-Homogeneous Boundary Value Problems and Applications.* Springer (1973).

[McL00] McLean, W.: *Strongly Elliptic Systems and Boundary Integral Equations.* Cambridge University Press (2000).

[Mi11] Mikhailov, S.E.: Traces, extensions and co-normal derivatives for elliptic systems on Lipschitz domains. *J. Math. Anal. and Appl.*, **378**, 324–342 (2011).

# Chapter 35
# Calderón–Zygmund Theory for Second-Order Elliptic Systems on Riemannian Manifolds

**D. Mitrea, I. Mitrea, M. Mitrea, and B. Schmutzler**

## 35.1 Background Assumptions and Basic Definitions

The aim of this chapter is to develop a Calderón–Zygmund theory for the layer potential operators naturally associated with second-order elliptic systems on Riemannian manifolds, which is effective in the treatment of boundary value problems in rough settings. Our main results are described in §35.2, while in §35.3 we illustrate the scope of this theory by presenting a number of concrete examples, of independent interest. We begin by introducing notation and making a series of basic assumptions of analytic and geometric nature.

Throughout the paper, we let $\mathscr{M}$ denote a compact, oriented, boundaryless Riemannian manifold of class $\mathscr{C}^2$ and real dimension $n \in \mathbb{N}$, $n \geq 2$. Also, we let $H^{s,p}$ stands for the $L^p$-based Sobolev space of (smoothness) order $s \in \mathbb{R}$, and denote by $H^{s,p}_{\rm loc}$ the local version of this scale.

**Hypothesis 1 (Analytic Assumptions).** *Consider a second-order differential operator $L : \mathscr{E} \to \mathscr{E}$ acting between sections of a given $\mathscr{C}^2$ Hermitian vector bundle $\mathscr{E} \to \mathscr{M}$, satisfying the following properties:*

D. Mitrea (✉) • M. Mitrea • B. Schmutzler
University of Missouri, Columbia, MO, USA
e-mail: mitread@missouri.edu; mitream@missouri.edu; brock.schmutzler@mail.missouri.edu

I. Mitrea
Temple University, Philadelphia, PA, USA
e-mail: imitrea@temple.edu



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_35

(i) *One has the quasi-factorization*

$$L = \widetilde{D}D + Q, \tag{35.1}$$

*where $\widetilde{D}$, $D$ are first-order differential operators*

$$D : \mathscr{E} \longrightarrow \mathscr{G}, \qquad \widetilde{D} : \mathscr{G} \longrightarrow \mathscr{E} \tag{35.2}$$

*for some Hermitian vector bundle $\mathscr{G} \to \mathscr{M}$ which, in any local coordinate chart $U$ on $\mathscr{M}$ and with respect to local trivializations of $\mathscr{E}, \mathscr{G}$, may be represented as*

$$\begin{aligned} &Du(x) = \sum_j A_j(x)\partial_j u(x) + B(x)u(x) \ \text{ where, for some } r > n, \\ &A_j \in H^{2,r}_{\rm loc}\big(U, \mathbb{C}^{{\rm rank}\mathscr{G}\times{\rm rank}\mathscr{E}}\big), \quad B \in H^{1,r}_{\rm loc}\big(U, \mathbb{C}^{{\rm rank}\mathscr{G}\times{\rm rank}\mathscr{E}}\big), \end{aligned} \tag{35.3}$$

*and*

$$\begin{aligned} &\widetilde{D}v(x) = \sum_j \widetilde{A}_j(x)\partial_j v(x) + \widetilde{B}(x)v(x) \ \text{ where, for some } r > n, \\ &\widetilde{A}_j \in H^{2,r}_{\rm loc}\big(U, \mathbb{C}^{{\rm rank}\mathscr{E}\times{\rm rank}\mathscr{G}}\big), \quad \widetilde{B} \in H^{1,r}_{\rm loc}\big(U, \mathbb{C}^{{\rm rank}\mathscr{E}\times{\rm rank}\mathscr{G}}\big), \end{aligned} \tag{35.4}$$

*while*

$$Q \in {\rm Hom}\big(\mathscr{E}, \mathscr{E}\big) \ \text{ has coefficients in } L^r \text{ for some } r > n. \tag{35.5}$$

(ii) *The operator $L$ is elliptic, in the sense that its principal symbol satisfies*

$${\rm Sym}(L, \xi) : \mathscr{E} \to \mathscr{E} \ \text{ is invertible for each } \ \xi \in T^*\mathscr{M} \setminus 0. \tag{35.6}$$

(iii) *The operator $L$ is invertible as a mapping*

$$L : H^{1,2}(\mathscr{M}, \mathscr{E}) \longrightarrow H^{-1,2}(\mathscr{M}, \mathscr{E}). \tag{35.7}$$

Throughout, we let $\langle \cdot, \cdot \rangle$ denote the real pointwise inner product in the various vector bundles (a pairing not involving any complex conjugation). Next, given a differential operator $R$ of order $m$, locally written as

$$Ru(x) = \Big( \sum_{|\gamma| \le m} a^{\alpha\beta}_{\gamma}(x)\partial^{\gamma} u_{\beta}(x) \Big)_{\alpha} \tag{35.8}$$

its real transposed is given by $R^{\top} v = \Big( \sum_{|\gamma|\le m} (-1)^{|\gamma|} \partial^{\gamma}\big(a^{\alpha\beta}_{\gamma} v_{\alpha}\big) \Big)_{\beta}$. Also, the

principal symbol of (35.8) is the mapping sending a section $u$ into

$$\mathrm{Sym}(R,\xi)u := \Big(i^m \sum_{|\gamma|=m} a_\gamma^{\alpha\beta}\xi^\gamma u_\beta\Big)_\alpha \qquad \forall\,\xi\in T^*\mathscr{M}, \tag{35.9}$$

where $i := \sqrt{-1}\in\mathbb{C}$. It follows that for $\xi\in T^*\mathscr{M}$,

$$\begin{aligned} &\mathrm{Sym}(R^\top,\xi) = (-1)^m\,\mathrm{Sym}(R,\xi)^\top \ \text{ and}\\ &\mathrm{Sym}(R_1R_2,\xi) = \mathrm{Sym}(R_1,\xi)\mathrm{Sym}(R_2,\xi), \end{aligned} \tag{35.10}$$

whenever the latter composition is meaningful.

In view of the quasi-factorization (35.1) and the subsequent assumptions on the operators $\widetilde{D},D$, it follows that $L$ may be locally written as

$$Lu(x) = \sum_{j,k}\partial_j\big(\mathbb{A}_{jk}(x)\partial_k u(x)\big) + \sum_j \mathbb{B}_j(x)\partial_j u(x) + \mathbb{V}(x)u(x), \tag{35.11}$$

with coefficients

$$\begin{aligned} &\mathbb{A}_{jk} := \widetilde{A}_jA_k \in \mathscr{C}^{1+\gamma}_{\mathrm{loc}}, \ \text{ for some } \ \gamma>0,\\ &\mathbb{B}_j := -\sum_k(\partial_k\widetilde{A}_k)A_j + \widetilde{A}_jB + \widetilde{B}A_j \in H^{1,r}_{\mathrm{loc}},\\ &\mathbb{V} := \sum_j \widetilde{A}_j\partial_j B + \widetilde{B}B + Q \in L^r_{\mathrm{loc}}. \end{aligned} \tag{35.12}$$

If $E_L$ denotes the Schwartz kernel of the inverse $L^{-1}$ of $L$ in (35.7), then

$$E_L \in \mathscr{D}'(\mathscr{M}\times\mathscr{M},\mathscr{E}\otimes\mathscr{E}) \cap \mathscr{C}^{1+\gamma}_{\mathrm{loc}}\big(\mathscr{M}\times\mathscr{M}\setminus\mathrm{diag},\mathscr{E}\otimes\mathscr{E}\big) \tag{35.13}$$

for some $\gamma>0$; see [MiMiTa01, Proposition 2.3, p. 15] in this regard.

Associated with the quasi-factorization (35.1), introduce the family of first-order differential operators indexed by sections in the cotangent bundle

$$\partial_\xi^L := (-i)\,\mathrm{Sym}\big(\widetilde{D},\xi\big)D, \qquad \xi\in T^*\mathscr{M}. \tag{35.14}$$

Whenever $L$ is as in Hypothesis 1, it follows that $L^\top$, the real transposed of $L$, also satisfies all conditions in Hypothesis 1. In particular, we now have the quasi-factorization

$$L^\top = D^\top\widetilde{D}^\top + Q^\top. \tag{35.15}$$

We shall denote by $E_{L^\top}$ the Schwartz kernel of the inverse $\big(L^\top\big)^{-1}$ of the operator $L^\top : H^{1,2}(\mathscr{M},\mathscr{E}) \to H^{-1,2}(\mathscr{M},\mathscr{E})$ (which continues to enjoy a regularity property

analogous to (35.13)). Also, associated with the quasi-factorization (35.15), we introduce the family of first-order differential operators indexed by sections in the cotangent bundle

$$\partial_{\xi}^{L^{\top}} := (-i)\,\mathrm{Sym}\big(D^{\top},\xi\big)\widetilde{D}^{\top}, \qquad \xi \in T^{*}\mathscr{M}. \tag{35.16}$$

We now turn to assumptions of a geometric nature.

**Hypothesis 2 (Geometric Assumptions).** *Let $\Omega \subset \mathscr{M}$ be an Ahlfors regular domain, with outward unit conormal $\nu \in T^{*}\mathscr{M}$ and surface measure $\sigma := \mathscr{H}^{n-1}\lfloor\partial\Omega$, satisfying a two-sided local John condition (cf. [HoMiTa10] ).*

Above, $\mathscr{H}^{n-1}$ denotes the $(n-1)$-dimensional Hausdorff measure induced by the geodesic distance on $\mathscr{M}$. With $\mathrm{dist}(x,y)$ denoting the geodesic distance between points $x,y \in \mathscr{M}$, we define the non-tangential approach region

$$\Gamma(x) := \big\{y \in \Omega : \mathrm{dist}(x,y) < 2\,\mathrm{dist}(y,\partial\Omega)\big\}, \qquad x \in \partial\Omega. \tag{35.17}$$

Also, consider the non-tangential maximal operator acting on a vector bundle-valued function $u$ defined in $\Omega$ according to

$$(\mathscr{N}u)(x) := \sup_{y\in\Gamma(x)} |u(y)|, \qquad x \in \partial\Omega, \tag{35.18}$$

and define the non-tangential boundary trace, whenever meaningful, as

$$\Big(u\Big|_{\partial\Omega}^{\mathrm{n.t.}}\Big)(x) := \lim_{\Gamma(x)\ni y\to x} u(y), \qquad x \in \partial\Omega. \tag{35.19}$$

Given $\Omega$ as in Hypothesis 2, we define

$$\Omega_{+} := \Omega \ \text{ and } \ \Omega_{-} := \mathscr{M}\setminus\overline{\Omega}. \tag{35.20}$$

Then the set $\Omega_{-}$ also satisfies all conditions specified in Hypothesis 2. In fact, $\partial\Omega_{-} = \partial\Omega_{+} = \partial\Omega$, and the outward unit conormal to $\Omega_{-}$ is $-\nu$.

By $L^{p}(\partial\Omega,\mathscr{E})$, $0 < p \le \infty$, we shall denote the space of measurable sections $f : \partial\Omega \to \mathscr{E}$ which are $p$-th power integrable with respect to the surface measure $\sigma$. Also, for $p \in (1,\infty)$ we let $L_{1}^{p}(\partial\Omega,\mathscr{E})$ stand for the $L^{p}$-based Sobolev space of order one on $\partial\Omega$, defined as the collections of functions from $L^{p}$ possessing first-order tangential derivatives along $\partial\Omega$ in $L^{p}$ (cf. [HoMiTa10, MiEtAl14], for more details). Here we remark that if $p'$ denotes the Hölder conjugate exponent of $p \in (1,\infty)$, then $L_{-1}^{p}(\partial\Omega,\mathscr{E}) := \big(L_{1}^{p'}(\partial\Omega,\mathscr{E})\big)^{*}$.

Moving on, we introduce boundary layer potentials, starting with the double-layer potential defined below.

**Definition 1 (Double Layers).** Assume Hypotheses 1-2. In this context, define the double-layer associated with the quasi-factorization of $L$ in (35.1) as the integral operator sending any $f\in L^1(\partial\Omega,\mathscr{E})$ into the function defined at each $x\in\mathscr{M}\setminus\partial\Omega$ by

$$\begin{aligned}\mathscr{D}_L f(x) &:= \int_{\partial\Omega}\Big\langle \big(I_x\otimes(-i)\,\mathrm{Sym}\big(D^\top,\nu(y)\big)\widetilde{D}_y^\top\big)E_L(x,y),f(y)\Big\rangle_{\mathscr{E}_y}\mathrm{d}\sigma(y)\\ &= \int_{\partial\Omega}\Big\langle \big(I_x\otimes\partial_{\nu(y)}^{L^\top}\big)E_L(x,y),f(y)\Big\rangle_{\mathscr{E}_y}\mathrm{d}\sigma(y),\end{aligned}\tag{35.21}$$

(where $I$ denotes the identity, here acting in the variable $x$, etc.). In addition, consider its principal value version on $\partial\Omega$ (in the sense of removing a geodesic ball centered at the singularity and taking the limit as the radius goes to zero) acting on some $f\in L^1(\partial\Omega,\mathscr{E})$ according to

$$\begin{aligned}K_L f(x) &:= \mathrm{P.V.}\int_{\partial\Omega}\Big\langle \big(I_x\otimes(-i)\,\mathrm{Sym}\big(D^\top,\nu(y)\big)\widetilde{D}_y^\top\big)E_L(x,y),f(y)\Big\rangle_{\mathscr{E}_y}\mathrm{d}\sigma(y)\\ &= \mathrm{P.V.}\int_{\partial\Omega}\Big\langle \big(I_x\otimes\partial_{\nu(y)}^{L^\top}\big)E_L(x,y),f(y)\Big\rangle_{\mathscr{E}_y}\mathrm{d}\sigma(y),\end{aligned}\tag{35.22}$$

at $\sigma$-a.e. $x\in\partial\Omega$.

We continue by considering the single-layer potential.

**Definition 2 (Single Layers).** Assuming Hypotheses 1-2, define the single-layer operator as the integral operator sending any $f\in L^1(\partial\Omega,\mathscr{E})$ into the function defined at each $x\in\mathscr{M}\setminus\partial\Omega$ by

$$\mathscr{S}_L f(x) := \int_{\partial\Omega}\big\langle E_L(x,y),f(y)\big\rangle_{\mathscr{E}_y}\mathrm{d}\sigma(y).\tag{35.23}$$

Moreover, define the boundary version of (35.23) by setting, at $\sigma$-a.e. $x\in\partial\Omega$,

$$S_L f(x) := \int_{\partial\Omega}\big\langle E_L(x,y),f(y)\big\rangle_{\mathscr{E}_y}\mathrm{d}\sigma(y).\tag{35.24}$$

## 35.2 Formulation of the Main Results

This section contains the principal results of this chapter, dealing with properties of the single-layer and double-layer potential operators associated with a second-order elliptic system, such as non-tangential maximal function estimates, jump formulas,

square-function estimates, Carleson measure estimates, operator identities, integral representation formulas, and Green-type identities. Our first theorem of this flavor concerns the double-layer potential operator.

**Theorem 1 (Properties of the Double Layer).** *Assume Hypotheses 1-2. Then for each integrability exponent $p \in (1,\infty)$ the following conclusions hold.*

(1) *Given any covariant derivative $\nabla$ on $\mathscr{E}$, there exists a constant $C \in (0,\infty)$ with the property that*

$$\|\mathscr{N}(\mathscr{D}_L f)\|_{L^p(\partial\Omega)} \leq C\|f\|_{L^p(\partial\Omega,\mathscr{E})}, \qquad \forall f \in L^p(\partial\Omega,\mathscr{E}), \tag{35.25}$$

$$\|\mathscr{N}(\nabla\mathscr{D}_L f)\|_{L^p(\partial\Omega)} \leq C\|f\|_{L^p_1(\partial\Omega,\mathscr{E})}, \qquad \forall f \in L^p_1(\partial\Omega,\mathscr{E}). \tag{35.26}$$

(2) *The following operators are well defined, linear, and bounded:*

$$K_L : L^p(\partial\Omega,\mathscr{E}) \to L^p(\partial\Omega,\mathscr{E}), \quad K_L : L^p_1(\partial\Omega,\mathscr{E}) \to L^p_1(\partial\Omega,\mathscr{E}). \tag{35.27}$$

(3) *For each $f \in L^p(\partial\Omega,\mathscr{E})$, one has*

$$(\mathscr{D}_L f)\Big|^{\rm n.t.}_{\partial\Omega_\pm} = \big(\pm\tfrac{1}{2}I + K_L\big)f \quad \text{at } \sigma\text{-a.e. point on } \partial\Omega. \tag{35.28}$$

(4) *For each $f \in L^p_1(\partial\Omega,\mathscr{E})$, the non-tangential traces*

$$\big(\nabla\mathscr{D}_L f\big)\Big|^{\rm n.t.}_{\partial\Omega_\pm} \quad \text{exist } \sigma\text{-a.e. on } \partial\Omega. \tag{35.29}$$

(5) *The conormal derivative of the double-layer does not jump across the boundary. That is, for every $f \in L^p_1(\partial\Omega,\mathscr{E})$, at $\sigma$-a.e. point on $\partial\Omega$, one has*

$$(-i)\mathrm{Sym}\,(\widetilde{D},\nu)\Big(D\mathscr{D}_L f\Big)\Big|^{\rm n.t.}_{\partial\Omega_+} = (-i)\mathrm{Sym}\,(\widetilde{D},\nu)\Big(D\mathscr{D}_L f\Big)\Big|^{\rm n.t.}_{\partial\Omega_-}. \tag{35.30}$$

*In particular, for every $f \in L^p_1(\partial\Omega)$ it is meaningful to abbreviate*

$$\partial^L_\nu \mathscr{D}_L f := (-i)\mathrm{Sym}\,(\widetilde{D},\nu)\Big(D\mathscr{D}_L f\Big)\Big|^{\rm n.t.}_{\partial\Omega_\pm}, \qquad \sigma\text{-a.e. on } \partial\Omega. \tag{35.31}$$

*Defined as such, the conormal derivative of the double-layer potential induces a bounded mapping*

$$\partial^L_\nu \mathscr{D}_L : L^p_1(\partial\Omega,\mathscr{E}) \longrightarrow L^p(\partial\Omega,\mathscr{E}). \tag{35.32}$$

(6) *Given a first-order differential operator $P:\mathscr{E}\to\mathscr{F}$, for some vector bundle $\mathscr{F}\to\mathscr{M}$, define the tangential derivative operator $\partial_{\tau_P}$ induced by $P$ on $\partial\Omega$ according to*

$$\begin{aligned}\partial_{\tau_P} &:= P-\mathrm{Sym}\,(P,\nu)\mathrm{Sym}\,(L,\nu)^{-1}\mathrm{Sym}\,(\widetilde{D},\nu)D\\ &= P+(-i)\mathrm{Sym}\,(P,\nu)\mathrm{Sym}\,(L,\nu)^{-1}\partial_\nu^L.\end{aligned}\tag{35.33}$$

*Then for every $f\in L^p_1(\partial\Omega,\mathscr{E})$ one has at $\sigma$-a.e. point on $\partial\Omega$*

$$\begin{aligned}\big(P\mathscr{D}_Lf\big)\Big|^{\mathrm{n.t.}}_{\partial\Omega_\pm} &= \partial_{\tau_P}\big(\pm\tfrac12 I+K_L\big)f\\ &\quad-(-i)\mathrm{Sym}\,(P,\nu)\mathrm{Sym}\,(L,\nu)^{-1}\partial_\nu^L\mathscr{D}_Lf.\end{aligned}\tag{35.34}$$

*As a consequence, for every $f\in L^p_1(\partial\Omega,\mathscr{E})$,*

$$\big(P\mathscr{D}_Lf\big)\Big|^{\mathrm{n.t.}}_{\partial\Omega_+}-\big(P\mathscr{D}_Lf\big)\Big|^{\mathrm{n.t.}}_{\partial\Omega_-}=\partial_{\tau_P}f\ \text{ at }\sigma\text{-a.e. point on }\partial\Omega.\tag{35.35}$$

(7) *There exists $C\in(0,\infty)$ such that for every $f\in L^2(\partial\Omega,\mathscr{E})$ one has the square-function estimate (hereafter* dV *denotes the volume element on $\mathscr{M}$)*

$$\int_{\mathscr{M}\setminus\partial\Omega}\big|\nabla(\mathscr{D}_Lf)(x)\big|^2\,\mathrm{dist}(x,\partial\Omega)\,\mathrm{dV}(x)\le C\int_{\partial\Omega}|f|^2\,\mathrm{d}\sigma.\tag{35.36}$$

(8) *Whenever $p\in(2,\infty)$, there exists $C\in(0,\infty)$ with the property that for every $f\in L^p(\partial\Omega,\mathscr{E})$ one has (abbreviating $\Delta(x,r):=B(x,r)\cap\partial\Omega$)*

$$\int_{\partial\Omega}\sup_{r>0}\left(\frac{1}{\sigma(\Delta(x,r))}\int_{B(x,r)\cap\Omega}\big|\nabla(\mathscr{D}_Lf)\big|^2\mathrm{dist}(\cdot,\partial\Omega)\,\mathrm{dV}\right)^{\frac{p}{2}}\mathrm{d}\sigma(x)\le C\int_{\partial\Omega}|f|^p\mathrm{d}\sigma.\tag{35.37}$$

*Moreover, corresponding to $p=\infty$, for each $f\in L^\infty(\partial\Omega,\mathscr{E})$ the measure $|(\nabla\mathscr{D}_Lf)(x)|^2\,\mathrm{dist}\,(x,\partial\Omega)\,\mathrm{dV}(x)$ is Carleson in $\Omega$ in the precise sense that*

$$\sup_{\substack{r>0\\ x\in\partial\Omega}}\left(\frac{1}{\sigma(\Delta(x,r))}\int_{B(x,r)\cap\Omega}\big|\nabla\mathscr{D}_Lf\big|^2\mathrm{dist}(\cdot,\partial\Omega)\,\mathrm{dV}\right)^{\frac12}\le C\|f\|_{L^\infty(\partial\Omega)}.\tag{35.38}$$

(9) *For each $f\in L^1(\partial\Omega,\mathscr{E})$ one has*

$$L(\mathscr{D}_Lf)=0\ \text{ in }\ \mathscr{M}\setminus\partial\Omega.\tag{35.39}$$

(10) *For a local frame* $\{X_j\}_j$ *in* $T\mathscr{M}$*, consisting of vector fields with continuous coefficients, define the tangential derivatives*

$$\partial_{\tau_{jk}} := \nu(X_j)\nabla_{X_k} - \nu(X_k)\nabla_{X_j}, \qquad j,k \in \{1,\dots,n\}, \tag{35.40}$$

*and view them as mappings from* $L^p_1(\partial\Omega,\mathscr{E})$ *into* $L^p(\partial\Omega,\mathscr{E})$*. Then, locally, for each fixed pair of indices* $j,k \in \{1,\dots,n\}$*, the commutator*

$$\big[\partial_{\tau_{jk}}, K_L\big] : L^p_1(\partial\Omega,\mathscr{E}) \longrightarrow L^p(\partial\Omega,\mathscr{E}) \tag{35.41}$$

*may be written as a linear combination of terms of the form* $[M_{\nu_\ell}, T_\ell]\partial_{\tau_{rs}}$ *plus a compact mapping from* $L^p_1(\partial\Omega,\mathscr{E})$ *into* $L^p(\partial\Omega,\mathscr{E})$*. Here,* $[M_{\nu_\ell}, T_\ell]$ *is the commutator between* $M_{\nu_\ell}$*, the operator of pointwise multiplication with the component* $\nu_\ell$ *of* $\nu$*, and a Calderón–Zygmund singular integral operator* $T_\ell$*, bounded on* $L^p(\partial\Omega)$*.*

(11) *With* $\mathrm{VMO}(\partial\Omega)$ *denoting the Sarason space of functions with vanishing mean oscillations on* $\partial\Omega$*, one has the implication*

$$\left.\begin{array}{l} \nu \in \mathrm{VMO}(\partial\Omega, T^*\mathscr{M}) \text{ and} \\ K_L \text{ compact on } L^p(\partial\Omega,\mathscr{E}) \end{array}\right\} \Rightarrow K_L \text{ compact on } L^p_1(\partial\Omega,\mathscr{E}). \tag{35.42}$$

Going further, we turn our attention to the single-layer potential operator.

**Theorem 2 (Properties of the Single Layer).** *Assume Hypotheses 1-2. Then for each integrability exponent* $p \in (1,\infty)$ *the following conclusions hold.*

(1) *Given any covariant derivative* $\nabla$ *on* $\mathscr{E}$*, there exists a constant* $C \in (0,\infty)$ *with the property that for each* $f \in L^p(\partial\Omega,\mathscr{E})$*, one has*

$$\|\mathscr{N}(\mathscr{S}_L f)\|_{L^p(\partial\Omega)} + \|\mathscr{N}(\nabla\mathscr{S}_L f)\|_{L^p(\partial\Omega)} \leq C\|f\|_{L^p(\partial\Omega,\mathscr{E})}. \tag{35.43}$$

*Moreover*

$$\|\mathscr{N}(\mathscr{S}_L f)\|_{L^p(\partial\Omega)} \leq C\|f\|_{L^p_{-1}(\partial\Omega,\mathscr{E})}, \qquad \forall f \in L^p_{-1}(\partial\Omega,\mathscr{E}). \tag{35.44}$$

(2) *The operators*

$$S_L : L^p(\partial\Omega,\mathscr{E}) \to L^p_1(\partial\Omega,\mathscr{E}), \quad S_L : L^p_{-1}(\partial\Omega,\mathscr{E}) \to L^p(\partial\Omega,\mathscr{E}), \tag{35.45}$$

*are well defined, linear, and bounded.*

(3) *For each* $f \in L^p(\partial\Omega,\mathscr{E})$*, the non-tangential traces*

$$\big(\nabla\mathscr{S}_L f\big)\Big|^{\text{n.t.}}_{\partial\Omega_\pm} \quad \text{exist } \sigma\text{-a.e. on } \partial\Omega. \tag{35.46}$$

(4) *For each* $f\in L^p(\partial\Omega,\mathscr{E})$, *one has*

$$\begin{aligned}\big(\partial_\nu^L\mathscr{S}_Lf\big)\Big|^{\rm n.t.}_{\partial\Omega_\pm}&:=(-i){\rm Sym}\,(\widetilde{D},\nu)\Big(D\mathscr{S}_Lf\Big)\Big|^{\rm n.t.}_{\partial\Omega_\pm}\\&=\Big(\mp\tfrac{1}{2}I+\big(K_{L^\top}\big)^\top\Big)f\quad\text{at }\sigma\text{-a.e. point on }\partial\Omega,\end{aligned}\tag{35.47}$$

*where* $\big(K_{L^\top}\big)^\top$ *denotes the* (*real*) *transposed of* $K_{L^\top}$, *the principal value of the double-layer potential associated with* $L^\top$ (*much as* $K_L$ *has been associated with* $L$ *in Definition 1, this time making use of the quasi-factorization* (35.15)).

(5) *Given a first-order differential operator* $P:\mathscr{E}\to\mathscr{F}$, *for some vector bundle* $\mathscr{F}\to\mathscr{M}$, *recall the tangential derivative operator* $\partial_{\tau_P}$ *induced by* $P$ *on* $\partial\Omega$ *as in* (35.33). *Then for every* $f\in L^p(\partial\Omega,\mathscr{E})$, *one has*

$$\begin{aligned}\big(P\mathscr{S}_Lf\big)\Big|^{\rm n.t.}_{\partial\Omega_\pm}&=i\,{\rm Sym}\,(P,\nu){\rm Sym}\,(L,\nu)^{-1}\Big(\mp\tfrac{1}{2}I+\big(K_{L^\top}\big)^\top\Big)f\\&\quad+\partial_{\tau_P}\big(S_Lf\big),\quad\text{at }\sigma\text{-a.e. point on }\partial\Omega.\end{aligned}\tag{35.48}$$

*As a consequence, given* $f\in L^p(\partial\Omega,\mathscr{E})$, *at* $\sigma$*-a.e. point on* $\partial\Omega$, *one has*

$$\big(P\mathscr{S}_Lf\big)\Big|^{\rm n.t.}_{\partial\Omega_+}-\big(P\mathscr{S}_Lf\big)\Big|^{\rm n.t.}_{\partial\Omega_-}=(-i){\rm Sym}\,(P,\nu){\rm Sym}\,(L,\nu)^{-1}f.\tag{35.49}$$

(6) *There exists* $C\in(0,\infty)$ *with the property that one has the square-function estimates*

$$\int_{\mathscr{M}\setminus\partial\Omega}\big|\nabla^2(\mathscr{S}_Lf)(x)\big|^2\,{\rm dist}(x,\partial\Omega)\,{\rm dV}(x)\leq C\int_{\partial\Omega}|f|^2\,{\rm d}\sigma\tag{35.50}$$

*for every* $f\in L^2(\partial\Omega,\mathscr{E})$, *and for every* $f\in L^2_{-1}(\partial\Omega,\mathscr{E})$

$$\Big(\int_{\mathscr{M}\setminus\partial\Omega}\big|\nabla(\mathscr{S}_Lf)(x)\big|^2\,{\rm dist}(x,\partial\Omega)\,{\rm dV}(x)\Big)^{\frac{1}{2}}\leq C\|f\|_{L^2_{-1}(\partial\Omega,\mathscr{E})}.\tag{35.51}$$

(7) *For each* $f\in L^1(\partial\Omega,\mathscr{E})$ *one has*

$$L(\mathscr{S}_Lf)=0\ \text{ in }\ \mathscr{M}\setminus\partial\Omega.\tag{35.52}$$

(8) *The following Fredholm property result holds:*

$$\begin{aligned}&\text{if }\nu\in{\rm VMO}(\partial\Omega,T^*\mathscr{M})\text{ and }K_L,K_{L^\top}\text{ are compact on }L^p(\partial\Omega,\mathscr{E})\\&\text{then }S_L:L^p(\partial\Omega,\mathscr{E})\to L^p_1(\partial\Omega,\mathscr{E})\text{ is a Fredholm operator.}\end{aligned}\tag{35.53}$$

We next discuss integral representation formulas involving volume and boundary layer potential operators.

**Theorem 3 (Layer Potential Integral Representation Formula).** *Assume Hypotheses 1-2. For a given function $u \in \mathscr{C}^1(\Omega,\mathscr{E})$, define the boundary conormal derivative as*

$$\partial_\nu^L u := (-i)\,\mathrm{Sym}(\widetilde{D},\nu)\big(Du\big)\Big|_{\partial\Omega}^{\mathrm{n.t.}} \quad \text{on } \partial\Omega, \tag{35.54}$$

*whenever this is meaningful in a pointwise, a.e. sense (with respect to the surface measure). Also, define the Newtonian (volume) potential $\Pi_L$ acting on a function $v \in L^1(\Omega,\mathscr{E})$ by*

$$\Pi_L v(x) := \int_\Omega \big\langle E_L(x,y), v(y)\big\rangle_{\mathscr{E}_y}\,\mathrm{dV}(y), \qquad x \in \Omega. \tag{35.55}$$

*Then every function satisfying*

$$\begin{aligned} &u \in \mathscr{C}^1(\Omega,\mathscr{E}), \quad Lu \in L^1(\Omega,\mathscr{E}), \quad \mathscr{N}(u), \mathscr{N}(Du) \in L^1(\partial\Omega)\\ &\text{and such that there exist } u\big|_{\partial\Omega}^{\mathrm{n.t.}} \text{ and } (Du)\big|_{\partial\Omega}^{\mathrm{n.t.}}\ \sigma\text{-a.e. on } \partial\Omega, \end{aligned} \tag{35.56}$$

*admits the integral representation formula*

$$u = \Pi_L(Lu) + \mathscr{D}_L\big(u\big|_{\partial\Omega}^{\mathrm{n.t.}}\big) - \mathscr{S}_L\big(\partial_\nu^L u\big) \quad \text{in } \Omega. \tag{35.57}$$

Our next theorem establishes a basic Green-type identity.

**Theorem 4 (Green's Identity).** *Assume Hypotheses 1-2. If $u,v \in \mathscr{C}^1(\Omega,\mathscr{E})$ are two functions such that, for $p,p' \in (1,\infty)$ with $1/p+1/p'=1$, one has*

$$\begin{aligned} &Lu \in L^{np/(n+p-1)}(\Omega,\mathscr{E}), \quad L^\top v \in L^{np/(np-n+1)}(\Omega,\mathscr{E}),\\ &\mathscr{N}(u), \mathscr{N}(Du) \in L^p(\partial\Omega),\ \mathscr{N}(v), \mathscr{N}(\widetilde{D}v) \in L^{p'}(\partial\Omega), \ \text{and}\\ &\text{there exist } u\big|_{\partial\Omega}^{\mathrm{n.t.}},\ (Du)\big|_{\partial\Omega}^{\mathrm{n.t.}},\ v\big|_{\partial\Omega}^{\mathrm{n.t.}},\ (\widetilde{D}v)\big|_{\partial\Omega}^{\mathrm{n.t.}},\ \sigma\text{-a.e. on } \partial\Omega, \end{aligned} \tag{35.58}$$

*then*

$$\begin{aligned} &\int_\Omega \big\langle Lu, v\big\rangle\,\mathrm{dV} - \int_\Omega \big\langle u, L^\top v\big\rangle\,\mathrm{dV}\\ &\qquad = \int_{\partial\Omega} \big\langle \partial_\nu^L u\,,\, v\big|_{\partial\Omega}^{\mathrm{n.t.}}\big\rangle\,\mathrm{d}\sigma - \int_{\partial\Omega} \big\langle u\big|_{\partial\Omega}^{\mathrm{n.t.}}\,,\, \partial_\nu^{L^\top} v\big\rangle\,\mathrm{d}\sigma. \end{aligned} \tag{35.59}$$

Finally, in our last result in this section we collect some fundamental operator identities involving boundary layer potentials and their transpositions.

**Theorem 5 (Operator Identities).** *Assume Hypotheses 1-2. Then the following boundary layer operator identities hold.*

(1) *For each $p\in(1,\infty)$ one has the intertwining formula*

$$S_L\big(K_{L^\top}\big)^\top = K_L S_L \ \text{ on } \ L^p(\partial\Omega,\mathscr{E}), \ \text{ and on } \ L^p_{-1}(\partial\Omega,\mathscr{E}). \tag{35.60}$$

(2) *If the exponents $p,p'\in(1,\infty)$ are such that $1/p+1/p'=1$, then*

$$\int_{\partial\Omega}\big\langle \partial_\nu^L \mathscr{D}_L f\,,\,g\big\rangle\,\mathrm{d}\sigma = \int_{\partial\Omega}\big\langle f\,,\,\partial_\nu^{L^\top}\mathscr{D}_{L^\top} g\big\rangle\,\mathrm{d}\sigma \tag{35.61}$$

*for every $f\in L^p_1(\partial\Omega,\mathscr{E})$ and $g\in L^{p'}_1(\partial\Omega,\mathscr{E})$. As a consequence, for each $p\in(1,\infty)$ the operator (35.32) further extends to a well-defined, linear, and bounded mapping*

$$\partial_\nu^L\mathscr{D}_L : L^p(\partial\Omega,\mathscr{E}) \longrightarrow L^p_{-1}(\partial\Omega,\mathscr{E}) \tag{35.62}$$

*whose transposed is the operator*

$$\big(\partial_\nu^L\mathscr{D}_L\big)^\top = \partial_\nu^{L^\top}\mathscr{D}_{L^\top} : L^{p'}_1(\partial\Omega,\mathscr{E}) \longrightarrow L^{p'}(\partial\Omega,\mathscr{E}). \tag{35.63}$$

(3) *For each $p\in(1,\infty)$ the mapping*

$$S_L : L^p(\partial\Omega,\mathscr{E}) \longrightarrow L^p(\partial\Omega,\mathscr{E}) \tag{35.64}$$

*is compact, and its (real) transposed is the operator*

$$(S_L)^\top = S_{L^\top} : L^{p'}(\partial\Omega,\mathscr{E}) \longrightarrow L^{p'}(\partial\Omega,\mathscr{E}), \quad 1/p+1/p'=1. \tag{35.65}$$

(4) *For each $p\in(1,\infty)$ the following operator identities hold on $L^p(\partial\Omega,\mathscr{E})$:*

$$\big(\tfrac12 I+K_L\big)\big(-\tfrac12 I+K_L\big) = S_L\big(\partial_\nu^L\mathscr{D}_L\big), \tag{35.66}$$

$$\Big(\tfrac12 I+\big(K_{L^\top}\big)^\top\Big)\Big(-\tfrac12 I+\big(K_{L^\top}\big)^\top\Big) = \big(\partial_\nu^L\mathscr{D}_L\big)S_L, \tag{35.67}$$

$$\big(K_{L^\top}\big)^\top\big(\partial_\nu^L\mathscr{D}_L\big) = \big(\partial_\nu^L\mathscr{D}_L\big)K_L. \tag{35.68}$$

## 35.3 Examples

On the geometrical side, classes of domains satisfying the conditions listed in Hypothesis 2 include Lipschitz domains or, more generally, domains locally given as the upper-graphs of continuous functions with gradients in BMO, the John–Nirenberg space of functions with bounded mean oscillations. This being said,

domains satisfying the conditions in Hypothesis 2 need not be locally of upper-graph type. Examples include the category of chord-arc domains satisfying a two-sided corkscrew condition in the plane and, in higher dimensions, two-sided NTA domains (in the sense of Jerison–Kenig [JeKe82]), with an Ahlfors regular boundary.

On the analytical side, consider first the case of the zero-th order perturbation

$$L := \Delta_{\rm LB} - V \tag{35.69}$$

of the Laplace–Beltrami operator $\Delta_{\rm LB}$ on the manifold $\mathscr{M}$, by a scalar function $V$. This Schrödinger type operator satisfies the quasi-factorization as in (35.1) with

$$D := \text{grad}, \quad \widetilde{D} := -D^{\top} = \text{div}, \quad Q := -V. \tag{35.70}$$

In turn, this quasi-factorization yields the conormal (cf. (35.16) with $\xi = \nu$)

$$\partial_{\nu}^{L^{\top}} = (-i)\,\text{Sym}\big(D^{\top},\nu\big)\widetilde{D}^{\top} = \langle \nu, \text{grad}\rangle = \partial_{\nu}, \tag{35.71}$$

the ordinary covariant derivative in the direction of $\nu$. Via the recipe from Definition 1, this conormal then produces (with $E$ denoting the Schwartz kernel of $L^{-1}$) the principal value double-layer potential

$$(Kf)(x) := \text{P.V.} \int_{\partial\Omega} \partial_{\nu(y)}\big[E(x,y)\big] f(y)\,d\sigma(y) \quad x \in \partial\Omega, \tag{35.72}$$

which has been studied in [MiTa99] in the context of Lipschitz domains.

Another natural example is obtained starting with the Hodge-Laplacian $\Delta_{\rm HL} = -(\delta d + d\delta)$, acting on $l$-forms for some fixed $l \in \{0,1,\ldots,n\}$, and then considering $L := \Delta_{\rm HL} - V$ where $V$ is a scalar potential. A quasi-factorization of $L$ as in (35.1) is then obtained by taking

$$D := \begin{pmatrix} \delta \\ d \end{pmatrix} \tag{35.73}$$

mapping sections of the vector bundle $\mathscr{E} := \Lambda^{l} T\mathscr{M}$ into sections of the vector bundle $\mathscr{G} := \Lambda^{l+1} T\mathscr{M} \oplus \Lambda^{l-1} T\mathscr{M}$, then taking

$$\widetilde{D} := -D^{\top} = -\begin{pmatrix} d & \delta \end{pmatrix} \tag{35.74}$$

and, finally, $Q := -V$. Indeed, in this scenario $\widetilde{D}D + Q = -d\delta - \delta d - V = L$, as wanted. Moreover, since

$$i\,\text{Sym}(D^{\top},\nu) = \big(i\,\text{Sym}(d,\nu) \;\; i\,\text{Sym}(\delta,\nu)\big) = (-\nu\wedge\cdot \;\; \nu\vee\cdot), \tag{35.75}$$

this quasi-factorization yields the conormal

$$\partial_{\nu}^{L^{\top}} = (-i)\,\text{Sym}\big(D^{\top},\nu\big)\widetilde{D}^{\top} = -\nu\wedge\delta + \nu\vee d. \tag{35.76}$$

As such, the specific format of the double layer associated with this factorization of the perturbed Hodge-Laplacian $L = \Delta_{\rm HL} - V$ is

$$K_L f(x) = \text{P.V.} \int_{\partial\Omega} \big\langle \nu(y) \vee d_y \Gamma_l(x,y) - \nu(y) \wedge \delta_y \Gamma_l(x,y), f(y) \big\rangle_y \, \mathrm{d}\sigma(y) \tag{35.77}$$

for each $f \in L^1(\partial\Omega, \Lambda^l T\mathscr{M})$ and $x \in \partial\Omega$, where $\Gamma_l$ is the Schwartz kernel of $L^{-1}$ on $l$-forms.

Another important quasi-factorization of the Hodge-Laplacian is offered by Weitzenböck's formula

$$\Delta_{\rm HL} = -\nabla^\top \nabla - \mathfrak{Ric}. \tag{35.78}$$

Here $\nabla$ denotes the Levi-Civita connection (or covariant derivative) on $\mathscr{M}$, whose action is extended to differential forms in a canonical fashion. Also, $\mathfrak{Ric}$ is the so-called Weitzenböck operator, a curvature term of order zero (depending linearly on the Riemann curvature, via real coefficients) that preserves $l$-forms, and is self-adjoint. In this scenario, the quasi-factorization of $L := \Delta_{\rm HL}$ as in (35.1) is satisfied with

$$\widetilde{D} := -\nabla^\top, \quad D := \nabla, \quad Q := -\mathfrak{Ric}. \tag{35.79}$$

Given that

$$\mathrm{Sym}(\nabla, \xi) u = i\xi \otimes u, \qquad \forall\, \xi \in T^*\mathscr{M}, \tag{35.80}$$

and

$$\mathrm{Sym}(\nabla^\top, \xi)(\eta \otimes u) = -i\langle \xi, \eta \rangle u, \qquad \forall\, \xi, \eta \in T^*\mathscr{M}, \tag{35.81}$$

this quasi-factorization yields the conormal

$$\begin{aligned} \partial_\nu^{L^\top} &= (-i)\,\mathrm{Sym}\big(D^\top, \nu\big)\widetilde{D}^\top = i \sum_j \mathrm{Sym}(\nabla^\top, \nu)\big(dx_j \otimes \nabla_{\partial_j}\big) \\ &= \sum_j \langle \nu, dx_j \rangle \nabla_{\partial_j} = \sum_j \nu_j^\sharp \nabla_{\partial_j} = \nabla_{\nu^\sharp} \end{aligned} \tag{35.82}$$

where $\nu^\sharp$ is the outward unit normal to $\Omega$ (i.e., the metric identification of the conormal $\nu \in T^*\mathscr{M}$ with a tangent vector). Granted this, formula (35.22) yields the following specific format of the double-layer potential associated with the above factorization of the Hodge-Laplacian $L = \Delta_{\rm HL}$:

$$K_L f(x) = \text{P.V.} \int_{\partial\Omega} \big\langle \big(I_x \otimes \nabla_{\nu^\sharp(y)}\big) E_L(x,y), f(y) \big\rangle_y \mathrm{d}\sigma(y) \tag{35.83}$$

at $\sigma$-a.e. $x \in \partial\Omega$, where $E_L$ is the Schwartz kernel of $L^{-1}$.

As a final conclusion, all results in §35.2 apply to the above double-layer potentials when considered in the geometric context described in Hypothesis 2.

**Acknowledgements** The first-named author was partially supported by the Simons Foundation grant # 200750. The second-named author was partially supported by the Simons Foundation grant # 318658; part of this work has been carried out while she was a von Neumann Fellow at the Institute for Advanced Study at Princeton, with partial support from Temple University. The third-named author was partially supported by the Simons Foundation grant # 281566, and a University of Missouri Research Leave.

## References

[JeKe82] D.S. Jerison and C.E. Kenig, *Boundary behavior of harmonic functions in nontangentially accessible domains*, Adv. in Math., 46 (1982), no. 1, 80–147.

[MiTa99] M. Mitrea and M. Taylor, *Boundary layer methods for Lipschitz domains in Riemannian manifolds*, Journal of Functional Analysis **163**, 181–251 (1999).

[MiMiTa01] D. Mitrea, M. Mitrea, and M. Taylor, *Layer Potentials, the Hodge Laplacian, and Global Boundary Problems in Nonsmooth Riemannian Manifolds*, Memoirs of the American Mathematical Society, March 2001, Volume 150, Number 713.

[HoMiTa10] S. Hofmann, M. Mitrea, and M. Taylor, *Singular Integrals and Elliptic Boundary Problems on Regular Semmes–Kenig–Toro Domains*, International Mathematics Research Notices, Vol. 2010, No. 14, pp. 2567–2865.

[MiEtAl14] D. Mitrea, Irina Mitrea, M. Mitrea, and M. Taylor, *Boundary Problems for the Hodge–Laplacian on Regular Semmes–Kenig–Toro Subdomains of Riemannian Manifolds*, book manuscript, 2014.

# Chapter 36
# The Regularity Problem in Rough Subdomains of Riemannian Manifolds

**M. Mitrea and B. Schmutzler**

## 36.1 Formulation of the Regularity Problem

Let $(\mathscr{M}, g)$ denote a compact, oriented, boundaryless Riemannian manifold of class $\mathscr{C}^2$ and real dimension $n \in \mathbb{N}$, $n \geq 2$. The convention of summing over repeated indices is used throughout. In particular, the local expression of the Riemannian metric tensor is $g = g_{jk}\, dx_j \otimes dx_k$. As is customary, we also use the symbol $g$ to denote $\det(g_{jk})$, and the inverse of $(g_{jk})$ is denoted by $(g^{jk})$. The Riemannian volume form dVol has the local expression $\mathrm{dVol} = \sqrt{g}\,\mathrm{d}x$, where $\mathrm{d}x$ is $n$-dimensional Lebesgue measure and $g = \det(g_{jk})$.

Throughout this paper, $\Omega \subset \mathscr{M}$ is a regular Semmes–Kenig–Toro (SKT) domain. This means that $\Omega$ is an open, nonempty subset of $\mathscr{M}$ that satisfies a two-sided local John condition, has Ahlfors regular boundary $\partial\Omega$, and its outward unit conormal $\nu : \partial\Omega \to T^*\mathscr{M}$ belongs to the Sarason space $\mathrm{VMO}(\partial\Omega, \sigma)$ consisting of functions with vanishing mean oscillations; see [MiEtAl14] for details. The measure $\sigma$ on $\partial\Omega$ is defined by setting $\sigma := \mathscr{H}^{n-1} \lfloor \partial\Omega$, where $\mathscr{H}^{n-1}$ denotes the $(n-1)$-dimensional Hausdorff measure (which depends on the metric $g$). Regular SKT domains provide the most general environment that supports compactness results such as Theorem 3, one of the key results of this paper.

For distinct points $x, y \in \mathscr{M}$, a rectifiable curve joining $x$ and $y$ is a curve $\gamma$ in $\mathscr{M}$ which has a Lipschitz parametrization $\gamma : [0,1] \to \mathscr{M}$ such that $\gamma(0) = x$ and $\gamma(1) = y$. The length of such a curve is defined by

$$L(\gamma) := \int_0^1 \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle}\, \mathrm{d}t. \tag{36.1}$$

M. Mitrea (✉) • B. Schmutzler
University of Missouri, Columbia, MO, USA
e-mail: mitream@missouri.edu; brock.schmutzler@mail.missouri.edu



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_36

The geodesic distance between distinct points $x, y \in \mathscr{M}$ is given by

$$\operatorname{dist}(x,y) := \inf\big\{L(\gamma) : \gamma \text{ is a rectifiable curve joining } x \text{ and } y\big\}. \tag{36.2}$$

Throughout the paper fix $\kappa > 0$ and define the non-tangential approach region

$$\Gamma(x) := \big\{y \in \Omega : \operatorname{dist}(x,y) < (1+\kappa)\operatorname{dist}(y,\partial\Omega)\big\}, \quad x \in \partial\Omega. \tag{36.3}$$

If $u : \Omega \to \mathbb{C}$, define the non-tangential maximal function by

$$(\mathscr{N}u)(x) := \sup_{y\in\Gamma(x)} |u(y)|, \quad x \in \partial\Omega, \tag{36.4}$$

and the non-tangential boundary trace by setting, whenever meaningful,

$$u\Big|^{\text{n.t.}}_{\partial\Omega}(x) := \lim_{\Gamma(x)\ni y\to x} u(y), \quad x \in \partial\Omega. \tag{36.5}$$

We now proceed to define an appropriate $L^p$-based Sobolev space of order one on $\partial\Omega$. The first step is to define such a space in the Euclidean context. Specifically, given any $\varphi \in \mathscr{C}^1_0(\mathbb{R}^n)$, set

$$\partial_{\tau_{jk}}\varphi := \nu_j\big(\partial_k\varphi\big)\Big|_{\partial\Omega} - \nu_k\big(\partial_j\varphi\big)\Big|_{\partial\Omega}, \quad \forall j,k \in \{1,\dots,n\}. \tag{36.6}$$

For any $p \in (1,\infty)$, the Euclidean Sobolev space $L^p_1(\partial\Omega)$ is then defined to be the collection of all functions $f \in L^p(\partial\Omega)$ with the property that, for each pair of indices $j,k \in \{1,\dots,n\}$, there exists $f_{jk} \in L^p(\partial\Omega)$ such that

$$\int_{\partial\Omega} \varphi f_{jk}\,\mathrm{d}\sigma = -\int_{\partial\Omega} (\partial_{\tau_{jk}}\varphi) f\,\mathrm{d}\sigma, \quad \forall\,\varphi \in \mathscr{C}^1_0(\mathbb{R}^n). \tag{36.7}$$

Such functions $f_{jk}$ are called tangential derivatives of $f \in L^p_1(\partial\Omega)$ and denoted by $\partial_{\tau_{jk}} f$. The Euclidean tangential gradient of $f \in L^p_1(\partial\Omega)$ is defined by

$$\nabla_{\tan} f := \big(\nu_k\,\partial_{\tau_{kj}} f\big)_{1\le j\le n}. \tag{36.8}$$

Finally, given a suitable subdomain $\Omega$ of the manifold $\mathscr{M}$, the corresponding Sobolev space $L^p_1(\partial\Omega)$, $1 < p < \infty$, is defined as the collections of functions which locally belong to the Euclidean version of this space (discussed above).

We are now in a position to formulate the boundary value problem which is the subject of this article. The Laplace–Beltrami operator is the second-order differential operator defined by

$$\Delta := \operatorname{div}\operatorname{grad}. \tag{36.9}$$

Given $u:\Omega\to\mathbb{C}$, the local coordinate expression for $\Delta u$ is

$$\Delta u=\frac{1}{\sqrt{g}}\,\partial_j\big(\sqrt{g}\,g^{jk}\partial_k u\big). \tag{36.10}$$

For each $p\in(1,\infty)$, the regularity problem $(R_p)$ for $\Delta$ is the task of finding a unique function $u\in\mathscr{C}^1(\Omega)$ that satisfies

$$(R_p)\begin{cases}\Delta u=0 \quad \text{in } \Omega,\\ \mathscr{N}u,\mathscr{N}(\nabla u)\in L^p(\partial\Omega),\\ u\big|^{\rm n.t.}_{\partial\Omega}=f\in L^p_1(\partial\Omega),\end{cases} \tag{36.11}$$

where $\nabla$ is the Levi–Civita connection associated with the metric $g$.

The main result in this paper, Theorem 1, pertains to the unique solvability of $(R_p)$. This complements work done in [HoMiTa10], where the Dirichlet problem has been considered. For a proof of Theorem 1, the reader is referred to §36.3.

**Theorem 1.** *The regularity problem $(R_p)$ is well-posed for each $p\in(1,\infty)$. Moreover, given $f\in L^p_1(\partial\Omega)$, the solution to $(R_p)$ is given by*

$$u:=\mathscr{D}\big[\big(\tfrac12 I+K\big)^{-1}f\big] \quad \text{in } \Omega, \tag{36.12}$$

*and satisfies*

$$\|\mathscr{N}u\|_{L^p(\partial\Omega)}+\|\mathscr{N}(\nabla u)\|_{L^p(\partial\Omega)}\lesssim\|f\|_{L^p_1(\partial\Omega)}. \tag{36.13}$$

## 36.2 Layer Potential Method

Assume $V\in L^\infty(\mathscr{M})$ is such that $V\geq 0$, $V\neq 0$ on $\mathscr{M}$, and $V=0$ near $\overline{\Omega}$. Set $L:=\Delta-V$. Then $L$ is invertible in the sense of pseudo-differential operator theory with inverse $L^{-1}\in OPS^{-2}_{\rm cl}$, where $L^{-1}:W^{-1,2}(\mathscr{M})\to W^{1,2}(\mathscr{M})$; see [MiTa99] for details. Let $E$ denote the Schwartz kernel of $L^{-1}$, i.e.,

$$(L^{-1}f)(x)=\int_{\mathscr{M}}E(x,y)f(y)\,{\rm dVol}(y),\quad x\in\mathscr{M},f\in W^{-1,2}(\mathscr{M}), \tag{36.14}$$

with

$$E\in\mathscr{D}'(\mathscr{M}\times\mathscr{M})\cap\mathscr{C}^\gamma(\mathscr{M}\times\mathscr{M}\setminus{\rm diag}),\quad \gamma<2. \tag{36.15}$$

The local coordinate expression for $E(x,y)$ is given by (cf. [MiEtAl14], p. 41)

$$E(x,y) = e_0(x,x-y) + e_1(x,y), \quad x,y \in \mathscr{M}, x \neq y. \tag{36.16}$$

The leading term $e_0(x,x-y)$ is $O(|x-y|)^{-(n-2)}$ and, for any $x,y \in \mathscr{M}$ with $x \neq y$, has the explicit form

$$e_0(x,x-y) = \frac{-1}{(n-2)\omega_{n-1}\sqrt{g(x)}} \Big[ g_{jk}(x)(x_j-y_j)(x_k-y_k) \Big]^{-\frac{n-2}{2}}. \tag{36.17}$$

A direct computation using the chain rule gives

$$\partial_{x_j}\partial_{y_m}\big[e_0(x,x-y)\big] = -\partial_{y_j}\partial_{y_m}\big[e_0(x,x-y)\big] + O\big(|x-y|^{-(n-1)}\big). \tag{36.18}$$

Another straightforward calculation using the chain rule yields

$$g^{\ell m}(x)\,\partial_{y_\ell}\partial_{y_m}\big[e_0(x,x-y)\big] \equiv 0. \tag{36.19}$$

The lower-order term $e_1(y,x)$ satisfies the following estimates. For each $\varepsilon > 0$ there is $C_\varepsilon \in (0,\infty)$ for which (cf. [MiEtAl14], p. 42)

$$|e_1(x,y)| \leq C_\varepsilon |x-y|^{-(n-3+\varepsilon)}, \tag{36.20}$$

$$|(\nabla_x e_1)(x,y)| + |(\nabla_y e_1)(x,y)| \leq C_\varepsilon |x-y|^{-(n-2+\varepsilon)}, \tag{36.21}$$

$$|(\nabla_x \nabla_y e_1)(x,y)| \leq C_\varepsilon |x-y|^{-(n-1+\varepsilon)}. \tag{36.22}$$

Let $p \in (1,\infty)$ and $f \in L^p(\partial\Omega)$. The double-layer potential is defined by

$$(\mathscr{D}f)(x) := \int_{\partial\Omega} \partial_{\nu(y)}\big[E(x,y)\big] f(y)\,\mathrm{d}\sigma(y), \quad x \in \Omega, \tag{36.23}$$

where $\partial_{\nu(y)}\big[E(x,y)\big]$ denotes the conormal derivative of $E$ with respect to $y$. The principal value double-layer potential is defined by

$$(Kf)(x) := \mathrm{P.V.} \int_{\partial\Omega} \partial_{\nu(y)}\big[E(x,y)\big] f(y)\,\mathrm{d}\sigma(y) \tag{36.24}$$

$$:= \lim_{\varepsilon \to 0^+} \int\limits_{\substack{y \in \partial\Omega \\ d(x,y) > \varepsilon}} \partial_{\nu(y)}\big[E(x,y)\big] f(y)\,\mathrm{d}\sigma(y), \quad x \in \partial\Omega, \tag{36.25}$$

where $d(x,y)$ is the geodesic distance defined in (36.2).

Below we collect some of the fundamental properties of the double-layer operator which have been established in [HoMiTa10] (cf., e.g., Theorem 5.6 on pp. 2770–2772, Corollary 3.28 on p. 2681, and Proposition 3.37 on p. 2680).

**Theorem 2.** *For each $p \in (1,\infty)$, the following statements hold:*

*(1) $K$ is compact (hence bounded) on $L^p(\partial\Omega)$ and bounded on $L^p_1(\partial\Omega)$.*
*(2) For any $f \in L^p(\partial\Omega)$,*

$$(\mathscr{D}f)\Big|^{\rm n.t.}_{\partial\Omega} = \big(\tfrac{1}{2}I + K\big)f \quad \sigma\text{-a.e. on } \partial\Omega. \tag{36.26}$$

*(3) There exists $C \in (0,\infty)$ such that*

$$\|\mathscr{N}(\mathscr{D}f)\|_{L^p(\partial\Omega)} \leq C\|f\|_{L^p(\partial\Omega)}, \quad \forall f \in L^p(\partial\Omega), \tag{36.27}$$

$$\big\|\mathscr{N}\big(\nabla\mathscr{D}f\big)\big\|_{L^p(\partial\Omega)} \leq C\|f\|_{L^p_1(\partial\Omega)}, \quad \forall f \in L^p_1(\partial\Omega). \tag{36.28}$$

*(4) For $f \in L^p_1(\partial\Omega)$, at $\sigma$-a.e. point on $\partial\Omega$, the trace $\big(\nabla\mathscr{D}f\big)\big|^{\rm n.t.}_{\partial\Omega}$ exists and*

$$\partial_{\tau_{jk}}(\mathscr{D}f)\Big|^{\rm n.t.}_{\partial\Omega} = \nu_j\big(\partial_k\mathscr{D}f\big)\Big|^{\rm n.t.}_{\partial\Omega} - \nu_k\big(\partial_j\mathscr{D}f\big)\Big|^{\rm n.t.}_{\partial\Omega}. \tag{36.29}$$

We are now ready to state and prove the main result of this section.

**Theorem 3.** *For each $p \in (1,\infty)$,*

$$K : L^p_1(\partial\Omega) \longrightarrow L^p_1(\partial\Omega) \quad \textit{is compact.} \tag{36.30}$$

*Proof.* Fix $p \in (1,\infty)$. To prove (36.30), we work locally in a small coordinate patch $U$ of an arbitrary fixed point $x_0 \in \partial\Omega$ and use the decomposition (36.57).

A few conventions about notation used throughout this proof are in order. Whenever convenient, we shall identify the portion of $\Omega$ contained in $U$ with its Euclidean image under the coordinate chart. In this scenario, we denote by $\nu^{\rm E} = (\nu^{\rm E}_j)_{1\leq j\leq n}$ the outward unit normal to $\partial\Omega$ with respect to the Euclidean metric in $\mathbb{R}^n$, and let $\sigma^{\rm E} := \mathscr{H}^{n-1}_{\mathbb{R}^n}\lfloor\partial\Omega$ be the surface measure induced by the flat Euclidean metric $\delta_{jk}$ on $\partial\Omega$. These are related to the manifold conormal $\nu$ and surface measure $\sigma$ (associated with the original Riemannian metric $g$ on $\mathscr{M}$) via explicit formulas (cf. [HoMiTa10], p. 2771)

$$\nu^{\rm E}_j(y) = \sqrt{G\big(y,\nu^{\rm E}(y)\big)}\,\nu_j(y), \quad \forall j \in \{1,\ldots,n\}, \tag{36.31}$$

$$\mathrm{d}\sigma^{\rm E}(y) = \Big[g(y)\,G\big(y,\nu^{\rm E}(y)\big)\Big]^{-\frac{1}{2}}\mathrm{d}\sigma(y), \tag{36.32}$$

where

$$G(y,\xi) := g^{jk}(y)\,\xi_j\xi_k, \quad y \in \partial\Omega,\ \xi \in \mathbb{R}^n. \tag{36.33}$$

The conormal derivative $\partial_{\nu(y)}$ is related to partial derivatives $\partial_{y_m}$ by

$$\partial_{\nu(y)} = \frac{g^{\ell m}(y)\,\nu_\ell^{\rm E}(y)}{\sqrt{G\big(y,\nu^{\rm E}(y)\big)}}\,\partial_{y_m}. \tag{36.34}$$

Now assume $f \in L^p_1(\partial\Omega)$ is supported in $\partial\Omega \cap U$. Using (36.31)-(36.32) and (36.34), for each $x \in \Omega \cap U$ we may write

$$(\mathscr{D}f)(x) = \int_{\partial\Omega} g^{\ell m}(y)\,\nu_\ell^{\rm E}(y)\,\partial_{y_m}\big[E(x,y)\big]f(y)\,\sqrt{g(y)}\,{\rm d}\sigma^{\rm E}(y). \tag{36.35}$$

Now fix an arbitrary point $z \in \partial\Omega \cap U$. Then for each $x \in \Omega \cap U$ split

$$(\mathscr{D}f)(x) = A_1 + f(z)A_2, \tag{36.36}$$

where

$$A_1 := \int_{\partial\Omega} g^{\ell m}(y)\,\nu_\ell^{\rm E}(y)\,\partial_{y_m}\big[E(x,y)\big]\big[f(y)-f(z)\big]\,\sqrt{g(y)}\,{\rm d}\sigma^{\rm E}(y), \tag{36.37}$$

$$A_2 := \int_{\partial\Omega} g^{\ell m}(y)\,\nu_\ell^{\rm E}(y)\,\partial_{y_m}\big[E(x,y)\big]\,\sqrt{g(y)}\,{\rm d}\sigma^{\rm E}(y). \tag{36.38}$$

Use the Divergence Theorem to write $A_2 = B_1 + B_2$, where

$$B_1 := \int_\Omega \partial_{y_\ell}\Big\{g^{\ell m}(y)\,\partial_{y_m}\big[E(x,y)\big]\Big\}\sqrt{g(y)}\,{\rm d}y, \tag{36.39}$$

$$B_2 := \int_\Omega g^{\ell m}(y)\,\partial_{y_m}\big[E(x,y)\big]\,\partial_{y_\ell}\Big\{\sqrt{g(y)}\Big\}\,{\rm d}y. \tag{36.40}$$

From (36.10) we see that

$$\begin{aligned}
&\partial_{y_\ell}\Big\{g^{\ell m}(y)\,\partial_{y_m}\big[E(x,y)\big]\Big\}\\
&\quad = \Delta_y\big[E(x,y)\big] - \frac{1}{\sqrt{g(y)}}\,g^{\ell m}(y)\,\partial_{y_\ell}\Big\{\sqrt{g(y)}\Big\}\,\partial_{y_m}\big[E(x,y)\big]\\
&\quad = {\rm Dirac}_x(y) - \frac{1}{\sqrt{g(y)}}\,g^{\ell m}(y)\,\partial_{y_m}\big[E(x,y)\big]\,\partial_{y_\ell}\Big\{\sqrt{g(y)}\Big\}.
\end{aligned} \tag{36.41}$$

Thus, by (36.39)-(36.40) and (36.41) it follows that $A_2=\sqrt{g(x)}$. Given any index $j\in\{1,\dots,n\}$, at each $x\in\Omega\cap U$ write

$$(\partial_{x_j}\mathscr{D}f)(x)=\mathrm{I}_j(x,z)+\mathrm{II}_j(x,z), \tag{36.42}$$

where

$$\begin{aligned}\mathrm{I}_j(x,z)&:=\int_{\partial\Omega}g^{\ell m}(y)\,\nu_\ell^{\mathrm{E}}(y)\,\partial_{x_j}\partial_{y_m}\big[E(x,y)\big]\big[f(y)-f(z)\big]\sqrt{g(y)}\,\mathrm{d}\sigma^{\mathrm{E}}(y),\\ \mathrm{II}_j(x,z)&:=f(z)\,\partial_{x_j}\Big\{\sqrt{g(x)}\Big\}.\end{aligned} \tag{36.43}$$

To handle $\mathrm{I}_j(x,z)$, first observe from (36.16),(36.17), (36.18), (36.20), and (36.21) that for each $j,m\in\{1,\dots,n\}$ it follows that

$$\begin{aligned}\partial_{x_j}\partial_{y_m}\big[E(x,y)\big]&=\partial_{x_j}\partial_{y_m}\big[e_0(x,x-y)\big]+R_{jm}(x,y)\\&=-\partial_{y_j}\partial_{y_m}\big[e_0(x,x-y)\big]+R_{jm}(x,y),\end{aligned} \tag{36.44}$$

where $R_{jm}(x,y)$ are residual terms (changing from line to line) that satisfy

$$R_{jm}(x,y)=O\big(|x-y|^{-(n-1+\varepsilon)}\big),\quad\forall\,\varepsilon>0. \tag{36.45}$$

Use (36.44) to re-write $\mathrm{I}_j(x,z)$ as

$$\begin{aligned}\mathrm{I}_j(x,z)&=-\int_{\partial\Omega}g^{\ell m}(y)\,\nu_\ell^{\mathrm{E}}(y)\,\partial_{y_j}\partial_{y_m}\big[e_0(x,x-y)\big]\big[f(y)-f(z)\big]\sqrt{g(y)}\,\mathrm{d}\sigma^{\mathrm{E}}(y)\\&\quad+\int_{\partial\Omega}g^{\ell m}(y)\,\nu_\ell^{\mathrm{E}}(y)\,R_{jm}(x,y)\big[f(y)-f(z)\big]\sqrt{g(y)}\,\mathrm{d}\sigma^{\mathrm{E}}(y)\\&=:\mathrm{I}_j^{(1)}(x,z)+\mathrm{I}_j^{(2)}(x,z).\end{aligned} \tag{36.46}$$

Since $-\nu_\ell^{\mathrm{E}}(y)\partial_{y_j}=\partial_{\tau_{j\ell}(y)}-\nu_j^{\mathrm{E}}(y)\partial_{y_\ell}$, we may further express $\mathrm{I}_j^{(1)}(x,z)$ as

$$\begin{aligned}\mathrm{I}_j^{(1)}(x,z)&=\int_{\partial\Omega}g^{\ell m}(y)\,\partial_{\tau_{j\ell}(y)}\partial_{y_m}\big[e_0(x,x-y)\big]\big[f(y)-f(z)\big]\sqrt{g(y)}\,\mathrm{d}\sigma^{\mathrm{E}}(y)\\&\quad-\int_{\partial\Omega}g^{\ell m}(y)\,\nu_j^{\mathrm{E}}(y)\,\partial_{y_\ell}\partial_{y_m}\big[e_0(x,x-y)\big]\big[f(y)-f(z)\big]\sqrt{g(y)}\,\mathrm{d}\sigma^{\mathrm{E}}(y)\\&=\int_{\partial\Omega}g^{\ell m}(y)\,\partial_{\tau_{j\ell}(y)}\partial_{y_m}\big[e_0(x,x-y)\big]\big[f(y)-f(z)\big]\sqrt{g(y)}\,\mathrm{d}\sigma^{\mathrm{E}}(y),\end{aligned} \tag{36.47}$$

making use of the cancelation property recorded in (36.19). To handle the integral in (36.47), integrate by parts on $\partial\Omega$ in order to relocate the boundary tangential derivative operator $\partial_{\tau_{j\ell}(y)}$ away from $\partial_{y_m}\big[e_0(y,x-y)\big]$, thus obtaining

$$\begin{aligned}
\mathrm{I}_j^{(1)}(x,z) &= \int_{\partial\Omega} \partial_{y_m}\big[e_0(x,x-y)\big]\,\partial_{\tau_{\ell j}(y)}\Big\{g^{\ell m}(y)\big[f(y)-f(z)\big]\sqrt{g(y)}\Big\}\,\mathrm{d}\sigma^{\mathrm{E}}(y)\\
&= \int_{\partial\Omega} \partial_{y_m}\big[e_0(x,x-y)\big]\,\partial_{\tau_{\ell j}}\big(g^{\ell m}\sqrt{g}\big)(y)\big[f(y)-f(z)\big]\,\mathrm{d}\sigma^{\mathrm{E}}(y)\\
&\quad+\int_{\partial\Omega} g^{\ell m}(y)\,\partial_{y_m}\big[e_0(x,x-y)\big]\big(\partial_{\tau_{\ell j}}f\big)(y)\sqrt{g(y)}\,\mathrm{d}\sigma^{\mathrm{E}}(y)\\
&=: \mathrm{I}_j^{(11)}(x,z)+\mathrm{I}_j^{(12)}(x).
\end{aligned}\tag{36.48}$$

The idea for handling the term $\mathrm{I}_j^{(12)}(x)$ is to reverse-engineer $E(x,y)$ starting from $e_0(x,x-y)$. Indeed, invoking (36.16) and (36.21), for each $m\in\{1,\dots,n\}$ we see that

$$\partial_{y_m}\big[e_0(x,x-y)\big]=\partial_{y_m}\big[E(x,y)\big]+Q_m(x,y),\tag{36.49}$$

where the residual term $Q_m(x,y)$ satisfies

$$Q_m(x,y)=O\big(|x-y|^{-(n-2+\varepsilon)}\big),\quad\forall\,\varepsilon>0.\tag{36.50}$$

We may then use (36.49) to rewrite $\mathrm{I}_j^{(12)}(x)$ as

$$\begin{aligned}
\mathrm{I}_j^{(12)}(x) &= \int_{\partial\Omega} g^{\ell m}(y)\,\partial_{y_m}\big[E(x,y)\big]\big(\partial_{\tau_{\ell j}}f\big)(y)\sqrt{g(y)}\,\mathrm{d}\sigma^{\mathrm{E}}(y)\\
&\quad+\int_{\partial\Omega} g^{\ell m}(y)Q_m(x,y)\big(\partial_{\tau_{\ell j}}f\big)(y)\sqrt{g(y)}\,\mathrm{d}\sigma^{\mathrm{E}}(y)\\
&=: \mathrm{I}_j^{(121)}(x)+\mathrm{I}_j^{(122)}(x).
\end{aligned}\tag{36.51}$$

Upon observing that

$$\partial_{\tau_{\ell j}}f=\nu_\ell^{\mathrm{E}}\big(\nabla_{\tan}^{\mathrm{E}}f\big)_j-\nu_j^{\mathrm{E}}\big(\nabla_{\tan}^{\mathrm{E}}f\big)_\ell,\tag{36.52}$$

we may further recast $\mathrm{I}_j^{(121)}(x)$ in the form

$$\mathrm{I}_j^{(121)}(x)=\mathrm{I}_j^{(1211)}(x)-\mathrm{I}_j^{(1212)}(x),\tag{36.53}$$

where

$$\mathrm{I}_j^{(1211)}(x) := \int_{\partial\Omega} \nu_\ell^{\mathrm{E}}(y) g^{\ell m}(y)\, \partial_{y_m}\big[E(x,y)\big] \big(\nabla_{\tan}^{\mathrm{E}} f\big)_j(y) \sqrt{g(y)}\, \mathrm{d}\sigma^{\mathrm{E}}(y), \tag{36.54}$$

$$\mathrm{I}_j^{(1212)}(x) := \int_{\partial\Omega} \nu_j^{\mathrm{E}}(y) g^{\ell m}(y)\, \partial_{y_m}\big[E(x,y)\big] \big(\nabla_{\tan}^{\mathrm{E}} f\big)_\ell(y) \sqrt{g(y)}\, \mathrm{d}\sigma^{\mathrm{E}}(y). \tag{36.55}$$

Finally, from (36.54) and (36.35) we recognize that

$$\mathrm{I}_j^{(1211)}(x) = \mathscr{D}(\nabla_{\tan}^{\mathrm{E}} f)_j(x). \tag{36.56}$$

Thus, given $j \in \{1,\dots,n\}$, at each $x \in \Omega \cap U$ we have the decomposition

$$\begin{aligned}(\partial_{x_j}\mathscr{D}f)(x) &= \mathscr{D}(\nabla_{\tan}^{\mathrm{E}} f)_j(x) - \mathrm{I}_j^{(1212)}(x) + \mathrm{I}_j^{(122)}(x) + \mathrm{I}_j^{(11)}(x,z) \\ &\quad + \mathrm{I}_j^{(2)}(x,z) + \mathrm{II}_j(x,z).\end{aligned} \tag{36.57}$$

This decomposition will be important for proving that $K : L_1^p(\partial\Omega) \to L_1^p(\partial\Omega)$ is compact, as shown below.

Continuing to work in Euclidean coordinates, for any $j,k \in \{1,\dots,n\}$, first write

$$\begin{aligned}\partial_{\tau_{jk}}(Kf)(z) &= \partial_{\tau_{jk}}(\tfrac{1}{2}f + Kf)(z) - \tfrac{1}{2}(\partial_{\tau_{jk}} f)(z) \\ &= \nu_j^{\mathrm{E}}(z)\big(\partial_k \mathscr{D}f\big)\Big|_{\partial\Omega}^{\mathrm{n.t.}}(z) - \nu_k^{\mathrm{E}}(z)\big(\partial_j \mathscr{D}f\big)\Big|_{\partial\Omega}^{\mathrm{n.t.}}(z) \\ &\quad - \tfrac{1}{2}(\partial_{\tau_{jk}} f)(z),\end{aligned} \tag{36.58}$$

where we have made use of (36.27), (36.26), (36.28), (36.29), and Theorem 13.3 in [MiEtAl14]. Second, observe that from (36.26) and

$$\partial_{\tau_{jk}} f = \nu_j^{\mathrm{E}}(\nabla_{\tan}^{\mathrm{E}} f)_k - \nu_k^{\mathrm{E}}(\nabla_{\tan}^{\mathrm{E}} f)_j, \tag{36.59}$$

we have

$$\begin{aligned}&\nu_j^{\mathrm{E}}(z)\,\big(\mathscr{D}(\nabla_{\tan}^{\mathrm{E}} f)_k\big)\Big|_{\partial\Omega}^{\mathrm{n.t.}}(z) - \nu_k^{\mathrm{E}}(z)\,\big(\mathscr{D}(\nabla_{\tan}^{\mathrm{E}} f)_j\big)\Big|_{\partial\Omega}^{\mathrm{n.t.}}(z) \\ &= \nu_j^{\mathrm{E}}(z)\Big(\tfrac{1}{2}(\nabla_{\tan}^{\mathrm{E}} f)_k + K(\nabla_{\tan}^{\mathrm{E}} f)_k\Big)(z) - \nu_k^{\mathrm{E}}(z)\Big(\tfrac{1}{2}(\nabla_{\tan}^{\mathrm{E}} f)_j + K(\nabla_{\tan}^{\mathrm{E}} f)_j\Big)(z) \\ &= \tfrac{1}{2}\big\{\nu_j^{\mathrm{E}}(\nabla_{\tan}^{\mathrm{E}} f)_k - \nu_k^{\mathrm{E}}(\nabla_{\tan}^{\mathrm{E}} f)_j\big\}(z) \\ &\qquad + \nu_j^{\mathrm{E}}(z)\big(K(\nabla_{\tan}^{\mathrm{E}} f)_k\big)(z) - \nu_k^{\mathrm{E}}(z)\big(K(\nabla_{\tan}^{\mathrm{E}} f)_j\big)(z)\end{aligned}$$

$$
\begin{aligned}
&= \tfrac{1}{2}(\partial_{\tau_{jk}} f)(z) + \big(K\big(\nu_j^{\rm E}(\nabla_{\rm tan}^{\rm E} f)_k\big)\big)(z) - \big(K\big(\nu_k^{\rm E}(\nabla_{\rm tan}^{\rm E} f)_j\big)\big)(z) \\
&\qquad + \Big(\big[M_{\nu_j^{\rm E}}, K\big](\nabla_{\rm tan}^{\rm E} f)_k\Big)(z) - \Big(\big[M_{\nu_k^{\rm E}}, K\big](\nabla_{\rm tan}^{\rm E} f)_j\Big)(z) \\
&= \tfrac{1}{2}(\partial_{\tau_{jk}} f)(z) + \big(K(\partial_{\tau_{jk}} f)\big)(z) \\
&\qquad + \Big(\big[M_{\nu_j^{\rm E}}, K_l\big](\nabla_{\rm tan}^{\rm E} f)_k\Big)(z) - \Big(\big[M_{\nu_k^{\rm E}}, K\big](\nabla_{\rm tan}^{\rm E} f)_j\Big)(z).
\end{aligned}
\tag{36.60}
$$

The above formula is relevant when assessing the contribution from $\mathrm{I}_j^{(1211)}(x)$, written as in (36.56), in the context of

$$
\nu_j^{\rm E}(z)\big(\partial_k \mathscr{D} f\big)\Big|_{\partial\Omega}^{\rm n.t.}(z) - \nu_k^{\rm E}(z)\big(\partial_j \mathscr{D} f\big)\Big|_{\partial\Omega}^{\rm n.t.}(z). \tag{36.61}
$$

Specifically, starting with (36.58), then recalling the decomposition (36.42) along with the subsequent analysis of the intervening pieces, for any given pair of indices $j,k \in \{1,\dots,n\}$, at $\sigma$-a.e. point $z \in \partial\Omega$ we obtain

$$
\begin{aligned}
\partial_{\tau_{jk}}(Kf)(z) &= \big(K(\partial_{\tau_{jk}} f)\big)(z) \\
&\quad + \Big(\big[M_{\nu_j^{\rm E}}, K\big](\nabla_{\rm tan}^{\rm E} f)_k\Big)(z) - \Big(\big[M_{\nu_k^{\rm E}}, K\big](\nabla_{\rm tan}^{\rm E} f)_j\Big)(z) \\
&\quad - \nu_j^{\rm E}(z)\,\mathrm{I}_k^{(1212)}\Big|_{\partial\Omega}^{\rm n.t.}(z) + \nu_k^{\rm E}(z)\,\mathrm{I}_j^{(1212)}\Big|_{\partial\Omega}^{\rm n.t.}(z) \\
&\quad + \nu_j^{\rm E}(z)\,\mathrm{I}_k^{(122)}(\cdot,z)\Big|_{\partial\Omega}^{\rm n.t.}(z) - \nu_k^{\rm E}(z)\,\mathrm{I}_j^{(122)}(\cdot,z)\Big|_{\partial\Omega}^{\rm n.t.}(z) \\
&\quad + \nu_j^{\rm E}(z)\,\mathrm{I}_k^{(11)}(\cdot,z)\Big|_{\partial\Omega}^{\rm n.t.}(z) - \nu_k^{\rm E}(z)\,\mathrm{I}_j^{(11)}(\cdot,z)\Big|_{\partial\Omega}^{\rm n.t.}(z) \\
&\quad + \nu_j^{\rm E}(z)\,\mathrm{I}_k^{(2)}(\cdot,z)\Big|_{\partial\Omega}^{\rm n.t.}(z) - \nu_k^{\rm E}(z)\,\mathrm{I}_j^{(2)}(\cdot,z)\Big|_{\partial\Omega}^{\rm n.t.}(z) \\
&\quad + \nu_j^{\rm E}(z)\,\mathrm{II}_k(\cdot,z)\Big|_{\partial\Omega}^{\rm n.t.}(z) - \nu_k^{\rm E}(z)\,\mathrm{II}_j(\cdot,z)\Big|_{\partial\Omega}^{\rm n.t.}(z).
\end{aligned}
\tag{36.62}
$$

The compactness of $K$ on $L_1^p(\partial\Omega)$ will be read off the decomposition in (36.62), considering each line individually. Specifically, the idea is to ensure that

$$
\partial_{\tau_{jk}}(Kf) = K(\partial_{\tau_{jk}} f) + C_{jk} f, \tag{36.63}
$$

for a compact operator $C_{jk} : L_1^p(\partial\Omega) \to L^p(\partial\Omega)$, from which the compactness of $K$ on $L_1^p(\partial\Omega)$ follows from standard functional analysis.

To implement this strategy, first observe that the assignment

$$L^p_1(\partial\Omega)\ni f\longmapsto K(\partial_{\tau_{jk}}f)\in L^p(\partial\Omega)\quad\text{is compact,}\tag{36.64}$$

thanks to part *(1)* of Theorem 2 and the fact that, for each $j,k\in\{1,\dots,n\}$,

$$\partial_{\tau_{jk}}: L^p_1(\partial\Omega)\longrightarrow L^p(\partial\Omega)\quad\text{is bounded.}\tag{36.65}$$

Next, (36.65) and Theorem 4 imply that for each $j,k\in\{1,\dots,n\}$ the mapping

$$L^p_1(\partial\Omega)\ni f\longmapsto\big[M_{\nu^{\rm E}_j},K\big](\nabla^{\rm E}_{\rm tan}f)_k\in L^p(\partial\Omega)\quad\text{is compact.}\tag{36.66}$$

This takes care of the second line of (36.62).

Regarding the third line of (36.62), note that from (36.55) we obtain (bearing in mind that the jump-terms produced after taking non-tangential boundary traces, within the context of formula (2.73) on p. 21 of [MiMiTa01], actually cancel in the combination considered in (36.67) below)

$$\begin{aligned}
&\nu^{\rm E}_j(z)\,{\rm I}^{(1212)}_k\Big|^{\rm n.t.}_{\partial\Omega}(z)-\nu^{\rm E}_k(z)\,{\rm I}^{(1212)}_j\Big|^{\rm n.t.}_{\partial\Omega}(z)\\
&={\rm P.V.}\int_{\partial\Omega}g^{\ell m}(y)\,\partial_{y_m}\big[E(x,y)\big]\,\nu^{\rm E}_j(z)\,\nu^{\rm E}_k(y)\big(\nabla^{\rm E}_{\rm tan}f\big)_\ell(y)\sqrt{g(y)}\,{\rm d}\sigma^{\rm E}(y)\\
&\quad-{\rm P.V.}\int_{\partial\Omega}g^{\ell m}(y)\,\partial_{y_m}\big[E(x,y)\big]\,\nu^{\rm E}_k(z)\,\nu^{\rm E}_j(y)\big(\nabla^{\rm E}_{\rm tan}f\big)_\ell(y)\sqrt{g(y)}\,{\rm d}\sigma^{\rm E}(y)\\
&={\rm P.V.}\int_{\partial\Omega}g^{\ell m}(y)\,\partial_{y_m}\big[E(x,y)\big]\,\nu^{\rm E}_j(z)\big[\nu^{\rm E}_k(y)-\nu^{\rm E}_k(z)\big]\times\\
&\qquad\qquad\times\big(\nabla^{\rm E}_{\rm tan}f\big)_\ell(y)\sqrt{g(y)}\,{\rm d}\sigma^{\rm E}(y)\\
&\quad-{\rm P.V.}\int_{\partial\Omega}g^{\ell m}(y)\,\partial_{y_m}\big[E(x,y)\big]\,\nu^{\rm E}_k(z)\big[\nu^{\rm E}_j(y)-\nu^{\rm E}_j(z)\big]\times\\
&\qquad\qquad\times\big(\nabla^{\rm E}_{\rm tan}f\big)_\ell(y)\sqrt{g(y)}\,{\rm d}\sigma^{\rm E}(y).
\end{aligned}\tag{36.67}$$

Each of the principal value singular integral operators on the right-hand side of (36.67) are of commutator-type, and thus amenable to treatment by Theorem 4. Mindful of (36.65), we conclude that the fourth line in (36.62) is also of a nature which is in line with the goal of verifying that (36.63) holds.

Next, recall ${\rm I}^{(122)}_{I,j}(x,z)$ defined in (36.51). Given the weakly singular nature of the integral kernel in ${\rm I}^{(122)}_{I,j}(\cdot,z)\big|^{\rm n.t.}_{\partial\Omega}$ (cf. (36.50)), it follows from Theorem 5 that for each $j,k\in\{1,\dots,n\}$ the mapping

$$L^p_1(\partial\Omega)\ni f\longmapsto\nu^{\rm E}_j(z)\,{\rm I}^{(122)}_k(\cdot,z)\Big|^{\rm n.t.}_{\partial\Omega}(z)\in L^p_z(\partial\Omega)\quad\text{is compact.}\tag{36.68}$$

Thus, the fourth line of (36.62) is also of the right nature according to (36.63).

Going further, recall $\mathrm{I}^{(11)}_{l,j}(x,z)$ from (36.48). In relation to this, Theorem 6 applies (with $\varepsilon = 0$) and gives that, for each $j,k \in \{1,\dots,n\}$, the assignment

$$L^p_1(\partial\Omega) \ni f \longmapsto \nu^{\mathrm{E}}_j(z)\,\mathrm{I}^{(11)}_k(\cdot,z)\Big|^{\mathrm{n.t.}}_{\partial\Omega}(z) \in L^p_z(\partial\Omega) \quad \text{is compact,} \tag{36.69}$$

taking care of the fifth line in (36.62). Now recall $\mathrm{I}^{(2)}_{l,j}(x,z)$ from (36.46). Aware of (36.45) and making use of the full strength of Theorem 6, we also have that for each $j,k \in \{1,\dots,n\}$ the operator

$$L^p_1(\partial\Omega) \ni f \longmapsto \nu^{\mathrm{E}}_j(z)\,\mathrm{I}^{(2)}_k(\cdot,z)\Big|^{\mathrm{n.t.}}_{\partial\Omega}(z) \in L^p_z(\partial\Omega) \quad \text{is compact.} \tag{36.70}$$

This clarifies matters with regard to the sixth line in (36.62). The nature of the seventh line in (36.62), involving $\mathrm{II}_j(x,z)$ originally defined in (36.43), is easily elucidated by relying on the fact that the following inclusions are well-defined, linear, and compact (cf. [MiEtAl14]):

$$L^p_1(\partial\Omega) \hookrightarrow L^r(\partial\Omega) \quad \text{if } 0 < r < \Big(\max\Big\{0, \tfrac{1}{p} - \tfrac{1}{n-1}\Big\}\Big)^{-1}, \tag{36.71}$$

$$L^p_1(\partial\Omega) \hookrightarrow L^\infty(\partial\Omega) \ \text{ if } n-1 < p < \infty. \tag{36.72}$$

With all lines of (36.62) accounted for according to (36.63), the final conclusion is that $K$ is compact on $L^p_1(\partial\Omega)$. This analysis yields (36.30), as desired.

## 36.3 The Proof of the Main Well-Posedness Result

This section is devoted to presenting the proof of Theorem 1. Existence follows from Theorem 3 and Fredholm theory. Indeed, since $K : L^p_1(\partial\Omega) \to L^p_1(\partial\Omega)$ is compact, it follows that $\frac{1}{2}I + K : L^p_1(\partial\Omega) \to L^p_1(\partial\Omega)$ is Fredholm of index zero. Moreover, the latter operator has a trivial null-space since in [HoMiTa10] it has been shown that $\frac{1}{2}I + K : L^p(\partial\Omega) \to L^p(\partial\Omega)$ is injective. Ultimately, this implies that $\frac{1}{2}I + K$ is invertible on $L^p_1(\partial\Omega)$. Granted this, given $f \in L^p_1(\partial\Omega)$, the function

$$u := \mathscr{D}\big[\big(\tfrac{1}{2}I + K\big)^{-1} f\big] \quad \text{in } \Omega \tag{36.73}$$

is a well-defined solution of $(R_p)$. Moreover, $u \in \mathscr{C}^\gamma(\Omega)$ for every $\gamma < 2$, by elliptic regularity. Furthermore, Theorem 2 provides estimate (36.13) (which, in particular, ensures that the solution $u$ depends continuously on the boundary datum $f$).

To prove that $u$ is unique, one may use the argument in Step 3 of the proof of Theorem 7.2 on pp. 2832–2837 in [HoMiTa10] by constructing a Green's function of the form

$$G(x,y) := E(x,y) - \mathscr{D}\Big(\big(\tfrac{1}{2}I+K\big)^{-1}\Big(E(x,\cdot)\Big|^{\text{n.t.}}_{\partial\Omega}\Big)\Big)(y), \quad x,y\in\Omega. \tag{36.74}$$

Altogether, this shows that $(R_p)$ is well-posed.

## 36.4 Auxiliary Results

**Theorem 4.** *Let $\Omega\subset\mathbb{R}^n$ be a regular* SKT *domain with compact boundary. Set $\sigma := \mathscr{H}^{n-1}\lfloor\partial\Omega$. Let $M(n)\in\mathbb{N}$ and $b:\mathbb{R}^n\times\big(\mathbb{R}^n\setminus\{0\}\big)\to\mathbb{R}$ be such that*

$$\begin{aligned} &b(x,z) \text{ is odd and positive homogeneous of degree } 1-n \text{ in } z,\\ &\text{while } (\partial_z^\alpha b)(x,z) \text{ is continuous and bounded on } \mathbb{R}^n\times S^{n-1}\\ &\text{for every multiindex } \alpha\in\mathbb{N}_0^n \text{ with length } |\alpha|\le M(n). \end{aligned} \tag{36.75}$$

*Define the integral operators $T$ on functions $f:\partial\Omega\to\mathbb{C}$ by*

$$(Tf)(x) := \text{P.V.}\int_{\partial\Omega}\big[\nu(x)-\nu(y)\big]b(x,x-y)f(y)\,d\sigma(y), \quad x\in\partial\Omega. \tag{36.76}$$

*Then for each $p\in(1,\infty)$ the operator $T:L^p(\partial\Omega)\to L^p(\partial\Omega)$ is compact.*

*Proof.* See Theorem 13.9 on p. 167 in [MiEtAl14].

**Theorem 5.** *Let $\Omega\subset\mathbb{R}^n$ be a bounded domain with compact Ahlfors regular boundary. Set $\sigma := \mathscr{H}^{n-1}\lfloor\partial\Omega$. Suppose $k:\partial\Omega\times\partial\Omega\setminus\text{diag}\to\mathbb{R}$ is $\sigma$-measurable and has the property that there exists $\varepsilon>0$ and $C\in(0,\infty)$ such that*

$$|k(x,y)|\le\frac{C}{|x-y|^{n-1-\varepsilon}}, \quad \forall\, x,y\in\partial\Omega,\ x\neq y. \tag{36.77}$$

*Define the integral operator $T$ acting on functions $f:\partial\Omega\to\mathbb{C}$ by*

$$(Tf)(x) := \int_{\partial\Omega}k(x,y)f(y)\,d\sigma(y), \quad x\in\partial\Omega. \tag{36.78}$$

*Then for each $p\in(1,\infty)$, the operator $T:L^p(\partial\Omega)\to L^p(\partial\Omega)$ is compact.*

*Proof.* This follows from Lemma 2.20 on p. 2608 of [HoMiTa10].

**Theorem 6.** *Let $\Omega\subset\mathbb{R}^n$ be an open set satisfying a two-sided local John condition and whose boundary is compact and Ahlfors regular. Set $\sigma := \mathscr{H}^{n-1}\lfloor\partial\Omega$. Assume $k:\partial\Omega\times\partial\Omega\setminus\text{diag}\to\mathbb{C}$ is a kernel that satisfies*

$$|k(x,y)| = O\big(|x-y|^{-(n-1+\varepsilon)}\big) \quad \text{for some } \varepsilon\in[0,1). \tag{36.79}$$

*Then the integral operator $T$ defined on functions $f : \partial\Omega \to \mathbb{C}$ by*

$$(Tf)(x) := \int_{\partial\Omega} k(x,y)\big[f(y) - f(x)\big]\,\mathrm{d}\sigma(y), \quad x \in \partial\Omega, \tag{36.80}$$

*has the property that $T : L^p_1(\partial\Omega) \to L^p(\partial\Omega)$ is compact for each $p \in (1,\infty)$.*

*Proof.* See Lemma 13.10 in [MiEtAl14].

**Acknowledgements** The first-named author was partially supported by Simons Foundation grant #281566 and a University of Missouri Research Leave.

## References

[MiTa99] M. Mitrea and M. Taylor, *Boundary Layer Methods for Lipschitz Domains in Riemannian Manifolds*, Journal of Functional Analysis **163**, 181–251 (1999).

[MiMiTa01] D. Mitrea, M. Mitrea, and M. Taylor, *Layer Potentials, the Hodge Laplacian, and Global Boundary Problems in Nonsmooth Riemannian Manifolds*, Memoirs of the American Mathematical Society, March 2001, Volume 150, Number 713.

[HoMiTa10] S. Hofmann, M. Mitrea, and M. Taylor, *Singular Integrals and Elliptic Boundary Problems on Regular Semmes–Kenig–Toro Domains*, International Mathematics Research Notices, Vol. 2010, No. 14, pp. 2567–2865.

[MiEtAl14] D. Mitrea, I. Mitrea, M. Mitrea, and M. Taylor, *Boundary Problems for the Hodge–Laplacian on Regular Semmes–Kenig–Toro Subdomains of Riemannian Manifolds*, book manuscript, 2014.

# Chapter 37
# A Collocation Method Based on the Central Part Interpolation for Integral Equations

**K. Orav-Puurand, A. Pedas, and G. Vainikko**

## 37.1 Integral Equation and Smoothness of the Solution

Consider the integral equation

$$u(x) = \int_0^1 [a(x,y)|x-y|^{-\nu} + b(x,y)]u(y)dy + f(x), \quad 0 \le x \le 1, \quad 0 < \nu < 1, \tag{37.1}$$

where $f \in C[0,1] \cap C^m(0,1)$, $a, b \in C^m([0,1] \times (0,1))$, $m \in \mathbb{N} = \{1,2,...\}$. By $C^m(\Omega)$ is meant the set of all $m$ times continuously differentiable functions on $\Omega$. By $C[0,1]$ is meant the Banach space of continuous functions $u : [0,1] \to \mathbb{R} = (-\infty, \infty)$ with the usual norm $\|u\|_\infty = \{\max |u(x)| : 0 \le x \le 1\}$.

Denote by $T$ the integral operator of equation (37.1):

$$(Tu)(x) = \int_0^1 [a(x,y)|x-y|^{-\nu} + b(x,y)]u(y)dy \quad 0 \le x \le 1, \quad 0 < \nu < 1. \tag{37.2}$$

We refer to [PeVa06a] for the proofs of the following two lemmas.

**Lemma 1.** *Let $T$ be defined by (37.2) with a fixed $\nu \in (0,1)$. Let $\lambda_0,\ \lambda_1 \in \mathbb{R}$, $\lambda_0 + \nu < 1$, $\lambda_1 + \nu < 1$. Assume that $a, b \in C([0,1] \times (0,1))$ and*

K. Orav-Puurand (✉) • A. Pedas • G. Vainikko
University of Tartu, Estonia
e-mail: kerli.orav-puurand@ut.ee; arvet.pedas@ut.ee; gennadi.vainikko@ut.ee



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_37

$$|a(x,y)|+|b(x,y)|\le cy^{-\lambda_0}(1-y)^{-\lambda_1},\quad (x,y)\in[0,1]\times(0,1),$$

*where $c=c(a,b)$ is a positive constant.*

*Then $T$ maps $C[0,1]$ into $C[0,1]$ and $T:C[0,1]\to C[0,1]$ is compact.*

For $m\in\mathbb{N}$, $\theta_0,\theta_1\in\mathbb{R}$, $\theta_0<1$, $\theta_1<1$, denote by $C^{m,\theta_0,\theta_1}(0,1)$ the weighted space of functions $u\in C[0,1]\cap C^m(0,1)$ such that

$$|u|_{m,\theta_0,\theta_1}:=\sum_{k=1}^{m}\sup_{0<x<1}\omega_{k-1+\theta_0}(x)\omega_{k-1+\theta_1}(1-x)\left|u^{(k)}(x)\right|<\infty.$$

Here

$$\omega_\rho(r)=\left\{\begin{array}{ll} 1 & \text{for } \rho<0\\ \frac{1}{1+|\log r|} & \text{for } \rho=0\\ r^\rho & \text{for } \rho>0\end{array}\right\},\quad r,\ \rho\in\mathbb{R},\ r>0.$$

Equipped with the norm

$$\|u\|_{C^{m,\theta_0,\theta_1}(0,1)}:=\|u\|_\infty+|u|_{m,\theta_0,\theta_1},\quad u\in C^{m,\theta_0,\theta_1}(0,1),$$

$C^{m,\theta_0,\theta_1}(0,1)$ it is a Banach space. Clearly, $C^m[0,1]\subset C^{m,\theta_0,\theta_1}(0,1)$ for any $m\in\mathbb{N}$, $\theta_0<1$, $\theta_1<1$.

Denote $\partial_x^k\partial_y^l=\left(\frac{\partial}{\partial x}\right)^k\left(\frac{\partial}{\partial y}\right)^l$, $k,l\in\mathbb{N}_0=\{0\}\cup\mathbb{N}$.

**Lemma 2.** *Let $T$ be defined by (37.2) with $\nu\in(0,1)$. Let $m\in\mathbb{N}$, $\lambda_0,\lambda_1\in\mathbb{R}$, $\lambda_0+\nu<1$, $\lambda_1+\nu<1$. Assume that $a,b\in C^m([0,1]\times(0,1))$ and satisfy*

$$\left|\partial_x^k\partial_y^l a(x,y)\right|+\left|\partial_x^k\partial_y^l b(x,y)\right|\le cy^{-\lambda_0-l}(1-y)^{-\lambda_1-l},\quad (x,y)\in[0,1]\times(0,1),$$

*with a positive constant $c=c(a,b)$ for all $k,l\in\mathbb{N}_0$ such that $k+l\le m$.*

*Then operator $T$ maps $C^{m,\theta_0,\theta_1}(0,1)$ with $\theta_0=\lambda_0+\nu$ and $\theta_1=\lambda_1+\nu$ into $C^{m,\theta_0,\theta_1}(0,1)$ and $T:C^{m,\theta_0,\theta_1}(0,1)\to C^{m,\theta_0,\theta_1}(0,1)$ is compact.*

Denote

$$\mathcal{N}(I-T)=\{u\in C[0,1]:u=Tu\}.$$

The following theorem is a consequence of Lemmas 1 and 2.

**Theorem 1.** *Assume the conditions of Lemma 2 and $\mathcal{N}(I-T)=\{0\}$. Let $f\in C^{m,\theta_0,\theta_1}(0,1)$, $\theta_0=\lambda_0+\nu$, $\theta_1=\lambda_1+\nu$.*

*Then equation (37.1) has a solution $u \in C^{m,\theta_0,\theta_1}(0,1)$ which is unique in $C[0,1]$. In particular, it holds for $0 < \lambda_0 + \nu < 1$, $0 < \lambda_1 + \nu < 1$ that*

$$\left|u^{(k)}(x)\right| \le cx^{1-\nu-\lambda_0-k}(1-x)^{1-\nu-\lambda_1-k}, \quad 0 < x < 1, \quad k = 1,\dots,m,$$

*where $c = c(u) > 0$ is a constant.*

## 37.2 Smoothing Transformation

Possible boundary singularities of the solution $u \in C^{m,\lambda_0+\nu,\lambda_1+\nu}(0,1)$ of equation (37.1) are generic, they occur for most of free terms $f$ even if $f$ has no boundary singularities. To suppress the singularities of the solution we perform in equation (37.1) the change of variables (cf. [MoSc98, PeVa06b, VaVa08, OrVa09, OrPeVa10, Or13])

$$x = \varphi(t), \quad y = \varphi(s), \quad 0 \le t \le 1, \quad 0 \le s \le 1, \tag{37.3}$$

where $\varphi : [0,1] \to [0,1]$ is defined by the formula

$$\begin{gathered}\varphi(t) = \frac{1}{c_*}\int_0^t \sigma^{p_0-1}(1-\sigma)^{p_1-1}d\sigma, \quad 0 \le t \le 1, \quad p_0, p_1 \in \mathbb{N}, \\ c_* = \int_0^1 \sigma^{p_0-1}(1-\sigma)^{p_1-1}d\sigma = \frac{\Gamma(p_0)\Gamma(p_1)}{\Gamma(p_0+p_1)},\end{gathered} \tag{37.4}$$

where $\Gamma$ is the Euler gamma function.

If $p_0 = p_1 = 1$ then $\varphi(t) = t$ for $0 \le t \le 1$. We are interested in transformations (37.4) with $p_0 > 1$ or/and $p_1 > 1$ since then this transformation possesses a smoothing property for functions $u(x)$ with singularities of derivatives of $u(x)$ at $x = 0$ or/and $x = 1$ (see Lemma 3 below, the proof of it can be found in [VaVa08]).

