Christian Constanda
Bardo E. J. Bodmann
Paul J. Harris
Editors

# Integral Methods in Science and Engineering

## Applications in Theoretical and Practical Research

Birkhäuser

Christian Constanda • Bardo E. J. Bodmann •
Paul J. Harris
Editors

# Integral Methods in Science and Engineering

## Applications in Theoretical and Practical Research

Birkhäuser

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

Bardo E. J. Bodmann
Engineering School
Federal University of Rio Grande do Sul
Porto Alegre, Rio Grande do Sul, Brazil

Paul J. Harris
Department of Mathematics
University of Brighton
Brighton, UK

ISBN 978-3-031-07170-6 ISBN 978-3-031-07171-3 (eBook)
https://doi.org/10.1007/978-3-031-07171-3

Mathematics Subject Classification: 45Exx, 45E10, 65R20, 45D05



This book is published under the imprint Birkhäuser, www.birkhauser-science.com by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

The international conferences on Integral Methods in Science and Engineering (IMSE) started in 1985 at the University of Texas–Arlington, and continued biennially in a variety of venues around the world, bringing together specialists who employ integration techniques as essential tools in their research. These procedures exhibit generality, elegance, and efficiency, all of which are essential ingredients in the work of a wide category of practitioners.

The dates and venues of the first 15 IMSE conferences are listed below.

1985, 1990: University of Texas–Arlington, TX, 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)
2014: Karlsruhe Institute of Technology, Germany
2016: University of Padova, Italy
2018: University of Brighton, UK

Due to the unfavorable world health conditions, the 2020 conference, scheduled to be held at the Steklov Mathematical Institute in St. Petersburg, Russia, had to be postponed. However, as an intermediate solution, a Symposium on the Theory and Applications of Integral Methods in Scientific Research was held online in July 2021. By making public their latest results, the participants in this event, all with long-standing IMSE credentials, have kept the flames of our common research interests burning bright, in anticipation of the more inclusive in-person meeting expected to take place, as planned, in St. Petersburg in the summer of 2022.

The peer-reviewed chapters of this volume, arranged alphabetically by first author's name, consist of 22 of the papers presented at the 2021 symposium. The editors would like to thank the reviewers for their help, Christopher Tominich at Birkhäuser–New York for his support of this project, and Saveetha Balasundaram and her production team for their courteous and professional handling of the publication process.

Tulsa, OK, USA — Christian Constanda
Porto Alegre, Brazil — Bardo E. J. Bodmann
Brighton, UK — Paul J. Harris
January 2022

# The International Steering Committee of IMSE

# Contents

# Contributors

**Luiz Felipe Fracasso Chaves Barcellos** Nuclear Studies Group, School of Engineering, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**Ricardo Carvalho de Barros** University of the State of Rio de Janeiro, Polytechnic Institute, Nova Friburgo, RJ, Brazil

**Wagner Queiroz Barros** Laboratory of Petroleum Engineering and Exploration, State University of Northern Rio de Janeiro, Macaé, RJ, Brazil

**Daniel Gustavo Benvenutti** Nuclear Studies Group, School of Engineering, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**Luiz Fernando Bez** Institute of Computing, Federal Fluminense University, Noterói, RJ, Brazil

**Bardo Ernst Josef Bodmann** Postgraduate Program in Mechanical Engineering, School of Engineering, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**Daniela Buske** Department of Mathematics and Statistics, Institute of Physics and Mathematics, Federal University of Pelotas, Capão do Leão, RS, Brazil

**David Chappell** Department of Physics and Mathematics, Nottingham Trent University, Nottingham, UK

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

**Leonardo Rodrigues da Costa Moraes** University of the State of Rio de Janeiro, Polytechnic Institute, Nova Friburgo, RJ, Brazil

**Mulugeta Alemayehu Dagnaw** Department of Mathematics, Debre Tabor University, Debre Tabor, Ethiopia

**Dale R. Doty** Department of Mathematics, The University of Tulsa, Tulsa, OK, USA

**Camila Fávero** Department of Mathematics and Statistics, Institute of Physics and Mathematics, Federal University of Pelotas, Capão do Leão, RS, Brazil

**Júlio César Lombaldo Fernandes** Institute of Mathematics and Statistics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**Débora Lídia Gisch** Institute of Mathematics and Statistics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**Delfina Gómez** Departamento Matemáticas, Estadística y Computación, Facultad de Ciencias, Universidad de Cantabria, Santander, Spain

**Glénio Aguiar Gonçalves** Department of Mathematics and Statistics, Institute of Physics and Mathematics, Federal University of Pelotas, Capão do Leão, RS, Brazil

**Evgeny A. Gorodnitskiy** St. Petersburg State University, St. Petersburg, Russia

**Igor A. Guz** School of Engineering, University of Aberdeen, Aberdeen, UK

**Paul J. Harris** University of Brighton, School of Architecture, Technology and Engineering, Brighton, UK

**Andreas Kleefeld** Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH, Jülich, Germany

**Renato Aloisio dos Santos Klein** Institute of Mathematics and Statistics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**Cibele Aparecida Ladeia** Institute of Mathematics and Statistics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**Rogério José Marczak** Mechanical Engineering Department, School of Engineering, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**André Meneghetti** Institute of Mathematics, Statistics and Physics, Federal University of Rio Grande, Rio Grandev, RS, Brazil

**Oleksandr Menshykov** School of Engineering, University of Aberdeen, Aberdeen, UK

**Marina Menshykova** School of Engineering, University of Aberdeen, Aberdeen, UK

**Sergey E. Mikhailov** Brunel University London, Department of Mathematics, Uxbridge, UK

**Dorina Mitrea** Department of Mathematics, Baylor University, Waco, TX, USA

**Irina Mitrea** Department of Mathematics, Temple University, Philadelphia, PA, USA

**Marius Mitrea** Department of Mathematics, Baylor University, Waco, TX, USA

**Ram S. Mohan** Department of Mechanical Engineering, The University of Tulsa, Tulsa, OK, USA

**Sergey A. Nazarov** Faculty of Mathematics and Mechanics, St. Petersburg State University, St. Petersburg, Russia
Institute of Problems of Mechanical Engineering RAS, St. Petersburg, Russia

**Rafael Orive-Illera** Departamento de Matemáticas, Facultad de Ciencias, Universidad Autónoma de Madrid, Madrid, Spain
Instituto de Ciencias Matemáticas, CSIC-UAM-UC3M-UCM, Madrid, Spain

**Aline R. Parigi** Federal Institute of Education, Science and Technology Farroupilha, São Vicente do Sul, RS, Brazil

**Maria V. Perel** St. Petersburg State University, St. Petersburg, Russia

**Álvaro Marcello Marco Peres** Laboratory of Petroleum Engineering and Exploration, State University of Northern Rio de Janeiro, Macaé, RJ, Brazil

**Maria-Eugenia Pérez-Martínez** Departamento Matemática Aplicada y Ciencias de la Computación, ETSI Caminos, Universidad de Cantabria, Santander, Spain

**Adolfo Puime Pires** Laboratory of Petroleum Engineering and Exploration, State University of Northern Rio de Janeiro, Macaé, RJ, Brazil

**Carlos Fresneda-Portillo** Department of Quantitative Methods, Universidad Loyola Andalucía, Dos Hermanas, Sevilla, Spain

**Régis Sperotto de Quadros** Department of Mathematics and Statistics, Institute of Physics and Mathematics, Federal University of Pelotas, Capão do Leão, RS, Brazil

**Jacob Rowbottom** Department of Physics and Mathematics, Nottingham Trent University, Nottingham, UK

**Jean Schmith** Polytechnic School, University of Vale do Rio dos Sinos, São Leopoldo, RS, Brazil

**Marcelo Schramm** Center of Engineering, Federal University of Pelotas, Pelotas, RS, Brazil

**Adalberto Schuck Jr.** Department of Electrical Engineering, School of Engineering, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**Cynthia Feijó Segatto** Institute of Mathematics and Statistics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**Malay Jignesh Shah** Department of Mechanical Engineering, The University of Tulsa, Tulsa, OK, USA

**Ovadia Shoham** Department of Petroleum Engineering, The University of Tulsa, Tulsa, OK, USA

**Fernando Carvalho da Silva** Department of Nuclear Engineering, Federal University of Rio de Janeiro, Centro de Tecnologia, Rio de Janeiro, RJ, Brazil

**Fernanda Tumelero** Graduate Program in Mechanical Engineering, School of Engineering, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**Marco Tullio Menna Barreto de Vilhena** Graduate Program in Mechanical Engineering, School of Engineering, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**Heloisa Robattini Zanetti** Institute of Mathematics and Statistics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

**Haoqing Zhao** Department of Petroleum Engineering, The University of Tulsa, Tulsa, OK, USA

**Maria Zimmermann** Medical Engineering and Technomathematics, FH Aachen Campus Jülich, Jülich, Germany

# Chapter 1
# Approximate Solution for One-Dimensional Compressible Two-Phase Immiscible Flow in Porous Media for Variable Boundary Conditions


**W. Q. Barros, A. P. Pires, and Á. M. M. Peres**

## 1.1 Introduction

In most petroleum reservoirs, there are at least two phases: oil and connate water. Usually, water is also injected to increase oil production and keep the reservoir pressure at some desired level. Oil displacement by injected water can be modeled by a system of two partial differential equations representing the mass conservation of each component and Darcy's law replacing momentum balance. For one-dimensional incompressible systems without mass transfer, the problem can be solved by the method of characteristics [BL42]. If the relative permeability curves are convex, the solution is given by a continuous two-phase saturation zone (rarefaction wave) followed by a discontinuity (shock). This solution was further expanded to include gravitational and capillary effects [SC59, FS59], to evaluate the pressure drop along porous medium [W52, JBN59], and for three-phase flow [IMPT92, GF97, AS09, CAFM16]. Analytical solutions for compressible two-phase problems are more difficult to develop because both pressure and saturation fields must be solved simultaneously. Approximate solutions were obtained for a two-zone system with constant saturation in each zone [HRM58, KMJ72]. Splitting the two-phase region in more segments improves the accuracy of the solution. The water saturation in each zone of this multi-region system is constant, and thus the velocity of water saturation front in the pressure solution can be neglected and a quasi-static approach can be used [AK89]. The authors of [BH90] proposed a different approximate solution superposing pressure transient effects on a previous saturation profile obtained by Buckley–Leverett solution. The authors of [TR97] generalized the theory for multiphase flow in a heterogeneous reservoir. In this approach, the

W. Q. Barros · A. P. Pires (✉) · Á. M. M. Peres
Universidade Estadual do Norte Fluminense, Macaé, RJ, Brazil
e-mail: adolfo.puime@gmail.com; alvaroperes@lenep.uenf.br



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_1

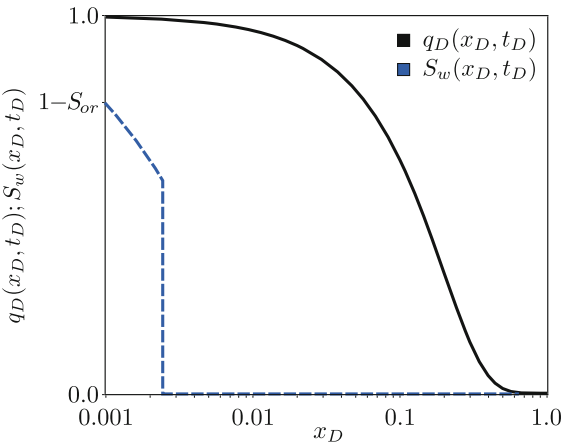



**Fig. 1.1** Typical water saturation and dimensionless flow rate profiles for constant injection rate for a fixed time

pressure and saturation zones move with different velocities, in which the saturation front is always within a steady-state flow-rate zone (Fig. 1.1). It is a simplified method to calculate the pressure profile for the problem of constant fluid injection, in which the saturation is obtained by the immiscible Buckley–Leverett problem and the flow rate by the single-phase compressible solution. The pressure solution is calculated integrating Darcy's equation [BTR98, PR03, PBR04, PBR06].

For constant boundary conditions, the Thompson–Reynolds conjecture provides good results when compared to numerical experiments. However, for non-constant boundary conditions, a new pressure perturbation along the reservoir appears and the conjecture cannot be applied. In this work, we present a new procedure to generalize the solution for non-constant boundary conditions. In Sect. 1.2, we derive the mathematical formulation and present an approximate solution. Next, we compare the solution with numerical results under different injection schedules and system compressibility (1.3). Finally, some conclusions are addressed (1.4).

## 1.2 Mathematical Model

In this work, it is considered a one-dimensional oil displacement by water in a homogeneous porous medium (Fig. 1.2). Additional hypothesis are:

- Immiscible and isothermal linear flow
- Constant cross-sectional area
- Negligible dispersion, gravitational and capillary effects
- Constant viscosity phases
- Constant phases and rock compressibility
- Darcy's law is valid

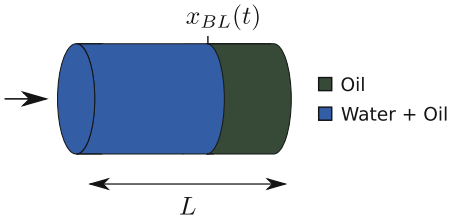



**Fig. 1.2** Representation of 1D water flooding

The velocity of each phase can be calculated using Darcy's law,

$$v_\pi = -\frac{K k_{r\pi}}{\mu_\pi} \frac{\partial P}{\partial x} \ , \tag{1.1}$$

where $K$ and $k_{r\pi}$ are the absolute and phase relative permeabilities, $\mu_\pi$ the phase viscosity, and $\frac{\partial P}{\partial x}$ the linear pressure gradient; the subscript $\pi$ denotes water $w$ or oil $o$ phase. Summing up the velocity for all phases and neglecting capillary effects, one gets

$$q_T\left(x,t\right) = -AK\lambda_T\left(x,t\right)\frac{\partial P\left(x,t\right)}{\partial x} \ , \tag{1.2}$$

where $A$ is the cross-sectional area, $q_T$ represents the total volumetric flow rate, and $\lambda_T$ is the total mobility of the phases ($\lambda_T = \frac{k_{rw}}{\mu_w} + \frac{k_{ro}}{\mu_o}$).

To determine the pressure profile along the porous medium length, we integrate Eq. 1.2 using a constant pressure external boundary condition

$$P\left(x = L, t\right) = P_i \ ,$$

in which $P_i$ denotes the initial pressure and $L$ is the core length, resulting in

$$P\left(x,t\right) - P_i = \frac{1}{AK}\int_x^L \frac{q_T\left(x',t\right)}{\lambda_T\left(x',t\right)} dx' .$$

Now, we introduce dimensionless time and space coordinates,

$$x_D = \frac{x}{L} \ , \tag{1.3}$$

$$t_D = \frac{q_{ref}\ t}{\left(1 - S_{wi} - S_{or}\right) AL\phi} \ , \tag{1.4}$$

where $q_{ref}$ is a reference flow rate, adopted as the first injection value, and $\phi$ is the rock porosity. The irreducible water saturation and residual oil saturation are

denoted by $S_{wi}$ and $S_{or}$, respectively. Thus, the pressure drop can be written in dimensionless variables as

$$P_D\left(x_D, t_D\right) = \int_{x_D}^{1} \frac{q_D\left(x'_D, t_D\right)}{\lambda_{TD}\left(x'_D, t_D\right)} dx'_D \ , \tag{1.5}$$

where

$$P_D\left(x_D, t_D\right) = \frac{KA\hat{\lambda}_o}{q_{ref}L}\left(P\left(x,t\right) - P_i\right) \ , \tag{1.6}$$

$$\lambda_{TD}\left(x_D, t_D\right) = \frac{\lambda_T\left(x,t\right)}{\hat{\lambda}_o} \ , \tag{1.7}$$

$$q_D\left(x_D, t_D\right) = \frac{q_T\left(x,t\right)}{q_{ref}} \ , \tag{1.8}$$

in which $\hat{\lambda}_o$ is oil mobility at water irreducible saturation. Equation 1.5 relates the flow rate and mobility profiles to the pressure drop change at a given position $x_D$. In this work, we solve this problem for the case of step-change internal boundary condition. Thus, an approximation can be obtained based on two key hypotheses:

1. The mobility profile can be obtained by the incompressible problem solution.
2. The total flow rate can be calculated considering two regions with fixed interface position for compressible flow.

The total flow rate is obtained from a linear partial differential equation. Thus, Eq. 1.5 applied for the internal boundary condition $P_D\left(x_D = 0, t_D\right) = P_{wD}\left(t_D\right)$ is written as

$$P_{wD}\left(t_D\right) = \int_{0}^{1} \frac{q_D\left(x'_D, t_D\right)}{\lambda_{TD}\left(x'_D, t_D\right)} dx'_D \ , \tag{1.9}$$

where

$$q_D\left(x_D, t_D\right) = \sum_{j=1}^{Nsteps} \left[q_{D_j}^{Inj} - q_{D_{j-1}}^{Inj}\right] q_{DC}\left(x_D, t_D - t_{D_{j-1}}\right) \ .$$

The last equation is the flow-rate superposition, in which $Nsteps$ is the number of flow-rate steps until $t_D$, $q_{D_j}^{Inj}$ is the injection flow rate in step $j$, and $t_{D_j}$ is the time when $q_{D_j}^{Inj}$ started. The terms inside parenthesis are the $(x_D, t_D)$ coordinates where $q_{DC}$ and $\lambda_{TD}$ are evaluated. The function $q_{DC}$ is the mathematical solution for the two-region problem under constant injection rate.

### 1.2.1 Approximation for $\lambda_{TD}(x_D, t_D)$

The mass conservation for simultaneous flow of oil and water in a linear porous media is modeled by the equations

$$\frac{\partial(\phi S_\pi \rho_\pi)}{\partial t} + \frac{\partial(\rho_\pi v_\pi)}{\partial x} = 0, \ \pi = w, o \ , \tag{1.10}$$

where $\rho_\pi$ is the phase density. Considering an incompressible system, we find

$$\frac{\partial S_\pi}{\partial t} + \frac{1}{\phi}\frac{\partial v_\pi}{\partial x} = 0, \ \pi = w, o \ .$$

Defining the normalized water saturation as

$$S_{nw} = \frac{S_w - S_{wi}}{1 - S_{wi} - S_{or}}, \ S_w \in [S_{wi}, 1 - S_{or}]$$

and applying the definitions of dimensionless variables (Eqs. 1.3, 1.4, and 1.8) together with Darcy's law (Eq. 1.1), we find [BL42]

$$\frac{\partial S_{nw}}{\partial t_D} + q_D(x_D = 0, t_D)\frac{\partial f_w}{\partial x_D} = 0 \ ,$$

in which $f_w$ defines the water fractional flow

$$f_w = \frac{\frac{k_{rw}}{\mu_w}}{\frac{k_{rw}}{\mu_w} + \frac{k_{ro}}{\mu_o}} \ .$$

For convex relative permeability curves, the derivative $\frac{df_w}{dS_{nw}}$ is not monotonic and the solution is not unique. To determine the most admissible solution, we apply the Lax [L57] and Oleinik [O57] stability criteria, and the solution is composed of a rarefaction wave followed by a shock. The shock must be a zero-diffusion limit of the solution given by traveling waves [L07]. The solution is given by

$$S_{nw} = \begin{cases} \frac{df_w}{dS_{nw}}^{-1}\left(\frac{1}{q_D(x_D=0,t_D)}\frac{x_D}{t_D}\right), & x_D \in \left(0, x_D^{BL}\right) \ , \\ 0, & x_D \in \left(x_D^{BL}, 1\right) \ , \end{cases} \tag{1.11}$$

where $\left(\frac{1}{q_D(x_D=0,t_D)}\frac{x_D}{t_D}\right)$ is the self-similar variable where the inverse of $\frac{df_w}{dS_{nw}}$ is evaluated. The shock position is denoted by $x_D^{BL}$ (Fig. 1.2) and is calculated solving the Rankine–Hugoniot ODE condition.

$$\frac{dx_D^{BL}}{dt_D} = q_D(x_D = 0, t_D)\frac{f_w^{BL}}{S_{nw}^{BL}}.$$

### 1.2.2 Approximation for $q_{DC}(x_D, t_D)$

Applying Darcy's law (Eq. 1.1) in the mass conservation (Eq. 1.10), we find

$$S_\pi \left( \phi \frac{\partial \rho_\pi}{\partial P} + \rho_\pi \frac{\partial \phi}{\partial P} \right) \frac{\partial P}{\partial t} + (\phi \rho_\pi) \frac{\partial S_\pi}{\partial t} - \left( \rho_\pi \frac{\partial \left( \frac{K k_{r\pi}}{\mu_\pi} \frac{\partial P}{\partial x} \right)}{\partial x} + \frac{K k_{r\pi}}{\mu_\pi} \frac{\partial \rho_\pi}{\partial P} \left( \frac{\partial P}{\partial x} \right)^2 \right) = 0, \ \pi = w, o \tag{1.12}$$

Using the rock and fluid compressibility definitions ($c_\phi = \frac{1}{\phi} \frac{\partial \phi}{\partial P}$ and $c_\pi = \frac{1}{\rho_\pi} \frac{\partial \rho_\pi}{\partial P}$) and summing for both phases, it is possible to derive

$$\frac{\partial \left( \lambda_T \frac{\partial P}{\partial x} \right)}{\partial x} + (c_w \lambda_w + c_o \lambda_o) \left( \frac{\partial P}{\partial x} \right)^2 = \frac{\phi c_t}{K} \frac{\partial P}{\partial t} ,$$

where $c_t$ is the total compressibility, given by

$$c_t(x,t) = c_\phi + c_o(1 - S_w(x,t)) + c_w S_w(x,t) .$$

For small pressure gradients and slightly compressible fluids, the quadratic term can be neglected. Thus, applying the dimensionless definitions (Eqs. 1.3, 1.4, and 1.5), we find the dimensionless PDE for the pressure in a compressible two-phase system,

$$\frac{1}{\lambda_{TD}} \frac{\partial \left( \lambda_{TD} \left( \frac{\partial P_D}{\partial x_D} \right) \right)}{\partial x_D} = \gamma_L \frac{\partial P_D}{\partial t_D} ,$$

where the term $\gamma_L$ is given by

$$\gamma_L(x_D, t_D) = \frac{q_{ref} L}{(1 - S_{wi} - S_{or}) K A} \frac{c_t}{\lambda_T} .$$

The terms $\gamma_L$ and $\lambda_{TD}$ depend on the saturation profile. To solve this equation, the domain is divided into two regions based on the shock position, and the saturation profile is considered constant in both zones

$$\begin{cases} \frac{\partial^2 P'_D}{\partial x_D^2} = \gamma_L^{IN} \frac{\partial P'_D}{\partial t_D}, & x_D \in \left(0, x_D^{BL}\right) , \\ \frac{\partial^2 P'_D}{\partial x_D^2} = \hat{\gamma}_L \frac{\partial P'_D}{\partial t_D}, & x_D \in \left(x_D^{BL}, 1\right) , \end{cases}$$

where $\gamma_L^{IN}$ is the average gamma in the region behind the shock, and $\hat{\gamma}_L$ is the gamma in the original oil condition. Note that $P'$ indicates the pressure for the two-zone problem. The internal boundary condition (I.B.C.) in dimensionless variables is given by

$$\lim_{x_D \to 0} \left( \frac{\partial P'_D}{\partial x_D} \right) = -\frac{1}{\lambda_{TD}^{IN}} \text{ (I.B.C.)} .$$

The initial condition (I.C.) and external boundary condition (E.B.C.) are

$$P'_D (x_D = 1, t_D) = 0 \text{ (E.B.C.)} ,$$

$$P'_D (x_D, t_D = 0) = 0 \text{ (I.C)} .$$

Thus, the equations that model the pressure in the inner zone are given by

$$\begin{cases} \frac{\partial^2 P'_D}{\partial x_D^2} = \gamma_L^{IN} \frac{\partial P'_D}{\partial t_D}, & x_D \in \left(0, x_D^{BL}\right) , \\ P'_D (x_D, t_D = 0) = 0 & \text{(I.C.)} , \\ \lim_{x_D \to 0} \left( \frac{\partial P'_D}{\partial x_D} \right) = -\frac{1}{\lambda_{TD}^{IN}} & \text{(I.B.C.)} . \end{cases}$$

The equations for the outer region are

$$\begin{cases} \frac{\partial^2 P'_D}{\partial x_D^2} = \hat{\gamma}_L \frac{\partial P'_D}{\partial t_D}, & x_D \in \left(x_D^{BL}, 1\right) , \\ P'_D (x_D, t_D = 0) = 0 & \text{(I.C.)} , \\ P'_D (x_D = 1, t_D) = 0 & \text{(E.B.C.)} . \end{cases}$$

The continuity of pressure and total flow rate at the interface of the two regions are used to close the problem.

$$\begin{gathered} \lim_{x_D \to x_D^{BL-}} P'_D (x_D, t_D) = \lim_{x_D \to x_D^{BL+}} P'_D (x_D, t_D) \\ \left( \lambda_{TD}^{IN} \frac{\partial P_D (x_D, t_D)}{\partial x_D} \right)_{x_D^{BL-}} = \left( \frac{\partial P_D (x_D, t_D)}{\partial x_D} \right)_{x_D^{BL+}} . \end{gathered}$$

The shock position $x_D^{BL} = x_D^{BL} (t_D)$ characterizes a moving internal condition. However, as the speed of this boundary is small, we may use a quasi-stationary assumption, in which the effect of a moving interface is neglected in the solution. However, the interface position is updated every time $t_D$ in order to evaluate the dimensionless pressure $P_D (x_D, t_D)$.

#### 1.2.2.1 Solution by the Laplace Transform

The quasi-stationary hypothesis allows one to solve the two-region problem by the Laplace transform. Applying the transform in the PDE and in both boundary conditions and using the initial condition, the system can be written for the inner zone as

$$\begin{cases} \frac{\partial^2 \bar{P'}_D}{\partial x_D^2} = \gamma_L^{IN} u \bar{P'}_D, & x_D \in \left(0, x_D^{BL}\right) , \\ \lim_{x_D \to 0} \left( \frac{\partial \bar{P'}_D}{\partial x_D} \right) = -\frac{1}{u \lambda_{TD}^{IN}} & \text{(I.B.C.)} \end{cases}$$

and for the outer zone as

$$\begin{cases} \frac{\partial^2 \bar{P'}_D}{\partial x_D^2} = \hat{\gamma}_L u \bar{P'}_D, & x_D \in \left(x_D^{BL}, 1\right) , \\ \bar{P'}_D \left(x_D = x_{Ds}, u\right) = 0 & \text{(E.B.C.)} . \end{cases}$$

The coupling condition in Laplace's domain is given by

$$\begin{gathered} \lim_{x_D \to x_D^{BL-}} \bar{P'}_D \left(x_D, u\right) = \lim_{x_D \to x_D^{BL+}} \bar{P'}_D \left(x_D, u\right) \\ \left( \lambda_{TD}^{IN} \frac{\partial \bar{P'}_D (x_D, u)}{\partial x_D} \right)_{x_D^{BL-}} = \left( \frac{\partial \bar{P'}_D (x_D, u)}{\partial x_D} \right)_{x_D^{BL+}} \end{gathered} .$$

The general solution is

$$\begin{aligned} \bar{P'}_D \left(x_D, u\right) &= A_0 e^{\sqrt{\gamma_L^{IN} u} x_D} + A_1 e^{-\sqrt{\gamma_L^{IN} u} x_D}, \quad \text{for } x_D \in \left(0, x_D^{BL}\right) , \\ \bar{P'}_D \left(x_D, u\right) &= A_2 e^{\sqrt{\hat{\gamma}_L u} x_D} + A_3 e^{-\sqrt{\hat{\gamma}_L u} x_D}, \quad \text{for } x_D \in \left(x_D^{BL}, 1\right) . \end{aligned}$$

Applying the boundary and coupling conditions, it is possible to write the following system of equations:

$$\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & e^{\hat{\alpha}_L x_{Ds}} & e^{-\hat{\alpha}_L x_{Ds}} \\ e^{\alpha_L^{IN} x_D^{BL}} & e^{-\alpha_L^{IN} x_D^{BL}} & -e^{\hat{\alpha}_L x_D^{BL}} & -e^{-\hat{\alpha}_L x_D^{BL}} \\ \lambda_{TD}^{IN} \alpha_L^{IN} e^{\alpha_L^{IN} x_D^{BL}} & -\lambda_{TD}^{IN} \alpha_L^{IN} e^{-\alpha_L^{IN} x_D^{BL}} & -\hat{\alpha}_o e^{\hat{\alpha}_L x_D^{BL}} & \hat{\alpha}_o e^{-\hat{\alpha}_L x_D^{BL}} \end{pmatrix} \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ A_3 \end{pmatrix} = \begin{pmatrix} -\frac{1}{\lambda_{TD}^{IN}} \frac{1}{u \alpha_L^{IN}} \\ 0 \\ 0 \\ 0 \end{pmatrix} ,$$

with $\alpha_L^{IN} = \sqrt{\gamma_L^{IN} u}$ and $\hat{\alpha}_L = \sqrt{\hat{\gamma}_L u}$. The coefficients $A_0$, $A_1$, $A_2$, and $A_3$ are calculated through

$$A_0 = \frac{1}{\lambda_{TD}^{IN} \alpha_L^{IN} u} \left( \frac{2\lambda_{TD}^{IN} \alpha_L^{IN} \left( e^{-\hat{\alpha}_L x_D^{BL}} - e^{\hat{\alpha}_L (x_D^{BL} - 2)} \right) - e^{-\alpha_L^{IN} x_D^{BL}} \Omega_L}{2\cosh\left(\alpha_L^{IN} x_D^{BL}\right) \Omega_L} \right) ,$$

$$A_1 = \frac{1}{\lambda_{TD}^{IN} \alpha_L^{IN} u} \left( \frac{e^{\alpha_L^{IN} x_D^{BL}} \Omega_L + 2\lambda_{TD}^{IN} \alpha_L^{IN} \left( e^{-\hat{\alpha}_L x_D^{BL}} - e^{\hat{\alpha}_L (x_D^{BL} - 2)} \right)}{2\cosh\left(\alpha_L^{IN} x_D^{BL}\right) \Omega_L} \right) ,$$

$$A_2 = -\frac{2e^{-2\hat{\alpha}_L}}{u\Omega_L} ,$$

$$A_3 = \frac{2}{u\Omega_L} ,$$

in which

$$\Omega_L = \left( \hat{\alpha}_L + \lambda_{TD}^{IN} \alpha_L^{IN} \right) \left( e^{\left(\alpha_L^{IN} - \hat{\alpha}_o\right) x_D^{BL}} + e^{\hat{\alpha}_L (x_D^{BL} - 2) - \alpha_L^{IN} x_D^{BL}} \right) ,$$
$$+ \left( \hat{\alpha}_L - \lambda_{TD}^{IN} \alpha_L^{IN} \right) \left( e^{-\left(\alpha_L^{IN} + \hat{\alpha}_L\right) x_D^{BL}} + e^{\hat{\alpha}_L (x_D^{BL} - 2) + \alpha_L^{IN} x_D^{BL}} \right) .$$

The coefficients $A_0$, $A_1$, $A_2$, and $A_3$ are time dependent because the interface position between the regions moves. Finally, we can apply Darcy's law (Eq. 1.2) in the two-zone pressure solution and obtain the approximated flow-rate profile

$$\bar{q}_{DC}(x_D, u) = \begin{cases} -\lambda_{TD}^{IN} \sqrt{\gamma_L^{IN} u} e^{\sqrt{\gamma_L^{IN} u} x_D} A_0 + \lambda_{TD}^{IN} \sqrt{\gamma_L^{IN} u} e^{-\sqrt{\gamma_L^{IN} u} x_D} A_1 \\ \qquad\qquad \text{for } x_D < x_D^{BL} , \\ -\sqrt{\hat{\gamma}_L u} e^{\sqrt{\hat{\gamma}_L u} x_D} A_2 + \sqrt{\hat{\gamma}_L u} e^{-\sqrt{\hat{\gamma}_L u} x_D} A_3 \\ \qquad\qquad \text{for } x_D > x_D^{BL} . \end{cases} \tag{1.13}$$

These equations are inverted to real space using Stehfest's algorithm [GS70]. When the water front position reaches the external core face, Eq. 1.13 for $x_D < x_D^{BL}$ is still valid; however, the terms $\gamma_L^{IN}$ and $\lambda_{TD}^{IN}$ must be averaged inside the core domain ($x_D \in (0, 1)$).

## 1.3 Model Validation

In this section, we apply the developed solution for a set of typical laboratory core flood experiment parameter sets (Table 1.1). The relative permeability curves were generated using the Corey model [C56],

$$\begin{cases} k_{rw} = k_{rw}^{S_{or}^w} (S_{nw})^{n_w} \ , \\ k_{ro} = k_{ro}^{S_{wi}} (S_{no})^{n_o^w} \ , \end{cases}$$

using properties shown in Table 1.2 and Fig. 1.3. The mobility ratio is given by $M = \frac{\hat{\lambda}_w}{\hat{\lambda}_o}$, where $\hat{\lambda}_w$ and $\hat{\lambda}_o$ denote the water mobility at residual oil saturation and the oil mobility at irreducible water saturation. For the data shown in Table 1.2, we have $M = 1.875$.

All solutions discussed in this section are compared to numerical results.

### *1.3.1 Injection Schedule 1*

The first case analyzed is an isochronal schedule composed of three increasing injection flow rates followed by a falloff (Table 1.3). To generate the approximate solution, the first step is solving the incompressible problem (Eq. 1.11) using the fractional flow shown in Fig. 1.3. Comparing the incompressible solution with the

**Table 1.1** Typical rock and fluid properties for core flood experiments

| | | |
|---|---|---|
| Core length | $L = 15$ | [cm] |
| Cross-sectional area | $A = 11.4$ | [cm$^2$] |
| Porosity | $\phi = 0.1$ | [-] |
| Absolute permeability | $K = 200$ | [mD] |
| Initial injection rate | $q_T^0 = 0.54$ | [cm$^3$/min] |
| Water viscosity | $\mu_w = 1.0$ | [cp] |
| Oil viscosity | $\mu_o = 5.0$ | [cp] |
| Rock compressibility | $c_r = 9.8E-6$ | [1/Kgf/cm2] |
| Water compressibility | $c_w = 1.0E-6$ | [1/Kgf/cm2] |
| Oil compressibility | $c_o = 4.0E-5$ | [1/Kgf/cm2] |

**Table 1.2** Relative permeability curves parameters

| |
|---|
| $S_{wi} = 0.20$ |
| $k_{ro}^{S_{wi}} = 0.80$ |
| $S_{or}^w = 0.20$ |
| $k_{rw}^{S_{or}^w} = 0.30$ |
| $n_w = 2.2$ |
| $n_o^w = 2.0$ |

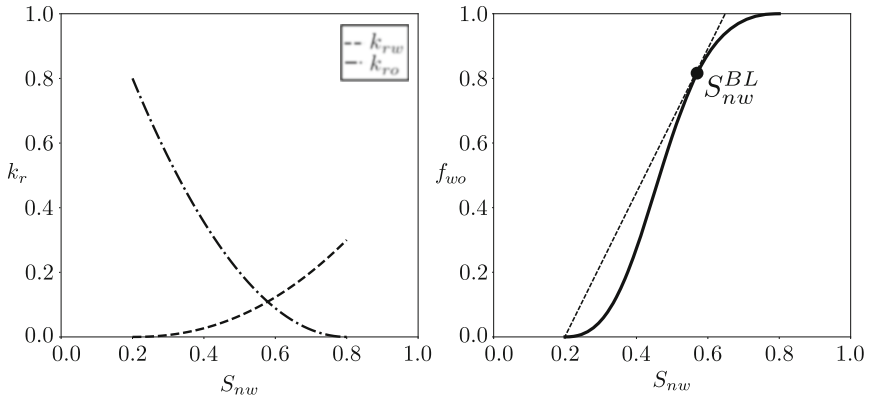



**Fig. 1.3** Relative permeability curves (left) and water fraction flow curve (right) for data shown in Tables 1.1 and 1.2

**Table 1.3** Injection schedule 1

| $t_D$ | $q_D^{INJ}$ |
|---|---|
| 0.00–0.05 | 1.0 |
| 0.05–0.10 | 2.0 |
| 0.10–0.15 | 3.0 |
| 0.15–0.20 | 0.0 |

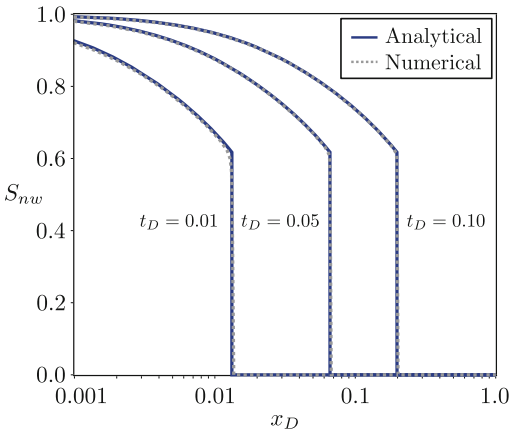



**Fig. 1.4** Analytical and numerical saturation profiles for schedule 1

numerical compressible solution (Fig. 1.4), it can be observed that the saturation profile matches for different injection times.

Once we have the saturation profile, we can solve Eq. 1.13 and obtain an approximate flow rate. In Fig. 1.5, three different $\Delta t_D$ after the first flow-rate change ($t_D = 0.05$) are compared. Note that the greatest difference between solutions

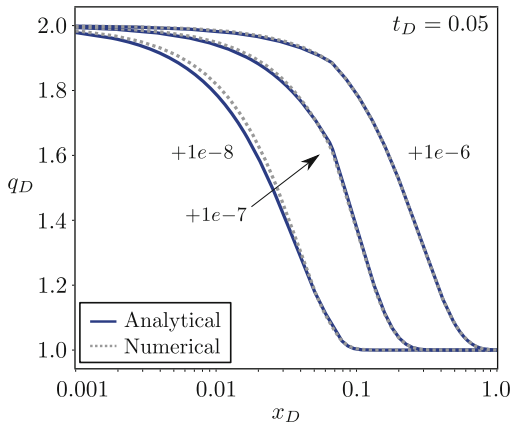



**Fig. 1.5** Analytical and numerical flow-rate profile for different $\Delta t_D$ after $t_D = 0.05$

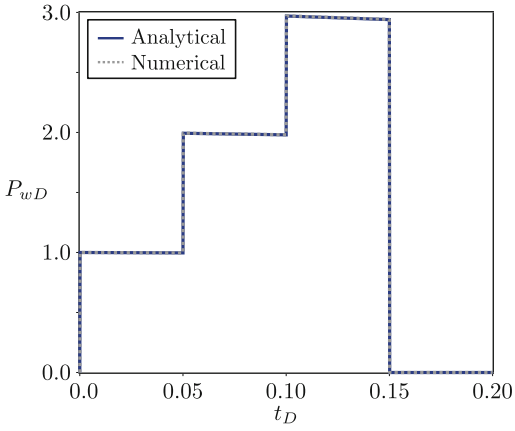



**Fig. 1.6** Analytical and numerical $P_{wD}$ solution for schedule 1

appears at small times ($\Delta t_D = 1e^{-8}$). After $\Delta t_D = 1e^{-7}$, the solutions present close agreement. Using the calculated $\lambda_{TD}$ and $q_D$, it is possible to integrate Equation 1.9 and obtain the final solution (Fig. 1.6).

### *1.3.2 Injection Schedule 2*

The second case changes the injection flow rate schedule (Table 1.4) using the same reservoir properties (Tables 1.1 and 1.2). Schedule 2 is composed of three isochronal decreasing flow rates, followed by a falloff. Figures 1.7 and 1.8 present the saturation and flow-rate profiles compared with the compressible numerical solutions. The presented profiles are calculated at three different $\Delta t_D$ after the falloff

**Table 1.4** Injection schedule 2

| $t_D$ | $q_{Di}$ |
|---|---|
| 0.00–0.05 | 1.0 |
| 0.05–0.10 | 2/3 |
| 0.10–0.15 | 1/3 |
| 0.15–0.20 | 0.0 |

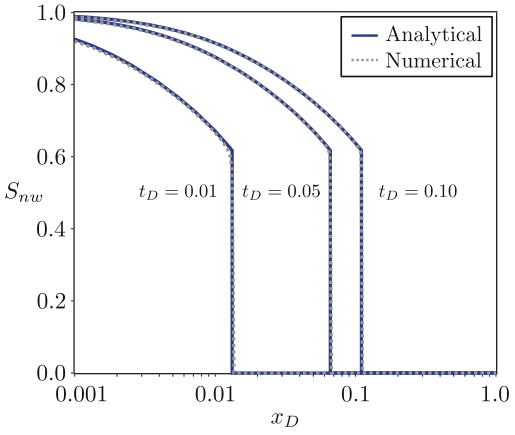



**Fig. 1.7** Analytical and numerical saturation profiles for different $t_D$ for schedule 2

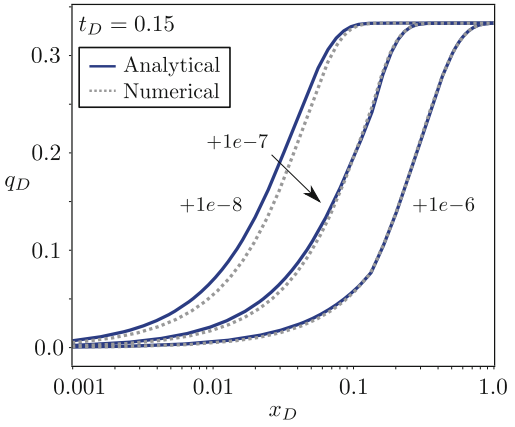



**Fig. 1.8** Analytical and numerical flow-rate profile for different $\Delta t_D$ after $t_D = 0.15$

($t_D = 0.15$). It can be noted that both solutions agree and can be used to build the pressure solution of the original problem (Fig. 1.9). Note that our approximation of $P_{wD}$ agrees with numerical simulation for all flow-rate steps.

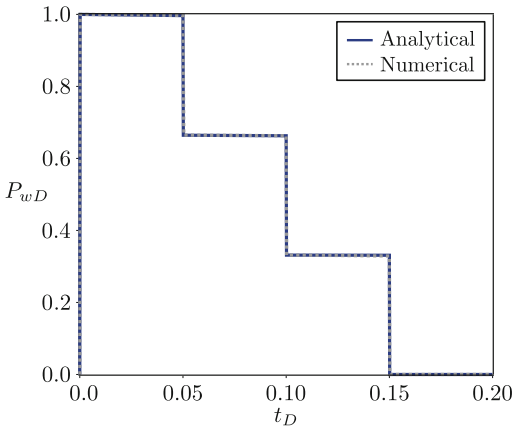



**Fig. 1.9** Analytical and numerical $P_{wD}$ solution for schedule 2

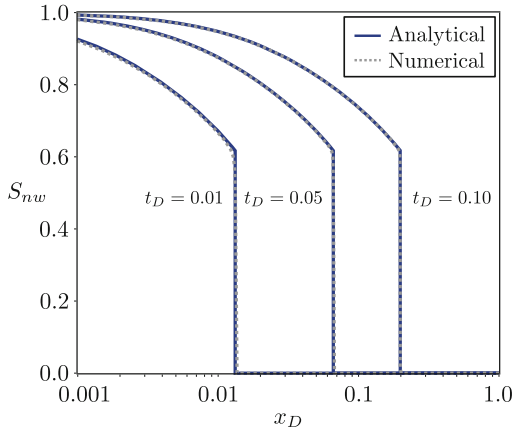



**Fig. 1.10** Analytical and numerical saturation profile for different $t_D$ for a more compressible system

### *1.3.3 Compressibility Effect*

Schedule 1 (Table 1.3) was used to analyze the compressibility effects in the results ($c_r = 1.0E-2$ 1/MPa and $c_o = 4.0E-3$ 1/Kgf/cm2 keeping all other properties constant. Even increasing the compressibility by a factor of 100, the incompressible and compressible saturation profiles still match (Fig. 1.10). As this system is much more compressible, it is expected that the flow rate propagates slower in the reservoir (Fig. 1.11). It can be noted that both solutions agree after $\Delta t_D = 1e^{-6}$. The pressure solution is presented in Fig. 1.12 showing the excellent agreement with numerical compressible simulation.

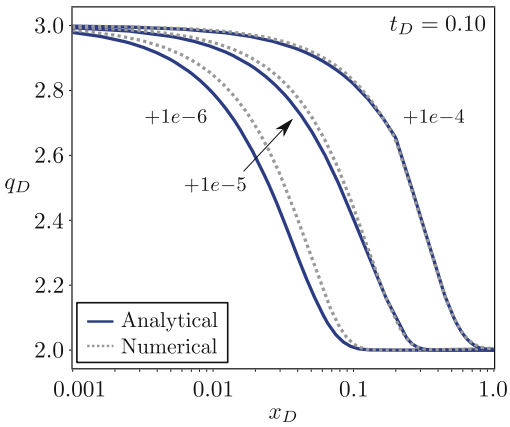



**Fig. 1.11** Analytical and numerical flow-rate profile for different $\Delta t_D$ after $t_D = 0.10$

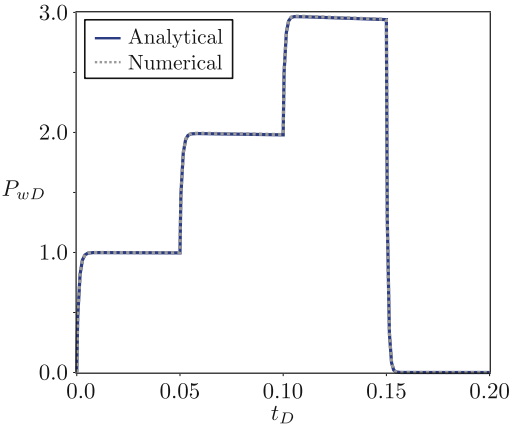



**Fig. 1.12** Analytical and numerical $P_{wD}$ solution for a more compressible system

## 1.4 Conclusion

This work presents a new solution for the pressure drop along a linear porous medium considering immiscible two-phase oil displacement and a step-rate variable boundary condition. The solution is calculated based on two main hypothesis:

1. The mobility profile can be determined by the incompressible problem solution.
2. The total flow rate can be calculated by a dual-zone compressible problem.

The model was tested for two different flow rate schedules, and the results were compared to numerical solutions with excellent agreement. The analytical solution built in this work can be used to model laboratory core flood experiments.

**Acknowledgments** The authors acknowledge Universidade Estadual do Norte Fluminense (UENF) for financial support. This study was also financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.

## References


[AK89] Abbaszadeh, M., & Kamal, M.: Pressure-transient testing of water-injection wells. SPE Reserv. Eng. **4**(01), 115–124 (1989). Society of Petroleum Engineers (SPE). https://doi.org/10.2118/16744-pa

[AS09] Azevedo, A.V., Souza, A.J., Furtado, F., Marchesin, D., Plohr, B.: The solution by the wave curve method of three-phase flow in virgin reservoirs. Trans. Porous Media **83**(01), 99–125 (2009). Springer Science and Business Media LLC. https://doi.org/10.1007/s11242-009-9508-9

[BH90] Bratvold, R.B., Horne, R.N.: Analysis of pressure-falloff tests following cold-water injection. SPE Form. Evaluation **5**(03), 293–302 (1990). Society of Petroleum Engineers (SPE). https://doi.org/10.2118/18111-pa

[BL42] Buckley, S., Leverett, M.: Mechanism of fluid displacement in sands. Trans. AIME **146**(01), 107–116 (1942). Society of Petroleum Engineers (SPE). https://doi.org/10.2118/942107-g

[BTR98] Banerjee, R., Thompson, L.G., Reynolds, A.C.: Injection/falloff testing in heterogeneous reservoirs. SPE Reser. Evaluation Eng. **1**(06), 519–527 (1998). Society of Petroleum Engineers (SPE). https://doi.org/10.2118/52670-pa

[CAFM16] Castañeda P., Abreu, E., Furtado, F., Marchesin, D.: On a universal structure for immiscible three-phase flow in virgin reservoirs, Comput. Geosci. **20**(01), 171–185 (2016). Springer Science and Business Media LLC. https://doi.org/10.1007/s10596-016-9556-5

[C56] Corey, A.T., Rathjens, C.H., Henderson, J.H., Wyllie, M.R.J.: Three-phase relative permeability. J. Petroleum Technol. **8**(11), 63–65 (1956). Society of Petroleum Engineers (SPE). https://doi.org/10.2118/737-g

[FS59] Fayers, F.J., Sheldon, J.W.: The effect of capillary pressure and gravity on two-phase fluid flow in a porous medium. Trans. AIME **216**(01), 147–155 (1959). Society of Petroleum Engineers (SPE). https://doi.org/10.2118/1089-g

[GF97] Guzmán, R.E., Fayers, F.J.: Mathematical properties of three-phase flow equations. SPE J. **2**(03), 291–300 (1997). Society of Petroleum Engineers (SPE). https://doi.org/10.2118/35154-pa

[GS70] Stehfest, H.: Algorithm 368: numerical inversion of laplace transforms. Commun. ACM **13**, 47–49 (1970)

[HRM58] Hazebroek, P., Rainbow, H., Matthews, C.S.: Pressure fall-off in water injection wells. Trans. AIME **213**(01), 250–260 (1958). Society of Petroleum Engineers (SPE). https://doi.org/10.2118/925-g

[IMPT92] Isaacson, E.L., Marchesin, D., Plohr, B.J., Temple, J.B.: Multiphase flow models with singular Riemann problems. Mat. Apl. Comput. **11**(02), 147–166 (1992). Sociedade Brasileira de Matemática Aplicada e Computacional

[JBN59] Johnson, E.F., Bossler, D.P., Neumann, V.O.: Calculation of relative permeability from displacement experiments. Trans. AIME **216**, 370–372 (1959). Society of Petroleum Engineers (SPE), SPE-1023G

[KMJ72] Kazemi, H., Merrill, L.S., Jargon, J.R.: Problems in interpretation of pressure fall-off tests in reservoirs with and without fluid banks. J. Petrol. Technol. **24**(09), 1147–1156 (1972). Society of Petroleum Engineers (SPE). https://doi.org/10.2118/3696-pa

[L07] Liu, T.: Hyperbolic and Viscous Conservation Laws. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 01. SIAM, Philadelphia (2007)

[L57] Lax, P.D.: Hyperbolic systems of conservation laws II. Commun. Pure Appl. Math. **10**(04), 537–566 (1957). Wiley. https://doi.org/10.1002/cpa.3160100406

[O57] Oleinik, O.A.: On the uniqueness of the generalized solution of the Cauchy problem for a non-linear system of equations occurring in mechanics. Uspekhi Mat. Nauk **12**, 169–176 (1957)

[PBR04] Peres, A.M.M., Boughrara, A.A., Reynolds, A.C.: Rate superposition for generating pressure falloff solutions for vertical and horizontal wells. SPE Annual Technical Conference and Exhibition, Society of Petroleum Engineers (SPE) (2004). https://doi.org/10.2118/90907-ms

[PBR06] Peres, A.M.M., Boughrara, A.A., Reynolds, A.C.: Rate superposition for generating pressure falloff solutions. SPE J. **11**(03), 364–374 (2006). Society of Petroleum Engineers (SPE). https://doi.org/10.2118/90907-pa

[PR03] Peres, A.M.M., Reynolds, A.C.: Theory and analysis of injectivity tests on horizontal wells. SPE J. **8**(02), 147–159 (2003). Society of Petroleum Engineers (SPE). https://doi.org/10.2118/84957-pa

[SC59] Sheldon, J.W., Cardwell, W.T.: One-dimensional, incompressible, noncapillary, two-phase fluid flow in a porous medium. Trans. AIME **216**(01), 290–296 (1959). Society of Petroleum Engineers (SPE). https://doi.org/10.2118/978-g

[TR97] Thompson, L., Reynolds, A.C.: Well testing for radially heterogeneous reservoirs under single and multiphase flow conditions. SPE Format. Evaluation **12**(01), 57–64 (1997). Society of Petroleum Engineers (SPE). https://doi.org/10.2118/30577-pa

[W52] Welge, H.J.: A simplified method for computing oil recovery by gas or water drive. J. Petrol. Technol. **4**(04), 91–98 (1952). Society of Petroleum Engineers (SPE). https://doi.org/10.2118/124-g

# Chapter 2
# On Pseudo-Cross Sections for Neutron Escape from a Domain by a Physical Monte Carlo Simulation


**D. G. Benvenutti, L. F. F. C. Barcellos, and B. E. J. Bodmann**

## 2.1 Introduction

Neutron transport in finite multiplicative media is relevant in a variety of applications, as for instance in medicine, industrial applications, energy production and many others. In these situations, nuclear interactions occur in the domain and neutrons escape from the domain constituting the physics of the system that can be dealt with by deterministic methods, such as the $P_n$ and $S_n$ approximations, or stochastic methods, such as the physical Monte Carlo method [La06]. It is noteworthy that these nuclear interactions are generally classified as absorption or scattering interactions and are quantified using the concept of nuclear cross sections, which are directly related to the probability for an interaction to occur such that each nuclear species in the medium has an associated cross section for each type of interaction [La66].

Concerning the deterministic treatment of this kind of problem, the necessary boundary conditions depend on the neutron flux in the domain, which in turn depend on the knowledge of the flux of escape neutrons across the boundary. Usually, the neutron flux is assumed to be zero at the boundaries or in regions close to these boundaries known as extrapolation distances [FeEtAl17, OlEtAl17, OlEtAl19, TuEtAl19]. Although this assumption is a reasonable approximation, it does not match exactly with reality. Differently, the application of the Monte Carlo method in the study of neutron transport in finite domains, such as in nuclear reactor cores, does not require the prior knowledge of the neutron flux on surfaces and allows to analyse and account for neutron escape in a simulation according to the microscopic physics of the problem in consideration [BaEtAl21].

D. G. Benvenutti (✉) · L. F. F. C. Barcellos · B. E. J. Bodmann
Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: daniel.benvenutti@ufrgs.com; bardo.bodmann@ufrgs.br



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_2

Thus, the present work reports on a novel approach in the study of neutron transport including escape from finite domains using a Monte Carlo simulation. More specifically, escape is treated in an analogous way to the absorption reaction but outside the domain of interest. Consequently, it has the same characteristics as a real interaction and can be quantified by an associated pseudo-cross section. To this end, neutron escape or leakage is interpreted as a pseudo-interaction and one may use the conception of reaction rates determined by a physical Monte Carlo simulation. This type of method makes use of neutron tracking while providing the position and energy of each neutron at the interaction vertices when located in the finite domain. Tallying allows then to compute pseudo-cross sections for neutron escape which may be cast in analytical functions, so that these can be applied in subsequent either deterministic or stochastic approaches.

## 2.2 The Physical Monte Carlo Simulation

The physical Monte Carlo simulation is a stochastic method used in the study of neutron transport, which allows the treatment of problems with complex geometries without the need for simplifications frequently applied in deterministic methods. On the contrary to analytical or numerical approaches, the physical Monte Carlo method is characterized by simulating the processes which constitute the Boltzmann transport equation, mimicking the real microscopic physical process at each instance in the history of each individual neutron. Associated reaction rates are based on probabilistic descriptions and provide the amplitudes for each specific event that may occur. The stochastic method like the transport equation is formulated in a seven-dimensional phase space (energy, oriented solid angle, position and time) as long as no simplifying assumptions are imposed. Thus, every individual neutron can be tallied using the track and interaction data recorded for posterior statistical analysis. From the simulation of a sufficiently large number of neutron histories and starting from an initial neutron population, it is possible to obtain physical quantities of interest, such as the neutron density, the scalar and angular flux and reaction rates, among others. One of the features of the physical Monte Carlo method is that it considers random processes that are not handled by directly solving the transport equation through deterministic methods [CaCa75].

One of the main advantages of the Monte Carlo method is that uncertainties due to the stochastic nature of the procedure can be reduced by increasing the number of simulated histories, in distinction to deterministic methods, where errors involved are systematic and cannot be reduced, since they come from necessary simplifying hypotheses, such as discretization of the independent variables or treating a three-dimensional problem in a dimensionally reduced fashion. A property of stochastic methods is that they provide results through tallies, e.g., the number of a given interaction over a period of time, in intervals $\Delta V$ $\Delta E$ and $\Delta\Omega$ that constitute the physical phase space and are previously defined. In practice, the size of these intervals (the resolution) depends on the number of simulated neutron histories,

and the larger the sampling of these histories, the more one may reduce the size of these intervals to obtain more detailed distributions without statistical fluctuations becoming dominant [LeMi84].

Developed in C++, the Monte Carlo simulator employed and adapted in this work was created as described in reference [CaEtAl11] and was successively restructured and optimized as reported in [CaEtAl13, BarEtAl17, BaEtAl21]. One of the differences to other neutron transport Monte Carlo codes is that in the present implementation the nuclear cross section data are parameterized by continuous functions implemented in the programme, while in most existing simulators the cross sections are determined by interpolating data from huge databases. In the simulator used in this work, the entire history of each neutron is assembled in the Monte Carlo steps, where they are generated by random events in the spirit of Markov chains. Resuming, at each Monte Carlo step, the neutron propagates from an initial position to an end position, and if the final position is outside the finite domain of interest, the neutron is said to have escaped and its history ends, while otherwise the neutron will induce an interaction at the final position located in the domain and a new Monte Carlo step describing a new neutron trajectory is initiated.

The history of two neutrons born from fission represented by $(n, f)$ as simulated with the Monte Carlo simulator described in detail in references [BarEtAl17, BaEtAl21] is exemplified in Fig. 2.1. One of the neutrons is absorbed in the finite medium by an absorption reaction represented by $(n, \gamma)$, while the second one undergoes successive scattering interactions, represented by $(n, e)$, until it escapes the finite medium. Here, $(n, z)$ represents the escape interaction and will be defined

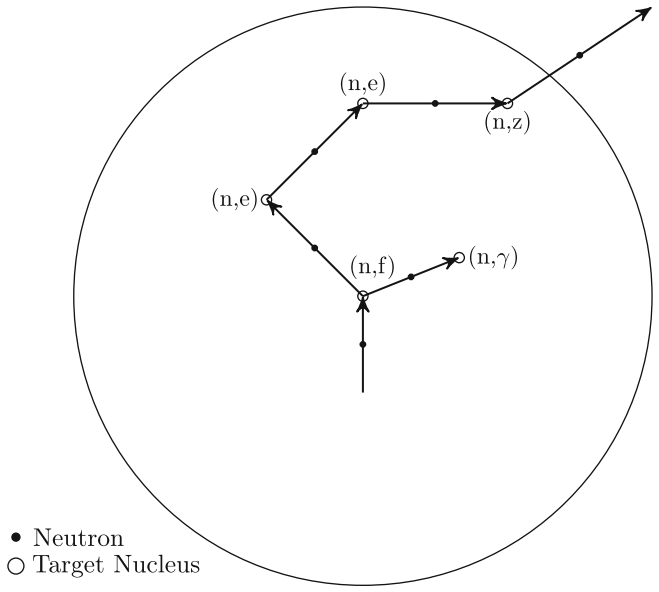



**Fig. 2.1** Possible histories of two neutrons in a finite domain

later. Note that each displacement the neutron performs characterizes one Monte Carlo step in the simulation process.

## 2.3 The Pseudo-Cross Section for Neutron Escape

Based on the simulation data of a neutron transport problem in a finite domain using the physical Monte Carlo method, we developed the methodology to obtain the pseudo-cross section for neutron escape. To this end, it is possible to start from the concept of reaction rates, which are defined as the number of interactions per unit volume and per time unit at position $\mathbf{r}$ and at time $t$. More formally, it is possible to define the reaction rate from a density function known as neutron angular density ($n(\mathbf{r}, \boldsymbol{\Omega}, E, t)$) that gives the number of neutrons per unit volume, per energy unit and per solid angle unit, at position $\mathbf{r}$, with velocity in the direction $\boldsymbol{\Omega}$ and with energy $E$ at time $t$ [BeGl70, La66]. Furthermore, this density function is defined in such a way that $n(\mathbf{r}, \boldsymbol{\Omega}, E, t)d\Omega dE$ represents the number of neutrons per unit volume at position $\mathbf{r}$ and time $t$, considering neutrons with energy in the infinitesimal range $dE$ around the energy $E$ and whose direction of motion points into the differential solid angle $d\Omega$ around $\boldsymbol{\Omega}$. Thus, the reaction rate of an interaction of type $i$ at position $\mathbf{r}$ and at a time $t$ can be written as

$$R(\mathbf{r}, t)_i = \int_0^\infty \Sigma_i(E)v(E) \int_{4\pi} n(\mathbf{r}, \boldsymbol{\Omega}, E, t)d\Omega dE \ .$$

Here, $\Sigma_i(E)$ represents the macroscopic cross section of interaction type $i$, $v(E)$ is the neutron speed and the integral $\int_{4\pi} n(\mathbf{r}, \boldsymbol{\Omega}, E, t)d\Omega$ corresponds to the total number of neutrons per unit volume and per energy unit at position $\mathbf{r}$ and at time $t$ for neutrons with energy $E$ and integrating over all directions.

An alternative way of writing the reaction rate, which will be used in this approach, is to consider energy-dependent density functions. Furthermore, in the present approach, a stationary case is assumed, i.e., $n(\mathbf{r}, \boldsymbol{\Omega}, E, t)$ does not vary with time during the time interval that is being simulated in the finite domain. Therefore, for a steady state case, the reaction rate of interaction type $i$, per energy unit, at position $\mathbf{r}$, at time $t$ and considering neutrons with energy $E$ is given by

$$R(\mathbf{r}, t, E)_i = \Sigma_i(E)v(E) \int_{4\pi} n(\mathbf{r}, \boldsymbol{\Omega}, E)d\Omega \ .$$

Now, two additional definitions that characterize types of interactions need to be made. First, an escape pseudo-interaction indicated by the subscript $z$ is understood when in the simulation a neutron escapes the finite domain and this pseudo-interaction is associated with the starting position with time stamp $t$ and the initial neutron energy in the respective Monte Carlo step. By construction, the escape

pseudo-interaction will have an associated reaction rate per energy unit written as

$$R(\mathbf{r}, t, E)_z = \Sigma_z(\mathbf{r}, E)v(E)\int_{4\pi} n(\mathbf{r}, \boldsymbol{\Omega}, E)d\Omega \ ,$$

where $\Sigma_z(\mathbf{r}, E)$ is the macroscopic pseudo-cross section for neutron escape.

Unlike the cross sections that characterize real interactions, which vary only with the neutron energy, the escape pseudo-cross section varies with both the kinetic energy of the neutron and its position in space, since neutrons closer to the boundaries of the finite domain are expected to have greater chances of escaping. Furthermore, the pseudo-cross section does not vary with time, since only steady state cases are being considered and it is assumed that this implies the reaction rate does not vary with time either. Note that these assumptions were validated in the simulations.

Usually, the total interactions in the physical system, represented by the subscript $t$, include any kind of real interaction, that is, scattering and absorption interactions. With the introduction of an escape pseudo-interaction, the set of all possible interactions in the domain shall include also escape. In this way and to avoid ambiguities, integral interactions represented by the letter $y$ are defined as the sum of any type of interaction that occurs in the domain, whether real interactions, such as scattering, or pseudo interactions, such as escape. Thus, the reaction rate of integral interactions per energy unit at time $t$, at position $\mathbf{r}$ and energy $E$ is

$$R(\mathbf{r}, t, E)_y = \Sigma_y(\mathbf{r}, E)v(E)\int_{4\pi} n(\mathbf{r}, \boldsymbol{\Omega}, E)d\Omega \ .$$

Here, $\Sigma_y(\mathbf{r}, E)$ is the integral macroscopic cross section and is, by definition, the sum of the total macroscopic cross section of the medium with the macroscopic pseudo-cross section for neutron escape previously introduced, i.e., $\Sigma_y(\mathbf{r}, E) = \Sigma_z(\mathbf{r}, E) + \Sigma_t(E)$.

Based on the reaction rate, it is possible to obtain the number of accumulated interactions up to time $t_0$. Thus, considering neutrons with energy $E$ and in the time interval $\Delta t = (t_0 - 0)$, the number of escape interactions per energy unit and per unit volume at $\mathbf{r}$ is

$$N(\mathbf{r}, E)_z = \int_0^{t_0} R(\mathbf{r}, E, t)_z dt = \Sigma_z(\mathbf{r}, E)\int_0^{t_0}\left(v(E)\int_{4\pi} n(\mathbf{r}, \boldsymbol{\Omega}, E)d\Omega\right)dt \ . \tag{2.1}$$

Similarly, the number of integral interactions per unit volume and per energy unit at $\mathbf{r}$, with $E$ and in the time interval $\Delta t$ is then

$$N(\mathbf{r}, E)_y = \int_0^{t_0} R(\mathbf{r}, E, t)_y dt = \Sigma_y(\mathbf{r}, E)\int_0^{t_0}\left(v(E)\int_{4\pi} n(\mathbf{r}, \boldsymbol{\Omega}, E)d\Omega\right)dt \ . \tag{2.2}$$

Note that using a steady state approach allows to remove $\Sigma_y(\mathbf{r}, E)$ and $\Sigma_z(\mathbf{r}, E)$ from the time integral in the above equation, so that the spectral scalar neutron flux, i.e., the integral $\int_0^{t_0} \left(v(E) \int_{4\pi} n(\mathbf{r}, \boldsymbol{\Omega}, E) d\Omega\right) dt$, appears in both Eqs. (2.1) and (2.2), respectively. Therefore, by manipulating these equations, it is possible to obtain the following relation:

$$\frac{N(\mathbf{r}, E)_z}{N(\mathbf{r}, E)_y} = \frac{\Sigma_z(\mathbf{r}, E)}{\Sigma_y(\mathbf{r}, E)} = \frac{\Sigma_z(\mathbf{r}, E)}{\Sigma_z(\mathbf{r}, E) + \Sigma_t(E)} .$$

Now, isolating $\Sigma_z(\mathbf{r}, E)$ from the above equation, one obtains the macroscopic pseudo-cross section for neutron escape.

$$\Sigma_z(\mathbf{r}, E) = \frac{N(\mathbf{r}, E)_z}{N(\mathbf{r}, E)_y} \Sigma_t(E) \left(1 - \frac{N(\mathbf{r}, E)_z}{N(\mathbf{r}, E)_y}\right)^{-1} . \tag{2.3}$$

Equation (2.3) allows to determine the pseudo-cross section for neutron escape for a given physical application. Evidently, variations in geometric properties or chemical composition in this application implies in different $\Sigma_z(\mathbf{r}, E)$.

Whenever necessary and in order to facilitate the post-processing of statistics, one last assumption can be made, which is the density functions $N(\mathbf{r}, E)_z$ and $N(\mathbf{r}, E)_y$ are separable and are described as the product of two simpler density functions. The first one represents the number of interactions per unit volume, which varies only with position $\mathbf{r}$, and the second one counts the number of interactions per energy unit, which varies only with the neutron energy $E$. Thus, Eq. (2.3) becomes

$$\Sigma_z(\mathbf{r}, E) = \frac{n_{z_r}(\mathbf{r}) n_{z_E}(E)}{n_{y_r}(\mathbf{r}) n_{y_E}(E)} \Sigma_t(E) \left(1 - \frac{n_{z_r}(\mathbf{r}) n_{z_E}(E)}{n_{y_r}(\mathbf{r}) n_{y_E}(E)}\right)^{-1} . \tag{2.4}$$

Here, $n_{z_r}(\mathbf{r})$ and $n_{y_r}(\mathbf{r})$ represent the number of escape and integral interactions per unit volume, while $n_{z_E}(E)$ and $n_{y_E}(E)$ represent the number of escape and integral interactions per energy unit, respectively. It is noteworthy that this simplification is not necessarily true especially near the surfaces of the finite domain, nevertheless it still allows to obtain satisfactory global results by the approximate approach.

Considering the case where the number of interactions per unit volume and the number of interactions per energy unit are not separable, in order to obtain the terms for the right-hand side in Eq. (2.3), as stated in Sect. 2.2, the intervals $\Delta V$ and $\Delta E$ that constitute the phase space were previously defined, therefore, both $N(\mathbf{r}, E)_y$ and $N(\mathbf{r}, E)_z$ are easily found from the Monte Carlo simulation of the neutron transport in the finite domain of interest. For example, let $C_z$ be the number of escape interactions in $\Delta V$ at $\mathbf{r}$ and in $\Delta E$ around $E$ obtained from the simulation considering a time interval $\Delta t$. From the definition of $N(\mathbf{r}, E)_z$, $C_z$ is given by

$$C_z = \int_{E_i}^{E_f} \int_D N(\mathbf{r}, E)_z dV dE .$$

Here, $E_i$ and $E_f$ define the total energy interval of interest and $D$ is the spatial domain limit, where the $\Delta E$ and $\Delta V$ intervals specify the resolution, respectively.

A way to approximate the number of escape interactions in the incremental intervals and also more convenient is to consider that $\Delta E$ and $\Delta V$ small enough so that $C_z \approx N(\mathbf{r}, E)_z \Delta V \Delta E$. From the discretization of the phase space, the $\Delta E$ and $\Delta V$ are known, so that the number of escape interactions per unit volume and per energy unit at $\mathbf{r}$ and considering neutrons with energy $E$ during $\Delta t$ can be written as $N(\mathbf{r}, E)_z \approx \frac{C_z}{\Delta V \Delta E}$. Since $\mathbf{r}$ and $E$ are arbitrary and all $\Delta V$ and $\Delta E$ that cover the phase space are known, it is possible to construct $N(\mathbf{r}, E)_z$ for the entire domain of interest, in this case, the entire energy spectrum and all positions in space. Note that by construction, $N(\mathbf{r}, E)_z$ is discrete, taking unique values for each interval $\Delta V$ and $\Delta E$ around $\mathbf{r}$ and $E$. The same idea can be applied to the integral interactions, so that $N(\mathbf{r}, E)_y \approx \frac{C_z}{\Delta V \Delta E}$.

Following a similar reasoning, one obtains $n_{z_r}(\mathbf{r})$, $n_{y_r}(\mathbf{r})$, $n_{z_r}(E)$ and $n_{y_r}(E)$, so that the remaining term needed to find the pseudo-cross section for neutron escape through Eqs. (2.3) or (2.4) is the total macroscopic cross section $\Sigma_t(E)$. This is obtained from the microscopic cross sections of the nuclides that make up the multiplicative medium and are provided in the nuclear data libraries.

As stated before, the idea is to obtain the pseudo-cross section in form of an analytical function. To this end, all the terms on the right-hand side of Eqs. (2.3) or (2.4) must be represented by analytical functions. Though these terms are obtained from the Monte Carlo data as discrete functions, therefore, a parametrization of these is proposed and determined through an optimization problem which consists of minimizing the weighted sum of squared residuals. Note that the model function must be nonlinear to parameterize the resonances present in $\Sigma_t(E)$ and accordingly in $N(\mathbf{r}, E)_z$ and $N(\mathbf{r}, E)_y$. By virtue of the model function being nonlinear, an iterative algorithm was used to solve the minimization problem, in the present treatise the Levenberg–Marquardt algorithm.

Basically, the model function is made up of sums and products of polynomial functions, window functions and rational functions. The rational functions used here served mainly to deal with the resonance regions of the cross sections and are described by

$$f(x) = a_0 \left(1 + \left(\frac{x - a_1}{a_2}\right)^2\right)^{-1},$$

where $a_0$, $a_1$ and $a_2$ are obtained from the fitting process. The choice of this type of function is due to the fact that individual resonances are physically described by the Breit–Wigner distribution, which has the form of a rational function. The other type of function used in the parametrization is a window function, which is not directly used in the curve fitting process, but serves to divide the dense resonance regions present in the cross sections into sub-regions in which the curve fitting process can be performed without resulting in divergences. In addition, this type of function is used to separate the regions of the cross section that do not have resonances from

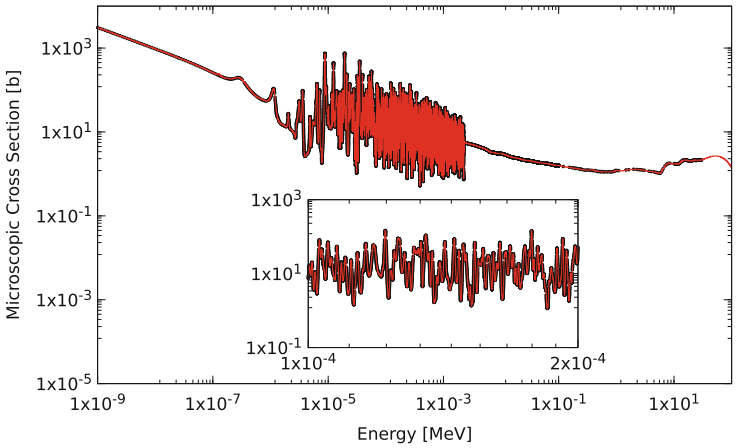



**Fig. 2.2** Microscopic fission cross section of uranium-235

those that do, because the fitting process of the regions that do not show resonances may be setup by simple polynomial functions. In this work, the window functions are constructed using hyperbolic tangent functions.

$$w(x) = \frac{1}{2}\left(\tanh(a_0(x - a_1)) + a_2 - \tanh(a_3(x - a_4))\right) \ .$$

Here, $a_0$ and $a_3$ are adjusted in such a way that the error between the cross section data and the parametrization function is minimized, while $a_1$, $a_2$ and $a_4$ limit the range in which the respective window function acts.

In Fig. 2.2, the microscopic fission cross section of the uranium-235 obtained directly from the database is represented by the black points and the analytic function composed of polynomials, rational and window functions that reproduces these points is represented by the red line. Furthermore, in the figure, a small range of the resonance region is shown to point out the large amount of resonances present in the cross section. These regions with dense variations are common in heavy nuclides and imply in pronounced oscillations in the total cross section of the medium, a necessary ingredient to obtain the pseudo-cross section for neutron escape through Eqs. (2.3) or (2.4).

## 2.4 Calculation of $\Sigma_z(\mathbf{r}, E)$ for a Spherical Case

In order to apply the proposed methodology to calculate the pseudo-cross section for neutron escape, neutron transport in a spherical finite multiplicative medium similar to early criticality experiments was simulated. The medium was composed of 72% of water, which plays the role of the moderator for fast neutrons, and 28%

of uranium dioxide, which is the nuclear fuel. Furthermore, the uranium present in the fuel was enriched to 2.5%, i.e., 2.5% of the uranium in the uranium dioxide is uranium-235. The spherical finite domain had a diameter of 76 cm, and due to the fact that the simulation was performed on a personal computer without large storage capacity the initial population amounted to $2 \times 10^4$ neutrons. These characteristics were chosen so that the multiplicative medium operated in a critical regime, which represented a situation where the neutron distribution in the domain did not vary with time, i.e., represented a steady state problem. For this setup, the respective pseudo-cross section for neutron escape was obtained by the methodology covered in Sect. 2.3.

As can be seen in the results further down, the number of initial neutrons used is relatively small and implies in some limitations, mainly due to the incremental intervals that constitute the phase space, which are not small enough to describe the neutron distributions with all its details in the domain. Despite this shortcoming, the main idea in this work is to present the methodology for obtaining the pseudo-cross section for neutron escape, so that these calculations for the present simulation may be understood as an instructive example for applications of the lined out new conception. Furthermore, even with these limitations, global and semi-quantitative properties of the pseudo-cross section can be determined. Nevertheless, for future applications, a larger ensemble with neutrons shall be used as the initial condition of the simulation.

Returning to the calculation of the pseudo-cross section of neutron escape for the present situation, $n_{z_r}(\mathbf{r})$, $n_{y_r}(\mathbf{r})$, $n_{z_E}(E)$ and $n_{y_E}(E)$ are determined from the generated data by the Monte Carlo simulation and $\Sigma_t(E)$ by the nuclear database so that one is ready to calculate the associated $\Sigma_z(\mathbf{r}, E)$ through the simplified way described by Eq. (2.4). In Fig. 2.3, the analytical function that parametrizes the total macroscopic cross section of the simulation medium is shown. By inspection, one observes that $\Sigma_t(E)$ is composed of several resonances, which was to be expected, since these stem from the resonances present in the microscopic cross sections of the nuclides that make up the medium, mainly not only from uranium for lower energies but also from oxygen for higher energies.

Two other necessary terms are the density functions that represent the number of interactions per unit volume, i.e., $n_{z_r}(\mathbf{r})$ and $n_{y_r}(\mathbf{r})$. In Fig. 2.4, these density functions are illustrated as black points representing the counts obtained directly from the simulation and by the blue and red solid lines that parametrize these points in form of analytical functions. Note that the graph is in linear scale and only interactions that occurred at a radius larger than 26 cm are presented, since almost no neutrons escaped from the central region of the finite domain. Furthermore, one may notice that the number of escape interactions is too small and practically constant up to approximately 70 cm. From this point on, which is already very close to the finite domain boundary, an exponential like growth sets in such that the closer the positions are to the surface, the larger is the number of escape interactions that approximate the number of integral interactions in the domain. This implies that

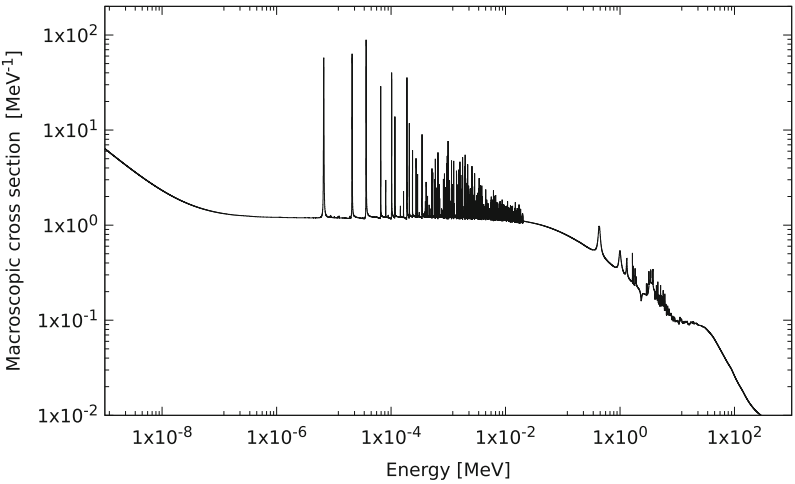



**Fig. 2.3** Total macroscopic cross section of the medium

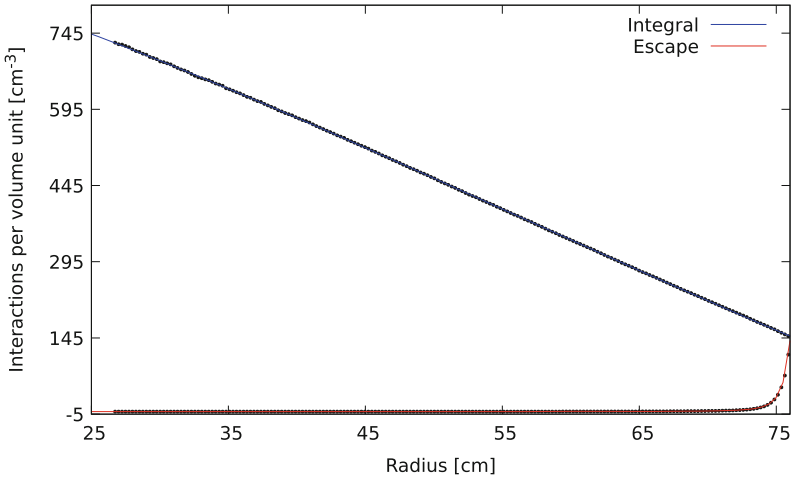



**Fig. 2.4** Number of escape and integral interactions per unit volume

the $\frac{n_{z_r}(\mathbf{r})}{n_{y_r}(\mathbf{r})}$ ratio tends to unity when $\mathbf{r}$ tends towards the radius that limits the finite multiplicative medium, in this case, 76 cm.

The ultimate functions needed to find the pseudo-cross section for neutron escape through Eq. (2.4) are $n_{z_E}(E)$ and $n_{y_E}(E)$. Again, in Fig. 2.5, these are illustrated as black points representing the counts obtained directly from the Monte Carlo simulation, while the blue and the red solid lines parametrize these points in an analytical fashion. One observes that the number of escape interactions per energy unit has the same behaviour as the number of integral interactions except for the

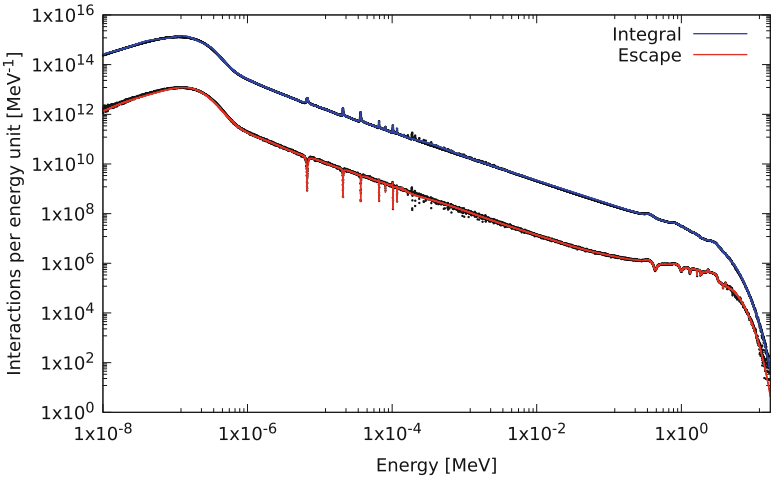



**Fig. 2.5** Number of escape and integral interactions per energy unit

peak regions. Furthermore, throughout the energy spectrum, the order of magnitude in the number of escape interactions varies between 1 and 10% compared to the order of magnitude of the integral interactions, which is an expected proportion for this kind of problem.

Other features that can be identified in Fig. 2.5 are the visible peaks and crevasses. With regard to the integral interactions, these peaks signify an increase in the number of interactions in certain narrow regions of the energy spectrum and are a direct consequence of the total cross section. On the other hand, for escape interactions, there are crevasses in the same regions of the energy domain and these represent a decrease in the number of escape interactions, which is also a direct consequence of $\Sigma_t(E)$, as discussed further down. Figure 2.5 also illustrates the limitations due to the low number of initial neutrons already mentioned in the beginning of this section. By inspection, one notices that the number of resonances in the total macroscopic cross section shown in Fig. 2.3 is considerably larger than those observed in Fig. 2.5. In fact, these peaks are also expected to appear in the number of interactions per energy unit as shown in Fig. 2.5, however, the sparse sampling made it impossible for the resonances to be sharply reproduced since the density of simulated data points is too low and it is impossible to identify these peaks with fidelity in the optimization process. As can be seen in the figure, only some resonances up to $1 \times 10^{-4}$ and around $1 \times 10^{0}$ were parameterized through the analytic function, while in the region with the most dispersed data, only their average was determined in the data fit process.

Another limitation is due to the use of considerably large incremental intervals $\Delta E$ such that the heights of the peaks in Fig. 2.5 are higher for integral interactions and consequently smaller for escape interactions, a property of averages. This effect

occurs because in the same $\Delta E$, the number of interactions per energy unit is influenced by regions in the energy spectrum with a high cross section, characterized by the extremes of the resonances, and regions with a lower cross section in the immediate vicinity of these extremes. In order to deal with these limitations, the total macroscopic cross section of the medium used to calculate the pseudo-cross section for neutron escape has been averaged so that in the energy region of the spectrum that corresponds to the interval in which the data is dispersed in Fig. 2.5 also the mean value of $\Sigma_t(E)$ was considered. Furthermore, for each of the $\Delta E$ that constitute the phase space, the mean of the total cross section was obtained from the integration of the analytic function which represents $\Sigma_t(E)$ as shown in Fig. 2.3 and considering as integration limits the respective intervals $\Delta E$.

Due to the analytic representation of all functions, the integration process is simple and fast. In this way, the shortcoming by the low initial number of neutrons can be compensated and $\Sigma_z(\mathbf{r}, E)$ can be obtained based on this approximation as illustrated in Fig. 2.6, where the total macroscopic cross section is shown according to the adopted energy resolution. It is evident that the above procedure does not need to be performed if $\Delta E$ is of compatible size with the resonance densities so that the total macroscopic cross section used to calculate the pseudo-cross section for neutron escape is the one illustrated in Fig. 2.3, which is of the same precision as the database.

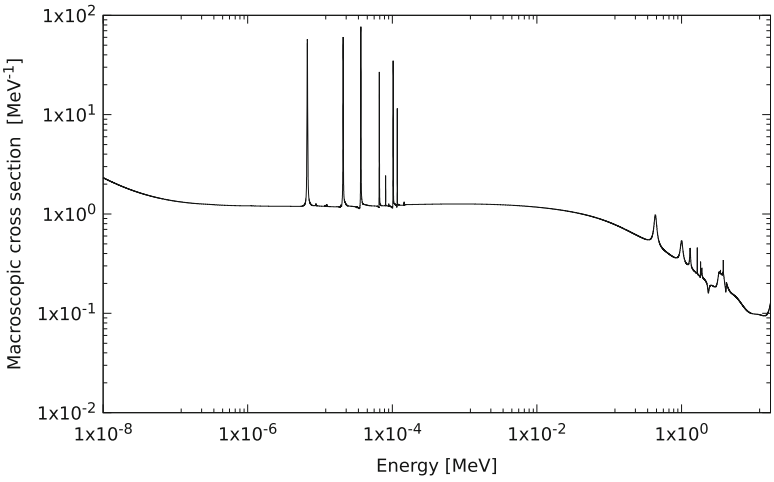



**Fig. 2.6** Total macroscopic cross section according to the adopted energy resolution

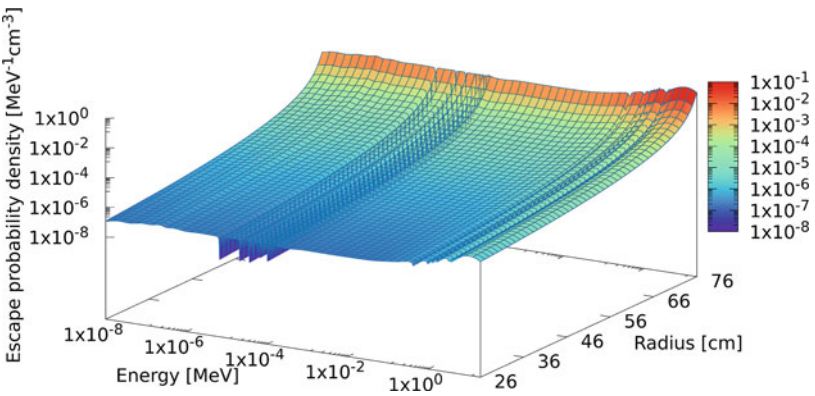



**Fig. 2.7** Escape probability density function

Before presenting the pseudo-cross section for the implemented simulation and based on the density functions that represent the number of interaction per volume unit and per energy unit, it is useful to determine the neutron escape probability as a function of the neutron position and its energy (shown in Fig. 2.7) upon introducing a normalization $\frac{N(\mathbf{r},E)_z}{N(\mathbf{r},E)_y}$. In agreement with expectation from experimental evidence, the escape probability increases as the neutron position gets closer to the boundaries of the finite domain and as the neutron's energy increases. Moreover, the escape probability significantly decreases for neutrons whose kinetic energy corresponds to the energy of one of the resonances present in the total macroscopic cross section. This result is also expected theoretically, because at these resonances the probability of a real interaction in the domain increases considerably and thus decreases the chance of the respective neutron to escape. The same explanation applies for the peaks and crevasses in the number of interactions per energy unit in Fig. 2.5.

Finally, with all terms on the right-hand side of Eq. (2.4) known and parameterized by analytic functions, it is now possible to calculate the function that represents the pseudo-cross section for simulated neutron escape from the finite domain as shown in Fig. 2.8. Note that the peaks remain visible which is expected since the pseudo-cross section depends directly on the escape probability density function and $\Sigma_t(E)$. In fact, peaks are expected to imply a decrease in $\Sigma_z(\mathbf{r}, E)$, however, as can be seen, the weaker oscillating behaviour may also be related to the aforementioned limitations and the procedure for dealing with these by the use of averages, as well as by anti-correlations between the resonances along event chains, thus decreasing the number of interactions per energy unit. Nevertheless, the general behaviour of the pseudo-cross section is as expected so that in principle the implemented simulation may be considered consistent and sound from the programming and physical point of view, respectively.

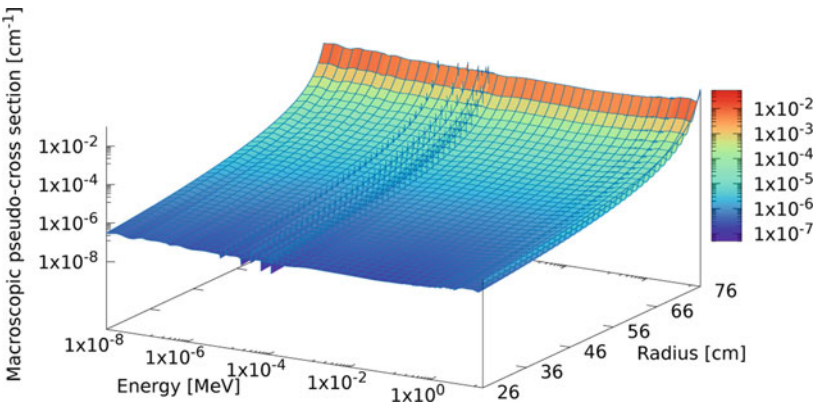



**Fig. 2.8** Macroscopic pseudo-cross section for neutron escape

## 2.5 Properties and Application of $\Sigma_z(\mathbf{r}, E)$

As discussed in Sect. 2.1, the escape can be quantified through the pseudo-cross section for neutron escape. An interesting property of $\Sigma_z(\mathbf{r}, E)$ is that neutrons that approach and interact near the surfaces of the finite domain tend to escape, implying that practically all interactions, except for some back scattering that redirects the neutron into the domain, are escape interactions. This can also be seen in Fig. 2.4 where the ratio between the number of escape interactions and integral interactions tends to unity as the interaction position approaches the boundary radius of the spherical multiplicative medium. Thus, note that

$$\frac{N(\mathbf{r}, E)_z}{N(\mathbf{r}, E)_y} \to 1 \implies \Sigma_z(\mathbf{r}, E) \to +\infty \, .$$

Based on this property, it is possible to use $\Sigma_z(\mathbf{r}, E)$ in deterministic and stochastic approaches considering initially infinite domains in a way that $\Sigma_z(\mathbf{r}, E)$ may account for the finite contours of a medium volume of interest. Thus, the introduction of $\Sigma_z(\mathbf{r}, E)$ in neutron transport problems may be justified the same way absorption cross sections are accounted for, i.e., leakage may be inserted directly in the transport equation for deterministic methods or added as a possible pseudo reaction in the random process that dice the type of interaction in stochastic methods.

This conception brings about an advantage in stochastic methods such as the physical Monte Carlo. Using the probability density functions for escape implies a reduction in the number of Monte Carlo steps in the simulations, since the execution of the last step of the neutron which escapes the domain (i.e., tracking) is no longer necessary. Moreover, these distributions may even simplify deterministic methods since there is no longer any need for neutron fluxes as boundary conditions.

Otherwise, the knowledge of the flux of neutron escape at the boundary may replace the commonly imposed zero flux condition or constructions that make use of an extrapolated distance.

## 2.6 Conclusion

In the present work, a new approach to quantify neutron escape from finite domains was presented by a methodology which results in a pseudo-cross section for neutron leakage and treats the latter on the same base as the physical interactions. Furthermore, with the calculation of this pseudo-cross section for a specific case, some global characteristics could be verified although it would have been desirable to implement the simulation starting with a larger ensemble, for instance, with $10^6$ neutrons. Highly populated initial ensembles allow to cast all density and distribution functions in analytical form so that the resulting pseudo-cross section is then crucial to improve, simplify and facilitate calculations and simulations involving the neutron transport in finite domains for stochastic as well as deterministic approaches.

Apart from the indicated limitations, we derived and showed how to use the pseudo-cross section conception for neutron escape in the study of neutron transport. Evidently, the higher the resolution, the better are the parametrizations and the more accurate are the reproduction of physical details such as the influence of the resonances on the densities and the probability functions. Nevertheless, the present approximation of $\Sigma_z(\mathbf{r}, E)$ by the use of averages in the spatial and energy intervals clearly showed the perspectives that arise so that in future simulations with power computing resources all the known details in $\Sigma_t(E)$ may be embedded in the calculation of the escape cross section and that the rich influence of the resonances and their effects on leakage may be evaluated in high fidelity.

Another type of study that can be done is the improvement of the methodology for obtaining the pseudo-cross section for neutron escape through the consideration of non-stationary problems and the dependence on the direction of motion of the neutrons in such a way that $\Sigma_z(\mathbf{r}, E)$ becomes more accurate and generalized. Finally, it is worth noting that this is a first approach in the study of neutron escape through a pseudo-cross section and that other properties and applications in different areas of particle transport can be derived using findings from this kind of approach.

## References

[BarEtAl17] Barcellos, L.F.F.C., Bodmann, B.E.J., de Queiroz Bogado Leite, S., e de Vilhena, M.T.: On a continuous energy Monte Carlo simulator for neutron transport: Optimisation with fission, intermediate and thermal distributions. In: Constanda, C., Riva, M.D., Lamberti, P.D., Musolino, P. (eds.) Integral Methods in Science and Engineering, vol. 2, pp. 1–10. Birkhäuser, Berlin (2017)

[BaEtAl21] Barcellos, L.F.F.C., Bodmann, B.E.J., e de Vilhena, M.T.: On a comparison of a neutron Monte Carlo transport simulation to a criticality benchmark experiment. Prog. Nucl. Energy **134**, 103652 (2021)

[BeGl70] Bell, G. I., e Glasstone, S.: Nuclear Reactor Theory. Van Nostrand Reinhold Company, New York (1970)

[CaEtAl11] de Camargo, D.Q., Bodmann, B.E.J., de Vilhena, M.T., e de Queiroz Bogado Leite, S.: A novel method for simulating spectral nuclear reactor criticality by a spatially dependent volume size control. In: Constanda, C., Harris, P.J. (eds.) Integral Methods in Science and Engineering: Computational and Analytic Aspects, pp. 33–45. Birkhäuser, Boston (2011)

[CaEtAl13] de Camargo, D.Q., Bodmann, B.E.J., de Vilhena, M.T., de Queiroz Bogado Leite, S., e Alvim, A.C.M.: A stochastic model for neutrons simulation considering the spectrum and nuclear properties with continuous dependence of energy. Prog. Nucl. Energy **69**, 59–63 (2013)

[CaCa75] Carter, L.L., e Cashwell, E.D.: Particle-Transport Simulation with the Monte Carlo Method, National Technical Information Service, Springfield (1975)

[FeEtAl17] Fernandes, J.C.L., Bodmann, B.E.J., de Vilhena, M.T.: On multi-group neutron transport in planar one dimensional geometry: A solution for a localized pulsed source. Ann. Nucl. Energy **101**, 552–558 (2017)

[La66] Lamarsh, J.R.: Introduction to Nuclear Reactor Theory. Addison-Wesley Publishing Company, Boston (1966)

[La06] Larsen, E.W.: An overview of neutron transport problems and simulation techniques. In: Graziani, F. (ed.) Computational Methods in Transport, pp. 513–533. Springer, Berlin (2006)

[LeMi84] Lewis, I.E., e Miller, Jr, W.F.: Computational Methods of Neutron Transport. Wiley, New York (1984)

[OlEtAl17] Oliveira, F.R., Bodmann, B.E.J., de Vilhena, M.T., e Carvalho, F.: On an analytical formulation for the mono-energetic neutron space-kinetic equation in full cylinder symmetry. Ann. Nucl. Energy **99**, 253–257 (2017)

[OlEtAl19] Oliveira, F.R., Fernandes, J.C.L., Bodmann, B.E.J., e de Vilhena, M.T.: On an analytical solution for the two energy group neutron space-kinetic equation in heterogeneous cylindrical geometry. Ann. Nucl. Energy **133**, 216–220 (2019)

[TuEtAl19] Tumelero, F., Bodmann, B.E.J., de Vilhena, M.T., e Lapa, M.F.: On the solution of the neutron diffusion kinetic equation in planar geometry free of stiffness with convergence analysis. Ann. Nucl. Energy **125**, 272–282 (2019)

# Chapter 3
# From a Unitary Symmetry Hypothesis to Dynamical Structures in Quantum Mechanics Models


**B. E. J. Bodmann**

## 3.1 Introduction

Mathematical modelling of physical systems in general starts from a specific principle such as dimensional analysis or some conservation law, and in cases where little is known about a system, a phenomenological *Ansatz* may define an irrevocable initial hypothesis. The formal description that constitutes the model is frequently given by a partial differential equation or equation system. However, conservation laws seem to be a safe justification for the construction of a model due to the fact that some symmetries are universal independent whether they apply to classical, statistical or quantum mechanics among other realms, where energy and momentum conservation are probably the most employed examples.

In classical models, the dynamical equation may have several symmetries such as space and time translations and their associated momentum and energy conservation, respectively, but non-homogeneous boundary conditions break some of the symmetries explicitly. Self-consistent quantum systems though at best are subject to a normalization condition such that finiteness of physical observables is guaranteed or if normalization is not possible, then expectation values in the spirit of the Gell-Mann–Low theorem shall exist [GeLo51, Mo07]. Nevertheless, the probability fields are defined in an infinite space so that in most cases there do not exist boundary conditions, and thus symmetries remain preserved in dynamical formulations or may be broken spontaneously [ArEtAl03]. In this sense, modelling quantum systems seems to be more simple than classical systems due to the symmetry argument. When considering conservation laws and following

B. E. J. Bodmann (✉)
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: bardo.bodmann@ufrgs.br



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_3

the reasoning of Noether's theorem symmetry transformations and their associated invariants play the principal role [No18, HaEtAl04].

Commonly, a dynamical equation (system) comes first, and only in a second step the symmetries of the equation (system) are analysed and identified. It is probably safe to say that even without the knowledge of a dynamical equation, the presence of (at least some) symmetries of a physical system of interest is easier to be recognized than determining them from the model. Hence, in the present discussion, an alternative modelling strategy is shown, which starts from a symmetry hypothesis and progressively leads to the ingredients that allow to construct the most general dynamical model inheriting the underlying symmetries. Evidently, such a procedure needs a principle which provides the connection between symmetry manifestations and dynamics, where in the present quantum system the model will define the dynamical vacuum structure.

Symmetries in differential geometry may be built by the use of the two fundamental forms of Gauss with invariants such as the rest mass squared, the eigen-time squared and an expression with the property of a phase of the type $\omega t - \mathbf{kr}$, which is a composition of space-time coordinates and their dual quantities (obtained by integral transforms) proportional to energy and momentum. While these invariants may be interpreted in a geometrical fashion for the purpose of dynamical modelling one needs a functional principle, i.e., the action integral together with the variational principle [Gr09]. Furthermore, one needs a fragment of the dynamical description as for instance a term compatible with free propagation of a quantum particle with constant energy–momentum, which triggers the cascade of symmetry breaking and symmetry restoring steps until in the end an invariant action arises. The aim of the forthcoming treatise is to show how to progressively construct a model compatible with an initial symmetry hypothesis.

## 3.2 The Symmetry Hypothesis

Observable quantities in quantum systems are typically computed as expectation values provided by sesqui-linear forms. The most simple example is that of a density, where a probability amplitude is multiplied by its adjoint so that the density remains invariant under a unitary transformation $U(1)$ [SaNa20]. Such an invariance may be related with charge conservation for instance since the continuity equation remains the same after the aforementioned transformation. Physical systems are not necessarily governed by only one type of interaction, where $U(1)$ could be associated with electromagnetism so that for the second type of interaction a second but different transformation shall be brought in which otherwise could not be differentiated from the unitary one and thus would not present anything new. A natural choice is the $SU(2)$ symmetry transformation which also complies with a charge conservation and could be similar to the weak charge of the weak interaction for instance [GeLe60].

Both the unitary $U(1)$ and also the $SU(2)$ are compatible with the continuity equation and thus impose no restriction in the question of mass terms. In a work [NaJo61], the presence of scalar terms like the mass terms in a quantum field model was associated with a spontaneous symmetry breaking of chiral symmetry, so that it seems plausible to add the chiral counterparts of the two unitary transformations ($U_5(1)$ and $SU_5(2)$) to the symmetry hypothesis and thus open up for the possibility to generate the particle mass terms from spontaneously broken $U_5(1) \otimes SU_5(2)$ symmetries. To this end, variations with respect to the combination of the symmetry transformations are applied to the action integral $\delta_+ \mathscr{A} = (\delta + \delta_5)\mathscr{A}$, which is the same applying the variations to the Lagrange densities since it is the Grassmann variables which are being varied while the coordinates remain the same and thus the variational operators commute with the integral operator and the invariance of the action integral is equivalent to determining the invariant Lagrange density.

## 3.3 The Free Particle Lagrange Density

As an initial attempt, one may consider the free particle Lagrange density and apply infinitesimal local unitary transformations to the Grassmann variables with the intent to determine the terms that explicitly break the combined $U(1) \otimes SU(2) \otimes U_5(1) \otimes SU_5(2)$ symmetries. Matter constituents are Fermions so that the free Dirac Lagrange density is that first fragment of the model.

$$\mathscr{L}_0 = \frac{\imath}{2}\left(\bar{\Psi}\gamma^\mu \partial_\mu \Psi - \partial_\mu \bar{\Psi}\gamma^\mu \Psi\right).$$

Here, $\Psi$, $\bar{\Psi} = \Psi^\dagger \gamma^0$ are spinors and adjoint spinors, the $\cdot^\dagger$ signifies the conjugate transposition, the $\gamma^\mu$ are the Dirac matrices and $\partial_\mu$ is the extension of the $\nabla$-operator to space–time $(\partial_t, \nabla)$. In the further, double occurring indices imply summation over the time ($\mu = 0$) and space indices ($\mu = 1, 2, 3$).

The combined unitary symmetry group transformation in finite and infinitesimal form is given by

$$\Psi \to e^{\imath(\Lambda_a + \gamma_5 \Lambda_{5a})\tau^a}\Psi\,, \qquad \delta_+\Psi = \imath(\Lambda_a + \gamma_5\Lambda_{5a})\tau^a\Psi\,,$$
$$\bar{\Psi} \to \bar{\Psi} e^{-\imath\tau_a(\Lambda_a - \gamma_5\Lambda_{5a})}\,, \qquad \delta_+\bar{\Psi} = -\imath\bar{\Psi}\tau_a(\Lambda_a - \gamma_5\Lambda_{5a})\,,$$

where $\Lambda_a$ and $\Lambda_{5a}$ are local Hermitian operators with independent local eigenvalues $\lambda_a$ and $\lambda_{5a}$, respectively. The adjective local means that the Hermitian operators as well as their eigenvalues are space–time dependent and in this sense are a manifestation of the functional character of interactions.

A complementary comment is in order here, the spinor does not only contain spin parity information but also contains pairs of spinors which are connected by the chiral operator, which is capable to transform the respective upper and lower

components into each other. Isospin in medium energy nuclear physics is one example where one state represents the proton and the other one a neutron, which are connected by $\beta$-decays, i.e. the weak interaction.

$$\Psi = \left(\psi_{+\frac{1}{2}}, \psi_{-\frac{1}{2}}\right)^T \qquad \tau_a = (\tau_0, \boldsymbol{\tau})^T .$$

Here, $\tau_0$ is the identity, $\boldsymbol{\tau}$ is a vector containing as components the three Pauli matrices and $\Lambda_a \tau^a$ is a quaternion [Pu12].

If we denote the total variation as a superposition by a unitary and chiral variation $\delta_+ = \underbrace{\delta}_{unitary} + \underbrace{\delta_5}_{chiral}$, then the variation of the free Dirac Lagrange density is

$$\delta_+ \mathscr{L}_0 = -\frac{1}{2}\bar{\Psi}\{\gamma^\mu, \tau_a \partial_\mu \Lambda^a\}\Psi - \frac{1}{2}\bar{\Psi}[\gamma^\mu, \gamma_5 \tau_a \partial_\mu \Lambda_5^a]\Psi \neq 0,$$

which because of the non-vanishing commutator ($[\cdot,\cdot]$) and anti-commutator ($\{\cdot,\cdot\}$) shows that local symmetry is explicitly broken. By inspection, one observes that the first term is a vector current density times a vector, i.e. the derivative of the Hermitian operator, while the second term is an axial vector current density times an axial vector, so that one at least has an indication what type of physical field one has to introduce together with its interaction to the Fermions in order to get one step closer to symmetry restoration.

## 3.4 The Interaction Lagrange Density

The necessary terms to be introduced may be understood as symmetry constraints $\mathscr{G}$ and may be added to $\mathscr{L}_0$ by the use of Lagrange multipliers thus ending up with a modified Lagrange density.

$$\mathscr{L}_1 = \mathscr{L}_0 + g\mathscr{G}_0 + g_5\mathscr{G}_1$$

The explicit form of $\mathscr{L}_1$ is given below.

$$\mathscr{L}_1 = \mathscr{L}_0 + g\bar{\Psi}\gamma^\mu \tau_a \Omega_\mu^a \Psi + \frac{\imath g_5}{2}\bar{\Psi}\left(\gamma^\mu \tau_a \Xi_\mu^a - \Xi_\mu^{a\dagger}\tau_a \gamma^\mu\right)\Psi.$$

Evidently, the vector current vector field interaction term and the axial vector current axial vector field interaction term restore the symmetry by compensating the terms which break local symmetry of the free particle Lagrange density $\delta\mathscr{L}_0 \neq 0$. Moreover, the additional terms together with the derivative terms from the free Dirac Lagrange density allow now to interpret them as a generalized gauge covariant derivative in analogy to the substantial derivative in fluid mechanics. In the earlier literature, the extended derivative was called “minimal substitution”, which appears

here as a symmetry constraint.

$$\partial_\mu \rightarrow \partial_\mu - \imath g \tau_a \Omega_\mu^a + g_5 \tau_a \Xi_\mu^a.$$

One may now use the fact that $\mathscr{L}_1$ has to be a genuine Lorentz scalar density of weight $w = -1$, which imposes some constraints $[\gamma^0, \Xi_\mu^{a(\dagger)}] = 0$ on the new field associated with the chiral transformation and reduces $\Xi$ to a block diagonal form.

$$\Xi_\mu^a = \begin{pmatrix} \xi_\mu^a & 0 \\ 0 & \eta_\mu^a \end{pmatrix}.$$

The necessity for the commutation relation between the parity operator (where an arbitrary phase was dropped) and the tensor fields $\Xi_\mu$ becomes apparent upon analysing the variation of the interacting Lagrange density $\delta_+ \mathscr{L}_1$ as described in detail next.

$$\begin{aligned}
\delta_+ \mathscr{L}_1 = & -\frac{1}{2}\bar{\Psi}\left\{\gamma^\mu, \tau_a \partial_\mu \Lambda^a\right\}\Psi - \frac{1}{2}\bar{\Psi}\left[\gamma^\mu, \gamma_5 \tau_a \partial_\mu \Lambda_5^a\right]\Psi \\
& + \frac{g}{2}\bar{\Psi}\left\{\gamma^\mu, \delta_+(\tau_a \Omega_\mu^a)\right\}\Psi + \frac{g_5}{2}\bar{\Psi}\left[\tau_a \Lambda^a, (\gamma^\mu \tau_b \Xi_\mu^b - \tau_b \Xi_\mu^{b\dagger}\gamma^\mu)\right]\Psi \\
& + \frac{\imath g_5}{2}\bar{\Psi}\left(\gamma^\mu \delta_+(\tau_a \Xi_\mu^a) - \delta_+(\tau_a \Xi_\mu^{a\dagger})\gamma^\mu\right)\Psi \\
& - \frac{g_5}{2}\bar{\psi}\left\{\gamma_5 \tau_a \Lambda_5^a, (\gamma^\mu \tau_b \Xi_\mu^b - \tau_b \Xi_\mu^{b\dagger}\gamma^\mu)\right\}\Psi.
\end{aligned} \tag{3.1}$$

The symmetry conditions of the variation are now based on the fact that individual terms have their characteristic behaviour under transformations (vector, axial vector and tensor), so that one shall group terms together with corresponding properties under the aforementioned transformations, which constitute the symmetry conditions similar to well-established gauge conditions in the field theoretical structure of the Standard Model of elementary particles [YaMi54].

As is the case in the well-established electromagnetic interaction with spinors the present extension to a symmetry transformation by groups defined by operators that follow a Lie algebra, the unitary transformation of the vector field establishes a local gauge symmetry if

$$g\delta\Omega_\mu^a = \partial_\mu \lambda^a \ .$$

Furthermore, due to the anti-commutation relation of the chiral operator and the Dirac matrices $\{\gamma^\mu, \gamma^5\} = 0$, the chiral variation of the polar vector field ($\delta_5(\Omega_\mu^a) = 0$) vanishes, and thus the first and the third terms of the right-hand side of Eq. (3.1) cancel. If we denote the transformation matrices $\boldsymbol{\tau}$ which form a Lie algebra by isotopic-spin operators, then isotopic-spin symmetry implies

$$\delta \Xi_\mu^b = 2\epsilon_{abc} \lambda^a \Xi_\mu^c$$

for $a, b, c \in \{1, 2, 3\}$ and where $\epsilon_{abc}$ is the totally antisymmetric Levi–Civita tensor of rank three. Thus, the unitary variation of the tensor fields compensates the fourth term on the right-hand side of Eq. (3.1), so that the last terms that shall match and establish symmetry are a local and global chiral symmetry condition. Here, the local chiral symmetry condition comes from the second term and the global one from the last term on the right-hand side of Eq. (3.1).

$$g_5\delta_5\Xi^a_\mu = -\imath\gamma_5\partial_\mu\lambda^a_5 + \imath g_5\tau_b\left(\gamma_5\lambda^a_5\Xi^b_\mu - \Xi^a_\mu\gamma_5\lambda^b_5\right).$$

What we have got so far is an interacting Lagrange density for spinor fields (i.e. Fermions), which is invariant under a combined $U(1)\otimes U_5(1)\otimes SU(2)\otimes SU_5(2)$ symmetry transformation, and however the model is not closed yet since there is a further need for a Boson Lagrange density so that the set of dynamical equations, which are derived from the Lagrange density using Hamilton's principle allow to compute the solutions for all involved fields except for a gauge degree of freedom still to be fixed.

## 3.5 The Interacting Boson Lagrange Density

The interacting Boson Lagrange density may be constructed using either formal arguments from differential geometry or one may copy from the theory of electromagnetic interaction. In either case, one considers some kind of parallel transport of momentum along an infinitesimal closed curve, which in analogy to Stokes law of electromagnetism induces in case of curvature an axial vector field in its interior. In order to construct from this reasoning a scalar density of weight $w = -1$, one may consider a flux term, which is sesqui-linear and if only $U(1)$ symmetry is considered which corresponds to charge conservation, then the resulting Lagrange density generates via Hamiltons principle the first two Maxwell equations, while the third and fourth equations are obtained from the first two upon applying the Hodge operator [Jo02, TeCh99]. In shorthand notation with $\tilde{D}_\mu = \partial_\mu - \imath g\tau_a\Omega^a_\mu + g_5\tau_a\Xi^a_\mu$ the resulting interacting Boson Lagrange density is

$$\begin{aligned}
\mathscr{L}_2 &= \frac{1}{32g^2}Tr_{SI}\left\{[\tilde{D}^\dagger_\nu, \tilde{D}^\dagger_\mu][\tilde{D}_\mu, \tilde{D}_\nu]\right\}, \\
&= \frac{1}{32g^2}Tr_{SI}\left\{\left(\imath g\tau^a(\partial_\nu\Omega_{a\mu} - \partial_\mu\Omega_{a\nu}) + g_5\tau^a(\partial_\nu\Xi^\dagger_{a\mu} - \partial_\mu\Xi^\dagger_{a\nu})\right.\right. \\
&\quad -g^2[\tau^a\Omega_{a\nu}, \tau^b\Omega_{b\mu}] + \imath gg_5[\tau^a\Omega_{a\nu}, \tau^b\Xi^\dagger_{b\mu}] \\
&\quad \left.\left.+\imath gg_5[\tau^a\Xi^\dagger_{a\nu}, \tau^b\Omega_{b\mu}] + g_5^2[\tau^a\Xi^\dagger_{a\nu}, \tau^b\Xi^\dagger_{b\mu}]\right) \right. \\
&\quad \left.\left(-\imath g\tau^c(\partial^\mu\Omega^\nu_c - \partial^\nu\Omega^\mu_c) + g_5\tau^c(\partial^\mu\Xi^\nu_c - \partial^\nu\Xi^\mu_c)\right.\right.
\end{aligned}$$

$$-g^2[\tau^c\Omega_c^\mu,\tau^d\Omega_d^\nu]-\imath g g_5[\tau^c\Omega_c^\mu,\tau^d\Xi_d^\nu]$$
$$-\imath g g_5[\tau^c\Xi_c^\mu,\tau^d\Omega_d^\nu]+g_5^2[\tau^c\Xi_c^\mu,\tau^d\Xi_d^\nu]\Big)\Big\},$$

where $Tr_{SI}\{\cdot\}$ denotes the trace over spin $S$ and isotopic-spin $I$ degrees of freedom.

Formally, the Lagrange density is setup by 36 terms, though as the analysis to follow will show, there are terms which naturally vanish and there are terms that impose restrictions such as to make sense from a physical point of view. Note, that some of the 36 terms are identical and some of them are related by Hermitian conjugation, so that for the cases of vanishing terms the number of expressions to be evaluated effectively reduces. In the following, we denote as "term $n$-$m$" the trace over spin and isotopic-spin degrees of freedom of the product between the $n$-th term of the first commutator and the $m$-th term of the second commutator.