**Lemma 3.** *Let $m \in \mathbb{N}$, $\theta_0, \theta_1 \in \mathbb{R}$, $\theta_0 < 1$, $\theta_1 < 1$. If $u \in C^{m,\theta_0,\theta_1}(0,1)$ and $v(t) = u(\varphi(t))$, with $\varphi$ defined by (37.4), then for $j = 1,\dots,m$, $0 < t < 1$,*

$$\left|v^{(j)}(t)\right| \le c\,\|u\|_{C^{m,\theta_0,\theta_1}(0,1)} \left\{\begin{array}{cc} t^{p_0-j}, & \theta_0 < 0 \\ t^{p_0-j}(1+|\log t|), & \theta_0 = 0 \\ t^{(1-\theta_0)p_0-j}, & \theta_0 > 0 \end{array}\right\} \times$$

$$\times \left\{\begin{array}{cc} (1-t)^{p_1-j}, & \theta_1 < 0 \\ (1-t)^{p_1-j}(1+|\log(1-t)|), & \theta_1 = 0 \\ (1-t)^{(1-\theta_1)p_1-j}, & \theta_1 > 0 \end{array}\right\}.$$

The following result is a consequence of Lemma 3.

**Theorem 2.** *Let* $m \in \mathbb{N}$, $0 < \nu < 1$, $\lambda_0, \lambda_0 \in \mathbb{R}$, $\lambda_0 < 1-\nu$, $\lambda_1 < 1-\nu$. *Let* $u \in C^{m,\nu+\lambda_0,\nu+\lambda_1}(0,1)$ *and* $v(t) = u(\varphi(t))$, *where* $\varphi$ *is defined by* (37.4) *with the parameters* $p_0, p_1 \in \mathbb{N}$ *satisfying*

$$p_i > \left\{ \begin{array}{ll} m & for \quad \lambda_i + \nu \le 0 \\ \frac{m}{1-\nu-\lambda_i} & for \quad 0 < \lambda_i + \nu < 1 \end{array} \right\}, \quad i = 0,\ i = 1. \tag{37.5}$$

*Then* $v \in C^m[0,1]$ *and*

$$v^{(j)}(0) = v^{(j)}(1) = 0, \quad j = 1, \dots, m. \tag{37.6}$$

It follows from (37.4) that $\varphi(0) = 0$, $\varphi(1) = 1$ and $\varphi$ is strictly increasing. Hence, for $s \neq t$ we have

$$\frac{\varphi(t) - \varphi(s)}{t-s} > 0, \quad |\varphi(t) - \varphi(s)|^{-\nu} = \left[ \frac{\varphi(t) - \varphi(s)}{t-s} \right]^{-\nu} |t-s|^{-\nu}.$$

After change of variables (37.3) equation (37.1) takes the form

$$v(t) = \int_0^1 K_\varphi(t,s) v(s) ds + f_\varphi(t), \quad 0 \le t \le 1, \quad 0 < \nu < 1, \tag{37.7}$$

where $f_\varphi(t) = f(\varphi(t))$,

$$K_\varphi(t,s) = \mathscr{A}(t,s)|t-s|^{-\nu} + \mathscr{B}(t,s), \tag{37.8}$$

$$\mathscr{A}(t,s) = a(\varphi(t), \varphi(s)) \Phi(t,s)^{-\nu} \varphi'(s), \quad \mathscr{B}(t,s) = b(\varphi(t), \varphi(s)) \varphi'(s),$$

and

$$\Phi(t,s) = \left\{ \begin{array}{ll} \frac{\varphi(t)-\varphi(s)}{t-s} & \text{for} \quad t \neq s \\ \varphi'(s) & \text{for} \quad t = s \end{array} \right\}, \quad 0 \le t, s \le 1;$$

the solutions of (37.1) and (37.7) are related by the equalities

$$v(t) = u(\varphi(t)), \quad u(x) = v(\varphi^{-1}(x)).$$

Under the conditions of Theorems 1 and 2 the solution $v(t) = u(\varphi(t))$ ($t \in [0,1]$) of (37.7) belongs $C^m[0,1]$ and satisfies (37.6). Continuing $v$ for $t < 0$ by the constant value $v(0)$ and for $t > 1$ by the constant value $v(1)$, the extended function belongs to $C^m(\mathbb{R})$. This circumstance is helpful for the 'central part' interpolation on the uniform grid by piecewise polynomials treated in next sections.

## 37.3 Central Part Interpolation by Polynomials

Given an interval $[a,b]$ $(a<b)$ and an integer $m\geq 2$, introduce the uniform grid consisting of $m$ points

$$x_i = a+\left(i-\frac{1}{2}\right)h, \quad i=1,\ldots,m, \quad h=\frac{b-a}{m}. \tag{37.9}$$

Denote by $\mathscr{P}_{m-1}$ the set of polynomials of degree not exceeding $m-1$ and by $\Pi_m$ the Lagrange interpolation projection operator assigning to any $v\in C[a,b]$ the polynomial $\Pi_m v\in\mathscr{P}_{m-1}$ that interpolates $v$ at points (37.9):

$$(\Pi_m v)(x)=\sum_{j=1}^{m} v(x_j)\prod_{\substack{k=1\\k\neq j}}^{m}\frac{x-x_k}{x_j-x_k}, \quad a\leq x\leq b.$$

**Lemma 4.** *In the case of interpolation knots (37.9) with $m\in\mathbb{N}$, $m\geq 2$, for $v\in C^m[a,b]$ it holds*

$$\max_{a\leq x\leq b}|v(x)-(\Pi_m v)(x)|\leq \theta_m h^m \max_{a\leq x\leq b}\left|v^{(m)}(x)\right|, \tag{37.10}$$

*with*

$$\theta_m=\frac{1\cdot 3\cdot\ldots\cdot(2m-1)}{2^m\,m!}=\frac{(2m)!}{2^m\,m!(2\cdot 4\cdot\ldots\cdot 2m)}\cong(\pi m)^{-\frac{1}{2}},$$

*where $\theta_m\cong\varepsilon_m$ means that $\theta_m/\varepsilon_m\to 1$ as $m\to\infty$.*

*Further, for $m=2k$, $k\geq 1$,*

$$\max_{x_k\leq x\leq x_{k+1}}|v(x)-(\Pi_m v)(x)|\leq \vartheta_m h^m \max_{a\leq x\leq b}\left|v^{(m)}(x)\right|, \tag{37.11}$$

*with*

$$\vartheta_m=2^{-2m}\frac{m!}{((m/2)!)^2}\cong\sqrt{2/\pi}\,m^{-\frac{1}{2}}2^{-m}, \tag{37.12}$$

*whereas for $m=2k+1$, $k\geq 1$,*

$$\max_{x_k\leq x\leq x_{k+2}}|v(x)-(\Pi_m v)(x)|\leq \vartheta_m h^m \max_{a\leq x\leq b}\left|v^{(m)}(x)\right|, \tag{37.13}$$

*with*

$$\vartheta_m=\frac{2\sqrt{3}}{9}\frac{(k!)^2}{(2k+1)!}\cong\frac{2\sqrt{6\pi}}{9}\,m^{-\frac{1}{2}}2^{-m}. \tag{37.14}$$

*Proof.* These estimates are consequences of the error formula

$$v(x)-(\Pi_m v)(x)=\frac{v^{(m)}(\xi)}{m!}(x-x_1)\dots(x-x_m),\ x\in[a,b],\ \xi=\xi(x)\in(a,b).$$

Comparing estimates (37.10), (37.11) and (37.13) we observe that in the central parts of $[a,b]$, the estimates for the error $v-\Pi_m v$ are approximately $2^m$ times more precise than on the whole interval. Although $m$ is fixed in our consideration, it is useful to know that in the central parts of $[a,b]$, the interpolation process on the uniform grid has also good stability properties as $m$ increases: in contrast to an exponential growth [Da77] of $\|\Pi_m\|_{\mathscr{L}(C[a,b],C[a,b])}$ as $m\to\infty$, it holds by the Runck's theorem (see [Da77, Ru61]) that

$$\|\Pi_m\|_{\mathscr{L}\left(C[a,b],C\left[\frac{a+b}{2}-rh^{1/2},\frac{a+b}{2}+rh^{1/2}\right]\right)}\le c_r\,(1+\log m),\quad rh^{\frac{1}{2}}\le\frac{b-a}{2},\tag{37.15}$$

where the constant $c_r$ depends only on $r>0$. It is known (see, e.g., [Da77]) that a logarithmic growth is the slowest one that holds for the norm of any projector $P_m: C[a,b]\to\mathscr{P}_{m-1}$ as $m\to\infty$ and, for example, the Chebyshev interpolation projectors have this slowest order of growth of norms.

## 37.4 Central Part Interpolation by Piecewise Polynomials

Introduce in $\mathbb{R}$ the uniform grid

$$\{jh: j\in\mathbb{Z}\},\quad h=\frac{1}{n},\quad n\in\mathbb{N}.$$

Let $m\in\mathbb{N}$, $m\ge 2$ be fixed. Given a function $v\in C[-\delta,1+\delta]$, $\delta>0$, we define a piecewise polynomial interpolant $\Pi_{h,m}v\in C[0,1]$ for $h=\frac{1}{n}<\frac{2\delta}{m}$ as follows. On every subinterval $[jh,(j+1)h]$, $0\le j\le n-1$, the function $\Pi_{h,m}v$ is defined independently of other subintervals as a polynomial $\Pi_{h,m}^{[j]}v\in\mathscr{P}_{m-1}$ of degree $\le m-1$ by the conditions

$$\Pi_{h,m}^{[j]}v(lh)=v(lh),\quad \text{for}\quad l\in\mathbb{Z}\quad\text{such that}\quad l-j\in\mathbb{Z}_m,\tag{37.16}$$

where $\mathbb{Z}_m=\left\{k\in\mathbb{Z}:-\frac{m}{2}<k\le\frac{m}{2}\right\}$. Observe that $\mathbb{Z}_m$ contains the following $m$ elements (integers):

$$\mathbb{Z}_m=\left\{-\frac{m}{2}+1,-\frac{m}{2}+2,\dots,\frac{m}{2}\right\}\quad\text{if } m \text{ is even},$$

$$\mathbb{Z}_m = \left\{ -\frac{m-1}{2}, -\frac{m-1}{2}+1, \ldots, \frac{m-1}{2} \right\} \quad \text{if } m \text{ is odd.}$$

For an 'interior' knot $jh$, $1 \le j \le n-1$, interpolation conditions (37.16) contain the condition $\left(\Pi_{h,m}^{[j-1]} v\right)(jh) = v(jh)$ as well as the condition $\left(\Pi_{h,m}^{[j]} v\right)(jh) = v(jh)$. Thus $\Pi_{h,m} v$ is uniquely defined at interior knots and $\Pi_{h,m} v$ is continuous on $[0,1]$. Namely, for the 'interior' knots $jh$, $1 \le j \le n-1$, interpolation conditions (37.16) yield

$$(\Pi_{h,m} v)(jh) = v(jh)$$

for $\Pi_{h,m} v$ as a function on $[(j-1)h, jh]$ as well as a function on $[jh, (j+1)h]$. The one side derivatives of the interpolant $\Pi_{h,m} v$ at the interior knots may be different.

Introduce the Lagrange fundamental polynomials $L_{k,m} \in \mathscr{P}_{m-1}$, $k \in \mathbb{Z}_m$, satisfying $L_{k,m}(l) = \delta_{k,l}$ for $l \in \mathbb{Z}_m$, where $\delta_{k,l}$ is the Kronecker symbol, $\delta_{k,l} = 0$ for $k \ne l$ and $\delta_{k,k} = 1$. An explicit formula for $L_{k,m}$ is given by

$$L_{k,m}(t) = \prod_{l \in \mathbb{Z}_m \setminus \{k\}} \frac{t-l}{k-l}, \quad k \in \mathbb{Z}_m. \tag{37.17}$$

We claim that

$$\left(\Pi_{h,m}^{[j]} v\right)(t) = \sum_{k \in \mathbb{Z}_m} v((j+k)h)\, L_{k,m}(nt-j), \quad t \in [jh, (j+1)h], \quad j = 0, \ldots, n-1. \tag{37.18}$$

Indeed, $\Pi_{h,m}^{[j]} v$ defined by (37.18) is really a polynomial of degree $\le m-1$ and it satisfies interpolation conditions (37.16): for $l$ with $l - j \in \mathbb{Z}_m$, it holds that

$$\left(\Pi_{h,m}^{[j]} v\right)(lh) = \sum_{k \in \mathbb{Z}_m} v((j+k)h)\, L_{k,m}(l-j) = \sum_{k \in \mathbb{Z}_m} v((j+k)h)\, \delta_{k,l-j}$$
$$= v((j+(l-j))h) = v(lh).$$

For $m = 2$, the interpolant $\Pi_{h,2} v$ is the usual piecewise linear function joining for $0 \le j \le n-1$ the pair of points

$$(jh, v(jh)) \in \mathbb{R}^2 \text{ and } ((j+1)h, v((j+1)h)) \in \mathbb{R}^2$$

by a straight line; $\Pi_{h,2} v$ does not use the values of $f$ outside $[0,1]$, and $\Pi_{h,2} v$ is a projection operator in $C[0,1]$, i.e. $\Pi_{h,2}^2 = \Pi_{h,2}$.

For $m \ge 3$, $\Pi_{h,m} v$ uses values of $v$ outside $[0,1]$. For $v \in C[0,1]$, $\Pi_{h,m} v$ obtains a sense after an extension of $v$ onto $[-\delta, 1+\delta]$ with $\delta \ge \frac{m}{2} h$ for even $m$ and $\delta \ge \frac{m-1}{2} h$ for odd $m$. In the case of functions $v \in C^m[0,1]$, satisfying the boundary conditions (cf. Theorem 2)

$$v^{(j)}(0) = v^{(j)}(1) = 0,\ j = 1,\ldots,m,$$

we are in a lucky situation since the simplest extension operator

$$E_\delta : C[0,1] \to C[-\delta, 1+\delta], \quad (E_\delta v)(t) = \left\{ \begin{array}{l} v(0), \quad -\delta \le t \le 0 \\ v(t), \quad 0 \le t \le 1 \\ v(1), \ 1 \le t \le 1+\delta \end{array} \right\} \tag{37.19}$$

maintains the $C^m$-smoothness of $v$. The operator

$$P_{h,m} := \Pi_{h,m} E_\delta : C[0,1] \to C[0,1] \tag{37.20}$$

is well defined and $P_{h,m}^2 = P_{h,m}$, i.e., $P_{h,m}$ is a projector in $C[0,1]$.

For $w_h \in \mathscr{R}(P_{h,m})$ (the range of $P_{h,m}$) we have

$$w_h = P_{h,m} w_h = \Pi_{h,m} E_\delta w_h,$$

and due to (37.18) we get for $t \in [jh, (j+1)h]$ $(j = 0,\ldots,n-1)$ that

$$w_h(t) = \sum_{k \in \mathbb{Z}_m} (E_\delta w_h)((j+k)h) L_{k,m}(nt - j) \tag{37.21}$$

where

$$(E_\delta w_h)(ih) = \left\{ \begin{array}{lll} w_h(ih) & \text{for} & i = 0,\ldots,n \\ w_h(0) & \text{for} & i < 0 \\ w_h(1) & \text{for} & i > n \end{array} \right\}.$$

Thus, $w_h \in \mathscr{R}(P_{h,m})$ is uniquely determined on $[0,1]$ by its knot values $w_h(ih)$, $i = 0,\ldots,n$. We conclude that $\dim \mathscr{R}(P_{h,m}) = n+1$. It is also clear that for a $w_h \in \mathscr{R}(P_{h,m})$ we have $w_h = 0$ if and only if $w_h(ih) = 0$, $i = 0,\ldots,n$.

For $v \in C[-\delta, 1+\delta]$, the interpolant $\Pi_{h,m} v$ is closely related to the central part interpolation of $v$ on the uniform grid treated in previous section. On $[jh, (j+1)h]$, the interpolant $\Pi_{h,m} v = \Pi_{h,m}^{[j]} v$ coincides with the polynomial interpolant $\Pi_m v$ constructed for $f$ on the interval $[a_j, b_j]$ where

$$a_j = \left(j - \frac{m-1}{2}\right) h, \ \ b_j = \left(j + \frac{m+1}{2}\right) h \quad \text{for even } m,$$

$$a_j = \left(j - \frac{m}{2}\right) h, \ \ b_j = \left(j + \frac{m}{2}\right) h \quad \text{for odd } m.$$

Moreover, $[jh,(j+1)h]$ is contained in the central part of $[a_j,b_j]$ on which the interpolation error can be estimated by (37.11) or (37.13). On this way we obtain the following result (cf. [OrVa09]).

**Lemma 5.** (i) *For* $v \in C^m[-\delta,1+\delta]$ $(m \geq 2, \delta > 0, h = \frac{1}{n})$,

$$\max_{0\leq t\leq 1} |v(t) - (\Pi_{h,m}v)(t)| \leq \vartheta_m h^m \max_{-\delta\leq t\leq 1+\delta} \left|v^{(m)}(t)\right|, \tag{37.22}$$

*with* $\vartheta_m$ *defined by* (37.12) *and* (37.14), *respectively, for even and odd m.*

(ii) *For* $v \in V^{(m)} := \left\{w \in C^m[0,1] : w^{(j)}(0) = w(j)(1) = 0,\ j = 1,\ldots,m\right\}$,

$$\max_{0\leq t\leq 1} |v(t) - (P_{h,m}v)(t)| \leq \vartheta_m h^m \max_{0\leq t\leq 1} \left|v^{(m)}(t)\right|. \tag{37.23}$$

*Proof.* The claim (i) is a direct consequence of Lemma 4. Further, to prove the estimate (37.23), we have $E_\delta v \in C^m[-\delta,1+\delta]$ for $v \in V^{(m)}$ and

$$\max_{-\delta\leq t\leq\delta} \left|(E_\delta v)^{(m)}(t)\right| = \max_{0\leq t\leq 1} \left|v^{(m)}(t)\right|, \quad (E_\delta v)(t) = v(t) \text{ for } 0 \leq t \leq 1.$$

Applying (37.22) to $E_\delta v$, it takes the form

$$\max_{0\leq t\leq 1} |(E_\delta v)(t) - (\Pi_{h,m}E_\delta v)(t)| \leq \vartheta_m h^m \max_{-\delta\leq t\leq 1+\delta} \left|(E_\delta v)^{(m)}(t)\right|,$$

$$\max_{0\leq t\leq 1} |v(t) - (P_{h,m}v)(t)| \leq \vartheta_m h^m \max_{0\leq t\leq 1} \left|v^{(m)}(t)\right|$$

completing the proof.

From (37.15), (37.19), (37.20) we obtain that the norms $\|P_{h,m}\|_{\mathscr{L}(C[0,1],C[0,1])}$ are uniformly bounded with respect to $n$, $h = \frac{1}{n}$:

$$\|P_{h,m}\|_{\mathscr{L}(C[0,1],C[0,1])} \leq c\,(1+\log m),$$

with a constant $c$ which is independent of $h$ (of $n$).

Together with (37.23), noticing that $V^{(m)}$ is dense in $C[0,1]$, Banach–Steinhaus theorem yields the following result.

**Lemma 6.** *For any* $v \in C[0,1]$,

$$\max_{0\leq t\leq 1} |v(t) - (P_{h,m}v)(t)| \to 0 \text{ as } n = \frac{1}{h} \to \infty.$$

## 37.5 Collocation Based on the Central Part Interpolation

We rewrite (37.7) in the operator form

$$v = T_\varphi v + f_\varphi, \tag{37.24}$$

with $T_\varphi$ defined by the formula

$$(T_\varphi v)(t) = \int_0^1 K_\varphi(t,s)v(s)ds, \quad 0 \le t \le 1, \tag{37.25}$$

where $K_\varphi(t,s)$ is given by the formula (37.8). Using the interpolation projector $P_{h,m}$ defined in (37.20), we approximate equation (37.24) by equation

$$v_h = P_{h,m}T_\varphi v_h + P_{h,m}f_\varphi. \tag{37.26}$$

This is the operator form of our piecewise polynomial collocation method based on a central part interpolation on the uniform grid.

**Theorem 3.** *Let the assumptions of Lemma 2 be fulfilled. Moreover, assume that* $f \in C^{m,\theta_0,\theta_1}(0,1)$, *with* $m \in \mathbb{N}$, $m \ge 2$, $\theta_0 = \lambda_0 + \nu$, $\theta_1 = \lambda_1 + \nu$. *Let* $\mathcal{N}(I-T) = \{0\}$ *or equivalently,* $\mathcal{N}(I-T_\varphi) = \{0\}$. *Finally, let* $\varphi$ *be defined by the formula (37.4) with parameters* $p_0, p_1 \in \mathbb{N}$ *satisfying (37.5).*

*Then equation (37.24) (equation (37.7)) has a unique solution* $v \in C[0,1]$ *and there exists an* $n_0$ *such that for* $n \ge n_0$, *the collocation equation (37.26) has a unique solution* $v_h$. *The accuracy of* $v_h$ *can be estimated by*

$$\|v - v_h\|_\infty \le ch^m \|v^{(m)}\|_\infty, \quad n = \frac{1}{h} \ge n_0, \tag{37.27}$$

*where* $c$ *is a positive constant not depending on* $n = \frac{1}{h}$ *and* $f$.

*Proof.* It follows from [VaVa08] that $\mathscr{A}, \mathscr{B} \in C([0,1]\times[0,1])$ and therefore $T_\varphi$ given by (37.25) is compact as an operator from : $C[0,1]$ into $C[0,1]$. Since $\mathcal{N}(I - T_\varphi) = \{0\}$, the bounded inverse $(I-T_\varphi)^{-1} : C[0,1] \to C[0,1]$ exists due to Fredholm alternative. Denote

$$\kappa := \left\|(I-T_\varphi)^{-1}\right\|_{\mathscr{L}(C[0,1],C[0,1])}.$$

The compactness of $T_\varphi : C[0,1] \to C[0,1]$ and the pointwise convergence $P_{h,m}$ to $I$ (the identity mapping) in $C[0,1]$ (see Lemma 6) imply the norm convergence

$$\varepsilon_h := \left\|T_\varphi - P_{h,m}T_\varphi\right\|_{\mathscr{L}(C[0,1],C[0,1])} \to 0 \text{ as } n \to \infty \quad (\text{as } h = \frac{1}{n} \to 0).$$

Hence there is an $n_0$ such that $\kappa\varepsilon_h < 1$ for $n > n_0$. We conclude that $I - P_{h,m}T_\varphi$ is invertible in $C[0,1]$ for $n \geq n_0$ and

$$\kappa_h := \left\|(I - P_{h,m}T_\varphi)^{-1}\right\|_{\mathscr{L}(C[0,1],C[0,1])} \to \kappa \text{ as } n \to \infty \quad (\text{as } h = \frac{1}{n} \to 0), \tag{37.28}$$

since,

$$\left\|(I - P_{h,m}T_\varphi)^{-1}\right\|_{\mathscr{L}(C[0,1],C[0,1])} \leq \frac{\kappa}{1 - \kappa\varepsilon_h} \to \kappa \text{ as } h \to 0.$$

This proves the unique solvability of the collocation equation (37.26) for $n \geq n_0$.

Let $v$ and $v_h$ be the solutions of (37.24) and (37.26), respectively. Then

$$(I - P_{h,m}T_\varphi)(v - v_h) = v - P_{h,m}T_\varphi v - P_{h,m}f_\varphi = v - P_{h,m}v,$$

$$v - v_h = (I - P_{h,m}T_\varphi)^{-1}(v - P_{h,m}v)$$

and

$$\|v - v_h\|_\infty \leq \kappa_h \|v - P_{h,m}v\|_\infty, \quad n = \frac{1}{h} \geq n_0. \tag{37.29}$$

By Theorem 1, for the solution $u$ of (37.1) we have $u \in C^{m,\theta_0,\theta_1}(0,1)$; by Theorem 2, for $v(t) = u(\varphi(t))$ we have $v \in C^m[0,1]$ and $v^{(j)}(0) = v^{(j)}(1) = 0$, $j = 1,\ldots,m$; by Lemma 5(ii),

$$\|v - P_{h,m}v\|_\infty \leq \vartheta_m h^m \|v^{(m)}\|_\infty.$$

Now (37.29) yields

$$\|v - v_h\|_\infty \leq \kappa_h \vartheta_m h^m \|v^{(m)}\|_\infty$$

that together with (37.28) implies (37.27). The proof is complete.

Numerical examples (omitted here) confirm the theoretical accuracy.

*Remark 1.* With respect to

$$u_h(x) := v_h(\varphi^{-1}(x)), \quad 0 \leq x \leq 1,$$

estimate (37.27) reads

$$\max_{0\leq x\leq 1} |u(x) - u_h(x)| = \max_{0\leq t\leq 1} |v(t) - v_h(t)| \leq c_1 h^m, \quad n = \frac{1}{h} \geq n_0,$$

where $c_1$ is a positive constant which does not depend on $n = \frac{1}{h}$.

## 37.6 Matrix Form of the Method

The solution $v_h$ of equation (37.26) belongs to $\mathscr{R}(P_{h,m})$, so the knot values

$$v_h(ih),\ i=0,\dots,n,$$

determine $v_h$ uniquely. Equation (37.26) is equivalent to a system of linear algebraic equation with respect to $v_h(ih)$, $i=0,\dots,n$, and our task is to write down this system.

For $w_h \in \mathscr{R}(P_{h,m})$ we have $w_h = 0$ if and only if $w_h(ih)=0$, $i=0,\dots,n$. Since $(P_{h,m}w)(ih) = w(ih)$, $i=0,\dots,n$, equation (37.26) is equivalent to the conditions

$$v_h(ih) = (T_\varphi v_h)(ih) + f_\varphi(ih), \quad i=0,\dots,n,$$

i.e., $v_h \in \mathscr{R}(P_{h,m})$ satisfies equation (37.24) (equation (37.7)) at the knots $ih$, $i=0,\dots,n$. Using for $v_h$ the representation (37.21) we obtain

$$\begin{aligned}(T_\varphi v_h)(ih) &= \int_0^1 K_\varphi(ih,s)v_h(s)ds = \sum_{j=0}^{n-1}\int_{jh}^{(j+1)h} K_\varphi(ih,s)v_h(s)ds \\ &= \sum_{j=0}^{n-1}\sum_{k\in\mathbb{Z}_m}\int_{jh}^{(j+1)h} K_\varphi(ih,s)L_{k,m}(ns-j)ds\,(E_\delta v_h)((j+k)h) \\ &= \sum_{j=0}^{n-1}\sum_{k\in\mathbb{Z}_m}\alpha_{i,j,k}\cdot\left\{\begin{array}{lll} v_h(0) & \text{for} & j+k\le 0 \\ v_h((j+k)h) & \text{for} & 1\le j+k\le n-1 \\ v_h(1) & \text{for} & j+k\ge n\end{array}\right\} \\ &= \sum_{l=0}^{n} b_{i,l}v_h(lh), \quad i=0,\dots,n,\end{aligned}$$

where for $k\in\mathbb{Z}_m$ we denoted

$$\alpha_{i,j,k} = \int_{jh}^{(j+1)h} K_\varphi(ih,s)L_{k,m}(ns-j)ds, \quad i=0,\dots,n, \quad j=0,\dots,n-1, \tag{37.30}$$

$$b_{i,l} = \left\{\begin{array}{ll} \sum_{k\in\mathbb{Z}_m}\sum_{\{j:0\le j\le n-1,\, j+k\le 0\}}\alpha_{i,j,k}, & \text{for } l=0 \\ \sum_{k\in\mathbb{Z}_m}\sum_{\{j:0\le j\le n-1,\, j+k=l\}}\alpha_{i,j,k}, & \text{for } l=1,\dots,n-1 \\ \sum_{k\in\mathbb{Z}_m}\sum_{\{j:0\le j\le n-1,\, j+k\ge n\}}\alpha_{i,j,k}, & \text{for } l=n\end{array}\right\}, \tag{37.31}$$

$$i,l=0,\dots,n.$$

Thus the matrix form of method (37.26) is given by

$$v_h(ih) = \sum_{l=0}^{n} b_{i,l} v_h(lh) + f_\varphi(ih), \quad i = 0, \ldots, n, \tag{37.32}$$

with $b_{i,l}$ defined by (37.30)–(37.31). Having determined $v_h(ih)$ $(i = 0, \ldots, n)$ through solving the system (37.32), the collocation solution $v_h(t)$ at any intermediate point $t \in [jh, (j+1)h]$, $j = 0, \ldots, n-1$, is given by

$$v_h(t) = \sum_{k \in \mathbb{Z}_m} \left\{ \begin{array}{ccc} v_h(0) & \text{for} & j+k \le 0 \\ v_h((j+k)h) & \text{for} & 1 \le j+k \le n-1 \\ v_h(1) & \text{for} & j+k \ge n \end{array} \right\} \cdot L_{k,m}(nt - j),$$

with $L_{k,m}$, $k \in \mathbb{Z}_m$, defined by (37.17).

**Acknowledgements** This work was supported by Estonian Science Foundation Grant No 9104 and by the institutional research funding IUT20-57 of the Estonian Ministry of Education and Research.

## References

[Da77] Daugavet, I. K.: Introduction to the Function Approximation Theory. Leningrad University Press, Leningrad (1977) (in Russian)

[MoSc98] Monegato, G., Scuderi, L.: High order methods for weakly singular integral equations with nonsmooth input functions. Math. Comput., **67**, 1493–1515 (1998)

[Or13] Orav-Puurand, K.: A Central Part Interpolation Scheme for Log-Singular Integral Equations. Mathematical Modelling and Analysis, **18(1)** 136–148 (2013)

[OrPeVa10] Orav-Puurand, K., Pedas, A., Vainikko, G.: Nyström type methods for Fredholm integral equations with weak singularities. Journal of Computational and Applied Mathematics, **234(9)**, 2848–2858 (2010)

[OrVa09] Orav-Puurand, K., Vainikko, G.: Central part interpolation schemes for integral equations. Numerical Functional Analysis and Optimization, **30**, 352–370 (2009)

[PeVa06a] Pedas, A., Vainikko, G.: Integral equations with diagonal and boundary singularities of the kernel. Z. Anal. Anwendungen, **25(4)**, 487–516 (2006)

[PeVa06b] Pedas, A., Vainikko, G.: Smoothing transformation and piecewise polynomial projection methods for weakly singular Fredholm integral equations. Commun. Pure Appl. Math, **5** 395–413 (2006)

[Ru61] Runck, P.: Über Konvergenzfragen bei Polynominterpolation mit equidistanten Knoten I, II. Journal für die reine und angewandte Mathematik, **208**; **210**, 51–69; 175–204 (1961; 1962)

[VaVa08] Vainikko, E., Vainikko, G.: A Spline product quasi-interpolation method for weakly singular Fredholm integral equations. SIAM Journal on Numerical Analysis, **46**, 1799–1820 (2008)

# Chapter 38
# Evolutional Contact with Coulomb Friction on a Periodic Microstructure

**J. Orlik and V. Shiryaev**

## 38.1 Statement of Quasi-Static Multi-Scale Contact Problem

**Assumption 38.1.1 (on geometry)** *We consider an $\varepsilon Y-$periodic domain $\Omega_\varepsilon \subset \mathbb{R}^n$ consisting of a connected domain $\Omega_0^\varepsilon$ and $s$ periodically distributed inclusions $\Omega_l^\varepsilon$, $l = 1,\dots,s$ with cracks on the interface between $\Omega_0^\varepsilon$ and the inclusions.*

*We denote the contact boundary of each domain by $S_\varepsilon^j$, $j = 0,\dots,N$ and the complete contact boundary by $S_\varepsilon$, which are Lipschitz continuous. We denote by $Y$ a unit periodicity cell and $0 < \varepsilon \ll 1$ is a scaling parameter. We suppose for $\Omega_\varepsilon$ a Lipschitz external boundary $\partial\Omega^\varepsilon$, which is decomposed in two parts $\partial\Omega_\varepsilon = \Gamma_{D\,\varepsilon} \cup \Gamma_{N\,\varepsilon}$, on which Dirichlet and Neumann boundary conditions are imposed, respectively. We denote by $j = 0$ domains touching the Dirichlet part of the boundary.*

Assume we have the symmetric bilinear form

$$\mathbf{a}^\varepsilon(e(u),e(v)) \doteq \int_{\Omega_\varepsilon^j} \sum_{\alpha,\beta,\gamma,\delta=1}^{3} a^\varepsilon_{\alpha\beta\gamma\delta}(x) e(u)_{\gamma\delta}(x)\, e(v)_{\alpha\beta}(x)\, dx,$$

where the tensor field $a^\varepsilon = (a^\varepsilon_{\alpha\beta\gamma\delta})$, $a^\varepsilon_{\alpha\beta\gamma\delta} \in L^\infty(\Omega^\varepsilon)$ has the usual properties of symmetry, boundedness (with constant $C_A$), and coercivity (with constant $\overline{\alpha}$) when operating on symmetric tensors of order two: $a^\varepsilon_{\alpha\beta\gamma\delta} = a^\varepsilon_{\beta\alpha\gamma\delta} = a^\varepsilon_{\alpha\beta\delta\gamma} = a^\varepsilon_{\gamma\delta\alpha\beta}, \overline{\alpha}\,\eta_{\alpha\beta}\eta_{\alpha\beta} \le a^\varepsilon_{\alpha\beta\gamma\delta}\,\eta_{\alpha\beta}\eta_{\gamma\delta} \le C_A\,\eta_{\alpha\beta}\eta_{\gamma\delta}$. Let $\mathscr{K}^\varepsilon$ be the convex set, defined for non-negative $g_\varepsilon^j \in H^{1/2}(S_\varepsilon^j)$, by $\mathscr{K}^\varepsilon \doteq \left\{ v \;\middle|\; v \in H^1(\Omega_\varepsilon^0;\Gamma_D), [v_\nu]_{S^j} \le g_\varepsilon^j \right\}$.

J. Orlik (✉) • V. Shiryaev
Fraunhofer ITWM, Kaiserslautern, Germany
e-mail: orlik@itwm.fhg.de; shiryaev@itwm.fhg.de



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_38

The vector fields $v$ are the admissible displacement fields with respect to the reference configuration $\Omega_\varepsilon$. We will denote by $[v]_{S^j_\varepsilon}$ the jump of the vector field across the surface $S^j_\varepsilon$. By standard trace theorems, these jumps belong to $H^{1/2}(S^j_\varepsilon)$.

The tensor field $\sigma^\varepsilon_{\alpha\beta}(v) \doteq \sum_{\gamma,\delta=1}^{3} a^\varepsilon_{\alpha\beta\gamma\delta} e(v)_{\gamma\delta}$ is the stress tensor associated with the deformation $e(v)$ (not to be confused with the surface measures $d\sigma$!).

The function $g^j_\varepsilon(x) = \varepsilon g^j(x/\varepsilon)$ is the original gap (in the reference configuration), and the corresponding inequality in the definition of $\mathscr{K}^\varepsilon$ represents the non-penetration condition. In case there is contact in the reference configuration, these functions are just 0. $f_\varepsilon \in L^2(\Omega_\varepsilon)$ represents the volume force. In the case of Coulomb friction, the dissipative term $\Psi_\varepsilon(v,[\dot v_\tau]) \doteq -\int_{S_\varepsilon} \mu\sigma_\nu(v)|[\dot v_\tau]|$ depends not only on the sliding $[\dot v_\varepsilon\ \tau]$, but also on a function $\sigma_\nu$ of the state variable $v$.

The strong formulation of the quasi-static contact problem reads: Find $u_\varepsilon$ in $\mathscr{K}^\varepsilon$ such that

$$\begin{cases} -\operatorname{div}\sigma_\varepsilon = \bar f_\varepsilon \quad \text{in } \Omega^*_\varepsilon, \\ [(\dot u_\varepsilon)_\nu]_{S_\varepsilon} - g^j_\varepsilon \le 0, \quad \sigma_\varepsilon(\nu)_\nu \le 0, \\ \sigma_\varepsilon(\nu)_\nu\big([(\dot u_\varepsilon)_\nu]_{S^0_\varepsilon} - g_\varepsilon\big) = 0 \\ \sigma_\varepsilon(\nu)_\tau \in \partial\Psi_\varepsilon(u,[(\dot u_\varepsilon)_\tau]_{S_\varepsilon}) \quad \text{on } S_\varepsilon, \\ \dot u_\varepsilon = \dot g \quad \text{on } \Gamma_{D\varepsilon}, \\ \sigma_\varepsilon\cdot\nu = 0 \quad \text{on } \Gamma_{N\varepsilon}, \\ u_\varepsilon(0,x) = u_{0\varepsilon}(x), \qquad x\in\Omega_\varepsilon, \end{cases}$$

where $\partial\Psi_\varepsilon$ denotes the subdifferential of $\Psi_\varepsilon$ (here taken in the sense of the $L^2(S_\varepsilon)$ duality), $\bar f^j = f^j_\varepsilon - \frac{\partial}{\partial y_h}\left(a_{jihk}\frac{\partial\chi_i(y)}{\partial y_k}\right)$, where $\bar f^j$ denotes the given volume force and $\chi$ denotes a $H^1(\Omega)$-extension of the Dirichlet values $g \in H^{1/2}(\Gamma_D)$, with $a_{jihk}\frac{\partial\chi_i(y)}{\partial y_k}n_h|_{\Gamma_N} = 0$, which exists due to the trace theorem.

We define $\sigma_\nu \in H^{-1/2}(S)$ as the normal component of the co-normal derivative on the contact interface of the solution of $\mathbf{div}\sigma(u) = -f$ in $Y\setminus S$, $u_\tau = 0$ on $S$, or

$$(\sigma_\nu(u), w_\nu)|_S = a(u,w) - (f,w), \quad w_\tau \in H^1(Y\setminus S),\ w_\tau = 0 \text{ on } S. \tag{38.1}$$

The weak formulation will be then as follows:

**Problem $\mathscr{P}''_\varepsilon$:** Find $u_\varepsilon \in \mathscr{K}^\varepsilon$ such that for every $v \in H^1(\Omega_\varepsilon;\Gamma_D)$

$$\mathbf{a}^\varepsilon(e(u_\varepsilon), e(v-\dot u_\varepsilon)) + \Psi_\varepsilon\big(u_\varepsilon, [v_\tau]_{S_\varepsilon} - [(\dot u_\varepsilon)_\tau]_{S_\varepsilon}\big) \ge (f_\varepsilon, (v-\dot u_\varepsilon)), \quad v\in\mathscr{K}^\varepsilon. \tag{38.2}$$

The problem can further be discretized in time similar to the formulation (3.4.10) from [EcJaKr05] and inequality (6) from [CoRo00]. We consider a partition of the

time interval $I_T$ with time steps of equal step size $\Delta t = T/L$. Let $t_l = l\Delta t$, $l = 0,\dots,L$. Let $u_l$ be an approximation for $u(t_l)$) and $\Delta u^{(l)} \equiv u_{(l)} - u_{(l-1)}$ be the time difference operator. The time discretized problem is obtained from (3.4.9) by replacing $u$ with $u^{(l)}$ and $\dot{u}^{(l)}$) with $\Delta u^{(l)}/\Delta t$. If the result is multiplied by the time step $\Delta t$, the following variational inequality is obtained:

**Problem** $\mathscr{P}'''_\varepsilon$**:** For $u_\varepsilon^{i-1} \in \mathscr{K}^\varepsilon$, find $u_\varepsilon^i \in \mathscr{K}^\varepsilon$ such that $\forall v \in H^1(\Omega_\varepsilon;\Gamma_D)$

$$\mathbf{a}^\varepsilon(e(u_\varepsilon^i), e(v-u_\varepsilon^i)) + \Psi_\varepsilon(u_\varepsilon^i, [v_\tau]_{S_\varepsilon} - [(u_\varepsilon^{i-1})_\tau]_{S_\varepsilon}) - \Psi_\varepsilon(u_\varepsilon^i, [(u_\varepsilon^i)_\tau]_{S_\varepsilon} - [(u_\varepsilon^{i-1})_\tau]_{S_\varepsilon}) \geq (f_\varepsilon, (v-u_\varepsilon^i)), \quad v \in \mathscr{K}^\varepsilon. \tag{38.3}$$

### *38.1.1 Auxiliary Inequalities*

**Lemma 1.** *There is a constant* $\gamma_0$ *such that*

$$\|[v]\|_{H^{1/2}(S)} \leq \gamma_0 \|\nabla v\|_{L^p(Y\setminus S)}$$

Proof is given by Poincaré–Wirtinger inequality in [CiDaOr13]. We define also the inverse trace-jump operator, i.e. an extension of a jump-function. It can be extended in the different ways. We consider an auxiliary problem (38.1), but put $f = 0$ and apply a given Dirichlet boundary and normal-jump values $\{v^0|_{\partial\Omega_D}, [v_n^j]|_{S_j}\} = g^j$, $j = 0,\dots,m$. Then, the existence of the extension and inverse trace-jump inequality comes from the preliminary estimate for such a problem $\sum_{j=0}^m \|v^j\|_{H^1(\Omega^j)} \leq \gamma_1 \sum_{j=0}^m \|g^j\|_{H^{1/2}(S^j)}$.

We define the space $H^{1,\alpha} = \{w \in H^1(\Omega_s); \|w\|_{H^{1,\alpha}} < +\infty\}$, for $0 < \alpha < 1$, where

$$\|w\|^2_{H^{1,\alpha}(\Omega_s)} = \|w\|^2_{H^1(\Omega_s)} + \int_{\mathbb{R}^n}\int_{\Omega_s} \frac{1}{|h|^{n+2\alpha}} \sum_{i=1}^n \left[\left(\frac{\partial w}{\partial x_i}\right)_{-h} - \left(\frac{\partial w}{\partial x}_i\right)\right]^2 dx\,dh,$$

with $v_{-h}(x) = v(x+h)$, for $\mathbf{x} \in \mathbb{R}^n$ and $h \in \mathbb{R}^n$.

We consider equally the space $H^\alpha(\mathbb{R}^n) = \{w \in L^2(\mathbb{R}^n); \|w\|_{H^\alpha} < +\infty\}$ with

$$\|w\|^2_{H^\alpha(\mathbb{R}^n)} = \|w\|^2_{L^2(\mathbb{R}^n)} + \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \frac{(w(x+h)-w(x))^2}{|h|^{n+2\alpha}} dx\,dh.$$

We consider $c_n(\alpha)$ such that $|\xi|^{2\alpha} c_n(\alpha) = \int_{\mathbb{R}^n} \frac{|e^{ih\cdot\xi}-1|^2}{|h|^{n+2\alpha}} dh$, $\forall \xi \in \mathbb{R}$ and denoting by $F[w]$ the Fourier transform of $w$, we have that

$$\|w\|^2_{H^\alpha(\mathbb{R}^n)} = \int_{\mathbb{R}^n} |F[w](\xi)|^2 (1 + c_n(\alpha)\,|\xi|^{2\alpha})\, d\xi.$$

We define $H^{-\alpha}(\mathbb{R}^n)$ as the dual space of $H^{\alpha}(\mathbb{R}^n)$. Thus its norm satisfies

$$\|w\|^2_{H^{-\alpha}(\mathbb{R}^n)} = \int_{\mathbb{R}^n} |F[w](\xi)|^2 \, (1 + c_n(-\alpha)\,|\xi\,|^{-2\alpha})^{-1} d\xi.$$

**Definition 1.** For a function $u$ defined on $\mathbb{R}^{n-1}$ and any $h \in \mathbb{R}^{n-1}$ we introduce the translation (shift) operator and the difference operator

$$S_h u = u_{-h} : \; x \mapsto u(x+h), \;\; x \in \mathbb{R}^{n-1},$$
$$\Delta_h : \; x \mapsto u(x+h) - u(x), \;\; x \in \mathbb{R}^{n-1}$$

Furthermore, we define by $||\cdot||'$ a semi-norm, for $\alpha, \beta > 0$ and $\alpha + \beta < 1$,

$$\int_{\mathbb{R}^n} \frac{||\Delta_h w||'^2_{H^{\beta}(\mathbb{R}^n)}}{|h|^{n+2\alpha}} \, dh = d_n(\alpha,\beta)||w||'^2_{H^{\alpha+\beta}(\mathbb{R}^n)}, \tag{38.4}$$

with $d_n(\alpha,\beta) = \frac{c_n(\alpha)c_n(\beta)}{c_n(\alpha+\beta)}$.

$$\int_{\mathbb{R}^n} \frac{||\Delta_h w||^2_{H^{-\beta}(\mathbb{R}^n)}}{|h|^{n+2\alpha}} \, dh = d^*_n(\alpha,\beta)||w||^2_{H^{\alpha-\beta}(\mathbb{R}^n)} + R_{\alpha,\beta}(u),$$

with $d^*_n(\alpha,\beta) = \frac{c_n(\alpha)c_n(\beta-\alpha)}{c_n(\beta)}$, $|R_{\alpha,\beta}(u)| \le c(\alpha,\beta)||w||^2_{H^{2(\alpha-\beta)}(\mathbb{R}^n)}$.

$$\int_{\mathbb{R}^{n-1}} \frac{||\Delta_h w||'^2_{H^{\beta}(\mathbb{R}^n)}}{|h'|^{n+2\alpha}} \, dh' = d_{n-1}(\alpha,\beta)||w||'^2_{H^{\alpha+\beta,\alpha}(\mathbb{R}^n)}, \tag{38.5}$$

with $h = (h', 0)$.

The following theorem is recalled from [CoRo00], it proves the existence of the solution to the incremental problem if the normal stress on the interface has a fixed point w.r.t. the time iterations.

**Theorem 1.** *Let tangential rigid displacements, $r_\tau$ be fixed on $\Gamma_1 \subset S$ with* meas$\Gamma_1 > 0$, *$\varepsilon$ be fixed. Furthermore, let coefficients $||\mu||_{L^\infty(S)} < \frac{\overline{\alpha}}{\gamma_0 \gamma_1 C_A}$, where $C_A$, $\overline{\alpha}$ are the constants from the continuity and coercivity condition for the elastic bilinear form, $\gamma_0$ and $\gamma_1$ are the constants from the jump and inverse jump inequalities and $f \in L^2(\Omega)$, $g \in H^{1/2}(\Gamma_D \cup S)$.*
*Then there exists a constant $C$ such that*

$$\left\|e(u^i)\right\|_{L^2(\Omega\setminus S)} + \|\sigma_\nu(u)\|_{H^{-1/2}(S)} \le C(\|\bar{f}\|_{L^2(\Omega} + \|g\|_{H^{1/2}(\Gamma_D,S)}). \tag{38.6}$$

*Proof.* For $v = 0$, (38.3) can be rewritten as

$$\mathbf{a}(e(u^i), e(u^i)) + \Psi_\varepsilon\left(u^i, [(u^i)_\tau]_S\right) \le (f, u^i)$$

For the Coulomb friction, the way to find a solution continuous in time and to get an estimate in some Hölder-spaces, the frictional term is shifted to the right-hand side and estimated from below (see Theorem 3.4.4 in [EcJaKr05] or Lemma 2.1 in [CoRo00]).

$$-\left(\mu\sigma_\nu(u^i)|[(u^i)_\tau]|\right)_S \ge -\|\mu\|_{H^{-1/2}(S)\to H^{-1/2}(S)}\|\sigma_\nu(u^i)\|_{H^{-1/2}(S)}\|[(u^i)_\tau]\|_{H^{1/2}(S)}.$$

Owing to the Korn's and jump inequalities,

$$\bar{\alpha}\left\|e(u^i)\right\|_{L^2(\Omega\setminus S)} \le \left\|f^i\right\|_{L^2(\Omega)} + \|g^i\|_{H^{1/2}(\Gamma_D,S)} + \|\mu\|_{H^{-1/2}(S)\to H^{-1/2}(S)}\,\gamma_0\left\|\sigma_\nu(u^i)\right\|_{H^{-1/2}(S)}.$$

Relation (38.1) together with the inverse jump theorem with the constant $\gamma_1$ enables us to obtain that

$$\|\sigma_\nu(u^i)\|_{H^{-1/2}(S)} \le \gamma_1(\|f^i\|_{L^2(\Omega)} + C_A\|e(u^i)\|_{L^2(\Omega\setminus S)}).$$

The statement of the lemma is obtained simply by multiple applications of the classical inequality $ab \le \ell\, a^2 + (1/(4\ell))b^2$ (for $\ell > 0$).

## 38.2 Scaling of Korn Inequalities and Trace Theorem via Unfolding

Let us now recall some formulas related to the unfolding operator (see for more details [CiEtAl12]) defined for any function $\phi$ Lebesgue measurable on $\Omega$, as $\mathscr{T}_\varepsilon(\phi)(x,y) = \phi\left(\varepsilon\left[\frac{x}{\varepsilon}\right]_Y + \varepsilon y\right)$ For $\varphi \in H^1(\Omega_\varepsilon^j)$, recall that $\nabla_y(\mathscr{T}_\varepsilon(\varphi)) = \varepsilon\mathscr{T}_\varepsilon(\nabla\varphi)$ on $\Omega \times Y^j$. In a similar way, $e_y(\mathscr{T}_\varepsilon(\varphi)) = \varepsilon\mathscr{T}_\varepsilon(e(\varphi))$. The following equalities hold:

$$\|\mathscr{T}_\varepsilon(u)\|_{L^2(\Omega\times Y^j)} = \sqrt{|Y|}\|u\|_{L^2(\Omega_\varepsilon^j)},\ \ \|\nabla_y\mathscr{T}_\varepsilon(u)\|_{L^2(\Omega\times Y^j)} = \varepsilon\sqrt{|Y|}\|\nabla u\|_{L^2(\Omega_\varepsilon^j)},$$

$$\|\mathscr{T}_\varepsilon(e(u))\|_{L^2(\Omega\times Y^j)} = \sqrt{|Y|}\|e(u)\|_{L^2(\Omega_\varepsilon^j)},\ \ \|\mathscr{T}_\varepsilon^b u_\varepsilon\|_{L^p(\Omega\times S^j)} = (\varepsilon|Y|)^{1/p}\|u\|_{L^p(S_\varepsilon^j)}.$$

The following inequality was proved in Prop. 5.1. in [CiDaOr13].

**Proposition 1.** *There exists a constant $C_1$ such that for all $u$ in $H^1(\Omega_\varepsilon^j)$, with $\Omega_\varepsilon^j$ non-locked periodic Lipschitz domains, $j=1,\dots,m$,*

$$\|u^j\|_{L^2(\Omega_\varepsilon^j)}+\varepsilon\|\nabla u^j\|_{L^2(\Omega_\varepsilon^j)}\le C_1\big(\|e(u^0)\|_{L^2(\Omega_\varepsilon^0)}+\varepsilon\|e(u^j)\|_{L^2(\Omega_\varepsilon^j)}$$
$$+\varepsilon^{1/2}\|g_\varepsilon^j\|_{L^1(\partial S_\varepsilon^j)}+\varepsilon^{-1/2}\big\|[u]_\tau\big\|_{L^1(S_\varepsilon^j)}\big).$$

While scaling of the trace theorem was done in [GaKnNe14]:

$$\varepsilon||u_\varepsilon||^2_{L^2(S_\varepsilon)}\le\gamma(||u_\varepsilon||^2_{L^2(\Omega_\varepsilon)})+\varepsilon^2||\nabla u_\varepsilon||^2_{L^2(\Omega_\varepsilon)})$$
$$\int_{S_\varepsilon}\int_{S_\varepsilon}\frac{|u_\varepsilon(x)-u_\varepsilon(y)|^2}{|x-y|^n}d\sigma_x d\sigma_y\le\gamma_2(\frac{1}{\varepsilon^2}||u_\varepsilon||^2_{L^2(\Omega_\varepsilon)})+\varepsilon^2||\nabla u_\varepsilon||^2_{L^2(\Omega_\varepsilon)}).$$

However for jumps, considering in this paper, we again get a better (see [CiDaOr13]) estimate via the semi-norm

**Lemma 2.** *We have*

$$|||[u_\varepsilon]|||^2_{L^2(S_\varepsilon)}\le\varepsilon\gamma_0||\nabla u_\varepsilon||^2_{L^2(\Omega_\varepsilon)}$$
$$\int_{S_\varepsilon}\int_{S_\varepsilon}\frac{|[u_\varepsilon](x)-[u_\varepsilon](y)|^2}{|x-y|^n}d\sigma_x d\sigma_y\le\gamma_2\varepsilon^2||\nabla u_\varepsilon||^2_{L^2(\Omega_\varepsilon)}.$$

And the following rule is valid [GaKnNe14] for the unfolding. For $u_\varepsilon\in H^s(S_\varepsilon)$ with $s\in(0,1)$, we have $T_\varepsilon^b u_\varepsilon\in H^s(S))$, and we have the equality

$$\int_\Omega\int_S\int_S\frac{|\mathscr{T}_\varepsilon^b u_\varepsilon(x,y)-\mathscr{T}_\varepsilon^b u_\varepsilon(x,z)|^2}{|y-z|^{n-1+2s}}d\sigma_y d\sigma_z\,dx$$
$$=\varepsilon^{1+2s}\int_{S_\varepsilon}\int_{S_\varepsilon}\frac{|u_\varepsilon(y)-u_\varepsilon(z)|^2}{|y-z|^{n-1+2s}}d\sigma_y d\sigma_z. \qquad (38.7)$$

We define also the inverse trace-jump operator and scale it.

**Lemma 3.**

$$\sum_{j=0}^m\|e(u_\varepsilon^j)\|_{L^2(\Omega_\varepsilon^j)}\le\frac{1}{\sqrt{\varepsilon}}\gamma_1\sum_{j=0}^m\|[u_\varepsilon^j]_n\|_{H^{1/2}(S_j)}.$$

*Proof.* We extend the jumps, set $v\equiv\mathscr{T}_\varepsilon u_\varepsilon$ and then pass to $u_\varepsilon$:

$$\varepsilon\sum_{j=0}^m\|e(u_\varepsilon^j)\|_{L^2(\Omega_\varepsilon^j)}\le\sqrt{\varepsilon}\gamma_1\sum_{j=0}^m\|[u_\varepsilon^j]_n\|_{H^{1/2}(S_j)}.$$

Let $\chi_\varepsilon(x) = \chi_1(x/\varepsilon)$. This was shown in [Mi14]: Denote by $\mathscr{F}_{x\to\xi}\left[\chi(x)\right]$ the Fourier transform. It can be scaled according to the following:

$$\mathscr{F}_{x\to\xi}\left[\chi_\varepsilon(x)\right] = -\frac{1}{4\,\pi}\int\limits_{\mathbb{R}^n}\chi_\varepsilon(x)\,e^{-2\pi i x\cdot\xi}\,dx$$

$$= -\frac{1}{4\,\pi}\int\limits_{\mathbb{R}^n}\chi_1(\tilde{x})\,e^{-2\pi i\varepsilon\tilde{x}\cdot\xi}\,\varepsilon^n\,d\tilde{x} = \varepsilon^n\hat{\chi}_1(\varepsilon\xi).$$

Let us consider a shifting $h_\varepsilon = \varepsilon h$ and scale the expression for definition of the constant $c_n(\alpha)$

**Lemma 4.** *For $\varepsilon$-periodic function*

$$\|\chi_1\left(\frac{x}{\varepsilon}\right)\|^2_{H^\alpha(\mathbb{R}^n)} = \varepsilon^n\int_{\mathbb{R}^n}(1+\varepsilon^{-2\alpha}c_{n,\xi}(\alpha)|\varepsilon\xi|^{2\alpha})|\hat{\chi}_1(\varepsilon\xi)|^2 d(\varepsilon\xi)$$

$$= \varepsilon^n\int_{\mathbb{R}^n}(1+\varepsilon^{-2\alpha}c_{n,\zeta}(\alpha)|\zeta|^{2\alpha})|\hat{\chi}_1(\zeta)|^2 d\zeta$$

$$c_n(\alpha) = |\varepsilon\xi|^{-2\alpha}\int_{\mathbb{R}^n}\frac{e^{ih(\varepsilon\xi)}-1}{|h|^{n+2\alpha}}\,dh,\;\; c_{n,\varepsilon}(\alpha) = \varepsilon^{n-2\alpha}c_n(\alpha).$$

We would like to estimate this constant for the semi-norm of the unfolding operator of a function defined on a periodic structure, like in (38.7). The purpose of this constant is to make the norm of the Bessel potential space equivalent to the one of the Sobolev–Slobodetski space with the same $\alpha$. From the scaling of the last one above, we may expect this constant to be $c_{n,\varepsilon}(\alpha) \approx \varepsilon^{-1-2\alpha}c_n(\alpha)$.

However, the proof will contain two steps: first in Lemma 4, we replace $\mathbb{R}^n$ by $S_\varepsilon \in \mathbb{R}^{n-1}$ by finite covering of the surface $S_\varepsilon$ and approximation of a function $u_\varepsilon$ on $S_\varepsilon$ by $\{\rho_j\}_{j\in\mathrm{i}} \in C^\infty$-partition of unity subordinate to the finite covering. It is known from, e.g., [CoRo00], that $\|u_\varepsilon\|_{H^{-1/2+\alpha}(S_\varepsilon)}$ is equivalent to $\sum_{j\in\mathrm{i}}\|\rho_j u_\varepsilon J_j\|_{H^{-1/2+\alpha}(\mathbb{R}^n)}$.

Let us now define the diffeomorphism as follows: at each point of the interface translate the global coordinate system to the point on the boundary and rotate it in such a way that $x_n$-direction coincides with the outer normal of the boundary $S_x$, like in [CoRo00] or [EcJaKr05]. Then, the transformation will be given by

$$\mathbf{S}_x(x):\ \bar{x}\mapsto a_x+\omega_x\bar{x},\;\; \bar{x}\in\mathbb{R}^{n-1}\;\text{ and }\; x_n\mapsto a_x+\omega_x\mathbf{S}'_x(\bar{x}).$$

Here $a_x$ is a bounded translation vector and $\omega_x$ is a bounded rotation matrix, depending just on the point $x$ on the surface and $\mathbf{S}_x \in C^{1,\beta}(R^n)$, since $S \in C^{1,\beta}$.

According to [CoRo00], there exists a finite covering and a partition of unity to approximate function $\mathscr{T}_\varepsilon\sigma_\nu$ on $S$, such that the estimate (38.8) will be valid for curved $S \in C^{1,\beta}(R)$.

The following tools are well known in the literature on shape derivatives (see, for example, [DeZo83]) and mathematical theory of nonlinear elasticity [Ci88].

For an arbitrary (sufficiently regular) domain $\Omega$, part (sufficiently regular) of its boundary $\Gamma$, and a (sufficiently regular) diffeomorphism $\mathbf{S}_x$, the volume integral transformation rule is

$$\int_{\mathbf{S}_x(\Omega)} \mathbf{f}(\mathbf{y})\, d\mathbf{y} = \int_{\Omega} \mathbf{f} \circ \mathbf{S}_x(\mathbf{x})\, |\det(\nabla \mathbf{S}_x(\mathbf{x}))|\, d\mathbf{x},$$

the surface integral transformation rule is

$$\int_{\mathbf{S}_x(\Gamma)} \mathbf{k}(\mathbf{s}_\mathbf{y})\, d\mathbf{s}_\mathbf{y} = \int_{\Gamma} \mathbf{k} \circ \mathbf{S}_x(\mathbf{s}_\mathbf{s})\, |\mathrm{cof}(\nabla \mathbf{S}_x(\mathbf{s}_\mathbf{x}))\nu|\, d\mathbf{s}_\mathbf{x},$$

the function's gradient and symmetrized gradient transformation rule is

$$\nabla \mathbf{u} \circ \mathbf{S}_x = \nabla(\mathbf{u} \circ \mathbf{S}_x)(\mathbf{S}_x)^{-1},$$

$$e(\mathbf{u}) \circ \mathbf{S}_x = \frac{1}{2}\left(\nabla(\mathbf{u} \circ \mathbf{S}_x)(\nabla \mathbf{S}_x)^{-1} + \left(\nabla(\mathbf{u} \circ \mathbf{S}_x)(\nabla \mathbf{S}_x)^{-1}\right)^T\right),$$

where cofactor matrix is defined as $\mathrm{cof}(A) = \det(A) A^{-T}$.

And the next step deals with the estimate for the Jacobi matrix for the transformation, depending on the small parameter $\varepsilon$:

**Lemma 5.** *Let* $\mathbf{S}_{x\,\varepsilon}(x) : \bar{x} \mapsto a_x + \omega_x \bar{x}, \ \bar{x} \in \mathbb{R}^{n-1}$ *and* $x_n \mapsto a_x + \omega_x \mathbf{S}'_{x\,\varepsilon}(\bar{x})$ *with* $\mathbf{S}_{x\,\varepsilon}(\bar{x}) = \varepsilon \mathbf{S}_x(\frac{\bar{x}}{\varepsilon})$. *Then the Jacobi and the cofactor matrix for this transformation, will be the finite matrices, depending just on the point on the surface x, but not on* $\varepsilon$.

*Proof.* The proof is based on a simple computation $\nabla\left(\varepsilon \mathbf{S}_x\left(\frac{x}{\varepsilon}\right)\right) = (\nabla_y \mathbf{S}_x)\left(\frac{x}{\varepsilon}\right)$. See also [ShOrPa] for more details.

Now we can replace $\mathbb{R}^n$ by $S_e \cap \varepsilon Y$ in Lemma 4 and sum up over all cells, $N \approx \varepsilon^{-n}$.

**Lemma 6.** *For real* $\alpha$,

$$||u_\varepsilon(x)||^2_{H^\alpha(S_\varepsilon)} = \varepsilon^{-1} \int_{S_\varepsilon} (1 + \varepsilon^{-2\alpha} c_{n-1,\zeta}(\alpha) |\zeta|^{2\alpha}) |\hat{u_\varepsilon}(\zeta)|^2 d\zeta,$$

$$c_{n,\varepsilon}(\alpha) = \varepsilon^{-1-2\alpha} c_n(\alpha).$$

*The unfolded semi-norm is then*

$$\|\mathscr{T}_\varepsilon u(x,y)\|'^2_{L^2(\Omega, H^\alpha(S))} = \varepsilon^{1+2\alpha} \|u_\varepsilon(x)\|'^2_{H^\alpha(S_\varepsilon)},$$

**Corollary 1.** *So, the scaling of the unfolding operator in the semi-norm of the Bessel potentials coincides with the one for the semi-norm in the Sobolev–Slobodetski spaces, but it allows to scale the norms also for negative* $\alpha$.

## 38.3 Boundedness of the Solution and the Normal Conormal Derivatives on the Contact Interface

Boundedness and compactness results for stationary periodic contact problems with the Tresca friction were obtained in [CiDaOr13] under some restrictions on the volume force given on the inclusions. In this section we fix a tangential rigid rotation on a piece of the boundary, in order to be able to use the Korn's inequality without traces and concentrate ourselves on the boundedness w.r.t. $\varepsilon$. Further, we estimate the normal component of the conormal derivative (tractions) on the contact interface by the solution and the normal traction from the previous step.

**Proposition 2.** *Let* $r_\tau$ *be fixed on* $\gamma_\varepsilon \subset S_\varepsilon$ *with* $\mathrm{meas}\gamma > 0$*, or the force on the inclusions will be orthogonal to the rigid displacements. Then there exists a fixed constant C such that for u element of a minimizing sequence*

$$\left\|e(u_\varepsilon)\right\|_{L^2(\Omega_\varepsilon\setminus S_\varepsilon)} \le C\big(\|\bar f_\varepsilon\|_{L^2(\Omega_\varepsilon)} + \varepsilon^{-1/2}\left\|(g^\varepsilon)\right\|_{L^2((S^j_\varepsilon)}\big),$$

*Proof.* We start with the usual estimate obtained for the variational inequality, by taking into account that $v = 0$ belongs to $\mathscr{K}^\varepsilon$,

$$\overline{\alpha}\sum_{j=0}^{m}\|e(u^j)\|^2_{L^2(\Omega^j_\varepsilon)} \le \int_{\Omega^0_\varepsilon}\bar f^0_\varepsilon\, u^0\,dx + \sum_{j=1}^{m}\int_{\Omega^j_\varepsilon} f^j_\varepsilon\, u^j\,dx.$$

Furthermore, the friction term is bounded from below by zero.

Owing to Korn's inequality

$$\overline{\alpha}\sum_{j=0}^{m}\|e(u^j)\|^2_{L^2(\Omega^j_\varepsilon)} \le C\left(\sum_{j=0}^{m}\|f^j_\varepsilon\|_{L^2(\Omega^j_\varepsilon)}\,\|e(u^j)\|_{L^2(\Omega^0_\varepsilon)} + \varepsilon\sum_{j=1}^{m}\|f^j_\varepsilon\|_{L^2(\Omega^j_\varepsilon)}\,\|e(u^j)\|_{L^2(\Omega^j_\varepsilon)} + \varepsilon^{-1/2}\|g^\varepsilon_j\|_{L^1(S^j_\varepsilon)}\right).$$

Estimate from (38.8) is obtained simply by multiple applications of the classical inequality $ab \le \ell\, a^2 + (1/(4\ell))b^2$ (for $\ell > 0$).

The next assertion is a corollary of the preceding estimate and was established in [GaKnNe14]. Here we modify it to account for jumps.

**Lemma 7.** *Owing to the preliminary estimate the following estimates hold:*

$$\frac{1}{\varepsilon^2|Y|}\|[\mathscr{T}^b_\varepsilon u_\varepsilon]\|^2_{L^2(\Omega\times S)} = \frac{1}{\varepsilon}\|[u_\varepsilon]\|^2_{L^2(S_\varepsilon)} \le C,$$

$$\frac{1}{\varepsilon^2}\|[\mathscr{T}^b_\varepsilon u_\varepsilon]\|^2_{L^2(\Omega,H^{1/2}(S)} = \frac{1}{\varepsilon^2}\|[\mathscr{T}^b_\varepsilon u_\varepsilon]\|^2_{L^2(\Omega\times S)} + \frac{1}{\varepsilon^2}\int_\Omega\int_S\int_S \frac{|[\mathscr{T}^b_\varepsilon u_\varepsilon(x,y)] - [\mathscr{T}^b_\varepsilon u_\varepsilon(x,z)]|^2}{|y-z|^n} d\sigma_y d\sigma_z dx \le C.$$

*Proof.* For the first statement we use just the definition of the unfolding operator on the boundary, the scaled jump theorem

$$\frac{1}{\varepsilon}\|[u_\varepsilon]\|^2_{L^2(S_\varepsilon)} \le \gamma\|\nabla u_\varepsilon\|^2_{L^2(\Omega_\varepsilon\setminus S)}$$

and the preliminary estimate from the previous theorem.

The second inequality can again be rescaled

$$\frac{1}{\varepsilon^2}\int_\Omega\int_S\int_S \frac{|[\mathscr{T}^b_\varepsilon u_\varepsilon(x,y)] - [\mathscr{T}^b_\varepsilon u_\varepsilon(x,z)]|^2}{|y-z|^n} d\sigma_y d\sigma_z dx = \int_{S_\varepsilon}\int_{S_\varepsilon} \frac{|[u_\varepsilon(y)] - [u_\varepsilon(z)]|^2}{|y-z|^n} d\sigma_y d\sigma_z \le \varepsilon^2\gamma_2||\nabla u_\varepsilon||^2_{L^2(\Omega_\varepsilon)} \le C.$$

**Proposition 3.** *Let in Prop. 2 additionally* $\|\mu\|_{L^\infty(S)} < \frac{\overline{\alpha}}{\gamma_0\gamma_1 C_A}$. *Then also*

$$\|\mathscr{T}_\varepsilon(\sigma_\nu(u_\varepsilon))\|_{L^2(\Omega,H^{-1/2}(S))} = \varepsilon^{1/2}\big\|\sigma_\nu(u_\varepsilon)\big\|_{H^{-1/2}(S_\varepsilon)} \le C_1\big(\|\bar{f}_\varepsilon\|_{L^2(\Omega_\varepsilon)} + \sqrt{\varepsilon}\big\|(g^\varepsilon)\big\|_{L^2((S^j_\varepsilon)}\big), \tag{38.8}$$

*Proof.* Like in (38.6) we estimate

$$\overline{\alpha}\sum_{j=0}^m \|e(u^j)\|^2_{L^2(\Omega^j_\varepsilon)} \le C\left(\sum_{j=0}^m \|f^j_\varepsilon\|_{L^2(\Omega^j_\varepsilon)}\,\|e(u^j)\|_{L^2(\Omega^0_\varepsilon)} + \varepsilon\sum_{j=1}^m \|f^j_\varepsilon\|_{L^2(\Omega^j_\varepsilon)}\,\|e(u^j)\|_{L^2(\Omega^j_\varepsilon)} + \varepsilon^{1/2}\|g^\varepsilon_j\|_{L^1(S^j_\varepsilon)}\right) + \|\mu_\varepsilon\|_{H^{-1/2}(S_\varepsilon)\to H^{-1/2}(S_\varepsilon)}\,\gamma_0\,\big\|\sigma_\nu(u^i_\varepsilon)\big\|_{H^{-1/2}(S_\varepsilon)}\,\big\|[u^i_\varepsilon)]_\tau\big\|_{H^{1/2}(S_\varepsilon)}.$$

Using the direct scaled jump theorem in the last term for estimating the interface tangential jump by the semi-norm in $H^1$, we arrive at the estimate

$$\overline{\alpha}\sum_{j=0}^{m}\|e(u^j)\|_{L^2(\Omega_\varepsilon^j)} \le C\left(\sum_{j=0}^{m}\|f_\varepsilon^j\|_{L^2(\Omega_\varepsilon^j)} + \varepsilon^{1/2}\|(g_j^\varepsilon)\|_{L^1(S_\varepsilon^j)}\right).$$
$$+\sqrt{\varepsilon}\,\|\mu_\varepsilon\|_{L^\infty(S)\to H^{-1/2}(S_\varepsilon)}\,\gamma_0\left\|\sigma_\nu(u_\varepsilon^i)\right\|_{H^{-1/2}(S_\varepsilon)}.$$

Relation (38.1) together with the inverse scaled jump theorem with the constant $\varepsilon^{-1/2}\gamma_1$ enables us to obtain that

$$\|\sigma_\nu(u^i)\|_{H^{-1/2}(S_\varepsilon)} \le \varepsilon^{-1/2}\,\gamma_1(\|f_\varepsilon^i\|_{L^2(\Omega_\varepsilon)} + C_A\|e(u_\varepsilon^i)\|_{L^2(\Omega_\varepsilon\setminus S_\varepsilon)}).$$

substituting the last inequality in the previous one provides the estimate.

## 38.4 Homogenization

The aim of this section now is to pass to the limit as $\varepsilon \to 0$ in our problem. We will obtain a limit "homogenized" problem that is given in Theorem 4 below. We omit here the rigid rotation of particles, which were considered in [CiDaOr13] in detail and whose convergence leads to the convergence of measures. Our main achievement in this section is an extension of the previous results to the convergence of the co-normal derivatives on the oscillating interface, which justify the convergence of the interface traces of the stresses and can be used for the convergence proof for the Coulomb friction.

As mentioned in the Introduction, we use for the proof the unfolding method and the results from the sections above.

For simplicity, the notation $W^1_{per}(Y^0)$ indicates the subspace of $Y$-periodic elements of $W^1(Y^0)$.

**Proposition 4.** *Up to a subsequence, there exists*

$$u^0 \in H^1(\Omega;\Gamma_D), \quad \widehat{u}^0 \in L^2(\Omega;W^1_{per}(Y^0))$$

*such that*

$$(i)\ \ \mathcal{T}_\varepsilon(u_e^0) \to u^0 \quad \textit{strongly in } L^2_{loc}(\Omega;H^1(Y^0)),$$

$$(ii)\ \ \mathcal{T}_\varepsilon(e(u_\varepsilon^0)) \rightharpoonup e(u^0)+e_y(\widehat{u}^0) \quad \textit{weakly in } L^2(\Omega\times Y^0),$$

$$(iii)\ \ \mathcal{T}_\varepsilon(a^\varepsilon e(u_\varepsilon^0)) \rightharpoonup a^0(e(u^0)+e_y(\widehat{u}^0)) \quad \textit{weakly in } L^2(\Omega\times Y^0),$$

$$(iv)\ \ \frac{1}{\varepsilon}[\mathcal{T}_\varepsilon(u_\varepsilon^0)]_{S^0} \rightharpoonup [\widehat{u}^0]_{S^0} \quad \textit{weakly in } L^2(\Omega;H^{1/2}(S^0)),$$

*and consequently*

$$[\widehat{u}_\nu^0]_{S^0} \le g^0 \text{ on } S^0.$$

The proof is given in [CiDaOr13].

**Proposition 5.** *Let* $\frac{\partial u_\varepsilon}{\partial \nu} \equiv A_\varepsilon \nabla u_\varepsilon \equiv \sigma_\nu(u_\varepsilon)$, *and let* $\mathscr{T}_\varepsilon^b\left(\frac{\partial u_\varepsilon}{\partial \nu}\right)$ *be bounded in the space* $L^2(\Omega, H^{-1/2}(S))$. *Then*

$$\mathscr{T}_\varepsilon^b\left(\frac{\partial u_\varepsilon}{\partial \nu}\right) \rightharpoonup a^0(x,y)(\nabla u_0 + \nabla_y \hat{u}(x,y))\nu_y(x,y) \quad \text{in } L^2(\Omega, H^{-1/2}(S)).$$

*Proof.* According to the Gauss identity,

$$(\sigma_\nu(u_\varepsilon), \phi_\varepsilon)_{S_\varepsilon} = a(u_\varepsilon, \phi_\varepsilon)_{\Omega_\varepsilon} - (f_\varepsilon, \phi_\varepsilon)_{\Omega_\varepsilon}$$

Applying the unfolding operators to both sides of the identity, we find that for all $\phi_\varepsilon \in H^1(\Omega^\varepsilon)$,

$$\frac{1}{\varepsilon|Y|}\int_\Omega\int_S \mathscr{T}_\varepsilon^b(\sigma_\nu(u_\varepsilon))\mathscr{T}_\varepsilon^b(\phi_\varepsilon)d\sigma_y dx$$
$$\overset{\text{Gauss in } Y}{=} \frac{1}{|Y|}\int_\Omega\int_Y \mathscr{T}_\varepsilon(f_\varepsilon)\mathscr{T}_\varepsilon(\phi_\varepsilon) + \mathscr{T}_\varepsilon(A_\varepsilon)\mathscr{T}_\varepsilon(\nabla u_\varepsilon)\mathscr{T}_\varepsilon(\nabla\phi_\varepsilon)dydx.$$

Let us choose $\phi_\varepsilon = \varepsilon\psi(x, \frac{x}{\varepsilon})$ and pass to the limit with respect to $\varepsilon \to 0$ in each term of the last expression,

$$\frac{1}{|Y|}\int_\Omega\int_Y a^0(\nabla u_0 + \nabla_y \hat{u})\nabla_y \psi dydx = \frac{1}{|Y|}\int_\Omega\int_S L(x,y)\psi(x,y)d\sigma_y dx,$$

and with respect to Green's formula the co-normal derivative $L(x,y) = a^0(\nabla u_0 + \nabla_y \hat{u})\nu_y(x,y)|_S$.

**Corollary 2.** *Let*

$$\mathscr{T}_\varepsilon\left(\frac{\partial u_\varepsilon}{\partial \nu}\right) \rightharpoonup a^0(x,y)(\nabla u_0 + \nabla_y \hat{u}(x,y))\nu_y(x,y) \quad \text{in } L^2(\Omega), H^{-1/2}(S)).$$

*and*

$$\frac{1}{\varepsilon}\mathscr{T}_\varepsilon[u_\varepsilon] \rightharpoonup [\hat{u}](x,y) \quad \text{in } L^2(\Omega), H^{1/2}(S)).$$

*If additionally one of the convergences is strong, then*

$$\lim_{\varepsilon\to 0}\int_{\Omega_\varepsilon}\frac{\partial u_\varepsilon}{\partial \nu}[u_\varepsilon](x)dx = \int_\Omega\int_S a^0(x,y)(\nabla u_0 + \nabla_y \hat{u}(x,y))\nu_y(x,y)[\hat{u}(x,y)]dxdy.$$

*Furthermore,*

$$\lim_{\varepsilon\to 0}\int_{\Omega_\varepsilon}\sigma_{\varepsilon\ \nu}|[u_\varepsilon]_\tau|(x)dx=\int_\Omega\int_S\sigma_\nu(x,y)|[\hat{u}]_\tau(x,y)|dxdy,$$

*where* $\sigma_{\varepsilon\ \nu}=\frac{\partial u_\varepsilon}{\partial\nu}\cdot\nu_\varepsilon(x))$ *and* $\sigma_\nu(x,y)=a^0(x,y)(\nabla u_0+\nabla_y\hat{u}(x,y))\nu_y(x,y)\cdot\nu_y(x,\xi)$

*Remark 1.* The strong convergence of the trace of the solution on the oscillating interface was proven in [GaKnNe14] by the shifting technique. So, for the static contact problem we can pass to Section 38.6.

However, for the quasi-static problem, it is necessary to estimate the conormal derivatives (normal interface tractions) by the normal tractions from the previous step and obtain a fixed point result as in [EcJaKr05, CoRo00] uniformly in $\varepsilon$.

## 38.5 Boundedness under Additional Regularity Assumptions

We assume a better regularity for the contact interface, friction coefficient, and the elastic coefficients in a $\delta$-neighborhood of the contact interface:

**Assumption 38.5.1** *$S\in C^{1,\beta}(\mathbb{R})$, $a_{ijkl}\in L^\infty(\cup_{j=0}^m Y_j)$, furthermore, $a_{ijkl}\in C^{0,\alpha}$, $0<\alpha<1/2$ in a $\delta$-neighborhood of $S$, $\mu\in C^1(S)$ with compact support in $S$. Furthermore, let coefficients $||\mu||_{L^\infty(S)}<\frac{\overline{\alpha}}{C_A\gamma_0\gamma_1}$, where $C_A$, $\overline{\alpha}$ are the constants from the continuity and coercivity condition for the elastic bilinear form, $\gamma_0$ and $\gamma_1$ are the constants from the jump and inverse jump inequalities, $\mathcal{T}_\varepsilon(f_\varepsilon)\in H^\alpha(\Omega\times Y)$, $\mathcal{T}_\varepsilon(g^\varepsilon)\in H^{1/2+\alpha}(\Gamma_D\cup S)$.*

Further, we estimate the normal component of the co-normal derivative in the space $H^{-1/2+\alpha}$ by shifting argument. This technique was used for the proof of the existence of the solution to the contact problems with Coulomb friction in [Ja83, Ro99, CoRo00].

Let us fix $\varepsilon$. For a fixed domain, the following assertion was proved in [CoRo00], Lemmas 2.5 and 2.6, and in [EcJaKr05], Sect. 1.7.2.

**Lemma 8.** *Let $u$ be a solution of the contact problem with $G\in H^{-1/2+\alpha}(S)\cap C^1(\mathbb{R})$ — normal friction traction on the contact interface known form the previous step. Then for an arbitrary $\delta>0$ we have*

$$\|\sigma_\nu(u)\|_{H^{-1/2+\alpha}(\mathbb{R}^{n-1})}\le(1+\delta)\left(\int_{\mathbb{R}^{n-1}}\frac{C_A a(u_{-h}-u,u_{-h}-u)}{c_{n-1}(\alpha)c_{n-1}\left(\frac{1}{2}-\alpha\right)|h|^{n-1+2\alpha}}dh\right)^{\frac{1}{2}}$$
$$+k_1(\delta,\alpha)\left[\left\|f^{i+1}\right\|_{L^2(\Omega)}+\|u\|_{H^1(\Omega_s)}\right].$$

*and*

$$\left(\frac{\bar{\alpha}a(u_{-h}-u,u_{-h}-u)}{2c_{n-1}(\alpha)c_{n-1}\left(\frac{1}{2}-\alpha\right)|h|^{n-1+2\alpha}}dh\right)^{\frac{1}{2}} \leq (1+\delta)\,\|\mu\|_{L^\infty(S)}\,\|G\|_{H^{-1/2+\alpha}(\mathbb{R}^{n-1})}$$
$$+k_2(\delta,\alpha)\left[\left\|f^{i+1}\right\|^2_{L^2(\Omega)}+\|u\|^2_{H^1(\Omega_s)}\right].$$

**Theorem 2.** *Under the regularity Assumption* (38.5.1)*,*

$$\varepsilon^{1/2}||\sigma_\nu(u_\varepsilon)||_{H^{-1/2+\alpha}(S_e)} \leq const(C_A,\mu,||\bar{f}_\varepsilon||_{L^2(\Omega)}+\sqrt{\varepsilon}||g_\varepsilon||_{H^{1/2}(S_e,\partial\Omega_D)}).$$

*Proof.* We just scale the previous Lemma with respect to the auxiliary scaling rules recalled and derived in the beginning of the paper. Owing to Lemma 6,

$$c_{n-1,\,\varepsilon}(\alpha)c_{n-1,\,\varepsilon}(1/2-\alpha)=\varepsilon^{-1}c_{n-1}(\alpha)c_{n-1}(1/2-\alpha).$$

Furthermore, we can scale constants in inequalities (38.4)–(38.5):

$$d_{n\,\varepsilon}(\alpha,\beta)=\frac{\varepsilon^{-2-2(\alpha+\beta)}}{\varepsilon^{-1-2(\alpha+\beta)}}d_n(\alpha,\beta)=\varepsilon^{-1}d_n(\alpha,\beta).$$

$$d^*_{n\,\varepsilon}(\alpha,\beta)=\frac{\varepsilon^{-2-2(\alpha+\beta-\alpha}}{\varepsilon^{-1-2(\beta)}}d^*_n(\alpha,\beta)=\varepsilon^{-1}d^*_n(\alpha,\beta).$$

We overcome from $\mathbb{R}^{n-1}$ to $S^\varepsilon$, according to Lemma 6. Application of scaled inequality (38.4) together with the scaled constants $c_n$ in the previous lemma and combination of it with Lemma 38.8 completes the proof.

**Theorem 3. (Equicontinuity of the solution)** *Let* $\|\mu\|_{L^\infty(S_\varepsilon)} < \overline{\alpha}/(C_A\gamma_0,\gamma_1)$*, where* $C_A$*,* $\overline{\alpha}$ *are the constants from the continuity and coercivity condition for the elastic bilinear form,* $f_\varepsilon \in C^1([0,T],L^2(\Omega_\varepsilon))$*,* $g^\varepsilon \in C^1([0,T],L^2(S_\varepsilon))$*,* $a_{ijkl} \in L^\infty(\Omega_\varepsilon)$*, furthermore,* $a_{ijkl} \in C^{0,\alpha}$*,* $0<\alpha<1/2$ *in a neighborhood of* $S_\varepsilon$*,* $\mu \in C^1(S_\varepsilon)$ *with compact support in* $S_\varepsilon$*, and the initial values* $u^\varepsilon_0$ *be stable. Then any solution of problem* $\mathscr{P}_\varepsilon$ *is Lipschitz continuous and the constant in the bound is independent of* $\varepsilon$*.*

Proof is similar to the one from [Or02, Or00].

## 38.6 Homogenized Problem

Let us now consider the following problem, called *corrector problem*: for every symmetric tensor $\mathscr{U}$ and for a.e. $x \in \Omega$, find $\chi(x,\cdot)$ in $\widehat{K}$ such that

$$\frac{1}{|Y|}\int_{Y^0} a^0\big(\mathscr{U}+e_y(\chi)\big)\big(e_y(W)-e_y(\chi)\big)\,dy$$
$$+\psi^0\big(a^0\big(\mathscr{U}+e_y(\chi)\big)_{\nu},[(W^0)_\tau]_{S^0}\big)-\psi^0\big(a^0\big(\mathscr{U}+e_y(\chi)\big)_{\nu},[(\chi)_\tau]_{S^0}\big)\ \geq 0,$$

for all $W$ in $\widehat{\mathscr{K}}$, the convex set defined by

$$\widehat{\mathscr{K}} \doteq \big\{w \mid\ w\in L^2(\Omega;H^1_{per}(Y^0)),\mathscr{M}_{Y^0}(w)=0,$$
$$[w]_{\nu|_{S^0}} \leq g^0 \text{ in } L^2(\Omega;H^{1/2}(S^0)).\big\}$$

This problem is equivalent to minimizing over the set $\widehat{K}$ the functional

$$\frac{1}{2|Y|}\int_{Y^*} a^0\big(\mathscr{U}+e_y(W)\big)\big(\mathscr{U}+e_y(W)\big)\,dy+\psi^0\big(a^0\big(\mathscr{U}+e_y(\chi)\big)_{\nu},[(W)_\tau]_{|S^0}\big). \tag{38.9}$$

The corrector problem (38.9) has at least one solution $\chi$.

**Theorem 4.** *Let $u_\varepsilon \in \mathscr{K}^\varepsilon$ be solution of Problem $\mathscr{P}'''_\varepsilon$ and $u^0 \in H^1(\Omega;\Gamma_D)$ be the limit function. Then there exists a strictly convex Carathéodory function $\mathscr{E}^{hom}$, defined on $\Omega\times M^S_3(\mathbb{R})$[1] such that $u^0$ is a minimizer over $H^1(\Omega;\Gamma_D)$ of the functional*

$$\int_\Omega \big(\mathscr{E}^{hom}(x,e(v))-Fv\big)\,dx.$$

*The functional $\mathscr{E}^{hom}$ is the minimum in (38.9), i.e.*

$$\mathscr{E}^{hom}(x,\mathscr{U})$$
$$\doteq \frac{1}{2|Y|}\int_{Y^*} a^0\big(\mathscr{U}+e_y(\chi)\big)\big(\mathscr{U}+e_y(\chi)\big)\,dy+\psi\big(a^0\big(\mathscr{U}+e_y(\chi)\big)_{\nu},[(\chi)_\tau]_{|S^0}\big)$$

*for every symmetric tensor $\mathscr{U}$ and for a.e. $x\in\Omega$, and the corrector function $\chi(x,\cdot)$ in $\widehat{K}$ is a solution of (38.9).*

And the following theorem extends the spacial convergences above to the uniform convergence in time. It can also be found as Prop. 4.3 in [MiTi07].

**Theorem 5.** *Let the right-hand side functions and elastic coefficients be Banach-valued or Sobolev-valued functions continuous in time $t\in[0,T]$, then it is possible to find a subsequence of the solution convergent for all time-points and this sequence will be continuous in $t\in[0,T]$.*

[1] This means that it is continuous with respect to its second argument and measurable with respect to its first one.

The proof, like in the original theorem, is based on the Arzela–Ascoli theorem for the weak topology and the Lebesgue's theorem of dominated convergence for the dissipative term, as the integrands are uniformly bounded and converge pointwise.

## References


[EcJaKr05] Eck, C., Jarušek, J., Krbec, M.: Unilateral contact problems variational methods and existence theorems. Springer (2005)

[GaKnNe14] Gahn, M., Knabner, P. & Neuss-Radu, M.: Homogenization of reaction-diffusion processes in a two-component porous medium with a nonlinear flux condition at the interface, and application to metabolic processes in cells. Preprint Angew. Math., Uni Erlangen, No. 384 (2014)

[CiEtAl12] Cioranescu, D., Damlamian, A., Donato, P., Griso, G., and Zaki, R.: The periodic unfolding method in domains with holes. SIAM J. of Math. Anal. Vol. 44, 2 (2012), 718–760

[MiTi07] Mielke, A., Timofte, A.M.: Two-scale homogenization for evolutionary variational inequalities via the energetic formulation. SIAM J. Math. Anal. (2007) WIAS Preprint 1172.

[CoRo00] Cocu M., Rocca, R.: Existence results for unilateral quasistaric contact problems with fricrtion and ahesion (2000)

[CiDaOr13] Cioranescu, D., Damlamian, A. & Orlik, J. : Homogenization via unfolding in periodic elasticity with contact on closed and open cracks. Asymptotic Analysis, Vol. 82, Issue 3–4 (2013)

[Ci88] Ciarlet, P.G.: Mathematical Elasticity: Three-dimensional elasticity. Volume 1. SIAM, Elsevier (1988)

[DeZo83] Delfour, M. C., Zolésio, J.-P.: Shapes and Geometries: Metrics, Analysis, Differential Calculus, and Optimization. Second Edition, SIAM (2011)

[Ja83] Jarušek J.: Contact problems with bounded friction coersive case. Czechoslovac Math. J. 33 (1983) pp. 237–261.

[Or00] Orlik, J.: Transmission and homogenization in hereditary viscoelasticity with ageing and shrinkage. PhD-Thesis Shaker Verlag, 2000.

[Ro99] Rocca, R.: Existence of a solution of a quasistatic problem of unilateral contact with local friction, C. R. Acad. Sci., Paris, Ser. I, 328 (1999) 1253–1258.

[Mi14] Mikhailov, S.E. : Dependence of Norms of Localized Boundary-Domain Integral Operators on Scaling, IMSE-Conference, Karlsruhe (2014)

[Or02] Orlik, J.: Homogenization for Viscoelasticity, Progress in industrial mathematics at ECMI 2000. Angelo Marcello Anile…ed.., Berlin, Heidelberg, New York,…: Springer, 618–624 (2002)

[ShOrPa] Shiryaev, V., Orlik, J., Panasenko, G.: Optimization of textile-like materials via homogenization and beam approximations, in preparation.

# Chapter 39
# Piecewise Polynomial Collocation for a Class of Fractional Integro-Differential Equations

**A. Pedas, E. Tamme, and M. Vikerpuur**

## 39.1 Fractional Integro-Differential Equation

We consider a linear fractional integro-differential equation of the form

$$(D_*^{\alpha_p} y)(t) + \sum_{i=0}^{p-1} h_i(t)(D_*^{\alpha_i} y)(t) + \int_0^t K(t,s)y(s)ds = f(t),\ 0 \le t \le b, \tag{39.1}$$

with

$$\gamma_0 y(0) + \sum_{k=1}^{l} \gamma_k y(b_k) = \gamma,\ 0 < b_1 < \cdots < b_l \le b,\ \gamma, \gamma_k \in \mathbb{R} := (-\infty, \infty). \tag{39.2}$$

Here

$$p \in \mathbb{N} = \{1, 2, \dots\}$$

and $D_*^{\alpha_i} y$ $(i = 0, \dots, p)$ are the Caputo fractional derivatives of $y$ of order $\alpha_i$ with

$$0 = \alpha_0 < \alpha_1 < \cdots < \alpha_p < 1; \tag{39.3}$$

in the case $\alpha_0 = 0$ we set $D_*^0 = I$ where $I$ is the identity mapping.

We assume that $f, h_i$ $(i = 0, \dots, p-1)$ and $K$ are some given continuous functions on $[0,b]$ and $\Delta$, respectively: $f, h_i \in C[0,b]$ $(i = 0, \dots, p-1)$ and $K \in C(\Delta)$,

$$\Delta = \{(t,s) : 0 \le s \le t \le b\}.$$

A. Pedas (✉) • E. Tamme • M. Vikerpuur
University of Tartu, Tartu, Estonia
e-mail: Arvet.Pedas@ut.ee; Enn.Tamme@ut.ee; azzo@ut.ee



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_39

The Caputo differential operator $D_*^\alpha$ of order $\alpha \in (0,1)$ is defined by (see, e.g., [Di10])

$$(D_*^\alpha y)(t) := (D^\alpha [y - y(0)])(t),\ t > 0.$$

Here $D^\alpha y$ is the Riemann–Liouville fractional derivative of $y$:

$$(D^\alpha y)(t) := \frac{d}{dt}(J^{1-\alpha} y)(t), \quad 0 < \alpha < 1, \quad t > 0,$$

with $J^\beta$, the Riemann–Liouville integral operator, defined by the formula

$$(J^\beta y)(t) := \frac{1}{\Gamma(\beta)} \int_0^t (t-s)^{\beta-1} y(s)\, ds, \quad t > 0, \quad \beta > 0, \tag{39.4}$$

where $\Gamma$ is the Euler gamma function.

It is well known (see, e.g. [BrPeVa01]) that $J^\beta$, $\beta > 0$, is linear, bounded, and compact as an operator from $L^\infty(0,b)$ into $C[0,b]$, and we have for any $y \in L^\infty(0,b)$ that (see, e.g. [KiSrTr06])

$$J^\beta y \in C[0,b], \quad (J^\beta y)(0) = 0, \quad \beta > 0, \tag{39.5}$$

$$D^\delta J^\beta y = D_*^\delta J^\beta y = J^{\beta-\delta} y, \quad 0 < \delta \le \beta, \tag{39.6}$$

with $J^0 = I$.

Fractional differential equations arise in various areas of science and engineering. In the last few decades theory and numerical analysis of fractional differential equations have received increasing attention (see, e.g. [Di10, KiSrTr06, Po99, AgBeHa10]). Some recent results about the numerical solution of fractional differential equations can be found in [Di10, PeTa11, FoMo11, FoMo14, DoBhEz11, PeTa12, PeTa14a, PeTa14b, MaHu14, YaPaFo14].

In this chapter, the numerical solution of (39.1)-(39.2) by piecewise polynomial collocation techniques is considered. We use an integral equation reformulation of the problem and special non-uniform grids reflecting the possible singular behavior of the exact solution. Our aim is to study the attainable order of the proposed algorithms in a situation where the higher-order (usual) derivatives of $h_i(t)\,(i = 0,\ldots,p-1)$ and $f(t)$ may be unbounded at $t = 0$. Our approach is based on some ideas and results of [PeTa12]. In particular, the case where (39.1)-(39.2) is an initial value problem ($\gamma_0 \neq 0$, $\gamma_1 = \cdots = \gamma_l = 0$), a boundary value problem ($\gamma_0 \neq 0, \gamma_l \neq 0, b_l = b$) or a terminal value problem ($\gamma_0 = \cdots = \gamma_{l-1} = 0, \gamma_l \neq 0, b_l = b$, see [FoMo11, FoMo14]) is under consideration.

## 39.2 Existence and Regularity of the Solution

In order to characterize the behavior of higher-order derivatives of a solution of equation (39.1), we introduce a weighted space of smooth functions $C^{q,\nu}(0,b]$ (cf., e.g., [BrPeVa01, Va93]). For given $q \in \mathbb{N}$ and $\nu \in \mathbb{R}$, $\nu < 1$, by $C^{q,\nu}(0,b]$ we denote the set of continuous functions $y : [0,b] \to \mathbb{R}$ which are $q$ times continuously differentiable in $(0,b]$ and such that for all $t \in (0,b]$ and $i = 1,\dots,q$ the following estimates hold:

$$\left|y^{(i)}(t)\right| \le c \begin{cases} 1 & \text{if } i < 1-\nu, \\ 1+|\log t| & \text{if } i = 1-\nu, \\ t^{1-\nu-i} & \text{if } i > 1-\nu. \end{cases}$$

Here $c = c(y)$ is a positive constant.

Clearly,

$$C^q[0,b] \subset C^{q,\nu}(0,b] \subset C^{m,\mu}(0,b] \subset C[0,b], \quad q \ge m \ge 1, \quad \nu \le \mu < 1. \tag{39.7}$$

Note also that a function of the form

$$y(t) = g_1(t)\, t^{\mu} + g_2(t)$$

is included in $C^{q,\nu}(0,b]$ if $\mu \ge 1-\nu > 0$ and $g_j \in C^q[0,b]$, $j = 1,2$.

In what follows we use an integral equation reformulation of (39.1)-(39.2). Let $y \in C[0,b]$ be such that $D_*^{\alpha_p} y \in C[0,b]$. Introduce a new unknown function $z := D_*^{\alpha_p} y$. Then (see [Di10, KiSrTr06])

$$y(t) = (J^{\alpha_p} z)(t) + c, \; 0 \le t \le b, \tag{39.8}$$

where $c$ is an arbitrary constant. Denote

$$\gamma_* = \sum_{k=0}^{l} \gamma_k.$$

The function $y$ given by (39.8) satisfies (39.2) if and only if (see (39.5))

$$c\gamma_* = \gamma - \sum_{k=1}^{l} \gamma_k (J^{\alpha_p} z)(b_k).$$

In the sequel, we assume that $\gamma_* \neq 0$. Therefore

$$c = \frac{1}{\gamma_*} \left( \gamma - \sum_{k=1}^{l} \gamma_k (J^{\alpha_p} z)(b_k) \right). \tag{39.9}$$

Thus, the function $y$ of the form (39.8) satisfies the conditions (39.2) if and only if

$$y(t) = (J^{\alpha_p} z)(t) + \frac{1}{\gamma_*}\left(\gamma - \sum_{k=1}^{l} \gamma_k (J^{\alpha_p} z)(b_k)\right), \quad 0 \le t \le b. \tag{39.10}$$

Substituting (39.10) into (39.1) and using (39.6), (39.4), we obtain for $z$ an operator equation of the form

$$z = Tz + g, \tag{39.11}$$

with an operator $T$, defined by formula

$$\begin{aligned}(Tz)(t) = &-\frac{h_0(t)}{\Gamma(\alpha_p)}\left[\int_0^t (t-s)^{\alpha_p-1} z(s)ds - \frac{1}{\gamma_*}\sum_{k=1}^{l}\gamma_k \int_0^{b_k} (b_k - s)^{\alpha_p-1} z(s)ds\right] \\ &- \sum_{i=1}^{p-1} \frac{h_i(t)}{\Gamma(\alpha_p - \alpha_i)} \int_0^t (t-s)^{\alpha_p-\alpha_i-1} z(s)ds \\ &- \frac{1}{\Gamma(\alpha_p)} \int_0^t K(t,s) \int_0^s (s-\tau)^{\alpha_p-1} z(\tau)d\tau ds \\ &+ \frac{1}{\gamma_* \Gamma(\alpha_p)} \sum_{k=1}^{l} \gamma_k \int_0^{b_k} (b_k - s)^{\alpha_p-1} z(s)ds \int_0^t K(t,s)ds, \; 0 \le t \le b,\end{aligned} \tag{39.12}$$

and

$$g(t) = f(t) - \frac{\gamma}{\gamma_*}\left(h_0(t) + \int_0^t K(t,s)ds\right), \; 0 \le t \le b. \tag{39.13}$$

We observe that equation (39.11) is a linear weakly singular Fredholm integral equation of the second kind with respect to $z$.

The existence and regularity of a solution to (39.1)-(39.2) is described by the following theorem which can be proved similarly to Theorem 2.1 in [PeTa12].

**Theorem 1.** *Assume that* $K \in C^q(\Delta)$, $h_i \in C^{q,\mu}(0,b]\,(i = 0,\ldots,p-1), f \in C^{q,\mu}(0,b]$, $q \in \mathbb{N}$, $\mu \in \mathbb{R}$, $\mu < 1$. *Moreover, assume that* $\gamma_* = \sum_{k=0}^{l} \gamma_k \ne 0$ *and problem* (39.1)-(39.2) *with* $f = 0$ *and* $\gamma = 0$ *has in* $C[0,b]$ *only the trivial solution* $y = 0$.

*Then problem* (39.1)-(39.2) *possesses a unique solution* $y \in C[0,b]$ *such that* $D_*^{\alpha_p} y \in C^{q,\nu}(0,b]$, *where*

$$\nu := \max\{\mu, 1 - \alpha_p + \alpha_{p-1}\}. \tag{39.14}$$

## 39.3 Numerical Method

Let $N \in \mathbb{N}$ and let $\Pi_N := \{t_0, \ldots, t_N\}$ be a partition (a graded grid) of the interval $[0,b]$ with the grid points

$$t_j := b\left(\frac{j}{N}\right)^r, \quad j = 0,1,\ldots,N, \tag{39.15}$$

where the grading exponent $r \in \mathbb{R},\ r \geq 1$. If $r = 1$, then the grid points (39.15) are distributed uniformly; for $r > 1$ the points (39.15) are more densely clustered near the left endpoint of the interval $[0,b]$.

For given integer $k \geq 0$ by $S_k^{(-1)}(\Pi_N)$ is denoted the standard space of piecewise polynomial functions :

$$S_k^{(-1)}(\Pi_N) := \left\{v : v\big|_{(t_{j-1},t_j)} \in \pi_k, j = 1,\ldots,N\right\}.$$

Here $v\big|_{(t_{j-1},t_j)}$ is the restriction of $v : [0,b] \to \mathbb{R}$ onto the subinterval $(t_{j-1},t_j) \subset [0,b]$ and $\pi_k$ denotes the set of polynomials of degree not exceeding $k$. Note that the elements of $S_k^{(-1)}(\Pi_N)$ may have jump discontinuities at the interior points $t_1,\ldots,t_{N-1}$ of the grid $\Pi_N$.

In every interval $[t_{j-1},t_j]$, $j = 1,\ldots,N$, we define $m \in \mathbb{N}$ collocation points $t_{j1},\ldots,t_{jm}$ by formula

$$t_{jk} := t_{j-1} + \eta_k(t_j - t_{j-1}), \quad k = 1,\ldots,m,\ j = 1,\ldots,N, \tag{39.16}$$

where $\eta_1 \ldots, \eta_m$ are some fixed (collocation) parameters which do not depend on $j$ and $N$ and satisfy

$$0 \leq \eta_1 < \eta_2 < \ldots < \eta_m \leq 1\,. \tag{39.17}$$

We look for an approximate solution $y_N$ to (39.1)-(39.2) in the form (cf. (39.10))

$$y_N(t) = (J^{\alpha_p} z_N)(t) + \frac{1}{\gamma_*}\left(\gamma - \sum_{k=1}^{l} \gamma_k (J^{\alpha_p} z_N)(b_k)\right), \quad 0 \leq t \leq b, \tag{39.18}$$

where $z_N \in S_{m-1}^{(-1)}(\Pi_N)$ $(m,N \in \mathbb{N})$ is determined by the following collocation conditions:

$$z_N(t_{jk}) = (Tz_N)(t_{jk}) + g(t_{jk}), \quad k = 1,\ldots,m,\ j = 1,\ldots,N. \tag{39.19}$$

Here $T, g$ and $t_{jk}$ are defined by (39.12),(39.13), and (39.16), respectively. If $\eta_1 = 0$, then by $z_N(t_{j1})$ we denote the right limit $\lim_{t \to t_{j-1}, t > t_{j-1}} z_N(t)$. If $\eta_m = 1$, then $z_N(t_{jm})$

denotes the left limit $\lim_{t\to t_j, t<t_j} z_N(t)$. Conditions (39.19) have an operator equation representation

$$z_N = \mathscr{P}_N T z_N + \mathscr{P}_N g \tag{39.20}$$

with an interpolation operator $\mathscr{P}_N = \mathscr{P}_{N,m} : C[0,T] \to S_{m-1}^{(-1)}(\Pi_N)$ defined for any $v \in C[0,b]$ by the following conditions:

$$\mathscr{P}_N v \in S_{m-1}^{(-1)}(\Pi_N),\ (\mathscr{P}_N v)(t_{jk}) = v(t_{jk}),\ k = 1,\ldots,m,\ j = 1,\ldots,N. \tag{39.21}$$

The collocation conditions (39.19) form a system of equations whose exact form is determined by the choice of a basis in $S_{m-1}^{(-1)}(\Pi_N)$. If $\eta_1 > 0$ or $\eta_m < 1$ then we can use the Lagrange fundamental polynomial representation:

$$z_N(t) = \sum_{\lambda=1}^{N} \sum_{\mu=1}^{m} c_{\lambda\mu} \varphi_{\lambda\mu}(t), \quad t \in [0,b], \tag{39.22}$$

where $\varphi_{\lambda\mu}(t) := 0$ for $t \notin [t_{\lambda-1}, t_\lambda]$ and

$$\varphi_{\lambda\mu}(t) := \prod_{i=1, i\neq\mu}^{m} \frac{t - t_{\lambda i}}{t_{\lambda\mu} - t_{\lambda i}} \quad \text{for} \quad t \in [t_{\lambda-1}, t_\lambda],\ \mu = 1,\ldots,m,\ \lambda = 1,\ldots,N.$$

Then $z_N \in S_{m-1}^{(-1)}(\Pi_N)$ and $z_N(t_{jk}) = c_{jk},\ k = 1,\ldots,m,\ j = 1,\ldots,N$. Searching the solution of (39.19) in the form (39.22), we obtain a system of linear algebraic equations with respect to the coefficients $c_{jk} = z_N(t_{jk})$:

$$c_{jk} = \sum_{\lambda=1}^{N} \sum_{\mu=1}^{m} (T\varphi_{\lambda\mu})(t_{jk}) c_{\lambda\mu} + g(t_{jk}), \quad k = 1,\ldots,m,\ j = 1,\ldots,N. \tag{39.23}$$

Note that this algorithm can be used also in the case if in (39.17) $\eta_1 = 0$ and $\eta_m = 1$. In this case we have

$$t_{jm} = t_{j+1,1} = t_j,\ c_{jm} = c_{j+1,1} = z_N(t_j), \quad j = 1,\ldots,N-1,$$

and hence in the system (39.23) there are $(m-1)N+1$ equations and unknowns.

## 39.4 Convergence Estimates

In this section we formulate two theorems about the convergence and convergence order of the proposed algorithms.

**Theorem 2.** (i) *Let $m \in \mathbb{N}$ and assume that the collocation points (39.16) with grid points (39.15) and arbitrary parameters $\eta_1, \ldots, \eta_m$ satisfying (39.17) are used. Assume that $h_i \in C[0,b]$ $(i = 0, \ldots, p-1)$, $f \in C[0,b]$ and $K \in C(\Delta)$. Moreover, assume that $\gamma_* = \sum_{k=0}^{l} \gamma_k \neq 0$ and the problem (39.1)-(39.2) with $f = 0$ and $\gamma = 0$ has in $C[0,b]$ only the trivial solution $y = 0$.*

*Then (39.1)-(39.2) has a unique solution $y \in C[0,b]$ such that $D_*^{\alpha_p} y \in C[0,b]$. Moreover, there exists an integer $N_0$ such that for all $N \geq N_0$ equation (39.20) possesses a unique solution $z_N \in S_{m-1}^{(-1)}(\Pi_N)$ and*

$$\|y - y_N\|_\infty \to 0 \quad \text{as} \quad N \to \infty$$

*where $y_N$ is defined by (39.18).*

(ii) *If, in addition, $h_i \in C^{m,\mu}(0,b]$ $(i = 0, \ldots, p-1)$, $f \in C^{m,\mu}(0,b]$, $K \in C^m(\Delta)$, with $\mu \in \mathbb{R}, \mu < 1$, then for all $N \geq N_0$ and $r \geq 1$ (given by (39.15)) the following error estimate holds:*

$$\|y - y_N\|_\infty \leq c \begin{cases} N^{-r(1-\nu)} & \text{for} \quad 1 \leq r < \frac{m}{1-\nu}, \\ N^{-m} & \text{for} \quad r \geq \frac{m}{1-\nu}. \end{cases}$$

*Here $c$ is a constant which does not depend on $N$, $\nu$ is given by formula (39.14) and*

$$\|v\|_\infty := \sup_{0<t<b} |v(t)|, \quad v \in L^\infty(0,b).$$

Note that on conditions of Theorem 2(ii) equation (39.20) has for sufficiently large $N$ a unique solution $z_N \in S_{m-1}^{(-1)}(\Pi_N)$ and

$$\|z - z_N\|_\infty \leq c \begin{cases} N^{-r(1-\nu)} & \text{for} \quad 1 \leq r < \frac{m}{1-\nu}, \\ N^{-m} & \text{for} \quad r \geq \frac{m}{1-\nu} \end{cases} \tag{39.24}$$

where $z = D_*^{\alpha_p} y$ and $c$ is a positive constant not depending on $N$.

It follows from Theorem 1 that in the case of sufficiently smooth $h_i$ $(i = 0, \ldots, p-1)$, $f$ and $K$, using sufficiently large values of the grid parameter $r$, for method (39.18),(39.20) by every choice of collocation parameters $0 \leq \eta_1 < \cdots < \eta_m \leq 1$ a convergence of order $O(N^{-m})$ can be expected. The following result shows that by a careful choice of parameters $\eta_1, \ldots, \eta_m$ it is possible to establish a faster convergence of this method.

**Theorem 3.** *Let the following conditions be fulfilled:*

(i) *$\mathscr{P}_N = \mathscr{P}_{N,m}$ $(N, m \in \mathbb{N})$ is defined by (39.21) where the interpolation nodes (39.16) with grid points (39.15) and parameters (39.17) are used;*

(ii) *the assumptions of Theorem 1 hold with $q := m + 1$;*

(iii) *the quadrature approximation*

$$\int_0^1 F(x)\,dx \approx \sum_{k=1}^{m} w_k F(\eta_k), \tag{39.25}$$

*with the knots* $\{\eta_k\}$ *satisfying* (39.17) *and appropriate weights* $\{w_k\}$ *is exact for all polynomials of degree* $m$.

*Then* (39.1)–(39.2) *has a unique solution* $y \in C[0,b]$ *such that* $D_*^{\alpha_p} y \in C^{q,\nu}(0,b]$. *There exists an integer* $N_0$ *such that, for* $N \geq N_0$, *equation* (39.20) *possesses a unique solution* $z_N \in S_{m-1}^{(-1)}(\Pi_N)$, *determining by* (39.18) *a unique approximation* $y_N$ *to* $y$, *the solution of* (39.1)-(39.2), *and the following error estimate holds:*

$$\|y - y_N\|_\infty \leq c \begin{cases} N^{-r(1+\alpha_p-\alpha_{p-1}-\nu)} & \text{for} \quad 1 \leq r < \frac{m+\alpha_p-\alpha_{p-1}}{1+\alpha_p-\alpha_{p-1}-\nu}, \\ N^{-m-(\alpha_p-\alpha_{p-1})} & \text{for} \quad r \geq \frac{m+\alpha_p-\alpha_{p-1}}{1+\alpha_p-\alpha_{p-1}-\nu}. \end{cases} \tag{39.26}$$

*Here* $r \in [1,\infty)$ *is the grading exponent of the grid* (*see* (39.15)), $\nu$ *is given by formula* (39.14) *and* $c$ *is a positive constant not depending on* $N$.

The proofs of Theorems 2 and 3 are based on Theorem 1 and are similar to the corresponding proofs of Theorems 4.1 and 4.2 in [PeTa12].

## 39.5 Numerical Illustration

We consider the following boundary value problem:

$$(D_*^{\frac{1}{2}} y)(t) + h_1(t)(D_*^{\frac{1}{4}} y)(t) + h_0(t)y(t) + \int_0^t K(t,s)y(s)ds = f(t), \quad 0 \leq t \leq 1, \tag{39.27}$$

with

$$y(0) + y(1) = 1,$$

where $K(t,s) := 1$ for $0 \leq s \leq t \leq 1$ and

$$h_0(t) := t^{\frac{1}{4}}, h_1(t) := 1, \quad 0 \leq t \leq 1,$$

$$f(t) := \frac{5\Gamma(\frac{7}{4})}{4\Gamma(\frac{9}{4})} t^{\frac{1}{4}} + \frac{3\Gamma(\frac{3}{4})}{2\sqrt{\pi}} t^{\frac{1}{2}} + t + \frac{4}{7} t^{\frac{7}{4}}, \quad 0 \leq t \leq 1.$$

This is a special problem of (39.1)–(39.2) with

$$p = 2, l = 1, \alpha_2 = \frac{1}{2}, \alpha_1 = \frac{1}{4}, b = 1, b_1 = 1, \gamma_0 = \gamma_1 = 1, \gamma = 1.$$

Clearly

$$h_0, h_1, f \in C^{q,\mu}(0,1],$$

with $\mu = \frac{3}{4}$ and arbitrary $q \in \mathbb{N}$.

To solve (39.27) by (39.18)–(39.20) we set

$$z := D_*^{\frac{1}{2}} y.$$

For $z$ we have equation (39.11) with $T$ and $g$ given by (39.12) and (39.13), respectively. Approximations $z_N \in S_{m-1}^{(-1)}(\Pi_N)$ for $m = 2$ and $N \in \mathbb{N}$ to the solution $z$ of equation (39.11) on the interval $[0, 1]$ are found by (39.19) using $m = 2$ and (39.16) with

$$\eta_1 = \frac{3 - \sqrt{3}}{6}, \quad \eta_2 = 1 - \eta_1,$$

the knots of the Gaussian quadrature formula (39.25). Actually,

$$z_N(t_{jk}) = c_{jk}, \quad k = 1, 2, \quad j = 1, \ldots, N,$$

and $z_N(t)$ for $t \in [0, 1]$ are determined by (39.23) and (39.22), respectively. After that the approximate solution $y_N$ for the boundary value problem (39.27) has been found by formula (39.18).

In the following tables (Tables 39.1 and 39.2), some results of numerical experiments for different values of the parameters $N$ and $r$ are presented. The errors $\varepsilon_N$ in the first table and the errors $\hat{\varepsilon}_N$ in the second table are calculated as follows:

$$\varepsilon_N := \max_{j=1,\ldots,N} \max_{k=0,\ldots,10} |y(\tau_{jk}) - y_N(\tau_{jk})|, \tag{39.28}$$

**Table 39.1** Numerical results for the problem (39.27).

| | $r = 1$ | | $r = 2$ | | $r = 5$ | | $r = 8$ | |
|---|---|---|---|---|---|---|---|---|
| $N$ | $\varepsilon_N$ | $\rho_N$ | $\varepsilon_N$ | $\rho_N$ | $\varepsilon_N$ | $\rho_N$ | $\varepsilon_N$ | $\rho_N$ |
| 16 | $4.88 \cdot 10^{-3}$ | 1.64 | $6.26 \cdot 10^{-4}$ | 3.68 | $2.72 \cdot 10^{-4}$ | 5.34 | $8.17 \cdot 10^{-4}$ | 4.76 |
| 32 | $2.96 \cdot 10^{-3}$ | 1.65 | $2.28 \cdot 10^{-5}$ | 2.74 | $4.86 \cdot 10^{-5}$ | 5.60 | $1.52 \cdot 10^{-4}$ | 5.36 |
| 64 | $1.78 \cdot 10^{-3}$ | 1.66 | $8.21 \cdot 10^{-5}$ | 2.79 | $8.54 \cdot 10^{-6}$ | 5.69 | $2.72 \cdot 10^{-5}$ | 5.59 |
| 128 | $1.07 \cdot 10^{-3}$ | 1.66 | $2.94 \cdot 10^{-5}$ | 2.80 | $1.49 \cdot 10^{-6}$ | 5.72 | $4.80 \cdot 10^{-6}$ | 5.67 |
| 256 | $6.44 \cdot 10^{-4}$ | 1.67 | $1.05 \cdot 10^{-5}$ | 2.81 | $2.61 \cdot 10^{-7}$ | 5.72 | $8.43 \cdot 10^{-7}$ | 5.69 |
| 512 | $3.86 \cdot 10^{-4}$ | 1.67 | $3.72 \cdot 10^{-6}$ | 2.82 | $4.58 \cdot 10^{-8}$ | 5.70 | $1.48 \cdot 10^{-7}$ | 5.69 |
| 1024 | $2.31 \cdot 10^{-4}$ | 1.67 | $1.32 \cdot 10^{-6}$ | 2.82 | $8.05 \cdot 10^{-9}$ | 5.69 | $2.60 \cdot 10^{-8}$ | 5.69 |
| | | 1.41 | | 2 | | 4.76 | | 4.76 |

**Table 39.2** Numerical results for the problem (39.27).

| | $r=1$ | | $r=2$ | | $r=5$ | | $r=8$ | |
|---|---|---|---|---|---|---|---|---|
| $N$ | $\hat{\varepsilon}_N$ | $\hat{\rho}_N$ | $\hat{\varepsilon}_N$ | $\hat{\rho}_N$ | $\hat{\varepsilon}_N$ | $\hat{\rho}_N$ | $\hat{\varepsilon}_N$ | $\hat{\rho}_N$ |
| 16 | $2.69\cdot 10^{-2}$ | 1.14 | $1.49\cdot 10^{-2}$ | 1.38 | $3.74\cdot 10^{-3}$ | 2.31 | $4.33\cdot 10^{-3}$ | 3.22 |
| 32 | $2.34\cdot 10^{-2}$ | 1.15 | $1.08\cdot 10^{-2}$ | 1.39 | $1.60\cdot 10^{-3}$ | 2.34 | $1.15\cdot 10^{-3}$ | 3.78 |
| 64 | $2.02\cdot 10^{-2}$ | 1.16 | $7.80\cdot 10^{-3}$ | 1.39 | $6.76\cdot 10^{-4}$ | 2.36 | $2.88\cdot 10^{-4}$ | 3.98 |
| 128 | $1.74\cdot 10^{-2}$ | 1.16 | $5.59\cdot 10^{-3}$ | 1.40 | $2.85\cdot 10^{-4}$ | 2.37 | $7.10\cdot 10^{-5}$ | 4.05 |
| 256 | $1.49\cdot 10^{-2}$ | 1.17 | $4.00\cdot 10^{-3}$ | 1.40 | $1.20\cdot 10^{-4}$ | 2.38 | $1.74\cdot 10^{-5}$ | 4.08 |
| 512 | $1.27\cdot 10^{-2}$ | 1.17 | $2.85\cdot 10^{-3}$ | 1.40 | $5.05\cdot 10^{-5}$ | 2.38 | $4.28\cdot 10^{-6}$ | 4.08 |
| 1024 | $1.08\cdot 10^{-2}$ | 1.17 | $2.02\cdot 10^{-3}$ | 1.41 | $2.12\cdot 10^{-5}$ | 2.38 | $1.05\cdot 10^{-6}$ | 4.07 |
| | | 1.19 | | 1.41 | | 2.38 | | 4 |

$$\hat{\varepsilon}_N := \max_{j=1,\ldots,N} \max_{k=0,\ldots,10} |z(\tau_{jk}) - z_N(\tau_{jk})|, \tag{39.29}$$

where

$$\tau_{jk} := t_{j-1} + k(t_j - t_{j-1})/10,\ k = 0,\ldots,10,\ j = 1,\ldots,N$$

with grid points $t_j$ given by (39.15). In (39.28) and (39.29) we have taken into account that the exact solution of (39.27) is

$$y(t) = t^{\frac{3}{4}}, \quad t \in [0,1],$$

and thus

$$z(t) = (D_*^{\frac{1}{2}} y)(t) = \frac{5\Gamma(\frac{7}{4})}{4\Gamma(\frac{9}{4})} t^{\frac{1}{4}}, \quad t \in [0,1].$$

The ratios

$$\rho_N := \frac{\varepsilon_{N/2}}{\varepsilon_N}, \quad \hat{\rho}_N := \frac{\hat{\varepsilon}_{N/2}}{\hat{\varepsilon}_N},$$

characterizing the observed convergence rate are also presented. Since

$$\alpha_2 = \frac{1}{2}, \alpha_1 = \frac{1}{4}, \mu = \frac{3}{4}, \nu = \max\{\mu, 1-\alpha_2+\alpha_1\} = \frac{3}{4},$$

we obtain from Theorem 3 (see (39.26)) that, for sufficiently large $N$,

$$\varepsilon_N \le c_0 \begin{cases} N^{-\frac{r}{2}} \text{ if } 1 \le r < \frac{9}{2}, \\ N^{-\frac{9}{4}} \text{ if } r \ge \frac{9}{2}, \end{cases} \tag{39.30}$$

and from (39.24) that

$$\hat{\varepsilon}_N \le c_1 \begin{cases} N^{-\frac{r}{4}} & \text{if } 1 \le r < 8, \\ N^{-2} & \text{if } r \ge 8. \end{cases} \tag{39.31}$$

Due to (39.30) the ratios $\rho_N$ for $r = 1$, $r = 2$, $r = 5$ and $r = 8$ ought to be approximatively $2^{\frac{1}{2}} \approx 1.41$, $2^1 = 2$, $2^{\frac{9}{4}} \approx 4.76$ and $2^{\frac{9}{4}} \approx 4.76$, respectively. These values are given in the last row of the first table. Due to (39.31) the ratios $\hat{\rho}_N$ for $r = 1$, $r = 2$, $r = 5$ and $r = 8$ ought to be approximatively $2^{\frac{1}{4}} \approx 1.19$, $2^{\frac{1}{2}} \approx 1.41$, $2^{\frac{5}{4}} \approx 2.38$ and $2^2 = 4$, respectively. These values are given in the last row of the second table. We can see from the tables that the actual rate of convergence of $z_N$ to $z$ is in good agreement with the estimate (39.24), but the convergence of $y_N$ to $y$ is faster than it is predicted by the theoretical estimate (39.26). This phenomenon is worth studying in a separate paper.

**Acknowledgements** This work was supported by Estonian Science Foundation Grant No. 9104 and by the institutional research funding IUT20-57 of the Estonian Ministry of Education and Research.

## References

[Di10] Diethelm, K.: The Analysis of Fractional Differential Equations. Lecture Notes in Mathematics, vol. **2004**, Springer, Berlin (2010)

[BrPeVa01] Brunner, H., Pedas, A., Vainikko, G.: Piecewise polynomial collocation methods for linear Volterra integro-differential equations with weakly singular kernels. SIAM J. Numer. Anal. **39**, 957–982 (2001)

[KiSrTr06] Kilbas, A. A., Srivastava, H. M., Trujillo, J. J.: Theory and Applications of Fractional Differential Equations. North-Holland Mathematics Studies, vol. **204**, Elsevier, Amsterdam (2006)

[Po99] Podlubny, I.: Fractional Differential Equations. Academic Press, San Diego (1999)

[AgBeHa10] Agarwal, R. P., Benchohra, M., Hamani, S.: A survey of existence results for boundary value problems of nonlinear fractional differential equations and inclusions. Acta. Appl. Math. **109**, 973–1033 (2010)

[PeTa11] Pedas, A., Tamme, E.: Spline collocation methods for linear multi-term fractional differential equations. J. Comput. Appl. Math. **236**, 167–176 (2011)

[FoMo11] Ford, N. J., Morgado, M. L.: Fractional boundary value problems: Analysis and numerical methods. Fract. Calc. Appl. Anal. **14**, 554–567 (2011)

[FoMo14] Ford, N. J., Morgado, M. L., Rebelo, M.: High order numerical methods for fractional terminal value problems. Comput. Methods Appl. Math. **14**, 55–70 (2014)

[DoBhEz11] Doha, E. H., Bhrawy, A. H., Ezz-Eldien, S. S.: A Chebyshev spectral method based on operational matrix for initial and boundary value problems of fractional order. Comput. Math. Appl. **62**, 2364–2373 (2011)

[PeTa12] Pedas, A., Tamme, E.: Piecewise polynomial collocation for linear boundary value problems of fractional differential equations. J. Comput. Appl. Math. **236**, 3349–3359 (2012)

[PeTa14a] Pedas, A., Tamme, E.: Numerical solution of nonlinear fractional differential equations by spline collocation methods. J. Comput. Appl. Math. **255**, 216–230 (2014)

[PeTa14b] Pedas, A., Tamme, E.: Spline collocation for nonlinear fractional boundary value problems. Appl. Math. Comput. **244**, 502-Ű513 (2014)

[MaHu14] Ma, X., Huang, C.: Spectral collocation method for linear fractional integro-differential equations. Appl. Math. Modelling **38**, 1434–1448 (2014)

[YaPaFo14] Yan, Y., Pal, K., Ford, N.: Higher order numerical methods for solving fractional differential equations. Bit Numer. Math. **54**, 555–584 (2014)

[Va93] Vainikko, G.: Multidimensional Weakly Singular Integral Equations. Lecture Notes in Mathematics, vol. **1549**, Springer, Berlin (1993)

# Chapter 40
# A Note on Transforming a Plane Strain First-Kind Fredholm Integral Equation into an Equivalent Second-Kind Equation

**S. Pomeranz**

## 40.1 Introduction

Methods to convert Fredholm integral equations of the first kind into equivalent Fredholm integral equations of the second kind are used to study issues of existence and uniqueness of solutions. For some examples applied to plane strain problems, see [Co95] and [Co00, Sec. 2.12]. In this paper, another technique to convert the Fredholm integral equation of the first kind that arises in a direct boundary integral formulation for the plane strain Dirichlet problem into an equivalent Fredholm integral equation of the second kind is developed. The technique presented in this paper generalizes work of Y. Yan and I.H. Sloan that was done for the scalar Laplace equation [YaSl88] to the plane strain system of displacement equations.

In Section 40.2, the plane strain boundary integral equations are developed. In Section 40.3, the Somigliana equations are expressed in a convenient form, and a Fredholm integral equation of the first kind is obtained. Adaptation of the work from [YaSl88] is implemented in Section 40.4 in formulating an equivalent Fredholm integral equation of the second kind. Results are summarized in Section 40.5.

## 40.2 Boundary Integral Equations for Plane Strain

The notation and presentation from [PaBrWr92, Sec. 6.2.1] is followed and is summarized here for completeness. Einstein summation notation is used, so that a repeated letter subscript in a term implies summation, and differentiation is denoted by commas within expressions.

S. Pomeranz (✉)
The University of Tulsa, 800 S. Tucker Drive, Tulsa, OK 74104, USA
e-mail: pomeranz@utulsa.edu



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_40

The system of plane strain equations for a linear, homogeneous, isotropic material, expressed in terms of the displacement components (Navier equations), is

$$G u_{j,kk} + \frac{G}{1-2\nu} u_{k,kj} + b_j = 0, \tag{40.1}$$

and the surface traction is given by

$$p_i = \frac{2G\nu}{1-2\nu} u_{k,k} n_i + G(u_{i,j} + u_{j,i}) n_j,$$

where $i,j,k = 1,2$; $\nu$ is Poisson's ratio; $G$ is the shear modulus; $u_i$ are the displacement components; $p_i$ are the surface traction components; and $b_i$ are the load components.

The boundary integral equations are obtained using a weighted residual method in which the weighting function is chosen to be $u^*$, the Kelvin fundamental solution for (40.1). Quantities associated with the fundamental solution are indicated with an asterisk superscript. The fundamental solution satisfies

$$G u^*_{lj,kk}(r) + \frac{G}{1-2\nu} u^*_{lk,kj}(r) = -\delta_{lj}\,\delta(r), \tag{40.2}$$

where $j,k,l = 1,2$; $\delta_{lj}$ is the Kronecker delta; $r$ is the distance between the load and field points; and $\delta(r)$ is the Dirac delta generalized function expressed as a function of $r$. The fundamental solution and the fundamental surface traction have components, respectively,

$$u^*_k = u^*_{1k} + u^*_{2k}, \tag{40.3}$$

$$p^*_k = p^*_{1k} + p^*_{2k}, \tag{40.4}$$

for $k = 1,2$. The double subscripts in (40.3) and (40.4) use the first subscript for the direction of the Dirac delta generalized function load and the second subscript for the direction of the resulting displacement or surface traction at the field point.

The plane strain fundamental solution and fundamental surface traction components are, respectively,

$$u^*_{ij} = \frac{1}{8\pi(1-\nu)G}\left((3-4\nu)\delta_{ij}\ln\left(\frac{1}{r}\right) + r_{,i}r_{,j}\right), \tag{40.5}$$

$$p^*_{ij} = \frac{-1}{4\pi(1-\nu)r}\left([(1-2\nu)\delta_{ij} + 2r_{,i}r_{,j}]\frac{\partial r}{\partial n} - (1-2\nu)(r_{,i}n_j - r_{,j}n_i)\right), \tag{40.6}$$

where $i,j = 1,2$, and $\mathbf{n} = (n_1, n_2)$ is the unit outward normal vector on the boundary [PaBrWr92, p. 227, Eq. 6.23].

For this Dirichlet problem there are two basic unknown quantities, the surface traction and the interior displacement. The surface traction can be determined first.

The surface traction is then used to obtain the interior displacement. The process on which we focus in this paper is that of determining the surface traction. To determine the surface traction, equation (40.1) is multiplied by the fundamental solution, $u^*$, and integrated over the two-dimensional domain. Green's second identity is used, and the boundary integral involving the Dirac delta generalized function from (40.2) is simplified. The two displacement components evaluated at a load point, $\mathbf{x}^i = (x_1^i, x_2^i)$, are described by the resulting Somigliana equations,

$$c_{lk}^i u_k^i = \int_\Gamma u_{lk}^{*i} p_k \, d\Gamma - \int_\Gamma p_{lk}^{*i} u_k \, d\Gamma + \int_\Omega u_{lk}^{*i} b_k \, da, \tag{40.7}$$

where $l,k = 1,2$; $d\Gamma$ is the differential element of arc length; and $da$ is the differential element of area. The domain is $\Omega$ and its boundary is $\Gamma$. If the evaluation (load) point $\mathbf{x}^i$ is a smooth boundary point, then $c_{lk}^i = \frac{1}{2}\delta_{lk}$. If $\mathbf{x}^i$ is an interior boundary point, then $c_{lk}^i = \delta_{lk}$.

In matrix-vector form, (40.7) becomes

$$c^i \mathbf{u}^i = \int_\Gamma u^{*i} \mathbf{p} \, d\Gamma - \int_\Gamma p^{*i} \mathbf{u} \, d\Gamma + \int_\Omega u^{*i} \mathbf{b} \, da,$$

with

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \qquad \mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix},$$

and

$$u^{*i} = \begin{pmatrix} u_{11}^{*i} & u_{12}^{*i} \\ u_{21}^{*i} & u_{22}^{*i} \end{pmatrix}, \qquad p^{*i} = \begin{pmatrix} p_{11}^{*i} & p_{12}^{*i} \\ p_{21}^{*i} & p_{22}^{*i} \end{pmatrix}.$$

## 40.3 Fredholm Integral Equation of the First Kind

The following material describes the technique to convert the Fredholm integral equation of the first kind, arising in the direct boundary integral formulation for the plane strain Dirichlet problem, into an equivalent Fredholm integral equation of the second kind. A Fredholm integral equation of the second kind is desired, for example, so that the Fredholm alternative can be applied to investigate existence and uniqueness of the solution for the surface traction. We adapt material from [YaSl88] for a scalar Dirichlet problem for Laplace's equation to apply to the vector-valued Dirichlet problem for plane strain, a system of equations. The same ideas can be applied to plane stress problems. The load term here should not depend on the unknown.

The Somigliana equations, expressed in notation from [PaBrWr92, p. 226, Eq. 6.25], are as given in (40.7). In matrix form, the fundamental displacement solution for the plane strain problem is

$$u^* = \frac{1}{8\pi(1-\nu)G}\begin{pmatrix} (3-4\nu)\ln\frac{1}{r} + (r_{,1})^2 & r_{,1}\,r_{,2} \\ r_{,2}\,r_{,1} & (3-4\nu)\ln\frac{1}{r} + (r_{,2})^2 \end{pmatrix},$$

which is equivalent to (40.5). The fundamental surface traction is (40.6). The boundary, $\Gamma$, must be a smooth, simple, closed curve [YaSl88, p. 562] and [Co00, p. 1 and p. 4]. We rearrange (40.7) in order to isolate the unknown surface traction components, $p_k$, $k = 1, 2$, and to more clearly observe that this is a Fredholm integral equation of the first kind,

$$\int_\Gamma u^{*i}_{lk} p_k d\Gamma = c^i_{lk} u^i_k + \int_\Gamma p^{*i}_{lk} u_k d\Gamma - \int_\Omega u^{*i}_{lk}\, b_k\, da. \tag{40.8}$$

Since the right-hand side of (40.8) is known, we can relabel the entire right-hand side as $f(x^i) = f(x) = (f_1(x), f_2(x))^T$, and (40.8) can be expressed as the system

$$\begin{aligned} &\frac{1}{8\pi(1-\nu)G}\int_\Gamma \left( \left[ (3-4\nu)\ln\left(\frac{1}{|x-y|}\right) + \frac{(x_1-y_1)^2}{|x-y|^2} \right] p_1(y) \right. \\ &\qquad \left. + \frac{(x_1-y_1)(x_2-y_2)}{|x-y|^2} p_2(y) \right) d\Gamma(y) = f_1(x), \\ &\frac{1}{8\pi(1-\nu)G}\int_\Gamma \left( \frac{(x_1-y_1)(x_2-y_2)}{|x-y|^2} p_1(y) \right. \\ &\qquad \left. + \left[ (3-4\nu)\ln\left(\frac{1}{|x-y|}\right) + \frac{(x_2-y_2)^2}{|x-y|^2} \right] p_2(y) \right) d\Gamma(y) = f_2(x) \end{aligned} \tag{40.9}$$

and, equivalently, in matrix-vector form, as

$$\int_\Gamma \begin{pmatrix} (-3+4\nu)\ln|x-y| + \frac{(x_1-y_1)^2}{|x-y|^2} & \frac{(x_1-y_1)(x_2-y_2)}{|x-y|^2} \\ \frac{(x_1-y_1)(x_2-y_2)}{|x-y|^2} & (-3+4\nu)\ln|x-y| + \frac{(x_2-y_2)^2}{|x-y|^2} \end{pmatrix} \begin{pmatrix} p_1(y) \\ p_2(y) \end{pmatrix} d\Gamma(y) = \begin{pmatrix} \hat{f}_1(x) \\ \hat{f}_2(x) \end{pmatrix},$$

where $x \in \Gamma$ and $\hat{f}(x) \equiv 8\pi(1-\nu)Gf(x)$.

We want to transform (40.9), a Fredholm integral equation of the first kind, into an equivalent Fredholm integral equation of the second kind. This will be done now by adapting a technique used for Laplace's equation in [YaSl88] and [AtHa01, p. 431, Sec. 12.3]. Also see [At97, Chpt. 7], [AtSl91, Secs. 1 and 2], [Ch83, Sec. 2], and [Sl91, Sec. 3].

## 40.4 Fredholm Integral Equation of the Second Kind

Let the function $v = (v_1, v_2)$ be a smooth, invertible, $2\pi$-periodic function that parameterizes the smooth boundary curve $\Gamma$. Specifically, $v : \mathbb{R}/2\pi\mathbb{Z} \to \Gamma \subset \mathbb{R}^2$ with $|v'(s)| = \sqrt{v_1'(s)^2 + v_2'(s)^2} \geq \rho > 0$, where the scalar parameter $s \in [-\pi, \pi]$, and $\rho$ is an arbitrary positive constant [YaSl88, p. 562, Sec. 4]. $\mathbb{Z}$ and $\mathbb{R}$ denote the integers and real numbers, respectively. The default amount of smoothness required of $v$ and other quantities stipulated as smooth is $C^\infty$, unless specified otherwise. Rewrite (40.9) in terms of this parameterization with $x = v(s)$, $y = v(\sigma)$, and $\sigma, s \in [-\pi, \pi]$. For convenience, define $w_i(\sigma) = |v'(\sigma)| p_i(v(\sigma))$ and $\bar{f}_i(s) = \hat{f}_i(v(s))$, for $i = 1, 2$, with $w = (w_1, w_2)^T$ and $\bar{f} = (\bar{f}_1, \bar{f}_2)^T$. We obtain

$$\begin{aligned}
&\int_{-\pi}^{\pi} \left( \left[ (-3 + 4\nu) \ln |v(s) - v(\sigma)| + \frac{(v_1(s) - v_1(\sigma))^2}{|v(s) - v(\sigma)|^2} \right] w_1(\sigma) \right. \\
&\qquad \left. + \frac{(v_1(s) - v_1(\sigma))(v_2(s) - v_2(\sigma))}{|v(s) - v(\sigma)|^2} w_2(\sigma) \right) d\sigma \\
&= \bar{f}_1(s), \\
&\int_{-\pi}^{\pi} \left( \frac{(v_1(s) - v_1(\sigma))(v_2(s) - v_2(\sigma))}{|v(s) - v(\sigma)|^2} w_1(\sigma) \right. \\
&\left. + \left[ (-3 + 4\nu) \ln |v(s) - v(\sigma)| + \frac{(v_2(s) - v_2(\sigma))^2}{|v(s) - v(\sigma)|^2} \right] w_2(\sigma) \right) d\sigma \\
&= \bar{f}_2(s).
\end{aligned} \tag{40.10}$$

The form of the system of equations (40.10) can be compared with that of the analogous scalar equation [YaSl88, p. 562, Eq. 12*]. From (40.10), define the symmetric integral matrix operator $K$,

$$\begin{aligned}
(Kw)(s) &= \int_{-\pi}^{\pi} k(s, \sigma) w(\sigma) \, d\sigma \\
&= \int_{-\pi}^{\pi} \begin{pmatrix} k_{11}(s, \sigma) & k_{12}(s, \sigma) \\ k_{21}(s, \sigma) & k_{22}(s, \sigma) \end{pmatrix} \begin{pmatrix} w_1(\sigma) \\ w_2(\sigma) \end{pmatrix} d\sigma \\
&= \bar{f}(s),
\end{aligned} \tag{40.11}$$

where $s \in [-\pi, \pi]$.

We introduce an invertible operator adapted from the corresponding operator for Laplace's equation in [YaSl88, pp. 560–561] and [AtHa01, Chap. 12], $A$, modified here for use with the plane strain equations. Our operator is represented using a diagonal $2 \times 2$ matrix, also denoted by $A$,

$$\begin{aligned} Kw(s) &= Aw(s) + (K-A)w(s) \\ &= Aw(s) + Bw(s) + Cw(s) \\ &= \bar{f}(s), \end{aligned} \tag{40.12}$$

where $K = A + B + C$.

Let

$$\begin{aligned} (Aw)(s) &= \int_{-\pi}^{\pi} a(s,\sigma)\, w(\sigma)\, d\sigma \\ &= \int_{-\pi}^{\pi} \begin{pmatrix} a_{11}(s,\sigma) & a_{12}(s,\sigma) \\ a_{21}(s,\sigma) & a_{22}(s,\sigma) \end{pmatrix} \begin{pmatrix} w_1(\sigma) \\ w_2(\sigma) \end{pmatrix} d\sigma, \end{aligned}$$

$$\begin{aligned} (Bw)(s) &= \int_{-\pi}^{\pi} b(s,\sigma)\, w(\sigma)\, d\sigma \\ &= \int_{-\pi}^{\pi} \begin{pmatrix} b_{11}(s,\sigma) & b_{12}(s,\sigma) \\ b_{21}(s,\sigma) & b_{22}(s,\sigma) \end{pmatrix} \begin{pmatrix} w_1(\sigma) \\ w_2(\sigma) \end{pmatrix} d\sigma, \end{aligned}$$

and

$$\begin{aligned} (Cw)(s) &= \int_{-\pi}^{\pi} c(s,\sigma)\, w(\sigma)\, d\sigma \\ &= \int_{-\pi}^{\pi} \begin{pmatrix} c_{11}(s,\sigma) & c_{12}(s,\sigma) \\ c_{21}(s,\sigma) & c_{22}(s,\sigma) \end{pmatrix} \begin{pmatrix} w_1(\sigma) \\ w_2(\sigma) \end{pmatrix} d\sigma, \end{aligned}$$

for $s \in [-\pi, \pi]$. As will be discussed in the following material, these integral operators have kernel components, respectively,

$$\begin{aligned} a_{11}(s,\sigma) &= a_{22}(s,\sigma) = (-3+4\nu)\ln|2e^{-1/2}\sin(\frac{s-\sigma}{2})|, \\ a_{12}(s,\sigma) &= a_{21}(s,\sigma) = 0, \end{aligned}$$

$$\begin{aligned} b_{11}(s,\sigma) &= b_{22}(s,\sigma) \\ &= \begin{cases} (-3+4\nu)\ln|\frac{\nu(s)-\nu(\sigma)}{2e^{-1/2}\sin(\frac{s-\sigma}{2})}|, & \text{if } s-\sigma \notin 2\pi\mathbb{Z}, \\ (-3+4\nu)\ln|e^{1/2}\nu'(s)|, & \text{if } s-\sigma \in 2\pi\mathbb{Z}, \end{cases} \\ b_{12}(s,\sigma) &= b_{21}(s,\sigma) = 0, \end{aligned}$$

and

$$c_{11}(s,\sigma) = \begin{cases} \frac{(v_1(s)-v_1(\sigma))^2}{|v(s)-v(\sigma)|^2}, & \text{if } s-\sigma \notin 2\pi\mathbb{Z}, \\ \frac{v_1'(s)^2}{v_1'(s)^2+v_2'(s)^2}, & \text{if } s-\sigma \in 2\pi\mathbb{Z}, \end{cases}$$

$$c_{12}(s,\sigma) = c_{21}(s,\sigma) = \begin{cases} \frac{(v_1(s)-v_1(\sigma))\,(v_2(s)-v_2(\sigma))}{|v(s)-v(\sigma)|^2}, & \text{if } s-\sigma \notin 2\pi\mathbb{Z}, \\ \frac{v_1'(s)\,v_2'(s)}{v_1'(s)^2+v_2'(s)^2}, & \text{if } s-\sigma \in 2\pi\mathbb{Z}, \end{cases} \tag{40.13}$$

$$c_{22}(s,\sigma) = \begin{cases} \frac{(v_2(s)-v_2(\sigma))^2}{|v(s)-v(\sigma)|^2}, & \text{if } s-\sigma \notin 2\pi\mathbb{Z}, \\ \frac{v_2'(s)^2}{v_1'(s)^2+v_2'(s)^2}, & \text{if } s-\sigma \in 2\pi\mathbb{Z}. \end{cases}$$

Therefore, we have for the kernel in (40.11),

$$\begin{pmatrix} k_{11}(s,\sigma)\; k_{12}(s,\sigma) \\ k_{21}(s,\sigma)\; k_{22}(s,\sigma) \end{pmatrix} =$$

$$\begin{pmatrix} (-3+4\nu)\ln|2e^{-1/2}\sin(\frac{s-\sigma}{2})| & \\ +(-3+4\nu)\ln|\frac{\nu(s)-\nu(\sigma)}{2e^{-1/2}\sin(\frac{s-\sigma}{2})}| & \frac{v_1(s)-v_1(\sigma))\,(v_2(s)-v_2(\sigma))}{|v(s)-v(\sigma)|^2} \\ +\frac{(v_1(s)-v_1(\sigma))^2}{|v(s)-v(\sigma)|^2} & \\ & \\ \frac{v_1(s)-v_1(\sigma))\,(v_2(s)-v_2(\sigma))}{|v(s)-v(\sigma)|^2} & (-3+4\nu)\ln|2e^{-1/2}\sin(\frac{s-\sigma}{2})| \\ & +(-3+4\nu)\ln|\frac{\nu(s)-\nu(\sigma)}{2e^{-1/2}\sin(\frac{s-\sigma}{2})}| \\ & +\frac{(v_2(s)-v_2(\sigma))^2}{|v(s)-v(\sigma)|^2} \end{pmatrix}$$

if $s-\sigma \notin 2\pi\mathbb{Z}$.

The following kernel matrix is included here only to emphasize the fact that functions $b_{ij}$ and $c_{ij}$, $i,j=1,2$, and all their derivatives have only isolated removable discontinuities that arise if $s-\sigma$ is an integer multiple of $2\pi$, and this is not problematic. The derivation of the new terms in the following matrix is given in the remaining material in this section:

$$\begin{pmatrix} k_{11}(s,\sigma)\; k_{12}(s,\sigma) \\ k_{21}(s,\sigma)\; k_{22}(s,\sigma) \end{pmatrix} =$$

$$\begin{pmatrix} (-3+4\nu)\ln|2e^{-1/2}\sin(\frac{s-\sigma}{2})| & \frac{\nu_1'(s)\nu_2'(s)}{\nu_1'(s)^2+\nu_2'(s)^2} \\ +(-3+4\nu)\ln|e^{1/2}\nu'(s)| & \\ +\frac{\nu_1'(s)^2}{\nu_1'(s)^2+\nu_2'(s)^2} & \\ \frac{\nu_1'(s)\nu_2'(s)}{\nu_1'(s)^2+\nu_2'(s)^2} & (-3+4\nu)\ln|2e^{-1/2}\sin(\frac{s-\sigma}{2})| \\ & +(-3+4\nu)\ln|e^{1/2}\nu'(s)| \\ & +\frac{\nu_2'(s)^2}{\nu_1'(s)^2+\nu_2'(s)^2} \end{pmatrix}$$

if $s-\sigma \in 2\pi\mathbb{Z}$.

Our operators $A$, $B$, and $C$ are adapted from the corresponding operators used in the scalar Laplace's equation problem in [YaSl88, p. 560, p. 563, and p. 567], $A$, $B$, and $F$. Integral operator $A$ has a weak singularity for $s-\sigma \in 2\pi\mathbb{Z}$. The associated improper integral exists. Integral operator kernels $b$ and $c$ can be defined at their singular points so as to be arbitrarily smooth, as will now be discussed.

Except for a multiplicative constant, our integral operator $A$ has nonzero kernel components that are identical to that of scalar operator $A$ in [YaSl88, p. 560]. It is proved there, using an equivalent Fourier series expansion, that $A$ is an invertible bounded linear operator. Therefore, our matrix operator $A$ inherits this property for its nonzero components, and we have an invertible bounded linear operator,

$$A : H^t[-\pi,\pi] \to H^{t+1}[-\pi,\pi],$$

where $t \in \mathbb{R}$ and $H^t[-\pi,\pi]$ is the standard Sobolev space that is defined as the completion of $C^\infty[-\pi,\pi]$ with respect to the Sobolev norm $||\cdot||_t$.

Operating with $A^{-1}$ on $K$ in (40.12) yields

$$\begin{aligned} (A^{-1}Kw)(s) &= (A^{-1}(A+B+C)w)(s) \\ &= \left((I+A^{-1}B+A^{-1}C)w\right)(s) \\ &= A^{-1}\bar{f}(s). \end{aligned}$$

Let $M = A^{-1}B + A^{-1}C$ and $\bar{\bar{f}} = A^{-1}\bar{f}$. We have a Fredholm integral equation of the second kind,

$$\left((I+M)w\right)(s) = \bar{\bar{f}}(s). \tag{40.14}$$

The following analysis of operators $B$ and $C$ is of a numerical nature and is a new/different approach to material that has been developed in a more analytic manner by other researchers. Except for a multiplicative constant, our integral operator $B$ has nonzero kernel components that are identical to that of the scalar operator $B$ in [YaSl88, p. 563]. We have, as in [YaSl88], that the kernel function $b_{11}(s,\sigma) = b_{22}(s,\sigma)$ is better behaved than the respective kernel function $k_{11}(s,\sigma) = k_{22}(s,\sigma)$. In order to more clearly observe that $b_{11}$ has a finite limit at

$s-\sigma \in 2\pi\mathbb{Z}$, Taylor-expand the logarithm factor for $\sigma$ close to $s$ (and more generally for $s-\sigma \in 2\pi\mathbb{Z}$ ), and take the limit (or use L'Hospital's rule):

$$\begin{aligned}\lim_{\sigma\to s} \ln\left|e^{1/2}\frac{v(s)-v(\sigma)}{2\sin\frac{s-\sigma}{2}}\right| &= \lim_{\sigma\to s} \ln\left|e^{1/2}\frac{v(s)-v(\sigma)}{s-\sigma}\right| \\ &= \ln|e^{1/2}v'(s)| \\ &= \ln\left(e^{1/2}\sqrt{v_1'(s)^2+v_2'(s)^2}\right)\end{aligned}$$

where $\sqrt{v_1'(s)^2+v_2'(s)^2} \geq \rho > 0$.

Investigating numerically using the computer algebra system (CAS) *Mathematica* (and mathematical induction), successive derivatives of $b(s,\sigma)$ with respect to $\sigma$ in the limit as $\sigma \to s$ are shown to be finite and such that the denominator of the simplified $n$th derivative is $(v_1'(s)^2+v_2'(s)^2)^n$, $n = 0,1,\ldots$. This demonstrates that $b_{11}(s,\sigma) \equiv b_{22}(s,\sigma)$ and their derivatives can be defined to be their respective limiting values at what originally were singular points, and $b_{11}$ and $b_{22}$ are then smooth for all $s,\sigma \in [-\pi,\pi]$, provided the parameterization $v$ is smooth and that $|v'(s)| = \sqrt{v_1'(s)^2+v_2'(s)^2} \geq \rho > 0$. Recall that these two conditions have been assumed.

The mathematical arguments from [YaSl88, p. 563, Props. 4.1 and 4.2] for the operator $B$ in that paper apply to the nonzero components of our operator $B$. Our operator kernel $b(s,\sigma) \in C^\infty([-\pi,\pi]\times[-\pi,\pi])$ and is $2\pi$-periodic in each variable. For any $w \in H^t[-\pi,\pi]$ and any $t \in \mathbb{R}$, the function $Bw(s)$ has derivatives of all orders, and $B : H^t[-\pi,\pi] \to H^{t+r}[-\pi,\pi]$ is a bounded linear operator for any $t \in \mathbb{R}$ and $r \in \mathbb{Z}$. Applying a Sobolev embedding theorem, we have that $H^{t+2}[-\pi,\pi]$ is compactly embedded in $H^{t+1}[-\pi,\pi]$, and, consequently, $B : H^t[-\pi,\pi] \to H^{t+1}[-\pi,\pi]$ is a compact operator.

It remains to consider the integral operator $C$, defined in (40.13). It turns out that the components comprising kernel $c$ are also well behaved in the limit as $\sigma - s \to 2\pi n$, $n \in \mathbb{Z}$ (and they are well behaved elsewhere). Therefore, the kernel $c$ can be treated similarly to the treatment of kernel $b$.

Using a CAS (e.g., *Mathematica*) to repeatedly apply L'Hospital's rule to a $c_{ij}$ term, $i,j = 1,2$, and its successively higher order derivatives, it is observed that each $c_{ij}$ and its derivatives have finite limits as $\sigma - s \to 2\pi n$, $n \in \mathbb{Z}$. A pattern is observed in which $2(n+1)$ applications of L'Hospital's rule show that $\partial^n c_{ij}/\partial\sigma^n(s,\sigma)$, $n = 0,1,\ldots$, has a finite limit at singular points. The denominator of the finite limiting expression for each $n$th partial derivative has the simplified form (up to a multiplicative constant) of $(v_1'(s)^2+v_2'(s)^2)^{n+1}$, which is nonzero by the previous assumption that $|v'(s)| = \sqrt{v_1'(s)^2+v_2'(s)^2} \geq \rho > 0$. Define each $\partial^n c_{ij}/\partial\sigma^n(s,s)$ to be the corresponding limit. A proof by induction concludes the argument that

$$c_{ij}(s,\sigma) \in C^\infty([-\pi,\pi]\times[-\pi,\pi]),$$

for each term $i,j = 1,2$. Therefore, we have the desired result that the behavior of the $c_{ij}$ terms is similar to that of the $b_{ij}$ terms in the sense that the operator $C : H^t[-\pi,\pi] \to H^{t+1}[-\pi,\pi]$ is a compact operator.

The integral operators $B$ and $C$ are each compact operators [YaSl88, p. 563], [Co00, pp. 3–4, Thm. 1.7]. The composition of bounded linear operator $A^{-1}$ with compact integral operators $B$ and $C$ preserves the compactness, so that $M = A^{-1}(B+C)$ in (40.14) is a compact operator on $H^t[-\pi,\pi] \to H^t[-\pi,\pi]$, $t \in \mathbb{R}$.

One motivation for transforming the original equation (40.9), a Fredholm integral equation of the first kind, into an equivalent Fredholm integral equation of the second kind, is to be able to apply the Fredholm alternative. This theorem can now be applied. Use the invertibility of the coordinate transformation to conclude that, since existence and uniqueness of the surface traction solution in the original coordinate system does not depend upon the specific parameterization, the result holds in the original Cartesian coordinates for $p(x)$.

### 40.4.1 Example of a Suitable Parameterization

For the special case in which the boundary curve, $\Gamma$, is a circle parameterized by $(x,y) = (v_1(t), v_2(t)) = (\cos t, \sin t)$, $t \in [-\pi,\pi]$, a direct computation (e.g., with *Mathematica*) shows that the $c_{ij}$ terms, $i,j = 1,2$, have the desired form. For example, for the $c_{11}(s,\sigma)$ term,

$$c_{11}(s,\sigma) = \frac{(v_1(s)-v_1(\sigma))^2}{|v(s)-v(\sigma)|^2}$$
$$= \frac{(\cos s - \cos\sigma)^2}{(\cos s-\cos\sigma)^2+(\sin s-\sin\sigma)^2} = \sin^2\left(\frac{s+\sigma}{2}\right),$$

if $s-\sigma \notin 2\pi\mathbb{Z}$. Therefore,

$$c_{11}(s,\sigma) \to \sin^2 s, \quad \text{as } \sigma \to s+2\pi n, \quad n \in \mathbb{Z}.$$

This function $c_{11}$ has a removable discontinuity when $s-\sigma$ is an integer multiple of $2\pi$, and $c_{11}$ can be defined at such points to be its limiting value to create a continuous function. Similarly, since all derivatives are constant multiples of $\cos 2s$ or $\sin 2s$ at such points, this implies that, with this definition, $\partial^n c_{11}/\partial s^n(s,\sigma)$ is arbitrarily smooth for all $n = 0,1,2,\ldots$.

## 40.5 Summary

By adapting the development for Laplace's equation presented in [YaSl88], we have demonstrated how to convert the Fredholm integral equation of the first kind arising in a direct boundary integral formulation for the plane strain (stress) Dirichlet problem into an equivalent Fredholm integral equation of the second kind.

**Acknowledgements** The author thanks Dr. Christian Constanda, the Charles W. Oliphant Endowed Chair in Mathematical Sciences at The University of Tulsa, for his assistance.

## References

[Co95] Constanda, C.: The Boundary Integral Equation Method in Plane Elasticity. Proc. Amer. Math. Soc. **123**, 3385–3396 (Nov. 1995)

[Co00] Constanda, C.: Direct and Indirect Boundary Integral Equations. Chapman & Hall/CRC, Boca Raton-London-New York-Washington, DC (2000)

[YaSl88] Yan, Y., Sloan, I.: On Integral Equations of the First Kind with Logarithmic Kernels. J. Integral Equations Appl. **1**, 549–579 (1988)

[PaBrWr92] Partridge, P., Brebbia, C., Wrobel, L.: The Dual Reciprocity Boundary Element Method. Computational Mechanics Publications and Elsevier Science Publishers Ltd., Southampton and Essex, UK (1992)

[AtHa01] Atkinson, K., Han, W.: Theoretical Numerical Analysis. Springer, New York (2001)

[At97] Atkinson, K.: The Numerical Solution of Integral Equations of the Second Kind. Cambridge University Press, Cambridge, UK (1997)

[AtSl91] Atkinson, K., Sloan, I.: The Numerical Solution of First Kind Logarithmic Kernel Integral Equations on Smooth Open Arcs. Math. Comp. **56**(193), 119–139 (1991)

[Ch83] Chandler, G.A.: Numerical Analysis of the Boundary Element Method. In: Gustafson, S.A., Womersley, R.S. (eds.) Mathematical Programming and Numerical Analysis Workshop: Proc. Centre Math. Analysis. **6**, pp. 211–230. Austral. Nat. Univ. (1983)

[Sl91] Sloan, I. H.: Error Analysis of Boundary Integral Methods. Acta Numerica **1**, 287–339 (1992)

# Chapter 41
# Asymptotic Analysis of the Steklov Spectral Problem in Thin Perforated Domains with Rapidly Varying Thickness and Different Limit Dimensions

**A. Popov**

## 41.1 Introduction

A rigorous method for constructing asymptotic approximations in thin domains was first proposed by Gol'denveizer [Go76, Go62]; it was further developed for thin domains of cylindrical type in [Dz72, Ca84, VaBu90, Na82]. These authors considered thin domains of cylindrical type and the main approach to asymptotic analysis was to make a special change of coordinates after which the scaled domain was independent of the small parameter. Then a small parameter appeared in the higher derivatives of the differential equations and the Lyusternik–Vishik method [ViLy67] was used.

These methods do not work for boundary value problems in thin perforated domains with rapidly changing thickness. Methods of homogenization theory were first used for thin domains by Panasenko and Reztsov [PaRe87] to investigate the three-dimensional elasticity problem in a thin inhomogeneous cylindrical plate. Mel'nik in [Me91] investigated elliptic and spectral problems with rapidly oscillating coefficients in thin perforated domains with rapidly changing thickness. The analysis of the asymptotic behavior of solutions to various boundary value problems in thin domains with rapidly changing thickness was also the subject of [AkNa04, CiCh02, Ko99, Ko85].

The monograph [Na02] contains a detailed presentation of the asymptotic theory of thin elastic plates and rods. The leading terms of the asymptotic solutions were considered, new methods for investigating boundary value and spectral problems were described.

A. Popov (✉)
Taras Shevchenko National University of Kyiv, 64/13 Volodymyrska Street, Kyiv 01601, Ukraine
e-mail: popov256@gmail.com



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_41

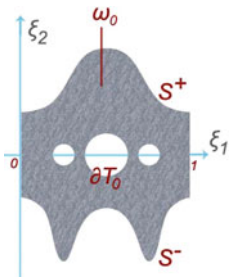


**Fig. 41.1** Periodicity cell $\omega_0$.

Steklov spectral problem in thin domain with non-smooth boundary was considered by Isakov in [Is88], where the leading terms of asymptotic expansion for eigenvalues were constructed.

In the papers mentioned above, asymptotic methods for boundary value problems in thin domains were developed separately depending on their limiting dimension (a thin plate or a thin bar). In this chapter, independently of the limit dimension of the thin perforated domain, we study the asymptotic behavior of eigenvalues and eigenfunctions of the Steklov spectral problem in such domains.

## 41.2 Description of a Thin Perforated Domain with Rapidly Oscillating Thickness and Statement of the Problem

Let $h_{\pm}^{(1)}(\xi'), h_{\pm}^{(2)}(\xi'), \dots, h_{\pm}^{(d)}(\xi')$ be smooth positive functions that are 1-periodic in all variables, where $\xi' := (\xi_1, \dots, \xi_{n-d}) \in \mathbb{R}^{n-d}$, $d, n \in \mathbb{N}$, $d < n$. We consider the following domain:

$$\omega = \left\{ \xi \in \mathbb{R}^n : \ \xi' \in (0,1)^{n-d}; \ -h_-^{(k)}(\xi') < \xi_{n-d+k} < h_+^{(k)}(\xi'), \ k = \overline{1,d} \right\}.$$

Let $T_0$ be a finite family of closed and disjoint domains with smooth boundary such that $T_0 \subset \omega$. With the help of $\omega$ and $T_0$ we define the following sets: $\omega_0 = \omega \setminus T_0$, $T_0^\varepsilon = \varepsilon \cdot T_0 = \{x : \ \varepsilon^{-1} x \in T_0\}$, $T^\varepsilon = \bigcup_{\mathbf{z}_0 \in \mathbb{Z}^n} \left(T_0^\varepsilon + \varepsilon \mathbf{z}_0\right)$, where $\mathbf{z}_0 = (z_1, \dots, z_{n-d}, 0, \dots, 0) \in \mathbb{Z}^n$, $\varepsilon$ is a small positive parameter (see Figures 41.1).

Let $\Omega$ be a bounded domain in $\mathbb{R}^{n-d}$ with smooth boundary. A model thin perforated domain with limiting dimension $n-d$ is defined as follows: $\Omega_\varepsilon^{n-d} = Q_\varepsilon \setminus T^\varepsilon$ (see Figures 41.2, 41.3), where

$$Q_\varepsilon = \Big\{ x = (x_1, \dots, x_{n-d}, x_{n-d+1}, \dots, x_n) \in \mathbb{R}^n : \\ x' := (x_1, \dots, x_{n-d}) \in \Omega, \\ -\varepsilon h_-^{(k)}\left(\tfrac{x'}{\varepsilon}\right) < x_{n-d+k} < \varepsilon h_+^{(k)}\left(\tfrac{x'}{\varepsilon}\right), \quad k = \overline{1,d} \Big\}.$$

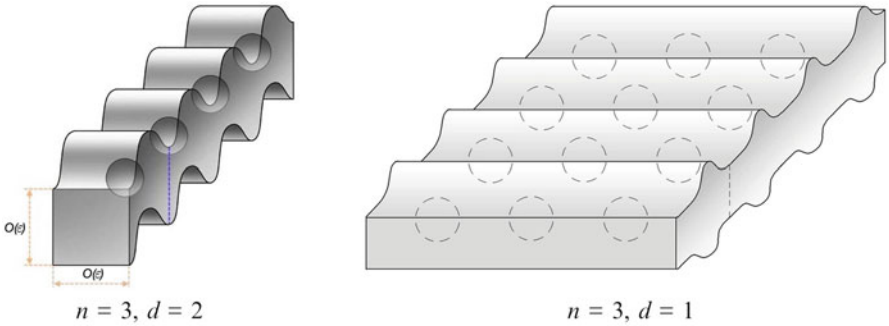


**Fig. 41.2** Examples of thin perforated domains $\Omega_\varepsilon^{n-d}$.

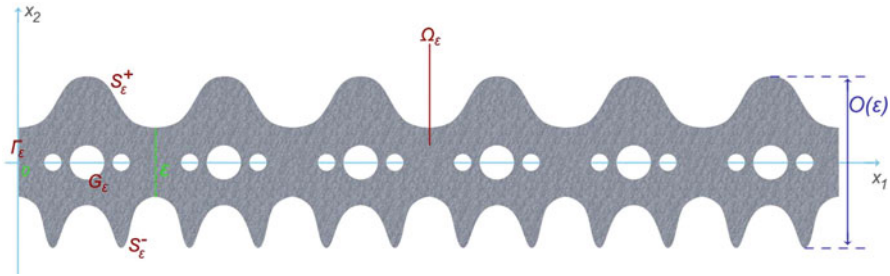


**Fig. 41.3** Thin perforated domain $\Omega_\varepsilon^{n-d}$, $n = 2$, $d = 1$.

Without loss of generality and in order to avoid additional technical difficulties, we assume that $\partial T_\varepsilon \cap \partial Q_\varepsilon = \emptyset$. For different parts of the boundary of $\Omega_\varepsilon^{n-d}$ we introduce the following notations:

$$S_\varepsilon^{\pm,i} = \Big\{ x : \, x' \in \Omega, \;\; x_{n-d+i} = \pm \varepsilon \, h_\pm^{(i)}\big(\tfrac{x'}{\varepsilon}\big), \\ x_{n-d+k} \in \Big( -\varepsilon \, h_-^{(k)}\big(\tfrac{x'}{\varepsilon}\big), \; \varepsilon \, h_+^{(k)}\big(\tfrac{x'}{\varepsilon}\big) \Big), \;\; k \in \{1,d\} \setminus \{i\} \Big\},$$

$$S_\varepsilon^{\pm} = \bigcup\nolimits_{i=1}^{d} S_\varepsilon^{\pm,i}, \quad G_\varepsilon = \partial T_\varepsilon \cap Q_\varepsilon, \quad \Gamma_\varepsilon = \partial \Omega_\varepsilon^{n-d} \setminus \big( S_\varepsilon^{\pm} \cup G_\varepsilon \big).$$

Let $L_\varepsilon \equiv \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} \Big( a_{ij}\big(\frac{x}{\varepsilon}\big) \frac{\partial}{\partial x_j} \Big)$ be a symmetric uniformly elliptic differential operator.

*Remark 1.* Here and in what follows summation over repeated indexes from 1 to $n$ is assumed. Also we will denote $\partial_{x_i} := \frac{\partial}{\partial x_i}$.

In the thin perforated domain $\Omega_\varepsilon^{n-d}$ we consider the following Steklov spectral problem:

$$\begin{cases} L_\varepsilon(u^\varepsilon) = 0 & \text{in } \Omega_\varepsilon^{n-d}, \\ \sigma_\varepsilon(u^\varepsilon) = \lambda(\varepsilon)\rho_\varepsilon u^\varepsilon & \text{on } G_\varepsilon, \\ \sigma_\varepsilon(u^\varepsilon) = 0 & \text{on } S_\varepsilon^\pm, \\ u^\varepsilon = 0 & \text{on } \Gamma_\varepsilon, \end{cases} \tag{41.1}$$

where $\lambda(\varepsilon)$ is the spectral parameter; the functions $\{a_{ij}(\xi)\}_{i,j=1}^n$ and $\rho(\xi)$ ($\xi \in \mathbb{R}^n$) are smooth and 1-periodic with respect to variables $\xi'$, $\rho$ is bounded by the positive constants: $0 < \rho_0 \le \rho \le \rho_1$; $\sigma_\varepsilon(u_\varepsilon) \equiv a_{ij}^\varepsilon \partial_{x_j} u_\varepsilon \nu_i\left(\frac{x}{\varepsilon}\right)$; $(\nu_1, \dots, \nu_n)$ is the unit outward normal to $\partial\Omega_\varepsilon^{n-d}$; $\rho_\varepsilon(x) := \rho(\frac{x}{\varepsilon})$, $a_{ij}^\varepsilon(x) := a_{ij}(\frac{x}{\varepsilon})$, $x \in \mathbb{R}^n$.

Our aim is to study asymptotic behavior of the eigenvalues and eigenfunctions of the problem (41.1) as $\varepsilon \to 0$.

## 41.3 An Auxiliary Integral Identity

Consider the following problem: find a function $N \in H^1_\sharp(\omega_0)$ such that

$$\begin{cases} L_{\xi\xi} N(\xi) = F_0(\xi) + \partial_{\xi_i} F_i(\xi), & \xi \in \omega_0, \\ \sigma_\xi(N(\xi)) = \Phi_0^\pm(\xi) + F_i(\xi)\nu_i(\xi), & \xi \in S^\pm, \\ \sigma_\xi(N(\xi)) = \Phi_1(\xi) + F_i(\xi)\nu_i(\xi), & \xi \in \partial T_0, \\ \langle N \rangle_{\omega_0} = 0. \end{cases} \tag{41.2}$$

Here $H^1_\sharp(\omega_0) := \{v \in H^1(\omega_0) :\ v - 1\text{-periodic with respect to } \xi'\}$;

$$L_{\xi\xi}(N) := \partial_{\xi_i}\left(a_{ij}(\xi)\,\partial_{\xi_j} N\right), \qquad \sigma_\xi(N) := a_{ij}(\xi)\partial_{\xi_j} N\,\nu_i(\xi);$$

$$\langle N \rangle_{\omega_0} = \frac{1}{|\omega_0|}\int\limits_{\omega_0} N(\xi)\,d\xi;$$

$(\nu_1(\xi), \dots, \nu_n(\xi))$ — outer normal to $\partial\omega_0$; $|\omega_0|$ — Lebesgue measure of domain $\omega_0$; $\{F_0, F_1, \dots, F_n\} \subset L^2(\omega_0)$; $\Phi_1 \in L^2(\partial T_0)$; $\Phi_0^\pm \in L^2(S^\pm)$; $S^\pm = \bigcup_{i=1}^d S^{\pm,i}$;

$$S^{\pm,i} = \Big\{\xi :\ \xi' \in [0,1]^{n-d},\ \ \xi_{n-d+i} = \pm h_\pm^{(i)}(\xi'),$$

$$\xi_{n-d+k} \in \left(-h_-^{(k)}(\xi'),\, h_+^{(k)}(\xi')\right),\ \ k \in \{1, \dots, d\} \setminus \{i\}\Big\}.$$

**Definition 1.** Function $N \in H^1_\sharp(\omega_0)$ is called a weak solution to the problem (41.2), if for any function $\psi \in H^1_\sharp(\omega_0)$

$$\int\limits_{\omega_0} a_{ij}\,\partial_{\xi_j} N\,\partial_{\xi_i}\psi\,d\xi = \int\limits_{\omega_0} \left(F_i\,\partial_{\xi_i}\psi - F_0\,\psi\right) d\xi + \int\limits_{\partial T_0} \Phi_1\,\psi\,d\sigma_\xi + \int\limits_{S^\pm} \Phi_0^\pm\,\psi\,d\sigma_\xi.$$

Similarly as in [BaPa84, p. 339] it can be shown that the problem (41.2) has a unique solution if and only if

$$\int\limits_{\omega_0} F_0(\xi)d\xi = \int\limits_{S^\pm} \Phi_0^\pm(\xi)\,d\sigma_\xi + \int\limits_{\partial T_0} \Phi_1(\xi)\,d\sigma_\xi. \tag{41.3}$$

Let $\psi_0 \in H^1_\sharp(\omega_0)$ be a weak solution (such that can be extended on $\Upsilon = \cup_{\mathbf{z}_0\in\mathbb{Z}^n}(\overline{\omega_0}+\mathbf{z}_0)$ ) to the following problem on the periodicity cell $\omega_0$:

$$\begin{cases} L_{\xi\xi}(\psi_0) = \Theta \text{ in } \omega_0, \\ \sigma_\xi(\psi_0) = \rho \text{ on } \partial T_0, \\ \sigma_\xi(\psi_0) = 0 \text{ on } S^\pm, \\ \langle\psi_0\rangle_{\omega_0} = 0, \end{cases}$$

where $\Theta = \widehat{\rho}\cdot|\omega_0|^{-1}$, $\widehat{\rho} = \int\limits_{\partial T_0} \rho\,d\sigma_x$. This problem satisfies solvability condition (41.3). Then $\varepsilon$-periodic function $\psi_0(\frac{x}{\varepsilon})$, $x \in \Omega_\varepsilon^{n-d}$ is a solution to the following problem:

$$\begin{cases} \partial_{x_i}\left(a_{ij}^\varepsilon(x)\,\partial_{x_j}\psi_0(\frac{x}{\varepsilon})\right) = \Theta\,\varepsilon^{-2}, & x \in \Omega_\varepsilon^{n-d}, \\ a_{ij}^\varepsilon(x)\,\partial_{x_j}\psi_0(\frac{x}{\varepsilon})\nu_i(\frac{x}{\varepsilon}) = \rho_\varepsilon\,\varepsilon^{-1}, & x \in G_\varepsilon, \\ a_{ij}^\varepsilon(x)\,\partial_{x_j}\psi_0(\frac{x}{\varepsilon})\nu_i(\frac{x}{\varepsilon}) = 0, & x \in S_\varepsilon^\pm, \\ \psi_0(\frac{x}{\varepsilon}) = 0, & x \in \Gamma_\varepsilon. \end{cases}$$

Multiplying the equation of this problem by an arbitrary function $\varphi \in H^1(\Omega_\varepsilon^{n-d})$ such that $\varphi|_{\Gamma_\varepsilon} = 0$ and integrating over the domain $\Omega_\varepsilon^{n-d}$ we obtain the following integral identity:

$$\Theta\,\varepsilon^{-1}\int\limits_{\Omega_\varepsilon^{n-d}} \varphi\,dx + \int\limits_{\Omega_\varepsilon^{n-d}} a_{ij}^\varepsilon(x)\,\partial_{\xi_j}\psi_0(\tfrac{x}{\varepsilon})\,\partial_{x_i}\varphi\,dx = \int\limits_{G_\varepsilon} \rho_\varepsilon\varphi\,d\sigma_x. \tag{41.4}$$

## 41.4 Equivalent Problem and Homogenized Problem

Let $H_\varepsilon := \left\{u \in H^1(\Omega_\varepsilon^{n-d}) \,:\, u|_{\Gamma_\varepsilon} = 0\right\}$ be a Hilbert space equipped with scalar product

$$\langle u, v\rangle_\varepsilon := \int\limits_{\Omega_\varepsilon^{n-d}} a_{ij}^\varepsilon \, \partial_{x_i} u \, \partial_{x_j} v \, dx, \qquad u, v \in H_\varepsilon.$$

We denote by $L^2(G_\varepsilon, \rho_\varepsilon)$ the weighted Lebesgue space with the scalar product

$$(u, v)_\varepsilon := \int\limits_{G_\varepsilon} \rho_\varepsilon \, u \, v \, d\sigma_x, \qquad u, v \in L^2(G_\varepsilon, \rho_\varepsilon).$$

Consider the following problem:

$$\begin{cases} L_\varepsilon(u^\varepsilon) = 0 & \text{in } \Omega_\varepsilon^{n-d}, \\ \sigma_\varepsilon(u^\varepsilon) = \rho_\varepsilon \, \varphi_\varepsilon & \text{on } G_\varepsilon, \\ \sigma_\varepsilon(u^\varepsilon) = 0 & \text{on } S_\varepsilon^{\pm}, \\ u^\varepsilon = 0 & \text{on } \Gamma_\varepsilon. \end{cases} \tag{41.5}$$

Multiplying the equation of problem (41.5) by an arbitrary function $\Psi_\varepsilon \in H_\varepsilon$ and integrating over $\Omega_\varepsilon^{n-d}$, we obtain the following identity:

$$\langle u^\varepsilon, \Psi_\varepsilon\rangle_\varepsilon = (\varphi_\varepsilon, B_\varepsilon \Psi_\varepsilon)_\varepsilon, \qquad \forall \Psi_\varepsilon \in H_\varepsilon, \tag{41.6}$$

where $B_\varepsilon : H_\varepsilon \to L^2(G_\varepsilon, \rho_\varepsilon)$ is the trace operator.

**Definition 2.** Function $u^\varepsilon \in H_\varepsilon$ is called a weak solution to the problem (41.5), if the identity (41.6) holds.

**Definition 3.** $\lambda(\varepsilon)$ is called an eigenvalue of problem (41.1), if there exists $u^\varepsilon \in H_\varepsilon$, $u^\varepsilon \neq 0$, such that

$$\langle u^\varepsilon, \Psi_\varepsilon\rangle_\varepsilon = \lambda(\varepsilon)(B_\varepsilon u^\varepsilon, B_\varepsilon \Psi_\varepsilon)_\varepsilon, \qquad \forall \Psi_\varepsilon \in H_\varepsilon; \tag{41.7}$$

and $u^\varepsilon$ is called an eigenfunction corresponding to the eigenvalue $\lambda(\varepsilon)$.

We define $A_\varepsilon := \varepsilon B_\varepsilon B_\varepsilon^*$, where $B_\varepsilon^*$ is conjugate to $B_\varepsilon$. By virtue of Proposition 1.1 from [Me94] $A_\varepsilon$ is self-adjoint, positive, and compact operator. It is easy to show that spectral problem for $A_\varepsilon$ is equivalent to the problem (41.1).

For every fixed $\varepsilon$, we can arrange the eigenvalues of the problem (41.1) in such a way that each eigenvalue is counted as many times as its multiplicity:

$$0 < \lambda_1(\varepsilon) < \lambda_2(\varepsilon) \le \lambda_3(\varepsilon) \le \cdots \le \lambda_m(\varepsilon) \le \cdots \to +\infty, \quad m \to +\infty. \tag{41.8}$$

Let us choose the respective eigenfunctions $u_m^\varepsilon \in H_\varepsilon$, $m \in \mathbb{N}$ such that

$$(B_\varepsilon u_m^\varepsilon, B_\varepsilon u_k^\varepsilon)_\varepsilon = \varepsilon^{d-1}\delta_{m,k} \qquad \forall m,k \in \mathbb{N}. \tag{41.9}$$

Similarly, as it was made in [Me94], we obtain homogenized problem for (41.1):

$$\begin{cases} \widehat{L}u + \Theta\,\lambda\, u = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{41.10}$$

where

$$\widehat{L}u = \sum_{p,q=1}^{n-d} \widehat{a}_{pq} \frac{\partial^2 u}{\partial x_p\, \partial x_q}; \qquad \widehat{a}_{pq} = \left\langle a_{pq} + \sum_{j=1}^{n} a_{pj} \frac{\partial N_q}{\partial \xi_j} \right\rangle_{\omega_0}, \quad p,q = \overline{1,n-d}.$$

Functions $N_p \in H^1_\sharp(\omega_0)$, $\quad p \in \{1,\dots,n-d\}$ are the solutions to such problems on the periodicity cell $\omega_0$:

$$\begin{cases} L_{\xi\,\xi}(N_p(\xi)) = -\partial_{\xi_i} a_{ip}(\xi), & \xi \in \omega_0, \\ \sigma_\xi(N_p(\xi)) = -a_{ip}(\xi)\nu_i(\xi), & \xi \in S^{\pm} \cup \partial T_0, \\ \langle N_p \rangle_{\omega_0} = 0. \end{cases} \tag{41.11}$$

## 41.5 Convergence Theorem

**Lemma 1 ([PoMe12]).** *There exists a linear operator* $\mathrm{P}_\varepsilon : H_\varepsilon \mapsto H_0^1(\Omega)$ *such that for any function* $u \in H_\varepsilon$

$$\|\mathrm{P}_\varepsilon u\|_{H^1(\Omega)} \le c\varepsilon^{-\frac{d}{2}} \|u\|_{H^1(\Omega_\varepsilon^{n-d})}. \tag{41.12}$$

**Lemma 2 ([PoMe12]).** *Suppose that the sequence* $\{u_\varepsilon\}_{\varepsilon>0} \subset H_\varepsilon$ *satisfies the inequality* $\sup_{\varepsilon>0} \|u_\varepsilon\|_{H^1(\Omega_\varepsilon^{n-d})} \le c\varepsilon^{\frac{d}{2}}$. *Then*

$$\varepsilon^{-d/2} \|u_\varepsilon - \mathrm{P}_\varepsilon u_\varepsilon\|_{L^2(\Omega_\varepsilon^{n-d})} \to 0 \quad \text{as} \quad \varepsilon \to 0.$$

Using Lemmas 1 and 2 we prove the following theorem.

**Theorem 1.** *Let* $\{\lambda_m(\varepsilon)\}_{m\in\mathbb{N}}$ *and* $\{\lambda_m\}_{m\in\mathbb{N}}$ *be the ordered sequences of eigenvalues and eigenfunctions of the problem* (41.1) *and* (41.10)*, respectively;* $\{u_m^\varepsilon\}_{m\in\mathbb{N}}$ *is the sequence of respective eigenfunctions of problem* (41.1)*, that are orthonormalized by condition* (41.9).

*Then for any $m \in \mathbb{N}$*

$$\lim_{\varepsilon \to 0} \frac{\lambda_m(\varepsilon)}{\varepsilon} = \lambda_m.$$

*There is a subsequence of the sequence $\{\varepsilon\}$, which is again denoted by $\{\varepsilon\}$, such that for any $m \in \mathbb{N}$*

$$\mathrm{P}_\varepsilon u_m^\varepsilon \to u_m \text{ weakly in } H_0^1(\Omega) \text{ as } \varepsilon \to 0,$$

*where $\{u_m\}_{m\in\mathbb{N}}$ — are the corresponding eigenfunctions of problem (41.10) such that*

$$\widehat{\rho} \int_\Omega u_m u_k \, dx' = \delta_{m,k}, \quad m,k \in \mathbb{N}.$$

*Proof.* Using the minimax principle for eigenvalues similarly as in [Me94] we deduce that for arbitrary $m \in \mathbb{N}$ there exists a positive constant $C_m$ (independent of $\varepsilon$) such that for $\varepsilon$ small enough the following estimate holds

$$C_0 \varepsilon \le \lambda_m(\varepsilon) \le C_m \varepsilon, \tag{41.13}$$

where $C_0$ is the positive constant independent of $\varepsilon$ and $m$.

From the relations (41.7) and (41.9) we have $\|u_m^\varepsilon\|^2_{H^1(\Omega_\varepsilon^{n-d})} \le c\varepsilon^d$, and hence $\|\mathrm{P}_\varepsilon u_m^\varepsilon\|_{H^1(\Omega_\varepsilon^{n-d})} \le C$. We denote

$$\gamma_{im}^\varepsilon(x') := \varepsilon^{-d} \int_{\Pi_\varepsilon^{(d)}} \chi_{\Omega_\varepsilon^{n-d}} \cdot a_{ij}^\varepsilon \cdot \partial_{x_j} (\mathscr{P}_\varepsilon u_m^\varepsilon) \, dx'', \quad i = 1, \dots, n-d.$$

where $\Pi_\varepsilon^{(d)} = \Big(-\varepsilon h_-^{(1)}(\frac{x'}{\varepsilon}),\, \varepsilon h_+^{(1)}(\frac{x'}{\varepsilon})\Big) \times \ldots \times \Big(-\varepsilon h_-^{(d)}(\frac{x'}{\varepsilon}),\, \varepsilon h_+^{(d)}(\frac{x'}{\varepsilon})\Big)$, $\chi_{\Omega_\varepsilon^{n-d}}$ is the characteristic function of domain $\Omega_\varepsilon^{n-d}$, $\mathscr{P}_\varepsilon : H^1(\Omega_\varepsilon^{n-d}) \mapsto H^1(Q_\varepsilon)$ is a uniformly bounded extension operator. Making use of the diagonal process, we can extract a subsequence from the sequence $\{\varepsilon\}$ (again denoting it by $\{\varepsilon\}$) for which the following limits are valid:

$$\frac{\lambda(\varepsilon)}{\varepsilon} \to \lambda_m^*, \tag{41.14}$$

$$\mathrm{P}_\varepsilon u_m^\varepsilon \to u_m \text{ weakly in } H^1(\Omega), \tag{41.15}$$

$$\gamma_{im}^\varepsilon \to \gamma_{im} \text{ weakly in } L^2(\Omega) \text{ as } \varepsilon \to 0.$$

Let us rewrite the identity (41.9) in the following form, using the formula (41.4) with $\varphi = u_k^\varepsilon \cdot u_m^\varepsilon$

$$\Theta\varepsilon^{-d} \int\limits_{\Omega_\varepsilon^{n-d}} u_k^\varepsilon\, u_m^\varepsilon\, dx + \varepsilon^{1-d} \int\limits_{\Omega_\varepsilon^{n-d}} a_{ij}^\varepsilon\, \partial_{\xi_j} \psi_0\big(\tfrac{x}{\varepsilon}\big)\, \partial_{x_i}\big(u_k^\varepsilon\, u_m^\varepsilon\big)\, dx = \delta_{m,k}. \tag{41.16}$$

It is easy to prove that the second term in the left-hand side of (41.16) vanishes as $\varepsilon \to 0$. Let us prove that

$$\Theta\varepsilon^{-d} \int\limits_{\Omega_\varepsilon^{n-d}} u_k^\varepsilon\, u_m^\varepsilon\, dx \to \widehat{\rho} \int\limits_{\Omega} u_m\, u_k dx'.$$

Indeed:

$$\int\limits_{\Omega_\varepsilon^{n-d}} u_k^\varepsilon\, u_m^\varepsilon\, dx \pm \int\limits_{\Omega_\varepsilon^{n-d}} u_k\, u_m\, dx \pm \int\limits_{\Omega_\varepsilon^{n-d}} u_k^\varepsilon\, u_m\, dx =$$

$$= \int\limits_{\Omega_\varepsilon^{n-d}} u_k\, u_m\, dx + \int\limits_{\Omega_\varepsilon^{n-d}} u_k^\varepsilon\, (u_m^\varepsilon - u_m)\, dx + \int\limits_{\Omega_\varepsilon^{n-d}} u_m\, (u_k^\varepsilon - u_k)\, dx.$$

Consider obtained terms, multiplied by $\Theta\varepsilon^{-d}$:

$$\Theta\varepsilon^{-d} \int\limits_{\Omega} u_k\, u_m \int\limits_{-h_-^{(1)}(\frac{x'}{\varepsilon})}^{h_+^{(1)}(\frac{x'}{\varepsilon})} \cdots \int\limits_{-h_-^{(d)}(\frac{x'}{\varepsilon})}^{h_+^{(d)}(\frac{x'}{\varepsilon})} d\xi'' dx' \quad \to \quad \widehat{\rho} \int\limits_{\Omega} u_m\, u_k\, dx', \quad \varepsilon \to 0,$$

by virtue of Corollary 1.7 [OlSh90, Chapter I]. Due to (41.15) and the Lemma 2

$$\Theta\varepsilon^{-d} \left| \int\limits_{\Omega_\varepsilon^{n-d}} u_k^\varepsilon\, (u_m^\varepsilon - u_m)\, dx \right| \le \Theta\varepsilon^{-d} \cdot \|u_k^\varepsilon\|_{L^2(\Omega_\varepsilon^{n-d})} \cdot \|u_m^\varepsilon - u_m\|_{L^2(\Omega_\varepsilon^{n-d})} \le$$

$$\le c_1\, \varepsilon^{-d/2} \|u_m^\varepsilon - \mathrm{P}_\varepsilon u_m^\varepsilon\|_{L^2(\Omega_\varepsilon^{n-d})} + c_2\, \|\mathrm{P}_\varepsilon u_m^\varepsilon - u_m\|_{L^2(\Omega)} \to 0,\ \varepsilon \to 0.$$

Similarly $\Theta\varepsilon^{-d} \int\limits_{\Omega_\varepsilon^{n-d}} u_m\, (u_k^\varepsilon - u_k)\, dx \to 0$. Passing to the limit in (41.16) as $\varepsilon \to 0$ we obtain

$$\widehat{\rho} \int\limits_{\Omega} u_k\, u_m\, dx' = \delta_{m,k}. \tag{41.17}$$

Now let us show that for any $v \in H_0^1(\Omega)$

$$\varepsilon^{-d}\lambda_m(\varepsilon)(B_\varepsilon u_m^\varepsilon, B_\varepsilon v)_\varepsilon \to \lambda_m^* \widehat{\rho} \int\limits_\Omega u_m v\, dx', \quad \varepsilon \to 0. \tag{41.18}$$

Using the formula (41.4) for $\varphi = u_m^\varepsilon \cdot v$ we have

$$\varepsilon^{-d}\lambda_m(\varepsilon)(B_\varepsilon u_m^\varepsilon, B_\varepsilon v)_\varepsilon = \varepsilon^{-d-1}\lambda_m(\varepsilon)\Theta \int\limits_{\Omega_\varepsilon^{n-d}} u_m^\varepsilon v\, dx + \\ + \varepsilon^{-d}\lambda_m(\varepsilon) \int\limits_{\Omega_\varepsilon^{n-d}} a_{ij}^\varepsilon \partial_{\xi_j}\psi_0\left(\tfrac{x}{\varepsilon}\right) \partial_{x_i}(u_m^\varepsilon v)\, dx.$$

Using the estimate for eigenvalues $\lambda_m(\varepsilon)$ it is easy to prove that second term in the right-hand side vanishes as $\varepsilon \to 0$. Similarly as above, using (41.14), we have

$$\frac{\lambda_m(\varepsilon)}{\varepsilon} \cdot \varepsilon^{-d}\Theta \int\limits_{\Omega_\varepsilon^{n-d}} u_m^\varepsilon v\, dx \to \lambda_m^* \widehat{\rho} \int\limits_\Omega u_m v\, dx', \quad \varepsilon \to 0.$$

We have proved that relation (41.18) holds.

We rewrite the relation (41.7) in the following way (with $\Psi_\varepsilon = v \in H_0^1(\Omega)$):

$$\int\limits_\Omega \sum_{i=1}^n \gamma_{im}^\varepsilon \cdot \frac{\partial v}{\partial x_i} dx' = \varepsilon^{-d}\lambda_m(\varepsilon) \int\limits_{G_\varepsilon} B_\varepsilon u_m^\varepsilon B_\varepsilon v\, d\sigma_x.$$

Passing to the limit as $\varepsilon \to 0$, we have

$$\int\limits_\Omega \sum_{p=1}^{n-d} \gamma_{pm} \frac{\partial v}{\partial x_p} dx' = \lambda_m^* \widehat{\rho} \int\limits_\Omega u_m v\, dx'. \tag{41.19}$$

Let us find the functions $\gamma_{pm}(x')$, $x' \in \Omega$. It follows from (41.11) that the function $N_p$ satisfies the following integral identity

$$\int\limits_{\Omega_\varepsilon^{n-d}} \left(a_{ip}^\varepsilon + a_{ij}^\varepsilon \partial_{\xi_j} N_p\right) \partial_{x_i} v\, u_m^\varepsilon\, dx + \int\limits_{\Omega_\varepsilon^{n-d}} a_{ip}^\varepsilon v\, \partial_{x_i} u_m^\varepsilon dx + \\ + \int\limits_{\Omega_\varepsilon^{n-d}} a_{ij}^\varepsilon v\, (\partial_{\xi_j} N_p)\, (\partial_{x_i} u_m^\varepsilon)\, dx = 0.$$

Here $p \in \{1,\dots,n-d\}$. Subtracting this identity from the integral identity (41.7) with the test-function $\varepsilon \cdot v(x') \cdot N_p(\frac{x}{\varepsilon})$, where $v$ is an arbitrary function from $H_0^1(\Omega)$, we get the following relation

$$\varepsilon^{-d} \int\limits_{\Omega_\varepsilon^{n-d}} \left( \left( a_{ip}^\varepsilon + a_{ij}^\varepsilon \partial_{\xi_j} N_p \right) \partial_{x_i} v \cdot u_m^\varepsilon + a_{ip}^\varepsilon \cdot v \cdot \partial_{x_i} u_m^\varepsilon \right) dx = \mathscr{O}(\varepsilon).$$

Similarly as above, passing to the limit in the last relation as $\varepsilon \to 0$, we obtain

$$|\omega_0| \int\limits_\Omega \sum_{i=1}^{n-d} \widehat{a}_{ip}\, u_m\, \partial_{x_i} v\, dx' + \int\limits_\Omega \gamma_{pm}\, v\, dx' = 0,$$

and hence

$$\gamma_{pm} = |\omega_0| \sum_{i=1}^{n-d} \widehat{a}_{ip}\, \partial_{x_i} u_m, \quad p \in \{1,\dots,n-d\}.$$

From (41.19) now we have

$$\int\limits_\Omega \sum_{p,q=1}^{n-d} \widehat{a}_{pq}\, \partial_{x_p} u_m\, \partial_{x_q} v\, dx' = \lambda_m^* \Theta \int\limits_\Omega u_m\, v\, dx'. \tag{41.20}$$

It follows from (41.20) that $u_m$ is an eigenfunction of the homogenized problem (41.10) and $\lambda_m^*$ is the corresponding eigenvalue. Moreover, since the eigenvalues of problem (41.1) are ordered so as to form the increasing sequence, we have by virtue of (41.17) that

$$0 < \lambda_1^* < \lambda_2^* \le \ldots \le \lambda_m^* \le \ldots, \qquad \lim_{m\to+\infty} \lambda_m^* = +\infty.$$

Let us show that $\lambda_m^* = \lambda_m, \forall m \in \mathbb{N}$. We assume the contrary. Let $\mathrm{w}_0$ be an eigenfunction of the homogenized problem (41.10), and the corresponding eigenvalue $\mu_0 \neq \lambda_m^*, \forall m \in \mathbb{N}$. Also we assume that

$$\widehat{\rho} \int\limits_\Omega \mathrm{w}_0^2\, dx' = 1, \qquad \int\limits_\Omega \mathrm{w}_0\, u_m\, dx' = 0.$$

Denote by $\mathrm{w}_\varepsilon \in H_\varepsilon$ the unique weak solution to the following problem:

$$\begin{cases} L_\varepsilon(\mathrm{w}_\varepsilon) = 0 & \text{in } \Omega_\varepsilon^{n-d}, \\ \sigma_\varepsilon(\mathrm{w}_\varepsilon) = \varepsilon\, \mu_0\, \rho_\varepsilon\, \mathrm{w}_0 & \text{on } G_\varepsilon, \\ \sigma_\varepsilon(\mathrm{w}_\varepsilon) = 0 & \text{on } S_\varepsilon^\pm, \\ \mathrm{w}_\varepsilon = 0 & \text{on } \Gamma_\varepsilon. \end{cases}$$

Similarly to the first part of the proof, it is easy to show that

$$\mathrm{P}_\varepsilon \mathrm{w}_\varepsilon \to \mathrm{w}_0 \text{ weakly in } H_0^1(\Omega) \text{ as } \varepsilon \to 0. \tag{41.21}$$

For simplicity, we can regard that $\lambda_1 = \mu_0$ and it means that $\lambda_1 < \lambda_1^*$. Define the function $\widetilde{\mathrm{w}}_\varepsilon := \mathrm{w}_\varepsilon - \varepsilon^{1-d}(\mathrm{w}_\varepsilon, u_1^\varepsilon)_\varepsilon \cdot u_1^\varepsilon$.

Since $(\widetilde{\mathrm{w}}_\varepsilon, u_1^\varepsilon)_\varepsilon = 0$, due to minimax principle we have

$$\varepsilon^{-d}\lambda_1(\varepsilon) \cdot (\widetilde{\mathrm{w}}_\varepsilon, \widetilde{\mathrm{w}}_\varepsilon)_\varepsilon \le \varepsilon^{-d}\langle \widetilde{\mathrm{w}}_\varepsilon, \widetilde{\mathrm{w}}_\varepsilon \rangle_\varepsilon.$$

Passing to the limit in this inequality with regard to (41.21), identity (41.4), and the properties of function $\mathrm{w}_0$, we obtain the contradiction: $\lambda_1^* \le \mu_0 = \lambda_1$.

To complete the proof, it suffices to observe that similar arguments are valid for any subsequence of the sequence $\{\varepsilon\}$ considered at the beginning of the proof.

## 41.6 Conclusions

We have combined asymptotic algorithms for studying spectral problems with rapidly oscillating coefficients in thin perforated domains with different limit dimensions. Convergence theorem for the eigenvalues and eigenfunctions of Steklov spectral problem in a thin perforated domains was proved. In particular, we showed that all the eigenvalues of the Steklov problem in such domains tend to zero as $\varepsilon \to 0$.

Under the assumption of certain symmetry condition on the coefficients of differential operators and on the geometry of thin domain it is possible to construct full asymptotic expansions for eigenfunctions and eigenvalues of the Steklov problem in $\Omega_\varepsilon^{n-d}$, similarly as it was performed in [PoMe12, MePo12] for thin perforated domains and in [Me94] for perforated cube.

## References

[Go76] Gol'denveizer, A. L.: The Theory of Elastic Thin Shells, Nauka, Moscow (1976)

[Go62] Gol'denveizer, A.L.: Derivation of an approximate theory of bending of a plate by the method of asymptotic integration of the equations of the theory of elasticity, Prikl. Mat. Meh. **26**, No. 4, pp. 668–686 (1962);

[Dz72] Dzhavadov, M.G.: Asymptotics of solutions of a boundary-value problem for second-order elliptic equations in thin domains, Differ. Urav. **4**, No. 10, pp. 1901–1909 (1968);

[Ca84] Caillerie, D.: Thin elastic and periodic plates, Math. Math. Appl. Sci. 6, pp. 159–191 (1984)

[VaBu90] Vasil'eva, A.B., Butuzov, V.F., Asymptotic Methods in the Theory of Singular Perturbations, Vyssh. Shkola, Moscow (1990)

[Na82] Nazarov, S.A., The structure of the solutions of elliptic boundary value problems in thin domains, Vestn. Leningr. Univ. Ser. Mat. Mekh. Astron., No. 2, pp. 65–68 (1982)

[ViLy67] Vishik, M.I., Lyusternik, L.A.: Regular degeneralization and boundary layer for linear differential equations with parameter, Usp. Mat. Nauk **12**, No. 5, 3–192 (1957)

[PaRe87] Panasenko, G.P., Reztsov, M.V.: Averaging a three-dimensional problem of elasticity theory in an inhomogeneous plate, Dokl. Akad. SSSR **294**, No. 5, 1061–1065 (1987);

[Me91] Mel'nik, T.A.: Averaging of elliptic equations describing processes in strongly inhomogeneous thin perforated domains with rapidly changing thickness, Akad. Nauk Ukr. SSR **10**, 15–18 (1991)

[Me94] Mel'nyk, T.A.: Asymptotic expansions of eigenvalues and eigenfunctions for elliptic boundary-value problems with rapidly oscillating coefficients in a perforated cube, Tr. Semin. Im. Petrovskogo **17**, 51–88 (1994);

[AkNa04] Akimova, E.A., Nazarov, S.A., Chechkin, G.A.: Asymptotics of the solution of the problem of deformation of an arbitrary locally periodic thin plate, Tr. Mosk. Mat. O.-va **65**, 3–34 (2004);

[CiCh02] Cioranescu, D., Chechkin, G.A.: Vibration of a thin plate with a 'rough' surface In: Nonlinear Partial Differential Equations and their Applications. Coll'ege de France Seminar. Volume XIV. Studies in Mathematics and its Applications, Elsevier, Amsterdam etc. pp. 147–169 (2002)

[Na02] Nazarov, S.A.: Asymptotic Analysis of Thin Plates and Bars, Vol. 1, Nauchnaya Kniga, Novosibirsk (2002)

[Is88] Isakov, R.V: Asymptotics of the spectral series of the Steklov problem for the Laplace equation in a 'thin' domain with nonsmooth boundary, Mat. Zametki, 44:5, pp. 694–696 (1988)

[Ko99] Kolpakov, A.G.: The governing equations of a thin elastic stressed beam with a periodic structure, Prikl. Mat. Mekh. *63*, No. 3, 513–523 (1999);

[Ko85] Korn, R.V., Vogelius, V.: A new model for thin plates with rapidly varying thickness.II: A convergence proof, Quart. Appl. Math. **18**, No. 1, 1–22, (1985)

[OlSh90] Oleinik, O.A., Shamaev, A.S., Yosifyan, G.A.: Mathematical Problems in the Theory of Strongly Inhomogeneous Elastic Media, Moscow Univ. Press, Moscow (1990)

[BaPa84] Bakhvalov, N.S., Panasenko, G.P.: Homogenization of Processes in Periodic Media, Nauka, Moscow (1984)

[MePo12] Mel'nyk, T.A., Popov, A.V.: Asymptotic analysis of the Dirichlet spectral problems in thin perforated domains with rapidly varying thickness and different limit dimensions. In: Roderick V. N. Melnik, Alexandra V. Antoniouk (eds.) Mathematics and Life Sciences, pp. 90–109. De Gruyter, Berlin. 89–107 (2012)

[PoMe12] Mel'nyk, T.A., Popov, A.V.: Asymptotic analysis of boundary-value and spectral problems in thin perforated regions with rapidly changing thickness and different limiting dimensions, Matem. Sbornik, 203:8, 97–124 (2012)

# Chapter 42
# Semi-Analytical Solution for Torsion of a Micropolar Beam of Elliptic Cross Section

**S. Potapenko**

## 42.1 Introduction

The theory of micropolar elasticity [Er66] was developed to account for discrepancies between the classical theory and experiments when the effects of material microstructure were known to significantly affect the body's overall deformation. The problem of torsion of micropolar elastic beams has been considered in [Sm70]-[Ie71]. However, the results in [Sm70] are confined to the simple case of a beam with circular cross-section while the analysis in [Ie71] overlooks certain differentiability requirements required to establish the rigorous solution of the problem (see, for example, [Sc89]). In neither case is there any attempt to quantify the influence of material microstructure on the beam's deformation.

The treatment of the torsion problem in micropolar elasticity requires the rigorous analysis of a Neumann-type boundary value problem in which the governing equations are a set of three second order coupled partial differential equations for three unknown antiplane displacement and microrotation fields. This is in contrast to the relatively simple torsion problem arising in classical linear elasticity in which a single antiplane displacement is found from the solution of a Neumann problem for Laplace's equation [TiGo70]. This means that in the case of a micropolar beam with non-circular cross-section it is extremely difficult (if not impossible) to find a closed-form analytical solution to the torsion problem.

In this paper, we use a simple, yet effective, numerical scheme based on an extension of Kupradze's method of generalized Fourier series [KuEtAl79] to approximate the solution of the problem of torsion of an elliptical micropolar beam.

S. Potapenko (✉)
University of Waterloo, Waterloo, ON N2L 3G1, Canada
e-mail: spotapen@uwaterloo.ca



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_42

Our numerical results demonstrate that the material microstructure does indeed have a significant effect on the torsional function and the subsequent warping of a typical cross-section.

## 42.2 Torsion of Micropolar Beams

Let $V$ be a domain in $\mathbb{R}^3$ occupied by a homogeneous and isotropic linearly elastic micropolar material with elastic constants $\lambda, \mu, \alpha, \beta, \gamma$ and $\kappa$ whose boundary is denoted by $\partial V$. The deformation of a micropolar elastic solid can be characterized by a displacement field of the form $U(x) = (u_1(x), u_2(x), u_3(x))^T$ and a microrotation field of the form $\Phi(x) = (\varphi_1(x), \varphi_2(x), \varphi_3(x))^T$ where $x = (x_1, x_2, x_3)$ is a generic point in $\mathbb{R}^3$ and a superscript $T$ indicates matrix transposition. We consider an isotropic, homogeneous, prismatic micropolar beam bounded by plane ends perpendicular to the generators. A typical cross-section $S$ is assumed to be a simply connected region bounded by a closed $C^2-$ curve $\partial S$ with outward unit normal $n = (n_1, n_2)$. Taking into account the basic relations describing the deformations of a homogeneous and isotropic, linearly elastic micropolar solid [No86], we can formulate the problem of torsion of a cylindrical micropolar beam (see, for example, [Sm70] and [Ie71]) as an interior Neumann problem of antiplane micropolar elasticity [PoScMi05]:

Find $u \in C^2(S) \cap C^1(S \cup \partial S)$ satisfying

$$L(\partial x)u(x) = 0, \quad x \in S, \tag{42.1}$$

such that

$$T(\partial x)u(x) = f(x) \quad x \in \partial S. \tag{42.2}$$

Here, $L(\partial x)$ is the $(3 \times 3)$ - matrix partial differential operator corresponding to the governing equations of torsion of a micropolar beam [Ie71], $u(x_1, x_2) = (\varphi_1(x_1, x_2), \varphi_2(x_1, x_2), u_3(x_1, x_2))^T$, $T(\partial x)$ is the boundary stress operator [Ie71] and $f = (\gamma n_1, \ \gamma n_{1,}..\mu(x_2 n_1 - x_1 n_2))^T$.

In [PoScMi05], the boundary integral equation method is used to prove existence and uniqueness results in the appropriate function spaces for the boundary value problem 42.1 and 42.2 . As part of this analysis, it is shown that the solution of 42.1 and 42.2 can be expressed in the form of an integral potential.

## 42.3 Generalized Fourier Series

Let $\partial S_*$ be a simple closed Liapunov curve such that $\partial S$ lies strictly inside the domain $S_*$ enclosed by $\partial S_*$, and let $\{x^{(k)} \in \partial S_*, k = 1, 2, ...\}$ be a countable set of points densely distributed on $\partial S_*$. We set $S_*^- = \mathbb{R}^2 \backslash \overline{S}_*$, denote by $D^{(i)}$ the columns

of the fundamental matrix $D$. [PoScMi05] and by $F^{(i)}$ the columns of matrix $F$ which form the basis of the set of rigid displacement and microrotations associated with 42.1 and 42.2. That is,

$$F = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{42.3}$$

The following result is fundamental to the numerical scheme used to approximate the solution of the micropolar torsion problem. Its proof proceeds as in [KuEtAl79].