## *3.5.1 The n-n Terms*

### 3.5.1.1 Term 1-1

This term reminds one on the completely antisymmetric tensor field term of electromagnetism except for the additional isotopic-spin degrees of freedom.

$$\frac{1}{32g^2}Tr_{SI}\ \Big\{\Big(\imath g\tau^a(\partial_\nu\Omega_{a\mu}-\partial_\mu\Omega_{a\nu})\Big)\Big(-\imath g\tau^c(\partial^\mu\Omega_c^\nu-\partial^\nu\Omega_c^\mu)\Big)\Big\}$$
$$=\frac{1}{4}(\partial_\nu\Omega_{a\mu}-\partial_\mu\Omega_{a\nu})(\partial^\mu\Omega^{a\nu}-\partial^\nu\Omega^{a\mu}).$$

Here, the spin trace supplies a factor of 4, whereas the isotopic-spin trace yields $Tr\{\tau^a\tau^c\}=2\delta^{ac}$.

### 3.5.1.2 Term 2-2

This term is the chiral counterpart of the completely antisymmetric tensor field of term 1-1 which yields the kinetic Lagrange density of the pseudo scalar field with the covariant derivative.

$$\frac{1}{32g^2}Tr_{SI}\left\{g_5\tau^a(\partial_\nu\Xi_{a\mu}^\dagger-\partial_\mu\Xi_{a\nu}^\dagger)\left(g_5\tau^c(\partial^\mu\Xi_c^\nu-\partial^\nu\Xi_c^\mu)\right)\right\}$$
$$=\frac{g_5^2}{16g^2}Tr_S\left\{(\partial_\nu\Xi_{a\mu}^\dagger-\partial_\mu\Xi_{a\nu}^\dagger)(\partial^\mu\Xi^{a\nu}-\partial^\nu\Xi^{a\mu})\right\}$$

$$
\begin{aligned}
&= \frac{g_5^2}{16g^2}\epsilon_{\alpha\beta\nu\mu} D^{\dagger\alpha}\chi^b u_{ba}^{\beta}\epsilon^{\gamma\delta\mu\nu} D_\gamma \chi_c u_\delta^{ca} \\
&= \frac{1}{2}\{D^{\dagger\alpha}\chi^a D_\alpha \chi_a\} \\
&= \frac{1}{2}\partial^\mu \chi^a \partial_\mu \chi_a + g^2 \Omega_b^\mu \Omega_\mu^b \chi^a \chi_a
\end{aligned}
$$

Here, the generalized gauge covariant derivative was decomposed into the usual gauge covariant derivative $D_\gamma = \partial_\mu - \imath g \tau_a \Omega_\mu^a$ and further the $2 \times 2$ blocks of the tensor field $\Xi$ were parameterized by coefficients and the Pauli matrices $\xi_\mu^a = \sum \alpha_{\mu b}^a \sigma^b$. Moreover, the following terms were interpreted in terms of a Stokes like identity for stationary problems, so that

$$
\tau^a \left(\partial^\mu \alpha_a^{\nu b} - \partial^\nu \alpha_a^{\mu b}\right) = \tau^a \epsilon^{\alpha\beta\mu\nu} D_\alpha \chi_a u_\beta^b
$$

and

$$
\left(\alpha_a^{\mu b}\alpha_c^{\nu d} - \alpha_a^{\nu b}\alpha_c^{\mu d}\right) = \epsilon^{\alpha\beta\mu\nu}\chi_a\chi_c u_\alpha^b u_\beta^d \,,
$$

where $u_\alpha^b$ represents a four velocity carrying spin degrees of freedom and thus hides some physics relevant for larger energy scales or equivalently for smaller length scales as the ones relevant for the current model. Furthermore, a steady state implies that the substantial covariant derivative shall vanish $u_\beta^b D^\beta \to 0$. There is an additional constraint between the unitary and chiral coupling constant ($g_5^2 = \frac{g^2}{2}$) which guarantees that upon contracting spin degrees of freedom the Klein–Gordon Lagrange density may be recovered.

#### 3.5.1.3 Term 3-3

$$
\begin{aligned}
&\frac{1}{32g^2} Tr_{SI}\left\{-g^2[\tau^a\Omega_{a\mu}, \tau^b\Omega_{b\nu}]\left(-g^2[\tau^c\Omega_c^\mu, \tau^d\Omega_d^\nu]\right)\right\} \\
&= \frac{g^2}{8} Tr_I\left\{[\tau^a,\tau^b][\tau^c,\tau^d]\right\}\Omega_{a\mu}\Omega_{b\nu}\Omega_c^\mu\Omega_d^\nu \\
&= -g^2\epsilon^{abe}\epsilon^{cdf}\delta_{ef}\Omega_{a\mu}\Omega_{b\nu}\Omega_c^\mu\Omega_d^\nu \\
&= -g^2(\delta^{ac}\delta^{bd} - \delta^{ad}\delta^{bc})\Omega_{a\mu}\Omega_{b\nu}\Omega_c^\mu\Omega_d^\nu
\end{aligned}
$$

This is the four vector boson interaction term due to non-commutativity of isotopic spin, which turns also the free vector Boson Lagrange density a non-linear model contribution.

#### 3.5.1.4 Terms 4-4 and 5-5

The terms 4-4 and 5-5 are of identical structure except for space–time, spin and isotopic-spin indices and the first one is shown below.

$$\begin{aligned}
&\frac{1}{32g^2}Tr_{SI}\left\{\imath g g_5[\tau^a\Omega_{av},\tau^b\Xi^\dagger_{b\mu}](-\imath g g_5)[\tau^c\Omega^\mu_c,\tau^d\Xi^v_d]\right\}\\
&=-\frac{g_5^2}{4}\epsilon^{abe}\epsilon^{cdf}\delta_{ef}\Omega_{av}\Omega^\mu_c Tr_S\left\{\Xi^\dagger_{b\mu}\Xi^v_d\right\}\\
&=-g_5^2(\delta^{bd}\delta^{ac}-\delta^{bc}\delta^{ad})\Omega_{av}\Omega^\mu_c\alpha^\dagger_{b\mu k}\alpha^{vk}_d\\
&=-g_5^2(\alpha^{d\dagger}_{\mu k}\alpha^{vk}_d\Omega_{av}\Omega^{a\mu}-\Omega_{av}\Omega^{b\mu}\alpha^\dagger_{b\mu k}\alpha^{avk})
\end{aligned}$$

Note that this term together with the terms 4-5 and 5-4 is combined such as to provide an interaction term between the vector bosons and the pseudo scalar field.

#### 3.5.1.5 Term 6-6

$$\begin{aligned}
&\frac{1}{32g^2}Tr_{SI}\left\{g_5^2[\tau^a\Xi^\dagger_{a\mu},\tau^b\Xi^\dagger_{bv}]g_5^2[\tau^c\Xi^\mu_c,\tau^d\Xi^v_d]\right\}\\
&=\frac{g_5^4}{32g^2}Tr_{SI}\left\{[\tau^a\Xi^\dagger_{a\mu},\tau^b\Xi^\dagger_{bv}][\tau^c\Xi^\mu_c,\tau^d\Xi^v_d]\right\}\\
&=\frac{g_5^4}{16g^2}Tr_{SI}\left\{\tau^a\tau^b\tau^c\tau^d\Xi^\dagger_{a\mu}\Xi^\dagger_{bv}\Xi^\mu_c\Xi^v_d-\tau^a\tau^b\tau^d\tau^c\Xi^\dagger_{a\mu}\Xi^\dagger_{bv}\Xi^v_d\Xi^\mu_c\right\}\\
&=3g^2\Bigg(\delta^{a0}\delta^{b0}\delta^{c0}\delta^{d0}+\delta^{a0}\delta^{b0}\underbrace{\delta^{cd}}_{cd\neq0}+\delta^{b0}\delta^{c0}\underbrace{\delta^{ad}}_{ad\neq0}+\delta^{a0}\delta^{c0}\underbrace{\delta^{bd}}_{bd\neq0}\\
&\quad+\delta^{b0}\delta^{d0}\underbrace{\delta^{ac}}_{ac\neq0}+\delta^{a0}\delta^{d0}\underbrace{\delta^{bc}}_{bc\neq0}+\delta^{c0}\delta^{d0}\underbrace{\delta^{ab}}_{ab\neq0}\\
&\quad-\underbrace{(\delta^{ac}\delta^{bd}-\delta^{ad}\delta^{bc})}_{abcd\neq0}\Bigg)\chi_a\chi_b\chi_c\chi_d
\end{aligned}$$

### 3.5.2 Terms n-m ($n \neq m$) with Vanishing Spin Trace $Tr_S\{\Xi_a^{\mu(\dagger)}\} = 0$

In the following, the terms with vanishing spin trace $Tr\{\Xi_a^{\mu(\dagger)}\} = 0$ are shown.

#### 3.5.2.1 Term 1-2 and Hermitian Conjugate Term 2-1

This term is assumed to vanish such that no physical contradictions appear in the model.

$$\frac{1}{32g^2} Tr_{SI} \left\{ \left( \imath g \tau^a (\partial_\nu \Omega_{a\mu} - \partial_\mu \Omega_{a\nu}) \right) g_5 \tau^c (\partial^\mu \Xi_c^\nu - \partial^\nu \Xi_c^\mu) \right\}$$
$$= \frac{\imath g_5}{16g} (\partial_\nu \Omega_{a\mu} - \partial_\mu \Omega_{a\nu}) Tr_S \left\{ \partial^\mu \Xi^{a\nu} - \partial^\nu \Xi^{a\mu} \right\} = 0$$

Note that from the physical point of view, this term does not make sense since the incoming particle undergoes a process without a genuine vertex (interaction). Thus, upon implying $Tr_S\{\Xi_\mu^a\} = 0$ cures the problem, which is in agreement with the symmetry condition for the block matrices $\xi_\mu^a = -\eta_\mu^a$ due to the difference in parity of the upper and lower two components of the spinors.

#### 3.5.2.2 The Terms 1-4, 4-1, 1-5 and 5-1 Vanish Individually

$$\frac{1}{32g^2} Tr_{SI} \left\{ \left( \imath g \tau^a (\partial_\nu \Omega_{a\mu} - \partial_\mu \Omega_{a\nu}) \right) \left( -\imath g g_5 [\tau^c \Omega_c^\mu, \tau^d \Xi_d^\nu] \right) \right\}$$
$$= \frac{g_5}{32} (\partial_\nu \Omega_{a\mu} - \partial_\mu \Omega_{a\nu}) \Omega_c^\mu Tr_S\{\Xi_d^\nu\} Tr_I\{\tau^a [\tau^c, \tau^d]\}$$
$$= \frac{\imath g_5}{8} \epsilon^{cda} (\partial_\nu \Omega_{a\mu} - \partial_\mu \Omega_{a\nu}) \Omega_c^\mu Tr_S\{\Xi_d^\nu\} = 0$$

The cancellation of these terms stems from the fact that the trace over spin degrees of freedom is equal zero.

#### 3.5.2.3 Terms 2-3 and 3-2 Vanish Individually

$$\frac{1}{32g^2} Tr_{SI} \left\{ g_5 \tau^a (\partial_\nu \Xi_{a\mu}^\dagger - \partial_\mu \Xi_{a\nu}^\dagger) \left( -g^2 [\tau^c \Omega_c^\mu, \tau^d \Omega_d^\nu] \right) \right\}$$
$$= -\frac{\imath g_5}{8} \epsilon^{cda} Tr_S \left\{ \partial_\nu \Xi_{a\mu}^\dagger - \partial_\mu \Xi_{a\nu}^\dagger \right\} \Omega_c^\mu \Omega_d^\nu = 0$$

#### 3.5.2.4 Terms 3-4 and 4-3 Vanish Identically

$$\frac{1}{32g^2}Tr_{SI}\left\{-g^2[\tau^a\Omega_{a\mu},\tau^b\Omega_{bv}](-\imath gg_5)[\tau^c\Omega_c^\mu,\tau^d\Xi_d^v]\right\}$$
$$=-\frac{\imath gg_5}{4}\epsilon^{abe}\epsilon^{cdf}\delta_{ef}\Omega_{a\mu}\Omega_{bv}\Omega_c^\mu Tr_S\left\{\Xi_d^v\right\}=0$$

#### 3.5.2.5 Terms 3-5 and 5-3 Do Not Contribute to the Dynamics

$$\frac{1}{32g^2}Tr_{SI}\left\{-g^2[\tau^a\Omega_{a\mu},\tau^b\Omega_{bv}](-\imath gg_5)[\tau^c\Xi_c^\mu,\tau^d\Omega_d^v]\right\}$$
$$=-\frac{\imath gg_5}{4}\epsilon^{abe}\epsilon^{cdf}\delta_{ef}\Omega_{a\mu}\Omega_{bv}\Omega_d^v Tr_S\left\{\Xi_c^\mu\right\}=0$$

### *3.5.3 Terms n-m ($n \neq m$) with Vanishing Spin Trace*

In the following, the terms with vanishing spin trace are shown. These types are either of the form $Tr_S\{\partial_\mu\Xi_a^{\mu\dagger}\Xi_{b\mu}\Xi_{cv}\}=0$ or $Tr_S\{\Xi_a^{\mu\dagger}\Xi_{b\mu}\Xi_{cv}\}=0$.

#### 3.5.3.1 Term 2-6 and Hermitian Conjugate Term 6-2

$$\frac{1}{32g^2}Tr_{SI}\left\{g_5\tau^a(\partial_v\Xi_{a\mu}^\dagger-\partial_\mu\Xi_{av}^\dagger)\left(g_5^2[\tau^c\Xi_c^\mu,\tau^d\Xi_d^v]\right)\right\}$$
$$=\frac{g_5^3}{32g^2}Tr_{SI}\left\{(\partial_v\Xi_{a\mu}^\dagger-\partial_\mu\Xi_{av}^\dagger)(\Xi_c^\mu\Xi_d^v\tau^a\tau^c\tau^d-\Xi_d^v\Xi_c^\mu\tau^a\tau^d\tau^c)\right\}$$
$$=\frac{g_5^3}{32g^2}Tr_I\{\tau^a\tau^c\tau^d\}\epsilon_{\alpha\beta\nu\mu}D^{\dagger\alpha}\chi^b u_{ba}^\beta\epsilon^{\gamma\delta\eta\mu}\epsilon_{\gamma\delta}{}^{\lambda\nu}\chi_m\chi_n u_{\eta c}^m u_{\lambda d}^n$$
$$=-\frac{2g_5^3}{32g^2}Tr_I\{\tau^a\tau^c\tau^d\}D^{\dagger\alpha}\chi^b\chi_m\chi_n u_{ba}^\beta u_{\eta c}^m u_{\lambda d}^n\delta_{ba}\delta_c m\delta_\eta^\beta\epsilon_{\alpha\beta}{}^{\eta\lambda}=0$$

#### 3.5.3.2 Term 4-6 and Hermitian Conjugate Term 6-4

These terms vanish due to a zero spin trace as well as a zero isotopic-spin trace.

$$\begin{aligned}
&\frac{1}{32g^2} Tr_{SI} \left\{ \imath g g_5 [\tau^a \Omega_{av}, \tau^b \Xi^\dagger_{b\mu}] g_5^2 [\tau^c \Xi^\mu_c, \tau^d \Xi^\nu_d] \right\} \\
&= -\frac{g_5^3}{16g} \epsilon^{abe} \Omega_{av} Tr_{SI} \left\{ \tau_e \Xi^\dagger_{b\mu} \left( \tau^c \tau^d \Xi_{c\mu} \Xi_{dv} - \tau^d \tau^c \Xi_{dv} \Xi_{c\mu} \right) \right\} \\
&= -\frac{g_5^3}{16g} \epsilon^{abe} \Omega_{av} Tr_{SI} \left\{ \tau_e \tau^c \tau^d \Xi^\dagger_{b\mu} \Xi_{c\mu} \Xi_{dv} - \tau_e \tau^d \tau^c \Xi^\dagger_{b\mu} \Xi_{dv} \Xi_{c\mu} \right\} = 0
\end{aligned}$$

#### 3.5.3.3 Term 5-6 and Hermitian Conjugate Term 6-5

The same argument as for the previous terms holds.

$$\begin{aligned}
&\frac{1}{32g^2} Tr_{SI} \left\{ \imath g g_5 [\tau^a \Xi^\dagger_{av}, \tau^b \Omega_{b\mu}] g_5^2 [\tau^c \Xi^\mu_c, \tau^d \Xi^\nu_d] \right\} \\
&= -\frac{g_5^3}{16g} \epsilon^{abe} \Omega_{b\mu} Tr_{SI} \left\{ \tau_e \tau^c \tau^d \Xi^\dagger_{av} \Xi^\mu_c \Xi^\nu_d - \tau_e \tau^d \tau^c \Xi^\dagger_{av} \Xi^\nu_d \Xi^\mu_c \right\} = 0
\end{aligned}$$

#### 3.5.3.4 The Terms 1-6 and 6-1 Vanish Because of the Contraction of Space–Time Asymmetry

$$\begin{aligned}
&\frac{1}{32g^2} Tr_{SI} \left\{ \left( \imath g \tau^a (\partial_v \Omega_{a\mu} - \partial_\mu \Omega_{av}) \right) \left( g_5^2 [\tau^c \Xi^\mu_c, \tau^d \Xi^\nu_d] \right) \right\} \\
&= \frac{\imath g_5^2}{32g} (\partial_v \Omega_{a\mu} - \partial_\mu \Omega_{av}) Tr_{SI} \left\{ \tau^a \tau^c \tau^d (\Xi^\mu_c \Xi^\nu_d - \Xi^\nu_c \Xi^\mu_d) \right\} \\
&= -\frac{\imath g}{4} (\partial_v \Omega_{a\mu} - \partial_\mu \Omega_{av}) \delta^{\mu\nu} Tr_I \{ \tau^a \tau^c \tau^d \} \chi_c \chi_d = 0
\end{aligned}$$

Here, the last line only takes into account the contribution for $cd \neq 0$ and the fact that the last term in parenthesis is completely antisymmetric so that the contraction with respect to space–time indices vanishes.

#### 3.5.3.5 Terms 2-4, 2-5, 4-2 and 5-2 Vanish

These terms vanish due to the contraction of the completely antisymmetric terms in both space–time indices as in the previous term.

$$\frac{1}{32g^2}Tr_{SI}\left\{g_5\tau^a(\partial_\nu \Xi^\dagger_{a\mu}-\partial_\mu \Xi^\dagger_{a\nu})\left(-\imath g g_5[\tau^c \Omega^\mu_c,\tau^d \Xi^\nu_d]\right)\right\}$$
$$=\frac{g_5^2}{8g}\epsilon^{cda}Tr_S\{(\partial_\nu \Xi^\dagger_{a\mu}-\partial_\mu \Xi^\dagger_{a\nu})\Xi^\nu_d\}\Omega^\mu_c=0$$

#### 3.5.3.6 The Terms 3-6 and 6-3

These terms vanish because of a vanishing isotopic-spin trace.

$$\frac{1}{32g^2}Tr_{SI}\left\{-g^2[\tau^a\Omega_{a\mu},\tau^b\Omega_{b\nu}]g_5^2[\tau^c\Xi^\mu_c,\tau^d\Xi^\nu_d]\right\}$$
$$=-\frac{\imath g_5^2}{16}\epsilon^{abe}\Omega_{a\mu}\Omega_{b\nu}Tr_{SI}\left\{\tau_e\tau^c\tau^d\Xi^\mu_c\Xi^\nu_d-\tau_e\tau^d\tau^c\Xi^\nu_d\Xi^\mu_c\right\}=0$$

### 3.5.4 Terms n-m ($n \neq m$) Which Contribute to the Dynamics

#### 3.5.4.1 Terms 1-3 and 3-1 Are Equal

$$\frac{1}{32g^2}Tr_{SI}\left\{\left(\imath g\tau^a(\partial_\nu\Omega_{a\mu}-\partial_\mu\Omega_{a\nu})\right)\left(-g^2[\tau^c\Omega^\mu_c,\tau^d\Omega^\nu_d]\right)\right\}$$
$$=-\frac{\imath g}{8}Tr_I\left\{\tau^a[\tau^c,\tau^d]\right\}(\partial_\nu\Omega_{a\mu}-\partial_\mu\Omega_{a\nu})\Omega^\mu_c\Omega^\nu_d$$
$$=\frac{g}{2}\epsilon^{cda}(\partial_\nu\Omega_{a\mu}-\partial_\mu\Omega_{a\nu})\Omega^\mu_c\Omega^\nu_d$$

The isotopic-spin contribution is $Tr_I\left\{\tau^a[\tau^c,\tau^d]\right\}=4\imath\epsilon^{cda}$ For $c,d=0$ the commutator vanishes, for $a=0$ $c=d$ the commutator also vanishes, therefore only different combinations for $acd$ survive. This part contributes in form of a three point interaction part of the vector boson Lagrangian and is a manifestation of non-commutativity.

#### 3.5.4.2 Term 4-5 and 5-4

$$\frac{1}{32g^2}Tr_{SI}\left\{\imath g g_5[\tau^a\Omega_{av},\tau^b\Xi^\dagger_{b\mu}](-\imath g g_5)[\tau^c\Xi^\mu_c,\tau^d\Omega^v_d]\right\}$$
$$=-\frac{g_5^2}{4}\epsilon^{abe}\epsilon^{cdf}\delta_{ef}\Omega_{av}\Omega^v_d Tr_S\left\{\Xi^\dagger_{b\mu}\Xi^\mu_c\right\}$$

These terms may be combined with the terms 4-4 and 5-5 and they will form a term of the type $\Omega_{av}\Omega^v_d\chi^a\chi^d$.

## 3.6 The Resulting Dynamical Model Compatible with Unitary and Chiral Symmetries

One may now collect all the non-vanishing terms from the previous analysis and write the dynamical model in terms of pure fermionic contributions ($\mathscr{L}_\psi$) pure bosonic contributions ($\mathscr{L}_\Omega$ and $\mathscr{L}_\chi$) and the interaction terms between fermions and bosons ($\mathscr{L}_{\psi\Omega}$ and $\mathscr{L}_{\psi\chi}$) and between bosons ($\mathscr{L}_{\Omega\chi}$) as dictated by the initial symmetry hypothesis.

$$\mathscr{L}=\mathscr{L}_\psi+\mathscr{L}_{\psi\Omega}+\mathscr{L}_{\psi\chi}+\mathscr{L}_\Omega+\mathscr{L}_\chi+\mathscr{L}_{\Omega\chi}. \tag{3.2}$$

Here, the explicit expressions for the fermionic sector including the interaction terms are

$$\mathscr{L}_{\psi+\psi\Omega+\psi\chi}=\frac{\imath}{2}\{\bar\Psi\gamma^\mu\partial_\mu\Psi-\partial_\mu\bar\Psi\gamma^\mu\Psi\}+g\bar\Psi\gamma^\mu\tau^a\Omega_{a\mu}\Psi$$
$$+\frac{\imath g}{\sqrt{2}}\bar\Psi\gamma^\mu\gamma_5\tau^a\chi_a\Psi\ .$$

The boson sector that reminds one on the theory of electromagnetism however with a non-linearity is

$$\mathscr{L}_\Omega=\frac{1}{4}\left(\partial_v\Omega^a_\mu-\partial_\mu\Omega^a_v\right)\left(\partial^\mu\Omega^v_a-\partial^v\Omega^\mu_a\right)+g\epsilon^{abc}\left(\partial^v\Omega^\mu_a-\partial^\mu\Omega^v_a\right)\Omega^\mu_b\Omega^v_c$$
$$+g^2\underbrace{(\delta^{ac}\delta^{bd}-\delta^{ad}\delta^{bc})}_{\text{isovector only}}\Omega_{a\mu}\Omega_{bv}\Omega^\mu_c\Omega^v_d$$

and the bosonic sector from the pseudo scalar field is also non-linear with a quartic interaction similar to the Higgs sector in the Standard model.

$$\mathscr{L}_{\chi+\Omega\chi}=\frac{1}{2}\partial_\mu\chi^a\partial^\mu\chi_a-3g^2F_I^{abcd}\chi_a\chi_b\chi_c\chi_d$$
$$+g^2F_I^{abcd}\Omega_{\mu a}\Omega^\mu_b\chi_c\chi_d$$

$$
\begin{aligned}
F_I^{abcd} = {} & \delta^{a0}\delta^{b0}\delta^{c0}\delta^{d0} + \delta^{a0}\delta^{b0}\underbrace{\delta^{cd}}_{cd\neq 0} + \delta^{b0}\delta^{c0}\underbrace{\delta^{ad}}_{ad\neq 0} + \delta^{a0}\delta^{c0}\underbrace{\delta^{bd}}_{bd\neq 0} \\
& + \delta^{b0}\delta^{d0}\underbrace{\delta^{ac}}_{ac\neq 0} + \delta^{a0}\delta^{d0}\underbrace{\delta^{bc}}_{bc\neq 0} + \delta^{c0}\delta^{d0}\underbrace{\delta^{ab}}_{ab\neq 0} \\
& - \underbrace{(\delta^{ac}\delta^{bd} - \delta^{ad}\delta^{bc})}_{abcd\neq 0}\,.
\end{aligned}
$$

All the factors may be identified from the individual terms in the analysis section.

## 3.7 Conclusion

The present discussion showed how starting from an initial symmetry hypothesis one may derive the most general dynamical structure in agreement with underlying symmetries. In our procedure, a combined local unitary and chiral transformation was exploited to progressively construct a model which in the end shows the considered symmetries. Comparing the obtained structure to existing models, one observes the similarity to the Standard Model, and however, in the latter the Higgs sector was constructed so that its contribution could generate the mass terms of the constituting particles by means of a spontaneous symmetry breaking. Differently, the present approach generates the non-linear interaction terms as a symmetry consequence, and one may further think of a spontaneously broken symmetry in order to generate mass terms as a consequence of non-vanishing vacuum expectation values similar to the Higgs reasoning. However, this issue is beyond the considerations of the present work and possible scenarios in this direction will be explored in a future work.

One of the reasons why one should start from symmetry considerations rather than directly proposing a dynamical structure is related to the fact that symmetries are related to conservation laws, and in general it is easier to identify these laws beforehand than looking for transformations that leave the dynamical equations invariant and thus indicate existent symmetries. Furthermore, the possibility to explore the same underlying dynamics with respect to apparent or hidden (spontaneously broken) symmetries may open a pathway to get insights in some of the phenomena in nature which may be related to such mechanisms instead of introducing them into the model *ad hoc* . In the same manner as presented in this work, other symmetry groups may be explored in the same fashion in order to obtain a generic dynamical structure which then relates constituents of a physical system to their possible interactions.

## References

[ArEtAl03] Arodz, H., Dziarmaga, J., Zurek, W.H.: Patterns of Symmetry Breaking. NATO Science Series II: Mathematics, Physics and Chemistry, vol. 127. Springer, Berlin (2003)

[GeLe60] Gell-Mann, M., Levy, M.: The axial vector current in beta decay. Il Nuovo Cimento **16**(4), 705–726 (1960)

[GeLo51] Gell-Mann, M., Low, F.: Bound states in quantum field theory. Phys. Rev. **84**(2), 350–354 (1951)

[Gr09] Gray, C.: Principle of least action. Scholarpedia, **4**(12), 8291 (2009)

[HaEtAl04] Hanca, J., Tulejab, S., Hancova, M.: Symmetries and conservation laws: consequences of Noether's theorem. Am. J. Phys. **72**(4), 428–35 (2004)

[Jo02] Jost, J.: Riemannian Geometry and Geometric Analysis. Springer, Berlin (2002)

[Mo07] Molinari, L.G.: Another proof of Gell-Mann and Low's theorem. J. Math. Phys. **48**(5), 052113 (2007)

[NaJo61] Nambu, Y., Jona-Lasinio, G.: Dynamical model of elementary particles based on an analogy with superconductivity. Phys. Rev. **122**, 345–358 (1961)

[No18] Noether, E.: Invariante variationsprobleme. Nachr. Ges. Wiss. Goettingen **1918**, 235–257 (1918)

[Pu12] Pujol, J.: Hamilton, rodrigues, gauss, quaternions, and rotations: a historical reassessment. Commun. Math. Anal. **13**(2), 1–14 (2012)

[SaNa20] Sakurai, J.J., Napolitano, J.: Modern Quantum Mechanics, 3rd edn. Cambridge University Press, Cambridge (2020)

[TeCh99] Teixeira, F., Chew, W.C.: Differential forms, metrics, and the reflectionless absorption of electromagnetic waves. J. Electromagnet. Waves Appl. **13**(5), 665–686 (1999)

[YaMi54] Yang, C.N., Mills, R.L.: Conservation of isotopic spin and isotopic gauge invariance. Phys. Rev. **96**, 191–195 (1954)

# Chapter 4
# The Traction Boundary Value Problem for Thin Elastic Structures


**C. Constanda and D. Doty**

## 4.1 Introduction

In this chapter, we construct a method for approximating the solution of bending of a load-free, unbounded elastic plate with a hole, under Neumann-type conditions prescribed on the boundary of the hole and a given far-field behavior. The procedure is implemented by means of a generalized Fourier series method that makes use of a complete set of functions spanning the space of the solution. The members of this set are constructed from elements intrinsically tied to the analytic structure of the mathematical model.

Similar problems have been considered for finite plates with Dirichlet, Neumann, and Robin boundary conditions, and for an infinite plate with Dirichlet data on the boundary, in [CoDo17a, CoDo17b, CoDo18, CoDo19a, CoDo19b, CoDo19c, CoDo20].

## 4.2 The Mathematical Model

In the sequel, $S^+$ is a finite domain in $\mathbb{R}^2$ bounded by a simple, closed, $C^2$–curve $\partial S$, $S^- = \mathbb{R}^2 \setminus (S^+ \cup \partial S)$, $x(x_1, x_2)$ and $y(y_1, y_2)$ are generic points in $S^+$, $S^-$, or on $\partial S$, and $|x - y|$ is the distance between $x$ and $y$ in the Cartesian metric. For a matrix $M$, we denote by $M^{(i)}$ and $M_{(i)}$ its columns and rows, and by $M^{\mathsf{T}}$ its transpose. Additionally, $C^{0,\alpha}(\partial S)$ and $C^{1,\alpha}(\partial S)$, $\alpha \in (0, 1)$ are, respectively, the

C. Constanda (✉) · D. Doty
The University of Tulsa, Tulsa, OK, USA
e-mail: christian-constanda@utulsa.edu; dale-doty@utulsa.edu



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_4

spaces of Hölder continuous and Hölder continuously differentiable functions on $\partial S$, and $\langle \cdot\,, \cdot \rangle$ and $\| \cdot \|$ are the inner product and norm on $L_2(\partial S)$.

We assume that the three-dimensional region $(S^- \cup \partial S) \times [-h_0/2, h_0/2]$, where $h_0 = \text{const}$, is occupied by a (homogeneous and isotropic) material with Lamé constants $\lambda$ and $\mu$.

The model of bending of plates with transverse shear deformation consists of the following mathematical elements (see [Co16]):

(i) The displacements are characterized by a vector of the form $u = (u_1, u_2, u_3)^\mathsf{T}$, whose components are functions of $x_1$ and $x_2$.

(ii) The columns $f^{(i)}$ of the matrix

$$f = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -x_1 & -x_2 & 1 \end{pmatrix}$$

form a basis for the space of rigid displacements.

(iii) The system of partial differential equations governing the state of equilibrium when the body forces are negligible is written as

$$A(\partial_1, \partial_2)u(x) = 0,$$

where

$$A(\partial_1, \partial_2) = \begin{pmatrix} h^2\mu\Delta + h^2(\lambda+\mu)\partial_1^2 - \mu & h^2(\lambda+\mu)\partial_1\partial_2 & -\mu\partial_1 \\ h^2(\lambda+\mu)\partial_1\partial_2 & h^2\mu\Delta + h^2(\lambda+\mu)\partial_2^2 - \mu & -\mu\partial_2 \\ \mu\partial_1 & \mu\partial_2 & \mu\Delta \end{pmatrix},$$

$\partial_\alpha = \partial/\partial x_\alpha$, $\alpha = 1, 2$, $D(x, y)$ is a matrix of fundamental solutions, $h^2 = h_0^2/12$, and $\Delta = \partial_1^2 + \partial_2^2$ is the Laplacian.

(iv) An associated matrix of singular solutions that plays an important role in the study of the model is defined by

$$P(x,\ y) = \big(T(\partial_y)D(y,\ x)\big)^\mathsf{T}, \tag{4.1}$$

where

$$T(\partial) = T(\partial_1, \partial_2) = \begin{pmatrix} h^2(\lambda+2\mu)n_1\partial_1 + h^2\mu n_2\partial_2 & h^2\mu n_2\partial_1 + h^2\lambda n_1\partial_2 & 0 \\ h^2\lambda n_2\partial_1 + h^2\mu n_1\partial_2 & h^2\mu n_1\partial_1 + h^2(\lambda+2\mu)n_2\partial_2 & 0 \\ \mu n_1 & \mu n_2 & \mu n_\alpha\partial_\alpha \end{pmatrix}$$

is the boundary moment-force operator, $n(n_1, n_2)$ is the unit normal to $\partial S$ pointing outside $S^-$, and $n_\alpha \partial_\alpha = n_1\partial_1 + n_2\partial_2$.

## 4.3 Exterior Neumann Problem

Let $\mathscr{A}$ be the class of vector functions in $S^-$ with far-field expansion, as $r \to \infty$,

$$\begin{aligned} u_1(r,\,\theta) &= r^{-1}\big[m_0 \sin\theta + 2m_1 \cos\theta - m_0 \sin(3\theta) + (m_2 - m_1)\cos(3\theta)\big] \\ &\quad + r^{-2}\big[(2m_3 + m_4)\sin(2\theta) + m_5 \cos(2\theta) - 2m_3 \sin(4\theta) \\ &\qquad + 2m_6 \cos(4\theta)\big] \\ &\quad + r^{-3}\big[2m_7 \sin(3\theta) + 2m_8 \cos(3\theta) + 3(m_9 - m_7)\sin(5\theta) \\ &\qquad + 3(m_{10} - m_8)\cos(5\theta)\big] + O(r^{-4}), \end{aligned}$$

$$\begin{aligned} u_2(r,\,\theta) &= r^{-1}\big[2m_2 \sin\theta + m_0 \cos\theta + (m_2 - m_1)\sin(3\theta) + m_0 \cos(3\theta)\big] \\ &\quad + r^{-2}\big[(2m_6 + m_5)\sin(2\theta) - m_4 \cos(2\theta) + 2m_6 \sin(4\theta) \\ &\qquad + 2m_3 \cos(4\theta)\big] \\ &\quad + r^{-3}\big[2m_{10} \sin(3\theta) - 2m_9 \cos(3\theta) + 3(m_{10} - m_8)\sin(5\theta) \\ &\qquad + 3(m_7 - m_9)\cos(5\theta)\big] + O(r^{-4}), \end{aligned}$$

$$\begin{aligned} u_3(r,\,\theta) &= -(m_1 + m_2)\ln r - \big[m_1 + m_2 + m_0 \sin(2\theta) + (m_1 - m_2)\cos(2\theta)\big] \\ &\quad + r^{-1}\big[(m_3 + m_4)\sin\theta + (m_5 + m_6)\cos\theta - m_3 \sin(3\theta) \\ &\qquad + m_6 \cos(3\theta)\big] \\ &\quad + r^{-2}\big[m_{11} \sin(2\theta) + m_{12} \cos(2\theta) + (m_9 - m_7)\sin(4\theta) \\ &\qquad + (m_{10} - m_8)\cos(4\theta)\big] + O(r^{-3}), \end{aligned}$$

where $m_0, \ldots, m_{12}$ are constants.

We make the decomposition

$$D = D^{\mathscr{A}} + D^{\infty},$$

where

$$(D^{\mathscr{A}})^{(i)} \in \mathscr{A}, \quad AD^{\mathscr{A}} = AD^{\infty} = 0 \ \text{ in } S^+ \cup S^-.$$

At the same time, for the $P$ matrix we have

$$P^{\infty} = 0, \quad P = P^{\mathscr{A}}, \quad (P^{\mathscr{A}})^{(i)} \in \mathscr{A}.$$

The exterior Neumann problem consists in finding $u \in C^2(S^-) \cap C^1(\bar{S}^-)$ that satisfies

$$\begin{aligned} &Au = 0 \ \text{ in } S^-, \\ &Tu = \mathscr{N} \ \text{ on } \partial S, \\ &u \in \mathscr{A}, \end{aligned} \tag{4.2}$$

where $\mathscr{N}$ is a $3 \times 1$ vector function prescribed on $\partial S$.

**Theorem 4.1** *Problem* (4.2) *has a unique solution u for any* $\mathscr{N} \in C^{0,\alpha}(\partial S)$ *if and only if*

$$\left\langle \mathscr{N}, f^{(i)}\big|_{\partial S} \right\rangle = 0, \quad i = 1, 2, 3. \tag{4.3}$$

Since $u \in \mathscr{A}$, we write $u^{\mathscr{A}}$ instead of $u$, so problem (4.2) becomes

$$\begin{aligned} &Au^{\mathscr{A}} = 0 \ \text{ in } S^-, \\ &Tu^{\mathscr{A}} = \mathscr{N} \ \text{ on } \partial S. \end{aligned} \tag{4.4}$$

As a solution of (4.4), $u^{\mathscr{A}}$ admits the representation formulas

$$u^{\mathscr{A}}(x) = -\int_{\partial S} D(x, y) T u^{\mathscr{A}}(y)\, ds(y) + \int_{\partial S} P(x, y) u^{\mathscr{A}}(y)\, ds(y), \quad x \in S^-,$$

$$0 = -\int_{\partial S} D(x, y) T u^{\mathscr{A}}(y)\, ds(y) + \int_{\partial S} P(x, y) u^{\mathscr{A}}(y)\, ds(y), \quad x \in S^+,$$

which, with $u^{\mathscr{A}}\big|_{\partial S} = \psi$ unknown and $Tu^{\mathscr{A}} = \mathscr{N}$, take the form

$$u^{\mathscr{A}}(x) = -\int_{\partial S} D(x, y) \mathscr{N}(y)\, ds(y) + \int_{\partial S} P(x, y) \psi(y)\, ds(y) \quad x \in S^-, \tag{4.5}$$

$$0 = -\int_{\partial S} D(x, y) \mathscr{N}(y)\, ds(y) + \int_{\partial S} P(x, y) \psi(y)\, ds(y), \quad x \in S^+. \tag{4.6}$$

Let $\partial S_*$ be a simple, closed, $C^2$–curve lying strictly inside $S^+$, let $\{x^{(k)}\}_{k=1}^{\infty}$ be a set of points densely distributed on $\partial S_*$, and consider the vector functions

$$\varphi^{(jk)}(x) = \Big(T\big(D(x, x^{(k)})\big)^{\mathscr{A}}\Big)^{(j)} = T\Big(\big(D(x, x^{(k)})\big)^{\mathscr{A}}\Big)^{(j)}. \tag{4.7}$$

In view of (4.1), these functions have the alternative expression

$$\varphi^{(jk)}(x) = \Big(\big(P(x^{(k)}, x)\big)_{(j)}\Big)^{\mathsf{T}}. \tag{4.8}$$

**Theorem 4.2** *The set*

$$\mathscr{G} = \big\{\varphi^{(jk)},\ j = 1, 2, 3,\ \ k = 1, 2, \dots\big\} \tag{4.9}$$

*is linearly independent on $\partial S$ and complete in $L^2(\partial S)$.*

The elements of $\mathscr{G}$ are ordered as the sequence

$$\varphi^{(11)},\ \varphi^{(21)},\ \varphi^{(31)},\ \varphi^{(12)},\ \varphi^{(22)},\ \varphi^{(32)},\ \dots$$

and re-indexed:

$$\mathscr{G} = \big\{\varphi^{(1)},\ \varphi^{(2)},\ \varphi^{(3)},\ \varphi^{(4)},\ \varphi^{(5)},\ \varphi^{(6)}, \dots\big\},$$

where

$$\varphi^{(i)} = \varphi^{(jk)}, \quad j = 1, 2, 3,\ \ k = 1, 2, \dots,\ \ i = j + 3(k-1) = 1, 2, \dots.$$

Writing the representation formulas in the more compact form

$$u^{\mathscr{A}}(x) = -\big\langle D(x, \cdot), \mathscr{N}\big\rangle + \big\langle P(x, \cdot), \psi\big\rangle, \quad x \in S^-, \tag{4.10}$$

$$\big\langle P(x, \cdot), \psi\big\rangle = \big\langle D(x, \cdot), \mathscr{N}\big\rangle, \quad x \in S^+, \tag{4.11}$$

we intend to approximate $\psi$ (on $\partial S$) from (4.11), and then approximate $u^{\mathscr{A}}$ (in $S^-$) from (4.10).

Using definition (4.1) of $P$ and (4.8), from (4.11) we deduce that

$$\big\langle \varphi^{(i)}, \psi\big\rangle = \Big\langle \Big(\big(D(x^{(k)}, \cdot)\big)_{(j)}\Big)^{\mathsf{T}}, \mathscr{N}\Big\rangle.$$

We consider the (unique) expansion

$$\psi = \sum_{h=1}^{\infty} c_h \varphi^{(h)},$$

which we truncate to get the approximation

$$\psi^{(n)} = \left(u^{\mathscr{A}}\big|_{\partial S}\right)^{(n)} = \sum_{h=1}^{n} c_h \varphi^{(h)}.$$

For symmetry and ease of computation, we use the subsequence of $\psi^{(n)}$ with $n = 3N$, where $N$ is the number of points $x^{(k)}$ selected on $\partial S_*$. This leads to a nonsingular linear algebraic system for computing the coefficients $c_h$, namely

$$\sum_{h=1}^{n} c_h \langle \varphi^{(i)}, \varphi^{(h)} \rangle = \left\langle \left( \left( (D(x^{(k)}, x))^{\mathscr{A}} \right)_{(j)} \right)^{\mathsf{T}}, \mathscr{N} \right\rangle,$$
$$i = j + 3(k-1) = 1, 2, \ldots n.$$

Finally, we construct the function

$$\begin{aligned} u^{(n)}(x) &= (u^{\mathscr{A}})^{(n)}(x) \\ &= -\left\langle \left( D(x, \cdot) \right)^{\mathscr{A}}, \mathscr{N} \right\rangle + \left\langle P(x, \cdot), \psi^{(n)} \right\rangle, \quad x \in S^-. \end{aligned}$$

**Theorem 4.3** *The vector function $u^{(n)}$ is an approximation of the solution $u$ of problem (4.2) in the sense that $u^{(n)} \to u$ uniformly on any closed and bounded subdomain of $S^-$.*

## 4.4 First Numerical Example

Let $S^+$ be the disk of radius 1 centered at origin, and let the plate parameters (after rescaling and non-dimensionalization) be $h = 0.5$ and $\lambda = \mu = 1$.

We consider the boundary condition function (in polar coordinates on $\partial S$)

$$\mathscr{N}(x) = \begin{pmatrix} -\frac{3}{2}\cos\theta + 2\sin\theta - 8\cos(2\theta) - \sin(2\theta) + 3\cos(3\theta) \\ \quad - 3\sin(3\theta) + 3\cos(4\theta) + \sin(4\theta) - 3\cos(5\theta) + 6\sin(5\theta) \\ 2\cos\theta - \frac{3}{2}\sin\theta - 3\cos(2\theta) + 4\sin(2\theta) - 9\cos(3\theta) \\ \quad - 3\sin(3\theta) - \cos(4\theta) + 3\sin(4\theta) - 6\cos(5\theta) - 3\sin(5\theta) \\ -12\sin(2\theta) + 54\cos(3\theta) + 18\sin(3\theta) - 36\cos(4\theta) \\ \qquad\qquad + 72\sin(4\theta) \end{pmatrix}.$$

Direct verification shows that this function satisfies the solvability condition (4.3), that the exact solution of problem (4.2) generated by it is

$$u(x) = \begin{pmatrix} (x_1^2+x_2^2)^{-1}[-3x_1+2x_2] \\ \quad +(x_1^2+x_2^2)^{-2}[-5x_1^2+5x_2^2-6x_1^2x_2+2x_2^3] \\ \quad +(x_1^2+x_2^2)^{-3}[x_1^3-3x_1x_2^2+6x_1^4+8x_1^3x_2-36x_1^2x_2^2-8x_1x_2^3+6x_2^4] \\ \quad +(x_1^2+x_2^2)^{-4}[-3x_1^5+30x_1^4x_2+30x_1^3x_2^2-60x_1^2x_2^3-15x_1x_2^4+6x_2^5] \\ (x_1^2+x_2^2)^{-1}[2x_1-3x_2] \\ \quad +(x_1^2+x_2^2)^{-2}[-2x_1^2+2x_1x_2+2x_2^2+2x_1^3-6x_1x_2^2] \\ \quad +(x_1^2+x_2^2)^{-3}[-4x_1^3-3x_1^2x_2+12x_1x_2^2+x_2^3-2x_1^4+24x_1^3x_2 \\ \qquad\qquad +12x_1^2x_2^2-24x_1x_2^3-2x_2^4] \\ \quad +(x_1^2+x_2^2)^{-4}[-6x_1^5-15x_1^4x_2+60x_1^3x_2^2+30x_1^2x_2^3-30x_1x_2^4-3x_2^5] \\ 3+\frac{3}{2}\ln(x_1^2+x_2^2) \\ \quad +(x_1^2+x_2^2)^{-1}[-2x_1+x_2-4x_1x_2] \\ \quad +(x_1^2+x_2^2)^{-2}[-10x_1x_2+3x_1^3+3x_1^2x_2-9x_1x_2^2-x_2^3] \\ \quad +(x_1^2+x_2^2)^{-3}[18x_1^3+18x_1^2x_2-54x_1x_2^2-6x_2^3-x_1^4+8x_1^3x_2 \\ \qquad\qquad +6x_1^2x_2^2-8x_1x_2^3-x_2^4] \\ \quad +(x_1^2+x_2^2)^{-4}[-9x_1^4+72x_1^3x_2+54x_1^2x_2^2-72x_1x_2^3-9x_2^4] \end{pmatrix},$$

and that the class $\mathscr{A}$ coefficients of this solution are

$$m_0=2,\ \ m_1=-\frac{3}{2},\ \ m_2=-\frac{3}{2},\ \ m_3=-1,\ \ m_4=2,\ \ m_5=-5,\ \ m_6=3,$$
$$m_7=0,\ \ m_8=\frac{1}{2},\ \ m_9=2,\ \ m_{10}=-\frac{1}{2},\ \ m_{11}=-5,\ \ m_{12}=0.$$

We take the auxiliary curve $\partial S_*$ to be the circle of radius 1/2 centered at the origin. This seems a reasonable choice since having $\partial S_*$ too far from $\partial S$ makes the set $\mathscr{G}$ "less linearly independent," whereas positioning it too close to $\partial S$ increases the sensitivity of $\mathscr{G}$ to the singularities of matrices $D$ and $P$.

It is obvious that the accuracy of the approximation depends on the selection of the set of points $\{x^{(k)}\}$ on $\partial S_*$. For the sake of symmetry, we make a uniformly distributed choice; specifically, for $N=1,2,\ldots,$

$$\{x^{(k)}\}_{k=1}^{N}\Big|_{\text{Cartesian}} = \left\{\left(\frac{1}{2},\frac{2\pi k}{N}\right)\right\}_{k=1}^{N}\Bigg|_{\text{Polar}}.$$

The numerical computation method used in this example is row reduction.

## 4.5 Graphical Illustrations I

The graphs of the three components of $(u^{\mathscr{A}})^{(60)}$ computed from $\psi^{(60)}$ (with $N = 20$ points $x^{(k)}$ on $\partial S_*$) for

$$r \geq 1.01, \quad 0 \leq \theta < 2\pi,$$

together with graph of $\psi^{(60)}$, are shown in Fig. 4.1.

The influence of the singularities of $D(x, y)$ and $P(x, y)$ for $x \in S^-$ very close to $y \in \partial S$ is mitigated by increasing the floating-point accuracy in the vicinity of $\partial S$ but is never completely eliminated. The gap between the computed subdomain and $\partial S$ is filled by appropriate interpolation.

The graphs of the components of $(u^{\mathscr{A}})^{(60)}$ constructed from $\psi^{(60)}$ for

$$1.01 < r \leq 100, \quad 0 \leq \theta < 2\pi,$$

which illustrate the class $\mathscr{A}$ behavior of the solution away from the boundary, are displayed in Fig. 4.2.

Figure 4.3 contains the graphs of the components of the error $(u^{\mathscr{A}})^{(60)} - u^{\mathscr{A}}$. The approximation is 4–5 digits of accuracy near $\partial S$ but improves significantly away from the boundary.

The behavior of the relative error

$$\frac{\|\psi^{(3N)} - u^{\mathscr{A}}|_{\partial S}\|}{\|u^{\mathscr{A}}|_{\partial S}\|}$$

as a function of $N$ reflects the efficiency and accuracy of our computational procedure. The logarithmic plot of the size of the relative error in terms of $N$ can be seen in Fig. 4.4. This plot strongly suggests that the relative error improves exponentially as $N$ increases. Fitting a linear curve to the logarithmic data produces the model

$$3.38762 - 0.295589\, N.$$

The relative error may be modeled by

$$\frac{\|\psi^{(3N)} - u^{\mathscr{A}}|_{\partial S}\|}{\|u^{\mathscr{A}}|_{\partial S}\|} = 2441.3 \times 0.506304^N. \tag{4.12}$$

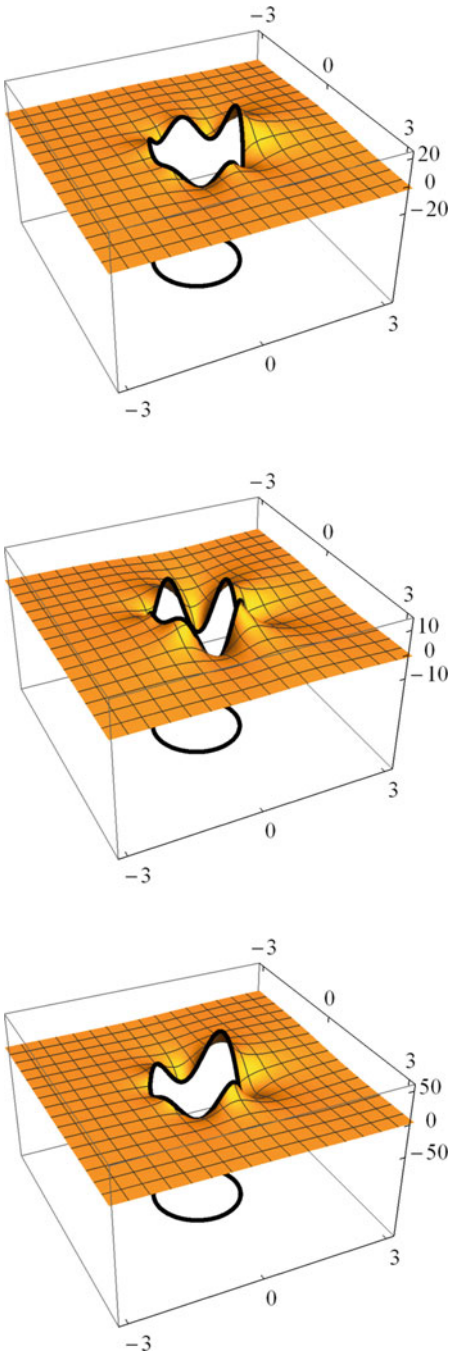



**Fig. 4.1** Graphs of the components of $(u^{\mathscr{A}})^{(60)}$ and $\psi^{(60)}$ for $r \geq 1.01$, $0 \leq \theta < 2\pi$

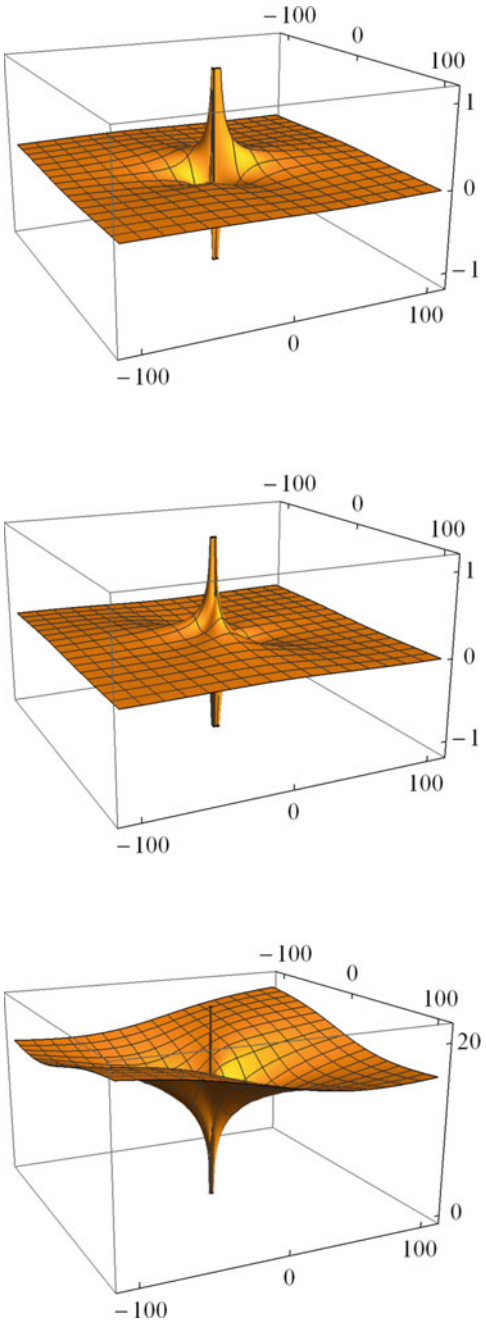



**Fig. 4.2** Graphs of the components of $(u^{\mathscr{A}})^{(60)}$ for $1.01 < r < 100,\ 0 \le \theta < 2\pi$

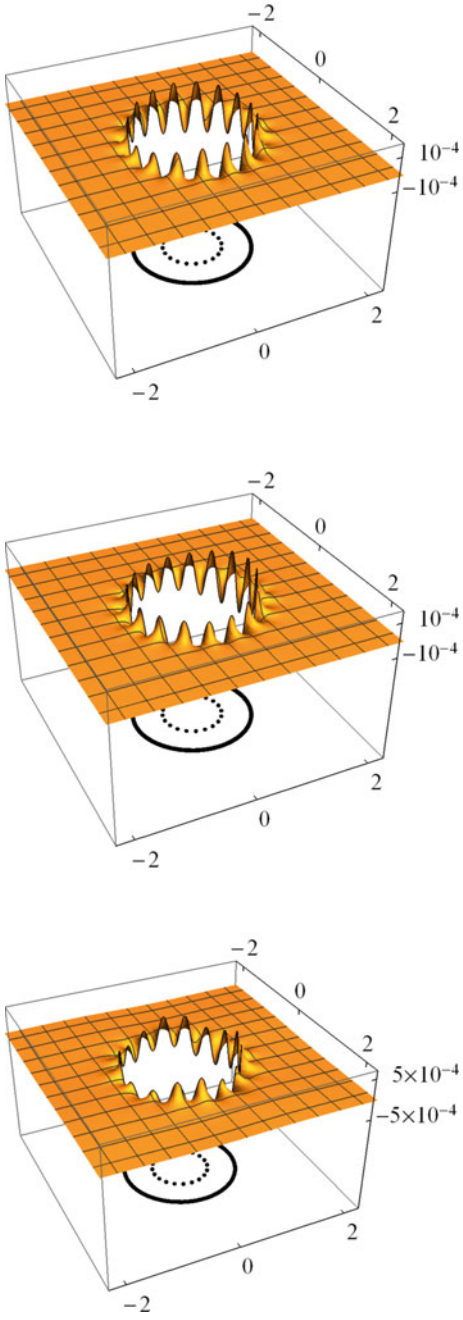



**Fig. 4.3** Graphs of the components of the error $(u^{\mathscr{A}})^{(60)} - u^{\mathscr{A}}$

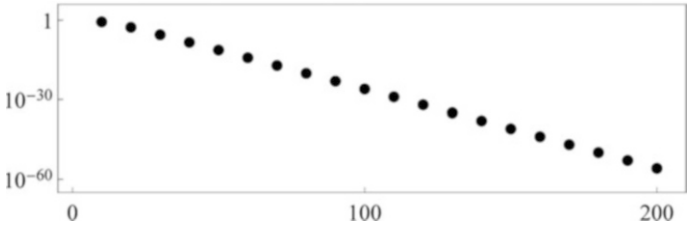



**Fig. 4.4** Logarithmic plot of the relative error as a function of $N$

## 4.6 Second Numerical Example

The first example, where the exact solution of the problem was known, validated the efficiency of our approximation method. We now solve the boundary value problem (4.2) with a data function $\mathscr{N}$ for which the exact solution is not known. Specifically, for the same domain $S^-$ we choose

$$\mathscr{N}(x) = \begin{pmatrix} 3\cos\theta - \sin\theta + 6\cos(2\theta) - 7\sin(2\theta) + 3\cos(3\theta) + 15\sin(3\theta) \\ \qquad - 3\cos(4\theta) + 2\sin(4\theta) - 6\cos(5\theta) - 9\sin(5\theta) \\ -\cos\theta + \sin\theta - \cos(2\theta) - 6\sin(2\theta) + 3\cos(3\theta) - 9\sin(3\theta) \\ \qquad - 2\cos(4\theta) - 3\sin(4\theta) + 9\cos(5\theta) - 6\sin(5\theta) \\ -6\cos(2\theta) - 6\sin(2\theta) - 54\cos(3\theta) + 36\sin(3\theta) - 72\cos(4\theta) \\ \qquad - 108\sin(4\theta) \end{pmatrix}.$$

The approximation is computed with the same auxiliary curve $\partial S_*$ and points $x^{(k)}$, and the same parameters as in Sect. 4.5, by means of the row reduction method.

## 4.7 Graphical Illustrations II

The graphs of the components of $u^{(75)}$, generated with 25 points $x^{(k)}$ on $\partial S_*$, for

$$r \geq 1.01, \quad 0 \leq \theta < 2\pi,$$

are shown in Fig. 4.5.

The graphs of the components of $u^{(75)}$ for

$$1.01 \leq r \leq 100, \quad 0 \leq \theta < 2\pi,$$

which indicate the class $\mathscr{A}$ behavior of the solution away from the origin, are displayed in Fig. 4.6.

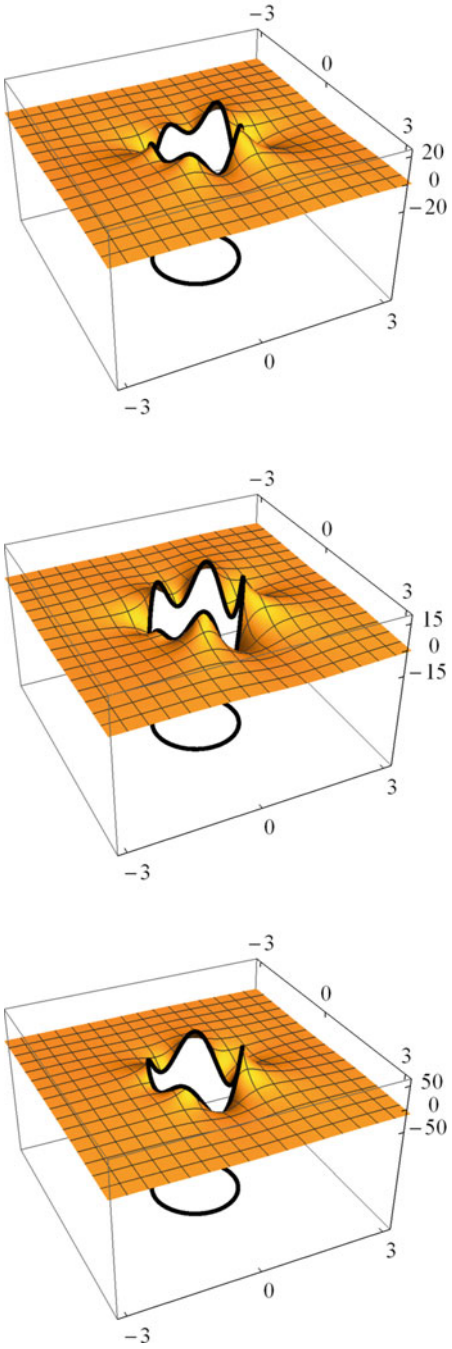



**Fig. 4.5** Graphs of the components of $(u^{\mathscr{A}})^{(60)}$ and $\psi^{(60)}$ for $r \geq 1.01$, $0 \leq \theta < 2\pi$

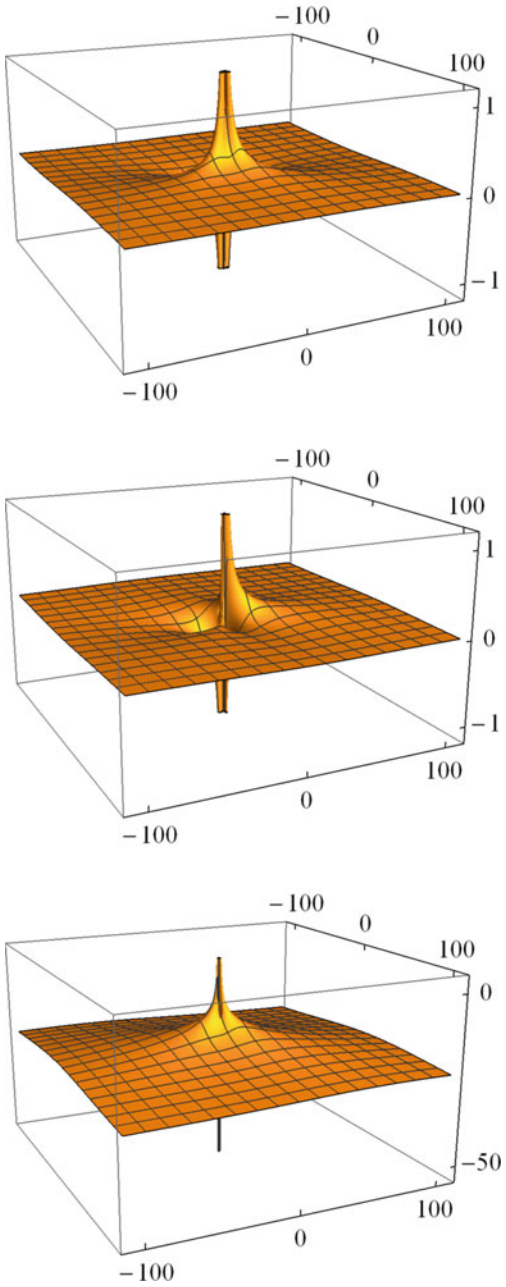



**Fig. 4.6** Graphs of the components of $(u^{\mathscr{A}})^{(75)}$ for $1.01 < r < 100,\ 0 \le \theta < 2\pi$

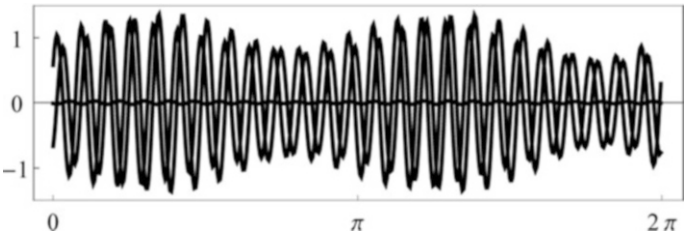



**Fig. 4.7** Components of $(T\tilde{u})^{(75)} - \mathscr{N}$ in polar coordinates

Since $u^{\mathscr{A}}\big|_{\partial S}$ is not known in this case, we cannot use (4.12) to estimate the error. Instead, we design a roundabout procedure that uses $\psi^{(75)}$ to construct the solution $\tilde{u}$ of an exterior Dirichlet problem, then compute $(T\tilde{u})^{(75)}$ by the method described in [CoDo20]. As can be seen in Fig. 4.7, $(T\tilde{u})^{(75)}$ is close to $\mathscr{N}$, which confirms that our technique is efficient.

A far smaller relative error is obtained if we take 200 points $x^{(k)}$ on $\partial S_*$; then

$$\frac{\|(T\tilde{u})^{(600)} - \mathscr{N}\|}{\|\mathscr{N}\|} = 1.41203 \times 10^{-52}.$$

## References

[Co16] Constanda, C.: Mathematical Methods for Elastic Plates. Springer, London (2016)

[CoDo17a] Constanda, C., Doty, D.: Bending of elastic plates: Generalized Fourier series method. In: Integral Methods in Science and Engineering: Theoretical Techniques, pp. 71–81. Birkhäuser, New York (2017)

[CoDo17b] Constanda, C., Doty, D.: The Neumann problem for bending of elastic plates. In: Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering CMMSE 2017, vol. II, pp. 619–622. Cádiz, Spain (2017)

[CoDo18] Constanda, C., Doty, D.: Bending of elastic plates with transverse shear deformation: the Neumann problem. Math. Methods Appl. Sci. **41**, 7130–7140 (2018). https://doi.org/10.1002/mma.4704

[CoDo19a] Constanda, C., Doty, D.: The Robin problem for bending of elastic plates. Math. Methods Appl. Sci. **42**, 5639–5648 (2019). https://doi.org/10.1002/mma.5286

[CoDo19b] Constanda, C., Doty, D.: Bending of plates with transverse shear deformation: The Robin problem. Comput. Math. Methods **1**, e1015 (2019). https://doi.org/10.1002/cmm4.1015

[CoDo19c] Constanda, C., Doty, D.: Bending of elastic plates: Generalized Fourier series method for the Robin problem. In: Integral Methods in Science and Engineering: Analytic Treatment and Numerical Approximations, pp. 97–110. Birkhäuser, New York (2019)

[CoDo20] Constanda, C., Doty, D.: Analytic and numerical solutions in the theory of elastic plates. Complex Var. Elliptic Equ. **65**, 40–56 (2020). https://doi.org/10.1080/17476933.2019.1636789

# Chapter 5
# Mapping Properties of Potential Operators Related to the 2D Compressible Stokes System in Weighted Sobolev Spaces


**M. A. Dagnaw and C. Fresneda-Portillo**

## 5.1 Introduction

The purpose of this chapter is providing the mapping properties on weighted Sobolev spaces of those boundary and domain integral operators that appear in the integral representation formula of the pressure and velocity solutions of the compressible Stokes system. These are required to prove further theorems related to boundary-domain integral equations in 2D. Let us remark that the results presented here build on the works for the compressible Stokes system in 2D [AyDa20] and the works on boundary-domain integral equations for the Stokes system in weighted Sobolev spaces in 3D [FrMi21].

## 5.2 Preliminaries

Let $\Omega := \Omega^+$ be a unbounded (exterior) simply connected domain in $\mathbb{R}^2$ and let $\Omega^- := \mathbb{R}^2 \setminus \overline{\Omega^+}$ be the complementary (bounded) subset of $\Omega$. The boundary $\partial\Omega$ is simply connected, closed, and $\mathscr{C}^2$–smooth for simplicity.

In what follows, $H^s(\Omega)$, $H^s(\partial\Omega)$ are the Bessel potential spaces, where $s \in \mathbb{R}$ is an arbitrary real number (see, e.g., [Mc00]). We recall that $H^s$ coincide with the Sobolev-Slobodetsky spaces $W_2^s$ for any non-negative $s$. For an open set $\Omega'$, we, as usual, denote $\mathscr{D}(\Omega') = C_0^\infty(\Omega')$, while $\mathscr{D}(\overline{\Omega'})$ is the restriction to $\overline{\Omega'}$ of the

M. A. Dagnaw
Injibara University, Injibara, Ethiopia

C. Fresneda-Portillo (✉)
Universidad Loyola Andalucía, Seville, Spain
e-mail: cfresneda@uloyola.es



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_5

space $\mathscr{D}(\mathbb{R}^2)$. In what follows we use the bold notation: $\boldsymbol{H}^s(\Omega) = [H^s(\Omega)]^2$ for 2-dimensional vector spaces.

We denote by $\widetilde{\boldsymbol{H}}^s(\Omega)$ the subspace of $\boldsymbol{H}^s(\mathbb{R}^2)$ defined as $\widetilde{\boldsymbol{H}}^s(\Omega) := \{\boldsymbol{g} : \boldsymbol{g} \in \boldsymbol{H}^s(\mathbb{R}^2), \operatorname{supp} \boldsymbol{g} \subset \overline{\Omega}\}$; similarly, $\widetilde{\boldsymbol{H}}^s(S_1) = \{\boldsymbol{g} \in \boldsymbol{H}^s(\partial\Omega), \operatorname{supp} \boldsymbol{g} \subset \overline{S_1}\}$ is the Sobolev space of functions having support in $S_1 \subset \partial\Omega$. We will use the following notation for derivative operators: $\partial_j = \partial_{x_j} := \frac{\partial}{\partial x_j}$ with $j = 1, 2$; $\nabla := (\partial_1, \partial_2)$.

Furthermore, to ensure unique solvability of the BVPs in exterior domains, we will need the *weighted Sobolev spaces*, see, e.g., [Ha71, AlAm00]. Let us first introduce the weighted Lebesgue space

$$L_2(\rho^{-1}; \Omega) = \{g : \rho^{-1} g \in L_2(\Omega)\},$$

where

$$\rho(\boldsymbol{x}) = (1 + |\mathbf{x}|^2)^{1/2} \ln(2 + |\mathbf{x}|^2).$$

Let $\mathscr{H}^1(\Omega)$ denote the following weighted Sobolev (Beppo-Levi) space

$$\mathscr{H}^1(\Omega) := \{g \in L_2(\rho^{-1}; \Omega) : \nabla g \in \boldsymbol{L}_2(\Omega)\}$$

endowed with the corresponding norm

$$\|g\|^2_{\mathscr{H}^1(\Omega)} := \|\rho^{-1} g\|^2_{L_2(\Omega)} + \|\nabla g\|^2_{L_2(\Omega)}.$$

The analogous vector counterpart of $\mathscr{H}^1(\Omega)$ reads

$$\mathscr{H}^1(\Omega) := \{\boldsymbol{g} \in \boldsymbol{L}_2(\rho^{-1}; \Omega) : \operatorname{grad} \boldsymbol{g} \in L_2(\Omega)^{3\times 3}\}.$$

It is well known that $\mathscr{D}(\overline{\Omega})$ is dense in $\mathscr{H}^1(\Omega)$, see, e.g., [Ha71]. If $\Omega$ is unbounded, then the seminorm

$$|\boldsymbol{g}|_{\mathscr{H}^1(\Omega)} := \|\nabla \boldsymbol{g}\|_{\boldsymbol{L}_2(\Omega)}$$

is equivalent to the norm $\|\boldsymbol{g}\|_{\mathscr{H}^1(\Omega)}$ in $\mathscr{H}^1(\Omega)$ [Li73, Chapter XI, Part B, §1]. If $\Omega^-$ is bounded, then $\mathscr{H}^1(\Omega^-) = \boldsymbol{H}^1(\Omega^-)$. If $\Omega'$ is a bounded subdomain of an unbounded domain $\Omega$ and $\boldsymbol{g} \in \mathscr{H}^1(\Omega)$, then $\boldsymbol{g} \in \boldsymbol{H}^1(\Omega')$.

Let $\widetilde{\mathscr{H}}^1(\Omega)$ be the completion of $\mathscr{D}(\Omega)$ in $\mathscr{H}^1(\mathbb{R}^2)$; it can be also characterised as $\widetilde{\mathscr{H}}^1(\Omega) = \{\boldsymbol{g} : \boldsymbol{g} \in \mathscr{H}^1(\mathbb{R}^2), \operatorname{supp} \boldsymbol{g} \subset \overline{\Omega}\}$. Let $\widetilde{\mathscr{H}}^{-1}(\Omega) := [\mathscr{H}^1(\Omega)]^*$ and $\mathscr{H}^{-1}(\Omega) := [\widetilde{\mathscr{H}}^1(\Omega)]^*$ be the corresponding dual spaces. Evidently, the space $\boldsymbol{L}_2(\rho; \Omega) \subset \mathscr{H}^{-1}(\Omega)$. Let us also consider the spaces $L^2_*(\Omega) = L^2(\Omega)/\mathbb{R} = \{q \in L^2(\Omega) : \int_\Omega q\, dx = 0\}$ and $\mathbf{H}^{-\frac{1}{2}}_{**}(\partial\Omega) := \{\boldsymbol{\rho} \in \mathbf{H}^s(\partial\Omega) : \langle \rho_i, 1\rangle_{\partial\Omega} = 0 \quad \text{for} \quad i = 1, 2\}$.

For any distribution $\boldsymbol{g}$ in $\widetilde{\mathscr{H}}^{-1}(\Omega)$, we have the following representation property (see [Ne01, Section 2.5]), $g_j = \sum_{i=1}^{2} \partial_i g_{ij} + g_j^0$, $g_{ij} \in L_2(\mathbb{R}^2)$, $g_j^0 \in L_2(\rho; \mathbb{R}^2)$ and $g_{ij}, g_j^0 = 0$ outside the domain $\Omega$, $i, j \in \{1, 2\}$. Consequently, $\mathscr{D}(\Omega)$ is dense in $\widetilde{\mathscr{H}}^{-1}(\Omega)$ and $\mathscr{D}(\mathbb{R}^2)$ is dense in $\mathscr{H}^{-1}(\mathbb{R}^2)$.

Let $\mu$ be the viscosity coefficient, $p$ the pressure field and $\boldsymbol{v}$ the velocity field. In this chapter, for an arbitrary couple $(p, \boldsymbol{v})$, the stress tensor operator, $\sigma_{ij}$, and the Stokes operator, $\mathscr{A}_j$, are defined for a compressible fluid as

$$\sigma_{ji}(p, \boldsymbol{v})(\boldsymbol{x}) := -\delta_i^j\, p(\boldsymbol{x}) + \mu(\boldsymbol{x}) \left( \frac{\partial v_i(\boldsymbol{x})}{\partial x_j} \frac{\partial v_j(\boldsymbol{x})}{\partial x_i} - \alpha \delta_i^j \operatorname{div} \boldsymbol{v}(\boldsymbol{x}) \right), \tag{5.1}$$

$$\begin{aligned} \mathscr{A}_j(p, \boldsymbol{v})(\boldsymbol{x}) :=& \frac{\partial}{\partial x_i} \sigma_{ji}(p, \boldsymbol{v})(\boldsymbol{x}) \\ =& \frac{\partial}{\partial x_i} \left( \mu(\boldsymbol{x}) \left( \frac{\partial v_j(\boldsymbol{x})}{\partial x_i} + \frac{\partial v_i(\boldsymbol{x})}{\partial x_j} - \alpha \delta_i^j \operatorname{div} \boldsymbol{v}(\boldsymbol{x}) \right) \right) - \frac{\partial p(\boldsymbol{x})}{\partial x_j}, \; j, i \in \{1, 2\}, \end{aligned} \tag{5.2}$$

where $\alpha = 1$ or $\alpha = \frac{2}{3}$ and $\delta_i^j$ is Kronecker symbol. Henceforth we assume the Einstein summation in repeated indices from 1 to 2 if not stated otherwise.

Throughout this chapter, we will assume the following condition to ensure boundedness properties of the integral operators introduced further on.

**Condition 5.1**

$$\mu \in \mathscr{C}^1(\mathbb{R}^2) \cap L_\infty(\mathbb{R}^2) \; : \; \rho \nabla \mu \in \boldsymbol{L}_\infty(\mathbb{R}^2).$$

*In addition, there exist constants $C_1$ and $C_2$ such that*

$$0 < C_1 < \mu(\boldsymbol{x}) < C_2. \tag{5.3}$$

The operator $\mathscr{A}$ acting on $(p, \boldsymbol{v}) \in L_2(\Omega) \times \mathscr{H}^1(\Omega)$ is well defined in the weak sense as long as the variable coefficient $\mu(\boldsymbol{x})$ is essentially bounded, i.e., $\mu \in L_\infty(\Omega)$. Indeed, in the sense of distributions the operator $\mathscr{A}$ is defined as

$$\langle \mathscr{A}(p, \boldsymbol{v}), \boldsymbol{u} \rangle_\Omega = -\mathscr{E}((p, \boldsymbol{v}), \boldsymbol{u}), \quad \forall \boldsymbol{u} \in \mathscr{D}(\Omega), \tag{5.4}$$

where

$$\mathscr{E}((p, \boldsymbol{v}), \boldsymbol{u}) := \int_\Omega E((p, \boldsymbol{v}), \boldsymbol{u})(\boldsymbol{x})\, dx, \tag{5.5}$$

and the function $E((p, \boldsymbol{v}), \boldsymbol{u})$ is defined as

$$E((p, \boldsymbol{v}), \boldsymbol{u})(\boldsymbol{x}) := \frac{1}{2}\mu(\boldsymbol{x})\left(\frac{\partial u_i(\boldsymbol{x})}{\partial x_j} + \frac{\partial u_j(\boldsymbol{x})}{\partial x_i}\right)\left(\frac{\partial v_i(\boldsymbol{x})}{\partial x_j} + \frac{\partial v_j(\boldsymbol{x})}{\partial x_i}\right) - \alpha\mu(\boldsymbol{x})\mathrm{div}\boldsymbol{v}(\boldsymbol{x})\,\mathrm{div}\boldsymbol{u}(\boldsymbol{x}) - p(\boldsymbol{x})\mathrm{div}\boldsymbol{u}(\boldsymbol{x}). \tag{5.6}$$

The bilinear form $\mathscr{E} : \left[L_2(\Omega) \times \mathscr{H}^1(\Omega)\right] \times \widetilde{\mathscr{H}}^1(\Omega) \to \mathbb{R}$ is evidently bounded. Thus, by the density of $\mathscr{D}(\Omega)$ in $\widetilde{\mathscr{H}}^1(\Omega)$, the operator

$$\mathscr{A} : L_2(\Omega) \times \mathscr{H}^1(\Omega) \to \mathscr{H}^{-1}(\Omega)$$

defined by (5.4) for any $\boldsymbol{u} \in \widetilde{\mathscr{H}}^1(\Omega)$ is also bounded and gives the weak form of operator (5.2).

We will also make use of the following space, (cf., e.g., [FrMi21]),

$$\mathscr{H}^{1,0}(\Omega; \mathscr{A}) := \{(p, \boldsymbol{v}) \in L_2(\Omega) \times \mathscr{H}^1(\Omega) : \mathscr{A}(p, \boldsymbol{v}) \in \boldsymbol{L}_2(\rho; \Omega)\},$$

endowed with the norm, $\|\cdot\|_{\mathscr{H}^{1,0}(\Omega;\mathscr{A})}$, where

$$\|(p, \boldsymbol{v})\|_{\mathscr{H}^{1,0}(\Omega;\mathscr{A})} := \left(\|p\|^2_{L_2(\Omega)} + \|\boldsymbol{v}\|^2_{\mathscr{H}^1(\Omega)} + \|\rho\mathscr{A}(p, \boldsymbol{v})\|^2_{\boldsymbol{L}_2(\Omega)}\right)^{1/2}.$$

Let us define also a space

$$\mathscr{H}^{1,0}_*(\Omega; \mathscr{A}) := \{(p, \boldsymbol{v}) \in L^2_*(\Omega) \times \mathscr{H}^1(\Omega) : \mathscr{A}(p, \boldsymbol{v}) \in \boldsymbol{L}_2(\rho; \Omega)\},$$

with the norm

$$\|(p, \boldsymbol{v})\|_{\mathscr{H}^{1,0}_*(\Omega;\mathscr{A})} := \left(\|p\|^2_{L^2_*(\Omega)} + \|\boldsymbol{v}\|^2_{\mathscr{H}^1(\Omega)} + \|\rho\mathscr{A}(p, \boldsymbol{v})\|^2_{\boldsymbol{L}_2(\Omega)}\right)^{1/2}.$$

Similar to [Mi11, Theorem 3.12], one can prove the following assertion.

**Theorem 5.2** *Let $\mu$ satisfy condition 5.1. Then the space $\mathscr{D}(\overline{\Omega}) \times \mathscr{D}(\overline{\Omega})$ is dense in $\mathscr{H}^{1,0}(\Omega; \mathscr{A})$.*

For sufficiently smooth functions $(p, \boldsymbol{v}) \in H^{s-1}(\Omega^\pm) \times \boldsymbol{H}^s(\Omega^\pm)$ with $s > 3/2$, we can define the classical traction (conormal derivative) operators, $\boldsymbol{T}^{c\pm} = \{T_i^{c\pm}\}_{i=1}^3$, on the boundary $\partial\Omega$ as

$$T_i^{c\pm}(p, \boldsymbol{v})(\boldsymbol{x}) := [\gamma^\pm\sigma_{ij}(p, \boldsymbol{v})(\boldsymbol{x})]n_j(\boldsymbol{x})$$
$$= -n_i(\boldsymbol{x})\gamma^\pm p(\boldsymbol{x}) + n_j(\boldsymbol{x})\mu(\boldsymbol{x})\gamma^\pm\left(\frac{\partial v_i(\boldsymbol{x})}{\partial x_j} + \frac{\partial v_j(\boldsymbol{x})}{\partial x_i} - \alpha\delta_i^j\mathrm{div}\boldsymbol{v}(\boldsymbol{x})\right), \ \boldsymbol{x} \in \partial\Omega, \tag{5.7}$$

where $n_j(\boldsymbol{x})$ denote the components of the unit normal vector $\boldsymbol{n}(\boldsymbol{x})$ to the boundary $\partial\Omega$ *directed outwards the exterior domain* $\Omega$. Moreover, $\boldsymbol{\gamma}^{\pm}$ denote the trace operators from inside and outside $\Omega$ which according to the trace theorem satisfy the mapping property $\gamma^{\pm} : \mathscr{H}^1(\Omega) \to \boldsymbol{H}^{1/2}(\partial\Omega)$.

Traction operators (5.7) can be continuously extended to the *canonical* traction operators $\boldsymbol{T}^{\pm} : \mathscr{H}^{1,0}(\Omega^{\pm}, \mathscr{A}) \to \boldsymbol{H}^{-1/2}(\partial\Omega)$ defined in the weak form (cf. [CMN13, FrMi21]), as

$$\langle \boldsymbol{T}^{+}(p, \boldsymbol{v}), \boldsymbol{w} \rangle_{\partial\Omega} := \int_{\Omega^{\pm}} \left[ \mathscr{A}(p, \boldsymbol{v}) \boldsymbol{\gamma}^{+}_{-1} \boldsymbol{w} + E\left((p, \boldsymbol{v}), \boldsymbol{\gamma}^{+}_{-1} \boldsymbol{w}\right) \right] dx$$
$$\forall\, (p, \boldsymbol{v}) \in \mathscr{H}^{1,0}(\Omega^{\pm}, \mathscr{A}),\ \forall\, \boldsymbol{w} \in \boldsymbol{H}^{1/2}(\partial\Omega),$$

where the operator $\boldsymbol{\gamma}^{+}_{-1} : \boldsymbol{H}^{1/2}(\partial\Omega) \to \mathscr{H}^1(\Omega)$ denotes a continuous right inverse of the trace operator $\boldsymbol{\gamma}^{+} : \mathscr{H}^1(\Omega) \to \boldsymbol{H}^{1/2}(\partial\Omega)$.

Furthermore, if $(p, \boldsymbol{v}) \in \mathscr{H}^{1,0}(\Omega, \mathscr{A})$ and $\boldsymbol{u} \in \mathscr{H}^1(\Omega)$, the following first Green identity holds, similar as in [FrMi21] for the 3D case,

$$\langle \boldsymbol{T}^{+}(p, \boldsymbol{v}), \boldsymbol{\gamma}^{+}\boldsymbol{u} \rangle_{\partial\Omega} = \int_{\Omega} [\mathscr{A}(p, \boldsymbol{v})\boldsymbol{u} + E\left((p, \boldsymbol{v}), \boldsymbol{u}\right)(\boldsymbol{x})] dx. \tag{5.8}$$

Applying identity (5.8) to the pairs $(p, \boldsymbol{v}), (q, \boldsymbol{u}) \in \mathscr{H}^{1,0}(\Omega, \mathscr{A})$ with exchanged roles and subtracting the one from the other, we arrive at the second Green identity,

$$\begin{aligned} &\langle \boldsymbol{T}^{+}(p, \boldsymbol{v}), \boldsymbol{\gamma}^{+}\boldsymbol{u} \rangle_{\partial\Omega} - \langle \boldsymbol{T}^{+}(q, \boldsymbol{u}), \boldsymbol{\gamma}^{+}\boldsymbol{v} \rangle_{\partial\Omega} \\ &= \int_{\Omega} \Big[ \mathscr{A}_j(p, \boldsymbol{v}) u_j - \mathscr{A}_j(q, \boldsymbol{u}) v_j + q \operatorname{div} \boldsymbol{v} - p \operatorname{div} \boldsymbol{u} \Big] dx. \end{aligned} \tag{5.9}$$

## 5.3 Parametrix and Remainder

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 functions $(\mathring{q}^k, \mathring{\boldsymbol{u}}^k)$, where summation in $k$ is not assumed, $\mathring{u}^k_j$ represent components of the incompressible velocity fundamental solution and $\mathring{q}^k$ represent the components of the pressure fundamental solution (see, e.g., [La69]. So for $r_0 > 0$, $\mathring{\mathbf{u}}^k$ and $\mathring{q}^k$ will have the form:

$$\mathring{q}^k(\boldsymbol{x}, \boldsymbol{y}) = \frac{-(x_k - y_k)}{2\pi |\mathbf{x} - \mathbf{y}|^2}, \tag{5.10}$$

$$\mathring{u}^k_j(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{4\pi} \left( \delta^k_j \log \frac{|\mathbf{x} - \mathbf{y}|}{r_0} - \frac{(x_j - y_j)(x_k - y_k)}{|\mathbf{x} - \mathbf{y}|^2} \right), \quad j, k \in \{1, 2\}. \tag{5.11}$$

Therefore, the couple $(\mathring{q}^k, \mathring{\boldsymbol{u}}^k)$ satisfies

$$\frac{\partial}{\partial x_k}\mathring{q}^k(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{2}\frac{\partial^2}{\partial x_k^2}\left(-\frac{1}{2\pi}\log|\mathbf{x}-\mathbf{y}|\right) = -\delta(\mathbf{x}-\mathbf{y}), \tag{5.12}$$

$$\mathring{\mathscr{A}}_j(\boldsymbol{x})(\mathring{q}^k(\boldsymbol{x},\boldsymbol{y}),\mathring{\boldsymbol{u}}^k(\boldsymbol{x},\boldsymbol{y})) = \sum_{i=1}^{2}\frac{\partial^2\mathring{u}_j^k(\boldsymbol{x},\boldsymbol{y})}{\partial x_i^2} - \frac{\partial\mathring{q}^k(\boldsymbol{x},\boldsymbol{y})}{\partial x_j} = \delta_j^k\delta(\boldsymbol{x}-\boldsymbol{y}), \tag{5.13}$$

$$\mathrm{div}_{\boldsymbol{x}}\mathring{\boldsymbol{u}}^k(\boldsymbol{x},\boldsymbol{y}) = 0. \tag{5.14}$$

Here and henceforth, $\delta(\cdot)$ is Dirac's distribution.

Let us denote $\mathring{\sigma}_{ij}(p,\boldsymbol{v}) := \sigma_{ij}(p,\boldsymbol{v})|_{\mu=1}$, $\mathring{T}_i^c(p,\boldsymbol{v}) := T_i^c(p,\boldsymbol{v})|_{\mu=1}$. Then by (5.1) the stress tensor of the fundamental solution reads as

$$\mathring{\sigma}_{ij}(\boldsymbol{x})(\mathring{q}^k(\boldsymbol{x},\boldsymbol{y}),\mathring{\boldsymbol{u}}^k(\boldsymbol{x},\boldsymbol{y})) = \frac{1}{\pi}\frac{(x_i-y_i)(x_j-y_j)(x_k-y_k)}{|\mathbf{x}-\mathbf{y}|^4},$$

and the classical boundary traction of the fundamental solution becomes

$$\begin{aligned}&\mathring{T}_i^c(\boldsymbol{x})(\mathring{q}^k(\boldsymbol{x},\boldsymbol{y}),\mathring{\boldsymbol{u}}^k(\boldsymbol{x},\boldsymbol{y}))\\ :=&\mathring{\sigma}_{ij}(\boldsymbol{x})(\mathring{q}^k(\boldsymbol{x},\boldsymbol{y}),\mathring{\boldsymbol{u}}^k(\boldsymbol{x},\boldsymbol{y}))\,n_j(\boldsymbol{x}) = \frac{1}{\pi}\frac{(x_i-y_i)(x_j-y_j)(x_k-y_k)}{|\mathbf{x}-\mathbf{y}|^4}n_i(\mathbf{x}).\end{aligned}$$

Let us define a pair of functions $(q^k,\boldsymbol{u}^k)_{k=1}^2$,

$$q^k(\boldsymbol{x},\boldsymbol{y}) = \frac{\mu(\boldsymbol{x})}{\mu(\boldsymbol{y})}\mathring{q}^k(\boldsymbol{x},\boldsymbol{y}) = \frac{\mu(\mathbf{x})}{\mu(\mathbf{y})}\frac{y_k-x_k}{2\pi|\mathbf{x}-\mathbf{y}|^2},\quad j,k\in\{1,2\}, \tag{5.15}$$

$$u_j^k(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{\mu(\boldsymbol{y})}\mathring{u}_j^k(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{4\pi\mu(\boldsymbol{y})}\left(\delta_j^k\log\frac{|\mathbf{x}-\mathbf{y}|}{r_0} - \frac{(x_j-y_j)(x_k-y_k)}{|\mathbf{x}-\mathbf{y}|^2}\right). \tag{5.16}$$

Then by (5.1),

$$\sigma_{ij}(\boldsymbol{x})(q^k(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{u}^k(\boldsymbol{x},\boldsymbol{y})) = \frac{\mu(\boldsymbol{x})}{\mu(\boldsymbol{y})}\mathring{\sigma}_{ij}(\boldsymbol{x})(\mathring{q}^k(\boldsymbol{x},\boldsymbol{y}),\mathring{\boldsymbol{u}}^k(\boldsymbol{x},\boldsymbol{y})), \tag{5.17}$$

$$\begin{aligned}T_i(\boldsymbol{x})(q^k(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{u}^k(\boldsymbol{x},\boldsymbol{y})) &:= \sigma_{ij}(\boldsymbol{x})(q^k(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{u}^k(\boldsymbol{x},\boldsymbol{y}))\,n_j(\boldsymbol{x})\\ &= \frac{\mu(\boldsymbol{x})}{\mu(\boldsymbol{y})}\mathring{T}_i(\boldsymbol{x})(\mathring{q}^k(\boldsymbol{x},\boldsymbol{y}),\mathring{\boldsymbol{u}}^k(\boldsymbol{x},\boldsymbol{y})).\end{aligned} \tag{5.18}$$

No summation in $k$ is assumed in (5.17) and (5.18).

Substituting (5.15)–(5.16) in the Stokes system with variable coefficient, (5.2) gives

$$\mathscr{A}_j(\boldsymbol{x})(q^k(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{u}^k(\boldsymbol{x},\boldsymbol{y})) = \delta_j^k\delta(\boldsymbol{x}-\boldsymbol{y}) + R_{kj}(\boldsymbol{x},\boldsymbol{y}), \tag{5.19}$$

where

$$R_{kj}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{\mu(\boldsymbol{y})}\frac{\partial\mu(\boldsymbol{x})}{\partial x_i}\mathring{\sigma}_{ij}(\boldsymbol{x})(\mathring{q}^k(\boldsymbol{x},\boldsymbol{y}),\mathring{\boldsymbol{u}}^k(\boldsymbol{x},\boldsymbol{y}))$$
$$= \frac{1}{\pi\mu(\boldsymbol{y})}\frac{\partial\mu(\boldsymbol{x})}{\partial x_i}\frac{(x_i-y_i)(x_j-y_j)(x_k-y_k)}{|\boldsymbol{x}-\boldsymbol{y}|^4} = \mathcal{O}(|\boldsymbol{x}-\boldsymbol{y}|)^{-1}) \tag{5.20}$$

is a weakly singular remainder and no summation in $k$ is assumed in (5.19)–(5.20). This implies that $(q^k,\boldsymbol{u}^k)$ is a parametrix of the operator $\mathscr{A}$. Let us keep in mind that we have not assumed summation on the index $k$ in (5.17)–(5.20).

Note that a parametrix is generally not unique (cf. [FrMi21] for BDIEs based on an alternative parametrix for a scalar PDE). The possibility to factor out $\frac{\mu(\boldsymbol{x})}{\mu(\boldsymbol{y})}$ in (5.17)–(5.18) and $\frac{\nabla\mu(\boldsymbol{x})}{\mu(\boldsymbol{y})}$ in (5.20) is due to the careful choice of the parametrix in form (5.15)-(5.16) and this essentially simplifies the analysis of parametrix-based potentials and BDIE systems further on.

## 5.4 Hydrodynamic Potentials

Let first $h$ and $\boldsymbol{h}$ be sufficiently smooth scalar and vector functions on $\overline{\Omega}$, e.g., $h\in\mathscr{D}(\overline{\Omega})$, $\boldsymbol{h}\in\mathscr{D}(\overline{\Omega})$. Let us define the parametrix-based Newton-type and remainder vector potentials for the velocity,

$$[\mathscr{U}\boldsymbol{h}]_k(\boldsymbol{y}) = \mathscr{U}_{kj}h_j(\boldsymbol{y}) := \int_\Omega u_j^k(\boldsymbol{x},\boldsymbol{y})h_j(\boldsymbol{x})dx,$$
$$[\mathscr{R}\boldsymbol{h}]_k(\boldsymbol{y}) = \mathscr{R}_{kj}h_j(\boldsymbol{y}) := \int_\Omega R_{kj}(\boldsymbol{x},\boldsymbol{y})h_j(\boldsymbol{x})dx,$$

and the scalar Newton-type and remainder potentials for the pressure,

$$[\mathscr{Q}h]_j(\boldsymbol{y}) = \mathscr{Q}_j h(\boldsymbol{y}) := \int_\Omega q^j(\boldsymbol{y},\boldsymbol{x})h(\boldsymbol{x})dx = -\int_\Omega q^j(\boldsymbol{x},\boldsymbol{y})h(\boldsymbol{x})dx, \tag{5.21}$$

$$\mathscr{Q}\boldsymbol{h}(\boldsymbol{y}) := \mathscr{Q}\cdot\boldsymbol{h}(\boldsymbol{y}) = \mathscr{Q}_j h_j(\boldsymbol{y}) = \int_\Omega q^j(\boldsymbol{y},\boldsymbol{x})h_j(\boldsymbol{x})dx = -\int_\Omega q^j(\boldsymbol{x},\boldsymbol{y})h_j(\boldsymbol{x})dx, \tag{5.22}$$

$$\mathscr{R}^\bullet\boldsymbol{h}(\boldsymbol{y}) = \mathscr{R}_j^\bullet h_j(\boldsymbol{y}) := -2\,\mathrm{p.v.}\int_\Omega \frac{\partial\mathring{q}^j(\boldsymbol{x},\boldsymbol{y})}{\partial x_i}\frac{\partial\mu(\boldsymbol{x})}{\partial x_i}h_j(\boldsymbol{x})dx - h_j\frac{\partial\mu}{\partial y_j} \tag{5.23}$$
$$= -2\left\langle\partial_i\mathring{q}^j(\cdot,\boldsymbol{y})\,,\,h_i\partial_j\mu\right\rangle_\Omega - 2h_i(\boldsymbol{y})\partial_i\mu(\boldsymbol{y}), \tag{5.24}$$

for $\boldsymbol{y}\in\mathbb{R}^2$. The integral in (5.23) is understood as a 2D strongly singular integral (in the sense of the Cauchy principal value). The bilinear form in (5.24) should be

understood in the sense of distributions, and the equality between (5.23) and (5.24) holds since

$$\begin{aligned}\left\langle \partial_i \mathring{q}^j(\cdot, \boldsymbol{y}), h_i \partial_j \mu \right\rangle_{\Omega} &= -\left\langle \mathring{q}^j(\cdot, \boldsymbol{y}), \partial_i (h_i \partial_j \mu) \right\rangle_{\Omega} + \left\langle n_i \mathring{q}^j(\cdot, \boldsymbol{y}), h_i \partial_j \mu \right\rangle_{\partial\Omega} \\ &= -\lim_{\epsilon \to 0} \left\langle \mathring{q}^j(\cdot, \boldsymbol{y}), \partial_i (h_i \partial_j \mu) \right\rangle_{\Omega_\epsilon} + \left\langle n_i \mathring{q}^j(\cdot, \boldsymbol{y}), h_i \partial_j \mu \right\rangle_{\partial\Omega} \\ &= \lim_{\epsilon \to 0} \left\langle \partial_i \mathring{q}^j(\cdot, \boldsymbol{y}), h_i \partial_j \mu \right\rangle_{\Omega_\epsilon} - \lim_{\epsilon \to 0} \left\langle n_i \mathring{q}^j(\cdot, \boldsymbol{y}), h_i \partial_j \mu \right\rangle_{\partial\Omega_\epsilon \setminus \partial\Omega} \\ &= \text{v.p.} \int_{\Omega} \frac{\partial \mathring{q}^j(\boldsymbol{x}, \boldsymbol{y})}{\partial x_i} \frac{\partial \mu(\boldsymbol{x})}{\partial x_i} h_j(\boldsymbol{x}) dx - h_j \frac{\partial \mu}{\partial y_j} ,\end{aligned}$$

where $\Omega_\epsilon = \Omega \setminus \bar{B}_\epsilon(\boldsymbol{y})$ and $B_\epsilon(\boldsymbol{y})$ is the ball of radius $\epsilon$ centred in $\boldsymbol{y}$, which implies that

$$\begin{aligned}&-2\left\langle \partial_i \mathring{q}^j(\cdot, \boldsymbol{y}),\ h_i \partial_j \mu \right\rangle_{\Omega} - 2h_i(\boldsymbol{y}) \partial_i \mu(\boldsymbol{y}) \\ &= -2\,\text{v.p.} \int_{\Omega} \frac{\partial \mathring{q}^j(\boldsymbol{x}, \boldsymbol{y})}{\partial x_i} \frac{\partial \mu(\boldsymbol{x})}{\partial x_i} h_j(\boldsymbol{x}) dx - h_j(\boldsymbol{y}) \frac{\partial \mu(\boldsymbol{y})}{\partial y_j} = \mathscr{R}^{\bullet} \boldsymbol{h}(\boldsymbol{y}).\end{aligned}$$

In addition, we will introduce the operators $\mathbf{U}, \mathbf{Q}, \mathbf{R}$, and $\mathbf{R}^{\bullet}$ whose definitions coincide, respectively, with the definition of the operators $\mathscr{U}, \mathscr{Q}, \mathscr{R}$, and $\mathscr{R}^{\bullet}$ with the sole difference that $\Omega = \mathbb{R}^3$.

Let us now define the parametrix-based velocity single-layer potential and double layer potential as follows:

$$[\boldsymbol{V}\boldsymbol{h}]_k(\boldsymbol{y}) = V_{kj} h_j(\boldsymbol{y}) := -\int_{\partial\Omega} u_j^k(\boldsymbol{x}, \boldsymbol{y}) h_j(\boldsymbol{x})\, dS(\boldsymbol{x}), \quad \boldsymbol{y} \notin \partial\Omega,$$

$$[\boldsymbol{W}\boldsymbol{h}]_k(\boldsymbol{y}) = W_{kj} h_j(\boldsymbol{y}) := -\int_{\partial\Omega} T_j^c(\boldsymbol{x}; q^k, \boldsymbol{u}^k)(\boldsymbol{x}, \boldsymbol{y}) h_j(\boldsymbol{x})\, dS(\boldsymbol{x}), \quad \boldsymbol{y} \notin \partial\Omega.$$

For the pressure we will need the following single-layer and double layer potentials:

$$\Pi^s \boldsymbol{h}(\boldsymbol{y}) = \Pi_j^s h_j(\boldsymbol{y}) := \int_{\partial\Omega} \mathring{q}^j(\boldsymbol{x}, \boldsymbol{y}) h_j(\boldsymbol{x}) dS(\boldsymbol{x}), \quad \boldsymbol{y} \notin \partial\Omega$$

$$\Pi^d \boldsymbol{h}(\boldsymbol{y}) = \Pi_j^d h_j(\boldsymbol{y}) := 2\int_{\partial\Omega} \frac{\partial \mathring{q}^j(\boldsymbol{x}, \boldsymbol{y})}{\partial n(\boldsymbol{x})} \mu(\boldsymbol{x}) h_j(\boldsymbol{x}) dS(\boldsymbol{x}), \quad \boldsymbol{y} \notin \partial\Omega.$$

It is easy to observe that the parametrix-based integral operators, with the variable coefficient $\mu$, can be expressed in terms of the corresponding integral operators for the constant- coefficient case, $\mu = 1$, marked by $\mathring{}$,

$$\mathscr{U}\boldsymbol{h} = \frac{1}{\mu} \mathring{\mathscr{U}} \boldsymbol{h}, \tag{5.25}$$

$$[\mathscr{R}\boldsymbol{h}]_k = \frac{-1}{\mu}\Big[\partial_j \mathring{\mathscr{U}}_{ki}(h_j\partial_i\mu) + \partial_i \mathring{\mathscr{U}}_{kj}(h_j\partial_i\mu) - \mathring{\mathscr{Q}}_k(h_j\partial_j\mu)\Big], \tag{5.26}$$

$$\mathscr{Q}h = \frac{1}{\mu}\mathring{\mathscr{Q}}(\mu h), \tag{5.27}$$

$$\mathscr{R}^{\bullet}\boldsymbol{h} = -2\partial_i \mathring{\mathscr{Q}}_j(h_j\partial_i\mu) - 2h_j\partial_j\mu, \tag{5.28}$$

$$\boldsymbol{V}\boldsymbol{h} = \frac{1}{\mu}\mathring{\boldsymbol{V}}\boldsymbol{h}, \qquad \boldsymbol{W}\boldsymbol{h} = \frac{1}{\mu}\mathring{\boldsymbol{W}}(\mu\boldsymbol{h}), \tag{5.29}$$

$$\Pi^s\boldsymbol{h} = \mathring{\Pi}^s\boldsymbol{h}, \qquad \Pi^d\boldsymbol{h} = \mathring{\Pi}^d(\mu\boldsymbol{h}). \tag{5.30}$$

We will further use (5.25)–(5.30) as definitions of the potentials in the left-hand sides of these relations, when the densities $h$ and $\boldsymbol{h}$ are more general functions or distributions on $\Omega$ or $\partial\Omega$.

Note that although the constant-coefficient velocity potentials $\mathring{\mathscr{U}}\boldsymbol{h}$, $\mathring{\boldsymbol{V}}\boldsymbol{h}$, and $\mathring{\boldsymbol{W}}\boldsymbol{h}$ are divergence-free in $\Omega^{\pm}$, the corresponding potentials $\mathscr{U}\boldsymbol{h}$, $\boldsymbol{V}\boldsymbol{h}$, and $\boldsymbol{W}\boldsymbol{h}$ are *not divergence-free for the variable coefficient* $\mu(\boldsymbol{y})$. Note also that by (5.10) and (5.21),

$$\mathring{\mathscr{Q}}_j h = \partial_j \mathscr{N}_{\Delta} h, \tag{5.31}$$

where

$$\mathscr{N}_{\Delta}h(\boldsymbol{y}) = -\frac{1}{2\pi}\int_{\Omega} \log\frac{|\boldsymbol{x}-\boldsymbol{y}|}{r_0} h(\boldsymbol{x})dx \tag{5.32}$$

is the harmonic Newton potential. Hence

$$\operatorname{div}\mathring{\mathscr{Q}}h = \partial_j\mathring{\mathscr{Q}}_j h = \Delta\mathscr{N}_{\Delta}h = -h. \tag{5.33}$$

Moreover, for the constant-coefficient potentials we have the following well-known relations,

$$\mathring{\mathscr{A}}(\mathring{\Pi}^s\boldsymbol{h}, \mathring{\boldsymbol{V}}\boldsymbol{h}) = \boldsymbol{0}, \quad \mathring{\mathscr{A}}(\mathring{\Pi}^d\boldsymbol{h}, \mathring{\boldsymbol{W}}\boldsymbol{h}) = \boldsymbol{0} \quad \text{in } \Omega^{\pm}, \tag{5.34}$$

$$\mathring{\mathscr{A}}(\mathring{\mathscr{Q}}\boldsymbol{h}, \mathring{\mathscr{U}}\boldsymbol{h}) = \boldsymbol{h}. \tag{5.35}$$

In addition, by (5.31) and (5.33),

$$\begin{aligned}\mathring{\mathscr{A}}_j((2-\alpha)h, -\mathring{\mathscr{Q}}h) &= -\partial_i\Big(\partial_i\mathring{\mathscr{Q}}_j h + \partial_j\mathring{\mathscr{Q}}_i h - \alpha\delta_i^j \operatorname{div}\mathring{\mathscr{Q}}h\Big) - (2-\alpha)\partial_j h \\ &= -(\Delta\mathring{\mathscr{Q}}_j h + \partial_j\operatorname{div}\mathring{\mathscr{Q}}h - \alpha\partial_j\operatorname{div}\mathring{\mathscr{Q}}h) - (2-\alpha)\partial_j h = 0.\end{aligned} \tag{5.36}$$

### 5.4.1 Mapping Properties

The following assertions are well known for the constant-coefficient case, see, e.g., Lemmas A.3 and A.4 in [KLMW16] and the references therein. Then by relations (5.25)-(5.30), we obtain their counterparts for the variable-coefficient case. Let us highlight that the operators $\boldsymbol{U}$, $\boldsymbol{Q}$, $Q$, $\boldsymbol{R}$, $R^{\bullet}$ are defined in the same way as $\mathscr{U}$, $\mathscr{Q}$, $\mathscr{Q}$, $\mathscr{R}$, and $\mathscr{R}^{\bullet}$ if we take $\Omega = \mathbb{R}^2$.

**Remark 5.3** *For sufficiently smooth $h$, the Newtonian volume potential over $\mathbb{R}^2$, cf. (5.32), is defined as*

$$N_{\Delta} h(\boldsymbol{y}) = \int_{\mathbb{R}^2} E_{\Delta}(\boldsymbol{x}, \boldsymbol{y})\, h(\boldsymbol{x})\, d\boldsymbol{x}, \tag{5.37}$$

*where*

$$E_{\Delta}(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{2\pi} \log \frac{|\boldsymbol{x} - \boldsymbol{y}|}{r_0}$$

*is the fundamental solution of the Laplace equation and moreover $\mathcal{N}_{\Delta} \Delta h = \Delta \mathcal{N}_{\Delta} h = -h$, i.e., the operator $N_{\Delta}$ is inverse to the Laplace operator $\Delta$. On the other hand, it is well known (see, e.g., [Ha71, Theorem III.2]) that the Laplace operator $\Delta : \mathscr{H}^1(\mathbb{R}^2) \to \mathscr{H}^{-1}(\mathbb{R}^2)$ has a continuous inverse, $\Delta^{-1} : \mathscr{H}^{-1}(\mathbb{R}^2) \to \mathscr{H}^1(\mathbb{R}^2)$ and thus $N_{\Delta} h = \Delta^{-1} h$ for any $h \in \mathscr{D}(\mathbb{R}^2)$. As remarked in [CMN13], due to the density of $\mathscr{D}(\mathbb{R}^2)$ in $\mathscr{H}^{-1}(\mathbb{R}^2)$ this provides a continuous extension of the operator $N_{\Delta}$ defined by (5.37) to the extended continuous Newtonian potential operator*

$$N_{\Delta} : \mathscr{H}^{-1}(\mathbb{R}^2) \to \mathscr{H}^1(\mathbb{R}^2). \tag{5.38}$$

**Theorem 5.4** *The following operators are continuous under condition 5.1,*

$$\boldsymbol{U} : \mathscr{H}^{-1}(\mathbb{R}^2) \to \mathscr{H}^1(\mathbb{R}^2), \tag{5.39}$$

$$\mathscr{U} : \widetilde{\mathscr{H}}^{-1}(\Omega) \to \mathscr{H}^1(\Omega), \tag{5.40}$$

$$\boldsymbol{Q} : L_2(\mathbb{R}^2) \to \mathscr{H}^1(\mathbb{R}^2), \tag{5.41}$$

$$\mathscr{Q} : L_2(\Omega) \to \mathscr{H}^1(\Omega), \tag{5.42}$$

$$Q : \mathscr{H}^{-1}(\mathbb{R}^2) \to L_2(\mathbb{R}^2), \tag{5.43}$$

$$\mathscr{Q} : \widetilde{\mathscr{H}}^{-1}(\Omega) \to L_2(\Omega), \tag{5.44}$$

$$\boldsymbol{R} : \boldsymbol{L}_2(\rho^{-1}; \mathbb{R}^2) \to \mathscr{H}^1(\mathbb{R}^2), \tag{5.45}$$

$$\mathscr{R} : \boldsymbol{L}_2(\rho^{-1}; \Omega) \to \mathscr{H}^1(\Omega), \tag{5.46}$$

$$R^\bullet : \boldsymbol{L}_2(\rho^{-1}; \mathbb{R}^2) \to L_2(\mathbb{R}^2), \tag{5.47}$$

$$\mathscr{R}^\bullet : \boldsymbol{L}_2(\rho^{-1}; \Omega) \to L_2(\Omega). \tag{5.48}$$

***Proof*** Let us consider relations (5.25) and (5.27). The continuity of operators $\boldsymbol{U}$, $\mathscr{U}$, $Q$ and $\mathscr{Q}$ in (5.39), (5.40), (5.43), and (5.44) then follows from the continuity of the corresponding operators $\mathring{\mathscr{U}}$, $\mathring{\boldsymbol{U}}$, $\mathring{\mathscr{Q}}$, and $\mathring{Q}$ provided in [KLMW16, Lemma A.3].

Let us prove now the continuity of operator (5.45), which follows if we prove the continuity of operators in the right hand side of (5.26). Let us note that by condition 5.1, $\mu$ and $\frac{1}{\mu}$ are bounded and act as multipliers in the space $\mathscr{H}^1(\Omega)$. In addition, condition 5.1 states that $\rho\partial_i\mu \in L_\infty(\mathbb{R}^2)$. Consequently, for any function $h_j \in L_2(\rho^{-1}; \mathbb{R}^2)$, we have that $h_j\partial_i\mu \in L_2(\mathbb{R}^2)$, see the proof of [CMN13, Theorem 4.1]. It is easy to prove that the operator $\nabla : L_2(\mathbb{R}^2) \to \mathscr{H}^1(\mathbb{R}^2)$ is continuous, which implies that $\nabla(h_j\partial_i\mu) \in \mathscr{H}^1(\mathbb{R}^2)$.

Let us prove continuity of the first operator in the right hand side of (5.26). First, we assume that $h_j\partial_i\mu$ in $\mathscr{D}(\mathbb{R}^2)$. Then

$$\partial_j \mathring{\mathscr{U}}_{ki}(h_j\partial_i\mu) = -\mathring{\mathscr{U}}_{ki}\partial_j(h_j\partial_i\mu). \tag{5.49}$$

By the density of $\mathscr{D}(\mathbb{R}^2)$ in $L_2(\mathbb{R}^2)$ and the continuity of operator $\mathring{\boldsymbol{U}} : \mathscr{H}^{-1}(\mathbb{R}^2) \to \mathscr{H}^1(\mathbb{R}^2)$, cf. (5.39), we can extend relation (5.49) from $h_j\partial_i\mu \in \mathscr{D}(\mathbb{R}^2)$ to $h_j\partial_i\mu \in L_2(\mathbb{R}^2)$. Then, the continuity of operator

$$\boldsymbol{h} \mapsto \partial_j \mathring{\mathscr{U}}_{ki}(h_j\partial_i\mu) : \boldsymbol{L}_2(\rho^{-1}; \mathbb{R}^2) \to \mathscr{H}^1(\mathbb{R}^2)$$

follows. The continuity of other two operators in the right hand side of (5.26) can be proved in a similar way. Consequently, operator (5.45) is continuous. The continuity of operator (5.45) implies the continuity of operator (5.46).

Taking into account (5.28), the continuity of operator (5.47) will follow from the continuity of the first operator in the right hand side of (5.28). Let $h_j \in L_2(\rho^{-1}; \mathbb{R}^2)$. Applying a similar density argument, as in the previous paragraph we can deduce $\partial_j\mathring{\mathscr{Q}}(h_j\partial_i\mu) = -\mathring{\mathscr{Q}}\partial_j(h_j\partial_i\mu)$. Since, $\partial_j(h_j\partial_i\mu) \in \mathscr{H}^{-1}(\mathbb{R}^2)$, the we have the inclusion $\partial_j\mathring{\mathscr{Q}}(h_j\partial_i\mu) \in L_2(\mathbb{R}^2)$ for any $h_j \in L_2(\rho^{-1}; \mathbb{R}^2)$, with the corresponding norm estimate. This implies the continuity of operator (5.47). Continuity of operator (5.48) is implied by the continuity of operator (5.47).

The mapping properties of operators (5.41) and (5.42) differ from the ones for operators (5.43) and (5.44) and need to be proved separately.

Let us consider $\phi \in \mathscr{D}(\mathbb{R}^2)$. Then by (5.31) and (5.32) we have

$$\begin{aligned}\mathring{\mathscr{Q}}_j\phi =& -\partial_j N_\Delta\phi = -\int_{\mathbb{R}^2} \frac{\partial E_\Delta}{\partial y_j}(\boldsymbol{x}, \boldsymbol{y})\phi(\boldsymbol{x})\, d\boldsymbol{x} = \int_{\mathbb{R}^2} \frac{\partial E_\Delta}{\partial x_j}(\boldsymbol{x}, \boldsymbol{y})\phi(\boldsymbol{x})\, d\boldsymbol{x} \\ =& -\int_{\mathbb{R}^2} E_\Delta(\boldsymbol{x}, \boldsymbol{y})\frac{\partial\phi(\boldsymbol{x})}{\partial x_j}\, d\boldsymbol{x} = -N_\Delta(\partial_j\phi).\end{aligned} \tag{5.50}$$

For any $h \in L_2(\mathbb{R}^2)$,

$$\|\partial_j h\|_{\mathscr{H}^{-1}(\mathbb{R}^n)} = \sup_{\xi \in \mathscr{D}(\mathbb{R}^2), \|\xi\|_{\mathscr{H}^1(\mathbb{R}^n)}=1} |\langle \partial_j h, \xi \rangle_{\mathbb{R}^2}| = \sup_{\xi \in \mathscr{D}(\mathbb{R}^2), \|\xi\|_{\mathscr{H}^1(\mathbb{R}^n)}=1} |\langle h, \partial_j \xi \rangle_{\mathbb{R}^2}|$$

$$\leq \sup_{\xi \in \mathscr{D}(\mathbb{R}^2), \|\xi\|_{\mathscr{H}^1(\mathbb{R}^n)}=1} \|h\|_{L_2(\mathbb{R}^2)} \|\partial_j \xi\|_{L_2(\mathbb{R}^2)} \leq \|h\|_{L_2(\mathbb{R}^2)}. \tag{5.51}$$

Due to the density of $\mathscr{D}(\mathbb{R}^2)$ in $\mathscr{H}^1(\mathbb{R}^2)$, this implies that $\partial_j h \in \mathscr{H}^{-1}(\mathbb{R}^n)$ and moreover the operator $\partial_j : L_2(\mathbb{R}^2) \to \mathscr{H}^{-1}(\mathbb{R}^2)$ is continuous.

As a result, the density of $\mathscr{D}(\mathbb{R}^3)$ in $L_2(\mathbb{R}^2)$ and the continuity of operator (5.38) in (5.50) imply that $\mathring{\mathscr{Q}}_j \phi = -\mathscr{N}_\Delta(\partial_j \phi) \in \mathscr{H}^1(\mathbb{R}^2)$ for any $\phi \in L_2(\mathbb{R}^2)$ and moreover, the operator $\mathring{\mathscr{Q}}_j : L_2(\mathbb{R}^2) \to \mathscr{H}^1(\mathbb{R}^2)$ is continuous. Then operator (5.41) and thus operator (5.42) are continuous as well.

**Theorem 5.5** *The following operators are continuous under condition 5.1*

$$\boldsymbol{V} : \boldsymbol{H}_{**}^{-1/2}(\partial\Omega) \to \mathscr{H}^1(\Omega), \tag{5.52}$$

$$\Pi^s : \boldsymbol{H}_{**}^{-1/2}(\partial\Omega) \to L_2(\Omega), \tag{5.53}$$

$$\boldsymbol{W} : \boldsymbol{H}^{1/2}(\partial\Omega) \to \mathscr{H}^1(\Omega), \tag{5.54}$$

$$\Pi^d : \boldsymbol{H}^{1/2}(\partial\Omega) \to L_2(\Omega). \tag{5.55}$$

***Proof*** Let us consider relations (5.29) and (5.30). The continuity of the operators $\boldsymbol{V}$, $\Pi^s$, $\boldsymbol{W}$, and $\Pi^d$ then follows from the continuity of the operators $\mathring{\boldsymbol{V}}$, $\mathring{\boldsymbol{W}}$, $\mathring{\Pi}^s$ and $\mathring{\Pi}^d$ which has already being proved in [Sa14, Proposition 7.2].

In the proofs further, second order derivatives of the coefficient $\mu(\boldsymbol{x})$ will appear and apart from Condition 5.1, we will sometimes need to assume the following additional condition.

**Condition 5.6**

$$\mu \in \mathscr{C}^2(\mathbb{R}^2) \ : \ \rho^2 \partial_j \partial_i \mu \in L_\infty(\mathbb{R}^2). \tag{5.56}$$

**Theorem 5.7** *The following operators are continuous under Conditions 5.1 and 5.6,*

$$\left(\Pi^s, \boldsymbol{V}\right) : \boldsymbol{H}_{**}^{-1/2}(\partial\Omega) \to \mathscr{H}^{1,0}(\Omega; \mathscr{A}), \tag{5.57}$$

$$\left(\Pi^d, \boldsymbol{W}\right) : \boldsymbol{H}^{1/2}(\partial\Omega) \to \mathscr{H}^{1,0}(\Omega; \mathscr{A}), \tag{5.58}$$

$$\left(\mathring{\mathscr{Q}}, \mathscr{U}\right) : \boldsymbol{L}_2(\rho; \Omega) \to \mathscr{H}^{1,0}(\mathbb{R}^2; \mathscr{A}), \tag{5.59}$$

$$\left(\mathscr{R}^\bullet, \mathscr{R}\right) : \mathscr{H}^1(\Omega) \to \mathscr{H}^{1,0}(\Omega; \mathscr{A}), \tag{5.60}$$

$$(((2-\alpha)\mu, -\mathscr{Q}) : L_2(\Omega) \to \mathscr{H}^{1,0}(\Omega; \mathscr{A}). \tag{5.61}$$

***Proof*** Let us consider first the single-layer potentials $(\Pi^s \boldsymbol{h}, \boldsymbol{V}\boldsymbol{h}) \in \mathscr{H}^1(\Omega) \times L_2(\Omega)$ for $\boldsymbol{h} \in \boldsymbol{H}^{-1/2}(\partial\Omega)$. Let us apply the operator $\mathscr{A}$ taking into consideration (5.29) and (5.30)

$$\mathscr{A}_j(\Pi^s \boldsymbol{h}, \boldsymbol{V}\boldsymbol{h}) = \mathscr{A}_j\left(\mathring{\Pi}^s \boldsymbol{h}, \frac{1}{\mu}\mathring{\boldsymbol{V}}\boldsymbol{h}\right)$$
$$= \mathring{\mathscr{A}}_j\left(\mathring{\Pi}^s \boldsymbol{h}, \mathring{V}_k \boldsymbol{h}\right) + \partial_k\left(\mu\left[\partial_j(1/\mu)\mathring{V}_k \boldsymbol{h} + \partial_k(1/\mu)\mathring{V}_j \boldsymbol{h} - \alpha\delta_j^k \partial_i(1/\mu)\mathring{V}_i \boldsymbol{h}\right]\right).$$

Now, the term $\mathring{\mathscr{A}}_j(\mathring{\Pi}^s \boldsymbol{h}, \mathring{V}_k \boldsymbol{h})$ vanishes and due Conditions 5.1 and 5.6, the last term belongs to $\boldsymbol{L}_2(\rho; \Omega)$ since $\mathring{\boldsymbol{V}}\boldsymbol{h} \in \mathscr{H}^1(\Omega)$, which implies the continuity of operator (5.57).

The same argument works for the double layer potential $\left(\boldsymbol{W}, \Pi^d\right)\boldsymbol{h}$ with $\boldsymbol{h} \in \boldsymbol{H}^{1/2}(\partial\Omega)$and implies the continuity of operator (5.58). In addition it works for the Newtonian potentials $(\mathscr{U}, \mathscr{Q})$ with the sole difference that $\mathring{\mathscr{A}}_j\left(\mathring{\mathscr{Q}}\boldsymbol{h}, \mathring{\mathscr{U}}_k \boldsymbol{h}\right) = h_j$ and $\boldsymbol{h} \in \boldsymbol{L}_2(\rho; \Omega)$. This implies the continuity of operator (5.59).

For operator (5.60), $\boldsymbol{h} \in \mathscr{H}^1(\Omega) \subset \boldsymbol{L}_2(\rho^{-1}; \Omega)$ and hence the operator $(\mathscr{R}^\bullet, \mathscr{R}) : \mathscr{H}^1(\Omega) \to L_2(\Omega) \times \mathscr{H}^1(\Omega)$ is continuous due to Theorem 5.4. Let us prove that $\mathscr{A}\,(\mathscr{R}^\bullet, \mathscr{R}) : \mathscr{H}^1(\Omega) \to \boldsymbol{L}_2(\rho; \Omega)$ is continuous. Indeed, by (5.2),

$$\mathscr{A}_j(\mathscr{R}^\bullet \boldsymbol{h}, \mathscr{R}\boldsymbol{h}) = \mathring{\mathscr{A}}_j(\mathscr{R}^\bullet \boldsymbol{h}, \mu\mathscr{R}\boldsymbol{h}) - 2\,\partial_i \mathbb{M}_{ij}(\mathscr{R}\boldsymbol{h}), \tag{5.62}$$

where

$$\mathbb{M}_{ij}(\mathbf{u}) := \frac{1}{2}(u_j \partial_i \mu + u_i \partial_j \mu) - \frac{\alpha}{2}\delta_{ij} u_l \partial_l \mu.$$

Hence due to Theorem 5.4 and Conditions 5.1 and 5.6, the operator $\partial_i \mathbb{M}_{ij}\mathscr{R} : \mathscr{H}^1(\Omega) \to L_2(\rho; \Omega)$ is continuous. Moreover, by (5.26), (5.28) and (5.35), $\mathring{\mathscr{A}}_j(\mathscr{R}^\bullet \boldsymbol{h}, \mu\mathscr{R}\boldsymbol{h}) = -2\,\partial_i \mathbb{M}_{ij}(\boldsymbol{h})$, hence by Conditions 5.1 and 5.6 the operator $\mathring{\mathscr{A}}_j(\mathscr{R}^\bullet, \mu\mathscr{R}) : \mathscr{H}^1(\Omega) \to L_2(\rho; \Omega)$ is continuous. Then (5.62) implies the continuity of operator $\mathscr{A}(\mathscr{R}^\bullet, \mu\mathscr{R}) : \mathscr{H}^1(\Omega) \to \boldsymbol{L}_2(\rho; \Omega)$ and hence of operator (5.60).

For operator (5.61) we proceed in a similar manner to obtain that

$$\mathscr{A}\,((2-\alpha)\mu h, -\mathscr{Q}h) = \mathscr{A}\left((2-\alpha)\mu h, -\frac{1}{\mu}\mathring{\mathscr{Q}}(\mu h)\right)$$
$$= \mathring{\mathscr{A}}_j\left((2-\alpha)\mu h, -\mathring{\mathscr{Q}}(\mu h)\right) + 2\,\partial_i \mathbb{M}_{ij}(\mathscr{Q}h) = 2\,\partial_i \mathbb{M}_{ij}(\mathscr{Q}h)$$

due to (5.36). By the continuity of operator (5.42) in Theorem 5.4 and due to Conditions 5.1 and 5.6, the operator $\partial_i \mathbb{M}_{ij}\mathscr{Q} : L_2(\Omega) \to \boldsymbol{L}_2(\rho; \Omega)$ is continuous, implying the continuity of operator (5.61).

Let us now define direct values on the boundary of the parametrix-based velocity single-layer and double layer potentials and introduce the notations for the conormal derivative of the latter, for sufficiently smooth scalar and vector functions $h$ and $\boldsymbol{h}$ on $\partial\Omega$, e.g., $h \in \mathscr{D}(\partial\Omega)$, $\boldsymbol{h} \in \mathscr{D}(\partial\Omega)$,

$$[\mathscr{V}\boldsymbol{h}]_k(\boldsymbol{y}) = \mathscr{V}_{kj}h_j(\boldsymbol{y}) := -\int_{\partial\Omega} u_j^k(\boldsymbol{x},\boldsymbol{y})h_j(\boldsymbol{x})\,dS(\boldsymbol{x}), \qquad \boldsymbol{y}\in\partial\Omega, \tag{5.63}$$

$$[\mathscr{W}\boldsymbol{h}]_k(\boldsymbol{y}) = \mathscr{W}_{kj}h_j(\boldsymbol{y}) := -\int_{\partial\Omega} T_j^c(\boldsymbol{x};q^k,\boldsymbol{u}^k)(\boldsymbol{x},\boldsymbol{y})h_j(\boldsymbol{x})\,dS(\boldsymbol{x}), \qquad \boldsymbol{y}\in\partial\Omega, \tag{5.64}$$

$$[\mathscr{W}'\boldsymbol{h}]_k(\boldsymbol{y}) = \mathscr{W}'_{kj}h_j(\boldsymbol{y}) := -\int_{\partial\Omega} T_j^c(\boldsymbol{y};q^k,\boldsymbol{u}^k)(\boldsymbol{x},\boldsymbol{y})h_j(\boldsymbol{x})\,dS(\boldsymbol{x}), \qquad \boldsymbol{y}\in\partial\Omega, \tag{5.65}$$

$$\mathscr{L}^{\pm}\boldsymbol{h}(\boldsymbol{y}) := \boldsymbol{T}^{\pm}(\Pi^d\boldsymbol{h},\boldsymbol{W}\boldsymbol{h})(\boldsymbol{y}), \qquad \boldsymbol{y}\in\partial\Omega. \tag{5.66}$$

Here $\boldsymbol{T}^{\pm}$ are the canonical derivative (traction) operators for the *compressible* fluid that are well defined due to Theorem 5.7.

Similar to the potentials in the domain, we can also express the boundary operators in terms of their counterparts with the constant coefficient $\mu = 1$,

$$\mathscr{V}\boldsymbol{h} = \frac{1}{\mu}\mathring{\mathscr{V}}\boldsymbol{h}, \qquad \mathscr{W}\boldsymbol{h} = \frac{1}{\mu}\mathring{\mathscr{W}}(\mu\boldsymbol{h}), \tag{5.67}$$

$$[\mathscr{W}'\boldsymbol{h}]_k = [\mathring{\mathscr{W}}'\boldsymbol{h}]_k - \left(\frac{\partial_i\mu}{\mu}[\mathring{\mathscr{V}}\boldsymbol{h}]_k + \alpha\delta_i^k\frac{\partial_j\mu}{\mu}[\mathring{\mathscr{V}}\boldsymbol{h}]_j\right)n_i. \tag{5.68}$$

We will further use relations (5.67) and (5.68) as definitions of the potentials $\mathscr{V}\boldsymbol{h}$, $\mathscr{W}\boldsymbol{h}$, and $\mathscr{W}'\boldsymbol{h}$ when their densities $h$ and $\boldsymbol{h}$ are more general functions or distributions on $\partial\Omega$.

**Theorem 5.8** *Let $s\in\mathbb{R}$. Then the following operators are continuous under Conditions 5.1 and 5.6,*

$$\mathscr{V}:\boldsymbol{H}^s(\partial\Omega)\to\boldsymbol{H}^{s+1}(\partial\Omega), \qquad \mathscr{W}:\boldsymbol{H}^s(\partial\Omega)\to\boldsymbol{H}^{s+1}(\partial\Omega), \tag{5.69}$$

$$\mathscr{L}^{\pm}:\boldsymbol{H}^s(\partial\Omega)\to\boldsymbol{H}^{s-1}(\partial\Omega), \qquad \mathscr{W}':\boldsymbol{H}^s(\partial\Omega)\to\boldsymbol{H}^{s+1}(\partial\Omega). \tag{5.70}$$

*Moreover, the following operators are compact,*

$$\mathscr{V}:\boldsymbol{H}^s(\partial\Omega)\to\boldsymbol{H}^s(\partial\Omega), \tag{5.71}$$

$$\mathscr{W}:\boldsymbol{H}^s(\partial\Omega)\to\boldsymbol{H}^s(\partial\Omega), \tag{5.72}$$

$$\mathscr{W}':\boldsymbol{H}^s(\partial\Omega)\to\boldsymbol{H}^s(\partial\Omega). \tag{5.73}$$

***Proof*** As in Theorem 4.4 of [FrMi21], the continuity of operators in (5.69)–(5.70) follows from relations (5.67)–(5.68) and the continuity of the counterpart operators for the constant-coefficient case. Then, compactness of operators (5.71)–(5.73) is implied by the Rellich compactness embedding theorem.

**Theorem 5.9** *If $\boldsymbol{\tau} \in \boldsymbol{H}^{1/2}(\partial\Omega)$, $\boldsymbol{h} \in \boldsymbol{H}^{-1/2}(\partial\Omega)$, then the following relations hold on $\partial\Omega$ under Conditions 5.1 and 5.6:*

$$\gamma^{\pm}\boldsymbol{V}\boldsymbol{h} = \mathscr{V}\boldsymbol{h}, \qquad \gamma^{\pm}\boldsymbol{W}\boldsymbol{\tau} = \mp\frac{1}{2}\boldsymbol{\tau} + \mathscr{W}\boldsymbol{\tau} \tag{5.74}$$

$$\boldsymbol{T}^{\pm}(\Pi^{s}\boldsymbol{h}, \boldsymbol{V}\boldsymbol{h}) = \pm\frac{1}{2}\boldsymbol{h} + \mathscr{W}'\boldsymbol{h}. \tag{5.75}$$

***Proof*** The proof of the theorem directly follows from relations (5.29), (5.67)–(5.68) and the analogous jump properties for the counterparts of the operators for the constant-coefficient case of $\mu = 1$, see, e.g., [HsWe08, Lemma 5.6.5].

For bounded domains, we had compactness of the remainder operators $\mathscr{R}$ and $\mathscr{R}^{\bullet}$ implied by the Rellich compact embedding theorem, which does not hold for exterior (unbounded) domains considered in this chapter. To overcome this issue, we prove that for exterior domains the operators $\mathscr{R}$ and $\mathscr{R}^{\bullet}$ are limits of some sequences of compact operators and thus are also compact. We will require the following condition.

**Condition 5.10** $\lim_{|x|\to\infty} \rho(\boldsymbol{x})\nabla\mu(\boldsymbol{x}) = 0.$

The proof of the following assertion is similar to [CMN13, Lemma 7.4] for the corresponding scalar case.

**Lemma 5.11** *Let Conditions 5.1 and 5.10 hold. For any sufficiently large $\eta > 0$, (i) the operator $\mathscr{R}$ can be represented as $\mathscr{R} = \mathscr{R}_{s,\eta} + \mathscr{R}_{c,\eta}$, where $\|\mathscr{R}_{s,\eta}\|_{\mathscr{H}^1(\Omega)\to\mathscr{H}^1(\Omega)} \to 0$ as $\eta \to \infty$, while $\mathscr{R}_{c,\eta} : \mathscr{H}^1(\Omega) \to \mathscr{H}^1(\Omega)$ is compact;*

*(ii) the operator $\mathscr{R}^{\bullet}$ can be represented as $\mathscr{R}^{\bullet} = \mathscr{R}^{\bullet}_{s,\eta} + \mathscr{R}^{\bullet}_{c,\eta}$, where $\|\mathscr{R}^{\bullet}_{s,\eta}\|_{\mathscr{H}^1(\Omega)\to L_2(\Omega)} \to 0$ as $\eta \to \infty$, while $\mathscr{R}^{\bullet}_{c,\eta} : \mathscr{H}^1(\Omega) \to L_2(\Omega)$ is compact.*

**Theorem 5.12** *Let Conditions 5.1 and 5.10 hold. Then the following operators are compact,*

$$\mathscr{R} : \mathscr{H}^1(\Omega) \to \mathscr{H}^1(\Omega), \quad \mathscr{R}^{\bullet} : \mathscr{H}^1(\Omega) \to L_2(\Omega). \tag{5.76}$$

## 5.5 Conclusions

When we replace the parametrix $(q^k, \boldsymbol{u}^k)$, in the second Green identity, it is possible to obtain the following integral representation formula for any $(p, \boldsymbol{v}) \in \mathscr{H}^{1,0}(\Omega; \mathscr{A})$

$$p + \mathscr{R}^{\bullet}\boldsymbol{v} - \Pi^{s}\boldsymbol{T}^{+}(p, \boldsymbol{v}) + \Pi^{d}\boldsymbol{\gamma}^{+}\boldsymbol{v} = \mathring{\mathscr{Q}}\mathscr{A}(p, \boldsymbol{v}) + (2-\alpha)\mu\,\mathrm{div}\boldsymbol{v} \quad \text{in } \Omega, \tag{5.77}$$

$$\boldsymbol{v} + \mathscr{R}\boldsymbol{v} - \boldsymbol{V}\boldsymbol{T}^{+}(p, \boldsymbol{v}) + \boldsymbol{W}\boldsymbol{\gamma}^{+}\boldsymbol{v} = \mathscr{U}\mathscr{A}(p, \boldsymbol{v}) - \mathscr{Q}\,\mathrm{div}\boldsymbol{v} \quad \text{in } \Omega. \tag{5.78}$$

The proof follows the argument [FrMi19, Theorem 5.1]. We note that the solution is represented in terms of integral operators defined on the boundary and also on the domain. Using these identities, it is possible to derive integral equation systems, defined on the boundary and the domain, from a given boundary value problem. To study the existence and unique solvability of such integral equation systems, we will require the mapping properties analysed in this paper, see., e.g., [FrMi21] as an example in 3D.

The advantage of this method is that, sometimes, it is easier to study the equivalence between the BVP and the system of integral equations, and the existence of solution of such system than proving directly the existence of solution of the BVP directly.

## References

[AlAm00] Alliot, F., Amrouche, C.: Weak solutions for the exterior Stokes problem in weighted Sobolev spaces. Math. Methods Appl. Sci. **23**, 575–600 (2000)

[AyDa20] Ayele, T.G., Dagnaw, M.A.: Boundary-domain integral equation systems to the Dirichlet and Neumann problems for compressible Stokes equations with variable viscosity in 2D. Math. Methods Appl. Sci. **44**, 9876–9898 (2021). https://doi.org/10.1002/mma.6476

[CMN13] Chkadua, O., Mikhailov, S.E., Natroshvili, D.: Analysis of direct segregated boundary-domain integral equations for variable-coefficient mixed BVPs in exterior domains. Anal. Appl. **11**, 1350006, 33 pp. (2013). https://doi.org/10.1142/S0219530513500061

[FrMi19] Fresneda-Portillo, C., Mikhailov, S.E.: Analysis of boundary-domain integral equations to the mixed BVP for a compressible Stokes system with variable viscosity. Commun. Pure Appl. Anal. **18**, 3059–3088. https://doi.org/10.3934/cpaa.2019137

[FrMi21] Fresneda-Portillo, C., Mikhailov, S.E.: Boundary-domain integral equations equivalent to an exterior mixed BVP for the variable-viscosity compressible Stokes PDEs. Commun. Pure Appl. Anal. **20**, 1103–1133 (2021). https://doi.org/10.3934/cpaa.2021009

[Ha71] Hanouzet, B.: Espaces de Sobolev avec Poids. Application au probleme de Dirichlet dans un demi espace. Rendiconti del Seminario Matematico della Universita di Padova **46**, 227–272 (1971)

[HsWe08] Hsiao, G.C., Wendland, W.L.: Boundary Integral Equations. Springer, Berlin (2008). https://doi.org/10.1007/978-3-540-68545-6

[KLMW16] Kohr, M., Lanza de Cristoforis, M., Mikhailov, S.E., Wendland, W.L.: Integral potential method for transmission problem with Lipschitz interface in $\mathbb{R}^3$ for the Stokes and Darcy-Forchheimer-Brinkman PDE systems. Z. Angew. Math. Phys. **67**, 116, 30pp. (2016). https://doi.org/10.1007/s00033-016-0696-1

[La69] Ladyzhenskaya, O.A.: The Mathematical Theory of Viscous Incompressible Flow. Gordon & Breach, New York (1969)

[Li73] Lions, J.L., Magenes, E.: Non-Homogeneous Boundary Value Problems and Applications. Springer, Berlin (1973)

[Mc00] McLean, W.: Strongly Elliptic Systems and Boundary Integral Equations. Cambridge University Press, Cambridge (2000)

[Mi11] Mikhailov, S.E.: Traces, extensions and co-normal derivatives for elliptic systems on Lipschitz domains. J. Math. Anal. Appl. **378**, 324–342 (2011). https://doi.org/10.1016/j.jmaa.2010.12.027

[Ne01] Nedelec, J.: Acoustic and Electromagnetic Equations. Springer, New York (2001). https://doi.org/10.1007/978-1-4757-4393-7

[Sa14] Sayas, F.J., Selgas, V.: Variational views of stokeslets and stresslets. SeMA **63**, 65–90 (2014)

# Chapter 6
# Stochastic Effects of the Meander on the Dispersion of Pollutants in the Planetary Boundary Layer Under Low Wind Conditions


**C. Fávero, G. A. Gonçalves, D. Buske, and R. S. Quadros**

## 6.1 Introduction

Outdoor air pollution is responsible for some of three million deaths around the world [WHO16]. In addition to this problem, it brings considerable damage to ecosystems and thus economic losses. These are only some of the reasons why it is important to study the dispersion of pollutants in the planetary boundary layer of the Earth's atmosphere. To this end, mathematical models are widely used to estimate the concentration of pollutants in the planetary boundary layer and in domains with horizontal extensions of micrometeorological scales. For regulatory purposes, conventional models such as a Gaussian plume model are known to provide acceptable results for many stability conditions of the atmosphere, except for those where the wind speed is below $\sim$2 m/s, henceforth called low wind conditions [GoKr02].

Under these conditions, the diffusive process is dominated besides the turbulent diffusivity also by a spread in the plume due to meandering. So far, to the best of our knowledge, there do not exist pollution dispersion models in three dimensions which take into account effects due to meandering. In the present approach, we consider scenarios with wind speeds below 2 m/s and include in the lateral dispersion of the plume the effect of meandering based on the discussion in reference [AnEtAl06]. Once the diffusive process of dispersion together with the effect of meander is implemented in a model, one shall expect that results will approach better reality, an essential quality for simulation applications by the regulatory authorities.

Accordingly, the objective of the present study is to investigate the performance of a newly developed model with analytical solution of the time-dependent

C. Fávero · G. A. Gonçalves · D. Buske (✉) · R. S. Quadros
Federal University of Pelotas, Pelotas, RS, Brazil
e-mail: camilafavero@msn.com



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_6

three-dimensional advection–diffusion equation. The proposed model contemplates turbulent diffusive parameters for the three spatial dimensions, in contrast to models that commonly neglect the longitudinal turbulent diffusivity in comparison to the larger longitudinal wind speed in the advection term, an approximation no longer valid for low wind conditions. In addition, the effect of meandering in the wind field is taken care of upon inserting fluctuations in the longitudinal and transverse wind speeds, respectively.

## 6.2 The Advection–Diffusion Model

We start our developments from the full space and time-dependent advection–diffusion equation in Cartesian coordinates, where without restricting generality the mean wind direction is aligned with the $x$-axis and the $z$ coordinate extends from the ground level up to the planetary boundary height. As a simplification, we assume the terrain to be flat, so that the vertical component of the wind field may be neglected (see, for instance, reference [Ar99]).

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} + v\frac{\partial C}{\partial y} = \frac{\partial}{\partial z}\left(K_z \frac{\partial C}{\partial z}\right).$$

Here, $C$ is the average concentration, $u$ is the mean wind speed in the $x$ direction, and $K_z$ is the eddy diffusivity in the vertical direction, where usually besides mechanical also thermal forcings generate turbulence and convection.

$$K_z \frac{\partial C}{\partial z} = 0 \qquad \text{at} \quad z = h \text{ and } z = z_0.$$

Here, $h$ is the boundary layer height and $z_0 > 0$ is the average ground level determined by the root-mean-square value of the surface roughness. The advection–diffusion equation is subject to boundary conditions that prescribe zero flux at ground level and at the top of the planetary boundary layer and assume a clean atmosphere as initial condition and the pollution source with emission rate $Q$ is a point source located at height $z = H_s$ which starts operating at $t = 0$.

$$uC(0, x, y, z) = Q(0)\delta(x)F(y)\delta(z - H_s)$$

$$u \lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} C(t, x, y, z)\delta(x)\, dx = Q(t)F(y)\delta(z - H_s) \lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} \delta(x)\, dx.$$

To obtain a solution for the advection–diffusion equation, the method of variable separation was used. Initially, $u$ and $v$ were considered constant and turbulent

diffusivity $K_z = K_z(z)$ in the vertical direction only.

$$\frac{d}{dt}T_{\kappa\lambda} = (\kappa - \lambda)T_{\kappa\lambda}$$

$$u\frac{d}{dx}X_{\alpha\lambda} = (\alpha + \lambda)X_{\alpha\lambda}$$

$$v\frac{d}{dy}Y_{\kappa} = -\kappa Y_{\kappa}$$

$$\frac{d}{dz}\left(K(z)\frac{d}{dz}Z_{\alpha}\right) = \alpha Z_{\alpha}.$$

For details of the solution derivation, see reference [GoEtAl18].

### 6.2.1 A Time-Dependent Solution

In the present approach and for convenience, the concentration $C$ is factorized by three functions, the factors containing one of the horizontal coordinates are time dependent in view of the extension by the meandering, and the third factor contains the coordinates of the mean wind velocity direction and the vertical component, where turbulent diffusion is the only driving force for dispersion.

$$C(t, x, y, z) = \psi(t, y)\varphi(t, x)\xi(x, z).$$

The function $\xi(x, z)$ is the result of the solution of the diffusive term by the integral transform approach thoroughly outlined in reference [BuEtAl12], while $\psi(t, y)$ and $\varphi(t, x)$ are given by the spectral composition

$$\psi(t, y) = \int_0^{\infty} A(\kappa)\psi_{\kappa}(t, y)\, d\kappa\ ,$$

$$\varphi(t, x) = \int_0^{\infty} B(\lambda)\varphi_{\lambda}(t, x)\, d\lambda\ .$$

To determine the coefficients $A(\kappa)$ and $B(\lambda)$, one makes use of the initial and the source conditions, which establish the following equalities:

$$uC(0, x, y, z) = Q(0)\delta(x)F(y)\delta(z - H_s) = u\psi(0, y)\varphi(0, x)\xi(x, z)$$

$$u \lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} C(t, x, y, z)\delta(x)\, dx = Q(t)F(y)\delta(z - H_s) \lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} \delta(x)\, dx$$
$$= u\psi(t, y) \lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} \varphi(t, x)\xi(x, z)\delta(x)\, dx.$$

Upon replacing the generic terms by explicit forms of the eigenfunctions after having solved the equations above, and further using the initial condition at $t = 0$ and the source condition at $x = 0$, the results are given by

$$\psi(t, y) = \int_0^{\infty} \mathscr{L}^{-1}\{F(y), y \to \eta\} e^{-\eta(y - vt)}\, d\eta = F(y - vt)\ ,$$

$$\varphi(t, x) = \int_0^{\infty} \mathscr{L}^{-1}\{Q(t), t \to \lambda\} e^{-\lambda(t - \frac{x}{u})}\, d\lambda = Q\left(t - \frac{x}{u}\right)\ ,$$

where $\mathscr{L}^{-1}$ has the form of a Laplace transform. Then, the solution that expresses the concentration flux of a substance emitted by a point source is given by

$$uC(t, x, y, z) = F(y - vt)Q\left(t - \frac{x}{u}\right)\xi(x, t)\ .$$

Assuming that $F$ and $Q$ are represented by Dirac delta functionals, it is possible to find solutions with different forms of time-dependent sources by superimposing instantaneous contributions to the total concentration. If the source has the time evolution of a Heaviside function $H(t)$, then the solution is

$$C(t, x, y, z) = \int_0^{\infty} H(t - \tau)C(t, \tau, x, y, z)\, d\tau$$
$$= \frac{2}{\sqrt{16\pi^2 K_y K_x t^2}} \int_{t_0}^{t} uQ(\tau) e^{-\frac{(u(t-\tau)-x)^2}{\frac{4K_x x}{u}}} e^{-\frac{(v(t-\tau)-y)^2}{\frac{4K_y x}{u}}}\, d\tau\ \xi(x, t)\ .$$

With this finding, the expression for the concentration computation is suitable for micrometeorological conditions that vary in a time interval, which in principle can be arbitrarily chosen, but for practical applications is usually determined by the inverse sampling frequency of the data acquisition system in air quality monitoring installations. Thus, the plume will be composed of segments $C_j$, for each interval

$$Q(t) = \frac{2Qu}{\pi D} \int_{t_0}^{t} e^{-\frac{(A\tau^2 + B\tau + C)}{D}}\, d\tau = \frac{2Qu}{\pi} \sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} \frac{1}{D_j} e^{-\frac{(A\tau^2 + B\tau + C)}{D}}\, d\tau,$$

where $A = u^2 + v^2$, $B = 2u(x - ut) + 2v(y - vt)$, $C = (x - ut)^2 + (y - vt)^2$ and $D = 16(\frac{x}{u})^2 K_x K_y$. The index $j$ indicates the $j$-th interval, where the total time

interval covered extends from $t_0 = 0$ to $t_N = t$. The solution for this equation is given below:

$$C(t,x,y,z) = \sum_{j=1}^{N} \frac{Q_j u_j}{\sqrt{\pi D_j A_j}} e^{-\frac{4C_j A_j + B_j^2}{4A_j D_j}} \times$$
$$\times \left( \text{erf}\left( \frac{2A_j t_j + B_j}{2\sqrt{A_j D_j}} \right) - \text{erf}\left( \frac{2A_j t_{j-1} + B_j}{2\sqrt{A_j D_j}} \right) \right) \xi(x,t).$$

### 6.2.2 *Fluctuations in the Horizontal Wind Velocity Field*

To relate the Eulerian autocorrelation functions with the influence of the meandering due to the low wind velocity, the following equations were used [AnEtAl06], which introduce fluctuations in the horizontal wind velocity components and thus giving the model a stochastic character.

$$u'(t + \Delta t) = u(t) - (pu' + qv')\, dt + \sigma_u \sqrt{2pdt}\xi_u$$
$$v'(t + \Delta t) = v(t) - (qu' + pv')\, dt + \sigma_v \sqrt{2pdt}\xi_v.$$

Here, the terms $\xi_u$ and $\xi_v$ are random Gaussian variables with zero mean and variance equal to unity. For the purpose of generating numerical results, $p$ and $q$ were estimated according to the following expressions proposed by [Fr53]:

$$p = \frac{1}{(m^2+1)T},$$
$$q = \frac{m}{(m^2+1)T},$$

where $T$ and $m$ are determined according to the prescription in reference [CaEtAl06].

$$T_* = \frac{2\pi(m^2+1)T}{m}$$
$$m = \frac{T_* + \sqrt{T_*^2 - 16\pi^2 T^2}}{4\pi T}.$$

Experimental data show that the value of the meander's period represented by $T_*$ is approximately 2000 s regardless of the stability class of the atmosphere's boundary layer. Experimental values for $p$ and $q$ were also used in the performed simulations and will be presented in the next section.

## 6.3 The INEL Experiment

Reference [SaDi74], known as the INEL *(Idaho National Engineering Laboratory)* report, provides the results of a series of 14 diffusion experiments conducted under stable micrometeorological boundary layer conditions with slow winds over a flat terrain. Because of the wind direction variability, a 360° sampling grid was necessary. Arcs were defined at distances of 100, 200, and 400 m from the center of the grid, and collectors were positioned in intervals of 6° along each arc amounting to a total of 180 sampling points. An $SF_6$ tracer was released at a height of 1.5 m, and the collectors were mounted at 0.76 m above the ground level. Every hour the average concentrations were read out as determined by electron capture chromatography. These data were used to evaluate the model, and thus simulations were carried out using the runs of the INEL experiment which presented wind speeds below 1 m/s as shown in Table 6.1. The meander specifications by the values for $p$ and $q$ are presented in Table 6.2 and were based on the data recorded in the INEL experiment and calculated by [St17].

**Table 6.1** Measured data from the INEL experiment for respective mean wind speeds ($u$) and standard deviations of the wind direction ($\sigma_\theta$)

| Run | Quantity | 2 m | 4 m | 8 m | 16 m |
|---|---|---|---|---|---|
| 4 | $u$ (m/s) | 0.7 | 1.2 | - | 1.5 |
| | $\sigma_\theta$(°) | 13.6 | 12.0 | 7.7 | 11.5 |
| 5 | $u$ (m/s) | 0.8 | 0.9 | 1.2 | 2.2 |
| | $\sigma_\theta$(°) | 28.4 | 28.4 | 22.3 | 16.6 |
| 7 | $u$ (m/s) | 0.6 | 0.9 | 0.4 | 0.5 |
| | $\sigma_\theta$(°) | 23.9 | 22.3 | 34.4 | 20.1 |
| 8 | $u$ (m/s) | 0.5 | 0.8 | 0.6 | 1.2 |
| | $\sigma_\theta$(°) | 19.6 | 72.1 | 25.5 | 15.3 |
| 9 | $u$ (m/s) | 0.5 | 0.5 | 0.9 | 1.6 |
| | $\sigma_\theta$(°) | 21.4 | 17.9 | 14.6 | 13.9 |
| 12 | $u$ (m/s) | 0.7 | 1.1 | 1.1 | 1.6 |
| | $\sigma_\theta$(°) | 28.8 | 60.2 | 92.6 | 74.2 |

**Table 6.2** Calculated values $p$ and $q$ for the respective runs of the INEL experiment

| Run | $p_u(s^{-1})$ | $q_u(s^{-1})$ | $p_v(s^{-1})$ | $q_v(s^{-1})$ |
|---|---|---|---|---|
| 4 (4 m) | 0.007556 | 0.030121 | 0.006515 | 0.012815 |
| 7 | 0.008302 | 0.021958 | 0.002489 | 0.009445 |
| 8 | 0.004295 | 0.009744 | 0.001799 | 0.008073 |
| 9 | 0.003513 | 0.010404 | 0.001335 | 0.006510 |
| 10 (4 m) | 0.002996 | 0.011516 | 0.001003 | 0.008080 |
| 11 | 0.001321 | 0.006060 | 0.001471 | 0.005338 |
| 12 | 0.012110 | 0.031390 | 0.010650 | 0.011440 |
| 13 | 0.001338 | 0.007407 | 0.001040 | 0.005696 |
| 14 | 0.002029 | 0.004640 | 0.001326 | 0.005390 |

## 6.4 Results and Discussion

So far, a theoretical treatise and the referenced experimental evidences were presented apart, so that in the next step the appropriateness of the model against data from measurement is in order. To this end, three different considerations were put on as to insert fluctuations in the model and in agreement with the data from the experimental runs summarized in Table 6.3.

To further validate the model, we also used traditional statistical indices proposed by the author of reference [Ha89]. The results for the three scenarios can be seen in Table 6.4. By inspection, one may assert that the model reproduces reasonably well the observed concentrations once the results indicate an acceptable correlation factor (COR $\gtrsim 0.8$), as well as a reasonably small normalized mean square error (NMSE $\lesssim 0.4$) and standard deviation difference (FS $\lesssim 0.2$). If the results are analyzed by arc, it is possible to infer that the simulation fidelity is best at the arc nearest to the source, what is welcome for dispersion under low wind conditions because pollutant concentrations propagate off less from the point of emission. Comparing the statistical indices in Table 6.4 between the evaluated scenarios, no significant conclusion may be drawn, which may be rooted in the fact that the underlying model is a deterministic one although modified by a stochastic component. Thus, without accounting for fluctuations by a turbulent diffusive variance in the model, the additional effects by meandering may not be revealed with contrast, since all of the cases generated satisfactory results.

At this point, it is appropriate to point out that the statistical evaluation of the presented simulations is right from the beginning limited by the fact that simulated mean concentration values are compared to single samples from experiment, which belong to an unknown distribution so that differences between prediction and

**Table 6.3** Simulated scenarios

| Simulation | Fluctuations at $u$ and $v$ | Parameters $p$ and $q$ |
|---|---|---|
| $C_{p1}$ | $u'(t+\Delta t) = \rho_u u'(t) + \sigma_u(1-\rho_u{}^2)^{1/2}\chi$ | Absent |
| | $v'(t+\Delta t) = \rho_v v'(t) + \sigma_v(1-\rho_v{}^2)^{1/2}\chi$ | |
| $C_{p2}$ | $u'(t+\Delta t) = u(t) - (pu' + qv')dt + \sigma_u\sqrt{2pdt}\xi_u$ | Section 6.2.2 |
| | $v'(t+\Delta t) = v(t) - (qu' + pv')dt + \sigma_v\sqrt{2pdt}\xi_v$ | |
| $C_{p3}$ | $u'(t+\Delta t) = u(t) - (pu' + qv')dt + \sigma_u\sqrt{2pdt}\xi_u$ | Table 6.2 |
| | $v'(t+\Delta t) = v(t) - (qu' + pv')dt + \sigma_v\sqrt{2pdt}\xi_v$ | |

**Table 6.4** Traditional statistical evaluation of the model

| Simulation | NMSE | COR | FS |
|---|---|---|---|
| Simulation $C_{p1}$ | 0, 30 | 0, 81 | 0, 01 |
| Simulation $C_{p2}$ | 0, 27 | 0, 82 | 0, 03 |
| Simulation $C_{p3}$ | 0, 25 | 0, 82 | 0, 03 |
| Simulation $C_{p1} < 1$ m/s | 0, 23 | 0, 88 | 0, 14 |
| Simulation $C_{p2} < 1$ m/s | 0, 24 | 0, 87 | 0, 17 |
| Simulation $C_{p3} < 1$ m/s | 0, 25 | 0, 87 | 0, 20 |

**Table 6.5** Experimentally observed concentrations ($C_o$) and predicted ones by the model ($C_{p1}$, $C_{p2}$ e $C_{p3}$) for the arcs of 100 m, 200 m, and 300 m of the INEL experiment. The concentrations were normalized by emission rate ($C/Q$)

| Exp. | Dist.(m) | $C_o$ | $C_{p1}$ | $C_{p2}$ | $C_{p3}$ |
|---|---|---|---|---|---|
| 4 | 100 | 5.81 | 4.80 | 4.71 | 4.70 |
| | 200 | 2.99 | 2.23 | 2.23 | 2.17 |
| | 400 | 1.47 | 1.11 | 1.12 | 1.06 |
| 5 | 100 | 1.36 | 1.64 | 1.60 | - |
| | 200 | 0.87 | 0.59 | 0.63 | - |
| | 400 | 0.30 | 0.30 | 0.32 | - |
| 7 | 100 | 1.26 | 2.10 | 2.11 | 2.10 |
| | 200 | 0.71 | 0.77 | 0.81 | 0.76 |
| | 400 | 0.33 | 0.39 | 0.40 | 0.40 |
| 8 | 100 | 0.59 | 1.24 | 1.33 | 1.40 |
| | 200 | 0.32 | 0.37 | 0.42 | 0.41 |
| | 400 | 0.33 | 0.15 | 0.15 | 0.17 |
| 9 | 100 | 1.09 | 2.75 | 2.69 | 2.71 |
| | 200 | 0.57 | 1.19 | 1.21 | 1.23 |
| | 400 | 0.39 | 0.59 | 0.63 | 0.63 |
| 10 | 100 | 2.41 | 2.02 | 2.05 | 2.06 |
| | 200 | 1.80 | 0.52 | 0.57 | 0.51 |
| | 400 | 0.71 | 0.24 | 0.25 | 0.25 |
| 11 | 100 | 2.32 | 1.41 | 1.42 | 1.89 |
| | 200 | 1.09 | 0.30 | 0.34 | 0.43 |
| | 400 | 1.10 | 0.10 | 0.10 | 0.12 |
| 12 | 100 | 2.00 | 2.13 | 2.17 | 2.63 |
| | 200 | 1.77 | 0.98 | 1.01 | 1.04 |
| | 400 | 0.99 | 0.50 | 0.47 | 0.51 |
| 13 | 100 | 3.19 | 4.21 | 4.17 | 4.16 |
| | 200 | 2.30 | 1.22 | 1.22 | 1.25 |
| | 400 | 1.37 | 0.60 | 0.60 | 0.61 |
| 14 | 100 | 2.81 | 2.65 | 2.75 | 2.70 |
| | 200 | 1.59 | 0.79 | 0.79 | 0.79 |
| | 400 | 0.30 | 0.38 | 0.38 | 0.39 |

observation are to be expected. Moreover, the introduced stochastic component in the model should be implemented in such a way as to avoid random numbers close to the recently generated one so that more distant values are more likely than neighboring ones. In the present implementation, we evaluated the performance with unpredictably varying fluctuations in the horizontal wind velocity components. Nevertheless, it is well known from reports in the literature that the meander has a significant effect on pollutant dispersion, especially close to the source [OeEtAl06] and according to [AnEtAl06] and [ShEtAl02], if the effect is not correctly represented, then the concentration values are typically overestimated. However, one has to question the validity of Fick's closure in the vicinity of sources, which was one of the premises that lead to the advection–diffusion model.

Nonetheless, it is possible to see in the results presented in Table 6.5 that for this study the values of observed and predicted concentrations show a fairly good agree-

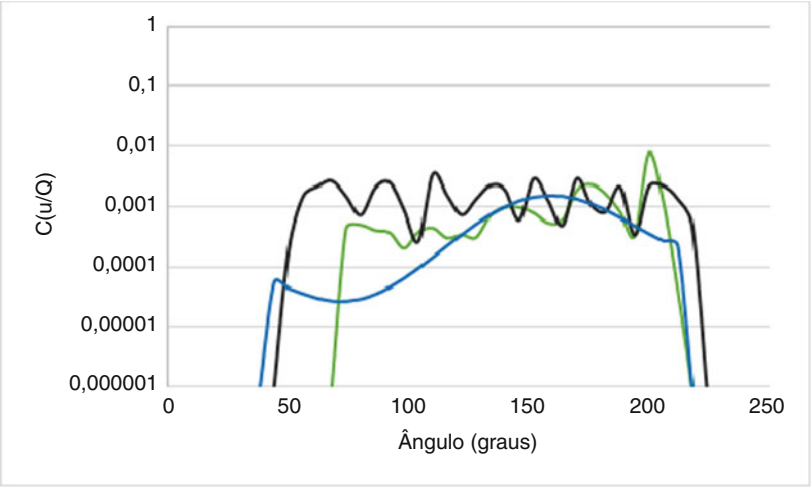



**Fig. 6.1** Scatter plot of observed concentrations (green line), predicted with the presence of meander (black line) and predicted without the presence of meander (blue line) for the 100 m arc of run 11 of the INEL experiment

ment. In order to take a more detailed look into comparisons between prediction and observation involving meandering, we show concentrations of pollutants for almost all low wind condition runs of the INEL experiment shown in Table 6.5. There are no calculated values of $p$ and $q$ for run 5 so that the corresponding $C_{p3}$ values were not calculated. Figure 6.1 shows the data for run 11, the purely deterministic simulation together with the simulation with the stochastic component which shall mimic some aspects of meandering. The purely deterministic model has apparently no pronounced oscillations in the angular distribution, while the simulation with the stochastic component shows a similar pattern with fluctuations compared to the experimental data. One effect that was to be expected did not occur, namely the increase in the spread of the angular distribution in comparison to the purely deterministic results although both simulations show a larger spread of angular values anyway. It may well be that in order to reveal meandering effects, periods with a more stable wind direction shall be chosen so that the afore discussed distributional effects become more apparent.

## 6.5 Conclusion

The solution in analytical form presented by the model facilitates the understanding and description of the physical phenomena involved in the problem, since it explicitly considers all involved parameters, either physical or phenomenological. Thus, the concentration can be obtained at any time and requiring little computational

effort. Recalling the fact that the employed model is Eulerian as obtained from the advection–diffusion equation, the insertion of new diffusive coefficients in the horizontal directions is a necessary adaptation for dispersion simulation in low wind conditions due to the presence of a new effect, namely meandering. In addition, the formulation of the latter was implemented by inserting stochastic variables, which in turn makes the mathematical representation approach the phenomenon observed in nature but still not fully understood. Furthermore, in comparison to the approaches reported in the literature, a formal advance was accomplished by avoiding the usually necessary numerical inversion of the Laplace transform in the time variable.

## References

[AnEtAl06] Anfossi, D., Alessandrini, S., Casteli, S.T., Ferrero, E.,Oettl, D., Degrazia, G.: Tracer dispersion simulation in low wind speed conditions with a new 2D Langevin equation system. Atmos. Environ. **40**, 7234–7245 (2006)

[Ar99] Arya, S.P.: Air Pollution Meteorology and Dispersion. Oxford University Press, New York (1999)

[BuEtAl12] Buske, D., Vilhena, M.T., Tirabassi, T., Bodmann, B.: Air pollution steady-state advection-diffusion equation: the general three-dimensional solution. J. Environ. Prot. **3**(9A), 1124–1134 (2012), https://doi.org/10.4236/jep.2012.329131

[CaEtAl06] Carvalho, C.J, Degrazia, A.G., Vilhena, T.M., Magalhaes, G.S., Goulart G.A., Anfossi, D., Acevedo C.O., Moraes L.L.O.: Parameterization of meandering phenomenon in a stable atmospheric boundary layer. Physica A **368**, 247–256 (2006)

[Fr53] Frenkiel, F.: Flow field of homogeneous turbulence. Adv. Appl. Mech. **3**, 61 (1953)

[GoEtAl18] Goncalves, G.A., Buske, D., Quadros, R.S., Weymar, G.J.: A new approach to solve the time-dependent three-dimensional advection-diffusion equation applied to model air pollution dispersion in the planetary boundary layer. Int. J. Develop. Res. **8**(5), 20535–20543 (2018)

[GoKr02] Goyal, P., Krishna, T.V.B.P.S.R: Dispersion of pollutants in convective low wind: a case study of Delhi. Atmos. Environ. **36**, 2071–2079 (2002)

[Ha89] Hanna, S.R.: Confidence limit for air quality models as estimated by bootstrap and jackknife resampling methods. Atmos. Environ. **23**, 1385–1395 (1989)

[OeEtAl06] Oettl, D., Goulart, A., Degrazia, G., Anfossi, D.: A new hypothesis on meandering atmospheric flows in low wind speed conditions. Atmosp. Environ. **39**, 1739–1748 (2006)

[SaDi74] Sagendorf, J.F., Dickson, C.R.: Diffusion under low wind-speed, inversion conditions. U. S. National Oceanics and Atmospherics Administration, Technical Memorandum, ERL ARL-52 (1974)

[ShEtAl02] Sharan, M., Yadav, A.K., Modani, M.: Simulation of short-range diffusion experiment in low wind convective conditions. Atmosp. Environ. **36**, 1901–1906 (2002)

[St17] Stefanello, M.B.: Development of a Lagrangian model to estimate dispersion of passive scalars in meandering conditions of the horizontal wind. (Master Dissertation, in Portuguese) Federal University of Santa Maria, Santa Maria, RS (2017)

[WHO16] World Health Organization: Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease, p. 121. World Health Organization, Geneva (2016)

# Chapter 7
# Asymptotics for the Spectrum of a Floquet-Parametric Family of Homogenization Problems Associated with a Dirichlet Waveguide


**D. Gómez, S. A. Nazarov, R. Orive-Illera, and M.-E. Pérez-Martínez**

## 7.1 Introduction

In this chapter, we consider a parametric family of spectral problems for the Laplace operator in a rectangular perforated domain $\varpi^\varepsilon$. The perforations are periodically placed along the ordinate axis at a distance $O(\varepsilon)$ between them, where $\varepsilon$ is a small parameter $\varepsilon \ll 1$, see Fig. 7.1a. We impose Dirichlet conditions on the boundary of the perforation and on the horizontal sides of the rectangle, while we impose *quasi-periodicity* conditions on the lateral sides containing the so-called *Floquet parameter* $\eta \in [-\pi, \pi]$. This parametric family arises as the model problem of a spectral problem posed in an unbounded strip periodically perforated by a string of holes, which is referred to as *perforation string*, cf. Fig. 7.1b. For each $\eta \in [-\pi, \pi]$, the spectral problem in the periodicity cell $\varpi^\varepsilon$ is itself a homogenization problem, and we study the asymptotic behavior of the eigenvalues and eigenfunctions as $\varepsilon \to 0$. In this way, we revisit the spectral problem for the Dirichlet–Laplace operator in a perforated waveguide addressed in [NaOrPe19a], providing new results that complement those.

The setting of the perturbation spectral problem is in Sect. 7.1.1; the homogenized problem is in Sect. 7.1.2, while the state of the art is in Sect. 7.1.3. Our aim is to study the asymptotic behavior of the spectrum as $\varepsilon \to 0$ at the same time that we

D. Gómez · M.-E. Pérez-Martínez (✉)
Universidad de Cantabria, Santander, Spain
e-mail: gomezdel@unican.es; meperez@unican.es

S. A. Nazarov
St. Petersburg State University, St. Petersburg, Russia

R. Orive-Illera
Universidad Autónoma de Madrid, Madrid, Spain
e-mail: rafael.orive@icmat.es



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_7

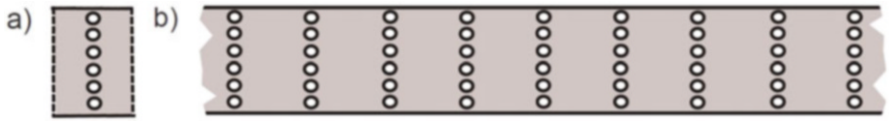



**Fig. 7.1** (**a**) The perforated domain $\varpi^\varepsilon$. (**b**) The perforated strip $\Pi^\varepsilon$

provide precise bounds for convergence rates which are uniform in both parameters $\varepsilon$ and $\eta$. This is in Sect. 7.3. Some preliminary results obtained in [NaOrPe19a] and [GoEtAl21] are stated in Sect. 7.2.

### *7.1.1 The Parametric Family of Homogenization Spectral Problems*

Let $\omega$ be a domain in the plane $\mathbb{R}^2$ which is bounded by a smooth simple closed curve $\partial\omega$ and has the compact closure $\overline{\omega} = \omega \cup \partial\omega \subset \varpi^0$, where $\varpi^0$ is the rectangle

$$\varpi^0 = (-1/2, 1/2) \times (0, H). \tag{7.1}$$

We introduce the perforated domain $\varpi^\varepsilon$, see Fig. 7.1a, obtained from $\varpi^0$ by removing the family of holes

$$\omega^\varepsilon(k) = \{x : \varepsilon^{-1}(x_1, x_2 - \varepsilon k H) \in \omega\}, \quad k = 0, \dots, N-1,$$

which are distributed periodically along the ordinate $x_2$-axis. Each hole is homothetic to $\omega$ of ratio $\varepsilon$ and translation of $\varepsilon\omega = \omega^\varepsilon(0)$; namely,

$$\varpi^\varepsilon = \varpi^0 \setminus \overline{\omega^\varepsilon} \quad \text{where} \quad \omega^\varepsilon = \bigcup_{k=0}^{N-1} \omega^\varepsilon(k). \tag{7.2}$$

Here, $\varepsilon$ is a small positive parameter and $N$ is a big natural number, both related by $N = \varepsilon^{-1}$. The period is $\varepsilon H$ with $\varepsilon \ll 1$.

In the domain $\varpi^\varepsilon$, we consider the spectral problem defined by the equations

$$-\Delta U^\varepsilon(x;\eta) = \Lambda^\varepsilon(\eta) U^\varepsilon(x;\eta), \quad x \in \varpi^\varepsilon, \tag{7.3}$$

$$U^\varepsilon(x;\eta) = 0, \quad x \in \Gamma^\varepsilon, \tag{7.4}$$

$$U^\varepsilon(1/2, x_2;\eta) = e^{i\eta} U^\varepsilon(-1/2, x_2;\eta), \quad x_2 \in (0, H), \tag{7.5}$$

$$\frac{\partial U^\varepsilon}{\partial x_1}(1/2, x_2;\eta) = e^{i\eta} \frac{\partial U^\varepsilon}{\partial x_1}(-1/2, x_2;\eta), \quad x_2 \in (0, H), \tag{7.6}$$

where

$$\Gamma^\varepsilon = \partial\varpi^\varepsilon \setminus \{\pm 1/2\} \times (0, H),$$

$\eta$ is the dual variable, namely, *the Floquet parameter*. $\Lambda^\varepsilon(\eta)$ and $U^\varepsilon(\cdot;\eta)$, respectively, denote the eigenvalues and eigenfunctions which depend on both the perturbation parameter and the Floquet parameter. Conditions (7.5)–(7.6) are the so-called *quasi-periodicity conditions* on the lateral sides $\{\pm 1/2\} \times (0, H)$ of $\varpi^\varepsilon$.

The variational formulation of the spectral problem (7.3)–(7.6) reads: find $\Lambda^\varepsilon(\eta)$ and $U^\varepsilon(\cdot;\eta) \in H^{1,\eta}_{per}(\varpi^\varepsilon;\Gamma^\varepsilon)$, $U^\varepsilon(\cdot;\eta) \neq 0$ satisfying

$$\left(\nabla U^\varepsilon(\cdot;\eta), \nabla V\right)_{\varpi^\varepsilon} = \Lambda^\varepsilon(\eta)\left(U^\varepsilon(\cdot;\eta), V\right)_{\varpi^\varepsilon} \qquad \forall V \in H^{1,\eta}_{per}(\varpi^\varepsilon;\Gamma^\varepsilon), \tag{7.7}$$

where $H^{1,\eta}_{per}(\varpi^\varepsilon;\Gamma^\varepsilon)$ denotes the subspace of $H^1(\varpi^\varepsilon)$ of functions which satisfy the quasi-periodicity conditions (7.5)–(7.6) and vanish on $\Gamma^\varepsilon$, and $(\cdot,\cdot)_{\varpi^\varepsilon}$ denotes the scalar product in $L^2(\varpi^\varepsilon)$.

As is well known (cf. [NaOrPe19a], Ch. 10 in [BiSo80], Ch. 13 in [ReSi78], and Ch. 4 in [SaSa89]), problem (7.7) has a discrete spectrum constituting the monotone unbounded sequence of eigenvalues

$$0 < \Lambda^\varepsilon_1(\eta) \leq \Lambda^\varepsilon_2(\eta) \leq \cdots \leq \Lambda^\varepsilon_m(\eta) \leq \cdots \to \infty, \qquad \text{as } m \to \infty, \tag{7.8}$$

which are repeated according to their multiplicities. Also, the corresponding eigenfunctions $\{U^\varepsilon_m(\cdot;\eta)\}_{m=1}^\infty$ are assumed to form an orthonormal basis in $L^2(\varpi^\varepsilon)$. Furthermore, the function

$$\eta \in [-\pi, \pi] \mapsto \Lambda^\varepsilon_m(\eta) \tag{7.9}$$

is continuous and $2\pi$-periodic. This last assertion is due to the fact that problem (7.3)–(7.6) is the model problem associated with a waveguide, which is referred to as the *Dirichlet strip*, and has been recently considered in the literature (cf. (7.20), Fig. 7.1b), [NaOrPe19a], and [NaOrPe19b]). For the sake of completeness, in order to outline the interest of the problem under consideration (7.3)–(7.6), as well as its properties, we introduce briefly this waveguide in Sect. 7.1.3.

### 7.1.2 The Homogenized Problem

For each $\eta \in [-\pi, \pi]$, the homogenized problem of (7.3)–(7.6) reads

$$-\Delta U^0(x;\eta) = \Lambda^0(\eta) U^0(x;\eta), \quad x \in \widetilde{\varpi}^0, \tag{7.10}$$

$$U^0(x;\eta) = 0, \quad x \in \Gamma_{lu0}, \tag{7.11}$$

$$U^0(1/2, x_2; \eta) = e^{i\eta} U^0(-1/2, x_2; \eta), \quad x_2 \in (0, H), \tag{7.12}$$

$$\frac{\partial U^0}{\partial x_1}(1/2, x_2; \eta) = e^{i\eta} \frac{\partial U^0}{\partial x_1}(-1/2, x_2; \eta), \quad x_2 \in (0, H), \tag{7.13}$$

where $\widetilde{\varpi}^0$ and $\Gamma_{lu0}$ denote

$$\widetilde{\varpi}^0 := (-1/2, 0) \times (0, H) \cup (0, 1/2) \times (0, H)$$

and

$$\Gamma_{lu0} := \{x \,:\, x_1 \in (-1/2, 1/2),\ x_2 \in \{0, H\}\} \cup \{x \,:\, x_1 = 0,\ x_2 \in (0, H)\}, \tag{7.14}$$

respectively, $\Lambda^0(\eta)$ is the spectral parameter, and $U^0(\cdot; \eta)$ is the corresponding eigenfunction.

The variational formulation of the spectral problem (7.10)–(7.13) reads: find $\Lambda^0(\eta)$ and $U^0(\cdot; \eta) \in H^{1,\eta}_{per}(\varpi^0; \Gamma_{lu0})$, $U^0(\cdot; \eta) \neq 0$ satisfying

$$\left(\nabla U^0(\cdot; \eta), \nabla V\right)_{\widetilde{\varpi}^0} = \Lambda^0(\eta) \left(U^0(\cdot; \eta), V\right)_{\widetilde{\varpi}^0} \qquad \forall V \in H^{1,\eta}_{per}(\varpi^0; \Gamma_{lu0}), \tag{7.15}$$

where $H^{1,\eta}_{per}(\varpi^0; \Gamma_{lu0})$ denotes the subspace of $H^1(\varpi^0)$ of functions which satisfy the quasi-periodicity conditions (7.12)–(7.13) and vanish on $\Gamma_{lu0}$. Similarly to (7.7), problem (7.15) has a discrete spectrum $\{\Lambda^0_m(\eta)\}_{m=1}^{\infty}$ with corresponding eigenfunctions $\{U^0_m(\cdot; \eta)\}_{m=1}^{\infty}$, which form an orthogonal basis in $L^2(\varpi^0)$.

Comparing the homogenization problem (7.3)–(7.6) with other homogenization problems having Dirichlet conditions on the boundary of the perforations, we see that it differs only in the quasi-periodicity boundary conditions on the lateral sides, and one can easily guess the homogenized problem (7.10)–(7.13), see, for instance, [LoEtAl98]. However, in this case, one can show that the eigenvalues coincide with those of the Dirichlet problem

$$\begin{aligned} &-\Delta U^0(x) = \Lambda^0 U^0(x), \quad x \in \upsilon, \quad \upsilon := (0, 1) \times (0, H), \\ &U^0(x) = 0, \quad x \in \partial\upsilon, \end{aligned} \tag{7.16}$$

and consequently, do not depend on $\eta$ (cf. [NaOrPe19a]).

Problem (7.16) has a discrete spectrum which forms the increasing sequence of eigenvalues

$$0 < \Lambda^0_1 < \Lambda^0_2 \leq \cdots \leq \Lambda^0_m \leq \cdots \to \infty, \quad \text{as } m \to \infty, \tag{7.17}$$

repeated according to their multiplicities. In addition, the eigenpairs of (7.16) can be computed explicitly

$$\Lambda^0_{np} = \pi^2 \left( n^2 + \frac{p^2}{H^2} \right), \quad U^0_{np}(x) = \frac{2}{\sqrt{H}} \sin(n\pi x_1) \sin(p\pi x_2/H), \quad p, n \in \mathbb{N}. \tag{7.18}$$

Note that the eigenvalues $\Lambda^0_{np}$ are numerated with two indexes and must be reordered in order to obtain the increasing sequence (7.17); the corresponding eigenfunctions $U^0_{np}$ are normalized in $L^2(\upsilon)$. Also, we note that if $H^2$ is an irrational number, all the eigenvalues are simple.

As noticed in [NaOrPe19a], extending by quasi-periodicity the eigenfunctions $U^0_m(\cdot;\eta)$,

$$u^0_m(x;\eta) = \begin{cases} U^0_m(x;\eta), & x_1 \in (0, 1/2), \\ e^{i\eta} U^0_m(x_1 - 1, x_2;\eta), & x_1 \in (1/2, 1), \end{cases} \tag{7.19}$$

we obtain a smooth function in the rectangle $\upsilon$, and the pair $(\Lambda^0_m(\eta),\, u^0_m(\cdot,\eta))$ satisfies (7.16).

The orthogonality of $\{U^0_m(\cdot;\eta)\}_{m=1}^{\infty}$ in $L^2(\varpi^0)$ implies that the functions in (7.19), $\{U^0_m(\cdot;\eta)\}_{m=1}^{\infty}$, form an orthogonal basis in $L^2(\upsilon)$, and this shows that the set $\{\Lambda^0_m(\eta)\}_{m=1}^{\infty}$ coincides with $\{\Lambda^0_m\}_{m=1}^{\infty}$ in the sequence (7.17) for any $\eta \in [-\pi, \pi]$.

By (7.18) and (7.19), we compute the eigenvalues and eigenfunctions of (7.10)–(7.13):

$$U^0_{np}(x,\eta) = \begin{cases} \frac{2}{\sqrt{H}} \sin(n\pi x_1) \sin(p\pi \frac{x_2}{H}), & x_1 \in (0, 1/2), \\ \frac{2e^{-i\eta}}{\sqrt{H}} \sin(n\pi(x_1 + 1)) \sin(p\pi \frac{x_2}{H}), & x_1 \in (-1/2, 0), \end{cases}$$

is the eigenfunction corresponding to $\Lambda^0_{np} = \pi^2 \left( n^2 + \frac{p^2}{H^2} \right)$ with $p, n \in \mathbb{N}$.

### 7.1.3 The Dirichlet Strip and Some Background

For convenience, we introduce here a problem closely related to (7.3)–(7.6): a Dirichlet problem for the Laplace operator in a strip with periodic dense transversal perforations by identical holes of diameter $\varepsilon$.

Extending $\varpi^\varepsilon$ (cf. (7.2) and Fig. 7.1a) by periodicity along the $x_1$-axis, we create the unbounded perforated strip $\Pi^\varepsilon$ (see Fig. 7.1b):

$$\Pi^\varepsilon = \mathbb{R} \times (0, H) \setminus \bigcup_{j \in \mathbb{Z}} \bigcup_{k=0}^{N-1} \overline{\omega^\varepsilon(j,k)},$$

where $\omega^\varepsilon(j,k) = \{x : \varepsilon^{-1}(x_1 - j, x_2 - \varepsilon k H) \in \omega\}$ with $j \in \mathbb{Z}$, $k = 0, 1, \ldots, N-1$. In the waveguide $\Pi^\varepsilon$, we consider the Dirichlet spectral problem

$$\begin{cases} -\Delta u^\varepsilon(x) = \lambda^\varepsilon u^\varepsilon(x), & x \in \Pi^\varepsilon, \\ u^\varepsilon(x) = 0, & x \in \partial\Pi^\varepsilon. \end{cases} \tag{7.20}$$

Then, applying the Floquet–Bloch–Gelfand transform

$$u^\varepsilon(x) \to U^\varepsilon(x;\eta) = \frac{1}{\sqrt{2\pi}} \sum_{n\in\mathbb{Z}} e^{-in\eta} u^\varepsilon(x_1 + n, x_2),$$

see, for instance, [Ge50], [ReSi78], [Sk85], [Ku93], and [CoPlVa94], problem (7.20) converts into an $\eta$-parametric family of spectral problems in the periodicity cell $\varpi^\varepsilon$, namely, into the parametric family of boundary value problems (7.3)–(7.6), see Fig. 7.1a.

The spectrum of the operator on the Hilbert space $L^2(\Pi^\varepsilon)$ associated with problem (7.20) is given by

$$\sigma^\varepsilon = \bigcup_{m\in\mathbb{N}} B_m^\varepsilon, \tag{7.21}$$

where

$$B_m^\varepsilon = \{\Lambda_m^\varepsilon(\eta) : \eta \in [-\pi, \pi]\}. \tag{7.22}$$

As a consequence of the previously mentioned continuity of $\Lambda_m^\varepsilon(\eta)$, cf. (7.9), the sets $B_m^\varepsilon$ are closed, connected, and bounded intervals of the real positive axis $\overline{\mathbb{R}_+}$.

Results (7.21) and (7.22) for the spectrum of the boundary value problem (7.20) are well known in the framework of the Floquet–Bloch–Gelfand theory (see the above references). The segments $B_m^\varepsilon$ and $B_{m+1}^\varepsilon$ may intersect or be disjoints so that a spectral gap may become open between them. Recall that a spectral gap is a non-empty interval which is free of the spectrum but has both end points in the spectrum.

Therefore, studying the asymptotic behavior of the spectrum of (7.3)–(7.6) becomes essential to detect the band gap structure of the spectrum (7.21). In this respect, an extensive asymptotic analysis of the spectral bands (7.22) has been performed in [NaOrPe19a]. In particular, we have obtained asymptotic formulas for the end points of the spectral bands (7.22) and show that $\sigma^\varepsilon$ has a long number of short bands of length $O(\varepsilon)$ which alternate with wide gaps of width $O(1)$, while we can guarantee that indeed there are open gaps corresponding with $B_m^\varepsilon$ and $B_{m+1}^\varepsilon$ only when the limit eigenvalue $\Lambda_m^0$ in the sequence (7.17) is simple, cf. Fig. 7.2 (on the right), and this strongly depends on $H$.

We note that the explicit formulas (7.18) are of great interest to draw the *limit dispersion curves* for different values of $H$ and, after obtaining bounds for discrepancies of the type (7.42) (cf. also (7.28)), they also allow us to draw possible

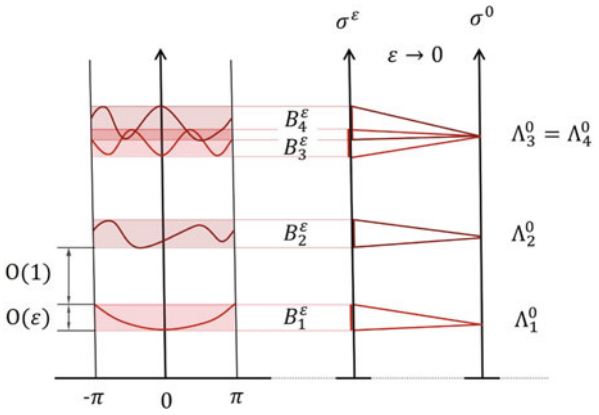



**Fig. 7.2** On the left: a sketch of possible dispersion curves in the axis $(\eta, \Lambda)$ for the problem in the waveguide $\Pi^\varepsilon$. On the right: a sketch of the possible distribution of the spectral bands $B^\varepsilon$

configurations of the perturbed dispersion curves associated with (7.20), cf. Fig. 7.2 (on the left). Recall that these curves are the graphs of $\Lambda_m^\varepsilon(\eta)$, for $\eta \in [-\pi, \pi]$. On account of (7.18), the limiting dispersion curves are independent of $\eta$.

We refer to [BaPe18] for a very different perturbed waveguide with limiting dispersion curves independent of the Floquet parameter and to [GoEtAl22a] and [GoEtAl22b] for the geometry of the waveguide here considered but with Neumann conditions instead of Dirichlet. Also, we refer to [GoEtAl22a] and [GoEtAl22b] for further references and an extensive comparison between the behaviors of the spectral bands when we change Dirichlet by Neumann conditions both in (7.20) and in (7.10)–(7.13). As a matter of fact, in the case of the Neumann strip, we find long bands, of order $O(1)$, which are separated from each other by short spectral gaps of order $O(\varepsilon)$. Moreover, it should be mentioned that, as a consequence of the fact that the limiting dispersion curves are not constant in the case of the Neumann strip, the asymptotic analysis is much more complicated and delicate, and in particular, it becomes multiscale in several variables, not only in the geometrical ones but also in the Floquet parameter.

Finally, let us observe that opening gaps in [NaOrPe19a] implies a thorough asymptotic analysis to obtain corrector terms of order $O(\varepsilon)$ that improves the uniform bounds (7.42). For the sake of brevity, we avoid defining the correctors here which involves introducing some boundary layer problems and the so-called *polarization matrix*. We refer to [NaOrPe19a] and [NaOrPe19b] in this connection.

## 7.2 Preliminary Results

Let us introduce here some estimates for the eigenvalues of the perturbation problem that improves that in [NaOrPe19a] and a couple of theorems whose proofs are in [NaOrPe19a]. The results of these theorems are improved in Sect. 7.3.

**Lemma 7.1** *For each fixed $m$, there are constants $\varepsilon_0 < 1$, $K_m(\eta)$, and $C_m$ such that*

$$0 < K_m(\eta) \le \Lambda_m^\varepsilon(\eta) \le C_m \qquad \forall \eta \in [-\pi, \pi], \quad \varepsilon \le \varepsilon_0. \tag{7.23}$$

***Proof*** To obtain the lower bound in (7.23) with $K_m(\eta) \equiv C$ independent of $m$ and $\eta$, it suffices to consider (7.7) for the eigenpair $(\Lambda_1^\varepsilon(\eta),\ U_1^\varepsilon(\cdot;\eta))$ and apply the Poincaré inequality in $H^1(\varpi^0)$ once that $U_1^\varepsilon(\cdot;\eta)$ is extended by zero in $\overline{\omega}^\varepsilon$, cf. (7.1) and (7.2). However, we can also obtain better bounds depending on $\eta$ that somehow could isolate the branches $\{\Lambda_m^\varepsilon(\eta) : \eta \in [-\pi, \pi]\}$.

Indeed, let us consider $\{\Lambda_m^*(\eta)\}_{m=1}^\infty$ to be the sequence of eigenvalues of the following problem in $\varpi^0$:

$$\begin{aligned}
&-\Delta U_m^*(x;\eta) = \Lambda_m^*(\eta) U_m^*(x;\eta), \quad x \in \varpi^0,\\
&U_m^*(x;\eta) = 0, \quad x \in \Gamma_{lu},\\
&U_m^*(1/2, x_2;\eta) = e^{i\eta} U_m^*(-1/2, x_2;\eta), \quad x_2 \in (0, H),\\
&\frac{\partial U_m^*}{\partial x_1}(1/2, x_2;\eta) = e^{i\eta} \frac{\partial U_m^*}{\partial x_1}(-1/2, x_2;\eta), \quad x_2 \in (0, H),
\end{aligned} \tag{7.24}$$

where we have denoted by $\Gamma_{lu}$ lower and upper basis of the rectangle $\varpi^0$, namely,

$$\Gamma_{lu} := \{x \ : \ x_1 \in (-1/2, 1/2),\ x_2 \in \{0, H\}\}, \tag{7.25}$$

cf. (7.14) to compare, and by $\{U_m^*(\cdot;\eta)\}_{m=1}^\infty$ the eigenfunctions.

Using the minimax principle,

$$\Lambda_m^*(\eta) = \min_{E_m \subset H^{1,\eta}_{per}(\varpi^0;\Gamma_{lu})} \max_{V \in E_m, V \neq 0} \frac{(\nabla V, \nabla V)_{\varpi^0}}{(V, V)_{\varpi^0}},$$

where the minimum is computed over the set of subspaces $E_m$ of $H^{1,\eta}_{per}(\varpi^0;\Gamma_{lu})$ with dimension $m$.

Consider the subspace $E_m^\varepsilon$ of $H^{1,\eta}_{per}(\varpi^\varepsilon;\Gamma^\varepsilon)$ with dimension $m$, of the eigenfunctions $U_k^\varepsilon(\cdot;\eta)$ of (7.3)–(7.6) associated with the eigenvalues $\Lambda_k^\varepsilon(\eta)$ in the sequence (7.8) with $k \le m$. These eigenfunctions have been taken to be orthonormal in $L^2(\varpi^\varepsilon)$ and are extended by 0 inside the holes, they are still denoted by $U_m^\varepsilon(\cdot;\eta)$ and orthonormal in $L^2(\varpi^0)$, and we take the particular subspace of dimension $m$ of $H^{1,\eta}_{per}(\varpi^0;\Gamma_{lu})$ to be the span $E_m^* = \left[U_1^\varepsilon(\cdot;\eta), U_2^\varepsilon(\cdot;\eta), \cdots U_m^\varepsilon(\cdot;\eta)\right]$. Then, we can write

$$\Lambda_m^*(\eta) \le \max_{V \in E_m^*, V \neq 0} \frac{(\nabla V, \nabla V)_{\varpi^0}}{(V, V)_{\varpi^0}} = \max_{V \in E_m^*, \|V\|_{L^2(\varpi^0)} = 1} (\nabla V, \nabla V)_{\varpi^0}\ .$$

For each $V \in E_m^*$, with $\|V\|_{L^2(\varpi^0)} = 1$, we write $V = \sum_{i=1}^m \alpha_i^\varepsilon(\eta) U_i^\varepsilon(\cdot;\eta)$ for certain constants $\alpha_i^\varepsilon(\eta)$. On account of the abovementioned orthonormality, these constants satisfy

$$\|V\|^2_{L^2(\varpi^0)} = \sum_{i=1}^{m} (\alpha_i^\varepsilon(\eta))^2 = 1.$$

Similarly, because of the extension by zero, the orthonormality, and (7.7), for the gradients, we can write

$$\|\nabla V\|^2_{L^2(\varpi^0)} = \sum_{i=1}^{m} (\alpha_i^\varepsilon(\eta))^2 \|\nabla U_i^\varepsilon(\cdot;\eta)\|^2_{L^2(\varpi^0)} = \sum_{i=1}^{m} (\alpha_i^\varepsilon(\eta))^2 \Lambda_i^\varepsilon(\eta) \le \Lambda_m^\varepsilon(\eta),$$

which gives

$$\Lambda_m^*(\eta) \le \Lambda_m^\varepsilon(\eta), \quad \forall \eta \in [-\pi, \pi], \quad m \ge 1.$$

Therefore, the left-hand side of (7.23) holds for $K_m(\eta) = \Lambda_m^*(\eta)$ the eigenvalue of the mixed problem (7.24).

Finally, the precise constant $C_m$ on the right-hand sides of (7.23) has been obtained in [NaOrPe19a], related to the $m$-th eigenvalue of a Dirichlet problem in any fixed rectangle $(\alpha, \beta) \times (0, H)$, with $0 < \alpha < \beta < 1/2$. □

The first convergence result is given in Theorem 7.1 below. It shows the somehow expected convergence of the spectrum with conservation of the multiplicity in homogenization theory. Also, the convergence of the corresponding eigenfunctions is stated. The proof in [NaOrPe19a] has been performed adapting standard techniques in homogenization and spectral perturbation theory: see, for instance, Ch. 3 in [OlShYo92] for a general framework and [LoEtAl98] for its application to spectral problems in perforated domains with different boundary conditions.

**Theorem 7.1** *Let us consider the spectral problem* (7.3)–(7.6) *and the sequence of eigenvalues* (7.8). *Then, for any* $\eta \in [-\pi, \pi]$, *we have the convergence*

$$\Lambda_m^\varepsilon(\eta) \to \Lambda_m^0, \quad \text{as } \varepsilon \to 0, \tag{7.26}$$

*where* $\Lambda_m^0$ *are the set of eigenvalues in the sequence* (7.17) *of the Dirichlet problem* (7.16). *In addition, for each sequence, we can extract a subsequence, still denoted by* $\varepsilon$, *such that the extension by zero of the eigenfunctions* $\{U_m^\varepsilon(\cdot;\eta)\}_{m=1}^\infty$ *normalized in* $L^2(\varpi^\varepsilon)$, $\{\widehat{U}_m^\varepsilon(\cdot;\eta)\}_{m=1}^\infty$, *converges toward the eigenfunctions of* (7.10)–(7.13) *in* $L^2(\varpi^0)$, *which form an orthonormal basis of* $L^2(\varpi^0)$.

As a consequence of the asymptotic analysis in [NaOrPe19a], we state the following result:

**Theorem 7.2** *Let* $m \in \mathbb{N}$, *and let* $\Lambda_m^0$ *be an eigenvalue of the Dirichlet problem* (7.16) *in the sequence* (7.17). *There is at least one eigenvalue* $\Lambda_p^\varepsilon(\eta)$ *of problem*

(7.3)–(7.6), *with* $p = p(\varepsilon, \eta, m) \geq m$, *satisfying*

$$|\Lambda_p^\varepsilon(\eta) - \Lambda_m^0| \leq c_m \varepsilon, \quad \forall \varepsilon \leq \varepsilon_m, \ \eta \in [-\pi, \pi], \tag{7.27}$$

*where* $\varepsilon_m$ *and* $c_m$ *are certain positive constants that are independent of* $\eta$ *and* $\varepsilon$.

The proof of Theorem 7.2 can be found in [NaOrPe19a], based on a lemma on almost eigenvalues and eigenfunctions from the spectral perturbation theory, cf. [ViLu57]. It involves the construction of approximations to eigenpairs of the perturbation problem by means of asymptotic expansions from the solutions of the homogenized problem and a boundary layer problem in an unbounded perforated strip, namely, in the "unit periodicity cell" for the homogenization problem (7.3)–(7.6) (cf. also [NaOrPe19b]).

In the next section, we show that the index $p$ provided by Theorem 7.2 coincides with $m$, cf. Theorem 7.4. Although the bound (7.27) with $p = m$ has been used to detect spectral gaps in [NaOrPe19a], we think that the proof in Sect. 7.3 of this work may clarify that in [NaOrPe19a].

*Remark 7.1* It should be noted that bounds (7.23) can be improved as follows: for each fixed $m$, there are positive constants $\varepsilon_0 < 1$, $\theta < 1$, $k_m$, and $c_m$ independent of $\varepsilon$ and $\eta$, such that

$$\Lambda_m^0 - k_m \varepsilon^{2\theta} \leq \Lambda_m^\varepsilon(\eta) \leq \Lambda_m^0 + c_m \varepsilon \qquad \forall \eta \in [-\pi, \pi], \quad \varepsilon \leq \varepsilon_0. \tag{7.28}$$

The proof of (7.28) can be obtained using the reasoning of [GoEtAl21] (Sect. 7.3) with minor modifications. This implies using the max–min principle, Hardy inequality, the normalization procedure used to obtain the left-hand side inequality in (7.23) (applied to both finite-dimensional spaces of eigenfunctions of the perturbation and homogenized problem), weighted estimates in Sobolev spaces and some cut-off functions vanishing in $\varepsilon$-neighborhoods of the perforation string. This result allows a simplification of the proof of Theorem 7.3 related to the eigenvalues. However, the bounds (7.28) are associated with the homogenization of perforated domains along lines with Dirichlet boundary conditions in the perforations (see [GoEtAl21] and [GoEtAl22b] for other boundary conditions), and the suitable bounds cannot be obtained in many problems of perturbed waveguides, see [GoEtAl21], [GoEtAl22a], and [GoEtAl22b] to compare. In contrast, the technique developed in Theorem 7.3 can be applied to many problems even when the limit dispersion curves depend on $\eta$, cf. [GoEtAl22a].

Also, it should be emphasized that the result in Theorem 7.4 improves the bound (7.28) providing the precise value of $\theta = 1/2$. □

## 7.3 Convergence and Convergence Rates for Eigenvalues

The first approach to the asymptotics for eigenpairs of (7.3)–(7.6) is given by Theorem 7.1, when the parameter $\eta$ is fixed. Theorem 7.3 below also allows a certain perturbation of this parameter and therefore improves the result in Theorem 7.1.

**Theorem 7.3** *Let us consider the spectral problem* (7.3)–(7.6) *and the sequence of eigenvalues* (7.8). *Then, for each sequence* $\{(\varepsilon_r, \eta_r)\}_{r=1}^{\infty}$ *such that* $\varepsilon_r \to 0$ *and* $\eta_r \to \widehat{\eta} \in [-\pi, \pi]$, *as* $r \to \infty$, *we have the convergence*

$$\Lambda_m^{\varepsilon_r}(\eta_r) \to \Lambda_m^0, \qquad as\ r \to \infty, \tag{7.29}$$

*where* $\Lambda_m^0$ *are the set of eigenvalues of the Dirichlet problem* (7.16) *in the sequence* (7.17). *In addition, we can extract a subsequence, still denoted by* $\varepsilon_r$, *such that the extension by zero of the eigenfunctions* $\{U_m^{\varepsilon_r}(\cdot;\eta_r)\}_{m=1}^{\infty}$ *normalized in* $L^2(\varpi^{\varepsilon_r})$, $\{\widehat{U}_m^{\varepsilon_r}(\cdot;\eta_r)\}_{m=1}^{\infty}$, *converges toward the eigenfunctions of* (7.10)–(7.13) *in* $L^2(\varpi^0)$, *which form an orthonormal basis of* $L^2(\varpi^0)$.

***Proof*** Let us consider $\Lambda_m^{\varepsilon_r}(\eta_r)$ and $U_m^{\varepsilon_r}(\cdot;\eta_r) \in H^{1,\eta_r}_{per}(\varpi^{\varepsilon_r}; \Gamma^{\varepsilon_r})$ the eigenpair of (7.7); namely, for fixed $(\eta_r, \varepsilon_r)$ and $m = 1, 2, \cdots$, they satisfy

$$\left(\nabla U_m^{\varepsilon_r}(\cdot;\eta_r), \nabla V\right)_{\varpi^{\varepsilon_r}} = \Lambda_m^{\varepsilon_r}(\eta_r)\left(U_m^{\varepsilon_r}(\cdot;\eta_r), V\right)_{\varpi^{\varepsilon_r}}, \quad V \in H^{1,\eta}_{per}(\varpi^{\varepsilon_r}; \Gamma^{\varepsilon_r}). \tag{7.30}$$

Taking $V = U_m^{\varepsilon_r}(\cdot;\eta_r)$, (7.30) reads

$$\|\nabla U_m^{\varepsilon_r}(\cdot;\eta_r)\|^2_{L^2(\varpi^{\varepsilon_r})} = \Lambda_m^{\varepsilon_r}(\eta_r)\|U_m^{\varepsilon_r}(\cdot;\eta_r)\|^2_{L^2(\varpi^{\varepsilon_r})}.$$

Let us extend the eigenfunctions by zero inside the holes. Then, using (7.23), the normalization $\|U_m^{\varepsilon_r}(\cdot;\eta_r)\|_{L^2(\varpi^{\varepsilon_r})} = 1$, and the Poincaré inequality, for each $m$, we get a uniform bound for the eigenvalues and eigenfunctions in $H^1(\varpi^0)$. Indeed, the inequalities

$$\min_{\eta\in[-\pi,\pi]} K_m(\eta) \le \Lambda_m^{\varepsilon_r}(\eta_r) \le C_m \quad \text{and} \quad \|U_m^{\varepsilon_r}(\cdot;\eta_r)\|_{H^1(\varpi^{\varepsilon_r})} \le c_m \tag{7.31}$$

hold for constants $c_m$ and $C_m$ which do not depend on $\varepsilon_r$ and $\eta_r$.

Hence, for each fixed $m$, we can extract a subsequence of $\varepsilon_r$ and $\eta_r$, still denoted by $r$ such that

$$(\eta_r, \varepsilon_r) \to (\widehat{\eta}, 0), \ \text{ as } r \to \infty, \tag{7.32}$$

and

$$\Lambda_m^{\varepsilon_r}(\eta_r) \to \widehat{\Lambda}_m^0, \qquad \widehat{U}_m^{\varepsilon_r}(\cdot;\eta_r) \rightharpoonup \widehat{U}_m^0 \text{ in } H^1(\varpi^0) - weak, \ \text{ as } r \to \infty, \tag{7.33}$$

for a certain positive $\widehat{\Lambda}_m^0$ and a certain function $\widehat{U}_m^0 \in H^1(\varpi^0)$ which vanishes on the lower and upper bases of $\varpi^0$, namely on $\Gamma_{lu}$, cf. (7.4) and (7.25). Let us prove that $\widehat{U}_m^0$ also vanishes along the line $\{x_1 = 0\} \cap \varpi^0$.

Indeed, we use the Poincaré inequality on the domains $\varpi^0 \setminus \overline{\omega}$ and $\varpi^0$, cf. (7.1),

$$\|U\|_{L^2(\varpi^0\setminus\overline{\omega})} \leq C\|\nabla U\|_{L^2(\varpi^0\setminus\overline{\omega})} \quad \forall U \in H^1(\varpi^0 \setminus \overline{\omega};\ \Gamma_{lu} \cup \partial\omega),$$

and

$$\|U\|_{L^2(\varpi^0)} \leq C\|\nabla U\|_{L^2(\varpi^0)} \quad \forall U \in H^1(\varpi^0;\Gamma_{lu}).$$

We deduce

$$\varepsilon_r^{-1}\left\|U_m^{\varepsilon_r}(\cdot;\eta_r)\right\|^2_{L^2(\{|x_1|\leq\varepsilon_r/2\}\cap\varpi^0)} \leq C\varepsilon_r\left\|\nabla U_m^{\varepsilon_r}(\cdot;\eta_r)\right\|^2_{L^2(\{|x_1|\leq\varepsilon_r/2\}\cap\varpi^0)}, \tag{7.34}$$

where $C$ is a constant independent of $r$ and $m$. Now, taking limits in (7.34) as $r \to \infty$, or equivalently as $\varepsilon_r \to 0$, we get $\widehat{U}_m^0 = 0$ on $\{x_1 = 0\}\cap\varpi^0$ (cf., e.g., [MaKh06] and (7.31)) as it has been announced.

Therefore, the limit function in (7.33) satisfies $\widehat{U}_m^0 \in H^1(\varpi^0;\Gamma_{lu0})$, cf. (7.14). Let us prove that it also satisfies the quasi-periodicity conditions on the lateral sides of $\varpi^0$:

$$\widehat{U}_m^0(1/2,x_2) = e^{i\widehat{\eta}}\widehat{U}_m^0(-1/2,x_2) \text{ and } \frac{\partial\widehat{U}_m^0}{\partial x_1}(1/2,x_2) = e^{i\widehat{\eta}}\frac{\partial\widehat{U}_m^0}{\partial x_1}(-1/2,x_2). \tag{7.35}$$

To do this, notice that the change $V_m^{\varepsilon_r}(\cdot;\eta_r) = U_m^{\varepsilon_r}(\cdot;\eta_r)e^{-\mathrm{i}\eta_r x_1}$ converts the Laplacian into the differential operator

$$-\Big(\frac{\partial}{\partial x_1} + \mathrm{i}\eta_r\Big)\Big(\frac{\partial}{\partial x_1} + \mathrm{i}\eta_r\Big) - \frac{\partial^2}{\partial x_2^2},$$

and the $\eta_r$ quasi-periodicity condition for $U_m^{\varepsilon_r}(\cdot;\eta_r)$ becomes a periodicity condition for $V_m^{\varepsilon_r}(\cdot;\eta_r) \in H^1_{per}(\varpi^0;\Gamma_{lu})$. Consequently, since the convergence (7.33) holds, we also have a bound for $\widehat{V}_m^{\varepsilon_r} \in H^1_{per}(\varpi^0;\Gamma_{lu})$ which holds uniformly in $\eta_r$ and $\varepsilon_r$, and consequently a convergence of $\widehat{V}_m^{\varepsilon_r}(\cdot;\eta_r)$ ($V_m^{\varepsilon_r}(\cdot;\eta_r)$ extended by zero inside the holes) toward a function $\widehat{V}_m^0(\cdot;\eta_r) \in H^1_{per}(\varpi^0;\Gamma_{lu0})$ holds in the weak topology of $H^1(\varpi^0;\Gamma_{lu0})$. Then, we obtain $\widehat{V}_m^0 = \widehat{U}_m^0 e^{-\mathrm{i}\widehat{\eta}x_1}$, as a consequence of the convergence

$$\|\widehat{U}_m^{\varepsilon_r}(\cdot;\eta_r)e^{-\mathrm{i}\eta_r x_1} - \widehat{U}_m^0 e^{-\mathrm{i}\widehat{\eta}x_1}\|_{L^2(\varpi^0)} \to 0 \text{ as } r \to \infty.$$

To verify the last convergence, it suffices to consider

$$\|\widehat{U}_m^{\varepsilon_r}(\cdot;\eta_r)e^{-\mathrm{i}\eta_r x_1} - \widehat{U}_m^0 e^{-\mathrm{i}\widehat{\eta}x_1}\|_{L^2(\varpi^0)}$$
$$\leq \|\big(\widehat{U}_m^{\varepsilon_r}(\cdot;\eta_r) - \widehat{U}_m^0\big)e^{-\mathrm{i}\eta_r x_1}\|_{L^2(\varpi^0)} + \|\widehat{U}_m^0\big(e^{-\mathrm{i}\eta_r x_1} - e^{-\mathrm{i}\widehat{\eta}x_1}\big)\|_{L^2(\varpi^0)},$$

the convergence (7.33), the smoothness of the exponential function, and the convergence of $\eta_r$ toward $\widehat{\eta}$ as $r \to \infty$.

Thus, we have $\widehat{U}_m^0 = \widehat{V}_m^0 e^{\mathrm{i}\widehat{\eta}x_1}$, with $\widehat{V}_m^0 \in H^1_{per}(\varpi^0)$, and this already implies (7.35). Consequently, we have shown that $\widehat{U}_m^0 \in H^{1,\widehat{\eta}}_{per}(\varpi^0;\Gamma_{lu0})$ and depends on $\widehat{\eta}$. Also, the normalization of the eigenfunctions $\widehat{U}_m^{\varepsilon_r}(\cdot;\eta_r)$ in $L^2(\varpi^0)$ and the convergence (7.33) provides $\widehat{U}_m^0 \neq 0$.

In addition, by taking limits in the variational formulation (7.30) for the test functions $V \in \mathscr{C}_0^\infty((-1/2,0)\times(0,H))$ and for $V \in \mathscr{C}_0^\infty((0,1/2)\times(0,H))$, we obtain the partial differential equation

$$-\Delta\widehat{U}_m^0 = \widehat{\Lambda}_m^0\widehat{U}_m^0 \quad \text{for } x \in \widetilde{\varpi}^0. \tag{7.36}$$

All of this together allows us to identify $(\widehat{\Lambda}_m^0, \widehat{U}_m^0)$ with an eigenpair of the boundary value problem (7.10)–(7.13), cf. also (7.15).

Note that the extracted subsequence and limits, cf. (7.32) and (7.33), may depend on $m$. However, using a diagonalization argument, for each sequence of $r$, we can extract another subsequence of $r$, still denoted by $r$ but independent of $m$, such that (7.33) holds $\forall m \in \mathbb{N}$. Hence, by construction, we have obtained an increasing sequence of eigenvalues of (7.10)–(7.13)

$$0 < \widehat{\Lambda}_1^0 \leq \widehat{\Lambda}_2^0 \leq \cdots \leq \widehat{\Lambda}_m^0 \leq \cdots. \tag{7.37}$$

In what follows, we prove that the sequence $\{\widehat{\Lambda}_m^0\}_{m=1}^\infty$ converges toward infinity as $m \to \infty$, while the whole sequence coincides with that in (7.17).

Indeed, from the orthonormality of $U_m^{\varepsilon_r}(\cdot;\eta_r)$ in $L^2(\varpi^{\varepsilon_r})$, we get the orthonormality of $\widehat{U}_m^0 := \widehat{U}_m^0(\cdot;\widehat{\eta})$ in $L^2(\varpi^0)$ just writing

$$(\widehat{U}_m^{\varepsilon_r}(\cdot;\eta_r), \widehat{U}_p^{\varepsilon_r}(\cdot;\eta_r))_{\varpi^0} = \delta_{m,n}, \quad \forall m,n \in \mathbb{N},$$

and taking limits as $r \to \infty$. This confirms that the sequence (7.37) converges toward infinity as $m \to \infty$.

Let us prove that the sequence (7.37) coincides with that in (7.17). Since for each $(\varepsilon_r, \eta_r)$, we have a spectral problem with the corresponding spectrum (7.8) and the eigenfunctions forming an orthonormal basis of $L^2(\varpi^{\varepsilon_r})$, we can follow the idea of Section 3.1 in [OlShYo92] or Section III.9.1 in [At84] to show the convergence of the whole sequence of eigenvalues $\{\Lambda_m^{\varepsilon_r}(\eta_r)\}_{m=1}^\infty$ toward those of (7.10)–(7.13) with conservation of the multiplicity and that the set $\{\widehat{U}_m^0\}_{m=1}^\infty$ forms a basis of $L^2(\varpi^0)$. The fact that the eigenvalues $\widehat{\Lambda}_m^0$ do not depend on $\widehat{\eta}$ is due to

the identification performed by means of the change (7.19). However, since we are dealing with a double perturbation, the technique must be adapted and, for the sake of completeness, we provide here the whole proof.

We proceed by contradiction, assuming that there is some $\Lambda^*$ eigenvalue of (7.10)–(7.13) in the sequence (7.17) which is not in the sequence (7.37). Therefore, for some $m \in \mathbb{N}$,

$$\Lambda^* < \widehat{\Lambda}^0_{m+1}.$$

Let $U^*(\cdot;\widehat{\eta}) \in H^{1,\eta}_{per}(\varpi^0;\Gamma_{lu0})$ be a corresponding eigenfunction that is orthogonal to the constructed sequence of eigenfunctions $\{\widehat{U}^0_l(\cdot;\widehat{\eta})\}_{l=1}^{\infty}$. Then, we consider the function $U^{\varepsilon_r}_*(\cdot;\eta_r) \in H^{1,\eta_r}_{per}(\varpi^{\varepsilon_r};\Gamma^{\varepsilon_r})$, solution of the problem

$$\left(\nabla U^{\varepsilon_r}_*(\cdot;\eta_r), \nabla V\right)_{\varpi^{\varepsilon_r}} = \Lambda^* \left(U^*(\cdot;\widehat{\eta}), V\right)_{\varpi^{\varepsilon_r}} \qquad \forall V \in H^{1,\eta_r}_{per}(\varpi^{\varepsilon_r};\Gamma^{\varepsilon_r}).$$

Applying the Poincaré inequality, we obtain that the extension by zero of $U^{\varepsilon_r}_*(\cdot;\eta_r)$ inside the holes, $\{\widehat{U}^{\varepsilon_r}_*(\cdot;\eta_r)\}_r$, constitutes a sequence uniformly bounded in $H^1(\varpi^0)$. Therefore, up to a subsequence, still denoted by $r$,

$$\widehat{U}^{\varepsilon_r}_*(\cdot;\eta_r) \rightharpoonup U^*(\cdot;\widehat{\eta}) \text{ in } H^1(\varpi^0) - weak, \text{ as } r \to \infty. \tag{7.38}$$

Note that to show the convergence (7.38), we need to rewrite the argument above, cf. (7.30)–(7.36), with minor modifications.

From $U^{\varepsilon_r}_*(\cdot;\eta_r)$, we construct a new function $W^{\varepsilon_r}_*(\cdot;\eta_r)$ orthogonal to the set $\{U^{\varepsilon_r}_l(\cdot;\eta_r)\}_{l=1}^m$ in the space $L^2(\varpi^{\varepsilon_r})$ as follows:

$$W^{\varepsilon_r}_*(\cdot;\eta_r) = U^{\varepsilon_r}_*(\cdot;\eta_r) - \sum_{l=1}^{m} (U^{\varepsilon_r}_*(\cdot;\eta_r), U^{\varepsilon_r}_l(\cdot;\eta_r))_{\varpi^{\varepsilon_r}} U^{\varepsilon_r}_l(\cdot;\eta_r).$$

In addition, from the above convergence for eigenfunctions, (7.38), the orthogonality of the limit eigenfunctions in $L^2(\varpi^0)$, and the assumption performed on the orthogonality of $U^*(\cdot;\widehat{\eta})$ to the limit eigenfunctions, we can write

$$(U^{\varepsilon_r}_*(\cdot;\eta_r), U^{\varepsilon_r}_l(\cdot;\eta_r))_{\varpi^{\varepsilon_r}} \to 0, \quad \text{as } r \to \infty, \quad l = 1, 2, \cdots, m, \tag{7.39}$$

$$\widehat{W}^{\varepsilon_r}_*(\cdot;\eta_r) \rightharpoonup U^*(\cdot;\widehat{\eta}) \text{ in } H^1(\varpi^0) - weak, \text{ as } r \to \infty, \tag{7.40}$$

$\widehat{W}^{\varepsilon_r}_*(\cdot;\eta_r)$ being the extension by zero on the holes of $W^{\varepsilon_r}_*(\cdot;\eta_r)$, and

$$(\nabla W^{\varepsilon_r}_*(\cdot;\eta_r), \nabla W^{\varepsilon_r}_*(\cdot;\eta_r))_{\varpi^{\varepsilon_r}} \to \Lambda^*(U^*(\cdot;\widehat{\eta}), U^*(\cdot;\widehat{\eta}))_{\varpi^0}, \quad \text{as } r \to \infty. \tag{7.41}$$

Then, since for each $\varepsilon_r$, we have constructed a function $W_*^{\varepsilon_r}(\cdot;\eta_r) \in \{V \in H^{1,\eta_r}_{per}(\varpi^{\varepsilon_r};\Gamma^{\varepsilon_r});\ (V,(U_l^{\varepsilon_r}(\cdot;\eta_r))_{\varpi^{\varepsilon_r}} = 0,\ l = 1,2,\cdots,m\}$, we can apply the Rayleigh principle, see, for instance, Section I.7 in [SaSa89],

$$\Lambda^{\varepsilon_r}_{m+1}(\eta_r) = \inf_V \frac{(\nabla V,\nabla V)_{\varpi^{\varepsilon_r}}}{(V,V)_{\varpi^{\varepsilon_r}}},$$

where the infimum is computed over the elements of the space

$$\{V \in H^{1,\eta_r}_{per}(\varpi^{\varepsilon_r};\Gamma^{\varepsilon_r}) \,:\, (V,\ U_l^{\varepsilon_r}(\cdot;\eta_r))_{\varpi^{\varepsilon_r}} = 0,\ l = 1,2,\cdots,m\}.$$

Consequently,

$$\Lambda^{\varepsilon_r}_{m+1}(\eta_r) \le \frac{(\nabla W_*^{\varepsilon_r}(\cdot;\eta_r),\nabla W_*^{\varepsilon_r}(\cdot;\eta_r))_{\varpi^{\varepsilon_r}}}{(W_*^{\varepsilon_r}(\cdot;\eta_r),W_*^{\varepsilon_r}(\cdot;\eta_r))_{\varpi^{\varepsilon_r}}},$$

and taking limits as $r \to \infty$, from (7.33) and (7.39)–(7.41), we already obtain

$$\widehat{\Lambda}^0_{m+1} \le \Lambda^*,$$

which contradicts our assumption, and we have proved that all the eigenvalues of the homogenized problem in (7.17) are in the sequence $\{\widehat{\Lambda}^m_0\}_{m=1}^\infty$.

Also, this confirms the fact that the set of limiting eigenfunctions $\{\widehat{U}^0_m(\cdot;\widehat{\eta})\}_{m=1}^\infty$ in (7.33) forms an orthogonal basis in $L^2(\varpi^0)$ and the sets of limiting eigenvalues (7.37) and (7.17) coincide and are independent on the Floquet parameter. Therefore, the theorem is proved. □

**Theorem 7.4** *Let $m \in \mathbb{N}$, and let $\Lambda^0_m$ be an eigenvalue of the Dirichlet problem (7.16) in the sequence (7.17). There exist positive $\varepsilon_m$ and $c_m$ independent of $\eta$ and $\varepsilon$ such that, for any $\varepsilon \in (0,\varepsilon_m]$, the eigenvalue $\Lambda^\varepsilon_m(\eta)$ of problem (7.3)–(7.6) in the sequence (7.8) meets the estimate*

$$|\Lambda^\varepsilon_m(\eta) - \Lambda^0_m| \le c_m\varepsilon, \quad \forall \varepsilon \le \varepsilon_m,\ \eta \in [-\pi,\pi]. \tag{7.42}$$

***Proof*** Let us recall Theorem 7.2 that provides (7.27) for a certain $p(\varepsilon,\eta,m) \ge m$. Here, without any restriction, we can assume that $\Lambda^0_{m+1} > \Lambda^0_m$, otherwise $p(\varepsilon,\eta,m) \ge m+1$ also. Let us show that $p(\varepsilon,\eta,m) = m$, and consequently the result of the statement holds. We proceed by contradiction, denying (7.42).

This implies that there is $\eta^*$ such that the estimate (7.42) does not hold. That is, for this $\eta^*$, we can find an $\varepsilon_{\eta^*} \le \varepsilon_m$ for which $p(\varepsilon_{\eta^*},\eta^*,m) \ge m+1$ (and, obviously, strictly greater than $m+1$ depending on whether the multiplicity of $\Lambda^0_m$ be greater than 1). First of all, we observe that the numbers $\varepsilon_{\eta^*}$ that we can find must range in a finite set $\{\varepsilon_{\eta^*,1},\varepsilon_{\eta^*,2},\cdots\varepsilon_{\eta^*,k_{\eta^*}}\}$, because, otherwise, we can take a subsequence $\{\varepsilon_{\eta^*,l}\}_{l=1}^\infty$, $\varepsilon_{\eta^*,l} \to 0$ as $l \to \infty$, for which $p(\varepsilon_{\eta^*,l},\eta^*,m) \ge m+1$.

Then, from (7.27), we write

$$\Lambda_{m+1}^{\varepsilon_{\eta^*,l}}(\eta^*) \leq \Lambda_{p(\varepsilon_{\eta^*,l},\eta^*,m)}^{\varepsilon_{\eta^*,l}}(\eta^*) \leq \Lambda_m^0 + c_m \varepsilon_{\eta^*,l},$$

and taking limits, as $l \to \infty$, we get a contradiction, see the convergence (7.26): for fixed $\eta^*$, we have

$$\Lambda_{m+1}^0 \leq \Lambda_m^0. \tag{7.43}$$

Note that the limit is independent of $\eta$.

Consequently, for each $\eta^*$ such that (7.42) does not hold, we associate the finite set $\{\varepsilon_{\eta^*,l}\}_{l=1}^{k_{\eta^*}}$ for which $p(\varepsilon_{\eta^*,l}, \eta^*, m) \geq m+1$. In addition, we note that if there is only one $\eta^*$ for which (7.42) does not hold, taking

$$\varepsilon_m^* = \min(\varepsilon_m, \varepsilon_{\eta^*,1}, \varepsilon_{\eta^*,2}, \cdots \varepsilon_{\eta^*,k_{\eta^*}}),$$

the inequality (7.42) holds for $\varepsilon \leq \varepsilon_m^*$, and the same occurs if there is only a finite number of $\eta^*$ for which (7.42) does not hold.

Therefore, we deduce that there is at least one subsequence $\{\eta_r^*\}_{r=1}^{\infty}$ that converges toward some $\widehat{\eta} \in [-\pi, \pi]$ as $r \to \infty$ such that (7.42) is not satisfied for $\varepsilon_{\eta_r^*,1}, \varepsilon_{\eta_r^*,2}, \cdots \varepsilon_{\eta_r^*,k_{\eta_r^*}}$, $r = 1, 2, \cdots$, while (7.27) holds. Without any restriction, we can assume that there is also a subsequence of $\varepsilon_{\eta_r^*}$ converging toward zero as $r \to \infty$. Indeed, let us explain the last assertion in further detail. For the set $\mathscr{J} := \{\eta^* \in [-\pi, \pi] : \text{(7.42) is not satisfied}\} \subset [-\pi, \pi]$, we consider the associated set of parameters constructed above: $\mathscr{E} := \{\varepsilon_{\eta^*,1}, \varepsilon_{\eta^*,2}, \cdots \varepsilon_{\eta^*,k_{\eta^*}}\}_{\eta^* \in \mathscr{J}}$. Either $\mathscr{E}$ has a lower bound $\varepsilon_m^{**} > 0$ or we can extract a sequence $\{\varepsilon_{\eta_r^*}\}_{r=1}^{\infty}$ converging toward zero as $r \to \infty$, each one associated with a certain value $\eta_r^* \in \mathscr{J}$. In the first case, (7.42) holds for $\varepsilon \leq \varepsilon_m^* := \min(\varepsilon_m^{**}, \varepsilon_m)$ and the proof is ended. In the second case, since the sequence $\{\eta_r^*\}_{r=1}^{\infty}$ is bounded from above and from below, we can construct a subsequence, still denoted by $r$, such that

$$(\eta_r^*, \varepsilon_{\eta_r^*}) \to (\widehat{\eta}, 0) \text{ as } r \to \infty.$$

To show that this last assertion leads us to a contradiction, we note that from (7.27) we can write that the corresponding sequence of eigenvalues satisfies

$$\Lambda_{m+1}^{\varepsilon_{\eta_r^*}}(\eta_r^*) \leq \Lambda_{p(\varepsilon_{\eta_r^*},\eta_r^*,m)}^{\varepsilon_{\eta_r^*}}(\eta_r^*) \leq \Lambda_m^0 + c_m \varepsilon_{\eta_r^*}.$$

Taking limits as $r \to \infty$, from the convergence (7.29), we get again the contradiction (7.43). Therefore, the result of the theorem holds true. □

**Acknowledgments** The work has been partially supported by MICINN through PGC2018-098178-B-I00, PID2020-114703GB-I00 and Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000904-S).

## References


[At84] Attouch, H.: Variational Convergence for Functions and Operators. Applicable Mathematics Series. Pitman, London (1984)

[BaPe18] Bakharev, F.L., Pérez, E.: Spectral gaps for the Dirichlet-Laplacian in a 3-D waveguide periodically perturbed by a family of concentrated masses. Math. Nachr. **291**(4), 556–575 (2018)

[BiSo80] Birman, M.Sh., Solomjak, M.Z.: Spectral Theory of Selfadjoint Operators in Hilbert Spaces. [Translated from the 1980 Russian original by S. Khrushchëv and V. Peller], Mathematics and its Applications (Soviet Series). D. Reidel Publishing, Dordrecht (1987)

[CoPlVa94] Conca, C., Planchard, J., Vanninathan, M.: Fluids and Periodic Structures. RAM: Research in Applied Mathematics, vol. 38. Wiley, Chichester; Masson, Paris (1995)

[Ge50] Gelfand, I.M.: Expansion in characteristic functions of an equation with periodic coefficients (Russian). Doklady Akad. Nauk SSSR **73**, 1117–1120 (1950)

[GoEtAl21] Gómez, D., Nazarov, S.A., Orive-Illera, R., Pérez-Martínez, M.-E.: Remark on justification of asymptotics of spectra of cylindrical waveguides with periodic singular perturbations of boundary and coefficients. J. Math. Sci. (N.Y.) **257**(5), 597–623 (2021); in Russian, Problems Math. Analy. **111**, 43–66 (2021)

[GoEtAl22a] Gómez, D., Nazarov, S.A., Orive-Illera, R., Pérez-Martínez, M.-E.: Spectral gaps in a double-periodic perforated Neumann waveguide. To appear in Asymptot. Anal. (2022) 57 pp.

[GoEtAl22b] Gómez, D., Nazarov, S.A., Orive-Illera, R., Pérez-Martínez, M.-E.: Asymptotic stability of the spectrum of a parametric family of homogenization problems associated with a perforated waveguide. Submitted (2022)

[Ku93] Kuchment, P.: Floquet Theory for Partial Differential Equations. Birkhäuser Verlag, Basel (1993)

[LoEtAl98] Lobo, M., Oleinik, O.A., Pérez, M.E., Shaposhnikova, T.A.: On homogenization of solutions of boundary value problems in domains perforated along manifolds. Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **25**(3–4), 611–629 (1998)

[MaKh06] Marchenko, V.A., Khruslov, E.Y.: Homogenization of Partial Differential Equations. Progress in Mathematical Physics, vol. 46. Birkhäuser, Boston (2006)

[NaOrPe19a] Nazarov, S.A., Orive-Illera, R., Pérez-Martínez, M.-E.: Asymptotic structure of the spectrum in a Dirichlet-strip with double periodic perforations. Netw. Heterog. Media **14**(4), 733–757 (2019)

[NaOrPe19b] Nazarov, S.A., Orive-Illera, R., Pérez-Martínez, M.-E.: On the polarization matrix for a perforated strip. In: Integral Methods in Science and Engineering: Analytic Treatment and Numerical Approximations, pp. 267–281. Birkhäuser, New York (2019)

[OlShYo92] Oleinik, O.A., Shamaev, A.S., Yosifian, G.A.: Mathematical Problems in Elasticity and Homogenization. North-Holland, Amsterdam (1992)

[ReSi78] Reed, M., Simon, B.: Methods of Modern Mathematical Physics. IV. Analysis of Operators. Academic Press, New York-London (1978)

[SaSa89] Sanchez-Hubert, J., Sanchez-Palencia, E.: Vibration and Coupling of Continuous Systems. Asymptotic Methods. Springer, Berlin (1989)

[Sk85] Skriganov, M.M.: Geometric and arithmetic methods in the spectral theory of multidimensional periodic operators. A translation of Trudy Mat. Inst. Steklov **171**, 3–122 (1985); Proc. Steklov Inst. Math. **171**(2) (1987)

[ViLu57] Vishik, M.I., Lusternik, L.A.: Regular degeneration and boundary layer for linear differential equations with small parameter. Amer. Math. Soc. Transl. **20**(2), 239–364 (1962)

# Chapter 8
# The Wavelet-Based Integral Formula for the Solutions of the Wave Equation in an Inhomogeneous Medium: Convergence of Integrals


**E. A. Gorodnitskiy and M. V. Perel**

## 8.1 Introduction

We study the initial-boundary value problem for the wave equation in the half-plane $\mathbb{R}^2_+ = \{(x, z) : x \in \mathbb{R}, z \geq 0\}$

$$\partial_t^2 u - c^2(x, z)\Delta u = 0, \quad c(x, 0) = 1, \tag{8.1}$$

$$u(t, x, 0) = f(t, x), \tag{8.2}$$

where $\partial_t^2 u$ is the second derivative with respect to $t$. To obtain an unique solution we add the condition

$$\int\limits_{\mathbb{R}^2_+} dxdz|u|^2 \underset{t\to-\infty}{\to} 0 \tag{8.3}$$

and some restrictions on $f(t, x)$. We treat here an integral representation of the solution of (8.1)–(8.3) in terms of localized solutions, which was presented in [GoEtAl16]. We name the constituent localized solutions in the representation the elementary ones. Here we give some results for justification of this representation. The best-known integral representation of a solution in a homogeneous medium is given by the Fourier integral. If the medium is inhomogeneous, the numerical methods are used. In the case $f(t, x) = A(t, x)e^{i\frac{\Phi(t,x)}{\varepsilon}}$, $\varepsilon << 1$, the computational cost of numerical calculations is high. Asymptotic methods [BaBu09, MaFe81] are more appropriate for the problem and they may provide a qualitative description of

E. A. Gorodnitskiy · M. V. Perel (✉)
Saint Petersburg State University, Saint Petersburg, Russia
e-mail: eugy@yandex.ru; m.perel@spbu.ru



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_8

the field. If $f(t,x)=\sum_j A_j(t,x)e^{i\frac{\Phi_j(t,x)}{\varepsilon_j}}$, $\varepsilon_j << 1$, where $\varepsilon_j$ are of different order, or $f(t,x)$ is a multi-scaled function found experimentally, the processing of data should be done first and then each component should be treated by an appropriate asymptotic method. The integral formula, which is studied here, contains built-in data processing because it is based on the Poincaré affine wavelet analysis [AnEtAl06], which may be applied for image and signal processing. It yields the decomposition of the solution in exact localized solutions, which may have an asymptotic approximation.