**Theorem 1.** *The set*

$$\{F^{(i)}, \theta^{(jk)}, i,j = 1,2,3, k = 1,2,...\}, \tag{42.4}$$

*where the $F^{(i)}$ are the columns of matrix 42.3 and*

$$\theta^{(jk)}(x) = T(\partial x) D^{(j)}(x, x^{(k)}),$$

*is linearly independent of $\partial S$ and fundamental in $L^2(\partial S)$.*

If we now introduce the new sequence $\{\eta^{(n)}\}_{n=1}^{\infty}$ obtained from 1 by means of a Gram-Schmidt orthonormalization process, and use the integral representation (Somigliana) formula for the solution of a boundary value problem [PoScMi05], then, as in [Co90] we can derive the approximate solution for the torsion problem in the form of generalized Fourier series:

$$u^{(n)}(x) = \widetilde{q}_3 \widetilde{F}^{(3)} - \sum_{r=1}^{n} q_r \int_{\partial S} P(x,y) \eta^{(r)}(y) ds(y) + G(x), \quad x \in S. \tag{42.5}$$

Here, the first term on the right-hand side is a rigid displacement independent of $n$, the Fourier coefficients $q_r$ are computed by means of the procedure discussed in [KuEtAl79] and [Co90], $P(x,y)$ is a matrix of singular solutions [PoScMi05] and $G(x)$ is given by

$$G(x) = \int_{\partial S} D(x,y) f(y) ds(y), \quad x \in \mathbb{R}^2 \backslash \partial S.$$

Since $\widetilde{q}_3$ cannot be determined in terms of the boundary data of the problem we can conclude that the solution is unique up to an arbitrary rigid displacement/microrotation which is consistent with the results obtained in [PoScMi05].

This numerical method is extremely attractive in that it inherits all the advantages of the boundary integral equation method and, as in the following example, can be shown to produce accurate, fast-converging and effective results.

## 42.4 Example: Torsion of an Elliptic Beam

Firstly, to verify the numerical method, it is a relatively simple matter to show that for the problem of a *circular* micropolar beam, the numerical scheme produces results which converge rapidly to the exact solution established in [Sm70] (that the cross-section does not warp, i.e. that the material microstructure is insignificant in the torsion of a *circular* micropolar bar). Of more interest however is the case of an elliptical micropolar bar [TiGo70] which, to the authors' knowledge, remains absent from the literature.

As an example, consider the torsion of a micropolar beam of elliptical cross-section in which the elastic constants take the following values : $\alpha = 3, \beta = 6, \gamma = 2, \kappa = 1$, and $\mu = 1$. The domain $S$ is bounded by the ellipse

$$x_1 = \cos t, \qquad x_2 = 1.5 \sin t.$$

As an auxiliary contour $\partial S_*$ we take a confocal ellipse

$$x_1 = 1.1 \cos t, .....x_2 = 1.6 \sin t.$$

Using the Gauss quadrature formula with 16 ordinates to evaluate the integrals over $\partial S$ and following the computational procedure discussed in [KuEtAl79] and [Co90], the approximate solution 42.5 is found to converge to eight decimal place accuracy for $n = 53$ terms of the series. Numerical values are presented below for representative points $(0,0), (0.25,0.25), (0.25,0.5)$ and $(0.5,0.75))$ inside the elliptical cross-section (see Table 42.1).

Note that if we compare the values of the out-of-plane displacement or torsional function $u_3$, with those obtained in the case of a classical elastic elliptic beam (these are 0, 0.02403812, 0.09615251, 0.14422874 at the same points - based on the exact solution for the warping function [TiGo70]), we conclude that there is up to a 15 percent difference at certain points. (In addition to the results presented in 42.1 we also considered several other points lying within the boundaries of the ellipse and arrived at a similar conclusion.)

In contrast to the case of a circular micropolar beam for which the cross-section remains flat [Sm70] (as in the classical case [TiGo70]), there is a significant difference in the torsional function for an elliptic beam made of micropolar material when compared to the same beam in which the microstructure is ignored (i.e., the classical case [TiGo70]).

**Table 42.1** Approximate Solution of M1icropolar Beam with Elliptic Cross-section with $n = 53$ in 42.5.

| Point in Cross-Section | | $(0,0)$ | $(0.25,0.25)$ | $(0.5,0.5)$ | $(0.5,0.75)$ |
|---|---|---|---|---|---|
| Microrotation about $x_1-$ axis | $\varphi_1$ | 0.74431942 | 1.17355112 | 1.24343810 | 1.82784247 |
| Microrotation about $x_2-$ axis | $\varphi_2$ | 0.48152259 | 0.97222035 | 1.11246544 | 1.36181203 |
| Antiplane Displacement | $u_3$ | 0.00006160 | 0.02139392 | 0.08461420 | 0.12380739 |

This method used here is easily extended, with only minor changes in detail, to the analysis of torsion of micropolar beams of any (smooth) cross-section where we again expect a significant contribution from the material microstructure.

## References


[Co90] Constanda, C.: *A Mathematical Analysis of Bending of Plates with Transverse Shear Deformation,* Longman Scientific & Technical, Harlow, England (1990).

[Er66] Eringen, A.C.: Linear theory of micropolar elasticity. *J. Math. Mech*, **15**, 909–923 (1966).

[Ie71] Iesan, D.: Torsion of Micropolar Elastic Beams. *Int. J. Engng. Sci*. **9**, 1047–1060 (1971).

[KuEtAl79] Kupradze, V.D., Gegelia, T.J., Basheleishvili, M.O., and Burchuladze, T.V.: *Three-Dimensional Problems of the Mathematical Theory of Elasticity and Thermoelasticity,* Elsevier Science Ltd, Amsterdam, Netherlands (1979).

[No86] Nowacki, W.: *Theory of Asymmetric Elasticity,* Polish Scientific Publishers, Warszawa, Poland (1986).

[PoScMi05] Potapenko, S., Schiavone, P. & Mioduchowski, A.: Antiplane Shear Deformations in a Linear Theory of Elasticity with Microstructure. *Journal of Applied Mathematics and Physics*(ZAMP), **56**, 516–528 (2005).

[Sc89] Schiavone, P.: On existence theorems in the theory of extensional motions of thin micropolar plates. *Int. J. Engng. Sci*. **27**, 1129–1133 (1989).

[Sm70] Smith, A.C.: Torsion and Vibrations of Cylinders of a Micro-polar Elastic Solid. *Recent Advances in Engineering Science* (edited by A.C. Eringen), **5**, 129–137 (1970).

[TiGo70] Timoshenko, S. & Goodier, J.: *Theory of Elasticity,* McGraw-Hill, New York, NY (1970).

# Chapter 43
# L1 Regularized Regression Modeling of Functional Connectivity

**M. Puhl, W.A. Coberly, S.J. Gotts, and W.K. Simmons**

## 43.1 Introduction

A network is referred to as 'dense' when there are a large number of connections between nodes in the network. A sparse network alternatively has a small number of connected nodes. At times it is beneficial to 'sparsify' a dense network to make the data easier to interpret. The brain itself is a very dense network with brain regions representing the nodes of the network, and the neurological pathways between regions representing the connections. We chose to investigate the statistical method known as the Least Absolute Selection and Shrinkage Operator (LASSO), as proposed by Tibshirani et. al. [Ti96], as a feature selection tool to be applied to functional connectivity data. This method is useful in cases when the number of subjects is significantly less than the number of variables. A shrinkage parameter causes a number of variables to be shrunk to zero, creating a sparser network. In this chapter, we analyze data from 86 social regions of the brain of 60 subjects that were identified as either neuro-typical disorder (TD) or autism spectrum disorder (ASD).

M. Puhl (✉) • W.A. Coberly
University of Tulsa, Tulsa, OK, USA
e-mail: maria-puhl@utulsa.edu; coberly@utulsa.edu

S.J. Gotts
Laboratory of Brain and Cognition National Institute of Mental Health Intramural Research Program, Bethesda, MD, USA
e-mail: gottss@mail.nih.gov

W.K. Simmons
Laureate Institute for Brain Research, Tulsa, OK, USA

University of Tulsa, Tulsa, OK, USA
e-mail: wksimmons@laureateinstitute.org



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_43

This created a network with 3656 pairwise correlations as predictor variables. At the same time, LASSO fits the remaining variables to a model which can be used to 'predict' whether a subject belongs to the TD or ASD classification.

## 43.2 MRI and fMRI

### *43.2.1 MRI*

A magnetic resonance imaging (MRI) scanner generates a magnetic field many times more powerful than the natural magnetic field of the earth. Atomic particles naturally 'spin' on their own around a central axis. This act of naturally spinning is called 'precession.' In the natural state, the nuclei in the body will precess in random directions resulting in a net magnetization of the body being zero. By exposing the body to the MRI scanner's intense magnetic field, the body's hydrogen atoms will align their spin axes with the magnetic field. Injecting additional energy in the system, in the form of radiofrequency (RF) pulses, at the right frequency (the 'resonance frequency'), causes the protons to absorb the energy and change the direction of their spin relative to the magnetic field. When the RF pulse is removed, the protons release that energy and return to their initial spin states. Radiofrequency coils in the MRI scanner can detect the energy emitted by the protons as they return to their normal spin state. By using a sequence of gradient pulses and small perturbations to the main magnetic field in the $x$, $y$, and $z$ directions, it is possible to identify the resonance energy at individual locations in the space that is being imaged [La08]. Importantly, the type of tissue in which the protons are embedded influences their rate of return to the natural spin state, and thus the resonance energy. As a result, it is possible to use this information to create images of the tissue structure of the brain. These images are typically very high resolution ($< 1$ mm in the $x$, $y$, and $z$ directions, producing a volume with resolution $< 1\text{mm}^3$) and can take from 5 minutes to an hour to collect, depending on the size of the image's field of view, the quality of the image, and a number of other factors. Structural MRI scans are an excellent way to non-invasively image the brain at high resolution (though not nearly high enough to image individual neurons). In the context of most neuroscience research, structural MRIs are collected to identify the physical anatomy of a research participant's brain, and detect anatomical abnormalities.

### *43.2.2 MRI Image Processing*

The signal that comes in from the system of gradient pulses can be approximately expressed as the Fourier transformation of the spin density at a single point in the frequency domain. This frequency domain is commonly referred to as $k$-space in the neuro-imaging field. If we let $M(x,y)$ be the spin density at the point $(x,y)$,

and $(k_x(t_j), k_y(t_j))$ be the point in the frequency domain at which the Fourier transformation is measured at time $t_j$, then we can express this measurement of the MR signal at the $j^{th}$ time point as [Li08]

$$S(t_j) \approx \int_x \int_y M(x,y) e^{\left(-2\pi i(k_x(t_j)x + k_y(t_j)y)\right)} dydx$$

Once this data is obtained, the inverse Fourier transformation will allow the data to be transformed into the image space, where most data analysis is performed. It is important to subtract the mean from *k*-space, or else the leading Fourier coefficient will dominate the image, causing a large bright spot in the center of the final image. Other important preprocessing steps that are done on the MR image involve removing thermal and system noise artifacts that are always present. A type of artifact that might need to be removed is noise caused by fluctuations in the strength of the MR signal over time. These appear as a stochastic process, and thus can be easily removed from the data. Other types of noise that are present are subject and task related. A person inside an MR scanner lying as still as possible will still move slightly due to breathing, heartbeat, and other physiological movements. These cannot be avoided. Depending on the task performed by the subject while in the system, additional noise artifacts may present themselves.

### *43.2.3 fMRI*

Functional magnetic resonance imaging (fMRI) is based on the same physical principles of structural MRI, but adjustments in the nature of the RF pulses mean that it is able to image changes in the ratio of oxygenated to de-oxygenated hemoglobin. The decay rate of the RF signal emitted by protons as they return to their normal spin direction relative to the magnetic field is related to small local inhomogeneities in the field. The larger the inhomogeneities, the smaller the signal received by the RF coils. Deoxygenated hemoglobin is more paramagnetic than oxygenated hemoglobin, thereby producing greater inhomogeneities in the local field. When neurons in a particular brain region become active, their increased metabolic activity causes a cascade of physiological processes resulting in an influx of oxygenated blood to the region. The increase in the relative amount of oxygenated hemoglobin results in a spike in the local signal intensity that can be seen in the fMRI image. As a result, a series of fMRI images taken over time can provide a record of changes (over time) in local metabolic activity. In the past few decades, this so-called blood oxygen level dependent (BOLD) contrast imaging has become the primary tool for studying the relationship between brain and cognition. fMRI scans may be used to map changes in brain activity, either during task-related scans or resting-state scans. Task related scans involve imaging the brain while a subject performs a specified cognitive task or behavior, such as cycling back and forth between resting for 15 seconds and continuously tapping a finger for 15 seconds, resulting in increased activity in the motor cortex contralateral to the hand used

in the task. Task-related scans can help determine which areas of the brain are activated during the performance of a given task. Resting state scans, in contrast, are scans that are taken while the subject is not involved in a specific task [FoRa07]. Generally, a research participant is simply asked to lie quietly in the scanner with eyes open, and to try not to think about anything in particular. These scans can be used to determine which brain regions exhibit correlated intrinsic activity. If the resting-state activity of two regions is observed to be reliably correlated, it is said that the two regions exhibit 'functional connectivity.' Importantly, there is a strong correlation in the brain between functional connectivity and underlying structural connectivity [HoSp09].

BOLD fMRI images are taken quickly, over a period of about two seconds, but with much lower spatial resolution than structural MRI images (usually a resolution of approximately 2–3 $mm^3$). In the present study, each resting state scan lasted 8 minutes and 10 seconds, with a temporal sampling resolution of 1 brain volume each 3.5 seconds. For details on the specific imaging parameters used in the fMRI data collection and pre-processing algorithms applied to the data, see Gotts et al. [GoSi12].

## 43.3 Explanation of LASSO

The LASSO (Least Absolute Shrinkage and Selection Operator) can be described as a constraint on the sum of absolute values of the model parameters with the sum constrained by a given constant as an upper bound or a 'penalty term' [Wh05]. The LASSO can be applied to a variety of statistical modeling methods, the most common is as an alternative to the method of least squares in linear regression [JaWi05].

The LASSO creates a subset of predictor variables. Because it is a shrinkage method, LASSO 'shrinks', or reduces, some coefficients of the predictor variables to zero. It does this by imposing a penalty based on their size, the above-mentioned penalty term [HaTi09]. Shrinking the coefficients results in an overall reduction in the variance of the coefficients [JaWi05]. The LASSO method will thus produce an easier to interpret model by eliminating a number of predictor variables. One would expect that the LASSO will perform well in a situation where a relatively small number of predictor values have substantial coefficients and the remaining predictors have coefficients that are zero or near zero [JaWi05]. The problem is that for some real data sets, the number of predictors is not known at the beginning of the data analysis [JaWi05]. Another problem is that highly correlated features sometimes present a problem in the LASSO algorithm. A group of variables are typically examined at one time. If a number of these variables are very highly correlated with one another, the algorithm will choose one to keep and shrink the rest. This means that while the features selected by the LASSO algorithm are significant, some equally as significant features may have been lost as well. Because of this, it might be useful to test a number of different methods on the data set and then use cross validation to determine which outcome is the best. Elastic-Net

regression which combines an $L_1$ (LASSO) and $L_2$ (Ridge) attempts to avoid this problem. This method is not examined in this paper, but will be looked at in the future.

### 43.3.1 Linear Regression LASSO

The Linear Least Squares LASSO is given by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

This equation is simply the residual sum of squares equation plus an additional constraint , $\lambda \sum_{j=1}^{p} |\beta_j|$. This constraint is known as the penalty term [JaWi05], and is in fact an $L_1$ penalty [HaTi09]. Recall that the $L_1$ norm, $\|\cdot\|_1$, of a coefficient vector is given by $\|\beta\|_1 = \sum |\beta_j|$. A similar method, known as 'Ridge Regression,' uses an $L_2$ constraint [JaWi05]. The $L_1$ penalty causes some coefficients to shrink to exactly 0 when $\lambda$ is sufficiently large [JaWi05]. The $\beta_j$ must cause a significant impact to 'survive' this form of continuous subset selection [HaTi09]. This means that the selection of $\lambda$ is crucial in the outcome of the model. We will investigate the selection methods for $\lambda$ in section 43.5.1. It is worth noting that the intercept term $\beta_0$ has no interaction with the penalty term. This is because the intercept is simply a measure of the mean value of the response if $\boldsymbol{X} = 0$. A simple way to estimate $\beta_0$ is to center the input matrix, $\boldsymbol{X}$, at mean zero, and then $\beta_0$ can be estimated by $\widehat{\beta_0} = \bar{y} = \sum_{i=1}^{n} y_i / n$ [JaWi05].

### 43.3.2 Logistic Regression and the LASSO

#### 43.3.2.1 Logistic Regression Review

Logistic regression models the probability that a response variable, $Y$, will belong to a particular category. In the simplest case, this is modeled with $Y = 1$ equating to a 'does belong to group' and $Y = 0$ equating to a 'does not belong to group.' We then attempt to model a relationship between $p(X) = \Pr[Y = 1|X]$ and $X$.

Recall that the linear regression model represented this probability using the function $p(X) = \beta_0 + \beta X$, where $\beta$ is a vector of coefficients and $X$ is the matrix of predictor variables. Logistic regression models $p(X)$ using a function that will give outputs between 0 or 1 for all values of $X$, given by the following equation:

$$p(X) = \frac{e^{\beta_0 + \beta X}}{1 + e^{\beta_0 + \beta X}} \tag{43.1}$$

From equation 43.1, we can create a relationship between the logistic and linear regression problems.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta X \tag{43.2}$$

Equation 43.2 is known as the log-odds or 'logit' link. Clearly the $\beta_0 + \beta X$ is the linear regression model. This link function is a one-to-one transformation from the linear regression model to the logistic model. To estimate $\beta$ we use the maximum likelihood method to fit the logistic regression model. The likelihood function is given by

$$L(\beta) = \prod_{i:y_i=1} p(x_i) \prod_{j:y_j=0} (1-p(x_j))$$

which can be shown to be

$$L(\beta|y_1,\ldots y_n) = \prod_{i=1}^{n} \left(\frac{1}{1+e^{-\beta^T x_i'}}\right)^{y_i} \left(\frac{e^{-\beta^T x_i'}}{1+e^{-\beta^T x_i}}\right)^{1-y_i}$$

Taking the log of both sides will give us the joint log-likelihood for $\beta$.

$$\ell(\beta) = \sum_{i=1}^{n} \left[ y_i \ln\left(\frac{1}{1+e^{-\beta^T x_i'}}\right) + (1-y_i)\ln\left(\frac{e^{-\beta^T x_i'}}{1+e^{-\beta^T x_i}}\right)\right]$$

$$= -\sum_{i=1}^{n} \left[(1-y_i)\beta^T x_i' + \ln\left(1+e^{-\beta^T x_i'}\right)\right]$$

We choose the estimates, $\hat{\beta}_0$, $\hat{\beta}$, as the values that maximize $\ell(\beta)$ [JaWi05].

#### 43.3.2.2 Logistic LASSO

If we now apply our $L_1$ penalty to an ordinary logistic regression, we will have the LASSO Logistic Regression Model [Wh05].

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} \left[(1-y_i)\beta^T \boldsymbol{x}_i' + \ln\left(1+e^{-\beta^T \boldsymbol{x}_i'}\right)\right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

The addition of the $L_1$ constraint onto the logistic regression model can be seen as adding a Lagrangian penalty to the joint log-likelihood of the model parameters [LeLe06].

## 43.4 Autism Spectrum Disorder

Autism spectrum disorder (ASD) is a developmental disorder characterized by enduring deficits in social communication and interactions with others. Individuals with ASD often exhibit intense idiosyncratic interests, as well as engage in overt repetitive behaviors. To meet the criteria for an ASD diagnosis, all of these symptoms must be present from early childhood (prior to 2 years of age) and must interfere with broad areas of the individual's normal functioning in society (e.g., school, occupation, normal social interactions, etc.). As the name would imply, the symptoms of ASD present along a spectrum with varying levels of impairment or disability. The adolescents in the present study were generally fairly high functioning. They were in their late teens, with normal intelligence, and were overwhelmingly male (as is generally characteristic of the population of individuals diagnosed with ASD). Refer to Gotts et al. [GoSi12] for details on the specific demographics of the ASD and neurotypical (TD) control participants in the present study.

## 43.5 Method

Computations were done in R. The data used is originally from Gotts et al. [GoSi12], and represents the functional connectivity between 96 brain regions. Preprocessing steps were completed with AFNI. The data was slice time corrected, normalized, transformed to the Talairach & Tournoux volume, and basic ANATICOR procedures for removing noise were completed. For detailed preprocessing steps, see Gotts et. al. [GoSi12]. The parcellations of the brain region were based on gray and white matter boundaries, using the labels provided by freesurfer (http://freesurfer.net) corresponding to the Desikan & Killany atlas. Of these regions, we used only subcortical and gray matter regions that were present throughout all 60 subjects. This left 91 regions. An additional 5 regions corresponding to the corpus callosum ROIs were excluded. This left us with 86 total regions analyzed in this paper. The data is given in $86 \times 86$ correlation matrices. Each column and row represents an individual brain region and an element $a_{ij}$ of the matrix represents the functional connectivity between region $i$ and region $j$. First, we transformed the upper triangular part of the matrices into a vector, column-wise. We applied a Fisher transformation to the values. There are 31 data sets corresponding to ASD subjects and 29 data sets corresponding to TD subjects. We create a new matrix, $3656 \times 60$, where the first row indicates 1 for ASD or 0 for TD. The remaining rows are the brain region pairs (each value corresponds to a pair of brain regions) and the columns are the individual subject cases. We used the 'glmnet' package in R (http://cran.r-project.org/web/packages/glmnet) to perform the Logistic LASSO. This package was created by Tibshirani et.al, who initially developed the idea of the LASSO.

### 43.5.1 Tuning Parameter Selection

The proper selection of $\lambda$ is a very important step when using LASSO to analyze a data set. As $\lambda$ increases, there will be more and more 'shrinkage' in the data. In other terms, a smaller value of $\lambda$ will result in a model with more coefficients than a larger choice of $\lambda$. This idea can be shown in a path diagram such as Figure 43.1(a).

Cross validation is usually used to select $\lambda$. Cross validation is a method of assessing how well a model can be generalized to an independent data set, and aims to avoid overfitting the model [JaWi05]. We compute deviance for a number of values of $\lambda$ and choose the value of $\lambda$ with the minimum deviance. [JaWi05]. The deviance of different values of $\lambda$ is seen in Figure 43.1(b). The graph also shows upper and lower standard deviation values for each $\lambda$ in the cross validation trials. Lower values of binomial deviance will result in a 'better' model. For our model, we chose $\lambda = 0.0305$. This $\lambda$ was chosen as it is the largest value of $\lambda$ that gives a minimum value of binomial deviance and the graph appears to have a relatively small rate of change with respect to changes in $\lambda$ in this area. This $\lambda$ value was also chosen because as seen in Figure 43.1(a), a large number of variables exit the model at this $\lambda$.

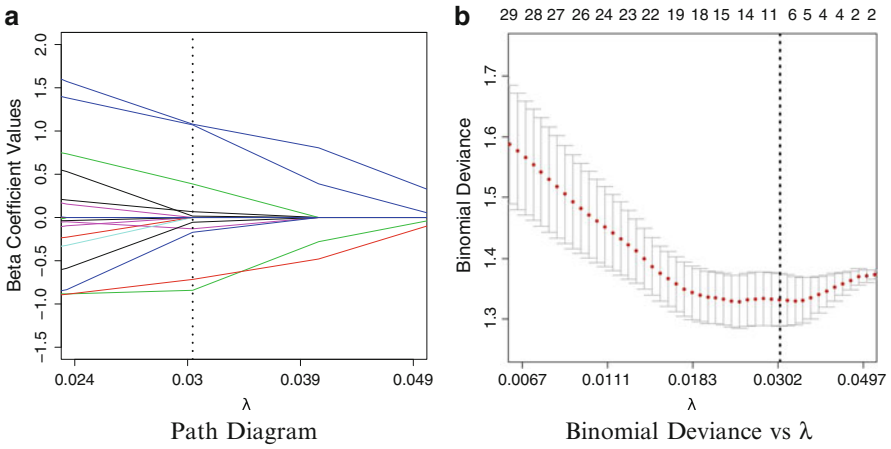


**Fig. 43.1** Note that $\lambda = 0.0305$ is marked in both graphs. **(a)** Each line represents a different variable and the value of the $\beta$ coefficient at each value of $\lambda$. Note that as $\lambda$ increases, more variables are shrunk to zero. **(b)** The amount of deviance explained by each value of $\lambda$, with the number of coefficients at each $\lambda$ value on the top axis.

## 43.6 Results

Using the value $\lambda = 0.0305$ creates a model with 10 coefficients. Recall that originally there were 3655 possible variables, so a large number of coefficients were shrunk to zero. We can look at how well this model 'predicts' the data set, in Figure 43.2. We selected a threshold value of 0.5. A subject with the probability of belonging to ASD being greater than 0.5 will be classified as 'ASD.' A subject with the probability belonging to ASD being less than 0.5 will be classified as 'TD.' Any subjects appearing in the shaded gray regions were misclassified. As seen in Figure 43.2, we have a number of subjects being misclassified in both the ASD and TD cases. In fact, 85% of the data set is predicted correctly. This may be misleading since this model is trained on the full data set. To more accurately evaluate how well the model performs we used a leave one out cross validation method to evaluate how well the model would work on test data. In table 43.1 we see the accuracy of the model, that is, the percentage of subjects correctly identified after performing the leave one out cross validation. We also look at the sensitivity (probability of being predicted as ASD when actually ASD) and specificity (probability of being predicted as TD when actually TD) of the model.

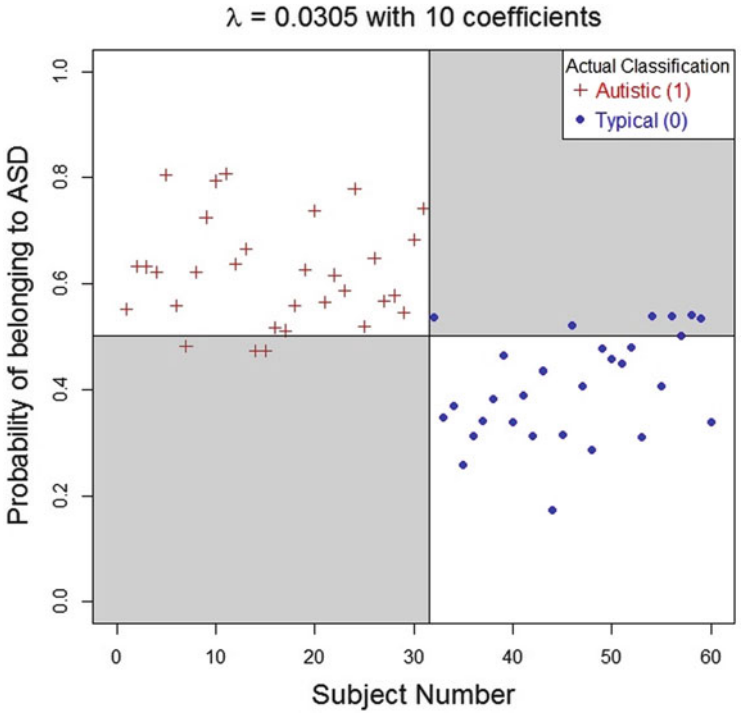


**Fig. 43.2** Results of using the model on the training data to predict if a subject was ASD or TD.

**Table 43.1** Accuracy (total correctly classified), Sensitivity (the probability of being predicted as ASD when actually ASD), and Specificity (probability of being predicted as TD when actually TD) of the model after leave one out cross validation.

| | |
|---|---|
| Accuracy | 65% |
| Sensitivity | 64.7% |
| Specificity | 65.38% |

**Table 43.2** Each row represents the connection between the two regions listed. The final column is the value of the $\beta$ coefficient in the model.

| $\beta$ **Coefficients** | | |
|---|---|---|
| **ROI Pairs** | | $\beta$ value |
| Right-Pallidum | ctx-lh-rostralanteriorcingulate | 1.08 |
| Right-Putamen | ctx-rh-superiortemporal | 1.07 |
| ctx-lh-cuneus | ctx-rh-isthmuscingulate | −0.84 |
| ctx-rh-inferiorparietal | ctx-rh-rostralanteriorcingulate | −0.72 |
| ctx-rh-caudalanteriorcingulate | ctx-rh-insula | 0.39 |
| ctx-lh-pericalcarine | ctx-rh-precentral | −0.17 |
| ctx-rh-entorhinal | ctx-rh-middletemporal | −0.13 |
| Right-Thalamus-Proper | Right-Hippocampus | 0.07 |
| ctx-lh-entorhinal | ctx-rh-parahippocampal | −0.06 |
| ctx-lh-bankssts | ctx-rh-frontalpole | 0.02 |

We finally look at the actual predictor variables the model chooses. Recall that each predictor variable represents a connection between two brain regions. These predictor variables are chosen and 'fit' to the model in a logistic regression sense, giving us coefficient values for each variable. The coefficient values of $\beta$ can be compared since all the $\beta$ coefficients come from a correlation matrix, and thus have the same scale. The value of the $\beta$ coefficient represents the strength it has in the given model. A larger $|\beta|$ will affect the model more than a smaller $|\beta|$. In Table 43.2 the predictor variables are given, along with their $\beta$ coefficient value.

## 43.7 Discussion

The regions identified by the model appear to have good face validity when compared both with the prior literature on the neural bases of autism and the findings published by Gotts et al. [GoSi12], Anderson et al. [AnNi11], and Di Martino et al. [DiKe11] using different analysis approaches on resting-state data in ASD. For example, from Table 43.2 we see that four of the five connections in the present study with the highest absolute beta coefficients involved regions of the cingulate gyrus (e.g., isthmus of the cingulate, caudal anterior cingulate, rostral anterior cingulate). This structure, which stretches along the midline immediately dorsal to the corpus callosum, is known to play important roles in many aspects of social cognition. For example, the posterior cingulate (of which the isthmus of the cingulate is a part) is a node in the brain's default mode network and has been shown to exhibit abnormal activity and connectivity in autism [LeSh14]. Likewise, the rostral anterior cingulate, which appeared in two of the top six connections, is an important visceromotor region that controls the affective modulation of autonomic functions. Likewise, the caudal (mid) cingulate, along with the insula, are regions that have been repeatedly shown to underlie interception and visceral autonomic responses to encountered stimuli [Cr09, CrHa13]. In fact, many of the connections between these two regions are by way of the unique von Economo neurons. These are unusually shaped neurons found only in hominid primates, elephants, and whales, that some have speculated may underlie elements of social awareness [CaGe14], which is a central element of ASD psychopathology. It is encouraging that our model highlighted connections among these regions as being particularly informative of functional connectivity differences in ASD. Future research will need to further explore how aberrant connectivity among these regions may contribute to the pathophysiology of ASD, and whether it will be possible to use measurements of the connection strengths among these regions as a diagnostic marker for autism.

**Acknowledgements** This work was supported by the National Institute of Mental Health, Division of Intramural Research Programs, project ZIA MH002920-06, by a NARSAD Young Investigator Award to W.K. Simmons, and by an NIMH grant (K01MH096175-01) to W.K. Simmons.

## References

[AnNi11] Anderson, J., Nielsen, J., Froehlich, A., DuBray, M., Druzgal, T. J., Cariello, A., Cooperrider, J., Zielinski, B., Ravichandran, C., Fletcher, P.T., Alexander, A., Bigler, E., Lange, N., Lainhart, J.: Functional connectivity magnetic resonance imaging classification of autism. Brain: A Journal of Neurology, **134**, 3742–3754 (2011)

[CaGe14] Cauda, F., Geminiani, G., Vercelli, A.: Evolutionary appearance of von Economo's neurons in the mammalian cerebral cortex. Frontiers in Human Neuroscience, **8.104**, 1–11 (2014)

[Cr09] Craig, A. D. B.: How do you feel now? The anterior insula and human awareness. Nature Reviews Neuroscience **10**, 59–70(2009)

[CrHa13] Critchley, H., Harrison, N.,: Visceral influences on brain and behavior. Neuron **77**, 624–638 (2013)

[DiKe11] Di Martino, A., Kelly, C., Grzadzinski, R., Zuo, X., Maarten, M., Mairena, M. A., Lord, C., Castellanos, F. X., Milham, M. P.: Aberrant Striatal Functional Connectivity in Children with Autism. Bilogical Psychiatry, **10.1016**, 847–856 (2011)

[FoRa07] Fox, M., Raichle, M.: Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nature Reviews Neuroscience, **8**, 700–711 (2007)

[GoSi12] Gotts, S., Simmons, W.K., Milbury, L., Wallace, G., Cox, R., Martin, A.: Fractionation of social brain circuits in autism spectrum disorders. Brain: A Journal of Neurology, **135**, 2711–2725 (2012)

[HaTi09] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data mining, Inference, and Prediction. Springer, New York (2009)

[HoSp09] Honey, CJ., Sporns, O., Cammoun, L., Gigandet X., Thiran, JP., Meuli R., Hagmann, P.: Predicting human resting-state functional connectivity from structural connectivity. Proceedings of the National Academy of Sciences, **106.6**, 2035–2040 (2009)

[JaWi05] James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning with Applications in R. Springer, New York (2013)

[La08] Lazar, N.: The Statistical Analysis of Functional MRI Data. Springer Science + Business Media, New York (2008)

[LeLe06] Lee, S., Lee, H., Abbeel, P., Ng, A.: Efficient $L_1$ Regularized Logistic Regression. Proceedings of the 21st National Conference on Artificial Intelligence (2006)

[LeSh14] Leech, R., Sharp, D. J.: The role of the posterior cingulate cortex in cognition and disease. Brain, **137**, 12–32 (2014)

[Li08] Lindquist, M.: The Statistical Analysis of fMRI Data. Statistical Science **23:4**, 439–464 (2008)

[Ti96] Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, Series B **58.1**, 267–288 (1996)

[Wh05] Wheeler, G.: The Lasso Logistic Regression Model: Modifications to Aid Causality Assessment for Adverse Events Following Immunization. London School of Hygiene and Tropical Medicine (2010)

# Chapter 44
# Automatic Separation of Retinal Vessels into Arteries and Veins Using Ensemble Learning

**N. Ramezani, H. Pourreza, and O. Khoshdel Borj**

## 44.1 Introduction

Diabetes is a persistent and life-threatening systemic disease causing abnormal increase in the level of glucose in the blood. After a time interval which is not so long, this high level of glucose starts damaging blood vessels. The damages may have negative effects on nervous systems, heart, kidneys, and other organs of the body. Rapid progress of the diabetes is one of the major challenges of today's medical care. The number of the people affected by the disease is dangerously increasing. As the early treatment methods can slow down its progress, early diagnosis of the disease is very important. The medical image analysis is a research field which has recently attracted lots of attention by scientists and physicians. The aim of this chapter is to develop and improve computer means which can help the physicians to diagnose and treat the disease. So far the best and the most effective treatments for diabetes have been made only in the early stages of the disease. Therefore, early diagnosis of the disease through continuous control of the patient is very important. The lowest expenses of such care and controls can be done with the technology of receiving digital images from retina. This technology can use high techniques of image processing which can detect retinal abnormalities. Anyway, for success in treatment, early diagnosis of the disease is essential. Regular and early tests are the only way for optimal treatment. By continuous control of the disease and constant study of the eyes such as comparing images taken in different periods of time which can be hardly done manually and can effectively be done by a computer-based approach, the disease can be treated in the quickest time possible. Changes in the vessels structures can have completely different impacts in the arteries and veins. Different diseases have a variety of impacts on the arteries

N. Ramezani (✉) • H. Pourreza • O. Khoshdel Borj
Islamic Azad University, Av. Emamieh, Mashhad, Iran
e-mail: NafiseRamezani1985@gmail.com; Hpourreza@ieee.org; Khoshdel.omid@gmail.com



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_44

and veins. For example, some diseases cause arterial bleeding and some others cause phlebitis. Moreover, one of the first signs of diabetic retinopathy is the reduction of ratio between the arteries and veins. These observations lead us to a strategy for developing a reliable classification technique for separation arteries and veins. Primarily we want to have a vessel classification around optic disc. Then using the structure of retinal vessels and tracking techniques, the technique can be extended to the regions out of this area where little or no data does exist for discriminating between vein and artery. Due to process of imaging and the convex shape of the retina, the retinal images generally have heterogeneous clearance and include a large variety of contrasts and local brightness. Thus the preprocessing step is essential in order to improve the results of segmentation.

### *44.1.1 Preprocessing Retinal Images*

Numerous algorithms have been proposed for improving contrast and correcting illumination, any of which has its own advantages on the special targets. Older techniques normalized the illumination of the image by omitting the low frequencies of illumination by using a high-pass filter. Other techniques were presented with the special target of retinal images. Normalizing the background using a large median filter is done to extract slow changes in illumination and then subtracting it from the main image. In order to estimate and correct the brightness intensity in the retinal images, a method has been proposed in [MaMi11]. It uses HSV color environment for better combination of brightness and chromatic information. It then uses a brightness model on the retinal background which eliminates many of the disadvantages of the previous methods specially when there is a big damage in the retina. The strategies which estimate correction from the whole image fail to create discrimination between brightness varieties due to different characteristics of brightness which leads to a total smoothing of the image's brightness changes. Nonlinear local adaptive filters will reduce the general differences between dark and bright characteristics even if they are capable to produce better local contrast and they won't guarantee reduction in brightness variety. Choosing vessels as qualifiers of brightness changes has its own disadvantages. Firstly the vessels have not been distributed throughout the whole image. For instance, there is no vessels in macular region that will lead to a highly un-crowded data in this region, making it difficult to estimate the brightness. Secondly, there is a high variety in reflection between arteries and veins. The different brightness patterns they offer, makes it extremely hard to have an assured estimation.

### 44.1.2 Vessel Segmentation

Automated segmenting of blood vessels in retinal images can assist physicians to control more people and recognize abnormal vessels caused by ophthalmic or systemic diseases. A number of methods have been proposed for vessels segmentation, neither of which, however, has shown a sufficient convincing result. Vessels segmentation is the problem of diagnosing a special line. Thus many of the blood vessels extraction algorithms are based on the line detection techniques. In [FrEtAl11] radial mapping method is used to locate central lines of the vessels which include small vessels as well. The main idea of this method is that if the pixel belongs to the vessel segment, the mapped curve of the pixel has typically a peak. Then the total gradient is used to extract the main structure of the vessels and the final segmentation is resulted of their combination. This method has been implemented on STARE data collection. The results have shown that this algorithm is able to detect in low contrasts and small vessels while it has well reduced the computation costs. In [PaEtAl10], the input image is primarily processed by Gaussian filter and matched filter response is gained which is used as starting points for iterative improved adaptive local threshold for blood vessels extraction. The results indicate that this method is appropriate for detecting large and small blood vessels. Paper [SeLa11] deals with neural network and wavelet improvement combination which includes wavelet transformation in preprocess phase and neural network in extraction phase. The topologies of different neural networks have been implemented on DRIVE image data collection with the aim of finding an effective combination method and SCG (Scaled Conjugate Gradient ) strategy with curve area ROC equals to 98% (the best result in neural network topologies) was gained. Paper [MaEtAl11] presents a function for total assessment of segmenting methods. This function is based on the characteristics of bound vessel structures in one block with indices like area and length with design sensitive to the anatomical characteristics of the vessel. A comparison between this method and other methods of segmenting assessment showed that the method of [MaEtAl11] has a higher degree of adaptation with the human's visual quality and thus can be used to increase the quality of segmenting in retinal images.

### 44.1.3 Separation of Retinal Vessels

In spite of high number of strategies for vessels segmentation, the discrimination between artery and vein is still an open debate. These vessels almost offer an identical texture, color, and form so there exists no algorithm that can appropriately overcome this problem. Moreover, this feature differs in different patients [SaEtAl12]. On the one hand, high color variety in the image due to light reflection and, on the other hand, some biological characteristics like skin color can produce different color patterns in the retina. As a result, correct and rapid labeling for retinal

vessels structure is essential for automated detection. Researches on classification techniques of retinal vessels are divided into two groups: Methods based on vessels characteristics and methods based on tracking. In [MuEtAl10], only color information is used to classify vein and artery. The target segment is classified to artery and vein in the train step using linear discriminant analysis (LDA). The rate of accuracy of classification considering the pixels of central line is 88.2%. In order to calculate the ratio of width of the artery to vein, a technique has been proposed to improve retinal vessels classification which uses the strategy of vessels clustering and tracking process based on the shortest route [PeEtAl10]. A new method is proposed in [VaEtAl10a] based on vessels segmentation strategy for automated classification retinal vessels in which K-means algorithm is used for feature vectors of each region. This algorithm computes the center of each cluster for the classes of vein and artery in each region and then using Euclidean distance assigns each feature vector to one class. 87% of the vessels of 58 images were classified correctly. The goal in paper [VaEtAl10b] is to develop an automated methodology for classifying retinal vessels which can also consider the effect of non-uniform brightness. In paper [NiEtAl10] a system with supervisor is applied. A method has been proposed that is able to isolate the retinal structure tree to a two-dimension color image in a way that they are on a data collection of 15 images to accuracy rate of 92% and 87% of correct detection of vessels pixels in comparison with manual labeling [JoEtAl11]. An automated method has been proposed to determine the ratio of artery size to vein for which the emphasis has been on extraction and selection of two main vessels. The sensitivity of this method for main vessels in the desired region is 87% while 93% of them have appropriately been classified to artery and veins [MuEtAl11]. Dashtbozorg in [DaMeCa14] offered an automatic method based on the analysis of a graph generated from the retinal vasculature for A/V classification, Accuracy value of 89.8% is achieved for the images of the VICAVR databases. Malek applied neural classification method for segmented vessels, which were extracted by match filters to classified vessels [MaTo13]. In [ReEtAl13], Relan divided retinal vessels in two groups: arteries or veins, using a Gaussian Mixture Model, an Expectation-Maximization (GMM-EM) unsupervised classifier, and a quadrant-pairwise method based on color features. All mentioned methods face problems when confronted with damaged retinal images and offer a low rate of accuracy. Note that patient retinal images are often affected by disease consequences.

## 44.2 Proposed Method

The data set applied in this chapter is VICAVR used to calculate artery-to-vein ratio. This data set has 58 images (normal, abnormal images) so far. In this chapter, primarily, Retinex improvement method is used to increase contrast and correct brightness of input images. Then combination of local entropy thresholding and modified co-occurrence matrix is done for segmenting blood vessels and after that

bifurcations and crossovers have been removed from vessel structure, then optic disc has been detected and region of interest is defined followed by Feature extraction and finally vessels clustering using Ensemble Learning is done.

### *44.2.1 Retinex Method*

Retinex theory [VaEtAl10b] is used in this research in order to make a model conforming to human's visual system of retinal image. In image model, the color component of the image is the multiplication of brightness and reflection. The basis of Retinex method is that the brightness changes occur slowly. Thus its frequency spectrum is in low frequencies. Therefore, the brightness is estimated and the output image is the result of subtracting estimated image with the main image. In [VaEtAl10b] two methods of single-scale retinex (SSR) and multi scale retinex (MSR) have been used. The illumination is estimated using a Gaussian form and, then, it is subtracted from the original image in Single-Scale Retinex (SSR). This is given by [VaEtAl10b]

$$R_i(xy) = logI_i(xy)log\big(F(xy) * I_i(xy)\big)$$

which $R_i(x,y)$ is output, $I_i(x,y)$ is the original image in the $i_th$ spectral band, $'*'$ shows the convolution operation, and $F(x,y)$ is a surround function [VaEtAl10b]:

$$F(xy) = k * (e^{(}\frac{-(x^2)+y^2}{(\partial^2)})),$$

where $\partial$ is the scale, controls extend of Gaussian surround, and

$$k = \frac{1}{(sum_x(sum_yF(x,y)))}$$

The Multi-Scale Retinex $(MSR)$ is simply a weighted sum of several different SSR outputs as follows [VaEtAl10b]:

$$R_MSRi = (sum_{n=1}^{N} W_nR_ni),$$

where $R_MSRi$ is output in the $ith$ color component,$N$ is the number of scales, $R_ni$ is the SSR output in the ith color component on the nth scale, and $w_n$ is the weight of the output of the nth scale.The results of implementation indicate that choosing weight $w_n = \frac{1}{3}$ for all scales with $N = 3$ is appropriate for all images. In Figure 44.1 sample and preprocess image from VICAVR data set has been shown by MSR method.

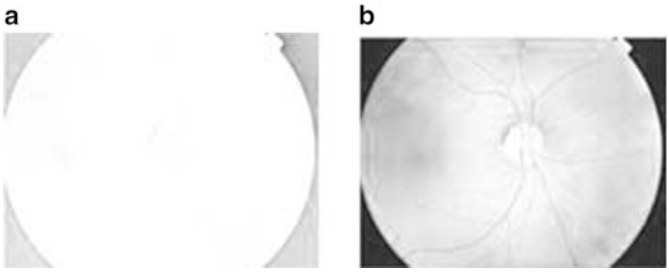


**Fig. 44.1** **(a)** Original image from VICAVR dataset. **(b)** Preprocessed image.


**Fig. 44.2** Modified co-occurrence matrix. Left: Computing modified co-occurrence matrix. Center: Original co-occurrence matrix in normalized logarithm scale. Right: Modified co-occurrence matrix in normalized logarithm scale [FrEtAl11].

As can be seen, the image brightness has been improved and an almost uniform picture has been presented.

### *44.2.2 Segmentation by Local Entropy Method Based on Thresholding*

An effective local entropy method based on thresholding for vessels extraction of retinal image is used in this study for preprocessed images. Methods proposed in [ChGuFr07, RoJi11] are an improvement to method [PaPa89] which uses thresholding algorithm based on local entropy for extracting retinal vessels. The vessels extracted by local thresholding are not often complete and some details in the structures are missed. In [PaPa89] two changes have been proposed to improve vessels extraction. Primarily some changes have been made in the definition of concurrent matrix so as to increase local entropy. The image concurrent matrix shows the transfer between brightness intensity of corresponding pixels. With change in the definition of concurrent matrix of the structures, similar spectral structures are retained and their variety is reduced. Figure 44.2 compares the original and modified co-occurrence matrix.

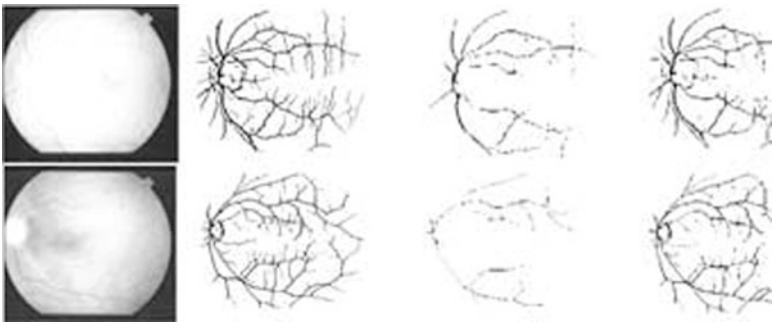

**Fig. 44.3** Vessel segmentation. Left: Original image from VICAVR dataset. Center: Segmentation on pre-processed image by applying combination of modified co-occurrence matrix and local entropy thresholding [PaPa89]. Right: Local entropy thresholding [ChGuFr07]. (d) Entropy thresholding [RoJi11].

As can be seen, both these matrices have similar structures that are significant for a creditable thresholding. Although modified co-occurrence matrix has larger entropy in spite of much smaller standard deviation which is more probable for local entropy thresholding. Then it uses sparse background for optimal selection of threshold. Choosing optimal threshold aims at maximizing background local entropy. The more the local entropy is, the more the balance between backgrounds will be. The results gained from [PaPa89] and combination with Retinex improvement method (proposed in this paper) with two methods of [ChGuFr07, RoJi11] have been compared in Figure 44.3.

The method of [PaPa89] in combination with Retinex method as proposed in this paper shows a higher efficiency and lower time of computation. In order to improve classification in next step, retinal vessel tree bifurcations and crossovers detected and removed from segmented vessels by [CaEtAl11]. After that, optic disc should be detected from retinal image [EaPo12] and then for next step the paired vessels are selected at distances of 2 to 2.5 times of optic disc radial with respect to the its' center. Due to the discriminant features of vessels in this region, vessel classification would be more accurate. In this research, the above-mentioned method is used for segmenting vessels on the green component of the image which produce suitable results even for the images of a damaged retina.

### *44.2.3 Feature Extraction*

In order to reduce the effect of non-uniform brightness and classified vessels, the image is divided into four overlapping regions. Then the vectors turn with 20 degree angle between $0^{\cdot}$ to $180^{0}$ [MaEtAl11]. Different 20 degree turns cause formation of overlapping regions and better compare of neighbor vessels under the effect of brightness almost uniform contrast. This rotation angle has been considered for

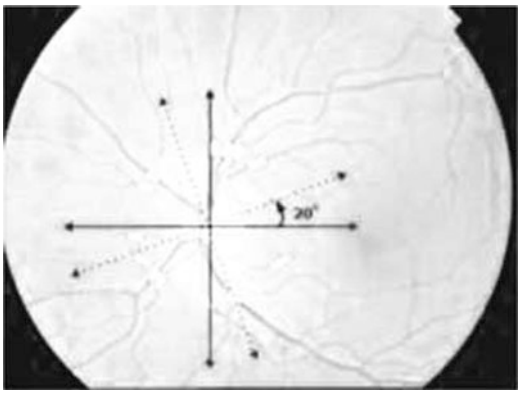


**Fig. 44.4** Dividing retinal images to overlapping regions [MaEtAl11].

| Nr. | Feature description |
|---|---|
| 1-3 | Normalized Mean Hue, Saturation and intensity across the vessel |
| 4-5 | Normalized Mean Red and Green plane intensities across the vessel |
| 6-8 | Standard deviation of Hue, Saturation and Intensity across the vessel |
| 9-10 | Standard deviation of Red and Green plane intensities across the vessel |
| 11-13 | Normalized Hue, Saturation and Intensity under centerline pixel |
| 14-15 | Normalized Red and Green plane intensity under the centerline pixel |
| 16-19 | Normalized highest and lowest intensity in the Red and Green plane across the vessel |
| 20-27 | Intensity under the centerline pixel in a Gaussian blurred ($\sigma$=2,4,8,16) version of the Red and Green plane |

**Fig. 44.5** Complete set of features extracted for each centerline pixel [NiEtAl10].

a compromise between time and the preciseness of the recommended method. It is obvious that minimizing the rotation angle increases the preciseness of feature extraction. While it increases the time of algorithm as well due to enlarging feature vector. In this way different regions are gained as shown in Figure 44.4.

Then the feature vector in each region is extracted. In this study, like in the method in [NiEtAl10], twenty-seven local characteristics are extracted based on Figure 44.5 for all recognized pixels in the center of the vessel.

### *44.2.4 Separation of Retinal Vessels Using Ensemble Learning*

Considering the above-mentioned problems of retinal images including lighting and non-uniform contrast, the impact of factors like age, the color of eyes as well as great similarity between artery and vein, it is not possible to apply one definite classifier

to all images and it reduces the accuracy of classifier. Thus in order to improve the results of classification and as a result, better separating the vessels into artery and vein, using Ensemble Learning method is proposed. In this method, by a combined use of a number of successful and efficient classifier, the final result of classification is improved. Even though this method takes a high time of computation, for using several classifiers, it can appropriately promote the accuracy of classification. For the same reason, using learning methods can result an appropriate separation rate for all images. Considering different papers each used different features and one special type of classifier for those features, Learning method can be used for combining different classifiers with local and feature vector appropriate for each classifier in order to increase the rate of separation accuracy. Any of these classifiers has to be efficient in an acceptable degree and act complementarily. For any of the extracted features, an appropriate classifier is selected. (Keeping the results of valid papers in the mind). Then each feature vector is assigned to the classifier special to it and finally, the results of different classifiers are combined with each other.For LDA Classifier [MuEtAl11, MuEtAl10, NiEtAl10] and for SVM [LiSh10, ChYe09, NaEtAl07, SeLa11] were studied. In Figure 44.6, each feature group, based on (44.5), is assigned to its suitable classifier based on available papers.

Thus in the present study, two methods of LDA and SVM (Support Vector Machine) will be used for different feature vectors of segmented vessels. By applying any of these classifiers to its related features, the possibility of belonging to any of the classes of artery or vein is assigned to each vessel segment. At the end, by performing Ensemble Learning and voting from different classifiers, it is determined to which class each vessel segment belongs to. After recognizing vessels segment inside the desired region and labeling them, this labeling spreads by tracking technique and all the vessels inside retinal image are separated. In Figure 44.7 , the result of applying the proposed method on a sample retinal image in different stages are shown.

**Fig. 44.6** Selecting appropriate Classifier for different feature vector.

| Classifier | Number of Features |
|---|---|
| LDA | 1-3 |
| LDA | 4-5 |
| LDA | 6-8 |
| LDA | 9-10 |
| SVM | 11-13 |
| SVM | 14-15 |
| SVM | 16-19 |
| SVM | 20-27 |

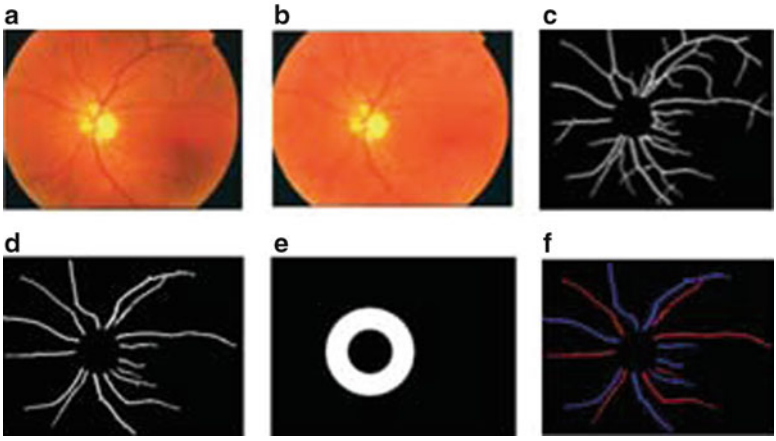


**Fig. 44.7** **(a)** Original image. **(b)** Preprocessed image. **(c)** Vessel segmentation. **(d)** Vessel structure without bifurcations and crossovers. **(e)** Region of interest around optic disc. **(f)** Vessel classification to artery (red) and vein (blue).

## 44.3 Conclusions

Separating retinal images in two groups of arteries and vein is a vital task to recognize the stage of disease in diabetes, it helps physicians to determine the manner and time of laser surgery. The present study proposed an effective method for dividing retinal vessels. The presented method for vessels separation is based on Ensemble Learning whose main objective is using several efficient classifiers and complementary for classifying the features of segmented vessels. The results of vessels separation have been compared with a number of other papers. The rate of accuracy for VICAVR retinal images equals 95.5% which is the highest value as compared with other papers and appropriate for damaged retinal images. One of the major advantages which have increased the efficiency of the proposed method in comparison with existing methods is using Ensemble Learning based on which several classifiers are trained to do classification. Moreover, using multi-purpose features and in several regions for retinal vessels has increased the accuracy of proposed method. Applying efficient methods of preprocessing has also helped vessels segmentation that has brought about increased accuracy of vessels separation.

## References

[MaMi11] Mahloojifar, A. and Miri, M.S.: Retinal Image Analysis Using Curvelet Transform and Multistructure Elements Morphology by Reconstruction. IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 58, NO. 5, MAY (2011)

[FrEtAl11] Fraz, M.M., Barman, S.A., Remagnino, P., Hoppe, A., Basit, A., Uyyanonvara, B., Rudnicka, A.R., and Owen, C.G.: An approach to localize the retinal blood vessels using bit planes and centerline detection. Computer Methods and Programs in Biomedicine, (2011)

[PaEtAl10] Palomera-Pérez, M.A., Martinez-Perez, M.E., Benitez-Pérez, H., and Ortega-Arjona, J.L.: Parallel Multiscale Feature Extraction and Region Growing: Application in Retinal Blood Vessel Detection. Information Technology in Biomedicine, IEEE Transactions on, 14(2): p. 500–506 (2010)

[SeLa11] Selvathi, D. and Lalitha Vaishnavi, P.: Gabor wavelet based blood vessel segmentation in retinal images using kernel classifiers. In Signal Processing, Communication, Computing and Networking Technologies (ICSCCN), International Conference on (2011)

[MaEtAl11] Marin, D., Aquino, A., Gegundez-Arias, M.E., and Bravo, J.M.: A New Supervised Method for Blood Vessel Segmentation in Retinal Images by Using Gray-Level and Moment Invariants-Based Features. Medical Imaging, IEEE Transactions. 30(1): p. 146–158 (2011)

[MuEtAl10] Muramatsu, C., Hatanaka, Y., Iwase, T., Hara, T., and Fujita, H: Automated detection and classification of major retinal vessels for determination of diameter ratio of arteries and veins. SPIE (2010)

[PeEtAl10] Penedo, M.G. , Saez, M., Vázquez, S.G., Cancela, B., and Barreira, N.: On the Automatic Computation of the Arterio-Venous Ratio in Retinal Images: Using Minimal Paths for the Artery/Vein Classification. IEEE,digital image computing (2010)

[VaEtAl10a] Vázquez, S.G., Barreira, N., Penedo, M.G., Penas, M., and Pose-Reino, A.: Automatic classification of retinal vessels into arteries and veins. In 7th International Conference Biomedical Engineering (BioMED) (2010)

[VaEtAl10b] VÃązquez, S.G., Penedo, M.G., Barreira, N., Penas, M., and Pose-Reino, A.: Using Retinex Image Enhancement to Improve the Artery/Vein Classification in Retinal Images, in Image Analysis and Recognition, A.Campilho and M. Kamel, Editors. Springer Berlin / Heidelberg (2010)

[NiEtAl10] Niemeijer, M., Xu, X., Dumitrescu, A.V., Gupta, P., van Ginneken. B., Folk, J.C., and Abramoff, M.D.: Automated Measurement of the Arteriolar-to-Venular Width Ratio in Digital Color Fundus Photographs. Medical Imaging, IEEE Transactions. 30(11): p. 1941–1950 (2010)

[JoEtAl11] Joshi, V.S., Reinhardt, J.M, Garvin, M.K, and Abramoff, M.D. : Automated method for the identification and analysis of vascular tree structures in retinal vessel network. SPIE (2011)

[MuEtAl11] Muramatsu, C., Hatanaka, Y., Iwase, T., Hara, T., and Fujita, H.: Automated selection of major arteries and veins for measurement of arteriolar-to-venular diameter ratio on retinal fundus images. Computerized Medical Imaging and Graphics. 35(6): p. 472–480 (2011)

[DaMeCa14] Dashtbozorg, B., Mendonca, A.M., and Campilho, A.: An Automatic Graph-Based Approach for Artery/Vein Classification in Retinal Images. Image Processing, IEEE Transactions. 23(3): p. 1073–1083 (2014)

[MaTo13] Malek, J. and Tourki, R.: Blood vessels extraction and classification into arteries and veins in retinal images. In Systems, Signals and Devices (SSD),10th International Multi-Conference on (2013)

[ReEtAl13] Relan, D., MacGillivray, T., Ballerini, L., and Trucco, E.: Retinal vessel classification: Sorting arteries and veins. In Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE (2013)

[ChGuFr07] Chanwimaluang, T., Guoliang, F., and Fransen, S.R.: Correction to Hybrid Retinal Image Registration. Information Technology in Biomedicine, IEEE Transactions. 11(1): p. 110–110 (2007)

[RoJi11] Rothaus, K. and Jiang, X.: Classification of Arteries and Veins in Retinal Images using Vessel Profile Features. AIP Conference Proceedings. 1(1), 371: p. 9–18 (2011)

[PaPa89] Pal, N.R. and Pal, S.K.: Entropic thresholding. Signal Processing. 16(2): p. 97–108 (1989)

[SaEtAl12] Saez, M., González-Vázquez, S., González-Penedo, M., Barcelá, M.A., Pena-Seijo, M., Coll de Tuero, G., and Pose-Reino, A.: Development of an automated system to classify retinal vessels into arteries and veins. Elsevier, Computer Methods and Programs in Biomedicine (2012)

[NaEtAl07] Narasimha-Iyer, H., Beach, J.M., Khoobehi, B., and Roysam, B.: Automatic Identification of Retinal Arteries and Veins From Dual-Wavelength Images Using Structural and Functional Features. IEEE,biomedical engineering (2007)

[ChYe09] Chin-Chen, C., Yen-Chang, C., and Chia-Chen, L.: A New Classification Mechanism for Retinal Images. In Information Technology and Computer Science(ITCS). International Conference on (2009)

[LiSh10] Lili, X.U. and Shuqian, L.: A novel method for blood vessel detection from retinal images. In Xu and Luo BioMedical Engineering OnLine (2010)

[CaEtAl11] Calvo, D., Ortega, M., Penedo, M.G., and Rouco, J.: Automatic detection and characterisation of retinal vessel tree bifurcations and crossovers in eye fundus images. Computer Methods and Programs in Biomedicine. 10: p. 28–38 (2011)

[EaPo12] Eadgahi, M.G.F. and Pourreza, H.: Localization of hard exudates in retinal fundus image by mathematical morphology operations. In Computer and Knowledge Engineering (ICCKE), 2nd International eConference on (2012)

# Chapter 45
# Study of Extreme Brazilian Meteorological Events

**H.M. Ruivo, F.M. Ramos, H.F. de Campos Velho, and G. Sampaio**

## 45.1 Introduction

Today, there is increasing scientific evidence that extreme climate and weather phenomena could become more frequent under a warmer planet [IP07]. This picture has been gradually emerging, since the first IPCC Assessment report in 1990, from a series of studies based on an increasing amount of data, which comprehensively covers the relevant atmospheric, land, ice, and ocean variables, computed or measured at different time intervals and spatial resolutions. These data sets come from remote instruments in satellites and in situ sensor networks, or are the outputs of computer simulations and reanalyzes [OvEtAl11]. Among the challenges generated by this deluge of data is the development of better technologies to store, distribute, analyze, and visualize their information content [HeTaTo10, Fo06].

Data mining, a computational process of discovering patterns in large data sets, extracts information and transforms it into an understandable structure for further use, in order to facilitate a better interpretation of existing data [FaEtAl96]. Here we present an innovative data mining approach to investigate the climatic causes of extreme events such as the Santa Catarina 2008 tragedy, and the Amazon droughts of 2005 and 2010. Our approach comprises two main steps of knowledge extraction, applied successively in order to reduce the complexity of the original data set, and identify a much smaller subset of climatic variables that may explain the event being studied. In the first step, we follow along the lines of [RuSaRa14], and apply a class comparison technique commonly used as a tool to analyze large data sets

H.M. Ruivo • F.M. Ramos • H.F. de Campos Velho (✉) • G. Sampaio
National Institute for Space Research, Av. dos Astronautas 1758,
São José dos Campos, SP, Brazil
e-mail: heloisa.musetti@cptec.inpe.br; fernando.ramos@inpe.br; haroldo@lac.inpe.br;
gilvan.sampaio@inpe.br



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_45

of genome-wide studies. This step results in a series of p-value spatial fields that identify which climatic variables behave differently across pre-defined classes of precipitation intensity. The second step consists of a decision tree (DT) learning algorithm used as a predictive model to map the set of statistically most significant climate variables identified in the previous step to classes of precipitation intensity. In the present context, the final result identifies a small subset of climatological variables that may explain or even forecast the extreme event in study.

The remainder of this chapter is organized as follows. Section 45.2 presents the methodology and data sets used in this investigation. Section 45.3 presents our results, while in Section 45.4 we draw some conclusions and discuss further developments.

## 45.2 Methodology

The data mining approach here employed comprises two main steps of knowledge extraction: class comparison, and decision trees. These methods are applied successively to reduce the complexity of the original data set and identify a much smaller subset of climatic variables that may explain the event being studied.

### *45.2.1 Class Comparison*

Class comparison methods are used for comparing two or more pre-defined classes in a data set. Here, we apply the class-comparison to time series of climatic grid box values or indices, but not to entire fields. The objective is to determine which variables in our data set behave differently across pre-defined classes of precipitation intensity ('high,' 'neutral,' and 'low,' for example). The 'no-difference' case corresponds to a null hypothesis. The classes are defined in such a way so as to capture in the correct class the main episodes of drought or extreme precipitation that occurred during the period being evaluated.

There are several methods for checking whether differences in variable values are statistically significant [Si03]. The F-test is a generalization of the well-known t-test, which measures the distance between two samples in units of standard deviation. Large absolute values of the F-statistic suggest that the observed differences among classes are not due to chance, and that the null hypothesis can therefore be rejected.

Supposing there are $J_1$ data points of class 1 and $J_2$ data points of class 2, the t-test score is computed as

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{s_p^2 \left(\frac{1}{J_1} + \frac{1}{J_2}\right)}} \tag{45.1}$$

where $s_p^2 = \frac{(J_1-1)s_1^2+(J_2-1)s_2^2}{J_1+J_2-2}$ and for i=1,2, $s_i^2 = \frac{1}{J_i-1}\sum_{j=1}^{J_i}(x_{ij}-\bar{x_i})^2$, where $\bar{x_1}$ is the mean of samples class 1 and $\bar{x_2}$ is the mean of samples class 2.

For more than two classes, an F-statistic shall be computed. In this case, the alternative to the null hypothesis is that at least one of the classes has a distribution that is different from the others. The t-test and F-test scores may be converted into probabilities, known as `p-values`. A `p-value` is the probability that one would observe under the null hypothesis a t-statistic (or F-statistic) as large as or larger than the one computed from the data. Both the t-test and F-test assume that the means are normally distributed, which may not hold, particularly when the number of data points is small. In this case, one could use the non-parametric counterparts of these tests, such as the Wilcoxcon test, the Kruskal–Wallis, or a permutation method.

The probability of observing an F-statistic as large as or larger than the one computed from the data is called a 'p-value.' It is a measure of statistical significance in the sense that one expects to observe, under the null hypothesis, p-values less than 0.01 only 1% of the time. Permutations methods, which do not rely on data normality assumptions, are commonly used for computing p-values [Si03, HaEtAl07]. For this, after calculating t-test scores for each variable, the class labels of the $J_1$ and $J_2$ are randomly permuted, so that a random $J_2$ of the samples are temporarily labeled as class 1, and the remaining $J_2$ samples are labeled as class 2. Using these temporarily labels, a new t-test score is calculated, say $t^*$. The labels are then reshuffled many times again, with a $t^*$ being computed at each permutation. The p-value from the permutation t-test is given by

$$p - value = \frac{1 + \text{number of random permutation where } |t^*| \geq |t|}{1 + \text{number of random permutation}} .$$

### 45.2.2 Decision Tree

The decision tree (DT) algorithm used here is the J4.8, from the WEKA package [WiFr00]. The J4.8 is a Java implementation of the C4.5 algorithm, which belongs to a succession of DT learners developed by Hunt and others in the late 1950s and early 1960s [Hu62]. DTs are tree-like recursive structures made of leafs, labeled with a class value, and test nodes with two or more outcomes, each linked to a sub-tree.

The input to a DT algorithm consists of a collection of training cases, each having a tuple of values for a fixed set of attributes (independent variables) and a class attribute (de-pendent variable). The goal is to generate a map that relates an attribute value to a given class. The classification task is performed following down from the root the path dictated by the successive test nodes, placed along the tree, until a leaf containing the predicted class.

Usually, DT learners use the divide-and-conquer strategy to construct a suitable tree from a training set. For this, the problem is successively divided into smaller

sub-problems until each subgroup addresses only one class, or until one of the classes shows a clear majority not justifying further divisions. Most algorithms attempt to build the smallest trees without loss of predictive power. To this end, the J4.8 algorithm relies on a partition heuristic that maximizes the 'in-formation gain ratio,' the amount of information generated by testing a specific attribute. This approach permits to identify the attributes with the greatest discrimination power among classes, and select those that will generate a tree that is both simple and efficient.

The information gain is measured in terms of Shannon's entropy reduction. Given a set A with two classes P and N, the information content (in bits) of a message that identifies the class of a case in A is then

$$I(p,n) = -\frac{p}{p+n} log_2 \left(\frac{p}{p+n}\right) - \frac{n}{p+n} log_2 \left(\frac{n}{p+n}\right)$$

where $p$ is the total number of objects belonging to class $P$, and $n$ is the total number of objects into the classes $N$. If $A$ is partitioned into subsets $A_1, A_2, \ldots, A_v$ by a given test $T$, the information gained is given by

$$G(A;T) = I(A) - \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(A_i),$$

where $A_i$ has $p_i$ objects from the class $P$, and $n_i$ from the class $N$. The algorithm chooses the test $T$ that maximizes the information gain ratio $G(A;T)/P(A;T)$, with

$$P(A;T) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} log_2 \frac{p_i + n_i}{p+n} .$$

## 45.3 Results

The climatic causes of the Santa Catarina 2008 tragedy and the Amazon droughts of 2005 and 2010 are investigated. The entire data sets used in the analysis can be freely downloaded from the Web. Surface- and pressure-level atmospheric fields have a spatial resolution of $2.5^o$ x $2.5^o$ and were extracted from NCEP/NCAR Reanalyzes [KaEtAl96]. Sea Surface Temperatures (SSTs) on a $2^o$ x $2^o$ grid were obtained from the NOAA Optimum Interpolation SST Analysis, version 2 [ReEtAl02]. The objective of this study design is to determine which variables in our data set behave differently across pre-defined classes of precipitation intensity. The 'no-difference' case corresponds to the null hypothesis for the applications considered here.

**Table 45.1** Data set used in this study.

| Variable | Unit | Observation | Number of time series |
|---|---|---|---|
| Sea surface temperature - SST | $^0C$ | Surface | 144 |
| Sea level pressure - SLP | Pa | 1000 hPa | 169 |
| Air temperature | $^0C$ | Surface | 169 |
| Specific humidity | g/kg | 850, 1000 hPa | 338 |
| Omega | Pa/s | 100, 200, 300, 400, 500, 600, 850, 1000 hPa | 1521 |
| Geopotential height | m | 1000 hPa | 169 |
| Zonal wind | m/s | 200, 500, 850 hPa | 507 |
| Meridional wind | m/s | 200, 500, 850 hPa | 507 |
| Cloud cover | % | Surface | 169 |

### *45.3.1 Extreme Rainfall Over Santa Catarina: Class Comparison*

The data set used in this study comprises 3,693 time series (Table 45.1). Gridded data cover a region delimited by latitudes $20^o$S and $50^o$S, and longitudes $30^o$W and $60^o$W. Since the episode of extreme rainfall in Santa Catarina was an event of short duration, pentad-averaged anomalies were used in the analysis.

The goal is to identify variables that might correlate with observed differences among classes of precipitation in the region of Blumenau (red dot in Figures 45.2 and 45.3), one of the most affected areas by the 2008 disaster. To this end, we analyzed 12 years (January 1999 up to December 2010) of pentad averages, comprising 3,693 environmental variables. Precipitation data in the region of Blumenau (Figure 45.1) is an average of five measurement stations of Brazilian National Water Agency (Agência Nacional de Águas, ANA) [Si10].

For classification purposes, the pentads of this time series were divided into three classes of precipitation intensity: 'strong,' 'moderate,' and 'light' rainfall. The standard t-test (solving 45.1) was applied, as recommended for applications with two classes: 'strong' (precipitation greater than 8), and 'moderate' (precipitation between 0 and 8). Results for the most significant variables identified by this procedure are presented in Figures 45.2 and 45.3. These results represent p-value fields, where coherent spatial patterns of low p-values indicate the existence of a possible links between omega and zonal/meridional wind anomalies, at different levels, and the precipitation intensity in the region of Blumenau (Figure 45.1). The isolines in Figure 45.2 correspond to omega anomalies averaged over the period November 22 up to 26, 2008, the period of most intense precipitation in Blumenau (delimited by the red bars in Figure 45.1). The wind fields in Figure 45.3 are also anomalies averages over the same period.

Regions with darker shades indicate the grid parameters with lower p-values. A p-value $< 0.01$, for example, indicates probability lower than 1% of being a

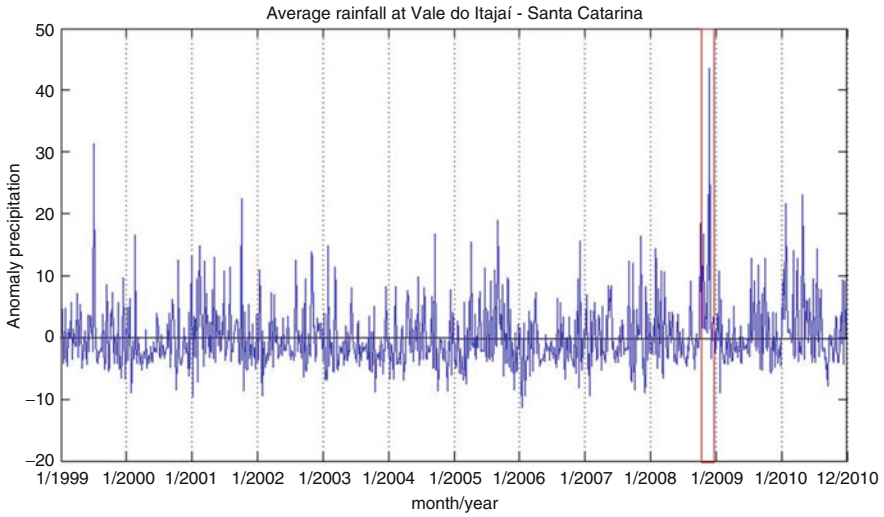


**Fig. 45.1** Average rainfall in Santa Catarina, Brazil.

false positive. Figures 45.2 and 45.3 show a dense dark area of low p-values for omega at different levels, which extends from the South Atlantic Ocean up the coast of Santa Catarina, and includes in its extreme west the area of Blumenau. During the extreme rainfall episode, we also observe (see the isolines) that omega values are negative over the continent (upward vertical motion) and positive over the ocean (downward vertical movement). It is well known that upward vertical motion over the continent can result in precipitation. This precipitation is fed by moisture transported from the ocean to the continent by easterly winds that predominated in the area in late November (see Figure 45.3). According to [Di00], the location of a blocking anticyclone on the Atlantic Ocean (with winds that rotate in anti-clockwise on the Southern Hemisphere) determined the occurrence of easterly winds on large part of the South Region coast, resulting in a large scale moisture transport from the ocean to the continent, particularly over the Itajaí valley.

### *45.3.2 Amazon Droughts: Class Comparison*

This analysis has used climatological data covering the period from January 1999 up to December 2010. Monthly anomalies were computed relative to the mean values over the period. The entire data set used in this illustrative study comprises 44,269 time series. The dataset also includes time series of the El Niño Southern Oscillation (ENSO) indices [NO07], the North Atlantic Oscillation (NAO) index (http://ossfoundation.us/projects/environment/global-warming/north-atlantic-oscillation-nao). Gridded data cover a region delimited by latitudes $40^o$N and $40^o$S and longitudes $140^o$W and $0^o$W.

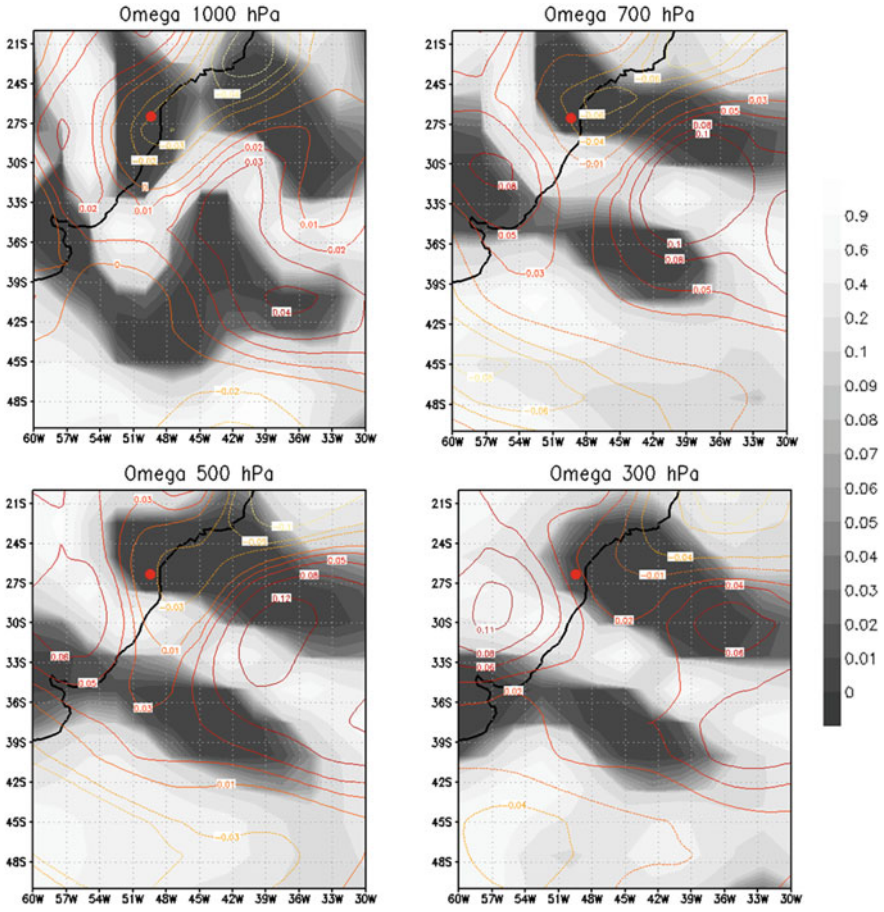


**Fig. 45.2** Representation in p-values of the climatic variable influence omega (1000, 700, 500, and 300 hPa) in Santa Satarina flood.

Class comparison was based on a time series of monthly accumulated precipitation anomalies [HuBo11], averaged over the area delimited by latitudes $4^oS$ and $8^oS$ and longitudes $68^oW$ and $72^oW$. This time series was used as proxy of drought in our analysis. This region, located in the south-western Amazon (indicated by a red square in Figure 45.4), was strongly affected by the droughts of 2005 and 2010 [LeEtAl11]. In this time series, the range of anomalies was split into 3 sub-classes: 'dry,' 'neutral,' and 'wet'. To this end, the interval is divided between the highest and the lowest precipitation anomaly into three parts, assigning the upper and lower 37% bins to the 'wet' and 'dry'classes, respectively, and the remaining 26% to the 'neutral' class. The results represent class comparison between 'dry' and 'neutral' classes.

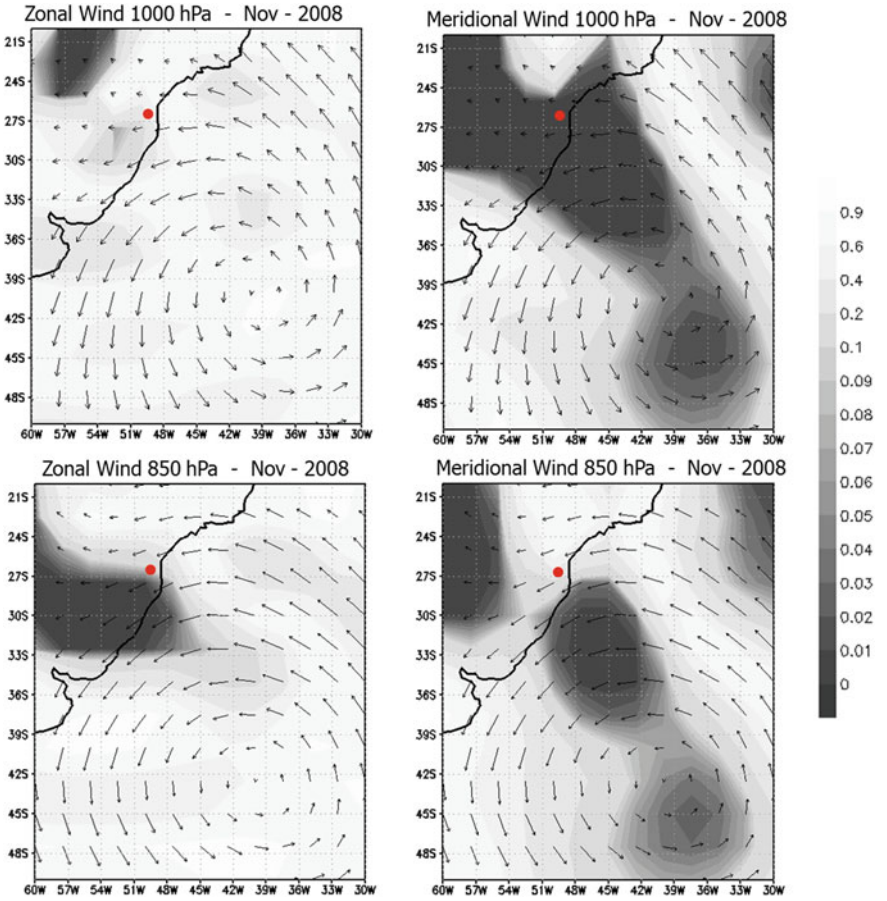


**Fig. 45.3** Representation in p-values of the climatic variables influences zonal and meridional wind (1000 and 850 hPa) in Santa Catarina flood.

One of the results is presented in the left side of Figure 45.4, that shows that the rainfall deficits in the South-Western Amazon region is linked with the widespread increase of the SST in the tropical North Atlantic, spanning from the coast of West Africa to the Caribbean. More results can be found in Ruivo et al. [RuSaRa14].

The Atlantic influence over the Amazon is modulated by seasonal and inter-annual variations in the strength and position of the intertropical convergence zone (ITCZ), following changes in the SST. This scenario, supported in the right side of Figure 45.4 presents the p-value field for the omega (vertical velocity) anomaly at 500 hP for August–September of 2009 and 2010, along with omega anomaly fields in two sub-areas in the region. These two years are used here as paradigms of years with accumulated precipitation above and below the climatic average, respectively. Negative anomalies indicate upward motion of the ITCZ. Note that the northward

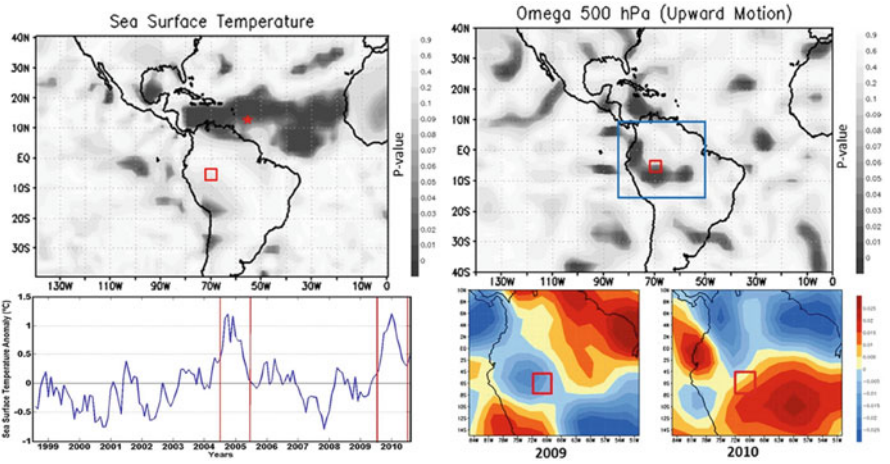


**Fig. 45.4** p-value field: left - sea-surface temperature anomaly, below: SST anomaly temporal evolution at 12.5$^o$N-55.5$^o$W (red star); right - omega (upward motion) anomaly at 500 hPa, below: pressure difference between grid points 15$^o$N-50$^o$W (red star) and 5$^o$S-60$^o$W (blue full circle).

shift of the downward branch of the Atlantic Hadley cell, favoring subsidence across the western and southern Amazon, is clearly captured in Figure 45.4. Weaker upward motion results in reduced convective development and rainfall.

### *45.3.3 Extreme Rainfall Over Santa Catarina: Decision Tree*

The decision tree with the J4.8 algorithm was created with confidence factor used for pruning (0.25), and number of instances per leaf (8). Several tests were performed: with fixed number of attributes (meteorological variable for different coordinates are considered different attribute) with smallest p-values. The best result was obtained with the 5 different climatological variables, considering 10 different coordinates for each variable, with smallest p-values (total 50 attributes). To this goal, the precipitation time series were divided over the area of Blumenau (red dot) in two classes: 'light' (values below the median), and 'strong' (values above the median), corresponding to episodes of low and high precipitation, respectively. The training set comprised data from 2000 up to 2006. The years of 1999, 2007, 2008, 2009, and 2010 were used to evaluate the tree performance. Figure 45.1 shows two rainfall intense episodes: July 1999, and November 2008. The event at July 1999 was less intense than November 2008.

The resulting tree, displayed in Figure 45.5, left side, has 7 leafs (4 'strong' and 3 'light') and 6 decision nodes. The variable with the highest information gain is omega at 500 hPa, and at coordinates 50$^o$W and 25$^o$S. As expected, these coordinates are as near to the disaster zone as the limited spatial resolution of the

gridded data permits. Note that all but one decision nodes are also associated with omega, at different pressure levels but always in the vicinity of the affected area. These results highlight the key role played in the episode of extreme rainfall in Santa Catarina 2008 by the vertical transport of the moisture, brought from the ocean by sustained easterly winds. As a predictor, the tree was able to forecast 100% of the cases of extreme rainfall during the evaluation years (1999, 2007–2010), including the episode occurred in July 2008.

### *45.3.4 Amazon Droughts: Decision Tree*

The decision tree was generated using 120 variables with lower p-values identified by the class-comparison methodology described in the previous section. To this end, the proxy precipitation anomaly time series was divided into two classes according to the median: 'dry' (values below the median), and 'wet' (values above the median). The training set comprised data from 1999 to 2004. The period from 2005 to 2010 was used for evaluating the predictive performance of the tree. The resulting tree has 7 leafs (4 'dry' and 3 'wet') and 6 decision notes (Figure 45.5, right side). Surprisingly, the variable with the highest information gain is the zonal wind at 200 hPa, at coordinates 72.5$^o$W and 25$^o$N.

This variable, together with a large area of zonal wind anomalies in North Atlantic, has indeed a very low p-value. This result supports recent claims [Ch00, CoZeYo10, MaEtAl11, MaEtAl11] that the recent episodes of intense drought in the Amazon are linked to the northwest displacement of the ITCZ. In 2010, for example, the ITCZ was displaced approximately five degrees northward from its climatic position [MaEtAl11]. Overall, the tree had hit rate of 83%, misclassifying only two months during the extreme drought periods of 2005 and 2010.

## 45.4 Conclusions

In this chapter, two techniques for data mining were used to investigate the climatic causes of two kinds of extreme events occurred in Brazil during the last decade: the Santa Catarina 2008 extreme rainfall tragedy and the Amazon droughts of 2005 and 2010. In both cases, our results are in good agreement with analyses published in the literature. The class comparison methodology was able to greatly reduce the size of the original data set, from the order of thousands of variables to a few tenths. The decision trees generated from the results of the class-comparison step were able to correctly classify/predict a high percentage of cases of extreme rainfall in Santa Catarina (100%) and of drought in the Amazon (83%). Overall, the data mining procedure here introduced has shown to be a promising approach in the investigation of climatic extreme events and the extraction of knowledge from large and complex data sets.

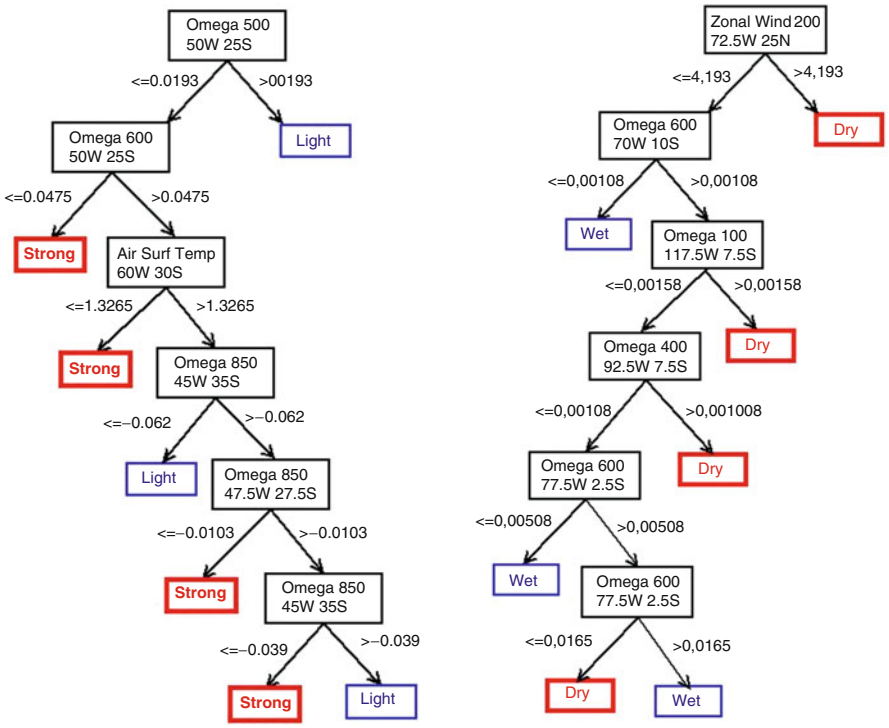


**Fig. 45.5** Decision trees: left - Extreme rainfall over Santa Catarina; right - Amazon droughts.

**Acknowledgements** This work was supported by grants from Brazil's CAPES, Ministry of Education, and CNPq, Ministry of Science and Technology. Analyses were performed using BRB-ArrayTools developed by Dr. Richard Simon and BRB-ArrayTools Development Team.

## References

[Di00] Dias, M.A.F.S.: As chuvas de novembro de 2008 em Santa Catarina: um estudo de caso visando a melhoria do monitoramento e da previsão de eventos extremos. http://pt.slideshare.net/comissaosantacatarina/defesa-civil-sc. Accessed 21 Feb 2009 (2008).

[IP07] IPCC: Cambio climático 2007: Informe de síntesis. Grupo Intergobernamental de Expertos sobre el Cambio Climático [Equipo de redacción principal: Pachauri, R.K. y Reisinger, A. (directores de la publicacion)] Ginebra, Suiza 104 (2007).

[OvEtAl11] Overpeck, T.J., Meehl, A.G., Sandrine, B., and Easterling, D.R.: Climate Data Challenges in the 21st Century. Science **331**, 700–702 (2011)

[HeTaTo10] Hey, T., Tansley, S., and Tolle, K.: The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/. Accessed 04 Nov 2011 (2010)

[Fo06] Foster, I.: A two-way street to science's future. Nature. **440**, pp 419 (2006)

[FaEtAl96] Fayyad, U., Piatesky-Shapiro, G., Smyth, P., and Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining. California The MIT Press **560** (1996).

[GaEtAl14] Ganguly, A.R., Kodra, E.A., Banerjee, A., Boriah, S., Chatterjee, S., and Choudhary A.: Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques, Nonlinear Processes in Geophysics Discussions. DOI = 10.5194/npgd-1-51-2014, **1** 51–96. (2014)

[RuSaRa14] Ruivo, H. M., Sampaio, G., and Ramos, F.M.: Knowledge extraction from large climatological data sets using a genome-wide analysis approach: application to the 2005 and 2010 Amazon droughts. Climatic Change; pp 1–15 (2014).

[Si03] Simon, R.M.: Design and analysis of DNA microarray investigations. Springer **209**, (2003).

[HaEtAl07] Hardin, J., Mitani, A., Hicks, L., and VanKoten, B.: A robust measure of correlation between two genes on a microarray. BMC Bioinformatics 8:220. (2007).

[WiFr00] Witten, I.H. and Frank, E.S.: Data mining: Practical machine learning tools and techniques with java implementation. Morgan Kaufmann Publishers. (2000).

[Hu62] Hunt, E.B.: Concept learning:An information processing problem. New York: Wiley (1962)

[KaEtAl96] Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., and Deaven, D.: The NCEP/NCAR 40-Year Reanalyses Project. Bull Amer Meteor Soc **77**, 437–471. (1996)

[ReEtAl02] Reynolds, R.W., Rayner, N.A., Smith, T.M., Stokes, D.C., and Wang, W.: An improved in situ and satellite SST analysis for climate. Journal of Climate **15**, 1609–1625. (2002)

[NO07] NOAA - Earth System Research Laboratory: Multivariate ENSO index (MEI). U.S. Department of Commerce, National Oceanic and Atmospheric Administration. http://www.cdc.noaa.gov/people/klaus.wolter/MEI/, Accessed 19 jul. 2010 (2007)

[HuBo11] Huffman, G.J. and Bolvin, D.T.: TRMM and Other Data Precipitation Data Set Documentation. Laboratory for Atmospheres, NASA Goddard Space Flight Center and Science Systems and Applications Inc. ftp://meso.gsfc.nasa.gov/pub/trmmdocs/3B42_3B43_doc.pdf.

[LeEtAl11] Lewis, S.L., Brando, P.M., Phillips, O.L., Van der Heijden, G.M.F., and Nepstad, D.: The 2010 Amazon Drought. Science 331:554–554. doi: 10.1126/science.1200807. (2011)

[Si10] Sistema Nacional de Informa cões sobre Recursos Hídricos (SNIRH) - Agencia Nacional de Águas (ANA). Available in: http://ana.gov.br/portalsnirh/. Accessed March 2010

[Ch00] Chao, W.C.: Multiple quasi equilibria of the ITCZ and the origin of monsoon onset. Journal of the atmospheric sciences,**57**(5), 641–652. (2000).

[CoZeYo10] Cook, B., Zeng, N., and Yoon, J.-H.: Climatic and ecological future of the Amazon: likelihood and causes of change. Earth Syst. Dynam. Discuss. (2010)

[MaEtAl11] Marengo, J.A., Tomasella, J., Alves, L.M., Soares, W.R., Rodriguez, D.A.: The drought of 2010 in the context of historical droughts in the Amazon region. Geophys Res Lett 38 L12703, doi:10.1029/2011GL047436. (2011)

[MaEtAl08] Marengo, J.A., Nobre, C.A., Tomasella, J., Cardoso, M.F., and Oyama, M.C.: Hydroclimatic and ecological behaviour of the drought of Amazonia in 2005. Philos Trans R Soc B pp 363:1773–1778. (2008)

# Chapter 46
# The Neutron Point Kinetics Equation: Suppression of Fractional Derivative Effects by Temperature Feedback

**M. Schramm, A.C.M. Alvim, B.E.J. Bodmann, M.T.B. Vilhena, and C.Z. Petersen**

## 46.1 Introduction

Fractional point kinetics has been discussed recently as one of the novel approaches that describes the short-term evolution of neutron densities as well as precursor concentrations in nuclear reactor theory. Kinetics may be derived from an original transport problem introducing simplifications that allow to decouple the time from spatial degrees of freedom. One of the motivations to extend the traditional point kinetics by additional terms that contain a fractional derivative is to improve the solution in the sense to compensate effects due to the simplifications mentioned above. Works on the fractional derivative point kinetics equation found in the literature [EsEtAl08, EsEtAl11, NeNe07, SaPa12] treat the extended kinetics problem, whereas in the present work we consider additionally temperature feedback on the reactivity [ElMa14, Na11, SiEtAl14], which mimics influences of thermohydraulics on neutronics and thus may be considered beside being a novelty also a more realistic model in comparison with the previous works.

It is noteworthy that in all the published literature on the topic, a variety of fractional derivative implementations are presented together with solutions for the extended point kinetics equation but nothing is said justifying, why a fractional

M. Schramm (✉) • B.E.J. Bodmann • M.T.B. Vilhena
Federal University of Rio Grande do Sul, Av. Osvaldo Aranha, 99/4, Porto Alegre 90046-900, RS, Brazil
e-mail: marceloschramm@hotmail.com; bardo.bodmann@ufrgs.br; vilhena@math.ufrgs.br

A.C.M. Alvim
Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil
e-mail: alvim@con.ufrj.br

C.Z. Petersen
Federal University of Pelotas, Pelotas, RS, Brazil
e-mail: claudio.petersen@ufpel.edu.br



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_46

derivative in point kinetics should make sense from a physical point of view. In the further we give some plausibility arguments for the use of a fractional derivative, however show in the end that such an extension is meaningless since its effects are suppressed by the temperature feedback.

In order to understand some properties that may be connected to physical characteristics, it is instructive to interpret the fractional differential [De03, OlSp74, Tr03]. The latter may be expressed in terms of the proper derivative term times a volume with fractional dimension. This fact implies in a specific scaling behavior if a standard length is changed, in other words the scaling of the fractional differential scales like a volume with fractional dimension. Scaling laws are intricately connected to fractal structures, that upon changing scales self-similarity appears. The multiplicative effects due to nuclear chain reactions could be interpreted in terms of self-similarity, since at the average each neutron in any generation and ramification that represents the fission process contributes effectively with the same multiplicity.

One may now establish a connection to a geometrical construction that represents multiplicative effects, such as the circumference length of the construction steps of a Koch curve. From one step to the subsequent one the addition of new edges adds new segments to the curve while the infinite limit creates a structure with Hausdorff dimension $\frac{\log(4)}{\log(3)}$, i.e. a fractal. Besides the fractional dimension one may also associate an effective time scale to the construction steps, that may characterize a generation lifetime, which is the second new parameter that defines the fractional derivative. Once having made plausible the use of a fractional derivative the discussion that follows, focusses on the question what influence these new degrees of freedom impose on the solution of point kinetics with temperature feedback.

## 46.2 The Fractional Derivative Model for Neutron Point Kinetics

The classical neutron point kinetics equations with temperature feedback effects has a known solution in an analytical representation (see [SiEtAl14] for details). There, the proposed methodology starts from the classical solution [SiEtAl14] and obtains an analytical representation for the fractional neutron point kinetics equations with temperature feedback using the decomposition method [Ad94]. This method consists in expanding the neutron density and $p$ precursor concentrations in a series ($n = \sum_{r=0}^{\infty} n_r$ and $C_i = \sum_{r=0}^{\infty} C_{ir}$, $i \in \{1, \ldots, p\}$) in which each term $n_r$, or $C_{ir}$ may be determined from a recursive set of linear differential equations as shown in the next section. The classical solution may be used as the recursion initialization by incorporating the fractional derivative terms that extends the usual time-derivative $\left(\tau^{\kappa} \frac{d^{\kappa+1} n}{dt^{\kappa+1}}, \left(\tau^{\kappa}\left(\frac{1}{l} - \frac{1-\beta}{\Lambda}\right) \frac{d^{\kappa} n}{dt^{\kappa}}\right.\right.$ and $\left.\left.\tau^{\kappa} \lambda_i \frac{d^{\kappa} C_i}{dt^{\kappa}}\right)\right)$ in the source terms together with the nonlinear terms $\frac{\alpha H}{\Lambda} n \int_0^t n(\hat{t})\, d\hat{t}$.

The problem as it stands needs the evaluation of an infinite number of terms or, equivalently, equations. In case of convergence, however, the series of $n$ and $C_i$ may be truncated according to a predefined numerical precision. In order to analyze convergence, a Lyapunov inspired criterion is used that determines the recursion depth until stable convergence (in our case exponential convergence) occurs. Details of this procedure are not discussed in the present contribution but may be found in [SiEtAl14]. The fractional neutron point kinetics equation with precursor contributions and temperature feedback effects is

$$\tau^\kappa\left(\frac{d}{dt}+\left(\frac{1}{\ell}-\frac{1-\beta}{\Lambda}\right)\right)\frac{d^\kappa n}{dt^\kappa}+\frac{dn}{dt}-\frac{\rho-\beta}{\Lambda}n-\sum_{i=1}^{p}\lambda_i\left(C_i+\tau^\kappa\frac{d^\kappa C_i}{dt^\kappa}\right)=0\,,$$

$$\frac{dC_i}{dt}-\frac{\beta_i}{\Lambda}n+\lambda_i C_i=0\,, \tag{46.1a}$$

$$\rho=\rho_0-\alpha(T-T(0)) \quad \text{and} \quad \frac{dT}{dt}-Hn=0\,. \tag{46.1b}$$

Here $\tau$ and $\kappa$ are the so-called fractional parameter, i.e. time scale and fractional derivative order, respectively. Further, $\beta=\sum\beta_i$ is the total delayed neutron fraction, whereas $\beta_i$ is the neutron fraction of the $i$-th precursor group, $\lambda_i$ is the decay constant of the $i$-th delayed neutron precursor, $\ell$ is the mean neutron lifetime, $\Lambda$ is the prompt neutron generation time, $\rho=\rho(T,t)$ is the reactivity, $\rho_0=\rho_0(t)$ is the apparent reactivity, $\alpha$ is the temperature coefficient, and $H$ is a parameter that mimics thermohydraulic effects on the core temperature $T=T(t)$.

It is noteworthy that the fractional point kinetics model with temperature feedback is a generalization of existing models. For $\tau=0$ and $\alpha=0$ one recovers the classical (linear) point kinetics model, whereas for $\tau=0$ and $\alpha\neq 0$ one reduces to the classical non-linear point kinetics model with temperature feedback as discussed in [SiEtAl14], while for $\tau\neq 0$ and $\alpha=0$ the model is identical to the fractional point kinetics model. In the further we study some implications of effects due to the fractional derivative as well as with temperature feedback on reactivity.