Other wavelet-based integral formulas for solutions of the wave equation in a homogeneous medium were presented and studied in [PeSi03, PeSi06, PeSi07, PeSi09, PeEtAl10], where wavelets constructed with the similitude group were applied. In a homogeneous medium, the affine Poincaré wavelet analysis was earlier used in [Pe09, PeGo12, GoPe17]. The elementary solutions in homogeneous medium were the exact solutions named the Gaussian wave packets in [KiPe99, KiPe00]. Our idea was to decompose the boundary data in wavelets and each wavelet is a boundary datum for an elementary solution in the medium. In [GoEtAl12, GoEtAl16], (see also, [PeGo19]), we studied propagation in an inhomogeneous medium and used as elementary solutions asymptotic high-frequency ones called quasiphotons [BaUl81], [Ra82]. The quasiphotons were given by an explicit formula for high frequencies. They represent wave packets localized according to the Gaussian law near a point moving along a semiclassical trajectory. The simplest packets have a "footprint" on the boundary of the form

$$\psi(t,x)=e^{-i\varkappa t-\left(\frac{t^2+x^2}{2}\right)}. \tag{8.4}$$

We introduce an exact solution, an exact quasiphoton, which is a solution of (8.1)–(8.3) for $f(t,x)=\psi(t,x)$.

We aim to show that an integral representation from [GoEtAl12, GoEtAl16] gives an exact solution and it can be used not only in a high-frequency regime. We propose here to use exact quasiphotons as elementary solutions for decomposition of solutions of (8.1)–(8.3). We believe that hybrid methods can be based on this decomposition formula: in the high-frequency case, asymptotic formulas for quasiphotons as elementary solutions may be applied; for other parameters, the numerical methods for exact quasiphotons can be used. To develop a rigorous approach to an integral representation, we should formulate the well-posed initial-boundary value problem for the wave equation in a half-plane on a semi-infinite time interval. Then a problem we solve and a problem for elementary solutions will be well-posed. In particular, we should find a priori estimates for norms of solutions given in terms of norms of boundary data. By using these estimates, we should find the dependence of norms of exact quasiphotons on parameters. The integral representation is an integral in the space of parameters, on which the solutions depend. The convergence of the integral should be studied.

Below we give the outline of the paper. First, we give some facts about the Poincaré affine wavelet analysis and the integral decomposition of solutions in the homogeneous medium. Then we formulate results for the well-posed problem for the wave equation in a half-plane on a semi-infinite time interval, details will be in a separate paper. The estimates of norms of elementary solutions in terms of parameters are presented in [GoPe21]. From these estimates, it follows that to study the convergence of the integral formula, it is necessary to consider the fourfold integral over the parameters of the wavelet transform multiplied by the powers of the parameters. In the present paper, we prove the convergence of such an integral.

## 8.2 Preliminary Considerations by the Fourier Transform

Suppose $f(t,x) \in \mathbb{L}_2(\mathbb{R}^2)$. This function can be expanded into the Fourier integral

$$f(t,x) = \int\limits_{\mathbb{R}^2} e^{i(k_x x - \omega t)}\, \hat{f}(\omega, k_x)\, d\omega\, dk_x,$$

where $\hat{f}(\omega, k_x)$ is the Fourier transform of this function. Let us find a solution to the wave equation (8.1) in the homogeneous medium ($c = 1$) satisfying the boundary condition (8.2):

$$u(t,x,z) = \int\limits_{\mathbb{R}^2} e^{i(-\omega t + k_x x + k_z z)}\, \hat{f}(\omega, k_x)\, d\omega\, dk_x, \qquad k_z = \sqrt{\omega^2 - k_x^2}, \tag{8.5}$$

where the branch of the square root in the definition of $k_z$ is fixed by the condition $k_z > 0$, if $\omega^2 > k_x^2$. The negative $k_z$ corresponds to the second solution.

It is easy to see that the solutions given by the formula (8.5) are divided into two classes: if $\omega^2 > k_x^2$, $k_z$ is real, then the solutions propagate in the direction $z$, if $\omega^2 < k_x^2$, $k_z$ is imaginary, then the solutions vary exponentially. In this paper, we will discuss solutions, which propagate.

We assume that $f \in \mathbb{L}_2(\mathbb{R}^2)$ and that $\hat{f}(\omega, k_x) \neq 0$ only if $\omega > |k_x|$. We denote by $D_1$ the domain $\omega > |k_x|$ and by $\mathscr{D}_1$ the class of functions with the support of their Fourier transform lying in $D_1$. Moreover, we assume that $\hat{f}(\omega, k_x) = 0$ if

$$\omega^2 - k_x^2 \leq \varepsilon^2. \tag{8.6}$$

This condition means that the Fourier integral of the wavefield contains plane waves propagating under not very small angles to the boundary with $k_z \geq \varepsilon$.

If the boundary data are multi-scaled, then the representation in terms of plane waves is not effective.

## 8.3 Some Facts from the Poincaré Affine Wavelet Analysis

This article is devoted to the expansion of solutions of the wave equation based on the mathematical techniques of the Poincaré affine wavelet analysis [AnEtAl06]. Let us list some facts we use. Let be $\vec{\chi} = (t, x)^{\mathrm{T}}$, $\vec{\sigma} = (\omega, k_x)^{\mathrm{T}}$. Such two-dimensional vectors form the Minkowski space with the pseudo-Euclidean scalar product

$$(\vec{\chi}_1, \vec{\chi}_2)_m = t_1 t_2 - x_1 x_2.$$

The subscript $m$ in the notation of inner product $(\cdot, \cdot)_m$ is introduced to distinguish the pseudo-Euclidean inner product from the ordinary Euclidean one.

The Fourier transform $\hat{f}(\omega, k_x) \equiv \hat{f}(\vec{\sigma})$ of the function $f(t, x)$ is defined as follows:

$$\hat{f}(\vec{\sigma}) = \frac{1}{(2\pi)^2} \int\limits_{\mathbb{R}^2} d^2\vec{\chi}\, f(\vec{\chi}) e^{i(\vec{\chi},\vec{\sigma})_m} = \frac{1}{(2\pi)^2} \int\limits_{\mathbb{R}^2} d^2\vec{\chi}\, f(\vec{\chi}) e^{i\omega t - i k_x x}.$$

We give here formulas for the Poincaré decomposition of functions $f \in \mathscr{D}_1$. Let us choose two functions from this space $\zeta(\vec{\chi}), \psi(\vec{\chi}) \in \mathscr{D}_1$ and call these functions *mother wavelets*. Let us construct *a family of wavelets* $\zeta_{a,\phi,\vec{\chi}_s}(\vec{\chi})$, $\psi_{a,\phi,\vec{\chi}_s}(\vec{\chi})$, applying Lorentz transforms, shifts, and dilations to the mother wavelets:

$$\zeta_{a,\phi,\vec{\chi}_s}(\vec{\chi}) = \frac{1}{a}\zeta\left(\boldsymbol{\Lambda}_{-\phi}\frac{\vec{\chi} - \vec{\chi}_s}{a}\right), \qquad \psi_{a,\phi,\vec{\chi}_s}(\vec{\chi}) = \frac{1}{a}\psi\left(\boldsymbol{\Lambda}_{-\phi}\frac{\vec{\chi} - \vec{\chi}_s}{a}\right),$$

where the matrix of hyperbolic rotations is defined by the formula

$$\boldsymbol{\Lambda}_\phi = \begin{pmatrix} \cosh\phi & -\sinh\phi \\ -\sinh\phi & \cosh\phi \end{pmatrix}.$$

We define the *affine Poincaré wavelet transform*, which in what follows, will be called for brevity the wavelet transform, by the formula

$$\left(\mathscr{W}_\zeta f\right)(a, \phi, \vec{\chi}_s) = \int\limits_{\mathbb{R}} d^2\vec{\chi}\, f(\vec{\chi})\overline{\zeta_{a,\phi,\vec{\chi}_s}(\vec{\chi})},$$

where $\overline{\zeta}$ means the complex conjugation of $\zeta$. The wavelet transform is a convolution, so it can be found in the Fourier domain

$$\left(\mathscr{W}_\zeta f\right)(a, \phi, \vec{\chi}_s) = (2\pi)^2 a \int\limits_{D_1} d^2\vec{\sigma}\, \hat{f}(\vec{\sigma})\overline{\hat{\zeta}(a\Lambda_{-\phi}\vec{\sigma})} e^{i(\vec{\sigma},\vec{\chi}_s)_m}. \tag{8.7}$$

If the wavelet transform of a function is known, then the function itself can be recovered by the formula

$$f(\vec{\chi}) = \frac{1}{c_{\zeta\psi}} \int_0^\infty \frac{da}{a^3} \int_{\mathbb{R}^2} d^2\vec{\chi}_s \int_{\mathbb{R}} d\phi \left(\mathscr{W}_\zeta f\right)(a, \phi, \vec{\chi}_s)\psi_{a,\phi,\vec{\chi}_s}(\vec{\chi}),$$

where the constant $c_{\zeta\psi}$ is defined as follows:

$$c_{\zeta\psi} = (2\pi)^2 \int_{D_1} d^2\vec{\sigma} \frac{\overline{\hat{\zeta}(\vec{\sigma})}\hat{\psi}(\vec{\sigma})}{|\omega^2 - k_x^2|}.$$

The condition

$$0 < c_{\zeta\psi} < \infty$$

is called the admissibility condition, it imposes restrictions on the choice of a pair of wavelets $\zeta, \psi$.

## 8.4 Integral Representation of Solutions in the Homogeneous Medium

We proposed in [Pe09, PeGo12] to find solutions of (8.1) and (8.2) in a homogeneous medium as an integral superposition of solutions, which we name elementary ones. To construct a family of elementary solutions we should choose one solution, the so-called mother solution, which satisfies the same problem (8.1)–(8.3) but with $f(t, x) = \psi(t, x)$, $\psi(t, x) \in \mathscr{D}_1$. We denote this solution $\Psi(t, x, z) \equiv \Psi(\vec{\chi}, z)$. We determine a family of elementary solutions as

$$\Psi_{(a,\phi,\vec{\chi}_s)}(\vec{\chi}, z) = \frac{1}{a}\Psi(\boldsymbol{\Lambda}_{-\phi}\frac{\vec{\chi} - \vec{\chi}_s}{a}, \frac{z}{a}). \tag{8.8}$$

Integral representation of solutions of (8.1)–(8.3) reads

$$u(\vec{\chi}, z) = \frac{1}{c_{\zeta\psi}} \int_0^\infty \frac{da}{a^3} \int_{\mathbb{R}} d\phi \int_{\mathbb{R}^2} d^2\vec{\chi}_s \left(\mathscr{W}_\zeta f\right)(a, \phi, \vec{\chi}_s)\Psi_{(a,\phi,\vec{\chi}_s)}(\vec{\chi}, z)$$

if $f \in \mathscr{D}_1$. We use the notation $\Psi_\nu(\vec{\chi}, z) \equiv \Psi_{(a,\phi,\vec{\chi}_s)}(\vec{\chi}, z) \equiv \Psi_\nu(t, x, z)$ for elementary solutions.

## 8.5 Initial-Value Problem for the Wave Equation in a Half-Plane on a Semi-Infinite Time Interval

The family of solutions $\Psi_\nu(t,x,z)$ cannot be constructed by formulas (8.8) in the case of variable wave speed. We determined them from the initial-boundary value problem for the wave equation in a half-plane on a semi-infinite time interval. This problem arose for $u$ itself, for the mother solution $\Psi$ and the family of solutions $\Psi_\nu$. Here we formulate conditions under which such a problem is well-posed and give a priori estimates for norms of solutions.

We need some notation.

We use domains: $\Pi_{-\infty,T} = \{(t,x,z) : t \in (-\infty,T), (x,z) \in \mathbb{R}^2_+\}$, $\mathbb{R}^2_+ = \{(x,z) : x \in \mathbb{R}, z \geq 0\}$, $\pi_{-\infty,T} = \{(t,x) : t \in (-\infty,T), x \in \mathbb{R}\}$. We will have $\Omega = \Pi_{-\infty,T}$, $\Omega = \pi_{-\infty,T}$, or $\Omega = \mathbb{R}^2_+$. Let $u, v \in \mathbb{H}^l(\Omega)$, then

$$(u,v)^{(l)}_\Omega = \sum_{s=0}^{l} \int\limits_\Omega \sum_{\varkappa_0+\varkappa_1+\varkappa_2=s} \partial_t^{\varkappa_0}\partial_x^{\varkappa_1}\partial_z^{\varkappa_2} u \quad \partial_t^{\varkappa_0}\partial_x^{\varkappa_1}\partial_z^{\varkappa_2}\overline{v}\, dxdzdt,$$

where $\varkappa_j$, $j = 0,1,2$, may be from 0 to $s$. If $u \in \mathbb{H}^l(\Omega)$, then $\|u\|^{(l)}_\Omega = \sqrt{(u,u)^{(l)}_\Omega}$. If superscript is omitted or equal to zero, then $u \in \mathbb{L}_2(\Omega)$. For functions $u = u(t,x,z)$, we denote

$$\|u(t,\cdot,\cdot)\|^{(l)}_{\mathbb{R}^2_+} = \left( \sum_{s=0}^{l} \int\limits_{\mathbb{R}^2_+} \sum_{\varkappa_1+\varkappa_2=s} \left|\partial_x^{\varkappa_1}\partial_z^{\varkappa_2} u(t,x,z)\right|^2 dxdz \right)^{1/2},$$

and

$$\|u\|_{\Pi_{-\infty,T}} = \left( \int\limits_{-\infty}^{T} dt \|u(t,.,.)\|^2_{\mathbb{R}^2_+} \right)^{1/2}.$$

We say that $u \in \mathscr{L}_{1,2}(\Pi_{-\infty,T})$ if

$$\|u\|_{\mathscr{L}_{1,2}(\Pi_{-\infty,T})} = \int\limits_{-\infty}^{T} dt \|u(t,.,.)\|_{\mathbb{R}^2_+} < \infty.$$

We will use also spaces with the norm

$$\|u\|^{(k)}_{\mathscr{L}_{1,2}(\Pi_{-\infty,T})} = \int\limits_{-\infty}^{T} dt\,(T-t)^k \|u(t,.,.)\|_{\mathbb{R}^2_+}.$$

We introduce the energy functional class $H^1_{en}(\Pi_{-\infty,T})$ as follows: the function $\Psi \in H^1_{en}(\Pi_{-\infty,T})$ if $\Psi \in H^1(\Pi_{-\infty,T})$ and it satisfies the equation in the sense of integral identity, see [La13]; the functions $\Psi(t,.,.)$ and $\Psi_t(t,.,.)$ are defined for every $t$ and are continuous as functions of $t$ in the classes $H^1(\mathbb{R}^2_+)$, $\mathscr{L}_2(\mathbb{R}^2_+)$, respectively.

Now we describe spaces for a boundary function $\psi = \psi(t,x)$. Let $\psi(t,x) \in H^2(\pi_{-\infty,T})$. Then the norm $\|\psi(t,.)\|$ in $H^2(\mathbb{R})$ exists for almost every $t$. We define the norm comprising the second derivative with respect to $t$:

$$\left(\|\psi(t,.)\|^{(2)}_{\mathbb{R}}\right)^2 = \int\limits_{\mathbb{R}} dx\,\left(|\psi|^2 + |\partial_t^2\psi|^2 + |\partial_x^2\psi|^2\right).$$

We denote

$$\|\!|\psi|\!\|^{(k,2)}_{\mathscr{L}_{1,2}(\pi_{-\infty,T})} = \int\limits_{-\infty}^{T} d\tau\,(t-\tau)^k \|\!|\psi(t,.)|\!\|^{(2)}_{\mathbb{R}} < \infty,$$

$k = 0, 1, 2.$

**Theorem 8.1**

- *Let the speed $c$ satisfy the conditions:*

$$0 < c_{min} \le c(x,z) \le c_{max} < \infty,$$
$$|\vec{\nabla} c(x,z)| \le c_1 < \infty.$$

- *Let $\psi \in H^2(\mathbb{R}^2)$ and*

$$\|\!|\psi|\!\|^{(k,2)}_{\mathscr{L}_{1,2}(\pi_{-\infty,T})} < \infty, \quad k = 0, 1, 2,$$

-

$$\int\limits_{\mathbb{R}} dx |\psi|^2 \underset{t\to-\infty}{\to} 0\,.$$

*If these conditions are satisfied, the solution of the problem (8.1)–(8.3) with $f = \psi(t,x)$ denoted $\Psi(t,x,z)$ such that $\Psi \in H^1_{en}(\Pi_{-\infty,T}) \cap \mathscr{L}_{1,2}(\Pi_{-\infty,T})$ and $\Psi_t, \Psi_x, \Psi_z \in \mathscr{L}_{1,2}(\Pi_{-\infty,T})$ exists, is unique and stable with respect to small variations of the boundary data, the following estimates are valid:*

$$\|\Psi(t,.,.)\|^{(1)}_{\mathbb{R}^2_+} \le C_1 \left( \|\!|\psi|\!\|^{(1,2)}_{\mathscr{L}_{1,2}(\pi_{-\infty,t})} + \|\!|\psi|\!\|^{(0,2)}_{\mathscr{L}_{1,2}(\pi_{-\infty,t})} + \|\psi(t,.)\|^{(1)}_{\mathbb{R}} \right),$$

$$\|\Psi_t(t,.,.)\|_{\mathbb{R}^2_+} \le C_2 \left( \|\!|\psi|\!\|^{(0,2)}_{\mathscr{L}_{1,2}(\pi_{-\infty,t})} + \|\partial_t \psi(t,.)\|_{\mathbb{R}} \right),$$

$$(\|\Psi\|^{(1)}_{\Pi_{-\infty,t}})^2 \le C_3 \left( \|\!|\psi|\!\|^{(1,2)}_{\mathscr{L}_{1,2}(\pi_{-\infty,t})} \right.$$
$$\left. \left( \|\!|\psi|\!\|^{(2,2)}_{\mathscr{L}_{1,2}(\pi_{-\infty,t})} + \|\!|\psi|\!\|^{(0,2)}_{\mathscr{L}_{1,2}(\pi_{-\infty,t})} \right) + (\|\psi\|^{(1)}_{\pi_{-\infty,t}})^2 \right),$$

*where $C_j$, $j = 1,2,3$ are some positive constants.*

## 8.6 Integral Representation for Solutions in the Inhomogeneous Medium

We study propagating solutions of the problem (8.1)–(8.3). Let the speed $c(x,z)$ and the boundary function $f(t,x)$ satisfy the conditions listed in the theorem 8.1. We assume additionally that $f, \psi \in \mathscr{D}_1$.

Let we know the solution $\Psi(t,x,z)$ of the problem (8.1)–(8.3), which exists according to the theorem 8.1.

Determine a family of solutions

$$\frac{1}{c^2(x,z)} \partial_t^2 \Psi_\nu - \partial_x^2 \Psi_\nu - \partial_z^2 \Psi_\nu = 0, \quad \nu = (a, \phi, \vec{\chi}_s),$$
$$\Psi_\nu(t,x,z)|_{z=0} = \psi_\nu(t,x),$$
$$\int\limits_{\mathbb{R}^2_+} dx dz |\Psi_\nu|^2 \underset{t \to -\infty}{\to} 0 .$$

It is easy to check that $\psi_\nu(t,x) \in \mathscr{D}_1$. We assume that $\psi_\nu(t,x)$ satisfies conditions of the theorem 8.1. This is true for a particular case of $\psi$ of the form (8.4).

If the family of solutions is constructed, the integral representation of solutions of (8.1)–(8.3) can be found

$$u(\vec{\chi}, z) = \frac{1}{c_{\zeta\psi}} \int\limits_0^\infty \frac{da}{a^3} \int\limits_{\mathbb{R}} d\phi \int\limits_{\mathbb{R}^2} d^2 \vec{\chi}_s \left( \mathscr{W}_\zeta f \right)(a, \phi, \vec{\chi}_s) \Psi_\nu(\vec{\chi}, z), \tag{8.9}$$

where $\Psi_\nu(\vec{\chi}, z) \equiv \Psi_\nu(t, x, z)$, $\nu = (a, \phi, \vec{\chi}_s)$. This representation was introduced formally in [GoEtAl16]. In [GoPe21], we estimate norms of exact quasiphotons from the family $\Psi_{(a,\phi,\vec{\chi}_s)}(\vec{\chi}, z)$ as functions of $a$, $\cosh\phi$, $\vec{\chi}_s$ and find a power law in these parameters. The convergence of the integral on the right-hand side of (8.9) is proved below in Theorem 8.2.

## 8.7 On Convergence of Integrals

We say that the function $\zeta(t, x)$ belongs to the Schwartz class, $\zeta(t, x) \in \mathscr{S}$, if $\zeta \in C^\infty$, and the function $\zeta$ and its derivatives decrease faster than any power of $t$ and $x$ for $t^2 + x^2 \to \infty$, that is, $(t^2 + x^2)^{m/2} \partial_t^k \partial_x^j \zeta(t, x) \underset{t^2+x^2\to\infty}{\to} 0$, for all non-negative $m, k, j$. It is known that the Fourier transform maps the Schwarz class into the Schwarz class.

**Lemma 8.1** *Let $f(t, x), \zeta(t, x) \in \mathscr{S}$, $f, \zeta \in \mathscr{D}_1$. Both functions $\hat{f}(\vec{\sigma})$ and $\hat{\zeta}(\vec{\sigma})$ vanish for $0 \le \omega^2 - k_x^2 \le \varepsilon^2$.*

*Then the following integral converges:*

$$J = \int_0^\infty \frac{da}{a^3} \int_{\mathbb{R}} d\phi U[f, \zeta](a, \phi)(\cosh\phi)^q a^l < \infty, \tag{8.10}$$

*for any $l$ and $q$, where*

$$U[f, \zeta](a, \phi) = a \int_{D_1} d^2\vec{\sigma} \left| \hat{f}(\vec{\sigma}) \hat{\zeta}\left(a \boldsymbol{\Lambda}_{-\phi} \vec{\sigma}\right) \right|.$$

***Proof*** The integral (8.10) reads

$$J = \int_0^\infty \frac{da}{a^2} \int_{\mathbb{R}} d\phi \int_{D_1} d^2\vec{\sigma} \left| \hat{f}(\vec{\sigma}) \hat{\zeta}\left(a \boldsymbol{\Lambda}_{-\phi} \vec{\sigma}\right) \right| (\cosh\phi)^q a^l$$

$$= \int_{D_1} d^2\vec{\sigma} |\hat{f}(\vec{\sigma})| \int_0^\infty \frac{da}{a^{2-l}} \int_{\mathbb{R}} d\phi (\cosh\phi)^q \left| \hat{\zeta}\left(a \boldsymbol{\Lambda}_{-\phi} \vec{\sigma}\right) \right|.$$

To calculate the inner integral we introduce new variable of integration $\vec{\sigma}'$ instead of $a$ and $\phi$:

$$\vec{\sigma}' = a \Lambda_{-\phi} \vec{\sigma}.$$

The Jacobian determinant is determined by the formula:

$$\left|\frac{D(a,\phi)}{D(\omega',k_x')}\right| = \frac{1}{a(\omega^2-k_x^2)}.$$

We express $a$ and $\phi$ through the variables $\vec{\sigma}$, $\vec{\sigma}'$. We take into account the fact that

$$a^2(\omega^2-k_x^2) = (\omega')^2-(k_x')^2,$$

denote $(\omega')^2-(k_x')^2 \equiv (\rho')^2$, $\omega^2-k_x^2 \equiv \rho^2$, and therefore

$$a = \rho'/\rho.$$

We get also that

$$\cosh\phi = \frac{\omega'\omega - k_x'k_x}{\rho\rho'} \le 2\frac{\omega'\omega}{\rho\rho'}.$$

If $q > 0$, we obtain that (8.10) is reduced to the form:

$$J \le 2^q \int\limits_{D_1} d^2\vec{\sigma}\,\frac{|\hat{f}(\vec{\sigma})|\omega^q}{\rho^{q+l-1}} \int\limits_{D_1} d^2\vec{\sigma}'\,\frac{\overline{|\hat{\zeta}(\vec{\sigma}')|}(\omega')^q}{(\rho')^{q-l+3}}.$$

Both integrals of the product converge for any $q$ and $l$, as it follows from the fact that

$$\varepsilon < \rho < \omega, \qquad \varepsilon < \rho' < \omega',$$

see (8.6). Indeed, if $q+l-1 > 0$, then $1/\rho^{q+l-1} < 1/\varepsilon^{q+l-1}$. The first integral is reduced to the integral of $|\hat{f}(\vec{\sigma})|\omega^q$, which converges if $f \in \mathscr{S}$. For $q+l-1 < 0$, the first integral can be estimated as follows

$$\int\limits_{D_1} d^2\vec{\sigma}\,|\hat{f}(\vec{\sigma})|\omega^q(\omega^2-k_x^2)^{|q+l-1|/2} \le \int\limits_{D_1} d^2\vec{\sigma}\,|\hat{f}(\vec{\sigma})|\omega^q\omega^{1-l-q}.$$

It converges because $\hat{f}(\vec{\sigma})$ decays faster than any power of $\omega$ as a function from the class $\mathscr{S}$. The second integral can be analyzed in a similar way. If $q < 0$, $\cosh^q\phi < 1$ and $J$ is estimated analogously.

**Theorem 8.2** *Let $f(t,x)$, $\zeta(t,x) \in \mathscr{S}$, $f,\zeta \in \mathscr{D}_1$. Both functions, $\hat{f}(\vec{\sigma})$ and $\hat{\zeta}(\vec{\sigma})$, vanish for $0 \le \omega^2-k_x^2 \le \varepsilon^2$.*

*Then the following integral converges:*

$$J = \int\limits_{\mathbb{R}^3} dx_s dt_s d\phi \int\limits_0^\infty \frac{da}{a^3} \left| \left(\mathscr{W}_f \zeta\right)(a, \phi, \vec{\chi}_s) \right| (\cosh\phi)^q \, a^l |t_s|^m |x_s|^n < \infty, \tag{8.11}$$

*for any* $l$, $q$, $m \geq 0$, *and* $n \geq 0$.

***Proof*** If $x_s$ and $t_s$ are bounded, the convergence of $J$ follows from lemma 8.1. In fact, after rewriting the wavelet transform in the form of (8.7), and estimating its modulus we obtain an integral, which does not depend on $t_s$ and $x_s$. Inserting it into (8.11) and integrating the result in bounded limits over $t_s$ and $x_s$, we obtain the integral (8.10), which converges.

Let now $x_s$ and $t_s$ belong to infinite intervals. The properties of the Fourier transform yield

$$(it_s)^m (-ix_s)^n \left(\mathscr{W}_\zeta f\right)(a, \phi, \vec{\chi}_s) = a \int\limits_{D_1} d^2\vec{\sigma} \; V_{m,n}(a, \phi, \vec{\sigma}) e^{-i(\vec{\sigma}, \vec{\chi}_s)_m}, \tag{8.12}$$

$$V_{m,n}(a, \phi, \vec{\sigma}) = (2\pi)^2 \frac{\partial^{m+n} (\hat{f}(\vec{\sigma}) \overline{\hat{\zeta}(a\boldsymbol{\Lambda}_{-\phi}\vec{\sigma})})}{\partial\omega^m \partial k_x^n}.$$

The function $V_{m,n}(a, \phi, \vec{\sigma})$ is the sum of the terms obtained by differentiating the product. To estimate the integral (8.11), it suffices to estimate integrals of the form

$$I = a \int\limits_{D_1} d^2\vec{\sigma} \; |\partial_\omega^{m_1} \partial_{k_x}^{n_1} \hat{f}(\vec{\sigma})| \; |\partial_\omega^{m_2} \partial_{k_x}^{n_2} \overline{\hat{\zeta}(\vec{\sigma}')}| \Big|_{\vec{\sigma}' = a\boldsymbol{\Lambda}_{-\phi}\vec{\sigma}}. \tag{8.13}$$

At each differentiation of $\zeta$, the multiplier $a\cosh\phi$ or $a\sinh\phi$ will arise, and the derivative of $\zeta$ by its argument will also appear. This derivative is a function of $a\boldsymbol{\Lambda}_{-\phi}\vec{\sigma}$, belonging to the Schwartz class. Finally, the integral (8.13) is reduced to the integral

$$I \leq a^{j+1} \cosh^j\phi \int\limits_{D_1} d^2\vec{\sigma} \left| g(\vec{\sigma}) \overline{\eta(a\boldsymbol{\Lambda}_{-\phi}\vec{\sigma})} \right| = a^{j+1} \cosh^j\phi \, U[g, \eta](a, \phi),$$

$$g = \partial_\omega^{m_1} \partial_{k_x}^{n_1} \hat{f}(\vec{\sigma}), \eta = \partial_{\omega'}^{m_2} \partial_{k_x'}^{n_2} \overline{\hat{\zeta}(\vec{\sigma}')} \Big|_{\vec{\sigma}' = a\boldsymbol{\Lambda}_{-\phi}\vec{\sigma}}, \tag{8.14}$$

where $j = m_2 + n_2$. Thus, $|g(t, x)| = |t^{m_1} x^{n_1} f(t, x)|$, and $|\eta(t, x)| = |t^{m_2} x^{n_2}|$ $|\zeta(t, x)|$. The differentiation does not change the support of the function and multiplication to powers of arguments does not take the function out of the Schwartz class. Therefore, lemma 8.1 can be applied to the integral obtained by substitution of (8.14) into (8.11) and the obtained integral converges. By using (8.12) and (8.14),

we find that

$$|\left(\mathscr{W}_\zeta f\right)(a,\phi,\vec{\chi}_s)| \leq \frac{\text{Const}}{(|x_s|^n+1)(|t_s|^m+1)} \sum_{\substack{m_1+m_2=m,\\ n_1+n_2=n}} U[t^{m_1}x^{n_1}f, t^{m_2}x^{n_2}\zeta](a,\phi)a^{j+1}\cosh^j\phi.$$

Substituting this expression into the integral (8.11) and applying the lemma 8.1, we get a converging integral.

## References


[AnEtAl06] Antoine, J.-P., Murenzi, R., Vandergheynst, P., Ali, S.T.: Two-Dimensional Wavelets and their Relatives. Cambridge University Press, Cambridge (2006)

[BaBu09] Babich, V.M., Buldyrev, V.S.: Asymptotic Methods in Short-Wavelength Diffraction Theory. Alpha Science International, Oxford (2009)

[BaUl81] Babich, V.M., Ulin, V.V.: The complex space-time ray method and "quasi-photons" (Russian). In: Mathematical questions in the theory of wave propagation. 12 Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **117**, 5–12 (1981). J. Soviet Math. **24**, 269–274 (1984)

[GoEtAl12] Gorodnitskiy, E.A., Perel, M.V., Geng, Y., Wu, R.S.: Poincaré wavelet techniques in depth migration. In: 2012 Proceedings of the International Conference Days on Diffraction, pp. 104–110. IEEE, Saint Petersburg (2012)

[GoEtAl16] Gorodnitskiy, E., Perel, M.V., Geng, Y., Wu, R.-S.: Depth migration with Gaussian wave packets based on Poincaré wavelets. Geophys. J. Int. **205**, 314–331 (2016)

[GoPe17] Gorodnitskiy, E.A., Perel, M.V.: Decompositions in Gaussian beams by wavelet methods. In: 2017 Progress in Electromagnetics Research Symposium-Spring (PIERS), pp. 1482–1487. IEEE, Piscataway (2017)

[GoPe21] Gorodnitskiy, E.A., Perel, M.V.: Rigorous mathematical formulation for quasiphotons. A priori estimates. In: 2021 Days on Diffraction (DD): Proceedings of the International Conference. IEEE, Saint-Petersburg (2021)

[KiPe99] Kiselev, A.P., Perel, M.V.: Gaussian wave packets. Opt. Spectrosc. **86**, 307–309 (1999)

[KiPe00] Kiselev, A.P., Perel, M.V.: Highly localized solutions of the wave equation. J. Math. Phys. **41**, 1034–1955 (2000)

[La13] Ladyzhenskaya, O.A.: The Boundary Value Problems of Mathematical Physics, vol. 49. Springer, New York (2013)

[MaFe81] Maslov, V.M., Fedoryuk, M.V.: Semiclassical Approximation in Quantum Mechanics, D. Reidel, Dordrecht (1981)

[Pe09] Perel, M.V.: Integral representation of solutions of the wave equation based on Poincaré wavelets. In: Proceedings of the International Conference Days on Diffraction, Saint-Petersburg, pp. 159–161 (2009)

[PeGo12] Perel, M., Gorodnitskiy, E.: Integral representations of solutions of the wave equation based on relativistic wavelets. J. Phys. A Math. Theor. **45**, 385203 (2012)

[PeGo19] Perel, M.V., Gorodnitskiy, E.A.: Decomposition of solutions of the wave equation into Poincaré wavelets. In: Integral Methods in Science and Engineering, pp. 343–352. Birkhäuser, Cham (2019)

[PeSi03] Perel, M.V., Sidorenko, M.S.: Wavelet analysis in solving the Cauchy problem for the wave equation in three-dimensional space. In: Cohen, G.C. (ed.) Mathematical and

Numerical Aspects of Wave Propagation: Waves 2003, pp.794–798. Springer, Berlin (2003)

[PeSi06] Perel, M.V., Sidorenko, M.S.: Wavelet analysis for the solutions of the wave equation. In: Proceedings of the International Conference Days on Diffraction, Saint-Petersburg, pp. 208–217 (2006)

[PeSi07] Perel, M.V., Sidorenko, M.S.: New physical wavelet 'Gaussian wave packet'. J. Phys. A Math. Theor. **40**, 3441 (2007)

[PeSi09] Perel, M.V., Sidorenko, M.S.: Wavelet-based integral representation for solutions of the wave equation. J. Phys. A Math. Theor. **42**, 375211 (2009)

[PeEtAl10] Perel, M., Sidorenko, M., Gorodnitskiy, E.: Multiscale investigation of solutions of the wave equation. In: Integral Methods in Science and Engineering, vol. 2, pp. 291–300. Birkhäuser, Boston (2010)

[Ra82] Ralston, J.: Gaussian beams and the propagation of singularities. Stud. Partial Differ. Equs. **23**, 206–248 (1982)

# Chapter 9
# Modelling the Spread of a Disease in an Epidemic Through a Country Divided into Geographical Regions


**P. J. Harris and B. E. J. Bodmann**

## 9.1 Introduction

The global spread of the SAR-CoV-2 virus (which is more commonly referred to as the Covid-19 virus) [RiAl20, SaEtAl20] has led to a renewed interest in the mathematical modelling of the spread of diseases [WuEtAl20]. Experimental evidence of how a disease spreads across a country or state shows that the spread will depend on the distribution of the people within the country being considered (see [AlEtAl, Ne15, ReEtAl13] for example). Generally, a disease will spread quickly through a densely populated city and slowly through a sparsely populated rural area. On other words, the spread of the disease is dependent on the population density of a region rather than just the population.

This chapter will consider the regional model for the spread of an infectious disease across a country or region developed in Harris and Bodmann [HaBo21] which uses the population density in each region rather than just the population. This model contains a number of new parameters, most notably the proportion of infected people who are not diagnosed with the disease and a parameter which controls how many people travel between the different geographical regions. In this chapter, we will investigate the effect that changing these parameters has on the predicted spread of a disease across the main part of the United Kingdom. We will then apply the model to simulate the spread of COVID-19 in the United Kingdom where we will simulate the effects of events like lockdowns by making appropriate changes to

P. J. Harris (✉)
The University of Brighton, Brighton, UK
e-mail: p.j.harris@brighton.ac.uk

B. E. J. Bodmann
Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil
e-mail: bardo.bodmann@ufrgs.br



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_9

some of the parameters in the model to simulate the reduction in transmission and travel that result from a lockdown.

## 9.2 Mathematical Model

In developing our model of the spread of a disease through a country divided into a number of regions, we made the following assumptions:

- We assume that the population is uniformly distributed within each region considered.
- We assume that births and deaths from causes other than the disease can be neglected. This means the total number of individuals in the system remains constant.
- Whilst people can move between regions within the country, we assume that no one enters or leaves the country.
- Anyone who recovers from the disease is then immune and cannot become reinfected.

The starting point for our analysis is the susceptible–infected–recovered (SIR) model [He00]

$$\begin{aligned} \frac{dS}{dt} &= -\lambda SI \\ \frac{dI}{dt} &= \lambda SI - \mu_{IR} I \\ \frac{dR}{dt} &= \mu_{IR} I \end{aligned} \tag{9.1}$$

where $S$, $I$ and $R$ denote the number of individuals who are susceptible to the disease, who are infected by the disease and who have recovered from the disease, respectively. Furthermore, $\lambda$ is the infection rate and $\mu_{IR}$ is the recovery rate.

The basic SIR model (9.1) can be extended to include categories such as carriers (denoted $C$) and those that have died from the disease (denoted $D$). Here, we define carriers as individuals who are infected with the disease but are not diagnosed because they are asymptomatic or because they mistake the symptoms for a different disease. In addition, we will use the term infected to denote those individuals who have been diagnosed and know that they are infected. The modified system of differential equations is [HaBo21]

$$\begin{aligned}
\frac{dS}{dt} &= -\lambda S(I+C) \\
\frac{dI}{dt} &= \lambda\beta S(I+C) - \mu_{IR}I - \mu_{ID}I \\
\frac{dC}{dt} &= \lambda(1-\beta)S(I+C) - \mu_{CR}C - \mu_{CD}C \\
\frac{dR}{dt} &= \mu_{IR}I + \mu_{CR}C \\
\frac{dD}{dt} &= \mu_{ID}I + \mu_{CD}C
\end{aligned} \tag{9.2}$$

where $\beta$ is the proportion of infected individuals who are diagnosed and not just carriers and $\mu_{ID}$ is the proportion of infected individuals who die from the disease. Similarly, $\mu_{CR}$ and $\mu_{CD}$ are the proportions of carriers who recover or die.

There are some drawbacks to using (9.2) to simulate the spread of a disease through a country. Firstly, it assumes that the population is uniformly distributed over the whole country when in reality a large proportion of the population is usually concentrated in cities with relatively few people living in rural areas. Secondly, in an outbreak, a disease is usually confined to a small area of the country and spreads out from that area. To illustrate the first drawback, consider the four countries described in Table 9.1. If we apply the SIR model (9.1) with the same parameters to each of the regions in Table 9.1, then the results for Regions A and D will be the same, as will the results for Regions B and C as they have the same populations. However, if apply the SIR model (9.1) to the population densities rather than the populations, then the proportions of the population who become infected are the same for regions A and C as they have the same population densities, whilst the proportion of the population who become infected is higher for Region B as it has a higher population density. The proportion of the population who become infected in Region D is smaller as Region D has the smallest population density. These results are illustrated in Fig. 9.1 where the curves for Regions A and C are superimposed.

Consider a country that can be divided into a number of geographical regions. For the UK, this can be the counties or administrative areas. We assume that the population is uniformly distributed within each region. Let the $i$th element of the vectors **S**, **I**, **C**, **R** and **D** be the number of people in each category in the $i$th region. Then, the differential equations for a single region (9.2) can be extended

**Table 9.1** The area and populations of four example isolated countries

| Region | Population | Initial infected | Area (km$^2$) | Population density (km$^{-2}$) |
|---|---|---|---|---|
| A | 1,000,000 | 10 | 1000 | 1000 |
| B | 2,000,000 | 20 | 1000 | 2000 |
| C | 2,000,000 | 20 | 2000 | 1000 |
| D | 1,000,000 | 10 | 2000 | 500 |

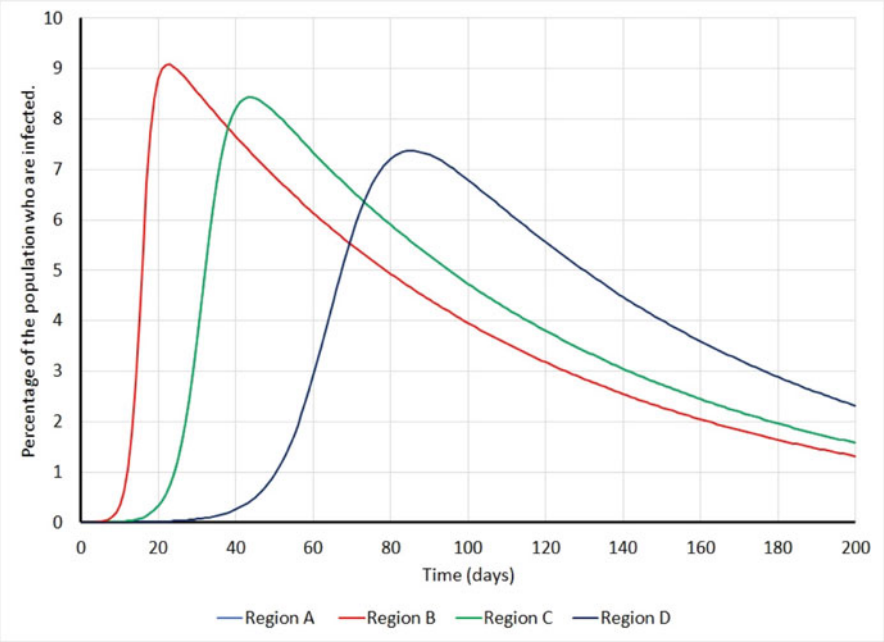



**Fig. 9.1** The proportion of the population who are infected in each of the regions

to consider a country divided into regions by writing in the equations in vector form as [HaBo21]

$$\begin{aligned}
\frac{d\mathbf{S}}{dt} &= -\lambda\mathbf{F} &&+ \alpha T\mathbf{S} \\
\frac{d\mathbf{I}}{dt} &= \lambda\beta\mathbf{F} - \mu_{IR}\mathbf{I} - \mu_{ID}\mathbf{I} &&+ \alpha T\mathbf{I} \\
\frac{d\mathbf{C}}{dt} &= \lambda(1-\beta)\mathbf{F} - \mu_{CR}\mathbf{C} - \mu_{CD}\mathbf{C} &&+ \alpha T\mathbf{C} \\
\frac{d\mathbf{R}}{dt} &= \mu_{IR}\mathbf{I} + \mu_{CR}\mathbf{C} &&+ \alpha T\mathbf{R} \\
\frac{d\mathbf{D}}{dt} &= \mu_{ID}\mathbf{I} + \mu_{CD}\mathbf{C}
\end{aligned}$$

where $F_i = S_i(I_i + C_i)$ and the movement matrix $T$ will be discussed later. Here, $\alpha$ is a parameter which controls the rate at which people move between regions.

However, as shown above, we need to use the population density rather than the population in the differential equation. Harris and Bodmann [HaBo21] developed a model which takes the population density into account and obtained the modified

system of differential equations

$$\begin{aligned}
\frac{d\mathbf{S}}{dt} &= -\lambda A^{-1}\mathbf{F} &&+ \alpha T\mathbf{S} \\
\frac{d\mathbf{I}}{dt} &= \lambda\beta A^{-1}\mathbf{F} - \mu_{IR}\mathbf{I} - \mu_{ID}\mathbf{I} &&+ \alpha T\mathbf{I} \\
\frac{d\mathbf{C}}{dt} &= \lambda(1-\beta)A^{-1}\mathbf{F} - \mu_{CR}\mathbf{C} - \mu_{CD}\mathbf{C} &&+ \alpha T\mathbf{C} \\
\frac{d\mathbf{R}}{dt} &= \mu_{IR}\mathbf{I} + \mu_{CR}\mathbf{C} &&+ \alpha T\mathbf{R} \\
\frac{d\mathbf{D}}{dt} &= \mu_{ID}\mathbf{I} + \mu_{CD}\mathbf{C}
\end{aligned}$$

where $A$ is a diagonal matrix with $A_{ii}$ being the area of the $i$th region.

The movement matrix $T$ which models people moving between regions is given by

$$T_{ij} = \begin{cases} (PM)_{ij} & i \neq j \\ -\sum_{k=1,k\neq i}^{N} (PM)_{ki} & i = j \end{cases}$$

where

$$M_{ij} = \max\left(1 - \frac{d_{ij}}{d_{\max}}, 0\right) \qquad i \neq j$$

$d_{ij}$ is the distance of Region $i$ from Region $j$, $d_{\max}$ is the maximum travel distance and $P$ is a diagonal matrix with

$$P_{ii} = \frac{S_i + I_i + C_i + R_i}{\sum_{i=1}^{N}(S_i + I_i + C_i + R_i)}$$

Here, the notation $(PM)_{ij}$ denotes the element at the $(i, j)$ position in the matrix product $PM$. This choice of the $T$ matrix will ensure that as the populations from the different regions mix, the total population of each region remains the same.

The system of differential equations are solved using an iterative predictor–corrector Crank–Nicholson method [At89, CrNi47]. When calculating the number of people in each category in each region for each day, a sequence of time-steps $h_0 > h_1 > h_2 > \cdots > h_n$ is used to ensure that the calculations have converged to a predetermined accuracy.

## 9.3 Numerical Results

The results presented here are for the UK, where the population data for 2018 is freely available from the UK Office of National Statistics and the initial conditions for the infection are that on day 0 there are 20 carriers in London (unless stated otherwise) and no infected people or carriers in the rest of the country.

In the numerical results presented in this chapter, we will look at the effect of changing some of these parameters. We will primarily look at varying some of the new parameters in the model, such as $\alpha$ and $\beta$ as the effect of changing some of the other parameters has been explored in the previous work on these equations.

We first investigate the effect of changing the distance parameter $\alpha$. To obtain these results, we used $\lambda = 7 \times 10^{-5}$, $\beta = 0.67$, $\mu_{IR} = \mu_{CR} = 0.0714$, $\mu_{ID} = 0.01$ and $\mu_{CD} = 0$. The reason for making $\mu_{CD} = 0$ is that ultimately we are going to apply the model to the spread of COVID-19 in the UK, and since COVID-19 is a notifiable disease in the UK, it effectively means that all cases where the patients die of the disease must have been diagnosed. Figure 9.2 shows the proportion of the population in each region who are either infected or carriers at the time when the total number of people are infected or carriers is maximised. When $\alpha = 0$, meaning that there is no movement of the population across the country, the disease is confined to London, and as $\alpha$ increases, the disease spreads to a larger number of regions in the country. Figure 9.3 shows the total number of people who are susceptible, known to be infected or carriers, who have recovered and who have died in the whole country for the different values of $\alpha$ considered. The graphs in Fig. 9.3 show that changing $\alpha$ has a relatively small effect on the overall number

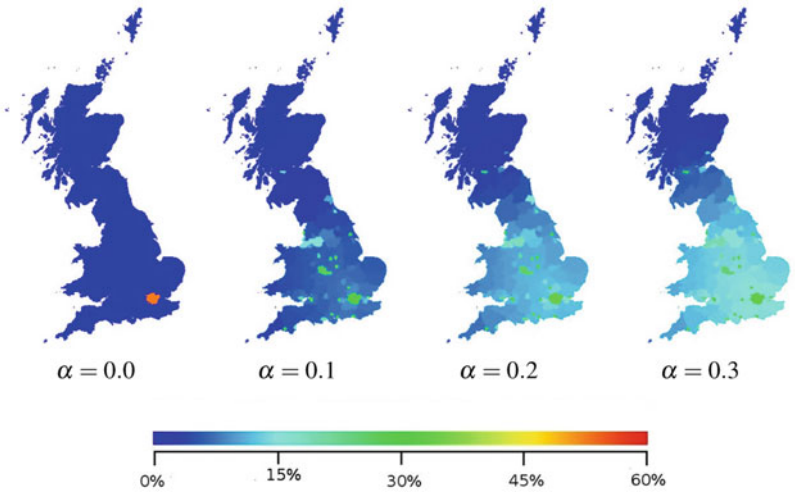



**Fig. 9.2** The proportion of the population who are infected or carriers for each region when the maximum number of people in the whole country are infected

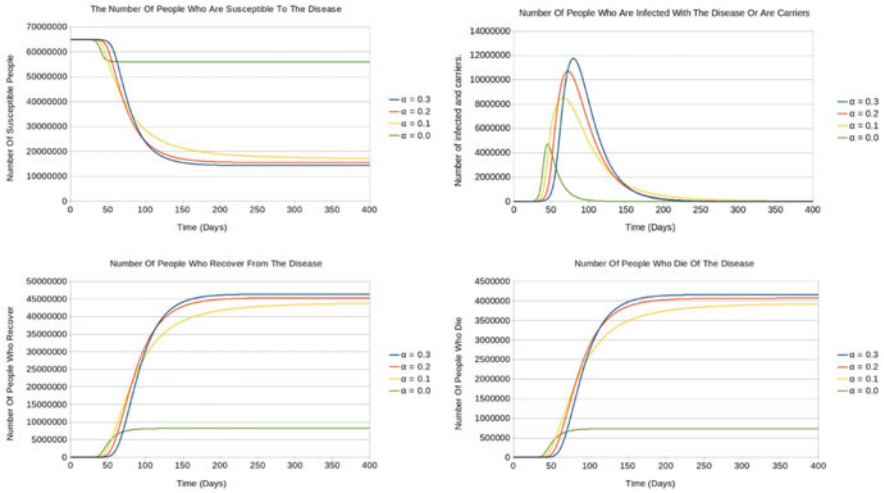



**Fig. 9.3** The number of susceptible people (top left), infected and carriers (top right), recovered people (bottom left) and people who have died (bottom right) as a function of time for different values of the distance parameter $\alpha$

of people affected by the disease but that when there is more movement of people between the regions then the distribution of the people affected by the disease are more spread out over the whole country.

We now investigate the effect of changing $\beta$ which is the proportion of those infected who are diagnosed. For this case, we used $\lambda = 7 \times 10^{-5}$, $\alpha = 0.1$, $\mu_{IR} = \mu_{CR} = 0.0714$, $\mu_{ID} = 0.01$ and $\mu_{CD} = 0$ to calculate our results. Figure 9.4 shows the total number of people who are susceptible, infected or carriers, who have recovered and who have died in the whole country for the different values of $\beta$ considered. These results show that changing the proportion of people who are diagnosed does not have a large effect on the number of susceptible people, people known to be infected and carriers, but it does affect the number of people who die from the disease. This is because a larger proportion of the people infected with the disease are being diagnosed and in the results presented here we have assumed that only diagnosed people can die from the disease and that all the carriers eventually recover.

We also used the model to simulate what happens when the initial infection is located in different cities. In the results presented here, we considered the initial infection to be in London, Birmingham, Manchester and Glasgow using the parameters $\lambda = 4 \times 10^{-5}$, $\beta = 0.67$, $\mu_{IR} = \mu_{CR} = 0.0714$, $\mu_{ID} = 0.01$, $\mu_{CD} = 0$, $\alpha = 0.1$. Figure 9.5 shows the percentage of people who are either infected or are carriers in each region on the day when there is the maximum number of people who are infected or are carriers. These results show that the initial location of the infection has a negligible effect on the number of people who eventually become infected with the disease.

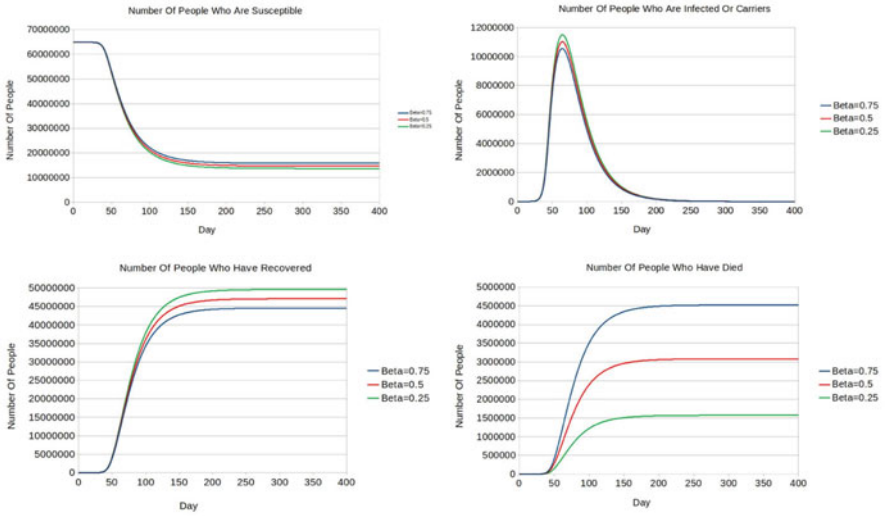



**Fig. 9.4** The number of susceptible people (top left), infected and carriers (top right), recovered people (bottom left) and people who have died (bottom right) as a function of time for different proportions of infected people who are diagnosed

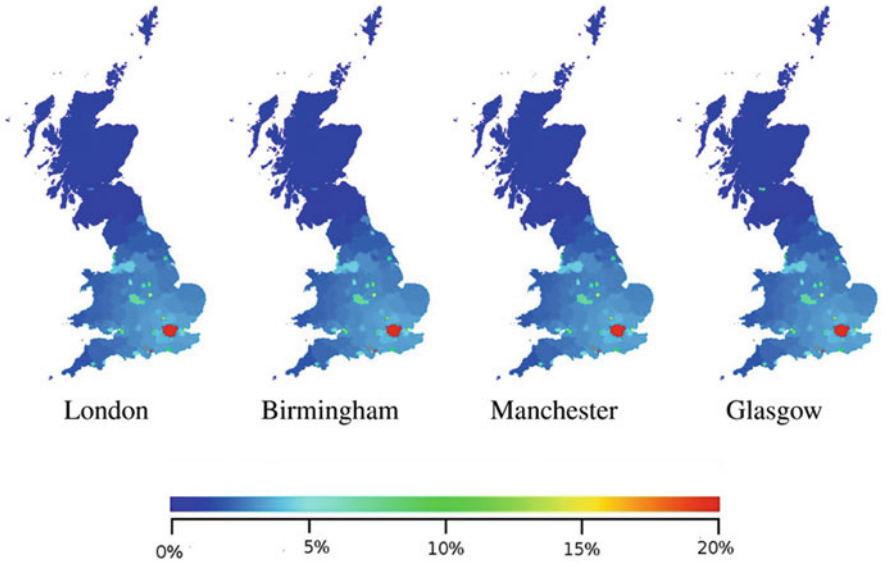



**Fig. 9.5** The percentage of individuals who are infected or are carriers in each region on the day when the maximum number of individuals are infected or are carriers. From left to right, the maps are for the initial infection located in London, Birmingham, Manchester and Glasgow

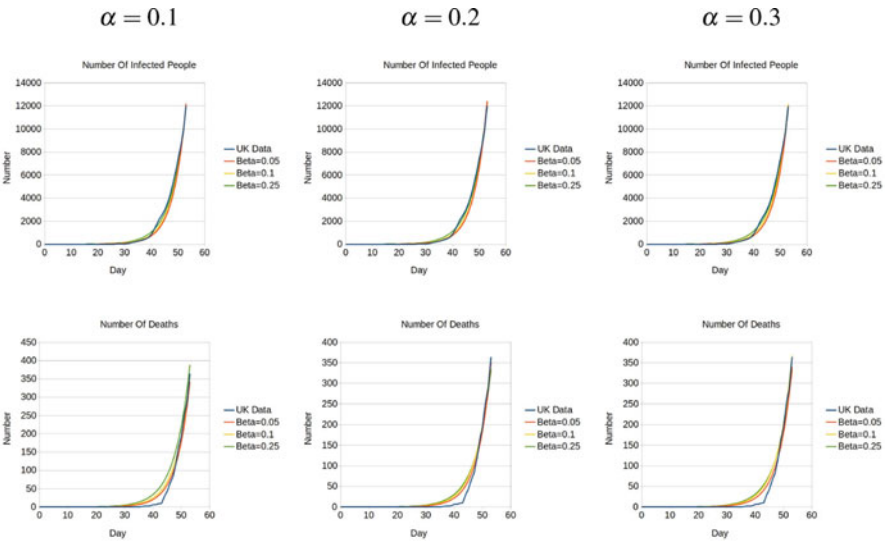



**Fig. 9.6** The simulated and observed numbers of people who have been infected and who have died of COVID-19 in the UK for different values of the parameters $\alpha$ and $\beta$

We can now apply the model to the spread of COVID-19 in the UK. The data for the daily number of infections and deaths from 30 January 2020 is freely available. For the first 53 days, there was no lockdown or restrictions in the UK. We can find $\lambda$ and $\mu_{ID}$ that approximately fit to the data for the first 53 days. We use three different values of $\alpha$ to model different rates of people moving between regions and three values of $\beta$ since the level of testing in the UK at this time was unknown. We used $\mu_{IR} = \mu_{CR} = 0.0714$ which corresponds to it taking 14 days for a person to cease to be infectious, and $\mu_{CD} = 0$ as all recorded deaths occur in people who have been tested (death recorded within 28 days of a positive test). Figure 9.6 shows the parameter which approximately fit the model to the observed data for the UK, and the values of the fitted parameters are shown in Table 9.2. As expected, changing $\alpha$ and $\beta$ does not have major effect on the value of $\mu_{ID}$ since the number of infected people is approximately the same for all the values of $\alpha$ and $\beta$ and so $\mu_{ID}$ will be the same to produce the same number of deaths. However, the results in Table 9.2 show that increasing $\alpha$ for a fixed value of $\beta$ causes $\lambda$ to increase, but increasing $\beta$ for a fixed $\lambda$ causes $\lambda$ to decrease. In other words, if there is more movement of people between the regions, then the infection rate $\lambda$ has to increase to give the same level of infections. However, if a greater proportion of people are diagnosed, then the infection rate has to decrease to give the same number of infections, as expected.

Using the fitted values of $\lambda$ and $\mu_{ID}$, we can use the model to predict the number of people who will be infected with the disease and the number that will die from the disease. Using the values given in Table 9.2 for $\alpha = 0.3$, the results for different values of $\beta$ are shown in Fig. 9.7.

**Table 9.2** The approximate values of the parameters $\lambda$ and $\mu_{ID}$ that approximately model the spread of COVID-19 in the UK for different values of the parameters $\alpha$ and $\beta$

| $\alpha$ | $\beta$ | $\lambda$ | $\mu_{ID}$ |
|---|---|---|---|
| 0.1 | 0.05 | $5.91 \times 10^{-5}$ | 0.006 |
| | 0.10 | $5.67 \times 10^{-5}$ | 0.006 |
| | 0.25 | $5.37 \times 10^{-5}$ | 0.006 |
| 0.2 | 0.05 | $6.60 \times 10^{-5}$ | 0.006 |
| | 0.10 | $6.35 \times 10^{-5}$ | 0.006 |
| | 0.25 | $6.04 \times 10^{-5}$ | 0.005 |
| 0.3 | 0.05 | $7.21 \times 10^{-5}$ | 0.006 |
| | 0.10 | $6.97 \times 10^{-5}$ | 0.006 |
| | 0.25 | $6.64 \times 10^{-5}$ | 0.005 |

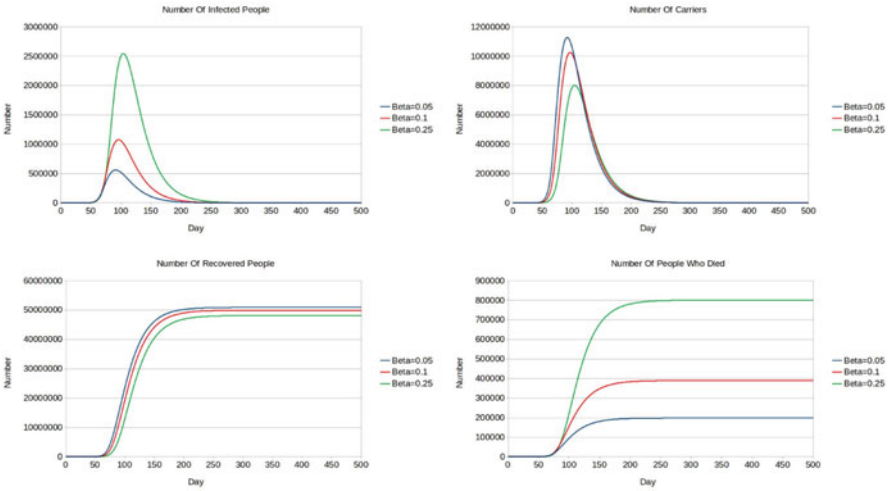



**Fig. 9.7** The predicted number of people who are infected (top left), who are carriers (top right), who recover (bottom left) and who die (bottom right) from COVID-19 in the UK for the case $\alpha = 0.3$

The main process for controlling the spread of COVID-19 in the UK was the use of lockdowns. Lockdowns can be simulated in the model by making the infection rate $\lambda$ and $\alpha$ which controls the rate at which people move between the different regions. Here, we simulated the effects of a lockdown lasting from day 53 to day 151 (which is 14 weeks in duration) where either $\lambda$ is reduced by 90% (which we can call reduced transmission) or $\alpha$ is reduced by 90% (which we shall call reduced travel) or both were reduced by 90%. The results of these calculations for $\alpha = 0.1$ and $\beta = 0.1$ are shown in Fig. 9.8. The results seem to show that provided the transmission is reduced, there is little point in reducing the travel as the curves for reduced travel and both are similar in all of the graphs shown in Fig. 9.8. Furthermore, the graph showing the simulated number of people dying from COVID-19 in the UK shows that the lockdown does not significantly reduce the number of deaths and merely delays them.

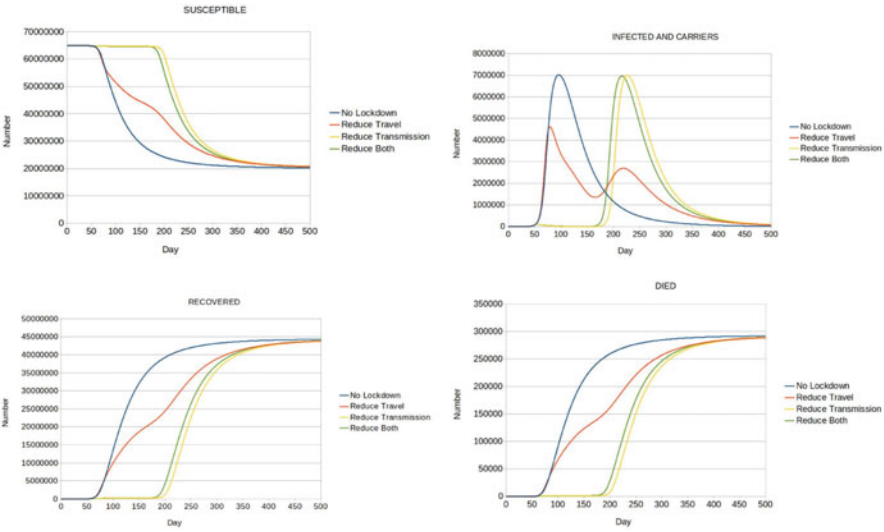



**Fig. 9.8** The predicted number of people who are susceptible (top left), who are infected (top right), who recover (bottom left) and who die (bottom right) from COVID-19 in the UK for $\alpha = 0.1$ and $\beta = 0.1$ when there are different types of lockdown between days 53 and 151

## 9.4 Conclusions

This chapter has presented a method for modelling the spread of a disease across a country divided into different geographical regions each with different population densities. The results show that the proportion of the population who become infected with the disease is greatest in the densely populated cities and that the proportion of the population infected is much lower in sparsely populated rural areas.

One of the important results shown in this chapter is that if the proportion of people travelling increases, then the proportion of people who are infected by the disease in any given area decreases, but more regions are affected by the disease and that the total number of people who have the disease across the whole country does not change significantly. The exception is that if travel is stopped completely, then the disease cannot spread to other regions, but the region where the disease is present is very badly affected by it. In reality, there will always be a small amount of travel between adjacent regions and so the disease will spread.

The results presented in this work also show that changing the initial location of the infection does not significantly affect the final number of people infected or the location with the highest proportion of infected people, although the time at which the peak in the number of infections will be different.

Finally, we have shown that for appropriate values of the parameters the model can be made to match real-world data. Here, we found parameters such that the total number of cases predicted by the model is approximately the same as those

recorded by the UK Government for COVID-19 in the days before the lockdown was imposed in the UK. However, it should be noted that whilst the overall numbers were the same, there was no attempt to fit the simulated values to the regional values that are known, and an area of future research could be to find the parameters that do give a better regional fit. Using the fitted values, we then simulated the effects of some different lockdown regimes on the spread of COVID-19 in the UK, and the results show that enforcing lockdowns that restrict transmission and/or travel only delay the inevitable spread of the disease. Furthermore, the graphs in Fig. 9.8 show that reducing transmission has much more significant effect in reducing the number of people infected with the disease than reducing travel between the regions.

## References

[AlEtAl] Alirol, E., Getaz, L., Stoll, B., Chappuis, F., Loutan, L.: Urbanisation and infectious diseases in a globalised world, Lancet Infect. Diseases **11**, 131–141 (2011)

[At89] Atkinson, K.E.: An Introduction to Numerical Analysis, 2nd edn. Wiley, New York (1989)

[CrNi47] Crank, J., Nicolson, P.: A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type, Math. Proc. Camb. Philos. Soc. **43**, 50–67 (1947)

[HaBo21] Harris P.J., Bodmann, B.E.J: A mathematical model for simulating the spread of a disease through a country divided into geographical regions with different populations densities. Accepted for publication in Math. Biol. (2022).

[He00] Hethcote, H.W.: The mathematics of infectious diseases. SIAM Rev. **42**, 599–653 (200)

[Ne15] Neiderud, C-J.: How urbanization affects the epidemiology of emerging infectious diseases. Infect. Ecol. Epidemiol. **5**, 27060 (2015)

[ReEtAl13] Reyes, R., Ahn, R., Thurber, K., Burke, T.F.: Urbanization and infectious diseases: General principles, historical perspectives, and contemporary challenges. In: Fong, I.W. (ed.) Challenges in Infectious Diseases, pp. 123–146. Springer, New York (2013)

[RiAl20] Riou, J., Althaus, C.L.: Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. Eurosurveillance **25**, 7–11 (2020)

[SaEtAl20] Sanche, S., Lin, Y.T., Xu, C., Romero-Severson, E., Hengartner, N., Ke, R.: High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. Emerg. Infect. Diseases **26**, 1470–1477 (2020)

[WuEtAl20] Wu, J.T., Leung, K., Leung, G.M.: Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. LANCET **395**, 689–697 (2020)

# Chapter 10
# Computing Elastic Interior Transmission Eigenvalues


**A. Kleefeld and M. Zimmermann**

## 10.1 Introduction

Non-destructive testing is an important tool to check whether a given object is homogeneous or not without destroying it. Interior transmission eigenvalues (ITEs) may have the potential to serve as an indicator whether an object is homogeneous or not due to a monotonicity result. If the object is not homogeneous, they might indicate where and how large the inhomogeneity is. Hence, they can be seen as a "fingerprint" of a given object. Therefore, it is of great interest to numerically calculate them for arbitrary domains to high accuracy.

They also play an important role in the theory for scattering problems. Precisely, algorithms such as the (general) linear sampling method or the factorization method to reconstruct the scattering object from the scattered field are not theoretically justified for such eigenvalues. Usually, time-harmonic acoustic, electromagnetic, or elastic scattering problems are considered. Recent work is now focusing on the latter one as we do here, too.

Unfortunately, the resulting system of partial differential equations, containing two Navier equations, are coupled by transmission conditions and lead therefore to a non-self-adjoint and non-elliptic problem. However, one can cope with this problem. Existing methods like the inside–outside duality method [Pe16] do not report numerical results, the method of fundamental solutions only works well for small perturbations of a circle [KlPi20], and variants of the finite element method

A. Kleefeld (✉)
Forschungszentrum Jülich GmbH, Jülich Supercomputing Centre, Jülich, Germany
e-mail: a.kleefeld@fz-juelich.de

M. Zimmermann
Fachhochschule Aachen Campus Jülich, Medical Engineering and Technomathematics, Jülich, Germany
e-mail: m.zimmermann@fh-aachen.de



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_10

only work well for polygonal domains [ChLiWa20, JiLiSu18, JiLiSu20, XiJi18, XiJiGe18, XiJiZh21, YaHaBi20, YaEtAl20].

An alternative is to use the boundary element method which works very well for domains with smooth boundaries to obtain numerical results to high accuracy. However, it can only be used for constant coefficients and the fundamental solution needs to be known. Luckily this is the case for the Navier equation, but one needs to solve a nonlinear eigenvalue problem which can be done with Beyn's algorithm [Be12] as done for the acoustic transmission problem [Kl13]. A first attempt has been made in [We18], but certain integral operators were too complicated to be approximated. An improvement is given in [Zi21] fully avoiding this integral operator by using a difference of Dirichlet-to-Neumann maps which has been successfully applied to the acoustic transmission problem in [CaKr17]. However, the numerical approximation of the singular integrals to high accuracy is complicated. Here, we use an approach which fully avoids the numerical calculation of singular integrals.

The existence of a countable number of real ITEs is known [BeCaGu13], but the existence of complex ITEs is still open, but with our approach we are able to give numerical results indicating that they do exist.

The chapter is organized as follows: first, we present the elastic interior transmission problem. Next, we illustrate how to solve it with the boundary element method using a difference of Dirichlet-to-Neumann maps. Then, the resulting integral equation is approximated by the boundary element collocation method, and the emerging nonlinear eigenvalue problem is solved with Beyn's algorithm. Numerical results are given to show the correct approximations for two test cases. Finally, numerical results are reported for a variety of domains and compared with existing results. A short summary and an outlook are given at the end.

## 10.2 Elastic Transmission Eigenvalue Problem

Let $D \subset \mathbb{R}^2$ be a bounded open domain that is simply connected. Its boundary $\partial D$ is given parametrically by $\mathbf{p}(\theta)$ with $\theta \in [0, 2\pi]$. We assume that $\partial D$ is a simple, closed curve with finite length satisfying $\mathbf{p}(0) = \mathbf{p}(2\pi)$, $\mathbf{p} \in C^2([0, 2\pi])$, and $\mathbf{p}'(\theta) \neq \mathbf{0}$ for all $\theta \in [0, 2\pi]$.

Time-harmonic elastic scattering with frequency $\omega$ can be described by the Navier equation

$$\mu\, \Delta \mathbf{u} + (\lambda + \mu)\, \text{grad div}\, \mathbf{u} + \omega^2\, \rho\, \mathbf{u} = \mathbf{0} \;\text{ in } D \subset \mathbb{R}^2\,, \tag{10.1}$$

where $\mathbf{u}(\mathbf{x}) = (u_1(\mathbf{x}), u_2(\mathbf{x}))^\top$ is the displacement field at the point $\mathbf{x} = (x_1, x_2)^\top \in \mathbb{R}^2$. Here, the parameter $\rho > 0$ is the mass density of the medium and assumed to be constant. The parameters $\lambda$ and $\mu$ are the Lamé parameters and describe the elastic material. They satisfy the conditions $\mu > 0$ and $2\mu + \lambda > 0$ ([Mc00, p. 297 ff.]).

Assume now that $D$ with mass density $\rho_1$ is contained in a medium with mass density $\rho_0$ with $\rho_1 > \rho_0$. Is there an incident field satisfying the Navier equation

that does not scatter? This leads to the elastic interior transmission problem: find $\omega^2$ and a nontrivial solution $(\mathbf{u}, \mathbf{v})$ such that

$$\mu\,\Delta\mathbf{u} + (\lambda+\mu)\,\operatorname{grad}\operatorname{div}\mathbf{u} + \omega^2\,\rho_0\,\mathbf{u} = \mathbf{0} \qquad \text{in } D\,, \tag{10.2}$$

$$\mu\,\Delta\mathbf{v} + (\lambda+\mu)\,\operatorname{grad}\operatorname{div}\mathbf{v} + \omega^2\,\rho_1\,\mathbf{v} = \mathbf{0} \qquad \text{in } D\,, \tag{10.3}$$

$$\mathbf{u} = \mathbf{v} \qquad \text{on } \partial D\,, \tag{10.4}$$

$$\mathbf{T}(\mathbf{u}) = \mathbf{T}(\mathbf{v}) \text{ on } \partial D \tag{10.5}$$

is satisfied, where

$$\mathbf{T}(\mathbf{f}) = \lambda\,\operatorname{div}(\mathbf{f})\,\nu \;+\; 2\,\mu\,(\nu^\top \operatorname{grad})\,\mathbf{f} \;+\; \mu\,\operatorname{div}(\mathrm{Q}\,\mathbf{f})\,\mathrm{Q}\,\nu$$

with the normalized vector $\nu = (\nu_1, \nu_2)^\top$ on $\partial D$ pointing into the exterior of $D$ and the matrix

$$\mathrm{Q} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Then, the parameter $\omega$ is an elastic interior transmission eigenvalue (EITE). The existence of real EITEs is known [BeCaGu13], and however the existence of complex EITEs is still open.

We will use boundary integral equations to solve the problem at hand. The matrix-valued fundamental solution is given by

$$\begin{aligned}\mathrm{K}_\omega(\mathbf{x},\mathbf{y}) = {}& \frac{\mathrm{i}}{4\mu} H_0^{(1)}(k_s\|\mathbf{x}-\mathbf{y}\|)\mathrm{I}_2 \\ & + \frac{\mathrm{i}}{4\omega^2}\,\operatorname{grad}_{\mathbf{x}}\operatorname{grad}_{\mathbf{x}}^\top \left[H_0^{(1)}(k_s\|\mathbf{x}-\mathbf{y}\|) - H_0^{(1)}(k_p\|\mathbf{x}-\mathbf{y}\|)\right] \in \mathbb{C}^{2x2}\,,\end{aligned}$$

where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$ with $\mathbf{x} \neq \mathbf{y}$, $\|\cdot\|$ denotes the Euclidean norm, and $\mathrm{I}_2$ is the $2 \times 2$ identity matrix. The function $H_0^{(1)}$ is the Hankel function of the first kind of order 0. The parameters $k_p$ and $k_s$ are the wave numbers of the shear and the pressure wave, respectively. They are given by

$$k_s^2 = \frac{\omega^2}{\mu} \quad \text{and} \quad k_p^2 = \frac{\omega^2}{\lambda+2\mu}\,.$$

The elastic single-layer operator defined by

$$\mathbf{u}(\mathbf{x}) = (\mathrm{SL}_\omega\mathbf{g})\,(\mathbf{x}) = \int_{\partial D} \mathrm{K}_\omega(\mathbf{x},\mathbf{y})\mathbf{g}(\mathbf{y})\,\mathrm{d}s(\mathbf{y})\,, \quad \mathbf{x} \in D,$$

as well as the elastic double-layer operator defined by

$$\mathbf{u}(\mathbf{x}) = (\mathrm{DL}_\omega \mathbf{h})(\mathbf{x}) = \int_{\partial D} \left[\mathbf{T}_\mathbf{y}\left(\mathrm{K}_\omega(\mathbf{x},\mathbf{y})\right)\right]^\top \mathbf{h}(\mathbf{y})\, \mathrm{d}s(\mathbf{y})\,, \quad \mathbf{x} \in D,$$

with unknown functions $\mathbf{g}$ and $\mathbf{h}$ solve the Navier equation (10.1). Note that the traction of a matrix is applied to each column. The unknown functions $\mathbf{g}$ and $\mathbf{h}$ are then determined by letting the point $\mathbf{x} \in D$ approach the boundary and using the given boundary condition incorporating the jump conditions of the elastic boundary layer operators defined by

$$\begin{aligned}
(\mathrm{S}_\omega \mathbf{g})(\mathbf{x}) &= \int_{\partial D} \mathrm{K}_\omega(\mathbf{x},\mathbf{y})\mathbf{g}(\mathbf{y})\, \mathrm{d}s(\mathbf{y})\,, \quad \mathbf{x} \in \partial D\,,\\
\left(\mathrm{D}_\omega^\top \mathbf{g}\right)(\mathbf{x}) &= \int_{\partial D} \mathbf{T}_\mathbf{y}\left(\mathrm{K}_\omega(\mathbf{x},\mathbf{y})\right)\mathbf{g}(\mathbf{y})\, \mathrm{d}s(\mathbf{y})\,, \quad \mathbf{x} \in \partial D\,,\\
(\mathrm{D}_\omega \mathbf{h})(\mathbf{x}) &= \int_{\partial D} \left[\mathbf{T}_\mathbf{y}\left(\mathrm{K}_\omega(\mathbf{x},\mathbf{y})\right)\right]^\top \mathbf{h}(\mathbf{y})\, \mathrm{d}s(\mathbf{y})\,, \quad \mathbf{x} \in \partial D\,.
\end{aligned}$$

The first operator is the elastic boundary single-layer operator, the second operator is the traction of the elastic boundary single-layer operator, and the third operator is the elastic boundary double-layer operator. To solve (10.2)–(10.5), we use the idea given in [CaKr17]. The following ansatz

$$\mathbf{u} = \mathrm{SL}_{\omega\sqrt{\rho_0}}\mathbf{g} \quad \text{and} \quad \mathbf{v} = \mathrm{SL}_{\omega\sqrt{\rho_1}}\mathbf{h}$$

solves (10.2) and (10.3) in $D$. The functions $\mathbf{g}$ and $\mathbf{h}$ are unknown. Letting the point approach the boundary yields

$$\mathbf{u} = \mathrm{S}_{\omega\sqrt{\rho_0}}\mathbf{g} \quad \text{and} \quad \mathbf{v} = \mathrm{S}_{\omega\sqrt{\rho_1}}\mathbf{h} \quad \text{on } \partial D\,.$$

Taking the traction along with the jump conditions yields

$$\mathbf{T}(\mathbf{u}) = \left(\frac{1}{2}\mathrm{I} + \mathrm{D}^\top_{\omega\sqrt{\rho_0}}\right)\mathbf{g} \quad \text{and} \quad \mathbf{T}(\mathbf{v}) = \left(\frac{1}{2}\mathrm{I} + \mathrm{D}^\top_{\omega\sqrt{\rho_1}}\right)\mathbf{h} \quad \text{on } \partial D\,,$$

where I denotes the identity operator. Combining the last two equations gives

$$\mathbf{T}(\mathbf{u}) = \left(\frac{1}{2}\mathrm{I} + \mathrm{D}^\top_{\omega\sqrt{\rho_0}}\right)\mathrm{S}^{-1}_{\omega\sqrt{\rho_0}}\mathbf{u} \text{ and} \tag{10.6}$$

$$\mathbf{T}(\mathbf{v}) = \left(\frac{1}{2}\mathrm{I} + \mathrm{D}^\top_{\omega\sqrt{\rho_1}}\right)\mathrm{S}^{-1}_{\omega\sqrt{\rho_1}}\mathbf{v} \quad \text{on } \partial D\,, \tag{10.7}$$

where we assume that $\omega^2\rho_0$ and $\omega^2\rho_1$ are not eigenvalues of the operator $\Delta^* := \mu\,\Delta\mathbf{u} + (\lambda+\mu)\,\mathrm{grad}\,\mathrm{div}\,\mathbf{u}$ with boundary condition $\mathbf{u} = \mathbf{0}$. Because of the boundary

condition (10.4), we can replace $\mathbf{v}$ by $\mathbf{u}$ in (10.7). Next, we take the difference of (10.6) and (10.7) and apply the boundary condition (10.5) yielding

$$\underbrace{\left[\left(\frac{1}{2}\mathrm{I}+\mathrm{D}^{\top}_{\omega\sqrt{\rho_0}}\right)\mathrm{S}^{-1}_{\omega\sqrt{\rho_0}}-\left(\frac{1}{2}\mathrm{I}+\mathrm{D}^{\top}_{\omega\sqrt{\rho_1}}\right)\mathrm{S}^{-1}_{\omega\sqrt{\rho_1}}\right]}_{=:\mathrm{N}(\omega)}\mathbf{u}=\mathbf{0}\quad\text{on }\partial D\,.$$

Then, the solution of the nonlinear eigenvalue problem $\mathrm{N}(\omega)\mathbf{u}=\mathbf{0}$, $\mathbf{u}\neq\mathbf{0}$ will be a solution of (10.2)–(10.5). However, we will consider the transpose of this equation to avoid the use of the traction of the elastic single-layer operator. Hence, we consider

$$\underbrace{\left[\mathrm{S}^{-1}_{\omega\sqrt{\rho_0}}\left(\frac{1}{2}\mathrm{I}+\mathrm{D}_{\omega\sqrt{\rho_0}}\right)-\mathrm{S}^{-1}_{\omega\sqrt{\rho_1}}\left(\frac{1}{2}\mathrm{I}+\mathrm{D}_{\omega\sqrt{\rho_1}}\right)\right]}_{=:\mathrm{M}(\omega)}\mathbf{u}=\mathbf{0}\quad\text{on }\partial D$$

and need to solve the problem

$$\mathrm{M}(\omega)\mathbf{u}=\mathbf{0}\,,\quad \mathbf{u}\neq\mathbf{0}$$

assuming $\omega^2\rho_0$ and $\omega^2\rho_1$ are not eigenvalues of the operator $\Delta^*$ with boundary condition $\mathbf{u}=\mathbf{0}$.

## 10.3 The Discretization of the Operators $\frac{1}{2}\mathrm{I}+\mathrm{D}_\omega$ and $\mathrm{S}_\omega$

In this section, we illustrate how to solve a given boundary integral equation with the boundary element collocation method which we will also use later to approximate the operators $\mathrm{S}_\omega$ and $\frac{1}{2}\mathrm{I}+\mathrm{D}_\omega$ for a given $\omega$. As an illustrative example, we want to solve the problem $\Delta^*\mathbf{u}+\omega^2\mathbf{u}=0$ in $\mathbb{R}^2\backslash\overline{D}$ with the boundary conditions $\mathbf{u}=\mathbf{f}$, where $\mathbf{f}$ is a given function defined on the boundary. The frequency $\omega$ is given as well. Using the double-layer ansatz $\mathbf{u}=\mathrm{DL}_\omega\mathbf{h}$ in $\mathbb{R}^2\backslash\overline{D}$ together with the jump condition yields the boundary integral equation of the second kind

$$\frac{1}{2}\mathbf{h}+\mathrm{D}_\omega\mathbf{h}=\mathbf{f}\,. \tag{10.8}$$

Now, we illustrate how to solve this equation numerically. First, we define for a given even $n$ the equidistant angles $\theta_j=2\pi(j-1)/n$, $j=1,\ldots,n$. With this, we define the nodes $\mathbf{v}_j=\mathbf{p}(\theta_j)$. Next, we define the line segments $\Delta_i\subset\partial D$, where the $i$-th segment has the starting point $\mathbf{v}_{2i-1}$ and the end point $\mathbf{v}_{2i+1}$ and a point in between $\mathbf{v}_{2i}$, $i=1,\ldots,n/2$. Note that $\mathbf{v}_{n+1}=\mathbf{v}_1$ since $\partial D$ is closed. Hence, the given boundary $\partial D$ can be written as the union of all $\Delta_i$. Therefore, Eq. (10.8) can

be written as

$$\frac{1}{2}\mathbf{h}(\mathbf{x}) + \sum_{i=1}^{n/2} \int_{\Delta_i} \left[\mathbf{T}_{\mathbf{y}} \left(\mathrm{K}_\omega(\mathbf{x},\mathbf{y})\right)\right]^\top \mathbf{h}(\mathbf{y})\, \mathrm{d}s(\mathbf{y}) = \mathbf{f}(\mathbf{x})\,, \quad \mathbf{x} \in \partial D\,.$$

It can be shown that there exists a bijective map $\mathbf{m}_i : \sigma = [0,1] \to \Delta_i$ for each $i = 1, \ldots, n/2$. Using a change of variables yields

$$\frac{1}{2}\mathbf{h}(\mathbf{x}) + \sum_{i=1}^{n/2} \int_{\sigma} \left[\mathbf{T}_{\mathbf{m}_i(s)} \left(\mathrm{K}_\omega(\mathbf{x},\mathbf{m}_i(s))\right)\right]^\top \mathbf{h}(\mathbf{m}_i(s)) J_i(s)\, \mathrm{d}s(s) = \mathbf{f}(\mathbf{x})\,, \quad \mathbf{x} \in \partial D\,,$$

where $J_i(s) = \|\partial_s \mathbf{m}_i(s)\|$ is the Jacobian. The map $\mathbf{m}_i$ is approximated by a quadratic interpolation polynomial $\widetilde{\mathbf{m}}_i(s) = \sum_{j=1}^{3} \mathbf{m}_i(q_j) L_j(s)$ with the Lagrange basis function $L_1(s) = (1-s)(1-2s)$, $L_2(s) = 4s(1-s)$, and $L_3(s) = s(2s-1)$ and $q_1 = 0$, $q_2 = 1/2$, and $q_3 = 1$. Note that $\mathbf{m}_i(q_j)$ selects the corresponding nodes $\mathbf{v}_{2i-1}$, $\mathbf{v}_{2i}$, and $\mathbf{v}_{2i+1}$. We approximately obtain

$$\frac{1}{2}\mathbf{h}(\mathbf{x}) + \sum_{i=1}^{n/2} \int_{\sigma} \left[\mathbf{T}_{\widetilde{\mathbf{m}}_i(s)} \left(\mathrm{K}_\omega(\mathbf{x},\widetilde{\mathbf{m}}_i(s))\right)\right]^\top \mathbf{h}(\widetilde{\mathbf{m}}_i(s)) \widetilde{J}_i(s)\, \mathrm{d}s(s) \approx \mathbf{f}(\mathbf{x})\,, \quad \mathbf{x} \in \partial D\,,$$

where $\widetilde{J}_i(s) = \|\partial_s \widetilde{\mathbf{m}}_i(s)\|$ is the Jacobian. We define for a given $0 < \alpha < 1/2$ the collocation nodes $\widetilde{\mathbf{v}}_{i,k} = \widetilde{\mathbf{m}}_i(\widetilde{q}_k)$ for $i = 1, \ldots, n/2$ and $k = 1, 2, 3$ with $\widetilde{q}_1 = \alpha$, $\widetilde{q}_2 = 1/2$, and $\widetilde{q}_3 = 1 - \alpha$. We now approximate each component of the unknown function $\mathbf{h}$ by a quadratic interpolation polynomial using the three nodes $\widetilde{q}_k$ and the three Lagrange basis functions

$$\widetilde{L}_1(s) = \frac{1-s-\alpha}{1-2\alpha}\frac{1-2s}{1-2\alpha}\,,\ \widetilde{L}_2(s) = 4\frac{s-\alpha}{1-2\alpha}\frac{1-s-\alpha}{1-2\alpha}\,,\ \widetilde{L}_3(s) = \frac{s-\alpha}{1-2\alpha}\frac{2s-1}{1-2\alpha}\,.$$

Precisely, we use

$$\mathbf{h}(\widetilde{\mathbf{m}}_i(s)) \approx \sum_{k=1}^{3} \mathbf{h}(\widetilde{\mathbf{m}}_i(\widetilde{q}_k)) \widetilde{L}_k(s) = \sum_{k=1}^{3} \mathbf{h}(\widetilde{\mathbf{v}}_{i,k}) \widetilde{L}_k(s),$$

and therefore, we obtain

$$\begin{aligned}\frac{1}{2}\mathbf{h}(\mathbf{x}) + \sum_{i=1}^{n/2}\sum_{k=1}^{3} \int_{\sigma} \left[\mathbf{T}_{\widetilde{\mathbf{m}}_i(s)} \left(\mathrm{K}_\omega(\mathbf{x},\widetilde{\mathbf{m}}_i(s))\right)\right]^\top \widetilde{J}_i(s) \widetilde{L}_k(s)\, \mathrm{d}s(s) \mathbf{h}(\widetilde{\mathbf{v}}_{i,k}) - \mathbf{f}(\mathbf{x}) \\ \approx \mathbf{r}(x)\,, \quad \mathbf{x} \in \partial D\,,\end{aligned}$$

with the residual $\mathbf{r}(x)$. We force the residual to be zero at the collocation nodes $\widetilde{\mathbf{v}}_{j,\ell}$, which leads to the linear system of size $3n \times 3n$

$$\frac{1}{2}\mathbf{h}(\widetilde{\mathbf{v}}_{j,\ell}) + \sum_{i=1}^{n/2}\sum_{k=1}^{3} {}^{\omega}A_{(i,k),(j,\ell)}\mathbf{h}(\widetilde{\mathbf{v}}_{i,k}) = \mathbf{f}(\widetilde{\mathbf{v}}_{j,\ell}) \tag{10.9}$$

with

$${}^{\omega}A_{(i,k),(j,\ell)} = \int_{\sigma} \left[\mathbf{T}_{\widetilde{\mathbf{m}}_i(s)}\left(\mathrm{K}_{\omega}(\widetilde{\mathbf{v}}_{j,\ell}, \widetilde{\mathbf{m}}_i(s))\right)\right]^{\top} \widetilde{J}_i(s)\widetilde{L}_k(s)\,\mathrm{d}s(s) \in \mathbb{C}^{2\times 2}$$

since the $(i,k)$, $(j,\ell)$-entry is a $2 \times 2$ matrix. All four elements of the $2 \times 2$ matrix are

$${}^{\omega}A^{(1,1)}_{(i,k),(j,\ell)} = \int_{\sigma} t^{(1,1)}_{i,j,\ell}(s)\widetilde{J}_i(s)\widetilde{L}_k(s)\,\mathrm{d}s(s)\,,\ {}^{\omega}A^{(1,2)}_{(i,k),(j,\ell)} = \int_{\sigma} t^{(1,2)}_{i,j,\ell}(s)\widetilde{J}_i(s)\widetilde{L}_k(s)\,\mathrm{d}s(s)$$

$${}^{\omega}A^{(2,1)}_{(i,k),(j,\ell)} = \int_{\sigma} t^{(2,1)}_{i,j,\ell}(s)\widetilde{J}_i(s)\widetilde{L}_k(s)\,\mathrm{d}s(s)\,,\ {}^{\omega}A^{(2,2)}_{(i,k),(j,\ell)} = \int_{\sigma} t^{(2,2)}_{i,j,\ell}(s)\widetilde{J}_i(s)\widetilde{L}_k(s)\,\mathrm{d}s(s)$$

with

$$\begin{aligned}
\mathbf{t}^{(1,1)}_{i,j,\ell}(s) = {} & \frac{c_1}{\|\mathbf{d}_{i,j,\ell}(s)\|}\left[(-\lambda-2\mu)\nu_1(y)d^{(1)}_{i,j,\ell}(s) - \mu\nu_2(y)d^{(2)}_{i,j,\ell}(s)\right] \\
& + \frac{c_2}{\|\mathbf{d}_{i,j,\ell}(s)\|^3}\left[(-\lambda-2\mu)\nu_1(y)\left(d^{(1)}_{i,j,\ell}(s)\right)^3\right. \\
& \left. - \lambda\nu_1(y)d^{(1)}_{i,j,\ell}(s)\left(d^{(2)}_{i,j,\ell}(s)\right)^2 - 2\mu\nu_2(y)\left(d^{(1)}_{i,j,\ell}(s)\right)^2 d^{(2)}_{i,j,\ell}(s)\right] \\
& + \frac{c_3}{\|\mathbf{d}_{i,j,\ell}(s)\|^4}\left[(-\lambda-4\mu)\nu_1(y)d^{(1)}_{i,j,\ell}(s) - \mu\nu_2(y)d^{(2)}_{i,j,\ell}(s)\right. \\
& \left. + 4\mu\left(\nu_1(y)\left(d^{(1)}_{i,j,\ell}(s)\right)^3 + \nu_2(y)\left(d^{(1)}_{i,j,\ell}(s)\right)^2 d^{(2)}_{i,j,\ell}(s)\right)\right] \\
\mathbf{t}^{(1,2)}_{i,j,\ell}(s) = {} & \frac{c_1}{\|\mathbf{d}_{i,j,\ell}(s)\|}\left[-\lambda\nu_2(y)d^{(1)}_{i,j,\ell}(s) - \mu\nu_1(y)d^{(2)}_{i,j,\ell}(s)\right] \\
& + \frac{c_2}{\|\mathbf{d}_{i,j,\ell}(s)\|^3}\left[(-\lambda-2\mu)\nu_2(y)d^{(1)}_{i,j,\ell}(s)\left(d^{(2)}_{i,j,\ell}(s)\right)^2\right. \\
& \left. - \lambda\nu_2(y)\left(d^{(1)}_{i,j,\ell}(s)\right)^3 - 2\mu\nu_1(y)\left(d^{(1)}_{i,j,\ell}(s)\right)^2 d^{(2)}_{i,j,\ell}(s)\right] \\
& + \frac{c_3}{\|\mathbf{d}_{i,j,\ell}(s)\|^4}\left[(-\lambda-2\mu)\nu_2(y)d^{(1)}_{i,j,\ell}(s) - \mu\nu_1(y)d^{(2)}_{i,j,\ell}(s)\right.
\end{aligned}$$

$$
\begin{aligned}
&+ 4\mu\left(\nu_1(y)\left(d^{(1)}_{i,j,\ell}(s)\right)^2 d^{(2)}_{i,j,\ell}(s) + \nu_2(y) d^{(1)}_{i,j,\ell}(s)\left(d^{(2)}_{i,j,\ell}(s)\right)^2\right)\Bigg]\\
\mathbf{t}^{(2,1)}_{i,j,\ell}(s) &= \frac{c_1}{\|\mathbf{d}_{i,j,\ell}(s)\|}\left[-\lambda\nu_1(y) d^{(2)}_{i,j,\ell}(s) - \mu\nu_2(y) d^{(1)}_{i,j,\ell}(s)\right]\\
&+ \frac{c_2}{\|\mathbf{d}_{i,j,\ell}(s)\|^3}\Bigg[(-\lambda-2\mu)\nu_1(y)\left(d^{(1)}_{i,j,\ell}(s)\right)^2 d^{(2)}_{i,j,\ell}(s)\\
&- \lambda\nu_1(y)\left(d^{(2)}_{i,j,\ell}(s)\right)^3 - 2\mu\nu_2(y) d^{(1)}_{i,j,\ell}(s)\left(d^{(2)}_{i,j,\ell}(s)\right)^2\Bigg]\\
&+ \frac{c_3}{\|\mathbf{d}_{i,j,\ell}(s)\|^4}\Bigg[(-\lambda-2\mu)\nu_1(y) d^{(2)}_{i,j,\ell}(s) - \mu\nu_2(y) d^{(1)}_{i,j,\ell}(s)\\
&+ 4\mu\left(\nu_1(y)\left(d^{(1)}_{i,j,\ell}(s)\right)^2 d^{(2)}_{i,j,\ell}(s) + \nu_2(y) d^{(1)}_{i,j,\ell}(s)\left(d^{(2)}_{i,j,\ell}(s)\right)^2\right)\Bigg]\\
\mathbf{t}^{(2,2)}_{i,j,\ell}(s) &= \frac{c_1}{\|\mathbf{d}_{i,j,\ell}(s)\|}\left[(-\lambda-2\mu)\nu_2(y) d^{(2)}_{i,j,\ell}(s) - \mu\nu_1(y) d^{(1)}_{i,j,\ell}(s)\right]\\
&+ \frac{c_2}{\|\mathbf{d}_{i,j,\ell}(s)\|^3}\Bigg[(-\lambda-2\mu)\nu_2(y)\left(d^{(2)}_{i,j,\ell}(s)\right)^3\\
&- \lambda\nu_2(y)\left(d^{(1)}_{i,j,\ell}(s)\right)^2 d^{(2)}_{i,j,\ell}(s) - 2\mu\nu_1(y) d^{(1)}_{i,j,\ell}(s)\left(d^{(2)}_{i,j,\ell}(s)\right)^2\Bigg]\\
&+ \frac{c_3}{\|\mathbf{d}_{i,j,\ell}(s)\|^4}\Bigg[(-\lambda-4\mu)\nu_2(y) d^{(2)}_{i,j,\ell}(s) - \mu\nu_1(y) d^{(1)}_{i,j,\ell}(s)\\
&+ 4\mu\left(\nu_1(y) d^{(1)}_{i,j,\ell}(s)\left(d^{(2)}_{i,j,\ell}(s)\right)^2 + \nu_2(y)\left(d^{(2)}_{i,j,\ell}(s)\right)^3\right)\Bigg],
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{d}_{i,j,\ell}(s) &= \widetilde{\mathbf{v}}_{j,\ell} - \widetilde{\mathbf{m}}_i(s)\\
c_1 &= -\frac{\mathrm{i}k_s}{4\mu}H_1^{(1)}(k_s\|\mathbf{d}_{i,j,\ell}(s)\|)\\
&- \frac{\mathrm{i}}{4\omega^2\|\mathbf{d}_{i,j,\ell}(s)\|}\Bigg[2\frac{k_p H_1^{(1)}(k_p\|\mathbf{d}_{i,j,\ell}(s)\|) - k_s H_1^{(1)}(k_s\|\mathbf{d}_{i,j,\ell}(s)\|)}{\|\mathbf{d}_{i,j,\ell}(s)\|}\\
&- k_p^2 H_0^{(1)}(k_p\|\mathbf{d}_{i,j,\ell}(s)\|) + k_s^2 H_0^{(1)}(k_s\|\mathbf{d}_{i,j,\ell}(s)\|)\Bigg],\\
c_2 &= \frac{\mathrm{i}}{2\omega^2\|\mathbf{d}_{i,j,\ell}(s)\|}\Bigg[k_s^2 H_0^{(1)}(k_s\|\mathbf{d}_{i,j,\ell}(s)\|) - k_p^2 H_0^{(1)}(k_p\|\mathbf{d}_{i,j,\ell}(s)\|)\\
&+ 2\frac{k_p H_1^{(1)}(k_p\|\mathbf{d}_{i,j,\ell}(s)\|) - k_s H_1^{(1)}(k_s\|\mathbf{d}_{i,j,\ell}(s)\|)}{\|\mathbf{d}_{i,j,\ell}(s)\|}\Bigg]
\end{aligned}
$$

$$+\frac{\mathrm{i}}{4\omega^2}\left[k_s^3 H_1^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)-k_p^3 H_1^{(1)}(k_p\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right],$$

$$c_3=\frac{\mathrm{i}}{2\omega^2\,\|\mathbf{d}_{i,j,\ell}(s)\|}\left[k_s H_1^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)-k_p H_1^{(1)}(k_p\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right]$$

$$+\frac{\mathrm{i}}{4\omega^2}\left[k_p^2 H_0^{(1)}(k_p\,\|\mathbf{d}_{i,j,\ell}(s)\|)-k_s^2 H_0^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right].$$

The four integrals in $A_{(i,k),(j,\ell)}$ have to be evaluated numerically which is done with an automatic integration routine using adaptive quadrature (refer to the software package QUADPACK). However, when $(i,k)=(j,\ell)$, a singularity is present. In this case, we use a singularity subtraction of the form

$$\begin{aligned}{}^{\omega}A_{(i,k),(i,k)} &= \int_\sigma \left[\mathbf{T}_{\widetilde{\mathbf{m}}_i(s)}\left(\mathrm{K}_\omega(\widetilde{\mathbf{v}}_{i,k},\widetilde{\mathbf{m}}_i(s))-\mathrm{K}_0(\widetilde{\mathbf{v}}_{i,k},\widetilde{\mathbf{m}}_i(s))\right)\right]^\top \widetilde{J}_i(s)\widetilde{L}_k(s)\,\mathrm{d}s(s)\\ &\quad+\underbrace{\int_\sigma \left[\mathbf{T}_{\widetilde{\mathbf{m}}_i(s)}\left(\mathrm{K}_0(\widetilde{\mathbf{v}}_{i,k},\widetilde{\mathbf{m}}_i(s))\right)\right]^\top \widetilde{J}_i(s)\widetilde{L}_k(s)\,\mathrm{d}s(s)}_{{}^{0}A_{(i,k),(i,k)}}\\ &=: \mathrm{int}_{i,k}^{\mathrm{smooth}}+\mathrm{int}_{i,k}^{\mathrm{singular}}.\end{aligned}$$

The integrand of $\mathrm{int}_{i,k}^{\mathrm{smooth}}$ is smooth and converges to the $2\times 2$ zero matrix, say $Z_2$. Therefore, we directly set $\mathrm{int}_{i,k}^{\mathrm{smooth}}$ equal to $Z_2$. Next, we consider $\mathrm{int}_{i,k}^{\mathrm{singular}}$. We use the fact that for $\phi=1$, we have $\mathrm{D}_0\phi(x)=-\frac{1}{2}I_2$ for all $x\in\partial D$. Hence, we approximately have

$$\sum_{i=1}^{n/2}\sum_{k=1}^{3}{}^{0}A_{(i,k),(j,\ell)}\approx-\frac{1}{2}I_2\qquad\forall\,(j,\ell),\tag{10.10}$$

and therefore we can find the diagonal matrix entry ${}^{0}A_{(i,k),(i,k)}=\mathrm{int}_{i,k}^{\mathrm{singular}}$ by enforcing (10.10) to be exact. Hence, we never have to integrate over a singularity, but we need to additionally compute the $2\times 2$ matrices ${}^{0}A_{(i,k),(j,\ell)}$ for all $(i,k)\neq(j,\ell)$. For a given $\mathbf{f}$ and $\omega$, the linear system (10.9) is solved directly for $\mathbf{h}$. Likewise, we can discretize $\mathbf{u}(\mathbf{x})=\mathrm{DL}_\omega\mathbf{h}(\mathbf{x})$ to compute the solution at any $\mathbf{x}\in\mathbb{R}^2\backslash\overline{D}$. Precisely, we have

$$\mathbf{u}(\mathbf{x})=\mathrm{DL}_\omega\mathbf{h}(\mathbf{x})\approx\sum_{i=1}^{n/2}\sum_{k=1}^{3}{}^{\omega}\widetilde{A}_{(i,k),x}\mathbf{h}(\widetilde{\mathbf{v}}_{i,k})=:\mathbf{u}_n(\mathbf{x})$$

with

$$\widetilde{A}_{(i,k),x}=\int_\sigma\left[\mathbf{T}_{\widetilde{\mathbf{m}}_i(s)}\left(\mathrm{K}_\omega(x,\widetilde{\mathbf{m}}_i(s))\right)\right]^\top\widetilde{J}_i(s)\widetilde{L}_k(s)\,\mathrm{d}s(s)\in\mathbb{C}^{2\times 2}.$$

**Table 10.1** Numerical results to test the discretization of the double-layer operator for $\omega = 1$ and $\omega = \mathrm{i}$

| $n$ | $e_n^{(1)}$ | $\mathrm{EOC}^{(1)}$ | $e_n^{(\mathrm{i})}$ | $\mathrm{EOC}^{(\mathrm{i})}$ |
|---|---|---|---|---|
| 10 | $1.310\,39_{-4}$ | | $7.746\,88_{-6}$ | |
| 20 | $2.932\,42_{-5}$ | 2.16 | $2.067\,50_{-6}$ | 1.91 |
| 40 | $2.007\,61_{-6}$ | 3.87 | $5.850\,69_{-8}$ | 5.14 |
| 80 | $2.606\,33_{-7}$ | 2.95 | $7.619\,38_{-9}$ | 2.94 |

*Example 10.1* Consider the solution of the Navier equation $\Delta^*\mathbf{u} + \omega^2\mathbf{u} = 0$ in $\mathbb{R}^2\backslash\overline{D}$ with $\mathbf{u} = \mathbf{f}$ on $\partial\Omega$, where the boundary of the domain $\Omega$ is given parametrically by $\mathbf{p}(\theta) = (2\cos(\theta), \sin(\theta))$ (an ellipse). The Lamé parameters are chosen to be $\lambda = 1$ and $\mu = 1$. The frequency $\omega$ is given by 1 and i and we used $\alpha = (1 - \sqrt{3/5})/2$. The first column of the fundamental solution with $\mathbf{y} = (0, 0)^\top$ satisfies $\Delta^*\mathbf{u} + \omega^2\mathbf{u} = 0$ and is used as a reference solution. The boundary function $\mathbf{f}$ is chosen to be the first column of the fundamental solution restricted to the given boundary. We compute the solution at $\mathbf{x} = (3, 3)^\top$ using the double-layer ansatz $\mathbf{u}(\mathbf{x}) = \mathrm{DL}_\omega\mathbf{h}(\mathbf{x})$, and test therefore the operator $\frac{1}{2}\mathrm{I} + \mathrm{D}_\omega$ since we need to compute

$$\frac{1}{2}\mathbf{h} + \mathrm{D}_\omega\mathbf{h} = \mathbf{f}$$

in order to obtain $\mathbf{h}$. In Table 10.1, we list the absolute error $e_n^{(\omega)} = \|\mathbf{u} - \mathbf{u}_n\|$ for various choices of $n$ as well as the estimated order of convergence $\mathrm{EOC}^{(\omega)} = \log\left(e_n^{(\omega)}/e_{2n}^{(\omega)}\right)/\log(2)$. As we can see in Table 10.1, we obtain a convergence order of at least two.

In a similar fashion, we can solve the problem $\Delta^*\mathbf{u} + \omega^2\mathbf{u} = 0$ in $\mathbb{R}^2\backslash\overline{D}$ with the boundary conditions $\mathbf{u} = \mathbf{f}$, where $\mathbf{f}$ is a given function defined on the boundary. The frequency $\omega$ is given as well. Using the single-layer ansatz $\mathbf{u} = \mathrm{SL}_\omega\mathbf{g}$ in $\mathbb{R}^2\backslash\overline{D}$ yields the boundary integral equation of the first kind

$$\mathrm{S}_\omega\mathbf{g} = \mathbf{f}. \tag{10.11}$$

Using the same strategy as explained before yields the linear system of size $3n \times 3n$

$$\sum_{i=1}^{n/2}\sum_{k=1}^{3} {}^{\omega}B_{(i,k),(j,\ell)}\mathbf{g}(\widetilde{\mathbf{v}}_{i,k}) = \mathbf{f}(\widetilde{\mathbf{v}}_{j,\ell}) \tag{10.12}$$

with

$${}^{\omega}B_{(i,k),(j,\ell)} = \int_\sigma \mathrm{K}_\omega(\widetilde{\mathbf{v}}_{j,\ell}, \widetilde{\mathbf{m}}_i(s))\widetilde{J}_i(s)\widetilde{L}_k(s)\,\mathrm{d}s(s) \in \mathbb{C}^{2\times 2}$$

since the $(i,k)$, $(j,\ell)$-entry is a $2\times 2$ matrix. All four elements of the $2\times 2$ matrix are

$$
{}^{\omega}B^{(1,1)}_{(i,k),(j,\ell)} = \int_\sigma u^{(1,1)}_{i,j,\ell}(s)\widetilde{J}_i(s)\widetilde{L}_k(s)\,\mathrm{d}s(s)\,,\quad {}^{\omega}B^{(1,2)}_{(i,k),(j,\ell)} = \int_\sigma u^{(1,2)}_{i,j,\ell}(s)\widetilde{J}_i(s)\widetilde{L}_k(s)\,\mathrm{d}s(s)
$$

$$
{}^{\omega}B^{(2,1)}_{(i,k),(j,\ell)} = \int_\sigma u^{(2,1)}_{i,j,\ell}(s)\widetilde{J}_i(s)\widetilde{L}_k(s)\,\mathrm{d}s(s)\,,\quad {}^{\omega}B^{(2,2)}_{(i,k),(j,\ell)} = \int_\sigma u^{(2,2)}_{i,j,\ell}(s)\widetilde{J}_i(s)\widetilde{L}_k(s)\,\mathrm{d}s(s)
$$

with

$$
\begin{aligned}
u^{(1,1)}_{i,j,\ell}(s) &= \frac{\mathrm{i}}{4\mu}H_0^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)\\
&\quad+\frac{\mathrm{i}}{4\omega^2}\,\frac{k_pH_1^{(1)}(k_p\,\|\mathbf{d}_{i,j,\ell}(s)\|)-k_sH_1^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)}{\|\mathbf{d}_{i,j,\ell}(s)\|}\\
&\quad+\frac{(\,d^{(1)}_{i,j,\ell}(s))^2}{\|\mathbf{d}_{i,j,\ell}(s)\|^2}\left(\frac{\mathrm{i}}{4\omega^2}\left[k_p^2H_0^{(1)}(k_p\,\|\mathbf{d}_{i,j,\ell}(s)\|)-k_s^2H_0^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right]\right.\\
&\quad\left.+\frac{\mathrm{i}}{2\omega^2\,\|\mathbf{d}_{i,j,\ell}(s)\|}\left[k_sH_1^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)-k_pH_1^{(1)}(k_p\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
u^{(2,1)}_{i,j,\ell}(s) &= \frac{d^{(1)}_{i,j,\ell}(s)d^{(2)}_{i,j,\ell}(s)}{\|\mathbf{d}_{i,j,\ell}(s)\|^2}\left(\frac{\mathrm{i}}{4\omega^2}\left[k_p^2H_0^{(1)}(k_p\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right.\right.\\
&\quad\left.-\,k_s^2H_0^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right]+\frac{\mathrm{i}}{2\omega^2\,\|\mathbf{d}_{i,j,\ell}(s)\|}\left[k_sH_1^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right.\\
&\quad\left.\left.-\,k_pH_1^{(1)}(k_p\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right]\right)\\
u^{(1,2)}_{i,j,\ell}(s) &= \frac{d^{(1)}_{i,j,\ell}(s)d^{(2)}_{i,j,\ell}(s)}{\|\mathbf{d}_{i,j,\ell}(s)\|^2}\left(\frac{\mathrm{i}}{4\omega^2}\left[k_p^2H_0^{(1)}(k_p\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right.\right.\\
&\quad\left.-\,k_s^2H_0^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right]+\frac{\mathrm{i}}{2\omega^2\,\|\mathbf{d}_{i,j,\ell}(s)\|}\left[k_sH_1^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right.\\
&\quad\left.\left.-\,k_pH_1^{(1)}(k_p\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
u^{(2,2)}_{i,j,\ell}(s) &= \frac{\mathrm{i}}{4\mu}H_0^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)\\
&\quad+\frac{\mathrm{i}}{4\omega^2}\,\frac{k_pH_1^{(1)}(k_p\,\|\mathbf{d}_{i,j,\ell}(s)\|)-k_sH_1^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)}{\|\mathbf{d}_{i,j,\ell}(s)\|}
\end{aligned}
$$

$$+\frac{(d^{(2)}_{i,j,\ell}(s))^2}{\|\mathbf{d}_{i,j,\ell}(s)\|^2}\left(\frac{\mathrm{i}}{4\omega^2}\left[k_p^2 H_0^{(1)}(k_p\,\|\mathbf{d}_{i,j,\ell}(s)\|)-k_s^2 H_0^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right]\right.$$
$$\left.+\frac{\mathrm{i}}{2\omega^2\,\|\mathbf{d}_{i,j,\ell}(s)\|}\left[k_s H_1^{(1)}(k_s\,\|\mathbf{d}_{i,j,\ell}(s)\|)-k_p H_1^{(1)}(k_p\,\|\mathbf{d}_{i,j,\ell}(s)\|)\right]\right)$$

For a given function $\mathbf{f}$ and frequency $\omega$, the linear system (10.12) is solved directly for $\mathbf{g}$. We discretize $\mathbf{u}(\mathbf{x}) = \mathrm{SL}_\omega\mathbf{g}(\mathbf{x})$ to compute the solution at any $\mathbf{x}\in\mathbb{R}^2\backslash\overline{D}$. Precisely, we have

$$\mathbf{u}(\mathbf{x}) = \mathrm{SL}_\omega\mathbf{g}(\mathbf{x}) \approx \sum_{i=1}^{n/2}\sum_{k=1}^{3} {}^{\omega}\widetilde{B}_{(i,k),x}\,\mathbf{g}(\widetilde{\mathbf{v}}_{i,k}) =: \mathbf{u}_n(\mathbf{x})$$

with

$$\widetilde{B}_{(i,k),x} = \int_\sigma \mathrm{K}_\omega(x,\widetilde{\mathbf{m}}_i(s))\,\widetilde{J}_i(s)\widetilde{L}_k(s)\,\mathrm{d}s(s) \in \mathbb{C}^{2\times 2}\,.$$

*Example 10.2* Consider again the solution of the Navier equation $\Delta^*\mathbf{u}+\omega^2\mathbf{u} = 0$ in $\mathbb{R}^2\backslash\overline{D}$ with $\mathbf{u} = \mathbf{f}$ on $\partial\Omega$, where the boundary of the domain $\Omega$ is given parametrically by $\mathbf{p}(\theta) = (2\cos(\theta),\sin(\theta))$ (an ellipse). We use the same parameters as before (refer to Example 10.1). We again compute the solution at $\mathbf{x} = (3,3)^\top$, but with a single-layer ansatz $\mathbf{u}(\mathbf{x}) = \mathrm{SL}_\omega\mathbf{g}(\mathbf{x})$ and test therefore the operator $\mathrm{S}_\omega$ since we need to compute

$$\mathrm{S}_\omega\mathbf{g} = \mathbf{f}$$

to obtain $\mathbf{g}$. In Table 10.2, we list the absolute error $e_n^{(\omega)}$ for various choices of $n$ including the estimated order of convergence $\mathrm{EOC}^{(\omega)}$. As we can see in Table 10.2, we obtain a convergence order of at least two.

**Table 10.2** Numerical results to test the discretization of the single-layer operator for $\omega = 1$ and $\omega = \mathrm{i}$

| $n$ | $e_n^{(1)}$ | $\mathrm{EOC}^{(1)}$ | $e_n^{(\mathrm{i})}$ | $\mathrm{EOC}^{(\mathrm{i})}$ |
|---|---|---|---|---|
| 10 | $1.475\,50_{-4}$ | | $1.310\,82_{-5}$ | |
| 20 | $1.354\,72_{-5}$ | 3.45 | $1.396\,89_{-6}$ | 3.23 |
| 40 | $2.156\,46_{-7}$ | 5.97 | $2.035\,61_{-8}$ | 6.10 |
| 80 | $9.554\,28_{-8}$ | 1.17 | $9.413\,91_{-9}$ | 1.11 |

## 10.4 Solving the Nonlinear Eigenvalue Problem

Beyn's algorithm [Be12] is used to solve the nonlinear eigenvalue problem of the form

$$\mathrm{M}(\omega)\mathbf{u} = \mathbf{0}\,, \mathbf{u} \neq 0$$

with $\mathrm{M}(\omega) \in \mathbb{C}^{m\times m}$. Therefore, the user specifies a smooth contour $\gamma$ in the complex plane and integrates over the resolvent. We will use a circle with radius $R$ centered at $c$ as the contour $\gamma$ given parametrically by $\phi(t) = c + R\mathrm{e}^{t\,\mathrm{i}}$ and $\phi'(t) = R\mathrm{i}\mathrm{e}^{t\,\mathrm{i}}$. With Keldysh's theorem, one can reduce the nonlinear eigenvalue problem to a linear eigenvalue problem of size $n(\gamma)$ which is much smaller than $m$. To be more specific, one has to compute the two integrals

$$\mathrm{A}_0 = \frac{1}{2\pi\,\mathrm{i}}\int_\gamma \mathrm{M}^{-1}(\omega)\hat{\mathrm{V}}\,\mathrm{d}s(\omega)\,, \qquad \mathrm{A}_1 = \frac{1}{2\pi\,\mathrm{i}}\int_\gamma \omega\mathrm{M}^{-1}(\omega)\hat{\mathrm{V}}\,\mathrm{d}s(\omega)\,,$$

where $\hat{\mathrm{V}} \in \mathbb{C}^{m\times\ell}$ with $m \gg \ell \geq n(\gamma)$ is a random matrix. The parameter $\ell$ has to be chosen such that it is greater than the number of possible eigenvalues $n(\gamma)$ (including multiplicities), but as small as possible to reduce computational work. Of course, the two integrals have to be computed numerically. We will use the trapezoidal rule yielding

$$\mathrm{A}_{0,N} = \frac{1}{\mathrm{i}\,N}\sum_{j=0}^{N-1} \mathrm{M}^{-1}(\phi(t_j))\hat{\mathrm{V}}\phi'(t_j)\,, \quad \mathrm{A}_{1,N} = \frac{1}{\mathrm{i}\,N}\sum_{j=0}^{N-1} \phi(t_j)\mathrm{M}^{-1}(\phi(t_j))\hat{\mathrm{V}}\phi'(t_j)\,.$$

The parameter $N$ is specified by the user, and with this we define the equidistant nodes $t_j = 2\pi j/N$, $j = 0,\ldots,N$. The parameter $N$ can be chosen small since the trapezoidal rule converges exponentially. Next, a (reduced) singular value decomposition of $\mathrm{A}_{0,N} = \mathrm{V}\Sigma\mathrm{W}^{\mathrm{H}}$ is computed, where $\mathrm{V} \in \mathbb{C}^{m\times\ell}$, $\Sigma \in \mathbb{C}^{\ell\times\ell}$, and $\mathrm{W} \in \mathbb{C}^{\ell\times\ell}$. Then, a rank test on the diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_\ell)$ is performed which indicates how many eigenvalues including multiplicities are contained within the chosen contour $\gamma$. We will use $\varepsilon = 10^{-2}$ and compute $n(\gamma)$ such that $\sigma_1 \geq \ldots \geq \sigma_{n(\gamma)} > \varepsilon > \sigma_{n(\gamma)+1} \geq \ldots \geq \sigma_\ell$ is satisfied. With this, we construct the three matrices $\mathrm{V}_0 = (\mathrm{V}_{ij})_{1\leq i\leq m, 1\leq j\leq n(\gamma)}$, $\Sigma_0 = (\Sigma_{ij})_{1\leq i\leq n(\gamma), 1\leq j\leq n(\gamma)}$, and $\mathrm{W}_0 = (\mathrm{W}_{ij})_{1\leq i\leq \ell, 1\leq j\leq n(\gamma))}$. Finally, we compute $n(\gamma)$ eigenvalues, say $\omega_i$, and eigenvectors $\mathbf{s}_i$ of the new matrix $\mathrm{B} = \mathrm{V}_0^{\mathrm{H}}\mathrm{A}_{1,N}\mathrm{W}_0\Sigma_0^{-1} \in \mathbb{C}^{n(\gamma)\times n(\gamma)}$. The $i$-th nonlinear eigenvector $\mathbf{u}_i$ is given by $\mathrm{V}_0\mathbf{s}_i$.

## 10.5 Numerical Results

In this section, we present numerical results for the computation of elastic interior transmission eigenvalues for a variety of two-dimensional domains. Let $\theta \in [0, 2\pi]$. The first domain $D_1$ under consideration is a disk with radius $r_1 = 1/2$ having the parametrization $\mathbf{p}_1(\theta) := (r_1 \cos(\theta), r_1 \sin(\theta))^\top$. The second domain $D_2$ is an ellipse with semi-axis $a_2 = 1$ and $b_2 = 1/2$. Its parametrization is given by $\mathbf{p}_2(\theta) := (a_2 \cos(\theta), b_2 \sin(\theta))^\top$. The third parametrization is given by $\mathbf{p}_3(\theta) := (3\cos(\theta)/4 + 3\cos(2t)/10, \sin(\theta))^\top$ and represents the "deformed ellipse" (kite) domain $D_3$. The unit square $D_4$ is the fourth domain under consideration.

For comparison, we will use the parameters $\varrho_1 = 1$ and $\varrho_2 = 4$ and the Lamé parameters $\mu = 1/16$ and $\lambda = 1/4$ which have been used in a variety of papers before. Furthermore, we use $N = 24$, $\ell = 20$, $\varepsilon = 10^{-2}$, and $R = 1/4$ within the Beyn algorithm. The parameter $c$ and the number of faces $n_f$ depend on the considered domain and are listed separately. The parameter $\alpha$ is chosen to be $(1 - \sqrt{3/5})/2$ for all the following numerical results.

At first, we consider $D_1$ and compute the first seven real elastic interior transmission eigenvalues using $n_f = 40$ and $c = 1.5$ for $\omega_1$, $\omega_2$, and $\omega_3$ and $n_f = 40$ and $c = 2.1$ for $\omega_4$, $\omega_5$, $\omega_6$, and $\omega_7$ with the boundary element method (BEM). We compare our results with the method of fundamental solutions (MFS) [KlPi20] since those results are accurate up to ten digits accuracy for $D_1$. Additionally, we compare our results with different finite element methods (FEMs) [YaEtAl20, XiJi18, JiLiSu18]. Note that the second, fourth, and sixth eigenvalues have multiplicity two. In [XiJi18], the first two eigenvalues are listed, and in [YaEtAl20, JiLiSu18], the first six eigenvalues are computed. In Table 10.3, we list the first seven eigenvalues and highlight the correct number of digits in bold. The eigenvalues obtained with the MFS are used for comparison. All reported digits are correct and therefore not highlighted in bold.

As we can see, our numerical results are accurate up to five digits accuracy using only $n_f = 40$ faces. The first eigenvalue is accurate up to six digits. The numerical results for the FEM methods are only accurate up to two to three digits with the exception of the first eigenvalue which is accurate up to four digits. The used mesh

**Table 10.3** Numerical results for the first seven real elastic interior transmission eigenvalues for a disk with radius 1/2

| ITE | BEM | FEM [YaEtAl20] | FEM [XiJi18] | FEM [JiLiSu18] | MFS [KlPi20] |
|---|---|---|---|---|---|
| $\omega_1$ | **1.451304** | **1.45**2482 | **1.451**948 | **1.455**078 | 1.451304028 |
| $\omega_2$ | **1.704645** | **1.706**023 | **1.705**370 | **1.709**214 | 1.704638247 |
| $\omega_3$ | **1.704645** | **1.706**023 | | **1.709**214 | |
| $\omega_4$ | **1.984551** | **1.986**143 | | **1.989**630 | 1.984530256 |
| $\omega_5$ | **1.984552** | **1.986**146 | | **1.989**630 | |
| $\omega_6$ | **2.269152** | **2.2**70963 | | **2.2**74992 | 2.269112085 |
| $\omega_7$ | **2.269152** | | | | |

**Table 10.4** Numerical results for the first four real elastic interior transmission eigenvalues for an ellipse with semi-axis 1 and 1/2

| ITE | BEM | MFS [KlPi20] |
|---|---|---|
| $\omega_1$ | **1.296**681 | 1.296728137 |
| $\omega_2$ | **1.302**814 | 1.302785814 |
| $\omega_3$ | **1.540**775 | 1.540896035 |
| $\omega_4$ | **1.565**173 | 1.565151107 |

**Table 10.5** Numerical results for the first four real elastic interior transmission eigenvalues for the kite domain

| ITE | BEM | MFS [KlPi20] |
|---|---|---|
| $\omega_1$ | **0.947**495 | 0.947 |
| $\omega_2$ | **1.047**398 | 1.047 |
| $\omega_3$ | **1.111**190 | 1.111 |
| $\omega_4$ | **1.235**261 | 1.235 |

size in [XiJi18] is $h = 1/160$, in [JiLiSu18] is $h = 1/80$, and in [YaEtAl20] is $h \approx 0.03125$. Note that in the preprint [XiJiGe18], $h = 0.0125$ was used and yields **1.45**6 for the first eigenvalue. In sum, our numerical results are much more accurate than the ones given by FEM. However, the best results are given by the MFS.

Next, we consider the ellipse $D_2$. We use $n_f = 40$ and $c = 1.4$ and compare our numerical results given in Table 10.4 with the MFS for the first four real interior transmission eigenvalues. The numerical results of the MFS are accurate with ten digits and serve again as reference values. They are not highlighted in bold. Unfortunately, no numerical results are available for the FEM method.

As we can see, we are able to obtain four digits accuracy. The fourth eigenvalue is accurate up to five digits accuracy. All eigenvalues are simple. Hence, the BEM method is a good alternative for the MFS and offers good flexibility in terms of using general domains. This is shown with the next domain $D_3$.

The numerical results for the first four elastic interior transmission eigenvalues for the kite are given in Table 10.5 using $n_f = 40$ faces and $c = 0.9$ for $\omega_1$ and $n_f = 40$ and $c = 1.1$ for $\omega_2$, $\omega_3$, and $\omega_4$ along with the numerical results obtained with the MFS. The eigenvalues obtained with the MFS are correct to four digits accuracy and not highlighted in bold.

We obtain at least four digits accuracy with the BEM for $\omega_1$, $\omega_2$, $\omega_3$, and $\omega_4$. For $\omega_2$, we obtain five digits accuracy. Hence, the results are equal to or better than the ones of the MFS. Therefore, the BEM method offers the flexibility to use it for more general domains with a smooth boundary. Unfortunately, no numerical results are reported with the FEM for such domains.

Of course, the FEM is much better suited for polygonal domains such as the unit square. We finally compare our method with the FEM (the accuracy is not known, but at least five digits) and the MFS (five digits accuracy). We use $n_f = 46$ and $c = 1.5$ for the first eigenvalue and $n_f = 46$ and $c = 1.8$ for the other eigenvalues to obtain the numerical results that are given in Table 10.6.

Our numerical results are better than the ones given in [XiJi18] ($h = 0.00625$) and [XiJiGe18] ($h = 0.0125$). Moreover, the results are comparable with the MFS. However, the numerical results reported in [YaEtAl20] ($h \approx 0.03125$)

**Table 10.6** Numerical results for the first five real elastic interior transmission eigenvalues for the unit square

| ITE | BEM | FEM [YaEtAl20] | FEM [JiLiSu18] | FEM [XiJiGe18] | FEM [XiJi18] | MFS [KlPi20] |
|---|---|---|---|---|---|---|
| $\omega_1$ | **1.3938**92 | 1.393877 | 1.393874 | 1.393879 | **1.39**4419 | **1.3938** |
| $\omega_2$ | **1.6182**64 | 1.618299 | 1.618296 | | **1.61**9008 | **1.6182** |
| $\omega_3$ | **1.6183**89 | 1.618299 | 1.618296 | | | |
| $\omega_4$ | **1.8020**89 | 1.802042 | 1.802032 | | | **1.8020** |
| $\omega_5$ | **1.9361**87 | 1.936138 | 1.936134 | | | **1.936**2 |

**Table 10.7** Numerical results for one complex-valued elastic interior transmission eigenvalues for the circle with radius $1/2$ and the unit square

| Domain | BEM |
|---|---|
| Circle | $1.987\,189 + 0.283\,146\mathrm{i}$ |
| Unit square | $1.865\,629 + 0.291\,766\mathrm{i}$ |

and [JiLiSu18] ($h \approx 0.025$) are better as expected. The same is true for FEM [YaHaBi20] using $m = 26$.

Finally, note that we can easily compute complex-valued elastic interior transmission eigenvalues by selecting a corresponding contour in the complex plane, although the existence of them is still an open question. Using $c = 2 + \mathrm{i}/2$ and $n_f = 40$ for $D_1$ and $n_f = 46$ for $D_4$ yields the results reported in Table 10.7.

## 10.6 Summary and Outlook

We presented an algorithm to compute interior elastic transmission eigenvalues in two dimensions with the boundary element collocation method in combination with a nonlinear eigenvalue solver. We are able to obtain good results for a circle and an ellipse which outperforms various finite element methods. However, the method of fundamental solutions beats the boundary element method in accuracy. The situation is different for polygonal domains such as a square. The best method in accuracy is the finite element method. However, for various domains with a smooth boundary, the boundary element method is the one for which the best accuracy can be obtained.

The Python program is available at

https://github.com/kleefeld80/elastic-ite-bem

and has been developed and tested under Windows 10 with Python version 3.8. All numerical results reported within this chapter have been obtained with Python version 3.9.4 under Windows 10 and can be reproduced using the `runall.py` script.

No numerical results are reported for the three-dimensional case. Hence, the next step would be to use the presented algorithm to numerically calculate interior elastic transmission eigenvalues in three dimensions with the boundary element method in a similar fashion as presented in [Kl13] for interior acoustic transmission eigenvalues.

## References


[Be12] Beyn, W.-J.: An integral method for solving nonlinear eigenvalue problems. Linear Algebra Appl. **436**, 3839–3863 (2012)

[BeCaGu13] Bellis, C., Cakoni, F., Guzina, B.: Nature of the transmission eigenvalue spectrum for elastic bodies. IMA J. Appl. Math. **78**, 895–923 (2013)

[CaKr17] Cakoni, F., Kress, R.: A boundary integral equation method for the transmission eigenvalue problem, Appl. Anal. **96**, 23–38 (2017)

[ChLiWa20] Chang, W.-C., Lin, W.-W., Wang, J.-N.: Efficient methods of computing interior transmission eigenvalues for the elastic waves. J. Comput. Phys. **407**, 109227 (2020)

[JiLiSu18] Ji, X., Li, P., Sun J.: Computation of transmission eigenvalues for elastic waves (2018). arXiv 1802.03687, 1–16

[JiLiSu20] Ji, X., Li, P., Sun, J.: Computation of interior elastic transmission eigenvalues using a conforming finite element and the secant method, Results Appl. Math. **5**, 100083 (2020)

[Kl13] Kleefeld, A.: A numerical method to compute interior transmission eigenvalues, Inverse Problems **29**, 104012 (2013)

[KlPi20] Kleefeld, A., Pieronek, L.: Elastic transmission eigenvalues and their computation via the method of fundamental solutions. Appl. Anal. **100**, 3445–3462 (2021). https://doi.org/10.1080/00036811.2020.1721473

[Mc00] McLean, W.: Strongly Elliptic Systems and Boundary Integral Equations. Cambridge University Press, Cambridge (2000)

[Pe16] Peters, S.: The inside-outside duality for elastic scattering problems. Appl. Anal. **96**, 48–69 (2016)

[We18] Weger, A.-C.: Numerische Berechnung von elastischen Streuproblemen in 2D. Jül Report **4413**, 1–118 (2018)

[XiJi18] Xi, Y., Ji, X.: A lowest-order mixed finite element method for the elastic transmission eigenvalue problem (2018). arXiv 1812.0851, 1–16

[XiJiGe18] Xi, Y., Ji, X., Geng, H.: A $C^0$IP method of transmission eigenvalues for elastic waves. J. Comput. Phys. **374**, 237–248 (2018)

[XiJiZh21] Xi, Y., Ji, X., Zhang S.: A simple low-degree optimal finite element scheme for the elastic transmission eigenvalue problem (2021). arXiv 2101.10783, 1–17

[YaHaBi20] Yang, Y., Han, J., Bi, H.: $H^2$-conforming methods and two-grid discretizations for the elastic transmission eigenvalue problem. Commun. Comput. Phys. **28**, 1366–1388 (2020)

[YaEtAl20] Yang, Y., Han, J., Bi, H., Li, H., Zhang, Y.: Mixed methods for the elastic transmission eigenvalue problem. Appl. Math. Comput. **374**, 125081 (2020)

[Zi21] Zimmermann, M.: Numerische Berechnung von elastischen Transmissionseigenwerten. Master Thesis (2021)

# Chapter 11
# A Novel Solution of the Multi-Group Neutron Diffusion Equation by the Hankel Transform Formalism


**R. A. S. Klein and J. C. L. Fernandes**

## 11.1 Introduction

The neutron multi-group equation is frequently used in applications for nuclear reactors. The division in energy groups has been used for a long time to develop more detailed solutions, since the separation by their speed or energy not only facilitates obtaining a better approximate model but also describes the diffusive process with more physical properties [Oz01, No21]. In addition, it is also common to use approaches with different types of geometry [Ma17, Ol19, Ma21], which provide some insight in the influences of the specific boundaries on neutronics. Nuclear reactor cores have different types of geometric approximations and one of the most used is one with axial symmetry in cylindrical coordinates. The choice of a specific coordinate system in general depends on the reactor type and the characteristics to be analysed.

In the course of time, several attempts were used to solve the neutron flux problem in reactor cores, among the most classic ones are procedures, which make use of integral transforms. This method has proven to be effective over many years of research and some representative works may be found in references [Du06, Vi08, Fe13]. Hence, in this work, we develop a methodology to solve the neutron diffusion equation analytically by a finite integral transform technique. In this line, recently Fernandes et al. [Fe11] solved the neutron diffusion equation in cylindrical geometry for a model with two energy groups using the Hankel transform in infinite space, and after constraining the solution to a finite domain, the Parseval identity was employed. In a similar solution procedure, the authors of reference [Gl03] solved the neutron transport equation in cylindrical geometry,

R. A. S. Klein · J. C. L. Fernandes (✉)
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: renato.klein@ufrgs.br; julio.lombaldo@ufrgs.br



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_11

while considering isotropic scattering and using the Hankel transform together with the Parseval identity. Thus, due to the promising results of these works and the fact that the approximation $S_2$ of the Boltzmann transport equation reduces to the diffusion equation, in the present work, we focus on the derivation of an analytical formulation for the fast and thermal neutron flux in the diffusion equation and exploring the finite Hankel transform. The derived solutions for different sources in cylindrical geometry are relevant for nuclear fuel element assembly calculations of reactor cores, as for example in pressurized water reactor core simulations.

## 11.2 Problem Formulation

We consider initially a steady-state problem with two energy groups in the neutron diffusion equation for a homogenized reactor core given by

$$-D_1 \triangle_r \phi_1 + \Sigma_{R1}\phi_1 = S_1 + \frac{1}{k_{eff}} \nu\chi_1 \Sigma_{f2}\phi_2 + \Sigma_{12}\phi_2$$

$$-D_2 \triangle_r \phi_2 + \Sigma_{R2}\phi_2 = \frac{1}{k_{eff}} \nu\chi_2 \Sigma_{f1}\phi_1 + S_2 + \Sigma_{21}\phi_1 \ ,$$

where $\phi_g$ is the neutron flux, $D_g$ is the diffusion coefficient for the group $g$, $\triangle_r$ is the Laplacian operator in cylinder coordinates explicitly given by $\triangle_r = \partial_r^2 + \frac{1}{r}\partial_r$, $\Sigma_{Rg}$ is the removal cross section of group $g$, $k_{eff}$ is the effective multiplication factor from nuclear reactor theory, $\nu$ is the average number of neutrons emitted by fission, $\Sigma_{fg}$ is the fission cross section, $\chi_g$ is the integrated spectrum for neutrons of group $g$ and $\Sigma_{gg'}$ is the scattering cross section from $g$ into group $g'$. The term $S_g$ is the source term of group $g$ which represents the term $\frac{1}{k_{eff}} \nu\chi_g \Sigma_{fg}\phi_g$ responsible for neutron multiplication, i.e. a manifestation of a chain reaction. The symmetry and boundary conditions for this problem defined individually for each energy group $g$ are

$$\left.\frac{\partial \phi_g}{\partial r}\right|_{r=0} = 0 \ \text{ and } \ \left.\phi_g\right|_{r=R} = 0. \tag{11.1}$$

In order to apply the finite Hankel transform to the previous equations, where as an idealization $R$ represents the extrapolated distance for the same problem as given in reference [La66]

$$H_0\{f(r)\} = \int_0^R r f(r) J_0(r\xi_i)\, dr \ ,$$

where $\xi_i$ is the i-$th$ root of $J_0(R\xi) = 0$, and the inversion of the finite Hankel transform is given by

$$H_0^{-1}\{f(\xi_i)\} = \frac{2}{R^2}\sum_{i=1}^{\infty} f(\xi_i)\frac{J_0(r\xi_i)}{J_1^2(R\xi_i)} \ .$$

Now, using the property

$$H_0\{-D_g \triangle_r \phi_g\} = -D_g\left(-\xi_i^2\bar{\phi}_g(\xi_i) - R\xi\phi_g(R)J_1'(R\xi)\right)$$

and further applying the extrapolated distance boundary condition $\phi_g(R) = 0$, the finite Hankel transform of this operator term is

$$H_0\{-D_g \triangle_r \phi_g\} = D_g\xi_i^2\bar{\phi}_g(\xi_i) \ .$$

The Hankel transform of the source terms is given by

$$H_0\{S_g\} = \bar{S}_g = \int_0^R r S_g J_0(r\xi_i)\, dr \ .$$

After application of the finite Hankel transform, one obtains a system of equations,

$$\begin{pmatrix} D_1\xi_i + \Sigma_{R1} & -\left(\frac{1}{k_{eff}}\chi_1\nu\Sigma_{f2} + \Sigma_{12}\right) \\ -\left(\frac{1}{k_{eff}}\chi_2\nu\Sigma_{f1} + \Sigma_{21}\right) & D_2\xi_i + \Sigma_{R2} \end{pmatrix} \begin{pmatrix} \bar{\phi}_1 \\ \bar{\phi}_2 \end{pmatrix} = \begin{pmatrix} \bar{S}_1 \\ \bar{S}_2 \end{pmatrix} \ ,$$

which is a matrix equation which represents the multi-group problem and may be cast in compact form

$$M(\xi_i)\bar{\Phi} = \bar{S}(\xi_i) \ .$$

The solution of this equation system is formally given by

$$\bar{\Phi} = M^{-1}(\xi_i)\bar{S}(\xi_i) \ .$$

For convenience, we introduce now the shorthand notations

$$A_1(\xi_i) = D_1\xi_i^2 + \Sigma_{R1} \ , \tag{11.2}$$

$$A_2(\xi_i) = D_2\xi_i^2 + \Sigma_{R2} \ , \tag{11.3}$$

and

$$C = \underbrace{\left(\frac{1}{k_{eff}}\chi_1 \nu\Sigma_{f2} + \Sigma_{12}\right)}_{=p_1}\underbrace{\left(\frac{1}{k_{eff}}\chi_2 \nu\Sigma_{f1} + \Sigma_{21}\right)}_{=p_2} \tag{11.4}$$

so that the determinant of matrix $M$ in compact form is

$$Det(M)(\xi_i) = A_1(\xi_i)A_2(\xi_i) - C \ .$$

With these conventions, one may write the transformed solution as

$$\bar{\phi}_1(\xi_i) = \frac{A_2(\xi_i)}{Det(M)(\xi_i)}\bar{S}_1 + \frac{p_1}{Det(M)(\xi_i)}\bar{S}_2 \ ,$$
$$\bar{\phi}_2(\xi_i) = \frac{A_1(\xi_i)}{Det(M)(\xi_i)}\bar{S}_2 + \frac{p_2}{Det(M)(\xi_i)}\bar{S}_1 \ .$$

These expressions are not the final solution yet, since they depend strongly on the choice for the sources terms. Nevertheless, using the definition of the inversion of the finite Hankel transform, we obtain for each energy group

$$\phi_1(r) = H_0^{-1}\{\bar{\phi}_1\} = \frac{2}{R^2}\sum_{i=1}^{\infty}\frac{A_2(\xi_i)\bar{S}_1}{Det(M)(\xi_i)}\frac{J_0(r\xi_i)}{J_1^2(R\xi_i)} + \frac{2}{R^2}p_1\sum_{i=1}^{\infty}\frac{\bar{S}_2}{Det(M)(\xi_i)}\frac{J_0(r\xi_i)}{J_1^2(R\xi_i)} \ ,$$

$$\phi_2(r) = H_0^{-1}\{\bar{\phi}_2\} = \frac{2}{R^2}\sum_{i=1}^{\infty}\frac{A_1(\xi_i)\bar{S}_2}{Det(M)(\xi_i)}\frac{J_0(r\xi_i)}{J_1^2(R\xi_i)} + \frac{2}{R^2}p_2\sum_{i=1}^{\infty}\frac{\bar{S}_1}{Det(M)(\xi_i)}\frac{J_0(r\xi_i)}{J_1^2(R\xi_i)} \ .$$

This couple of equations may be used to elaborate the solution of the multi-group neutron problem for a steady-state diffusion problem, here shown for two energy groups.

## 11.3 Solution by the Infinite Hankel Transform

The initial problem in cylindrical coordinates is well defined inside the spatial domain with $r \in [0, R]$ and it was natural to choose the finite Hankel transform to solve the problem successfully. We now point out and discuss the consequences if one solves the same problem but using the infinite Hankel transform, which by definition is

$$H_0\{f(r); r \to \xi\} = \int_0^{\infty} rf(r)J_0(r\xi)\, dr,$$

and the inversion has the form

$$H_0^{-1}\{f(\xi);\xi\to r\}=\int_0^\infty \xi f(\xi)J_0(r\xi)\,d\xi\ . \tag{11.5}$$

This version of the transform is usually applied for half-open domains. As a matter of fact, the flux profile for both groups is well known as $r$ goes to infinity, and we can define an extrapolated distance for both fluxes at $R$ where these vanish and moreover redefine (11.1) for this kind of problems.

Since the property of the Hankel transform for the operator $\triangle_r$ is the same as for the finite Hankel transform, therefore,

$$H_0\{-D_g\triangle_r\phi_g\}=D_g\xi^2\bar{\phi}_g(\xi),$$

and the equation system after applying the Hankel transform can be written in matrix form as

$$\begin{pmatrix} D_1\xi+\Sigma_{R1} & -\left(\frac{1}{k_{eff}}\chi_1\nu\Sigma_{f2}+\Sigma_{12}\right)\\ -\left(\frac{1}{k_{eff}}\chi_2\nu\Sigma_{f1}+\Sigma_{21}\right) & D_2\xi+\Sigma_{R2}\end{pmatrix}\begin{pmatrix}\bar{\phi}_1\\ \bar{\phi}_2\end{pmatrix}=\begin{pmatrix}\bar{S}_1\\ \bar{S}_2\end{pmatrix}\ .$$

Using now the definitions (11.2), (11.3) and (11.4), the matrix equation reads

$$\begin{pmatrix} A_1(\xi) & -p_1\\ -p_2 & A_2(\xi)\end{pmatrix}\begin{pmatrix}\bar{\phi}_1\\ \bar{\phi}_2\end{pmatrix}=\begin{pmatrix}\bar{S}_1\\ \bar{S}_2\end{pmatrix}$$

with formal solution given by

$$\bar{\Phi}(\xi)=\frac{1}{Det(M)(\xi)}\begin{pmatrix}A_2(\xi)\bar{S}_1+p_1\bar{S}_2\\ A_1(\xi)\bar{S}_2+p_2\bar{S}_1\end{pmatrix}\ .$$

If we focus now our attention on the inversion problem for $\phi_1$, we need to investigate the first term of the last equation using Eq. (11.5).

$$\bar{\phi}_1=\frac{1}{Det(M)(\xi)}\left(A_2(\xi)\bar{S}_1(\xi)+p_1\bar{S}_2(\xi)\right).$$

Using the inversion theorem, one gets

$$\phi_1(x)=\int_0^\infty \xi\left(\frac{A_2(\xi)J_0(r\xi)}{Det(M)(\xi)}\right)\bar{S}_1(\xi)\,d\xi+p_1\int_0^\infty \xi\left(\frac{\bar{S}_2(\xi)}{Det(M)(\xi)}\right)J_0(r\xi)\,d\xi\ .$$

The first term of the solution is clearly more complicated to solve, so that to this end we split the fast flux $\phi_1(x)=\phi_1^{(1)}(x)+\phi_1^{(2)}(x)$ and consider the following theorems.

**Theorem 11.1 (Hankel Inversion)** *If $\sqrt{r'}f(r')$ is piecewise continuous and absolutely integrable along the real axis, then if $\gamma \geq -\frac{1}{2}$, $f_\gamma(\xi) = H_\gamma\{f(r')\}$, then*

$$\int_0^\infty \xi f_\gamma(\xi) J_\gamma(r'\xi)\, d\xi = \frac{1}{2}(f(r'+) + f(r'-)).$$

**Theorem 11.2 (Parseval Relation)** *If the functions $f(r')$ and $g(r')$ satisfy the conditions of Theorem 11.1 and if $f_\gamma(\xi)$ and $g_\gamma(\xi)$ are the Hankel transforms of order $\gamma \geq -\frac{1}{2}$, respectively, then*

$$\int_0^\infty r' f(r') g(r') dr' = \int_0^\infty \xi\, \bar{f}_\gamma(\xi) \bar{g}_\gamma(\xi)\, d\xi.$$

These two theorems are essential so that this alternative procedure may be applied. Upon substituting $\bar{f}_0(\xi)$ and $\bar{g}_0(\xi)$ with $\frac{A_2(\xi)J_0(r\xi)}{Det(M)(\xi)}$ and $\bar{S}_1$, respectively, and using Theorem 11.2, one obtains

$$\phi_1^{(1)}(x) = \int_0^\infty \xi \left(\frac{A_2(\xi)J_0(r\xi)}{Det(M)(\xi)}\right) \bar{S}_1\, d\xi = \int_0^\infty r' H_0^{-1}\left\{\frac{A_2(\xi)J_0(r\xi)}{Det(M)(\xi)}\right\} S_1(r')\, dr'\,.$$

In other words, we need to calculate $f(r')$.

$$f(r') = H_0^{-1}\left\{\frac{A_2(\xi)J_0(r\xi)}{Det(M)(\xi)}\right\} = \int_0^\infty \xi \frac{A_2(\xi)J_0(r\xi)}{Det(M)(\xi)} J_0(r'\xi)\, d\xi\,. \tag{11.6}$$

Recalling that $Det(M)(\xi) = A_1(\xi)A_2(\xi) - C$ and that the physically meaningful nuclear parameter set satisfies the following condition, $0 < \frac{C}{A_1(\xi)A_2(\xi)} < 1$ for all $\xi \in [0, \infty)$, we can expand the term $\frac{A_2}{Det(M)(\xi)}$

$$\begin{aligned}\frac{A_2(\xi)}{A_1(\xi)A_2(\xi) - C} &= \frac{1}{A_1(\xi)} \frac{1}{1 - \frac{C}{A_1(\xi)A_2(\xi)}} \\ &= \frac{1}{A_1(\xi)}\left(1 + \left(\frac{C}{A_1(\xi)A_2(\xi)}\right) + \left(\frac{C}{A_1(\xi)A_2(\xi)}\right)^2 + \ldots\right).\end{aligned}$$

Indeed, for all nuclear parameter sets known in the literature, they comply with $\frac{C}{A_1(\xi)A_2(\xi)} << 1$ for $\xi \geq 0$. After evaluating different kinds of parameter sets, one may obtain an estimate for the order of magnitude $\mathcal{O}\left(\frac{C}{A_1A_2}\right) = 10^{-3}$ and use this as a maximum for all values of $\xi$, so that one may safely take only the first term of the expansion.

$$\frac{A_2(\xi)}{A_1(\xi)A_2(\xi) - C} \approx \frac{1}{A_1(\xi)}. \tag{11.7}$$

Consequently, Eq. (11.6) simplifies to

$$f(r') = \int_0^\infty \xi \frac{J_0(r\xi)}{A_1(\xi)} J_0(r'\xi)\, d\xi \ ,$$

so that by the definition of $A_1(\xi) = D_1\xi^2 + \Sigma_{R1} = D_1\left(\xi^2 + \sqrt{\frac{\Sigma_{R1}}{D_1}}^2\right)$, (11.3) may be explicitly written as

$$\begin{aligned} f(r') &= \frac{1}{D_1}\int_0^\infty \xi \frac{J_0(r\xi)}{\xi^2 + (\sqrt{\alpha_1})^2} J_0(r'\xi)\, d\xi \\ &= \begin{cases} \frac{1}{D_1} I_0(\sqrt{\alpha_1}r') K_0(\sqrt{\alpha_1}r) \ , & 0 < r' < r \\ \frac{1}{D_1} I_0(\sqrt{\alpha_1}r) K_0(\sqrt{\alpha_1}r') \ , & r' < r < \infty. \end{cases} \end{aligned}$$

Here, $\alpha_1 = \frac{\Sigma_{R1}}{D_1}$ and $I_0$ and $K_0$ are the modified Bessel functions of zero order. Then, we can write the final expression for $\phi_1^{(1)}$, i.e. the solution, using the fact that there is no source outside the cylinder

$$\begin{aligned} \phi_1^{(1)}(r) = &\frac{K_0(\sqrt{\alpha_1}r)}{D_1}\int_0^r r' I_0(\sqrt{\alpha_1}r') S_1(r')\, dr' \\ &+ \frac{I_0(\sqrt{\alpha_1}r)}{D_1}\int_r^R r' K_0(\sqrt{\alpha_1}r') S_1(r') dr' \ , \end{aligned}$$

and for $\phi_1^{(2)}$, we use only the definition of the Hankel transform to obtain

$$\phi_1^{(2)}(r) = p_1 S_2(r) \ .$$

By a similar procedure, we obtain the solution for $\phi_2(r)$ completing this way the entire solution of this problem using the infinite Hankel transform approach.

## 11.4 Results

We elaborated the general solutions in the previous sections, which for specific applications need the definitions of the parameter set and sources, respectively. Due to the fact that by virtue the specific source terms dominate the found solutions, in this section we present the influence of these source terms on the solution for a steady-state diffusion problem. The employed nuclear parameter sets are listed in Table 11.1, where for all cases we used $R = 5$, $k_{eff} = 0.95$ and $\nu = 2.5$ in the simulations.

**Table 11.1** Nuclear parameter sets

| | $D_1$ | $D_2$ | $S_0^{(1)}$ | $S_0^{(2)}$ | $\Sigma_{R1}$ | $\Sigma_{R2}$ | $\Sigma_{12}$ | $\Sigma_{21}$ | $\Sigma_{f1}$ | $\Sigma_{f2}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Set 1 | 1.43 | 0.39 | 4 | 0.0 | 0.029 | 0.104 | 0.015 | 0.00000 | 0.0041 | 0.0077 |
| Set 2 | 1.43 | 0.39 | 4 | 0.1 | 0.029 | 0.104 | 0.015 | 0.00825 | 0.0041 | 0.0077 |
| Set 3 | 1.43 | 0.51 | 4 | 0.1 | 0.052 | 0.081 | 0.015 | 0.00825 | 0.0041 | 0.0077 |
| Set 4 | 1.13 | 0.39 | 4 | 0.1 | 0.052 | 0.081 | 0.015 | 0.00825 | 0.0051 | 0.0081 |

We consider cases with different sources and compare results from the application of the finite and infinite Hankel transform, respectively. To this end, we consider for all cases a dominant source with fast neutrons and one case with no thermal neutron source and three cases with a weak thermal neutron source. A further differentiation stems from different removal cross sections for the fast and the thermal neutron group. The last set is distinct in comparison to all other ones because of an increased fission cross section in the fast and the thermal neutron group. For the first case, only the fast neutron source contributes,

$$S_1(r) = S_0^{(1)} H(R - r) \; .$$

Upon applying the finite Hankel transform, the source term is

$$\begin{aligned} \bar{S}_1(\xi_i) = H_0\{S_1\} &= \int_0^R r S_0^{(1)} H(R - r) J_0(r\xi_i) \, dr \\ &= S_0^{(1)} \int_0^R r J_0(r\xi_i) \, dr \\ &= S_0^{(1)} \frac{R}{\xi_i} J_1(R\xi_i) \; , \end{aligned}$$

and then using the final expression for the scalar neutron flux yields

$$\phi_1(r) = \frac{2}{R} S_0^{(1)} \sum_{i=1}^{\infty} \frac{A_2(\xi_i)}{\xi_i} \frac{1}{Det(M)(\xi_i)} \frac{J_0(r\xi_i)}{J_1(R\xi_i)} + \frac{2}{R} p_1 S_0^{(2)} \sum_{i=1}^{\infty} \frac{1}{\xi_i} \frac{1}{Det(M)(\xi_i)} \frac{J_0(r\xi_i)}{J_1(R\xi_i)} .$$

The procedure to obtain $\phi_2$ works in close analogy to the one for $\phi_1$. We obtained the following results for the selected parameter sets specified in Table 11.1 (Figs. 11.1, 11.2, 11.3 and 11.4).

By inspection of the obtained results, one observes qualitative agreement with what is expected from operational experience for processes inside a nuclear reactor core using this type of geometry. Quantitative properties are the flat current density at the origin, i.e. the flux with null derivative at $r = 0$ represents a symmetry condition. Furthermore, the vanishing flux at the outer boundary drags the flux from the maximum value at the center of the domain to decreasingly smaller values with

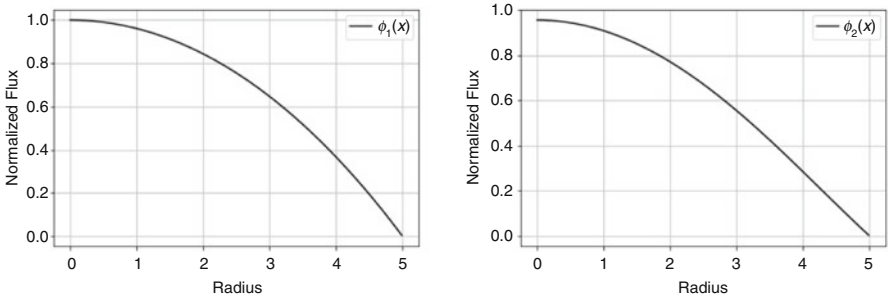



**Fig. 11.1** The scalar neutron flux for the fast and thermal energy group $\Phi_1$ and $\Phi_2$ for parameter set 1

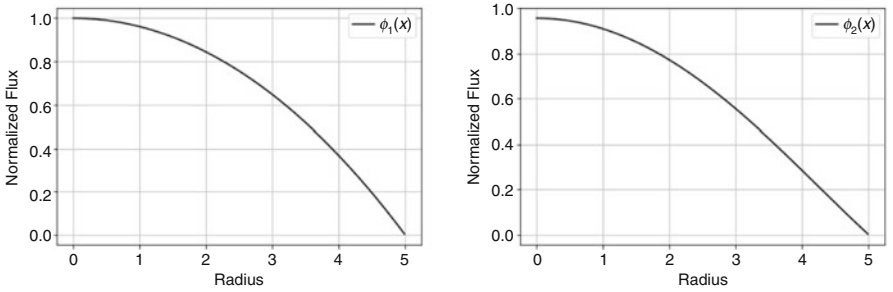



**Fig. 11.2** The scalar neutron flux for the fast and thermal energy group $\Phi_1$ and $\Phi_2$ for parameter set 2

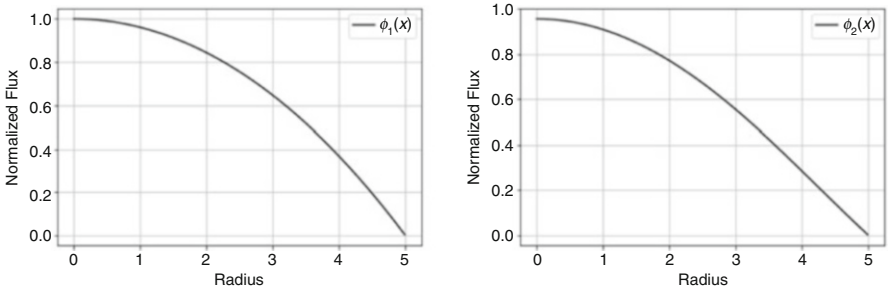



**Fig. 11.3** The scalar neutron flux for the fast and thermal energy group $\Phi_1$ and $\Phi_2$ for parameter set 3

increasing radius. As a systematic feature for all parameter sets, the fast flux always shows a somewhat larger concavity than the thermal flux. In order to provide a quantitative comparison between the solutions from the finite and infinite Hankel transforms, a table with the numerical values for the normalized solutions $\phi_2$ using both types of integral transforms is shown. Note that our findings agree fairly well with results in the literature [Da11].

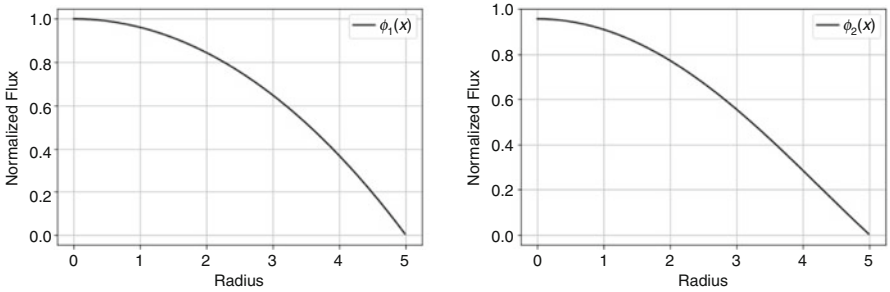



**Fig. 11.4** The scalar neutron flux for the fast and thermal energy group $\Phi_1$ and $\Phi_2$ for parameter set 4

**Table 11.2** Solution $\phi_2$ using the finite Hankel (FHT) and the infinite Hankel Transform (IHT) for parameter set 1

| $r/R$ | FHT | IHT |
|---|---|---|
| 0.0 | 1.000000 | 1.000000 |
| 0.1 | 0.987633 | 0.990253 |
| 0.2 | 0.950761 | 0.943312 |
| 0.3 | 0.890091 | 0.885213 |
| 0.4 | 0.806854 | 0.791032 |
| 0.5 | 0.702894 | 0.670122 |
| 0.6 | 0.580790 | 0.543311 |
| 0.7 | 0.444037 | 0.407630 |
| 0.8 | 0.297278 | 0.281963 |
| 0.9 | 0.146613 | 0.149313 |
| 1.0 | 0.000000 | 0.000000 |
| $N = 50$ | | |

| $r/R$ | FHT | IHT |
|---|---|---|
| 0.0 | 1.000000 | 1.000000 |
| 0.1 | 0.987633 | 0.990253 |
| 0.2 | 0.950761 | 0.943312 |
| 0.3 | 0.890091 | 0.885213 |
| 0.4 | 0.806854 | 0.791032 |
| 0.5 | 0.702894 | 0.670122 |
| 0.6 | 0.580790 | 0.543311 |
| 0.7 | 0.444037 | 0.407630 |
| 0.8 | 0.297278 | 0.281963 |
| 0.9 | 0.146613 | 0.149313 |
| 1.0 | 0.000000 | 0.000000 |
| $N = 100$ | | |

| $r/R$ | FHT | IHT |
|---|---|---|
| 0.0 | 1.000000 | 1.000000 |
| 0.1 | 0.987633 | 0.990253 |
| 0.2 | 0.950761 | 0.943312 |
| 0.3 | 0.890091 | 0.885213 |
| 0.4 | 0.806854 | 0.791032 |
| 0.5 | 0.702894 | 0.670122 |
| 0.6 | 0.580790 | 0.543311 |
| 0.7 | 0.444037 | 0.407630 |
| 0.8 | 0.297278 | 0.281963 |
| 0.9 | 0.146613 | 0.149313 |
| 1.0 | 0.000000 | 0.000000 |
| $N = 500$ | | |

In Table 11.2, results for $\phi_2$ using the finite and the infinite Hankel transform are shown. Comparing the solutions for the finite Hankel transform for truncations at $N = 50$, $N = 100$ and $N = 500$ shows stability of the obtained solution, so that $N = 50$ already provides a solution with six significant digits. However, for solution by the finite Hankel transform, it is not obvious where to truncate the series in order to obtain an acceptable solution, which depend on the cumbersome task of determining the roots of the Bessel functions of order 0 and order 1. From the comparison of the solution $\phi_2$ on the one hand by the finite and on the other hand by the infinite Hankel transform shows that the latter provides solutions fairly close to the ones by the finite integral transform. The advantage of the infinite Hankel transform over the finite case is that that there is no need to determine the lowest truncation of the series, which provides an acceptable solution. Besides having solved the stationary problem, where the found solution has value on its own right, the stationary case commonly provides the initial condition for a transient problem.

## 11.5 Conclusion

In the reported discussion, we presented and compared two integral methods to solve the stationary problem of two energy group neutron diffusion in cylinder geometry. Both methods, the finite and the infinite Hankel transform, generated comparable and acceptable results for the considered problems (parameter sets 1 to 4). While the finite Hankel transform seems to be the more natural tool to derive the solution due to the finite domain in consideration, the infinite sum of the analytical solution imposes the problem to determine truncation such that the approximate solution represents the exact solution to a prescribed accuracy. This task does not appear when the infinite Hankel transform is used, where it is the computation of the integrals that represents the challenge, and however numerical schemes for well-behaved integrands are usually no issue. All implemented simulations showed that both methods provide solutions with acceptable quality, but that the infinite integral transform is the simpler method especially due to the necessity to have a sufficiently large number $N$ of terms in the series of the solution by the finite Hankel transform.

From the computational point of view, the source code for the implementation was written in Python 3.8 for both integral transforms and ran on a simple home computer, Intel(R) Core(TM) i3-4150 CPU @ 3.50 GHz (64-bit operating system) with Microsoft Windows 10 operational system. For the solution by the infinite Hankel transform the CPU, time amounted to a few seconds, while the finite Hankel transform provided also a solution in a small but larger computational time, however with increasing tendency for increasing $N$. Our findings allow to conject that the solution by the infinite Hankel transform in principle opens pathways to increase the problem setup, such as to include more energy groups and allow for heterogeneous domains, which designs these new cases closer to the ones of real reactor cores.

## References

[BeGl70] Bell, G.I., Glasstone, S.: Nuclear Reactor Theory. Van Nostrand Reinhold Company, New York (1970)

[Da11] Dababneh, S.: An alternative solution of the neutron diffusion equation in cylindrical symmetry. Ann. Nuclear Energy **38**, 1040–1043 (2011)

[Du06] Dulla, S.: Space asymptotic method for the study of neutron propagation. Ann. Nuclear Energy **33**, 932–940 (2006)

[Fe11] Fernandes, J.C.L.: An analytical solution of multi-group neutron diffusion equation for a cylindrical symmetry using Hankel Transform, Thesis, Brazil (2011)

[Fe13] Fernandes, J.C.L.: On a comparative analysis of the solutions of the kinetic neutron diffusion equation by the Hankel transform formalism and the spectral method. Progr. Nuclear Energy **69**, 71–76 (2013)

[Gl03] Goncalves, G.A.: Analytical solution of neutron transport equation in cylindrical geometry, Thesis, Brazil (2003)

[La66] Lamarsh, J.R.: Introduction to Nuclear Reactor Theory. Addison Wesley Publishing, New York (1966)

[LeMi84] Lewis, E.E., Miller Jr., W.F.: Computational Methods of Neutron Transport. Wiley, New York (1984)

[Ma17] Manish, R.: Solution of neutron diffusion equation in 2d polar coordinates using nodal integral method. Ann. Nuclear Energy **105**, 69–78 (2017)

[Ma21] Manish, R.: Nodal Integral method for multi-group neutron diffusion equation in three dimensional cylindrical coordinate system. Ann. Nuclear Energy **151**, 107–114 (2021)

[No21] Di Nora, V.A.: Optimization of multi-group energy structures for diffusion analyses of sodium-cooled fast reactors assisted by simulated annealing – Part I: Methodology demonstration. Ann. Nuclear Energy**155**, 108–183 (2021)

[Ol19] Oliveira, F.R.: On an analytical solution for the two energy group neutron space-kinetic equation in heterogeneous cylindrical geometry. Ann. Nuclear Energy **133**, 216–220 (2019)

[Oz01] Ozgener, H.A.: A multi-region boundary element method for multigroup neutron diffusion calculations. Ann. Nuclear Energy **28**, 585–616 (2001)

[Oz73] Ozisik, M.N.: Radiative Transfer and Interaction with Conductions and Convection. Wiley, New York (1973)

[Vi08] Vilhena, M.T.B.: An analytical solution for the general perturbative diffusion equation by integral transform techniques. Ann. Nuclear Energy **35**, 2410–2413 (2008)

# Chapter 12
# A Simple Numerical Scheme to Obtain Reflectivity and Transmissivity of an Isotropically Scattering Slab


**C. A. Ladeia, H. R. Zanetti, D. L. Gisch, M. Schramm, and J. C. L. Fernandes**

## 12.1 Introduction

The radiative transfer equation has numerous applications such as radiation transport in the atmosphere, nuclear reactors, buildings, biological tissues, and vegetation, among others [ClOz83, Mo13, PiFu13, HoEtAl20]. Originally, the model has been formulated as an integro-differential equation, whose analytical solution is practically impossible to obtain for general cases, so that numerical iterative methods were developed over the years to obtain approximate solutions [ClOz83, LiWu96, CrEtAl17]. In order to simulate the essence of radiation transport, it is desirable to have a precise and reliable numerical model to solve the linear radiative transfer equation. The application of numerical methods in transport theory is the conventional approach, and it has been explored in classic textbooks such as [LewMi84, KaEtAl09] only to some essential extent, and there is practically no convergence analysis for iterative methods presented in their texts. However, many practical applications require reliable information of radiative fluxes (or partial fluxes), or at least their reflected and transmitted fractions [ClOz83, LiWu96], which allow to make contact with the experimental sector.

In this chapter, we focus on the linear radiative transfer model, considering a passive medium with no thermal contributions, which is an initial approximation for application problems. To this end, we consider a one-dimensional slab geometry and compute some fundamental properties for applications of radiative transfer, namely reflectivity and transmissivity. Our developments are based on the discrete

C. A. Ladeia (✉) · H. R. Zanetti · D. L. Gisch · J. C. L. Fernandes
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: cibele.ladeia@ufrgs.br; julio.lombaldo@ufrgs.br

M. Schramm
Federal University of Pelotas, Capão do Leão, RS, Brazil
e-mail: mschramm@ufpel.edu.br



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_12

ordinate method in the angular variable [Ch50] and a modified version of the finite volume method [LaEtAl20] in the spatial variable. As usual in transport theory, the solution is obtained by a line-by-line iterative method, and we present a simple but necessary condition for its convergence based on norm operations. As a case study, we compute values for reflectivity and transmissivity of isotropically scattering slabs.

## 12.2 The Radiative Transfer Equation in Slab Geometry

We consider the radiative transfer equation in slab geometry [Oz73],

$$\mu \frac{\partial}{\partial \tau} I(\tau, \mu) + I(\tau, \mu) = \frac{\omega(\tau)}{2} \int_{-1}^{1} P(\mu, \mu') I(\tau, \mu') \mathrm{d}\mu' \,, \tag{12.1}$$

where $\tau \in [0, L]$, $\tau$ is the optical depth variable, $\mu \in [-1, 1]$, $\mu = \cos\theta$, $\theta$ is the direction angle, and consequently $\mu'$ is the direction cosine of the incident rays, and $I = I(\tau, \mu)$ is the radiation intensity. Further, $\omega$ is the single scattering albedo; $P$ is the phase function and may be approximated as a truncated series with Legendre polynomials [Ch50].

$$P(\mu, \mu') = \sum_{\ell=0}^{\mathscr{L}} \beta_\ell \mathscr{P}_\ell(\mu) \mathscr{P}_\ell(\mu').$$

The boundary conditions of Eq. (12.1) are given for the forward and backward directions, respectively.

$$I(0, \mu) = f_1(\mu) + \epsilon_1 I_{b1}(T) - 2\rho_1 \int_{-1}^{0} I(0, \mu') \mu' \mathrm{d}\mu'$$

for $\mu > 0$ and

$$I(L, \mu) = f_2(\mu) + \epsilon_2 I_{b2}(T) + 2\rho_2 \int_{0}^{1} I(L, \mu') \mu' \mathrm{d}\mu'$$

for $\mu < 0$. Here, $f_1(\mu)$ and $f_2(\mu)$ are the external irradiation incident on the surfaces at $\tau = 0$ and $\tau = L$, respectively. Similarly, $\epsilon_1$ and $\epsilon_2$ are the emissivities, and $\rho_1$ and $\rho_2$ are the diffusive reflectivities on the surfaces with $\tau = 0$ and $\tau = L$, respectively. Although the medium is passive, at the two surfaces, one shall consider thermal radiation contributions $I_{b1}(T)$ and $I_{b2}(T)$, which are intensities due to black-body radiation for a temperature $T$.

## 12.3 Discrete Ordinate Method

So far the problem was formulated using continuous variables, and in order to cast the problem in discrete form, we follow the original work of Chandrasekhar [Ch50] and use the so-called $S_M$ approach in the polar angle, defining the intensity in the now discrete direction by $I_m(\tau) = I(\tau, \mu_m)$. The discrete ordinate version of Eq. (12.1) is

$$\mu_m \frac{\mathrm{d}}{\mathrm{d}\tau} I_m + I_m = \frac{\omega(\tau)}{2} \sum_{m'=1}^{M} w_{m'} P(\mu_m, \mu_{m'}) I_{m'} \tag{12.2}$$

for $m = 1, 2, \ldots, M$, where $\mu_m$ and $w_m$ are the (crescent) abscissas and weights of the Gauss–Legendre quadrature. For convenience, we consider only even values for $M$, such that $M/2$ is an integer. The discrete ordinates form of the boundary conditions is then

$$I_m(0) = f_1(\mu) + \epsilon_1 I_{b1}(T) - 2\rho_1 \sum_{m'=1}^{M/2} \mu_{m'} w_{m'} I_{m'}(0) \tag{12.3}$$

for $m = M/2 + 1, M/2 + 2, \ldots, M$ and

$$I_m(L) = f_2(\mu) + \epsilon_2 I_{b2}(T) + 2\rho_2 \sum_{m'=M/2+1}^{M} \mu_{m'} w_{m'} I_{m'}(L) \tag{12.4}$$

for $m = 1, 2, \ldots, M/2$. One of the reasons to adopt a discrete form of the original equations is that these may be cast in matrix form, so that one may make use of available techniques to solve the approximate equations.

## 12.4 Spatial Discretization

Different from the conventional $S_M$ approximation, we also discretize the spatial variable $0 \leq \tau \leq L$ in equally spaced nodes $\{\tau_i\}_{i=0}^{N}$, wherein $\tau_i = i\Delta\tau$ for $i = 0, 1, \ldots, N$ and $\Delta\tau = L/N$. One may construct a discretized version of the problem (12.2) upon using average values for each respective interval or formally upon applying the operator

$$\frac{1}{\Delta\tau} \int_{\tau_i}^{\tau_{i+1}} [\,\cdot\,]\,\mathrm{d}\tau$$

for $i = 0, 1, \ldots, N-1$. For the integrals, we used the trapezoidal rule and dropping the error term [LaEtAl20] so that Eq. (12.2) becomes

$$\mu_m \left( \frac{1}{\Delta\tau} I_m^{i+1} - \frac{1}{\Delta\tau} I_m^i \right) + \left( \frac{1}{2} I_m^{i+1} + \frac{1}{2} I_m^i \right) = \frac{\omega(\tau)}{4} \sum_{m'=1}^{M} w_{m'} P(\mu_m, \mu_{m'}) \left( I_{m'}^{i+1} + I_{m'}^i \right)$$

or in a shorthand notation

$$B_m^i I_m^{i+1} + A_m^i I_m^i = S_m^i \,, \tag{12.5}$$

where $I_m^i = I(\tau_i, \mu_m)$ and

$$A_m^i = -\frac{\mu_m}{\Delta\tau} + \frac{1}{2} \,, \tag{12.6}$$

$$B_m^i = \frac{\mu_m}{\Delta\tau} + \frac{1}{2} \,, \tag{12.7}$$

$$S_m^i = \frac{\omega(\tau)}{4} \sum_{m'=1}^{M} w_{m'} P(\mu_m, \mu_{m'}) \left( I_{m'}^{i+1} + I_{m'}^i \right) . \tag{12.8}$$

The boundary conditions (12.4) and (12.3) in discrete form are then given by

$$I_m^0 = f_1(\mu) + \epsilon_1 I_{b1}(T) - 2\rho_1 \sum_{m'=1}^{M/2} w_{m'} \mu_{m'} I_{m'}^0 \tag{12.9}$$

for $m = M/2 + 1, M/2 + 2, \ldots, M$ and

$$I_m^N = f_2(\mu) + \epsilon_2 I_{b2}(T) + 2\rho_2 \sum_{m'=M/2+1}^{M} w_{m'} \mu_{m'} I_{m'}^N \tag{12.10}$$

for $m = 1, 2, \ldots, M/2$.

This system is solved in an iterative scheme, updating the left-hand side values $I_m^{i+1}$ and using the previous (old) values in the right-hand side so that (12.5) reads

$$I_m^{i+1} = \frac{S_m^i - A_m^i I_m^i}{B_m^i} \tag{12.11}$$

or

$$I_m^i = \frac{S_m^i - B_m^i I_m^{i+1}}{A_m^i} \tag{12.12}$$

depending on whether the iterative process is directed top down or bottom up. Further, we used the stopping criterion

$$E \leq E_{\max},$$

where

$$E = \max_{\substack{m=1,\ldots,M \\ i=0,\ldots,N}} \frac{\left|I_m^i - \left(I_m^i\right)_{\text{old}}\right|}{|I_m^i| + |\left(I_m^i\right)_{\text{old}}|}. \tag{12.13}$$

Here, $\left(I_m^i\right)_{\text{old}}$ represents the values of $I_m^i$ in the previous iteration, and $E_{\max}$ is a measure for the relative maximum difference. The iterative process is described in the following pseudocode.

```
Get the problem data.
Choose the numerical data: N, M, and E_max.
Compute the method quantities, like τ_i and μ_m.
Compute A_m^i and B_m^i using (12.6) and (12.7).
Guess I_m^i for all m = 1,2,...,M and i = 0,1,...,N.
Set E = E_max + 1 (to enter the while loop).
While E > E_max
    for m = 1,2,...,M/2
        update I_m^N using (12.10);
        for i = N-1,N-2,...,0
            compute S_m^i using (12.8);
            update I_m^i using (12.12);
    for m = M/2+1,M/2+2,...,M
        update I_m^0 using (12.9);
        for i = 0,1,...,N-1
            compute S_m^i using (12.8);
            update I_m^{i+1} using (12.11);
    compute the relative error E with (12.13).
```

In order to calculate the reflectivity and transmissivity, we define the backward and forward radiation fluxes as

$$q^-(\tau) = \int_{-1}^{0} I(\tau,\mu)\mu \mathrm{d}\mu = \sum_{m'=1}^{M/2} w_{m'}\mu_{m'} I_{m'}^i,$$

$$q^+(\tau) = \int_{0}^{1} I(\tau,\mu)\mu \mathrm{d}\mu = \sum_{m'=M/2+1}^{M} w_{m'}\mu_{m'} I_{m'}^i,$$

and then the coefficients may be calculated from $q^-(0)$, $q^+(0)$, and $q^+(L)$. For the finite slab geometry, where $f_1(\mu) + \varepsilon_1 I_{b1}(T) \neq 0$ and $f_2(\mu) + \varepsilon_2 I_{b2}(T) = 0$, the reflectivity is the ratio between the total outgoing flux at $\tau = 0$ and the total

incoming flux at $\tau = 0$.

$$\mathscr{R} = \frac{\int_{-1}^{0} I(0,\mu)\mu \mathrm{d}\mu}{\int_{0}^{1} I(0,\mu)\mu \mathrm{d}\mu} = \frac{q^{-}(0)}{q^{+}(0)}.$$

Likewise, the transmissivity is the ratio between the total outgoing flux at $\tau = L$ and the total incoming flux at $\tau = 0$.

$$\mathscr{T} = \frac{\int_{0}^{1} I(L,\mu)\mu \mathrm{d}\mu}{\int_{0}^{1} I(0,\mu)\mu \mathrm{d}\mu} = \frac{q^{+}(L)}{q^{+}(0)}.$$

## 12.5 Numerical Results

To check if the present methodology is appropriate, we determine numerical values for reflectivity $\mathscr{R}$ and transmissivity $\mathscr{T}$ depending on a selection of values for albedos $\omega = \{0.2, 0.8, 0.995\}$, without surface emissivity and reflectivity $\epsilon_1 = \epsilon_2 = 0$ $\rho_1 = \rho_2 = 0$, and for isotropic incident radiation through the boundary with $\tau = 0$, thus $f_1(\mu) = 1$ and $f_2(\mu) = 0$ at $\tau = L$. Further, six values for the optical thickness were considered $L = \{0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$, and the phase function ($P(\mu, \mu') = 1$) was set up for isotropic scattering with $\mathscr{L} = 0$. The spatial and angular mesh from discretization was implemented with $N = 50$ and $M = 100$. Figure 12.1 depicts the physical scenario in the domain with depth $L$ and the total radiative fluxes across the respective boundaries ($q^{+}(0)$, $q^{+}(L)$, $q^{-}(0)$) at $\tau = 0$ and $\tau = L$, respectively. It is noteworthy that reflectivity $\mathscr{R}$ and transmissivity $\mathscr{T}$ in the present case are merely a consequence of the physical properties of the medium, which is dominated by isotropic scattering.

The computed values for the reflectivity and transmissivity from the total radiative fluxes are shown in Figs. 12.2 and 12.3, respectively. On observing the effect of the homogeneous medium with constant albedo everywhere and isotropic scattering, the higher the albedo value the higher is reflectivity and the lower is transmissivity for all optical thicknesses. For lower albedos, one observes in Fig. 12.2 a saturation for larger optical thicknesses that may be understood as a balance between attenuation by absorption and the compensation by the isotropic source $f_1(\mu)$ on the boundary at $\tau = 0$, so that deeper regions contribute less to $q^{-}(0)$ through scattering. In Fig. 12.3, one observes the attenuation effects with increasing optical thickness, which is to be expected by the Beer–Lambert–Bouguer law.

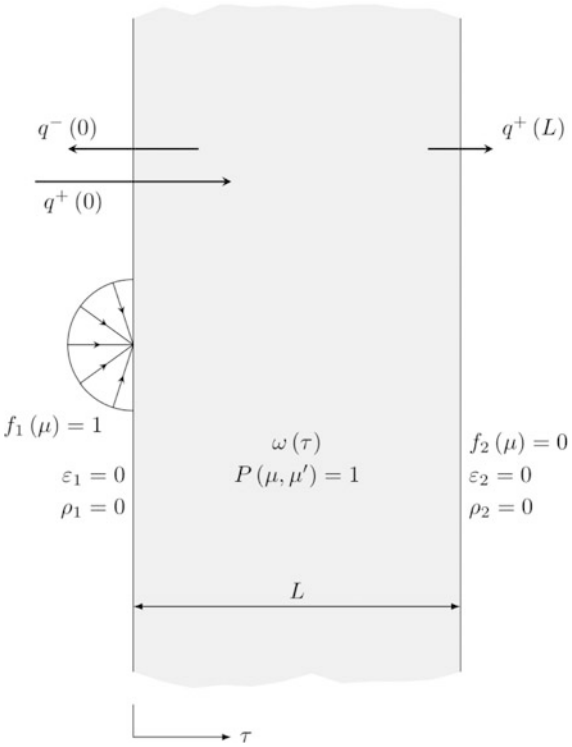



**Fig. 12.1** Physical scenario with medium and boundary properties and the respective total radiative fluxes across the boundaries

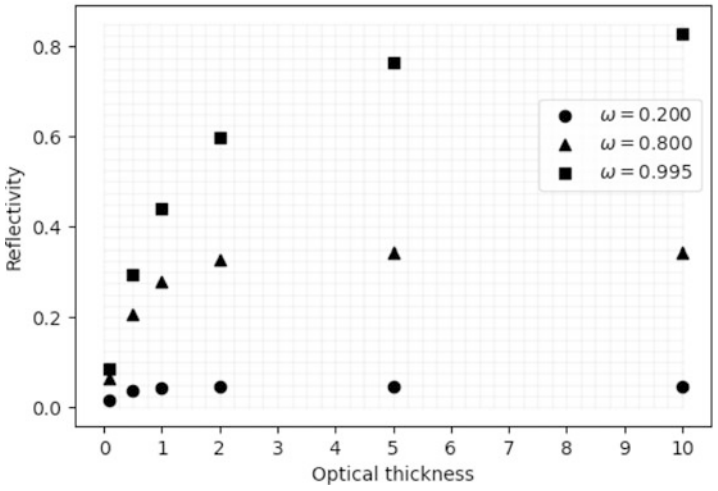



**Fig. 12.2** Reflectivity along the optical thickness for different albedos

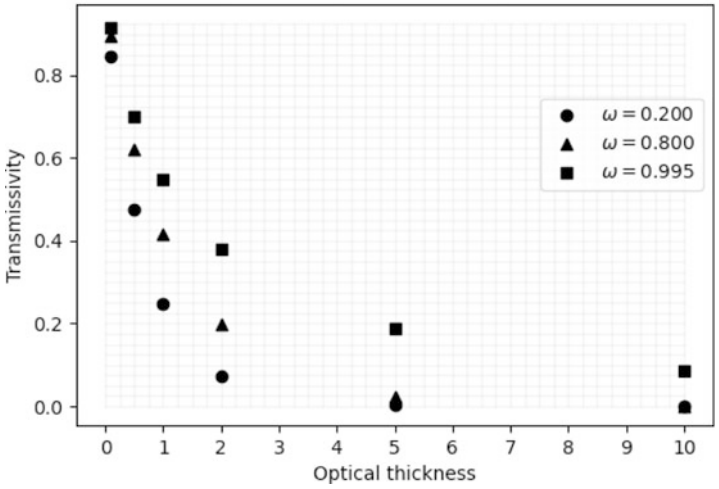



**Fig. 12.3** Transmissivity along the optical thickness for different albedos

## 12.6 A Convergence Criterion

The iterative scheme may be written in matrix notation

$$\boldsymbol{L}\boldsymbol{I}_k = -\boldsymbol{U}\boldsymbol{I}_{k-1} + \boldsymbol{b}\,, \tag{12.14}$$

where $k$ is the iteration index, while the exact solution $\boldsymbol{I}^*$ satisfies

$$\boldsymbol{L}\boldsymbol{I}^* = -\boldsymbol{U}\boldsymbol{I}^* + \boldsymbol{b}\,. \tag{12.15}$$

The sequence $\{\boldsymbol{I}_k\}$ converges to the exact solution if

$$\boldsymbol{I}_k - \boldsymbol{I}^* \to \boldsymbol{0} \tag{12.16}$$

for all $k$ greater than certain value. One standard way of assuring (12.16) is taking the norm in order to get semi-positive differences only, so if

$$\left\|\boldsymbol{I}_k - \boldsymbol{I}^*\right\| \to 0\,,$$

then (12.16) is true and the sequence converges to the exact solution $\boldsymbol{I}^*$. Here, $\|\cdot\|$ represents the maximum norm.

Upon subtracting (12.15) from (12.14), we obtain

$$\boldsymbol{I}_k - \boldsymbol{I}^* = -\boldsymbol{L}^{-1}\boldsymbol{U}\left(\boldsymbol{I}_{k-1} - \boldsymbol{I}^*\right)\,,$$

**Table 12.1** Matrix norms of $\boldsymbol{L}$ and $\boldsymbol{U}$ for all test cases

| | $\omega = 0,2$ | | $\omega = 0,8$ | | $\omega = 0,995$ | |
|---|---|---|---|---|---|---|
| $L$ | $\|\boldsymbol{L}\|$ | $\|\boldsymbol{U}\|$ | $\|\boldsymbol{L}\|$ | $\|\boldsymbol{U}\|$ | $\|\boldsymbol{L}\|$ | $\|\boldsymbol{U}\|$ |
| 0.1 | 979.96938 | 0.24993 | 980.71916 | 0.99971 | 980.96284 | 1.24338 |
| 0.5 | 196.19382 | 0.24993 | 196.94360 | 0.99971 | 197.18727 | 1.24338 |
| 1 | 98.22187 | 0.24993 | 98.97165 | 0.99971 | 99.21533 | 1.24338 |
| 2 | 49.23590 | 0.24993 | 49.98568 | 0.99971 | 50.22936 | 1.24338 |
| 5 | 19.84432 | 0.24993 | 20.59410 | 0.99971 | 20.83777 | 1.24338 |
| 10 | 10.04712 | 0.24993 | 10.79690 | 0.99971 | 11.04058 | 1.24338 |

and therefore, we have an expression for the $k$-th error starting from an initial guess $\boldsymbol{I}_0$,

$$\boldsymbol{I}_k - \boldsymbol{I}^* = \left(-\boldsymbol{L}^{-1}\boldsymbol{U}\right)^k \left(\boldsymbol{I}_0 - \boldsymbol{I}^*\right) \tag{12.17}$$

for $k = 1, 2, 3, \ldots$.

Applying the maximum norm in (12.17) and using some norm operations, we get

$$\left\|\boldsymbol{I}_k - \boldsymbol{I}^*\right\| \leq \left\|\boldsymbol{L}^{-1}\boldsymbol{U}\right\|^k \left\|\boldsymbol{I}_0 - \boldsymbol{I}^*\right\| ,$$

where it becomes evident that $\left\|\boldsymbol{I}_k - \boldsymbol{I}^*\right\| \to 0$ as $k \to \infty$ if $\left\|\boldsymbol{L}^{-1}\boldsymbol{U}\right\| < 1$. Further, using some norm product operations, we get that if

$$\left\|\boldsymbol{L}^{-1}\right\| \|\boldsymbol{U}\| < 1, \tag{12.18}$$

then the scheme is convergent. Moreover, upon multiplying both sides in (12.18) by $\|\boldsymbol{L}\|$ and using $\left\|\boldsymbol{L}^{-1}\right\| \|\boldsymbol{L}\| \geq 1$, then the iterative scheme converges only if

$$\|\boldsymbol{L}\| > \|\boldsymbol{U}\| .$$

Table 12.1 shows $\|\boldsymbol{L}\|$ and $\|\boldsymbol{U}\|$ for the test cases.

## 12.7 Final Remarks and Conclusion

In the present work, the linear radiative problem in slab geometry was solved by a numerical scheme using a spatial and angular mesh. The reflectivities and transmissivities for a selection of albedos of an isotropically scattering medium were determined for an isotropic incident radiation. The obtained results are in agreement with expectations from experimental physics, which may be used as an indication that the computational implementations are consistent. The codes were written in the

Python programming language version 3.8, which proved to be quick and effective for all simulated cases. The numerical results were computed on a standard personal computer, and execution times for all simulated cases were of the order of $10^0$ $s$, thus reaffirming the effectiveness and applicability of the methodology.

As a highlight of this work, a convergence criterion was implemented, which to the best of our knowledge is usually absent in the literature of transport theory. Quite often benchmark results are used as reference solutions, however without a desirable justification for their claim of high precision. Despite the presented convergence criterion is required, it is not sufficient to assure general convergence of this iterative scheme; however, it may be considered a starting point toward a rigorous convergence criterion for problems in transport theory that make use of source iterations, a usual technique in the field. As a continuation of the present treatise, we will develop a sufficient convergence criterion for linear cases, which will set then the base for a genuine convergence criterion for the nonlinear radiative transfer equation.

## References

[Ch50] Chandrasekhar, S.: The Radiative Transfer. Oxford University Press, New York (1950)

[ClOz83] Clements, T.B., Ozisik, M.N.: Effects of stepwise variation of albedo on reflectivity and transmissivity of an isotropically scattering slab. Int. J. Heat Mass Trans. **26**, 1419–1426 (1983)

[CrEtAl17] Cromianskim, S.R, Camargo, M., Rodrigues, P., Barichello, L.B.: Avaliação de Propriedades Radiativas em Meios Homogêneos Unidimensionais: Reflectância e Transmitância. TEMA **18**, 531–547 (2017)

[HoEtAl20] Howell, J.R., Menguç, M.P., Daun, K., Siegel, R.: Thermal Radiation Heat Transfer. CRC Press, Boca Raton (2020)

[KaEtAl09] Kanschat, G., Meinköhn, E., Rannacher, R., Wehrse, R. (eds.): Numerical Methods in Multidimensional Radiative Transfer. Springer, Berlin (2009)

[LaEtAl20] Ladeia, C.A., Schramm, M., Fernandes, J.C.L.: A simple numerical scheme to linear radiative transfer in hollow and solid spheres. Semin., Ciênc. Exatas Tecnol. **41**, 21–30 (2020)

[LewMi84] Lewis, E.E., Miller, W.F.: Computational Methods of Neutron Transport. Wiley, New York (1984)

[LiWu96] Liou, B.-T., Wu, C.-Y.: Composite discrete-ordinate solutions for radiative transfer in a two-layer medium with Fresnel interfaces. Taylor Francis **30**, 739–751 (1996)

[Mo13] Modest, M.F.: Radiative Heat Transfer. Academic, New York (2013)

[Oz73] Ozisik, M.N.: Radiative Transfer and Interaction with Conduction and Convection. Wiley, New York (1973)

[PiFu13] Picca, P., Furfaro, R.: Analytical discrete ordinate method for radiative transfer in dense vegetation canopies. J. Quant. Spectrosc. Radiat. Trans. **118**, 60–69 (2013)

# Chapter 13
# A Unified Integral Equation Formulation for Linear and Geometrically Nonlinear Analysis of Thick Plates: Derivation of Equations


**R. J. Marczak**

## 13.1 Introduction

Numerical solutions for geometrically nonlinear bending of moderately thick plates are well reported in the literature. Among the conventional numerical methods used to solve this type of problem, the boundary element method (BEM) has been receiving relatively little attention on the subject, in spite of the excellence of the results obtained with the method for linear problems [We82, KaTe88] [RaEtAl97, Ra15]. Many reasons have contributed to prevent the general application of the BEM in nonlinear problems. The generality of the finite element method is obviously one of them, but some mathematical aspects inherent to integral equation methods have contributed as well. As one of these aspects, one could mention the so-called convective (or free) terms that arise in derivative integral equations, as these terms are sometimes misunderstood or even missing from the equations.

The objective of this chapter is to outline the deduction of the convective terms appearing in integral equations for large displacement analysis of Mindlin and Reissner plate models. There are only few works exploring the solution of geometrically nonlinear thick-plate bending problems using the BEM [XiEtAl90, Vi90, Ji91, XiQui93, SuEtAl94, Ra98]. However, most of them do not present the derivation of the free terms, and, in addition to the best of the author's knowledge, no one shows results for maximum transverse displacement far beyond the plate thickness magnitude. The present work aims to outline a clear and didactic derivation of such terms, as they are quite common in nonlinear applications using boundary integral equation methods.

R. J. Marczak (✉)
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: rato@mecanica.ufrgs.br



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_13

The Mindlin and the Reissner plate theories are very well-known structural models. In his celebrated work, E. Reissner [Re44] started from a stress field and a mixed variational principle to obtain the equilibrium equations. The Hencky–Bollé–Mindlin (or simply Mindlin, as it is generally known) plate model [Bo47, Mi51] can be more easily obtained departing from a kinematical point of view, where the Kirchhoff–Love normality (thin plate) condition is relaxed.

$$U_\alpha\left(x_1, x_2, x_3\right) = u_\alpha\left(x_1, x_2\right) + x_3 \psi_\alpha\left(x_1, x_2\right) \tag{13.1a}$$

$$U_3\left(x_1, x_2, x_3\right) = u_3\left(x_1, x_2\right). \tag{13.1b}$$

In all expressions throughout this chapter, Greek indices range from 1 to 2, while Latin indices range from 1 to 3. Here, $\mathbf{u}$ contains the membrane (in-plane) and transverse plate displacements, respectively (i.e., $u_\alpha$ and $u_3$), while $\psi_\alpha$ are the plate rotations. All variables are referred to the plate's middle surface. If taken pointwise across the thickness, the displacement field of the Reissner model is more complex than postulated in Eq. (13.1). However, the middle surface fields remain valid for this model if it is interpreted as a weighed mean value of the displacement field across the thickness $h$.

$$\psi_\alpha^{\text{Mindlin}} = \frac{12}{h^3}\int_{-h/2}^{h/2} \psi_\alpha^{\text{Reissner}}(x_1, x_2, x_3)\, x_3\, dx_3$$

$$u_3^{\text{Mindlin}} = \frac{3}{2h}\int_{-h/2}^{h/2} u_3^{\text{Reissner}}(x_1, x_2, x_3)\left[1 - \left(\frac{2x_3}{h}\right)^2\right] dx_3.$$

The in-plane displacements are included in Eq. (13.1) because the two-dimensional elasticity behavior will be superimposed on the plate bending equations, aiming the derivation of equilibrium equations for geometrically nonlinear bending problems. These are found to be written in terms of resultant stresses following the reasoning of reference [Fu65].

$$N_{\alpha\beta,_\beta} + q_\alpha = 0 \tag{13.2a}$$

$$\left(N_{\alpha\beta} u_{3,_\alpha}\right)_{,\beta} + Q_{\alpha,_\alpha} + q_3 = 0 \tag{13.2b}$$

$$M_{\alpha\beta,_\beta} - Q_\alpha + m_\alpha = 0. \tag{13.2c}$$

Here, $N_{\alpha\beta}$ are the in-plane (membrane) forces, $Q_\alpha$ are the shear forces, and $M_{\alpha\beta}$ are the bending moments. The symbols $q_\alpha$ and $q_3$ stand for in-plane and transverse loadings, respectively, while $m_\alpha$ are the distributed moments. Equation (13.2) can be recovered in terms of displacements through the stress–displacement relations.

$$N_{\alpha\beta} = C\frac{1-\nu}{2}\left[u_{\alpha,_\beta} + u_{\beta,_\alpha} + u_{3,_\alpha} u_{3,_\beta} + \frac{2\nu}{1-\nu}\left(u_{\gamma,_\gamma} + \frac{1}{2}u_{3,_\gamma} u_{3,_\gamma}\right)\delta_{\alpha\beta}\right] \tag{13.3a}$$

$$M_{\alpha\beta} = D\frac{1-\nu}{2}\left[u_{\alpha,\beta} + u_{\beta,\alpha} + \frac{2\nu}{1-\nu}u_{\gamma,\gamma}\delta_{\alpha\beta}\right] \tag{13.3b}$$

$$Q_\alpha = D\lambda^2\frac{1-\nu}{2}\left[u_\alpha + u_{3,\alpha}\right]. \tag{13.3c}$$

Further, $C = \frac{Eh}{(1-\nu^2)}$, $D = \frac{Eh^3}{12(1-\nu^2)}$, $\lambda^2 = \frac{12\kappa^2}{h^2}$, and $\kappa^2$ is the shear stress correction factor. In comparison to the plate theory commonly used, the only visible difference in Eq. (13.3) is the expression for the moments, which has an additional term in the Reissner plate model.

$$M_{\alpha\beta}^{\text{Reissner}} = \text{ R.H.S. of Eq. (13.3b)} + \frac{\nu}{(1-\nu)\lambda^2}\, q_3\, \delta_{\alpha\beta}. \tag{13.4}$$

In order to unify the equilibrium equations in the same computational model, a plate model factor ($m_f$) is employed [WeBa90].

$$M_{\alpha\beta} = D\frac{1-\nu}{2}\left[\psi_{\alpha,\beta} + \psi_{\beta,\alpha} + \frac{2\nu}{1-\nu}\psi_{\gamma,\gamma}\delta_{\alpha\beta}\right] + m_f\ q_3\,\delta_{\alpha\beta}\ , \tag{13.5}$$

where

$$m_f = \frac{\nu}{(1-\nu)\lambda^2} \qquad \text{for the Reissner model,} \tag{13.6a}$$

$$m_f = 0 \qquad \text{for the Mindlin model.} \tag{13.6b}$$

Equation (13.2) describes moderately thick-plate bending problems for large displacements and a moderately large rotations regime [Fu65]. In view of Eq. (13.5), they can be used regardless of the plate model considered, including the classical Kirchhoff–Love model. The presence of the nonlinear terms in Eq. (13.3) is a consequence of relevant higher-order terms kept in the Green–Lagrange strain tensor. Both the linear and nonlinear contributions can be further evidenced by writing,

$$N_{\alpha\beta} = N_{\alpha\beta}^l + N_{\alpha\beta}^n\ , \tag{13.7a}$$

$$Q_\alpha = Q_\alpha^l + Q_\alpha^n\ , \tag{13.7b}$$

where

$$N_{\alpha\beta}^l = C\frac{1-\nu}{2}\left[\bar{u}_{\alpha,\beta} + \bar{u}_{\beta,\alpha} + \frac{2\nu}{1-\nu}\bar{u}_{\gamma,\gamma}\delta_{\alpha\beta}\right]\ , \tag{13.8a}$$

$$N_{\alpha\beta}^n = C\frac{1-\nu}{2}\left[u_{3,\alpha}u_{3,\beta} + \frac{\nu}{1-\nu}u_{3,\gamma}u_{3,\gamma}\delta_{\alpha\beta}\right]\ , \tag{13.8b}$$

$$Q_\alpha^l = D\lambda^2 \frac{1-\nu}{2}\left(u_\alpha + u_{3,\alpha}\right) , \tag{13.8c}$$

$$Q_\alpha^n = N_{\alpha\beta}\, u_{3,\beta} . \tag{13.8d}$$

Upon substituting these into the equilibrium equations, one obtains the (coupled) Navier equations of the problem, where the nonlinear terms are added to the loading terms in a general system.

$$\begin{bmatrix} {}^m\mathbf{L} & \mathbf{0} \\ \mathbf{0} & {}^f\mathbf{L} \end{bmatrix} \begin{pmatrix} {}^m\mathbf{u} \\ {}^f\mathbf{u} \end{pmatrix} = \begin{pmatrix} {}^m\widehat{\mathbf{q}} \\ {}^f\widehat{\mathbf{q}} \end{pmatrix}. \tag{13.9}$$

Here, ${}^m\mathbf{L}$ is the differential operator of the linear membrane equilibrium problem, ${}^f\mathbf{L}$ is the linear bending operator, ${}^m\mathbf{u} = \{u_1\ u_2\}^T$ are the in-plane displacements, and ${}^f\mathbf{u} = \{\psi_1\ \ \psi_2\ \ u_3\}^T$ are the plate displacements. The membrane-bending coupling is implicit in the corresponding pseudo-loadings ${}^m\hat{\mathbf{q}}$ and ${}^f\hat{\mathbf{q}}$.

$${}^m\widehat{q}_\alpha = -\,{}^m\mathscr{F}_{\alpha\beta}(\partial_Q)\,{}^m q_\beta^l(Q) + {}^m q_\alpha^n(Q) \tag{13.10a}$$

$${}^f\widehat{q}_i = -\,{}^f\mathscr{F}_{ij}(\partial_Q)\,{}^f q_j^l(Q) + {}^f q_i^n(Q). \tag{13.10b}$$

The complete expressions of the terms used in Eqs. (13.9) and (13.10a)–(13.10b) are as follows.

$${}^m\mathbf{L}(\partial_Q) = C\frac{1-\nu}{2}\begin{bmatrix} \Delta + \dfrac{1+\nu}{1-\nu}\dfrac{\partial^2}{\partial x_1^2} & \dfrac{1+\nu}{1-\nu}\dfrac{\partial^2}{\partial x_1 \partial x_2} \\ \dfrac{1+\nu}{1-\nu}\dfrac{\partial^2}{\partial x_1 \partial x_2} & \Delta + \dfrac{1+\nu}{1-\nu}\dfrac{\partial^2}{\partial x_2^2} \end{bmatrix} \tag{13.11a}$$

$${}^f\mathbf{L}(\partial_Q) = D\frac{1-\nu}{2}\begin{bmatrix} \Delta - \lambda^2 + \dfrac{1+\nu}{1-\nu}\dfrac{\partial^2}{\partial x_1^2} & \dfrac{1+\nu}{1-\nu}\dfrac{\partial^2}{\partial x_1 \partial x_2} & -\lambda^2\dfrac{\partial}{\partial x_1} \\ \dfrac{1+\nu}{1-\nu}\dfrac{\partial^2}{\partial x_1 \partial x_2} & \Delta - \lambda^2 + \dfrac{1+\nu}{1-\nu}\dfrac{\partial^2}{\partial x_2^2} & -\lambda^2\dfrac{\partial}{\partial x_2} \\ \lambda^2\dfrac{\partial}{\partial x_1} & \lambda^2\dfrac{\partial}{\partial x_2} & \lambda^2\Delta \end{bmatrix} \tag{13.11b}$$

$$ {}^{m}\mathbf{F}(\partial_Q) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{13.12a} $$

$$ {}^{f}\mathbf{F}(\partial_Q) = \begin{bmatrix} 1 & 0 & m_f \dfrac{\partial}{\partial x_1} \\ 0 & 1 & m_f \dfrac{\partial}{\partial x_2} \\ 0 & 0 & 1 \end{bmatrix} \tag{13.12b} $$

$$ {}^{m}\mathbf{q}^{l}(Q) = \left\{ q_1,\, q_2 \right\}^T \tag{13.13a} $$

$$ {}^{m}\mathbf{q}^{n}(Q) = C\frac{1-\nu}{2} \begin{Bmatrix} (u_{3,1}u_{3,\alpha})_{,\alpha} + \dfrac{\nu}{1-\nu}(u_{3,\gamma}u_{3,\gamma})_{,1} \\ (u_{3,2}u_{3,\beta})_{,\beta} + \dfrac{\nu}{1-\nu}(u_{3,\gamma}u_{3,\gamma})_{,2} \end{Bmatrix} \tag{13.13b} $$

$$ {}^{f}\mathbf{q}^{l}(Q) = \left\{ m_1,\, m_2,\, q_3 \right\}^T \tag{13.14a} $$

$$ {}^{f}\mathbf{q}^{n}(Q) = D\frac{1-\nu}{2} \left\{ 0,\, 0,\, (N_{\alpha\beta}u_{3,\beta})_{,\alpha} \right\}^T . \tag{13.14b} $$

Equations (13.2) and (13.3)—with Eq. (13.5) replacing equation (13.3b)—are taken herein as a starting point for an incremental integral formulation. Using the weighted residual method [BrEtAl84], the following Somigliana identities for boundary variables are obtained [XiEtAl90, Ra98, Ra15],

$$ {}^{m}C_{\alpha\beta}(p)\,{}^{m}u_{\beta}(p) + \int_{\Gamma} {}^{m}T_{\alpha\beta}(q,p)\,{}^{m}u_{\beta}(q)\,d\Gamma_q = \int_{\Gamma} {}^{m}U_{\alpha\beta}(q,p)\,{}^{m}t_{\beta}(q)\,d\Gamma_q $$
$$ + \int_{\Omega} {}^{m}V_{\alpha\beta}(Q,p)\,{}^{m}q_{\beta}(Q)\,d\Omega_Q - \int_{\Omega} {}^{m}U_{\alpha\beta,\gamma}(Q,p)N^{n}_{\beta\gamma}(Q)\,d\Omega_Q + {}^{m}v_{\alpha}(p) \tag{13.15} $$

and

$$ {}^{f}C_{ij}(p)\,{}^{f}u_{j}(p) + \int_{\Gamma} {}^{f}T_{ij}(q,p)\,{}^{f}u_{j}(q)\,d\Gamma_q = \int_{\Gamma} {}^{f}U_{ij}(q,p)\,{}^{f}t_{j}(q)\,d\Gamma_q $$
$$ + \int_{\Omega} {}^{f}V_{ij}(Q,p)\,{}^{f}q_{j}(Q)\,d\Omega_Q - \int_{\Omega} {}^{f}U_{i3,\beta}(Q,p)N_{\alpha\beta}(Q)\,{}^{f}u_{3,\alpha}(Q)\,d\Omega_Q + {}^{f}v_{i}(p)\,, \tag{13.16} $$

where the $m$ and $f$ prefixes refer to the membrane and the bending problem, respectively, and the non-integral terms ${}^{m}v_{\beta}$ and ${}^{f}v_{i}$ were included to account for concentrated loads inside the domain [KaSa85]. The symbols $p$ and $q$ denote source (collocation) and field points, where lower case letters indicate boundary points

and upper case letters indicate domain points, respectively. The corresponding displacement (${}^{m}U_{ij}$ and ${}^{f}U_{ij}$), traction (${}^{m}T_{ij}$ and ${}^{f}T_{ij}$), and the other fundamental solution tensors can be found elsewhere ([We82, Ra98]). Equations (13.15) and (13.16) are easily particularized for internal points upon substituting ${}^{m}C_{\alpha\beta} = \delta_{\alpha\beta}$ and ${}^{f}C_{ij} = \delta_{ij}$.

From Eqs. (13.15) and (13.16), it is evident that the evaluation of the derivatives of the transverse displacement ($u_3$) is required. They are present in the nonlinear membrane forces in the last integral of Eq. (13.15) and also in the last integral of Eq. (13.16). These terms are partially responsible for the membrane-bending coupling. In domain methods such as finite elements, it is typical to employ the derivatives of the shape functions, i.e., $u_{i,\alpha} = \phi_{i,\alpha} u_i$, where $\phi_i$ are the shape functions. Despite being simple, this approach may generate poor results when the global shape function is not able to represent accurately the gradients of the displacement field. Similar approaches can be used for boundary elements, but the use of higher-order domain cells becomes mandatory for acceptable results (see, for instance, [Vi90]). In the case of employing the boundary element method, there is no need to assume an *a priori* interpolated form for the displacement derivatives since equations (13.15) and (13.16) are already a strong form of the displacement field. Therefore, a more rigorous solution can be obtained by differentiation of these integral equations with respect to the coordinates $x_\alpha(P)$. The procedure leads to the six additionally required integral equations for $\psi_{\beta,\alpha}$ and $u_{3,\alpha}$.

Assuming that the displacement derivatives are required only at internal points, the differentiation of Eqs. (13.15) and (13.16) is straightforward as all their kernels become regular. However, the differentiation of the last two integrals on the right-hand side of both equations is rather tedious because the tensors ${}^{m}V_{\beta\gamma,\alpha}$, ${}^{f}V_{3i,\alpha}$, ${}^{m}U_{\alpha\beta,\gamma}$ and ${}^{f}U_{i3,\beta}$ have weak singularities when $Q \equiv P$. Taking into account the dimension of the corresponding integration domains, one can show that the integral containing ${}^{f}\mathbf{V}$ is singular only in the case of Reissner's plate model, while ${}^{m}\mathbf{V}$ is always regular [WeBa90]. Unfortunately, the differentiation of integrals containing singular kernels does not obey the classical calculus rules, and they must be treated by means of the Leibnitz formula [Mi62, Bu78]. The formal derivation of such derivative integral equations produces the so-called convective terms [BrEtAl84], which must be added to the final expressions for $\bar{u}_{\beta,\alpha}(P)$ and $u_{3,\alpha}(P)$.

$$
\begin{aligned}
\bar{u}_{\beta,\alpha}(P) - \int_\Gamma {}^{m}T_{\beta\gamma,\alpha}(q,P)\bar{u}_\gamma(q)\,d\Gamma_q &= -\int_\Gamma {}^{m}U_{\beta\gamma,\alpha}(q,P)\bar{t}_\gamma(q)\,d\Gamma_q \\
&- \int_\Omega {}^{m}V_{\beta\gamma,\alpha}(Q,P)q_\gamma(Q)\,d\Omega_Q + \int_\Omega {}^{m}U_{\beta\gamma,\delta\alpha}(Q,P)N^n_{\gamma\delta}(Q)\,d\Omega_Q \\
&+ N^n_{\gamma\delta}(P)\int_{\Gamma_1'} {}^{m}U_{\beta\gamma,\delta}(Q,P)\,r_{,\alpha}(P)\,d\Gamma_{Q_1} \\
&- q_\gamma(P)\int_{\Gamma_1'} {}^{m}V_{\beta\gamma}(Q,P)\,r_{,\alpha}(P)\,d\Gamma_{Q_1} - {}^{m}v_{\beta,\alpha}(P)
\end{aligned}
\tag{13.17}
$$

$$u_{3,\alpha}(P) - \int_{\Gamma} {}^{f}T_{3i,\alpha}(q,P)u_i(q)\,d\Gamma_q = -\int_{\Gamma} {}^{f}U_{3i,\alpha}(q,P)t_i(q)\,d\Gamma_q$$
$$- \int_{\Omega} {}^{f}V_{3i,\alpha}(Q,P)q_i(Q)\,d\Omega_Q + \int_{\Omega} {}^{f}U_{33,\alpha\gamma}(Q,P)N_{\beta\gamma}(Q)u_{3,\beta}(Q)\,d\Omega_Q$$
$$+ N_{\beta\gamma}(P)u_{3,\beta}(P)\int_{\Gamma_1'} {}^{f}U_{33,\gamma}(Q,P)\,r_{,\alpha}(P)\,d\Gamma_{Q_1}$$
$$- m_f\ q_i(P)\int_{\Gamma_1'} {}^{f}V_{3i}(Q,P)\,r_{,\alpha}(P)\,d\Gamma_{Q_1} - {}^{f}v_{3,\alpha}(P). \tag{13.18}$$

A negative sign was added to all the integrals as the derivatives are assumed to be taken with respect to $x_\alpha(P)$. The integrals on $\Gamma_1'$ in Eqs. (13.17) and (13.18) are the aforementioned convective terms, and $\Gamma_1'$ stands for a unit circle centered in $P$, where the derivation of the former is the objective of the present work. In the further, the main goal is to solve the analytical expressions for all four convective terms.

$${}^{f}c^{N}_{\alpha\beta}(P) = N_{\beta\gamma}(P)u_{3,\beta}(P)\int_{\Gamma_1'} {}^{f}U_{33,\gamma}(Q,P)\,r_{,\alpha}(P)\,d\Gamma_{Q_1} \tag{13.19a}$$

$${}^{f}c^{q}_{\alpha}(P) = m_f\ q_i(P)\int_{\Gamma_1'} {}^{f}V_{3i}(Q,P)\,r_{,\alpha}(P)\,d\Gamma_{Q_1} \tag{13.19b}$$

$${}^{m}c^{N}_{\alpha\beta}(P) = N^{n}_{\gamma\delta}(P)\int_{\Gamma_1'} {}^{m}U_{\beta\gamma,\delta}(Q,P)\,r_{,\alpha}(P)\,d\Gamma_{Q_1} \tag{13.19c}$$

$${}^{m}c^{q}_{\alpha\beta}(P) = q_\gamma(P)\int_{\Gamma_1'} {}^{m}V_{\beta\gamma}(Q,P)\,r_{,\alpha}(P)\,d\Gamma_{Q_1}. \tag{13.19d}$$

## 13.2 Derivation of the Convective Terms

This section details the analytical exposition of Eq. (13.19) following the steps described in reference [BrEtAl84]. Once these terms are obtained, the set of derivative integral equations for the translational displacements are completed. An inspection of Eqs. (13.15) and (13.16) reveals that the candidate terms that give origin to the convective terms are

$$I_i^N = \int_{\Omega} {}^{f}U_{i3,\beta}(Q,P)N_{\alpha\beta}(Q)u_{3,\alpha}(Q)\,d\Omega_Q\ , \tag{13.20a}$$

$$I_i^q = \int_{\Omega} {}^{f}V_{ij}(Q,P)q_j(Q)\,d\Omega_Q\ , \tag{13.20b}$$

$$J_\alpha^N = \int_\Omega {}^m U_{\alpha\beta,\delta}(Q, P) N_{\beta\delta}^n(Q)\, d\Omega_Q \ , \tag{13.20c}$$

$$J_\alpha^q = \int_\Omega {}^m V_{\alpha\beta}(Q, P) q_\beta(Q)\, d\Omega_Q \ , \tag{13.20d}$$

where its derivation with respect to the coordinate axes leads to a general form for Eq. (13.19).

$$\frac{\partial I_i^N}{\partial x_\gamma\,(P)} = {}^f c_{\alpha\beta}^N(P) \tag{13.21a}$$

$$\frac{\partial I_i^q}{\partial x_\gamma\,(P)} = {}^f c_\alpha^q(P) \tag{13.21b}$$

$$\frac{\partial J_\alpha^N}{\partial x_\gamma\,(P)} = {}^m c_{\alpha\beta}^N(P) \tag{13.21c}$$

$$\frac{\partial J_\alpha^q}{\partial x_\gamma\,(P)} = {}^m c_{\alpha\beta}^q(P). \tag{13.21d}$$

In order to keep the notation simpler, for convenience the prefixes $m$ and $f$ will be suppressed in the next paragraphs. In order to recover the complete representation of all expressions, one may consult equation (13.21).

**Evaluation of** $\dfrac{\partial I_i^N}{\partial x_\gamma\,(P)}$

Equation (13.20a) may be expressed as the limit

$$I_i^N = \lim_{\epsilon\to 0} \int_{\Omega-\Omega_\epsilon} U_{i3,\alpha}(Q, P) M_\alpha(Q)\, d\Omega_Q \ , \tag{13.22}$$

where $M_\alpha(Q) = N_{\alpha\beta}(Q) u_{3,\beta}(Q)$ and $\Omega_\epsilon$ is a unit circle centered at the source point $P$. The boundary of $\Omega_\epsilon$ is denoted $\overline{\Gamma}_\epsilon$. Consequently, Eq. (13.20a) may be expressed by

$$\frac{\partial I_i^N}{\partial x_\gamma\,(P)} = \lim_{\epsilon\to 0} \left( \frac{\partial}{\partial x_\gamma} \int_{\Omega-\Omega_\epsilon} U_{i3,\alpha}(Q, P) M_\alpha(Q)\, d\Omega_Q \right) . \tag{13.23}$$

Using a polar coordinate system $(\bar{r}, \bar{\theta})$ with origin at $P \equiv o$ as depicted in Fig. 13.1, $U_{i3,\alpha}$ is rewritten considering only its strongly singular part.

$$U_{i3,\alpha} = \frac{1}{r(\bar{r}, \bar{\theta})} \Lambda_{i3,\alpha}(\phi). \tag{13.24}$$

Figure 13.1a shows the case with $r(\bar{r},\bar{\theta}) = \bar{r}$ and $\phi(\bar{r},\bar{\theta}) = \bar{\theta}$; however, if the source $P$ is perturbed by a Cartesian increment $\Delta x_\alpha$, the parameters $r$ and $\phi$ differ from $\bar{r}$ and $\bar{\theta}$, respectively, and the boundary $\overline{\Gamma}_\epsilon$ changes as well (see Fig. 13.1b). This shows that $\overline{\Gamma}_\epsilon$ is dependent on the load point location, so that for convenience one may cast equation (13.23) in polar coordinate system representation.

$$\frac{\partial I_i^N}{\partial x_\gamma} = \int_0^{2\pi} \lim_{\epsilon\to 0}\left(\frac{\partial}{\partial x_\gamma}\int_{\bar{\epsilon}}^{R(\bar{\theta})} \frac{\Lambda_{i3,_\alpha}(\phi)}{r} M_\alpha(Q)\,\bar{r}\,d\bar{r}\right) d\bar{\theta}. \tag{13.25}$$

One should note that in Eq. (13.25) the integration limits vary with the integration variable, and when this dependence holds, the Leibnitz formula shall be used [SoRe58].

$$\frac{d}{d\alpha}\int_{\phi_1(\alpha)}^{\phi_2(\alpha)} f(x,\alpha)\,dx = \int_{\phi_1(\alpha)}^{\phi_2(\alpha)} \frac{\partial f(x,\alpha)}{\partial\alpha}\,dx - f(\phi_1,\alpha)\frac{d\phi_1}{d\alpha} + f(\phi_2,\alpha)\frac{d\phi_2}{d\alpha}. \tag{13.26}$$

Applying Eq. (13.26) directly to Eq. (13.25) yields

$$\begin{aligned}\frac{\partial}{\partial x_\gamma}\int_{\bar{\epsilon}}^{R(\bar{\theta})} \frac{\Lambda_{i3,_\alpha}(\phi)}{r} M_\alpha(Q)\,\bar{r}\,d\bar{r} &= \int_{\bar{\epsilon}}^{R(\bar{\theta})} \frac{\partial}{\partial x_\gamma}\left(\frac{\Lambda_{i3,_\alpha}(\phi)}{r}\right) M_\alpha(Q)\,\bar{r}\,d\bar{r} \\ &- \frac{\Lambda_{i3,_\alpha}(\phi)}{r(\bar{\epsilon},\bar{\theta})} M_\alpha(P)\,\bar{\epsilon}\frac{d\bar{\epsilon}}{dx_\gamma} + \frac{\Lambda_{i3,_\alpha}(\phi)}{r(R,\bar{\theta})} M_\alpha(P)\,R\frac{dR}{dx_\gamma}\,.\end{aligned} \tag{13.27}$$

Due to the fact that the origin of the coordinate system coincides with the source point $P$ before the imposition of $\Delta x_\alpha$, and it remains there after the application of the increment, only $\bar{\epsilon}$ changes with $x_\alpha$, while $R$ does not. As a consequence, the last term on the right-hand side of Eq. (13.27) vanishes. Moreover, taking into account that $r(\bar{\epsilon},\bar{\theta}) = \epsilon = \bar{\epsilon}$ when $P \equiv o$, one obtains

$$\begin{aligned}\frac{\partial I_i^N}{\partial x_\gamma} &= \int_0^{2\pi} \lim_{\epsilon\to 0}\left[\int_\epsilon^{R(\phi)} \frac{\partial}{\partial x_\gamma}\left(\frac{\Lambda_{i3,_\alpha}(\phi)}{r}\right) M_\alpha(Q)\,r\,dr\right] d\phi \\ &- M_\alpha(P)\int_0^{2\pi} \Lambda_{i3,_\alpha}(\phi)\,\cos(r,x_\gamma)\,d\phi\,.\end{aligned} \tag{13.28}$$

Now it is instructive to investigate the existence of the first integral on the right-hand side of (13.28). Noting that

$$\frac{\partial}{\partial x_\gamma}\left(\frac{\Lambda_{i3,_\alpha}(\phi)}{r}\right) M_\alpha(Q)\,r = r^2\frac{\partial}{\partial x_\gamma}\left(\frac{\Lambda_{i3,_\alpha}(\phi)}{r}\right) M_\alpha(Q)\,\frac{1}{r} \tag{13.29}$$

and defining $\bar{\Lambda}_{i3,\alpha\gamma}(\phi) = r^2 \frac{\partial}{\partial x_\gamma}\left(\frac{\Lambda_{i3,\alpha}(\phi)}{r}\right)$, the term

$$\int_0^{2\pi} \lim_{\epsilon\to 0}\left[\int_\epsilon^R \frac{\partial}{\partial x_\gamma}\left(\frac{\Lambda_{i3,\alpha}(\phi)}{r}\right) M_\alpha(P)\, r\, dr\right] d\phi$$

can be added and subtracted from Eq. (13.29), resulting in

$$\begin{aligned}
&\int_0^{2\pi} \lim_{\epsilon\to 0}\left[\int_\epsilon^R \frac{\partial}{\partial x_\gamma}\left(\frac{\Lambda_{i3,\alpha}(\phi)}{r}\right) M_\alpha(Q)\, r\, dr\right] d\phi \\
&\quad = \int_0^{2\pi} \lim_{\epsilon\to 0}\left\{\bar{\Lambda}_{i3,\alpha\gamma}(\phi)\int_\epsilon^R [M_\alpha(Q) - M_\alpha(P)]\,\frac{1}{r}\, dr\right\} d\phi \\
&+ M_\alpha(P)\int_0^{2\pi} \bar{\Lambda}_{i3,\alpha\gamma}(\phi)\ln(R)\, d\phi - \lim_{\epsilon\to 0}\left[M_\alpha(P)\ln\epsilon \int_0^{2\pi} \bar{\Lambda}_{i3,\alpha\gamma}(\phi) d\phi\right].
\end{aligned} \tag{13.30}$$

All the integrals in Eq. (13.30) are limited, provided that the membrane-bending coupling satisfies the Hölder condition in $P$.

$$\|M_\alpha(Q) - M_\alpha(P)\| \le A r^\alpha \quad , \qquad A, \alpha > 0.$$

Due to the tensor $\bar{\Lambda}_{i3,\alpha m}$ satisfying the property $\int_0^{2\pi} \bar{\Lambda}_{i3,\alpha\gamma}(\phi)\, d\phi = 0$, the last two terms in Eq. (13.30) vanish. In addition, the first integral on the right-hand side is convergent since

$$\lim_{\epsilon\to 0}\left[\bar{\Lambda}_{i3,\alpha m}(\phi)\int_\epsilon^R \frac{A r^\alpha}{r}\, dr\right] = \lim_{\epsilon\to 0}\left[\frac{A\, r^{\alpha+2}}{\alpha - 1}\ln(r)\frac{\partial}{\partial x_\gamma}\left(\frac{\Lambda_{i3,\alpha}}{r}\right)\right]_\epsilon^R < \infty\,,$$

which completes the demonstration.

Now $\partial I_i^N/\partial x_\gamma$ can be transformed back into Cartesian coordinates,

$$\begin{aligned}
\frac{\partial I_i^N}{\partial x_\gamma} &= -\int_\Omega \frac{\partial U_{i3,\alpha}(Q,P)}{\partial x_\gamma} N_{\alpha\beta}(Q)\, u_{3,\beta}(Q)\, d\Omega_Q \\
&\quad - N_{\alpha\beta}(P)\, u_{3,\beta}(P)\int_{\Gamma_1'} U_{i3,\alpha}\, r_{,\gamma}\, d\Gamma'\,,
\end{aligned} \tag{13.31}$$

where the first integral shall be interpreted in terms of the Cauchy principal value (CPV). The second term on the right-hand side of (13.31) is the convective contribution, as it appears from a change in the position of the source point. In the present work, the interest remains in the development of the convective term particularized for $i = 3$ according to Eq. (13.19a).

Since the exterior normal of $\Gamma_1'$ points to the center of the circle $r_{,\alpha} = -n_\alpha$, one can write the convective term as

$$ {}^f\mathbf{c}^N(P) = N_{\alpha\beta}(P) \int_{\Gamma_1'} U^s_{33,\alpha}\, r_{,\gamma}\, d\Gamma' = -N_{\alpha\beta}(P) \int_{\Gamma_1'} U^s_{33,\alpha}\, n_\gamma\, d\Gamma' \,, \tag{13.32} $$

with $U^s_{33,\alpha}$ containing only the singular part of $U_{33,\alpha}$. In the present case,

$$ U^s_{33,\alpha} = \frac{-1}{\pi D(1-\nu)\lambda^2} \frac{r_{,\alpha}}{r}, $$

thus validating the representation (13.24). Using $d\Gamma = r\, d\phi$, then Eq. (13.32) is analytically defined by

$$ {}^f c^N_{\alpha\beta}(P) = \frac{-1}{\pi D(1-\nu)\lambda^2} \left[ \int_{2\pi}^{0} n_\gamma n_\alpha\, d\phi \right] N_{\gamma\beta}(P) \,. $$

Recalling (Fig. 13.1) that $n_1 = -\cos\phi$, $n_2 = -\sin\phi$ and using elementary trigonometric integrals, the following result is obtained.

$$ {}^f c^N_{\alpha\beta}(P) = \frac{-N_{\alpha\beta}(P)}{D(1-\nu)\lambda^2}. \tag{13.33} $$

This non-integral term is added to Eq. (13.18) replacing thus the first integral on $\Gamma_1'$. Note that the correction is necessary only in the singular case ($P \equiv Q$). A comparison to findings in the literature shows that Eq. (13.33) is in agreement with the results obtained by Xiao-Yan et al. [XiEtAl90].