It is convenient to reduce the original problem (46.1) with four equations to a two equation system containing the neutron population $n$ (also sometimes identified with the power level) and the precursor group concentrations. To this end equation (46.1b) for the temperature feedback is integrated and inserted into the reactivity relation which is then substituted into the equation (46.1a) for the neutron density.

$$T=H\int_0^t n(\hat{t})\,d\hat{t}+T(0) \quad \Rightarrow \quad \rho=\rho_0-\alpha H\int_0^t n(\hat{t})\,d\hat{t}$$

The obtained two equation system reads then

$$\frac{dn}{dt}-\frac{\rho_0-\beta}{\Lambda}n-\sum_{i=1}^{p}\lambda_i C_i=\underbrace{-\frac{\alpha H}{\Lambda}n\int_0^t n(\hat{t})\,d\hat{t}}_{N}$$

$$\underbrace{-\tau^{\kappa}\left(\frac{d}{dt}-\left(\frac{1}{\ell}-\frac{1-\beta}{\Lambda}\right)\right)\frac{d^{\kappa}n}{dt^{\kappa}}+\tau^{\kappa}\sum_{i=1}^{p}\lambda_i\frac{d^{\kappa}C_i}{dt^{\kappa}}}_{F}\,, \tag{46.2a}$$

$$\frac{dC_i}{dt}-\frac{\beta_i}{\Lambda}n+\lambda_i C_i=0\,. \tag{46.2b}$$

In equation (46.2a) the shorthand notation $N$ stands for the non-linear contribution and $F$ for the fractional derivative contribution.

## 46.3 The Recursive Scheme

Upon substituting the series $n=\sum_r n_r$ and $C_i=\sum_r C_{ir}$ into equation (46.1a) and reshuffling terms, then equation (46.2) reads

$$\begin{aligned}\sum_{r=0}^{\infty}\left[\frac{dn_r}{dt}-\frac{\rho_0-\beta}{\Lambda}n_r-\sum_{i=1}^{p}\lambda_i C_{i_r}\right]=\sum_{r=0}^{\infty}\Bigg[&\underbrace{-\frac{\alpha H}{\Lambda}\left(\sum_{q=0}^{\infty}n_q\right)\int_0^t n_r(\hat{t})\,d\hat{t}}_{N_r}\\&\underbrace{-\tau^{\kappa}\frac{d^{\kappa+1}n_r}{dt^{\kappa+1}}-\tau^{\kappa}\left(\frac{1}{\ell}-\frac{1-\beta}{\Lambda}\right)\frac{d^{\kappa}n_r}{dt^{\kappa}}+\tau^{\kappa}\sum_{i=1}^{p}\lambda_i\frac{d^{\kappa}C_{ir}}{dt^{\kappa}}}_{F_r}\Bigg]\,,\\ \sum_{r=0}^{\infty}\left[\frac{dC_{ir}}{dt}-\frac{\beta_i}{\Lambda}n_r+\lambda_i C_{ir}\right]=0\,.\end{aligned} \tag{46.3}$$

The recursive scheme may now be set up, using the artificial degrees of freedom introduced by the expansion of $n$ and $C_i$, respectively, transforming the non-linear problem into a set of nonhomogeneous linear equations, where the first problem to be solved is the classical neutron point kinetics problem [Ga13, SiEtAl14, PaMaGo09], that shall obey the initial condition of the original problem. Thus all the remaining equations of the recursive scheme have zero initial conditions.

$$\begin{aligned}\frac{dn_r}{dt}-\frac{\rho_0-\beta}{\Lambda}n_r-\sum_{i=1}^{p}\lambda_i C_{ir}&=S_r\,,\\ \frac{dC_{ir}}{dt}-\frac{\beta_i}{\Lambda}n_r+\lambda_i C_{ir}&=0\,,\end{aligned} \tag{46.4}$$

for $r = 0, 1, \ldots$ and $\delta_{ij}$ the Kronecker symbol. Here the source term is defined by the findings of the previous recursions and thus are known.

$$S_r = (\delta_{r0} - 1)(N_{r-1} + F_{r-1})$$

For convenience one may cast (46.4) in matrix form

$$\frac{d\mathbf{Y}_r}{dt} - \mathbf{A}\mathbf{Y}_r = \mathbf{S}_r \,,$$

$$\mathbf{A} = \begin{pmatrix} \frac{\rho_0 - \beta}{\Lambda} & \lambda_1 & \lambda_2 & \cdots & \lambda_p \\ \frac{\beta_1}{\Lambda} & -\lambda_1 & 0 & \cdots & 0 \\ \frac{\beta_2}{\Lambda} & 0 & -\lambda_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\beta_p}{\Lambda} & 0 & 0 & \cdots & -\lambda_p \end{pmatrix}, \mathbf{S}_r = \begin{pmatrix} S_r \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \mathbf{Y}_r = \begin{pmatrix} n_r \\ C_{1r} \\ C_{2r} \\ \vdots \\ C_{pr} \end{pmatrix},$$

with known solution

$$\mathbf{Y}_r = \exp\left(\int_0^t \mathbf{A}\,(\hat{t})\; d\hat{t}\right)\mathbf{Y}_r(0) + \int_0^t \exp\left(\int_0^{\hat{t}} \mathbf{A}\,(\bar{t})\; d\bar{t}\right)\mathbf{S}_r\,(t-\hat{t})\; d\hat{t}\,.$$

and initial condition

$$\mathbf{Y}_r(0) = \delta_{r0}\mathbf{Y}(0)$$

## 46.4 Numerical Implementation

In order to obtain numerical results from the afore derived solution, we introduce some simplifications in the reactivity and in the source terms, respectively. These simplifications involve constant approximations that still provide a solution that remains within an acceptable error range but avoids excessive computing time. To this end we discretize the time interval for the matrix $\mathbf{A}$ and the source term $\mathbf{S}_r$, so that the time evolution is constructed using the results from the previous time interval as initial condition. Thus we end up with constant matrices $\bar{\mathbf{A}}$ and $\bar{\mathbf{S}}_r$. Note that $\bar{\mathbf{A}}$ has distinct eigenvalues, since the operator of the differential equation is self-adjoint, and one can factorize $\bar{\mathbf{A}}\mathbf{V}\mathbf{D}\mathbf{V}^{-1}$ and $\exp(\bar{\mathbf{A}}t) = \mathbf{V}\exp(\mathbf{D}t)\,\mathbf{V}^{-1}$, with $\mathbf{V}$ the eigenvector matrix of $\bar{\mathbf{A}}$ and $\mathbf{D}$ the diagonal matrix with associated eigenvalues. Thus, the final form of $\mathbf{Y}_r$ is

$$\mathbf{Y}_r = \mathbf{V}\exp(\mathbf{D}t)\,\mathbf{V}^{-1}\mathbf{Y}_r(0) + \mathbf{V}\mathbf{D}^{-1}\left(\exp(\mathbf{D}t) - \mathbf{I}\right)\mathbf{V}^{-1}\bar{\mathbf{S}}_r. \tag{46.5}$$

By virtue, the only time dependent part in matrix **A** is the apparent reactivity $\rho_0$ that depends on the history of the neutron population from $t = 0$ up to the actual time (see equation (46.2)). Since we are using an analytical continuation in our numerical procedure, the reactivity has to be computed in two steps, first considering the solution up to the last time step and afterwards correcting the integral over last time interval using the solution for $n$ in this time interval. More specifically, the integrals are solved using the trapezoidal rule for the reactivity as well as the nonlinearity and fractional derivative, still to be discussed.

$$\bar{\rho} = \frac{1}{2}\left(\rho(t-\Delta t)+\rho(t)\right)$$

$$\bar{S}_r = -\frac{1}{2}\left(N_{r-1}(t-\Delta t)+N_{r-1}(t)+F_{r-1}(t-\Delta t)+F_{r-1}(t)\right)$$

To calculate the fractional derivative $F_r$ one faces one dilemma, namely the fact that there does not exist a unique definition for this kind of derivative. Definitions that may be found in the literature are the Riemann–Liouville, the Caputo and Grünwald-Letnikov's definitions [De03, OlSp74, Tr03]. By inspection one verifies that the Riemann–Liouville definition of fractional derivatives has a consistent form with respect to classical limits, and moreover it preserves homogeneity of homogeneous functions, where the exponential function is one striking example. Also integral transforms maintain their classically established properties, here classically means properties defined for the usual derivatives. The Riemann–Liouville definition of the fractional derivative operator acting on a function $f = f(x)$ is defined by [OlSp74]

$$\frac{d^\nu f}{dx^\nu} = \frac{1}{\Gamma(m-\nu)}\frac{d^m}{dx^m}\int_{-\infty}^{x}\frac{f(\xi)}{(x-\xi)^{1+\nu-m}}\,d\xi$$

for $\nu \in \mathbb{R}$, $m \in \mathbb{Z}$, and $\nu < m < \nu + 1$. Here, $\nu$ is the fractional derivative order, $\Gamma(x)$ is the gamma function, $m$ is an integer. One of the principal properties of this definition is that it keeps the transcendental character of a function after application of the fractional derivative. For instance, application of this operator on the exponential functions yields

$$\frac{d^\nu}{dx^\nu}\exp(ax) = a^\nu \exp(ax),$$

$$\frac{d^\nu}{dx^\nu}K = \frac{d^\nu}{dx^\nu}K\exp(0x) = 0,$$

where $a$ and $K$ are real constants. Note that the exponential function and its relation to the fractional derivative properties is essential because of the construction of the

solution terms $\mathbf{Y}_r$ containing the exponential function. Using this definition, the numerical values of the terms for $F_r$ may be uniquely computed.

## 46.5 Results

In the sequel, five cases of reactivity input are evaluated, for a constant positive and negative reactivity, for a linearly increasing reactivity with time and a sinusoidal case with the temperature feedback switched off ($\alpha \equiv 0$) and last not least, reactivity driven by the temperature feedback. For all cases the same set of nuclear parameter was used and is shown in table 46.1.

The initial condition valid for all cases may be calculated from the parameter in table 46.1 and is

$$\mathbf{Y}(0) = \left(1 \ \frac{\beta_1}{\Lambda\lambda_1} \ \frac{\beta_2}{\Lambda\lambda_2} \ \cdots \ \frac{\beta_p}{\Lambda\lambda_p}\right)^T. \tag{46.6}$$

For the first four cases, combinations with fractional derivatives (FD) $\kappa = 0.9$, 1.0, 1.1 and $\tau = 0, 10^{-6}, 10^{-4}\ s$ are shown in figures 46.1–46.4, where $P/[MW]$ signifies the power level that is proportional to the neutron density $n$. For constant positive reactivity ($\rho = 0.003$, time step $\Delta t = 0.1\ s$), only for small times visible differences occur in the case with the largest $\tau = 10^{-4}\ s$, where for negative constant reactivity ($\rho = -0.003$, time step $\Delta t = 0.1\ s$, and the sinusoidal reactivity change ($\rho = 0.003\sin(\pi t)$, time step $\Delta t = 0.01\ s$) no significant differences appear in comparison with the classical point kinetics solution. In the linear reactivity case ($\rho = 0.003t$, time step $\Delta t = 0.01\ s$), only for the largest $\tau$ and for large times ($t \sim 2\ s$) a significant difference between the different solutions are found.

In the case considering temperature feedback with $\alpha = 5 \times 10^{-5}\ K^{-1}$, $H = 0.05\ Ks/MW$, $\rho_0 = 0.003$ and time step $\Delta t = 0.001\ s$ any contribution due to the fractional derivative was completely suppressed. Figures 46.5–46.7 show the power level, the reactivity evolution with time and the temperature for the classical model ($\tau = 0$), and the fractional model with $\tau = 10^{-4}\ s$ and $\kappa = 0.9$. Unexpectedly, both curves practically coincide for the three quantities, power level, reactivity and temperature over the whole time interval up to 10 $s$.

**Table 46.1** Nuclear parameter set.

| $\Lambda = 10^{-5}\ s$ | | $\ell = 0.00024\ s$ | | | $\beta = 0.007$ | |
|---|---|---|---|---|---|---|
| $\beta_i$ | 0.000266 | 0.001491 | 0.001316 | 0.002849 | 0.000896 | 0.000182 |
| $\lambda_i[s^{-1}]$ | 0.0127 | 0.0317 | 0.115 | 0.311 | 1.4 | 3.87 |

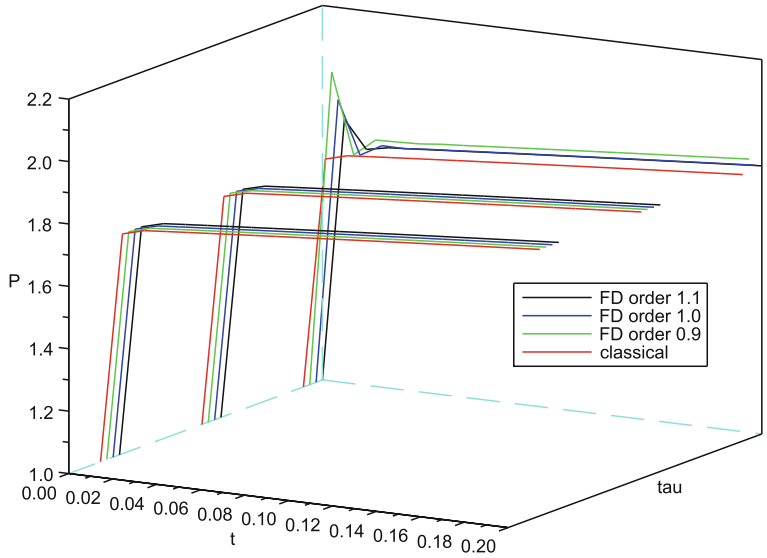


**Fig. 46.1** Case 1: Power level for $\rho = 0.003$.

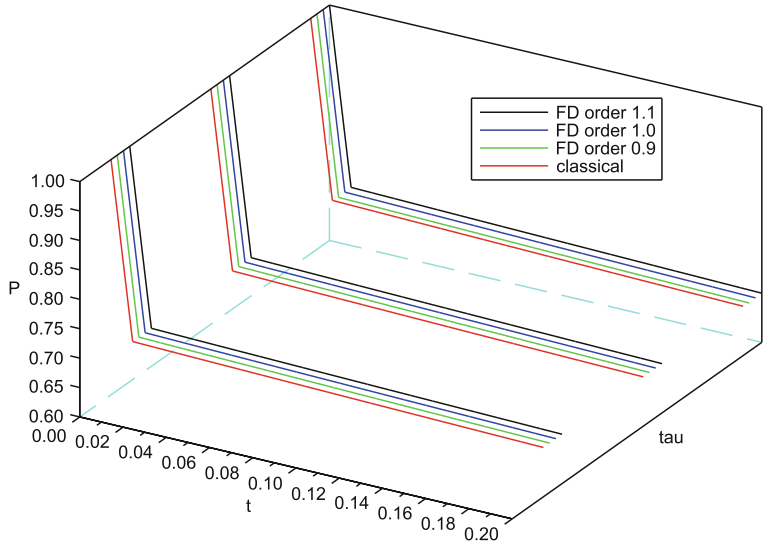


**Fig. 46.2** Case 2: Power level for $\rho = -0.003$.

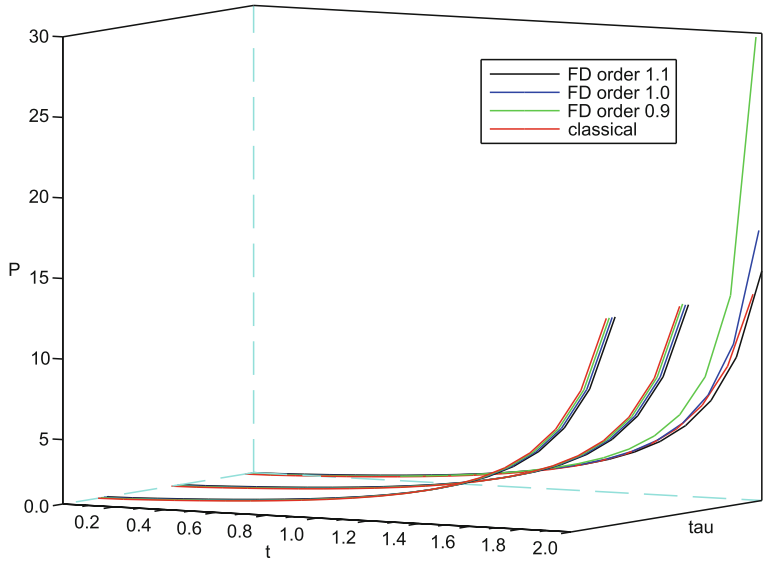


**Fig. 46.3** Case 3: Power level for $\rho = 0.003t$.

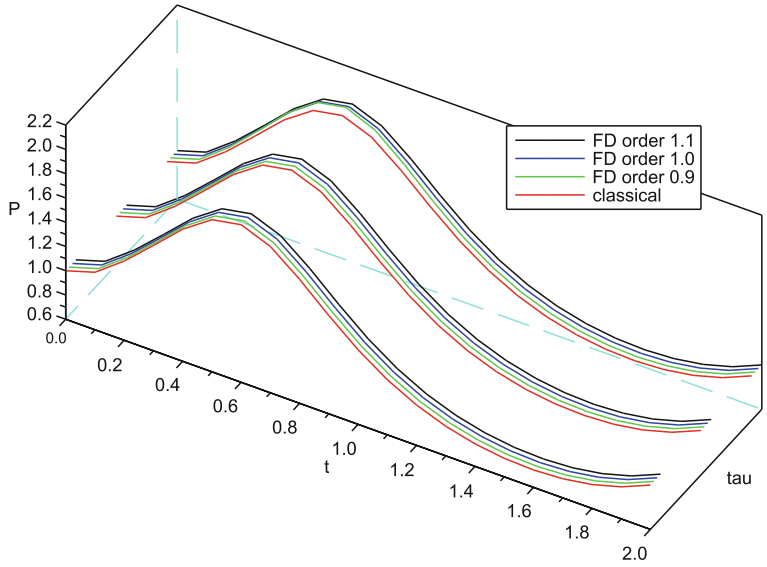


**Fig. 46.4** Case 4: Power level for $\rho = 0.003\sin(\pi t)$.

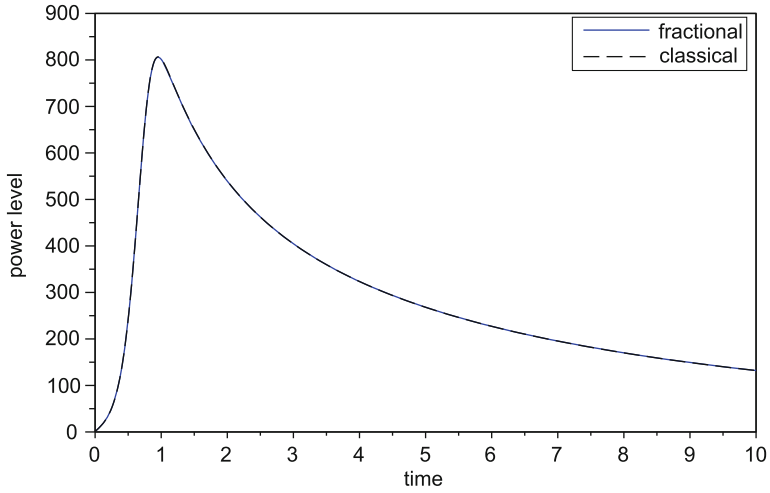


**Fig. 46.5** Case 5: Power level evolution with time, considering temperature feedback.

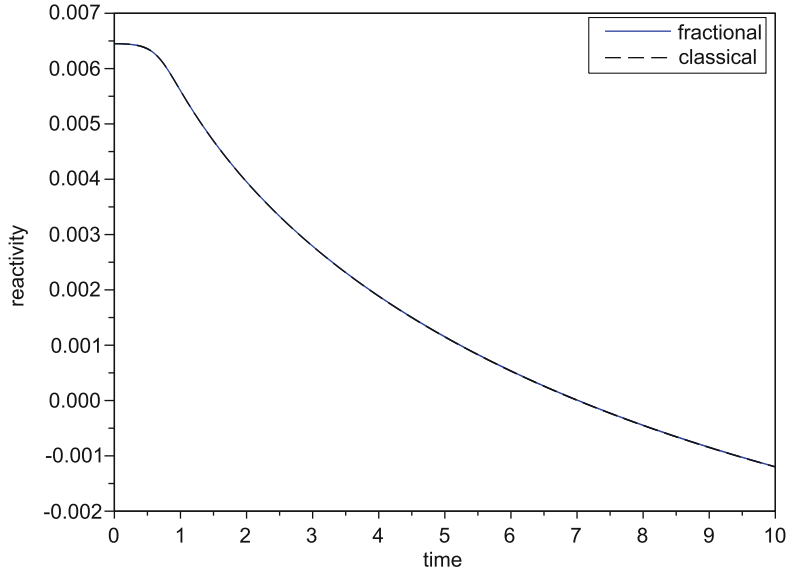


**Fig. 46.6** Case 5: Reactivity evolution with time, considering temperature feedback.

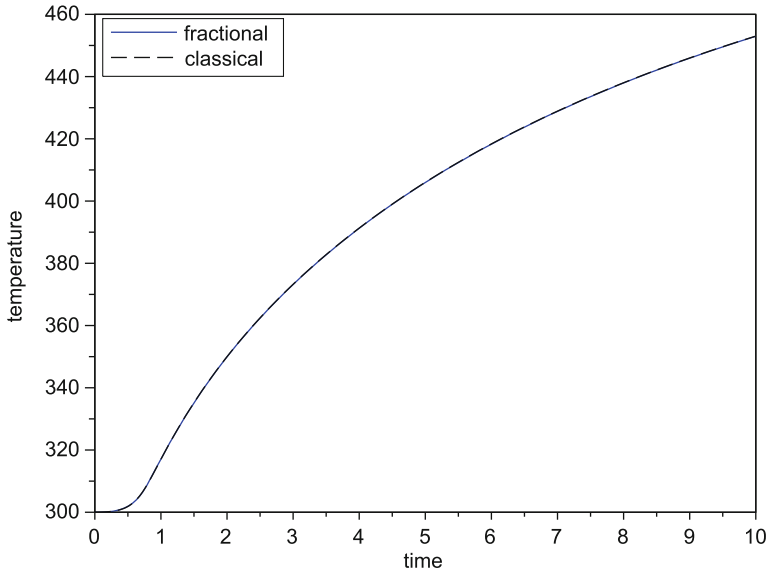


**Fig. 46.7** Case 5: Temperature evolution with time, considering temperature feedback.

## 46.6 Conclusions

As shown in the results, within the range of the parameters considered the influence of the fractional derivative is practically completely suppressed as an effect of the temperature feedback. An extensive comparison of results with and without feedback, that are not shown in this contribution clearly identified the temperature influence on reactivity as the cause for suppression.

This statement can be confirmed due to the fact that we determined the solution in a closed form, without using simplifications or approximations along the derivation of the solution. Moreover, a genuine analysis of convergence not presented here showed us that the obtained solution is close to the true solution. Formally, the solutions are determined by a recursive scheme, where recursion initialization satisfies the initial conditions, all remaining initial conditions are null. The recursion scheme is truncated according to a prescribed precision which may be evaluated by a convergence criterion, that indicates if more recursions are in order.

A further noteworthy aspect is that there does not exist a unique definition for the fractional derivative. However, if such a derivative is considered an extension to the traditional derivative, certain properties of functions shall be preserved, one of them is the homogeneity. By direct inspection one notices that the only consistent definition compatible with this criterion is the Riemann–Liouville definition, used in this work. Further, integral transform properties and its inversion were found to be free of contradictions only for the latter definition.

Concluding, we have shown that the extension of the classical model with a fractional derivative accompanied by two additional parameters is not only more cumbersome to be solved but also does not contribute with significant improvements in comparison with the classical point kinetic results. Although, works known from the literature have shown some changes that occur, the present model considering additional temperature feedback revealed that none of those effects survived.

## References


[Ad94] Adomian, G.: Solving frontier problems of physics: the decomposition method. Kluwer Academic Publishers, Aarhus C., Denmark (1994)

[De03] Debnath, L.: Recent applications of fractional calculus to science and engineering International Journal of Mathematics and Mathematical Sciences **2003**, 3413–3442 (2003)

[ElMa14] El Tokhy, M., Mahmoud, I.I.: Parameter analysis of the neutron point kinetics equations with feedback temperature effects. Annals of Nuclear Energy **68**, 228–233 (2014)

[EsEtAl08] Espinosa-Paredes, G., Morales-Sandoval, J., Vásquez-Rodríguez, R. and Espinosa-Martínez, E.-G.: Constitutive laws for the neutron density current. Annals of Nuclear Energy **35**, 1963–1967 (2008)

[EsEtAl11] Espinosa-Paredes, G., Polo-Labarrios, M., Espinosa-Martínez, E.-G., and Valle-Galegos, E.: Fractional neutron point kinetics equations for nuclear reactor dynamics, Annals of Nuclear Energy **38**, 207–330 (2011)

[Ga13] Ganapol, B.D.: A highly accurate algorithm for the solution of the point kinetics equations. Annals of Nuclear Energy **62**, 564–571 (2013)

[Na11] Nahla, A.A.: Taylor's series method for solving the nonlinear point kinetics equations. Nuclear Engineering and Design **241**, 1502–1505 (2011)

[NeNe07] Nec, Y., Nepomnyashchy, A.A.: Turing instability in sub-diffusive reaction-diffusion system. Journal of Physics A: Mathematical and Theoretical **40(49)**, 14687–14702 (2007)

[OlSp74] Oldham, K., Spanier, J.: The Fractional Calculus. Academic Press, New York, USA (1974)

[PaMaGo09] Palma, D.A.P., Martinez, A.S., GonÃğalves, A.: Analytical solution of point kinetics equations for linear reactivity variation during the start-up of a nuclear reactor. Annals of Nuclear Energy **36**, 1469–1471 (2009)

[SaPa12] Saha Ray, S., Patra, A.: An explicit finite difference scheme for numerical solution of fractional neutron point kinetic equation. Annals of Nuclear Energy **41**, 61–66 (2012)

[SiEtAl14] Silva, J.J.A., Alvim, A.C.M., Vilhena, M.T.M.B., Bodmann, B.E.J., Petersen, C.Z.: On a closed-form solution of the point kinetics equations with reactivity feedback of temperature. International Journal of Nuclear Energy Science and Technology **8(2)**, 131–140 (2014)

[SiEtAl14] da Silva, M.W., Leite, S.B., Vilhena, M.T.M.B., Bodmann, B.E.J.: On an analytical representation for the solution of the neutron point kinetics equation free of stiffness. Annals of Nuclear Energy **71**, 97–102 (2014)

[Tr03] Trenčevski, K.: New Approach to The Fractional Derivatives. International Journal of Mathematics and Mathematical Sciences **2003(5)**, 315–325 (2003)

# Chapter 47
# Comparison of Analytical and Numerical Solution Methods for the Point Kinetics Equation with Temperature Feedback Free of Stiffness

**J.J.A. Silva, A.C.M. Alvim, B.E.J. Bodmann, and M.T.B. Vilhena**

## 47.1 Introduction

The nuclear point kinetics equations with temperature feedback are a stiff system of nonlinear differential equations that determine the neutron density and delayed precursor concentrations. From the time evolution of these quantities one may simulate the power density of a nuclear reactor. Computing solutions of the point kinetics equations provide information on the dynamics of nuclear reactor operation and are useful for an understanding of power variations when the control rods are adjusted (see, for instance, references [Ab09, AbHa03, AbHa02, AbNa02, ChEtAl07, KiAl04, NaZa10, Na10, PeNiRa06, PeEtAl11, SaPa13, TaJaHa10, SiEtAl14]. As pointed out by many authors, the system of point kinetics equations continues to be the crucial set of equations. Although its range of applicability has been severely restricted by the increasing importance of optimal power loosely coupled reactor cores, they still remain useful in terms of preliminary studies, especially when control aspects are of concern. The presence of temperature feedback is useful to provide an estimate of the transient behavior of reactor power and other system variables of tightly coupled reactor cores.

In this chapter, first the usual neutron point kinetics equations are solved using a diagonalization decomposition method that is free of stiffness applying a semi-analytical and two numerical approaches. Since, it is the temperature feedback

J.J.A. Silva (✉) • A.C.M. Alvim
Federal University of Rio de Janeiro, Av. Horácio Macedo 2030, Rio de Janeiro 21941-914, RJ, Brazil
e-mail: shaolin.jr@gmail.com; alvim@nuclear.ufrj.br

B.E.J. Bodmann • M.T.B. Vilhena
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: bardo.bodmann@ufrgs.br; vilhena@pq.cnpq.br



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_47

which drives reactivity, that is not considered in the first part of the study, we prescribe by hand a linear reactivity evolution with time. Second, the point reactor kinetics equation are considered in the presence of a Newtonian temperature feedback, which are reduced to a second order nonlinear differential equation, where an analytical as well as a numerical decomposition procedure is applied to determine the solution. To this end the solution is expanded in a series of functions and the nonlinear term is handled using Adomian's decomposition method. On the one hand, the recursion system is implemented using analytical corrections while, on the other hand, the explicit and implicit Euler method is used to determine the time evolution in a specific interval. Upon substituting the expansions into the original equations, a recursive linear system is built from the originally nonlinear problem, which is then solved. In the sequel, we present solutions in numerical as well as semi-analytical form for the point kinetics equations with and without temperature feedback in the presence of one delayed neutron precursor group. The initial condition is given by some steady state power level and a Newtonian feedback model is being assumed for the fuel temperature equations.

## 47.2 Neutron Point Kinetics Equations with Temperature Feedback

Our starting point is the point kinetics equations and one group precursors as reported in ref. [AbNa11, NaZa10, SiEtAl14], where $n(t)$ is the time-dependent neutron population, $C(t)$ is the concentration of delayed neutrons precursors, $T(t)$ is the time dependent temperature of the nuclear core, $\rho(T)$ is the temperature-dependent reactivity, $\beta$ is the delayed neutron fraction, $L$ is the prompt neutron generation time, and $\lambda$ is the average decay constant of the precursors.

$$\frac{dn(t)}{dt} = \left(\frac{\rho(T)-\beta}{L}\right) n(t) + \lambda C(t) \tag{47.1}$$

$$\frac{dC(t)}{dt} = \frac{\beta}{L} n(t) - \lambda C(t) \tag{47.2}$$

Equations (47.2) and (47.1) constitute the usual point kinetics equation system, which is extended by a perturbation in form of a temperature feedback, where perturbation signifies a change of the nuclear system's configuration. This change leads to an altered specific heat flow manifest in a change in the core temperature. As the source of heat production are fission and decay processes, the thermal change rate shall be proportional to the neutron density, which we assume to be linear and the proportionality constant $H$ signifies a parameter for the influence of the change of heat flow on the rate of temperature change.

$$\frac{dT(t)}{dt} = Hn(t) \tag{47.3}$$

Note that the linear relation between temperature change rate and neutron density represents a feedback mechanism. The necessary initial conditions are defined considering the reactor at equilibrium ( $\frac{dn}{dt}\big|_{t=0} = 0$), with known initial power density and temperature ($n(0)$, $T(0)$) and initial concentration of delayed neutrons precursors

$$C(0) = \frac{1}{\lambda}\left(\frac{\beta - \rho}{L}\right) n(0)\ . \tag{47.4}$$

The variation of reactivity with temperature is given by

$$\rho(T) = \rho(0) - \alpha\,(T - T(0))\ , \tag{47.5}$$

where $\rho(0)$ is the initial reactivity and $\alpha$ is the fuel temperature reactivity coefficient.

## 47.3 The Conventional Neutron Point Kinetics Equation

In the further we will analyze and compare the analytical and the numerical solutions for the usual point kinetics problem using a recursive diagonalization decomposition procedure as shown in [WoEtAl14]. In both cases the recursion steps are evaluated by an analytical procedure as well as an explicit and implicit Euler method. To this end we consider the equation system (47.1) and (47.2) and consider $H \equiv 0$ in equation (47.3), which from the physical point of view implies $\alpha \equiv 0$ in (47.5). Such a simplification would leave the reactivity constant or in turn requires an explicitly prescribed time dependence for reactivity based on observation or simple model cases.

The discussion that follows in this section shows a method applicable to analytical, semi-analytical up to numerical approaches circumventing stiffness that may be present when present time scales differ by various orders in magnitude. Since stiffness cannot be decoupled from the context of available computational resources and their arithmetic precision, the present line out is to show the formal procedure but without challenging stiffness limits with today's computing technology. The consideration of an increased number of neutron precursor concentrations ($\sim$6) would represent an adequate problem to that question. For simplicity we consider only one precursor group but without changing the structure of the solution steps. The neutron density, precursor concentration equation system (47.1) and (47.2) may be cast in matrix representation as

$$\frac{d}{dt}\begin{pmatrix} n \\ C \end{pmatrix} = \begin{pmatrix} \frac{\rho(t)-\beta}{L} & \lambda \\ \frac{\beta}{L} & -\lambda \end{pmatrix}\begin{pmatrix} n \\ C \end{pmatrix} \tag{47.6}$$

Here, for simplicity we omitted the explicit time dependency of $n$ and $C$. The time dependence of the reactivity may have origin in various sources, as insertion or removal of control rods and actions that involve aspects of thermohydraulics, each one with its specific formulation for reactivity. If moderate changes are applied, one may approximate the reactivity evolution by a linear function with time $\rho(t) = at$ starting from initial reactivity equal zero, where $a$ is a coefficient. If one considers a general reactivity function with time, it is convenient to separate the matrix on the right-hand side of equation (47.6) into a diagonal part and one with the residual contributions.

$$\frac{d}{dt}\begin{pmatrix} n \\ C \end{pmatrix} = \left(\begin{pmatrix} \frac{\rho(t)-\beta}{L} & 0 \\ 0 & -\lambda \end{pmatrix} + \begin{pmatrix} 0 & \lambda \\ \frac{\beta}{L} & 0 \end{pmatrix}\right)\begin{pmatrix} n \\ C \end{pmatrix}$$

In shorthand this reads $\frac{d}{dt}X = (D + W)X$, where $X$ is the vector containing the unknown variables, $D$ is the diagonal part of the coefficient matrix, and $W$ consists in the remaining terms. A recursive scheme may now defined upon introducing artificial degrees of freedom by expanding $n$ and $C$ or equivalently $X$ in a series. For numerical purposes these are truncated after $R$ terms.

$$n = \sum_{j=0}^{R-1} n_j \qquad C = \sum_{j=0}^{R-1} C_j$$

These new degrees of freedoms, without accompanying constitutive equations, may be used to define a recursive scheme, i.e. we can split the original system in a number of simpler systems, choosing the diagonal matrix and the initial conditions as recursion initialization, while the remaining equations of the recursion scheme have zero initial conditions:

$$\frac{d}{dt}\begin{pmatrix} n_0 \\ C_0 \end{pmatrix} = \begin{pmatrix} \frac{\rho(t)-\beta}{L} & 0 \\ 0 & -\lambda \end{pmatrix}\begin{pmatrix} n(0) \\ C(0) \end{pmatrix}$$

$$\frac{d}{dt}\begin{pmatrix} n_{j>0} \\ C_{j>0} \end{pmatrix} = \begin{pmatrix} 0 & \lambda \\ \frac{\beta}{L} & 0 \end{pmatrix}\begin{pmatrix} n_{j-1} \\ C_{j-1} \end{pmatrix}$$

This way, the first system is solved directly due to the absence of non-diagonal elements, and incorporates all the initial conditions. The other systems are solved recursively, using the results of the previously determined terms.

## 47.4 Results

The algorithms for the solution were implemented in program codes using the traditional explicit and implicit Euler method, followed by the diagonalization decomposition method (DDM) using an explicit and an implicit version. The numerical results were compared using the following parameter set.

$$\beta = 0.0065 \qquad L = 1 \times 10^{-5}\ s \qquad \lambda = 0.008\ s^{-1} \qquad n(0) = 1\ MW$$

Assuming that the system is initially in a given steady state, we can determine the initial concentration of precursor concentrations according to (47.4) rearranging (47.1) and considering $\frac{d}{dt}n(t)\big|_{t=0} = 0$. Two cases of linear reactivity coefficients were tested ($a = 0.25$ and $a = 0.50$), each with two different values for time steps ($\Delta t = 10^{-5}\ s$ and $\Delta t = 10^{-6}\ s$), for all the above methods. In the cases, where the decomposition was used, we truncated the expansion after the eleventh term, which contributed at most to a change in the twenty-fourth digit.

The results obtained are shown in tables 47.1 (for a linear reactivity coefficient $a = 0.25$) and 47.2 (for $a = 0.5$). The first two columns show the findings from the explicit methods without and with diagonalization decomposition, whereas the third and fourth columns of tables show the values obtained by implicit methods, without and with the use of the decomposition. Table 47.3 gives an estimate on the order of magnitude of the contribution of each of the decomposition terms in the calculated power output series. Note that the orders of magnitudes of each contribution were the same for all calculations, which was to be expected for a linear system, without real stiffness.

**Table 47.1** Power ($MW$) for $a = 0.25$.

| $\Delta t = 10^{-6}\ s$ | Time ($s$) | Power ($MW$) | | | |
|---|---|---|---|---|---|
| | | Explicit | DDM(E) | Implicit | DDM(I) |
| | 0.0025 | 1.050185 | 1.050185 | 1.050166 | 1.050166 |
| | 0.0050 | 1.150238 | 1.150238 | 1.150202 | 1.150202 |
| | 0.0075 | 1.280032 | 1.280032 | 1.279984 | 1.279984 |
| | 0.0100 | 1.442750 | 1.442750 | 1.442690 | 1.442690 |
| $\Delta t = 10^{-5}\ s$ | Time ($s$) | Power ($MW$) | | | |
| | | Explicit | DDM(E) | Implicit | DDM (I) |
| | 0.0025 | 1.050120 | 1.050120 | 1.049934 | 1.049934 |
| | 0.0050 | 1.150191 | 1.150191 | 1.149834 | 1.149834 |
| | 0.0075 | 1.279987 | 1.279987 | 1.279507 | 1.279507 |
| | 0.0100 | 1.442691 | 1.442691 | 1.442092 | 1.442092 |

**Table 47.2** Power ($MW$) for $a = 0.50$.

| $\Delta t = 10^{-6}\ s$ | Time ($s$) | Power ($MW$) | | | |
|---|---|---|---|---|---|
| | | Explicit | DDM(E) | Implicit | DDM(I) |
| | 0.0025 | 1.103630 | 1.103630 | 1.103593 | 1.103593 |
| | 0.0050 | 1.335361 | 1.335361 | 1.335284 | 1.335284 |
| | 0.0075 | 1.700609 | 1.700609 | 1.700491 | 1.700491 |
| | 0.0100 | 2.293542 | 2.293542 | 2.293375 | 2.293375 |
| $\Delta t = 10^{-5}\ s$ | Time ($s$) | Power ($MW$) | | | |
| | | Explicit | DDM(E) | Implicit | DDM (I) |
| | 0.0025 | 1.103479 | 1.103479 | 1.103109 | 1.103109 |
| | 0.0050 | 1.335188 | 1.335188 | 1.334415 | 1.334415 |
| | 0.0075 | 1.700319 | 1.700319 | 1.699140 | 1.699140 |
| | 0.0100 | 2.292916 | 2.292916 | 2.291249 | 2.291249 |

**Table 47.3** Order of magnitude of each decomposition term.

| DDM Terms | $1^{st}$ | $2^{nd}$ | $3^{rd}$ | $4^{th}$ | $5^{th}$ | $6^{th}$ | $7^{th}$ | $8^{th}$ | $9^{th}$ | $10^{th}$ | $11^{th}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Contribution $10^{x}$) | 0 | 0 | $-5$ | $-4$ | $-9$ | $-9$ | $-14$ | $-14$ | $-19$ | $-19$ | $-24$ |

## 47.5 The Neutron Point Kinetics Equation with Temperature Feedback

In this section we analyze the nonlinear system due to reactivity driven by a temperature feedback (see [SiEtAl14]). The solution is obtained using the Adomian decomposition method, where the time dependence of each recursion step was determined by step-wise constant source terms. The resulting simplified equations were solved using a selection of methods, the traditional Adomian method, the numerical implementation by the authors of reference [NaZa10], an analytical inversion, a Crank–Nicholson implementation and an explicit time step method.

The nonlinear second-order differential equation to be solved is obtained upon substituting (47.5), (47.2) and equation (47.3) in the differentiated equation (47.1). The nonlinearity may be hidden as a source term of known functions in a recursive set of linear differential equations for the neutron population:

$$\underbrace{\frac{d^2n}{dt^2} + \left(\lambda + \frac{\beta - \rho(0) + \alpha T(0)}{L}\right)\frac{dn}{dt} + \lambda\frac{\alpha T(0) - \rho(0)}{L}n}_{linear} = \underbrace{\frac{\alpha T}{L}\left(\frac{dn}{dt} + \lambda n\right) + \frac{\alpha H}{L}n^2}_{non-linear} \tag{47.7}$$

For convenience we introduce the abbreviations for the constant expressions $a = \lambda + (\beta - \rho(0) + \alpha T(0))/L$ and $b = \lambda(\alpha T(0) - \rho(0))/L$, then, the system to be solved is

$$\frac{d^2n}{dt^2} + a\frac{dn}{dt} + bn = S\,, \tag{47.8}$$

where $S$ is the nonlinearity, i.e. the right-hand side of equation (47.7) and the temperature evolution may be determined by equation (47.3).

### 47.5.1 The Decomposition Method

The initial conditions for $\frac{d}{dt}n(t)|_{t=0} = 0$ and $C(0)$ are the same as for the conventional neutron point kinetics problem of the previous section, namely the system is initially in a steady state and $C(0)$ is given by equation (47.4). As already introduced before, the nonlinearity stems from the Newtonian cooling model, which was also adopted in the works of references [NaZa10] and [SiEtAl14]. Note that the nonlinearity has two contributions an explicit (quadratic) one and the term $T\left(\frac{dn}{dt} - \lambda n\right)$ that by virtue of $T = T(n)$ is in general of unknown nonlinear type. The quadratic nonlinearity is handled using standard Adomian functional polynomials, the implicit nonlinear contribution is treated as shown below. Therefore, the nonlinearity is rewritten as

$$S = -\frac{\alpha}{L}(P + A) \qquad \text{where} \quad P = T\left(\frac{dn}{dt} - \lambda n\right) \quad \text{and} \quad A = Hn^2\,.$$

As already put into practice before, the neutron density is expanded in a series $n = \sum_{j=0}^{R-1} n_j$ together with the temperature $T = \sum_{j=0}^{R-1} T_j$, which again for numerical reasons is truncated at $R$ according to a desired precision to be tested by a posterior analysis (not shown in this contribution but see reference [SiEtAl14]). Again the artificial degrees of freedoms introduced are now used to define a recursive scheme in a fashion, that the source terms are always determined from combinations of the solutions obtained by all the previous recursion steps. Choosing $P_0 = 0$ and $A_0 = 0$ results in a homogeneous equation for $j = 0$,

$$\frac{d^2n_0(t)}{dt^2} + b\frac{dn_0(t)}{dt} + cn_0(t) = 0\,,$$

for which the solution is well known [SiEtAl14]. With this result, one can evaluate the temperature by integrating equation (47.3).

$$T_0(t) = T(0) + H\int_0^t n(\tau)\,d\tau$$

With known $n_0(t)$ and $T_0(t)$, one can evaluate the terms for $j = 1$.

$$\frac{d^2 n_1(t)}{dt} + b\frac{dn_1(t)}{dt} + cn_1(t) = \underbrace{-\frac{\alpha}{L}(P_1 + A_1)}_{S_1}$$

As the right-hand side of equation depends on $n_0$ and $T_0$, we may introduce an approximation and set a fixed time step, $\omega$, and consider $P_1(\omega)$ and $A_1(\omega)$ as 'pseudo-constants.' For an adequately chosen $\omega$ the error in the integrals involving the $A_j$ and $P_j$ will be sufficiently small so that the solution is still within an acceptable precision (to be verified in a posterior convergence analysis). This way, all the equations with $j > 0$ are non-homogeneous equations, where the right-hand side is constant and thus easy to solve.

### 47.5.2 Expansion of the $P_j$ and $A_j$

These terms carry the products of the variables, $T\left(\frac{dn}{dt} - \lambda n\right)$. In their construction it is convenient that each of the $P_j$ depends only on $n_p$ and $T_p$ for which $p < j$ holds. The following construction is one possibility.

$$P_1 = T_0\left(\frac{dn_0}{dt} - \lambda n_0\right)$$

$$P_{j>1} = \sum_{p=0}^{j} T_j\left(\frac{dn_p}{dt} - \lambda n_p\right) + \sum_{p=0}^{j} T_p\left(\frac{dn_j}{dt} - \lambda n_j\right)$$

The expansion of the nonlinear terms in Adomian polynomials is unique, although there are numerous ways to group together terms of such an expansion [Ad94]. The scheme of a 'fast conversion' expansion (accelerated polynomials) for $n^2$ are

$$A_1 = H(n_0^2)\,,$$

$$A_{j>1} = H\left(n_j^2 + 2n_j \sum_{p=0}^{j-1} n_p\right)\,,$$

that we adopt in our present implementation.

## 47.6 Results

All the solutions were implemented in a program code, except the solutions from [NaZa10], which were extracted from the aforementioned reference. The initial conditions and nuclear parameters were identical for all the cases, except for the initial reactivity, which assumed values of $\rho(0) = 0.2\beta$, $\rho(0) = 0.5\beta$, and $\rho(0) = 0.8\beta$.

$$\beta = 0.0065 \qquad L = 1 \times 10^{-5}\ s \qquad \lambda = 0.07741\ s^{-1} \qquad H = 0.5\frac{K}{MW\ s}$$

$$\alpha = 5 \times 10^{-5}\ K^{-1} \qquad n(0) = 10\ MW \qquad \left.\frac{dn}{dt}\right|_{t=0} = 0 \qquad T(0) = 300\ K$$

The initial delayed neutrons precursors concentration is given by equation 47.4. The numerical parameters used for all the cases were an integration time step $\omega = 0.01\ s$ and a simulated time interval $t_{max} = 100\ s$. The results obtained for the power and temperature in three different cases are shown in tables 47.4 to 47.9.

## 47.7 Conclusions

In this chapter, a comparison between (semi-)analytical and numerical methods for the traditional neutron point kinetic problem was elaborated, with the goal to circumvent stiffness problems in cases where the dynamics has characteristic time scales that extend over several orders in magnitude. The method that is useful for analytical as well as numerical schemes made use of a diagonalization decomposition of the matrix differential equation system that together with an

**Table 47.4** Power ($MW$) for initial reactivity $\rho(0) = 0.2\beta$.

| Time (s) | Adomian | Nahla- | Analytical | Crank- | Time |
|---|---|---|---|---|---|
| | | Zayed | Inversion | Nicholson | Explicit |
| 0 | 10.00000 | 10.00000 | 10.00000 | 10.00000 | 10.00000 |
| 10 | 11.27005 | 11.27252 | 11.27057 | 11.27057 | 11.27057 |
| 20 | 12.09304 | 12.09509 | 12.09411 | 12.09411 | 12.09411 |
| 30 | 12.40054 | 12.40206 | 12.40198 | 12.40198 | 12.40198 |
| 40 | 12.21855 | 12.21948 | 12.22011 | 12.22011 | 12.22011 |
| 50 | 11.6379 | 11.63786 | 11.63898 | 11.63898 | 11.63898 |
| 60 | 10.77603 | 10.77559 | 10.77700 | 10.77700 | 10.77700 |
| 70 | 9.74997 | 9.74926 | 9.75078 | 9.75078 | 9.75078 |
| 80 | 8.65758 | 8.65667 | 8.65820 | 8.65820 | 8.65820 |
| 90 | 7.57179 | 7.57075 | 7.57223 | 7.57223 | 7.57223 |
| 100 | 6.54159 | – | 6.54185 | 6.54185 | 6.54185 |

**Table 47.5** Temperature ($K$) for initial reactivity $\rho(0) = 0.2\beta$.

| Time (s) | Adomian | Nahla-Zayed | Analytical Inversion | Crank-Nicholson | Time Explicit |
|---|---|---|---|---|---|
| 0 | 300 | 300 | 300 | 300 | 300 |
| 10 | 305.3325 | 305.333 | 305.33304 | 305.333 | 305.333 |
| 20 | 311.1939 | 311.195 | 311.19507 | 311.1951 | 311.1951 |
| 30 | 317.3387 | 317.341 | 317.34067 | 317.3407 | 317.3407 |
| 40 | 323.5109 | 323.515 | 323.51506 | 323.5151 | 323.5151 |
| 50 | 329.4892 | 329.494 | 329.49409 | 329.4941 | 329.4941 |
| 60 | 335.1017 | 335.107 | 335.10725 | 335.1073 | 335.1073 |
| 70 | 340.2376 | 340.244 | 340.24382 | 340.2438 | 340.2438 |
| 80 | 344.8403 | 344.847 | 344.84712 | 344.8471 | 344.8471 |
| 90 | 348.896 | 348.904 | 348.90326 | 348.9033 | 348.9033 |
| 100 | 352.4213 | - | 352.42873 | 352.4287 | 352.4287 |

**Table 47.6** Power ($MW$) for initial reactivity $\rho(0) = 0.5\beta$.

| Time (s) | Adomian | Nahla-Zayed | Analytical Inversion | Crank-Nicholson | Time Explicit |
|---|---|---|---|---|---|
| 0 | 10.00000 | 10.00000 | 10.00000 | 10.00000 | 10.00000 |
| 10 | 18.19257 | 18.22015 | 18.19535 | 18.19535 | 18.19534 |
| 20 | 26.71911 | 26.74348 | 26.72623 | 26.72623 | 26.72622 |
| 30 | 31.88565 | 31.90059 | 31.89489 | 31.89489 | 31.89488 |
| 40 | 32.58814 | 32.59413 | 32.59608 | 32.59608 | 32.59608 |
| 50 | 30.09699 | 30.09671 | 30.10183 | 30.10183 | 30.10183 |
| 60 | 26.07722 | 26.07449 | 26.08038 | 26.08038 | 26.08038 |
| 70 | 21.69675 | 21.69170 | 21.69731 | 21.69731 | 21.69731 |
| 80 | 17.57909 | 17.57298 | 17.57798 | 17.57798 | 17.57798 |
| 90 | 13.99046 | 13.98410 | 13.98837 | 13.98837 | 13.98838 |
| 100 | 10.99761 | 10.99148 | 10.99506 | 10.99506 | 10.99506 |

**Table 47.7** Temperature ($K$) for initial reactivity $\rho(0) = 0.5\beta$.

| Time (s) | Adomian | Nahla-Zayed | Analytical Inversion | Crank-Nicholson | Time Explicit |
|---|---|---|---|---|---|
| 0 | 300 | 300 | 300 | 300 | 300 |
| 10 | 306.9508 | 306.953 | 306.9531 | 306.9531 | 306.9531 |
| 20 | 318.248 | 318.256 | 318.2562 | 318.2562 | 318.2562 |
| 30 | 333.0838 | 333.100 | 333.1004 | 333.1004 | 333.1004 |
| 40 | 349.3676 | 349.391 | 349.3924 | 349.3924 | 349.3924 |
| 50 | 365.1325 | 365.162 | 365.1628 | 365.1628 | 365.1628 |
| 60 | 379.2091 | 379.243 | 379.2431 | 379.2431 | 379.2431 |
| 70 | 391.1504 | 391.1860 | 391.1861 | 391.1861 | 391.1860 |
| 80 | 400.9504 | 400.987 | 400.9863 | 400.9863 | 400.9863 |
| 90 | 408.818 | 408.854 | 408.8534 | 408.8534 | 408.8534 |
| 100 | 415.0403 | 415.076 | 415.0746 | 415.0746 | 415.0746 |

**Table 47.8** Power ($MW$) for initial reactivity $\rho(0) = 0.8\beta$.

| Time (s) | Adomian | Nahla-Zayed | Analytical Inversion | Crank-Nicholson | Time Explicit |
|---|---|---|---|---|---|
| 0 | 10.00000 | 10.00000 | 10.00000 | 10.00000 | 10.00000 |
| 10 | 64.48478 | 64.79652 | 64.56067 | 64.56067 | 64.56055 |
| 20 | 83.23222 | 83.30637 | 83.31814 | 83.31814 | 83.31809 |
| 30 | 70.43784 | 70.46077 | 70.48758 | 70.48758 | 70.48759 |
| 40 | 53.89034 | 53.89474 | 53.91555 | 53.91555 | 53.91557 |
| 50 | 39.75314 | 39.74448 | 39.7597 | 39.7597 | 39.75972 |
| 60 | 28.82673 | 28.81245 | 28.82374 | 28.82374 | 28.82375 |
| 70 | 20.71048 | 20.69477 | 20.70331 | 20.70331 | 20.70333 |
| 80 | 14.79781 | 14.78265 | 14.78942 | 14.78942 | 14.78943 |
| 90 | 10.5368 | 10.52308 | 10.52864 | 10.52864 | 10.52864 |
| 100 | 7.48594 | 7.47445 | 7.47865 | 7.47865 | 7.47865 |

**Table 47.9** Temperature ($K$) for initial reactivity $\rho(0) = 0.8\beta$.

| Time (s) | Adomian | Nahla-Zayed | Analytical Inversion | Crank-Nicholson | Time Explicit |
|---|---|---|---|---|---|
| 0 | 300 | 300 | 300 | 300 | 300 |
| 10 | 317.2932 | 317.305 | 317.3071 | 317.3071 | 317.3071 |
| 20 | 356.6443 | 356.699 | 356.7019 | 356.7019 | 356.7019 |
| 30 | 395.5235 | 395.611 | 395.6125 | 395.6125 | 395.6124 |
| 40 | 426.5540 | 426.669 | 426.6687 | 426.6687 | 426.6687 |
| 50 | 449.8289 | 449.955 | 449.9537 | 449.9537 | 449.9536 |
| 60 | 466.8441 | 466.973 | 466.9703 | 466.9703 | 466.9703 |
| 70 | 479.1235 | 479.250 | 479.2474 | 479.2474 | 479.2474 |
| 80 | 487.9211 | 488.044 | 488.0410 | 488.0410 | 488.0410 |
| 90 | 494.1959 | 494.315 | 494.3117 | 494.3117 | 494.3117 |
| 100 | 498.6588 | 498.774 | 498.7708 | 498.7708 | 498.7708 |

expansion of the neutron density and precursor concentration opened pathways for an apparently stable recursive scheme. We are aware of the fact that at the present state of the work one would like to have genuine convergence criterion, but postpone this task to a future work. Nevertheless, exhaustive tests with examples have shown us that at least for the range of parameters that make sense from the nuclear physics point of view the presented implementations show consensus, which could indicate that a true convergence in the considered cases exist.

In the subsequent consideration a nonlinear problem was focused on, again comparing the effectiveness of another decomposition method, namely Adomian's prescription for nonlinear deterministic and stochastic systems. To this end, the neutron point kinetics equation was extended by a temperature feedback, which introduced a nonlinearity into the previous system because of its influence on reactivity and its dependence on the neutron density. Again numerical findings and semi-analytical evaluations showed fairly good agreement and one can come to the conclusion that a methodology that originally was developed with the intuition

to consolidate an analytical approach for this type of problems seems to prove useful also for numerical approaches. This is especially desirable due to the fact that appearing integrals that contain the nonlinear source terms may in many cases not be evaluated analytically. Although one may claim that the closed integral forms represent the solution as an analytical expression, but as a consequence of the present findings also numerical approximations are promising and seem to maintain the solutions within a reasonable precision despite stiffness and/or nonlinearities. Thus, it seems an interesting perspective, that a reasoning that led to the development of solutions in form of analytical representation are also useful beyond the original scope and may improve numerical approaches that otherwise would suffer from the limitations as, for instance, the Lax–Milgram theorem for stiff problems. It is noteworthy, that for nonlinear problems no equivalent theorem seems to be available so far to the scientific community.

## References

[AbHa02] Aboanber, A.E., Hamada, Y.M.: PWS: an efficient system for solving space-independent nuclear reactor dynamics. Annals of Nuclear Energy **29**, 2159–2172 (2002).

[AbNa02] Aboanber, A.E., Nahla, A.A.: Solution of the point kinetics equations in the presence of Newtonian temperature feedback by Padé approximations via the analytical inversion method. Journal of Physics A **35**, 9609–9627 (2002).

[AbNa11] Abdallah, A., Nahla, A.A.: An efficient technique for the point reactor kinetics equations with Newtonian temperature feedback effects. Annals of Nuclear Energy **38**, 307–330 (2011).

[AbHa03] Aboanber, A.E., Hamada, Y.M.: Power series solution (PWS) of nuclear reactor dynamics with Newtonian temperature feedback. Annals of Nuclear Energy **30**, 1111–1122 (2003).

[Ab09] Aboanber, A.E.: Exact solution for the non-linear two point kinetic model of reflected reactors. Progress in Nuclear Energy **51**, 719–726 (2009).

[Ad94] Adomian, G.: Solving Frontier Problems of Physics: The Decomposition Method. Kluwer Academic Publishers, Dordrecht, The Netherlands (1994).

[ChEtAl07] Chen, W., Guo, L., Zhu, B. and Li, H.: Accuracy of analytical methods for obtaining supercritical transients with temperature feedback. Progress in Nuclear Energy **49**, 290–302 (2007).

[KiAl04] Kinard, M., Allen, E.J.: Efficient numerical solution of the point kinetics equations in nuclear reactor dynamics. Annals of Nuclear Energy **31**, 1039–1051 (2004).

[Na10] Nahla, A.A.: Analytical solution to solve the point kinetics equations. Nuclear Engineering and Design **240**, 1622–1629 (2010).

[NaZa10] Nahla, A.A., Zayed E.M.E.: Solution to the non-linear point nuclear reactor kinetics Equations. Progress in Nuclear Energy **52**, 743–746 (2010).

[PeNiRa06] Peinetti, F., Nicolino, C., Ravetto, P.: Kinetics of a point reactor in the presence of reactivity oscillations. Annals of Nuclear Energy **33**, 1189–1195 (2009).

[PeEtAl11] Petersen, C.Z., Dulla, S., Vilhena, M.T.M.B. and Ravetto, P.: An analytical solution of the point kinetics equations with time-variable reactivity by the decomposition method. Progress in Nuclear Energy **53**, 1091–1094 (2011).

[SaPa13] Saha Ray, S., Patra, A.: Numerical solution for stochastic point-kinetics equations with sinusoidal reactivity in dynamical system of nuclear reactor. International Journal of Nuclear Energy Science and Technology **7**(3), 231–242 (2013).

[SiEtAl14] Silva, J.J.A., Alvim, A.C.M., Vilhena, M.T.M.B., Bodmann, B.E.J., Petersen, C.Z.: On a closed-form solution of the point kinetics equations with reactivity feedback of temperature. Int. J. of Nuclear Energy Science and Technology **8**(2), 131–145 (2014).

[TaJaHa10] Tashakor, S., Jahanfarnia, G., Hashemi-Tilehnoee, M.: Numerical solution of the point kinetics equations with fuel burn-up and temperature feedback. Annals of Nuclear Energy **37**, 265–269 (2010).

[WoEtAl14] Wollmann da Silva, M., Bogado Leite, S., Vilhena, M.T., Bodmann, B.E.J.: On an analytical representation for the solution of the neutron point kinetics equation free of stiffness. Annals of Nuclear Energy **71**, 97–102 (2014).

# Chapter 48
# The Wind Meandering Phenomenon in an Eulerian Three-Dimensional Model to Simulate the Pollutants Dispersion

**V.C. Silveira, D. Buske, and G.A. Degrazia**

## 48.1 Introduction

Usually in stable conditions in the Planetary Boundary Layer (PBL), when the wind velocity magnitude is low, horizontal wind low-frequency oscillations are observed. These horizontal wind oscillations characterize the wind meandering phenomenon.

The aim of the present work is to develop a new model to simulate the atmospheric pollutants dispersion taking into account the meandering in low wind speed. To accomplish this, we present a new analytical solution for the three-dimensional advection–diffusion equation. The advection–diffusion equation is solved combining the Laplace transform and Generalized Integral Laplace Transform Technique (GILTT) [BuEtAl08, MoEtAl09].

The importance of the present study to describe the pollutant dispersion in low wind conditions lies in the fact that such conditions occur frequently and are of crucial importance in air pollution evaluations. In such conditions, traditional models employed to calculate the contaminants concentration can be not adapted to predict the pollutants dispersion.

A characteristic of the wind meandering is the presence of observed negative lobules in the autocorrelation function (ACF) associated with the horizontal components of the wind vector [AnEtAl05]. In this study we employ an Eulerian dispersion model to investigate the transport phenomenon of contaminants caused by the wind meandering.

V.C. Silveira (✉) • G.A. Degrazia
Federal University of Santa Maria, Av. Roraima 1000, Santa Maria 97105-900, RS, Brazil
e-mail: viliamcardoso@gmail.com; gervasiodegrazia@gmail.com

D. Buske
Federal University of Pelotas, Pelotas, Brazil
e-mail: danielabuske@gmail.com



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_48

To evaluate the present methodology we used observed concentration data in stable conditions of low wind speed that were measured in the classical diffusion experiment called Idaho National Engineering Laboratory (INEL) [SaDi74].

## 48.2 Analytical Solution

The advection–diffusion equation is written as

$$\overline{u}\frac{\partial \overline{c}}{\partial x}+\overline{v}\frac{\partial \overline{c}}{\partial y}+\overline{w}\frac{\partial \overline{c}}{\partial z}=\frac{\partial}{\partial x}\left(K_x\frac{\partial \overline{c}}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_y\frac{\partial \overline{c}}{\partial y}\right)+\frac{\partial}{\partial z}\left(K_z\frac{\partial \overline{c}}{\partial z}\right) \tag{48.1}$$

where the wind speed and the wind direction of the INEL experiments were utilized to calculate the u and v wind components. The u- and v-components are given by

$$u = Vsin(\theta)$$

$$v = Vcos(\theta)$$

where $V$ is the horizontal wind velocity magnitude and $\theta$ is the angle. The boundary and source conditions are

$$K_x\frac{\partial \overline{c}(L_x,y,z)}{\partial x}=K_y\frac{\partial \overline{c}(x,0,z)}{\partial y}=K_y\frac{\partial \overline{c}(x,L_y,z)}{\partial y}=$$
$$=K_z\frac{\partial \overline{c}(x,y,0)}{\partial z}=K_z\frac{\partial \overline{c}(x,y,h)}{\partial z}=0$$
$$\overline{u}\,\overline{c}(0,y,z)=Q\delta(y-y_o)\delta(z-H_s) \tag{48.2}$$

Using the integral transform technique in the y variable and expanding the pollutant concentration yields

$$\overline{c}(x,y,z)=\sum_{n=0}^{N}\frac{\overline{c_n}(x,z)\zeta_n(y)}{N_n^{\frac{1}{2}}} \tag{48.3}$$

Replacing the equation (48.3) in the equation (48.1) and applying the operator

$$\frac{1}{N_m^{\frac{1}{2}}}\int_0^{L_y}()\zeta_m dy \tag{48.4}$$

we can write the equation

$$\begin{aligned}&\alpha_{n,m}\overline{u}\frac{\partial\overline{c_n}}{\partial x}+\beta_{n,m}\overline{v}\,\overline{c_n}+\alpha_{n,m}\overline{w}\frac{\partial\overline{c_n}}{\partial z}=\\&=\alpha_{n,m}\frac{\partial}{\partial x}\left(K_x\frac{\partial\overline{c_n}}{\partial x}\right)+\alpha_{n,m}\frac{\partial}{\partial z}\left(K_z\frac{\partial\overline{c_n}}{\partial z}\right)-\alpha_{n,m}{\lambda_n}^2K_y\overline{c_n}\end{aligned}\tag{48.5}$$

The matrices $\alpha_{n,m}$ and $\beta_{n,m}$ are

$$\alpha_{n,m}=\frac{1}{N_n^{\frac{1}{2}}N_m^{\frac{1}{2}}}\int_0^{L_y}\zeta_n(y)\zeta_m(y)dy=\begin{cases}0,\ \mathrm{m}\neq\mathrm{n}\\1,\ \mathrm{m}=\mathrm{n}\end{cases}$$

and

$$\begin{aligned}&\beta_{n,m}=\frac{1}{N_n^{\frac{1}{2}}N_m^{\frac{1}{2}}}\int_0^{L_y}\zeta_n'(y)\zeta_m(y)dy=\\&=\begin{cases}\frac{2n^2}{L_y(n^2-m^2)}\left[cos(n\pi)cos(m\pi)-1\right],\ \mathrm{m}\neq\mathrm{n}\\0,\ \mathrm{m}=\mathrm{n}\end{cases}\end{aligned}$$

The solution of the problem (48.5) is

$$\overline{c_n}(x,z)=\sum_{i=0}^{I}\overline{c_{n,i}}(x)\,\varsigma_i(z)\tag{48.6}$$

Replacing the equation (48.6) in the equation (48.5), we can write the equation (48.5) in matrix form as

$$Y''(x)+FY'(x)+GY(x)=0\tag{48.7}$$

The matrices $F$ and $G$ are given, respectively, by $F=B^{-1}D$ and $G=B^{-1}E$. The matrices $B$, $D$ and $E$ are written as

$$b_{i,j}=\alpha_{n,m}\int_0^h K_x\,\varsigma_i(z)\,\varsigma_j(z)dz$$

$$d_{i,j}=-\alpha_{n,m}\int_0^h\overline{u}\,\varsigma_i(z)\,\varsigma_j(z)dz+\alpha_{n,m}\int_0^h K_x'\,\varsigma_i(z)\,\varsigma_j(z)dz$$

$$e_{i,j}=-\alpha_{n,m}\int_0^h\overline{w}\,\varsigma_i'(z)\,\varsigma_j(z)dz+\alpha_{n,m}\int_0^h K_z'\,\varsigma_i'(z)\,\varsigma_j(z)dz-$$

$$-\alpha_{n,m}\lambda_i^2 \int_0^h K_z\, \varsigma_i(z)\, \varsigma_j(z)dz - \alpha_{n,m}\lambda_i^2 \int_0^h K_y\, \varsigma_i(z)\, \varsigma_j(z)dz -$$

$$-\beta_{n,m} \int_0^h \overline{v}\, \varsigma_i(z)\, \varsigma_j(z)dz$$

To solve the problem described by the equation (48.7) we apply order reduction:

$$Z'(x) + H.Z(x) = 0 \tag{48.8}$$

The $H$ matrix has the block form

$$H = \begin{bmatrix} 0 & -I \\ G & F \end{bmatrix}$$

The transformed problem provided by the equation (48.8) is solved by the Laplace transform technique and diagonalization:

$$Z(x) = XM(x)\,\xi$$

$M(x)$ is the diagonal matrix with elements $e^{-d_i x}$, $X$ is an eigenvector matrix of the $H$ matrix, $\xi = X^{-1}Z(0)$, $X^{-1}$ is the inverse matrix of the eigenvector of the $H$ matrix and $Z(0)$ is the initial condition. In this model the w component of the wind speed is zero.

For the source condition, replacing the equation (48.3) in the equation (48.2) and applying the operator of the equation (48.4), we obtain the formulation

$$\begin{aligned} &\overline{u} \sum_{n=0}^{N} \overline{c}_n(0,z) \frac{1}{N_n^{\frac{1}{2}}} \frac{1}{N_m^{\frac{1}{2}}} \int_0^{L_y} \zeta_n(y)\zeta_m(y)dy \\ &= Q\delta(z - H_s) \frac{1}{N_m^{\frac{1}{2}}} \int_0^{L_y} \delta(y - y_o)\zeta_m(y)dy \end{aligned} \tag{48.9}$$

Replacing the equation (48.6) in the equation (48.9) yields

$$\overline{u} \sum_{i=0}^{I} \overline{c}_{n,i}(0)\varsigma_i(z) \frac{1}{N_n^{\frac{1}{2}}} \frac{1}{N_m^{\frac{1}{2}}} \int_0^{L_y} \zeta_n(y)\zeta_m(y)dy = \frac{Q\zeta_i(y_o)}{N_m^{\frac{1}{2}}} \delta(z - H_s) \tag{48.10}$$

Applying the following operator in the equation (48.10), we find that

$$\int_0^h (\,)\varsigma_j(z)dz$$

results in the source conditions

$$Y(0) = \frac{Q\zeta_i(y_o)\varsigma_j(H_s)}{\overline{u}\sqrt{L_y}h} \quad para \quad ((i=j)\ e\ (m=n)) = 0$$

$$Y(0) = \frac{Q\zeta_i(y_o)\varsigma_j(H_s)}{\overline{u}\sqrt{\frac{L_y}{2}\frac{h}{2}}} \quad para \quad ((i=j)\ e\ (m=n)) \neq 0$$

where $Y(0)$ is the column vector containing the components $\{\overline{c_{n,i}}(0)\}$.

## 48.3 Parameterization of the Turbulence

In the present dispersion model, to parameterize the turbulence effects we employ the algebraic eddy diffusivities that depend on the source distance. In the stable conditions [DeViMo96] proposed the following algebraic formulation to the eddy diffusivities ($K_\alpha$):

$$K_\alpha = \frac{2\sqrt{\pi}0.64u_* h a_i^2 (1-z/h)^{\alpha_1}(z/h)X^*}{[2\sqrt{\pi}0.64(z/h)+16a_i(f_m)_i(1-z/h)^{\alpha_1/2}X^*]^2} *$$

$$*\frac{[2\sqrt{\pi}0.64a_i^2(z/h)+8a_i(f_m)_i(1-z/h)^{\alpha_1/2}X^*]}{[2\sqrt{\pi}0.64(z/h)+16a_i(f_m)_i(1-z/h)^{\alpha_1/2}X^*]^2}$$

where the $X^*$ is the dimensionless source distance

$$X^* = \frac{xw_*}{\overline{u}h}$$

with $\alpha = x, y, z$, $i = u, v, w$, $h$ is the height of stable boundary layer, $\alpha_1$ is a constant that depends on the evolution of the stable boundary layer, $(f_m)_i$ is the frequency of spectral peak provided by the relation

$$(f_m)_i = (f_m)_{n,i}\left(1+3.7\frac{z}{\Lambda}\right)$$

where $(f_m)_{n,i}$ is the frequency of spectral peak in the surface for neutral conditions $[(f_m)_{n,w} = 0.33;\ (f_m)_{n,v} = 0.22;\ (f_m)_{n,u} = 0.045]$, z is the height above the ground, $\Lambda$ is the local Monin–Obukhov length expressed by

$$\Lambda = L\left(1-\frac{z}{h}\right)^{(1.5\alpha_1-\alpha_2)}$$

with $[\alpha_1 = 1.5; \alpha_2 = 1]$ and $a_i$ is given by the ratio

$$a_i = \frac{(2.7c_i)^{1/2}}{(f_m)_{n,i}^{1/3}}$$

where $c_{v,w}$ and $c_u$ are, respectively, given by $[c_{v,w} = 0.36; c_u = 0.27]$.

## 48.4 Wind Profile

The wind speed profile is described by the power law [PaDu84]

$$\frac{\overline{V}}{\overline{V}_1} = \left(\frac{z}{z_1}\right)^{\alpha}$$

where $V$ is provided to the $u$ and $v$ components, $\overline{V}$ and $\overline{V}_1$ are wind mean velocities at the heights $z$ and $z_1$ and $\alpha$ is a constant ($\alpha = 0.1$).

In this study the wind speed profile is also described by a similarity law:

$$\overline{V} = \frac{u_*}{k}\left[ln\left(\frac{z}{z_0}\right) - \psi_m\left(\frac{z}{L}\right)\right] \; for\; z \leq z_b$$

and

$$\overline{V} = \overline{V}(z_b)\; for\; z > z_b$$

$z_b = min[|L|, 0, 1h]$, $k$ is the von Kármán constant ($\approx 0.4$) and $z_0$ is the terrain roughness.

The stability function is given by the Businger relationship:

$$\psi_m\left(\frac{z}{L}\right) = -4.7\frac{z}{L}, \;\; \frac{1}{L} \geq 0$$

## 48.5 Experimental Data

The INEL low wind speed diffusion experiments were accomplished in a flat and uniform terrain. The pollutant $SF_6$ was collected in arcs of 100, 200 and 400 m of the emission point at the height of 0.76 m above of the ground. The pollutant was released of a height of 1.5 m above of the ground-level. The wind in 2 m was obtained of the experiment. The roughness length for the INEL experiments was of 0.005 m.

**Table 48.1** The observed and calculated meteorological parameters of the INEL experiment.

| Exp. | $\bar{u}(2m)(ms^{-1})$ | $u_*(ms^{-1})$ | $L(m)$ | $h(m)$ |
|---|---|---|---|---|
| 4 | 0.7 | 0.047 | 2.4 | 13 |
| 5 | 0.8 | 0.053 | 3.1 | 16 |
| 6 | 1.2 | 0.08 | 7.1 | 30 |
| 7 | 0.6 | 0.04 | 1.8 | 11 |
| 8 | 0.5 | 0.033 | 1.2 | 8 |
| 9 | 0.5 | 0.033 | 1.2 | 8 |
| 10 | 1.1 | 0.073 | 5.9 | 26 |
| 11 | 1.4 | 0.093 | 9.6 | 37 |
| 12 | 0.7 | 0.047 | 2.4 | 13 |
| 13 | 1.0 | 0.067 | 4.9 | 23 |
| 14 | 1.0 | 0.067 | 4.9 | 23 |

The Monin–Obukhov length, friction velocity and height of the PBL were not measured in the INEL experiments, but were calculated from the empirical formulations.

The Monin–Obukhov length was calculated utilizing the relation [Za90]

$$L = 1100u_*^2$$

The friction velocity is calculated as [Ve80]:

$$u_* = \frac{k\bar{u}(z_r)}{ln(z_r/z_o)}$$

where $z_r = 2$ m (reference height).

To calculate $h$ we use the expression [Zi72]

$$h = 0.4\left(\frac{u_*L}{f_c}\right)^{1/2}$$

The meteorological parameters of the INEL experiment are showed in Table 48.1:

## 48.6 Statistical Indexes

The statistical indexes describe the model performance to reproduce the observed concentrations. In our analysis we use the following Normalized Mean Square Error ($NMSE$), Correlation Coefficient ($COR$), Factor of 2 ($FA2$), Fractional Bias ($FB$), and Standard Fractional Bias ($FS$) (see [HaPa89]).

The Normalized Mean Square Error ($NMSE$) represents all deviations between the simulated and observed concentrations:

$$NMSE = \frac{1}{N}\frac{\sum(c_o - c_p)^2}{\overline{c_o}\,\overline{c_p}}$$

The Correlation Coefficient ($COR$) represents the association degree between the simulated and observed concentrations:

$$COR = \frac{(\sum c_o\, c_p) - N\overline{c_o}\,\overline{c_p}}{\sqrt{(\sum c_o^2 - N\overline{c_o}^2)(\sum c_p^2 - N\overline{c_p}^2)}}$$

The Factor of 2 ($FA2$) represents the data fraction (% normalized to one):

$$FA2 = \frac{\overline{c_p}}{\overline{c_o}}, \quad 0,5 \leq FA2 \leq 2$$

The Fractional Bias ($FB$) represents the tendency of the model to overestimate or underestimate the observed concentrations:

$$FB = \frac{\overline{c_o} - \overline{c_p}}{0,5(\overline{c_o} + \overline{c_p})}$$

The Standard Fractional Bias ($FS$):

$$FS = \frac{\sigma_o - \sigma_p}{0,5(\sigma_o + \sigma_p)}$$

where the $\sigma_o$ and $\sigma_p$ are the observed and simulated standard deviation. The better results are obtained when the COR and FA2 are nearest of one and the NMSE, FB and FS are close to zero.

## 48.7 Results

Figure 48.1 shows the negative lobule in the autocorrelation function that characterizes the wind meandering of the horizontal components.

Figure 48.2 shows the scatter diagram of the observed and simulated concentrations using the wind power law. Figure 48.3 shows the scatter diagram of the observed and simulated concentrations using the wind similarity law. The better results are obtained with the wind power law.

To test the very low wind speed influence in the simulation of contaminant concentrations, we selected only cases in that the wind velocity is lesser than $1m/s$. On the other hand, Figure 48.4 shows the scatter diagram of the observed and

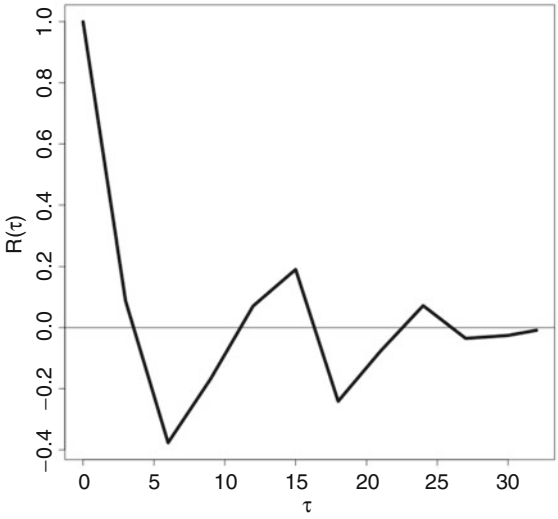


**Fig. 48.1** Autocorrelation function using database INEL.

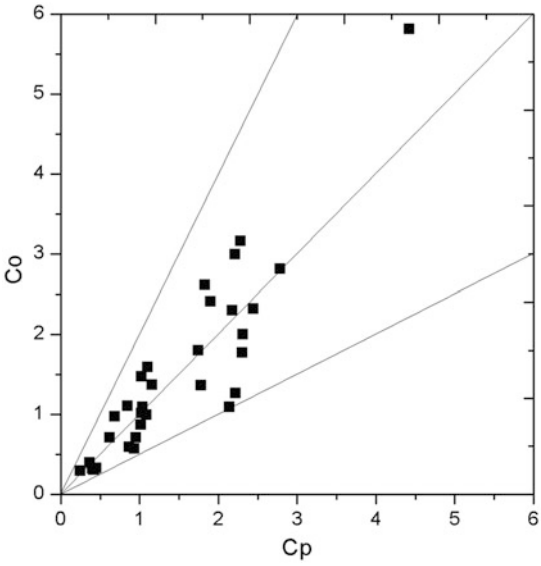


**Fig. 48.2** Scatter diagram of the observed ($c_o$) and simulated ($c_p$) concentrations data by 3D-GILTT method using wind power law.

simulated concentrations using the wind power law with $V < 1m/s$ and Figure 48.5 shows scatter diagram of the observed and simulated concentrations using the wind similarity law with $V < 1m/s$. Better results are obtained when we do not consider only the horizontal wind lesser than $1m/s$.

Table 48.2 shows the statistical performance of the present model. The model simulates satisfactorily the observed concentrations using the wind power law and wind similarity law without considering the very low wind speed. The NMSE, FB and FS are nearest to zero and COR and FA2 are one or are nearest to one.

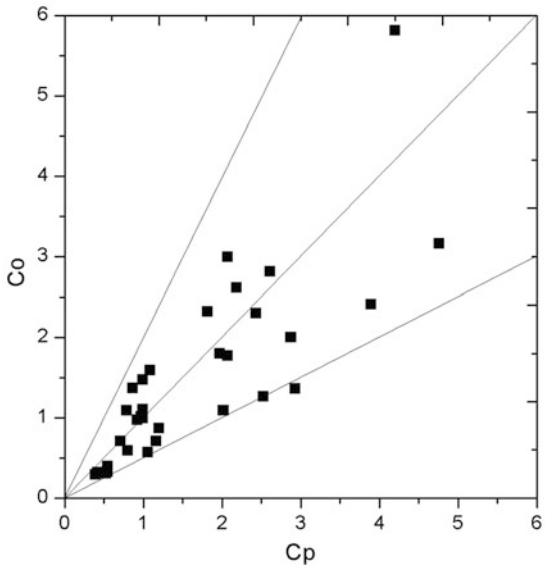


**Fig. 48.3** Scatter diagram of the observed ($c_o$) and simulated ($c_p$) concentrations data by 3D-GILTT method using wind similarity law.

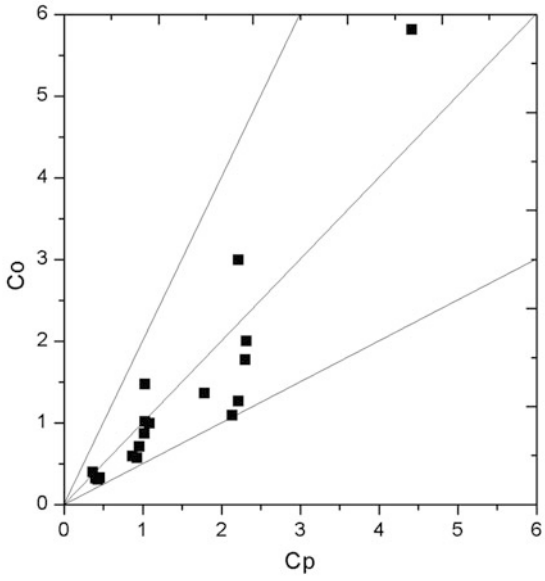


**Fig. 48.4** Scatter diagram of the observed ($c_o$) and simulated ($c_p$) concentrations data by 3D-GILTT method using the wind power law ($V < 1m/s$).

Table 48.3 shows the results of the observed and simulated concentrations using power and similarity wind.

**Table 48.2** Statistical performance of the present model.

| Simulation | NMSE | COR | FA2 | FB | FS |
|---|---|---|---|---|---|
| Wind power law | 0.11 | 0.91 | 1.00 | 0.03 | 0.22 |
| Wind similarity law | 0.20 | 0.82 | 0.94 | −0.10 | 0.00 |
| Wind power law (V<1m/s) | 0.16 | 0.91 | 1.00 | −0.09 | 0.25 |
| Wind similarity law (V<1m/s) | 0.28 | 0.83 | 0.89 | −0.16 | 0.22 |

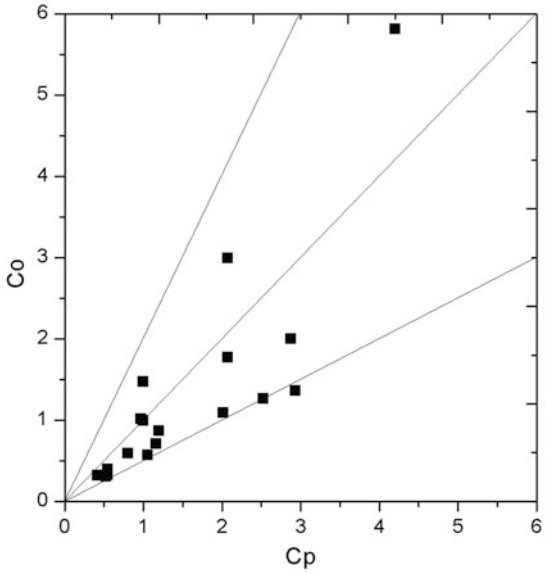


**Fig. 48.5** Scatter diagram of the observed ($c_o$) and simulated ($c_p$) concentrations data by 3D-GILTT method using the wind similarity law ($V < 1m/s$).

## 48.8 Conclusions

It is very difficult to simulate the pollutants dispersion in low wind speed stable conditions. In such conditions the horizontal wind has not a predominant direction and the models cannot be adapted to simulate the pollutants dispersion.

The results generated by the present dispersion model show a good agreement between the observed and simulated concentrations. The statistical indexes show a good performance, with COR and FA2 close to one and NMSE, FB and FS close to zero. Better results are obtained when we do not consider a very low wind speed ($V < 1m/s$).

**Acknowledgements** The authors wish to thank CAPES, CNPq, and FAPERGS for partial financial support of this work.

**Table 48.3** The observed ($c_o$) and simulated ($c_p$) concentrations using the INEL experiment and wind power law and wind similarity law.

| Exp. | Dis. (m) | $c_o$ | $c_p$ (wind power law) | $c_p$ (wind similarity law) |
|---|---|---|---|---|
| 4 | 100 | 5.81 | 4.42 | 4.20 |
| | 200 | 2.99 | 2.21 | 2.07 |
| | 400 | 1.47 | 1.03 | 1.00 |
| 5 | 100 | 1.36 | 1.79 | 2.93 |
| | 200 | 0.87 | 1.02 | 1.20 |
| | 400 | 0.30 | 0.43 | 0.53 |
| 6 | 100 | 2.61 | 1.84 | 2.19 |
| | 200 | 0.97 | 0.69 | 0.93 |
| | 400 | 0.29 | 0.25 | 0.39 |
| 7 | 100 | 1.26 | 2.22 | 2.53 |
| | 200 | 0.71 | 0.96 | 1.16 |
| | 400 | 1.01 | 1.03 | 0.97 |
| 8 | 100 | 0.59 | 0.87 | 0.81 |
| | 200 | 0.32 | 0.41 | 0.42 |
| | 400 | 0.33 | 0.45 | 0.55 |
| 9 | 100 | 1.09 | 2.14 | 2.02 |
| | 200 | 0.57 | 0.93 | 1.06 |
| | 400 | 0.39 | 0.37 | 0.55 |
| 10 | 100 | 2.41 | 1.90 | 3.89 |
| | 200 | 1.80 | 1.75 | 1.97 |
| | 400 | 0.71 | 0.63 | 0.71 |
| 11 | 100 | 2.32 | 2.44 | 1.82 |
| | 120 | 1.09 | 1.04 | 0.79 |
| | 400 | 1.10 | 0.85 | 1.00 |
| 12 | 100 | 2.00 | 2.31 | 2.88 |
| | 200 | 1.77 | 2.30 | 2.07 |
| | 400 | 0.99 | 1.09 | 1.00 |
| 13 | 100 | 3.16 | 2.28 | 4.76 |
| | 200 | 2.30 | 2.18 | 2.44 |
| | 400 | 1.37 | 1.16 | 0.87 |
| 14 | 100 | 2.81 | 2.79 | 2.61 |
| | 200 | 1.59 | 1.11 | 1.09 |
| | 400 | 0.30 | 0.42 | 0.46 |

## References

[AnEtAl05] Anfossi, D., Oettl, D., Degrazia, G., Goulart, A.: An analysis of sonic anemometer observations in low wind speed conditions. Boundary-Layer Meteorology, 179–203, 114 (2005).

[BuEtAl08] Buske, D., Vilhena, M. T., Moreira, D. M., Bodmann, B.: An analytical solution for the steady-state two-dimensional diffusion-advection-deposition model by the GILTT approach. Integral Methods in Science and Engineering: Techniques and Applications, Organized by: C. Constanda; S. Potapenko - Birkhauser, Boston, 27–36, (2008).

[DeViMo96] Degrazia, G. A., Vilhena, M. T., Moraes, O. L. L.: An algebraic expression for the eddy diffusivities in the stable boundary layer: a description of near-source diffusion. Il Nuovo Cimento, 399–403, 19C (1996).

[HaPa89] Hanna, S., Paine, R. J.: Hibrid plume dispersion model (HPDM) development and evaluation. Journal of Applied Meteorology, 206–224, 28 (1989).

[MoEtAl09] Moreira, D., Vilhena, M. T., Buske, D., Tirabassi, T.: The state-of-art of the GILTT method to simulate pollutant dispersion in the atmosphere. Atmospheric Research, 1–17, 92 (2009).

[PaDu84] Panosfsky, H. A., Dutton, J. A.: Atmospheric Turbulence. New York: John Wiley & Sons (1984).

[SaDi74] Sagendorf, J. F., Dickson, C. R.: Diffusion under low wind-speed, inversion conditions. Technical Memorandum ERL ARL-52, U. S. National Oceanics and Atmospherics Administration (1974).

[Ve80] Venkatram, A.: Estimating the Monin-Obukhov length in the stable boundary layer for dispersion calculations. Boundary Layer Meteorology, 481–485, 19 (1980).

[Za90] Zannetti, P.: Air Pollution Modelling. Computational Mechanics Publications, Southampton, 444pp, (1990).

[Zi72] Zilitinkevith, S. S.: On the determination of the height of the Ekman boundary layer. Boundary Layer Meteorology, 141–145, 3 (1972).

# Chapter 49
# Semilinear Second-Order Ordinary Differential Equations: Distances Between Consecutive Zeros of Oscillatory Solutions

**Tadie**

## 49.1 Preliminaries

Oscillation criteria for semilinear differential equations have been largely investigated in the literature, both in one-dimensional and in multi-dimensional cases. But the arrangements of the zeros for two different solutions or the diameters of two consecutive zeros of a solution are rare to find in the literature. This note gives some lights for the one-dimensional case, i.e. for equations of the type

$$\left(a(t)\phi_\alpha(u')\right)' + c(t)\phi_\alpha(u) + f(t,u) = 0; \quad t \geq 0$$

where $a \in C^1(\mathbb{R},\ (0,\ \infty))$ with $a' \geq 0$; $\exists T, m > 0 \quad c \in C([T,\ \infty),\ (m,\ \infty)), f \in C(\mathbb{R}^2,\ \mathbb{R})$ and for some $\alpha > 0,\ \phi_\alpha(S) := |S|^{\alpha-1}S$.

**Definition 1.** Let $h \in C(E)$ where $E$ denotes $\mathbb{R}$ or $\mathbb{R}^n$. $h$ will be said to be

(i) (weakly) oscillatory in $E$ if $\forall T > 0, \quad h$ has a zero in $\Omega_T := \{x \in E\,;\, |x| > T\}$;
(ii) strongly oscillatory if it has a nodal set in any $\Omega_T$, $\forall T > 0$, where a nodal set is any nontrivial connected and bounded component of the support $D(h)$ of $h$.
(iii) A differential equation will be said to be oscillatory if any of its nontrivial and bounded solutions is oscillatory.
(iv) Therefore a function $w$ will be said not to be oscillatory if either there are $\mu, R > 0$ such that $|w| > \mu$ in $\Omega_R$ or $\liminf_{t\nearrow\infty} |w(t)| = 0$.

Tadie (✉)
Universitet Copenhagen, Universitet Sparken 5, 2100 Copenhagen, Denmark
e-mail: tadietadie@yahoo.com



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_49

(v) Two distinct oscillatory functions $u_1$ and $u_2$ will be said to have alternating zeros in $\Omega_T$ say, if in $\Omega_T$, between any two consecutive zeros of $u_i$ lies a zero of $u_j$ where $i \neq j$ and $i,j \in \{1,\ 2\}$.

In the sequel, unless indicated otherwise, $\phi$ will denote $\phi_\alpha$ with an $\alpha > 0$.
We will be dealing with equations of the types

$$\begin{cases} (i) \quad P(y) := \left\{ a(t)\phi(y') \right\}' + q(t)\phi(y) = 0, \quad t \in \mathbb{R} \quad \text{and} \\ (ii) \quad K(u) := \left\{ A(t)\phi(u') \right\}' + Q(t)\phi(u) + F(t,u,u') = 0, \end{cases}$$

where $a, A \in C^1(\mathbb{R},\ (0,\ \infty))\ ,\ (0,\ \infty))\ ,\ q \in C(\mathbb{R},\ \mathbb{R})\ ,\ C \in C(\mathbb{R},\ \mathbb{R})$ and for some $\alpha > 0,\ t \in \mathbb{R},\ \phi(t) := \phi_\alpha(t) = |t|^{\alpha-1}t$; with the properties that $t\phi(t) = |t|^{\alpha+1}$ and $t\phi'(t) = \alpha\phi(t)$. In (i) we have a semilinear equation and in (ii) a perturbed semilinear one.

In the sequel, these general hypotheses will be set on the coefficients of the semilinear parts of the equations:

**(H)**: h1) the numerical functions $a$ and $A$ are continuously differentiable in their respective domains, strictly positive and non-decreasing;

h2) $Q$ and $q$ are continuous in their respective arguments and eventually strictly positive (i.e. $\exists m,\ T > 0\ ;\ Q, q > m$ in $\Omega_T$).

By means of some comparison methods based on some Picone-type identities (inequalities) we will investigate the arrangement of the zeros of some distinct oscillatory solutions. We recall here some Picone's formulas which will be often referred to: given two equations for $i = 1,2$ and $\phi := \phi_\alpha$

$$K_i(u_i) := \left\{ a_i(t)\phi(y_i') \right\}' + q_i(t)\phi(y) + f_i(t,u_i,u_i') = 0, \quad t \in \mathbb{R},$$

we easily have (see, e.g., [Ta12]) the following Picone-type identity:

$$\begin{cases} \Psi(y_1,\ y_2)\ := \left\{ y_1 a_1(t)\phi(y_1') - y_1\phi(\ \frac{y_1}{y_2}\ )a_2(t)\phi(y_2') \right\}' = \\ a_2(t)\zeta_\alpha(y_1,y_2) + \left[ a_1(t) - a_2(t) \right] |y_1'|^{\alpha+1} + \\ \left[ q_2(t) - q_1(t) \right] |y_1|^{\alpha+1} + |y_1|^{\alpha+1} \left[ \dfrac{f_2(t,y_2,y_2')}{\phi(y_2)} - \dfrac{f_1(t,y_1,y_1')}{\phi(y_1)} \right] \end{cases} \tag{49.1}$$

where $\forall \gamma > 0$, the two-form function $\zeta_\gamma$ is defined $\forall u, v \in C^1(\mathbb{R},\ \mathbb{R})$ by

$$(Z1): \qquad \zeta_\gamma(u,v) \begin{cases} = |u'|^{\gamma+1} - (\gamma+1)u'\phi_\gamma(\dfrac{uv'}{v}) + \gamma v' \dfrac{u}{v}\phi_\gamma(\dfrac{uv'}{v}) \\ = |u'|^{\gamma+1} - (\gamma+1)u'\phi_\gamma(\dfrac{uv'}{v}) + \gamma|\dfrac{uv'}{v}|^{\gamma+1} \end{cases}$$

which is strictly positive for non-constant $u \neq \lambda v, \quad \forall \lambda \in \mathbb{R}$.

*Remark 1.* (R1) An important tool in our investigation is the following:
Let $u$ and $v$ be oscillatory functions in $\Omega_T$ such that $v$ has a zero, $v_1$ in a D(u) and $u_1 \in D(u)$ is the singularity of $u$. A suitable translation of $v$ like $V(t) := v(t+\xi)$, $\xi \in \mathbb{R}$ can be chosen such that $V'(u_1) = u'(u_1) = 0$. Because $V$ and $v$ have nodal sets of the same sizes, this in some cases provides a clear comparison like $diam(v) = diam(D(V)) \leq diam(D(u))$.

(R2) We will show that in certain conditions on the coefficients of the equations of $u$ and $v$, wherever their nodal sets $D(u)$ and $D(v)$ overlap, a translation of one of them $D(u)$, say, can be chosen such that $D(u_1) \subset D(v)$ where $u_1$ is a translation of $u$.

## 49.2 Comparison of Diameters of Overlapping Nodal Sets

It is well established that under the hypotheses (H), any semi-linear equation

$$\left\{a(t)\phi_\alpha(u')\right\}' + q(t)\phi_\alpha(u) = 0; \quad t \geq 0$$

is strongly oscillatory (see, e.g., [Ta12] and the references therein).

Consider in $t \geq 0$ the oscillatory equations

$$\begin{cases} (i) \quad \left\{a(t)\phi(y')\right\}' + q(t)\phi(y) = 0 \\ (ii) \quad and \quad \left\{A(t)\phi(z')\right\}' + Q(t)\phi(z) = 0 \end{cases} \tag{49.2}$$

where the coefficients satisfy (H).

As in (49.1) the solutions $y$ and $z$ satisfy (wherever $z \neq 0$)

$$\begin{cases} \Psi(y,z) = \left\{ya(t)\phi(y') - y\phi(\frac{y}{z})A(t)\phi(z')\right\}' = \\ = A(t)\zeta_\alpha(y,z) + \left[a(t) - A(t)\right]|y'|^{\alpha+1} + \left[Q(t) - q(t)\right]|y|^{\alpha+1} \end{cases} \tag{49.3}$$

and wherever $y \neq 0$

$$\begin{cases} \Psi(z,y) = \left\{zA(t)\phi(z') - z\phi(\frac{z}{y})a(t)\phi(y')\right\}' = \\ = a(t)\zeta_\alpha(z,y) - \left[a(t) - A(t)\right]|z'|^{\alpha+1} - \left[Q(t) - q(t)\right]|z|^{\alpha+1}. \end{cases} \tag{49.4}$$

**Theorem 1.** *Let $y$ and $z$ be two solutions of (49.2), oscillatory in some $\Omega_T$.*

*(1) (i) If $a \equiv A$ and $Q \equiv q$ in $\Omega_T$ then if two nodal sets $D(y)$ and $D(z)$ overlap, in $D(y) \bigcup D(z)$ either they coincide or between two consecutive zeros of $y$ lies a zero of $z$ and reversely, i.e.*

*(ii) if two such solutions are distinct, then one of the nodal sets is a translation of the other, i.e. $\exists \xi \in \mathbb{R} \setminus \{0\}; \quad y(t) = z(t+\xi)$.*

*(2) For the solutions of (49.2), assume that $a \geq A$ and $Q \geq q$ (but not both equal) in some $\Omega_T$. Then if $D(y)$ and $D(z)$ overlap, the diameter of $D(z)$ is smaller than that of $D(y)$.*

*Proof.* (1) (i) Assume that $D(y) := (u_1, u_2)$ and $D(z) := (v_1, v_2)$ are two overlapping nodal sets of the solutions with $a \equiv A$ and $Q \equiv q$ in $t > T$ . From (49.4), if $y > z$ in $D(z)$, then $0 = \int_{D(z)} \Psi(y,z)dt = \int_{D(z)} a(t)\zeta_\alpha(z,y)dt > 0$ which is absurd; $y$ has to have a zero inside $D(z)$. Similarly from (49.3) we get that $z$ has a zero inside $D(y)$.

(ii) Let $y$ and $z$ be two solutions with $z(s) = y(s) = 0$ for some $s > T$. If say, we assume that the next zero of $y$ is strictly smaller than that of $z$, then $z$ would not have a zero between the two consecutive zeros of $y$.

(2) As seen before, from (49.3) $z$ has a zero inside $D(y)$. We choose $z$ such that for some $\xi \in D(y), \quad z'(\xi) = y'(\xi) = 0$. Let $D(y) := (u_1, u_2)$. The integration over $(u_1, \xi)$ of (49.3) shows that $z$ has to have a zero in $(u_1, \xi)$ and its integration over $(\xi, u_2)$ shows that $z$ has to have a zero in $(\xi, u_2)$ . Therefore $D(z) \subset D(y)$.

## 49.3 Cases of Semilinear Equations with Perturbations

Now we consider equations of the types

$$\begin{cases} (i) \quad \Big\{a(t)\phi(y')\Big\}' + q(t)\phi(y) + Y = 0 \\ (ii) \quad \textit{and} \qquad \Big\{A(t)\phi(z')\Big\}' + Q(t)\phi(z) + Z = 0 \end{cases} \tag{49.5}$$

where $a, A, Q, q$ satisfy (H) , $Y$ and $Z$ being continuous functions in $\Omega_T \times \mathbb{R}$. Oscillation criteria for these equations are established in the literature (e.g. [Ta10, Ta11]).

As in (49.1) and (49.3), wherever $z \neq 0$ in $D(y)$,

$$\begin{cases} \Psi(y,z) = \Big\{ya(t)\phi(y') - y\phi(\frac{y}{z})A(t)\phi(z')\Big\}' = \\ = A(t)\zeta_\alpha(y,z) + \Big[a(t) - A(t)\Big]|y'|^{\alpha+1} + \Big[Q(t) - q(t)\Big]|y|^{\alpha+1} \\ +|y|^{\alpha+1}\Big\{\frac{Z}{\phi(z)} - \frac{Y}{\phi(y)}\Big\}. \end{cases} \tag{49.6}$$

**Theorem 2.** *Let $a, A, Q$ and $q$ be as in (H). Assume that*

$$\forall (t,\, S) \in \Omega_T \times \mathbb{R} \qquad SZ(t,S) \geq 0 \quad \textit{and} \quad SY(t,S) \leq 0,$$

*and*

$$a \geq A \quad \textit{and} \quad Q \geq q \quad (\textit{but not both equal }) \textit{ in} \quad \Omega_T.$$

*Then for two solutions $y$ and $z$ of (49.5) with overlapping nodal sets $D(y)$ and $D(z)$, the diameter of $D(z)$ is smaller than that of $D(y)$.*

*Proof.* Because under the hypotheses the second member of (49.6) is strictly positive on $D(y)$, the proof is similar to the last one.