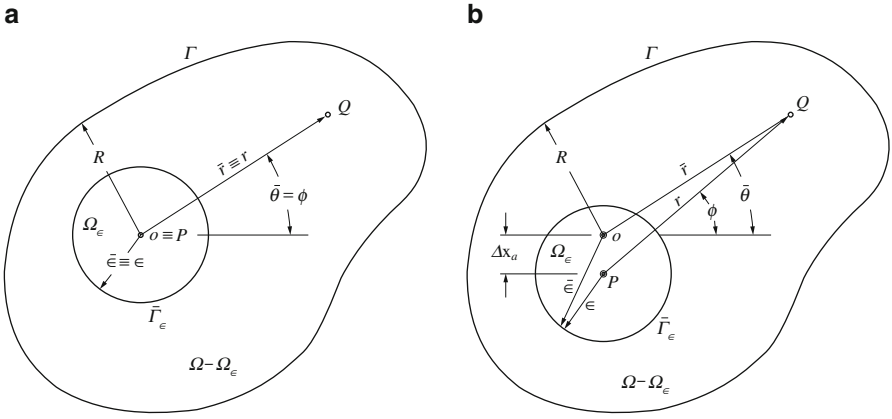



**Fig. 13.1** Definition of the boundary $\overline{\Gamma}_\varepsilon$ around the source point. (**a**) Initial configuration, (**b**) the effect of an increment $\Delta x_\alpha$ applied to the source point coordinates

**Evaluation of** $\dfrac{\partial I_i^q}{\partial x_\gamma\,(P)}$

The fundamental solution tensor used to take into account domain bending loadings in both the Mindlin and the Reissner plate models is given by ([WeBa90])

$$ {}^f\mathbf{V} = {}^f\mathbf{U} - m_f\,{}^f\widetilde{\mathbf{U}} = {}^f\mathbf{U} - m_f \begin{bmatrix} 0\;0\;U_{11,1} + U_{12,2} \\ 0\;0\;U_{21,1} + U_{22,2} \\ 0\;0\;U_{31,1} + U_{32,2} \end{bmatrix} . $$

Following the procedure outlined in the previous section, Eq. (13.20b) is written in terms of a limit

$$ I_i^q = \lim_{\epsilon\to 0} \int_{\Omega-\Omega_\epsilon} V_{ij}(Q,P)\, q_j(Q)\, d\Omega_Q \,, \tag{13.34} $$

so that its derivative results in

$$ \frac{\partial I_i^N}{\partial x_\gamma\,(P)} = \lim_{\epsilon\to 0} \left( \frac{\partial}{\partial x_\gamma} \int_{\Omega-\Omega_\epsilon} U_{ij}(Q,P) q_j(Q)\, d\Omega_Q \right. \tag{13.35} $$

$$ \left. + m_f \frac{\partial}{\partial x_\gamma} \int_{\Omega-\Omega_\epsilon} \widetilde{U}_{ij}(Q,P) q_j(Q)\, d\Omega_Q \right) . \tag{13.36} $$

Now, the treatment has to be carried out for the Reissner model ($m_f = 1$), otherwise ${}^f\mathbf{V} = {}^f\mathbf{U}$, and since $\mathbf{U} = O(\ln r)$, the first integral does not manifest strong singularities after the differentiation and will not provide convective terms. The second integral deserves a more careful inspection. Since the interest is in the derivative of the plate transverse displacement, Eq. (13.35) is particularized, considering from the outset only the necessary terms.

$$ \frac{\partial I_3^N}{\partial x_\gamma\,(P)} = \lim_{\epsilon\to 0} \left( \frac{\partial}{\partial x_\gamma} \int_{\Omega-\Omega_\epsilon} U_{3\alpha,\alpha}(Q,P) q_3(Q)\, d\Omega_Q \right) . \tag{13.37} $$

However, since $U_{3\alpha,\alpha}$ is regular on $\Omega$, it is not possible to apply the representation

$$ U_{3\alpha,\alpha} = \frac{1}{r(\bar{r},\bar{\theta})}\, \Lambda_{3\alpha,\alpha}(\phi), \tag{13.38} $$

and consequently, there is no convective contribution, as expected.

$$ {}^f c^q_{\alpha\beta}(P) = 0. \tag{13.39} $$

**Evaluation of** $\dfrac{\partial J_\alpha^N}{\partial x_\gamma\,(P)}$

In the case of Eq. (13.20c), one may follow the same spirit as lined out in the previous two paragraphs.

$$J_\alpha^N = \lim_{\epsilon\to 0}\int_{\Omega-\Omega_\epsilon} U_{\alpha\beta,\delta}(Q,P)N_{\beta\delta}^n(Q)\,d\Omega_Q. \tag{13.40}$$

The first step is to write the integral as a limit,

$$\frac{\partial J_\alpha^N}{\partial x_\gamma\,(P)} = \lim_{\epsilon\to 0}\left(\frac{\partial}{\partial x_\gamma}\int_{\Omega-\Omega_\epsilon} U_{\alpha\beta,\delta}(Q,P)N_{\beta\delta}^n(Q)\,d\Omega_Q\right), \tag{13.41}$$

and then introducing

$$U_{\alpha\beta,\delta} = \frac{1}{r(\bar r,\bar\theta)}\,\Lambda_{\alpha\beta,\delta}(\phi), \tag{13.42}$$

one arrives at an expression that may be solved by the use of the Leibnitz formula.

$$\frac{\partial J_\alpha^N}{\partial x_\gamma} = -\int_\Omega \frac{U_{\alpha\beta,\delta}}{\partial x_\gamma}N_{\beta\delta}^n(Q)\,d\Omega_Q - N_{\beta\delta}^n(P)\int_{\Gamma_1'} U_{\alpha\beta,\delta}\,r_{,\gamma}\,d\Gamma'. \tag{13.43}$$

Here, the first integral shall again be interpreted in the CPV sense, provided the nonlinear membrane forces satisfy the Hölder condition on $P$.

$$\|N_{\beta\delta}^n(Q) - N_{\beta\delta}^n(P)\| \le Ar^\alpha \quad , \qquad A,\alpha > 0. \tag{13.44}$$

Upon analyzing Eq. (13.43), one identifies the expected convective term,

$${}^m\mathbf{c}^N(P) = N_{\beta\delta}^n(P)\int_{\Gamma_1'} U_{\alpha\beta,\delta}\,r_{,\gamma}\,d\Gamma' = -N_{\beta\delta}^n(P)\int_{\Gamma_1'} U_{\alpha\beta,\delta}\,n_\gamma\,d\Gamma', \tag{13.45}$$

where $U_{\alpha\beta,\delta}$ is $O(r^{-1})$, and consequently, the analytical representation of Eq. (13.45) is

$$\begin{aligned}{}^m\mathbf{c}^N(P) = &\frac{1}{8\pi G\,(1-\nu)}\left\{\int_{2\pi}^0\left[(3-4\nu)\,r_{,\gamma}\,\delta_{\alpha\beta} - r_{,\alpha}\,\delta_{\beta\gamma}\right.\right.\\ &\left.\left. -r_{,\beta}\,\delta_{\alpha\gamma} + 2r_{,\alpha}r_{,\beta}r_{,\gamma}\right]d\phi\right\}N_{\beta\delta}^n(P)\,.\end{aligned} \tag{13.46}$$

Finally, using the relations $n_1 = -\cos\phi$ , $n_2 = -\sin\phi$ and elementary integrals of trigonometric powers leads to the following expression:

$$
\begin{aligned}
{}^{m}c^{N}_{\alpha\beta}(P) =& \frac{-1}{8G(1-\nu)}\Big[(3-4\nu)\,\delta_{\alpha\delta}\delta_{\beta\gamma} - \delta_{\alpha\beta}\delta_{\gamma\delta} - \delta_{\alpha\gamma}\delta_{\beta\delta} \\
&+ \frac{1}{4}\delta_{\alpha\beta}\delta_{\gamma\delta}\left(1+2\delta_{\alpha\gamma}\right)\Big] N^{n}_{\gamma\delta}(P). \qquad (13.47)
\end{aligned}
$$

**Evaluation of** $\dfrac{\partial J^{q}_{\alpha}}{\partial x_{\gamma}\,(P)}$

In this case, ${}^{m}\mathbf{V} = {}^{m}\mathbf{U}$, and since $\mathbf{U} = O\,(\ln r)$, no convective term is involved,

$$
{}^{m}c^{q}_{\alpha\beta}(P) = 0\ .
$$

## 13.3 Summary of the Results

All the relevant expressions obtained in the previous sections can be summarized as follows:

$$
\begin{aligned}
{}^{m}c^{N}_{\alpha\beta}(P) =& \frac{-1}{8G(1-\nu)}\Big[(3-4\nu)\,\delta_{\alpha\delta}\delta_{\beta\gamma} - \delta_{\alpha\beta}\delta_{\gamma\delta} - \delta_{\alpha\gamma}\delta_{\beta\delta} \\
&- \frac{1}{4}\delta_{\alpha\beta}\delta_{\gamma\delta}\left(1+2\delta_{\alpha\gamma}\right)\Big] N^{n}_{\gamma\delta}(P)\ , \qquad (13.48a)
\end{aligned}
$$

$$
{}^{m}c^{q}_{\alpha\beta}(P) = 0\ , \qquad (13.48b)
$$

$$
{}^{f}c^{N}_{\alpha\beta}(P) = -\frac{\delta_{\alpha\gamma}}{D(1-\nu)\lambda^{2}}\, N_{\gamma\beta}(P)\ , \qquad (13.48c)
$$

$$
{}^{f}c^{q}_{\alpha}(P) = 0\ . \qquad (13.48d)
$$

These equations are subject to the conditions

$$
\begin{aligned}
\|M_{\alpha}(Q) - M_{\alpha}(P)\| &\le Ar^{\alpha}\ , \qquad A, \alpha > 0\ ; \\
\|N^{n}_{\beta\delta}(Q) - N^{n}_{\beta\delta}(P)\| &\le Br^{\beta}\ , \qquad B, \beta > 0\ ,
\end{aligned}
$$

so that finally Eqs. (13.17) and (13.18) may be cast in their final form.

$$
\begin{aligned}
{}^{m}u_{\beta,\alpha}(P) - \int_{\Gamma} {}^{m}T_{\beta\gamma,\alpha}(q,P)\,{}^{m}u_{\gamma}(q)\,d\Gamma_q &= -\int_{\Gamma} {}^{m}U_{\beta\gamma,\alpha}(q,P)\,{}^{m}t_{\gamma}(q)\,d\Gamma_q \\
&- \int_{\Omega} {}^{m}V_{\beta\gamma,\alpha}(Q,P)\,{}^{m}q_{\gamma}(Q)\,d\Omega_Q + \int_{\Omega} {}^{m}U_{\beta\gamma,\delta\alpha}(Q,P)N^{n}_{\gamma\delta}(Q)\,d\Omega_Q \\
&+ {}^{m}c^{N}_{\alpha\beta}(P) - {}^{m}v_{\beta,\alpha}(P)
\end{aligned}
\tag{13.49}
$$

$$
\begin{aligned}
{}^{f}u_{3,\alpha}(P) - \int_{\Gamma} {}^{f}T_{3i,\alpha}(q,P)\,{}^{f}u_{i}(q)\,d\Gamma_q &= -\int_{\Gamma} {}^{f}U_{3i,\alpha}(q,P)\,{}^{f}t_{i}(q)\,d\Gamma_q \\
&- \int_{\Omega} {}^{f}V_{3i,\alpha}(Q,P)\,{}^{f}q_{i}(Q)\,d\Omega_Q + \int_{\Omega} {}^{f}U_{33,\alpha\gamma}(Q,P)N_{\gamma\beta}(Q)\,{}^{f}u_{3,\beta}(Q)\,d\Omega_Q \\
&+ {}^{f}c^{N}_{\alpha\beta}(P)u_{3,\beta}(P) - {}^{f}v_{3,\alpha}(P).
\end{aligned}
\tag{13.50}
$$

Note that Eqs. (13.49) and (13.50) are valid for interior points, and consequently, attention shall be paid to the singularities $O\left(1/r^2\right)$ in the integrals on the left-hand side, and $O\left(1/r\right)$ and $O\left(1/r^2\right)$ for the first and third integrals on the right-hand side. For boundary points, their limit to the boundary must be taken in order to obtain the corresponding geometric factors, i.e., the $C$ matrix. In that case, the integrals on the left-hand side must be interpreted in the Hadamard sense, which demonstrates the hyper-singular character of these equations, while all remaining integrals are interpreted employing the CPV.

Moreover, using any traditional collocation-type process ([BrEtAl84]), Eqs. (13.15), (13.16), (13.49), and (13.50) lead to the following set of algebraic equations:

- Membrane (2D elasticity) problem:

$$
{}^{m}\mathbf{H}\,{}^{m}\mathbf{u} = {}^{m}\mathbf{G}\,{}^{m}\mathbf{t} + {}^{m}\mathbf{B} + {}^{m}\mathbf{f}. \tag{13.51}
$$

- Bending problem:

$$
{}^{f}\mathbf{H}\,{}^{f}\mathbf{u} = {}^{f}\mathbf{G}\,{}^{f}\mathbf{t} + {}^{f}\mathbf{B}\,\bar{\mathbf{u}}_3 + {}^{f}\mathbf{f}. \tag{13.52}
$$

- In-plane displacement derivatives:

$$
\mathbf{u}'_{\beta} + {}^{\beta}\mathbf{H}\,{}^{m}\mathbf{u} = {}^{\beta}\mathbf{G}\,{}^{m}\mathbf{t} + {}^{\beta}\mathbf{B} + {}^{\beta}\mathbf{f}. \tag{13.53}
$$

- Transverse displacement derivatives:

$$
\mathbf{u}'_{3} + {}^{3}\mathbf{H}\,{}^{f}\mathbf{u} = {}^{3}\mathbf{G}\,{}^{f}\mathbf{t} + {}^{3}\mathbf{B}\,\mathbf{u}'_{3} + {}^{3}\mathbf{f}, \tag{13.54}
$$

where

$$\mathbf{u}'_{\beta} = \left\{ {}^{m}u_{\beta,1} \, , {}^{m}u_{\beta,2} \right\}^{T} \quad \text{and} \quad \mathbf{u}'_{3} = \left\{ {}^{f}u_{3,1} \, , {}^{f}u_{3,2} \right\}^{T} . \tag{13.55}$$

## 13.4 Conclusions

This chapter presented a compilation of the relevant integral equations for linear and geometrically nonlinear bending, as well as elastic stability of moderately thick plates. The hyper-singular derivative integral equations for the displacement field were presented, including the corresponding convective terms. The resulting integral equations can be used to solve geometrically nonlinear bending problems, as well as in-plane extension, linear bending, and stability problems by particularization. Domain discretization is assumed for the domain integrals whenever necessary.

## References

[Bo47] Bolle, L.: Contribution au probleme lineaire de flexion d'une plaque elastique. Bull. Techn. de la Suisse Romande **21**, 281–285 (1947)

[BrEtAl84] Brebbia, C.A., Telles, J.C.F., Wrobel, L.C.: Boundary Element Techniques: Theory and Applications in Engineering. Springer, Heidelberg (1984)

[Bu78] Bui, H.D.: Some remarks about the formulation of three-dimensional thermoelasto-plastic problems by integral equations. Int. J. Solids Struct. **14**, 935–939 (1978)

[Fu65] Fung, Y.C.: Foundations of Solid Mechanics. Prentice-Hall, Hoboken (1965)

[Ji91] Jianqiao, Y.: Non-linear bending analysis of plates and shells by using a mixed spline boundary element and finite element method. Int. J. Num. Meth. Eng. **31**, 1283–1294 (1991)

[KaSa85] Kamiya, N., Sawaki, Y.: An efficient BEM for some inhomogeneous and nonlinear problems. In: Brebbia, C.A., Maier, G. (eds.) Proceedings of the 7th International Sem. BEM, pp. 13–68. Villa Olmo (1985)

[KaTe88] Karam, V.J., Telles, J.C.F.: On boundary elements for Reissner's plate theory. Eng. Analy. **5**, 21–27 (1988)

[Ra15] Marczak, R.J.: Revisiting some developments of boundary elements for thick plates in Brazil. Lat. Am. J. Solids Struct. **12**, 948–979 (2015)

[Ra98] Marczak, R.J.: A boundary element formulation for linear and non-linear bending of plates. In: Idhelson, S. (ed.) Computational Mechanics - New Trends and Application. International Association for Computational Mechanics (1998)

[Mi51] Mindlin, R.D.: Influence of rotatory inertia and shear on flexural motions of isotropic, elastic plates. J. Appl. Mech. **18**, 31–38 (1951)

[Mi62] Mikhlin, S.G.: Singular integral equations. Amer. Math. Soc. Transl., Ser. 1 **10**, 84–198 (1962).

[RaEtAl97] Rashed, Y.F., Aliabadi, M.H., Brebbia, C.A.: On the evaluation of the stresses in the BEM for Reissner plate-bending problems. Appl. Math. Modelling **21**, 155–163 (1997)

[Re44] Reissner, E.: On the theory of bending of elastic plates. J. Math. Phys. **23**, 184–191 (1944)

[SoRe58] Sokolnikoff, I.S., Redheffer, R.M.: Mathematical of Physics & Modern Engineering. McGraw-Hill, New York (1958)

[SuEtAl94] Sun, Y.B., He, X.Q., Qin, Q.H.: A new procedure for the nonlinear analysis of Reissner plate by boundary element method. Comput. Struct. **53**(3), 649–652 (1994)

[Vi90] Vilmann, O.: The boundary element method applied in mindlin plate bending analysis. PhD Thesis, Department of Structural Engineering, Technical University of Denmark, Denmark (1990)

[We82] van der Weeen, F.: Application of the boundary integral equation method to Reissner's plate model. Int. J. Num. Meth. Eng. **18**, 1–10 (1982)

[WeBa90] Westphal Jr., T., de Barcellos, C.S.: Applications of the boundary element method to Reissner's and Mindlin's plate models. In: Tanaka, M., Brebbia, C.A., Honma, T. (eds.) Proceedings of the 12th International Conference on BEM, vol. 1, pp. 467–477. Sapporo, Japan (1990)

[XiEtAl90] Xiao-Yan, L., Mao-Kwang, H., Xiuxi, W.: Geometrically nonlinear analysis of a Reissner type plate by the boundary element method. Comput. Struct. **37**, 911–916 (1990)

[XiQui93] Xiao-Qiao, H., Qing-Hua, Q.: Nonlinear analysis of Reissner's plate by variational approaches and boundary element methods. Appl. Math. Modelling **17**, 149–155 (1993)

# Chapter 14
# On Viscous Fluid Flow in Curvilinear Coordinate Systems


**A. Meneghetti, B. E. J. Bodmann, and M. T. M. B. Vilhena**

## 14.1 Introduction

Many theoretical problems and their applications such as fluid dynamics scenarios are formulated in a specific coordinate system, which is frequently the Cartesian coordinate system. However, depending on the topography of the physical domain, more specifically, the geometry of the domain boundaries, another choice might be advantageous. The question as to what is the most adequate system is one of the principal issues in the Theory of General Relativity and is based on the mathematical framework of differential geometry. While this theory relates the geometry of space-time with its energy–momentum content [We72], some of the ideas of curved space may be exported to other realms as for instance mechanical engineering. In engineering and especially fluid mechanics, we can use a similar methodology, but instead of using geodesics representing the geometric properties of a physical system, one may define the curvilinear coordinate system from the geometry of the physical domain and its boundaries.

If a coordinate axis is interpreted as a solution of a geodesic equation, then the affine connection introduces terms due to curvature in the curvilinear coordinate system. In other words, we locally shear, twist, stretch, or compress the domain such that the boundaries have simple geometries, which constitutes the principal difference to the conceptions of general relativity. As a consequence, the differential operators of the dynamical equations are changed by the addition of new terms, see for instance reference [So64]. The apparent disadvantage of obtaining larger

A. Meneghetti (✉)
Federal University of Rio Grande, Rio Grande, RS, Brazil
e-mail: andremeneghetti@furg.br

B. E. J. Bodmann · M. T. M. B. Vilhena
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: bardo.bodmann@ufrgs.br



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_14

equations after the coordinate transformation is effectively compensated by simpler, in our work plane parallel boundaries. Evidently, one has to accept some restrictions with respect to shapes that define the orography, imposed by the diffeomorph conformal character of the transformation. Conditions that are mandatory will be defined in order to establish a unique relation between the original and new coordinate systems. Thus, the novelty of the present work is the procedure to solve the curvilinear Navier–Stokes equation, which is then transformed back into the original Cartesian coordinate system. Three-dimensional case studies are presented as numerical implementations of this methodology.

## 14.2 Transformation of the Coordinate System

As mentioned before, the curvilinear boundary guides the construction of the coordinate transformation that simplifies the boundaries of the transformed problem. Let $x^1 \times x^2 \times x^3$ be the Cartesian coordinate system and $\xi^1 \times \xi^2 \times \xi^3$ a generalised coordinate system; then we define the general transformation $T$ by $T : \xi^\alpha = \xi^\alpha(\{x^i\})$. Here, $i, \alpha \in \{1, 2, 3\}$, $\{x^i\} = \{x^1, x^2, x^3\} \in \Omega \subset \mathbb{R}^3$, where $\Omega$ is an open set. According to [MeEtAl17], $T$ is a diffeomorph conformal transformation if and only if the functions $\xi^\alpha$ are of class $C^1(\Omega)$ and $|J| = \left|\frac{\partial \xi^\alpha}{\partial x^i}\right| \neq 0$ in $\Omega$, where $|J|$ is the Jacobian determinant of the transformation $T$. Thus, we guarantee that the solution of the transformed problem obtained in the new coordinate system can in the end be inverted and presented in the original coordinate system. In addition to its mathematical characteristics, this property ensures that conservation laws are preserved, as shown in [We72]. As already mentioned in the introduction, we use the curvilinear contours of the domain in which the problem is structured to define the new coordinate system, more precisely the $\xi^1$, $\xi^2$, and $\xi^3$ system, which by virtue is a curvilinear coordinate system.

## 14.3 The Transformed Navier–Stokes Equation

We consider an incompressible model and define as a starting point the dimensionless Navier–Stokes equation (14.1),which together with the continuity equation is given in Cartesian coordinates (see for instance reference [ScGe17]). In order to turn the developments more compact, in the further the Einstein summation convention is understood in all equations that follow.

$$\frac{\partial u^m}{\partial t} + u^i \frac{\partial u^m}{\partial x^i} = -\frac{\partial P}{\partial x^m} + \frac{\delta_i^j}{Re} \frac{\partial^2 u^m}{\partial x^i \partial x^j} \qquad \frac{\partial u^i}{\partial x^i} = 0. \tag{14.1}$$

Here, $i, j, m \in \{1, 2, 3\}$, $u^m$ is the velocity in the $x^m$ direction, $P$ is the pressure, $Re$ is the Reynolds number, and the Kronecker symbol $\delta_i^j = 1$ for $i = j$ and zero otherwise. As a convenient simplification, we neglect effects due to the force of gravity.

Using algebraic manipulations in the Eq. (14.1), it is possible to obtain the Poisson equation for pressure (14.2), where $D = \frac{\partial u^i}{\partial x^i}$.

$$\delta_i^j \frac{\partial^2 P}{\partial x^i \partial x^j} = -\frac{\partial^2 (u^i u^j)}{\partial x^i \partial x^j} + \frac{\delta_i^j}{Re} \frac{\partial^2 D}{\partial x^i \partial x^j} - \frac{\partial D}{\partial t}. \tag{14.2}$$

Recalling that Eq. (14.1) is represented in a Cartesian coordinate system, one may now use a coordinate transformation and obtain the transformed equation in the generalised coordinate system. The variables of both coordinate systems are made compatible through the affine connection (the manifestation of curvature), which as a consequence of coordinate changes enter in the dynamical equations. More precisely, differential operators are altered by the addition of new terms. The transformed Navier–Stokes equations as well as the transformed Poisson equation for the pressure are presented by Eqs. (14.3) and (14.4), where $D = \frac{\partial u^i}{\partial \xi^\alpha} \frac{\partial \xi^\alpha}{\partial x^i}$.

$$\frac{\partial u^m}{\partial t} + u^i \frac{\partial u^m}{\partial \xi^\alpha} \frac{\partial \xi^\alpha}{\partial x^i} = -\frac{\partial P}{\partial \xi^\alpha} \frac{\partial \xi^\alpha}{\partial x^m} + \frac{\delta_i^j}{Re} \left( \frac{\partial u^m}{\partial \xi^\alpha} \frac{\partial^2 \xi^\alpha}{\partial x^i \partial x^j} + \frac{\partial^2 u^m}{\partial \xi^\alpha \partial \xi^\beta} \frac{\partial \xi^\alpha}{\partial x^i} \frac{\partial \xi^\beta}{\partial x^j} \right) \tag{14.3}$$

$$\begin{aligned} \delta_i^j \left( \frac{\partial P}{\partial \xi^\alpha} \frac{\partial^2 \xi^\alpha}{\partial x^i \partial x^j} \frac{\partial^2 P}{\partial \xi^\alpha \partial \xi^\beta} \frac{\partial \xi^\alpha}{\partial x^i} \frac{\partial \xi^\beta}{\partial x^j} \right) &= -\frac{\partial (u^i u^j)}{\partial \xi^\alpha} \frac{\partial^2 \xi^\alpha}{\partial x^i \partial x^j} + \frac{\partial^2 (u^i u^j)}{\partial \xi^\alpha \partial \xi^\beta} \frac{\partial \xi^\alpha}{\partial x^i} \frac{\partial \xi^\beta}{\partial x^j} \\ &+ \frac{1}{Re} \left( \delta_i^j \frac{\partial (D)}{\partial \xi^\alpha} \frac{\partial^2 \xi^\alpha}{\partial x^i \partial x^j} + \delta_i^j \frac{\partial^2 (D)}{\partial \xi^\alpha \partial \xi^\beta} \frac{\partial \xi^\alpha}{\partial x^i} \frac{\partial \xi^\beta}{\partial x^j} \right) - \frac{\partial D}{\partial t}. \end{aligned} \tag{14.4}$$

## 14.4 Numerical Solution

In the further, the sequence of steps to obtain a numerical solution of Eqs. (14.3) and (14.4) is lined out. For convenience, we regroup terms in Eq. (14.3) and end up with Eq. (14.5), with the new abbreviations defined in (14.6).

$$\frac{\partial u^m}{\partial t} = D_\alpha \frac{\partial u^m}{\partial \xi^\alpha} + D_{3+\alpha} \frac{\partial^2 u^m}{\partial (\xi^\alpha) \partial (\xi^\alpha)} + D_{4+\alpha+\beta} \frac{\partial^2 u^m}{\partial \xi^\alpha \partial \xi^\beta} + \phi^m. \tag{14.5}$$

$$D_\alpha = u^i \frac{\partial \xi^\alpha}{\partial x^i} + \frac{1}{Re} \frac{\partial^2 \xi^\alpha}{\partial x^i \partial x^i} \qquad D_{3+\alpha} = \frac{1}{Re} \left(\frac{\partial \xi^\alpha}{\partial x^i}\right)\left(\frac{\partial \xi^\alpha}{\partial x^i}\right)$$

$$D_{4+\alpha+\beta} = \left(\frac{|\epsilon_{\gamma\alpha\beta}|}{2}\right)\left(\frac{2}{Re}\right) \frac{\partial \xi^\alpha}{\partial x^i} \frac{\partial \xi^\beta}{\partial x^i} \qquad \phi^m = -\frac{\partial P}{\partial \xi^\alpha} \frac{\partial \xi^\alpha}{\partial x^m}. \quad (14.6)$$

Equation (14.5) is nonlinear since the expressions $D$ depend on velocities, so that the equation needs to be approximated by an iterative scheme. To this end, we use the implicit finite difference method in Eq. (14.5), see [Ho01] and [Bo15], and end up with the approximate equation (14.7).

$$\begin{aligned}
&\frac{(u^m)^{n+1}_{i,j,l}(u^m)^n_{i,j,l}}{\Delta t} = \\
&(D_1)^{n+1}_{i,j,l} \frac{(u^m)^{n+1}_{i+1,j,l} - (u^m)^{n+1}_{i-1,j,l}}{2\Delta\xi^1} + (D_2)^{n+1}_{i,j,l} \frac{(u^m)^{n+1}_{i,j+1,l} - (u^m)^{n+1}_{i,j-1,l}}{2\Delta\xi^2} \\
&+(D_3)^{n+1}_{i,j,l} \frac{(u^m)^{n+1}_{i,j,l+1} - (u^m)^{n+1}_{i,j,l-1}}{2\Delta\xi^3} + (D_4)^{n+1}_{i,j,l} \frac{(u^m)^{n+1}_{i+1,j,l} - 2(u^m)^{n+1}_{i,j,l} + (u^m)^{n+1}_{i-1,j,l}}{\Delta(\xi^1)^2} \\
&+(D_5)^{n+1}_{i,j,l} \frac{(u^m)^{n+1}_{i,j+1,l} - 2(u^m)^{n+1}_{i,j,l} + (u^m)^{n+1}_{i,j-1,l}}{\Delta(\xi^2)^2} \\
&+(D_6)^{n+1}_{i,j,l} \frac{(u^m)^{n+1}_{i,j,l+1} - 2(u^m)^{n+1}_{i,j,l} + (u^m)^{n+1}_{i,j,l-1}}{\Delta(\xi^3)^2} \qquad (14.7) \\
&+(D_7)^{n+1}_{i,j,l} \frac{(u^m)^{n+1}_{i+1,j+1,l} - (u^m)^{n+1}_{i-1,j+1,l} - (u^m)^{n+1}_{i+1,j-1,l} + (u^m)^{n+1}_{i-1,j-1,l}}{4\Delta\xi^1\Delta\xi^2} \\
&+(D_8)^{n+1}_{i,j,l} \frac{(u^m)^{n+1}_{i+1,j,l+1} - (u^m)^{n+1}_{i-1,j,l+1} - (u^m)^{n+1}_{i+1,j,l-1} + (u^m)^{n+1}_{i-1,j,l-1}}{4\Delta\xi^1\Delta\xi^3} \\
&+(D_9)^{n+1}_{i,j,l} \frac{(u^m)^{n+1}_{i,j+1,l+1} - (u^m)^{n+1}_{i,j-1,l+1} - (u^m)^{n+1}_{i,j+1,l-1} + (u^m)^{n+1}_{i,j-1,l-1}}{4\Delta\xi^2\Delta\xi^3} + (\phi^m)^{n+1}_{i,j,l}.
\end{aligned}$$

Upon regrouping terms $u^m$ with the corresponding space and time indices ($i$, $j$, $k$ and $n$), one obtains Eq. (14.8),

$$\begin{aligned}
&\left(E^{000}\right)^{n+1}_{i,j,l} (u^m)^{n+1}_{i,j,l} + \left(E^{100}\right)^{n+1}_{i,j,l} (u^m)^{n+1}_{i+1,j,l} + \left(E^{-100}\right)^{n+1}_{i,j,l} (u^m)^{n+1}_{i-1,j,l} \\
&+ \left(E^{010}\right)^{n+1}_{i,j,l} (u^m)^{n+1}_{i,j+1,l} + \left(E^{0-10}\right)^{n+1}_{i,j,l} (u^m)^{n+1}_{i,j-1,l} + \left(E^{001}\right)^{n+1}_{i,j,l} (u^m)^{n+1}_{i,j,l+1} \\
&+ \left(E^{00-1}\right)^{n+1}_{i,j,l} (u^m)^{n+1}_{i,j,l-1} + \left(E^{110}\right)^{n+1}_{i,j,l} (u^m)^{n+1}_{i+1,j+1,l} \\
&+ \left(E^{-110}\right)^{n+1}_{i,j,l} (u^m)^{n+1}_{i-1,j+1,l} + \left(E^{1-10}\right)^{n+1}_{i,j,l} (u^m)^{n+1}_{i+1,j-1,l}
\end{aligned}$$

$$
\begin{aligned}
&+\left(E^{-1-10}\right)_{i,j,l}^{n+1}(u^m)_{i-1,j-1,l}^{n+1}+\left(E^{101}\right)_{i,j,l}^{n+1}(u^m)_{i+1,j,l+1}^{n+1} \\
&+\left(E^{-101}\right)_{i,j,l}^{n+1}(u^m)_{i-1,j,l+1}^{n+1}+\left(E^{10-1}\right)_{i,j,l}^{n+1}(u^m)_{i+1,j,l-1}^{n+1} \\
&+\left(E^{-10-1}\right)_{i,j,l}^{n+1}(u^m)_{i-1,j,l-1}^{n+1}+\left(E^{011}\right)_{i,j,l}^{n+1}(u^m)_{i,j+1,l+1}^{n+1} \\
&+\left(E^{0-11}\right)_{i,j,l}^{n+1}(u^m)_{i,j-1,l+1}^{n+1}+\left(E^{01-1}\right)_{i,j,l}^{n+1}(u^m)_{i,j+1,l-1}^{n+1} \\
&+\left(E^{0-1-1}\right)_{i,j,l}^{n+1}(u^m)_{i,j-1,l-1}^{n+1}=(u^m)_{i,j,l}^{n}+(\phi^m)_{i,j,l}^{n}\,,
\end{aligned}
\tag{14.8}
$$

which may be cast in a matrix equation (14.9).

$$
\mathbf{E}^{n+1}\mathbf{u}^{n+1}=\mathbf{u}^{n}-\mathbf{u}_c^{n+1}+\boldsymbol{\phi}^{n}. \tag{14.9}
$$

Note that the vectors $\mathbf{u}^n$ and $\boldsymbol{\phi}^n$ are known because they are evaluated in the previous time step $n$. However, the matrix $\mathbf{E}^{n+1}$ and the vector $\mathbf{u}_c^{n+1}$ depend on $\mathbf{u}^{n+1}$ in the time step $n+1$ to be determined, and the vector $\mathbf{u}_c^{n+1}$ is constructed using the values of the nodes that belong to the domain boundary. In order to work around the problem with the nonlinearity and the contours, an approximation by iteration is employed. At each time step $n+1$, we construct $\bar{\mathbf{u}}^{\bar{n}}$, which yields an estimate value for the velocity component advanced in time by iteration $\bar{\mathbf{u}}^{\bar{n}} \rightarrow \mathbf{u}^{n+1}$, and moreover, Eq. (14.9) is approximated by (14.10).

$$
\bar{\mathbf{E}}^{\bar{n}}\mathbf{u}^{n+1}=\mathbf{u}^{n}-\bar{\mathbf{u}}_c^{\bar{n}}+\boldsymbol{\phi}^{n}. \tag{14.10}
$$

For each time stamp $n+1$, the approximation starts with $\bar{n}=0$ and $\bar{\mathbf{u}}^{\bar{n}}$ is set to $\mathbf{u}^n$. Equation (14.10) is solved, and one obtains the preliminary values for the vector $\mathbf{u}^{n+1}$. In the next iteration, for $\bar{n}=1$, one assumes that $\bar{\mathbf{u}}^{\bar{n}}=\mathbf{u}^{n+1}$, and Eq. (14.10) is again solved and a new estimated value for vector $\mathbf{u}^{n+1}$ is calculated. Note that in each iteration both the $\bar{\mathbf{u}}_c^{\bar{n}}$ vector and the $\bar{\mathbf{E}}^{\bar{n}}$ matrix are updated. The process is repeated until the vectors $\bar{\mathbf{u}}^{\bar{n}}$ and $\mathbf{u}^{n+1}$ "converge" according to a pre-established stopping criterion. In each step, the matrix equation (14.10) is solved using the standard Gauss–Seidel method ([Ho01, Bo15]).

Moreover, for each iteration in $\bar{n}$, the vector $\mathbf{u}_\mathbf{c}^{\mathbf{N}}$ is updated too, which is built by nodes on the domain boundary and is defined by the parametrised surfaces. In case that no variations occur on the surface, then the nodes are kept without updates. For those that vary, except in the main direction (the direction of the inlet flow), the update is done using the neighbouring node. In the main direction, in this work defined along $\xi^1$, the update follows Eq. (14.8). As an inconvenience, there appear nodes outside the mesh, so that this shortcoming has to be corrected by the use of homogeneous boundary conditions $\frac{\partial u^m}{\partial \xi^1}=0$.

$$\frac{u_{i+1}^m - u_i^m}{\Delta \xi^1} \approx \frac{\partial u^m}{\partial \xi^1} = 0 \Rightarrow u_{i+1}^m \approx u_i^m.$$

The Poisson equation for the pressure (14.4) was solved proceeding in an analogue way. More specifically, the equation was initially approximated using the implicit finite difference method, and the discretised equation was rearranged and cast in the matrix form. Due to the Poisson equation being linear, the iteration defined at each time step will be used only to approximate the vector formed by the nodes belonging to the contours. As already mentioned, the transformation $T$ is defined from the geometry of the curvilinear domain, so that in order to solve the equations by the finite difference method, the transformation is setup by the meshes constructed in the curvilinear domain (see ref. [MeEtAl17]).

## 14.5 Numerical Simulations

In this section, two simulations are presented using the dimensionless Navier–Stokes equations in its two-dimensional and three-dimensional form.

**Simulation 1**
A two-dimensional duct has the top and bottom described by Eq. (14.11).

$$\begin{aligned} f_s(x^1) &= 1 - 0.125\left(\tanh\left(8(x^1 - 6.875)\right) - \tanh\left(8(x^1 - 8.125)\right)\right) \\ f_i(x^1) &= -0.249\left(\tanh\left(8(x^1 - 3.75)\right) - \tanh\left(8(x^1 - 5)\right)\right). \end{aligned} \tag{14.11}$$

Now, a fluid flow inside this duct is considered, with inlet horizontally from the left to the right, with Reynolds number $Re = 100$ and in a domain $(x^1, x^2) \in [0, 10] \otimes [0, 1]$. Thus the flow is modelled by the dimensionless transformed Navier–Stokes equations (14.3) and (14.4), adapted to the two-dimensional case, subject to the following initial and boundary conditions:

- The initial conditions are given by a horizontal flux $u^1 = 1, u^2 = 0$ and a constant pressure $P = 1$ in the domain.
- The boundary conditions for the inlet in the domain are a horizontal flux: $u^1 = 8\xi^2(1-\xi^2)$, $u^2 = 0$ and a vanishing pressure gradient $\frac{\partial P}{\partial \xi^1} = 0$, i.e., a mechanical equilibrium.
- The exit of the domain to the right is given by vanishing velocity field gradients $\frac{\partial u^1}{\partial \xi^1} = 0$, $\frac{\partial u^2}{\partial \xi^1} = 0$ and a prescribed pressure $P = 1$.
- On the top and bottom boundary, no-slip conditions are assumed ($u^1 = 0$ and $u^2 = 0$) together with a vanishing pressure gradient $\frac{\partial P}{\partial \xi^2} = 0$.

In the following, some results obtained in this simulation are shown, where the velocity and the pressure fields are presented in the original Cartesian coordinate

system, i.e., after the inverse coordinate transformation. In Figs. 14.1, 14.2, 14.3 and 14.4, the horizontal axis ($x^1$) corresponds to the initial direction of the flow, while the vertical axis ($x^2$) corresponds to the cross flow. Figure 14.1 shows the local speed in the domain $(x^1, x^2) \in [0, 10] \otimes [0, 1]$, while for the locations $x^1 \in \{0.3, 2.0, 4.0, 6.0, 8.4\}$, the two-dimensional vector field is given. Due to the no-slip condition, the velocity close to the top and bottom boundaries approaches zero, whereas in the centre of the two-dimensional duct, the velocity assumes larger values. Further, as was to be expected, the velocity of the fluid flow is largest at the narrowing of the vertical dimension of the domain around $x^1 = 7.5$.

The details according to the change in direction of the fluid flow and its associated velocity field close to the widening and narrowing regions in the vertical direction

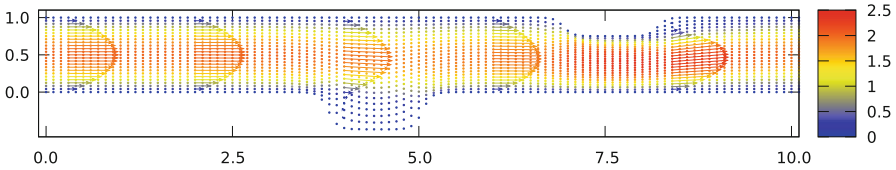



**Fig. 14.1** Local speed of the fluid flow ($\sqrt{(u^1)^2 + (u^2)^2}$) in the Cartesian coordinate system. The two-dimensional velocity vector field is shown for the coordinates $x^1 = 0.3$, 2.0, 4.0, 6.0, and 8.4

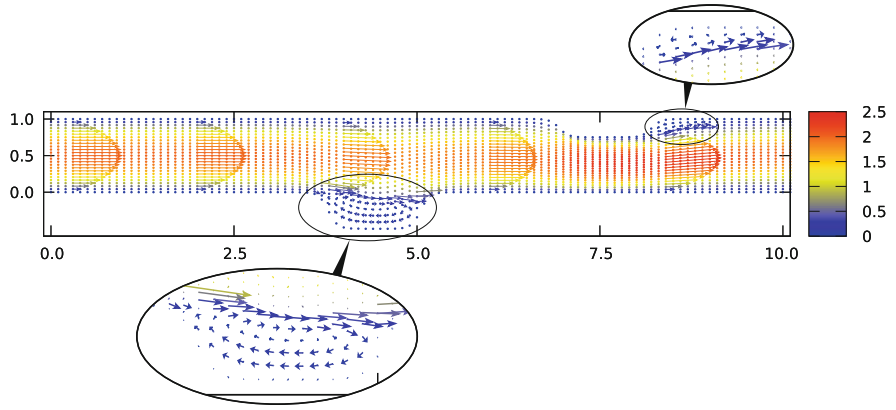



**Fig. 14.2** Zoom images showing details of Fig. 14.1 in the neighbourhood of the concavities at the bottom and top boundaries, respectively. The colours follow the same scale as in Fig. 14.1

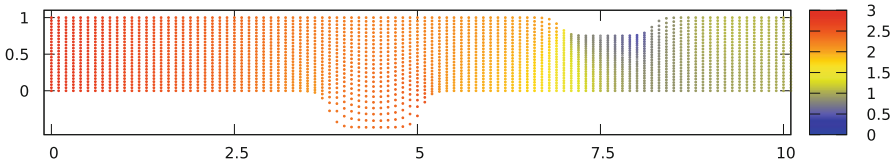



**Fig. 14.3** Local pressure distribution $P$ in the domain

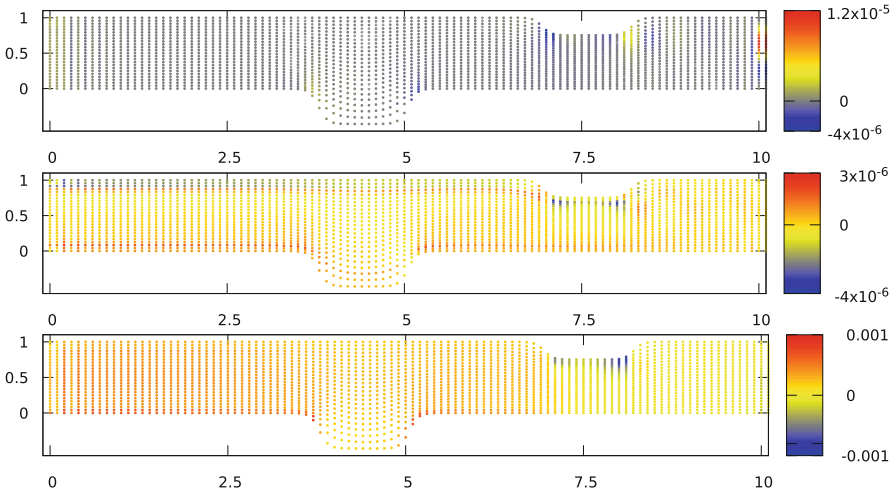



**Fig. 14.4** Residuals of the obtained numerical solution for $u^1$ (top), $u^2$ (centre) and for $P$ (bottom)

are shown in Fig. 14.2. Considering the speed $u = \sqrt{(u^1)^2 + (u^2)^2}$, the regions of the flow field with $u \gtrsim 1$ follow the horizontal direction, i.e., the direction of the incoming fluid flow. Close to the boundaries, one observes the effect of the no-slip condition. Inside the indentation and shortly after the bulge, one observes an inversion of the flow direction, i.e., a counter flow, which is physically expected for flows that pass around obstacles. It is noteworthy that the velocity of the fluid flow almost doubles below the bulge located on the top of the boundary.

Figure 14.2 displays some details of the flow velocity vector field in the vicinity of the concavities at the bottom and the top, respectively.

The pressure distribution, shown in Fig. 14.3, follows the expected behaviour, where globally there is a decreasing tendency from upstream to downstream. In the region below the concavity on the top boundary, one observes a pressure drop caused by the increase of the flow velocity. This may also be verified in Fig. 14.2, where the effect of the concavity on the velocity field becomes apparent. Close to the curved boundary, the no-slip condition dominates, but there is a pronounced increase of the velocity in the vertical direction, where an increase between 2.0 and 2.5 in comparison to the inlet velocity was found.

By inspection of Figs. 14.1, 14.2 and 14.3, one verifies that the discussed numerical solutions obtained by the presented methodology are qualitatively in agreement with physical intuition. In the further, we employ a second argument to validate the quality of our findings. To this end, the solutions for the velocity field and the pressure field are inserted back into Eqs. (14.3) and (14.4), where in each equation the absolute difference between the left-hand side and the right-hand side is used as a measure for the quality of the obtained numerical solution and in the further referred to as residual. These differences are shown in Fig. 14.4.

Generally, the largest differences of the left- and right-hand sides occur in the region of the concavities and at the end of the domain. The order of magnitude of the components of the velocity fields is considerably small (of the order $10^{-6}$), so that the numerical velocity field may be considered a good approximation for the true solution. The slightly larger differences at the end of the domain are a result of the imposed condition by the pressure, where the points outside the domain have to be estimated for the numerical calculations. This does not occur on the left boundary where the pressure gradient vanishes, and thus the same pressure value may be used for points outside the domain in the numerical evaluation of the solution.

Although the order of magnitude of the differences in the vertical velocity component is considerably small, also the numerical values of $u^2$ are small. This is the reason why the distribution of the difference between the left- and right-hand sides of the Navier–Stokes equation shows a less homogeneous pattern for the cross flow velocity component. In the first half of the domain $[0, \sim 5]$ on the top boundary, there is an oscillation of the difference from positive values close to the boundary to negative ones, while the central part is reasonably homogeneous. At the bottom, there is only a negative difference close to the boundary. In the second half of the domain, the larger differences are still closer to the boundary but less pronounced. This asymmetric upstream downstream behaviour depends on where the first concavity is located, on the top or bottom boundary. Since the numerical implementation is symmetrical with respect to the top and bottom boundaries, an inversion of the $x^2$ coordinate (collocating the concavity on the bottom to the upper location and the one on the top to the lower boundary) inverts also the distribution of the differences. The upstream downstream asymmetry is also visible in the differences for the pressure equation. The upstream half of the domain has larger negative differences on the lower boundary, while in the downstream located half of the domain the values are also negative and of the order of $10^{-3}$. In the latter half of the domain, the differences tend towards numerical values closer to zero. Although the presented validation is from the mathematical point of view a necessary but not sufficient condition for convergence, the discussion of the obtained results indicates that the found solutions for the velocity and the pressure field are acceptable because from the physical point of view they look sound.

**Simulation 2**

In this simulation, a domain consisting of $x^1 \in [0, 10]$, $x^2 \in [0, 1]$ and the curvilinear boundaries on the top and bottom with $f_s$ and $f_i$ are considered and are given by Eq. (14.12). The sketch of the domain is shown in Fig. 14.5.

$$\begin{aligned} x^3 &= f_s(x^1, x^2) = 1 \\ x^3 &= f_i(x^1, x^2) = 0.1\, g(x^1)\, h(x^2). \end{aligned} \tag{14.12}$$

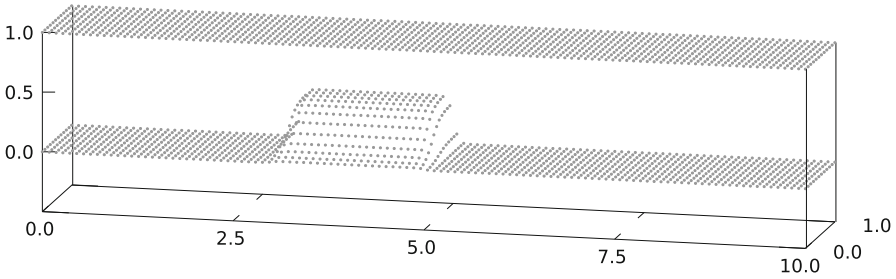



**Fig. 14.5** Illustration of the domain with curvilinear bottom boundary

Here,

$$g(x^1) = \tanh\Big(15(x^1 - 3)\Big) - \tanh\Big(15(x^1 - 5.0)\Big) \ ,$$
$$h(x^2) = \tanh\Big(4(x^2 - 0.4)\Big) - \tanh\Big(4(x^2 - 2)\Big) \ .$$

We assume again that the incoming fluid flow is aligned with the direction of the positive $x^1$ axis. In this simulation, the fluid obeys the three-dimensional Navier–Stokes and Poisson equations with Reynolds number $Re = 100$. Further, we assume the following initial and boundary conditions:

- Horizontal incoming flow with $u^1 = 1$, $u^2 = 0$, and $u^3 = 0$, and homogeneous pressure $P = 1$.
- At the upstream domain surface with $x^1 = 0$, a parabolic velocity profile $u^1 = 8 \times x^3(1 - x^3)$ is understood in agreement with the no-slip condition on the surfaces and with normal vectors perpendicular to the incoming flow direction. Further, no cross fluxes $u^2 = 0$, $u^3 = 0$ and pressure gradients $\frac{\partial P}{\partial \xi^1} = 0$ are assumed.
- At the downstream surface with $x^1 = 10$, vanishing velocity gradients for all components $\frac{\partial u^1}{\partial \xi^1} = 0$, $\frac{\partial u^2}{\partial \xi^1} = 0$ and $\frac{\partial u^3}{\partial \xi^1} = 0$ are defined together with a prescribed pressure $P = 1$.
- At the front and back surfaces with $x^2 = 0$ and $x^2 = 1$, respectively, no velocity gradients for all components $\frac{\partial u^1}{\partial \xi^2} = 0$, $\frac{\partial u^2}{\partial \xi^2} = 0$, $\frac{\partial u^3}{\partial \xi^2} = 0$ and null pressure gradients $\frac{\partial P}{\partial \xi^2} = 0$ are considered.
- At the top and bottom boundaries with $x^3 = f_s$ and $x^3 = f_i$, respectively, the no-slip condition $u^1 = 0$, $u^2 = 0$, $u^3 = 0$ applies, and a vanishing pressure gradient $\frac{\partial P}{\partial \xi^3} = 0$ normal to the surfaces is understood.

From the coordinate transformation, a mesh structure with 111 partitions along the $x^1$ axis ($0 \leq i \leq 111$), 11 partitions along the $x^2$ axis ($0 \leq j \leq 11$), and 24 partitions along the $x^3$ axis ($0 \leq k \leq 24$) was constructed, totalling in $33{,}600 = 112 \times 12 \times 25$ nodes defined in the domain. Figure 14.6 shows a slice in the domain, parallel to the plane $x^1 \times x^3$ and with $j = 11$, that is, the set of

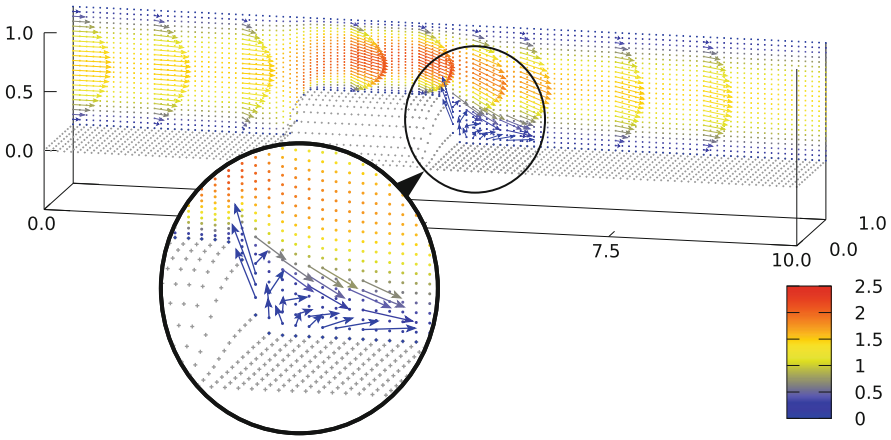



**Fig. 14.6** Vertically parabolic shaped vector velocity field for a slice at $j = 11$ and its deformations around the maximum concavity

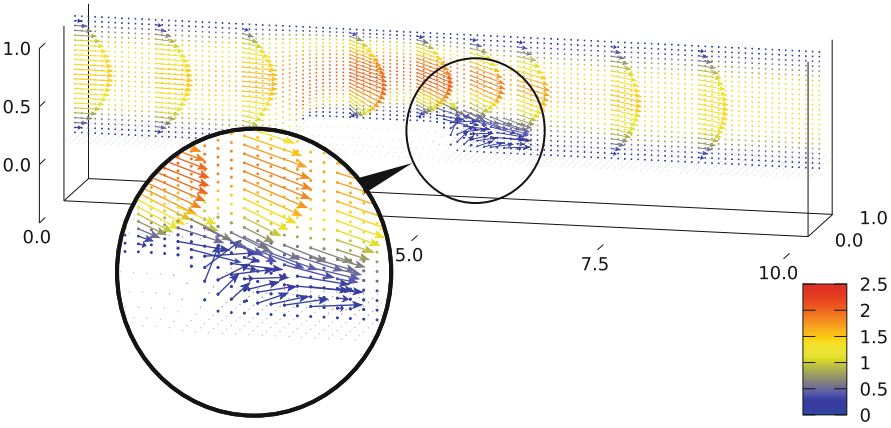



**Fig. 14.7** Vertically parabolic shaped vector velocity field for a slice at $j = 5$ and its deformations around the less pronounced concavity

discrete points identified by the indices $(i, 11, k)$, where $i \in \{0, 1, \ldots, 111\}$ and $k \in \{0, 1, \ldots, 24\}$ for which the vector velocity field $\mathbf{u} = (u^1, u^2, u^3)$ is shown, and the colours indicate the speed of the flow $u = \|\mathbf{u}\| = \sqrt{(u^1)^2 + (u^2)^2 + (u^3)^2}$. As expected, in the areas of the narrowing, the speed increases and the velocity field conforms to the curvilinear domain and the horizontal movement of the fluid. As already shown in the two-dimensional case, also in this case, one observes the formation of vortices immediately after the bump.

The three-dimensional bump is irregular and decreases in the direction from $y = 1$ to $y = 0$. Figure 14.6 shows the slice at $j = 11$ where the bump has its largest extension, and one consequence is the formation of a well-defined vortex, while Fig. 14.7 shows the slice at $j = 5$ where the bump is less salient. This implies

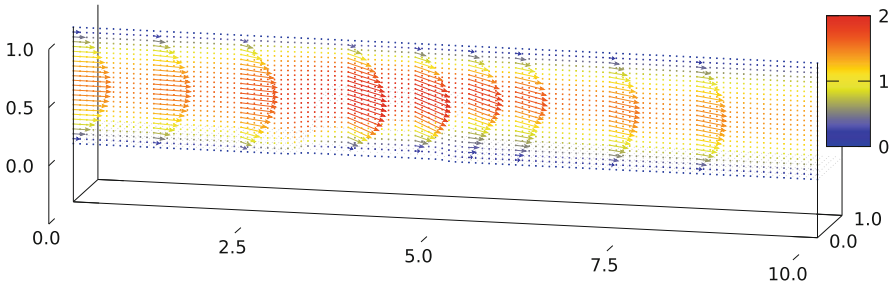



**Fig. 14.8** Vertically parabolic shaped vector velocity field for a slice at $j = 0$ and its deformations around the shallow end of the concavity

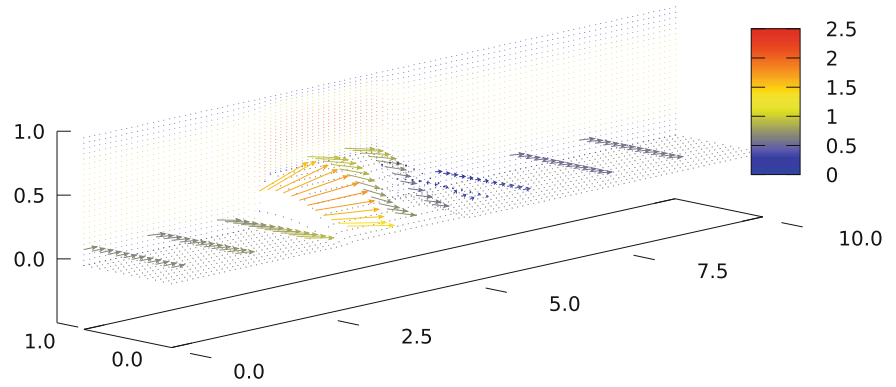



**Fig. 14.9** Vertical speed profile in the plane with $j = 11$ and evolution of the vector field close to the lower boundary ($k = 3$)

smoother velocity profile changes and consequently a weaker formation of vortices. Further, Fig. 14.8 shows the shallow end of the bump, at $j = 0$, where there is only a spurious change in the speed profile.

Initially, the fluid moves in the direction of the $x^1$ axis, but as soon as it interacts with the bump, the flow characteristics change. This fact can be seen in more detail in Fig. 14.9. In this figure, the vector field in the vicinity of the lower boundary for $k = 3$ is shown indicating the direction and intensity of the velocity field. One may observe the changes in the vector field profile as soon as the fluid interacts with the bump, which has as an effect an increase in the flow speed and a diversion of the movement around the concavity.

A complementary plot (Fig. 14.10) shows the velocity vector field in $x^2$-$x^3$ planes for the positions at $i = 0$, $i = 15$, $i = 34$, $i = 48$, $i = 57$, $i = 80$, and $i = 100$. As expected, one observes the predominant parabolic profile in the direction of the $x^1$ axis. Collision with the bump near $x^1 = 3$ reduces the duct's cross section in the vertical direction, causing the flow to move upwards and increasing the velocity at this location. From approximately $x^1 = 5$ on, where the concavity vanishes, the

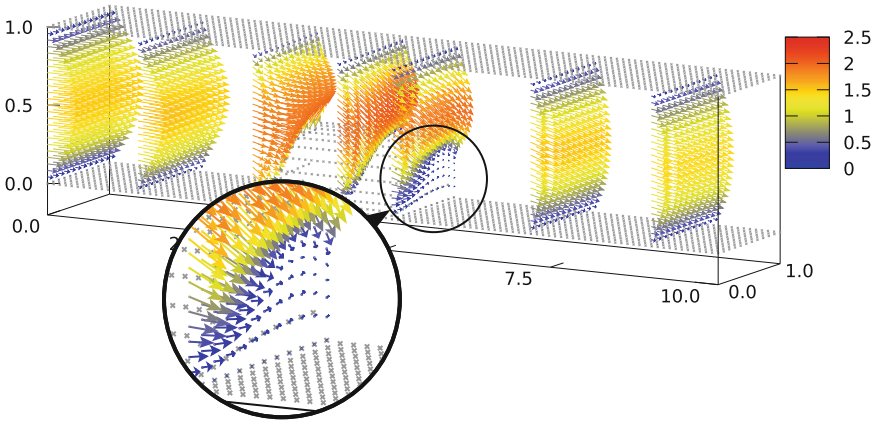



**Fig. 14.10** Three-dimensional vector velocity field (numerical values are reduced by a factor of three) at the positions $i = 0$, $i = 15$, $i = 34$, $i = 48$, $i = 57$, $i = 80$, and $i = 100$. Zoom into the region after the bump of the lower boundary at $x^1 = 5$

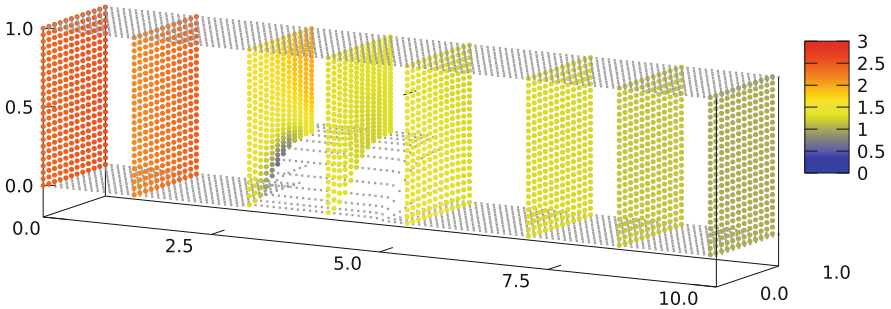



**Fig. 14.11** Cross-sectional pressure profile $P$ for the three-dimensional simulation at the positions $i = 0$, $i = 15$, $i = 34$, $i = 48$, $i = 57$, $i = 80$, and $i = 100$

flow converges against the main direction due to the increase of the cross section in the vertical direction. The formation of vortices occurs and is more accentuated in the limit $y = 1$. Comparable experiments, which make use of the PIV (Particle Image Velocimetry) technique, show exactly these details [Ad05].

The pressure field $P$ with its visible effects due to the change in the cross section is shown in Fig. 14.11 for the positions $i = 0$, $i = 15$, $i = 34$, $i = 48$, $i = 57$, $i = 80$, and $i = 100$. While in the inlet ($x^1 < 2$) and the outlet regions ($x^1 > 7$) the pressure profiles across the $x^2$–$x^3$ plane seem to be to a good approximation homogeneous, in the region of the bump, the effect on the pressure becomes apparent. Although identifying the pressure equation associated to the Navier–Stokes equations is an arduous task, the Poisson equation used in the present discussion to model the pressure contribution to the flow provides results compatible with experimentally trained intuition.

**Table 14.1** Residuals for the components of the velocity fields and the pressure

| Residual | Minimum | Maximum |
|---|---|---|
| $u_r^1$ | $1.2393 \times 10^{-11}$ | $9.4690 \times 10^{-04}$ |
| $u_r^2$ | $3.7872 \times 10^{-12}$ | $6.1993 \times 10^{-04}$ |
| $u_r^3$ | $1.2816 \times 10^{-12}$ | $3.1434 \times 10^{-04}$ |
| $P_r$ | $4.3617 \times 10^{-06}$ | $4.6000 \times 10^{-3}$ |

A quantitative criterion beyond mere intuition is provided by the residual test already used in the two-dimensional case. Table 14.1 shows the minimum and maximum absolute values of the residuals for the three components of the vector velocity field $u_r^1$, $u_r^2$, $u_r^3$ and the pressure $P_r$ field, respectively. In all cases, the residual values attest an acceptable solution for the velocity vector field as well as the pressure distribution. Thus, the resolution of the discretisation of the numerical method was sufficient and did not compromise the quality of the numerical results, which in general is not known beforehand but was evaluated by the obtained results. Our findings and posterior error analysis allow us to conclude that if the model is adequate to simulate the flow phenomenon, our found velocity vector and pressure field represent an acceptable description within the numerical and arithmetic uncertainties.

## 14.6 Conclusions

Flow problems in real scenarios generally have complex curvilinear boundaries, which provide challenges for numerical as well as (semi-)analytical approaches. Quite often, these problems are discussed considering idealised (simplified) boundaries only. Hence, in the present work, the authors made a step into a direction where a class of curvilinear boundaries may be taken into account, but after a diffeomorph conformal coordinate transformation, these simplify to plane parallel boundaries, evidently at the cost of additional terms in the differential operators of the dynamical equation. From the numerical point of view, this fact does not introduce complications into the algorithmic solver, so that as a benefit the matrix system that solves the equation may be set up in the same way as is done for a simple plane parallel boundary problem. The coordinate transform technique has the advantage in comparison to irregular mesh methods that in the latter it is not straightforward to preserve conservation laws, which is guaranteed with the present method. In this line, we showed by two simulations how the method works and made plausible by an error analysis that the obtained solutions are fairly close to the true solutions.

Although of numerical architecture, this solution may be considered a benchmark for other approaches, especially (semi-)analytical ones, which are the focus of our future efforts. Once the surfaces of a three-dimensional domain of interest are parametrised, they provide the basis for the construction of the coordinate

transformation. This reasoning of using curvilinear boundaries to define the new coordinate system that is being used to derive the solution is a new aspect for solving more realistic scenarios in fluid flow problems. Nevertheless, the authors of the present work are aware of the fact that the developed approach imposes restrictions on implementable environmental reliefs for flow simulations, although the discussion so far shows promising perspectives for future developments.

## References

[Ad05] Adrian, R.J.: Twenty years of particle image velocimetry. Exp. Fluids **39**(2), 159–169 (2005). Springer Science and Business Media. https://doi.org/10.1007/s00348-005-0991-7

[Bo15] Bortoli, A.: Modeling and Simulation of Reactive Flows. Elsevier, Amsterdam (2015)

[Ho01] Hoffman, J.: Numerical Methods for Engineers and Scientists. Dekker, New York (2001)

[MeEtAl17] Meneghetti, A., Bodmann, B.E.J., Vilhena, M.T.: A new diffeomorph conformal methodology to solve flow problems with complex boundaries by an equivalent plane parallel problem. In: Integral Methods in Science and Engineering. Vol.1: Theoretical Techniques, pp. 205–214. Birkhäuser, New York (2017). https://doi.org/10.1007/978-3-319-59384-5_18

[ScGe17] Schlichting, H., Gersten, K.: Boundary-Layer Theory. Springer, Heidelberg (2017)

[So64] Sokolnikoff, I.: Tensor Analysis, Theory and Applications to Geometry and Mechanics of Continua. Wiley, New York (1964)

[We72] Weinberg, S.: Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity, Wiley, New York (1972)

# Chapter 15
# Impact Loading of Interface Cracks: Effects of Cracks Closure and Friction


**O. Menshykov, M. Menshykova, and I. A. Guz**

## 15.1 Introduction

It is well known, e.g., see [GuMe03] and [MeGu08a], that the cracks' closure and friction under any type of loading shall be taken into account when the fracture mechanics problems for cracked materials are considered. The simplification of the problem by neglecting the contact between the opposite cracks' faces leads to the quantitative and qualitative change of the results.

The main reason for the neglecting the cracks' faces contact is the complexity of the problem solution. Linear crack problems for normal and oblique time-harmonic loading with consideration of the cracks' closure were solved in [MeWe05] and [MeGu08a]. For the oblique loading, the effects of friction according to the Coulomb friction law were taken into account, and the problem was solved using boundary integral equation method. The solutions of the contact problems for penny-shaped and elliptical cracks in homogeneous material under harmonic loading were presented in [GuMe03, MeGu06, MeGu08b]. Boundary integral equation method was also used by [FoGo21] in order to simulate the elastic wave propagation in layered piezoelectric phononic crystals. Dynamic contact and crack propagation problems were recently solved in [ZhDu21].

With the growing industrial usage of various composite materials, the problems for cracks situated at the interface between two materials are of high interest. Matbuly [Ma06] considered an interfacial crack under shear loading and derived the singular system of integral equations using the variables' separation technique. Men'shikov et al. [MeGu07] presented the expressions for the integral kernels and the numerical solution for a penny-shaped interface crack under normal tension–

O. Menshykov · M. Menshykova (✉) · I. A. Guz
School of Engineering, University of Aberdeen, Aberdeen, UK
e-mail: o.menshykov@abdn.ac.uk; m.menshykova@abdn.ac.uk; i.guz@abdn.ac.uk



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_15

compression loading problem. The validation of the boundary integral equations method for the harmonic loading of the interface crack was done in [MeGu08c]. The application of the boundary integral equations for the case of time-harmonic loading of the crack situated at the bi-material interface problem was presented in [MeGu09a]. The results obtained for a penny-shaped crack under dynamic loading were compared with the ones obtained for the static case. The solution of the problem for shear wave incidence on the interface linear crack is presented in [MeGu09b], where the system of boundary integral equations for displacements and tractions can also be found. The detailed review of the interface crack problems is given in [GuMe13]; please note that the problems mentioned in the current paragraph were solved neglecting the effects of the friction and cracks' closure.

The linear interface crack closure problem for the case of harmonic loading was considered in [MeGu10] and [MeGu11]. The system of hyper-singular integral equations for boundary displacements and tractions was derived from the dynamic Somigliana identity, and the comparison of the results obtained with and without contact interaction of the crack's faces was presented for 2-D and for 3-D cases, see also [MeGu12]. The detailed investigation of the algorithm convergence for the problem solution is done in [MeGu11], where the effects of frequency on the distribution of the stress intensity factors were also studied.

The cracked materials under transient dynamic loading were considered in [MeGu16]. To solve the problem, the boundary integral equations in frequency domain were used, and for different stress pulses, the dynamic stress intensity factors were obtained. Basu and Mandal [BaMa16] considered the problem of the impact torsional loading for the case of penny-shaped crack situated within the elastic layer. The problem for two-dimensional crack under transient dynamic loading was solved by [WuZh09a], the comparison of two hyper-singular time-domain boundary element methods was carried out, and the analysis of dynamic stress intensity factors was presented. The factors that influence the dynamic distribution of stresses for the case of the saw-tooth shock pulse were investigated in [ZhSh20].

The problem for the linear interface crack under impact loading neglecting the crack's closure was solved in [MeGu20a], and the effects of the material properties on the stress intensity factors were analyzed. Interface cracks in layered anisotropic solids were considered by using the time-domain boundary element method in [WuZh09b]. Orthotropic materials with interfacial cracks under impact loading (normal and shear pulses) were investigated in [LiRu05]. Impact of a torsional load was also considered in [KaBa18] for the case of a penny-shaped interface crack.

Finally, the problem for linear crack subjected to normal impact loading was solved in [MeGu20b] taking the crack's closure into account. The solution of the problem was obtained for different friction coefficients and stress pulses. The convergence of the algorithm was analyzed. The problem of oblique Heaviside compression loading of a linear crack was considered in [MeGu20c]. The calculation of the contact forces is done, and the dependence of the solution on the friction coefficient was presented.

In the current study, the 2-D problem for a linear interface crack under the Heaviside normal shear pulse is solved using the boundary integral equations method in the frequency domain accounting for the cracks' closure and friction, the components of the solution are presented, and the iterative process convergence is discussed.

## 15.2 Statement of the Problem and Boundary Integral Formulation

Let us consider a two-dimensional isotropic linearly elastic bi-material under external transient loading. The bi-material contains a linear interface crack of a finite length and without any initial opening, and the Heaviside shear pulse propagates normally to the surface of the crack.

For each isotropic domain, the equations of motion and the generalized Hooke's law lead to the linear Lamé equations of elastodynamics for the displacement field with the appropriate boundary (continuity conditions for stresses and displacements, the Sommerfeld radiation condition at the infinity, and the tractions at the crack's surface defined by the external loading) and initial (zero deformations at the initial moment) conditions.

Thus, the components of the displacement field in both domains, $\Omega^{(1)}$ and $\Omega^{(2)}$, could be represented by the boundary displacements and tractions at the interface (at domains' boundaries), $\Gamma^{(1)}$ and $\Gamma^{(2)}$, using the Somigliana dynamic identity with the appropriate fundamental solutions $U_{ii}^{(m)}(\mathbf{x}, \mathbf{y}, t-\tau)$ and $W_{ij}^{(m)}(\mathbf{x}, \mathbf{y}, t-\tau)$, see [AlBrPa94, MeGu08c, MeGu11] and [MeGu16]:

$$u_j^{(m)}(\mathbf{x}, t) = \int\limits_T \int\limits_{\Gamma(m)} (p_i^{(m)}(\mathbf{y}, \tau) U_{ij}^{(m)}(\mathbf{x}, \mathbf{y}, t-\tau) - u_i^{(m)}(\mathbf{y}, \tau) W_{ij}^{(m)}(\mathbf{x}, \mathbf{y}, t-\tau)) d\mathbf{y} d\tau, \tag{15.1}$$

$$\mathbf{x} \in \Omega^{(m)}, \quad t \in T, \quad j, m = 1, 2.$$

Similar representation may be obtained for the tractions by applying the differential operator to (15.1), and the boundary integral equations for the limiting case at the domains' boundaries have the following form (assuming the smoothness of the boundary displacements and tractions):

$$\frac{1}{2} u_j^{(m)}(\mathbf{x}, t) = \int\limits_T \int\limits_{\Gamma(m)} (p_i^{(m)}(\mathbf{y}, \tau) U_{ij}^{(m)}(\mathbf{x}, \mathbf{y}, t-\tau) - u_i^{(m)}(\mathbf{y}, \tau) W_{ij}^{(m)}(\mathbf{x}, \mathbf{y}, t-\tau)) d\mathbf{y} d\tau, \tag{15.2}$$

$$\frac{1}{2}u_j^{(m)}(\mathbf{x},t)=\int\limits_T\int\limits_{\Gamma(m)}(p_i^{(m)}(\mathbf{y},\tau)K_{ij}^{(m)}(\mathbf{x},\mathbf{y},t-\tau)-u_i^{(m)}(\mathbf{y},\tau)F_{ij}^{(m)}(\mathbf{x},\mathbf{y},t-\tau))d\mathbf{y}d\tau, \tag{15.3}$$

where $\mathbf{x}\in\Gamma^{(m)},\quad t\in T.$

Because of the crack's faces closure, the traction vector at the crack's surface can be represented as the superposition of the predefined traction caused by the external loading and the contact force that appears at the crack's surface (in the contact zone that changes in time due to the dynamic loading). The length (in a 2-D case) and shape (in a 3-D case) of the contact zone are unknown beforehand and depend on the parameters of the external loading (type of the loading, its direction, magnitude, frequency, etc.), mechanical properties of the bi-material, and the friction conditions at the crack's surface and must be determined during the solution process.

To take the crack closure and friction into account, the Signorini unilateral constraints (ensuring that there is no interpenetration of the opposite crack faces, the normal component of the contact force is unilateral and present in the contact zone only) and the Coulomb friction law (the contacting crack faces do not move in the tangential direction, while they are held by the friction unless the slipping happens) are applied to the normal and tangential components of the displacement jump (displacement discontinuity), $[\mathbf{u}(\mathbf{x},t)]=\mathbf{u}^{(1)}(\mathbf{x},t)-\mathbf{u}^{(2)}(\mathbf{x},t)$, and contact forces, see [GuZo02, GuMe13, MeGu20b]:

$$[u_n(\mathbf{x},t)]\geq 0,\quad q_n(\mathbf{x},t)\geq 0,\quad [u_n(\mathbf{x},t)]q_n(\mathbf{x},t)=0, \tag{15.4}$$

$$|\mathbf{q}_\tau(\mathbf{x},t)|<k_\tau q_n(\mathbf{x},t)\Rightarrow\frac{\partial[\mathbf{u}_\tau(\mathbf{x},t)]}{\partial t}=0, \tag{15.5}$$

$$|\mathbf{q}_\tau(\mathbf{x},t)|=k_\tau q_n(\mathbf{x},t)\Rightarrow\frac{\partial[\mathbf{u}_\tau(\mathbf{x},t)]}{\partial t}=-\frac{\mathbf{q}_\tau(\mathbf{x},t)}{|\mathbf{q}_\tau(\mathbf{x},t)|}\left|\frac{\partial[\mathbf{u}_\tau(\mathbf{x},t)]}{\partial t}\right|. \tag{15.6}$$

Let us approximate the external transient dynamic loading and the components of the solution by the Fourier exponential series with the appropriate number of the Fourier coefficients (may be quite high for the impact pulses), as it was suggested in [GuMe13, MeGu20a] and [MeGu20b]. That will allow to use the solution approach previously developed by authors for cracked materials under harmonic loading in the frequency domain.

In particular, the Heaviside impact pulse can be approximated, for example, by the repeating "steep and long" trapezoidal stress pulse, [MeGu16]:

$$\sigma(t)=\sigma^*\left\{\begin{array}{c}\frac{t}{t^*}(H(t)-H(t-t^*))+(H(t-t^*)-H(t-t^*-t_d))\\ +(2-\frac{t-t_d}{t^*}(H(t-t^*-t_d)-H(t-2t^*-t_d))\end{array}\right\},$$

where $c_2^{(1)}t^*=0.1\ \ and\ \ c_2^{(1)}t_d=12.$

Thus, as it was mentioned above, the components of the solution at the crack surface can be approximated by the following exponential Fourier time series:

$$f(\bullet,t)=Re\left\{\sum_{k=-\infty}^{+\infty}f^k(\bullet)e^{i\omega_k t}\right\},\quad f^k(\bullet)=\frac{\omega}{2\pi}\int_0^T f(\bullet,t)e^{-i\omega_k t}dt, \tag{15.7}$$

where $\omega_k=2\pi k/T$ and $i$ is the imaginary unit.

Fundamental solutions in the frequency domain have the following form, see [AlBrPa94, MeGu09b]:

$$U_{12}^{(m)}(\mathbf{x},\mathbf{y},\omega_k)=U_{21}^{(m)}(\mathbf{x},\mathbf{y},\omega_k)=0,$$

$$U_{11}^{(m)}(\mathbf{x},\mathbf{y},\omega_k)=\frac{1}{2\pi\mu^{(m)}}\left[K_0(l_{2,k}^{(m)})+\frac{1}{l_{2,k}^{(m)}}\left(K_1(l_{2,k}^{(m)})-\frac{c_2^{(m)}}{c_1^{(m)}}K_1(l_{1,k}^{(m)})\right)\right],$$

$$U_{22}^{(m)}(\mathbf{x},\mathbf{y},\omega_k)=\frac{1}{2\pi\mu^{(m)}}\left[\begin{array}{c}(\frac{c_2^{(m)}}{c_1^{(m)}})^2K_2(l_{1,k}^{(m)})-K_2(l_{2,k}^{(m)})+K_0(l_{2,k}^{(m)})\\ +\frac{1}{l_{2,k}^{(m)}}\left(K_1(l_{2,k}^{(m)})-\frac{c_2^{(m)}}{c_1^{(m)}}K_1(l_{1,k}^{(m)})\right)\end{array}\right],$$

$$W_{11}^{(m)}(\mathbf{x},\mathbf{y},\omega_k)=W_{22}^{(m)}(\mathbf{x},\mathbf{y},\omega_k)=0,$$

$$W_{12}^{(m)}(\mathbf{x},\mathbf{y},\omega_k)=\frac{1}{2\pi r}\frac{\delta r}{\delta y_1}\left[\begin{array}{c}l_{2,k}^{(m)}K_1(l_{2,k}^{(m)})-2K_2(l_{2,k}^{(m)})\\ +2\left(\frac{c_2^{(m)}}{c_1^{(m)}}\right)^2K_2(l_{1,k}^{(m)})\end{array}\right],$$

$$W_{21}^{(m)}(\mathbf{x},\mathbf{y},\omega_k)=\frac{1}{2\pi r}\frac{\delta r}{\delta y_1}\left[\begin{array}{c}-\frac{\lambda^{(m)}\mu^{(m)}}{(\lambda^{(m)}+2\mu^{(m)})^2}l_{1,k}^{(m)}K_1(l_{1,k}^{(m)})\\ -2K_2(l_{2,k}^{(m)})+2\left(\frac{c_2^{(m)}}{c_1^{(m)}}\right)^2K^2(l_{1,k}^{(m)})\end{array}\right],$$

where $l_{1,k}^{(m)}=i\omega_k r/c_1^{(m)}$, $l_{2,k}^{(m)}=i\omega_k r/c_2^{(m)}$; $c_1^{(m)}=\sqrt{(\lambda^{(m)}+2\mu^{(m)})/\rho^{(m)}}$ and $c_2^{(m)}=\sqrt{\mu^{(m)}/\rho^{(m)}}$, and $r=|x_1-y_1|$ is the distance between the loading and observation points.

For every Fourier coefficient number, $k$, the appropriate system of linear algebraic equations can be obtained from the boundary integral equations (15.2) and (15.3) and then solved numerically, so the Fourier representations of the components of the solution (15.7) with the finite number of the Fourier coefficients can be found.

During the numerical solution, divergent integrals of various orders (weakly singular, singular, and hyper-singular ones) that depend on the type and order of the

approximation shall be regularized and computed. In the current study, the simplest piecewise-constant approximation was used, as it successfully proved its efficiency for two-dimensional problems compared, for example, with the Galerkin method, see [MeWe05].

In order to take the contact constraints (15.4)–(15.6) into account, the iterative correction algorithm based on the orthogonal projections on the sets of constraints shall be used. The detailed description, studies on the numerical convergence, and the comparison of the iterative algorithms applicable to homogeneous and layered materials are given in [GuMe13, MeGu11, MeGu20b]. In the current study, the algorithm presented in [MeGu11] is used. The references above also contain the detailed analysis of the numerical convergence of the iterative algorithm for different loading conditions and material properties.

In particular, according to [MeGu16], for the impact loading of a homogeneous cracked material, at least 30 Fourier coefficients should be used to adequately approximate the components of the solution and the external pulse. For linear interface cracks, additional details of the numerical convergence analysis were also presented in [MeGu20a] with the recommended number of Fourier coefficients being equal to 50. Thus, in this chapter, for the consistency, 50 Fourier coefficients were used to represent the external loading and the components of the numerical solution.

## 15.3 Numerical Results and Conclusions

For the validation of the numerical solution, the linear interface crack of the length $2L$ under the normally incident Heaviside shear pulse of amplitude $\sigma_0$ (with the normalized wave number $k_2^{(2)} L = \omega L / c_2^{(2)} = 0.01$) was considered.

The following mechanical properties of the bi-material ($\nu^{(1)} = 0.1$, $E^{(1)} = 29 GPa$, and $\nu^{(2)} = 0.49$, $E^{(2)} = 400 GPa$) were used in order to satisfy the model constraint, see [Co90] and [CoDu80]:

$$\beta = \frac{\mu^{(2)}(\kappa^{(1)} - 1) - \mu^{(1)}(\kappa^{(2)} - 1)}{\mu^{(2)}(\kappa^{(1)} - 1) + \mu^{(1)}(\kappa^{(2)} - 1)} = 0.5, \qquad \kappa^{(m)} = 3 - 4\upsilon^{(m)}.$$

The normalized normal components of the displacement discontinuity and the contact force, $2\mu_0[u_n]/\sigma_0 L$ and $q_n/\sigma_0$ accordingly, at the crack's surface when the stable quasi-static solution is achieved (after some time since the shear pulse is applied to the crack) are presented in Figs. 15.1 and 15.2 disregarding the friction and taking it into account (for the friction coefficient $k_\tau = 0.0$ and $k_\tau = 1.0$).

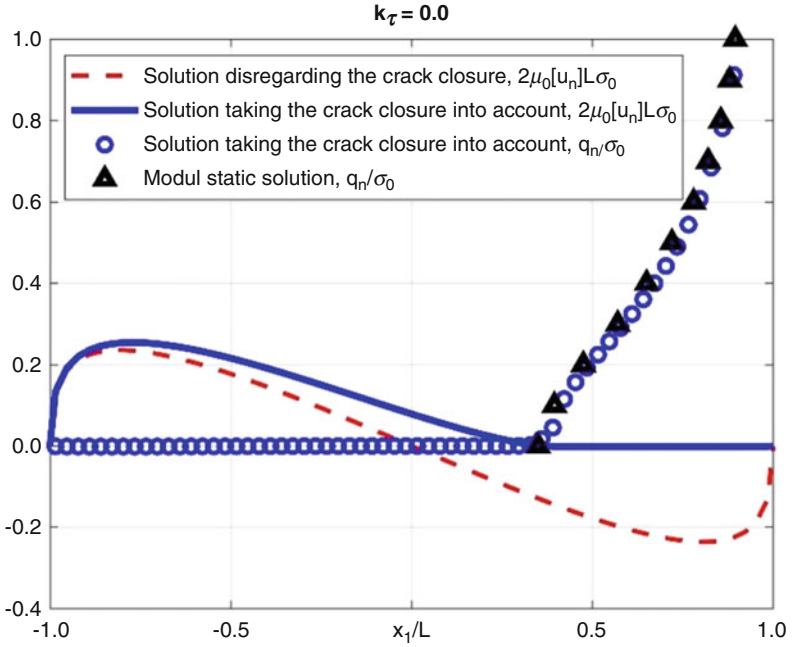



**Fig. 15.1** Normal contact forces and the displacement jump at the crack surface disregarding friction

Please note that:

$$\mu_0 = \mu^{(1)} \frac{1-\gamma_2}{1+\kappa^{(1)}}, \quad \gamma_2 = \frac{a_1}{2} - a_2,$$

$$a_1 = \frac{\mu^{(1)} - \mu^{(2)}}{\mu^{(1)} + \kappa^{(1)}\mu^{(2)}}, \quad a_2 = \frac{\kappa^{(1)}\mu^{(2)} - \kappa^{(2)}\mu^{(1)}}{2(\mu^{(2)} + \kappa^{(2)}\mu^{(1)})}.$$

The normal and tangential components of the displacement jump and contact forces at the crack surface plotted against the iteration number are presented in Figs. 15.3, 15.4, 15.5 and 15.6 for $k_\tau = 1.0$. One can clearly see that both components of the solution are gradually changing till the final solution is found.

Please note also that the convergence of the iteration process takes much longer than 100 iterations (as presented in Figs. 15.3, 15.4, 15.5 and 15.6 for illustration purposes only). In particular, the small region of the crack's faces interpenetration is still visible in Fig. 15.3 next to the crack tip $(L,0)$, and the correction of the components, especially, of the components of the contact force, has not been fully completed (as the forces are still significantly changing with each iteration step).

The results in Figure 1 have been presented after the convergence had been achieved (after 1000 iterations). Finally, it is worth to mention that the convergence

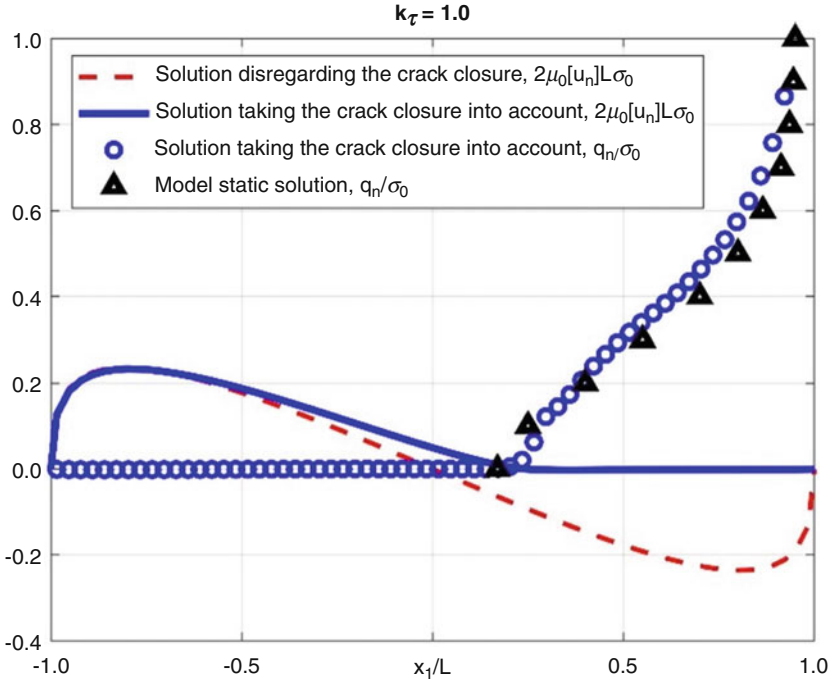



**Fig. 15.2** Normal contact forces and the displacement jump at the crack surface with friction

rate can be changed by the choice of the iterative coefficients, and the values recommended in [MeGu11] have been used.

Please note that, as it was also shown in [MeGu21], after the correction, the contact constraints (15.4)–(15.6) are satisfied on the entire surface of the crack. The most importantly, there is no interpenetration of the crack's opposite faces, and the friction significantly affects the distribution of displacements and tractions, as well as the length of the contact zone; and the Sommerfeld radiation conditions are satisfied at the infinity (the displacements and forces slowly but surely decrease at the bonding interface with the increase of the distance from the crack), so the iterative process effectively corrects the solution.

The contact forces and the size of the contact zone are compared with the model static solution by Comninou and Dundurs [CoDu80]. As one can see, the results are in a very good agreement, complementing the results presented in [MeGu21] for the case of "slow" harmonic shear loading of the interface crack.

Thus, the crack's closure and friction significantly change the distribution of the displacements and tractions at the bonding interface, inevitably affecting the distribution of the dynamic stress intensity factors in the vicinity of the crack's tips. The stress intensity factors (the opening and the transverse shear modes) can be computed using the following asymptotic formulas:

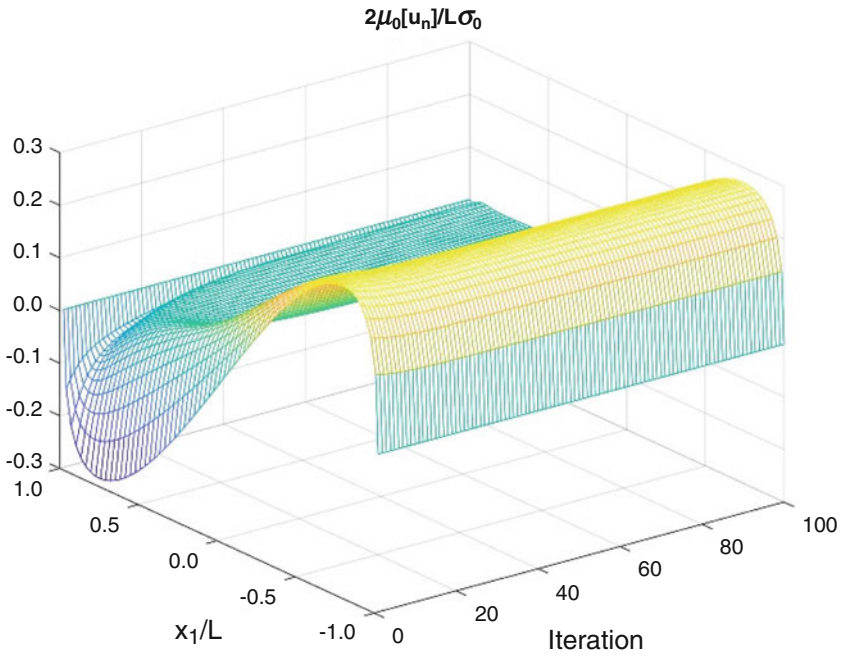



**Fig. 15.3** Normal displacement jump at the crack surface

$$K_I = \max_t \lim_{r\to 0} |p_n^*(R+r,t)|\sqrt{2\pi r}, \tag{15.8}$$

$$K_{II} = \max_t \lim_{r\to 0} |p_\tau^*(R+r,t)|\sqrt{2\pi r}, \tag{15.9}$$

where $p_n^*(L+r,t)$ and $p_\tau^*(L+r,t)$ are the normal and tangential components of the traction vector at the bonding interface and r is the distance from the crack tip. The appropriate representations of the stress intensity factors computed through the displacement discontinuity (very similar to asymptotic representations (15.8) and (15.9)) may also be used. The computation and analysis of the stress intensity factors for different mechanical properties of the bi-material and different directions of the loading will be the next step of the current research study.

As a conclusion, it shall be added that the proposed approach may be extended to three-dimensional fracture mechanics problems for cracked materials under arbitrary dynamic loading, and the special attention shall be paid to the coupling oscillation singularities in the vicinity of the crack's front, e.g., see [Co90], [Os19].

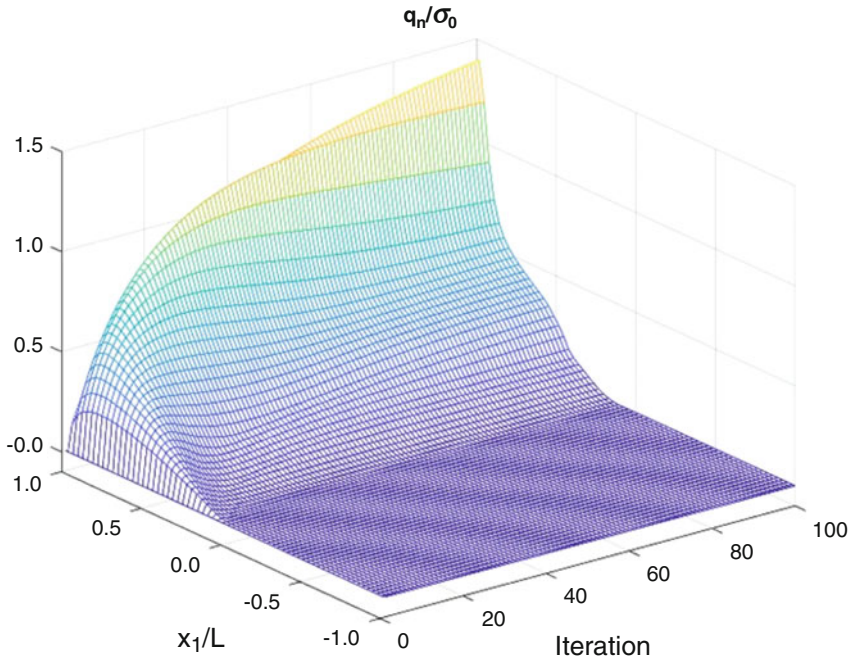



**Fig. 15.4** Normal contact forces at the crack surface

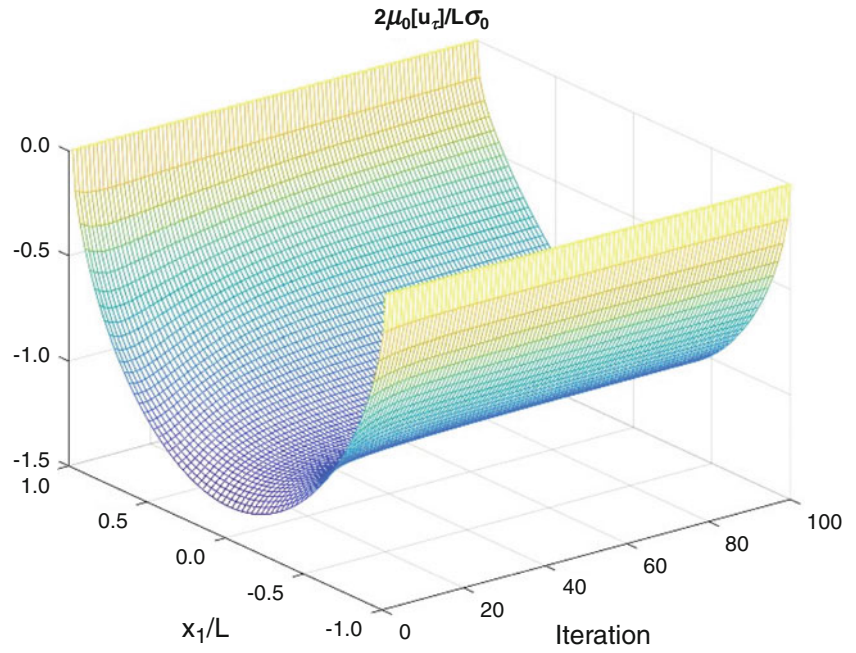



**Fig. 15.5** Shear displacement jump at the crack surface

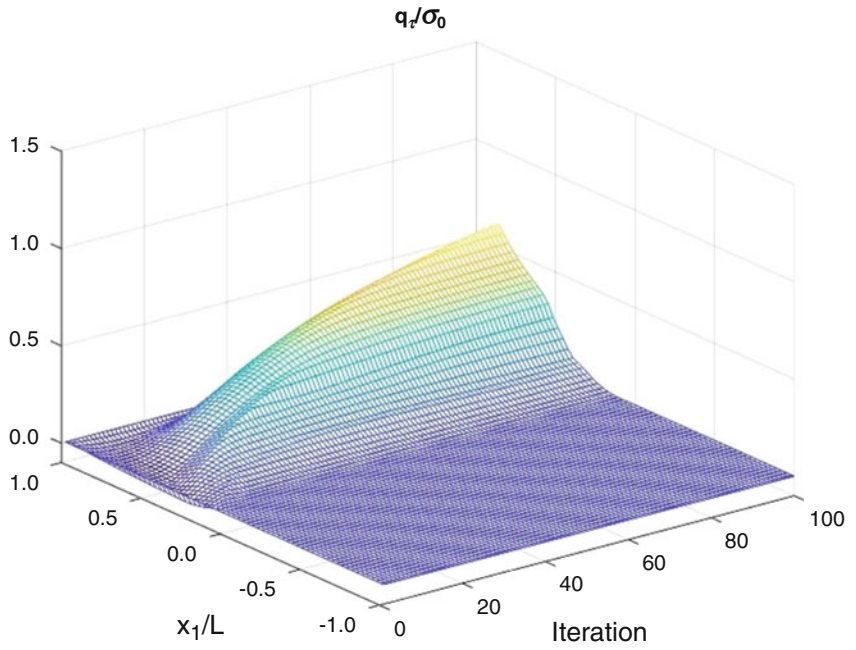



**Fig. 15.6** Shear contact forces at the crack surface

## References

[AlBrPa94] Aliabadi, M.H., Brebbia, C.A., Parton, V.Z.: Static and Dynamic Fracture Mechanics. Computational Mechanics Publications, Southampton (1994)

[BaMa16] Basu, S., Mandal, S.C.: Impact of torsional load on a penny-shaped crack in an elastic layer sandwiched between two elastic half-spaces. Int. J. Appl. Comput. Math. **2**, 533–543 (2016)

[Co90] Comninou, M.: An overview of interfacial cracks. Eng. Fract. Mech. **37**(1), 197–208 (1990)

[CoDu80] Comninou, M., Dundurs, J.: Effect of friction on the interface crack loaded in shear. J. Elasticity **10**, 203–212 (1980)

[FoGo21] Fomenko, S.I., Golub, M.V., Doroshenko, O.V., Wang, Y., Zhang, C.: An advanced boundary integral equation method for wave propagation analysis in a layered piezoelectric phononic crystal with a crack or an electrode. J. Comput. Phys. **447**, 110669 (2021)

[GuMe03] Guz, A.N., Menshykov, O.V., Zozulya, V.V.: Surface contact of elliptical crack under normally incident tension-compression wave. Theor. Appl. Fract. Mech. **40**(3), 285–291 (2003)

[GuMe13] Guz, A.N., Guz, I.A., Men'shikov, A.V., Men'shikov, V.A.: Three-dimensional problems in the dynamic fracture mechanics of materials with interface cracks (Review). Int. Appl. Mech. **49**(1), 1–61 (2013)

[GuZo02] Guz, A.N., Zozulya, V.V.: Elastodynamic unilateral contact problems with friction for bodies with cracks. Int. Appl. Mech. **38**, 895–932 (2002)

[KaBa18] Karan, S., Basu, S., Mandal, S.C.: Impact of a torsional load on a penny-shaped crack sandwiched between two elastic layers embedded in an elastic medium. Acta Mech. **229**, 1759–1772 (2018)

[LiRu05] Lira-Vergara, E., Rubio-Gonzalez, C.: Dynamic stress intensity factor of interfacial finite cracks in orthotropic materials. Int. J. Fract. **135**, 285–309 (2005)

[Ma06] Matbuly, M.S.: Analytical solution for an interfacial crack subjected to dynamic anti-plane shear loading. Acta Mech. **184**, 77–85 (2006)

[MeGu06] Menshykov, O., Guz, I.: Contact interaction of crack faces under oblique incidence of a harmonic wave. Int. J. Fract. **139(1)**, 145–152 (2006)

[MeGu07] Men'shikov, V.A., Men'shikov, A.V., Guz, I.A.: Interfacial crack between elastic half-spaces under harmonic loading. Int. Appl. Mech. **43(8)**, 865–873 (2007)

[MeGu08a] Menshykov, O.V., Menshykova, M.V., Guz, I.A.: Effect of friction of the crack faces for a linear crack under an oblique harmonic loading. Int. J. Eng. Sci. **46(5)**, 438–458 (2008)

[MeGu08b] Menshykov, O.V., Menshykov, V.A., Guz, I.A.: The contact problem for an open penny-shaped crack under normally incident tension-compression wave. Eng. Fract. Mech. **75**(5), 1114–1126 (2008)

[MeGu08c] Menshykov, O.V., Guz, I.A., Menshykov, V.A.: Boundary integral equations in elastodynamics of interface cracks. Phil. Trans. R. Soc. A Math. Phys. Eng. Sci. **366**(1871), 1835–1839 (2008)

[MeGu09a] Menshykov, O.V., Menshykov, V.A., Guz, I.A.: Elastodynamics of a crack on the bi-material interface. Eng. Analy. Bound. Elem. **33**(3), 294–301 (2009)

[MeGu09b] Menshykova, M.V., Menshykov, O.V., Guz, I.A.: Linear interface crack under plane shear wave. CMES–Comput. Modell. Eng. Sci. **48**(2), 107–120 (2009)

[MeGu10] Menshykova, M.V., Menshykov, O.V., Guz, I.A.: Modelling crack closure for an interface crack under harmonic loading. Int. J. Fract. **165**(1), 127–134 (2010)

[MeGu11] Menshykova, M.V., Menshykov, O.V., Guz, I.A.: An iterative BEM for the dynamic analysis of interface crack contact problems. Eng. Analy. Bound. Elem. **35**(5), 735–749 (2011)

[MeGu12] Menshykov, O.V., Menshykova, M.V., Guz I.A.: 3-D elastodynamic contact problem for an interface crack under harmonic loading. Eng. Fract. Mech. **80**, 52–59 (2012)

[MeGu16] Menshykova, M.V., Menshykov, O.V., Guz, I.A., Wuensche, M., Zhang, C.: A boundary integral equation method in the frequency domain for cracks under transient loading. Acta Mech. **227**(11), 3305–3314 (2016)

[MeGu20a] Menshykov, O.V., Menshykova, M.V., Guz I.A.: Boundary integral equations in the frequency domain for interface linear cracks under impact loading. Acta Mech. **231**, 3461–3471 (2020)

[MeGu20b] Menshykov, O.V., Menshykova, M.V., Guz I.A.: Effects of crack closure and friction for linear crack under normal impact. Eng. Analy. Bound. Elem. **115**, 1–9 (2020)

[MeGu20c] Menshykov, O.V., Menshykova, M.V., Guz I.A.: Contact problems for cracks under impact loading. Proc. Struct. Integr. **28**, 1621–1628 (2020)

[MeGu21] Menshykov, V.A., Menshykov, O.V., Guz, I.A.: Contact problems for interface cracks under harmonic shear loading. In: Scipedia – Fracture, Damage and Failure Mechanics, vol. 100 (2021). https://doi.org/10.23967/wccm-eccomas.2020.104

[MeWe05] Menshykov, O.V., Menshykova, M.V., Wendland, W.L.: On the use of Galerkin method to solve the fracture mechanics problem for a linear crack under normal loading. Int. Appl. Mech. **41**(11), 1324–1329 (2005)

[Os19] Ostrik, V.I.: Contact of faces of a rectilinear crack under complex loading and various contact conditions. Acta Mech. **230**, 3741–3758 (2019)

[WuZh09a] Wuensche, M., Zhang, Ch., Garcia-Sanchez, F., Saez, A., Sladek, J., Sladek, V.: On two hypersingular time-domain BEM for dynamic crack analysis in 2D anisotropic elastic solids. Comput. Methods Appl. Mech. Eng. **198**, 2812–2824 (2009)

[WuZh09b] Wuensche, M., Zhang, Ch., Sladek, J., Sladek, V., Hirose, S., Kuna, M.: Transient dynamic analysis of interface cracks in layered anisotropic solids under impact loading. Int. J. Fract. **157**, 131–147 (2009)

[ZhDu21] Zhang, P., Du, C., Zhao, W., Sun, L.: Dynamic crack face contact and propagation simulation based on the scaled boundary finite element method. Comput. Methods Appl. Mech. Eng. **385**, 114044 (2021)

[ZhSh20] Zhang, X., Shi, Y., Pan, G.: Dynamic stress control of bi-material structure subjected to sawtooth shock pulse based on interface characteristics. Mech. Res. Commun. **107**, 103558 (2020)

# Chapter 16
# Periodic Solutions in $\mathbb{R}^n$ for Stationary Anisotropic Stokes and Navier-Stokes Systems


**S. E. Mikhailov**

## 16.1 Introduction

Analysis of Stokes and Navier-Stokes equations is an established and active field of research in the applied mathematical analysis, see, e.g., [CF88, Ga11, RRS16, Se15, So01, Te95, Te01] and references therein. In [KMW20, KMW21a, KMW21b] this field has been extended to the transmission and boundary-value problems for stationary Stokes and Navier-Stokes equations of anisotropic fluids, particularly, with relaxed ellipticity condition on the viscosity tensor. In this chapter, we present some further results in this direction considering periodic solutions to the stationary Stokes and Navier-Stokes equations of anisotropic fluids, with an emphasis on solution regularity.

First, the solution uniqueness and existence of a stationary, anisotropic (linear) Stokes system with constant viscosity coefficients in a compressible framework are analysed on $n$-dimensional flat torus in a range of periodic Sobolev (Bessel-potential) spaces. By employing the Leray-Schauder fixed point theorem, the linear results are used to show existence of solution to the stationary anisotropic (non-linear) Navier-Stokes incompressible system on torus in a periodic Sobolev space for $n \in \{2, 3\}$. Then the solution regularity results for stationary anisotropic Navier-Stokes system on torus are established for $n \in \{2, 3\}$.

S. E. Mikhailov (✉)
Brunel University London, Uxbridge, UK
e-mail: sergey.mikhailov@brunel.ac.uk



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_16

## 16.2 Anisotropic Stokes and Navier-Stokes Systems

Let $\mathfrak{L}$ denote a second order differential operator in the component-wise divergence form,

$$(\mathfrak{L}\mathbf{u})_k := \partial_\alpha\big(a_{kj}^{\alpha\beta} E_{j\beta}(\mathbf{u})\big), \quad k = 1, \dots, n,$$

were $\mathbf{u} = (u_1, \dots, u_n)^\top$, $E_{j\beta}(\mathbf{u}) := \frac{1}{2}(\partial_j u_\beta + \partial_\beta u_j)$ are the entries of the symmetric part $\mathbb{E}(\mathbf{u})$ of $\nabla\mathbf{u}$ (the gradient of $\mathbf{u}$), and $a_{kj}^{\alpha\beta}$ are constant components of the tensor viscosity coefficient $\mathbb{A} := \left(a_{kj}^{\alpha\beta}\right)_{1\le i,j,\alpha,\beta\le n}$, cf. [Duf78].

Here and further on, the Einstein summation convention in repeated indices from 1 to $n$ is used unless stated otherwise.

The following symmetry conditions are assumed (see [OSY92, (3.1),(3.3)]),

$$a_{kj}^{\alpha\beta} = a_{\alpha j}^{k\beta} = a_{k\beta}^{\alpha j}. \tag{16.1}$$

In addition, we require that tensor $\mathbb{A}$ satisfies the (relaxed) ellipticity condition in terms of all *symmetric* matrices in $\mathbb{R}^{n\times n}$ with *zero matrix trace*, see [KMW21a, KMW21b]. Thus, we assume that there exists a constant $C_{\mathbb{A}} > 0$ such that,

$$a_{kj}^{\alpha\beta}\zeta_{k\alpha}\zeta_{j\beta} \ge C_{\mathbb{A}}^{-1}|\boldsymbol{\zeta}|^2, \ \forall\, \boldsymbol{\zeta} = (\zeta_{k\alpha})_{k,\alpha=1,\dots,n} \in \mathbb{R}^{n\times n}$$
$$\text{such that } \boldsymbol{\zeta} = \boldsymbol{\zeta}^\top \text{ and } \sum_{k=1}^{n}\zeta_{kk} = 0, \tag{16.2}$$

where $|\boldsymbol{\zeta}|^2 = \zeta_{k\alpha}\zeta_{k\alpha}$, and the superscript $\top$ denotes the transpose of a matrix.

The tensor $\mathbb{A}$ is endowed with the norm

$$\|\mathbb{A}\| := \max\left\{|a_{kj}^{\alpha\beta}| : k, j, \alpha, \beta = 1 \dots, n\right\}.$$

Symmetry conditions (16.1) lead to the following equivalent form of the operator $\mathfrak{L}$

$$(\mathfrak{L}\mathbf{u})_k = \partial_\alpha\big(a_{kj}^{\alpha\beta}\partial_\beta u_j\big), \quad k = 1, \dots, n. \tag{16.3}$$

Let us also define the Stokes operator $\boldsymbol{\mathcal{L}}$ as

$$\boldsymbol{\mathscr{L}}(\mathbf{u}, p) := \mathfrak{L}\mathbf{u} - \nabla p. \tag{16.4}$$

Let $\mathbf{u}$ be an unknown vector field, $p$ be an unknown scalar field, $\mathbf{f}$ be a given vector field and $g$ be a given scalar field defined in $\mathbb{T}$. Then the equations

$$-\boldsymbol{\mathscr{L}}(\mathbf{u}, p) = \mathbf{f}, \ \operatorname{div}\mathbf{u} = g \text{ in } \mathbb{T} \tag{16.5}$$

determine the *anisotropic stationary Stokes system with viscosity tensor coefficient* $\mathbb{A} = \left(A^{\alpha\beta}\right)_{1\le\alpha,\beta\le n}$ *in a compressible framework*.

In addition, the following nonlinear system

$$-\mathscr{L}(\mathbf{u}, p) + (\mathbf{u}\cdot\nabla)\mathbf{u} = \mathbf{f}\,,\ \ \operatorname{div}\mathbf{u} = g \text{ in } \mathbb{T} \tag{16.6}$$

is called the *anisotropic stationary Navier-Stokes system with viscosity tensor coefficient* $\mathbb{A} = \left(A^{\alpha\beta}\right)_{1\le\alpha,\beta\le n}$ *in a compressible framework*. If $g = 0$ in (16.5) and (16.6), then these equations are reduced, respectively, to the *incompressible anisotropic stationary Stokes and Navier-Stokes systems*.

In the *isotropic case*, the tensor $\mathbb{A}$ reduces to

$$a_{kj}^{\alpha\beta} = \lambda\delta_{k\alpha}\delta_{j\beta} + \mu\left(\delta_{\alpha j}\delta_{\beta k} + \delta_{\alpha\beta}\delta_{kj}\right),\ 1\le i, j, \alpha, \beta \le n\,, \tag{16.7}$$

where $\lambda$ and $\mu$ are real constant parameters with $\mu > 0$ (cf., e.g., Appendix III, Part I, Section 1 in [Te01]), and (16.3) becomes

$$\mathfrak{L}\mathbf{u} = (\lambda+\mu)\nabla\operatorname{div}\mathbf{u} + \mu\Delta\mathbf{u}. \tag{16.8}$$

Then it is immediate that condition (16.2) is fulfilled (cf. [KMW21b]) and thus our results apply also to the Stokes and Navier-Stokes systems in the *isotropic case*. Assuming $\lambda = 0$, $\mu = 1$ we arrive at the classical mathematical formulations of isotropic Stokes and Navier-Stokes systems.

## 16.3 Some Function Spaces on Torus

Let us introduce some function spaces on torus and periodic function spaces (see, e.g., [Agm65, p.26], [Agr15], [McL91], [RT10, Chapter 3], [RRS16, Section 1.7.1], and [Te95, Chapter 2], for more details).

Let $n \ge 1$ be an integer and $\mathbb{T}$ be the $n$-dimensional flat torus that can be parametrized as the semi-open cube $\mathbb{T} = [0, 1)^n \subset \mathbb{R}^n$, cf. [Zy02, p. 312]. In what follows, $\mathcal{D}(\mathbb{T}) = \mathcal{C}^\infty(\mathbb{T})$ denotes the space of infinitely smooth real or complex functions on the torus. As usual, $\mathbb{N}$ denotes the set of natural numbers, $\mathbb{N}_0$ the set of natural numbers complemented by 0, and $\mathbb{Z}$ the set of integers.

Let $\boldsymbol{\xi} \in \mathbb{Z}^n$ denote the $n$-dimensional vector with integer components. We will further need also the set

$$\dot{\mathbb{Z}}^n := \mathbb{Z}^n \setminus \{\mathbf{0}\}.$$

Extending the torus parametrisation to $\mathbb{R}^n$, it is often useful to identify $\mathbb{T}$ with the quotient space $\mathbb{R}^n \setminus \mathbb{Z}^n$. Then the space of functions $\mathcal{C}^\infty(\mathbb{T})$ on the torus can be identified with the space of $\mathbb{T}$-periodic (1-periodic) functions $\mathcal{C}^\infty_\# = \mathcal{C}^\infty_\#(\mathbb{R}^n)$ that

consists of functions $\phi \in \mathcal{C}^\infty(\mathbb{R}^n)$ such that

$$\phi(\mathbf{x}+\boldsymbol{\xi}) = \phi(\mathbf{x}) \quad \forall\, \boldsymbol{\xi} \in \mathbb{Z}^n.$$

Similarly, the Lebesgue space on the torus $L_p(\mathbb{T})$, $1 \le p \le \infty$, can be identified with the periodic Lebesgue space $L_{p\#} = L_{p\#}(\mathbb{R}^n)$ that consists of functions $\phi \in L_{p,\mathrm{loc}}(\mathbb{R}^n)$, which satisfy the periodicity condition for a.e. $\mathbf{x}$.

The space dual to $\mathcal{D}(\mathbb{T})$, i.e., the space of linear bounded functionals on $\mathcal{D}(\mathbb{T})$, called the space of torus distributions is denoted by $\mathcal{D}'(\mathbb{T})$ and can be identified with the space of periodic distributions $\mathcal{D}'_\#$ acting on $\mathcal{C}^\infty_\#$.

The toroidal/periodic Fourier transform mapping a function $g \in \mathcal{C}^\infty_\#$ to a set of its Fourier coefficients $\hat{g}$ is defined as (see, e.g., [RT10, Definition 3.1.8])

$$\hat{g}(\boldsymbol{\xi}) = [\mathcal{F}_{\mathbb{T}} g](\boldsymbol{\xi}) := \int_{\mathbb{T}} e^{-2\pi i \mathbf{x}\cdot\boldsymbol{\xi}} g(\mathbf{x}) d\mathbf{x}, \quad \boldsymbol{\xi} \in \mathbb{Z}^n.$$

and can be generalised to the Fourier transform acting on a distribution $g \in \mathcal{D}'_\#$.

For any $\boldsymbol{\xi} \in \mathbb{Z}^n$, let $|\boldsymbol{\xi}| := (\sum_{j=1}^n \xi_j^2)^{1/2}$ be the Euclidean norm in $\mathbb{Z}^n$ and let us denote

$$\rho(\boldsymbol{\xi}) := (1 + |\boldsymbol{\xi}|^2)^{1/2}.$$

Evidently,

$$\frac{1}{2}\rho(\boldsymbol{\xi})^2 \le |\boldsymbol{\xi}|^2 \le \rho(\boldsymbol{\xi})^2 \quad \forall\, \boldsymbol{\xi} \in \dot{\mathbb{Z}}^n. \tag{16.9}$$

Similar to [RT10, Definition 3.2.2], for $s \in \mathbb{R}$ we define the *periodic/toroidal Sobolev (Bessel-potential) spaces* $H^s_\# := H^s_\#(\mathbb{R}^n) := H^s(\mathbb{T})$, which consist of the torus distributions $g \in \mathcal{D}'(\mathbb{T})$, for which the norm

$$\|g\|_{H^s_\#} := \|\rho^s \widehat{g}\|_{\ell_2} := \left( \sum_{\xi \in \mathbb{Z}^n} \rho(\boldsymbol{\xi})^{2s} |\widehat{g}(\boldsymbol{\xi})|^2 \right)^{1/2} \tag{16.10}$$

is finite, i.e., the series in (16.10) converges. Here $\|\cdot\|_{\ell_2}$ is the standard norm in the space of square summable sequences. By Ruzhansky and Turunen [RT10, Proposition 3.2.6], $H^s_\#$ are Hilbert spaces.

For $g \in H^s_\#$, $s \in \mathbb{R}$, and $m \in \mathbb{N}_0$, let us consider the partial sums

$$g_m(\mathbf{x}) = \sum_{\boldsymbol{\xi} \in \mathbb{Z}^n, |\boldsymbol{\xi}| \le m} \hat{g}(\boldsymbol{\xi}) e^{2\pi i \mathbf{x}\cdot\boldsymbol{\xi}}.$$

Evidently, $g_m \in \mathcal{C}^\infty_\#$, $\hat{g}_m(\boldsymbol{\xi}) = \hat{g}(\boldsymbol{\xi})$ if $|\boldsymbol{\xi}| \le m$ and $\hat{g}_m(\boldsymbol{\xi}) = 0$ if $|\boldsymbol{\xi}| > m$. This implies that $\|g - g_m\|_{H^s_\#} \to 0$ as $m \to \infty$ and hence we can write

$$g(\mathbf{x}) = \sum_{\boldsymbol{\xi} \in \mathbb{Z}^n} \hat{g}(\boldsymbol{\xi}) e^{2\pi i \mathbf{x}\cdot\boldsymbol{\xi}}, \tag{16.11}$$

where the Fourier series converges in the sense of norm (16.10). Moreover, since $g$ is an arbitrary distribution from $H^s_\#$, this also implies that the space $\mathcal{C}^\infty_\#$ is dense in $H^s_\#$ for any $s \in \mathbb{R}$ (cf. [RT10, Exercise 3.2.9]).

There holds the compact embedding $H^t_\# \hookrightarrow H^s_\#$ if $t > s$, embeddings $H^s_\# \subset \mathcal{C}^m_\#$ if $m \in \mathbb{N}_0$, $s > m + n/2$, and moreover, $\bigcap_{s\in\mathbb{R}} H^s_\# = \mathcal{C}^\infty_\#$ (cf. [RT10, Exercises 3.2.10, 3.2.10 and Corollary 3.2.11]). Note also that the torus norms on $H^s_\#$ are equivalent to the corresponding standard (non-periodic) Bessel potential norms on $\mathbb{T}$ as a cubic domain, see, e.g., [Agr15, Section 13.8.1].

By (16.10), $\|g\|^2_{H^s_\#} = |\widehat{g}(\mathbf{0})|^2 + |g|^2_{H^s_\#}$, where

$$|g|_{H^s_\#} := \|\rho^s \widehat{g}\|_{\ell_2} := \left( \sum_{\boldsymbol{\xi} \in \dot{\mathbb{Z}}^n} \rho(\boldsymbol{\xi})^{2s} |\widehat{g}(\boldsymbol{\xi})|^2 \right)^{1/2}$$

is the seminorm in $H^s_\#$.