## 49.4 Some Applications

In this section, we take $a \equiv A = 1$. From (49.4) of [Ta12], the generalized Sine equation (where $\phi := \phi_\alpha;\ \alpha \geq 1$) is

$$\Big(\phi_\alpha(u')\Big)' + \alpha\phi(u) = 0$$

whose solution is $S_\alpha(t)$ and it satisfies

$$S_\alpha(t+\pi_\alpha) = -S_\alpha(t); \quad \pi_\alpha = 2\pi\left[(\alpha+1)\sin\{\frac{\pi}{\alpha+1}\}\right]^{-1}.$$

For any $\lambda \in \mathbb{R}$, easy calculations lead to

$$\begin{cases} \Sigma(t) := S_\alpha(\lambda t) \Longrightarrow \Big\{\phi(\Sigma')\Big\}' + \lambda^{\alpha+1}\alpha\phi(\Sigma) = 0 \\ \text{and} \qquad \Sigma(t+\dfrac{\pi_\alpha}{\lambda}) = -\Sigma(t). \end{cases}$$

(see [JaKu99, Ta12]). Define for any $K > 0$ $U_K(t)$ to be a nontrivial oscillatory solution of

$$\Big(\phi(U')\Big)' + K^{\alpha+1}\alpha\phi(U) = 0$$

and $V$ that of

$$\Big(\phi(V')\Big)' + Q(t)\phi(V) = 0$$

in some $\Omega_T$.

From Theorem 2.1, assume that the continuous function $Q \in C(\Omega_T)$ satisfies for some $M_1 > M_0 > 0$

$$\begin{cases} \alpha M_0^{\alpha+1} < Q(t) < \alpha M_1^{\alpha+1} \quad \forall t \in \Omega_T \quad \text{then} \\ U_{M_1} \quad \text{has a zero in any nodal set of } V \text{ and} \\ V \quad \text{has a zero in any nodal set of } U_{M_0}. \end{cases}$$

We then have the following result:

**Theorem 3.** *Let $y$ be an oscillatory solution of*

$$\begin{cases} (i) \quad \left(\phi(y')\right)' + q(t)\phi(y) = 0; \quad \text{in } \Omega_T \\ (ii) \quad \text{with } \alpha M_0^{\alpha+1} < q(t) < \alpha M_1^{\alpha+1} \text{ in } \Omega_T. \end{cases} \tag{49.7}$$

*(a) Then if $x_1 < x_2$ are two consecutive zeros of $y$ in $(a, b) \subset \Omega_T$, then*

$$\frac{\pi_\alpha}{M_1} \leq (x_2 - x_1) \leq \frac{\pi_\alpha}{M_0}. \tag{49.8}$$

*Consequently if $q$ is unbounded above then the increasing sequence of consecutive roots of the solution satisfies*

$$\lim_{n\to\infty} (x_{n+1} - x_n) = 0. \tag{49.9}$$

*(b) The conclusions in a) still hold if (49.7)(i) is replaced by*

$$\left(\phi(y')\right)' + q(t)\phi(y) + F(t,y) = 0; \quad \text{in } \Omega_T$$

*provided that* $\forall S \in \mathbb{R}, \quad SF(t,S) \leq 0 \text{ in } \Omega_T.$

*Proof.* (a) From Theorem 2.1, the diameter of any nodal set of the solution of (49.7)(i) is smaller than that of an overlapping nodal set of $U_{M_0}$ and that of the nodal set of $U_{M_1}$ is smaller than that of an overlapping nodal set of the solution of (49.7)(i). This leads to (49.8). If $q(t)$ is unbounded above by setting $\alpha M_0^{\alpha+1}$ ( $M_1^{\alpha+1}$ ) $= \min_{[a,\,b]} q(t)$ ( $= \max_{[a,\,b]} q(t)$ ) , (49.9) is obtained.
(b) This is similarly obtained via the equation (49.3).

As an application to the last result, we have here some results related to the Bessel functions (see [GeKr07]).

Consider for $a, c, k, \alpha > 0$ and a parameter $p \in \mathbb{R}$ the equation

$$kt^2u''(t) + \left(ct^2 + a - \alpha p^2\right)u = 0 \quad \text{in } (0\,,\,\infty)$$

which can take the form

$$\begin{cases} u'' + q_p(t)u = 0, \quad t > 0; \\ q_p(t) := \dfrac{c}{k} + \dfrac{a - \alpha p^2}{kt^2}. \end{cases} \tag{49.10}$$

We know that if $\exists m, t_0 > 0$ such that $q_p(t) > m$ in $\Omega_{t_0}$, any nontrivial and bounded solution of (49.10) is oscillatory (see [Ta10]-[Ta12]).

Easily we observe that such an $m$ exists if

$$t^2 > \frac{\alpha p^2 - a}{c - km} \quad \text{and clearly} \quad \lim_{t \nearrow \infty} q_p(t) = \frac{c}{k}.$$

Also with

$$\begin{cases} (i) \quad M_p^+ \ (M_p^-) := \dfrac{c}{k} + \dfrac{a - \alpha p^2}{km^2} \quad \text{if} \quad \alpha p^2 > a \quad (\alpha p^2 < a), \\ (ii) \quad \dfrac{c}{k} \le q_p(t) < M_p^+ \quad \text{if} \quad \alpha p^2 < a \\ (iii) \quad \text{and } M_p^- < q_p(t) \le \dfrac{c}{k} \quad \text{if} \quad \alpha p^2 \ge a. \end{cases}$$

We then have the following result:

**Theorem 4.** *Let $a, c, k, \alpha > 0$ and a parameter $p \in \mathbb{R}$ be such that there are $m, \mu > 0$ satisfying*

$$\mu^2 > \frac{\alpha p^2 - a}{c - km}.$$

*Then any bounded and nontrivial solution of (49.10) in $\Omega_\mu$ is oscillatory. Let $\{x_k\}_{k \in \mathbb{N}}$ denote the increasing sequence of the zeros of such a solution.*

*(1) If $p^2 = \dfrac{a}{\alpha}$, then $\forall m \in \mathbb{N}, \quad x_{m+1} - x_m = \pi \sqrt{\left[\dfrac{k}{c}\right]}$.*

*(2) For all $m \in \mathbb{N}$, $\quad \pi \sqrt{\left[\dfrac{k}{c}\right]} < x_{m+1} - x_m \le \pi \left(\sqrt{M_p^-}\right)^{-1}$ if $p^2 > \dfrac{a}{\alpha}$.*

*(3) For all $m \in \mathbb{N}$, $\quad \pi \left(\sqrt{M_p^+}\right)^{-1} \le x_{m+1} - x_m < \pi \sqrt{\left[\dfrac{k}{c}\right]}$ if $p^2 < \dfrac{a}{\alpha}$.*

*(4) In any case, $\quad \lim_{m \to \infty} (x_{m+1} - x_m) = \pi \left(\sqrt{M_p^-}\right)^{-1}$.*

## References

[JaKu99] J. Jaros and T. Kusano: A Picone-type identity for second order half-linear differential equations. Acta Math. Univ. Comenianae 68 (1999), 137–151

[GeKr07] George F. Simons, Steven G. Krankz: Differential Equations, Theory, Technique and Practice. Mc Graw Hill Higher Education (2007)

[Ta10] Tadié: Oscillation criteria for semilinear elliptic equations with a damping term in $\mathbb{R}^n$. Electronic J. of Differential Equations, 2010, no. 51, 1–5.

[Ta11] Tadié: Oscillation criteria for damped Quasilinear second-order Elliptic Equations. Electronic J. of Differential Equations, 2011, No. 151, 1–11.

[Ta12] Tadié : On Strong Oscillation Criteria for Bounded Solutions for Some Quasilinear Second-Order Elliptic Equations. Communications in Mathematical Analysis, 13, No. 2, 15–26 (2012).

# Chapter 50
# Oscillation Criteria for some Third-Order Linear Ordinary Differential Equations

**Tadie**

## 50.1 Preliminaries

In this chapter, for $t > t_0 \geq 0$ we investigate in $t > t_0$ some oscillation criteria for problems of the type

$$\begin{cases} (i) \quad u'''(t) + c(t)u'(t) + h(t,u) = 0; \quad u(t_0) = u''(t_0) = 0; \\ (ii) \quad \text{where} \quad c \in C^1(\mathbb{R},\, (0,\, \infty)), \quad h \in C(\mathbb{R} \times \mathbb{R},\, \mathbb{R}) \\ (iii) \quad \text{with } h(.,\, 0) = 0 \text{ and } \forall s \in \mathbb{R} \setminus \{0\}, \quad sh(t,s) > 0. \end{cases} \tag{50.1}$$

Some similar but not quite related problems can be found in [Wo02] and [OuWo04]. For hypotheses, in the sequel

(H):

(h1) $h(t,S) = q(t)f(S)$ *where* $f \in C(\mathbb{R}), \quad Sf(S) > 0 \quad \forall S \neq 0$ *and* $f(0) = 0$;
(h2) *Eventually* $c, q \in C(\Omega_{t_0},\, (m,\, \infty))$ *for some* $m > 0,\, t_0 \geq 0$ *and* $\Omega_S := (S,\, \infty)$.

This work contains no non-oscillation results. The strategy used to get results here lies on the fact that the integration of (50.1)(i) over $(t_0,\, t)$ gives

$$\begin{cases} (i) \quad u''(t) + c(t)u(t) + H(t,u) = 0,\, t \geq t_0; \\ (ii) \quad \text{where} \quad H(t,u) := \int_{t_0}^{t} \Big( h(s,u) - c'(s)u(s) \Big) ds \end{cases} \tag{50.2}$$

Tadie (✉)
Universitet Copenhagen, Universitet Sparken 5, 2100 Copenhagen, Denmark
e-mail: tadietadie@yahoo.com



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_50

and some oscillation criteria for (50.2)(i) can be found, e.g., in [Ta12, Ta11]. Among those criteria we recall that:

**(C1)** If $\exists m, T > 0$ such that $c(t) > m$, then $y''(t) + c(t)y(t) = 0$ is oscillatory.
**(C2)** With $c$ being as in (C1), if a continuous function $h(t,u)$ is nonnegative in $\Omega_T$ then $y''(t) + c(t)y(t) + h(t,y) = 0$ is oscillatory.
**(C3)** With $c$ and $h$ as in (C1) and (C2), $\forall \phi \in C^1(\Omega_T)$,

$y''(t) + c(t)y(t) + \phi'(t)y'(t) + h(t,y) = 0$ is oscillatory.

## 50.2 Equations with Constant Coefficients

It is obvious that $\forall \omega > 0$ ,

$$y'''(t) + \omega y'(t) = 0 \quad in\ [0, \infty); \quad y(t_0) = y''(t_0) = 0 \tag{50.3}$$

has oscillatory solutions . Examples of such solutions are $\sin\sqrt{\omega}t$ and $\cos\sqrt{\omega}t$.

In fact, for any positive constants $\omega$, $T$, $A$, no regular solution of

$$\begin{cases} (i) \quad y'''(t) + \omega y'(t) = 0, \quad t > T \geq 0; \\ (ii) \quad 0 < y < A \quad \text{in } \Omega_T := (T, \infty); \\ (iii) \quad y(T) = y''(T) = 0 \end{cases}$$

can exist.

Any solution $y$ of (50.3) satisfies for $t \geq t_0$

$$\begin{cases} (i) \quad y''(t) + \omega y(t) = 0; \\ \text{after multiplying (i) by } y' \text{ , the integration over } (t_0,\ t) \text{ gives} \\ (ii) \quad y'(t)^2 + \omega y(t)^2 = y'(t_0)^2. \end{cases} \tag{50.4}$$

As a consequence, any non-trivial solution $y$ of (50.3) and its first derivative are bounded with $y'(t_0) \neq 0$. Also successive derivatives of $y$ give for $n \geq 3$

$$\begin{cases} y^{(n+1)} + \omega y^{(n-1)} = 0, \\ \{y^{(n)}(t)\}^2 + \omega\{y^{(n-1)}(t)\}^2 = \{y^{(n)}(t_0)\}^2 \\ \text{and} \quad y^{(n)}(t_0) \neq 0 \text{ for any non-trivial solution } y. \end{cases}$$

We then have the following result:

**Theorem 1.** *Given any $\omega > 0$ and $t_0 \in \mathbb{R}$, the problem*

$$y'''(t) + \omega y'(t) = 0, \quad t \geq t_0; \quad y(t_0) = y''(t_0) = 0 \tag{50.5}$$

*has an infinite number of oscillatory solutions; but if in addition $y'(t_0) \neq 0$ is prescribed, such a solution is unique.*

*Proof.* It is clear that if $z$ solves (50.5) so would do $\lambda z, \quad \forall \lambda \in \mathbb{R}$. But if in addition $z'(t_0)$ in prescribed, (50.4)(ii) shows that the corresponding solution would be unique. In fact if the condition $y'(t_0)$ is added to (50.5) and there are two solutions $z$ and $y$, the function $W := z - y$ satisfies $W''' + \omega W' = 0$; $W(t_0) = W'(t_0) = W''(t_0) = 0$ and as in (50.4)(ii), we get $W'(t)^2 + \omega W(t)^2 = 0$ and $W \equiv 0$ in $\Omega_{t_0}$.

Consider for some $t_0 \geq 0$ and $c, M > 0$ the problem

$$y'''(t) + cy'(t) + M = 0, \quad t > t_0; \tag{50.6}$$

$$y(t_0) = y''(t_0) = 0. \tag{50.7}$$

For any non-trivial solution $y \in C^3(\Omega_{t_0})$ of (50.6)-(50.7), after multiplying (50.6) by $y''$ and integrating the result over $(t_0, t)$ we get

$$\begin{cases} (i) \quad y''(t)^2 + cy'(t)^2 + 2My(t) = cy'(t_0)^2 \\ (ii) \quad \text{hence if } y > 0 \text{ in any } D \subset \Omega_{t_0} \\ y''(t)^2 + cy'(t)^2 \leq cy'(t_0)^2 \quad \text{and in } D \quad y'(t_0) \neq 0. \end{cases}$$

**Theorem 2.** *Any bounded and non-trivial solution $y$ of the problem (50.6)-(50.7) is oscillatory. This holds even if $M$ is any continuous and non-negative function. For such a solution, $y'(t_0) \neq 0$. Moreover, this holds also for $M < 0$ or if $M$ is replaced by a negative function.*

*Proof.* Let $y$ be any bounded and non-trivial solution of (50.6)-(50.7). The integration over $(t_0, t)$ of (50.6) gives $y''(t) + cy(t) + M(t - t_0) = 0, \quad t > t_0$ and any bounded and non-trivial solution of this equation is oscillatory by (C2).

If $M = -M_1 < 0$, the equation above reads $y''(t) + cy(t) - M_1(t - t_0) = 0, t > t_0$ and with an oscillatory solution of $z'' + cz = 0$, a Picone-type formula reads

$$\left(zz' - \frac{z^2 y'}{y}\right)' = \left\{z' - \frac{zy'}{y}\right\}^2 + \frac{z^2}{y} M_1(t - t_0) \tag{50.8}$$

and because the functions $y$ and $z$ are assumed bounded in $C^1(\Omega_{t_0})$, the second term cannot be zero for large $t$. So if we assume that $y \neq 0$ in some $\Omega_T$ say, the integration over any nodal set $D(z) \subset \Omega_T$ of (50.8) gives a contradiction. Therefore $y$ cannot remain nonzero in any $t > T > 0$. We note here that even if $y \to 0$ at $\infty$ the term $\dfrac{y'}{y} \not\to \infty$.

**Theorem 3.** *If the function* $\phi \in C(\mathbb{R})$ *satisfies*

$$\forall S \in \mathbb{R}, \quad S\phi(S) \geq 0 \qquad \text{or otherwise} \quad \exists \mu > 0, \quad |\phi(S)| > \mu,$$

*then the equation*

$$y'''(t) + cy'(t) + \phi(y) = 0,\ t > t_0; \qquad y''(t_0) = y(t_0) = 0$$

*is oscillatory.*

*Proof.* We proceed as before and similarly here, the oscillatory function $z$ being in use, (50.8) is

$$\left(zz' - \frac{z^2y'}{y}\right)' = \left\{z' - \frac{zy'}{y}\right\}^2 + \frac{z^2}{y}\int_{t_0}^{t} \phi(y(s))ds$$

leading to the same conclusions.

## 50.3 Equations with Variable Coefficients

In this section, we consider problems of the type

$$u'''(t) + c(t)u'(t) + q(t)f(u) = 0 \quad \text{in } \Omega_{t_0}; \quad u(t_0) = u''(t_0) = 0.$$

where $c,\ q, f$ are as displayed in (H).

**Lemma 1.** *Let* $a \geq 0$ *be a constant. Then if* $\exists m > 0, \quad c \in C(\Omega_a,\ (m,\ \infty))$ *, any bounded and non-trivial solution u of*

$$u'''(t) + c(t)u'(t) = 0, \quad t \in \Omega_a; \quad u(a) = u''(a) = 0 \tag{50.9}$$

*has an oscillatory second derivative* $u''$.

*Proof.* Let for some $b > 0 \quad v$, be an oscillatory solution of

$$v''' + bv' = 0 \quad t \in \Omega_a$$

as in the Theorem 1 and $u$ such a solution of (50.9). Then, proceeding as before,

$$\begin{cases} (i) & (v'''u'' - v''u''') + b(u''v' - u'v'') + (b-c)v''u' = 0 \quad \text{hense} \\ (ii) & \dfrac{u''}{v''}\left(\dfrac{v''}{u''}\right)' = c(t)\dfrac{u'}{u''} - b\dfrac{v'}{v''}. \end{cases} \tag{50.10}$$

Assume that there is such a solution $u$ such that for some $T > t_0$, $u'' > 0$ in $\Omega_T$. Let $[\alpha, \beta]$ be a nodal set of $v''$ inside which $v'' > 0$. For $s := \alpha + \dfrac{\beta - \alpha}{6}$ and $s < t < \beta$, from (50.10)(ii)

$$\left| \frac{v''(t)u''(s)}{u''(t)v''(s)} \right| = \exp\left[ -b \int_s^t \frac{v'}{v''} d\tau \right] \exp\left( \int_s^t c(\tau) \frac{u'}{u''} d\tau \right). \tag{50.11}$$

In $[s, \beta]$ the left-hand side of (50.11) is finite but as $\dfrac{v'}{v''} < 0$ near $\beta$ and tends to $-\infty$, the right-hand side is unbounded there. Therefore $u''$ has to be oscillatory.

Because $u''$ is oscillatory, $u'''$ is also oscillatory. Also as $c(.) > 0$ in $\Omega_T$, $u'$ has to be oscillatory as $u'''(t) = -c(t)u'(t)$. We have the following result:

**Theorem 4.** *For some $t_0 \geq 0$, consider the problem*

$$\begin{cases} (i) \quad u'''(t) + c(t)u'(t) = 0 \quad \text{in } \Omega_{t_0}; \quad u(t_0) = u''(t_0) = 0 \\ (ii) \quad \text{where} \quad c \in C^1(\Omega_{t_0}, (m, \infty)) \text{ for some } m > 0. \end{cases}$$

*Then any bounded and non-trivial solution of the problem has oscillatory first, second, and third derivatives. Moreover, if, in addition,*

$$c \text{ is monotone decreasing in some } \Omega_T \text{ or } c' \in L^1(\Omega_T) \tag{50.12}$$

*then any such a solution is oscillatory. Also if c is rather monotone increasing there, the same conclusion holds unless for that solution u,*

$$\liminf_{t \nearrow \infty} |u(t)| = 0. \tag{50.13}$$

*Proof.* For any such a solution $u$, that $u'$, $u''$ and $u'''$ are oscillatory is established in Lemma 1. The integration over $(t_0, t)$ of (50.9) gives

$$u''(t) + c(t)u(t) + F(t,u) = 0, \qquad \text{with } F(t,u) := -\int_{t_0}^t c'(s)u(s)ds.$$

With (50.12),

$$v''(t) + c(t)v = 0$$

is oscillatory (see [Ta14], Theorem 1.3). From these two equations, a version of Picone formula reads

$$\left( vv' - \frac{v^2}{u}u' \right)' = \left[ v' - \frac{v}{u}u' \right]^2 + \frac{v^2}{u}F(t,u).$$

If we assume that $u$ is not oscillatory, i.e. $u \neq 0$ in some $\Omega_T$, $\quad T \geq t_0$, the right-hand side of (50.13) is strictly positive as $uF(t,u) \geq 0$ if $c$ is decreasing there. In that case, integrating (50.13) over any nodal set $D(v) \subset \Omega_T$ of $v$ would lead to a contradiction (because the left-hand side would be 0); $u$ has to be oscillatory as it has to have a zero in any $D(v) \subset \Omega_T$. We reach the same conclusion if $c' \in L^1(\Omega_T)$; in fact we have for bounded $v$ and $u > \beta > 0$ $|\frac{v(t)}{u(t)} \int_{t_0}^t c'(s)u(s)ds| \leq \{|\frac{v(t)}{u(t)} \int_{t_0}^T c'(s)u(s)ds| + |\frac{v(t)}{u(t)} \int_T^t c'(s)u(s)ds|\}$ and the right-hand side is bounded above by a constant, $B > 0$, say. If the oscillatory function $v$ was taken to be a solution of

$$v''(t) + c(t)v - kB = 0$$

which is oscillatory for large $k > 0$, a Picone-type formula would read

$$\left(vv' - \frac{v^2}{u}u'\right)' = \left[v' - \frac{v}{u}u'\right]^2 + v(t)\left(kB - \frac{v}{u}\int_{t_0}^t c'(s)u(s)ds\right).$$

and the right-hand side is strictly positive, leading to the conclusion.

If $c$ is increasing in $\Omega_T$ and $\liminf_{t\to\infty} |u(t)| > \nu > 0$ there, then $t \mapsto |F(t,u)|$ is unbounded and the integration of (3.8) over $D(v)$ leads to a contradiction (as the right-hand side would be unbounded). So, any such a solution satisfying (50.12) and $\liminf_{t\to\infty} |u(t)| > 0$ would be oscillatory.

**Theorem 5.** *Assume that*

$$c,\ q \text{ and } f \text{ satisfy (H), with a monotone } c.$$

*Then any bounded and non-trivial solution $u$ of*

$$u'''(t) + c(t)u'(t) + q(t)f(u) = 0 \quad \text{in } \Omega_{t_0}; \quad u(t_0) = u''(t_0) = 0. \tag{50.14}$$

**(a)** *is oscillatory if $c \in C^1(\Omega_{t_0})$ and $c' \leq 0$ in some $\Omega_R$;*
**(b)** *is oscillatory even when $c' > 0$ in some $\Omega_T$ unless*

$$\liminf_{t\to\infty} |u(t)| = 0.$$

*Proof.* Let $u$ be a bounded non-trivial solution of (50.14). Then the integration over $(t_0,\ t)$ of the equation in (50.14) gives

$$\begin{cases} (i) \quad u''(t) + c(t)u(t) + G(t,u) = 0, \quad t \in \Omega_{t_0}; \text{where} \\ (ii) \quad G(t,u) := \int_{t_0}^t \left(q(s)f(u(s)) - c'(s)u(s)\right)ds. \end{cases}$$

Let $M \in C(\Omega_{t_0}, (k, \infty))$ for some large $k > 0$ and $v$ an oscillatory solution of

$$v'' + c(t)v - M(t) = 0, \ t \geq t_0.$$

Then $u$ being a bounded solution for (50.14), if we suppose that $u \neq 0$ in some $\Omega_a$ then for $t > s > a \geq t_0$

$$\begin{cases} \left(vv' - \dfrac{v^2}{u}u'\right)' = \left[v' - \dfrac{v}{u}u'\right]^2 + \dfrac{v^2}{u}G(t,u) \\ = \left[v' - \dfrac{v}{u}u'\right]^2 + \dfrac{v^2}{u}\int_s^t q(\tau)f(u(\tau))d\tau + v(t)M(t) \\ -\dfrac{v^2}{u}\int_s^t c'(\tau)u(\tau)d\tau. \end{cases} \tag{50.15}$$

If we suppose that $u > 0$ in some $\Omega_T$ then the right side of (50.15) is strictly positive if $c' \leq 0$, leading to a contradiction.

Assume that $c' > 0$ eventually, in $\Omega_T$, say. If $u > \mu > 0$ in $\Omega_T$, then for some $\beta > 0$ and in any $D(v^+) \subset \Omega_T$ we have $\dfrac{v}{u}\int_s^t c'(\tau)u(\tau)d\tau > \dfrac{v}{u}\mu(\, c(t) - c/T)\,) > \beta[c(t) - c(T)]$.

We choose $M$ such that $M(t) > \beta[c(t) - c(T)]$ for $t > T$; this makes the right-hand side of (50.15) strictly positive and the integration over any $D(v) \subset \Omega_T$ of (50.15) would lead to a contradiction; hence, $u$ has to have a zero in any such $D(v)$.

## References


[Wo02] J.S.W. Wong : A nonoscillation theorem for Emden–Fowler equations. J.Math. Anal.Appl. **274** (2002), 746–754

[OuWo04] C.H. Ou and J.S.W. Wong: Oscillation and non-oscillation theorems for superlinear Emden-Fowler equations of the fourth order. Annali di Matematica **183**, 25–43 (2004)

[Ta14] Tadié: Semilinear second-order ordinary differential equations: distances between consecutive zeros of oscillatory solutions (in this volume)

[Ta12] Tadié: On strong oscillation criteria for bounded solutions for some quasilinear second-order elliptic equations. Communications in Mathematical Analysis vol. 13 No. 2, 15–26 (2012)

[Ta11] Tadié: Oscillation criteria for damped quasilinear second-order elliptic equations. Electronic J. of Differential Equations, 2011, no. 151, 1–11.

# Chapter 51
# Oscillation Criteria for some Semi-Linear Emden–Fowler ODE

**Tadie**

## 51.1 Preliminaries

Inspired from earlier works on oscillation criteria for semi-linear elliptic equations, we pinpoint here some straightforward and easy oscillation criteria for Emden–Fowler differential equations. We find out that for $\alpha \geq 0$, the equation

$$[|y'|^{\alpha-1}y']' + f(t,y) = 0$$

is oscillatory if for some $m, T > 0$ and $\beta \in [1, \alpha]$ $\exists q \in C([T, \infty), (m, \infty))$ such that

$$\forall t > T \text{ and } \forall s \in \mathbb{R}, \qquad f(t,s) \geq q(t)|s|^{\beta-1}s.$$

The main tools for our investigation are some version of Picone identities and comparison methods. We are considering equations of the type

$$\begin{cases} (i) & \left\{\phi(y')\right\}' + \Psi(t,y,y') = 0 \\ (ii) & \text{where } \forall S \in \mathbb{R} \text{ and some } \alpha \geq 0 \quad \phi(S) := \phi_\alpha(S) = |S|^{\alpha-1}S; \\ (iii) & \Psi \in C(\mathbb{R}^3, \mathbb{R}). \end{cases}$$

Tadie (✉)
Universitet Copenhagen, Universitet Sparken 5, 2100 Copenhagen, Denmark
e-mail: tadietadie@yahoo.com



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_51

Usually equations in these contexts have the form

$$\left\{a(t)\phi(y')\right\}' + \Psi(t,y,y') = 0$$

where for some $t_0 \geq 0, \quad a \in C^1([t_0, \infty))$ is strictly positive with $a' \geq 0$. Because of these conditions on $a$, in regard of oscillatory character, that equation is equivalent to

$$\left\{\phi(y')\right\}' + \frac{a'(t)}{a(t)}\phi(y') + \frac{\Psi(t,y,y')}{a(t)} = 0.$$

This is the reason why we take $a(t) \equiv 1$ in our study and extend the investigation to the equations with damping terms, $\phi(y')$, say. We set the following hypotheses :

**(H)**: the function $\Psi$ has the form

(H1) $\Psi(t,u,u') := f(t,u)$ where $\forall t \in \mathbb{R}$ and $u \neq 0, \quad uf(t,u) > 0$;
(H2) $\Psi(t,u,u') := g(t,u') + f(t,u)$ whith $f$ as in (H1) and $g \in C(\mathbb{R}^2, \mathbb{R})$.

It is worth recalling the following:

**Definition 1.** With $\Omega_T := (T, \infty) \quad$ where $T \geq 0$

(1) A function $v$ will be said to be oscillatory if $\forall R > 0, \quad v$ has zero in $\Omega_R$.
(2) A function $v$ will be said to be strongly oscillatory if $\forall R > 0, \quad v$ has a nodal set in $\Omega_R$ where a nodal set of $v$ is here any interval $D(v) := (t_1, t_2)$ such that $v \neq 0$ in $D(v)$ and $v(t_1) = v(t_2) = 0$.
(3) An equation will be said to be oscillatory (strongly oscillatory) if any bounded (bounded in $C^1$) and non-trivial solution of the equation is oscillatory (respectively, strongly oscillatory).
(4) The function $\phi_\gamma$ satisfies $\quad S\phi_\gamma(S) = |S|^{\gamma+1}$ and $S\phi'_\gamma(S) = \gamma\phi_\gamma(S)$.

Here, a solution of the equation in $\Omega_R$ will be an element of $C^1(\overline{\Omega_R}) \bigcap C^2(\Omega_R)$ which satisfies the equation. Also we will often mention that a solution $w$ is bounded in $E$, say, if it is bounded in $C^1(E)$ i.e., there is $M > 0$ such that $|w|_{C^1(E)} := \max_{\overline{E}} \{|w(t)|, |w'(t)|\} < M$.

For our investigations, we use similar approaches as those in [Ta09a, Ta09b, Ta10, Ta07], based on Picone-type identities. In [WaXi00], a similar equation is considered, with $\Psi(t,y) = q(t)f(y)$ with integrable $q$. Other results on the field can be found in [OuWo04, Pa11] for the fourth-order equations.

## 51.2 Equations Without Damping Terms

Define for any $\gamma \geq 1$ and $u, v \in C^1(\mathbb{R}, \mathbb{R})$

$$\zeta_\gamma(u,v) = |u'|^{\gamma+1} - (\gamma+1)u'\phi(\frac{u}{v}v') + \gamma|\frac{u}{v}v'|^{\gamma+1} \qquad (Z),$$

which is strictly positive if $u$ and $v$ are distinct and non-constant and zero only if $u = \lambda v$ for some $\lambda \in \mathbb{R}$. We consider here equations of the type

$$\left\{\phi_\alpha(y')\right\}' + f(t,y) = 0; \quad \alpha \geq 0 \tag{51.1}$$

where $f$ satisfies (H1). We recall here that if a continuous $F(s,v) \geq 0$ and $q \in C(\Omega_T,\ (m,\ \infty))$ for some $T, m > 0$ then $\quad \forall \alpha \geq 0$

$$\left\{\phi_\alpha(y')\right\}' + q(t)\phi_\alpha(y) + F(t,y) = 0$$

is strongly oscillatory ( see [Ta09b] ). This remains true if $F < 0$ and large enough. We then have the following result:

**Theorem 1.** *Assume that for some* $m, T > 0$ *and* $\phi = \phi_\alpha$

$$\exists q \in C(\Omega_T,\ (m,\ \infty)); \quad \left(\frac{f(t,s)}{\phi(s)} - q(t)\right) \geq 0 \text{ in } \Omega_T \times \mathbb{R}. \tag{51.2}$$

*Then* (51.1) *is strongly oscillatory.*

*In particular, if*

$$(*) \qquad \exists k > 0 \quad \text{such that} \quad \forall s \neq 0 \qquad \frac{f(t,s)}{\phi(s)} > k$$

*then (51.1) is oscillatory.*

*Proof.* If $y$ is a non-trivial solution of (51.1) and (51.2) holds, let $z$ be an oscillatory solution of $\{\phi(z')\}' + q(t)\phi(z) = 0 \quad$ in $\Omega_T$.

A version of Picone's identity reads for those functions whenever $y \neq 0$

$$\left\{z\phi(z') - z\phi\left(\frac{z}{y}y'\right)\right\}' = \zeta_\alpha(z,y) + |z|^{\alpha+1}\left[\frac{f(t,y)}{\phi(y)} - q(t)\right]. \tag{51.3}$$

The integration of both sides of (51.3) over any $D(z) \subset \Omega_T$ leads to an absurdity because the left side would be zero and the right strictly positive by (51.2). Therefore $y$ has to have a zero in any nodal set $D(z) \subset \Omega_T$ ; $y$ cannot be non-zero in any $\Omega_R$.

The proof is completed by the fact that when $(*)$ holds, using instead $\{\phi(z')\}' + k\phi(z) = 0$ in $\Omega_T$ , (51.3) becomes

$$\left\{z\phi(z') - z\phi\left(\frac{z}{y}y'\right)\right\}' = \zeta_\alpha(z,y) + |z|^{\alpha+1}\left(\frac{f(t,y)}{\phi(y)} - k\right)$$

whose right-hand side is strictly positive.

**Corollary 1.** *Consider the equation*

$$\begin{cases} (i) & \left\{\phi_\alpha(y')\right\}' + f(t,y) + h(t,y) = 0 \\ (ii) & \text{where } f \text{ satisfies (H1) and} \\ (iii) & h \in C(\mathbb{R}^2, \mathbb{R}) \quad \text{satisfies } sh(t,s) \geq 0 \text{ in } \Omega_T \times \mathbb{R}. \end{cases} \tag{51.4}$$

*Then under the condition* (51.2)*, the equation* (51.4)*(i) is strongly oscillatory.*

*Proof.* We consider $z$, an oscillatory non-trivial solution of $\phi(z')' + q(t)\phi(z) = 0$. The proof follows from the fact that for (51.4)(i), (51.3) becomes

$$\left\{z\phi(z') - z\phi\left(\frac{z}{y}y'\right)\right\}' = \zeta_\alpha(z,y) + |z|^{\alpha+1}\left[\frac{f(t,y)}{\phi(y)} - q(t) + \frac{h(t,y)}{\phi(y)}\right].$$

**Lemma 1.** *Assume that the constants* $\alpha, \beta \geq 1$ *and* $q \in C(\Omega_T, (m, \infty))$ *for some* $T, m > 0$*. Then any bounded and non-trivial solution of*

$$\phi_\alpha(y')' + q(t)\phi_\beta(y) = 0 \quad \text{is oscillatory if} \quad \alpha \geq \beta. \tag{51.5}$$

*Proof.* For a bounded and non-trivial solution $v$ of (51.5) in $\Omega_T$, say, define

$$y(t) := \begin{cases} (i) & \dfrac{v(t)}{|v|_\infty} \quad \text{if } |v|_\infty > 1 \\ (ii) & v(t) \quad \text{if } |v|_\infty \leq 1. \end{cases}$$

Then $y$ satisfies for $\phi := \phi_\alpha$

$$\textbf{(**)} \quad \begin{cases} |y(t)| \leq 1 \quad \text{and} \quad \phi(y')' + Q(t)\phi_\beta(y) = 0 \quad \text{in } \Omega_T \\ \text{where } Q(t) := [|v|_\infty]^{\beta-\alpha} q(t) \text{ for (i) or } q(t) \text{ for (ii).} \end{cases}$$

Let $z$ be a non-trivial, bounded, strongly oscillatory solution of $\phi(z')' + Q(t)\phi(z) = 0$ in $\Omega_T$. Then if $y \neq 0$ in any $D(z)$ ,

$$\left(z\phi(z') - z\phi\left(\frac{z}{y}y'\right)\right)' = \zeta_\alpha(z,y) + |z|^{\alpha+1}Q(t)\left\{|y|^{\beta-\alpha} - 1\right\},$$

which would be strictly positive there as $|y| \leq 1$ there and $\alpha \geq \beta$. Therefore $y \neq 0$ cannot hold in any $D(z)$.

The last results lead to the following:

**Theorem 2.** *Assume that for some* $T, m > 0$, *if* $f \in C(\Omega_T \times \mathbb{R},\ \mathbb{R})$ *satisfies* $sf(t,s) > 0 \quad \forall s \neq 0$ *and* $\forall t > T$

$$\left( \frac{f(t,s)}{\phi_\beta(s)} - q(t) \right) \geq 0 \quad \text{for some } \beta \geq 0 \text{ and } q \in C(\Omega_T,\ (m,\ \infty)).$$

*Then* $\forall \alpha \geq \beta$ *any non-trivial and bounded solution of*

$$\left\{ \phi_\alpha(u') \right\}' + f(t,u) = 0, \quad t > T \tag{51.6}$$

*is oscillatory.*

For the proof of the theorem we need the next lemma.

An equation is said to be homogenous when whenever $u$ is its solution , so is $\lambda u, \quad \forall \lambda \in \mathbb{R}$. In such a case, any of its non-trivial solution $v$ can freely be supposed to be normal meaning here that $|v|_{C^1} \leq 1$.

**Lemma 2.** *Let* $v \in C^2(\Omega_T)$ *be a locally bounded solution of the nonhomogenous equation*

$$\phi(v')' + f(t,v) = 0, \quad t > T.$$

*Then if* $J \subset \Omega_T$ *is bounded*

$$\begin{cases} \exists \lambda := \lambda_J > 0 \quad \text{and } y \in C^2(\bar{J}) \text{ with } |y|_{C^1(J)} \leq 1 \text{ solution of} \\ \phi(y')' + f_\lambda(t,y) = 0, \quad t \in J; \quad \text{where } f_\lambda(t,s) := \lambda^{-\alpha} f(t, \lambda y); \end{cases}$$

$\lambda_J := |v|_{C^1(\bar{J})}$ *and* $y$ *will be called the normalized solution of the equation in* $J$.

*Proof.* The function

$$y(t) := \begin{cases} \dfrac{v(t)}{\lambda} & \text{if } \lambda > 1 \\ v(t) & \text{if } \lambda \leq 1 \end{cases}$$

satisfies in $J$

$$|y(t)|,\ |y'(t)| \in [0,\ 1] \quad \text{and} \quad \left\{ \phi(y') \right\}' + \lambda^{-\alpha} f(t, \lambda y) = 0.$$

It is obvious that if $v$ is bounded in the whole $\Omega_T$, we take $\lambda := |v|_{C^1(\Omega_T)}$.

*Proof of Theorem 2*. We take $J := \Omega_T$ and let $v \in C^2(\Omega_T)$ be a bounded solution of (51.6), with $\lambda := |v|_{C^1(\Omega_T)}$ and $y$ its corresponding normalized solution in $\Omega_T$. Let $z$ be a non-trivial oscillatory solution of

$$\left\{\phi_\beta(z')\right\}' + \lambda^{-\alpha} q(t)\phi_\beta(z) = 0 \quad t > T.$$

If $y \neq 0$ in $\Omega_T$, then

$$\begin{aligned}
&\left\{z\phi_\beta(z') - z\phi_\beta(\frac{z}{y})\phi(y')\right\}' \\
&= |z'|^{\beta+1} - \mu^\alpha q(t)|z|^{\beta+1} - (1+\beta)z'\phi_\beta(\frac{z}{y}y')\frac{\phi(y')}{\phi_\beta(y')} \\
&\quad + \beta\frac{z}{y}y'\phi_\beta(\frac{z}{y}y')\frac{\phi(y')}{\phi_\beta(y')} + \lambda^\beta\mu^\alpha|z|^{\alpha+1}\frac{f(t,\lambda y)}{\phi_\beta(\lambda y)} - \mu^\alpha q(t)|z|^{\beta+1} \\
&= |z'|^{\beta+1}\left[1 - |y'|^{\alpha-\beta}\right] + |y'|^{\alpha-\beta}\zeta_\beta(z,y) \\
&\quad + |z|^{\beta+1}\mu^\alpha\left\{\lambda^\beta\frac{f(t,\lambda y)}{\phi_\beta(\lambda y)} - q(t)\right\} > 0
\end{aligned} \tag{51.7}$$

if $\alpha \geq \beta$ , $|y'| \leq 1$ and $\dfrac{f(t,\lambda y)}{\phi_\beta(\lambda y)} - q(t) \geq 0$ in $\Omega_T$. The integration of (51.7) over any $D(z) \subset \Omega_T$ leads then to a contradiction whence $y$ has to have a zero in any such a $D(z)$.

## 51.3 Problems with Damping Terms

Consider now in some $\Omega_T$, $m > 0$ and $b \in C(\Omega_T)$ the equation

$$\begin{cases}
(i) \quad \left\{\phi(y')\right\}' + b(t)\phi(y') + f(t,y) = 0 \\
(ii) \quad \text{where } f \text{ satisfies (H1) and in } \Omega_T \times \mathbb{R} \\
\qquad \exists q \in C(\Omega_T,\ (m,\ \infty)); \quad \left(\dfrac{f(t,y)}{\phi(y)} - q(t)\right) \geq 0; \\
(iii) \quad \exists k \in C(\Omega_T) \text{ bounded and } B \in C^1(\Omega_T) \text{ such that} \\
\quad B'(t) = b(t) + k(t).
\end{cases} \tag{51.8}$$

**Theorem 3.** *Under the conditions (ii) and (iii) above, bounded and non-trivial solutions of* (51.8)*(i) are oscillatory*

*(1) if* $k \equiv 0$*;*
*(2) if* $k \not\equiv 0$ *but bounded, unless* $\liminf_{t\nearrow\infty} |y(t)| = 0$.

*Proof.* Let for some large positive continuous function $M$ $z$ be a strongly oscillatory solution of

$$\phi(z')' + q(t)\phi(z) - M(t) = 0;\ t > T$$

and $y$ a bounded and non-trivial solution of (51.8)(i). A version of Picone identity for the two solutions reads whenever $y \neq 0$

$$\begin{cases} \left\{ z\phi(z') - z\phi(\frac{z}{y}y') - B(t)z\phi(\frac{z}{y}y') \right\}' \\ = \zeta_\alpha(z,y) + |z|^{\alpha+1}\left[\frac{f(t,y)}{\phi(y)} - q(t)\right] \\ +z\left(M(t) - k(t)\phi(\frac{z}{y}y')\right) - B(t)\left(z\phi(\frac{z}{y}y')\right)'. \end{cases} \tag{51.9}$$

Let $D(z^+) \subset \Omega_T$ be a nodal set of $z^+$. If $y \neq 0$ in $D(z^+)$, then the integration over $D(z) := D(z^+)$ of (51.9) gives $0 = \int_{D(z)} \left[\zeta_\alpha(z,y) + |z|^{\alpha+1}\left[\frac{f(t,y)}{\phi(y)} - q(t)\right] + z\left(M(t) - k(t)\phi(\frac{z}{y}y')\right)\right]dt - \int_{D(z)} B(t)\left(z\phi(\frac{zy'}{y})\right)'dt$ which holds even if $B$ is replaced by $B_1(t) := B(t) + \lambda$ for any $\lambda \in \mathbb{R}$, i.e.

$$\begin{cases} \forall \lambda \in \mathbb{R} \quad 0 = \int_{D(z)} \left[\zeta_\alpha(z,y) + |z|^{\alpha+1}\left[\frac{f(t,y)}{\phi(y)} - q(t)\right]\right. \\ \left. +z\left(M(t) - k(t)\phi(\frac{z}{y}y')\right)\right]dt - \int_{D(z)} \{B(t) + \lambda\}\left(z\phi(\frac{z}{y}y')\right)'dt. \end{cases} \tag{51.10}$$

That can hold only if each integrand is null in $D(z)$. But the first integrand which is $\left\{\zeta_\alpha(z,y) + |z|^{\alpha+1}\left[\frac{f(s,y)}{\phi(y)} - q(s)\right] + z\left(M(t) - k(t)\phi(\frac{z}{y}y')\right)\right\}$ is strictly positive if

(1) $k \equiv 0$ (even if $M \equiv 0$ ) thus in this case $y$ has to have a zero in any $D(z^+) \subset \Omega_T$;
(2) if $y > \nu > 0$ in $\Omega_T$ and $k$ bounded, $M$ can be chosen such that $M(t) > |k(t)\phi(\frac{z}{y}y')|$ in $\Omega_R$ for some $R \geq T$ and (51.10) would not hold. Therefore (51.10) would hold only if $y > \nu > 0$ in $\Omega_R$ fails for some $R \geq T$.

Consider now in some $\Omega_T$, $m>0$ and $b\in C(\Omega_T)$ the equation

$$\begin{cases} (i) \quad \left\{\phi(y')\right\}' + b(t)\phi(y') + f(t,y) = 0 \\ (ii) \quad \text{where for some } \beta \geq 0 \quad f \text{ satisfies (H1) and} \\ \quad \exists q \in C(\Omega_T,\ (m,\ \infty)); \quad \left(\dfrac{f(t,y)}{\phi_\beta(y)} - q(t)\right) \geq 0 \text{ in } \Omega_T \times \mathbb{R}\ ; \\ (iii) \quad \exists B \in C^1(\Omega_T); \quad B'(t) = b(t) \text{ in } \Omega_T. \end{cases} \tag{51.11}$$

To proceed as for the proof of Theorem 2, we skip the details and will consider a non-trivial and bounded solution $y$ of (51.11)(i) satisfying $|y(t)|, |y'(t)| \leq 1$ in $\Omega_T$ which contains few nodal sets $D(z)$ where $z$ is a strong oscillatory solution of

$$\begin{cases} \phi_\beta(z')' + q(t)\phi_\beta(z) = 0 \quad t \geq T \\ \text{with } |z(t)|,\ |z'(t)| \leq 1 \quad \text{in } \Omega_T. \end{cases} \tag{51.12}$$

**Theorem 4.** *Under the conditions (51.11)(i)–(51.11)(iii), where $\phi := \phi_\alpha$; $\alpha \geq 0$, any non-trivial and locally bounded solution of (51.11)(i) is oscillatory if $\alpha \geq \beta$.*

*Proof.* Without loss of generality, let $y$ be a non-trivial and bounded solution of (51.11)(i) in $\Omega_T$ $|y|_{C^1(\bar{J})} \leq 1$. We assume that $z$, a strongly oscillatory solution in (51.12). If $y \neq 0$ in any $D(z) \subset \Omega_T$

$$\begin{cases} \left\{z\phi_\beta(z') - z\phi_\beta(\frac{z}{y})\phi(y') - B(t)z\phi_\beta(\frac{z}{y})\phi(y')\right\}' \\ = |z'|^{\beta+1} - |Z|^{\beta+1}q(t) - (\beta+1)z'\phi_\beta(\frac{z}{y}y')\dfrac{\phi(y')}{\phi_\beta(y')} + \beta y'\dfrac{z}{y}\phi_\beta(\frac{z}{y}y')\dfrac{phi(y')}{\phi_\beta(y')} \\ +|z|^{\beta+1}\left(b(t)\dfrac{\phi(y')}{\phi_\beta(y')} + \dfrac{f(t,y)}{\phi_\beta(y)}\right) - b(t)z\phi_\beta(\frac{z}{y})\phi(y') - B(t)\left(z\phi_\beta(\frac{z}{y})\phi(y')\right)' \\ = |z'|^{\beta+1}\left[1 - |y'|^{\alpha-\beta}\right] + |y'|^{\alpha-\beta}\zeta_\beta(z,y) + |z|^{\beta+1}\left\{\dfrac{f(t,y)}{\phi_\beta(y)} - q(t)\right\} \\ -B(t)\left(z\phi_\beta(\frac{z}{y})\phi(y')\right)'. \end{cases}$$

As $|y'| \leq 1$ in $J$, we have $|z'|^{\beta+1}\left[1 - |y'|^{\alpha-\beta}\right] + |y'|^{\alpha-\beta}\zeta_\beta(z,y) + |z|^{\beta+1}\left\{\dfrac{f(t,y)}{\phi_\beta(y)} - q(t)\right\} > 0$ in $D(z)$ and the proof is completed as that of Theorem 2; i.e., $y$ has to have a zero in any such a $D(z)$.

## References

[OuWo04] C.H. Ou, J.S.W. Wong: Oscillation and non-oscillation theorems for superlinear Emden–Fowler equations of the fourth order. Annali di Matematica **183**, 25–43 (2004)

[Pa11] P. Pietramala: A Note on a Beam Equation with Nonlinear Boundary Conditions. Hindawi Publ. Corporation, Boundary Value Problems, 2011, article ID 376782.

[Ta09a] Tadié: Comparison results for quasilinear elliptic equations via Picone-type identity. Part II: cases where the coefficient of the principal part depends on the unknown function. J. Nonlinear Anal., T.M.A., 71 (2009), e601-e606.

[Ta09b] Tadié: Comparison results for semilinear elliptic equations via Picone-type identities. Electronic J. of Differential Equations 2009, no. 67, 1–7.

[Ta10] Tadié: Oscillation criteria for semilinear elliptic equations with a damping term in $\mathbb{R}^n$. Electronic J. of Differential Equations, 2010, no. 51, 1–5.

[Ta07] Tadié: Sturmian comparison results for quasilinear elliptic equations in $\mathbb{R}^n$, Electronic J. of Differential Equations, 2007, no. 26, 1–8.

[WaXi00] Wan-Tong Li and Xiaohu Li: Oscillation criteria for second-order nonlinear differential equations with integrable coefficient. Applied Mathematics Letters 13 (2000), 1–6.

# Chapter 52
# Analytic Representation of the Solution of Neutron Kinetic Transport Equation in Slab-Geometry Discrete Ordinates Formulation

**F.K. Tomaschewski, C.F. Segatto, R.C. Barros, and M.T.B. Vilhena**

## 52.1 Introduction

Presented in this chapter is an analytical representation for the solution of slab-geometry neutron kinetics equations in one-speed discrete ordinates ($S_N$) transport formulation with one group of delayed neutron precursors. The basic idea involves the following steps: (i) the neutron angular flux and the concentration of delayed neutron precursors are expanded in truncated series of unknown functions, (ii) by substituting these expansion representations into the $S_N$ kinetics equations, a set of recursive systems of first-order ordinary differential equations results. It is assumed that the first equation of the system has no source and is the only one which satisfies the initial conditions. The remaining equations of the recursive systems satisfy initial condition equal to zero and the source term is written in terms of the previous step solution. At each step of the recursive system, the $S_N$ kinetics equations are solved by using the $TLTS_N$ method, whose essence is to apply the double Laplace transform technique [ToSeVi13, To12]. To achieve this goal, the Laplace transformation in the time variable is first applied, and then, the resulting equation is solved by the conventional $LTS_N$ method, which consists of the application of the Laplace transformation to the $S_N$ equations, yielding a system of $N$ linear algebraic equations in the complex parameter 's' [Ba92, BaVi97, OrViCa04]. This system is then solved for the transformed angular flux and the inverse transformation is performed analytically, thus obtaining an expression for the angular flux of particles migrating in the N discrete directions of the $S_N$ model. The $LTS_N$ method is

F.K. Tomaschewski (✉) • C.F. Segatto • M.T.B. Vilhena
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: fernandasls_89@hotmail.com; cynthia.segatto@ufrgs.br; marco.vilhena@ufrgs.br

R.C. Barros
State University of Rio de Janeiro, Nova Friburgo, RJ, Brazil
e-mail: rcbarros@pq.cnpq.br



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_52

completely free from all spatial truncation errors. By this procedure, the solution is written in terms of a line integral in the time variable, which, in the present method, is evaluated by the Gaver–Stehfest numerical scheme [Ga66, St70, St70a]. In the next section, this decomposition method is described. In section 52.3, the TLTS$_N$ method for the solution of the recursive system is presented, and in Section 52.4 numerical results for two model problems are given. Concluding this chapter, section 52.5 offers a brief discussion of the results with suggestions for future work.

## 52.2 Decomposition Method

The general formulation of one-speed, slab-geometry S$_N$ kinetics problems with one group of delayed neutron precursors and isotropic scattering are

$$\begin{cases} \dfrac{1}{v}\dfrac{\partial \psi_m}{\partial t} + \mu_m \dfrac{\partial \psi_m}{\partial x} + \sigma_t \psi_m - \dfrac{\sigma_s}{2}\phi = (1-\beta)\dfrac{\nu\sigma_f}{2}\phi + \lambda C, \quad m = 1:N. \\ \dfrac{\partial C}{\partial t} = \dfrac{\beta\nu\sigma_f}{2}\phi - \lambda C \end{cases} \tag{52.1}$$

with appropriate initial conditions

$$\begin{aligned} \psi_m(x,0) &= \phi_0(x), \quad m = 1:N \\ C(x,0) &= \beta\frac{\nu\sigma_f}{2\lambda}\phi_0(x) \end{aligned} \tag{52.2}$$

and boundary condition

$$\begin{aligned} \psi_m(0,t) &= f_m, \qquad \mu_m > 0, \\ \psi_m(L,t) &= g_m, \qquad \mu_m < 0. \end{aligned} \tag{52.3}$$

Here, $m = 1:N$, where $N$ is the order of the angular quadrature set, $f_m$ and $g_m$ are prescribed incoming angular fluxes, which are assumed to be independent of time; $\psi_m(x,t)$ is the angular flux of neutrons migrating in the discrete angular direction $\mu_m$ at time $t$ as a function of $0 < x < L$; $w_n$ are the quadrature weights; $C(x,t)$ is the concentration of delayed neutron precursors with radioactive decay constant $\lambda$, $\phi = \sum_{n=1}^{N} \psi_n w_n$ is the scalar flux and $\phi_0(x)$ is the scalar flux profile at $t = 0$. Moreover, $\sigma_t$ , $\sigma_s$ and $\sigma_f$ are, respectively, the total, scattering and fission macroscopic cross sections, $\nu$ the total average number of neutrons emitted in each fission event and $\beta$ is the delayed neutron fraction, which defends on the nuclear fuel [LeMi84].

The present decomposition method is based on the expansions of the angular flux and the precursor concentration in series of unknown functions

$$\begin{aligned} \psi_m(x,t) &= \sum_{k=0}^{\infty} \psi_m^{(k)}(x,t) \\ C(x,t) &= \sum_{k=0}^{\infty} C^{(k)}(x,t). \end{aligned} \tag{52.4}$$

In order to use these analytical representations for computer numerical applications, these series are truncated by using prescribed stopping criteria in the numerical scheme. In other words, Eq.(52.4) become

$$\begin{aligned} \psi_m(x,t) &= \sum_{k=0}^{M} \psi_m^{(k)}(x,t) \\ C(x,t) &= \sum_{k=0}^{M} C^{(k)}(x,t). \end{aligned} \tag{52.5}$$

By substituting Eq.(52.5) into Eq.(52.1), we obtain

$$\begin{aligned} &\frac{1}{v}\frac{\partial}{\partial t}\sum_{k=0}^{M}\psi_m^{(k)} + \mu_m \frac{\partial}{\partial x}\sum_{k=0}^{M}\psi_m^{(k)} + \sigma_t \sum_{k=0}^{M}\psi_m^{(k)} - \frac{\sigma_s}{2}\sum_{k=0}^{M}\phi^{(k)} = \\ &\qquad (1-\beta)\frac{\nu\sigma_f}{2}\sum_{k=0}^{M}\phi^{(k)} + \lambda\sum_{k=0}^{M}C^{(k)}, \\ &\frac{\partial}{\partial t}\sum_{k=0}^{M}C^{(k)} = \frac{\beta\nu\sigma_f}{2}\sum_{k=0}^{M}\phi^{(k)} - \lambda\sum_{k=0}^{M}C^{(k)}, \end{aligned} \tag{52.6}$$

where we have defined $\phi^{(k)} = \sum_{n=1}^{N}\psi_n^{(k)}w_n$. A simple count indicates that the $S_N$ kinetics problem 52.1–52.3 has been reduced to a system of $N+1$ partial differential equations in $M(N+1)$ unknown expansion functions $\psi_m^{(k)}$ and $C^{(k)}$, $m = 1:N$, $k = 1:M$. Once the recursive system of equations to determine the unknown functions is not unique, we make the following choice of the initial equation for $\phi_m^{(0)}$ and $C^{(0)}$:

$$\begin{cases} \dfrac{1}{v}\dfrac{\partial\psi_m^{(0)}}{\partial t} + \mu_m\dfrac{\partial\psi_m^{(0)}}{\partial x} + \sigma_t\psi_m^{(0)} - \dfrac{\sigma_s}{2}\phi^{(0)} = 0, \quad m = 1:N. \\ \dfrac{\partial C^{(0)}}{\partial t} = \dfrac{\beta\nu\sigma_f}{2}\phi^{(0)} - \lambda C^{(0)} \end{cases} \tag{52.7}$$

with initial conditions

$$\begin{aligned} \psi_m^{(0)}(x,0) &= \phi_0(x), \quad m = 1:N \\ C^{(0)}(x,0) &= \beta\frac{\nu\sigma_f}{2\lambda}\phi_0(x) \end{aligned} \tag{52.8}$$

and boundary conditions

$$\begin{aligned} \psi_m^{(0)}(0,t) &= f_m, \qquad \mu_m > 0 \\ \psi_m^{(0)}(L,t) &= g_m, \qquad \mu_m < 0. \end{aligned} \tag{52.9}$$

Following this scheme, the $k$th recursive system of partial differential equations for $\psi_m^{(k)}$ and $C^{(k)}$ appears as

$$\begin{cases} \dfrac{1}{v}\dfrac{\partial \psi_m^{(k)}}{\partial t} + \mu_m \dfrac{\partial \psi_m^{(k)}}{\partial x} + \sigma_t \psi_m^{(k)} - \dfrac{\sigma_s}{2}\phi^{(k)} = (1-\beta)\dfrac{\nu\sigma_f}{2}\phi^{(k-1)} + \lambda C^{(k-1)}, \\ \dfrac{\partial C^{(k)}}{\partial t} = \dfrac{\beta\nu\sigma_f}{2}\phi^{(k)} - \lambda C^{(k)}, \quad m = 1:N, \;\; k = 1:M. \end{cases} \tag{52.10}$$

Each solution of each recursive system has to satisfy Eq.(52.6) in addition, according to Eqs.(52.8) and (52.9), the solution for $k = 0$ is required to satisfy initial and boundary conditions to the $S_N$ problem. Therefore, functions $\psi_m^{(k)}$ and $C^{(k)}$, $k > 0$, satisfy zero initial conditions. We remark that the present formulation for building the recursive systems is not unique. We have chosen the procedure, as described in this chapter, since it is convenient for the application of the $TLTS_N$ method [ToSeVi13, To12], that we briefly describe in the next section.

## 52.3 The $TLTS_N$ Solution

In this section, we present the $TLTS_N$ method to solve the recursive systems of partial differential equations, as we described in the previous section. Therefore, let us consider the time dependent, slab-geometry $S_N$ transport equations

$$\frac{1}{v}\frac{\partial}{\partial t}\psi_m(x,t) + \mu_m \frac{\partial}{\partial x}\psi_m(x,t) + \sigma_t \psi_m(x,t) = \frac{\sigma_s}{2}\sum_{n=1}^{N}\psi_m(x,t)w_n + S_m(x,t) \tag{52.11}$$

where we have defined $S(x,t)$ as the source term. With initial condition

$$\psi_m(x,0) = \phi(x) \tag{52.12}$$

and boundary conditions

$$\begin{aligned} \psi_m(0,t) &= f_m, \qquad \mu_m > 0 \\ \psi_m(L,t) &= g_m, \qquad \mu_m < 0. \end{aligned} \tag{52.13}$$

Now we apply the Laplace transformation to Eq.(52.11), in the time variable. The result is

$$\mu_m \frac{d}{dx}\overline{\psi}_m(x,p) + \sigma_t^p \overline{\psi}_m(x,p) = \frac{\sigma_s}{2}\sum_{n=1}^{N}\overline{\psi}_n(x,p)w_n + \overline{R}_m(x,p), \tag{52.14}$$

with the boundary conditions

$$\begin{aligned} \overline{\psi}_m(0,p) &= f_m/p, \qquad \mu_m > 0 \\ \overline{\psi}_m(L,p) &= g_m/p, \qquad \mu_m < 0. \end{aligned} \tag{52.15}$$

Here, $\overline{\psi}_m(x,p)$ denotes the transformed angular flux. Moreover we have defined

$$\sigma_t^p = \sigma_t + \frac{p}{v} \quad \text{and} \quad \overline{R}_m(x,p) = \frac{1}{v}\phi(x) + \overline{S}(x,p).$$

Furthermore, we write Eq.(52.14) in the matrix form

$$\frac{d}{dx}\overline{\mathbf{P}}(x,p) - \mathbf{A}(p)\overline{\mathbf{P}}(x,p) = \overline{\mathbf{R}}(x,p) \tag{52.16}$$

where $\mathbf{A}(p)$ is the $N \times N$ square matrix whose entries are

$$a_{ij}(p) = \begin{cases} \dfrac{\sigma_s w_j}{2\mu_i} - \dfrac{\sigma_t^p}{\mu_i} & \text{if } i = j, \\ \dfrac{\sigma_s w_j}{2\mu_i} & \text{if } i \neq j. \end{cases} \tag{52.17}$$

In addition, $\overline{\mathbf{P}}(x,p)$ and $\overline{\mathbf{R}}(x,p)$ are $N$-dimensional vectors defined as

$$\overline{\mathbf{P}}(x,p) = \begin{bmatrix} \overline{\mathbf{P}}_1(x,p) \\ \overline{\mathbf{P}}_2(x,p) \end{bmatrix} = \begin{bmatrix} \overline{\Psi}_1(x,p) \\ \vdots \\ \overline{\Psi}_{N/2}(x,p) \\ \overline{\Psi}_{N/2+1}(x,p) \\ \vdots \\ \overline{\Psi}_N(x,p) \end{bmatrix}$$

and

$$\overline{\mathbf{R}}(x,p) = \left[\frac{\overline{R}_1}{\mu_1}, \ldots, \frac{\overline{R}_N}{\mu_N}\right]^T,$$

with boundary conditions

$$\overline{\mathbf{P}}_1(x,p) = \mathbf{f}/p \quad \text{and} \quad \overline{\mathbf{P}}_2(x,p) = \mathbf{g}/p \tag{52.18}$$

where $\overline{\mathbf{P}}_1(x,p)$ and $\overline{\mathbf{P}}_2(x,p)$ are $N/2$-dimensional sub-vectors of the transformed angular flux vector $\overline{\mathbf{P}}(x,p)$. The $N/2$ entries of $\overline{\mathbf{P}}_1(x,p)$ are the transformed angular fluxes in directions $\mu_m > 0$ and the $N/2$ entries of $\overline{\mathbf{P}}_2(x,p)$ are the transformed angular fluxes in directions $\mu_m < 0$.

Furthermore, by applying the $\text{LTS}_\text{N}$ method to Eq.(52.16), we obtain

$$\begin{aligned}\overline{\mathbf{P}}(x,p) &= \begin{pmatrix}\overline{\mathbf{P}}_1(x,p)\\ \overline{\mathbf{P}}_2(x,p)\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{X_{11}}(p) & \mathbf{X_{12}}(p)\\ \mathbf{X_{21}}(p) & \mathbf{X_{22}}(p)\end{pmatrix}\begin{pmatrix}\mathbf{e}^{\mathbf{D}^+(x-L)} & \mathbf{0}\\ \mathbf{0} & \mathbf{e}^{\mathbf{D}^- x}\end{pmatrix}\begin{pmatrix}\xi_1(p)\\ \xi_2(p)\end{pmatrix} + \begin{pmatrix}\overline{\mathbf{H}}_1(x,p)\\ \overline{\mathbf{H}}_2(x,p)\end{pmatrix}\end{aligned} \tag{52.19}$$

where $\mathbf{D}^+$ and $\mathbf{D}^-$ are $N/2$-order diagonal matrices whose entries, are respectively the positive and negative eigenvalues, and $\mathbf{X}(p)$ is the matrix whose columns are eigenvectors of $\mathbf{A}(p)$. In addition, we write the particular solution $\overline{\mathbf{H}}(x,p)$ as

$$\begin{aligned}&\overline{\mathbf{H}}(x,p) =\\ &\begin{pmatrix}\overline{\mathbf{H}}_1(x,p)\\ \overline{\mathbf{H}}_2(x,p)\end{pmatrix} = \begin{pmatrix}\mathbf{X_{11}}(p) & \mathbf{X_{12}}(p)\\ \mathbf{X_{21}}(p) & \mathbf{X_{22}}(p)\end{pmatrix}\begin{pmatrix}\int_L^x \mathbf{e}^{\mathbf{D}^+(x-\xi)}\sum_{j=1}^2 \mathbf{Z_{1j}}(p)\mathbf{R}_j(\xi)d\xi\\ \int_0^x \mathbf{e}^{\mathbf{D}^-(x-\xi)}\sum_{j=1}^2 \mathbf{Z_{1j}}(p)\mathbf{R}_j(\xi)d\xi\end{pmatrix}\end{aligned} \tag{52.20}$$

where $\mathbf{Z}(p) = \mathbf{X}^{-1}(p)$. At this point we apply boundary conditions (52.18) at $x = 0$ and $x = L$, and we determine the unknown sub-vectors $\xi_1$ and $\xi_2$ by solving the resulting system of linear equations. Now, we are able to determine the angular flux profile by applying the inverse Laplace transform to the transformed angular flux (52.19)

$$\mathbf{P}(x,t) = \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} \bar{\mathbf{P}}(x,p)e^{pt}dp. \tag{52.21}$$

Due to operation difficulties to solve analytically the integral given in Eq.(52.21), we choose to approximate it by use of the Gaver–Stehfest algorithm [Ga66, St70, St70a], which yields an approximate solution by use of the following expression

$$f(t) = \frac{\ln 2}{t}\sum_{i=1}^{N} V_i f\left(\frac{\ln 2}{t}i\right),$$

where $N$ is an even integer and $V_i$ is defined as

$$V_i = (-1)^{N/2+i} \sum_{k=\left[\frac{i+1}{2}\right]}^{Min(i,N/2)} \frac{k^{N/2}(2k)!}{(N/2-k)!(k)!(k-1)!(i-k)!(2k-i)!}.$$

Before concluding this section, we remark that solving the recursive systems represented in Eq.(52.10) by the $TLTS_N$ method turned out to be very inefficient, as it required a large number of iterations when $[\sigma_s + (1-\beta)\nu\sigma_f] > \sigma_t$. Therefore, the fission source term has been partitioned in two terms, so the resulting recursive systems appear as

$$\frac{1}{v}\frac{\partial}{\partial t}\psi_m(x,t) + \mu_m \frac{\partial}{\partial x}\psi_m(x,t) + \sigma_t \psi_m(x,t) - \frac{\sigma_s + \alpha(1-\beta)\nu\sigma_f}{2} \sum_{n=1}^{N} \psi_n(x,t) w_n$$

$$= (1-\alpha)(1-\beta)\frac{\nu\sigma_f}{2} \sum_{n=1}^{N} \psi_n(x,t) w_n + \lambda C(x,t), \quad m = 1:N \tag{52.22}$$

such that $\sigma_s + \alpha(1-\beta)\nu\sigma_f < \sigma_t$. Following this acceleration technique reduced significantly the number of iterations to reach the stopping criterion.

## 52.4 Numerical Results

In this section, we show two numerical experiments to a model problem or a homogeneous slab. Before describing the model problem, we define reactivity as the quantity

$$\rho = \frac{k_{\text{eff}} - 1}{k_{\text{eff}}}, \tag{52.23}$$

where $k_{\text{eff}}$ is the effective multiplication factor which is defined as the dominant eigenvalue of the steady-state $S_N$ problem

$$\mu_m \frac{d}{dx}\psi_{m,0}(x) + \sigma \psi_{m,0}(x) = \frac{\sigma_s + \nu\sigma_f/k_{\text{eff}}}{2}\phi_0(x). \tag{52.24}$$

It is common in nuclear reactor physics to divide reactivity $\rho$ by the delayed neutron fraction $\beta$, in which case the reactivity is referred to as dollar (\$) [LeMi84].

The homogeneous slab has length $L = 10$ cm, the thermal neutron speed is $v = 2.2 \times 10^5$ cm/s and to model this problem we used vacuum boundary conditions at $x = 0$ and $x = 10$ with the Gauss–Legendre $S_{10}$ angular quadrature set and

**Table 52.1** Macroscopic cross section to the model problem

| Macroscopic | cross section |
|---|---|
| $\sigma_t$ | 1 $cm^{-1}$ |
| $\sigma_s$ | 0.1 $cm^{-1}$ |
| $\nu\sigma_f$ | 0.925 $cm^{-1}$ |

**Table 52.2** Numerical results for the scalar flux.

| | | Scalar | flux | (neutrons/$cm^2$s) |
|---|---|---|---|---|
| Reactivity ($) | Time (s) | x = 2.0 cm | x = 5.0 cm | x = 7.0 cm |
| 0 | 0 | 0.32202E +17 | 0.47659E +17 | 0.40573E +17 |
| | 0.1 | 0.32205E +17 | 0.47665E +17 | 0.40577E +17 |
| | 1 | 0.32205E +17 | 0.47664E +17 | 0.40577E +17 |
| | 10 | 0.32205E +17 | 0.47665E +17 | 0.40578E +17 |
| −0.1 | 0.1 | 0.29257E +17 | 0.43301E +17 | 0.36862E +17 |
| | 1 | 0.29074E +17 | 0.43030E +17 | 0.36632E +17 |
| | 10 | 0.27307E +17 | 0.40415E +17 | 0.34405E +17 |
| 0.001 | 0.1 | 0.32237E +17 | 0.47713E +17 | 0.40618E +17 |
| | 1 | 0.32239E +17 | 0.47716E + 017 | 0.40620E +17 |
| | 10 | 0.32262E +17 | 0.47750E +17 | 0.40650E +17 |

macroscopic cross section as shown in Table 52.1. Our first numerical experiment consists of varying the reactivity for $U_{92}^{235}$ fuel ($\beta = 0.0065$, $\lambda = 0.076666$) which generates power $P = 10$Mw; that is, we keep the system critical for 10 seconds by dividing the fission cross section by $k_{\text{eff}} = 1.00010842619$ ($\rho = 0$\$). As we see in Table 52.2 the scalar flux at $x = 2$, $x = 5$ and $x = 7$ are kept constant, apart from reduced computation rounding errors. Then, we add a negative reactivity ($\rho = -0.1$\$) and Table 52.2 shows that the scalar flux decreases as a function of time, as predicted theoretically. On the other hand, by adding a positive reactivity ($\rho = 0.001$\$) the scalar flux increases as a function of time. Figures 52.1, 52.2 and 52.3 display the scalar flux profiles within the slab for these three tests.

The second numerical experiment shows numerical results generated by the described decomposition method, by adding at $t = 0$ a fixed negative reactivity $\rho = -0.00065$ [Eq.(52.23)] in distinct fission chain reacting systems composed of three types of fuel: $U_{92}^{235}$ ($\beta = 0.0065$, $\lambda = 0.076666$), $U_{92}^{233}$($\beta = 0.0026$, $\lambda = 0.0542527$) and $Pu_{94}^{239}$($\beta = 0.0021$, $\lambda = 0.0648618$). Table 52.3 displays the numerical results for the scalar flux at the same positions as in the previous numerical experiment for $t = 1$s. Since $\rho < 0$, the scalar flux decreases as a function of time. We remark that, as expected, the scalar flux decreased much more for the care of $Pu_{94}^{239}$-fuel. This is due to the fact that the delayed neutron fraction $\beta$ for the $Pu_{94}^{239}$-fuel is smaller than $\beta$ for the other. As a result, we conclude that the delayed neutrons play an important role in nuclear reactor control. Figure 52.4 shows the scalar flux profiles at $t = 1$ s for these three types of fuel.

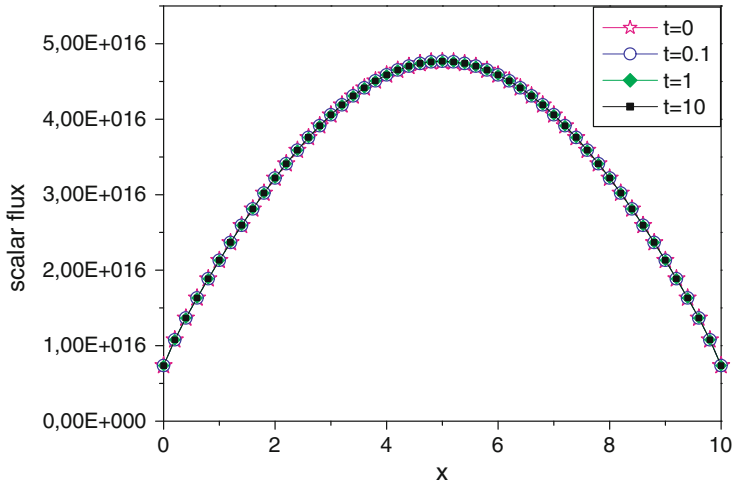


**Fig. 52.1** Scalar flux profile ($\rho = 0$).

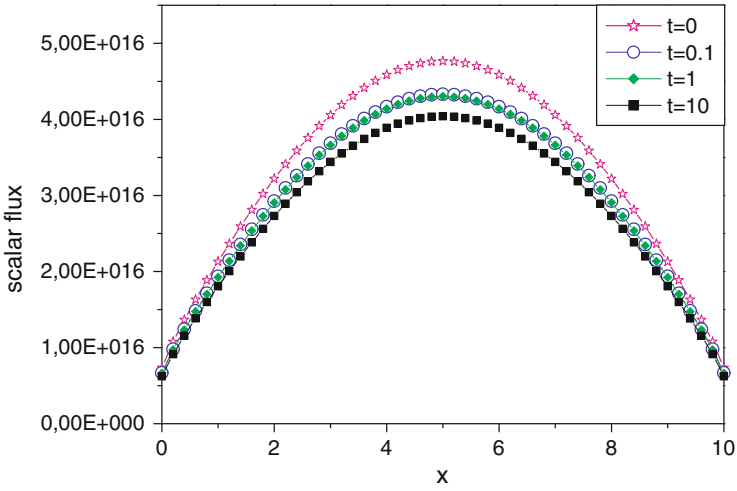


**Fig. 52.2** Scalar flux profiles at $0 \leq t \leq 10$ s ($\rho = -0.1\$$).

## 52.5 Concluding Remarks

Regarding the contribution of this work we would like to emphasize that the proposed recursive system of equations allows us to solve the $S_N$ neutron transport equation with neutron fission source in a slab by the $LST_N$ version with real eigenvalues, bearing in mind that for this kind of problem the eigenvalues are purely imaginary. Another important feature of this recursion scheme consists in the acceleration of the solution convergence, in the sense that we can obtain

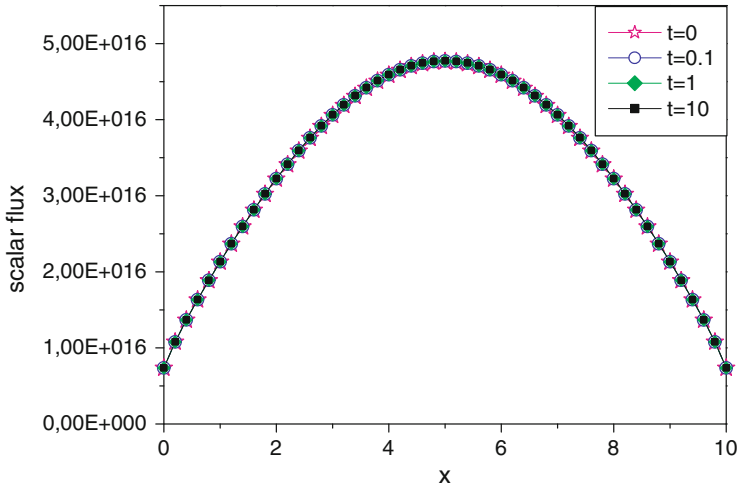


**Fig. 52.3** Scalar flux profiles at $0 \leq t \leq 10$ s ($\rho = 0.001\$$).

**Table 52.3** Numerical results for the scalar flux.

| | Scalar | flux | (neutrons/cm$^2$s) |
|---|---|---|---|
| Fuel | x = 2.0 cm | x = 5.0 cm | x = 7.0 cm |
| $U_{92}^{235}$ | 0.29074E +17 | 0.43030E +17 | 0.36632E +17 |
| $U_{92}^{233}$ | 0.25488E +17 | 0.37723E +17 | 0.32114E +17 |
| $Pu_{94}^{239}$ | 0.24222E +17 | 0.35850E +17 | 0.30519E +17 |

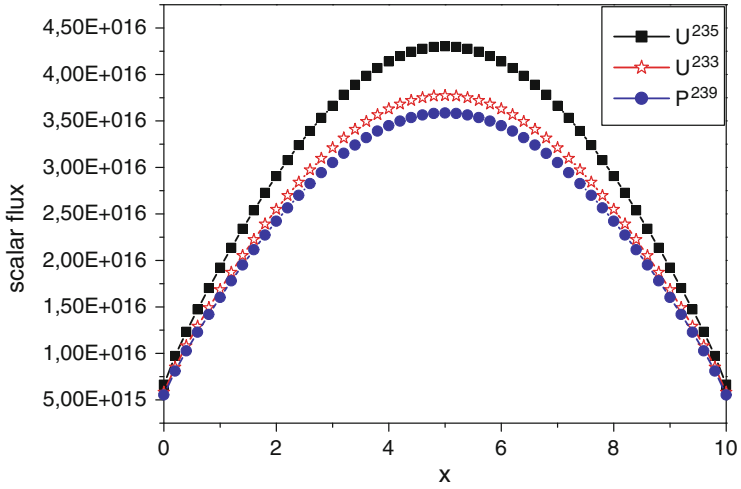


**Fig. 52.4** Scalar flux for three different fuels.

any prescribed accuracy for the results with a drastic reduction of the number of solutions for the respective recursive equations. Finally, it is noteworthy that this

work paves the way for applications of the discussed methodology in accelerator driven nuclear reactor concepts, observing that the pulsed neutron source can be replaced by an idealized neutron fission source. Motivated by the good results obtained, we focus our future attention in this direction.

**Acknowledgements** The authors CFS, RCB, and MTV wish to thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for their partial financial support of this work.

## References

[Ba92] Barichello, L.B.: *Formulação Analítica para Solução do Problema de Ordenada Discreta Unidimensional.* Tese de Doutorado pelo Programa de Pós-Graduação em Engenharia Mecânica (PROMEC), Universidade Federal do Rio Grande do Sul (1992).

[BaVi97] Batistela, C., Vilhena, M.: *Cálculo de Criticalidade pelo Método $LTS_N$*. In XI ENFIR-Encontro Nacional de Física de Reatores e Termo-hidráulica. Poços de Caldas, M. G., Brasil. volume 1, pp. 226–231 (1997a).

[Ga66] Gaver, D.P., Jr.: *Observing Stochastic Processes and Approximate Transform Inversion.* Operations Research, volume 14, pp. 444–459 (1966).

[LeMi84] Lewis, E.E., Miller, W.F., Jr.: *Computational Method of Neutron Transport.* John Wiley and Sons, New York (1984).

[OrViCa04] Orengo, G., Vilhena, M., G. C. O., Caldeira, G., G. G. A.: *Recent Advances in the $LTS_N$ Method for Criticality Calculations in a Slab Geommetry.* Annals of Nuclear Energy. volume 31, pp. 2195–2202 (2004).

[St70] Stehfest, H.: *Numerical Inversion of Laplace Transforms Algorithm 368.* Communications of the ACM, volume 13(1), pp. 47–49 (1970).

[St70a] Stehfest, H.: *Numerical Inversion of Laplace Transforms Remark on Algorithm 368.* Communications of the ACM, volume 13(10), pp. 624 (1970).

[To12] Tomaschewski, F.K.: *Solução da Equação $S_N$ Multigrupo de Transporte Dependente do Tempo em Meio Heterogêneo.* Dissertação de Mestrado pelo Programa de Pós-Graduação em Matemática Aplicada (PPGMAp), Universidade Federal do Rio Grande do Sul (2012).

[ToSeVi13] Tomaschewski, F.K., Segatto, C.F., Vilhena, M.T.: *A Genuine Analytical Solution for the $S_N$ Multi-Group Neutron Equation in Planar Geometry.* Integral Methods in Science and Engineering: Progress in Numerical and Analytic Techniques, C. Constanda, B. Bodmann, H. F. Campos Velho. (Org.), 1ed.Basel: Springer/BirkhÄauser, vol. XXII, p. 311–320 (2013).

# Chapter 53
# New Constructions in the Theory of Elliptic Boundary Value Problems

**V.B. Vasilyev**

## 53.1 Introduction

How are potentials constructed for boundary value problems? One takes a fundamental solution of the corresponding differential operator *in whole space* $\mathbf{R}^m$, and with its help constructs the potentials according to boundary conditions. Further, one studies their boundary properties, and with the help of potentials reduces the boundary value problem to an equivalent integral equation on the boundary. The formulas for integral representation of solution of the boundary value problem were obtained for separate cases only (a ball, a half-space, such places, where one has explicit form for a Green function). Thus, an ideal result for a boundary value problem even with a smooth boundary is its reduction to an equivalent Fredholm equation and obtaining the existence and uniqueness theorem (without knowing how the solution looks) [Ag57, Fa88, Ke94, MiMiTa01, HsWe08]. We would like to show that potentials can arise from another point of view, without using fundamental solution, but using factorization idea and they obviously should take into account the boundary geometry. A smooth boundary is a hyper-plane locally (there is a Poisson formula for the Dirichlet problem, see also [Es81]), first type of non-smooth boundary is a conical surface.

V.B. Vasilyev (✉)
Lipetsk State Technical University, Moskovskaya 30, Lipetsk 398600, Russia
e-mail: vbv57@inbox.ru



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_53

## 53.2 Operators, Equations, and Wave Factorization

We consider an elliptic pseudo-differential equation in a multi-dimensional cone and starting wave factorization concept we add some boundary conditions. For the simplest cases explicit formulas for solution are given like layer potentials for a classical case.

Let's go to studying solvability of pseudo-differential equations [Va00a, Va11, Va10]

$$(Au_+)(x) = f(x),\ x \in C^a_+, \tag{53.1}$$

in the space $H^s(C^a_+)$, where $C^a_+$ is $m$-dimensional cone

$$C^a_+ = \{x \in \mathbf{R}^m : x = (x_1, ..., x_{m-1}, x_m), x_m > a|x'|, a > 0\},\ x' = (x_1, ..., x_{m-1}),$$

$A$ is pseudo-differential operator ($\tilde{u}$ denotes the Fourier transform of $u$)

$$u(x) \longmapsto \int\limits_{\mathbf{R}^m} e^{ix\cdot\xi} A(\xi)\tilde{u}(\xi)d\xi,\ \ x \in \mathbf{R}^m,$$

with the symbol $A(\xi)$ satisfying the condition

$$c_1 \le |A(\xi)(1+|\xi|)^{-\alpha}| \le c_2.$$

(Such symbols are elliptic [Es81] and have the order $\alpha \in \mathbf{R}$ at infinity.)

By definition, the space $H^s(C^a_+)$ consists of distributions from $H^s(\mathbf{R}^m)$, whose support belongs to $\overline{C^a_+}$. The norm in the space $H^s(C^a_+)$ is induced by the norm from $H^s(\mathbf{R}^m)$. The right-hand side $f$ is chosen from the space $H^{s-\alpha}_0(C^a_+)$, which is space of distributions $S'(C^a_+)$, admitting the continuation on $H^{s-\alpha}(\mathbf{R}^m)$. The norm in the space $H^{s-\alpha}_0(C^a_+)$ is defined by

$$||f||^+_{s-\alpha} = \inf ||lf||_{s-\alpha},$$

where *infimum* is chosen from all continuations $l$.

Further, we define a special multi-dimensional singular integral by the formula

$$(G_m u)(x) = \lim_{\tau\to 0+} \int\limits_{\mathbf{R^m}} \frac{u(y', y_m)dy'dy_m}{(|x'-y'|^2 - a^2(x_m - y_m + i\tau)^2)^{m/2}}$$

(we omit a certain constant, see [Va00a]). Let us recall, this operator is multi-dimensional analogue of the one-dimensional Cauchy type integral, or Hilbert transform.

We also need some notations before definition.

The symbol $\overset{*}{C}{}^a_+$ denotes a conjugate cone for $C^a_+$:

$$\overset{*}{C}{}^a_+ = \{x \in \mathbf{R}^m : x = (x', x_m), ax_m > |x'|\},$$

$C^a_- \equiv -C^a_+$, $T(C^a_+)$ denotes radial tube domain over the cone $C^a_+$, i.e. domain in a complex space $\mathbf{C}^m$ of type $\mathbf{R}^m + iC^a_+$.

To describe the solvability picture for the equation (53.1) we will introduce the following definition.

**Definition 1.** Wave factorization for the symbol $A(\xi)$ is called its representation in the form

$$A(\xi) = A_{\neq}(\xi)A_{=}(\xi),$$

where the factors $A_{\neq}(\xi), A_{=}(\xi)$ must satisfy the following conditions:

1) $A_{\neq}(\xi), A_{=}(\xi)$ are defined for all admissible values $\xi \in \mathbf{R}^m$, without may be, the points $\{\xi \in \mathbf{R}^m : |\xi'|^2 = a^2\xi_m^2\}$;
2) $A_{\neq}(\xi), A_{=}(\xi)$ admit an analytical continuation into radial tube domains $T(\overset{*}{C}{}^a_+), T(\overset{*}{C}{}^a_-)$, respectively, with estimates

$$|A^{\pm 1}_{\neq}(\xi + i\tau)| \le c_1(1 + |\xi| + |\tau|)^{\pm\kappa},$$

$$|A^{\pm 1}_{=}(\xi - i\tau)| \le c_2(1 + |\xi| + |\tau|)^{\pm(\alpha-\kappa)}, \ \forall \tau \in \overset{*}{C}{}^a_+ .$$

The number $\kappa \in \mathbf{R}$ is called index of wave factorization.

The class of elliptic symbols admitting the wave factorization is very large. There are the special chapter in the book [Va00a] and the paper [Va00b] devoted to this question, there are examples also for certain operators of mathematical physics.

Everywhere below we will suppose that the mentioned wave factorization does exist, and the sign $\sim$ will denote the Fourier transform, particularly $\tilde{H}(D)$ denotes the Fourier image of the space $H(D)$.

## 53.3 After the Wave Factorization

Now we will consider the equation (53.1) for the case $\kappa - s = n + \delta, n \in \mathbf{N}, |\delta| < 1/2$, only. A general solution can be constructed in the following way. We choose an arbitrary continuation $lf$ of the right-hand side on a whole space $H^{s-\alpha}(\mathbf{R}^m)$ and introduce

$$u_-(x) = (lf)(x) - (Au_+)(x).$$

After wave factorization for the symbol $A(\xi)$ with preliminary Fourier transform we write

$$A_{\neq}(\xi)\tilde{u}_+(\xi)+A_{=}^{-1}(\xi)\tilde{u}_-(\xi)=A_{=}^{-1}(\xi)\tilde{f}(\xi).$$

One can see that $A_{=}^{-1}(\xi)\tilde{f}(\xi)$ belongs to the space $\tilde{H}^{s-\kappa}(\mathbf{R}^m)$, and if we choose the polynomial $Q(\xi)$, satisfying the condition

$$|Q(\xi)|\sim(1+|\xi|)^n,$$

then $Q^{-1}(\xi)A_{=}^{-1}(\xi)\tilde{f}(\xi)$ will belong to the space $\tilde{H}^{-\delta}(\mathbf{R}^m)$.

Further, according to the theory of multi-dimensional Riemann problem [Va00a], we can decompose the last function on two summands (jump problem):

$$Q^{-1}A_{=}^{-1}\tilde{f}=f_++f_-,$$

where $f_+\in\tilde{H}(C_+^a)$,$f_-\in\tilde{H}(\mathbf{R}^m\setminus C_+^a)$.

So, we have

$$Q^{-1}A_{\neq}\tilde{u}_++Q^{-1}A_{=}^{-1}\tilde{u}_-=f_++f_-,$$

or

$$Q^{-1}A_{\neq}\tilde{u}_+-f_+=f_--Q^{-1}A_{=}^{-1}\tilde{u}_-$$

In other words,

$$A_{\neq}\tilde{u}_+-Qf_+=Qf_--A_{=}^{-1}\tilde{u}_-.$$

The left-hand side of the equality belongs to the space $\tilde{H}^{-n-\delta}(C_+^a)$, and right-hand side is from $\tilde{H}^{-n-\delta}(\mathbf{R}^m\setminus C_+^a)$, hence

$$F^{-1}(A_{\neq}\tilde{u}_+-Qf_+)=F^{-1}(Qf_--A_{=}^{-1}\tilde{u}_-),$$

where the left-hand side belongs to the space $H^{-n-\delta}(C_+^a)$, and the right-hand side belongs to the space $H^{-n-\delta}(\mathbf{R}^m\setminus C_+^a)$, that's why we conclude immediately that it is a distribution supported on $\partial C_+^a$.

The main tool now is to define the form of the distribution.

We denote $T_a$ the bijection operator transferring $\partial C_+^a$ into hyperplane $x_m=0$, more precisely, it is transformation $\mathbf{R}^m\longrightarrow\mathbf{R}^m$ of the following type

$$\begin{cases} t_1=x_1,\\ \ldots\ldots\ldots\\ t_{m-1}=x_{m-1},\\ t_m=x_m-a|x'|. \end{cases}$$

Then the function

$$T_a F^{-1}(A_{\neq}\tilde{u}_+ - Qf_+)$$

will be supported on the hyperplane $t_m = 0$ and belongs to $H^{-n-\delta}(\mathbf{R}^m)$. Such distribution is a linear span of Dirac mass-function and its derivatives [GeSh59] and looks as the following sum

$$\sum_{k=0}^{n-1} c_k(t')\delta^{(k)}(t_m).$$

It is left to think, what is operator $T_a$ in Fourier image. Explicit calculations give a simple answer:

$$FT_a u = V_a \tilde{u},$$

where $V_a$ is something like a pseudo-differential operator with symbol $e^{-ia|\xi'|\xi_m}$, and, further, one can construct the general solution of our pseudo-differential equation (53.1).

We need some connections between the Fourier transform and the operator $T_a$:

$$(FT_a u)(\xi) = \int\limits_{\mathbf{R}^m} e^{-ix\cdot\xi} u(x_1, ..., x_{m-1}, x_m - a|x'|)dx =$$

$$= \int\limits_{\mathbf{R}^m} e^{-iy'\xi'} e^{-i(y_m + a|y'|)\xi_m} u(y_1, ..., y_{m-1}, y_m)dy =$$

$$= \int\limits_{\mathbf{R}^{m-1}} e^{-ia|y'|\xi_m} e^{-iy'\xi'} \hat{u}(y_1, ..., y_{m-1}, \xi_m)dy',$$

where $\hat{u}$ denotes the Fourier transform on the last variable, and the Jacobian is

$$\frac{D(x_1, x_2, ..., x_m)}{D(y_1, y_2, ..., y_m)} = \begin{vmatrix} 1 & 0 & \cdots 0 & 0 \\ 0 & 1 & \cdots 0 & 0 \\ 0 & 0 & \cdots 1 & 0 \\ & \cdots\cdots\cdots\cdots & & \\ \frac{ay_1}{|y'|} & \frac{ay_2}{|y'|} & \ldots\ldots\frac{ay_{m-1}}{|y'|} & 1 \end{vmatrix} = 1.$$

If we define a pseudo-differential operator by the formula

$$(Au)(x) = \int\limits_{\mathbf{R}^m} e^{ix\xi} A(\xi)\tilde{u}(\xi)d\xi,$$

and the direct Fourier transformation

$$\tilde{u}(\xi) = \int\limits_{\mathbf{R}^m} e^{-ix\xi} u(x)dx,$$

then we have the following relation formally at least

$$(FT_a u)(\xi) = \int\limits_{\mathbf{R}^{m-1}} e^{-ia|y'|\xi_m} e^{-iy'\xi'} \hat{u}(y_1, \ldots, y_{m-1}, \xi_m)dy. \tag{53.2}$$

In other words, if we denote the $(m-1)$-dimensional Fourier transform ($y' \to \xi'$ in distribution sense) of function $e^{-ia|y'|\xi_m}$ by $E_a(\xi', \xi_m)$, then the formula (53.2) will be the following

$$(FT_a u)(\xi) = (E_a * \tilde{u})(\xi),$$

where the sign $*$ denotes a convolution for the first $m-1$ variables, and the multiplier for the last variable $\xi_m$. Thus, $V_a$ is a combination of a convolution operator and the multiplier with the kernel $E_a(\xi', \xi_m)$. It is very simple operator, and it is bounded in Sobolev–Slobodetski spaces $H^s(\mathbf{R}^m)$.

Notice that distributions supported on conical surface and their Fourier transforms were considered in [GeSh59], but the author did not find the multi-dimensional analogue of theorem on a distribution supported in a single point in all issues of this book.

*Remark 1.* One can wonder why we can't use this transform in the beginning to reduce the conical situation (53.1) to hyperplane one, and then to apply Eskin's technique [Es81]. Unfortunately, this is impossible because $T_a$ is non-smooth transformation, but even for smooth transformation we obtain the same operator $A$ with some additional compact operator. Obtaining the invertibility conditions for such operator is a very serious problem.

## 53.4 General Solution

The following result is valid (it follows from considerations of Section 53.3).

**Theorem 1.** *A general solution of the equation* (53.1) *in Fourier image is given by the formula*

$$\tilde{u}_+(\xi) = A_{\neq}^{-1}(\xi)Q(\xi)G_m Q^{-1}(\xi)A_{=}^{-1}(\xi)\tilde{lf}(\xi) +$$

$$+A_{\neq}^{-1}(\xi)V_{-a}F\left(\sum_{k=1}^{n} c_k(x')\delta^{(k-1)}(x_m)\right),$$

*where* $c_k(x') \in H^{s_k}(\mathbf{R}^{m-1})$ *are arbitrary functions,* $s_k = s - \kappa + k - 1/2$, $k = 1, 2, \ldots, n$, *lf is an arbitrary continuation f on* $H^{s-\alpha}(\mathbf{R}^m)$.

Starting from this representation one can suggest different statements of boundary value problems for the equation (53.1).

### 53.4.1 Another Singularity

It may be that the singularity point will be different from considered one. This matter will influence the structure of the operator $V_a$. So, if we consider an another $m$-dimensional cone, for example $C_+^{\vec{a}} = \{x \in \mathbf{R}^m : x = (x_1,...,x_{m-1},x_m), x_m > \sum_{k=1}^{m-1} a_k|x_k|,\ a_k > 0,\ k = 1,2,...,m-1\}$, $\vec{a} = (a_1,...,a_{m-1})$, then we need certain corrections for our studies, in general it will be the same. Namely, we need to define a special multi-dimensional singular integral by the formula

$$(G_m u)(x) = (2i)^{m-1} \lim_{\tau \to 0+} \int_{\mathbf{R^m}} \prod_{j=1}^{m-1} \frac{a_j (x_m - y_m + i\tau)^{m-2}}{(x_j - y_j)^2 - a_j^2 (x_m + y_m + i\tau)^2} u(y) dy$$

(for details see also [Va00a]). Such operator corresponds to the Fourier multiplier (characteristic function, or indicator) of the pyramid $C_+^{\vec{a}}$.

The Jacobian for this transformation $T_{\vec{a}}$ is

$$\frac{D(x_1,x_2,...,x_m)}{D(y_1,y_2,...,y_m)} = \begin{vmatrix} 1 & 0 & \cdots 0 & 0 \\ 0 & 1 & \cdots 0 & 0 \\ 0 & 0 & \cdots 1 & 0 \\ & \cdots\cdots\cdots & & \\ -a\ sign(y_1) & -a\ sign(y_2) & ....-a\ sign(y_{m-1}) & 1 \end{vmatrix} = 1,$$

and the argument continues as above. This kind of singularity is considered in a forthcoming paper by the author, to appear in *Adv. Dyn. Syst. Appl.*

## 53.5 Boundary Conditions: Simplest Version, the Dirichlet Condition

We consider a very simple case, when $f \equiv 0,\ a = 1,\ n = 1$. Then the formula from the theorem takes the form

$$\tilde{u}_+(\xi) = A_{\neq}^{-1}(\xi) V_{-1} \tilde{c}_0(\xi').$$

We consider the following construction separately. According to the Fourier transform our solution is

$$u_+(x) = F^{-1}\{A_{\neq}^{-1}(\xi) V_{-1} \tilde{c}_0(\xi')\}.$$

Let's suppose we choose the Dirichlet boundary condition on $\partial C^1_+$ for unique identification of an unknown function $c_0$, i.e.

$$(Pu)(y) = g(y),$$

where $g$ is given function on $\partial C^1_+$, $P$ is restriction operator on the boundary, so we know the solution on the boundary $\partial C^1_+$. Thus,

$$T_1 u(x) = T_1 F^{-1}\{A^{-1}_{\neq}(\xi) V_{-1}\tilde{c}_0(\xi')\},$$

so we have

$$FT_1 u(x) = FT_1 F^{-1}\{A^{-1}_{\neq}(\xi) V_{-1}\tilde{c}_0(\xi')\} = V_1\{A^{-1}_{\neq}(\xi) V_{-1}\tilde{c}_0(\xi')\}, \tag{53.3}$$

and we know $(P'T_1 u)(x') \equiv v(x')$, where $P'$ is the restriction operator on the hyperplane $x_m = 0$.

The relation between the operators $P'$ and $F$ is well-known [Es81]:

$$(FP'u)(\xi') = \int\limits_{-\infty}^{+\infty} \tilde{u}(\xi', \xi_m) d\xi_m.$$

Returning to the formula (53.3) we obtain the following

$$\tilde{v}(\xi') = \int\limits_{-\infty}^{+\infty} \{V_1\{A^{-1}_{\neq}(\xi) V_{-1}\tilde{c}_0(\xi')\}\}(\xi', \xi_m) d\xi_m, \tag{53.4}$$

where $\tilde{v}(\xi')$ is given function. Hence, the equation (53.4) is an integral equation for determining $c_0(x')$.

The Neumann boundary condition leads to analogous integral equation (see below).

## 53.6 Conical Potentials

We consider the particular case where $f \equiv 0, n = 1$. The formula for general solution of the equation (53.1) takes the form

$$\tilde{u}_+(\xi)) = A^{-1}_{\neq}(\xi) V_{-a} F\{c_0(x')\delta^{(0)}(x_m)\},$$

and after Fourier transform (for simplicity we write $\tilde{c}$ instead of $V_{-1}\tilde{c}_0$),

$$\tilde{u}_+(\xi) = A^{-1}_{\neq}(\xi)\tilde{c}(\xi'), \tag{53.5}$$

or equivalently the solution is

$$u_+(x) = F^{-1}\{A_{\neq}^{-1}(\xi)\tilde{c}(\xi')\}.$$

Then we apply the operator $T_a$ to formula (53.5)

$$(T_a u_+)(t) = T_a F^{-1}\{A_{\neq}^{-1}(\xi)\tilde{c}(\xi')\}$$

and the Fourier transform

$$(FT_a u_+)(\xi) = FT_a F^{-1}\{A_{\neq}^{-1}(\xi)\tilde{c}(\xi')\}.$$

If the boundary values of our solution $u_+$ are known on $\partial C_+^a$, it means that the following function is given:

$$\int_{-\infty}^{+\infty} (FT_a u_+)(\xi) d\xi_m.$$

So, if we denote

$$\int_{-\infty}^{+\infty} (FT_a u_+)(\xi) d\xi_m \equiv \tilde{g}(\xi'),$$

then for determining $\tilde{c}(\xi')$ we have the following equation:

$$\int_{-\infty}^{+\infty} (FT_a F^{-1})\{A_{\neq}^{-1}(\xi)\tilde{c}(\xi')\} d\xi_m = \tilde{g}(\xi'), \tag{53.6}$$

This is a convolution equation, and if evaluating the inverse Fourier transform $\xi' \to x'$, we'll obtain the conical analogue of layer potential.

### 53.6.1 Studying the Last Equation

Now we try to determine the form of the operator $FT_a F^{-1}$ (see above Sec. 3). We write

$$(FT_a F^{-1}\tilde{u})(\xi) = (FT_a u)(\xi) = \int_{\mathbf{R}^{m-1}} e^{-ia|y'|\xi_m} e^{-iy'\cdot\xi'} \hat{u}(y', \xi_m) dy', \tag{53.7}$$

where $y' = (y_1, \ldots y_{m-1})$, $\hat{u}$ is the Fourier transform of $u$ on last variable $y_m$.

We denote the convolution operator with symbol $A_{\neq}^{-1}(\xi)$ by letter $a$, so that by definition

$$(a*u)(x) = \int\limits_{\mathbf{R}^m} a(x-y)u(y)dy,$$

or, for Fourier images,

$$F(a*u)(\xi) = A_{\neq}^{-1}(\xi)\tilde{u}(\xi).$$

As above, we denote $\hat{a}(x',\xi_m)$ the Fourier transform of convolution kernel $a(x)$ on the last variable $x_m$. The integral in (53.6) takes the form (according to (53.7))

$$\int\limits_{\mathbf{R}^{m-1}} e^{-ia|y'|\xi_m} e^{-iy'\cdot\xi'}(\hat{a}*c)(y',\xi_m)dy',$$

Taking into account the properties of the convolution operator and the Fourier transform we have the following representation (see Section 53.3)

$$E_a * (A_{\neq}^{-1}(\xi)\tilde{c}(\xi')),$$

or, in more detail,

$$\int\limits_{\mathbf{R}^{m-1}} E_a(\xi'-\eta',\xi_m)A_{\neq}^{-1}(\eta',\xi_m)\tilde{c}(\eta')d\eta'.$$

Then the equation (53.6) takes the form

$$\int\limits_{\mathbf{R}^{m-1}} K_a(\eta',\xi'-\eta')\tilde{c}(\eta')d\eta' = \tilde{g}(\xi'), \tag{53.8}$$

where

$$K_a(\eta',\xi') = \int\limits_{-\infty}^{+\infty} \frac{E_a(\xi',\xi_m)}{A_{\neq}(\eta',\xi_m)}d\xi_m.$$

So, the integral equation (53.8) is an equation for determining $\tilde{c}(\xi')$. This is a conical analogue of the double-layer potential.

Suppose that we solved this equation and constructed the inverse operator $L_a$, so that $L_a\tilde{g} = \tilde{c}$. By the way, we'll note the unique solvability condition for the equation (53.8) (i.e. existence of bounded operator $L_a$) is necessary and sufficient for unique solvability for our Dirichlet boundary value problem. Using the formula (53.5) we obtain

$$\tilde{u}_+(\xi) = A_{\neq}^{-1}(\xi)(L_a\tilde{g})(\xi'),$$

or, relabeling,

$$\tilde{u}_+(\xi) = A_{\neq}^{-1}(\xi)\tilde{d}_a(\xi').$$

Then

$$u_+(x',x_m) = \int\limits_{\mathbf{R}^{m-1}} W(x'-y',x_m)d_a(y')dy', \tag{53.9}$$

where $W(x',x_m) = F^{-1}_{\xi\to x}(A_{\neq}^{-1}(\xi))$.

Formula (53.9) is the analogue of the Poisson integral for a half-space.

## 53.7 Comparison with the Half-Space Case for the Laplacian

For the half-space $x_m > 0$ we have the following (see Eskin's book [Es81]):

$$\tilde{u}_+(\xi) = \frac{\tilde{c}(\xi')}{\xi_m + i|\xi'|}.$$

If we have the Dirichlet condition on the boundary, then the function

$$\tilde{g}(\xi') = \int\limits_{-\infty}^{+\infty} \tilde{u}_+(\xi)d\xi_m$$

is given.

From the formula above we have

$$\tilde{g}(\xi') = \tilde{c}(\xi') \int\limits_{-\infty}^{+\infty} \frac{d\xi_m}{\xi_m + i|\xi'|},$$

and we need to calculate the last integral only.

For this case we can use the residue technique and find that the last integral is equal to $-\pi i$. Thus,

$$\tilde{u}_+(\xi) = -\frac{\tilde{g}(\xi')}{\pi i(\xi_m + i|\xi'|)}.$$

Consequently, our solution $u_+(x)$ is the convolution (for first $(m-1)$ variables) of the given function $g(x')$ and the kernel defined by inverse Fourier transform of

function $(\xi_m + i|\xi'|)^{-1}$ (up to a constant). The inverse Fourier transform on variable $\xi_m$ leads to the function $e^{-x_m|\xi'|}$, and further, the inverse Fourier transform $\xi' \to x'$ leads to Poisson kernel

$$P(x', x_m) = \frac{c_m x_m}{(|x'|^2 + x_m^2)^{m/2}},$$

$c_m$ is certain constant defined by Euler $\Gamma$-function.

Thus, for the solution of the Dirichlet problem in half-space $\mathbf{R}^m_+$ for the Laplacian with given Dirichlet data $g(x')$ on the boundary $\mathbf{R}^{m-1}$ we have the integral representation

$$u_+(x', x_m) = \int\limits_{\mathbf{R}^{m-1}} P(x' - y', x_m) g(y') dy'.$$

## 53.8 Oblique Derivative Problem

We go back to formula (53.5). We can write

$$\xi_k \tilde{u}_+(\xi) = \xi_k A_{\neq}^{-1}(\xi) \tilde{c}(\xi),$$

or equivalently according to Fourier transform properties

$$\frac{\partial u_+}{\partial x_m} = F^{-1}\{\xi_k A_{\neq}^{-1}(\xi) \tilde{c}(\xi)\},$$

for arbitrary fixed $k = 1, 2, ..., m$.

Further, we apply the operator $T_a$ and work as above. Our considerations will be the same, and in all places instead of $A_{\neq}^{-1}(\xi)$ will stand $\xi_k A_{\neq}^{-1}(\xi)$. We call this situation the oblique derivative problem, because $\frac{\partial}{\partial x_k}$ related to conical surface is not normal derivative exactly.

*Remark 2.* Some words on the Neumann problem. If we try to give normal derivative of our solution on conical surface different from origin, then we have the boundary value problem with variable coefficients because the boundary condition varies from one point to another one on conical surface. We need additional localization for such points to reduce it to the case of constant coefficients and consider corresponding model problem in $\mathbf{R}^m_+$. Roughly speaking, we would say that the solution looks locally different in dependence on the type of boundary point. In other words, local principle permits to work with symbols and boundary conditions independent of the space variable.

## 53.9 Conclusions

It seems that to solve explicitly the simplest boundary value problems in domains with conical point, we need to use another potentials different from classical single-layer and double-layer potentials. This fact will be shown for the Laplacian with Dirichlet condition on a conical surface by direct calculations in a future paper.

**Acknowledgements** This work was completed when the author was a DAAD stipendiat and hosted by Institute of Analysis and Algebra, Technical University of Braunschweig. He wishes to thank the DAAD and Prof. Dr. Volker Bach for their support.

## References

[Ag57] Agmon, S.: Multiple layer potentials and the Dirichlet problem for higher order elliptic equations in the plane. Commun. Pure Appl. Math. **10**, 179–239 (1957)

[Fa88] Fabes, E.: Layer potential methods for boundary value problems on lipschitz domains. Lect. Notes Math. **1344**, 55–80 (1988)

[Ke94] Kenig, K.: Harmonic Analysis Techniques for Second Order Elliptic Boundary Value Problems. CBMS Reg. Conf. Ser. Math. AMS, Providence (1994)

[MiMiTa01] Mitrea, D., Mitrea, M., Taylor M.: Layer potentials, the Hodge Laplacian and global boundary problems in nonsmooth Riemannian manifolds Mem. Amer. Math. Soc. **150**, No. 713 (2001)

[HsWe08] Hsiao, G., Wendland, W.: Boundary Integral Equations. Springer-Verlag, Berlin-Heidelberg (2008)

[Es81] Eskin, G.: Boundary Value Problems for Elliptic Pseudodifferential Equations. AMS, Providence (1981)

[Va00a] Vasil'ev, V.B.: Wave Factorization of Elliptic Symbols: Theory and Applications. Introduction to the Theory of Boundary Value Problems in Non-smooth Domains. Kluwer Academic Publishers, Dordrecht-Boston-London (2000)

[Va00b] Vasil'ev, V.B.: Wave factorization of elliptic symbols. Math. Notes **68**, 556–568 (2000)

[Va11] Vasilyev, V.B.: Elliptic equations and boundary value problems in non-smooth domains. Pseudo Differential Operators: Analysis, Applications and Computations. Eds. Rodino L., Wong M.W., Zhu H. Operator Theory: Advances and Applications, 2011, V.213. Birkhauser, Basel. P.105–121.

[Va10] Vasilyev, V.B.: Multipliers of Fourier Integrals, Pseudo Differential Equations, Wave Factorization, Boundary Value Problems. Editorial URSS, Moscow, 2nd edition (2010) (in Russian)

[GeSh59] Gel'fand, I.M. and Shilov, G.E.: Distributions and Operations with them. Fizmatgiz, Moscow (1959) (in Russian)

# Chapter 54
# Optimal Control of Partial Differential Equations by Means of Stackelberg Strategies: An Environmental Application

**M.E. Vázquez-Méndez, L.J. Alvarez-Vázquez, N. García-Chan, and A. Martínez**

## 54.1 Mathematical Formulation of the Physical Problem

A sewage depuration system consists of a small number of treatment plants—collecting the wastewater from an urban area—where purification treatments are applied to discharge the final effluent—through an outfall—at some point of a domain occupied by shallow water: river, lake, estuary, etc. The location of these discharges and the intensities of the treatments in each plant are crucial points in order to protect the ecosystem in the water domain. Moreover, these plants can be built and/or managed by different agents (industries, municipalities, local governments, etc.), which causes a wide range of options to define a final optimal strategy.

When a unique agent is concerned, the situation can be formulated as an optimal control problem of partial differential equations [MaRoVa00, AlEtAl02], but if the system is managed by several organizations, the problem becomes multi-objective, and the concept of *optimal strategy* depends on the particularities of the stakeholders. Usually, these multi-objective problems are approached from a cooperative viewpoint (Pareto-optimal solutions) [AlEtAl10] or from a non-cooperative one (Nash equilibria) [GaMuVa09]. However, in this chapter, we propose an alternative, more realistic, hierarchical framework: in charge of the

M.E. Vázquez-Méndez (✉)
University of Santiago de Compostela, Lugo, Spain
e-mail: miguelernesto.vazquez@usc.es

L.J. Alvarez-Vázquez • A. Martínez
University of Vigo, Vigo, Spain
e-mail: lino@dma.uvigo.es; aurea@dma.uvigo.es

N. García-Chan
University of Guadalajara, Guadalajara, Mexico
e-mail: nestorg.chan@red.cucei.udg.mx



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_54

construction of the plant there is a higher organism (for instance, a regional government), which pursues global goals, and the plant manager (responsible for determining the intensity of treatments to be developed) is a subordinate entity (for instance, a council government), pursuing only local targets. A strategy good enough for the upper entity (the *leader*) and the subordinate (the *follower*) is said a Stackelberg strategy [St52]. This type of strategies is widely used in economics. Nevertheless, its application to multi-objective optimal control problems governed by partial differential equations has been, as far as we know, very limited.

So, let us consider a domain $\Omega \subset \mathbb{R}^2$, with a smooth enough boundary $\partial\Omega$, occupied by shallow water in which are released, through submarine outfalls, wastewater discharges. The environmental impact of these discharges can be controlled through the concentration level of fecal coliforms (FC) at each point of the domain. The concentration $\rho(x,t)$ of FC at point $x \in \Omega$ and time $t \in [0,T]$ can be obtained by solving the problem [GaMuVa09]

$$\left.\begin{array}{ll} \dfrac{\partial \rho}{\partial t} + \vec{u}\cdot\nabla\rho - \beta\Delta\rho + \kappa\rho = \dfrac{1}{h}\left(m(t)\delta_b(x) + \displaystyle\sum_{j=1}^{N_P} m_j(t)\delta_{c_j}(x)\right) & \text{in } \mathscr{O}, \\ \rho(x,0) = \rho_0(x) & \text{in } \Omega, \\ \dfrac{\partial \rho}{\partial n} = 0 & \text{on } \Gamma, \end{array}\right\} \tag{54.1}$$

where $h(x,t)$ and $\vec{u}(x,t)$ are, respectively, the height and the horizontal velocity of water, $\beta > 0$ is a viscosity coefficient, $\kappa$ is an experimental coefficient related to FC loss rate, $b, c_1, \ldots, c_{N_P} \in \Omega$ are the points where wastewater is discharged and $m(t), m_1(t), \ldots, m_{N_P}(t)$ are, respectively, the mass flow of FC discharged at those point, $\delta_p(x)$ represents the Dirac measure located at point $p \in \Omega$, $\rho_0(x) \in \mathscr{C}(\bar{\Omega})$ gives the concentration of FC at initial time, and, finally, $\mathscr{O} = \Omega \times (0,T)$ and $\Gamma = \partial\Omega \times (0,T)$.

Let us assume a depuration system consisting of $N_P$ plants already operating (that is, points $c_1, \ldots, c_{N_P} \in \Omega$, and discharges $m_1(t), \ldots, m_{N_P}(t)$ are fixed and known) and of a new plant, located at a point $a \in \partial\Omega$. This plant will be connected, through a submarine outfall to be built, with a discharge point $b \in \Omega$ (to be determined), and at that point will be released, at each time $t \in (0,T)$, a FC flow $m(t)$ (also to be determined).

The choice of $b \in \Omega$ and $m(t) \in L^\infty(0,T)$ (the controls in our problem) will be done trying to optimize two objectives:

1. First, the organization responsible for choosing the new point (hereinafter the leader) seeks the length of the outfall and the concentration of FC in a certain protected area $A_L \subset \Omega$ to be as small as possible. Consequently, its goal is to minimize the functional

$$J_L(b,m) = \frac{1}{2}\|b-a\|^2 + \frac{\varepsilon_L}{T|A_L|}\int_0^T \int_{\bar{A}_L} \rho(x,t)\,dxdt, \tag{54.2}$$

where $\varepsilon_L$ is a penalty parameter and $|A_L|$ denotes the area of $A_L$.

2. Second, the organization in charge of managing the plant (hereinafter the follower) seeks to reduce economic costs, both from the treatments to be performed in the depuration plant and from the penalties to pay if the concentration of FC in its small influence area, $A_F \subset \Omega$, is greater than a certain threshold $\sigma_F$. Thus, its aim is to minimize the functional

$$J_F(b,m) = \int_0^T f(m(t))\,dt + \frac{\varepsilon_F}{2}\int_0^T \int_{\bar{A}_F} (\rho(x,t) - \sigma_F)_+^2\,dxdt, \tag{54.3}$$

where $f : (0,\infty) \longrightarrow \mathbb{R}$ measures treatment cost, $\varepsilon_F$ is a penalty parameter, and $(\cdot)_+$ denotes the positive part function $(y)_+ = \max\{y,0\}$.

Evidently the problem presents technological constraints limiting both the points where the outfall can be located, and the treatment intensities that can be applied in the plant. So, we denote by $U_{ad} \subset \Omega$ the set of admissible points for placing the outfall, and, for given values $0 < \underline{m} < \overline{m}$, we define as $M_{ad} = \{m \in L^\infty(0,T) : \underline{m} \leq m(t) \leq \overline{m}$ a.e. in $(0,T)\}$ the set of feasible discharges.

The ideal (or *utopic*) solution consists of finding a control $(b^I, m^I) \in U_{ad} \times M_{ad}$ minimizing simultaneously above functionals $J_L$ and $J_F$. However, this control rarely exists, due to the opposite character of both functionals. To address and solve this type of multi-objective problems there exist different strategies. Since the organization responsible for the location of the outfall acts as leader, choosing first, and the manager of the plant, knowing the decision of the leader, will act accordingly, seeking the discharge that favors its own interests more, in this work we will use Stackelberg strategies [St52] to find the optimal solution of the problem. To do this, we first consider the *follower problem*:

$$\text{For a given } b \in U_{ad}, \text{ find } \min_{m \in M_{ad}} J_F(b,m). \tag{54.4}$$

Let us assume that, for each $b \in U_{ad}$, the problem (54.4) has a unique solution denoted by $m_b \in M_{ad}$, and consider the *leader problem*:

$$\min_{b \in U_{ad}} J_L(b, m_b), \text{ where } m_b \in M_{ad} \text{ is the solution of (54.4).} \tag{54.5}$$

**Definition 1.** We say that $(b^*, m^*) \in U_{ad} \times M_{ad}$ is a Stackelberg strategy (solution of our optimal control problem) if and only if

1. $m^*$ is the best response of the follower to the leader choice $b^*$, that is, $m^*$ is the solution of the problem (54.4) for $b^* \in U_{ad}$, or, equivalently, $m^* = m_{b^*}$.
2. $b^*$ is the best choice of the leader, that is, $b^*$ is the solution of the problem (54.5).

In the next sections we will analyze in detail the problems (54.4) and (54.5). In particular, we will prove that, for each $b \in U_{ad}$, the problem (54.4) admits a unique solution (which guarantees that the above definition is correct), and we will demonstrate that our problem admits, at least, one Stackelberg strategy (the problem (54.5) admits, at least, one solution). Finally, in order to characterize the optimal

solutions, we will derive a first order optimality system. In these results we will obtain expressions for the gradients of the objective functions in both problems, which will be very useful for the numerical resolution of the problem (to be addressed in a forthcoming paper).

## 54.2 Analysis of the Follower Problem

We begin this section studying the existence and regularity of solutions to the state system (54.1). A solution of (54.1) can be defined by transposition techniques in the following sense (see [MaRoVa00] for details):

**Definition 2.** For given $r,s \in [1,2)$, $\frac{2}{r}+\frac{2}{s}>3$, we say that $\rho \in L^r(0,T;W^{1,s}(\Omega))$ is a solution of problem (54.1) if, for each $\Phi \in \mathscr{C}^1(\bar{\Omega}\times[0,T])$ such that $\Phi(\cdot,T)=0$, it verifies that

$$\int_0^T \int_\Omega \left(-\frac{\partial \Phi}{\partial t}\rho + \beta \nabla \Phi \nabla \rho + \vec{u}\Phi\nabla\rho + \kappa\Phi\rho\right) dxdt = \int_\Omega \Phi(x,0)\rho_0(x)\,dx + \int_0^T \frac{1}{h(b,t)}\Phi(b,t)m(t)\,dt + \sum_{j=1}^{N_P}\int_0^T \frac{1}{h(c_j,t)}\Phi(c_j,t)m_j(t)\,dt.$$

Then we have the following result (cf. [MaRoVa00] and [AlEtAl02]):

**Theorem 1.** *Let $\Omega$ be a bounded domain with a smooth enough boundary $\partial\Omega$. We consider $\vec{u} \in [L^\infty(0,T;W^{1,\infty}(\Omega))]^2$ and $h \in \mathscr{C}(\bar{\Omega}\times[0,T])$ satisfying $h(x,t)\geq \alpha > 0$, $\forall (x,t)\in \bar{\Omega}\times[0,T]$. Then*

1. *There exists a unique function $\rho \in [L^r(0,T;W^{1,s}(\Omega)) \cap L^2(0,T;L^2(\Omega))]$ with $\frac{\partial \rho}{\partial t} \in L^r(0,T;(W^{1,s'}(\Omega))')$, solution of (54.1) and verifying*

$$\int_0^T \langle -\frac{\partial \Phi}{\partial t} - \beta\Delta\Phi - div(\Phi\,\vec{u}) + \kappa\Phi, \rho\rangle\,dt = \int_\Omega \Phi(x,0)\rho_0(x)dx + \int_0^T \frac{1}{h(b,t)}\Phi(b,t)m(t)dt + \sum_{j=1}^{N_P}\int_0^T \frac{1}{h(c_j,t)}\Phi(c_j,t)m_j(t)dt,$$

*for all $\Phi \in L^2(0,T;H^2(\Omega))\cap H^1(0,T;L^2(\Omega))$ such that*

$$\beta\frac{\partial\Phi}{\partial n} + \Phi\vec{u}\cdot\vec{n} = 0 \ \text{ on } \partial\Omega\times(0,T), \quad \Phi(\cdot,T)=0 \ \text{ in } \Omega.$$

2. *If there exists a closed set $E\subset\Omega$ such that $\Omega\backslash E$ is a smooth enough domain, $\{b,c_1,\ldots,c_{N_P}\}\subset E$ and $\bar{A}_L\cup\bar{A}_F\subset\Omega\backslash E$, then $\rho_{|(\bar{A}_L\cup\bar{A}_F)\times[0,T]} \in \mathscr{C}((\bar{A}_L\cup\bar{A}_F)\times[0,T])$, so that the functional*

$$\begin{aligned} F : U_{ad} \times M_{ad} &\longrightarrow \mathscr{C}((\bar{A}_L \cup \bar{A}_F) \times [0,T]) \\ (b,m) &\longrightarrow F(b,m) = \rho_{|(\bar{A}_L \cup \bar{A}_F) \times [0,T]} \end{aligned}$$

*is well defined, continuous and, moreover,*

*a. for each $b \in U_{ad}$, the function*

$$F(b,\cdot) : m \in M_{ad} \longrightarrow F(b,m) \in \mathscr{C}((\bar{A}_L \cup \bar{A}_F) \times [0,T])$$

*is affine, and Gâteaux (also Frechet) differentiable;*

*b. if $h \in \mathscr{C}([0,T];\mathscr{C}^1(\Omega))$, then, for each $m \in M_{ad}$, the function*

$$F(\cdot,m) : b \in U_{ad} \longrightarrow F(b,m) \in \mathscr{C}((\bar{A}_L \cup \bar{A}_F) \times [0,T])$$

*is Gâteaux differentiable.*

Let us analyze now the cost functional $J_F$, formally introduced in (54.3). Under the hypotheses of Theorem 1 we have that $J_F$ is well defined in $U_{ad} \times M_{ad}$, and can be written as $J_F(b,m) = \Theta(m) + H_F(F(b,m))$, with

- $\Theta : m \in M_{ad} \longrightarrow \Theta(m) \in \mathbb{R}$, given by $\Theta(m) = \displaystyle\int_0^T f(m(t))dt$,
- $H_F : \rho \in \mathscr{C}((\bar{A}_L \cup \bar{A}_F) \times [0,T]) \longrightarrow H_F(\rho) \in \mathbb{R}$, defined by

$$H_F(\rho) = \frac{\varepsilon_F}{2} \int_0^T \int_{\bar{A}_F} (\rho(x,t) - \sigma_F)_+^2 \, dxdt.$$

Bearing in mind that, since $M_{ad}$ is a bounded subset of $L^\infty(0,T)$, on $M_{ad}$ the weak* topology of $L^\infty(0,T)$ and the weak topology of $L^2(0,T)$ are equivalent, we have the following result.

**Theorem 2.** *Under the hypotheses of Theorem 1, if $f \in \mathscr{C}(0,\infty)$ is strictly convex, then, for each $b \in U_{ad}$, the problem (54.4) has a unique solution $m_b$. Moreover, if $f \in \mathscr{C}^1(0,\infty)$, then $m_b \in M_{ad}$ is the solution of (54.4) if and only if there exist $\rho \in [L^r(0,T;W^{1,s}(\Omega)) \cap L^2(0,T;L^2(\Omega))]$ ($r,s \in [1,2)$, $\frac{2}{r} + \frac{2}{s} > 3$), solution of the state system (54.1), and $q \in W^{1,\infty}([0,T];L^\infty(\Omega)) \cap L^\infty([0,T];W^{2,\infty}(\Omega))$, solution of the adjoint system*

$$\left.\begin{array}{ll} -\dfrac{\partial q}{\partial t} - \beta \Delta q - div(q\vec{u}) + \kappa q = \varepsilon_F \chi_{\bar{A}_F}(\rho - \sigma_F)_+ & in\ \mathscr{O}, \\ q(x,T) = 0 & in\ \Omega, \\ \beta \dfrac{\partial q}{\partial n} + q(\vec{u} \cdot \vec{n}) = 0 & on\ \Gamma, \end{array}\right\} \tag{54.6}$$

*satisfying the optimality condition*

$$\int_0^T \left( f'(m_b(t)) + \frac{1}{h(b,t)} q(b,t) \right) (m(t) - m_b(t))\, dt \geq 0, \quad \forall m \in M_{ad}, \tag{54.7}$$

*where $\chi_{\bar{A}_F}$ denotes the characteristic function of set $\bar{A}_F$.*

*Proof.* If the function $f$ is continuous and strictly convex, then so it the functional $\Theta$. In addition, $H_F$ is clearly continuous, convex and Gâteaux differentiable. Thus, as for each $b \in U_{ad}$ the function $F(b,\cdot)$ is continuous and affine (see Theorem 1), the functional

$$J_F(b,\cdot) : m \in M_{ad} \longrightarrow J_F(b,m) = \Theta(m) + H_F(F(b,m)) \in \mathbb{R} \tag{54.8}$$

is continuous and strictly convex (and, consequently, lower semicontinuous in the weak topology of $L^2(0,T)$—see, for instance, Corollary 3.9 of [Br11]). Furthermore, $M_{ad}$ is bounded, convex and closed in $L^2(0,T)$ (and therefore also weakly closed; see, for example, Theorem 3.7 of [Br11]). A classic result of optimization in Banach spaces (see, for instance, [Ce71]) ensures that the problem (54.4) has a solution. Finally, this solution is unique since the functional (54.8) is strictly convex.

So, we denote $m_b \in M_{ad}$ the unique solution of the problem (54.4) for a given $b \in U_{ad}$. We try now to characterize it. First, we note that, given $(b, m_b) \in U_{ad} \times M_d$, the existence of the function $\rho \in [L^r(0,T; W^{1,s}(\Omega)) \cap L^2(0,T; L^2(\Omega))]$ solution of (54.1) is guaranteed by Theorem 1. Moreover, $\rho_{|(\bar{A}_L \cup \bar{A}_F) \times [0,T]} \in \mathscr{C}((\bar{A}_L \cup \bar{A}_F) \times [0,T])$; therefore, there is a unique $q \in W^{1,\infty}([0,T]; L^\infty(\Omega)) \cap L^\infty([0,T]; W^{2,\infty}(\Omega))$ satisfying (54.6) (see, for instance, Section 9 of Chapter IV of [LaSoUr68]).

On the other hand, if $f \in \mathscr{C}^1(0,\infty)$, then $\Theta$ is Gâteaux differentiable (see [GaMuVa09]) and

$$\langle D\Theta(m_b), m \rangle = \int_0^T f'(m_b(t)) m(t) dt, \quad \forall m \in M_{ad}.$$

Moreover, since $F(b,\cdot)$ is affine, continuous and Gâteaux differentiable (see Theorem 1) and $H_F$ is Gâteaux differentiable, the composition $H_F \circ F(b,\cdot)$ is also Gâteaux differentiable, and

$$\begin{aligned} &\langle D(H_F \circ F(b,\cdot))(m_b), m \rangle \\ &\quad = \varepsilon_F \int_0^T \int_{A_F} (\rho(x,t) - \sigma_F)_+ \delta\rho(x,t)\, dx dt, \quad \forall m \in M_{ad}, \end{aligned}$$

where $\delta\rho(x,t)$ is the solution of the linearized problem

$$\left.\begin{array}{ll} \dfrac{\partial \delta\rho}{\partial t} + \vec{u} \cdot \nabla \delta\rho - \beta \Delta \delta\rho + \kappa \delta\rho = \dfrac{1}{h} m(t) \delta_b(x) & \text{in } \mathscr{O}, \\ \delta\rho(x,0) = 0 & \text{in } \Omega, \\ \dfrac{\partial \delta\rho}{\partial n} = 0 & \text{on } \Gamma. \end{array}\right\} \tag{54.9}$$

Then the functional $J_F(b,\cdot)$ is Gâteaux differentiable, and

$$\langle DJ_F(b,\cdot)(m_b),m\rangle = \int_0^T f'(m_b(t))m(t)\,dt$$
$$+\varepsilon_F \int_0^T \int_{A_F} (\rho(x,t)-\sigma_F)_+ \delta\rho(x,t)\,dxdt, \quad \forall m \in M_{ad}.$$

It is worth mentioning here that the last term of the previous expression can be simplified by adjoint techniques. In effect, multiplying the first equation of (54.9) by the solution $q$ of (54.6), and integrating over $\Omega \times (0,T)$, we obtain

$$\int_0^T \int_\Omega q\Big[\frac{\partial \delta\rho}{\partial t} + \vec{u}\cdot\nabla\delta\rho - \beta\Delta\delta\rho + \kappa\delta\rho\Big]dxdt = \int_0^T \frac{1}{h(b,t)} q(b,t)m(t)\,dt.$$

Taking into account boundary and initial conditions of system (54.9), using Green's formula, and integrating by parts, we have

$$\int_0^T \int_\Omega \left(-\frac{\partial q}{\partial t} - div(q\vec{u}) - \beta\Delta q + \kappa q\right)\delta\rho\,dxdt + \int_\Omega q(x,T)\delta\rho(x,T)\,dx$$
$$+\int_0^T \int_{\partial\Omega} \left(\beta\frac{\partial q}{\partial n} + q\vec{u}\cdot\vec{n}\right)\delta\rho\,d\gamma dt = \int_0^T \frac{1}{h(b,t)} q(b,t)m(t)\,dt$$

and, because of $q(x,t)$ is the solution of (54.6), it becomes

$$\varepsilon_F \int_0^T \int_{A_F} (\rho(x,t)-\sigma_F)_+ \delta\rho(x,t)\,dxdt = \int_0^T \frac{1}{h(b,t)} q(b,t)m(t)\,dt$$

Thus, we obtain the expression

$$\langle DJ_F(b,\cdot)(m_b),m\rangle = \int_0^T f'(m_b(t))m(t)\,dt + \int_0^T \frac{1}{h(b,t)} q(b,t)m(t)\,dt \tag{54.10}$$

Finally, since $M_{ad}$ is convex and $J_F(b,\cdot)$ is Gâteaux differentiable and strictly convex, we know that (see, for instance, Chapter IV of [Ce71]) $m_b \in M_{ad}$ is the solution of problem (54.4) if and only if

$$\langle DJ_F(b,\cdot)(m_b), m-m_b\rangle \geq 0, \quad \forall m \in M_{ad}$$

and then, by using the expression (54.10), we obtain (54.7) and conclude the proof.

## 54.3 Analysis of the Leader Problem

Theorem 2 ensures that the problem (54.5) is well posed, and allows us to define the functional

$$T : b \in U_{ad} \longrightarrow T(b) = m_b \in M_{ad}, \tag{54.11}$$

where, as above commented, $m_b$ is the solution of the problem (54.4). In order to prove that the function $T$ is continuous, it will be very useful the following Lemma, whose demonstration can be seen in [GaMuVa09].

**Lemma 1.** *Let $b \in U_{ad}$ and let $\{m^n\} \subset M_{ad}$ be a sequence such that $\{m^n\} \overset{*}{\rightharpoonup} m$ (convergence in the weak* topology of $L^\infty(0,T)$). Under the hypotheses of Theorem 1 we have that*

$$\{F(b,m^n)\} \to F(b,m) \ \text{ in } \mathscr{C}((\bar{A}_L \cup \bar{A}_F) \times [0,T]).$$

Thus, we can prove here the following result:

**Theorem 3.** *Let us suppose that $f \in \mathscr{C}(0,\infty)$ is strictly convex and that the hypotheses of Theorem 1 are satisfied. Then the functional $T$, given by (54.11), is continuous considering in $M_{ad}$ the weak* topology of $L^\infty(0,T)$.*

*Proof.* Let $\{b^n\} \subset U_{ad}$ and $b \in U_{ad}$ be such that $\{b^n\} \to b$. We are going to prove that, necessarily, $\{T(b^n)\} \overset{*}{\rightharpoonup} T(b) \in M_{ad}$.

Since $\{T(b^n)\} \subset M_{ad}$ is bounded in $L^\infty(0,T)$, there exist a subsequence $\{T(b^{n'})\}$ and an element $\hat{m} \in L^\infty(0,T)$ such that (see, for instance, Corollary 3.30 of [Br11]):

$$\{T(b^{n'})\} \overset{*}{\rightharpoonup} \hat{m} \tag{54.12}$$

Since $M_{ad}$ is closed in the weak* topology of $L^\infty(0,T)$ (as a consequence of the fact that $M_{ad}$ is closed in $L^2(0,T)$ and convex, so also weakly closed in $L^2(0,T)$—cf., for example, Theorem 3.7 of [Br11]), we have that $\hat{m} \in M_{ad}$. If we prove now that $\hat{m} = T(b)$ then, by the uniqueness of limit, we will have that the whole sequence $\{T(b^n)\}$ (not only the subsequence) converges to $T(b)$, which will conclude the demonstration.

In order to see that $\hat{m} = T(b)$, by the definition of $T$, it suffices to prove that

$$J_F(b,\hat{m}) \leq J_F(b,m), \quad \forall m \in M_{ad}. \tag{54.13}$$

Considering (54.12) we can say that

- Since $\Theta$ is lower semicontinuous in the weak* topology of $L^\infty(0,T)$ (it is continuous and convex, so it is lower semicontinuous in the weak topology of $L^2(0,T)$, which coincides with the weak* one of $L^\infty(0,T)$), then

$$\Theta(\hat{m}) \leq \lim_{n' \to \infty} \inf \Theta(T(b^{n'})). \tag{54.14}$$

- Since $H_F$ is continuous, from Lemma 1,

$$H_F(F(b,\hat{m})) = \lim_{n' \to \infty} H_F(F(b, T(b^{n'}))). \tag{54.15}$$

Moreover, taking into account that $\{b^n\} \to b$ and that $F(\cdot, m)$ is continuous for any $m \in M_{ad}$, we have

$$H_F(F(b,T(b^{n'}))) = \lim_{n \to \infty} H_F(F(b^n, T(b^{n'}))). \tag{54.16}$$

Using now (54.14), (54.15) and (54.16), and bearing in mind that $F$ is continuous, we obtain

$$\begin{aligned} J_F(b,\hat{m}) &= \Theta(\hat{m}) + H_F(F(b,\hat{m})) \\ &\leq \lim_{n' \to \infty} \inf \Theta(T(b^{n'}) + \lim_{n' \to \infty} H_F(F(b^{n'}, T(b^{n'}))) \\ &= \lim_{n' \to \infty} J_F(b^{n'}, T(b^{n'})) \\ &\leq \lim_{n' \to \infty} J_F(b^{n'}, m), \quad \forall m \in M_{ad}, \end{aligned} \tag{54.17}$$

where the last inequality is a direct consequence of the definition of $T$. Again, taking into account the continuity of $H_F$ and $F(\cdot, m)$, we have

$$\begin{aligned} \lim_{n' \to \infty} J_F(b^{n'}, m) &= \Theta(m) + \lim_{n' \to \infty} H_F(F(b^{n'}, m)) \\ &= \Theta(m) + H_F(F(b,m)) = J_F(b,m) \end{aligned} \tag{54.18}$$

Finally, combining (54.17) and (54.18) we deduce (54.13), which concludes the proof.

If we consider the composition of functions

$$\begin{array}{ccccc} U_{ad} & \overset{1_{U_{ad}} \times T}{\longrightarrow} & U_{ad} \times M_{ad} & \overset{J_L}{\longrightarrow} & \mathbb{R} \\ b & \longrightarrow & (b, m_b) & \longrightarrow & J_L(b, m_b) \end{array}$$

and define $J = J_L \circ (1_{U_{ad}} \times T)$, then the problem (54.5) can be rewritten as

$$\min_{b \in U_{ad}} J(b) \tag{54.19}$$

**Theorem 4.** *Let us assume that $f \in \mathscr{C}(0,\infty)$ is strictly convex and that the hypotheses of Theorem 1 are satisfied. Then the functional $J$ is continuous.*

*Proof.* The functional $J = J_L \circ (1_{U_{ad}} \times T)$ is given by

$$J(b) = \frac{1}{2}\|b-a\|^2 + H_L(F(b,T(b))), \tag{54.20}$$

where

$$H_L : \rho \in \mathscr{C}((\bar{A}_L \cup \bar{A}_F) \times [0,T]) \longrightarrow H_L(\rho) = \frac{\varepsilon_L}{T|A_L|}\int_0^T \int_{\bar{A}_L} \rho(x,t)\,dxdt \in \mathbb{R}$$

is trivially continuous. Then the continuity of $J$ follows from the continuity of $F$ and $T$ (Theorems 1 and 3) and from Lemma 1:

Let $\{b^n\} \subset U_{ad}$ such that $\{b^n\} \to b^*$. From Theorem 3, $\{T(b^n)\} \overset{*}{\rightharpoonup} T(b^*)$, and from Lemma 1 we obtain that $\{F(b,T(b^n))\} \to F(b,T(b^*))$, $\forall b \in U_{ad}$. Then, from the continuity of $F$ and $H$, we have

$$\lim_{n\to\infty} H_L(F(b^n,T(b^n))) = H_L(F(b^*,T(b^*)))$$

and, finally,

$$\begin{aligned}\lim_{n\to\infty} J(b^n) &= \lim_{n\to\infty}\left(\frac{1}{2}\|b^n-a\|^2 + H_L(F(b^n,T(b^n)))\right)\\ &= \frac{1}{2}\|b^*-a\|^2 + H_L(F(b^*,T(b^*))) = J(b^*),\end{aligned}$$

which ensures the continuity of $J$.

As a direct consequence of above result, we obtain the existence of, at least, one Stackelberg strategy for our problem:

**Corollary 1.** *Under the hypotheses of Theorem 1, if $f \in \mathscr{C}(0,\infty)$ is strictly convex, and $U_{ad} \in \Omega$ is closed, then the problem (54.19)—or, equivalently, the problem (54.5)—admits at least one solution.*

Assuming additional regularity on functions $f$ and $h$ we are able to derive a first-order optimality condition for problem (54.19). With this idea in mind, and searching for a new simpler expression for functional $J$, given by (54.20), we introduce the following problem (adjoint state for the leader problem):

$$\left.\begin{aligned} -\frac{\partial p}{\partial t} - \beta\Delta p - div(p\vec{u}) + \kappa p &= \frac{1}{T|A_L|}\chi_{\bar{A}_L} && \text{in } \mathscr{O},\\ p(x,T) &= 0 && \text{in } \Omega,\\ \beta\frac{\partial p}{\partial n} + p(\vec{u}\cdot\vec{n}) &= 0 && \text{on } \Gamma. \end{aligned}\right\} \tag{54.21}$$

This problem admits a unique solution (see [LaSoUr68], Ch. IV, Sect. 9) $p \in W^{1,\infty}([0,T];L^\infty(\Omega)) \cap L^\infty([0,T];W^{2,\infty}(\Omega))$. Then, from (54.1) and (54.21), using Green's formula and integration by parts, we obtain the equalities

$$\begin{aligned}
H_L(\rho) &= \frac{\varepsilon_L}{T|A_L|}\int_0^T\int_{\bar{A}_L}\rho(x,t)\,dxdt = \varepsilon_L\int_0^T\int_\Omega \frac{1}{T|A_L|}\chi_{\bar{A}_L}\rho(x,t)\,dxdt \\
&= \varepsilon_L\int_0^T\int_\Omega\left(-\frac{\partial p}{\partial t} - \beta\Delta p - div(p\vec{u}) + \kappa p\right)\rho(x,t)\,dxdt \\
&= \varepsilon_L\left[\int_0^T\int_\Omega\left(\frac{\partial\rho}{\partial t} + \vec{u}\cdot\nabla\rho - \beta\Delta\rho + \kappa\rho)\right)p(x,t)\,dxdt\right. \\
&\quad + \int_0^T\int_{\partial\Omega}\left(-\rho\, p\,\vec{u}\cdot\vec{n} + \beta\, p\,\frac{\partial\rho}{\partial n} - \beta\,\rho\,\frac{\partial p}{\partial n}\right)d\gamma dt \\
&\quad \left. - \int_\Omega \rho(x,T)p(x,T)\,dx + \int_\Omega \rho(x,0)p(x,0)\,dx\right] \\
&= \varepsilon_L\left[\int_0^T\int_\Omega \frac{p}{h}(x,t)\left(m_b(t)\delta_b(x) + \sum_{j=1}^{N_P} m_j(t)\delta_{c_j}(x)\right)dxdt\right. \\
&\quad \left. + \int_\Omega \rho_0(x)p(x,0)\,dx\right]
\end{aligned}$$

In this way, expression (54.20) turns into

$$J(b) = \frac{1}{2}\|b-a\|^2 + \varepsilon_L\left(\int_0^T \frac{p(b,t)}{h(b,t)}m_b(t)\,dt + C\right), \tag{54.22}$$

where

$$C = \sum_{j=1}^{N_P}\int_0^T \frac{p(c_j,t)}{h(c_j,t)}m_j(t)\,dt + \int_\Omega \rho_0(x)p(x,0)\,dx.$$

In order to establish the optimality condition, we need to recall the following basic property of real functions:

**Lemma 2.** *If $f \in \mathscr{C}^2(0,\infty)$ is strictly convex, and $D \subset \mathbb{R}$ denotes the image of $f'$ (that is, $D = \{y \in \mathbb{R} : \exists x \in (0,\infty) \,/\, y = f'(x)\}$), then $f'$ is invertible, and the inverse function $g = (f')^{-1} \in \mathscr{C}^1(D)$ verifies:*

$$g'(f'(x)) = \frac{1}{f''(x)}, \quad \forall x \in (0,\infty).$$

Then we have the following result, which states a condition that must be satisfied by any Stackelberg strategy for our problem.

**Theorem 5.** *Let us assume that all the hypotheses in Theorem 1 are satisfied, and, furthermore, that $U_{ad}$ is a convex set, $h \in \mathscr{C}([0,T];\mathscr{C}^1(\bar{\Omega}))$ and $f \in \mathscr{C}^2(0,\infty)$ is strictly convex. Then, if $(b,m_b) \in U_{ad} \times \mathring{M}_{ad}$ is a Stackelberg strategy for our problem, there exist $\rho \in [L^r(0,T;W^{1,s}(\Omega)) \cap L^2(0,T;L^2(\Omega))]$ $(r,s \in [1,2)$, $\frac{2}{r}+\frac{2}{s}>3)$ and $q \in W^{1,\infty}([0,T];L^\infty(\Omega)) \cap L^\infty([0,T];W^{2,\infty}(\Omega))$ satisfying (54.1), (54.6) and*

$$f'(m_b(t)) + \frac{1}{h(b,t)} q(b,t) = 0, \quad \forall t \in (0,T). \tag{54.23}$$

*In addition, if $p \in W^{1,\infty}([0,T];L^\infty(\Omega)) \cap L^\infty([0,T];W^{2,\infty}(\Omega))$ is the solution of (54.21) then, for all $b^* = (b_1^*, b_2^*) \in U_{ad}$, it verifies*

$$\sum_{i=1}^{2} \left( b_i - a_i + \varepsilon_L \int_0^T \left( \frac{h\frac{\partial p}{\partial b_i} - p\frac{\partial h}{\partial b_i}}{h^2}(b,t)\, m_b(t) \right.\right. \tag{54.24}$$

$$\left.\left. - \frac{p\left(h\frac{\partial q}{\partial b_i} - q\frac{\partial h}{\partial b_i}\right)}{h^3}(b,t)\, \frac{1}{f''(m_b(t))} \right) dt \right) (b_i^* - b_i) \geq 0.$$

*Proof.* Since $m_b \in \mathring{M}_{ad}$, the existence of $\rho$ and $q$ satisfying (54.1), (54.6) and (54.23) can be directly obtained from Theorem 2. Moreover, since $f \in \mathscr{C}^2(0,\infty)$ is strictly convex, we know that $f'$ admits an inverse $g = (f')^{-1}$, allowing us to clear $m_b(t)$ in (54.23) and write

$$m_b(t) = g\left(-\frac{q}{h}(b,t)\right), \quad \forall t \in (0,T).$$

Taking this expression to (54.22), we have

$$J(b) = \frac{1}{2}\|b-a\|^2 + \varepsilon_L \left( \int_0^T \frac{p}{h}(b,t)\, g\left(-\frac{q}{h}(b,t)\right) dt + C \right),$$

from where, applying Lemma 2 and the chain rule, we obtain

$$\frac{\partial J}{\partial b_i}(b) = b_i - a_i$$

$$+ \varepsilon_L \int_0^T \left( \frac{h\frac{\partial p}{\partial b_i} - p\frac{\partial h}{\partial b_i}}{h^2}(b,t)\, m_b(t) - \frac{p\left(h\frac{\partial q}{\partial b_i} - q\frac{\partial h}{\partial b_i}\right)}{h^3}(b,t)\, \frac{1}{f''(m_b(t))} \right) dt.$$

Finally, if $(b, m_b)$ is a Stackelberg strategy for our problem, then $b \in U_{ad}$ is a solution of (54.19) and, consequently,

$$\sum_{i=1}^{2} \left( \frac{\partial J}{\partial b_i}(b) \right) (b_i^* - b_i) \geq 0, \quad \forall b^* = (b_1^*, b_2^*) \in U_{ad},$$

which leads to (54.24), and concludes the proof.

**Acknowledgements** This work was supported by Project MTM2012-30842 of M.E.C. (Spain) and FEDER. The third author also thanks the support from Sistema Nacional de Investigadores SNI-52768 and Programa de Mejoramiento del Profesorado PROMEP/103.5/13/6219 (Mexico).

## References

[AlEtAl10] Alvarez-Vázquez, L.J., García-Chan, N., Martínez, A., Vázquez-Méndez, M.E.: Multi-objective Pareto-optimal control: An application to waste-water management. Comput. Optim. Appl. **46**, 135–157 (2010)

[AlEtAl02] Alvarez-Vázquez, L.J, Martínez, A., Rodríguez, C., Vázquez-Méndez, M.E.: Mathematical analysis of the optimal location of wastewater outfalls. IMA J. Appl. Math. **67**, 23–39 (2002)

[Br11] Brezis, H.: Functional analysis, Sobolev spaces and partial differential equations. Springer, New York (2011)

[Ce71] Cea, J.: Optimisation: Theorie et algorithmes. Dunod, Paris (1971)

[GaMuVa09] García-Chan, N., Muñoz-Sola, R., Vázquez-Méndez, M.E.: Nash equilibrium for a multiobjective control problem related to wastewater management. ESAIM: Control Optim. Calc. Var. **15**, 581–588 (2009)

[LaSoUr68] Ladyzenskaja, O.A., Solonnikov, V.A., Uralceva, N.N.: Linear and quasilinear equations of parabolic type. Amer. Math. Soc., Providence (1968)

[MaRoVa00] Martínez, A., Rodríguez, C., Vázquez-Méndez, M.E.: Theoretical and numerical analysis of an optimal control problem related to wastewater treatment. SIAM J. Control Optim. **38**, 1534–1553 (2000)

[St52] Stackelberg, H.: The theory of market economy. Oxford University Press, Oxford (1952)

# Chapter 55
# An Overview of the Modified Buckley–Leverett Equation

**Y. Wang**

## 55.1 Introduction

The classical Buckley–Leverett (BL) equation [BuLe42] is a simple model for two-phase fluid flow in a porous medium. One application is secondary recovery by water-drive in oil reservoir simulation. In one space dimension the equation has the standard conservation form

$$\begin{aligned} u_t + (f(u))_x &= 0 \qquad \text{in} \qquad Q = \{(x,t) : x > 0, t > 0\} \\ u(x,0) &= 0 \qquad\qquad x \in (0,\infty) \\ u(0,t) &= u_B \qquad\qquad t \in [0,\infty) \end{aligned} \tag{55.1}$$

with the flux function $f(u)$ being defined as $f(u) = 0$; $u < 0$, $f(u) = 1$, $u > 1$, and

$$f(u) = \frac{u^2}{u^2 + M(1-u)^2} \qquad\qquad 0 \le u \le 1.$$

In this context, $u : \bar{Q} \to [0,1]$ denotes the water saturation (e.g. $u = 1$ means pure water), $u_B$ is a constant which indicates water saturation at $x = 0$, and $M > 0$ is the water/oil viscosity ratio.

The classical BL equation (55.1) is a prototype for conservation laws with convex-concave flux functions. The graph of $f(u)$ and $f'(u)$ with $M = 2$ is given in Figure 55.1. It has been well studied (see [Le92] for an introduction).

Y. Wang (✉)
University of Oklahoma, 601 Elm Avenue, Norman, OK 73019, USA
e-mail: wang@math.ou.edu



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_55

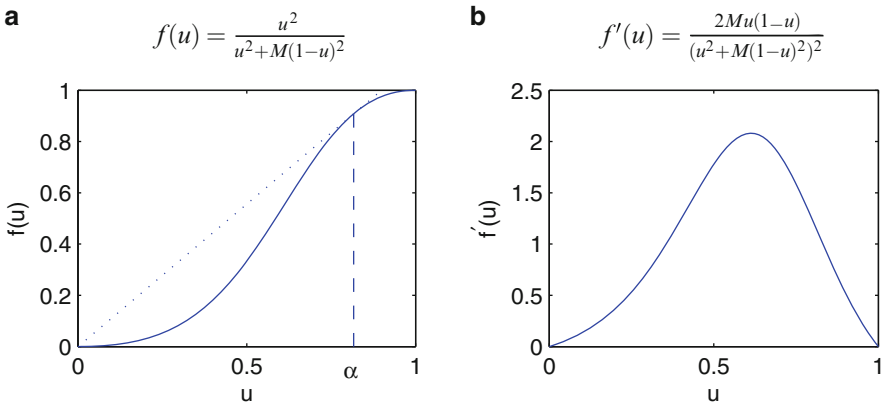


**Fig. 55.1** $f(u)$ and $f'(u)$ with $M = 2$.

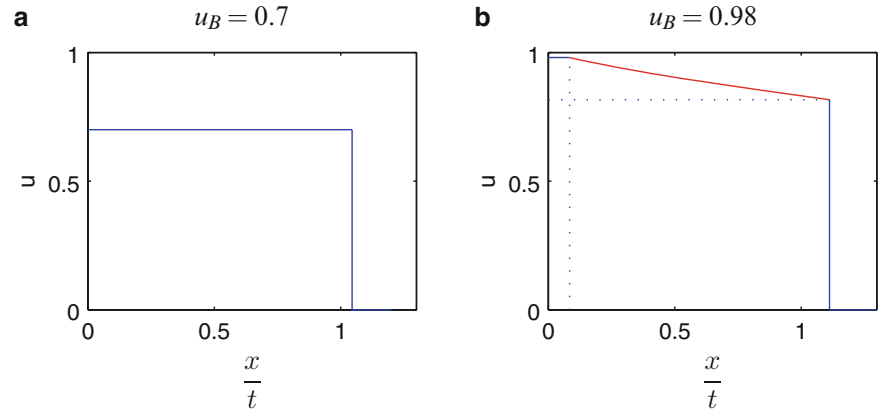


**Fig. 55.2** The entropy solution of the classical BL equation ($M = 2$, $\alpha = \sqrt{\frac{2}{3}} \approx 0.8165$). (**a**) $0 < uB = 0.7 \leq \alpha$, the solution consists of one shock at $\frac{x}{t} = \frac{f(u_B)}{u_B}$; (**b**) $\alpha < uB = 0.98 < 1$, the solution consists of a rarefaction between $u_B$ and $\alpha$ for $f'(u_B) < \frac{x}{t} < f'(\alpha)$ and a shock at $\frac{x}{t} = \frac{f(\alpha)}{\alpha}$.

Let $\alpha$ be the solution of $f'(u) = \frac{f(u)}{u}$, i.e., $\alpha = \sqrt{\frac{M}{M+1}}$. The entropy solution of the classical BL equation can be classified into two categories:

1. If $0 < u_B \leq \alpha$, the entropy solution has a single shock at $\frac{x}{t} = \frac{f(u_B)}{u_B}$.
2. If $\alpha < u_B < 1$, the entropy solution contains a rarefaction between $u_B$ and $\alpha$ for $f'(u_B) < \frac{x}{t} < f'(\alpha)$ and a shock at $\frac{x}{t} = \frac{f(\alpha)}{\alpha}$.

These two types of solutions are shown in Figure 55.2 for $M = 2$.

In either case, the entropy solution of the classical BL equation (55.1) is a non-increasing function of $x$ at any given time $t > 0$. However, the experiments of two-phase flow in porous medium reveal complex infiltration profiles, which may involve overshoot, i.e., profiles may not be monotone [Di04].

To address this issue, the modified Buckley–Leverett equation (MBL) has been derived [HaGr90, HaGr93, VaPePo07, VaMiPo00, VaMiPo02, Wa10, WaKa13]

$$u_t + (f(u))_x = \varepsilon u_{xx} + \varepsilon^2 \tau u_{xxt} \tag{55.2}$$

Van Dujin et al. [VaPePo07] showed that the value $\tau$ is critical in determining the type of the solution profile. In particular, for certain Riemann problems, the solution profile of (55.2) is not monotone when $\tau$ is larger than the threshold value $\tau_*$, where $\tau_*$ was numerically determined to be 0.61 [VaPePo07]. The non-monotonicity of the solution profile is consistent with the experimental observations [Di04].

The classical BL equation (55.1) is hyperbolic, and the numerical schemes for hyperbolic equations have been well developed (for example, [Le92, Le02, CoEtAl98, CoKaSh00, NeTa90, KuLe00], and [LiTa98]). The MBL equation (55.2), however, is pseudo-parabolic, and we will illustrate how to extend the central schemes [NeTa90, KuLe00] to solve (55.2) numerically. Unlike the finite domain of dependence for the classical BL equation (55.1), the domain of dependence for the MBL equation (55.2) is infinite. This naturally raises the question for the choice of computational domain. To answer this question, we will first study the MBL equation equipped with two types of domains and corresponding boundary conditions. One is the half-line problem $x \in [0, \infty)$, and the other one is the finite interval boundary value problem $x \in [0, L]$.

The organization of this chapter is as follows. Section 55.2 will bring forward the exact theory comparing the solutions of the half-line problem and finite interval problem, similar to a study carried out for BBM equation [BoEtAl05, BoLu95]. The difference between the solutions of these two types of problems decays exponentially with respect to the length of the interval $L$ for practically interesting initial profiles. This provides a theoretical justification for the choice of the computational domain. In section 55.3, high order central schemes will be developed for MBL equation in finite interval domain. We provide a detailed derivation on how to extend the central schemes [NeTa90, KuLe00] for conservation laws to solve the MBL equation (55.2). The idea of adopting numerical schemes originally designed for hyperbolic equations to pseudo-parabolic equations is not restricted to central type schemes only ([XuSh08, XuSh09]). The numerical results in section 55.4 show that the water saturation profile strongly depends on the dispersive parameter $\tau$ value as studied in [VaPePo07]. For $\tau > \tau_*$, the MBL equation (55.2) gives non-monotone water saturation profiles for certain Riemann problems as suggested by experimental observations [Di04]. Section 55.5 gives the conclusion of the paper and the possible future directions.

## 55.2 The Half-Line Boundary Value Problem Versus the Finite Interval Boundary Value Problem

Let $u(x,t)$ be the solution to the half-line problem

$$\begin{aligned} u_t + (f(u))_x &= \varepsilon u_{xx} + \varepsilon^2 \tau u_{xxt} \qquad \text{in} \qquad Q = \{(x,t) : x > 0, t > 0\} \\ u(x,0) &= u_0(x) \qquad\qquad x \in [0,\infty) \\ u(0,t) = g_u(t), \quad \lim_{x\to\infty} u(x,t) &= 0 \qquad\qquad t \in [0,\infty) \\ u_0(0) &= g_u(0) \qquad\qquad \text{compatibility condition} \end{aligned} \tag{55.3}$$

and let $v(x,t)$ be the solution to the finite interval boundary value problem

$$\begin{aligned} v_t + (f(v))_x &= \varepsilon v_{xx} + \varepsilon^2 \tau v_{xxt} \qquad \text{in} \qquad \widetilde{Q} = \{(x,t) : x \in (0,L), t > 0\} \\ v(x,0) &= v_0(x) \qquad\qquad x \in [0,L] \\ v(0,t) = g_v(t), \quad v(L,t) &= h(t) \qquad\qquad t \in [0,\infty) \\ v_0(0) = g_v(0), \quad v_0(L) &= h(0) \qquad\qquad \text{compatibility condition.} \end{aligned} \tag{55.4}$$

We consider

$$u_0(x) = \begin{cases} v_0(x) & \text{for } x \in [0,L] \\ 0 & \text{for } x \in [L,+\infty) \end{cases}, \qquad g_u(t) = g_v(t) \equiv g(t), \qquad h(t) \equiv 0,$$

The goal of this section is to develop an estimate of the difference between $u$ and $v$ on the spatial interval $[0,L]$ at a given finite time $t$. The main result of this section is

**Theorem 1 (The main Theorem).** *If $u_0(x)$ satisfies*

$$u_0(x) = \begin{cases} C_u & x \in [0,L_0] \\ 0 & x > L_0 \end{cases} \tag{55.5}$$

*where $L_0 < L$ and $C_u$, are positive constants, then*

$$\| u(\cdot,t) - v(\cdot,t) \|_{H^1_{L,\varepsilon,\tau}} \le D_{1;\varepsilon,\tau}(t) e^{-\frac{\lambda L}{\varepsilon\sqrt{\tau}}} + D_{2;\varepsilon,\tau}(t) e^{-\frac{\lambda(L-L_0)}{\varepsilon\sqrt{\tau}}}$$

*for some $0 < \lambda < 1$, $D_{1;\varepsilon,\tau}(t) > 0$ and $D_{2;\varepsilon,\tau}(t) > 0$, where*

$$\| Y(\cdot,t) \|_{H^1_{L,\varepsilon,\tau}} := \sqrt{\int_0^L Y(x,t)^2 + (\varepsilon\sqrt{\tau} Y_x(x,t))^2 \, dx}$$

To prove theorem 1, we first derive the implicit solution formulae for the half-line problem and the finite interval boundary value problem in section 55.2.1. In section 55.2.2, we use Gronwall's inequality multiple times to obtain the desired result in theorem 1.

### 55.2.1 Implicit Solutions

The implicit solution formulas are in integral form, which are derived by separating the $x$-derivative from the $t$-derivative, and formally solving a first order linear ODE in $t$ and a second order nonhomogeneous ODE in $x$. The details can be found in [Wa10, WaKa14]. The implicit solution formula for the half-line problem (55.3) is

$$\begin{aligned} u(x,t) = &-\frac{1}{2\varepsilon^2\tau\sqrt{\tau}}\int_0^t\int_0^{+\infty}\left(e^{-\frac{x+\xi}{\varepsilon\sqrt{\tau}}}-e^{-\frac{|x-\xi|}{\varepsilon\sqrt{\tau}}}\right)u(\xi,s)e^{-\frac{t-s}{\varepsilon\tau}}\,d\xi\,ds\\ &+\frac{1}{2\varepsilon^2\tau}\int_0^t\int_0^{+\infty}\left(e^{-\frac{x+\xi}{\varepsilon\sqrt{\tau}}}+\operatorname{sgn}(x-\xi)e^{-\frac{|x-\xi|}{\varepsilon\sqrt{\tau}}}\right)f(u)e^{-\frac{t-s}{\varepsilon\tau}}\,d\xi\,ds\\ &+\left(g(t)-e^{-\frac{t}{\varepsilon\tau}}g(0)\right)e^{-\frac{x}{\varepsilon\sqrt{\tau}}}+e^{-\frac{t}{\varepsilon\tau}}u_0(x). \end{aligned}$$

The implicit solution formula for the finite interval boundary value problem (55.4) is

$$\begin{aligned} v(x,t) = &-\frac{1}{2\varepsilon^2\tau\sqrt{\tau}(e^{\frac{2L}{\varepsilon\sqrt{\tau}}}-1)}\int_0^t\int_0^L\left(e^{\frac{x+\xi}{\varepsilon\sqrt{\tau}}}+e^{\frac{2L-(x+\xi)}{\varepsilon\sqrt{\tau}}}-e^{\frac{|x-\xi|}{\varepsilon\sqrt{\tau}}}\right.\\ &\left.-e^{\frac{2L-|x-\xi|}{\varepsilon\sqrt{\tau}}}\right)v(\xi,s)e^{-\frac{t-s}{\varepsilon\tau}}\,d\xi\,ds\\ &-\frac{1}{2\varepsilon^2\tau(e^{\frac{2L}{\varepsilon\sqrt{\tau}}}-1)}\int_0^t\int_0^L\left(e^{\frac{x+\xi}{\varepsilon\sqrt{\tau}}}-e^{\frac{2L-(x+\xi)}{\varepsilon\sqrt{\tau}}}+\operatorname{sgn}(x-\xi)e^{\frac{|x-\xi|}{\varepsilon\sqrt{\tau}}}\right.\\ &\left.-\operatorname{sgn}(x-\xi)e^{\frac{2L-|x-\xi|}{\varepsilon\sqrt{\tau}}}\right)f(v)e^{-\frac{t-s}{\varepsilon\tau}}\,d\xi\,ds\\ &+c_1(t)\phi_1(x)+c_2(t)\phi_2(x)+e^{-\frac{t}{\varepsilon\tau}}v_0(x), \end{aligned}$$

where

$$c_1(t) = g(t)-e^{-\frac{t}{\varepsilon\tau}}g(0), \qquad c_2(t) = h(t)-e^{-\frac{t}{\varepsilon\tau}}h(0), \tag{55.6}$$

$$\phi_1(x) = \frac{e^{\frac{L-x}{\varepsilon\sqrt{\tau}}}-e^{\frac{-L+x}{\varepsilon\sqrt{\tau}}}}{e^{\frac{L}{\varepsilon\sqrt{\tau}}}-e^{-\frac{L}{\varepsilon\sqrt{\tau}}}}, \qquad \phi_2(x) = \frac{e^{\frac{x}{\varepsilon\sqrt{\tau}}}-e^{-\frac{x}{\varepsilon\sqrt{\tau}}}}{e^{\frac{L}{\varepsilon\sqrt{\tau}}}-e^{-\frac{L}{\varepsilon\sqrt{\tau}}}}. \tag{55.7}$$

### 55.2.2 Comparisons

We will prove in this section that the solution $u(x,t)$ to the half-line problem can be approximated as accurately as one wants by the solution $v(x,t)$ to the finite interval boundary value problem as stated in Theorem 1.

The idea of the proof is to decompose $u(x,t)$ ($v(x,t)$ respectively) into two parts: $U(x,t)$ and $u_L(x,t)$ ($V(x,t)$ and $v_L(x,t)$ respectively). $u_L(x,t)$ ($v_L(x,t)$ respectively) consists of terms involving the initial condition $u_0(x)$ ($v_0(x)$ respectively) and the boundary conditions $g(t)$ ($g(t)$ and $h(t)$ respectively) for the governing equation (55.3)((55.4) respectively). $U(x,t)$ ($V(x,t)$ respectively) enjoys zero initial condition and boundary conditions while satisfying a slightly different equation than (55.3)((55.4) respectively). We estimate the difference between $u(\cdot,t)$ and $v(\cdot,t)$ by estimating the differences between $u_L(\cdot,t)$ and $v_L(\cdot,t)$, $U(\cdot,t)$ and $V(\cdot,t)$, then applying the triangle inequality.

#### Definitions

We first decompose $u(x,t)$ as sum of two terms $U(x,t)$ and $u_L(x,t)$, such that

$$u(x,t) = U(x,t) + u_L(x,t) \qquad x \in [0,+\infty)$$

where

$$u_L = e^{-\frac{t}{\varepsilon\tau}} u_0(x) + c_1(t) e^{-\frac{x}{\varepsilon\sqrt{\tau}}} + \left( u(L,t) - c_1(t) e^{-\frac{L}{\varepsilon\sqrt{\tau}}} - e^{-\frac{t}{\varepsilon\tau}} u_0(L) \right) \phi_2(x)$$

and $c_1(t)$ and $\phi_2(x)$ are given in (55.6) and (55.7) respectively. Then $U$ satisfies an equation slightly different from the equation $u$ satisfies in (55.3):

$$\begin{aligned} U_t - \varepsilon U_{xx} - \varepsilon^2 \tau U_{xxt} &= \left( u_t - \varepsilon u_{xx} - \varepsilon^2 \tau u_{xxt} \right) - \left( (u_L)_t - \varepsilon (u_L)_{xx} - \varepsilon^2 \tau (u_L)_{xxt} \right) \\ &= -\left( f(u) \right)_x + \frac{1}{\varepsilon\tau} u_L(x,t) \end{aligned} \tag{55.8}$$

In addition, $U(x,t)$ has zero initial condition and boundary conditions at $x = 0$ and $x = L$, i.e.,

$$U(x,0) = 0, \qquad U(0,t) = 0, \qquad U(L,t) = 0. \tag{55.9}$$

Similarly, for $v(x,t)$, let

$$v(x,t) = V(x,t) + v_L(x,t) \qquad x \in [0,L]$$

where

$$v_L = e^{-\frac{t}{\varepsilon\tau}} v_0(x) + c_1(t)\phi_1(x) + c_2(t)\phi_2(x)$$

and $c_1(t)$, $c_2(t)$ and $\phi_1(x)$, $\phi_2(x)$ are given in (55.6) and (55.7) respectively. Then $V$ satisfies an equation slightly different from the equation $v$ satisfies in (55.4):

$$\begin{aligned} V_t - \varepsilon V_{xx} - \varepsilon^2 \tau V_{xxt} &= \big(v_t - \varepsilon v_{xx} - \varepsilon^2 \tau v_{xxt}\big) - \big((v_L)_t - \varepsilon (v_L)_{xx} - \varepsilon^2 \tau (v_L)_{xxt}\big) \\ &= -\big(f(v)\big)_x + \frac{1}{\varepsilon\tau} v_L(x,t) \end{aligned} \tag{55.10}$$

In addition, $V(x,t)$ has zero initial condition and boundary conditions at $x = 0$ and $x = L$, i.e.,

$$V(x,0) = 0, \qquad V(0,t) = 0, \qquad V(L,t) = 0. \tag{55.11}$$

Since, in the end, we want to study the difference between $U(x,t)$ and $V(x,t)$, we define

$$W(x,t) = V(x,t) - U(x,t) \qquad \text{for} \qquad x \in [0,L].$$

Because of (55.8) and (55.10), we have

$$W_t - \varepsilon W_{xx} - \varepsilon^2 \tau W_{xxt} = -\big(f(v) - f(u)\big)_x + \frac{1}{\varepsilon\tau}(v_L - u_L).$$

In lieu of (55.9) and (55.11), $W(x,t)$ also has zero initial condition and boundary conditions at $x = 0$ and $x = L$, i.e.,

$$W(x,0) = 0, \qquad W(0,t) = 0, \qquad W(L,t) = 0.$$

### Propositions

First, we find the maximum difference of $\| u_L(\cdot,t) - v_L(\cdot,t) \|_\infty$, then we will derive $\| u_L(\cdot,t) - v_L(\cdot,t) \|_{H^1_{L,\varepsilon,\tau}}$ and $\| W(\cdot,t) \|_{H^1_{L,\varepsilon,\tau}} = \| U(\cdot,t) - V(\cdot,t) \|_{H^1_{L,\varepsilon,\tau}}$. Combining these two, we will get an estimate for $\| u(\cdot,t) - v(\cdot,t) \|_{H^1_{L,\varepsilon,\tau}}$. The proof of the propositions can be found in [Wa10, WaKa14].

**Proposition 1.** *If $u_0(x)$ satisfies (55.5), then*

$$\| u_L - v_L \|_\infty \le E_{1;\varepsilon,\tau}(t) e^{-\frac{\lambda L}{\varepsilon\sqrt{\tau}}} + E_{2;\varepsilon,\tau}(t) e^{-\frac{\lambda(L-L_0)}{\varepsilon\sqrt{\tau}}}$$

*where $E_{1;\varepsilon,\tau}(t) = |c_1(\cdot)|_\infty + a_\tau e^{\frac{b_\tau t}{\varepsilon\tau}}$ and $E_{2;\varepsilon,\tau}(t) = c_\tau \frac{t}{\varepsilon\tau} e^{\frac{(b_\tau - 1)t}{\varepsilon\tau}}$, $a_\tau, b_\tau, c_\tau$ are $\tau$-dependent constants.*

**Proposition 2.** *If $u_0(x)$ satisfies (55.5), and $E_{1;\varepsilon,\tau}(t), E_{2;\varepsilon,\tau}(t)$ are as in proposition 1, then*

$$\| u_L(\cdot,t) - v_L(\cdot,t) \|_{H^1_{L,\varepsilon,\tau}} \le \sqrt{5L} \left( E_{1;\varepsilon,\tau}(t) e^{-\frac{\lambda L}{\varepsilon\sqrt{\tau}}} + E_{2;\varepsilon,\tau}(t) e^{-\frac{\lambda(L-L_0)}{\varepsilon\sqrt{\tau}}} \right).$$

**Proposition 3.** *If $u_0(x)$ satisfies (55.5), then*

$$\| W(\cdot,t) \|_{H^1_{L,\varepsilon,\tau}} \le \gamma_{1;\varepsilon,\tau}(t) e^{-\frac{\lambda L}{\varepsilon\sqrt{\tau}}} + \gamma_{2;\varepsilon,\tau}(t) e^{-\frac{\lambda(L-L_0)}{\varepsilon\sqrt{\tau}}}$$

*where the coefficients are given by*

$$\gamma_{1;\varepsilon,\tau}(t) = e^{\frac{(M+1)^2 t}{2M\varepsilon\sqrt{\tau}}} \left( \frac{(M+1)^2\sqrt{\tau}}{2M} + 1 \right) \sqrt{L} \left( \frac{t}{\varepsilon\tau} |c_1(\cdot)|_\infty + \frac{a_\tau}{b_\tau} (e^{\frac{b_\tau t}{\varepsilon\tau}} - 1) \right)$$

$$\gamma_{2;\varepsilon,\tau}(t) = e^{\frac{(M+1)^2 t}{2M\varepsilon\sqrt{\tau}}} \left( \frac{(M+1)^2\sqrt{\tau}}{2M} + 1 \right) \sqrt{L} c_\tau \cdot \cdot \left( \frac{t}{\varepsilon\tau(b_\tau - 1)} e^{\frac{(b_\tau - 1)t}{\varepsilon\tau}} - \frac{1}{(b_\tau - 1)^2} (e^{\frac{(b_\tau - 1)t}{\varepsilon\tau}} - 1) \right).$$

## Proof of Theorem 1

*Proof (Proof of the Main Theorem 1).* By triangle inequality, We have that

$$\begin{aligned} \| u(\cdot,t) - v(\cdot,t) \|_{H^1_{L,\varepsilon,\tau}} &\le \| W(\cdot,t) \|_{H^1_{L,\varepsilon,\tau}} + \| v_L(\cdot,t) - u_L(\cdot,t) \|_{H^1_{L,\varepsilon,\tau}} \\ &\le D_{1;\varepsilon,\tau}(t) e^{-\frac{\lambda L}{\varepsilon\sqrt{\tau}}} + D_{2;\varepsilon,\tau}(t) e^{-\frac{\lambda(L-L_0)}{\varepsilon\sqrt{\tau}}} \end{aligned}$$

where by propositions 2 and 3

$$\begin{aligned} D_{1;\varepsilon,\tau}(t) &= \gamma_{1;\varepsilon,\tau}(t) + \sqrt{5L} E_{1;\varepsilon,\tau}(t) \\ &= e^{\frac{(M+1)^2 t}{2M\varepsilon\sqrt{\tau}}} \left( \frac{(M+1)^2\sqrt{\tau}}{2M} + 1 \right) \sqrt{L} \left( \frac{t}{\varepsilon\tau} |c_1(\cdot)|_\infty + \frac{a_\tau}{b_\tau} (e^{\frac{b_\tau t}{\varepsilon\tau}} - 1) \right) \\ &\quad + \sqrt{5L} (|c(\cdot)|_\infty + a_\tau e^{\frac{b_\tau t}{\varepsilon\tau}}), \\ D_{2;\varepsilon,\tau}(t) &= \gamma_{2;\varepsilon,\tau}(t) + \sqrt{5L} E_{2;\varepsilon,\tau}(t) \\ &= e^{\frac{(M+1)^2 t}{2M\varepsilon\sqrt{\tau}}} \left( \frac{(M+1)^2\sqrt{\tau}}{2M} + 1 \right) \sqrt{L} c_\tau \cdot \\ &\quad \cdot \left( \frac{t}{\varepsilon\tau(b_\tau - 1)} e^{\frac{(b_\tau - 1)t}{\varepsilon\tau}} - \frac{1}{(b_\tau - 1)^2} (e^{\frac{(b_\tau - 1)t}{\varepsilon\tau}} - 1) \right) \\ &\quad + \sqrt{5L} c_\tau \frac{t}{\varepsilon\tau} e^{\frac{(b_\tau - 1)t}{\varepsilon\tau}} \end{aligned}$$

This theorem shows that if $\frac{\lambda L}{\varepsilon\sqrt{\tau}}$ and $\frac{\lambda(L-L_0)}{\varepsilon\sqrt{\tau}}$ converge to infinity, then the solution $v(x,t)$ of the finite interval boundary value problem converges to the solution $u(x,t)$ of the half-line problem in the sense of $\| \cdot \|_{H^1_{L,\varepsilon,\tau}}$. This can be achieved

either by letting $L \to \infty$ or $\varepsilon \to 0$. For example, in the extreme case, $\varepsilon = 0$, the half line problem (55.3) becomes hyperbolic and the domain of dependence is finite, so, certainly, one only needs to consider the finite interval boundary value problem. This is consistent with the main theorem in the sense that for a fixed final time $t$, if $\lambda L > b_\tau t$ and $\lambda(L - L_0) > (b_\tau - 1)t$, i.e., $L > \max(\frac{b_\tau t}{\lambda}, \frac{(b_\tau - 1)t}{\lambda})$, then $\| u(\cdot,t) - v(\cdot,t) \|_{H^1_{L,\varepsilon,\tau}} \le D_{1;\varepsilon,\tau}(t) e^{-\frac{\lambda L}{\varepsilon\sqrt{\tau}}} + D_{2;\varepsilon,\tau}(t) e^{-\frac{\lambda(L-L_0)}{\varepsilon\sqrt{\tau}}} \to 0$ as $\varepsilon \to 0$.

Theorem 1 gives a theoretical justification for using the solution of the finite interval boundary value problem to approximate the solution of the half-line problem with appropriate choice of $L$ and $\varepsilon$. Hence in the next chapter, the numerical scheme designed to solve the MBL equation (55.2) is given for finite interval boundary value problem.

## 55.3 Numerical Schemes

In this section, we show how to apply the central schemes [NeTa90] originally designed for hyperbolic conservation laws to numerically solve the MBL equation (55.2), which is of pseudo-parabolic type. We first collect all the terms with time derivative and rewrite MBL equation (55.2) as

$$(u - \varepsilon^2 \tau u_{xx})_t + (f(u))_x = \varepsilon u_{xx}. \tag{55.12}$$

By letting

$$w = u - \varepsilon^2 \tau u_{xx} \quad \Longleftrightarrow \quad u = (I - \varepsilon^2 \tau \partial_{xx})^{-1} w,$$

MBL equation (55.12) can be written as

$$w_t + (f(u))_x = \varepsilon u_{xx}. \tag{55.13}$$

As in [NeTa90], at each time level, we first reconstruct a piecewise linear approximation of the form

$$L_j(x,t) = w_j(t) + (x - x_j)\frac{w_j'}{\Delta x}, \qquad x_{j-\frac{1}{2}} \le x \le x_{j+\frac{1}{2}}. \tag{55.14}$$

Second-order accuracy is guaranteed if the so-called vector of numerical derivative $\frac{w_j'}{\Delta x}$, which will be given later, satisfies

$$\frac{w_j'}{\Delta x} = \frac{\partial w(x_j,t)}{\partial x} + O(\Delta x).$$

We denote the staggered piecewise-constant functions $\bar{w}_{j+\frac{1}{2}}(t)$ as

$$\bar{w}_{j+\frac{1}{2}}(t) = \frac{1}{\Delta x}\int_{x_j}^{x_{j+1}} w(x,t)\,dx.$$

Evolve the piecewise linear interplant (55.14) by integrating (55.13) over $[x_j, x_{j+1}] \times [t, t+\Delta t]$

$$\begin{aligned}\bar{w}_{j+\frac{1}{2}}(t+\Delta t) =& \frac{1}{2}[w_j(t)+w_{j+1}(t)] + \frac{1}{8}[w_j' - w_{j+1}'] \\ & - \lambda\left[f(u_{j+1}(t+\frac{\Delta t}{2}) - f(u_j(t+\frac{\Delta t}{2}))\right] \\ & + \frac{\varepsilon}{\Delta x}\left[\int_t^{t+\Delta t}\int_{x_j}^{x_{j+1}} \frac{\partial^2 u(x,s)}{\partial x^2}\, dx\, ds\right].\end{aligned} \tag{55.15}$$

Nessyahu and Tadmor in [NeTa90] have introduced many ways to estimate the derivatives, so, we won't reproduce them here. Instead, we will focus on the last integral in (55.15). There are many ways to numerically calculate the integral $\int_t^{t+\Delta t}\int_{x_j}^{x_{j+1}} \frac{\partial^2 u(x,s)}{\partial x^2}\, dx\, ds$. We will show two ways to do this in the following two subsections, all of them achieve second order accuracy.

### 55.3.1 Trapezoid Scheme

In this scheme, we use the trapezoid rule to calculate the integral numerically as follows:

$$\begin{aligned}\int_t^{t+\Delta t}\int_{x_j}^{x_{j+1}} \frac{\partial^2 u(x,s)}{\partial x^2}\, dx\, ds &= \Delta x \int_t^{t+\Delta t} (\bar{u}_{xx})_{j+\frac{1}{2}}(s)\, ds \\ &= \frac{\Delta x \Delta t}{2}\left((\bar{u}_{xx})_{j+\frac{1}{2}}(t) + (\bar{u}_{xx})_{j+\frac{1}{2}}(t+\Delta t)\right)\end{aligned}$$

with $\mathcal{O}(\Delta t^3)$ error. The flow chart of the trapezoid scheme is given in table 55.1.

**Table 55.1** Flow chart for Trapezoid Scheme.

| | Trapezoid Scheme |
|---|---|
| `Calculate` | $\bar{w}_{j+\frac{1}{2}}(t) = \frac{1}{2}(w_j(t)+w_{j+1}(t)) + \frac{1}{8}(w_j' - w_{j+1}')$ |
| `Solve` | $(I-\varepsilon^2\tau\partial_{xx})\bar{u}_{j+\frac{1}{2}}(t) = \bar{w}_{j+\frac{1}{2}}(t)$ `for` $\bar{u}_{j+\frac{1}{2}}(t)$ |
| `Calculate` | $w_j(t+\frac{\Delta t}{2}) = w_j(t) + (\varepsilon\frac{\Delta^2 u_j}{\Delta x} - f_j')\frac{\lambda}{2}$ |
| `Solve` | $(I-\varepsilon^2\tau\partial_{xx})u_j(t+\frac{\Delta t}{2}) = w_j(t+\frac{\Delta t}{2})$ `for` $u_j(t+\frac{\Delta t}{2})$ |
| `Solve` | $\left(I-(\varepsilon^2\tau+\frac{\varepsilon\Delta t}{2})\partial_{xx}\right)\bar{u}_{j+\frac{1}{2}}(t+\Delta t)$ |
| | $= \left(I-(\varepsilon^2\tau-\frac{\varepsilon\Delta t}{2})\partial_{xx}\right)\bar{u}_{j+\frac{1}{2}}(t)$ |
| | $-\lambda\left[f(u_{j+1}(t+\frac{\Delta t}{2})) - f(u_j(t+\frac{\Delta t}{2}))\right]$ `for` $\bar{u}_{j+\frac{1}{2}}(t+\Delta t)$ |

**Table 55.2** Flow chart for Midpoint Scheme.

| | Midpoint Scheme |
|---|---|
| `Calculate` | $\bar{w}_{j+\frac{1}{2}}(t) = \frac{1}{2}(w_j(t)+w_{j+1}(t))+\frac{1}{8}(w'_j-w'_{j+1})$ |
| `Calculate` | $w_j(t+\frac{\Delta t}{2}) = w_j(t)+(\varepsilon\frac{\Delta^2 u_j}{\Delta x}-f'_j)\frac{\lambda}{2}$ |
| `Solve` | $(I-\varepsilon^2\tau\partial_{xx})u_j(t+\frac{\Delta t}{2}) = w_j(t+\frac{\Delta t}{2})$ `for` $u_j(t+\frac{\Delta t}{2})$ |
| `Calculate` | $\bar{w}_{j+\frac{1}{2}}(t+\frac{\Delta t}{2}) = \frac{1}{2}(w_j(t+\frac{\Delta t}{2})+w_{j+1}(t+\frac{\Delta t}{2}))$ |
| | $+\frac{1}{8}(w'_j(t+\frac{\Delta t}{2})-w'_{j+1}(t+\frac{\Delta t}{2}))$ |
| `Solve` | $(I-\varepsilon^2\tau\partial_{xx})\bar{u}_{j+\frac{1}{2}}(t+\frac{\Delta t}{2}) = \bar{w}_{j+\frac{1}{2}}(t+\frac{\Delta t}{2})$ `for` $\bar{u}_{j+\frac{1}{2}}(t+\frac{\Delta t}{2})$ |
| `Solve` | $(I-\varepsilon^2\tau\partial_{xx})\bar{u}_{j+\frac{1}{2}}(t+\Delta t)=\bar{w}_{j+\frac{1}{2}}(t)-\lambda[f(u_{j+1}(t+\frac{\Delta t}{2})-f(u_j(t+\frac{\Delta t}{2}))]$ |
| | $+\varepsilon\Delta t(\bar{u}_{xx})_{j+\frac{1}{2}}(t+\frac{\Delta t}{2})$ `for` $\bar{u}_{j+\frac{1}{2}}(t+\Delta t)$ |

### *55.3.2 Midpoint Scheme*

In this scheme, we use the midpoint rule to calculate the integral numerically as follows:

$$\int_t^{t+\Delta t}\int_{x_j}^{x_{j+1}}\frac{\partial^2 u(x,s)}{\partial x^2}\,dx\,ds = \Delta x\int_t^{t+\Delta t}(\bar{u}_{xx})_{j+\frac{1}{2}}(s)\,ds$$
$$= \Delta x\Delta t(\bar{u}_{xx})_{j+\frac{1}{2}}(t+\frac{\Delta t}{2})$$

The flow chart of the midpoint scheme is given in table 55.2.

## 55.4 Computational Results

In this section, we show the numerical solutions to the MBL equation (55.2) with the initial condition

$$u_0(x)=\begin{cases}u_B & \text{if } x=0\\ 0 & \text{if } x>0\end{cases}$$

and the Dirichlet boundary condition. Duijn et al. [VaPePo07] numerically provided a bifurcation diagram (Figure 55.3) of MBL (55.2) equation as the dispersive parameter $\tau$ and the post-shock value $u_B$ of the initial condition vary. The solution of (55.2) has been proven to display qualitatively different profiles for parameter values $(\tau,u_B)$ falling in different regimes of the bifurcation diagram. In particular, for every fixed $\tau$ value, there are two critical $u_B$ values, namely, $\bar{u}$ and $\underline{u}$. From the bifurcation diagram (Figure 55.3), it is clear that, when $\tau<\tau_*$, $\bar{u}=\underline{u}=\alpha$. For a fixed $\tau$ value, the solution has three different profiles.

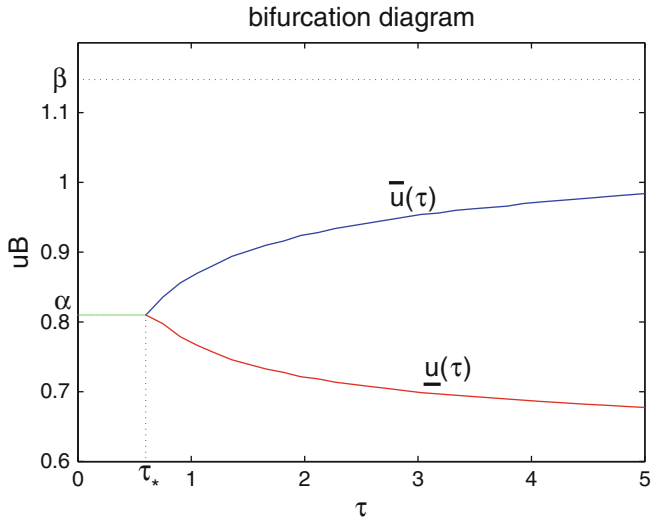


**Fig. 55.3** The bifurcation diagram of the MBL equation (55.2) with the bifurcation parameters $(\tau, u_B)$.

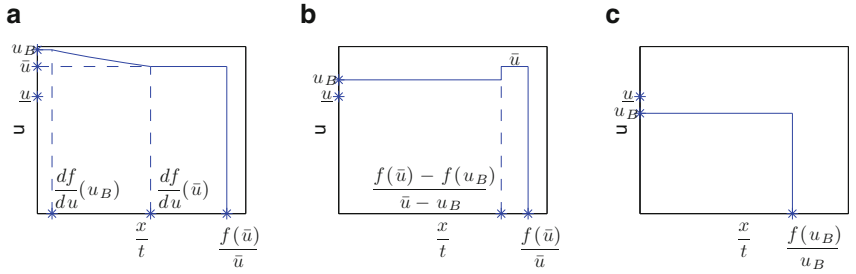


**Fig. 55.4** Given a fixed $\tau$, the three qualitatively different solution profiles due to different values of $u_B$. In particular, when $\tau > \tau_*$ and $\underline{u} < u_B < \bar{u}$, the solution profiles (Figure (b)) displays non-monotonicity, which is consistent with the experimental observations ([Di04]). Figures (**a**), (**b**) and (**c**) are demonstrative figures.

(a) If $u_B \in [\bar{u}, 1]$, the solution contains a plateau value $u_B$ for $0 \le \frac{x}{t} \le \frac{df}{du}(u_B)$, a rarefaction wave connection $u_B$ to $\bar{u}$ for $\frac{df}{du}(u_B) \le \frac{x}{t} \le \frac{df}{du}(\bar{u})$, another plateau value $\bar{u}$ for $\frac{df}{du}(\bar{u}) < \frac{x}{t} < \frac{f(\bar{u})}{\bar{u}}$, and a shock from $\bar{u}$ down to 0 at $\frac{x}{t} = \frac{f(\bar{u})}{\bar{u}}$ (see Figure 55.4(a)).

(b) If $u_B \in (\underline{u}, \bar{u})$, the solution contains a plateau value $u_B$ for $0 \le \frac{x}{t} < \frac{f(\bar{u})-f(u_B)}{\bar{u}-u_B}$, a shock from $u_B$ up to $\bar{u}$ at $\frac{x}{t} = \frac{f(\bar{u})-f(u_B)}{\bar{u}-u_B}$, another plateau value $\bar{u}$ for $\frac{f(\bar{u})-f(u_B)}{\bar{u}-u_B} < \frac{x}{t} < \frac{f(\bar{u})}{\bar{u}}$, and a shock from $\bar{u}$ down to 0 at $\frac{x}{t} = \frac{f(\bar{u})}{\bar{u}}$ (see Figure 55.4(b)). The solution may exhibit a damped oscillation near $u = u_B$.

**Table 55.3** 9 pairs of $(\tau, u_B)$ values with either fixed $\tau$ value or fixed $u_B$ value used for the numerical results given in Figure 55.5 (examples 1–6).

| $(\tau, u_B)$ | Example 4 | Example5 | Example 6 |
|---|---|---|---|
| Example 1 | $(0.2, 0.9)$ | $(1, 0.9)$ | $(5, 0.9)$ |
| Example 2 | $(0.2, \alpha)$ | $(1, \alpha)$ | $(5, \alpha)$ |
| Example 3 | $(0.2, 0.75)$ | $(1, 0.75)$ | $(5, 0.75)$ |

(c) If $u_B \in (0, \underline{u}]$, the solution consists a single shock connecting $u_B$ and 0 at $\frac{x}{t} = \frac{f(u_B)}{u_B}$ (see Figure 55.4(c)). It may exhibit oscillatory behavior near $u = u_B$.

Notice that when $\tau > \tau_*$ and $\underline{u} < u_B < \bar{u}$, the solution profiles (55.4(b)) displays non-monotonicity, which is consistent with the experimental observations ([Di04]).

In the numerical computation we show below, we will therefore test the accuracy and capability of central schemes for different parameter values ($\tau$ and $u_B$) that fall into various regimes of the bifurcation diagram, and therefore display qualitatively different solution profiles. The numerical experiments were carried out for $M = 2$, $\varepsilon = 0.001$ and $T = 4000 \times \varepsilon$, i.e. $\tilde{T} = 4000$ to get the asymptotic solution profiles, and $\Delta x$ was chosen to be $\frac{\varepsilon}{10}$ and $\lambda = \frac{\Delta t}{\Delta x}$ was chosen to be 0.1. The scheme used in the computation is the second-order trapezoid scheme as shown in section 55.3.1. The midpoint scheme delivers similar computational results, hence is omitted here. The solution profiles at $\frac{T}{4}$ (blue), $\frac{2*T}{4}$ (green), $\frac{3*T}{4}$ (magenta) and $T$ (black) are chosen to demonstrate the time evolution of the solutions. The red dashed lines are used to denote the theoretical shock locations and plateau values for comparison purpose.

We start with $\tau > 0$. Based on the bifurcation diagram (Figure 55.3), we choose three representative $u_B$ values, i.e. $u_B = 0.9 > \alpha$, $u_B = \alpha = \sqrt{\frac{M}{M+1}} = \sqrt{\frac{2}{3}}$ (for $M = 2$) and $u_B = 0.75 < \alpha$. For each fixed $u_B$, we choose three representative $\tau$ values, i.e. $\tau = 0.2 < \tau_* \approx 0.61$, $\tau = 1 > \tau_*$ with $u_B = 0.75 < \underline{u}_{\tau=1} < u_B = \alpha < \bar{u} < u_B = 0.9$, and $\tau = 5$ with $u_B = 0.75, \alpha, 0.9 \in [\underline{u}_{\tau=5}, \bar{u}_{\tau=5}]$. We use this 9 pairs of $(\tau, u_B)$ values given in Table 55.3 to validate the solution profiles with the demonstrative solution profiles given in Figure 55.4.

*Example 1.* $(\tau, u_B) = (0.2, 0.9), (\tau, u_B) = (1, 0.9), (\tau, u_B) = (5, 0.9)$.

When $u_B = 0.9 > \alpha$ is fixed, we increase $\tau$ from 0.2 to 1 to 5 (Figure 55.5(a), 55.5(b) , 55.5(c)), the dispersive effect starts to dominate the solution profile. When $\tau = 0.2$ (Figure 55.5(a)), the solution profile is similar to the classical BL equation solution (see Figure 55.2(b)), with a rarefaction wave for $\frac{x}{t} \in [f'(u = 0.9), f'(u = \alpha) = f'(u = \bar{u}_{\tau=0.2})]$ and a shock from $u = \alpha$ to $u = 0$ at $\frac{x}{t} = f'(\alpha)$. This corresponds to Figure 55.4(a) with $\frac{df}{du}(\bar{u}_{\tau=0.2} = \alpha) = \frac{f(\bar{u}_{\tau=0.2})}{\bar{u}_{\tau=0.2}} = \frac{f(\alpha)}{\alpha}$. When $\tau = 1$ (Figure 55.5(b)), the rarefaction wave is between $\frac{x}{t} \in [f'(u = 0.9), f'(u = \bar{u}_{\tau=1})]$ and the solution remains at the plateau value $u = \bar{u}_{\tau=1}$ for $\frac{x}{t} \in [f'(u = \bar{u}_{\tau=1}), \frac{f(\bar{u}_{\tau=1})}{\bar{u}_{\tau=1}}]$ and

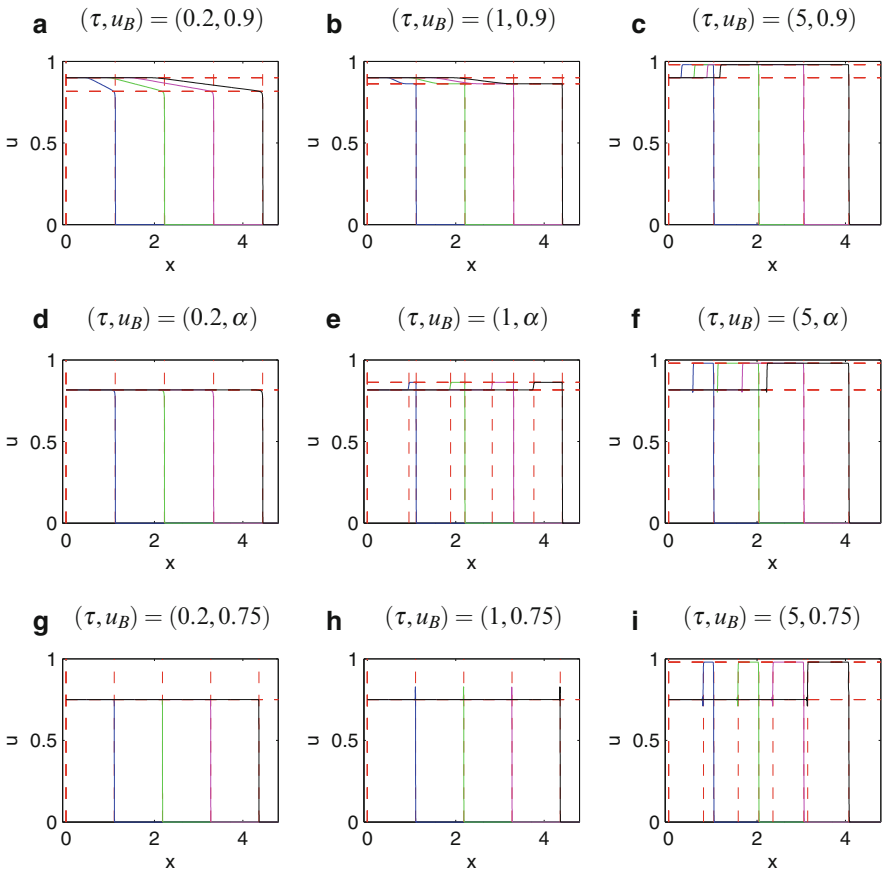


**Fig. 55.5** Numerical solutions to MBL equation with parameter settings fall in different regimes of the bifurcation diagram (Figure 55.3). The color coding is for different time: $\frac{1}{4}T$ (blue), $\frac{2}{4}T$ (green), $\frac{3}{4}T$ (magenta) and $T$ (black). The results are discussed in examples 1 – 6. In figures (d)–(f), $\alpha = \sqrt{\frac{M}{M+1}} = \sqrt{\frac{2}{3}}$ for $M = 2$.

the shock occurs at $\frac{x}{t} = \frac{f(\bar{u}_{\tau=1})}{\bar{u}_{\tau=1}}$. This corresponds to Figure 55.4(a) with $u_B = 0.9 > \bar{u}_{\tau=1} \approx 0.86$. When $\tau = 5$ (Figure 55.5(c)), the solution displays the first shock from $u = 0.9$ to $u = \bar{u}_{\tau=5}$ at $\frac{x}{t} = \frac{f(\bar{u}_{\tau=5}) - f(u_B)}{\bar{u}_{\tau=5} - u_B}$, and then remains at the plateau value $u = \bar{u}_{\tau=5}$ for $\frac{x}{t} \in [\frac{f(\bar{u}_{\tau=5}) - f(u_B)}{\bar{u}_{\tau=5} - u_B}, \frac{f(\bar{u}_{\tau=5})}{\bar{u}_{\tau=5}}]$ and the second shocks occurs at $\frac{x}{t} = \frac{f(\bar{u}_{\tau=5})}{\bar{u}_{\tau=5}}$. This corresponds to Figure 55.4(b) with $\underline{u}_{\tau=5} \approx 0.68 < u_B = 0.9 < \bar{u}_{\tau=5} \approx 0.98$. Notice that as $\tau$ increases, the rarefaction region shrinks and the plateau region enlarges.

*Example 2.* $(\tau, u_B) = (0.2, \alpha), (\tau, u_B) = (1, \alpha), (\tau, u_B) = (5, \alpha)$.

When $u_B = \alpha$ is fixed, we increase $\tau$ from 0.2 to 1 to 5 (Figure 55.5(d), (e), (f)), the dispersive effect starts to dominate the solution profile. When $\tau = 0.2$, the solution displays one single shock at $\frac{x}{t} = \frac{f(\alpha)}{\alpha}$. For both $\tau = 1$ and $\tau = 5$, the solution has two shocks, one at $\frac{x}{t} = \frac{f(\bar{u}_{\tau=1(\tau=5 \text{ respectively})}) - f(\alpha)}{\bar{u}_{\tau=1(\tau=5 \text{ respectively})} - \alpha}$, and another one at $\frac{x}{t} = \frac{f(\bar{u}_{\tau=1(\tau=5 \text{ respectively})})}{\bar{u}_{\tau=1(\tau=5 \text{ respectively})}}$. For both $\tau = 1$ and $\tau = 5$ (Figures 55.5(e), (f)), the solutions correspond to Figure 55.4(b), which are consistent with the experimental observations. Notice that as $\tau$ increases from 1 to 5, i.e., the dispersive effect increases, the inter-shock interval length increases at every fixed time (compare Figure 55.5(e) with Figure 55.5(f)). In addition, for fix $\tau = 1$ ($\tau = 5$ respectively), as time progresses, the inter-shock interval length increases in the linear fashion (see Figure 55.5(e) (Figure 55.5(f) respectively) ).

*Example 3.* $(\tau, u_B) = (0.2, 0.75), (\tau, u_B) = (1, 0.75), (\tau, u_B) = (5, 0.75)$.

When $u_B = 0.75 <= \alpha$ is fixed, we increase $\tau$ from 0.2 to 1 to 5 (Figure 55.5(g), (h), (i)), the dispersive effects start to dominate the solution profile in the similar fashion as $u_B = 0.9$ and $u_B = \alpha$. Notice that when $\tau = 1$, since $u_B = 0.75$ is very close to $\underline{u}_{\tau=1}$, the solution displays oscillation at $\frac{x}{t} = \frac{f(u_B)}{u_B}$ (Figure 55.5(h)). If we increase $\tau$ further to $\tau = 5$, the dispersive effect is strong enough to create a plateau value at $\bar{u} \approx 0.98$ (see Figure 55.5(i)).

*Example 4.* $(\tau, u_B) = (0.2, 0.9), (\tau, u_B) = (0.2, \alpha), (\tau, u_B) = (0.2, 0.75)$.

Now, we fix $\tau = 0.2$, decrease $u_B$ from 0.9 to $\alpha$, to 0.75 (Figures 55.5(a), (d), (g)). If $u_B > \alpha$ the solution consists a rarefaction wave connecting $u_B$ down to $\alpha$, then a shock from $\alpha$ to 0, otherwise, the solution consists a single shock from $u_B$ down to 0. In all cases, since $\tau = 0.2 < \tau_*$, regardless of the $u_B$ value, the solution will not display non-monotone behavior, due to the lack of dispersive effect.

*Example 5.* $(\tau, u_B) = (1, 0.9), (\tau, u_B) = (1, \alpha), (\tau, u_B) = (1, 0.75)$.

Now, we fix $\tau = 1$, decrease $u_B$ from 0.9 to $\alpha$, to 0.75 (Figures 55.5(b), (e), (h)). If $u_B = 0.9 > \bar{u}_{\tau=1}$, the solution consists of a rarefaction wave connecting $u_B$ and $\bar{u}$, and a shock connecting $\bar{u}$ down to 0 (Figure 55.5(b)). Even if $\underline{u} < u_B < \bar{u}$, because $\tau = 1 > \tau_*$, the solution still has a chance to increase to the plateau value $\bar{u}$ as seen in Figure 55.5(e). But, if $u_B$ is too small, for example, $u_B = 0.75 < \underline{u}$, the solution does not increase to $\bar{u}$ any more, instead, it consists of a single shock connecting $u_B$ down to 0 (Figure 55.5(h)).

*Example 6.* $(\tau, u_B) = (5, 0.9), (\tau, u_B) = (5, \alpha), (\tau, u_B) = (5, 0.75)$.

Now, we fix $\tau = 5$, decrease $u_B$ from 0.9 to $\alpha$, to 0.75 (Figures 55.5(c), (f), (i)). For all three $u_B$, they are between $\underline{u}_{\tau=5}$ and $\bar{u}_{\tau=5}$, hence all increase to the plateau value $\bar{u}_{\tau=5} \approx 0.98$ before dropping to 0. Notice that as $u_B$ decreases, the inter-shock interval length decreases at every fixed time (compare Figures 55.5(c), (f) and (i)). This shows that when the dispersive effect is strong ($\tau > \tau_*$), the bigger $u_B$ is, the bigger region the solution stays at the plateau value.

More numerical examples can be found in [WaKa13].

## 55.5 Conclusions

We proved that the solution to the infinite domain problem can be approximated by that of the bounded domain problem. This provides a theoretical justification for using finite domain to calculation of the numerical solution of the MBL equation (55.2). We also extended the classical central scheme originally designed for the hyperbolic systems to solve the MBL equation, which is of pseudo-parabolic type. The numerical solutions for qualitatively different parameter values $\tau$ and initial conditions $u_B$ show that the jump locations are consistent with the theoretical calculation and the plateau heights are consistent with the numerically obtained values given in [VaPePo07]. In particular, when $\tau > \tau_*$, for $u_B \in (\underline{u}, \bar{u})$, the numerical solutions give non-monotone water saturation profiles, which is consistent with the experimental observations.

**Acknowledgements** This research was supported in part by a Faculty Investment Program and a Junior Faculty Fellow Program grant from the Research Council and College of Arts and Sciences of the University of Oklahoma Norman Campus.