For any $s \in \mathbb{R}$, let us also introduce the space $\dot{H}^s_\# := \{g \in H^s_\# : \langle g, 1\rangle_{\mathbb{T}} = 0\}$. The definition implies that if $g \in \dot{H}^s_\#$, then $\widehat{g}(\mathbf{0}) = 0$ and

$$\|g\|_{\dot{H}^s_\#} = \|g\|_{H^s_\#} = |g|_{H^s_\#} = \|\rho^s \widehat{g}\|_{\ell_2}\,. \tag{16.12}$$

Denoting $\dot{\mathcal{C}}^\infty_\# := \{g \in \mathcal{C}^\infty_\# : \langle g, 1\rangle_{\mathbb{T}} = 0\}$, then $\bigcap_{s\in\mathbb{R}} \dot{H}^s_\# = \dot{\mathcal{C}}^\infty_\#$.

The corresponding spaces of $n$-component vector functions/distributions are denoted as $\mathbf{H}^s_\# := (H^s_\#)^n$, etc.

Note that the norm $\|\nabla(\cdot)\|_{\mathbf{H}^0_\#}$ is an equivalent norm in $\dot{H}^1_\#$. Indeed, by (16.11)

$$\nabla g(\mathbf{x}) = 2\pi i \sum_{\boldsymbol{\xi} \in \dot{\mathbb{Z}}^n} \boldsymbol{\xi} e^{2\pi i \mathbf{x}\cdot\boldsymbol{\xi}} \hat{g}(\boldsymbol{\xi}), \quad \widehat{\nabla g}(\boldsymbol{\xi}) = 2\pi i \boldsymbol{\xi} \hat{g}(\boldsymbol{\xi})$$

and then (16.9) and (16.12) imply

$$\begin{aligned} 2\pi^2 \|g\|^2_{H^1_\#} &= 2\pi^2 \|g\|^2_{\dot{H}^1_\#} = 2\pi^2 |g|^2_{H^1_\#} \le \|\nabla g\|^2_{\mathbf{H}^0_\#} \\ &\le 4\pi^2 |g|^2_{H^1_\#} = 4\pi^2 \|g\|^2_{\dot{H}^1_\#} = 4\pi^2 \|g\|^2_{H^1_\#} \quad \forall\, g \in \dot{H}^1_\#. \end{aligned} \tag{16.13}$$

The vector counterpart of (16.13) takes form

$$2\pi^2\|\mathbf{v}\|^2_{\mathbf{H}^1_\#} = 2\pi^2\|\mathbf{v}\|^2_{\dot{\mathbf{H}}^1_\#} \le \|\nabla\mathbf{v}\|^2_{(H^0_\#)^{n\times n}} \le 4\pi^2\|\mathbf{v}\|^2_{\dot{\mathbf{H}}^1_\#} = 4\pi^2\|\mathbf{v}\|^2_{\mathbf{H}^1_\#} \quad \forall\,\mathbf{v}\in\dot{\mathbf{H}}^1_\#. \tag{16.14}$$

We will further need also the first Korn inequality

$$\|\nabla\mathbf{v}\|^2_{(L_{2\#})^{n\times n}} \le 2\|\mathbb{E}(\mathbf{v})\|^2_{(L_{2\#})^{n\times n}} \quad \forall\,\mathbf{v}\in\mathbf{H}^1_\# \tag{16.15}$$

that can be easily proved by adapting, e.g., the proof in [McL00, Theorem 10.1]) to the periodic Sobolev space.

Let us define the Sobolev spaces of divergence-free functions/distributions,

$$\dot{\mathbf{H}}^s_{\#\sigma} := \left\{\mathbf{w}\in\dot{\mathbf{H}}^s_\# : \operatorname{div}\mathbf{w}=0\right\}, \quad s\in\mathbb{R},$$

endowed with the same norm (16.10).

## 16.4 Stationary Anisotropic Stokes System on Flat Torus

In this section, we generalise to the isotropic and anisotropic (linear) Stokes systems in compressible framework and to a range of Sobolev spaces the analysis, available in [Te95, Section 2.2]

For the unknowns $(\mathbf{u},\,p)\in\dot{\mathbf{H}}^s_\#\times\dot{H}^{s-1}_\#$ and the given data $(\mathbf{f},\,g)\in\dot{\mathbf{H}}^{s-2}_\#\times\dot{H}^{s-1}_\#$, $s\in\mathbb{R}$, let us consider the Stokes system

$$-\mathscr{L}(\mathbf{u},\,p) = \mathbf{f}, \tag{16.16}$$

$$\operatorname{div}\mathbf{u} = g, \tag{16.17}$$

that should be understood in the sense of distributions, i.e.,

$$-\langle\mathscr{L}(\mathbf{u},\,p),\boldsymbol{\phi}\rangle_{\mathbb{T}} = \langle\mathbf{f},\boldsymbol{\phi}\rangle_{\mathbb{T}} \quad \forall\,\boldsymbol{\phi}\in(\mathcal{C}^\infty_\#)^n, \tag{16.18}$$

$$\langle\operatorname{div}\mathbf{u},\phi\rangle_{\mathbb{T}} = \langle g,\phi\rangle_{\mathbb{T}} \quad \forall\,\phi\in\mathcal{C}^\infty_\#. \tag{16.19}$$

For $\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n$, let us employ $\bar{e}_{\boldsymbol{\xi}}(\mathbf{x}) = e^{-2\pi i x\cdot\boldsymbol{\xi}}$ as $\phi$ in (16.19) and $\bar{e}_{\boldsymbol{\xi}}(\mathbf{x})$, multiplied by the unit coordinate vector, as $\boldsymbol{\phi}$ in (16.18). Then recalling (16.3) and (16.4), we arrive for each $\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n$ at the following algebraic system for the Fourier coefficients, $\hat{u}_j(\boldsymbol{\xi})$, $k=1,2,\dots,n$, and $\hat{p}(\boldsymbol{\xi})$.

$$4\pi^2\xi_\alpha a^{\alpha\beta}_{kj}\xi_\beta\hat{u}_j(\boldsymbol{\xi}) + 2\pi i\xi_k\hat{p}(\boldsymbol{\xi}) = \hat{f}_k(\boldsymbol{\xi}) \quad \forall\,\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n,\ k=1,2,\dots,n \tag{16.20}$$

$$2\pi i\xi_j\hat{u}_j(\boldsymbol{\xi}) = \hat{g}(\boldsymbol{\xi}) \quad \forall\,\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n. \tag{16.21}$$

The $(n+1)\times(n+1)$ matrix, $\mathfrak{S}(\boldsymbol{\xi})$, of system (16.20)–(16.21) is in fact the principal symbol of the anisotropic Stokes system (16.16)–(16.17) that was analysed in [KMW21b, Lemma 15] to prove that the Stokes system is elliptic in the sense of Agmon–Douglis–Nirenberg. It was, particularly proved that the matrix $\mathfrak{S}$ is nonsingular if $\boldsymbol{\xi}\neq 0$ and hence the solution of system (16.20) and (16.21) can be represented in terms of the inverse matrix $\mathfrak{S}^{-1}(\boldsymbol{\xi})$ as

$$\begin{pmatrix}\widehat{\mathbf{u}}(\boldsymbol{\xi})\\ \hat{p}(\boldsymbol{\xi})\end{pmatrix}=\mathfrak{S}^{-1}(\boldsymbol{\xi})\begin{pmatrix}\widehat{\mathbf{f}}(\boldsymbol{\xi})\\ \hat{g}(\boldsymbol{\xi})\end{pmatrix}\quad\forall\,\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n. \tag{16.22}$$

Moreover, using the estimates for the matrix, obtained in that lemma proof, and implementing to the algebraic system the variant of Babuska–Brezzi theory given in Theorem 2.34 and Remark 2.35(i) in [EG04], see also [KMW21b, Theorem 10], we obtain the following estimates for the solution of the algebraic system (16.20)–(16.21),

$$|\widehat{\mathbf{u}}(\boldsymbol{\xi})|\leq C_{uf}\frac{|\widehat{\mathbf{f}}(\boldsymbol{\xi})|}{|2\pi\boldsymbol{\xi}|^2}+C_{ug}\frac{|\hat{g}(\boldsymbol{\xi})|}{2\pi|\boldsymbol{\xi}|}, \tag{16.23}$$

$$|\hat{p}(\boldsymbol{\xi})|\leq C_{pf}\frac{|\widehat{\mathbf{f}}(\boldsymbol{\xi})|}{2\pi|\boldsymbol{\xi}|}+C_{pg}|\hat{g}(\boldsymbol{\xi})|\quad\forall\,\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n, \tag{16.24}$$

where $C_{uf}=2C_{\mathbb{A}}$, $C_{ug}=C_{pf}=1+2C_{\mathbb{A}}\|\mathbb{A}\|$, $C_{pg}=\|\mathbb{A}\|(1+2C_{\mathbb{A}}\|\mathbb{A}\|)$.

*Remark 16.1* For the isotropic case (16.7), due to (16.8), system (16.20)–(16.21) reduces to

$$4\pi^2\Big[(\lambda+\mu)\boldsymbol{\xi}(\boldsymbol{\xi}\cdot\widehat{\mathbf{u}}(\boldsymbol{\xi}))+\mu|\boldsymbol{\xi}|^2\widehat{\mathbf{u}}(\boldsymbol{\xi})\Big]+2\pi i\boldsymbol{\xi}\hat{p}(\boldsymbol{\xi})=\widehat{\mathbf{f}}(\boldsymbol{\xi}),\quad\forall\,\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n, \tag{16.25}$$

$$2\pi i\boldsymbol{\xi}\cdot\widehat{\mathbf{u}}(\boldsymbol{\xi})=\hat{g}(\boldsymbol{\xi})\quad\forall\,\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n. \tag{16.26}$$

Taking scalar product of Eq. (16.25) with $\boldsymbol{\xi}$ and employing (16.26), we obtain

$$\hat{p}(\boldsymbol{\xi})=\frac{\boldsymbol{\xi}\cdot\widehat{\mathbf{f}}(\boldsymbol{\xi})}{2\pi i|\boldsymbol{\xi}|^2}+(\lambda+2\mu)\hat{g}(\boldsymbol{\xi}),\quad\forall\,\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n, \tag{16.27}$$

and substituting this back to (16.25), we get

$$\widehat{\mathbf{u}}(\boldsymbol{\xi})=\frac{1}{4\pi^2\mu|\boldsymbol{\xi}|^2}\left[\widehat{\mathbf{f}}(\boldsymbol{\xi})-\boldsymbol{\xi}\frac{\boldsymbol{\xi}\cdot\widehat{\mathbf{f}}(\boldsymbol{\xi})}{|\boldsymbol{\xi}|^2}\right]+\boldsymbol{\xi}\frac{\hat{g}(\boldsymbol{\xi})}{2\pi i|\boldsymbol{\xi}|^2},\quad\forall\,\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n \tag{16.28}$$

(cf. [Te95, Section 2.2] for the case $s=1$, $g=0$, $\lambda=0$, and $\mu=1$). Expressions (16.27) and (16.28) evidently satisfy estimates (16.23) and (16.24). □

The anisotropic Stokes system (16.16) and (16.17) can be re-written as

$$S\begin{pmatrix}\mathbf{u}\\ p\end{pmatrix}=\begin{pmatrix}\mathbf{f}\\ g\end{pmatrix},$$

where

$$S\begin{pmatrix}\mathbf{u}\\ p\end{pmatrix}:=\begin{pmatrix}-\mathscr{L}(\mathbf{u},p)\\ \operatorname{div}\mathbf{u}\end{pmatrix},$$

and for any $s\in\mathbb{R}$,

$$S:\dot{\mathbf{H}}^s_{\#}\times\dot{H}^{s-1}_{\#}\to\dot{\mathbf{H}}^{s-2}_{\#}\times\dot{H}^{s-1}_{\#} \tag{16.29}$$

is a linear continuous operator.

Now we are in the position to prove the following assertion.

**Theorem 16.11** *Let condition* (16.2) *hold.*

*(i) For any* $(\mathbf{f},g)\in\dot{\mathbf{H}}^{s-2}_{\#}\times\dot{H}^{s-1}_{\#}$, $s\in\mathbb{R}$, *the anisotropic Stokes system* (16.16)–(16.17) *in torus* $\mathbb{T}$ *has a unique solution* $(\mathbf{u},p)\in\dot{\mathbf{H}}^s_{\#}\times\dot{H}^{s-1}_{\#}$, *where*

$$\mathbf{u}(\mathbf{x})=\sum_{\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n}e^{2\pi i x\cdot\boldsymbol{\xi}}\widehat{\mathbf{u}}(\boldsymbol{\xi}),\quad p(\mathbf{x})=\sum_{\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n}e^{2\pi i x\cdot\boldsymbol{\xi}}\hat{p}(\boldsymbol{\xi}) \tag{16.30}$$

*with* $\widehat{\mathbf{u}}(\boldsymbol{\xi})$ *and* $\hat{p}(\boldsymbol{\xi})$ *given by* (16.22). *In addition, there exists a constant* $C=C(C_{\mathbb{A}},n)>0$ *such that*

$$\|\mathbf{u}\|_{\dot{\mathbf{H}}^s_{\#}}+\|p\|_{\dot{H}^{s-1}_{\#}}\le C\left(\|\mathbf{f}\|_{\dot{\mathbf{H}}^{s-2}_{\#}}+\|g\|_{\dot{H}^{s-1}_{\#}}\right) \tag{16.31}$$

*and operator* (16.29) *is an isomorphism.*

*(ii) Moreover, if* $(\mathbf{f},g)\in(\dot{\mathcal{C}}^\infty_{\#})^n\times\dot{\mathcal{C}}^\infty_{\#}$ *then* $(\mathbf{u},p)\in(\dot{\mathcal{C}}^\infty_{\#})^n\times\dot{\mathcal{C}}^\infty_{\#}$.

***Proof***

(i) Expressions (16.22) supplemented by the relations $\widehat{\mathbf{u}}(\mathbf{0})=\mathbf{0}$, $\hat{p}(\mathbf{0})=0$ imply the uniqueness. From estimates (16.23) and (16.24) we obtain the estimate

$$\begin{aligned}\|\mathbf{u}\|_{\dot{\mathbf{H}}^s_{\#}}&=\left(\sum_{\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n}\rho(\boldsymbol{\xi})^{2s}|\widehat{\mathbf{u}}(\boldsymbol{\xi})|^2\right)^{1/2}\\&\le\frac{C_{uf}}{4\pi^2}\left(\sum_{\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n}\rho(\boldsymbol{\xi})^{2s}\frac{|\widehat{\mathbf{f}}(\boldsymbol{\xi})|^2}{|\boldsymbol{\xi}|^4}\right)^{1/2}+\frac{C_{ug}}{2\pi}\left(\sum_{\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n}\rho(\boldsymbol{\xi})^{2s}\frac{|\hat{g}(\boldsymbol{\xi})|^2}{|\boldsymbol{\xi}|^2}\right)^{1/2}\end{aligned}$$

$$= \frac{C_{uf}}{4\pi^2}\left(\sum_{\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n}\rho(\boldsymbol{\xi})^{2(s-2)}|\widehat{\mathbf{f}}(\boldsymbol{\xi})|^2\frac{\rho(\boldsymbol{\xi})^4}{|\boldsymbol{\xi}|^4}\right)^{1/2}$$

$$+ \frac{C_{ug}}{2\pi}\left(\sum_{\boldsymbol{\xi}\in\dot{\mathbb{Z}}^n}\rho(\boldsymbol{\xi})^{2(s-1)}|\hat{g}(\boldsymbol{\xi})|^2\frac{\rho(\boldsymbol{\xi})^2}{|\boldsymbol{\xi}|^2}\right)^{1/2}$$

$$\le \frac{C_{uf}}{2\pi^2}\|\mathbf{f}\|_{\dot{\mathbf{H}}^{s-2}_{\#}} + \frac{C_{ug}}{2\pi}\sqrt{2}\|g\|_{\dot{H}^{s-1}_{\#}}$$

and the similar estimate for $\|p\|_{\dot{H}^{s-1}_{\#}}$, which imply (16.31) and hence inclusions in the corresponding spaces.

(ii) The inclusion $(\mathbf{f}, g) \in (\dot{\mathcal{C}}^{\infty}_{\#})^n \times \dot{\mathcal{C}}^{\infty}_{\#}$ implies that $(\mathbf{f}, g) \in \dot{\mathbf{H}}^{s-2}_{\#} \times \dot{H}^{s-1}_{\#}$ for any $s \in \mathbb{R}$. Then by item (i), $(\mathbf{u}, p) \in \dot{\mathbf{H}}^{s}_{\#} \times \dot{H}^{s-1}_{\#}$ for any $s \in \mathbb{R}$ and hence $(\mathbf{u}, p) \in (\dot{\mathcal{C}}^{\infty}_{\#})^n \times \dot{\mathcal{C}}^{\infty}_{\#}$.

□

If $g = 0$ in (16.17), we can re-formulate the Stokes system (16.16)–(16.17) as one vector equation

$$-\mathscr{L}(\mathbf{u}, p) = \mathbf{f} \tag{16.32}$$

for the unknowns $(\mathbf{u}, p) \in \dot{\mathbf{H}}^{s}_{\#\sigma} \times \dot{H}^{s-1}_{\#}$ and the given data $\mathbf{f} \in \dot{\mathbf{H}}^{s-2}_{\#}$, $s \in \mathbb{R}$. Then Theorem 16.11 implies the following assertion.

**Corollary 16.1** *Let condition (16.2) hold.*

*(i) For any $\mathbf{f} \in \dot{\mathbf{H}}^{s-2}_{\#}$, $s \in \mathbb{R}$, the anisotropic Stokes equation (16.32) in torus $\mathbb{T}$ has a unique incompressible solution $(\mathbf{u}, p) \in \dot{\mathbf{H}}^{s}_{\#\sigma} \times \dot{H}^{s-1}_{\#}$, with $\widehat{\mathbf{u}}(\boldsymbol{\xi})$ and $\hat{p}(\boldsymbol{\xi})$ given by (16.22) and (16.30) (and particularly by (16.28), (16.27), and (16.30) for the isotropic case (16.7)) with $g = 0$. In addition, there exists a constant $C = C(C_{\mathbb{A}}, n) > 0$ such that*

$$\|\mathbf{u}\|_{\dot{\mathbf{H}}^{s}_{\#}} + \|p\|_{\dot{H}^{s-1}_{\#}} \le C\|\mathbf{f}\|_{\dot{\mathbf{H}}^{s-2}_{\#}}$$

*and the operator*

$$\mathscr{L} : \dot{\mathbf{H}}^{s}_{\#\sigma} \times \dot{H}^{s-1}_{\#} \to \dot{\mathbf{H}}^{s-2}_{\#}$$

*is an isomorphism.*

*(ii) Moreover, if $\mathbf{f} \in (\dot{\mathcal{C}}^{\infty}_{\#})^n$ then $(\mathbf{u}, p) \in (\dot{\mathcal{C}}^{\infty}_{\#})^n \times \dot{\mathcal{C}}^{\infty}_{\#}$.*

## 16.5 Stationary Anisotropic Navier-Stokes System with Constant Coefficients on Torus

### *16.5.1 Existence of a Weak Solution to Anisotropic Incompressible Navier-Stokes System on Torus*

In this section, we show the existence of a weak solution of the anisotropic Navier-Stokes system in the incompressible case with general data in $L^2$-based Sobolev spaces on the torus $\mathbb{T}$, for $n \in \{2, 3\}$. We use the well-posedness result established in Theorem 16.11 for the Stokes system on a torus and the following variant of the *Leray-Schauder fixed point theorem* (see, e.g., [GT01, Theorem 11.3]).

**Theorem 16.2** *Let $B$ denote a Banach space and $T : B \to B$ be a continuous and compact operator. If there exists a constant $M_0 > 0$ such that $\|x\|_B \leq M_0$ for every pair $(\mathbf{x}, \theta) \in B \times [0, 1]$ satisfying $\mathbf{x} = \theta T\mathbf{x}$, then the operator $T$ has a fixed point $\mathbf{x}_0$ (with $\|x_0\|_B \leq M_0$).*

Let us consider the Navier-Stokes system

$$-\mathscr{L}(\mathbf{u}, p) = \mathbf{f} - (\mathbf{u} \cdot \nabla)\mathbf{u}, \tag{16.33}$$

$$\operatorname{div} \mathbf{u} = 0, \tag{16.34}$$

for the couple of unknowns $(\mathbf{u}, p) \in \dot{\mathbf{H}}^1_{\#} \times \dot{H}^0_{\#}$ and the given data $\mathbf{f} \in \dot{\mathbf{H}}^{-1}_{\#}$. As for the Stokes system, the Navier-Stokes system (16.33) and (16.34) can be re-written as one vector equation

$$-\boldsymbol{\mathscr{L}}(\mathbf{u}, p) = \mathbf{f} - (\mathbf{u} \cdot \nabla)\mathbf{u} \tag{16.35}$$

for the unknowns $(\mathbf{u}, p) \in \dot{\mathbf{H}}^1_{\#\sigma} \times \dot{H}^0_{\#}$ and the given data $\mathbf{f} \in \dot{\mathbf{H}}^{-1}_{\#}$.

Let us denote the nonlinear operator as $\boldsymbol{B}$, i.e.,

$$\boldsymbol{B}\mathbf{w} := (\mathbf{w} \cdot \nabla)\mathbf{w}, \quad \forall\, \mathbf{w} \in \mathbf{H}^s_{\#},\ s \in \mathbb{R}. \tag{16.36}$$

**Theorem 16.3** *Let the operator $\boldsymbol{B} : \mathbf{w} \mapsto \boldsymbol{B}\mathbf{w}$ be defined by (16.36) and let $n \geq 2$.*

*(i) If $0 < s < n/2$ then*

$$\boldsymbol{B} : \dot{\mathbf{H}}^s_{\#\sigma} \to \dot{\mathbf{H}}^{2s-1-n/2}_{\#} \tag{16.37}$$

*is a well defined, continuous and bounded quadratic operator, i.e., there exists $C_{n,s} > 0$ such that*

$$\|\boldsymbol{B}\mathbf{w}\|_{\mathbf{H}^{2s-1-n/2}_{\#}} \leq C_{n,s}\|\mathbf{w}\|^2_{\mathbf{H}^s_{\#}} \quad \forall\, \mathbf{w} \in \mathbf{H}^s_{\#}. \tag{16.38}$$

*(ii) If $s > n/2$ then*

$$\boldsymbol{B} : \dot{\mathbf{H}}^s_{\#\sigma} \to \dot{\mathbf{H}}^{s-1}_{\#} \tag{16.39}$$

*is well defined, continuous and bounded quadratic operator, i.e., there exists $C_{n,s} > 0$ such that*

$$\|\boldsymbol{B}\mathbf{w}\|_{\mathbf{H}^{s-1}_{\#}} \le C_{n,s}\|\mathbf{w}\|^2_{\mathbf{H}^s_{\#}} \quad \forall\, \mathbf{w} \in \mathbf{H}^s_{\#}. \tag{16.40}$$

***Proof*** If a function $\mathbf{w}$ is periodic, then evidently the function $\boldsymbol{B}\mathbf{w}$ is periodic as well.

(i) Let $0 < s < n/2$. Due to Theorem 1(iii) in Section 4.6.1 of [RS96] and equivalence of the Bessel potential norms on square and norms (16.10) for the Sobolev spaces on torus, we have,

$$\|(\mathbf{v}_1 \cdot \nabla)\mathbf{v}_2\|_{\mathbf{H}^{2s-1-n/2}_{\#}} \le C_{n,s}\|\mathbf{v}_1\|_{\mathbf{H}^s_{\#}}\|\mathbf{v}_2\|_{\mathbf{H}^s_{\#}}, \quad \forall\, \mathbf{v}_1, \mathbf{v}_2 \in \mathbf{H}^s_{\#}. \tag{16.41}$$

for some constant $C_{n,s} > 0$. This particularly implies estimate (16.38).

Further, if $\mathbf{u} \in \dot{\mathbf{H}}^s_{\#\sigma}$ then

$$\langle \boldsymbol{B}\mathbf{u}, 1\rangle_{\mathbb{T}} = \langle \mathbf{u} \cdot \nabla)\mathbf{u}, 1\rangle_{\mathbb{T}} = -\langle (\operatorname{div}\mathbf{u})\mathbf{u}, 1\rangle_{\mathbb{T}} = \mathbf{0}$$

since $\operatorname{div}\mathbf{u} = 0$. Together with estimate (16.38) this implies that quadratic operator (16.37) is well defined and bounded.

Let $\mathbf{w}, \mathbf{w}' \in \dot{\mathbf{H}}^1_{\#\sigma}$. Then by (16.41) we obtain

$$\begin{aligned}
\left\|\boldsymbol{B}\mathbf{w} - \boldsymbol{B}\mathbf{w}'\right\|_{\mathbf{H}^{2s-1-n/2}_{\#}} &\le \left\|(\mathbf{w} \cdot \nabla)\mathbf{w} - (\mathbf{w}' \cdot \nabla)\mathbf{w}'\right\|_{\mathbf{H}^{2s-1-n/2}_{\#}} \\
&\le \left\|((\mathbf{w} - \mathbf{w}') \cdot \nabla)\mathbf{w} + (\mathbf{w}' \cdot \nabla)(\mathbf{w} - \mathbf{w}')\right\|_{\mathbf{H}^{2s-1-n/2}_{\#}} \\
&\le C_{n,s}\left\|\mathbf{w} - \mathbf{w}'\right\|_{\mathbf{H}^s_{\#}}\left(\|\mathbf{w}\|_{\mathbf{H}^s_{\#}} + \|\mathbf{w}'\|_{\mathbf{H}^s_{\#}}\right).
\end{aligned}$$

This estimate shows that operator (16.37) is continuous.

(ii) Let $s > n/2$. Due to Theorem 1(i) in Section 4.6.1 of [RS96] and equivalence of the Bessel potential norms and norms (16.10) for the Sobolev spaces on torus, we have,

$$\|(\mathbf{v}_1 \cdot \nabla)\mathbf{v}_2\|_{\mathbf{H}^{s-1}_{\#}} \le C_{n,s}\|\mathbf{v}_1\|_{\mathbf{H}^s_{\#}}\|\mathbf{v}_2\|_{\mathbf{H}^s_{\#}}, \quad \forall\, \mathbf{v}_1, \mathbf{v}_2 \in \mathbf{H}^s_{\#}.$$

for some constant $C_{n,s} > 0$. This particularly implies estimate (16.40) and then the boundedness of operator (16.39). By the same arguments as in item (i), one can prove that this operator is also well defined and continuous.

□

**Corollary 16.2** *Let $n \in \{2, 3\}$. Then the quadratic operator*

$$\boldsymbol{B} : \dot{\mathbf{H}}^1_{\#\sigma} \to \mathbf{H}^{-1}_{\#} \tag{16.42}$$

*is well defined, continuous, bounded and compact.*

***Proof*** Let $n = 3$. Due to Theorem 16.3(i), the operator $\boldsymbol{B} : \dot{\mathbf{H}}^1_{\#\sigma} \to \dot{\mathbf{H}}^{-1/2}_{\#}$ is well defined, continuous and bounded. On the other hand, the compactness of embedding $H^{-1/2}_{\#} \hookrightarrow H^{-1}_{\#}$ implies the compactness of embedding $\dot{H}^{-1/2}_{\#} \hookrightarrow \dot{H}^{-1}_{\#}$ and hence gives the compactness of operator (16.42) and thus the corollary claim for $n = 3$.

Let now $n = 2$. Then by Theorem 16.3(i), the operator $\boldsymbol{B} : \dot{\mathbf{H}}^s_{\#\sigma} \to \dot{\mathbf{H}}^{2s-2}_{\#}$ is well defined, continuous and bounded for any $s \in (1/2, 1)$. In addition, for $s \in (1/2, 1)$ we also have the compact embeddings $\dot{H}^1_{\#\sigma} \hookrightarrow \dot{H}^s_{\#\sigma}$ and $\dot{H}^{2s-2}_{\#} \hookrightarrow \dot{H}^{-1}_{\#}$ that lead to the corollary claim for $n = 2$. □

Next we show the existence of a weak solution of the Navier-Stokes equation.

**Theorem 16.4** *Let $n \in \{2, 3\}$ and suppose that condition (16.2) holds. If $\mathbf{f} \in \dot{\mathbf{H}}^{-1}_{\#}$, then the anisotropic Navier-Stokes equation (16.35) has a solution $(\mathbf{u}, p) \in \dot{\mathbf{H}}^1_{\#\sigma} \times \dot{H}^0_{\#}$.*

***Proof*** We will reduce the analysis of the nonlinear equation (16.35) to the analysis of a nonlinear operator in the Hilbert space $\dot{\mathbf{H}}^1_{\#\sigma}$ and show that this operator has a fixed-point due to the Leray-Schauder Theorem.

Nonlinear equation (16.35) can be re-written as

$$-\boldsymbol{\mathscr{L}}(\mathbf{u}, p) = \mathbf{f} - \boldsymbol{B}\mathbf{u}. \tag{16.43}$$

By Corollary 16.1, the linear operator

$$-\boldsymbol{\mathscr{L}} : \dot{\mathbf{H}}^1_{\#\sigma} \times \dot{H}^0_{\#} \to \dot{\mathbf{H}}^{-1}_{\#} \tag{16.44}$$

is an isomorphism. Its inverse operator, $-\boldsymbol{\mathscr{L}}^{-1}$, can be split into two operator components,

$$-\boldsymbol{\mathscr{L}}^{-1} = \begin{pmatrix} \boldsymbol{\mathscr{U}} \\ \mathcal{P} \end{pmatrix}$$

where $\boldsymbol{\mathscr{U}} : \dot{\mathbf{H}}^{-1}_{\#} \to \dot{\mathbf{H}}^1_{\#\sigma}$ and $\mathcal{P} : \dot{\mathbf{H}}^{-1}_{\#} \to \dot{H}^0_{\#}$ are linear continuous operators such that

$$-\boldsymbol{\mathscr{L}} \begin{pmatrix} \boldsymbol{\mathscr{U}} & \boldsymbol{\mathscr{F}} \\ \mathcal{P} & \mathscr{F} \end{pmatrix} = \boldsymbol{\mathscr{F}}$$

for any $\mathscr{F} \in \dot{\mathbf{H}}^{-1}_{\#}$. Applying the inverse operator, $-\mathscr{L}^{-1}$, to Eq. (16.43), we reduce it to the equivalent nonlinear system

$$\mathbf{u} = \mathbf{U}\mathbf{u}, \tag{16.45}$$

$$p = P\mathbf{u}, \tag{16.46}$$

where $\mathbf{U} : \dot{\mathbf{H}}^1_{\#\sigma} \to \dot{\mathbf{H}}^1_{\#\sigma}$ and $P : \dot{\mathbf{H}}^1_{\#\sigma} \to \dot{H}^0_{\#}$ are the nonlinear operators defined as

$$\mathbf{U}\mathbf{w} := \mathscr{U}(\mathbf{f} - \boldsymbol{B}\mathbf{w}), \tag{16.47}$$

$$P\mathbf{w} := \mathcal{P}(\mathbf{f} - \boldsymbol{B}\mathbf{w}) \tag{16.48}$$

for the fixed $\mathbf{f}$.

Since $p$ is not involved in (16.45), we will first prove the existence of a solution $\mathbf{u} \in \dot{\mathbf{H}}^1_{\#\sigma}$ to this equation. Then we use (16.46) as a representation formula for $p$, which gives the existence of the pressure field $p \in \dot{H}^0_{\#}$. In order to show the existence of a fixed point of the operator $\mathbf{U}$ and, thus, the existence of a solution of Eq. (16.45), we employ Theorem 16.2.

By Corollary 16.2, for $n \in \{2, 3\}$ the operator $\boldsymbol{B} : \dot{\mathbf{H}}^1_{\#\sigma} \to \mathbf{H}^{-1}_{\#}$ is bounded, continuous and compact. Since $\mathbf{f} \in \mathbf{H}^{-1}_{\#}$ is fixed and the operator $\mathscr{U} : \mathbf{H}^{-1}_{\#} \to \dot{\mathbf{H}}^1_{\#\sigma}$ is linear and continuos, definition (16.47) implies that the operator $\mathbf{U} : \dot{\mathbf{H}}^1_{\#\sigma} \to \dot{\mathbf{H}}^1_{\#\sigma}$ is also bounded, continuous, and compact.

Next, we show that there exists a constant $M_0 > 0$ such that if $\mathbf{w} \in \dot{\mathbf{H}}^1_{\#\sigma}$ satisfies the equation

$$\mathbf{w} = \theta\mathbf{U}\mathbf{w} \tag{16.49}$$

for some $\theta \in [0, 1]$, then $\|\mathbf{w}\|_{\dot{\mathbf{H}}^1_{\#\sigma}} \le M_0$. Let us denote

$$q := \theta P\mathbf{w}. \tag{16.50}$$

By applying the operator $-\mathscr{L}$ to Eqs. (16.49) and (16.50) and by using relations (16.47) and (16.48), we deduce that whenever the pair $(\mathbf{w}, \theta) \in \dot{\mathbf{H}}^1_{\#\sigma} \times \mathbb{R}$ satisfies Eq. (16.49), then the equation

$$-\mathscr{L}(\mathbf{w}, q) = \theta(\mathbf{f} - \boldsymbol{B}\mathbf{w}),$$

is also satisfied due to the isomorphism property of operator (16.44). This equation should be understood in the sense of distribution, i.e.,

$$\begin{aligned}\langle -\mathscr{L}(\mathbf{w}, q), \boldsymbol{\phi}\rangle_{\mathbb{T}} &= \left\langle a_{ij}^{\alpha\beta} E_{j\beta}(\mathbf{w}), E_{i\alpha}(\boldsymbol{\phi})\right\rangle_{\mathbb{T}} - \langle q, \operatorname{div}\boldsymbol{\phi}\rangle_{\mathbb{T}} \\ &= \theta\langle \mathbf{f} - \boldsymbol{B}\mathbf{w}, \boldsymbol{\phi}\rangle_{\mathbb{T}} \quad \forall\, \boldsymbol{\phi} \in (\mathcal{C}^\infty_{\#})^n. \end{aligned} \tag{16.51}$$

Taking into account that the space $(\mathcal{C}^\infty_\#)^n$ is dense in $\mathbf{H}^1_\#$ and the continuity of the dual products in (16.51) with respect to $\boldsymbol{\phi} \in \mathbf{H}^1_\#$, Eq. (16.51) should hold also for $\boldsymbol{\phi} = \mathbf{w} \in \dot{\mathbf{H}}^1_{\#\sigma}$. Then we obtain

$$\left\langle a^{\alpha\beta}_{ij} E_{j\beta}(\mathbf{w}),\, E_{i\alpha}(\mathbf{w}) \right\rangle_{\mathbb{T}} = \theta \langle \mathbf{f} - \boldsymbol{B}\mathbf{w}, \mathbf{w} \rangle_{\mathbb{T}}. \tag{16.52}$$

Since $\mathbf{w} \in \dot{\mathbf{H}}^1_{\#\sigma}$, relation (16.55) implies that $\langle \boldsymbol{B}\mathbf{w}, \mathbf{w} \rangle_{\mathbb{T}} = \langle (\mathbf{w} \cdot \nabla)\mathbf{w}, \mathbf{w} \rangle_{\mathbb{T}} = 0$. Then by using the norm equivalence (16.14), the Korn first inequality (16.15), the ellipticity condition (16.2), Eq. (16.52), and the Hölder inequality, we obtain for $\theta \geq 0$ that

$$\begin{aligned} \|\mathbf{w}\|^2_{\dot{\mathbf{H}}^1_\#} &\leq \frac{1}{2\pi^2} \|\nabla \mathbf{w}\|^2_{(L_{2\#})^{n\times n}} \leq \frac{1}{\pi^2} \|\mathbb{E}(\mathbf{w})\|^2_{(L_{2\#})^{n\times n}} \\ &\leq \frac{1}{\pi^2} C_{\mathbb{A}} \left\langle a^{\alpha\beta}_{ij} E_{j\beta}(\mathbf{w}),\, E_{i\alpha}(\mathbf{w}) \right\rangle_{\mathbb{T}} \leq \frac{\theta}{\pi^2} C_{\mathbb{A}} \|\mathbf{f}\|_{\dot{\mathbf{H}}^{-1}_\#} \|\mathbf{w}\|_{\dot{\mathbf{H}}^1_\#}. \end{aligned}$$

Hence, for $\theta \in [0, 1]$,

$$\|\mathbf{w}\|_{\dot{\mathbf{H}}^1_\#} \leq M_0 := \frac{1}{\pi^2} C_{\mathbb{A}} \|\mathbf{f}\|_{\dot{\mathbf{H}}^{-1}_\#}.$$

Therefore, the operator $\mathbf{U} : \dot{\mathbf{H}}^1_{\#\sigma} \to \dot{\mathbf{H}}^1_{\#\sigma}$ satisfies the hypothesis of Theorem 16.2 (with $B = \dot{\mathbf{H}}^1_{\#\sigma}$), and hence it has a fixed point $\mathbf{u} \in \dot{\mathbf{H}}^1_{\#\sigma}$, that is, $\mathbf{u} = \mathbf{U}\mathbf{u}$. Then with $p \in \dot{H}^0_\#$ as in (16.46), we obtain that the couple $(\mathbf{u}, p) \in \dot{\mathbf{H}}^1_{\#\sigma} \times \dot{H}^0_\#$ satisfies the nonlinear equation (16.35). □

### 16.5.2 *Solution Regularity for the Stationary Anisotropic Navier-Stokes System*

In this section, using the bootstrap argument we show that the regularity of a solution of the anisotropic incompressible Navier-Stokes system on $\mathbb{T}^n$, $n \in \{2, 3\}$, is completely determined by the regularity of its right-hand side, as for the Stokes system. To prove this we use the inclusions of the nonlinear term $\boldsymbol{B}\mathbf{u}$ given by Theorem 16.3 and the unique solvability of corresponding (linear) Stokes system.

**Theorem 16.5** *Let condition* (16.2) *hold. Let* $n \geq 2$ *and* $n/2 - 1 < s_1 < s_2$.

*(i) If* $(\mathbf{u}, p) \in \dot{\mathbf{H}}^{s_1}_{\#\sigma} \times \dot{H}^{s_1-1}_\#$ *is a solution of the anisotropic Navier-Stokes equation* (16.35) *with a right hand side* $\mathbf{f} \in \dot{\mathbf{H}}^{s_2-2}_\#$, *then* $(\mathbf{u}, p) \in \dot{\mathbf{H}}^{s_2}_{\#\sigma} \times \dot{H}^{s_2-1}_\#$.

*(ii) Moreover, if* $\mathbf{f} \in (\dot{\mathcal{C}}^\infty_\#)^n$ *then* $(\mathbf{u}, p) \in (\dot{\mathcal{C}}^\infty_\#)^n \times \dot{\mathcal{C}}^\infty_\#$.

***Proof***

(i) Let $(\mathbf{u}, p) \in \dot{\mathbf{H}}^{s_1}_{\#\sigma} \times \dot{H}^{s_1-1}_{\#}$ be a solution of (16.35) with $\mathbf{f} \in \dot{\mathbf{H}}^{s_2-2}_{\#}$. Then by Theorem 16.3, for the nonlinear term we have the inclusion $\boldsymbol{B}\mathbf{u} \in \dot{\mathbf{H}}^{t_1}_{\#}$ with $t_1 = 2s_1 - 1 - n/2$ if $s_1 < n/2$, with $t_1 = s_1 - 1$ if $s_1 > n/2$, and with any $t_1 \in (s_1 - 2, s_1 - 1)$ (and we can further use $t_1 = s_1 - 3/2$ for certainty) if $s_1 = n/2$. Hence the couple $(\mathbf{u}, p)$ satisfies the equation

$$-\mathscr{L}(\mathbf{u}, p) = \mathbf{f}^{(1)} \tag{16.53}$$

with $\mathbf{f}^{(1)} := \mathbf{f} - \boldsymbol{B}\mathbf{u} \in \dot{\mathbf{H}}^{s^{(1)}-2}_{\#}$, where $s^{(1)} = \min\{s_2, t_1 + 2\}$. By Corollary 16.1(i), the linear equation (16.53) has a unique solution in $\dot{\mathbf{H}}^{s}_{\#\sigma} \times \dot{H}^{s-1}_{\#}$ for any $s \le s^{(1)}$ and thus $(\mathbf{u}, p) \in \dot{\mathbf{H}}^{s^{(1)}}_{\#\sigma} \times \dot{H}^{s^{(1)}-1}_{\#}$. If $s^{(1)} = s_2$, which we call the case (a), this proves item (i) of the theorem.

Otherwise we have the case (b), when $s^{(1)} < s_2$, i.e., $s^{(1)} = t_1 + 2$, by the definition of $s^{(1)}$ and the theorem condition $s_1 > n/2 - 1$. Then we arrange an iterative process by replacing $s_1$ with $s^{(1)} = t_1 + 2$ on each iteration until we arrive at the case (a), thus proving item (i) of the theorem. Note that in the case (b),

$$s^{(1)} - s_1 \ge \delta := \min\{s_1 + 1 - n/2, 1, 1/2\} > 0$$

in the first iteration, and $\delta$ can only increase in the next iterations due to the increase of $s_1$. This implies that the iteration process will reach the case (a) and stop after a finite number of iterations.

(ii) If $\mathbf{f} \in (\dot{\mathcal{C}}^{\infty}_{\#})^n$, then for any $s_2 \in \mathbb{R}$ we have $\mathbf{f} \in \dot{\mathbf{H}}^{s_2-2}_{\#}$ and item (i) implies that $(\mathbf{u}, p) \in \dot{\mathbf{H}}^{s_2}_{\#\sigma} \times \dot{H}^{s_2-1}_{\#}$. Hence $(\mathbf{u}, p) \in (\dot{\mathcal{C}}^{\infty}_{\#})^n \times \dot{\mathcal{C}}^{\infty}_{\#}$.

□

Combining Theorems 16.4 and 16.5, we obtain the following assertion on existence and regularity of solution to the Navier-Stokes system on torus.

**Theorem 16.6** *Let $n \in \{2, 3\}$ and condition (16.2) hold.*

*(i) If $\mathbf{f} \in \dot{\mathbf{H}}^{s-2}_{\#}$, $s \ge 1$, then the anisotropic Navier-Stokes equation (16.35) has a solution $(\mathbf{u}, p) \in \dot{\mathbf{H}}^{s}_{\#\sigma} \times \dot{H}^{s-1}_{\#}$.*

*(ii) Moreover, if $\mathbf{f} \in (\dot{\mathcal{C}}^{\infty}_{\#})^n$ then (16.35) has a solution $(\mathbf{u}, p) \in (\dot{\mathcal{C}}^{\infty}_{\#})^n \times \dot{\mathcal{C}}^{\infty}_{\#}$.*

Note that in the *isotropic case* (16.7) with $\lambda = 0$, similar results for the Navier-Stokes system in torus as well as in domains of $\mathbb{R}^n$ are available, e.g., in [Ga11, RRS16, Se15, So01, Te01].

## 16.6 Some Auxiliary Results

The dense embedding of the space $(\mathcal{C}^\infty_\#)^n$ into $\mathbf{H}^1_\#$ and the divergence theorem imply the following identity for any $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in \mathbf{H}^1_\#$.

$$\begin{aligned}\langle(\mathbf{v}_1 \cdot \nabla)\mathbf{v}_2, \mathbf{v}_3\rangle_{\mathbb{T}} &= \int_{\mathbb{T}} \nabla \cdot (\mathbf{v}_1(\mathbf{v}_2 \cdot \mathbf{v}_3))\, d\mathbf{x} - \langle(\nabla \cdot \mathbf{v}_1)\mathbf{v}_3 + (\mathbf{v}_1 \cdot \nabla)\mathbf{v}_3, \mathbf{v}_2\rangle_{\mathbb{T}} \\ &= -\langle(\mathbf{v}_1 \cdot \nabla)\mathbf{v}_3, \mathbf{v}_2\rangle_{\mathbb{T}} - \langle(\nabla \cdot \mathbf{v}_1)\mathbf{v}_3, \mathbf{v}_2\rangle_{\mathbb{T}}\,. \end{aligned} \tag{16.54}$$

In view of (16.54) we obtain the identity

$$\langle(\mathbf{v}_1 \cdot \nabla)\mathbf{v}_2, \mathbf{v}_3\rangle_{\mathbb{T}} = -\langle(\mathbf{v}_1 \cdot \nabla)\mathbf{v}_3, \mathbf{v}_2\rangle_{\mathbb{T}} \quad \forall\, \mathbf{v}_1 \in \mathbf{H}^1_{\#\sigma},\ \mathbf{v}_2,\ \mathbf{v}_3 \in \mathbf{H}^1_\#\,,$$

and hence the well known formula

$$\langle(\mathbf{v}_1 \cdot \nabla)\mathbf{v}_2, \mathbf{v}_2\rangle_{\mathbb{T}} = 0 \quad \forall\, \mathbf{v}_1 \in \mathbf{H}^1_{\#\sigma},\ \mathbf{v}_2 \in \mathbf{H}^1_\#. \tag{16.55}$$

## References


[Agm65] Agmon, S.: Lectures on Elliptic Boundary Value Problems. Van Nostrand, New York (1965)

[Agr15] Agranovich, M.S.: Sobolev Spaces, Their Generalizations, and Elliptic Problems in Smooth and Lipschitz Domains. Springer, Berlin (2015)

[CF88] Constantin, P., Foias, C.: Navier-Stokes Equations. The University of Chicago Press, Chicago (1988)

[Duf78] Duffy, B.R.: Flow of a liquid with an anisotropic viscosity tensor. J. Nonnewton. Fluid Mech. **4**, 177–193 (1978)

[Ga11] Galdi, G.P.: An Introduction to the Mathematical Theory of the Navier–Stokes Equations. Steady-State Problems, 2nd edn. Springer, New York (2011)

[EG04] Ern, A., Guermond, J.L.: Theory and Practice of Finite Elements. Springer, New York (2004)

[GT01] Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order. Springer, Berlin (2001)

[KMW20] Kohr, M., Mikhailov, S.E., Wendland, W.L.: Potentials and transmission problems in weighted Sobolev spaces for anisotropic Stokes and Navier-Stokes systems with $L_\infty$ strongly elliptic coefficient tensor. Complex Var. Elliptic Equ. **65**, 109–140 (2020)

[KMW21a] Kohr, M., Mikhailov, S.E., Wendland, W.L.: Dirichlet and transmission problems for anisotropic Stokes and Navier-Stokes systems with $L_\infty$ tensor coefficient under relaxed ellipticity condition. Discrete Contin. Dyn. Syst. Ser. A. **41**, 4421–4460 (2021)

[KMW21b] Kohr, M., Mikhailov, S.E., Wendland, W.L.: Layer potential theory for the anisotropic Stokes system with variable $L_\infty$ symmetrically elliptic tensor coefficient. Math. Methods Appl. Sci. **44**, 9641–9674 (2021)

[McL91] McLean, W.: Local and global descriptions of periodic pseudodifferential operators. Math. Nachr. **150**, 151–161 (1991)

[McL00] McLean, W.: Strongly Elliptic Systems and Boundary Integral Equations. Cambridge University Press, Cambridge (2000)

[OSY92] Oleinik, O.A., Shamaev, A.S., Yosifian, G.A.: Mathematical Problems in Elasticity and Homogenization. Horth-Holland, Amsterdam (1992)
[RRS16] Robinson, J.C., Rodrigo, J.L., Sadowski, W.: The Three-Dimensional Navier–Stokes equations. Classical Theory. Cambridge University Press, Cambridge (2016)
[RS96] Runst, T., Sickel, W.: Sobolev Spaces of Fractional Order, Nemytskij Operators, and Nonlinear Partial Differential Equations. De Gruyter, Berlin (1996)
[RT10] Ruzhansky M., Turunen,V.: Pseudo-Differential Operators and Symmetries: Background Analysis and Advanced Topics. Birkhäuser, Basel (2010)
[Se15] Seregin, G.: Lecture Notes on Regularity Theory for the Navier-Stokes Equations. World Scientific, London (2015)
[So01] Sohr, H.: The Navier-Stokes Equations: An Elementary Functional Analytic Approach. Springer, Basel (2001)
[Te95] Temam, R.: Navier-Stokes Equations and Nonlinear Functional Analysis. SIAM, Philadelphia (1995)
[Te01] Temam, R.: Navier-Stokes Equations. Theory and Numerical Analysis. AMS Chelsea Edition. American Mathematical Society, Providence (2001)
[Zy02] Zygmund, A.: Trigonometric Series, vol. II, 3rd edn. Cambridge University Press, Cambridge (2002)

# Chapter 17
# Null-Solutions of Elliptic Partial Differential Equations with Power Growth

**D. Mitrea, I. Mitrea, and M. Mitrea**

## 17.1 Introduction and Statement of Main Result

Let $\mathbb{N}$ denote the collection of natural numbers and set $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. We shall typically assume that $n \in \mathbb{N}$ with $n \geq 2$. Define the length of any given multi-index $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$ as $|\alpha| := \sum_{j=1}^n \alpha_j$, and set $\alpha! := \alpha_1! \cdots \alpha_n!$. Corresponding to each $j \in \{1, \ldots, n\}$, denote by $e_j$ the multi-index in $\mathbb{N}_0^n$ of length one and with the $j$-th component equal to 1.

Next, fix $m, M \in \mathbb{N}$ and consider an $M \times M$ system $L$ of homogeneous differential operators of order $2m$, with constant (complex) coefficients in $\mathbb{R}^n$. Hence, there exists a family of coefficient tensors $A = \{A_{\alpha\beta}\}_{|\alpha|=|\beta|=m}$ consisting of matrices $A_{\alpha\beta}$ in $\mathbb{C}^{M\times M}$ indexed by multi-indices $\alpha, \beta \in \mathbb{N}_0^n$ with $|\alpha| = |\beta| = m$ such that the system $L$ may be expressed as

$$L = \sum_{|\alpha|=|\beta|=m} \partial^\alpha A_{\alpha\beta}\, \partial^\beta. \tag{17.1}$$

In this work we further assume that $L$ is weakly elliptic in the sense that its characteristic matrix

$$L(\xi) := \sum_{|\alpha|=|\beta|=m} A_{\alpha\beta}\, \xi^{\alpha+\beta} \quad \text{for} \quad \xi \in \mathbb{R}^n \tag{17.2}$$

D. Mitrea (✉) · M. Mitrea
Department of Mathematics, Baylor University, Waco, TX, USA
e-mail: Dorina_Mitrea@Baylor.edu; Marius_Mitrea@Baylor.edu

I. Mitrea
Department of Mathematics, Temple University, Philadelphia, PA, USA
e-mail: imitrea@temple.edu



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_17

is invertible in $\mathbb{R}^n \setminus \{0\}$, i.e.,

$$\det\Big[\sum_{|\alpha|=|\beta|=m} A_{\alpha\beta}\,\xi^{\alpha+\beta}\Big] \neq 0 \quad \text{for all} \quad \xi \in \mathbb{R}^n \setminus \{0\}. \tag{17.3}$$

Relevant for our work is the fact that such a weakly elliptic system $L$ has a fundamental solution $E_L$ of the sort described in Proposition 17.1.

The reader is reminded that an arbitrary set $\Omega \subset \mathbb{R}^n$ is said to be an exterior domain provided $\mathbb{R}^n \setminus \Omega$ is a compact set. The main result in this work, elucidating the asymptotic behavior at infinity for null-solutions of weakly elliptic systems in exterior domains with at most polynomial growth at infinity, is stated next.

**Theorem 17.1** *Fix $n, m, M \in \mathbb{N}$, $n \geq 2$. Let $L$ be a homogeneous $M \times M$ system of differential operators of order $2m$ in $\mathbb{R}^n$, with constant complex coefficients. Assume $L$ is weakly elliptic in the sense of (17.3). Let $\Omega \subseteq \mathbb{R}^n$ be an exterior domain and consider a function $u \in [\mathscr{C}^\infty(\Omega)]^M$ satisfying $Lu = 0$ in $\Omega$, for which there exists $N \in \mathbb{R}$ such that*

$$u(x) = o\big(|x|^N\big) \quad \text{as} \quad |x| \to \infty. \tag{17.4}$$

*Then there exists some polynomial $P$ in $\mathbb{R}^n$ satisfying $LP = 0$ in $\mathbb{R}^n$ and with the property that*

$$u - P = \begin{cases} O\Big(\frac{1}{|x|^{n-2m}}\Big) & \text{if } n \text{ is odd or if } n > 2m, \\ O\big(|x|^{2m-n}\ln|x|\big) & \text{if } n \text{ is even and } n \leq 2m, \end{cases} \quad \text{as } |x| \to \infty. \tag{17.5}$$

*In addition,*

$$\begin{cases} P = 0 & \text{if } N \in (-\infty, 0), \\ P = 0 & \text{if } N = 0 \text{ and either } n \text{ is odd, or } n > 2m, \\ \deg P \leq [N] & \text{if } N \in (0,\infty) \setminus \mathbb{N}, \\ \deg P \leq N-1 & \text{if } N \in \mathbb{N} \text{ and either } n \text{ is odd, or } n > 2m, \\ \deg P \leq N & \text{if } N \in \mathbb{N}_0 \text{ and } n \text{ is even and satisfies } n \leq 2m. \end{cases} \tag{17.6}$$

The regime where the above theorem is effective is $N > 2m - n$, since in such a scenario $u - P$ decays faster (or grows slower) than $u$ was originally assumed to decay (or grow). In particular, this shows that, in this range, the leading term in the asymptotic expansion at infinity of null-solutions of a weakly elliptic system in an exterior domain is always a polynomial. This is remarkable since these null-solutions can, globally, be very different from polynomial functions.

Even though the statement does not reflect it, our proof of Theorem 17.1 crucially relies on boundary layer potentials for higher-order weakly elliptic systems, as

introduced and studied in [MM13] and [HLM21]. Two other key ingredients in the proof are as follows: first, we shall employ a version of the Divergence Theorem for singular vector fields which has recently appeared in [MMM20]; second, we shall need precise estimates for fundamental solutions of higher-order weakly elliptic systems, of the sort obtained in [DM18]. For the reader's convenience, all these are reviewed in the next section.

Here, we conclude by including the following useful corollary, itself an immediate consequence of our main result.

**Corollary 17.1** *Fix $n, m, M \in \mathbb{N}$, with $n \geq 2$. Let $L$ be a homogeneous $M \times M$ system of differential operators of order $2m$ in $\mathbb{R}^n$, with constant complex coefficients which is weakly elliptic in the sense of (17.3). Suppose $\Omega$ is an exterior domain in $\mathbb{R}^n$.*

*For $N \in \mathbb{R}$ consider the boundary value problem*

$$\begin{cases} u \in [\mathscr{C}^\infty(\Omega) \cap \mathscr{C}^{m-1}(\overline{\Omega})]^M, \\ Lu = 0 \ \text{ in } \ \Omega, \\ (\partial^\gamma u)\big|_{\partial\Omega} = f_\gamma \in [\mathscr{C}^0(\partial\Omega)]^M, \ |\gamma| \leq m-1, \\ u(x) = o\big(|x|^N\big) \ \text{ as } \ |x| \to \infty. \end{cases} \tag{17.7}$$

*Then any solution $u$ of this boundary value problem is of the form*

$$u = P + u_0, \tag{17.8}$$

*where $P$ is a polynomial in $\mathbb{R}^n$ satisfying $LP = 0$ in $\mathbb{R}^n$ and (17.6), and $u_0$ is a solution of the boundary value problem*

$$\begin{cases} u_0 \in [\mathscr{C}^\infty(\Omega) \cap \mathscr{C}^{m-1}(\overline{\Omega})]^M, \quad Lu_0 = 0 \ \text{ in } \ \Omega, \\ (\partial^\gamma u_0)\big|_{\partial\Omega} = f_\gamma - (\partial^\gamma P)\big|_{\partial\Omega} \ \text{ for } \ |\gamma| \leq m-1, \end{cases} \tag{17.9}$$

*whose behavior at infinity is given by*

$$u_0(x) = \begin{cases} O\Big(\frac{1}{|x|^{n-2m}}\Big) & \text{if } n \text{ is odd, or if } n > 2m, \\ O\big(|x|^{2m-n} \ln |x|\big) & \text{if } n \text{ is even and } n \leq 2m, \end{cases} \quad \text{as } |x| \to \infty. \tag{17.10}$$

## 17.2 Preliminaries

We start by recalling that weakly elliptic systems as in (17.1) and (17.3) have fundamental solutions as described in the proposition below. Before stating it, the

reader is reminded that $\Delta$ stands for the Laplacian in $\mathbb{R}^n$ (with $\Delta_x$ indicating that this operator acts in the variable $x$), and $\mathscr{H}^{n-1}$ denotes the $(n-1)$-dimensional Hausdorff measure in $\mathbb{R}^n$. Also, 'dot' denotes the inner product of vectors in $\mathbb{R}^n$. Finally, $\mathscr{D}'$, $\mathscr{E}'$, $\mathscr{S}'$ stand, respectively, for the space of distributions, compactly supported distributions, and tempered distributions. For a proof of Proposition 17.1 see [DM18, Theorem 11.1, pp. 393–395].

**Proposition 17.1** *Fix $n, m, M \in \mathbb{N}$, with $n \geq 2$. Let $L$ be a homogeneous $M \times M$ system of differential operators of order $2m$ in $\mathbb{R}^n$, with constant complex coefficients and let $A = \{A_{\alpha\beta}\}_{|\alpha|=|\beta|=m}$ be a family of coefficient tensors associated with $L$ in the sense of (17.1). Suppose $L$ is weakly elliptic in the sense of (17.3).*

*Consider the $M \times M$ matrix $E_L$ defined at each point $x \in \mathbb{R}^n \setminus \{0\}$ by*

$$E_L(x) := \frac{\Delta_x^{(n-1)/2}}{4(2\pi\, i)^{n-1}} \left\{ \int_{S^{n-1}} \frac{|x \cdot \xi|^{2m-1}}{(2m-1)!} \big[L(\xi)\big]^{-1} d\mathscr{H}^{n-1}(\xi) \right\} \tag{17.11}$$

*if $n$ is odd,*

*and*

$$E_L(x) := \frac{-\Delta_x^{n/2}}{(2\pi\, i)^n} \left\{ \int_{S^{n-1}} \frac{(x \cdot \xi)^{2m}}{(2m)!} \ln |\langle x, \xi \rangle| \, \big[L(\xi)\big]^{-1} d\mathscr{H}^{n-1}(\xi) \right\} \tag{17.12}$$

*if $n$ is even.*

*In relation to the $\mathbb{C}^{M\times M}$-valued function (17.11)-(17.12), the following properties hold:*

*(1) With $\mathscr{S}'(\mathbb{R}^n)$ denoting the space of tempered distributions in $\mathbb{R}^n$, one has*

$$\begin{aligned} &E_L \in \big[\mathscr{C}^\infty(\mathbb{R}^n \setminus \{0\}) \cap L^1_{\rm loc}(\mathbb{R}^n) \cap \mathscr{S}'(\mathbb{R}^n)\big]^{M\times M} \\ &E_L(-x) = E_L(x) \ \ \text{for every point} \ \ x \in \mathbb{R}^n \setminus \{0\}. \end{aligned} \tag{17.13}$$

*In fact, each entry in $E_L$ is a real-analytic function in $\mathbb{R}^n \setminus \{0\}$ and an even tempered distribution in $\mathbb{R}^n$.*

*(2) If for each $y \in \mathbb{R}^n$ one denotes by $\delta_y$ Dirac's delta distribution with mass at $y$ in $\mathbb{R}^n$, then in the sense of distributions in $\mathbb{R}^n$ one has*

$$L_x\big[E_L(x-y)\big] = \delta_y(x)\, I_{M\times M}, \qquad \forall\, y \in \mathbb{R}^n, \tag{17.14}$$

*where $I_{M\times M}$ is the $M \times M$ identity matrix, and the subscript $x$ indicates that the operator $L$ is applied to each column of the matrix $E_L(x-y)$ in the variable $x$.*

*(3) For each $\alpha \in \mathbb{N}_0^n$ there exists $C_\alpha \in (0,\infty)$ such that for each point $x \in \mathbb{R}^n \setminus \{0\}$ one has*

$$\left|(\partial^\alpha E_L)(x)\right| \leq \begin{cases} \dfrac{C_\alpha}{|x|^{n-2m+|\alpha|}} & \text{if } n \text{ is odd, or if } n > 2m, \\ & \text{or if } |\alpha| \geq 2m-n+1, \\ \dfrac{C_\alpha(1+|\ln|x||)}{|x|^{n-2m+|\alpha|}} & \text{if } n \text{ is even, and } n \leq 2m, \\ & \text{and } \ 0 \leq |\alpha| \leq 2m-n. \end{cases} \tag{17.15}$$

*Thus, with the derivatives taken in the sense of distributions in $\mathbb{R}^n$,*

$$\begin{aligned} &\text{the components of the matrix } \partial^\alpha E_L \text{ belong to} \\ &L^1_{\rm loc}(\mathbb{R}^n) \text{ whenever } \alpha \in \mathbb{N}_0^n \text{ satisfies } |\alpha| \leq 2m-1. \end{aligned} \tag{17.16}$$

Having fixed a Lipschitz domain $\Omega \subset \mathbb{R}^n$, denote by $\nu = (\nu_j)_{1\leq j\leq n}$ its outward unit normal and by $\sigma := \mathscr{H}^{n-1}\lfloor\partial\Omega$ the corresponding surface measure on $\partial\Omega$. For each $m \in \mathbb{N}$, the boundary multi-trace of order $m-1$ of a vector-valued function $u \in [\mathscr{C}^{m-1}(\overline{\Omega})]^M$, denoted by $\mathrm{Tr}^{m-1}_{\partial\Omega}u$, is the family of functions indexed by multi-indices of length $\leq m-1$ given by

$$\mathrm{Tr}^{m-1}_{\partial\Omega}u := \left\{(\partial^\gamma u)\big|_{\partial\Omega}\right\}_{|\gamma|\leq m-1}. \tag{17.17}$$

For a coefficient tensor $A = (A_{\alpha\beta})_{|\alpha|=|\beta|=m}$ of $M \times M$ matrices with complex entries, the conormal derivative of a function $u \in [\mathscr{C}^{2m}(\overline{\Omega})]^M$ associated with $A$ is defined as

$$\partial^A_\nu u := \left\{\left(\partial^A_\nu u\right)_\delta\right\}_{|\delta|\leq m-1} \tag{17.18}$$

with the $\delta$-component given by the formula

$$\begin{aligned} \left(\partial^A_\nu u\right)_\delta := \sum_{\substack{|\alpha|=|\beta|=m \\ \delta+e_j\leq\alpha}} (-1)^{|\delta|}\,\frac{\alpha!|\delta|!(m-|\delta|-1)!}{m!\delta!(\alpha-\delta-e_j)!}\times \\ \times\,\nu_j A_{\alpha\beta}\Big(\partial^{\alpha+\beta-\delta-e_j}u\Big)\Big|_{\partial\Omega}. \end{aligned} \tag{17.19}$$

Single and double layer potentials are introduced next, including a summary of the their properties that are important for our subsequent analysis. For more on this topic, see [HLM21] and [MM13].

**Theorem 17.2** *Fix $n, m, M \in \mathbb{N}$, $n \geq 2$. Let $L$ be a homogeneous $M \times M$ system of differential operators of order $2m$ in $\mathbb{R}^n$, with constant complex coefficients and let $A = \{A_{\alpha\beta}\}_{|\alpha|=|\beta|=m}$ be a family of coefficient tensors associated with $L$ in the sense of (17.1). Suppose $L$ is weakly elliptic in the sense of (17.3). Let $\Omega$ be a Lipschitz domain in $\mathbb{R}^n$ with compact boundary. Denote by $\nu = (\nu_j)_{1\leq j\leq n}$ its outward unit normal and let $\sigma := \mathscr{H}^{n-1}\lfloor\partial\Omega$ be the surface measure on its boundary.*

*In this setting, let $\dot{f} = \{f_\delta\}_{|\delta|\leq m-1}$ be a family of $\mathbb{C}^M$-valued functions, indexed by multi-indices $\delta \in \mathbb{N}_0^n$ of length $\leq m-1$, such that*

$$f_\delta \in \big[L^1(\partial\Omega, \sigma)\big]^M \text{ for each } \delta \in \mathbb{N}_0^n \text{ with } |\delta| \leq m-1. \tag{17.20}$$

*Then the action of the single multi-layer potential operator $\dot{\mathscr{S}}$, and of the double multi-layer potential operator $\dot{\mathscr{D}}_A$, on $\dot{f}$, is defined according to*

$$(\dot{\mathscr{S}}\dot{f})(x) := \sum_{|\alpha|\leq m-1} (-1)^{|\alpha|} \int_{\partial\Omega} (\partial^\alpha E)(x-y) f_\alpha(y)\, d\sigma(y) \tag{17.21}$$

*and according to*

$$\begin{aligned}(\dot{\mathscr{D}}_A\dot{f})(x) := -\sum_{|\alpha|=|\beta|=m} \sum_{\delta+\gamma+e_j=\alpha} \frac{\alpha!}{|\alpha|!}\frac{|\delta|!}{\delta!}\frac{|\gamma|!}{\gamma!}\times \\ \times \int_{\partial_*\Omega} \nu_j(y)(\partial^{\beta+\gamma}E_L)(x-y)A_{\beta\alpha} f_\delta(y)\, d\sigma_*(y),\end{aligned} \tag{17.22}$$

*respectively, at each $x \in \Omega$.*

*These are well-defined $\mathbb{C}^M$-valued functions and satisfy the following properties*

$$\begin{aligned}&\dot{\mathscr{S}}\dot{f},\ \dot{\mathscr{D}}_A\dot{f} \in \big[\mathscr{C}^\infty(\Omega)\big]^M \text{ as well as} \\ &L(\dot{\mathscr{S}}\dot{f})(x) = 0 \text{ and } L(\dot{\mathscr{D}}_A\dot{f})(x) = 0 \text{ for all } x \in \Omega.\end{aligned} \tag{17.23}$$

The following combinatorial lemma has been proved in [HLM21].

**Lemma 17.1** *For every multi-index $\widetilde{\alpha} \in \mathbb{N}_0^n$ one has*

$$\sum_{\alpha+e_i=\widetilde{\alpha}} \frac{1}{\alpha!} = \frac{|\widetilde{\alpha}|}{\widetilde{\alpha}!}, \tag{17.24}$$

*where the sum is performed over all $\alpha \in \mathbb{N}_0^n$ and all $i \in \{1, \ldots, n\}$ with the property that $\alpha + e_i = \widetilde{\alpha}$.*

Lastly, we include a version of the Divergence Theorem for singular vector fields, which is a special case of [MMM20, Theorem 2.6, pp. 1299–1300]. An extensive account on this topic may be found in [MMM22].

**Theorem 17.3** *Fix $n \in \mathbb{N}$ and let $\Omega$ be an exterior Lipschitz domain in $\mathbb{R}^n$. Denote by $\nu$ the outward unit normal to $\Omega$ and consider the surface measure $\sigma := \mathscr{H}^{n-1}\lfloor\partial\Omega$. Assume the vector field*

$$\vec{F} = (F_1, \ldots, F_n) \in \big[\mathscr{D}'(\Omega)\big]^n \quad \text{has} \quad \operatorname{div}\vec{F} \in \mathscr{E}'(\Omega), \tag{17.25}$$

*where the divergence is considered in the sense of distributions in $\Omega$. In addition, suppose*

$$\begin{aligned} &\text{there exists a compact set } K \text{ contained in } \Omega \\ &\text{such that } \ \vec{F}\big|_{\Omega\setminus K} \in \big[\mathscr{C}^0(\overline{\Omega}\setminus K)\big]^n, \end{aligned} \tag{17.26}$$

*and*

$$\vec{F}(x) = o\big(|x|^{1-n}\big) \ \text{ as } \ |x| \to \infty. \tag{17.27}$$

*Then*

$$\big\langle \operatorname{div}\vec{F}\,,\,1\big\rangle_{\mathscr{E}(\Omega)} = \int_{\partial\Omega} \nu\cdot\big(\vec{F}\,\big|_{\partial\Omega}\big)\,d\sigma. \tag{17.28}$$

## 17.3 Proof of the Main Result

Here we present the proof Theorem 17.1.

***Proof*** Pick $R \in (0,\infty)$ sufficiently large so that $\partial\Omega$ is contained in $B(0,R)$ and set $\Omega_0 := \mathbb{R}^n \setminus \overline{B(0,R)}$. Fix an arbitrary point $x \in \Omega_0$ and select an arbitrary multi-index $\eta \in \mathbb{N}_0^n$. For these choices, consider the vector field

$$\vec{F}_\eta := \big(F_j^\eta\big)_{1\le j\le n} \tag{17.29}$$

whose $j$-th component, $j \in \{1,\ldots,n\}$, is defined for $y \in \Omega$ as

$$\begin{aligned} F_j^\eta(y) := -\sum_{|\alpha|=|\beta|=m}\ \sum_{\delta+\gamma+e_j=\alpha} \frac{\alpha!\,|\delta|!\,|\gamma|!}{m!\,\gamma!\,\delta!}\times \\ \times\,(-1)^{|\eta|+m+|\gamma|}\partial_y^{\eta+\beta+\gamma}\big[E_L(x-y)\big]A_{\beta\alpha}(\partial^\delta u)(y) \end{aligned}$$

$$- \sum_{|\alpha|=|\beta|=m} \sum_{\delta+\gamma+e_j=\alpha} \frac{\alpha!\,|\delta|\,!|\gamma|!}{m!\,\delta!\,\gamma!} \times \\ \times (-1)^{|\eta|+|\delta|} \partial_y^{\eta+\delta}\big[E_L(x-y)\big] A_{\alpha\beta} (\partial^{\beta+\gamma} u)(y). \tag{17.30}$$

Since $E_L(x-\cdot) \in \big[L^1_{\rm loc}(\Omega_0)\big]^{M\times M} \subseteq \big[\mathscr{D}'(\Omega_0)\big]^{M\times M}$, and for every $\theta \in \mathbb{N}_0^n$ we have $\partial^\theta u \in \big[\mathscr{C}^\infty(\Omega_0)\big]^M$, it follows that $\vec{F}_\eta \in \big[\mathscr{D}'(\Omega_0)\big]^{M\times n}$. In fact, based on (17.30), the hypotheses on $u$ and (17.13) we have

$$\vec{F}_\eta \in \big[\mathscr{C}^\infty\big(\overline{\Omega_0} \setminus \{x\}\big)\big]^{M\times n} \cap \big[\mathscr{D}'(\Omega_0)\big]^{M\times n}. \tag{17.31}$$

Next, we claim that

$$\operatorname{div} \vec{F}_\eta = (-1)^{|\eta|} u \cdot \partial^\eta \delta_x \in [\mathscr{E}'(\Omega_0)]^M, \tag{17.32}$$

where $\delta_x$ denotes the Dirac distribution with mass at $x$. To see why this claim is true, let $y \in \Omega_0$ be arbitrary and write

$$\operatorname{div} \vec{F}_\eta(y) = \sum_{j=1}^n \partial_j F_j^\eta(y) = I_a + I_b + II_a + II_b, \tag{17.33}$$

where

$$I_a := - \sum_{\substack{|\alpha|=|\beta|=m \\ \delta+\gamma+e_j=\alpha}} \frac{\alpha!\,|\delta|!\,|\gamma|!}{m!\,\gamma!\,\delta!} (-1)^{|\eta|+m+|\gamma|} \partial_y^{\eta+\beta+\gamma+e_j}\big[E_L(x-y)\big] A_{\beta\alpha} (\partial^\delta u)(y),$$

$$I_b := - \sum_{\substack{|\alpha|=|\beta|=m \\ \delta+\gamma+e_j=\alpha}} \frac{\alpha!\,|\delta|!\,|\gamma|!}{m!\,\gamma!\,\delta!} (-1)^{|\eta|+m+|\gamma|} \partial_y^{\eta+\beta+\gamma}\big[E_L(x-y)\big] A_{\beta\alpha} (\partial^{\delta+e_j} u)(y),$$

$$II_a := - \sum_{\substack{|\alpha|=|\beta|=m \\ \delta+\gamma+e_j=\alpha}} \frac{\alpha!\,|\delta|!\,|\gamma|!}{m!\,\delta!\,\gamma!} (-1)^{|\eta|+|\delta|} \partial_y^{\eta+\delta+e_j}\big[E_L(x-y)\big] A_{\alpha\beta} (\partial^{\beta+\gamma} u)(y),$$

$$II_b := - \sum_{\substack{|\alpha|=|\beta|=m \\ \delta+\gamma+e_j=\alpha}} \frac{\alpha!\,|\delta|!\,|\gamma|!}{m!\,\delta!\,\gamma!} (-1)^{|\eta|+|\delta|} \partial_y^{\eta+\delta}\big[E_L(x-y)\big] A_{\alpha\beta} (\partial^{\beta+\gamma+e_j} u)(y).$$

We proceed by rewriting the sums in the right-most side of (17.33) as follows. Starting with $I_a$ we have

$$
\begin{aligned}
I_a &= \sum_{|\alpha|=|\beta|=m} \sum_{\delta<\alpha} \Big( \sum_{\gamma+e_j=\alpha-\delta} \frac{1}{\gamma!} \Big) \frac{\alpha!\,|\delta|!\,(|\alpha-\delta|-1)!}{m!\,\delta!} \times \\
&\quad \times (-1)^{|\eta|-|\delta|} \partial_y^{\eta+\beta+\alpha-\delta} \big[ E_L(x-y) \big] A_{\beta\alpha}(\partial^\delta u)(y) \\
&= \sum_{|\alpha|=|\beta|=m} \sum_{\delta<\alpha} \frac{|\alpha-\delta|}{(\alpha-\delta)!} \cdot \frac{\alpha!\,|\delta|!\,(|\alpha-\delta|-1)!}{m!\,\delta!} \times \\
&\quad \times (-1)^{|\eta|-|\delta|} \partial_y^{\eta+\beta+\alpha-\delta} \big[ E_L(x-y) \big] A_{\beta\alpha}(\partial^\delta u)(y) \\
&= \sum_{|\alpha|=|\beta|=m} \sum_{\delta<\alpha} \frac{\alpha!\,|\delta|!\,|\alpha-\delta|!}{m!\,\delta!\,(\alpha-\delta)!} \times \\
&\quad \times (-1)^{|\eta|-|\delta|} \partial_y^{\eta+\beta+\alpha-\delta} \big[ E_L(x-y) \big] A_{\beta\alpha}(\partial^\delta u)(y) \\
&= \sum_{|\alpha|=|\beta|=m} \sum_{\delta+\gamma=\alpha,\,|\gamma|>0} \frac{\alpha!\,|\delta|!\,|\gamma|!}{m!\,\delta!\,\gamma!} \times \\
&\quad \times (-1)^{|\eta|-|\delta|} \partial_y^{\eta+\beta+\gamma} \big[ E_L(x-y) \big] A_{\beta\alpha}(\partial^\delta u)(y), \qquad (17.34)
\end{aligned}
$$

where for the first equality above we have used the fact that in the definition of $I_a$ we have $|\gamma| = |\alpha| - |\delta| - 1 = m - 1 - |\delta|$, for the second equality in (17.34) we have applied Lemma 17.1 to the effect that $\sum_{\gamma+e_j=\alpha-\delta} \frac{1}{\gamma!} = \frac{|\alpha-\delta|}{(\alpha-\delta)!}$ for each $\delta < \alpha$, and for the final equality in (17.34) we have denoted $\gamma := \alpha - \delta$.

Moving on to $I_b$ we write

$$
\begin{aligned}
I_b &= -\sum_{|\alpha|=|\beta|=m} \sum_{\gamma<\alpha} \Big( \sum_{\delta+e_j=\alpha-\gamma} \frac{1}{\delta!} \Big) \frac{\alpha!\,(|\alpha-\gamma|-1)!\,|\gamma|!}{m!\,\gamma!} \times \\
&\quad \times (-1)^{|\eta|+m+|\gamma|} \partial_y^{\eta+\beta+\gamma} \big[ E_L(x-y) \big] A_{\beta\alpha}(\partial^{\alpha-\gamma} u)(y) \\
&= -\sum_{|\alpha|=|\beta|=m} \sum_{\gamma<\alpha} \frac{|\alpha-\gamma|}{(\alpha-\gamma)!} \cdot \frac{\alpha!\,(|\alpha-\gamma|-1)!\,|\gamma|!}{m!\,\gamma!} \times \\
&\quad \times (-1)^{|\eta|+m+|\gamma|} \partial_y^{\eta+\beta+\gamma} \big[ E_L(x-y) \big] A_{\beta\alpha}(\partial^{\alpha-\gamma} u)(y) \\
&= -\sum_{|\alpha|=|\beta|=m} \sum_{\gamma<\alpha} \frac{\alpha!\,|\alpha-\gamma|!\,|\gamma|!}{m!\,\gamma!\,(\alpha-\gamma)!} \times \\
&\quad \times (-1)^{|\eta|+m+|\gamma|} \partial_y^{\eta+\beta+\gamma} \big[ E_L(x-y) \big] A_{\beta\alpha}(\partial^{\alpha-\gamma} u)(y)
\end{aligned}
$$

$$= - \sum_{|\alpha|=|\beta|=m} \sum_{\gamma+\delta=\alpha,\,|\delta|>0} \frac{\alpha!\,|\delta|!\,|\gamma|!}{m!\,\gamma!\,\delta!} \times$$

$$\times (-1)^{|\eta|-|\delta|} \partial_y^{\eta+\beta+\gamma} \big[E_L(x-y)\big] A_{\beta\alpha} (\partial^\delta u)(y), \tag{17.35}$$

where for the second equality in (17.35) we have employed Lemma 17.1 which implies that $\sum_{\delta+e_j=\alpha-\gamma} \frac{1}{\delta!} = \frac{|\alpha-\gamma|}{(\alpha-\gamma)!}$ for each $\delta < \alpha$, while for the last equality in (17.35) we have denoted $\delta := \alpha - \gamma$ and used the fact that $|\gamma| = m - |\delta|$.

A simple inspection of (17.34) and (17.35) reveals that all the terms in $I_a + I_b$ corresponding to $|\delta| > 0$ and $|\gamma| > 0$ cancel out. The terms in $I_a$ corresponding to $|\delta| = 0$ and the terms in $I_b$ corresponding to $|\gamma| = 0$ are the only ones that remain, hence

$$I_a + I_b = (-1)^{|\eta|} \Big( \sum_{|\alpha|=|\beta|=m} \partial_y^{\eta+\beta+\alpha} \big[E_L(x-y)\big] A_{\beta\alpha} \Big) u(y)$$

$$- (-1)^{|\eta|+m} \sum_{|\alpha|=|\beta|=m} \partial_y^{\eta+\beta} \big[E_L(x-y)\big] A_{\beta\alpha} (\partial^\alpha u)(y). \tag{17.36}$$

Observe that based on the properties of $E_L$ we further obtain

$$\begin{aligned}
\partial_y^{\eta+\beta+\alpha} \big[E_L(x-y)\big] A_{\beta\alpha} &= \Big( A_{\beta\alpha}^\top \partial_y^{\eta+\beta+\alpha} \big[E_L(x-y)^\top\big] \Big)^\top \\
&= \Big( A_{\beta\alpha}^\top \partial_y^{\eta+\beta+\alpha} \big[E_{L^\top}(x-y)\big] \Big)^\top \\
&= \partial_y^\eta \big[ L_y^\top \big( E_{L^\top}(x-y) \big) \big]^\top \\
&= \partial_y^\eta \big[ \delta_x I_{M\times M} \big]^\top \\
&= (\partial^\eta \delta_x) I_{M\times M}.
\end{aligned} \tag{17.37}$$

Now we combine (17.36) and (17.37) to conclude that

$$I_a + I_b = (-1)^{|\eta|} u(y) (\partial^\eta \delta_x)$$

$$- (-1)^{|\eta|+m} \sum_{|\alpha|=|\beta|=m} \partial_y^{\eta+\beta} \big[E_L(x-y)\big] A_{\beta\alpha} (\partial^\alpha u)(y). \tag{17.38}$$

Using the same circle of ideas, write $II_a$ as

$$\begin{aligned} II_a &= \sum_{|\alpha|=|\beta|=m} \sum_{\gamma<\alpha} \Big( \sum_{\delta+e_j=\alpha-\gamma} \frac{1}{\delta!} \Big) \frac{\alpha!\,(|\alpha-\gamma|-1)!|\gamma|!}{m!\,\gamma!} \times \\ &\quad \times (-1)^{m+|\eta|-|\gamma|} \partial_y^{\eta+\alpha-\gamma} \big[ E_L(x-y) \big] A_{\alpha\beta} (\partial^{\beta+\gamma} u)(y) \\ &= \sum_{|\alpha|=|\beta|=m} \sum_{\gamma<\alpha} \frac{|\alpha-\gamma|}{(\alpha-\gamma)!} \cdot \frac{\alpha!\,(|\alpha-\gamma|-1)!|\gamma|!}{m!\,\gamma!} \times \\ &\quad \times (-1)^{m+|\eta|-|\gamma|} \partial_y^{\eta+\alpha-\gamma} \big[ E_L(x-y) \big] A_{\alpha\beta} (\partial^{\beta+\gamma} u)(y) \\ &= \sum_{|\alpha|=|\beta|=m} \sum_{\gamma+\delta=\alpha,\,|\delta|>0} \frac{\alpha!\,|\delta|!|\gamma|!}{m!\,\gamma!\,\delta!} \times \\ &\quad \times (-1)^{|\eta|+|\delta|} \partial_y^{\eta+\delta} \big[ E_L(x-y) \big] A_{\alpha\beta} (\partial^{\beta+\gamma} u)(y) \end{aligned} \tag{17.39}$$

and $II_b$ as

$$\begin{aligned} II_b &= - \sum_{|\alpha|=|\beta|=m} \sum_{\delta<\alpha} \Big( \sum_{\gamma+e_j=\alpha-\delta} \frac{1}{\gamma!} \Big) \frac{\alpha!\,(|\alpha-\delta|-1)!\,|\delta|!}{m!\,\delta!} \times \\ &\quad \times (-1)^{|\eta|+|\delta|} \partial_y^{\eta+\delta} \big[ E_L(x-y) \big] A_{\alpha\beta} (\partial^{\beta+\alpha-\delta} u)(y) \\ &= - \sum_{|\alpha|=|\beta|=m} \sum_{\delta<\alpha} \frac{|\alpha-\delta|}{(\alpha-\delta)!} \cdot \frac{\alpha!\,(|\alpha-\delta|-1)!\,|\delta|!}{m!\,\delta!} \times \\ &\quad \times (-1)^{|\eta|+|\delta|} \partial_y^{\eta+\delta} \big[ E_L(x-y) \big] A_{\alpha\beta} (\partial^{\beta+\alpha-\delta} u)(y) \\ &= - \sum_{|\alpha|=|\beta|=m} \sum_{\gamma+\delta=\alpha,|\gamma|>0} \frac{\alpha!\,|\gamma|!\,|\delta|!}{m!\,\delta!\,\gamma!} \times \\ &\quad \times (-1)^{|\eta|+|\delta|} \partial_y^{\eta+\delta} \big[ E_L(x-y) \big] A_{\alpha\beta} (\partial^{\beta+\gamma} u)(y). \end{aligned} \tag{17.40}$$

An inspection of (17.39) and (17.40) reveals that the terms in $II_a$ corresponding to $|\gamma| > 0$ and the terms in $II_b$ corresponding to $|\delta| > 0$ cancel out, so

$$\begin{aligned} II_a + II_b = &\sum_{|\alpha|=|\beta|=m} (-1)^{|\eta|+m} \partial_y^{\eta+\alpha} \big[ E_L(x-y) \big] A_{\alpha\beta} (\partial^{\beta} u)(y) \\ &- \sum_{|\alpha|=|\beta|=m} (-1)^{|\eta|} \partial_y^{\eta} \big[ E_L(x-y) \big] A_{\alpha\beta} (\partial^{\beta+\alpha} u)(y) \end{aligned}$$

$$
\begin{aligned}
&= (-1)^{|\eta|+m} \sum_{|\alpha|=|\beta|=m} \partial_y^{\eta+\alpha}\big[E_L(x-y)\big]A_{\alpha\beta}(\partial^\beta u)(y) \\
&\qquad\qquad - (-1)^{|\eta|}\partial_y^{\eta}\big[E_L(x-y)\big](Lu)(y) \\
&= (-1)^{|\eta|+m} \sum_{|\alpha|=|\beta|=m} \partial_y^{\eta+\beta}\big[E_L(x-y)\big]A_{\beta\alpha}(\partial^\alpha u)(y), \qquad (17.41)
\end{aligned}
$$

where we have used the definition of $L$ in the second equality above, while for the last equality in (17.41) we have used the fact that $Lu = 0$ in $\Omega$.

Together, (17.33), (17.38), (17.41), and the arbitrariness of $x \in \Omega_0$ imply

$$
\operatorname{div} \vec{F}_\eta(y) = (-1)^{|\eta|} u(y)(\partial^\eta \delta_x) \text{ in } \big[\mathscr{D}'(\Omega_0)\big]^M \qquad (17.42)
$$

proving (17.32).

Let us now assume that the multi-index $\eta$ is such that

$$
\begin{cases} |\eta| \geq N & \text{if } n \text{ is odd, or if } n > 2m, \\ |\eta| > N & \text{if } n \text{ is even and } n \leq 2m. \end{cases} \qquad (17.43)
$$

Under this additional assumption, our second claim is that

$$
\vec{F}_\eta(y) = o\big(|y|^{1-n}\big) \text{ as } |y| \to \infty. \qquad (17.44)
$$

To justify (17.44), start by invoking interior estimates for the null-solution $u$ of $L$ (cf. [DM18, Theorem 11.7, p. 409] and [MMM22]) and conclude that the decay in (17.4) further improves to

$$
(\partial^\theta u)(y) = o\big(|y|^{N-|\theta|}\big) \text{ as } |y| \to \infty, \quad \text{for all } \theta \in \mathbb{N}_0^n. \qquad (17.45)
$$

To proceed, consider the case when either $n$ is odd, or $n > 2m$. Fix $\alpha, \beta, \delta, \gamma \in \mathbb{N}_0^n$ along with some $j \in \{1, \ldots, n\}$ satisfying $|\alpha| = |\beta| = m$ and $\delta + \gamma + e_j = \alpha$. Then the decay in (17.45), the estimates in (17.15), and the fact that by assumption $|\eta| \geq N$ (cf. (17.43)) imply

$$
\begin{aligned}
\big|\partial_y^{\eta+\beta+\gamma}\big[E_L(x-y)\big]\big|(\partial^\delta u)(y)| &= O\Big(\frac{1}{|y|^{n-2m+|\eta|+|\beta|+|\gamma|}}\Big) o\big(|x|^{N-|\delta|}\big) \\
&= o\Big(\frac{1}{|y|^{n-2m+|\eta|+m+m-1-N}}\Big) \\
&= o\Big(\frac{1}{|y|^{n-1}}\Big) \qquad (17.46)
\end{aligned}
$$

as $|y| \to \infty$, and

$$\begin{aligned}\left|\partial_y^{\eta+\delta}\big[E_L(x-y)\big]\right||(\partial^{\beta+\gamma}u)(y)| &= O\Big(\frac{1}{|y|^{n-2m+|\eta|+|\delta|}}\Big)o\big(|y|^{N-|\beta|-|\gamma|}\big)\\ &= o\Big(\frac{1}{|y|^{n-2m+|\eta|+m+m-1-N}}\Big)\\ &= o\Big(\frac{1}{|y|^{n-1}}\Big)\end{aligned} \tag{17.47}$$

as $|y| \to \infty$. A combination of (17.30), (17.46), and (17.47) gives (17.44) in the current case, namely whenever $|\eta| \geq N$ and if $n$ is odd, or if $n > 2m$.

It remains to consider the case when $n$ is even and $n \leq 2m$. This time, if $\alpha, \beta, \delta, \gamma \in \mathbb{N}_0^n$ together with $j \in \{1, \ldots, n\}$ are such that $|\alpha| = |\beta| = m$ and $\delta + \gamma + e_j = \alpha$, using the decay in (17.45), the corresponding estimates in (17.15), and the fact that now we assume $|\eta| > N$, we obtain

$$\begin{aligned}\left|\partial_y^{\eta+\beta+\gamma}\big[E_L(x-y)\big]\right||(\partial^{\delta}u)(y)| &= O\Big(\frac{\ln|y|}{|y|^{n-2m+|\eta|+|\beta|+|\gamma|}}\Big)o\big(|y|^{N-|\delta|}\big)\\ &= o\Big(\frac{1}{|y|^{n-1}}\Big)\end{aligned} \tag{17.48}$$

as $|y| \to \infty$, and

$$\begin{aligned}\left|\partial_y^{\eta+\delta}\big[E_L(x-y)\big]\right||(\partial^{\beta+\gamma}u)(y)| &= O\Big(\frac{\ln|y|}{|y|^{n-2m+|\eta|+|\delta|}}\Big)o\big(|y|^{N-|\beta|-|\gamma|}\big)\\ &= o\Big(\frac{1}{|y|^{n-1}}\Big)\end{aligned} \tag{17.49}$$

as $|y| \to \infty$. Thus, (17.30), (17.48), and (17.49) may be combined to conclude that (17.44) also holds if $n$ is even and $n \leq 2m$ provided $|\eta| > N$.

Having proved (17.31), (17.32), and (17.44), we see that whenever (17.43) holds, the vector field $\vec{F}_\eta$ satisfies the hypotheses of the Divergence Theorem 17.3. As such, formula (17.28) written for this vector field implies

$$\big\langle \operatorname{div}\vec{F}_\eta\,,\,1\big\rangle_{\mathscr{E}(\Omega_0)} = \int_{\partial\Omega_0} \nu\cdot\big(\vec{F}_\eta\,\big|_{\partial\Omega_0}\big)\,d\sigma. \tag{17.50}$$

Making use of (17.32) we can further express

$$\begin{aligned}\big\langle \operatorname{div}\vec{F}_\eta\,,\,1\big\rangle_{\mathscr{E}(\Omega_0)} &= \big\langle (-1)^{|\eta|}u\,\partial^\eta\delta_x\,,\,1\big\rangle_{\mathscr{E}(\Omega_0)}\\ &= \big\langle (-1)^{|\eta|}\partial^\eta\delta_x\,,\,u\big\rangle_{\mathscr{E}(\Omega_0)}\end{aligned}$$

$$= {}_{\mathscr{E}'(\Omega_0)}\big\langle \delta_x\,,\ \partial^\eta u\big\rangle_{\mathscr{E}(\Omega_0)}$$

$$= (\partial^\eta u)(x). \tag{17.51}$$

As for the right-hand side of (17.50),

$$\begin{aligned}\int_{\partial\Omega_0} \nu\cdot\big(\vec{F}_\eta\,\big|_{\partial\Omega_0}\big)\,d\sigma &= -\sum_{|\alpha|=|\beta|=m}\ \sum_{\delta+\gamma+e_j=\alpha}\frac{\alpha!\,|\delta|!\,|\gamma|!}{m!\,\gamma!\,\delta!}\times\\ &\qquad\times\partial^\eta\int_{\partial\Omega_0}\nu_j(y)(\partial^{\beta+\gamma}E_L)(x-y)A_{\beta\alpha}(\partial^\delta u)(y)\,d\sigma(y)\\ &\quad-\sum_{|\alpha|=|\beta|=m}\ \sum_{\delta+\gamma+e_j=\alpha}\frac{\alpha!\,|\delta|\,!|\gamma|!}{m!\,\delta!\,\gamma!}\times\\ &\qquad\times\partial^\eta\int_{\partial\Omega_0}\nu_j(y)(\partial^{\delta}E_L)(x-y)A_{\alpha\beta}(\partial^{\beta+\gamma}u)(y)\,d\sigma(y)\\ &=\partial^\eta\Big[\dot{\mathscr{D}}_A\big(\mathrm{Tr}^{m-1}_{\partial\Omega_0}u\big)-\dot{\mathscr{S}}\big(\partial^A_\nu u\big)\Big](x).\end{aligned} \tag{17.52}$$

Above, $\dot{\mathscr{D}}_A$, $\dot{\mathscr{S}}$ are, respectively, the double and single multi-layers associated as in Theorem 17.2 with the set $\Omega_0$. These act, respectively, on the arrays $\mathrm{Tr}^{m-1}_{\partial\Omega_0}u$, the multi-trace of order $m-1$ of $u$ on $\partial\Omega_0$ (cf. (17.17)), and the conormal derivative $\partial^A_\nu u$ of $u$ on $\partial\Omega_0$ (cf. (17.18)–(17.19)). Together, (17.50), (17.51), and (17.52), give

$$\partial^\eta\Big[u-\dot{\mathscr{D}}_A\big(\mathrm{Tr}^{m-1}_{\partial\Omega_0}u\big)+\dot{\mathscr{S}}\big(\partial^A_\nu u\big)\Big]=0\ \text{ in }\ \Omega_0. \tag{17.53}$$

Consequently, if we set

$$P_0:=u-\dot{\mathscr{D}}_A\big(\mathrm{Tr}^{m-1}_{\partial\Omega_0}u\big)+\dot{\mathscr{S}}\big(\partial^A_\nu u\big)\ \text{ in }\ \Omega_0, \tag{17.54}$$

then what we proved so far ensures that

$$\begin{aligned}&P_0\in\big[\mathscr{C}^\infty(\Omega_0)\big]^M\ \text{ and }\ \partial^\eta P_0=0\ \text{ in }\ \Omega_0\\ &\qquad\text{for all }\eta\in\mathbb{N}_0^n\text{ satisfying (17.43).}\end{aligned} \tag{17.55}$$

Via a Taylor series argument, it is not difficult to see that the conditions in (17.55) force $P_0$ to be locally a polynomial. The degree of this polynomial depends on the lower bound for $|\eta|$ as dictated by (17.43). In addition, since $\Omega_0$ is a connected set in $\mathbb{R}^n$, it follows that there exists a polynomial $P$ in $\mathbb{R}^n$ such that $P_0=P\big|_{\Omega_0}$ in $\Omega_0$. An inspection of (17.54) also implies $LP_0=0$ in $\Omega_0$ which forces $LP=0$ in $\Omega_0$.

Hence, the polynomial $LP$ which is defined in $\mathbb{R}^n$ vanishes in $\Omega_0$, thus necessarily $LP = 0$ in $\mathbb{R}^n$. In summary, so far we proved that

$$\begin{aligned} &\text{there exists some polynomial } P \text{ in } \mathbb{R}^n, \\ &\text{satisfying } \ LP = 0 \ \text{ in } \ \mathbb{R}^n, \ \text{ and such that} \\ &u = P + \dot{\mathscr{D}}_A\big(\mathrm{Tr}^{m-1}_{\partial\Omega_0} u\big) - \dot{\mathscr{S}}\big(\partial_\nu^A u\big) \ \text{ in } \ \Omega_0. \end{aligned} \tag{17.56}$$

The fact that $u$ satisfies (17.5) now follows from (17.56), (17.21), (17.22), and (17.15). In addition, the claim about $P$ made in (17.6) is a consequence of what we proved so far and a careful inspection of (17.55), while keeping in mind (17.43).

**Acknowledgments** The authors gratefully acknowledge partial support from the Simons Foundation (through grants # 426669, # 637481), as well as NSF (grant # 1900938).

## References

[HLM21] Hoepfner, G., Liboni, P., Mitrea, D., Mitrea, I., Mitrea, M.: Multi-layer potentials for higher order systems in rough domains. Anal. PDE **14**(4), 1233–1308 (2021)

[DM18] Mitrea, D.: Distributions, partial differential equations, and harmonic analysis, 2nd edn. Springer, Cham (2018)

[MMM20] Mitrea, D., Mitrea, I., Mitrea, M.: A sharp divergence theorem with nontangential traces. Not. AMS **67**(9), 1295–1305 (2020)

[MMM22] Mitrea, D., Mitrea, I., Mitrea, M.: Geometric Harmonic Analysis I–V, vol. 72–76, Developments in Mathematics, Springer, Cham (2022)

[MM13] Mitrea, I., Mitrea, M.: Multi-Layer Potentials and Boundary Problems for Higher-Order Elliptic Systems in Lipschitz Domains. Lecture Notes in Mathematics, vol. 2063. Springer, Berlin (2013)

# Chapter 18
# On the Use of the Adjoint Technique to the Estimation of Neutron Source Distributions in the Context of Subcritical Nuclear Reactors


**L. R. C. Moraes and R. C. Barros**

## 18.1 Introduction

A nuclear reactor characterized as subcritical possesses a core that cannot achieve criticality, in the sense, that the fission-chain reactions occurring with the fissile materials at the reactor core cannot hold without the action of an external stationary source of neutrons. Nowadays, there are different classes of such devices driven by external sources, being the accelerator-driven systems the most common in projects involving subcritical reactors [NiEtAl01]. Due to its inherent safety, as one just needs to switch off the external source to shut down the reactor, and other positive features, which also include substantial flexibility in fuel processing and managing, the subcritical reactors have been increasing research interest, although they possess a more complex structure in comparison to critical commercial nuclear reactors.

In this chapter we provide a more deep and general description of the methodology presented in the literature [LeEtAl20, LeEtAl21]. This methodology is based on the adjoint technique and is used to determine the neutron source distribution required to drive a subcritical system to a prescribed distribution of power. In other words, the forward transport equation to mathematically model the neutron migration within the reactor core together with the equation that is adjoint to it are correlated through a reciprocity relation (Sect. 18.2), leading to a relationship between the sources of neutrons and the power produced by the system. This relationship is centered on the construction of a special matrix, namely importance matrix (Sect. 18.3), composed of the solutions of the adjoint transport equations, which are interpreted here as measure of the importance that one neutron inserted

L. R. C. Moraes · R. C. Barros (✉)
Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro, RJ, Brazil
e-mail: lrcmoraes@iprj.uerj.br; ricardob@iprj.uerj.br



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_18

into the system has to the power generation. An illustrative example is given in Sect. 18.4 and we close this chapter with some concluding remarks in Sect. 18.5.

## 18.2 Mathematical Basis

We begin with considering the functional

$$\mathtt{F} = \int_V dV \int_0^\infty dE \int_{4\pi} Q^\dagger(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E)\Psi(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E)d\hat{\Omega} \equiv \langle Q^\dagger, \Psi \rangle, \tag{18.1}$$

which is linear with respect to both $\Psi$ and $Q^\dagger$ functions. These two functions are defined in the same phase space, that is composed by the spatial variable $\mathbf{r}$ (a point inside a convex volume $V$), the angular variable $\hat{\boldsymbol{\Omega}}$ (a direction of flight taken at the surface of an unit sphere, i.e., $|\hat{\boldsymbol{\Omega}}| = 1$), and the kinetic energy variable $E$. The function $\Psi$ in Eq. (18.1) is the neutron angular flux, which satisfies the linear Boltzmann equation (LBE), also referred to in this chapter as the forward neutron transport equation [PrLa10]. That is,

$$\begin{aligned} &\hat{\boldsymbol{\Omega}} \cdot \nabla\Psi(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) + \sigma_\mathrm{t}(\mathbf{r}, E)\Psi(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) = \\ &\int_0^\infty dE' \int_{4\pi} \sigma_\mathrm{s}(\mathbf{r}, \hat{\boldsymbol{\Omega}}' \cdot \hat{\boldsymbol{\Omega}}, E' \to E)\Psi(\mathbf{r}, \hat{\boldsymbol{\Omega}}', E')d\hat{\boldsymbol{\Omega}}' + \\ &\frac{\chi(\mathbf{r}, E)}{4\pi} \int_0^\infty dE' \int_{4\pi} \nu\sigma_\mathrm{f}(\mathbf{r}, E')\Psi(\mathbf{r}, \hat{\boldsymbol{\Omega}}', E')d\hat{\boldsymbol{\Omega}}' + Q(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E), \end{aligned} \tag{18.2a}$$

where $\sigma_\mathrm{t}$, $\sigma_\mathrm{s}$, and $\sigma_\mathrm{f}$ are the total, scattering, and fission macroscopic cross sections, $\nu$ and $\chi$ are the average number of neutrons released in a fission event and the fission spectrum, respectively, and $Q$ represents a neutron source. Equation (18.2a) is subjected to the boundary conditions

$$\Psi(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) = \Psi^b(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E), \quad \mathbf{r} \in \partial V, \ \hat{\boldsymbol{\Omega}} \cdot \mathbf{n} < 0, \ 0 < E < \infty, \tag{18.2b}$$

where $\mathbf{n}$ is the unit outward normal vector at $\mathbf{r} \in \partial V$. In addition, back to Eq. (18.1), $Q^\dagger$ is a function called adjoint source, whose meaning is tied to the definition of the functional $\mathtt{F}$. At this point we remark that the theoretical approach considered in this chapter is consistent with positive definite functionals, i.e., $\mathtt{F} > 0$ in Eq. (18.1).

Now, in order to analyze the constitution of the functional $\mathtt{F}$, we suppose the introduction of one neutron at position $\mathbf{r}$, migrating in direction $\hat{\boldsymbol{\Omega}}$ with energy $E$, inside the same system as $\Psi$ in Eq. (18.1) is defined. From this insertion an (average) increase $\Delta\Psi$ of the neutron angular flux will then promote a change (increase) $\Delta\mathtt{F}$ in the functional $\mathtt{F}$. We remark that this increase $\Delta\mathtt{F}$ may be produced either directly by the inserted neutron or through its precursors, in the case of a multiplying system.

The contribution $\Delta \mathtt{F}$ generated in the functional due to the insertion of a neutron into the system is defined as neutron importance [Ga87], also referred to as adjoint angular flux.

Based on the concept of neutron importance, each neutron inserted into the system contributes with a value $\Delta \mathtt{F}$ to the functional. Therefore, considering the two ways of which neutrons can be inserted into the system, i.e., through its boundaries or by the neutron source, we can rewrite the functional $\mathtt{F}$ accounting for the contribution of all neutrons inserted into the system. In this case, we have

$$\mathtt{F} = \langle \Psi^{\dagger}, Q \rangle + P_b \left[ \Psi^{\dagger}, \Psi \right], \tag{18.3a}$$

where we have defined

$$\langle \Psi^{\dagger}, Q \rangle = \int_V dV \int_0^{\infty} dE \int_{4\pi} \Psi^{\dagger}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) Q(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) d\hat{\boldsymbol{\Omega}} \tag{18.3b}$$

and

$$\begin{aligned} P_b \left[ \Psi^{\dagger}, \Psi \right] = & \int_0^{\infty} dE \int_{\mathbf{n}\cdot\hat{\boldsymbol{\Omega}}<0} d\hat{\boldsymbol{\Omega}} \int_{\partial V} |\mathbf{n} \cdot \hat{\boldsymbol{\Omega}}| \Psi^{\dagger}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) \Psi^{b}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) dA \\ & - \int_0^{\infty} dE \int_{\mathbf{n}\cdot\hat{\boldsymbol{\Omega}}>0} d\hat{\boldsymbol{\Omega}} \int_{\partial V} (\mathbf{n} \cdot \hat{\boldsymbol{\Omega}}) \Psi^{\dagger}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) \Psi(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) dA. \end{aligned} \tag{18.3c}$$

In Eq. (18.3) we have used the function $\Psi^{\dagger}$ to represent the importance of one neutron inserted into the system at $(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E)$, which is equal to the aforementioned contribution $\Delta \mathtt{F}$.

For the sake of simplicity, let us examine separately Eqs. (18.3b) and (18.3c). Regarding Eq. (18.3b), this term accounts in the functional $\mathtt{F}$ for the contribution of all neutrons inserted into the system by the neutron source. Here, $\Psi^{\dagger}$ gives the contribution $\Delta \mathtt{F}$ of one neutron inserted at $(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E)$ and $Q$ biases this contribution to the total number of neutrons inserted into the system by the neutron source at the same point. On the other hand, the term given in Eq. (18.3c) accounts in the functional for the contribution of neutrons that are inserted, for the first time, into the system through its structural boundaries. In Eq. (18.3c) the term

$$\int_0^{\infty} dE \int_{\mathbf{n}\cdot\hat{\boldsymbol{\Omega}}<0} d\hat{\boldsymbol{\Omega}} \int_{\partial V} |\mathbf{n} \cdot \hat{\boldsymbol{\Omega}}| \Psi^{\dagger}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) \Psi^{b}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) dA \tag{18.4a}$$

accounts for the contribution to the functional of all neutrons that enter the system through its boundaries. Also in Eq. (18.4a), $\Psi^{\dagger}$ provides the contribution $\Delta \mathtt{F}$ to the functional of one neutron inserted into the system at $(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E)$ and $\Psi^{b}$ (Eq. (18.2b)) biases this contribution with respect to all neutrons that enter the system through its boundaries. However, if there is a possibility that part of the neutrons

entering the system can be formed by neutrons that were previously inside the system, for example, when reflective boundary conditions are considered, then their contribution to the functional have been already accounted for, either by the neutron source term or even by the boundary term itself. In this case, neutrons that enter the system through its boundaries, and for some reason had been previously inside the system, should be removed in order to avoid considering the same contribution multiple times. This situation explains the term

$$-\int_0^\infty dE \int_{\mathbf{n}\cdot\hat{\boldsymbol{\Omega}}>0} d\hat{\boldsymbol{\Omega}} \int_{\partial V} (\mathbf{n}\cdot\hat{\boldsymbol{\Omega}})\Psi^\dagger(\mathbf{r},\hat{\boldsymbol{\Omega}},E)\Psi(\mathbf{r},\hat{\boldsymbol{\Omega}},E)dA \tag{18.4b}$$

that appears in Eq. (18.3c). If there is a possibility of a neutron leaving the system to return, the importance of this neutron ($\Psi^\dagger$, for $\hat{\boldsymbol{\Omega}}\cdot\mathbf{n}>0$ at the boundaries, i.e., $\mathbf{r}\in\partial V$) is different from zero. Therefore, in Eq. (18.4b), the product between $\Psi^\dagger$ and $\Psi$ represents the contribution to the functional of all neutrons that are re-entering the system. Again, as the minus sign in Eq. (18.4b) indicates, this contribution needs to be subtracted of Eq. (18.4a), since this term makes no distinction about the origin of the inserted neutron.

Equation (18.3a) is the well-known reciprocity relation [PrLa10]. Observing Eqs. (18.1), (18.2), and (18.3), it is reasonable to assume that $Q^\dagger$ is the source term of the equation whose solution is $\Psi^\dagger$. In fact, the equation centered on $\Psi^\dagger$ can be derived using different approaches, such as the operation reversal [Ga87]. The importance equation, also called adjoint transport equation, since it is the equation that is adjoint to Eq. (18.2a), appears as

$$\begin{aligned}
&-\hat{\boldsymbol{\Omega}}\cdot\nabla\Psi^\dagger(\mathbf{r},\hat{\boldsymbol{\Omega}},E)+\sigma_\mathrm{t}(\mathbf{r},E)\Psi^\dagger(\mathbf{r},\hat{\boldsymbol{\Omega}},E)= \\
&\int_0^\infty dE'\int_{4\pi}\sigma_\mathrm{s}(\mathbf{r},\hat{\boldsymbol{\Omega}}\cdot\hat{\boldsymbol{\Omega}}',E\to E')\Psi^\dagger(\mathbf{r},\hat{\boldsymbol{\Omega}}',E')d\hat{\boldsymbol{\Omega}}'+ \\
&\frac{\nu\sigma_\mathrm{f}(\mathbf{r},E)}{4\pi}\int_0^\infty dE'\int_{4\pi}\chi(\mathbf{r},E')\Psi^\dagger(\mathbf{r},\hat{\boldsymbol{\Omega}}',E')d\hat{\boldsymbol{\Omega}}'+Q^\dagger(\mathbf{r},\hat{\boldsymbol{\Omega}},E).
\end{aligned} \tag{18.5a}$$

Equation (18.5a) is subjected to the boundary conditions

$$\Psi^\dagger(\mathbf{r},\hat{\boldsymbol{\Omega}},E)=\Psi^{\dagger^b}(\mathbf{r},\hat{\boldsymbol{\Omega}},E),\quad \mathbf{r}\in\partial V,\ \hat{\boldsymbol{\Omega}}\cdot\mathbf{n}>0,\ 0<E<\infty, \tag{18.5b}$$

which possess a strict relation with the boundary conditions for the forward transport problem (Eq. (18.2b)) as will become apparent in the next section.

## 18.3 The Source Estimation Through the Adjoint Technique

Let us consider again the functional defined in Eq. (18.1). As briefly discussed in the previous section, the adjoint source keeps a strict relation with the functional $\mathtt{F}$. In fact, the definition of $Q^\dagger$ depends on the definition of the functional $\mathtt{F}$. When one requires $\mathtt{F}$ to represent a specific physical quantity, $Q^\dagger$ is automatically defined, such that $\mathtt{F}$ becomes this desired quantity. For example, if it is required for $\mathtt{F}$ to represent the reaction-rate of a given material at a specific point $\mathbf{r}^\star$, then $Q^\dagger$ becomes

$$Q^\dagger(\mathbf{r}, E) = \sigma_{\mathrm{x}}(\mathbf{r}, E)\delta(\mathbf{r} - \mathbf{r}^\star),$$

where $\sigma_{\mathrm{x}}(\mathbf{r}, E)$ is the reaction cross section of the material considered and $\delta$ is the Dirac delta function.

As we are interested in building an explicit relation between the power generated by a subcritical system and the neutron source distribution used to stabilize it, we assume that $\mathtt{F}$ represents the power generated by the system. That is,

$$\mathtt{F} = \mathtt{P}_{\mathrm{total}} = \int_V dV \int_0^\infty dE \int_{4\pi} \epsilon\sigma_{\mathrm{f}}(\mathbf{r}, E)\Psi(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E)d\hat{\Omega}, \tag{18.6a}$$

where $\epsilon$ is the average energy released in one fission event. From Eqs. (18.1) and (18.6a), one concludes that the adjoint source is

$$Q^\dagger(\mathbf{r}, E) = \epsilon\sigma_{\mathrm{f}}(\mathbf{r}, E). \tag{18.6b}$$

The adjoint angular flux associated with the adjoint source as in Eq. (18.6b) represents, in this case, the contribution that one neutron inserted into the system has to the generation of power by the whole subcritical system.

### *18.3.1 Boundary Conditions*

At this point we need to focus our attention on the boundary conditions considered in the neutron transport problem, in order to examine their relation with the adjoint boundary conditions and the impact of such conditions in the definition of the functional as presented in Eq. (18.3).

Thus, let us suppose that it is possible to know the origin of all neutrons entering the system through its boundaries. In this case, Eq. (18.2b) can be written as a sum of two different terms, i.e.,

$$\Psi^b(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) = \Psi^\star(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) + \Psi^{\star\star}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E), \quad \mathbf{r} \in \partial V, \ \hat{\boldsymbol{\Omega}} \cdot \mathbf{n} < 0, \ 0 < E < \infty, \tag{18.7}$$

where $\Psi^{\star}$ is the part of $\Psi^{b}$ composed of the neutrons that come into the system for the first time, and $\Psi^{\star\star}$ is the part of $\Psi^{b}$ composed of the neutrons that enter the system and, for some reason, had been previously inside the system. Substituting Eq. (18.7) into Eq. (18.3c), and recalling the definition of term $P_b\left[\Psi^{\dagger}, \Psi\right]$ given in the previous section, we write

$$P_b\left[\Psi^{\dagger}, \Psi\right] = \int_0^{\infty} dE \int_{\mathbf{n}\cdot\hat{\boldsymbol{\Omega}}<0} d\hat{\boldsymbol{\Omega}} \int_{\partial V} |\mathbf{n}\cdot\hat{\boldsymbol{\Omega}}| \Psi^{\dagger}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E)\Psi^{\star}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) dA \tag{18.8a}$$

and

$$\begin{aligned} &\int_0^{\infty} dE \int_{\mathbf{n}\cdot\hat{\boldsymbol{\Omega}}<0} d\hat{\boldsymbol{\Omega}} \int_{\partial V} |\mathbf{n}\cdot\hat{\boldsymbol{\Omega}}| \Psi^{\dagger}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E)\Psi^{\star\star}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) dA \\ &= \int_0^{\infty} dE \int_{\mathbf{n}\cdot\hat{\boldsymbol{\Omega}}>0} d\hat{\boldsymbol{\Omega}} \int_{\partial V} (\mathbf{n}\cdot\hat{\boldsymbol{\Omega}}) \Psi^{\dagger}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E)\Psi(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) dA. \end{aligned} \tag{18.8b}$$

In nuclear reactor physics problems, the angular flux of neutrons usually satisfies the vacuum and/or reflective boundary conditions. In the case of vacuum boundary conditions (VBC), it is assumed that no neutron can enter the system through its boundaries. In other words,

$$\Psi^{b}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) = 0, \quad \mathbf{r} \in \partial V, \ \hat{\boldsymbol{\Omega}}\cdot\mathbf{n} < 0, \ 0 < E < \infty.$$

Thus, for VBC, functions $\Psi^{\star}$ and $\Psi^{\star\star}$ are identically zero. Setting $\Psi^{\star}$ and $\Psi^{\star\star}$ as zero in Eqs. (18.8), we obtain

$$P_b\left[\Psi^{\dagger}, \Psi\right] = 0 \tag{18.9}$$

and, according to Eq. (18.8b) we conclude that

$$\Psi^{\dagger}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) = 0, \quad \mathbf{r} \in \partial V, \ \hat{\boldsymbol{\Omega}}\cdot\mathbf{n} > 0, \ 0 < E < \infty. \tag{18.10}$$

As no neutron can enter the system through its boundaries, considering VBC means that the contribution to the functional of neutrons that are inserted for the first time into the system is zero, as we notice from Eq. (18.9). Furthermore, from Eq. (18.10), we can also conclude that neutrons leaving the system have zero importance. As there is no possibility for a neutron leaving the system to return, once a neutron leaves the system, it will not contribute to the functional, hence it has zero importance.