## References

[BoLu95] Bona, J. L. and Luo, L.-H.: Initial-boundary value problems for model equations for the propagation of long waves. Lecture Notes in Pure and Appl. Math., **168**, 63, 65–94 (1995)

[BoEtAl05] Bona, J.L., Chen, H.-Q., Sun, S.M., and Zhang, B.-Y.: Comparison of quarter-plane and two-point boundary value problems: the BBM-equation. Discrete Contin. Dyn. Syst., **13(4)**, 921–940 (2005)

[BuLe42] Buckley, S.E. and Leverett, M.C.: Mechanism of fluid displacement in sands. Petroleum Transactions, AIME, **146**, 107–116 (1942)

[CoEtAl98] Cockburn, B., Johnson, C., Shu, C.-W., and Tadmor, E.: Advanced numerical approximation of nonlinear hyperbolic equations. Lecture Notes in Mathematics, Papers from the C.I.M.E. Summer School held in Cetraro, June 23–28, 1997, **1697** (1998)

[CoKaSh00] Cockburn, B., Karniadakis, G. E., and Shu, C.-W. (Eds.): Discontinuous Galerkin Methods: Theory, Computation and Applications. Lecture Notes in Computational Science and Engineering (2000)

[Di04] DiCarlo, D. A.: Experimental measurements of saturation overshoot on infiltration. Water Resources Research, **40**, 4215.1 – 4215.9 (2004)

[HaGr90] Hassanizadeh, S.M. and Gray, W.G.: Mechanics and thermodynamics of multiphase flow in porous media including interphase boundaries. Adv. Water Resour., **13**, 169–186 (1990)

[HaGr93] Hassanizadeh, S.M. and Gray, W.G.: Thermodynamic basis of capillary pressure in porous media. Water Resour. Res., **29**, 3389–3405 (1993)

[KuLe00] Kurganov, A. and Levy, D.: A third-order semidiscrete central scheme for conservation laws and convection-diffusion equations. SIAM J. Sci. Comput., **22(4)**, 1461–1488 (2000)

[Le92] LeVeque, R.J.: Numerical methods for conservation laws. Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel (1992)

[Le02] LeVeque, R.J.: Finite volume methods for hyperbolic problems. Cambridge University Press, 2002

[LiTa98] Liu, X.-D. and Tadmor, E.: Third order nonoscillatory central scheme for hyperbolic conservation laws. Numer. Math., **79(3)**, 397–425 (1998)

[NeTa90] Nessyahu, H. and Tadmor, E.: Nonoscillatory central differencing for hyperbolic conservation laws. J. Comput. Phys., **87(2)**, 408–463 (1990)

[VaMiPo00] Van Duijn, C.J., Mikelic, A., and Pop, I.S.: Effective Buckley-Leverett equations by homogenization. Progress in industrial mathematics at ECMI, 42–52 (2000).

[VaMiPo02] Van Duijn, C.J., Mikelic, A., and Pop, I.S.: Effective Equations for Two-Phase Flow with Trapping on the Micro Scale. SIAM Journal on Applied Mathematics, **62(5)**, 1531–1568 (2002)

[VaPePo07] Van Duijn, C.J., Peletier, L.A., and Pop, I.S.: A new class of entropy solutions of the Buckley-Leverett equation. SIAM J. Math. Anal., **39(2)**, 507–536 (2007)

[Wa10] Wang, Y. : Central schemes for the modified Buckley-Leverett equation. Ph.D. thesis, The Ohio State University (2010).

[WaKa13] Wang, Y. and Kao, C.-Y.: Central schemes for the modified Buckley-Leverett equation. Journal of Computational Science, **4(1)**, 12–23 (2013)

[WaKa14] Wang, Y. and Kao, C.-Y.: Bounded Domain Problem for the Modified Buckley-Leverett Equation. Journal of Dynamics and Differential Equations, **26(3)**,607–629 (2014)

[XuSh08] Xu, Y. and Shu, C-W.: A local discontinuous Galerkin method for the Camassa-Holm equation. SIAM J. Numer. Anal., **46(4)**, 1998–2021 (2008)

[XuSh09] Xu, Y. and Shu, C-W.: Local discontinuous Galerkin method for the Hunter-Saxton equation and its zero-viscosity and zero-dispersion limits. SIAM J. Sci. Comput., **31(2)**, 1249–1268 (2008/09)

# Chapter 56
# Influence of Stochastic Moments on the Solution of the Neutron Point Kinetics Equation

**M. Wollmann da Silva, B.E.J. Bodmann, M.T.B. Vilhena, and R. Vasques**

## 56.1 Introduction

The neutron point kinetics equations, which model the time-dependent behavior of nuclear reactors [AbHa03, HaAl05, He71, Sa89], are often used to understand the dynamics of nuclear reactor operations (e.g. power fluctuations caused by control rod motions during start-up and shut-down procedures). They consist of a system of coupled differential equations that model the interaction between (i) the neutron population, and (ii) the concentration of the delayed neutron precursors, which are radioactive isotopes formed in the fission process that decay through neutron emission. These equations are deterministic in nature, and therefore can provide only average values of the modeled populations. However, the actual dynamical process is stochastic: the neutron density and the delayed neutron precursor concentrations vary randomly with time.

To address this stochastic behavior, Hayes and Allen [HaAl05] have generalized the standard deterministic point kinetics equations. They have derived a system of stochastic differential equations that can accurately model the random behavior of the neutron density and the precursor concentrations in a point reactor. Due to the issue of stiffness, they numerically implement this system using a stochastic piecewise constant approximation method (Stochastic PCA).

Here, we present a study of the influence of stochastic fluctuations on the results of the neutron point kinetics equations. We reproduce the stochastic formulation

M. Wollmann da Silva (✉) • B.E.J. Bodmann • M.T.B. Vilhena • R. Vasques
Federal University of Rio Grande do Sul, Av. Osvaldo Aranha 99/4, Porto Alegre 90046-900, RS, Brazil
e-mail: milena.wollmann@ufrgs.br; bardobodmann@ufrgs.br; vilhena@math.ufrgs.br; richard.vasques@fulbrightmail.org



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_56

introduced in [HaAl05] and compute Monte Carlo numerical results for examples with constant and time-dependent reactivity, comparing these results with stochastic and deterministic methods found in the literature [HaAl05, Ra12, WoLe14].

The remainder of this work is organized as follows. In Section 56.2, we reproduce the derivation of the stochastic equations introduced in [HaAl05]. Section 56.3 starts with a short discussion on the numerical implementation of the stochastic models. In Section 56.3.1, we provide numerical results for examples with constant reactivity, for the cases of one and six precursor groups; and in Section 56.3.2, we present results for an example with linear reactivity and one precursor group. Finally, in Section 56.4, we discuss the stochastic fluctuations we have encountered, and address the future steps to be undertaken in order to accurately study them.

## 56.2 Stochastic Model Formulation

In this section, we reproduce the stochastic formulation introduced by Hayes and Allen. Following [HaAl05, He71], the time-dependent equations that describe the neutron density and the delayed neutron precursor concentrations are

$$\frac{\partial N}{\partial t} = Dv\nabla^2 N - (\Sigma_a - \Sigma_f)\, vN + [(1-\beta)\, k_\infty \Sigma_a - \Sigma_f]\, vN + \sum_i \lambda_i C_i + S_0, \tag{56.1a}$$

$$\frac{\partial C_i}{\partial t} = \beta_i k_\infty \Sigma_a vN - \lambda_i C_i, \tag{56.1b}$$

where $i = 1, 2, ..., m$, $v$ is the velocity, $N = N(r,t)$ is the neutron density at position $r$ and time $t$, and $C_i = C_i(r,t)$ is the concentration of the $i$-th type of precursor at position $r$ and time $t$. On the right-hand side of Eq. (56.1a) we have the following terms:

- $Dv\nabla^2 N$, representing the diffusion of neutrons.
- $(\Sigma_a - \Sigma_f)vN$, representing the capture of neutrons. Notice that the capture cross section is given by the difference between the absorption ($\Sigma_a$) and the fission ($\Sigma_f$) cross sections.
- $[(1-\beta)\, k_\infty \Sigma_a - \Sigma_f]\, vN$, representing the prompt-neutron contribution to the source. Here, $\beta = \sum_{i=1}^{m} \beta_i$ is the delayed-neutron fraction and $k_\infty$ is the infinite medium reproduction factor.
- $\sum_i \lambda_i C_i$, representing the rate of transformation from the neutron precursors to the neutron population, with $\lambda_i$ as the decay constant.
- $S_0(r,t)$, representing the external source.

Assuming that $N$ and $C_i$ are separable in time and space, we can write $N(r,t) = f(r)n(t)$ and $C_i(r,t) = g_i(r)c_i(t)$, where $n(t)$ and $c_i(t)$ represent the total neutron density and the total concentration of precursors of the $i$-th type at time $t$, respectively. Equations (56.1) now become

$$\frac{dn}{dt}(t) = Dv\frac{\nabla^2 f(r)}{f(r)}n(t) - (\Sigma_a - \Sigma_f)vn(t) + [(1-\beta)k_\infty\Sigma_a - \Sigma_f]vn(t) + \sum_i \lambda_i \frac{g_i(r)c_i(t)}{f(r)} + \frac{S_0(r,t)}{f(r)},$$

$$\frac{dc_i}{dt}(t) = \beta_i k_\infty \Sigma_a v \frac{f(r)n(t)}{g_i(r)} - \lambda_i c_i(t).$$

It is assumed that (i) $\frac{f(r)}{g_i(r)} = 1$; (ii) $f$ satisfies $\nabla^2 f + B^2 f = 0$; and (iii) $S_0$ has the same spatial dependence as $f$. If we write $q(t) = \frac{S_0(r,t)}{f(r)}$, the previous equations become

$$\frac{dn}{dt} = -DvB^2 n - (\Sigma_a - \Sigma_f)vn + [(1-\beta)k_\infty\Sigma_a - \Sigma_f]vn + \sum_i \lambda_i c_i + q, \tag{56.2a}$$

$$\frac{dc_i}{dt} = \beta_i k_\infty \Sigma_a vn - \lambda_i c_i. \tag{56.2b}$$

Furthermore, the terms in Eq. (56.2a) can be rearranged according to the type of neutron reaction:

$$\frac{dn}{dt} = \underbrace{-DvB^2 n - (\Sigma_a - \Sigma_f)vn}_{deaths} + \underbrace{(k_\infty\Sigma_a - \Sigma_f)vn}_{births} - \underbrace{\beta k_\infty \Sigma_a vn + \sum_i \lambda_i c_i}_{transformations} + q. \tag{56.3}$$

In order to simplify the notation, several parameters are now introduced. We define the absorption lifetime $l_\infty = \frac{1}{v\Sigma_a}$ and the diffusion length $L^2 = \frac{D}{\Sigma_a}$, and rewrite Eqs. (56.3) and (56.2b) as

$$\frac{dn}{dt} = \underbrace{\left[\frac{-L^2B^2 - \frac{(\Sigma_a - \Sigma_f)}{\Sigma_a}}{l_\infty}\right]}_{deaths} n + \underbrace{\left[\frac{k_\infty - \frac{\Sigma_f}{\Sigma_a}}{l_\infty}\right]}_{births} n - \underbrace{\frac{\beta k_\infty}{l_\infty} n + \sum_i \lambda_i c_i}_{transformations} + q, \tag{56.4a}$$

$$\frac{dc_i}{dt} = \frac{\beta_i k_\infty}{l_\infty} n - \lambda_i c_i. \tag{56.4b}$$

Defining the reproduction factor $k = \frac{k_\infty}{1+L^2B^2}$ and the neutron lifetime $l_0 = \frac{l_\infty}{1+L^2B^2}$, Eqs. (56.4) become

$$\frac{dn}{dt} = \left[-\frac{1}{l_0} + \frac{\Sigma_f}{\Sigma_a l_\infty}\right] n + \left[\frac{k}{l_0} - \frac{\Sigma_f}{\Sigma_a l_\infty}\right] n - \frac{\beta k}{l_0} n + \sum_i \lambda_i c_i + q, \tag{56.5a}$$

$$\frac{dc_i}{dt} = \frac{\beta_i k}{l_0} n - \lambda_i c_i. \tag{56.5b}$$

Next, we introduce the neutron generation time $l = \frac{l_0}{k}$. Substituting $l$ into Eqs. (56.5), we obtain

$$\frac{dn}{dt} = \left[-\frac{1}{kl} + \frac{\Sigma_f}{\Sigma_a l_\infty}\right] n + \left[\frac{1}{l} - \frac{\Sigma_f}{\Sigma_a l_\infty}\right] n - \frac{\beta}{l} n + \sum_i \lambda_i c_i + q,$$

$$\frac{dc_i}{dt} = \frac{\beta_i}{l} n - \lambda_i c_i.$$

Finally, we define reactivity $\rho = 1 - \frac{1}{k}$. Moreover, a simple algebraic calculation shows that $\frac{\Sigma_f}{\Sigma_a l_\infty} = \frac{\alpha}{l}$, where $\alpha = \frac{\Sigma_f}{\Sigma_a k_\infty} \approx \frac{1}{\nu}$ and $\nu$ is the number of neutrons per fission. Hence, the final deterministic system becomes

$$\frac{dn}{dt} = \underbrace{-\left[\frac{-\rho+1-\alpha}{l}\right] n}_{deaths} + \underbrace{\left[\frac{1-\alpha-\beta}{l}\right] n}_{births} + \underbrace{\sum_i \lambda_i c_i}_{transformations} + q, \tag{56.6a}$$

$$\frac{dc_i}{dt} = \frac{\beta_i}{l} n - \lambda_i c_i, \tag{56.6b}$$

for $i = 1, 2..., m$.

To derive the stochastic system, we first consider the case of just one precursor; that is, $\beta = \beta_1$. (The system will be generalized to $m$ precursors later.) Equations (56.6) for one precursor are written as

$$\frac{dn}{dt}(t) = \left\{-\left[\frac{-\rho+1-\alpha}{l}\right] + \left[\frac{1-\alpha-\beta}{l}\right]\right\} n(t) + \lambda_1 c_1(t) + q,$$

$$\frac{dc_1}{dt}(t) = \frac{\beta_1}{l} n(t) - \lambda_1 c_1(t).$$

We consider a time interval $\Delta t$ small enough to guarantee that the probability of more than one event occurring during $\Delta t$ is negligible. Let $[\Delta n, \Delta c_1]^T$ be a random vector variable that represents the changes in the neutron density and in the delayed neutron precursor concentration. The four possible events are

$$\begin{bmatrix} \Delta n \\ \Delta c_1 \end{bmatrix}_1 = \begin{bmatrix} -1 \\ 0 \end{bmatrix} = \text{death (capture)},$$

$$\begin{bmatrix} \Delta n \\ \Delta c_1 \end{bmatrix}_2 = \begin{bmatrix} -1+(1-\beta)\nu \\ \beta_1 \nu \end{bmatrix} = \begin{matrix} \text{birth (fission event and} \\ \text{production of delayed neutrons)}, \end{matrix}$$

$$\begin{bmatrix} \Delta n \\ \Delta c_1 \end{bmatrix}_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{matrix} \text{transformation of a delayed} \\ \text{neutron precursor to a neutron}, \end{matrix}$$

$$\begin{bmatrix} \Delta n \\ \Delta c_1 \end{bmatrix}_4 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \text{birth of a source neutron};$$

and the probabilities of these events (assuming $\alpha = \frac{1}{\nu}$) are

$$P_1 = \left( \frac{-\rho + 1 - \alpha}{l} \right) n\Delta t, \quad P_2 = \left( \frac{1}{\nu l} \right) n\Delta t, \quad P_3 = \lambda_1 c_1 \Delta t, \quad P_4 = q\Delta t.$$

It is also assumed that the neutron source produces neutrons randomly following a Poisson process with intensity $q$.

Finally, the mean change $E([\Delta n, \Delta c_1]^T)$ for the small time interval $\Delta t$ is given by

$$E\left( \begin{bmatrix} \Delta n \\ \Delta c_1 \end{bmatrix} \right) = \sum_{k=1}^{4} P_k \begin{bmatrix} \Delta n \\ \Delta c_1 \end{bmatrix}_k = \begin{bmatrix} \frac{p-\beta}{l} n + \lambda_1 c_1 + q \\ \frac{\beta_1}{l} n - \lambda_1 c_1 \end{bmatrix} \Delta t,$$

and the covariance of the change is given by

$$E\left( \begin{bmatrix} \Delta n \\ \Delta c_1 \end{bmatrix} \begin{bmatrix} \Delta n \Delta c_1 \end{bmatrix} \right) = \sum_{k-1}^{4} P_k \begin{bmatrix} \Delta n \\ \Delta c_1 \end{bmatrix} \begin{bmatrix} \Delta n \Delta c_1 \end{bmatrix}_k = \hat{B} \Delta t,$$

where $\hat{B}$ is defined as

$$\hat{B} = \begin{bmatrix} \gamma n + \lambda_1 c_1 + q & \frac{\beta_1}{l}(-1+(1-\beta)\nu)n - \lambda_1 c_1 \\ \frac{\beta_1}{l}(-1+(1-\beta)\nu)n - \lambda_1 c_1 & \frac{\beta_1^2 \nu}{l} n + \lambda_1 c_1 \end{bmatrix}$$

and $\gamma = \frac{-1-\rho+2\beta+(1-\beta)^2\nu}{l}$.

With the assumption that the changes are approximately normally distributed, the above results imply that, to $O((\Delta t)^2)$,

$$\begin{bmatrix} n(t+\Delta t) \\ c_1(t+\Delta t) \end{bmatrix} = \begin{bmatrix} n(t) \\ c_1(t) \end{bmatrix} + \hat{A} \begin{bmatrix} n(t) \\ c_1(t) \end{bmatrix} \Delta t + \begin{bmatrix} q \\ 0 \end{bmatrix} \Delta t + \hat{B}^{\frac{1}{2}} \sqrt{\Delta t} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix},$$

where $\eta_1, \eta_2 \sim \mathcal{N}(0,1)$, $\hat{B} = \hat{B}^{\frac{1}{2}} \cdot \hat{B}^{\frac{1}{2}}$, and $\hat{A} = \begin{bmatrix} \frac{p-\beta}{l} + \lambda_1 \\ \frac{\beta_1}{l} - \lambda_1 \end{bmatrix}$.

As $\Delta t \to 0$, the above equations yield the following Itô stochastic differential equation system [Hi01, RaPa13]:

$$\frac{d}{dt}\begin{bmatrix} n \\ c_1 \end{bmatrix} = \hat{A}\begin{bmatrix} n \\ c_1 \end{bmatrix} + \begin{bmatrix} q \\ 0 \end{bmatrix} + \hat{B}^{\frac{1}{2}}\frac{d\vec{W}}{dt}, \qquad \vec{W} = \begin{bmatrix} W_1(t) \\ W_2(t) \end{bmatrix}, \tag{56.7}$$

where $W_1(t)$ and $W_2(t)$ are Wiener processes. Equations (56.7) are the stochastic neutron point kinetics equations for one precursor group.

To generalize these equations to $m$ precursors, let

$$\hat{A} = \begin{bmatrix} \frac{\rho(t)-\beta}{l} & \lambda_1 & \lambda_2 & \dots & \lambda_m \\ \frac{\beta_1}{l} & -\lambda_1 & 0 & \dots & 0 \\ \frac{\beta_2}{l} & 0 & -\lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \frac{\beta_m}{l} & 0 & \dots & 0 & -\lambda_m \end{bmatrix}$$

and

$$\hat{B} = \begin{bmatrix} \zeta & a_1 & a_2 & \dots & a_m \\ a_1 & r_1 & b_{2,3} & \dots & b_{2,m+1} \\ a_2 & b_{3,2} & r_2 & \dots & b_{m,m+1} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_m & b_{m+1,2} & \dots & b_{m+1.m} & r_m \end{bmatrix},$$

where

$$\zeta = \gamma n + \sum_{i=1}^{m} \lambda_i c_i + q,$$

$$\gamma = \frac{-1-\rho+2\beta+(1-\beta)^2\nu}{l},$$

$$a_i = \frac{\beta_i}{l}(-1+(1-\beta)\nu)n - \lambda_i c_i,$$

$$b_{i,j} = \frac{\beta_{i-1}\beta_{j-1}\nu}{l}n,$$

$$r_i = \frac{{\beta_i}^2\nu}{l}n + \lambda_i c_i.$$

Using the same approach as before, but now for $m$ precursors, we obtain the Itô stochastic system:

$$\frac{d}{dt}\begin{bmatrix} n(t) \\ c_1(t) \\ c_2(t) \\ \vdots \\ c_m(t) \end{bmatrix} = \hat{A}\begin{bmatrix} n(t) \\ c_1(t) \\ c_2(t) \\ \vdots \\ c_m(t) \end{bmatrix} + \begin{bmatrix} q \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \hat{B}^{\frac{1}{2}}\frac{d\vec{W}}{dt}(t). \tag{56.8}$$

Note that if $\hat{B} = 0$, then Eq. (56.8) reduces to the standard deterministic point kinetics equations.

## 56.3 Numerical Results

We begin this section by briefly sketching the implementation of two approaches that address the stochastic behavior discussed in this work: (I) the Stochastic PCA model [HaAl05], and (II) the Euler–Maruyama approximation [Ra12]. Specific details of each implementation can be found in the references.

**(I)** *The Stochastic PCA model* is based on the system given in equation (56.8). For instance, assuming $m = 6$ delayed groups, this system can be written as

$$\frac{d\vec{x}}{dt} = A\vec{x} + B(t)\vec{x} + \vec{F}(t) + \hat{B}^{\frac{1}{2}}\frac{d\vec{W}}{dt}, \tag{56.9}$$

where $\hat{B}$ is already known and

$$A = \begin{bmatrix} \frac{-\beta}{l} & \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 & \lambda_5 & \lambda_6 \\ \frac{\beta_1}{l} & -\lambda_1 & 0 & 0 & 0 & 0 & 0 \\ \frac{\beta_2}{l} & 0 & -\lambda_2 & 0 & 0 & 0 & 0 \\ \frac{\beta_3}{l} & 0 & 0 & -\lambda_3 & 0 & 0 & 0 \\ \frac{\beta_4}{l} & 0 & 0 & 0 & -\lambda_4 & 0 & 0 \\ \frac{\beta_5}{l} & 0 & 0 & 0 & 0 & -\lambda_5 & 0 \\ \frac{\beta_6}{l} & 0 & 0 & 0 & 0 & 0 & -\lambda_6 \end{bmatrix}, \quad B(t) = \begin{bmatrix} \frac{\rho(t)}{l} & 0\,0\,0\,0\,0\,0 \\ 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0\,0\,0\,0\,0\,0 \end{bmatrix},$$

$$\vec{F}(t) = [q(t), 0, 0, 0, 0, 0, 0]^T, \quad \vec{x} = [n, c_1, c_2, c_3, c_4, c_5, c_6]^T.$$

The source function $q(t)$ and the reactivity function $\rho(t)$ are approximated by piecewise constant functions; in particular,

$$\rho(t) \approx \rho\left(\frac{t_i + t_{i+1}}{2}\right) = \rho_i, \quad \text{for } t_i \le t \le t_{i+1}$$

and

$$B(t) \approx B\left(\frac{t_i + t_{i+1}}{2}\right) = B_i, \quad \text{for } t_i \leq t \leq t_{i+1}.$$

Now, for $t_i \leq t \leq t_{i+1}$, equation (56.9) becomes

$$\frac{d\vec{x}}{dt} = A\vec{x} + B_i\vec{x} + \vec{F}(t) + \hat{B}^{\frac{1}{2}}\frac{d\vec{W}}{dt},$$

and using Itô's formula [KlPl92] we obtain

$$\frac{d}{dt}\left[e^{-(A+B_i)t}\vec{x}\right] = e^{-(A+B_i)t}\vec{F}(t) + e^{-(A+B_i)t}\hat{B}^{\frac{1}{2}}\frac{d\vec{W}}{dt}.$$

Finally, this equation is approximated using Euler's method, and the eigenvalues and eigenvectors of the matrix $(A + B_i)$ are computed using diagonalization.

**(II)** *The Euler–Maruyama approximation* performs the time-discrete approximation of an Itô process. Let $\{X_t\}$ be an Itô process on $t \in [t_0, T]$ that satisfies the stochastic differential equation $dX_t = a(t, X_t)dt + b(t, X_t)dW_t$, $X_{t_0} = X_0$. For a given time-discretization $t_0 < t_1 < t_2 < \ldots < t_N = T$, an Euler–Maruyama approximation is a continuous time stochastic process $\{Y(t), t_0 \leq t \leq T\}$ that satisfies the interactive scheme given by [KlPl92]

$$Y_{n+1} = Y_n + a(t_n, Y_n)\Delta t_{n+1} + b(t_n, Y_n)\Delta W_{n+1}, \quad n = 0, 1, \ldots, N-1,$$

where $Y_0 = X_0$, $Y_n = Y(t_n)$, $\Delta t_{n+1} = t_{n+1} - t_n$, and $\Delta W_{n+1} = W(t_{n+1}) - W(t_n)$. Each random number is given by $\Delta W_n = z_n\sqrt{\Delta t_n}$, where $z_n$ is chosen from a standard normal distribution $\mathcal{N}(0, 1)$. In this type of procedure the considered time intervals must be equidistant.

In the following sections we consider examples with constant and linear reactivity, and present Monte Carlo (MC) simulations for each one of them. We compare the MC estimates to the results obtained with the stochastic models previously discussed, as well as with the Deterministic Model [PeVi09, WoLe14].

In these MC simulations, we have chosen the time interval $\Delta t$ to be small enough such that the likelihood of more than one event taking place during $\Delta t$ is very small. This was achieved by considering the half-life time of the precursor groups, according to the time decay constants $\lambda_i$. The number of seeds used in the MC estimates for each case was large enough to guarantee that the statistical error of the mean values is less than 0.05% (with 95% confidence).

### 56.3.1 Constant Reactivity

In the following examples, we present the results for the mean values $E$ of the neutron density and the delayed neutron precursors concentration. In addition, we also present the standard deviations $\sigma$ of these quantities for the stochastic models.

In the first example, we reproduce a test case presented in [HaAl05], which assumes only one neutron precursor and simulates a step-reactivity insertion. Although it does not model an actual physical nuclear reactor problem, it provides simple computational solutions for comparison with Monte Carlo results. The parameters are $\lambda_1 = 0.1$, $\beta_1 = 0.005$, $\nu = 2.5$, $q = 200$, $l = \frac{2}{3}$, and $\rho = -\frac{1}{3}$, with equilibrium values for the initial condition: $\vec{x}(0) = [400, 300]^T$.

Table 56.1 shows that, while the standard deviations for both quantities are one order of magnitude smaller than their corresponding mean values, the standard deviation for the neutron density is still significant ($\approx 7\%$ of the mean). This suggests that a deterministic approach may not be sufficient for the computation of this quantity.

The next example (two scenarios) uses $m = 6$ delayed neutron precursor groups, and models step reactivity insertions for an actual nuclear reactor [ChAt85, HaAl05, KiAl04]. The first scenario models a prompt insertion with $\rho = 0.003$, whereas the second scenario models a prompt insertion with $\rho = 0.007$. In both scenarios the parameters are chosen as follows:

$$\lambda_i = [0.0127, 0.0317, 0.115, 0.311, 1.4, 3.87];$$
$$\beta_i = [0.000266, 0.001491, 0.001316, 0.002849, 0.000896, 0.000182];$$
$$\beta = 0.007; \quad \nu = 2.5; \quad q = 0; \quad l = 0.00002;$$

with an initial condition that assumes a source-free equilibrium:

$$\vec{x}(0) = 100 \left[ \frac{\beta_1}{\lambda_1 l}, \frac{\beta_2}{\lambda_2 l}, \frac{\beta_3}{\lambda_3 l}, \frac{\beta_4}{\lambda_4 l}, \frac{\beta_5}{\lambda_5 l}, \frac{\beta_6}{\lambda_6 l} \right]^T.$$

It is important to point out that the issue of stiffness arises when solving the stochastic models for these scenarios. This puts an additional constraint in the probability calculations. As in the previous example, an analysis of the standard

**Table 56.1** Results for one precursor group and reactivity $\rho = -1/3$.

| | Monte Carlo | Stochastic PCA | Euler-Maruyama approximation | Deterministic model |
|---|---|---|---|---|
| $E(n(2))$ | 400.032 | 395.32 | 412.23 | 412.13 |
| $\sigma(n(2))$ | 27.311 | 29.411 | 34.391 | – |
| $E(c_1(2))$ | 300.01 | 300.67 | 315.96 | 315.93 |
| $\sigma(c_1(2))$ | 7.807 | 8.3564 | 8.2656 | – |

**Table 56.2** Results for six precursor groups and reactivity $\rho = 0.003$.

| | Monte Carlo | Stochastic PCA | Euler-Maruyama approximation | Deterministic model |
|---|---|---|---|---|
| $E(n(0.1))$ | 183.04 | 186.31 | 208.6 | 200.005 |
| $\sigma(n(0.1))$ | 168.79 | 164.16 | 255.95 | – |
| $E(\sum_{i=1}^{6} c_i(0.1))$ | $4.478\times10^5$ | $4.491\times10^5$ | $4.498\times10^5$ | $4.497\times10^5$ |
| $\sigma(\sum_{i=1}^{6} c_i(0.1))$ | 1495.72 | 1917.2 | 1233.38 | – |

**Table 56.3** Results for six precursor groups and reactivity $\rho = 0.007$.

| | Monte Carlo | Stochastic PCA | Euler-Maruyama approximation | Deterministic model |
|---|---|---|---|---|
| $E(n(0.001))$ | 135.66 | 134.55 | 139.568 | 139,61 |
| $\sigma(n(0.001))$ | 93.376 | 91.242 | 92.042 | – |
| $E(\sum_{i=1}^{6} c_i(0.001))$ | $4.464\times10^5$ | $4.694\times10^5$ | $4.463\times10^5$ | $4.463\times10^5$ |
| $\sigma(\sum_{i=1}^{6} c_i(0.001))$ | 16.226 | 19.444 | 6.071 | – |

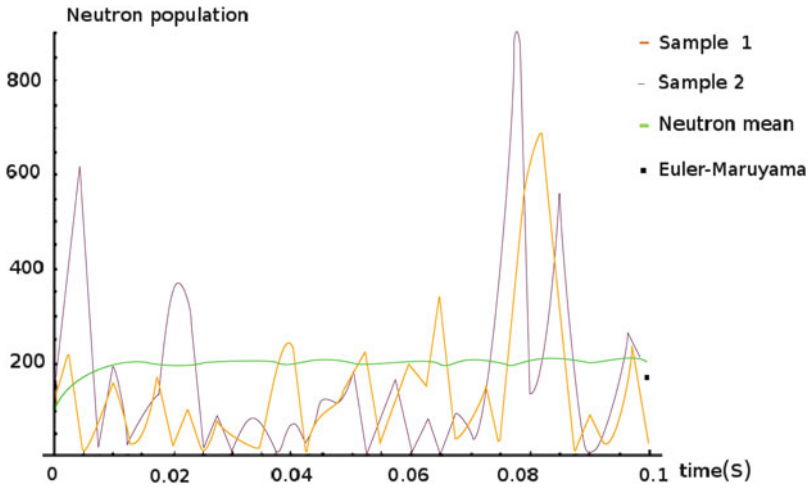


**Fig. 56.1** Neutron Density for six Precursor Groups with reactivity $\rho = 0.007$.

deviations in Tables 56.2 and 56.3 indicates that the stochastic effects need to be taken under consideration, since the values obtained for the mean and the standard deviation of the neutron density are of the same order of magnitude.

Besides the evaluation for a fixed time $t = 0.1s$ by the Euler-Maruyama approach, we also generate the time line (Figure 56.1) of the neutron density and compare two Monte Carlo realizations (Sample 1 and Sample 2) with the mean value of the neutron density after averaging over a sufficiently large set of samples.

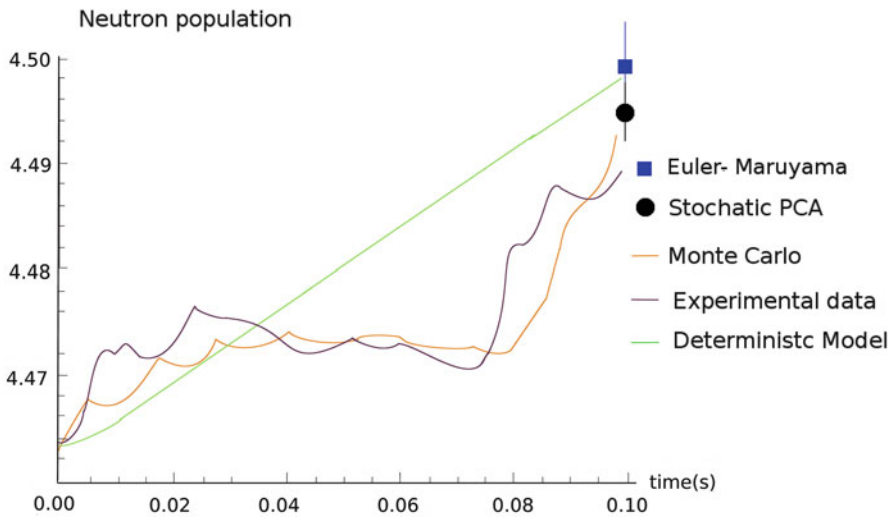


**Fig. 56.2** Neutron Density for one precursor group and reactivity $\rho(t) = 0.25t$.

### *56.3.2 Linear Reactivity*

The example discussed in this section is, to the best of our knowledge, the first study of this kind that considers time-dependent reactivity. We provide Monte Carlo results for an example with one precursor group and linear reactivity (see Figure 56.2) and compare our findings to experimental data [Ha60] as well as to the deterministic model prediction [PeVi09, WoLe14]. For the time $t = 0.1s$, the Stochastic PCA and Euler-Maruyama results are indicated. The parameter set used for this simulation is $\lambda_1 = 0.1$, $\beta_1 = 0.005$, $\nu = 2.5$, $l = 0.00001$ with time-dependent reactivity $\rho(t) = 0.25t$ and with initial condition $\vec{x}(0) = 100[1, \frac{\beta_1}{\lambda_1 l}]^T$.

We note that, while the deterministic model yields a curve with the correct qualitative behavior, it fails to provide any information on the stochastic fluctuations of the neutron population over time. Clearly, a model that can predict these fluctuations would be an improvement over the deterministic approach.

## 56.4 Discussion

From the phenomenological point of view, it is evident that one needs to take under consideration the stochastic effects in order to compute the neutron density. This is confirmed by the results of the simulations we have presented, where we see that the values for the mean and standard deviation of the neutron density can be of the same order of magnitude. The examples presented here also suggest that the fluctuations

in the precursor concentrations are small. This behavior arises from the stochastic nature of decay; specifically, from the property of time homogeneity inherent to the radioactive decay law.

The present work is the first one in a sequence, in which reactivity of time-dependent scenarios and the effects of stochastic moments are studied. This will be done by solving the stochastic equation in a hierarchic fashion: first, the deterministic part of the problem is solved, and then the solution is modified by including the stochastic moments. This contrasts with the procedures currently found in the literature, which make use of the roots of the inhour equation. One of the main difficulties encountered refers to the stiffness of the problem, which imposes severe restrictions on the calculation of the event probabilities. In a future work these issues will be addressed in an optimized solution procedure.

**Acknowledgements** M.T.B.V. and R.V. wish to thank CNPq, and M.W.d.S. wishes to thank CAPES, for financial support.

## References

[AbHa03] Aboander, A.E., Hamada, Y.M.: Power series solution (PWS) of nuclear reactor dynamics with Newtonian temperature feedback. Ann. Nucl. Energy **30**, 1111–1122 (2003)

[ChAt85] Chao, Y., Attard, A.: A resolution to the stiffness problem of reactor kinetics. Nucl. Sci. Eng. **90**, 40–46 (1985)

[Ha60] Hansen, G.E.: Assembly of fissionable material in the presence of a weak neutron source. Nucl. Sci. Eng. **8**, 709–719 (1960)

[HaAl05] Hayes, J.G., Allen, E.J.: Stochastic point-kinetics equations in nuclear reactor dynamics. Ann. Nucl. Energy **32**, 572–587 (2005)

[He71] Hetrick, D.L.: Dynamics of Nuclear Reactors. University of Chicago Press, Chicago (1971)

[Hi01] Higham, D.J.: An algorithmic introduction to numerical simulation of stochastic differential equations. SIAM Rev. **43**, 525–546 (2001)

[KiAl04] Kinard, M., Allen, E.J.: Efficient numerical solution of the point kinetics equations in nuclear reactor dynamics. Ann. Nucl. Energy **31**, 1039–1051 (2004)

[KlPl92] Kloeden, P.E. and Platen, E.: Numerical Solution of Stochastic Differential Equations, Springer, New York, 1992.

[PeVi09] Petersen, Z.C., Vilhena, M.T, Dulla, S., Ravetto, P.: An analytical solution of the point kinetics equations with time variable reactivity by the decomposition method. In: International Nuclear Atlantic Conference, pp. R16–43 (2009)

[Ra12] Ray, S. Saha: Numerical simulation of stochastic point kinetic equation in the dynamical system of nuclear reactor. Ann. Nucl. Energy **49**, 154–159 (2012)

[RaPa13] Ray, S. Saha, Patra, A.: Numerical solution of fractional stochastic neutron point kinetic equation for nuclear reactor dynamics. Ann. Nucl. Energy **54**, 154–161 (2013)

[Sa89] Sánchez, J.: On the numerical solution of the point reactor kinetics equations by generalized Runge-Kutta methods. Nucl. Sci. Eng. **103**, 94–99 (1989)

[WoLe14] Wollmann da Silva, M., Leite, S.B., Vilhena, M.T., Bodmann, B.E.J.: On an analytical representation for the solution of the neutron point kinetics equation free of stiffness. Ann. Nucl. Energy **71**, 97–102 (2014)

# Chapter 57
# The Hamilton Principle for Mechanical Systems with Impacts and Unilateral Constraints

**K. Yunt**

## 57.1 Introduction

An action integral is presented for Hamiltonian mechanics in canonical form with unilateral constraints and/or impacts. The transition conditions on generalized impulses and the energy are presented as variational inequalities, which are obtained by the application of discontinuous transversality conditions. The energetical behavior for elastic, plastic, and blocking type impacts is analyzed. A general impact equation is obtained by the stationarity conditions, which is compatible with the most general impact laws and is applicable to various impactive processes straightforwardly. The crux in achieving energetical behavior which conforms with the physics of the impactive process is shown to be the consistency conditions on the impact time variations.

Hamilton postulated in 1835 in his seminal works [Ha34, Ha35], that if a Lagrangian system occupies certain positions at fixed times $t_0$ and $t_f$, then it should move between these two positions along those admissible arcs $q(t) \in C_n^1[t_0, t_f]$, which make the action integral

$$J(q) = \int_{t_0}^{t_f} L(q(s), \dot{q}(s))\, ds$$

stationary. The Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is defined as $L = T(q, \dot{q}) - V(q)$, where $V(q)$ and $T(q, \dot{q})$ represent the potential and kinetic energy, respectively. The original form of Hamilton's principle deals with conservative systems with equality constraints, which are perfect bilateral constraints. Though the Hamiltonian mechanics was born later than Lagrangian mechanics, the domains of physics in

K. Yunt (✉)
General Control Design, Zürich, Switzerland
e-mail: kerimyunt@web.de



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_57

which its formalism is used even reach to more modern branches of physics, such as relativistic/quantum mechanics [Fe65]. The following Legendre transformation on the generalized velocities: $p = \partial_{\dot{q}} L$ where p is called the conjugate generalized momentum, yields the Hamiltonian canonical equations as the stationarity condition:

$$\dot{p} = -\partial_q H, \quad \dot{q} = \partial_p H$$

in smooth conservative motion.

The missing link in analytical mechanics which shows that general impactive processes are obtained by extremizing some sort of action integral for which momentum and energy are not necessarily conserved is recently presented in [Yu12] for elastic contact impacts in the Lagrangian formalism. In [Yu12] the conditions, under which general non-conserving impacts become a part of an extremizing solution for mechanical systems, which are scleronomic and holonomic, are investigated. The general momentum balance and the total energy change over a collisional impact for a mechanical scleronomic holonomic finite-dimensional Lagrangian system are obtained in the form of stationarity conditions of a modified action integral. The reference [Yu12] has been preceded/succeeded by many works such as [St65, FeEtAl03, Si81, PaGl98, PaGl00, KoTr91, LeAeGl09, PeMu12], which were not able to present impulsive action integrals for impacts without energy conservation.

In [Yu13] blocking as a dissipative impactive process is analyzed by the technique in [Yu12] and an impulsive action integral in the Lagrangian formalism is presented. In this work, by making use of the results obtained in [Yu12] and [Yu13] the impactive principle of stationary action for impactive processes are obtained by maximizing, for which momentum and energy are not necessarily conserved over the impact in the Hamiltonian formalism.

In this work, a smooth Riemannian configuration manifold $\mathscr{M}$, for which $q$ denotes the n-tuple of generalized local coordinates is considered. The kinetic metric associated with $\mathscr{M}$ is given by $M(q)$ at each $q$. The generalized velocity of the system $\dot{q}$ lives in the tangent space of the manifold $T_{\mathscr{M}}(q)$. If the motion of the system is constrained to a submanifold of $\mathscr{M}$ denoted by the admissible set $\mathscr{C}$, then the tangent space $T_{\mathscr{M}}(q)$ is subdivided into a pair of cones $T_{\mathscr{C}}(q)$ and $T^{\perp}_{\mathscr{C}}(q)$, which are orthogonal to each other in the kinetic metric. The cotangent space is denoted by $N_{\mathscr{C}}(q)$ and the cones $T_{\mathscr{C}}(q)$ and $T^{\perp}_{\mathscr{C}}(q)$ are subspaces of $T_{\mathscr{M}}(q)$. It is assumed that the constraint structure may differ in the pre-impact and post-impact phases. The total energy of the scleronomic holonomic Lagrangian system is given by its total mechanical energy $H(q,p) = T(q,p) + V(q)$. The differential measure of the total mechanical energy is given by $dH(q,p) = \partial_t H(q,p)\, dt + \partial_\sigma H(q,p)\, d\sigma$. The absolutely continuous part of the measure $dH$ is denoted by $\frac{dH}{dt}$. The singular part of $dH$ is represented as $\frac{dH}{d\sigma}$, where $d\sigma$ a regular Borel measure, and $\frac{dH}{d\sigma}$ is the Radon–Nikodym derivative of $dH$ with respect to $d\sigma$. The Lebesgue–Stieltjes integration of the differential measure of the total mechanical energy over the

impact time yields $\int_{\{t_s\}} dH = H^+ - H^- = T^+ - T^- = L^+ - L^-$. The regularity of the pre-impact and post-impact transition sets at the instant and position of impact is an assumption of local convexity. The irregularity of the constraint set at an instant of impact is visualized in mechanics in the form of inward/re-entrant corners as discussed in [Ma87] and [Gl02]. If at the location of impactive transition the regularity is not present either at pre-transition and/or post-transition state, in the sense that the contingent cone does not overlap with the tangent cone, then the obtained stationarity conditions are weakened to substationarity conditions, and the variational inequalities are termed as hemi-variational or quasi-variational inequalities. Having set the stage, the impulsive action integral becomes

$$J(q,p,t_s) = \int_{t_0}^{t_s^-} \langle p, \dot{q} \rangle - H(q,p)\, ds + \int_{t_s^+}^{t_f} \langle p, \dot{q} \rangle - H(q,p)\, ds = J_1(q,t_s) + J_2(q,t_s). \tag{57.1}$$

The following main theorem is proven in this work for generalized positions from the space of absolutely continuous functions $AC$ and for conjugate momenta from the space of locally bounded variation functions LBV:

**Theorem 1 (Main Theorem).** *If there exist arcs $\tilde{q} \in AC_n[t_0, t_f]$ and $\tilde{p} \in LBV_n[t_0, t_f]$, impact position $\tilde{q}(\tilde{t}_s)$, pre-impact and post-impact conjugate momenta $\tilde{p}(\tilde{t}_s^-)$ and $\tilde{p}(\tilde{t}_s^+)$ at an impact time $\tilde{t}_s$ because of a impactive process at multiple contacts/locations, which induces the system, which moves in $\mathscr{C}^-$, to evolve on the constraint $\mathscr{C}^+$ and if these arcs provide for the action integral in (57.1) a maximizer, then the following conditions hold:*

1. *The Hamilton canonical equations on $[t_0, t_f]$ in the almost everywhere sense*

$$\dot{\tilde{p}}_j = -\partial_{q_j} H(\tilde{q},\tilde{p}), \quad \dot{\tilde{q}}_j = \partial_{p_j} H(\tilde{q},\tilde{p}), \quad j = 1,2,\ldots,n. \tag{57.2}$$

2. *The conjugate momentum balance:*

$$\tilde{p}^+ - \tilde{p}^- = D^+(\tilde{q}(\tilde{t}_s))\, \tilde{\Lambda}^+ + D^-(\tilde{q}(\tilde{t}_s))\, \tilde{\Lambda}^-. \tag{57.3}$$

   *Here the matrices $D^-$ and $D^+$ are the pre-impact and post-impact generalized impulse direction. The vectors $\tilde{\Lambda}^-$ and $\tilde{\Lambda}^+$ are Lagrange multipliers/impulses for which $\tilde{\Lambda}^+ \in N_{\mathscr{C}^+}(q)$ and $\tilde{\Lambda}^- \in N_{\mathscr{C}^-}(q)$ hold.*
3. *A finite amount of energy is removed or added to the system.*

The theory of subgradients and variational inequalities has found in the recent decades many fields of application. The basics of nonsmooth variational analysis can be retrieved in the classical work of [Cl90] by Clarke.

## 57.2 Internal Boundary Variations (IBV) and Discontinuous Transversality Conditions (DTC)

If any isolated instant of discontinuity is considered as a boundary on the timeline, then this boundary constitutes an upper boundary for one segment of the interval whereas for the other segment a lower boundary and becomes an internal boundary. The pre-transition and post-transition variations are interrelated by the transition conditions. The discontinuous transversality conditions follow straightforwardly by the evaluation of the corresponding variational inequalities to the internal boundary variations. Several families of variational curves, which are parameterized by a nonnegative $\varepsilon$, are introduced in order to generate the variations

$$
\begin{aligned}
p(t,\varepsilon) &= p(t) + \varepsilon\,\hat{p}(t) = p(t) + \delta p(t),\, q(t,\varepsilon) = q(t) + \varepsilon\,\hat{q}(t) = q(t) + \delta q(t),\\
q(t_s^+,\varepsilon) &= q(t_s^+) + \varepsilon\,\hat{q}(t_s^+) = q(t_s^+) + \delta q(t_s^+),\\
q(t_s^-,\varepsilon) &= q(t_s^-) + \varepsilon\,\hat{q}(t_s^-) = q(t_s^-) + \delta q(t_s^-),\, t_s(\varepsilon) = t_s + \varepsilon\,\hat{t}_s = t_s + \delta t_s.
\end{aligned}
$$

The variations of the pre- and post-transition positions at fixed time $\hat{q}(t_s^+)$, $\hat{q}(t_s^-)$ are defined by Gâteaux derivatives and are related with the total variations in these entities $\hat{q}_s^+$, $\hat{q}_s^-$ at the internal boundary by the following affine relations as depicted in figure 57.1:

$$
\hat{q}(t_s^+) = \hat{q}_s^+ - \dot{q}(t_s^+)\,\hat{t}_s,\; \hat{q}(t_s^-) = \hat{q}_s^- - \dot{q}(t_s^-)\,\hat{t}_s. \tag{57.4}
$$

The variations of the post-transition and pre-transition positions are

$$
\delta q_s^+ = \delta q\left(t_s^+\right) + \dot{q}\left(t_s^+\right)\,\delta t_s,\; \delta q_s^- = \delta q\left(t_s^-\right) + \dot{q}\left(t_s^-\right)\,\delta t_s. \tag{57.5}
$$

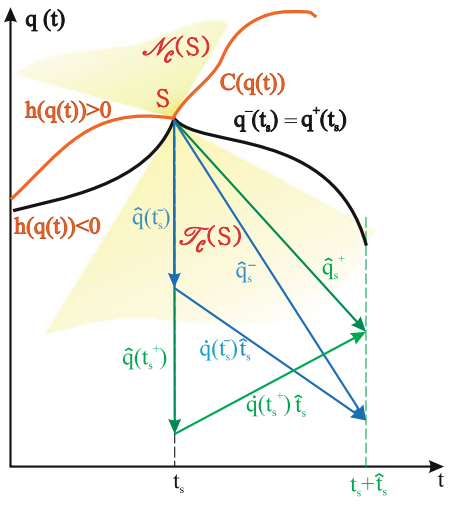


**Fig. 57.1** General decomposition of boundary variations at a reentrant corner.

The tangent cone to a set $\mathscr{C}$ at a given point x in the regular case is

$$T_{\mathscr{C}}(x) = \left\{ y \in K \,|\, \lim_{\tau \downarrow 0} \frac{d_{\mathscr{C}}(x+\tau y)}{\tau} = 0 \right\}.$$

Here $d_{\mathscr{C}}(x)$ denotes the distance function of the point $x$ to the set $\mathscr{C}$, where it takes the value zero, if and only if $x \in \mathscr{C}$. The tangent cone $T_{\mathscr{C}}(x)$ to a set $\mathscr{C} \subset K$ is polar to a nonempty convex cone $N_{\mathscr{C}}(x)$ in the dual space $K^*$:

$$N_{\mathscr{C}}(x) = \{ z \in K^* \,|\, \langle y, z \rangle \leq 0, \quad y \in T_{\mathscr{C}}(x) \}. \tag{57.6}$$

The allowable pre-impact and post-impact position variations are limited to

$$\delta q_s^+ \in T_{\mathscr{C}^+}\left(q(t_s^+)\right), \quad \delta q_s^- \in T_{\mathscr{C}^-}\left(q(t_s^-)\right).$$

The continuity of the positions requires the equality of the pre-impact and post-impact position variations:

$$\delta q_s^+ = \delta q_s^- = \delta q_s. \tag{57.7}$$

According to (57.7), $\delta q_s$ fulfills both the pre-impact and post-impact conditions

$$\delta q_s \in T_{\mathscr{C}^+} \wedge \delta q_s \in T_{\mathscr{C}^-} \Rightarrow \delta q_s \in T_{\mathscr{C}^+}(q_s) \cap T_{\mathscr{C}^-}(q_s).$$

The following set relations hold:

$$T_{\mathscr{C}^+}(q_s) \cap T_{\mathscr{C}^-}(q_s) \equiv T_{\mathscr{C}^+ \cap \mathscr{C}^-}(q_s),$$
$$N_{\mathscr{C}^+ \cap \mathscr{C}^-}(q_s) \equiv N_{\mathscr{C}^+}(q_s) \oplus N_{\mathscr{C}^-}(q_s).$$

Here $\oplus$ denotes the set addition. The equality (57.7) means that:

$$\delta q\left(t_s^+\right) - \delta q\left(t_s^-\right) = \left(\dot{q}\left(t_s^+\right) - \dot{q}\left(t_s^-\right)\right) \delta t_s$$

must hold in general.

The following condition is valid in phases of motion, where the generalized velocities are continuous: $\dot{q} \in T_{\mathscr{C}}(q)$.

At an instant of velocity jump, which may be accompanied by alteration in the pre-impact and post-impact constraint sets, it is necessary to distinguish among forward and backward dynamics:

$$\dot{q}^+ \in T_{\mathscr{C}^+}\left(q(t_s^+)\right), \quad \dot{q}^- \in T_{\mathscr{C}^-}\left(q(t_s^-)\right).$$

The allowable pre-impact and post-impact velocity variations are expressed as

$$\delta \dot{q}_s^+ \in T_{T_{\mathscr{C}^+}\left(q(t_s^+)\right)}\left(\dot{q}(t_s^+)\right), \quad \delta \dot{q}_s^- \in T_{T_{\mathscr{C}^-}\left(q(t_s^-)\right)}\left(\dot{q}(t_s^-)\right).$$

Analogously, the spaces of the variations of the pre-impact and post-impact conjugate momenta read:

$$\delta p^+ \in N^{\perp}_{T_{\mathscr{C}^+}\left(q(t_s^+)\right)}\left(p(t_s^+)\right), \quad \delta p^- \in N^{\perp}_{T_{\mathscr{C}^-}\left(q(t_s^-)\right)}\left(p(t_s^-)\right).$$

The set in $T_{\mathscr{C}^+\cap\mathscr{C}^-}\left(q(t_s^+)\right)$ should cover all possible candidate directions for the internal boundary variations, so that the obtained extremizing arc is the extremizer over all comparison curves. This property is guaranteed by the tangential regularity of the transition sets.

It is assumed that the constraint sets $\mathscr{C}^+$ and $\mathscr{C}^-$ are regular, and the impactive process does not impair thereby the regularity of the post-impact set.

## 57.3 Stationary Nature of the Impulsive Action Integral

The regularity of the functional in (57.1) is guaranteed by the regularity of the integrands and of the transition sets $\mathscr{C}^+$ and $\mathscr{C}^-$. The functional J is defined on Banach space and is assumed to be Lipschitz. The directional derivative of $f(x)$, if $f(x)$ is Lipschitz around $x$ is defined by

$$f^0(x;y) = \limsup_{\substack{x' \to x \\ \varepsilon \downarrow 0}} \frac{f(x'+\varepsilon y) - f(x')}{\varepsilon}. \tag{57.8}$$

The subdifferential is $\partial f(x) := \left\{z \in Z \,|\, f^0(x;y) \geq \langle z, y\rangle, \forall y \in X\right\}$. If $f$ is continuously differentiable, then $\partial f(x)$ consists of a single element, namely, $\partial f(x) = \{\nabla_x f(x)\}$. The function $y \to f^0(x;y)$ is finite, positively homogeneous, and subadditive on X, and satisfies $|f^0(x;y)| \leq K\|y\|$. The directional derivative $f^0(x;y)$ is as a function of y Lipschitz of rank K on X. The following relation holds: $-f^0(x;y) = f^0(x;-y)$ [Cl90]. If f attains a local minimum or maximum at $x$, then the zero vector is an element of the subdifferential: $0 \in \partial f(x)$. For every y in X, the directional majorizes the expression $f^0(x;y) = \max\{\langle \xi, y\rangle \, \forall \xi \in \partial f(x)\}$.

### *57.3.1 Proof of the Main Theorem*

Let $q(t) \in AC^1[t_0, t_f]$ be an arc and $t_s$ be an transition time for which the action integral $J(q(t), p(t), t_s)$ is well-defined and finite. The arc $\tilde{q}(t) \in AC^1[t_0, t_f]$ and $\tilde{t}_s \in \mathbb{R}$ is a weak local maximum for (57.1), if there exist $\varepsilon > 0$ and $\varepsilon_t > 0$ such that every $\hat{q}(t) \in AC^1[t_0, t_f]$ with $\|\hat{q}(t)\|_\infty + \|\hat{p}(t)\|_\infty < \varepsilon$ and $\|\hat{t}\| < \varepsilon_t$ gives rise to a well-defined objective value $J(\tilde{q}(t) + \varepsilon\hat{q}(t), \tilde{p}(t) + \varepsilon\hat{p}(t), \tilde{t}_s + \varepsilon_t\hat{t}_s)$, which satisfies

$$J(\tilde{q}(t) + \varepsilon\hat{q}(t), \tilde{p}(t) + \varepsilon\hat{p}(t), \tilde{t}_s + \varepsilon_t\hat{t}_s) \leq J(\tilde{q}(t), \tilde{p}(t), \tilde{t}_s). \tag{57.9}$$

If there exist arcs $\tilde{q}$ and $\tilde{p}$, transition position $\tilde{q}(\tilde{t}_s)$, pre-transition and post-transition conjugate momenta $\tilde{p}(\tilde{t}_s^-)$ and $\tilde{p}(\tilde{t}_s^+)$ at a transition time $\tilde{t}_s$, which all together maximize the functional in (57.1), such that the value functional assumes the finite value $\tilde{J}(\varepsilon = 0) = J(\tilde{q}, \tilde{t}_s)$, then the following variational inequality is fulfilled:

$$\sum_{\forall \hat{\psi}_j} \tilde{J}^0(\cdot; \varepsilon\, \hat{\psi}_j) \leq 0, \forall \hat{\psi}_j \in \{\hat{q}(t_s), \hat{t}_s\} \bigcup \{\hat{q}, \hat{p}\}, \tag{57.10}$$

since $J$ is subdifferentially regular at any extremal solution. The following one parameter functionals are used to transform (57.10):

$$G_i(\varepsilon) = \left\langle p + \varepsilon\, \hat{p}\,, \dot{q} + \varepsilon\, \hat{\dot{q}} \right\rangle - H_i(q + \varepsilon\, \hat{q}, p + \varepsilon\, \hat{p}), \quad i = 1, 2, \tag{57.11}$$

into the following stationarity condition:

$$\begin{aligned}
&\limsup_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \left[ \int_{t_0}^{t_s^- + \varepsilon \hat{t}} G_1(\varepsilon)\, ds - \int_{t_0}^{t_s^-} G_1(0)\, ds \right] \qquad (57.12)\\
&+ \frac{1}{\varepsilon} \left[ \int_{t_s^+ + \varepsilon \hat{t}}^{t_f} G_2(\varepsilon)\, ds - \int_{t_s^+}^{t_f} G_2(0)\, ds \right] =\\
&\limsup_{\varepsilon \to 0^+} \int_{t_0}^{t_s^-} \frac{G_1(\varepsilon) - G_1(0)}{\varepsilon}\, ds + \int_{t_s^+}^{t_f} \frac{G_2(\varepsilon) - G_2(0)}{\varepsilon}\, ds\\
&+ \int_{t_s^-}^{t_s^- + \varepsilon \hat{t}} \frac{G_1(\varepsilon)}{\varepsilon}\, ds + \int_{t_s^+ + \varepsilon \hat{t}}^{t_s^+} \frac{G_2(\varepsilon)}{\varepsilon}\, ds =\\
&\int_{t_0}^{t_s^-} \limsup_{\varepsilon \to 0^+} \frac{G_1(\varepsilon) - G_1(0)}{\varepsilon}\, ds + \int_{t_s^-}^{t_s^- + \varepsilon \hat{t}} \limsup_{\varepsilon \to 0^+} \frac{G_1(\varepsilon)}{\varepsilon}\, ds +\\
&\int_{t_s^+}^{t_f} \limsup_{\varepsilon \to 0^+} \frac{G_2(\varepsilon) - G_2(0)}{\varepsilon}\, ds + \int_{t_s^+ + \varepsilon \hat{t}}^{t_s^+} \limsup_{\varepsilon \to 0^+} \frac{G_2(\varepsilon)}{\varepsilon}\, ds\\
&= \int_{t_0}^{t_s^-} \langle \hat{p}, \dot{q} \rangle + \langle p, \hat{\dot{q}} \rangle - \left\langle \frac{\partial H_1}{\partial q}, \hat{q} \right\rangle - \left\langle \frac{\partial H_1}{\partial p}, \hat{p} \right\rangle ds\\
&+ \int_{t_s^+}^{t_f} \langle \hat{p}, \dot{q} \rangle + \langle p, \hat{\dot{q}} \rangle - \left\langle \frac{\partial H_2}{\partial q}, \hat{q} \right\rangle + \left\langle \frac{\partial H_2}{\partial p}, \hat{p} \right\rangle ds\\
&+ \left\langle p(t_s^-)\, \dot{q}(t_s^-) \right\rangle \hat{t} - \left\langle p(t_s^+), \dot{q}(t_s^+) \right\rangle \hat{t} + H_1(q(t_s^-), p(t_s^-))\, \hat{t}\\
&- H_2(q(t_s^+), p(t_s^+))\, \hat{t} \leq 0. \qquad (57.13)
\end{aligned}$$

The Raymond–Dubois lemma states that

$$\langle p(s), \hat{q}(s) \rangle |_a^b = \int_a^b \langle \dot{p}(s), \hat{q}(s) \rangle + \langle p(s), \hat{\dot{q}}(s) \rangle\, ds. \tag{57.14}$$

By inserting (57.14) and (57.5) in (57.12) the following is obtained:

$$\begin{aligned}&\int_{t_0}^{t_s^-} \left\langle \dot{q}(s) - \partial_p H_1 \, , \hat{p}(s) \right\rangle - \left\langle \partial_q H_1 + \dot{p}(s) \, , \hat{q}(s) \right\rangle ds \\ &+ \int_{t_s^+}^{t_f} \left\langle \dot{q}(s) - \partial_p H_2 \, , \hat{p}(s) \right\rangle - \left\langle \partial_q H_2 + \dot{p}(s) \, , \hat{q}(s) \right\rangle ds \\ &\left\langle p(t_s^-) - p(t_s^+) \, , \hat{q}_s \right\rangle + \left[ H_1^- - H_2^+ \right] \hat{t} \leq 0.\end{aligned} \tag{57.15}$$

in the limit. By the application of Fatou's lemma on the integral in (57.15) the Hamilton's canonical equations are obtained in the almost everywhere sense as given in (57.2).

The validity of (57.7) relates their directional derivatives under regularity of $J_2$ and $J_1$ at $(q(t_s), t_s)$:

$$J^0 \left( \cdot ; \varepsilon \, \hat{q}_s \right) = J_1^0 \left( \cdot ; \varepsilon \, \hat{q}_s^- \right) + J_2^0 \left( \cdot ; \varepsilon \, \hat{q}_s^+ \right). \tag{57.16}$$

The condition for a maximum is:

$$J^0 \left( \cdot ; \varepsilon \, \hat{q}_s \right) \leq 0, \quad \forall \hat{q}_s \in T_{(\mathscr{C}^+ \cap \mathscr{C}^-)} \left( \tilde{q} \right). \tag{57.17}$$

By making use of definition (57.6), this optimality condition is fulfilled for:

$$\tilde{p}(t_s^-) - \tilde{p}(t_s^+) \in N_{(\mathscr{C}^+ \cap \mathscr{C}^-)} \left( \tilde{q} \right). \tag{57.18}$$

The inclusion (57.18) is equivalently expressed as:

$$\tilde{p}(t_s^+) - \tilde{p}(t_s^-) = \tilde{D}^+ \tilde{\Lambda}^+ + \tilde{D}^- \tilde{\Lambda}^-. \tag{57.19}$$

The directional derivative of $J$ in the direction $\delta t_s$ is:

$$J^0 \left( \cdot ; \varepsilon \, \hat{t}_s \right) = J_1^0 \left( \cdot ; \varepsilon \, \hat{t}_s \right) + J_2^0 \left( \cdot ; \varepsilon \, \hat{t}_s \right) = \left( H^+ - H^- \right) \delta t_s, \tag{57.20}$$

where $H^-$ and $H^+$ denote the pre-impact and post-impact Hamiltonians, respectively. The variational inequality pertaining to the impact time variation yields the optimality condition:

$$J^0 \left( \cdot ; \varepsilon \, \hat{t}_s \right) = \left( H^+ - H^- \right) \delta t_s \leq 0. \tag{57.21}$$

### 57.3.2 The Consistency Conditions on the Time Variation in Different Scenarios

By the property, that boundary variations of the generalized positions consist of two components, which are independent of each other as in (57.5), it is assumed that

$$\delta q(t_s^+) \in T_{\mathscr{C}^+}\left(q(t_s^+)\right) \ \wedge \ \dot{q}(t_s^+)\,\delta t_s \in T_{\mathscr{C}^+}\left(q(t_s^+)\right), \tag{57.22}$$

$$\delta q(t_s^-) \in T_{\mathscr{C}^-}\left(q(t_s^-)\right) \ \wedge \ \dot{q}(t_s^-)\,\delta t_s \in T_{\mathscr{C}^-}\left(q(t_s^-)\right). \tag{57.23}$$

## 57.4 Elastic Rigid Body Collisions

A given vector-valued differentiable function $g(q)$ ($h(q) = -g(q)$) represents the shortest distances between the rigid bodies in the system and these distances are always nonnegative (non-positive) due to the impenetrability assumption. If elements of g become zero, then contact among rigid bodies occurs and the mechanical system reaches the boundary of the admissible set $\mathscr{C}$. At the instant of a multi-impact at m contacts, at which $g(q) = 0$ is valid, the pre-impact and post-impact transition sets are identical, so that for the generalized impulse directions $\mathscr{C}^- = \mathscr{C}^+$ holds, and the impact equation takes the well-known form

$$\tilde{p}^+ - \tilde{p}^- = D\left(\tilde{q}(\tilde{t}_s)\right)\left(\tilde{\Lambda}^+ + \tilde{\Lambda}^-\right). \tag{57.24}$$

In rigid body collisions one has a non-positive pre-impact relative velocities in the approach phase, and nonnegative post-impact relative velocities due to the rigidity or impenetrability condition, which are stated as $v^- = D^T(q)\,\dot{q}^-$ and $v^+ = D^T(q)\,\dot{q}^+$, respectively. The linear operator $D(q)$ is defined as $D(q) = \nabla_q g(q)$. Since in this case the impactive transition sets $\mathscr{C}^+$ and $\mathscr{C}^-$ coincide, the set is defined by $\mathscr{C} = \{q(t_s^+) | h(q(t_s)) \leq 0\}$, respectively, then the position variations at pre-impact and post-impact instants are in the tangent cone

$$T_{\mathscr{C}}\left(q\left(t_s^+\right)\right) = T_{\mathscr{C}} = \left\{\xi \,|D^T\left(q\left(t_s\right)\right)\xi \geq 0\right\}, \tag{57.25}$$

respectively, if $h(q(t_s)) = 0$. The normal cone of the set $\mathscr{C}$ is

$$N_{\mathscr{C}}\left(q(t_s)\right) = \left\{-D\left(q\left(t_s\right)\right)\xi \,|\, \xi \geq 0\right\},$$

respectively. For the variational parts pertaining to impact time variations, the inclusions in (57.22) and (57.23) mean that

$$v^+\delta t_s \geq 0, \quad v^-\delta t_s \geq 0. \tag{57.26}$$

Since $v^+ > 0$ and $v^- < 0$ need to hold because of rigidity assumption, then the impact time variations must vanish $\hat{t}_s = 0$, in order to maintain the consistency of the internal boundary variations. The unboundedness of the energy change $H^+ - H^-$ would contradict the assumption that the generalized velocities belong to the class of bounded variation functions.

## 57.5 Impactive Processes Arising Due to Blocking and Non-Smooth Constraints

The case is investigated by the author in [Yu13] by making use of DTC and its similarity to fully inelastic impacts is stated. At an instant of blocking, the directions characterized by $D^+$ are after the transition time, abruptly closed for evolution, which requires $D^+ \dot{q} = 0$ to hold.

Consider the motion of a mass particle without friction and gravity along an ideal holonomic constraint with a kink at the origin as shown in figure 57.2. The particle is supposed to move along the line $g_1(x,y) : y = 0$ until the origin, and to follow the line $g_2(x,y) : y - cx = 0$ beginning at the origin. It has a pre-impact velocity of $\dot{x}^-$. The impactive process is modelled by a release of the constraint $g_1(x,y)$ and blocking of constraint $g_2(x,y)$ at the kink. The pre-impact and post-impact generalized directions become $D^- = \begin{pmatrix} 0 & 1 \end{pmatrix}^T, \quad D^+ = \begin{pmatrix} -c & 1 \end{pmatrix}^T$.

In [Yu13] it is shown that the change in the conjugate momentum and the total mechanical energy for a releasing-blocking type transition is

$$p^+ - p^- = -D^+ G_{bb}^{-1} v^- = D^+ \Lambda^+ + D^- \Lambda^- \tag{57.27}$$

and

$$H^+ - H^- = -\frac{1}{2} \left\langle v^- , G_{bb}^{-1} v^- \right\rangle , \tag{57.28}$$

respectively, where the Delassus' operator $G_{bb}$ is: $G_{bb} = {}^+D^T M^{-1} D^+$ and $v^-$ denotes the pre-impact relative velocity with respect to the blocking constraint:

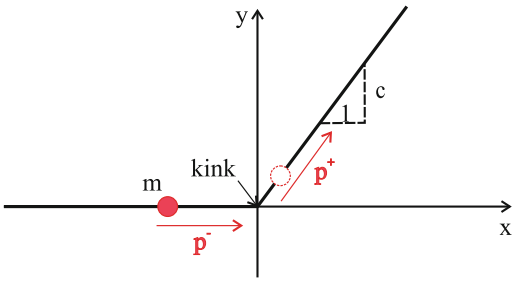


**Fig. 57.2** Point mass moving along a non-smooth constraint.

$v^- = D^+\dot{q}^-$. Substituting relevant entities to the example in the relations (57.27) and (57.28) the following is obtained:

$$H^+ - H^- = -\frac{mc^2}{2(1+c^2)}(\dot{x}^-)^2, \quad p^+ - p^- = \begin{pmatrix} -\frac{c^2}{(1+c^2)} \\ \frac{c}{(1+c^2)} \end{pmatrix} m\dot{x}^-$$

such that the conjugate momenta after the impact become

$$p_x^+ = \frac{m}{(1+c^2)}\dot{x}^-, \quad p_y^+ = \frac{mc^2}{(1+c^2)}\dot{x}^-$$

and the impulses are given by $\Lambda^- = 0, \quad \Lambda^+ = \frac{mc\dot{x}^-}{(1+c^2)}$. The decrease in energy is correctly characterized by the discontinuous transversality conditions, and the right-hand side of the impact equation describes an impactive process for which the change the constraint structure induces an impact. In the case of non-smooth constraints and blocking action, the impact equation is driven by the post-impact impulses.

## 57.6 Impacts Accompanied by Alteration in the Mass Structure

Consider, for example, a stellar or atomic scale collision process in which one of the participating objects disappears through energy emittance and/or the compressive phase impulses exceeds a certain threshold such that one 'bursts.' Due to the nonexistence of a post-impact constraint structure $\mathscr{C}^+$, one is only left with the condition (57.23), which translates in this case into

$$^-D^T\left(q(t_s^-)\right)\dot{q}_s^- \leq 0 \wedge \dot{q}(t_s^-)\,\delta t_s \in T_{\mathscr{C}^-} \Rightarrow \hat{t}_s \leq 0. \tag{57.29}$$

The condition (57.29) requires that $\hat{t}_s \leq 0$. If this is considered, together with the stationarity condition (57.21), then a decrease in the total mechanical energy is required: $H^- \geq H^+$. The energy decrease is here interpretable, as the activation energy required to dissolve the other particle in the form of compressive work in a Poisson type impact process and the conversion to other types of energies from mechanical energy. The impact equation here involves only the compression/pre-impact part:

$$\tilde{p}^+ - \tilde{p}^- = {}^-D(\tilde{q}(\tilde{t}_s))\,\tilde{\Lambda}^-.$$

## 57.7 Discussion and Conclusions

The introduced Hamiltonian framework is capable of providing an impact equation which is able to cope with impactive processes, for which

1. The energy is not conserved, and dissipative impacts become eligible.
2. The mass distribution may change abruptly.
3. The constraint structure due to addition and removal of constraints is altered suddenly.

The impact equation which is obtained by the proposed variational formulation is compatible with the most common impact equations such as Newton's or Poisson's impact law. The Poisson's impact law, which requires to distinguish between compression phase and decompression phase impulsive forces $\Lambda^-$ and $\Lambda^+$, and interrelates them by a coefficient of restitution in the form $\Lambda^+ = \varepsilon_p \Lambda^-$, is directly structurally adaptable to the impactive process due to the clear distinction between the pre-impact and post-impact phases. Different than in many preceding studies, which were limited to energy conserving impactive processes, the variational framework for the Hamiltonian formulation enables the implementation of arbitrary restitution coefficients with Newton's and Poisson's impact laws. The equation (57.8) is in comparison with the assumptions $\delta q(t_s^+) = \dot{q}(t_s^+)\,\delta t_s, \quad \delta q(t_s^-) = \dot{q}(t_s^-)\,\delta t_s$, which are used in the works [LeAeGl09] and [FeEtAl03], an improvement, because it enables nonconservative impactive processes to become extremizing arcs. It is straightforward to show that the conditions of the main theorem are valid also in the strong norm. A detailed account and a comparison between the Weierstrass–Erdmann conditions and discontinuous transversality conditions for impactive processes of finite-dimensional Lagrangian systems is given in [Yu12]. Irrespective whether the set $\mathscr{C}^+ \cap \mathscr{C}^-$ is tangentially regular at $q(t_s)$ or not, the generalized impulse exists in $N_{\mathscr{C}^+\cap\mathscr{C}^-}(q(t_s))$. The vast majority of the literature on the stationarity principle focuses o determining the stationarity conditions based on the classical rule of Fermat which requires: $\delta J = 0$ for all admissible variations. If the functional J is merely Lipschitz, then the stationarity conditions for a minimum and a maximum need to be investigated separately, which require $\delta J \geq 0$ or $\delta J \leq 0$, respectively. If the stationarity conditions for a non-smooth minimum is investigated in the presented case, then the impulse and energy balance equations become physically incorrect and by sign reversal.

The generalized directional derivative of Clarke does not require the existence of any limit in the vicinity of the point of interest, and involves only the behavior of the functional near the stationary point, which is in the sense of the analysis presented here, because of the discontinuity of the generalized conjugate momenta, the integrand does not exist at an instant of impact and the stationarity conditions are to be understood in the limiting sense.

## References


[Ha34] Hamilton, W. R.: On a general method in dynamics. Philos. Trans. R. Soc. Lond., 247–308 (1834)

[Ha35] Hamilton, W. R.: Second Essay On a General Method in Dynamics. Philos. Trans. R. Soc. Lond., 95–144 (1835)

[Fe65] Feynman, R. P.: The character of physical law. MIT Press, Cambridge (1965)

[Yu12] Yunt, K.: The Impulsive Action Integral for Rigid-Body Mechanical Systems With Impacts. J. Comput. Nonlinear Dyn. **7/3**, 031012-031012-9 (2012)

[St65] Stavrakova, N. E.: The principle of hamilton-ostrogradskii for systems with one-sided constraints. J. Appl. Math. Mech. **29/4**, 874–878 (1965)

[FeEtAl03] Fetecau, R. C., Marsden, J. E. and Ortiz, M. and West, M.: Nonsmooth Lagrangian mechanics and variational collision integrators. SIAM J. Appl. Dyn. Syst. **2/3**, 381–416 (2003)

[Si81] Sinitsyn, V. A.: On the Hamilton-Ostrogradskii principle in the case of impulsive motions of dynamic systems. J. Appl. Math. . Mech. **45/3**, 356–359 (1981)

[PaGl98] Panagiotopoulos, P.D., Glocker, Ch.: Analytical Mechanics. Addendum I: Inequality Constraints with Elastic Impacts. The Convex Case. ZAMM Z. Angew. Math. Mech. Z **78/4**, 219–229 (1998)

[PaGl00] Panagiotopoulos, P.D., Glocker, Ch.: Inequality constraints with elastic impacts in deformable bodies. The convex case. Arch. Appl. Mech. **70/5**, 349–365 (2000)

[KoTr91] Kozlov, V. V., Treshchëv, D. V.: Billiards: A Generic Introduction to the Dynamics of Systems with Impacts. American Mathematical Soc., XXXX (1991)

[LeAeGl09] Leine, R. I., Aeberhard, U., Glocker, C.: Hamilton's Principle as Variational Inequality for Mechanical Systems with Impact. J. Nonlinear Sci. **19/6**, 633–664 (2009)

[PeMu12] Pekarek, D., Murphey, T.D.: Variational nonsmooth mechanics via a projected Hamilton's principle. American Control Conference (ACC),IEEE, 1040–1046 (2012)

[Yu13] Yunt, K.: Analysis of Discrete Mechanical Systems With Blockable Directions. J. Appl. Mech. **7/3**, 031012-031012-9 (2012)

[Cl90] Clarke, F.H.: Optimisation and Nonsmooth Analysis. SIAM 5, (1990)

[Gl02] Glocker, Ch.: Impacts with global dissipation index at reentrant corners. In: Martins, J. A. C., Monteiro Marques, M. D. P. (eds.) Contact Mechanics, pp. 45–52. Springer, (2002)

[Ma87] May, H.O.: Generalized Variational Principles and Unilateral Constraints in Analytical Mechanics. In: Unilateral Problems in Structural Analysis-2: Proceedings of the Second Meeting on Unilateral Problems in Structural Analysis, pp. 221–237. Springer, (1987)

# Chapter 58
# Numerical Solutions and Their Error Bounds for Oscillatory Neural Networks

**B. Zubik-Kowal**

## 58.1 Introduction

In this chapter, we investigate a mathematical model introduced by Hoppensteadt and Izhikevich [HoIz99] written in the form of nonlinear Volterra integro-differential equations of convolution type:

$$\frac{dx_i}{dt}(t) = \Omega_i + \varepsilon \int_0^t \delta(t-s)\, a(s) \sum_{j=1}^{N} \sin\Big(x_j(s) - x_i(s)\Big) ds, \tag{58.1}$$

where $i = 1, 2, \ldots, N$, $t \in [0, T]$. Equations (58.1) are known as thalamo-cortical equations and describe a new architecture for a neurocomputer composed of $N$ oscillators.

The above system (58.1) generalizes the model by Kuramoto studied in [Ku84]. The $i$th solution $x_i(t)$ is the phase of the $i$th oscillator at time $t$. System (58.1) is modeled by employing principles of the human brain [HoIz99] and the $N$ oscillators generating different frequencies $\Omega_i$ ($\Omega_i \neq \Omega_j$ for $i \neq j$) are referred to as neurons and are forced by the thalamic input $a(t)$. We apply the input defined by

$$a(t) = \cos^2(\pi t),$$

for $t \geq 0$, and the convolution kernel defined by the step function

$$\delta(t) = \begin{cases} c, & \text{for } 0 \leq t \leq \tau, \\ 0, & \text{for } t > \tau, \end{cases}$$

B. Zubik-Kowal (✉)
Boise State University, 1910 University Drive, Boise, ID 83725, USA
e-mail: zubik@math.boisestate.edu



C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5_58

with $c, \tau > 0$ being positive parameters. The oscillators are homogeneously and weakly connected to their common medium and the strength of their connections is given in terms of the parameter $0 < \varepsilon \ll 1$. Numerical solutions to the system (58.1) will be illustrated in a later section of this chapter.

For problems of this type, the number of oscillators $N$ is usually taken to be a relatively large number, making (58.1) is a large system of strongly coupled differential equations. Difficulties associated with limitations on available resources inevitably arise in solving such problems [ZuVa99, Zu00, JaWeZu04, JaZu05], and parallelization can often serve as a resolution for some of them [JaZu06, Zu06, HoJa07, MiZu12]. In the next sections, we address the problem on how to solve the system introduced in this chapter using parallel environments.

The organization of this chapter is as follows. In Section 58.2, we describe dynamic iterations for the thalamo-cortical system. Error bounds showing the convergence of the successive iterates are derived in Section 58.3. Results of numerical experiments are presented in Section 58.4. We then finish with concluding remarks in Section 58.5.

## 58.2 Numerical Solutions

Solving thalamo-cortical systems (58.1) numerically is computationally expensive and robust and efficient numerical algorithms are vital for computer simulations. To solve (58.1) numerically in parallel computing environments, we apply the dynamic iteration

$$\frac{d}{dt}x_i^{(k+1)}(t) = \Omega_i + \varepsilon \int_0^t \delta(t-s)\, a(s) \sum_{j=1}^{N} \sin\left(x_j^{(k)}(s) - x_i^{(k)}(s)\right) ds, \tag{58.2}$$

where $k = 0, 1, 2, \ldots$, $i = 1, \ldots, N$, and $x_i^{(0)} : [0,T] \to \mathbb{R}$ are arbitrary starting functions. System (58.2) differs from static iterations by the fact that the successive iterates $x^{(k)} : [0,T] \to \mathbb{R}^N$ are vector functions.

The advantage of (58.2) over (58.1) is that each equation in (58.2) is separated from the other equations in the system (the right-hand side of (58.2) depends only on the previous iterate $x^{(k)}(s)$) while (58.1) is composed of strongly joined equations. Therefore, for each $k$, each equation in (58.2) can be solved on a separate processor working independently on $[0,T]$.

If $N$ processors are available, then each processor (the $i$th processor, say) is assigned to solve its corresponding, $i$th, equation. If only a smaller number of processors is available, then the system of $N$ equations can be divided into larger groups composed of more than one equation and each group is then assigned to a separate processor.

To investigate the convergence of the iteration process (58.2) we define the errors

$$e_i^{(k)}(t) = x_i^{(k)}(t) - x_i(t),$$
$$e^{(k)}(t) = \left[e_1^{(k)}(t), \ldots, e_N^{(k)}(t)\right]^T$$

and the maximum norm

$$\|e^{(k)}(t)\| = \max\{|e_i^{(k)}(t)| : 1 \le i \le N\}.$$

It can be seen from (58.1) and (58.2) that

$$\begin{aligned}\frac{d}{dt}e_i^{(k+1)}(t) &= \varepsilon \int_0^t \delta(t-s)\,a(s) \sum_{j=1}^N \sin\left(x_j^{(k)}(s) - x_i^{(k)}(s)\right) ds \\ &\quad - \varepsilon \int_0^t \delta(t-s)\,a(s) \sum_{j=1}^N \sin\left(x_j(s) - x_i(s)\right) ds \\ &= \varepsilon \int_0^t \delta(t-s)\,a(s) \sum_{j=1}^N \cos\left(\xi_{ij}^{(k)}(s)\right)\left(e_j^{(k)}(s) - e_i^{(k)}(s)\right) ds,\end{aligned} \tag{58.3}$$

where $\xi_{ij}^{(k)}(s)$ are between $x_j^{(k)}(s) - x_i^{(k)}(s)$ and $x_j(s) - x_i(s)$. We now integrate the last relation in (58.3) and get

$$\begin{aligned}e_i^{(k+1)}(t) &= e_i^{(k+1)}(0) \\ &\quad + \varepsilon \int_0^t \int_0^z \delta(z-s)\,a(s) \sum_{j=1}^N \cos\left(\xi_{ij}^{(k)}(s)\right)\left(e_j^{(k)}(s) - e_i^{(k)}(s)\right) ds dz.\end{aligned}$$

Since $x_i^{(k)}(0) = x_i(0)$, it can be seen from the above relation that

$$\begin{aligned}|e_i^{(k+1)}(t)| &\le 2\varepsilon(N-1) \int_0^t \int_0^z \delta(z-s)\,|a(s)| \sum_{j=1}^N \|e^{(k)}(s)\| ds dz \\ &\le 2\varepsilon(N-1) \int_0^t \int_0^z \delta(z-s) \sum_{j=1}^N \|e^{(k)}(s)\| ds dz.\end{aligned}$$

We now take the maximum for all $i = 1, \ldots, N$ on both sides and get

$$\|e^{(k+1)}(t)\| \le 2\varepsilon(N-1) \int_0^t \int_0^z \delta(z-s) \sum_{j=1}^N \|e^{(k)}(s)\| ds dz. \tag{58.4}$$

In the next section, we apply the recurrence relation (58.4) and use the step function $\delta$ to derive error bounds for $\|e^{(k)}(t)\|$.

## 58.3 Error Bounds

To simplify the next steps, we introduce the following notation:

$$\lambda = 2\varepsilon(N-1)$$

and

$$E(t) = \max_{s\in[0,t]} \|e^{(0)}(s)\|.$$

Then from (58.4) we get

$$\begin{aligned} \|e^{(k+1)}(t)\| &\le \lambda c \int_0^t \int_0^z \delta(z-s)\|e^{(k)}(s)\|dsdz \\ &\le \lambda c \int_0^t \int_0^z \|e^{(k)}(s)\|dsdz, \end{aligned} \tag{58.5}$$

which helps to prove the following result.

**Theorem 1.** *The k-th dynamic iteration $x^{(k)}(t)$ satisfies the following error bound*

$$\|e^{(k)}(t)\| \le E(t)\big(\lambda c\big)^k \frac{t^{2k} - \big(\max\{t-\tau,0\}\big)^{2k}}{(2k)!}, \tag{58.6}$$

*for $t \ge 0$ and $k = 1,2,3,\ldots$.*

Note that a consequence of Theorem 1, in particular the inequality (58.6), is that the dynamic iteration (58.2) converges for any starting vector function $x^{(0)}(t)$, $t \ge 0$.

*Proof.* We consider two cases: $t > \tau$ and $0 \le t \le \tau$. Suppose $t > \tau$. Then, from (58.5), we get

$$\begin{aligned} \|e^{(k)}(t)\| &\le \lambda c \int_0^t \int_0^z \delta(z-s)\|e^{(k-1)}(s)\|dsdz \\ &= -\lambda c \int_0^t \int_z^0 \delta(r)\|e^{(k-1)}(z-r)\|drdz \\ &= \lambda c \int_0^t \int_0^z \delta(s)\|e^{(k-1)}(z-s)\|dsdz \\ &= \lambda c \int_0^\tau \int_0^z \delta(s)\|e^{(k-1)}(z-s)\|dsdz \\ &+ \lambda c \int_\tau^t \int_0^z \delta(s)\|e^{(k-1)}(z-s)\|dsdz \end{aligned}$$

$$
\begin{aligned}
&= \lambda c \int_0^{\tau} \int_0^{z} \delta(s) \| e^{(k-1)}(z-s) \| ds dz \\
&+ \lambda c \int_{\tau}^{t} \int_0^{\tau} \delta(s) \| e^{(k-1)}(z-s) \| ds dz \\
&+ \lambda c \int_{\tau}^{t} \int_{\tau}^{z} \delta(s) \| e^{(k-1)}(z-s) \| ds dz
\end{aligned}
$$

and since the $\delta(s)$ function is identically equal to unity in the first two components and vanishes in the third component on the right-hand side of the above inequality, it can be seen that

$$
\begin{aligned}
\| e^{(k)}(t) \| &\leq \lambda c \int_0^{\tau} \int_0^{z} \| e^{(k-1)}(z-s) \| ds dz + \lambda c \int_{\tau}^{t} \int_0^{\tau} \| e^{(k-1)}(z-s) \| ds dz \\
&= -\lambda c \int_0^{\tau} \int_z^{0} \| e^{(k-1)}(r) \| dr dz - \lambda c \int_{\tau}^{t} \int_z^{z-\tau} \| e^{(k-1)}(r) \| dr dz
\end{aligned}
$$

and

$$
\| e^{(k)}(t) \| \leq \lambda c \int_0^{\tau} \int_0^{z} \| e^{(k-1)}(s) \| ds dz + \lambda c \int_{\tau}^{t} \int_{z-\tau}^{z} \| e^{(k-1)}(s) \| ds dz. \tag{58.7}
$$

Therefore, for $k = 1$, we get

$$
\begin{aligned}
\| e^{(1)}(t) \| &\leq \lambda c \int_0^{\tau} \int_0^{z} \| e^{(0)}(s) \| ds dz + \lambda c \int_{\tau}^{t} \int_{z-\tau}^{z} \| e^{(0)}(s) \| ds dz \\
&\leq \lambda c E(\tau) \int_0^{\tau} \int_0^{z} ds dz + \lambda c E(t) \int_{\tau}^{t} \int_{z-\tau}^{z} ds dz \\
&= \lambda c E(\tau) \frac{\tau^2}{2} + \lambda c E(t) \tau (t-\tau) \leq \lambda c E(t) \Big( t\tau - \frac{\tau^2}{2} \Big) \\
&= \lambda c E(t) \frac{1}{2} \big( t^2 - t^2 + 2t\tau - \tau^2 \big) = \lambda c E(t) \frac{1}{2} \Big( t^2 - (t-\tau)^2 \Big) \\
&= \lambda c E(t) \frac{1}{2} \Big( t^2 - \big( \max \{ 0, t-\tau \} \big)^2 \Big),
\end{aligned}
$$

which shows that (58.6) holds for $k = 1$. We now assume (58.6) holds for a certain $k$. Then, from the relation (58.7), we get

$$
\begin{aligned}
\| e^{(k+1)}(t) \| &\leq \lambda c \int_0^{\tau} \int_0^{z} \| e^{(k)}(s) \| ds dz + \lambda c \int_{\tau}^{t} \int_{z-\tau}^{z} \| e^{(k)}(s) \| ds dz \\
&\leq \lambda c \int_0^{\tau} \int_0^{z} (\lambda c)^k E(s) \frac{s^{2k}}{(2k)!} ds dz \\
&+ \lambda c \int_{\tau}^{t} \int_{z-\tau}^{z} (\lambda c)^k E(s) \frac{s^{2k}}{(2k)!} ds dz \\
&\leq (\lambda c)^{k+1} E(t) \Big( \int_0^{\tau} \int_0^{z} \frac{s^{2k}}{(2k)!} ds dz + \int_{\tau}^{t} \int_{z-\tau}^{z} \frac{s^{2k}}{(2k)!} ds dz \Big)
\end{aligned}
$$

$$= (\lambda c)^{k+1} E(t) \frac{1}{(2k+1)!} \Big( \int_0^{\tau} z^{2k+1} dz + \int_{\tau}^{t} z^{2k+1} - (z-\tau)^{2k+1} dz \Big)$$
$$= (\lambda c)^{k+1} E(t) \frac{1}{(2k+2)!} \Big( \tau^{2k+2} + t^{2k+2} - \tau^{2k+2} - (t-\tau)^{2k+2} \Big)$$
$$= (\lambda c)^{k+1} E(t) \frac{t^{2k+2} - \big(\max\{0, t-\tau\}\big)^{2k+2}}{(2k+2)!},$$

which shows that the result (58.6) holds for all $k = 1, 2, \ldots$ and $t > \tau$. We now consider $t$ such that $0 \le t \le \tau$. In this case, from (58.5), we get

$$\|e^{(k)}(t)\| \le \lambda c \int_0^t \int_0^z \delta(z-s) \|e^{(k-1)}(s)\| ds dz$$
$$= \lambda c \int_0^t \int_0^z \|e^{(k-1)}(s)\| ds dz, \tag{58.8}$$

for $k = 1, 2, 3, \ldots$, which implies that

$$\|e^{(1)}(t)\| \le \lambda c \int_0^t \int_0^z \|e^{(0)}(s)\| ds dz \le E(t) \lambda c \int_0^t \int_0^z ds dz = E(t) \lambda c \frac{t^2}{2!}.$$

and shows that the result (58.6) holds for $k = 1$. Assuming that (58.6) is satisfied for a certain $k$, from (58.8), we get

$$\|e^{(k+1)}(t)\| \le \lambda c \int_0^t \int_0^z \|e^{(k)}(s)\| ds dz \le \lambda c \int_0^t \int_0^z E(s) (\lambda c)^k \frac{s^{2k}}{(2k)!} ds dz$$
$$\le E(t) (\lambda c)^{k+1} \int_0^t \frac{z^{2k+1}}{(2k+1)!} dz = E(t) (\lambda c)^{k+1} \frac{t^{2k+2}}{(2k+2)!},$$

which shows that (58.6) holds for $t$ such that $0 \le t \le \tau$ and for all $k = 1, 2, 3, \ldots$. This finishes the proof of the theorem.

We now apply the dynamic iterations for numerical simulations with the model (58.1).

## 58.4 Numerical Experiments

Results of numerical experiments with the model (58.1) are presented in Figures 58.1, 58.2, and 58.3 with different numbers of equations $N$. To solve (58.1) we apply scheme (58.2) and use the numerical solutions $x_j^{(k)}(t)$ for computing the mean field activity function of the network defined by

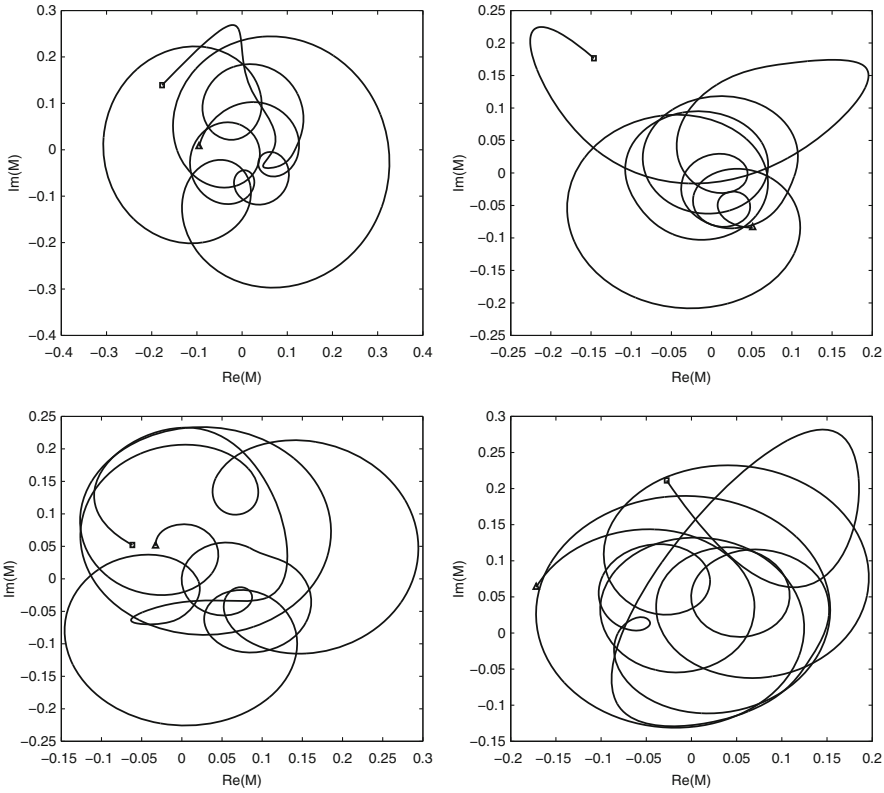


**Fig. 58.1** Numerical solutions with $N = 50$ for $t \in [0, 10]$ and random initial conditions.

$$M(t) = \frac{1}{N} \sum_{j=1}^{N} \exp(\imath x_j^{(k)}(t)), \tag{58.9}$$

where $\imath$ is the imaginary unit.

We solved system (58.1) on the time interval $[0, 10]$ with $N = 50$, $N = 100$, and $N = 150$ and initial conditions randomly distributed on the interval $[0, 2\pi]$. Figures 58.1, 58.2, and 58.3 illustrate the trajectories of the resulting mean field activity functions (58.9) in the complex plane, each subfigure for a separate random distribution of initial conditions. Figure 58.1 illustrates the trajectories with $N = 50$, Figure 58.2 with $N = 100$, and Figure 58.3 with $N = 150$. The numerical experiments were performed with the distinct frequencies of the oscillators defined by

$$\Omega_j = 2\pi \frac{j-1}{N-1},$$

for $j = 1, 2, \ldots, N$, and with strength of connections given by $\varepsilon = 0.01$.

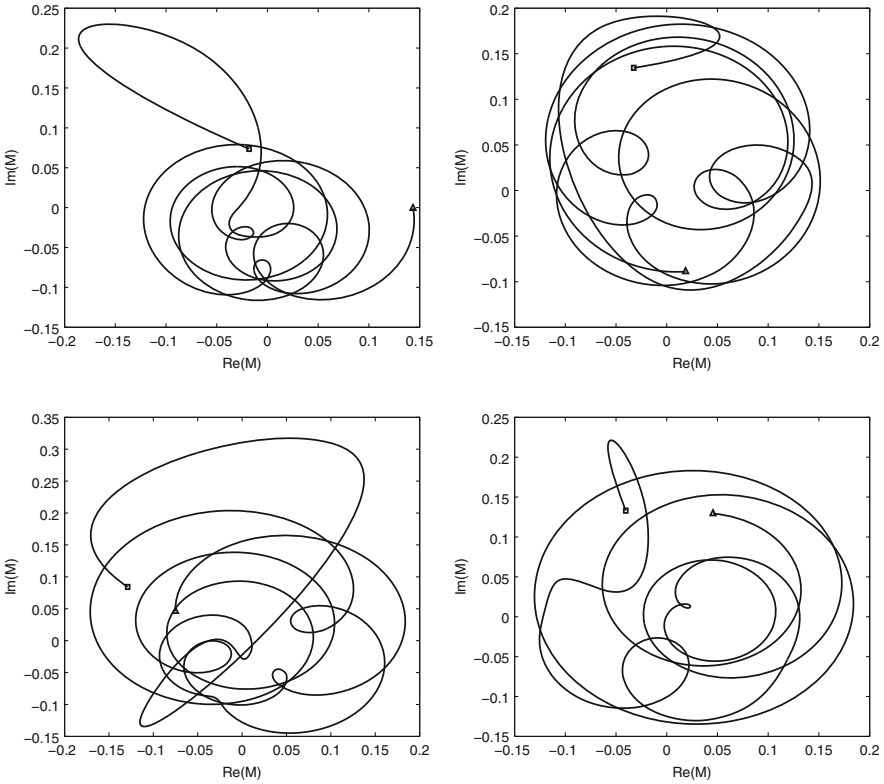


**Fig. 58.2** Numerical solutions with $N = 100$ for $t \in [0, 10]$ and random initial conditions.

## 58.5 Concluding Remarks

We have studied dynamic iterations for thalamo-cortical systems written in terms of Volterra integro-differential equations with the convolution kernel defined by a step function. The iteration process separates each equation from the other equations in the model and is suitable for efficient implementation in parallel environments, significantly reducing the computational time. The amount of processors working independently for each iteration is allowed to vary between 1 and $N$, depending on available resources.

We have derived error bounds for the iteration process and showed its convergence. The error bounds are derived using the kernel of the model and are written in terms of its parameters. The convergence of the dynamic iterations is illustrated by a sequence of numerical experiments with different numbers of equations $N$.

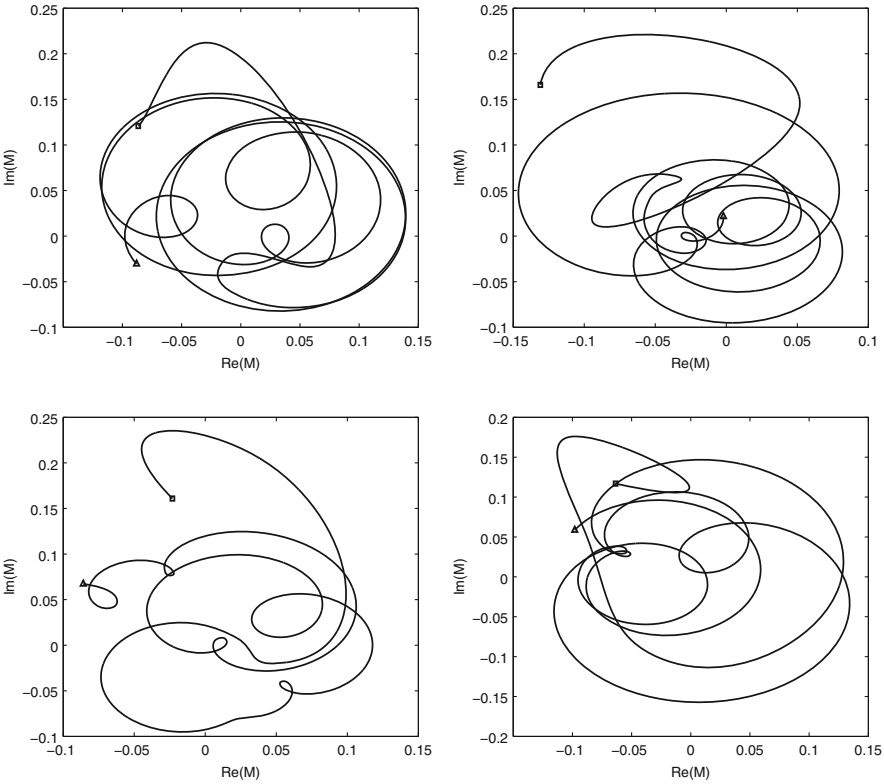


**Fig. 58.3** Numerical solutions with $N = 150$ for $t \in [0, 10]$ and random initial conditions.

# References

[HoIz99] Hoppensteadt, F.C., Izhikevich, E.M.: Oscillatory neurocomputers with dynamic connectivity, Phys. Rev. Lett. **82**, 2983–2986 (1999).

[HoJa07] Hoppensteadt, F.C., Jackiewicz, Z., Zubik-Kowal, B.: Numerical solution of Volterra integro-differential equations with rapidly vanishing convolution kernels, BIT Numerical Mathematics, **47**, 325–350 (2007).

[JaZu06] Jackiewicz, Z., Zubik-Kowal, B.: Spectral collocation and waveform relaxation methods for nonlinear delay partial differential equations, Appl. Numer. Math. **56**, 433–443 (2006).

[JaZu05] Jackiewicz, Z., Zubik-Kowal, B.: Spectral collocation and waveform relaxation methods with Gegenbauer reconstruction for nonlinear conservation laws. Comput. Methods Appl. Math. **5**, 51–71 (2005).

[JaWeZu04] Jackiewicz, Z., Welfert, B.D., Zubik-Kowal, B.: Spectral versus pseudospectral solutions of the wave equation by waveform relaxation methods. J. Sci. Comput. **20**, 1–28 (2004).

[Ku84] Kuramoto, Y.: Chemical Oscillations, Waves, and Turbulence, Springer, New York, (1984).

[MiZu12] Michaels, P., Zubik-Kowal, B.: Parallel computations and numerical simulations for nonlinear systems of Volterra integro-differential equations, Commun. Nonlinear Sci. Numer. Simulat. **17**, 3022–3030 (2012).

[Zu06] Zubik-Kowal, B.: Solutions for the cell cycle in cell lines derived from human tumors, Comput. Math. Methods Med. **7**, 215–228 (2006).

[Zu00] Zubik-Kowal, B.: Chebyshev pseudospectral method and waveform relaxation for differential and differential-functional parabolic equations, Appl. Numer. Math. **34**, 309–328 (2000).

[ZuVa99] Zubik-Kowal, B., Vandewalle, S.: Waveform relaxation for functional-differential equations, SIAM J. Sci. Comput. **21**, 207–226 (1999).

# Index





C. Constanda, A. Kirsch (eds.), *Integral Methods in Science and Engineering*,
DOI 10.1007/978-3-319-16727-5