In the reflective boundary conditions (RBC), it is considered that all neutrons leaving the system are reflected back into the system. In other words, the neutron angular fluxes in the incoming directions at the boundaries are settled as equal to the neutron angular fluxes in the outgoing directions at the same points. Thus,

$$\Psi^{b}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) = \Psi(\mathbf{r}, -\hat{\boldsymbol{\Omega}}, E), \quad \mathbf{r} \in \partial V, \ \hat{\boldsymbol{\Omega}}\cdot\mathbf{n} < 0, \ 0 < E < \infty.$$

As in the RBC the neutrons entering the system were all previously inside the system (inserted by the neutron source), we have $\Psi^{\star}$ equal to zero and

$$\Psi^{\star\star}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) = \Psi(\mathbf{r}, -\hat{\boldsymbol{\Omega}}, E), \quad \mathbf{r} \in \partial V, \ \hat{\boldsymbol{\Omega}} \cdot \mathbf{n} < 0, \ 0 < E < \infty. \tag{18.11}$$

Therefore, as no neutron enters the system, for the first time, the contribution to the functional of $P_b\left[\Psi^{\dagger}, \Psi\right]$ is also zero. Moreover, substituting Eq. (18.11) into Eq. (18.8b) we obtain the boundary conditions for the adjoint transport problem, associated with RBC for the forward neutron transport problem. That is,

$$\Psi^{\dagger}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E) = \Psi^{\dagger}(\mathbf{r}, -\hat{\boldsymbol{\Omega}}, E), \quad \mathbf{r} \in \partial V, \ \hat{\boldsymbol{\Omega}} \cdot \mathbf{n} > 0, \ 0 < E < \infty. \tag{18.12}$$

As all neutrons leaving the system are reflected back into the system, their contribution to the functional is equal to the contribution of neutrons entering the system, as we can infer from Eq. (18.12).

### *18.3.2 The Importance Matrix*

Considering the definition of the functional given in Eq. (18.6a) and $P_b\left[\Psi^{\dagger}, \Psi\right]$ as shown in Eq. (18.9) (valid for VBC and RBC), we can conclude from Eqs. (18.3) that

$$\mathtt{P}_{\text{total}} = \int_V dV \int_0^{\infty} dE \int_{4\pi} Q(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E)\Psi^{\dagger^{\text{total}}}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E)d\hat{\Omega}. \tag{18.13}$$

The superscript "total" that appears in $\Psi^{\dagger}$ is used to indicate that these functions are obtained by the solution of Eq. (18.5a) considering the adjoint source as given in Eq. (18.6b), and boundary conditions as defined in Eqs. (18.10) and/or (18.12).

To proceed, let us consider the spatial domain $V$ (subcritical system) as composed of $I$ subdomains $V_i$ (hereafter called regions), i.e.,

$$V = \bigcup_{i=1}^{I} V_i,$$

where the neutron source is uniform with respect to the spatial variables inside each region $V_i$. Moreover, we assume that it is given not only the total power generated by the system ($\mathtt{P}_{\text{total}}$) but also how this power is distributed. That is to say,

$$\mathtt{P}_{\text{total}} = \sum_{i=1}^{I} \mathtt{P}_i,$$

where the power generated in each region ($\mathtt{P}_i$) is considered to be known. Thus, we redefine the functional $\mathtt{F}$ as the power generated by the region $V_\mathrm{i}$ instead of the whole system as in Eq. (18.6a). In this case, we have

$$\mathtt{P}_i = \int_V dV \int_0^\infty dE \int_{4\pi} \epsilon\sigma_\mathrm{f}(\mathbf{r}, E)\zeta_i(\mathbf{r})\Psi(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E)d\hat{\Omega}, \tag{18.14a}$$

with

$$Q^\dagger(\mathbf{r}, E) = \epsilon\sigma_\mathrm{f}(\mathbf{r}, E)\zeta_i(\mathbf{r}), \tag{18.14b}$$

where

$$\zeta_i(\mathbf{r}) = \begin{cases} 1, & \text{if } \mathbf{r} \in V_i \\ 0, & \text{otherwise} \end{cases}.$$

Following a procedure which is analogous to the one that results in Eq. (18.13), we obtain

$$\mathtt{P}_i = \int_0^\infty dE \int_{4\pi} \sum_{k=1}^{I} Q_k(\hat{\boldsymbol{\Omega}}, E)\mathbb{V}_k\overline{\Psi}^{\dagger i}_k(\hat{\boldsymbol{\Omega}}, E)d\hat{\Omega}, \tag{18.15}$$

where we have defined

$$\overline{\Psi}^{\dagger \mathrm{i}}_k(\hat{\boldsymbol{\Omega}}, E) = \frac{1}{\mathbb{V}_k}\int_{V_\mathrm{k}} \Psi^{\dagger i}(\mathbf{r}, \hat{\boldsymbol{\Omega}}, E)dV,$$

with $\mathbb{V}_k$ standing for the volume of region $V_\mathrm{k}$. The subscript "$k$" in the neutron source is used to emphasize that this function is uniform with respect to the spatial variables inside each region $V_k$. Furthermore, the superscript "$i$" in $\Psi^\dagger$ is used to indicate that this function is obtained through the solution of Eq. (18.5a) considering the adjoint source as shown in Eq. (18.14b), and boundary conditions as in Eqs. (18.10) and/or (18.12).

Moreover, we use the energy multigroup formulation [PrLa10], wherein the energy variable is discretized in a finite number $G$ of contiguous energy groups, such that

$$E_\mathrm{min} = E_G < E_{G-1} < \cdots < E_g < E_{g-1} < \cdots < E_2 < E_1 = E_\mathrm{max},$$

where $E_\mathrm{min}$ is considered to be sufficiently small that neutrons with energy less than $E_\mathrm{min}$ are negligible and $E_\mathrm{max}$ is sufficiently large that neutrons with energy greater

than $E_{\max}$ are also negligible [PrLa10]. Assuming the neutron source as uniform with respect to the energy variable inside each energy group, Eq. (18.15) appears as

$$\mathtt{P}_i = \sum_{k=1}^{I} \sum_{g=1}^{G} Q_{k,g} \mathbb{V}_k \Phi_{k,g}^{\dagger^i}, \tag{18.16a}$$

where we have defined

$$\Phi_{k,g}^{\dagger^i} = \int_{E_g}^{E_{g-1}} dE \int_{4\pi} \overline{\Psi}_k^{\dagger^i}(\hat{\boldsymbol{\Omega}}, E) d\hat{\boldsymbol{\Omega}}. \tag{18.16b}$$

At this point we remark that we have also considered isotropic neutron sources in Eq. (18.16a). For $i = 1 : I$, we obtain the isotropic matrix equation

$$\mathbf{P} = \mathbf{L}^{\dagger} \mathbf{Q}, \tag{18.17a}$$

where $\mathbf{P}$ is an $I$-dimensional column vector composed of the prescribed power density distribution; $\mathbf{Q}$ is an $IG$-dimensional column vector composed of the uniform multigroup neutron sources; and $\mathbf{L}^{\dagger}$ is an $I \times IG$ matrix defined as

$$\mathbf{L}^{\dagger} = \begin{bmatrix} \mathbb{V}_1 \Phi_{1,1}^{\dagger^1} & \mathbb{V}_1 \Phi_{1,2}^{\dagger^1} & \mathbb{V}_1 \Phi_{1,3}^{\dagger^1} & \cdots & \mathbb{V}_1 \Phi_{1,G}^{\dagger^1} & \mathbb{V}_2 \Phi_{2,1}^{\dagger^1} & \cdots & \mathbb{V}_I \Phi_{I,G}^{\dagger^1} \\ \mathbb{V}_1 \Phi_{1,1}^{\dagger^2} & \mathbb{V}_1 \Phi_{1,2}^{\dagger^2} & \mathbb{V}_1 \Phi_{1,3}^{\dagger^2} & \cdots & \mathbb{V}_1 \Phi_{1,G}^{\dagger^2} & \mathbb{V}_2 \Phi_{2,1}^{\dagger^2} & \cdots & \mathbb{V}_I \Phi_{I,G}^{\dagger^2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbb{V}_1 \Phi_{1,1}^{\dagger^I} & \mathbb{V}_1 \Phi_{1,2}^{\dagger^I} & \mathbb{V}_1 \Phi_{1,3}^{\dagger^I} & \cdots & \mathbb{V}_1 \Phi_{1,G}^{\dagger^I} & \mathbb{V}_2 \Phi_{2,1}^{\dagger^I} & \cdots & \mathbb{V}_I \Phi_{I,G}^{\dagger^I} \end{bmatrix}. \tag{18.17b}$$

We observe that each row of matrix $\mathbf{L}^{\dagger}$ is composed of the solution of Eq. (18.5a) considering boundary conditions given by Eq. (18.10) and/or (18.12) and appropriate adjoint sources, cf. Eq. (18.14b) for $i = 1 : I$.

The matrix displayed in Eq. (18.17b) is referred to as importance matrix. Each element of this matrix represents the importance that all neutrons inserted into the system at a region $k$ with kinetic energy in group $g$ have to the generation of power in region $i$, cf. Eqs. (18.16). The calculation of the importance matrix allows us to explicitly correlate a prescribed distribution of power ($\mathbf{P}$) with the neutron source distribution ($\mathbf{Q}$) required to drive the subcritical system into the given level of power.

If we consider the subcritical system as composed only of multiplying regions, as well as $G = 1$, we can obtain a unique solution for Eq. (18.17a). That is,

$$\mathbf{Q} = \mathbf{L}^{\dagger^{-1}} \mathbf{P},$$

provided that $\mathbf{L}^{\dagger^{-1}}$ exists. However, if we consider that there are non-multiplying regions composing the subcritical system and/or $G > 1$, which accounts for most

of realistic situations, the linear system presented in Eq. (18.17a) does not have a unique solution. In particular, with respect to the existence of non-multiplying regions, the rows of Eq. (18.17b) associated with such regions have all elements equal to zero. As in a non-multiplying region there is no fissile material to produce fission, i.e., $\mathtt{P} = 0$, the importance that a neutron inserted in any place into the system has to the generation of power in this non-multiplying region is clearly zero. This situation can also be viewed, noticing that trivial solution of Eq. (18.5a) arises considering boundary conditions as Eqs. (18.10) and/or (18.12) and adjoint source equal to zero ($\sigma_{\mathrm{f}}(\mathbf{r}, E) = 0$ when $\zeta_i(\mathbf{r}) = 1$).

As the linear system presented in Eq. (18.17a) does not have a unique solution, auxiliary information must be given in order to determine a unique neutron source distribution. This situation actually gives to the present methodology an interesting flexibility, since different neutron source distributions can be generated (according to the auxiliary information), all of them driving the subcritical system to the prescribed distribution of power.

## 18.4 An Illustrative Example

To illustrate the use of the adjoint technique to estimate neutron source distributions driving a subcritical system to a prescribed power distribution, we consider a 2-energy group slab-geometry subcritical problem, as depicted in Fig. 18.1. The material parameters for this problem are displayed in Table 18.1. Moreover, we adopt $\nu = 2.5$ and $\epsilon = 200\,\mathrm{MeV}$. At this point we remark that the numerical values of the macroscopic cross sections presented in this section are fictitious. The goal here is just to illustrate the application of the present methodology.

The adjoint transport equation for slab-geometry problems appears in the multigroup discrete ordinates formulation ($\mathrm{S}_N$) [LeMi93] as

$$-\mu_m \frac{d}{dx}\Psi^{\dagger i}_{m,g}(x) + \sigma_{\mathrm{t}_g}\Psi^{\dagger i}_{m,g}(x) = \frac{1}{2}\sum_{g'=1}^{G}\sigma_{\mathrm{s}_{g\to g'}}\sum_{n=1}^{N}\Psi^{\dagger i}_{n,g'}(x)\omega_n + \frac{\nu\sigma_{\mathrm{f}_g}}{2}\sum_{g'=1}^{G}\chi_{g'}\sum_{n=1}^{N}\Psi^{\dagger i}_{n,g'}(x)\omega_n + Q^{\dagger}_g(x), \tag{18.18a}$$

$$m = 1:N,\ g = 1:G,\ x \in V_k\ \ (k = 1:5),$$

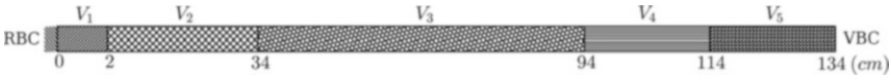



**Fig. 18.1** Geometry of the test problem

**Table 18.1** Material parameters of the test problem

| Material parameters | | $\sigma_{t_g}$ (cm$^{-1}$) | $\nu\sigma_{f_g}$ (cm$^{-1}$) | $\sigma_{s_{g\to g'}}$ (cm$^{-1}$) | | $\chi_g$ |
|---|---|---|---|---|---|---|
| | | | | $g'=1$ | $g'=2$ | |
| $V_1$ | $g=1$ | 2.31800E−01[a] | 0.00000E+00 | 1.82450E−01 | 1.56300E−02 | |
| | $g=2$ | 8.38000E−01 | 0.00000E+00 | 0.00000E+00 | 7.33100E−01 | |
| $V_2$ | $g=1$ | 2.36570E−01 | 1.12038E−02 | 2.09560E−01 | 1.45311E−02 | 1 |
| | $g=2$ | 8.22800E−01 | 1.04762E−01 | 0.00000E+00 | 7.06506E−01 | 0 |
| $V_3$ | $g=1$ | 2.31800E−01 | 8.66721E−03 | 1.84200E−01 | 1.63800E−02 | 1 |
| | $g=2$ | 8.71460E−01 | 1.88993E−01 | 0.00000E+00 | 7.42840E−01 | 0 |
| $V_4$ | $g=1$ | 3.37600E−01 | 0.00000E+00 | 3.03126E−01 | 1.01203E−03 | |
| | $g=2$ | 9.99500E−01 | 0.00000E+00 | 0.00000E+00 | 3.13126E−01 | |
| $V_5$ | $g=1$ | 1.78120E−01 | 0.00000E+00 | 8.00710E−02 | 3.43413E−03 | |
| | $g=2$ | 1.17616E+00 | 0.00000E+00 | 0.00000E+00 | 1.14458E−01 | |

[a] Read as $2.31800\times10^{-1}$

where the adjoint source for this case is

$$Q_g^{\dagger}(x) = \epsilon\sigma_{f_g}\zeta_i(x), \tag{18.18b}$$

with

$$\zeta_i(x) = \begin{cases} 1, & \text{if } x \in V_i \\ 0, & \text{otherwise} \end{cases}, \quad i = 1:5.$$

In Eq. (18.18a) $N$ is the number of discrete directions, $G$ is the number of energy groups, and $\omega_n$ is the weight of the angular quadrature. The superscript "$i$" is used to emphasize the relation between the adjoint angular flux and the adjoint source as defined in Eq. (18.18b). That is,

$$\Psi_{m,g}^{\dagger^i}(x) = \int_{E_g}^{E_{g-1}} \Psi^{\dagger^i}(x, \mu_m, E)dE.$$

Furthermore, we consider the macroscopic cross sections as uniform with respect to the spatial variable within each region $V_i$, viz Table 18.1.

As described in the previous section, we need to calculate the importance matrix in order to correlate the neutron source distribution and the power generated by the system. However, as we consider here a slab-geometry problem, the vector **P** presented in Eq. (18.17a) represents, in fact, the distribution of power per unit cross sectional area in accordance with Eqs. (18.16). Using the fact that for slab-geometry

problems the adjoint angular flux is considered to be uniform in the Y-Z plane, we rewrite Eq. (18.16a) as

$$\mathtt{P}_i = \sum_{k=1}^{I} \sum_{g=1}^{G} Q_{k,g} H_k A \Phi_{k,g}^{\dagger i}, \tag{18.19a}$$

where we have defined

$$\Phi_{k,g}^{\dagger i} = \frac{1}{H_k} \int_{x_{k-1}}^{x_k} \sum_{m=1}^{N} \Psi_{m,g}^{\dagger i}(x) \omega_m \, dx, \tag{18.19b}$$

with $H_k = x_k - x_{k-1}$, the width of region $V_k$. In Eq. (18.19a) the volume $\mathbb{V}_k$ is considered to be $\mathbb{V}_k = H_k A$, where $A$ is a constant cross sectional area. Dividing Eq. (18.19a) by $A$, we obtain

$$\widehat{\mathtt{P}}_i = \sum_{k=1}^{I} \sum_{g=1}^{G} Q_{k,g} H_k \Phi_{k,g}^{\dagger i}, \tag{18.20}$$

where $\widehat{\mathtt{P}}_i = \mathtt{P}_i / A$. At this point we remark that Eq. (18.20) holds due to the fact that we have considered that the system's volume has a constant cross sectional area $A$.

Furthermore, in order to calculate the importance matrix we solve Eq. (18.18a) 2 times (number of multiplying regions) considering the adjoint source properly, i.e., by setting $i = 2$ and $i = 3$ in Eq. (18.18b). To solve Eq. (18.18a) we use the Response Matrix (RM) method [LeEtAl21] considering $N = 16$. The RM method is free from spatial truncation errors, as it generates numerical results for the adjoint angular fluxes in multilayer slabs that agree with the numerical values obtained from the analytical solution of the energy multigroup adjoint $S_N$ problems. More details about the RM method can be found in reference [LeEtAl21]. Table 18.2 presents the importance matrix as generated for the problem depicted in Fig. 18.1.

As this is an underdetermined problem, we need additional information in order to obtain a unique solution (Eq. (18.17a)). Thus, we consider the neutron source distribution as

$$Q_{k,g} = \begin{cases} q_1, \text{ if } (k,g) = (2,1) \\ q_2, \text{ if } (k,g) = (3,2) \\ 0, \quad \text{otherwise} \end{cases}. \tag{18.21}$$

Now, as $\widehat{\mathtt{P}}_1 = \widehat{\mathtt{P}}_4 = \widehat{\mathtt{P}}_5 = 0$ (non-multiplying regions) and by setting $\widehat{\mathtt{P}}_2 = \widehat{\mathtt{P}}_3 = 0.5\,\text{MW/cm}^2$, considering the importance matrix as displayed in Table 18.2 and the

**Table 18.2** The importance matrix

| $\mathbf{L}^{\dagger}_{a,b}$ | | $b$ | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| $a$ | 1 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 |
| | 2 | 1.17797E−16[a] | 7.24480E−17 | 2.58976E−15 | 2.99135E−15 | 4.85430E−16 |
| | 3 | 1.55041E−17 | 7.13617E−18 | 6.75813E−16 | 6.62154E−16 | 2.85270E−15 |
| | 4 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 |
| | 5 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 |
| $\mathbf{L}^{\dagger}_{a,b}$ | | $b$ | | | | |
| | | 6 | 7 | 8 | 9 | 10 |
| $a$ | 1 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 |
| | 2 | 7.25905E−16 | 6.08541E−19 | 4.46152E−20 | 8.15136E−21 | 2.08975E−29 |
| | 3 | 6.30994E−15 | 1.09166E−16 | 1.58733E−17 | 1.44139E−18 | 7.30164E−27 |
| | 4 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 |
| | 5 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 | 0.00000E+00 |

[a] Read as $1.17797 \times 10^{-16}$

**Table 18.3** Power generated by the forward transport problem with $S_{16}$ and $G = 2$

| | | Region | | | | |
|---|---|---|---|---|---|---|
| | | $V_1$ | $V_2$ | $V_3$ | $V_4$ | $V_5$ |
| Power (MW/cm$^2$) | $g = 1$ | 0.00000 | 0.230728 | 0.097421 | 0.00000 | 0.00000 |
| | $g = 2$ | 0.00000 | 0.269272 | 0.402579 | 0.00000 | 0.00000 |
| | Total | 0.00000 | 0.500000 | 0.500000 | 0.00000 | 0.00000 |

neutron source distribution as given in Eq. (18.21), we obtain: $q_1 = 1.76144 \times 10^{+14}\ neutrons/cm^3 s$ and $q_2 = 6.03743 \times 10^{+13}$ neutrons/cm$^3$s. In other words,

$$Q_{k,g} = \begin{cases} 1.76144 \times 10^{+14}, & \text{if } (k, g) = (2, 1) \\ 6.03743 \times 10^{+13}, & \text{if } (k, g) = (3, 2) \\ 0, & \text{otherwise} \end{cases}. \tag{18.22}$$

In order to verify if the neutron source distribution shown in Eq. (18.22) in fact drives the subcritical system to the prescribed distribution of power, we solve the forward neutron transport equation for slab-geometry problems in the multigroup ($G = 2$) and discrete ordinates ($S_{16}$) formulations, considering the same transport problem as depicted in Fig. 18.1 and neutron source distribution as given in Eq. (18.22). Then, we calculate the power per unit area generated in each region, as can be seen in Table 18.3. Furthermore, Fig. 18.2 displays the neutron scalar flux generated by this forward transport problem.

We remark that we have also used the RM method to solve the forward transport problem. Observing Table 18.3, we conclude that the neutron source distribution correctly drove the subcritical system to the given power distribution.

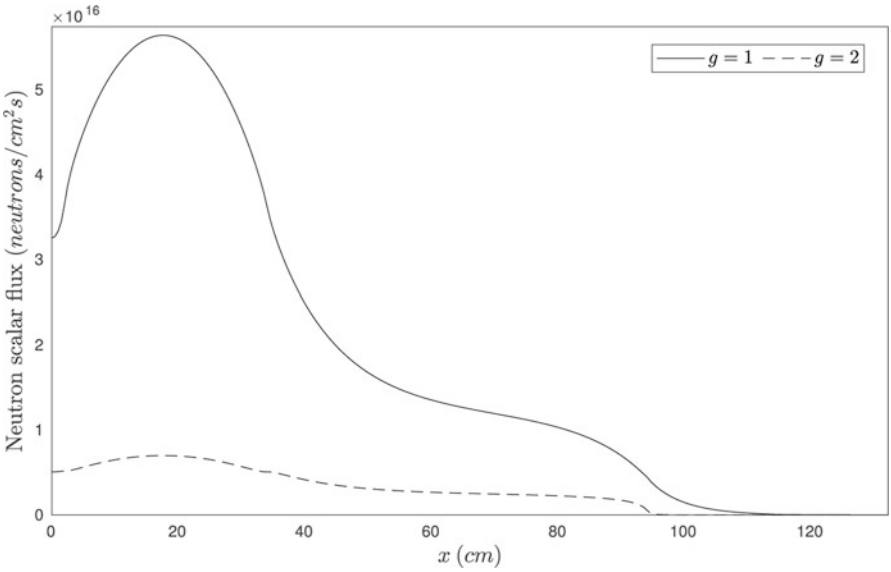



**Fig. 18.2** Neutron scalar flux generated by the forward transport problem with $S_{16}$ and $G = 2$

## 18.5 Concluding Remarks

Presented here is a more general description of the methodology given in [LeEtAl20, LeEtAl21], where the adjoint technique is used to estimate the neutron source distribution that drives a subcritical system to a prescribed distribution of power.

This methodology is based on a relation between a linear functional with respect to the neutron angular flux and the importance function. The importance function, also termed adjoint angular flux, represents the contribution that one neutron inserted into the system has to the generation of power. Under a practical viewpoint, the most important aspect of the present methodology lies on the calculation of the importance matrix. This matrix is composed of solutions of the adjoint transport equation, considering appropriate adjoint sources, which are related to the meaning of the functional. The importance matrix allows the direct correlation between the neutron source distribution and the power generated by the system.

Once the importance matrix is settled, one can obtain the desired neutron source distribution, with additional information, if necessary. This additional information, in fact, gives to the methodology an interesting flexibility, since it can change the arrangement of the estimated neutron source distribution. As an example, if we had considered in the previous section the neutron source distribution as

$$Q_{k,g} = \begin{cases} q_1, \text{ if } (k,g) = (1,2) \\ q_2, \text{ if } (k,g) = (3,1) \\ 0, \quad \text{otherwise} \end{cases},$$

we would have obtained, considering the same importance function: $q_1 = 5.82473 \times 10^{+15}$ neutrons/cm$^3$s and $q_2 = 1.60701 \times 10^{+14}$ neutrons/cm$^3$s. This alternative neutron source distribution drives the subcritical system to the generation of power displayed in Table 18.4. In addition, Fig. 18.3 displays the neutron scalar flux as generated by the forward transport problem with this new distribution source.

Analyzing Fig. 18.3, we can notice that this different additional information has led to a neutron source distribution that changed considerably the profile of the neutron scalar flux in the forward problem. However, despite the change in the shape of the neutron angular flux, which suggests a modification in the power generated by the system, the new neutron source distribution still drives the subcritical system to the prescribed power distribution, as can be seen in Table 18.4.

We conclude this chapter by pointing out that the interpretation of the adjoint angular flux as an importance function, and its subsequent use in the development of the offered methodology, is not unique. In fact, the theoretical approach considered in this work is referred to as heuristic approach [Ga87]. Other approaches, such

**Table 18.4** Power generated by the forward transport problem with the alternative neutron source distribution

| | | Region | | | | |
|---|---|---|---|---|---|---|
| | | $V_1$ | $V_2$ | $V_3$ | $V_4$ | $V_5$ |
| Power (MW/cm$^2$) | $g = 1$ | 0.00000 | 0.178961 | 0.133306 | 0.00000 | 0.00000 |
| | $g = 2$ | 0.00000 | 0.321039 | 0.366694 | 0.00000 | 0.00000 |
| | Total | 0.00000 | 0.500000 | 0.500000 | 0.00000 | 0.00000 |

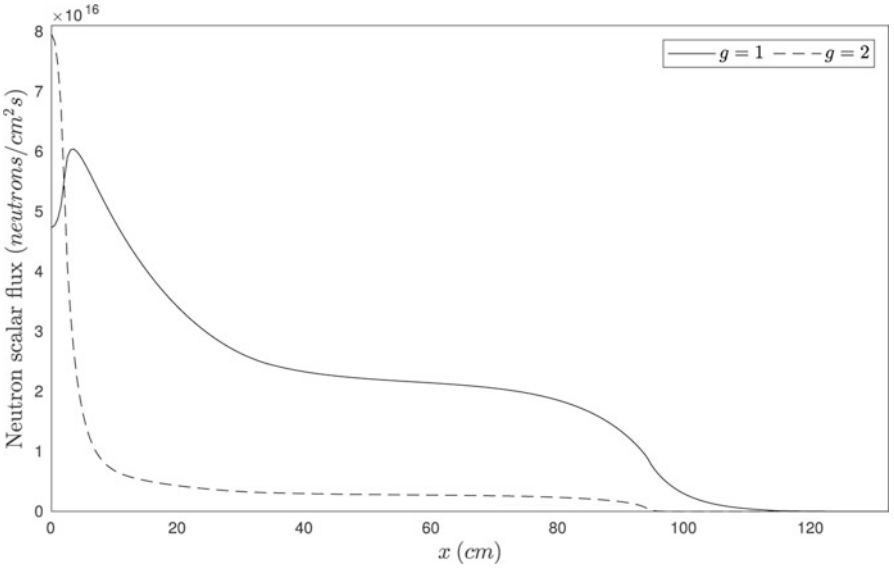



**Fig. 18.3** Neutron scalar flux generated by the forward transport problem with the alternative neutron source distribution

as the variational [Po67] and differential [Ob76, PrLa10] approaches, can also be used in the development of the present methodology. Nonetheless, regardless of the approach that is considered, the prescribed power density and the neutron source distribution will still be correlated as given in Eq. (18.17a).

**Acknowledgments** This study was financed in part by the Introducão de Aperfeicoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001, and Fundacão Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro—Brasil (FAPERJ). R.C. Barros also acknowledges support from Conselho Nacional de Desenvolvimento Científico e Tecnológico—Brasil (CNPq).

## References

[Ga87] Gandini, A.: Generalized perturbation theory (GPT) methods. A heuristic approach. In: Lewins, J., Becker, M. (eds.) Advances in Nuclear Science and Technology, vol. 19. Springer, Boston (1987)

[LeEtAl20] Moraes, L. R. C., Alves Filho, H., Barros, R. C.: Estimation of neutron sources driving prescribed power generations in subcritical systems using one-speed two-dimensional discrete ordinates formulations. Ann. Nucl. Energy **136**, 107053 (2020)

[LeEtAl21] Moraes, L. R. C., Mansur, R. S., Moura, C. A., Curbelo, J. P., Alves Filho, H., Barros, R. C.: A response matrix method for slab-geometry discrete ordinates adjoint calculations in energy-dependent neutral particle transport. J. Comput. Theor. Transp. **50**, 159–179 (2021)

[LeMi93] Lewis, E. E., Miller, W. F.: Computational Methods of Neutron Transport. American Nuclear Society, Illinois (1993)

[NiEtAl01] Nifenecker, N., David, S., Loiseaux, J. M., Meplana, O.: Basics of accelerator driven subcritical reactors. Nucl. Instrum. Methods Phys. Res. Sect. A Accelerators Spectrometers Detect. Assoc. Equip. **463**, 428–467 (2001)

[Ob76] Oblow, E.M.: Sensitivity theory from a differential viewpoint. Nucl. Sci. Eng. **59**, 187–189 (1976)

[Po67] Pomraning, G.: Variational principle for eigenvalue equations. J. Math. Phys. **8**, 149–160 (1967)

[PrLa10] Prinja, A.K., Larsen, E.W.: General principles of neutron transport. In: Cacuci, D.G. (ed.) Handbook of Nuclear Engineering, chap. 5. Springer, New York (2010)

# Chapter 19
# The Nodal $\mathrm{LTS}_N$ Solution and a New Approach to Determine the Outgoing Angular Flux at the Boundary in a Rectangular Domain


**A. R. Parigi, C. F. Segatto, B. E. J. Bodmann, and F. C. da Silva**

## 19.1 Introduction

One of the approaches to solve neutral particle transport in multiplicative or partially multiplicative media is the discrete ordinate method, also known as the $S_N$ *Ansatz* [Se95]. Recent literature has shown that this method is a convenient starting point for an analysis of the problem and for developments of approximate solutions in analytical representation with potential application in a variety of reactor problems. The basic idea is based upon discretizing the continuous angular variables in the transport equation, so that the transport equation becomes a system of coupled partial differential equations. A general derivation of the neutron transport equation based on microscopic dynamics may be found in references [BoEtAl83, Sp78], and the simplification to the $S_N$ equation is reported in [SeEtAl12].

If the dimension of the domain is two- or three-dimensional, a pathological problem shows up, namely the angular fluxes at the boundaries shall be provided so that a unique solution may be determined, which in the one-dimensional case is not necessary. This peculiarity is independent of the specific approach, whether numerical or (semi-)analytical, it rather stems from the fact that in the one-dimensional problem the boundary is not connected (i.e., simply two end points), whereas for domain dimensions larger than one the boundary of the typically

A. R. Parigi (✉)
Science and Technology Farroupilha, Federal Institute of Education, São Vicente do Sul, RS, Brazil
e-mail: aline.parigi@iffarroupilha.edu.br

C. F. Segatto · B. E. J. Bodmann
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: bardo.bodmann@ufrgs.br

F. C. da Silva
Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_19

convex domain is connected. In the further the $S_N$ equation is solved using a well-established method in the treatment of multidimensional transport, which is based on integrating out the transverse degree of freedom of the spatial variables in the transport equation. This procedure allows in principle to derive the analytical solution of the $S_N$ equation system, but using approximations for the transverse-leakage terms which arise from the integration step and the scattering source term, respectively.

A number of attempts to obtain analytical solutions already exist in the literature, where the ones relevant for the present discussion are derived using besides the discrete ordinate description also the Laplace integral transform. In this context, the $LTS_N$ method for two- and three-dimensional transport problems are reported in references [ZaEtAl95, ZaEtAl97, Ha02, Ha06], while in [HaEtAl03] the question of convergence of the discrete ordinate solution towards the continuous solution in the angular variables is addressed. Following in some part the reasoning of the previous works, in the further we present the two-dimensional nodal $LTS_N 2D$ solution using a new approach for the determination of the unknown angular fluxes on the contours, which were introduced *adhoc* in the aforementioned references and citations therein.

Recalling that the problem of the unknown angular flux on the boundary does not exist in the one-dimensional formulation of the $S_N$ equation, one may use this fact and construct the two-dimensional solution from the one-dimensional version. To this end the two-dimensional domain is segmented into narrow straight subdomains, so that in each strip one may consider a one-dimensional transport problem. Then the originally unknown angular fluxes in the boundaries are approximated by the one-dimensional $LTS_N$ solution in the contours of each line, which represents the strip. For a sufficiently large number of subdomains this approximation should already provide an acceptable solution for the two-dimensional problem and especially represent the "better physics" as compared to the necessary but nevertheless "arbitrary" assumptions for the angular flux at the respective domain boundaries as commonly adopted in the literature.

## 19.2 The Integrated $S_N$ Equations

Nuclear scenarios are typically analyzed in a rectangular domain, which might represent a nuclear fuel cell or a region of the nuclear reactor core. The reduction of an initially three-dimensional problem to a two-dimensional problem may be justified by the fact that the horizontal dimensions of a fuel element are typically very much smaller than the element height, so that one may consider approximate translational symmetry along the vertical axis and consequently the problem depends only on the horizontal coordinates $(x, y)$. Moreover, in the horizontal plane one may assume an approximate symmetry under reflection across the $x$ and $y$ axis, respectively, so that only one quarter of the geometric cross section of the problem shall be considered. Thus, this has implications on the type of boundary conditions, namely reflective conditions at $(0, y)$ and $(x, 0)$ due to symmetry and

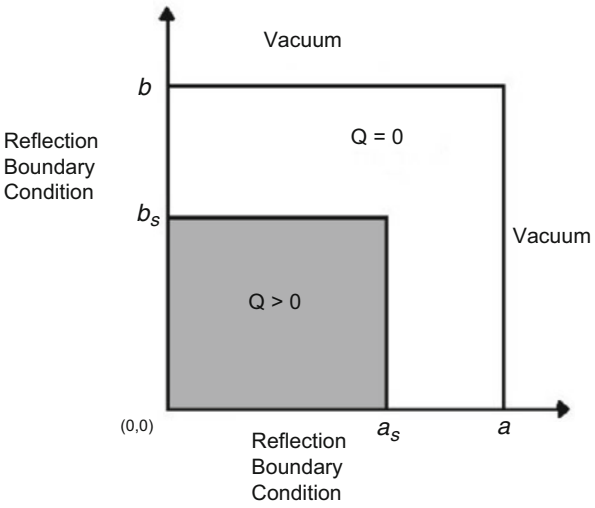



**Fig. 19.1** Rectangular domain $[0, a] \otimes [0, b]$ with neutron source region (gray) $[0, a_s] \otimes [0, b_s]$ and reflective and vacuum boundary conditions

vacuum boundary conditions at $(a, y)$ and $(x, b)$ (see Fig. 19.1). In this sense one considers an idealized neutrons transport problem defined in a rectangular domain $\Re = \{(x, y) \in [0, a] \otimes [0, b]\}$ with the aforementioned boundary conditions of vacuum and reflection type together with a multiplicative region representing a neutron source $\Re_Q = \{(x, y) \in [0, a_s] \otimes [0, b_s]\}$ with $a_s < a$ and $b_s < b$. As already made plausible the $S_N$ transport equation is defined in two-dimensional Cartesian geometry with sectionally homogeneous media, i.e., a region where fission occurs $(x, y) \in [0, a_s] \otimes [0, b_s]$, and a region $(x, y) \in [0, a] \otimes [0, b] \backslash [0, a_s] \otimes [0, b_s]$ where only absorption and scattering are the relevant nuclear reaction processes. Additionally discrete ordinates are understood, further we assume that the dominant collision process is described by isotropic scattering and for simplicity the energy dependence of the general transport equation was integrated out. In the transport equation, which we define as our starting point of the discussion to follow, the scattering integral was tacitly replaced by a quadrature approximation, where in this work the *Level Symmetric Quadrature- $LQ_N$* was used (for details see, for instance, reference [LeEtAl93]).

$$\mu_m \frac{\partial \Psi_m(x, y)}{\partial x} + \eta_m \frac{\partial \Psi_m(x, y)}{\partial y} + \sigma_t \Psi_m(x, y) = Q(x, y) + \frac{\sigma_s}{4} \sum_{n=1}^{M} \omega_n \Psi_m(x, y) \tag{19.1}$$

Here, $m = 1, \ldots, M$ are the discrete angular directions where the total number of directions $M = \frac{N(N+2)}{2}$ is related to the quadrature scheme $N$. $\Psi_m(x, y) \equiv \Psi_m(x, y, \hat{\Omega}_m)$ denotes the angular flux of particles at position $(x, y)$, which propagate in the direction $\hat{\Omega}_m$, where the discrete directions of motion $\hat{\Omega}_m = (\mu_m, \eta_m)$ are specified by the two angular cosines $\mu_m$ and $\eta_m$. The total macroscopic cross

section $\sigma_t$ and the isotropic scattering cross section $\sigma_s$ define the considered nuclear reactions, $\omega_m$ is the weight from the quadrature scheme associated with direction $m$ and $Q_(x, y) \equiv Q(x, y, \hat{\Omega}_m)$ signifies the neutron source at location $(x, y)$, which is different from zero in the subdomain where fission occurs and otherwise vanishes. The boundary conditions (19.2) of the considered scenario are the two reflective conditions followed by the two vacuum conditions.

$$\begin{aligned}
\Psi_m(0, y, \hat{\Omega}_m(\mu_m, \eta_m)) &= \Psi_m(0, y, \hat{\Omega}_m(-\mu_m, \eta_m)) \ , \quad \text{for} \mu_m > 0 \\
\Psi_m(x, 0, \hat{\Omega}_m(\mu_m, \eta_m)) &= \Psi_m(x, 0, \hat{\Omega}_m(\mu_m, -\eta_m)) \ , \quad \text{for} \eta_m > 0 \\
\Psi_m(a, y, \hat{\Omega}_m(\mu_m, \eta_m)) &= 0 \ , \quad \mu_m < 0 \\
\Psi_m(x, b, \hat{\Omega}_m(\mu_m, \eta_m)) &= 0 \ , \quad \eta_m < 0
\end{aligned} \tag{19.2}$$

As a next step towards solving Eq. (19.1) subject to the boundary conditions (19.2), the $S_N$ equation was integrated over the transverse direction $x$ from 0 to $a$ and multiplied by $\frac{1}{a}$ so that the dimension of a flux is maintained and we end up with a system of ordinary differential equations which depend on variable $y$ only.

$$\eta_m \frac{d\tilde{\Psi}_{xm}(y)}{dy} + \frac{\mu_m}{a} \left(\Psi_m(a, y) - \Psi_m(0, y)\right) + \sigma_t \tilde{\Psi}_{xm}(y) = Q_{xm}(y) + \frac{\sigma_s}{4} \sum_{n=1}^{M} \omega_n \tilde{\Psi}_{xn}(y) \tag{19.3}$$

Here the cross direction integrated flux and source term is

$$\tilde{\Psi}_{xm}(y) \equiv \frac{1}{a} \int_0^a \Psi_m(x, y) \, dx \ ,$$

$$Q_{xm}(y) \equiv \frac{1}{a} \int_0^a Q_m(x, y) \, dx \ ,$$

and $\Psi_m(0, y)$ and $\Psi_m(a, y)$ are the boundary conditions (19.2) for $m = 1, \ldots, M$. Proceeding in the same manner for the second spatial variable, Eq. (19.1) was integrated with respect to $y$ from 0 to $b$ and for dimensional reasons the integral was multiplied by $\frac{1}{b}$.

$$\mu_m \frac{d\hat{\Psi}_{ym}(x)}{dx} + \frac{\eta_m}{b} \left(\Psi_m(x, b) - \Psi_m(x, 0)\right) + \sigma_t \hat{\Psi}_{ym}(x) = Q_{ym}(x) + \frac{\sigma_s}{4} \sum_{n=1}^{M} \omega_n \hat{\Psi}_{yn}(x) \ , \tag{19.4}$$

where

$$\hat{\Psi}_{ym}(x) \equiv \frac{1}{b}\int_0^b \Psi_m(x, y)\, dy,$$

$$Q_{ym}(x) \equiv \frac{1}{b}\int_0^b Q_m(x, y)\, dy$$

and $\Psi_m(x, 0)$ and $\Psi_m(x, b)$ are the respective boundary conditions from Eq. (19.2), for $m = 1, \ldots, M$.

## 19.3 The Nodal $LTS_N$ Solution

In order to find the solutions of Eqs. (19.3) and (19.4) by the $LTS_N$ method, for convenience the set of equations was cast in matrix form,

$$\frac{d\tilde{\boldsymbol{\Psi}}_x(y)}{dy} - \mathbf{A}_y\tilde{\boldsymbol{\Psi}}_x(y) = \mathbf{Z}(y) \tag{19.5}$$

and

$$\frac{d\hat{\boldsymbol{\Psi}}_y(x)}{dx} - \mathbf{A}_x\hat{\boldsymbol{\Psi}}_y(x) = \mathbf{S}(x)\ , \tag{19.6}$$

where $\mathbf{A}_x$ and $\mathbf{A}_y$ are square matrices of order $M$ with the components given by

$$\mathbf{A}_x(i, j) = \begin{cases} -\frac{\sigma_t}{\mu_i} + \frac{\sigma_s\omega_j}{4\mu_i} & \text{if } i = j \\ \frac{\sigma_s\omega_j}{4\mu_i} & \text{if } i \neq j \end{cases}$$

and

$$\mathbf{A}_y(i, j) = \begin{cases} -\frac{\sigma_t}{\eta_i} + \frac{\sigma_s\omega_j}{4\eta_i} & \text{if } i = j \\ \frac{\sigma_s\omega_j}{4\eta_i} & \text{if } i \neq j \end{cases}\ .$$

Here, $\tilde{\boldsymbol{\Psi}}_x(y)$ and $\hat{\boldsymbol{\Psi}}_y(x)$ are vectors of order $M$, representing the angular flux averaged along the $x$ and $y$ degree of freedom, respectively, and the integrated source term vectors are

$$\mathbf{Z}(y) = \mathbf{N}^{-1}\mathbf{Q}_x(y) - \frac{1}{a}\mathbf{N}^{-1}\mathbf{M}\left(\boldsymbol{\Psi}(a, y) - \boldsymbol{\Psi}(0, y)\right)\ ,$$

$$\mathbf{S}(x) = \mathbf{M}^{-1}\mathbf{Q}_y(x) - \frac{1}{b}\mathbf{M}^{-1}\mathbf{N}\left(\boldsymbol{\Psi}(x, b) - \boldsymbol{\Psi}(x, 0)\right)\ .$$

The matrices $\mathbf{M}$ and $\mathbf{N}$ are diagonal of order $M$ with components containing the direction cosines $\mu_m$ and $\eta_m$. Further, $\boldsymbol{\Psi}(a, y)$, $\boldsymbol{\Psi}(0, y)$ are the boundary condition vectors of the two-dimensional problem at $x = a$ and $x = 0$, and $\boldsymbol{\Psi}(x, b)$, $\boldsymbol{\Psi}(x, 0)$ are the counterpart boundary condition vectors at $y = b$ and $y = 0$, respectively.

The next step to obtain the solution of (19.3) and (19.4) by the $LTS_N$ method is provided upon applying the Laplace transform in the respective spatial variable of the ordinary matrix differential equation (19.5) and (19.6) from which one obtains two linear matrix equation systems of order $M$, depending now on the dual complex variable $s$.

$$(s\mathbf{I} - \mathbf{A}_y)\bar{\tilde{\boldsymbol{\Psi}}}_x(s) = \bar{\mathbf{Z}}(s) + \tilde{\Psi}_x(0) \tag{19.7}$$

$$(s\mathbf{I} - \mathbf{A}_x)\bar{\hat{\boldsymbol{\Psi}}}_y(s) = \bar{\mathbf{S}}(s) + \hat{\Psi}_y(0) \tag{19.8}$$

Here, $\mathbf{I}$ is the usual identity matrix of order $M$ and

$$\bar{\hat{\boldsymbol{\Psi}}}_y(s) = \mathscr{L}\{\hat{\boldsymbol{\Psi}}_y(x)\}\,, \qquad \bar{\tilde{\boldsymbol{\Psi}}}_x(s) = \mathscr{L}\{\tilde{\boldsymbol{\Psi}}_x(y)\}\,,$$
$$\bar{\mathbf{S}}(s) = \mathscr{L}\{\mathbf{S}(x)\}\,, \qquad \bar{\mathbf{Z}}(s) = \mathscr{L}\{\mathbf{Z}(y)\}$$

are the Laplace transforms of the averaged angular fluxes and the cross direction integrated source terms, respectively. Now, the transformed problems (19.7) and (19.8) may be solved for non-singular matrices $(s\mathbf{I} - \mathbf{A})$,

$$\bar{\tilde{\boldsymbol{\Psi}}}_x(s) = \left(s\mathbf{I} - \mathbf{A}_y\right)^{-1}\left(\bar{\mathbf{Z}}(s) + \tilde{\boldsymbol{\Psi}}_x(0)\right) \tag{19.9}$$

and

$$\bar{\hat{\boldsymbol{\Psi}}}_y(s) = (s\mathbf{I} - \mathbf{A}_x)^{-1}\left(\bar{\mathbf{S}}(s) + \hat{\boldsymbol{\Psi}}_y(0)\right)\,. \tag{19.10}$$

Upon applying the inverse Laplace transform in (19.9) and (19.10) one obtains the solutions for the averaged angular fluxes in the original coordinates $y$ and $x$, respectively.

$$\tilde{\boldsymbol{\Psi}}_x(y) = \mathbf{Y}e^{\mathbf{E}^*(y)}\mathbf{Y}^{-1}\tilde{\boldsymbol{\Psi}}_x(0)$$
$$+\mathbf{Y}e^{\mathbf{E}y}\mathbf{Y}^{-1} * \left(\mathbf{N}^{-1}\mathbf{Q}_x(y) - \frac{1}{a}\mathbf{N}^{-1}\mathbf{M}\left(\boldsymbol{\Psi}(a, y) - \boldsymbol{\Psi}(0, y)\right)\right) \tag{19.11}$$

$$\hat{\boldsymbol{\Psi}}_y(x) = \mathbf{X}e^{\mathbf{D}^*(x)}\mathbf{X}^{-1}\hat{\boldsymbol{\Psi}}_y(0)$$
$$+\mathbf{X}e^{\mathbf{D}x}\mathbf{X}^{-1} * \left(\mathbf{M}^{-1}\mathbf{Q}_y(x) - \frac{1}{b}\mathbf{M}^{-1}\mathbf{N}\left(\boldsymbol{\Psi}(x, b) - \boldsymbol{\Psi}(x, 0)\right)\right) \tag{19.12}$$

In these two equations $*$ signifies the convolution operation, $\mathbf{D} = \text{diag}\{d_1, d_2, \ldots, d_M\}$ are the $M$ distinct eigenvalues $d_i$ of matrix $\mathbf{A}_x$, $\mathbf{E} = \text{diag}\{e_1, e_2, \ldots, e_M\}$ are the $M$ distinct eigenvalues $e_i$ of matrix $\mathbf{A}_y$. Further, the arguments of the exponential function of the homogeneous solutions are given by

$$\mathbf{D}^*(x) = \begin{cases} d_i x & \text{if } d_i < 0 \\ d_i(x-a) & \text{if } d_i > 0 \end{cases},$$

$$\mathbf{E}^*(y) = \begin{cases} e_i y & \text{if } e_i < 0 \\ e_i(y-b) & \text{if } e_i > 0 \end{cases}.$$

Last but not least, $\mathbf{X}$ is the matrix with eigenvectors of $\mathbf{A}_x$ and correspondingly $\mathbf{Y}$ is the matrix containing the eigenvectors of $\mathbf{A}_y$. Thus, the solutions (19.11) and (19.12) are determined except for the unknown angular fluxes at the boundary $\boldsymbol{\Psi}(0, y)$, $\boldsymbol{\Psi}(a, y)$, $\boldsymbol{\Psi}(x, 0)$, and $\boldsymbol{\Psi}(x, b)$, respectively.

At this point, instead of choosing boundary conditions which may or may not represent a physically sound scenario, we follow a different reasoning and attempt to construct the two-dimensional solution from the one-dimensional one, since as mentioned in the introduction in the one-dimensional case the boundaries are points only. To this end, we consider the rectangular domain covered with a discrete and finite set of narrow stripes, either oriented parallel to the $x$-axis with discrete $r_k = y_k$ and $y_k \in [0, b]$ as shown in Fig. 19.2 or parallel to the $y$-axis with $r_k = x_k$ and $x_k \in [0, a]$, where in either case $k = 1, \ldots, J$ with $J \in \mathbb{N}$. If the differences of $r_k - r_{k-1} << a$ or $b$ depending on the case in consideration, then the angular fluxes at the extreme ends of the stripe may be interpreted as the ones of a one-dimensional problem, so that the known angular fluxes from the one-dimensional problem shall be an acceptable approximation for the boundary values at either $x_k \in \{0, a\}$ or $y_k \in \{0, b\}$.

More specifically, along each line, we considered a heterogeneous and one-dimensional problem, so that the domain is divided into two regions, where one region contains a neutron source, whereas the second region has no neutron source. Note that even for lines that do not cross the region where the physical neutron source is located (the shaded region in Fig. 19.2) one has to admit a spurious source in order to avoid the trivial solution, which would arise otherwise. Thus, each one-dimensional problem is represented by the following equation $S_N$.

$$\gamma_m \frac{d\phi_{km}^{(i)}(z)}{dz} + \sigma_t \phi_{km}^{(i)}(z) = \frac{\sigma_s}{2} \sum_{n=1}^{\mathcal{N}} \phi_{kn}^{(i)}(z)\omega_n + q_k^{(i)}(z)$$

Here, $z$ represents the spatial variable, $\phi_{km}^{(i)}(z)$ is the one-dimensional angular flux of line $k$, in the direction $m$ and in region $i$ ($i \in \{1, 2\}$), where in region 1 the restriction $z < a_s$ or $b_s$ holds, so that the line crosses either the neutron source region if $r_k < b_s$ or $a_s$ and has a spurious source otherwise. Region 2 is the complement in the domain, where the medium is characterized by the nuclear reactions scattering

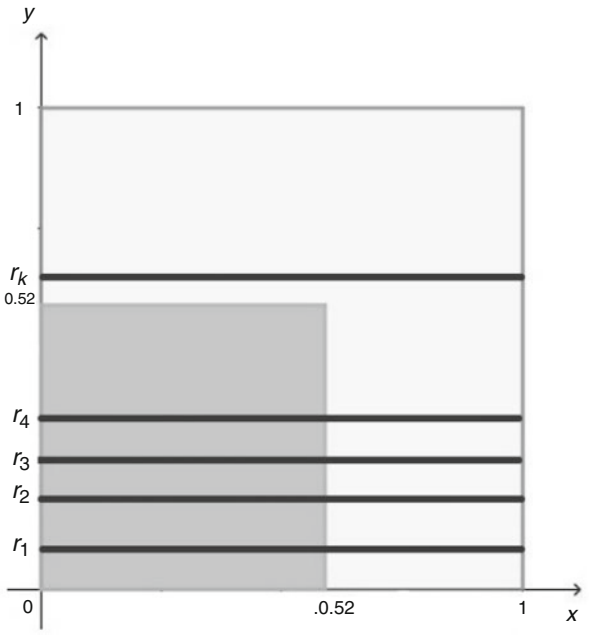



**Fig. 19.2** Approximation of the two-dimensional problem by an enumerable and finite set of one-dimensional problems

and absorption. In analogy to the two-dimensional problem $\gamma_m$ is the directional cosine and $\omega_n$ ($n = 1, \ldots, \mathcal{N}$) are the weights of the quadrature of order $\mathcal{N}$. The source term for region 1 is then given by

$$q_k^{(1)}(z) = \begin{cases} 1 & \forall z \in [0, 0.52] \text{ if } r_k \leq 0.52 \\ e^{-\beta(r_k - 0.52)} & \forall z \in [0, 0.52] \text{ if } y_k > 0.52 \end{cases},$$

while in region 2 the source term vanishes, $q_k{}^{(2)}(z) = 0$ with $z \in (0.52, 1]$, because here the medium is no longer multiplicative, or physically speaking does not contain nuclear fuel. The macroscopic scattering and total cross sections $\sigma_s$ and $\sigma_t$ of the one-dimensional problems are the same as for the two-dimensional case.

However, differently from the two-dimensional problem, where only the shaded region (see Fig. 19.2) has a non-vanishing source term, the one-dimensional neutron transport problem needs source terms for the lines with $r_k > b_s$ or $a_s$ and we assume them to be of the type $q(z)_k^{(1)} = e^{-\beta(y_k - b_s)}$, where $\beta$ is a constant to be determined *a priori* in a way that this term is sufficiently close to zero but finite to guarantee that the found solution does not coincide with the trivial one. Moreover, this term was crucial in order to preserve the reflexive boundary condition for the one-dimensional problem, which circumvents the characteristics and consequences of a homogeneous and non-multiplicative medium. For all the obtained results a

numerical value for $\beta = 25$ has proven to be suitable. Then, the boundary conditions for $z = 0$ are

$$\phi_k^{(1)}(0, \gamma_n) = \phi_k^{(1)}(0, -\gamma_n) \qquad \text{with } \gamma_n > 0 \text{ and } n = \{1, \ldots, \frac{\mathcal{N}}{2}\} ,$$

while at the interface between the regions with and without the source term (at $z = a_s$ or $b_s$) a continuity condition holds,

$$\phi_k^{(1)}(a_s \vee b_s, \gamma_n) = \phi_k^{(2)}(a_s \vee b_s, \gamma_n) ,$$

and at the outer boundary at $z = a$ or $b$ the fluxes vanish.

$$\phi_k^{(2)}(a \vee b, \gamma_n) = 0 \qquad \text{for } n = \{\frac{\mathcal{N}}{2} + 1, \ldots, \mathcal{N}\}$$

From the application of the $LTS_N$ method to the one-dimensional case one finds the solution (for details see reference [SeEtAl99])

$$\boldsymbol{\phi}^{(1)}(z) = \mathbf{B}^{(1)}(z)\boldsymbol{\xi}^{(1)} + \mathbf{H}^{(1)}(z) \qquad \text{for } z \in [0, a_s \vee b_s]$$

$$\boldsymbol{\phi}^{(2)}(z) = \mathbf{B}^{(2)}(z)\boldsymbol{\xi}^{(2)} \qquad \text{for } z \in [a_s \vee b_s, a \vee b],$$

where the matrices $\mathbf{B}$ are

$$\mathbf{B}^i(z) = \begin{cases} \mathbf{X}e^{\mathbf{D}z} & \text{if } \mathbf{D} < 0 \\ \mathbf{X}e^{\mathbf{D}(z-L_i)} & \text{if } \mathbf{D} > 0 \end{cases}$$

and depending on the orientation of the one-dimensional problem in the two-dimensional domain $L_i = a_s \vee b_s$ for $i = 1$ and $L_i = a \vee b_s$ for $i = 2$. The vector $\mathbf{H}$ contains the source term as follows:

$$H^{(1)}(z) = \mathbf{X} \begin{cases} \int_0^z e^{\mathbf{D}(z-\zeta)}\mathbf{X}^{-1}q^{(1)}(\zeta)\, d\zeta & \text{if } \mathbf{D} < 0 \\ \int_z^{L_s} e^{\mathbf{D}(z-\zeta)}\mathbf{X}^{-1}q^{(1)}(\zeta)\, d\zeta & \text{if } \mathbf{D} > 0 \end{cases} ,$$

and here $\mathbf{X}$ is the matrix containing the eigenvectors and the diagonal matrix D with the distinct eigenvalues of the $LTS_N$ matrix of the one-dimensional problem, which is the one-dimensional analogue to the matrices $\mathbf{A}$ in Eqs. (19.5) and (19.6).

To estimate the angular fluxes of the two-dimensional problem on the boundaries one employs the findings from the one-dimensional angular fluxes and uses the $LTS_N$ method associated with the *DNI* technique (dummy-node inclusion) [ChEtAl00] to construct the angular fluxes in the two-dimensional problem. This new procedure opened a pathway to interpolate the directions of the two-dimensional problem by the values of the one-dimensional directions. More specif-

ically, we considered the direction cosines $\hat{\Omega}_m$, $m \in [1, \frac{M}{2}]$ included them in the quadrature scheme and associated with them null weights, so that the quadrature scheme of the one-dimensional problem becomes

- $\gamma_i = \cos(\hat{\Omega}_i)$ and the weights $\omega_i = 0$ for $i = 1, \ldots, \frac{M}{4}$;
- $\gamma_{i+\frac{M}{4}} = \gamma_i$ and the weights $\omega_{i+\frac{M}{4}} = \omega_i$ for $i = 1, \ldots, \aleph$, where $\aleph$ is the total number of directions of the one-dimensional problem;
- $\gamma_{i+\frac{M}{2}+\mathscr{N}} = \cos(\hat{\Omega}_i)$ and the weights $\omega_{i+\frac{M}{4}} + \mathscr{N} = 0$ for $i = \frac{M}{4} + 1, \ldots, \frac{M}{2}$.

As a consequence, this allows to approximate the unknown angular fluxes at the boundaries by the solution of the one-dimensional problem calculated in the boundaries of the one-dimensional problem, that is, according to the boundary conditions the estimated $\Psi_m(0, y_k)$ are given by

$$\Psi_m(0, y_k) = \Psi_{m+\frac{M}{4}}(0, y_k) \equiv \phi^1_{km}(0)$$

$$\Psi_{m+\frac{M}{2}}(0, y_k) = \Psi_{m+\frac{3M}{4}}(0, y_k) \equiv \phi^1_{k(m+\frac{M}{4}+\mathscr{N})}(0)$$

for $m = 1, \frac{M}{4}$. Now, in agreement with the outer boundary conditions $\Psi_m(a, y_k)$ may be estimated

$$\Psi_m(a, y_k) = \Psi_{m+\frac{3M}{4}}(a, y_k) \equiv \phi^2_{km}(L)$$

with $m = 1, \frac{M}{4}$, where $\phi^1_k(0)$ and $\phi^2_k(L)$ represent the angular fluxes at the origin and end of the domain of each one-dimensional transport problem in the narrow rectangle aligned with $r_k$, respectively.

In the same manner one approximates the fluxes in $\Psi(x, 0)$ and $\Psi(x, b)$, so that finally the solutions (19.11) and (19.12) are completely determined and the only necessary information is to determine the vectors $\tilde{\Psi}_x(0)$ and $\hat{\Psi}_y(0)$. To this end, two linear systems with $M$ equations each are solved, which are associated with the Eqs. (19.11) and (19.12), respectively, obtained by estimating the same equations for the boundary values $x = a$ and $y = b$.

## 19.4 Numerical Results

Next some numerical results obtained from the solution presented in the previous section are shown, considering the domain described in Fig. 19.1 where $a = b = 1$, $Q(x, y) = 1$ for $0 \leq x \leq a_s = 0.52$ and $0 \leq y \leq b_s = 0.52$. Note that all the dimensions are given in multiples of mean-free-paths, $\sigma_t = 1.0\,\text{cm}^{-1}$ and three situations for $\sigma_s$, $\sigma_s = 0.5\,\text{cm}^{-1}$, $\sigma_s = 0.1\,\text{cm}^{-1}$, and $\sigma_s = 0.05\,\text{cm}^{-1}$ were analyzed. For all the cases of $\sigma_s$ and $N$ from a strong to a weaker scattering medium, we used twenty directions for the one-dimensional problems associated with $\frac{M}{2}$

**Table 19.1** Scalar Fluxes for the present method and comparison to the findings of reference [Si18]

| | | | $x = 0.5$ | | $x = 0.7$ | | $x = 0.98$ | |
|---|---|---|---|---|---|---|---|---|
| $\sigma_s$ | $N$ | Spatial grid for DD | $LTS_N$ | DD [Si18] | $LTS_N$ | DD [Si18] | $LTS_N$ | DD [Si18] |
| 0.5 | 2 | $50 \times 50$ | 0.280 | 0.312 | 0.211 | 0.216 | 0.137 | 0.112 |
| | 4 | $50 \times 50$ | 0.319 | 0.314 | 0.221 | 0.196 | 0.128 | 0.097 |
| | 6 | $50 \times 50$ | 0.325 | 0.314 | 0.218 | 0.188 | 0.123 | 0.095 |
| | 8 | $50 \times 50$ | 0.328 | 0.3315 | 0.214 | 0.184 | 0.121 | 0.095 |
| | 12 | $50 \times 50$ | 0.330 | 0.3316 | 0.211 | 0.181 | 0.119 | 0.095 |
| | 16 | $50 \times 50$ | 0.330 | 0.317 | 0.209 | 0.180 | 0.118 | 0.095 |
| 0.1 | 2 | $50 \times 50$ | 0.211 | 0.3224 | 0.151 | 0.147 | 0.094 | 0.071 |
| | 4 | $100 \times 100$ | 0.229 | 0.3223 | 0.146 | 0.127 | 0.094 | 0.059 |
| | 6 | $200 \times 200$ | 0.231 | 0.3223 | 0.140 | 0.120 | 0.075 | 0.058 |
| | 8 | $200 \times 200$ | 0.232 | 0.3224 | 0.137 | 0.117 | 0.073 | 0.058 |
| | 12 | $500 \times 500$ | 0.233 | 0.3225 | 0.133 | 0.114 | 0.072 | 0.058 |
| | 16 | $500 \times 500$ | 0.234 | 0.226 | 0.131 | 0.113 | 0.171 | 0.058 |
| 0.05 | 2 | $50 \times 50$ | 0.204 | 0.3216 | 0.145 | 0.141 | 0.090 | 0.068 |
| | 4 | $400 \times 400$ | 0.220 | 0.3215 | 0.139 | 0.122 | 0.076 | 0.056 |
| | 6 | $400 \times 400$ | 0.223 | 0.3215 | 0.134 | 0.115 | 0.071 | 0.055 |
| | 8 | $500 \times 500$ | 0.224 | 0.3216 | 0.130 | 0.111 | 0.069 | 0.055 |
| | 12 | $1000 \times 1000$ | 0.225 | 0.3217 | 0.126 | 0.109 | 0.068 | 0.055 |
| | 16 | $1000 \times 1000$ | 0.225 | 0.218 | 0.124 | 0.108 | 0.067 | 0.055 |

directions corresponding to the two-dimensional problem. The numerical results obtained by this novel methodology are presented in Table 19.1 and compared with those obtained by the DD method—*Diamond Difference* of reference [Si18].

## 19.5 Conclusions

In this work we presented the first results obtained by the proposition of a new methodology to determine the solution of a neutron transport problem with isotropic scattering, with a fixed source and in two-dimensional Cartesian geometry. The solution was found upon integrating the $S_N$ equations in the spatial variables followed by the application of the $LTS_N$ method. Since it is a characteristics of nodal methods to establish auxiliary equations to represent the transverse-leakage terms, we developed a new approach for the treatment of the unknown angular fluxes on the boundaries. To be more specific, we considered the two-dimensional domain covered by a set of straight subdomains, in a way that the unknown fluxes may be approximated by the angular fluxes at the ends of the domain by a one-dimensional transport problem where each subdomain corresponds to narrow strip of the coverage. Thus, the unknown angular fluxes at the boundary of the two-

dimensional domain were estimated by the solution of the one-dimensional $LTS_N$ problem calculated at the endpoint of each straight line.

This formulation allowed to decouple the solutions of the average angular fluxes in the $x$ and $y$ directions, making it possible to transform the resulting linear system, so that the LTSN method could be applied. Through this approach, the solution of integrated $S_N$ problems becomes equivalent to the solution of the one-dimensional $S_N$ problem and the comparison of the present approximations coincide fairly well with those obtained by other procedures reported in the literature and giving support to our reasoning. Nevertheless, in the literature the unknown fluxes at the boundary are commonly chosen *adhoc* without any additional physical justification so the we believe to have made a step towards a more consistent solution of the problem not only from a mathematical but also from a physical point of view.

As future steps, we will elaborate error estimates of the found solution in order to analyze the necessity of a refinement of the procedure. Furthermore, this idea of building the two-dimensional solution from one-dimensional one can be modified so that each strip of a line problem does not necessarily be oriented along the $x$ or $y$ axis but may be oriented in a way to contain a section with physical source term instead of introducing a spurious term. These results can then be compared with the ones obtained in this work, so that one may evaluate the progress of the extended method with the current results.

## References

[BoEtAl83] Boldrighini, C., Bunimovich, L.A., Sinai, Y.G.: On the Boltzmann equation for the Lorentz gas. J. Stat. Phys. **32**(3), 477–501 (1983)

[ChEtAl00] Chalhoub, E.S., Garcia, R.D.M.: The equivalence between two techniques of angular interpolation for the discrete-ordinates method. J. Quant. Spectroscopy Radiative Transf. **64**, 517–535 (2000)

[Ha02] Hauser, E.B.: LTSN formulation for transport problems without azimuthal symmetry and time dependent problems (in Portuguese). Thesis, Federal University of Rio Grande do Sul, Graduate Program in Mechanical Engineering, Porto Alegre, Brazil (2002)

[Ha06] Hauser, E.B.: Development of an analytical nodal method for discrete ordinate problems in two and three dimensional cartesian geometry in homogeneous e heterogeneous domains (in Portuguese). Thesis, Federal University of Rio Grande do Sul, Graduate Program in Mathematics, Porto Alegre, Brazil (2006)

[HaEtAl03] Hauser, E.B., Pazos, R., Vilhena, M., Barros, R.C.: Solution and study of nodal neutron transport equation applying the LTSN DiagExp method. Int. Nucl. Inf. Syst. **35**(9), 303–307 (2003)

[LeEtAl93] Lewis, E., Mille, W.: Computational Methods of Neutron Transport. Wiley, New York (1993)

[Se95] Segatto, C.F.: On the solution of the two dimensional neutron transport equation by the LTSN method for high orders in the angular quadratures: LTSN 2D-Diag and LTSN 2D-DiagExp (in Portuguese). Thesis, Federal University of Rio Grande do Sul, Graduate Program in Mechanical Engineering, Porto Alegre, Brasil (1995)

[SeEtAl99] Segatto, C.F., Vilhena, M.T., Gomes, M.: The one-dimensional LTSN solution in a slab with high degree of quadrature. Ann. Nucl. Energy **26**, 925–934 (1999)

[SeEtAl12] Segatto, C.F., Vilhena, M.T., Goncalez, T.T.: On the analytical solution of the neutron *SN* equation in a rectangle assuming an exponential exiting angular flux at the boundary. Int. J. Nucl. Energy Sci. Technol. **7**, 45–56 (2012)

[Si18] Silva, F.C.: Files with Numerical Results by a Nodal Approach. Private Communication (2018)

[Sp78] Spohn, H.: The Lorentz process converges to a random flight process. Commun. Math. Phys. **60**(3), 277–290 (1978)

[ZaEtAl95] Zabadal, J., Vilhena, M.T., Barichello, L.B.: Solution of the three-dimensional one group discrete ordinates problem by the LTSN method. Ann. Nucl. Energy **22**, 131–134 (1995)

[ZaEtAl97] Zabadal, J., Vilhena, M.T., Barichello, L.B.: An analytical solution for the two-dimensional discrete ordinate problem in a convex domain. Prog. Nucl. Energy **31**, 225–228 (1997)

# Chapter 20
# A Numerical Study of the Convergence of Two Hybrid Convolution Quadrature Schemes for Broadband Wave Problems


**J. Rowbottom and D. J. Chappell**

## 20.1 Introduction

In this chapter, we compare the performance of two recently proposed hybrid methods [RoCh21] for numerically solving the wave equation in two spatial dimensions. The convolution quadrature (CQ) method is employed for the time discretisation [Lu88, Lu94, Ch11], which can be used to transform the original time-domain problem into a system of frequency domain Helmholtz problems with complex wavenumbers [BaSa08, BeEtAl17, MaEtAl20]. For a range of wavenumbers that will be considered as low frequencies, the Helmholtz problems will be solved numerically using a piecewise constant collocation boundary element method (BEM). The remaining wavenumbers will be considered as the high frequencies, and the Helmholtz models will be replaced by one of two alternative high-frequency approximations, leading to the two hybrid schemes that we compare in this study.

The first high-frequency approximation will be based on a plane-wave approximation in which the amplitudes are approximated via dynamical energy analysis (DEA) with a Petrov-Galerkin discretization, as discussed in [ChEtAl21]. DEA is an approach for modelling wave energy densities at high frequencies that was first proposed just over 10 years ago [Ta09]. DEA is based on a linear integral operator model (like the BEM) of phase-space density transport along ray trajectories between positions on the boundary of a domain or sub-domain. Recent developments have seen the capability of DEA extended to three-dimensional [BaEtAl17] and industrial applications [HaEtAl19], as well as stochastic propagation through uncertain structures [ChTa14, BaCh20]. The phase terms will then be approximated by matching the solutions calculated via BEM with an expression for the plane-wave

J. Rowbottom · D. J. Chappell (✉)
Nottingham Trent University, Nottingham, UK
e-mail: david.chappell@ntu.ac.uk



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_20

approximation, as discussed in more detail in Sect. 20.4.1. The second high-frequency approximation will be based on an incident illumination approximation where only the direct contribution of the source term on the boundary is included. Numerical experiments are then discussed, investigating the convergence of both hybrid methods when the wave problems are driven by a plane wave travelling into the domain.

## 20.2 Convolution Quadrature for the Wave Equation: Summary

Let $\Omega \subset \mathbb{R}^2$ be a finite domain with boundary $\Gamma = \partial\Omega$. We consider the following initial-boundary value problem (IBVP) for the homogeneous wave equation:

$$\Delta\Phi - \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} = 0, \quad \text{in} \quad \Omega \times (0, T), \tag{20.1}$$

with initial conditions

$$\Phi(\cdot, 0) = \partial_t\Phi(\cdot, 0) = 0, \quad \text{in} \quad \Omega, \tag{20.2}$$

and Neumann boundary condition

$$\frac{\partial\Phi}{\partial\hat{\mathbf{n}}} = F \quad \text{on} \quad \Gamma \times (0, T), \tag{20.3}$$

for some $T > 0$. Here, we assume $F$ is a real-valued function of space and time, $c > 0$ is the wave speed and $\hat{\mathbf{n}}$ is the unit outward normal to the boundary. We will consider problems when the boundary $\Gamma$ corresponds to an interface with a vibrating structure that generates an inhomogeneous boundary condition $F$ [ChEtAl08, MaEtAl20].

We consider solving the IBVP of the wave equation (20.1)–(20.3) by reformulating it as a direct boundary integral equation

$$\Phi(\boldsymbol{x}, t) = (\mathscr{S}F)(\boldsymbol{x}, t) - (\mathscr{D}\Phi)(\boldsymbol{x}, t) \quad \text{in} \quad \Omega \times [0, T]. \tag{20.4}$$

Here, $\mathscr{S}$ and $\mathscr{D}$ are, respectively, the single and double layer potential operators

$$(\mathscr{S}F)(\boldsymbol{x}, t) := \int_0^T \int_\Gamma G(\boldsymbol{x} - \boldsymbol{y}, t - \tau)F(\boldsymbol{y}, \tau)\mathrm{d}\Gamma_{\boldsymbol{y}}\mathrm{d}\tau,$$

$$(\mathscr{D}\Phi)(\boldsymbol{x}, t) := \int_0^T \int_\Gamma \frac{\partial G}{\partial\hat{\mathbf{n}}_{\boldsymbol{y}}}(\boldsymbol{x} - \boldsymbol{y}, t - \tau)\Phi(\boldsymbol{y}, \tau)\mathrm{d}\Gamma_{\boldsymbol{y}}\mathrm{d}\tau,$$

where the fundamental solution $G$ is given by

$$G(\boldsymbol{x}, t) = \frac{H(t - \|\boldsymbol{x}\|/c)}{2\pi\sqrt{t^2 - \|\boldsymbol{x}\|^2/c^2}},$$

and $H$ is the Heaviside step-function. Moving (20.4) from the interior domain $\Omega$ to the boundary $\Gamma$, one then obtains the following direct time-domain boundary integral equation

$$\left(\frac{1}{2}I + K\right)\Phi(\boldsymbol{x}, t) = (VF)(\boldsymbol{x}, t) \quad \text{on} \quad \Gamma \times [0, T], \tag{20.5}$$

where $V$ and $K$ are, respectively, the traces of $\mathscr{S}$ and $\mathscr{D}$ on $\Gamma$.

Note that boundary integral operators $V$ and $K$ in (20.5) are time convolution operators. We will employ a time discretisation of (20.5) based on the BDF2 multistep scheme outlined in [BaSa08, BeEtAl17, MaEtAl20]. In doing so we split the time interval $[0, T]$ into $N$ steps of equal length $\Delta t = T/N$ and compute an approximate solution at the discrete time-steps $t_n = n\Delta t$. We will make use of the Laplace transforms of the operators $V$ and $K$ evaluated at a set of Laplace domain frequencies $\zeta_l, l =, \ldots, \widetilde{N} - 1$, which we denote by

$$(\widetilde{V}(\zeta_l)F)(\boldsymbol{x}) = \int_\Gamma G_l(\boldsymbol{x} - \boldsymbol{y})\, F(\boldsymbol{y}) \mathrm{d}\Gamma_{\boldsymbol{y}},$$
$$(\widetilde{K}(\zeta_l)u)(\boldsymbol{x}) = \int_\Gamma \frac{\partial G_l}{\partial \hat{\mathbf{n}}_{\boldsymbol{y}}}(\boldsymbol{x} - \boldsymbol{y})\, u(\boldsymbol{y}) \mathrm{d}\Gamma_{\boldsymbol{y}}.$$

Here, $G_l$ is the fundamental solution to the Helmholtz equation in two dimensions given by

$$G_l(\boldsymbol{x}) = -\frac{\mathrm{i}}{4} H_0^{(1)}(k_l \|\boldsymbol{x}\|),$$

with $H_0^{(1)}$ being the zeroth order Hankel function of the first kind and $k_l = \mathrm{i}\zeta_l/c$ the wavenumber. The choice of Laplace domain frequencies $\zeta_l$ relate to those used in a numerical approximation of Cauchy's integral formula applied to the inversion of a $Z$-transform, where the contour is taken as a circle of radius $\lambda < 1$ [BaSa08, BeEtAl17, MaEtAl20] and one obtains

$$\zeta_l = \frac{\gamma(\lambda e^{-2\pi \mathrm{i} l/\widetilde{N}})}{\Delta t}.$$

The function $\gamma(z) = \frac{1}{2}(z^2 - 4z + 3)$ is the quotient of the generating polynomials of the BDF2 multistep method. We allow the choices of $N$ and $\widetilde{N}$ to be decoupled

in order to potentially over-resolve in the Laplace domain for better accuracy as proposed in [BeEtAl17], although in this study we fix $\widetilde{N} = 2N$.

The result of this time discretization process can be expressed as a system of boundary integral equations for the Helmholtz equation (see [BaSa08, BeEtAl17, MaEtAl20] for details) given by

$$\frac{1}{2}u_l(\boldsymbol{x}) + \left(\widetilde{K}(\zeta_l)u_l\right)(\boldsymbol{x}) = \left(\widetilde{V}(\zeta_l)\widetilde{F}_l\right)(\boldsymbol{x}), \quad \boldsymbol{x} \in \Gamma, \tag{20.6}$$

where

$$u_l = \sum_{n=0}^{N-1} \Phi_n^{\lambda} \lambda^n e^{-2\pi i l n/\widetilde{N}}, \quad \widetilde{F}_l = \sum_{n=0}^{N-1} F(\cdot, t_n)\lambda^n e^{-2\pi i l n/\widetilde{N}},$$

are the Z-transforms of $\Phi^{\Delta t,\lambda}$ and $F$, respectively, and $\Phi_n^{\lambda}$ denotes the solution of (20.5) after it has been semi-discretised in time using CQ.

Once we have computed the Helmholtz solutions $u_l$ for $l = 0, 1, \ldots, \widetilde{N} - 1$, the discrete solution to the wave equation $\Phi_n^{\lambda}$ can then be approximated via a trapezoidal rule for the inverse Z-transform. The interior solution is also calculated by applying the same time and spatial discretisation to (20.4). In the next section we outline the splitting of these Helmholtz problems into low and high-frequency cases according to the index $l = 0, 1, \ldots, \widetilde{N} - 1$ and briefly describe the methods employed to approximate their solution in each case.

## 20.3 Hybrid Methods Framework

For the spatial discretisation we now either apply a piecewise constant collocation BEM to (20.6) or for a high-frequency region (to be specified in terms of $\mathrm{Re}(k_l)$), we employ a high-frequency approximation (HFA). In order to define this procedure we heuristically specify a threshold $k^*$ for which we employ the BEM when $|\mathrm{Re}(k_l)| \leq k^*$ and let $\eta \leq \widetilde{N}/2$ be the minimal integer valued index of the minimal $|\mathrm{Re}(k_l)| > k^*$, which is the region for which we apply the high-frequency approximation—see Fig. 20.1. We note that the indexing $l = 0, 1, \ldots, \widetilde{N}-1$ starts from $\mathrm{Re}(k_0) = 0$ at the bottom of the loop and runs clockwise. We specify the location of the wavenumber $k_\eta$ to be within the lower left quarter of the loop of possible $k_l$ values. Note that we are only required to solve $\widetilde{N}/2 + 1$ Helmholtz problems since the wavenumbers $k_l$, $l = 0, 1, \ldots, \widetilde{N} - 1$ occur in symmetric pairs and consequently the solutions arise in complex conjugate pairs.

The first HFA will be provided by a plane-wave approximation in which the amplitudes are determined using the DEA method detailed in [ChEtAl21] as described in [RoCh21]. The phases are constructed by performing a matching of the high-frequency approximation with the BEM results at the highest frequencies

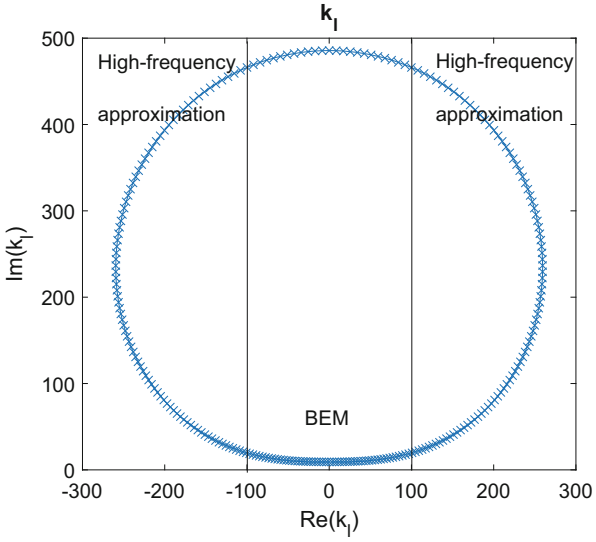



**Fig. 20.1** Example of a hybrid method implementation with the threshold $k^* = 100$ chosen to be the wavenumber where the method switches from the BEM to a high-frequency approximation

for which the BEM is applied, as discussed later in Sect. 20.4.1. The second HFA we consider is an incident illumination approximation where only the direct contribution of the source term on the boundary is included and reflected contributions are assumed to play an insignificant role. This approximation, therefore, relies on there being sufficient decay before any reflections occur. For wavenumbers which have a very large imaginary part, we expect that the incident illumination approximation will be a reasonable approach since the magnitude of $\mathrm{Im}(k_l)$ determines the decay rate of the plane waves as they propagate. The DEA numerical approach will be able to go beyond the incident illumination model in terms of the reflection order but will introduce additional sources of error due to the numerical discretisation procedures.

## 20.4 High-Frequency Approximations

In the following subsections, we outline two high-frequency approximations for solving the set of Helmholtz problems (20.6) for $\{l = 0, 1, \ldots, \widetilde{N}/2 : |\mathrm{Re}(k_l)| > k^*\}$.

### 20.4.1 DEA Based HFA

The first HFA is based on the fact that the solution to the Helmholtz equation

$$\Delta u + k^2 u = 0$$

for wavenumbers $k$ with large real part may be well described as a plane-wave superposition solution of the form

$$u(\boldsymbol{x}) = \sum_{\kappa=1}^{R} A_\kappa(\boldsymbol{x}, \omega) e^{\mathrm{i}\omega S_\kappa(\boldsymbol{x})}, \tag{20.7}$$

where $\omega = \mathrm{Re}(ck)$ is the angular frequency. In the method proposed here, the amplitude terms $A_\kappa$ in (20.7) are approximated using the direction preserving DEA method detailed in [ChEtAl21]. In particular, we make use of the following relationship between the stationary phase-space density $\rho$ (the variable approximated by DEA) and the amplitudes $A_\kappa$ and phases $S_\kappa$:

$$\rho(\boldsymbol{x}, \boldsymbol{p}) = \sum_{\kappa=1}^{R} A_\kappa^2(\boldsymbol{x}, \omega) \delta(\boldsymbol{p} - \nabla S_\kappa(\boldsymbol{x})). \tag{20.8}$$

Here $\boldsymbol{p} \in \mathbb{R}^2$ is the momentum vector whereby $|\boldsymbol{p}| = c^{-1}$. For a plane wave directed with angle $\Theta$ relative to the $x_1$ axis, then $c\boldsymbol{p} = (\cos(\Theta), \sin(\Theta))$. Therefore, the phase-space density $\rho$ is equal to the superposition of squares of the amplitudes $A_\kappa$ corresponding to rays travelling in directions defined by $S_\kappa$. The approximation of the phase terms $S_\kappa$ in our plane-wave superposition solution (20.7) will be calculated by setting the solution calculated via the BEM equal to the expression (20.7) at $l = \eta - 1$ and $l = \eta$, in which the amplitude terms have been determined from DEA, and the phase terms are the only unknowns in the expression to be determined.

We now discuss how to determine the phase terms $S_\kappa$ in (20.7), given that the amplitudes $A_\kappa$ for each direction $\kappa$ have been found by choosing the directions in the sum over $\kappa$ in (20.8) to correspond to those of the direction preserving DEA discretisation [ChEtAl21]. A new methodology introduced recently in [RoCh21] is applied to determine the unknown phase terms $S_\kappa$. In particular, we reconstruct the phase terms from a full wave solution calculated via the BEM at the maximal frequency before we switch to the high-frequency approximation and also at the lowest frequency at which we apply the high-frequency approximation. These frequency values are denoted $\omega_{\eta-1} = \mathrm{Re}(ck_{\eta-1})$ and $\omega_\eta = \mathrm{Re}(ck_\eta)$, respectively. The choice of $\omega_{\eta-1}$ and $\omega_\eta$ has been investigated numerically in [RoCh21]. We apply both the BEM and DEA to obtain a set of equations of the form

$$u_l(\boldsymbol{x}) = \sum_{\kappa=1}^{R} A_\kappa(\boldsymbol{x}, \omega_l) e^{\mathrm{i}\omega_l(\sin(\theta_\kappa)s/c + \gamma_\kappa^l)}, \tag{20.9}$$

for $l = \eta - 1$ and also for $l = \eta$. The left hand side of (20.9) is provided by the solutions calculated from the BEM and $s$ is the boundary arclength value corresponding to the Cartesian coordinates $\boldsymbol{x} \in \Gamma$; see Fig. 20.2. Note that the representation of the phase terms in (20.9) as linear functions

$$S_\kappa(\boldsymbol{x}) = \sin(\theta_\kappa)s/c + \gamma_\kappa^l$$

stems from the fact that the wave speed $c$ is assumed to be constant. In addition, $\theta_\kappa \in (-\pi/2, \pi/2)$ represents the direction relative to $-\hat{\mathbf{n}}_{\boldsymbol{x}}$ of a plane wave directed into $\Omega$ from $\boldsymbol{x}$ and $\gamma_\kappa^l$, $\kappa = 1, 2, \ldots, R$, $l = \eta, \eta + 1$ are a set of unknown constants to be determined by imposing (20.9) at a set of points $\boldsymbol{x} \in \Gamma$.

The phase reconstruction procedure must be performed at more than one frequency owing to the periodicity of the plane waves, and hence the non-uniqueness of the phase solution at a single frequency. The phase terms at $\omega_{\eta-1}$ and $\omega_\eta$ may then be related via

$$\gamma_\kappa^{\eta-1} + \frac{\sin(\theta_\kappa)s}{c} + \frac{2\pi\nu}{\omega_{\eta-1}} = \gamma_\kappa^{\eta} + \frac{\sin(\theta_\kappa)s}{c} + \frac{2\pi\nu}{\omega_\eta}. \tag{20.10}$$

Solving (20.10) for $\nu \in \mathbb{Z}$ allows us to recover a unique set of phase constants $\gamma_\kappa$ via

$$\gamma_\kappa = \gamma_\kappa^l + \frac{2\pi\nu}{\omega_l},$$

for either $l = \eta - 1$ or $l = \eta$. Once $\gamma_\kappa$ are known we calculate the solutions to the Helmholtz problems using (20.9) for all frequencies with absolute value larger than $|\omega_{\eta-1}|$.

The calculation of $\gamma_\kappa^{\eta-1}$ (and $\gamma_\kappa^{\eta}$) will be dependent on the numerical example we are considering. In our numerical examples we only consider polygonal domains and the values for $\gamma_\kappa^{\eta}$ will need to be calculated for each edge separately since the $\gamma_\kappa^{\eta}$ values will relate to different directions of propagation from each edge. We generate a system of equations of the form (20.9) by choosing $\boldsymbol{x} = \boldsymbol{x}_i$ for $i = 1, 2, \ldots, M$ as the collocation points from the BEM approximation (located in the centre of each of the $M$ boundary elements) in (20.9). The next task is to determine how many of the amplitudes $A_\kappa$ are non-zero at every collocation point $\boldsymbol{x}_i$ on a given edge, since this provides a reduction in the number of phase constants $\gamma_\kappa^l$ that we need to recover. The system of equations (20.9) can then be solved as a linear system

$$u_l(\boldsymbol{x}_i) = \sum_{\kappa=1}^{R} A_\kappa(\boldsymbol{x}_i, \omega_l) e^{i\omega_l(\sin(\theta_l)s_i/c)} v_\kappa^l,$$

for $i = 1, 2, \ldots, M$, $l = \eta - 1$ or $l = \eta$ and where $s_i$ is the arclength parameter for the point $\boldsymbol{x}_i$. The unknowns $v_\kappa^l = e^{i\omega_l\gamma_\kappa^l}$ may be determined using the Moore-

Penrose pseudo-inverse to obtain the least squares solution. Once each $v_\kappa^l$ term has been found, one can directly calculate the phase constants via $\gamma_\kappa^l = -i \log(v_\kappa^l)/\omega_l$.

### *20.4.2 Simple HFA*

In this section, we describe a simple high-frequency approximation (SHFA) based on the observation that the wavenumbers $k_l$ in the high-frequency range typically have large imaginary part. Since the DEA calculation includes a dissipative factor with exponential decay rate $2\,\mathrm{Im}(k_l)$ along each ray trajectory, then the only significant contributions to the DEA solution will come from very short ray trajectories. In this case, the solution for a wavenumber $k_l$ with a large enough imaginary part can be reasonably well approximated by simply rescaling the Z-transformed boundary data $\widetilde{F}_l$. In particular, we set

$$u_l(\boldsymbol{x}) = \frac{\widetilde{F}_l(\boldsymbol{x})}{ik_l \cos\left(\theta_0(\boldsymbol{x})\right)},$$

where $\theta_0(\boldsymbol{x})$ defines the direction of the source term at $\boldsymbol{x} \in \Gamma$ relative to the normal direction. For a boundary value problem with boundary data related to a plane wave entering the domain from one or more edges, then the angle $\theta_0$ can be found directly from the plane wave direction. In the next section, we will present numerical results for the interior solution produced using the hybrid methods described above in a variety of different examples.

## 20.5 Numerical Results

In this section we consider numerically solving the wave equation (20.1) with Neumann boundary conditions (20.3) via the two hybrid methods introduced in this chapter. We consider an inhomogeneous Neumann IBVP for the cases when $\Omega$ is a unit square or an L-shaped domain and the boundary data corresponds to a plane wave travelling into the domain, as depicted in Fig. 20.2. We define our Neumann boundary condition (20.3) to be

$$F(\boldsymbol{x}, t) = \begin{cases} W(x_2 \sin(\Theta_0) - ct) & \text{if } x_1 = 0, \\ W(x_1 \cos(\Theta_0) - ct) & \text{if } x_2 = 0 \text{ and } \Theta_0 > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{20.11}$$

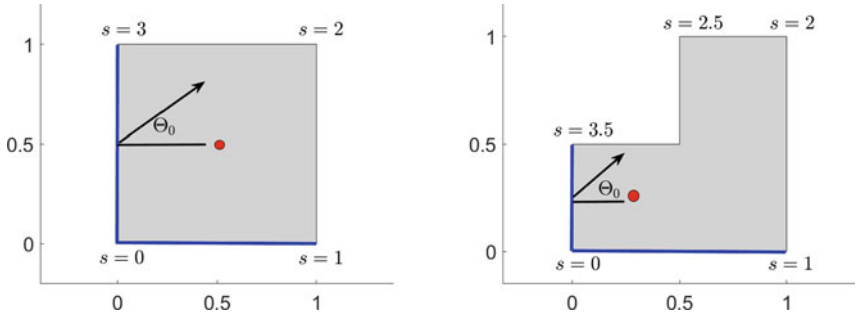



**Fig. 20.2** The domains considered in the numerical experiments for solving the homogeneous wave equation (20.1) showing the value of the boundary arclength $s$ at each vertex, the propagation direction $\Theta_0$ for the plane wave boundary data and the interior evaluation point as a red dot. The bold boundary lines indicate the positions where the plane wave may enter the domain and, therefore, provide inhomogeneous boundary data. (**a**) Unit square domain. (**b**) L-shaped domain

where $\Theta_0 \in [0, \pi/2)$ is a direction relative to the positive $x_1$ axis. Here $\boldsymbol{x} = (x_1, x_2)$ and the angle $\Theta_0$ is, in general, distinct from directions used in the DEA discretisation. However, for accuracy reasons we choose $\Theta_0$ to correspond to one of the DEA discretisation directions, which we note can be specified in a problem specific manner, and, therefore, this choice does not indicate a limitation of the method. We consider the case when the function $W$ takes the form of the normal derivative of a Gaussian pulse written as

$$W(x) = -\alpha(x + ct_0)(n_1 \cos(\Theta_0) + n_2 \sin(\Theta_0))e^{-\alpha(x+ct_0)^2},$$

for $x \in \mathbb{R}$ and where $\hat{\mathbf{n}} = (n_1, n_2)$ are the entries of the unit normal vector $\hat{\mathbf{n}}$. The parameters $t_0 > 0$ and $\alpha > 0$ control the position of the peak of the Gaussian pulse and its bandwidth, respectively. These parameters are chosen carefully to ensure that the initial conditions are approximately satisfied and the pulse has decayed sufficiently at $t = 0$. Throughout this section we choose $\alpha = 4096$, $t_0 = 0.1$ and $c = 1$ in order to obtain a broadband signal. We consider regular geometric domains and directions $\Theta_0$ for the boundary condition, as illustrated in Fig. 20.2, such that we will only need to use 8 global directions in the DEA implementation in order to include all possible propagation directions. Technical details regarding the implementation of the Neumann boundary condition (20.11) in the DEA scheme can be found in [RoCh21].

### 20.5.1 Square Domain

We now present the numerical results for the case when $\Omega$ is a unit square as shown in Fig. 20.2a. The error of the time-dependent interior solution is calculated via

$$\text{Error}(N) = \sqrt{\frac{\sum_{n=0}^{N-1} \left(\Phi(\boldsymbol{x}, t_n) - \Phi_n(\boldsymbol{x})\right)^2}{\sum_{n=0}^{N-1} \left(\Phi(\boldsymbol{x}, t_n)\right)^2}}, \tag{20.12}$$

and the estimated order of convergence (EOC) is given by

$$\text{EOC}(N) = \log_2(\text{Error}(N/2)/\text{Error}(N)).$$

We initially consider the case $\Theta_0 = 0$ where we can compare the numerical solution $\Phi_n$ against the exact solution $\Phi(\boldsymbol{x}, t_n)$ given by

$$\Phi(\boldsymbol{x}, t) = \frac{1}{2} e^{-\alpha(x_1 - c(t-t_0))^2} \tag{20.13}$$

for an infinite domain (in the $x_1$—direction). We consider only early times such that we do not observe any reflections and the solution matches (20.13). For the unit square example we can compare our numerical results against the exact solution given by (20.13) up to time $T = 1$, such that we do not observe any reflections.

Figure 20.3 shows a comparison between the exact and numerical interior solutions at $\boldsymbol{x} = (0.5, 0.5)$. We apply a high-frequency approximation whenever $|\text{Re}(k_l)| > 350$ and employ $M = 1024$ boundary elements to provide a good level of accuracy up to the BEM cut-off wavenumber $k^* = 350$. The plots compare the results of using the SHFA and the DEA based plane wave approximations with the exact solution up to $T = 1$, for $N = \widetilde{N}/2 = 4096$ time-steps. In this case both high-frequency approximations produce identical looking results matching the exact

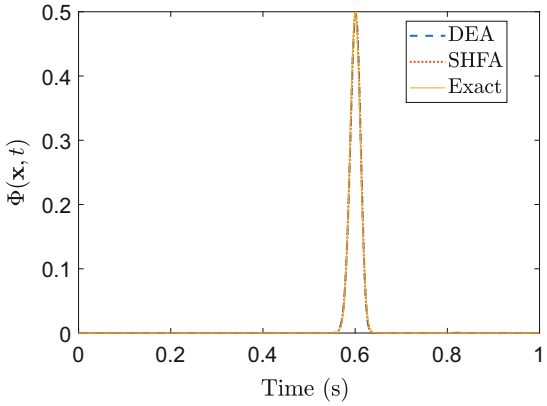



**Fig. 20.3** Interior solution to the wave equation at $\boldsymbol{x} = (0.5, 0.5)$ inside a unit square with boundary data (20.11) and parameters $\Theta_0 = 0, \alpha = 4096, t_0 = 0.1$, with $M = 1024$ boundary elements and $N = \widetilde{N}/2 = 4096$ time-steps. The high-frequency approximations are applied whenever $|\text{Re}(k_l)| > 350$

**Table 20.1** Errors and convergence rates for the interior solution in the unit square domain observed at the point $\boldsymbol{x} = (0.5, 0.5)$ with parameters $\Theta_0 = 0$, $\alpha = 4096$, $t_0 = 0.1$ and $T = 1$. The interior solutions were calculated numerically using the DEA and SHFA based hybrid CQ schemes whereby the high-frequency approximations were applied whenever $|\mathrm{Re}(k_l)| > 350$

| | | | | DEA | | SHFA | |
|---|---|---|---|---|---|---|---|
| $\widetilde{N}$ | $N$ | $M$ | $\eta$ | Error | EOC | Error | EOC |
| 1024 | 512 | 4 | 101 | 0.4518 | – | 0.4518 | – |
| 2048 | 1024 | 16 | 108 | 0.4629 | −0.04 | 0.4629 | −0.04 |
| 4096 | 2048 | 64 | 111 | 0.1372 | 1.75 | 0.1372 | 1.75 |
| 8192 | 4096 | 256 | 112 | 0.0142 | 3.27 | 0.0142 | 3.27 |
| 16384 | 8192 | 1024 | 112 | 0.0026 | 2.45 | 0.0026 | 2.45 |

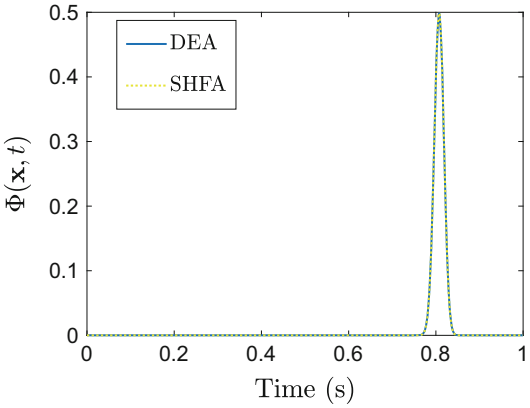



**Fig. 20.4** Interior solution to the wave equation at $\boldsymbol{x} = (0.5, 0.5)$ inside a unit square with boundary data (20.11) and parameters $\Theta_0 = \pi/4$, $\alpha = 4096$, $t_0 = 0.1$, with $M = 1024$ boundary elements and $N = \widetilde{N}/2 = 4096$ time-steps. The high-frequency approximations are applied whenever $|\mathrm{Re}(k_l)| > 350$

solution. In Table 20.1 we investigate the relative errors and convergence rates of the interior solutions observed at the point $\boldsymbol{x} = (0.5, 0.5)$, calculated via both hybrid methods, as we double the number of time-steps $N = \widetilde{N}/2$ and increase the number of boundary elements $M$ by a factor of four. The interior solutions were calculated for the parameters $\Theta_0 = 0$, $\alpha = 4096$ and $t_0 = 0.1$, with the high-frequency approximations again being implemented whenever $|\mathrm{Re}(k_l)| > 350$. The relative errors were computed via (20.12) using the exact solution (20.13) up until $T = 1$. From the table we observe that both hybrid methods produce identical results and convergence rates. On the last row of the table we achieve errors of less than 1% as there are enough boundary elements to model the highly oscillatory behaviour, and we also observe a convergence rate close to the expected second order.

We now investigate the same IBVP as discussed previously, but now we consider the case when the plane wave boundary data enters the domain at an angle of $\Theta_0 = \pi/4$. Figure 20.4 compares the interior solutions at the point $\boldsymbol{x} = (0.5, 0.5)$ computed using the DEA and SHFA high-frequency approximations. In this figure we observe that both solutions behave identically. Table 20.2 investigates the relative errors and convergence rates of the interior solutions, observed at the point $\boldsymbol{x} = (0.5, 0.5)$, calculated via both hybrid methods with parameters $\Theta_0 = \pi/4$, $\alpha =$

**Table 20.2** Errors and convergence rates for the interior solution on the unit square domain observed at the point $\boldsymbol{x} = (0.5, 0.5)$ with parameters $\Theta_0 = \pi/4$, $\alpha = 4096$, $t_0 = 0.1$ and $T = 1$. The interior solutions were calculated numerically using the DEA and SHFA based hybrid CQ schemes whereby the high-frequency approximations were applied whenever $|\mathrm{Re}(k_l)| > 350$

| | | | | DEA | | SHFA | |
|---|---|---|---|---|---|---|---|
| $\widetilde{N}$ | $N$ | $M$ | $\eta$ | Error | EOC | Error | EOC |
| 256 | 128 | 1024 | 80 | – | – | – | – |
| 512 | 256 | 1024 | 92 | 0.5919 | – | 0.5919 | – |
| 1024 | 512 | 1024 | 101 | 0.3769 | 0.65 | 0.3769 | 0.65 |
| 2048 | 1024 | 1024 | 108 | 0.1455 | 1.37 | 0.1455 | 1.37 |
| 4096 | 2048 | 1024 | 111 | 0.0397 | 1.87 | 0.0397 | 1.87 |
| 8192 | 4096 | 1024 | 112 | 0.0100 | 2.80 | 0.0100 | 2.80 |
| 16384 | 8192 | 1024 | 112 | 0.0025 | 2.00 | 0.0025 | 2.00 |

4096 and $t_0 = 0.1$, with the high-frequency approximations implemented whenever $|\mathrm{Re}(k_l)| > 350$. The errors were computed via (20.12) but using subsequent interior solutions as we double the number of time-steps. We investigate the error when doubling the number of time-steps for a fixed number of boundary elements $M = 1024$ and observe the expected second order convergence rate for BDF2 based CQ schemes with errors smaller than 1% for both hybrid methods. The errors for both methods are identical when comparing against subsequent interior solutions.

### *20.5.2 L-Shaped Domain*

We now present the numerical results for solving the same IBVP as above for the case when $\Omega$ is an L-shaped domain as shown in Fig. 20.2b. The DEA approximation process needs to be modified for non-convex domains such as the L-shape and we implement the DEA approximation on a sub-divided version of the domain where each of the (two) sub-domains is convex. In this case the sub-division was implemented by introducing an (artificial) internal interface connecting the vertices at $s = 1$ and $s = 3$ to form two convex quadrilateral sub-domains. The extension of the DEA approximation to multi-domains is discussed in more detail in [ChEtAl21]. We also note that we must omit any amplitudes associated with the internal interface from the DEA result and then reorder the degrees of freedom to be consistent with the low frequency BEM calculations before integrating into the CQ algorithm. For this example we can compare our numerical results against the exact solution given by (20.13) up to the time $t = 0.5$ so that we do not observe any reflections at the solution point $\boldsymbol{x} = (0.25, 0.25)$.

Figure 20.5 shows a comparison between the exact and numerical interior solutions at $\boldsymbol{x} = (0.25, 0.25)$. We again apply a high-frequency approximation whenever $|\mathrm{Re}(k_l)| > 350$ and use $N = \widetilde{N}/2 = 4096$ time-steps and $M = 1024$ boundary elements. We observe that both high-frequency approximations produce

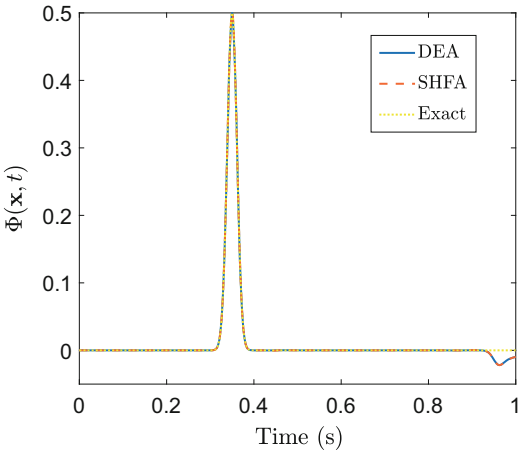



**Fig. 20.5** Interior solution to the wave equation at $\boldsymbol{x} = (0.25, 0.25)$ inside an L-shaped domain with boundary data (20.11) and parameters $\Theta_0 = 0$, $\alpha = 4096$, $t_0 = 0.1$, with $M = 1024$ boundary elements and $N = \widetilde{N}/2 = 4096$ time-steps. The high-frequency approximations are applied whenever $|\mathrm{Re}(k_l)| > 350$

**Table 20.3** Errors and convergence rates for the interior solution on the L-shaped domain observed at the point $\boldsymbol{x} = (0.25, 0.25)$ with parameters $\Theta_0 = 0$, $\alpha = 4096$, $t_0 = 0.1$ and $T = 0.5$. The interior solutions were calculated numerically using the DEA and SHFA based hybrid CQ schemes whereby the high-frequency approximations were applied whenever $|\mathrm{Re}(k_l)| > 350$

| | | | | DEA | | SHFA | |
|---|---|---|---|---|---|---|---|
| $\widetilde{N}$ | $N$ | $M$ | $\eta$ | Error | EOC | Error | EOC |
| 1024 | 512 | 8 | 101 | 0.2898 | – | 0.2898 | – |
| 2048 | 1024 | 32 | 108 | 0.1836 | 0.66 | 0.1836 | 0.66 |
| 4096 | 2048 | 128 | 111 | 0.0433 | 2.08 | 0.0433 | 2.08 |
| 8192 | 4096 | 512 | 112 | 0.0037 | 3.55 | 0.0037 | 3.55 |
| 16384 | 8192 | 2048 | 112 | 5.0609e−4 | 2.87 | 5.0609e−4 | 2.87 |

identical looking results up to $t = 0.9$. For $t > 0.9$ we observe that the numerical solutions deviate from the exact solution because the numerical solutions include contributions due to diffraction from the re-entrant corner at $s = 3$ and, therefore, the exact solution is not valid. In Table 20.3, we investigate the relative errors and convergence rates of the interior solutions observed at the point $\boldsymbol{x} = (0.25, 0.25)$, calculated via both hybrid methods, as we double the number of time-steps $N = \widetilde{N}/2$ and increase the number of boundary elements $M$ by a factor of four. The interior solutions were calculated for the parameters $\Theta_0 = 0$, $\alpha = 4096$ and $t_0 = 0.1$, with the high-frequency approximations being implemented whenever $|\mathrm{Re}(k_l)| > 350$. The relative errors were computed via (20.12) against the exact solution (20.13) up until $t = 0.5$, when the exact solution is valid. From the table we observe that both methods give the same error and convergence results, obtaining less than 1% error with $N \geq 4096$ time-steps and $M = 512$ boundary elements. We also approximately achieve the expected second order convergence for BDF2 based CQ schemes.

We now investigate the same IBVP as discussed previously, but consider the case when the plane wave boundary data enters the domain at an angle of $\Theta_0 = \pi/4$.

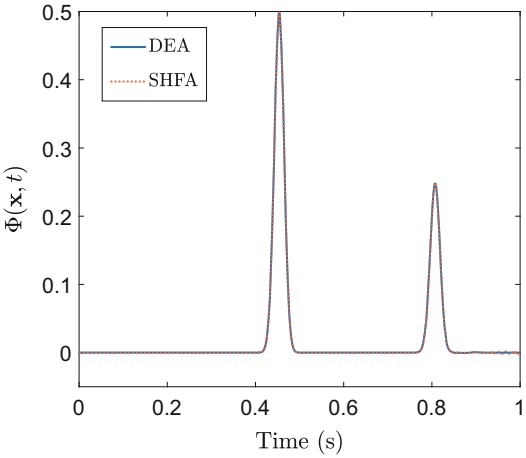



**Fig. 20.6** Interior solution to the wave equation at $\boldsymbol{x} = (0.25, 0.25)$ inside an L-shaped domain with boundary data (20.11) and parameters $\Theta_0 = \pi/4$, $\alpha = 4096$, $t_0 = 0.1$, with $M = 1024$ boundary elements and $N = \widetilde{N}/2 = 4096$ time-steps. The high-frequency approximations are applied whenever $|\mathrm{Re}(k_l)| > 350$

**Table 20.4** Errors and convergence rates for the interior solution on the L-shaped domain observed at the point $\boldsymbol{x} = (0.25, 0.25)$ with parameters $\Theta_0 = \pi/4$, $\alpha = 4096$, $t_0 = 0.1$ and $T = 1$. The interior solutions were calculated numerically for fixed $M = 1024$ using the DEA and SHFA based hybrid CQ schemes whereby the high-frequency approximations were applied whenever $|\mathrm{Re}(k_l)| > 350$

| | | | | DEA | | SHFA | |
|---|---|---|---|---|---|---|---|
| $\widetilde{N}$ | $N$ | $M$ | $\eta$ | Error | EOC | Error | EOC |
| 256 | 128 | 1024 | 80 | – | – | – | – |
| 512 | 256 | 1024 | 92 | 0.4338 | – | 0.4338 | – |
| 1024 | 512 | 1024 | 101 | 0.2346 | 0.89 | 0.2346 | 0.89 |
| 2048 | 1024 | 1024 | 108 | 0.0789 | 1.57 | 0.0789 | 1.57 |
| 4096 | 2048 | 1024 | 111 | 0.0209 | 1.92 | 0.0209 | 1.92 |
| 8192 | 4096 | 1024 | 112 | 0.0053 | 1.98 | 0.0053 | 1.98 |
| 16384 | 8192 | 1024 | 112 | 0.0013 | 2.03 | 0.0013 | 2.03 |

Figure 20.6 shows that the interior solution at $\boldsymbol{x} = (0.25, 0.25)$ is visually identical for each of the hybrid methods. Table 20.4 investigates the relative errors and convergence rates of the interior solutions, observed at the point $\boldsymbol{x} = (0.25, 0.25)$, calculated via the DEA and SHFA hybrid methods with parameters $\Theta_0 = \pi/4$, $\alpha = 4096$, $t_0 = 0.1$ and $T = 1$, with the high-frequency approximations being implemented whenever $|\mathrm{Re}(k_l)| > 350$. The errors were computed via (20.12) but using subsequent interior solutions as we double the number of time-steps. Table 20.4 investigates the error when doubling the number of time-steps for a fixed number of boundary elements $M = 1024$. In this table, we observe the expected second order convergence rate and errors smaller than 1% for both methods. Again the errors for both methods are identical when comparing against subsequent interior solutions.

## 20.6 Conclusion

We have described two hybrid CQ based discretisations of the wave equation for interior acoustic Neumann problems with broadband boundary data or source terms. We performed a series of numerical experiments to demonstrate the effectiveness of both hybrid approaches for the case of plane wave boundary data. The hybrid methods were able to provide faster computations than using CQ with BEM alone, while retaining the expected second order convergence behaviour for BDF2-based CQ schemes.

## References

[BaCh20] Bajars, J., Chappell, D.J.: Modelling uncertainties in phase-space boundary integral models of ray propagation. Commun. Nonlinear Sci. Numer. Simul. **80**, 104973 (2020)

[BaEtAl17] Bajars, J., Chappell, D.J., Søndergaard, N., Tanner, G.: Transport of phase space densities through tetrahedral meshes using discrete flow mapping. J. Comput. Phys. **328**, 95–108 (2017)

[BaSa08] Banjai, L., Sauter, S.: Rapid solution of the wave equation on unbounded domains. SIAM J. Numer. Anal. **47**, 229–247 (2008)

[BeEtAl17] Betcke, T., Salles, N., Smigaj, W.: Overresolving in the Laplace domain for convolution quadrature methods. SIAM J. Sci. Comput. **39**(1), A188–A213 (2017)

[Ch11] Chappell, D.: Convolution quadrature Galerkin boundary element method for the wave equation with reduced quadrature weight computation. IMA J. Numer. Anal. **31**(2), 640–666 (2011)

[ChEtAl08] Chappell, D.J., Harris, P.J., Henwood, D., Chakrabarti, R.: Modelling the transient interaction of a thin elastic shell with an exterior acoustic field. Int. J. Numer. Methods Eng. **75**(3), 275–290 (2008)

[ChEtAl21] Chappell, D.J., Crofts, J.J., Richter, M., Tanner, G.: A direction preserving discretization for computing phase-space densities. SIAM J. Sci. Comput. **44**(4), B884–B906 (2021)

[ChTa14] Chappell, D.J., Tanner, G.: A boundary integral formalism for stochastic ray tracing in billiards. Chaos **24**(4), 043137 (2014)

[HaEtAl19] Hartmann, T., Morita, S., Tanner, G., Chappell, D.J.: High-frequency structure-and air-borne sound transmission for a tractor model using dynamical energy analysis. Wave Motion **469**, 20130153 (2019)

[Lu88] Lubich, C.: Convolution quadrature and discretized operational calculus I. Numer. Math. **52**, 129–145 (1988)

[Lu94] Lubich, C.: On the multistep time discretization of linear initial-boundary value problems and their boundary integral equations. Numer. Math. **57**, 365–389 (1994)

[MaEtAl20] Mavaleix-Marchessoux, D., Bonnet, M., Chaillat, S., Leblé , B.: A fast boundary element method using the Z-transform and high-frequency approximations for large-scale three-dimensional transient wave problems. Int. J. Numer. Methods Eng. **121**(21), 4734–4767 (2020)

[RoCh21] Rowbottom, J., Chappell, D.J.: On hybrid convolution quadrature approaches for modelling time-domain wave problems with broadband frequency content. Int. J. Numer. Methods Eng. (in press)

[Ta09] Tanner, G.: Dynamical energy analysis - determining wave energy distributions in vibro-acoustical structures in the high-frequency regime. J. Sound Vib. **320**(4–5), 1023–1038 (2009)

# Chapter 21
# Analytical Reconstruction of the Nonlinear Transfer Function for a Wiener–Hammerstein Model


**J. Schmith, A. Schuck Jr., B. E. J. Bodmann, and P. J. Harris**

## 21.1 Introduction

Professional audio equipment that is traditionally being used by musicians either in the recording studio or in live performances is usually analog, where we specifically focus on tube amplifiers and effect pedals. Amplifiers of the various blends all have their characteristic sounds, which originate from details in the electronic architecture of the device as well as specific electro-acoustic features. Thus, the loudspeakers installed in the amplifier cabinets are one source of those sound characteristics due to the nonlinear conversion of an electronic signal into an acoustic response as reported in references [ScOl21, OlSc13]. Another source of nonlinear behavior is the tubes, which because of different kinds of fabricated models for a specific operation makes it hard to understand details of the reasons for their differences in sound signal reproduction.

Even today, tube amplifiers are the "holy grail" for guitarists, and the tubes are the nonlinear components that are present in the pre- or power stages of amplifiers, which are distinctively identified in distortion situations. Nevertheless, recent investments in the developments of digital amplifier simulators define a new paradigm for signal amplification and modulation for practical usage. It is noteworthy that although there is a tremendous progress in the technical evolution

J. Schmith (✉)
University of the Vale do Rio dos Sinos, São Leopoldo, RS, Brazil
e-mail: jschmith@unisinos.br

A. Schuck Jr. · B. E. J. Bodmann
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: bardo.bodmann@ufrgs.br

P. J. Harris
University of Brighton, Brighton, UK
e-mail: p.j.harris@brighton.ac.uk



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_21

of digital audio devices for simulating tube amplifiers by algorithms, it is safe to say that the simulation fidelity of tube-driven devices is still a challenge [MoEtAl15, EiZo16, EiEtAl17, EiZo18, EiZo18a]. An incontestable example astonishingly close to the original amplifier that is being simulated is the Kemper Amplifier Profiler™ [DuEtAl20, Ke15], which makes use of a Volterra model type of approach[Og07, Zo11] for its digital implementation of signal amplification reconstruction and reproduction. Nevertheless, there is still need for improvements on the one side for high-fidelity amplifier response reproduction and on the other side for optimization in the computational approach in order to allow for more complex sound creations with a combination of amplifiers and effects.

To this end, we propose a different approach to the problem where input and output signals are superimposed and produce a Lissajous curve, which contains besides the nonlinearity of the Wiener–Hammerstein model [SjSc12, RoEtAl14, ScEtAl14] also the influence of the linear equalization before and after the nonlinear block, respectively. A nice feature of this type of approach is that the analytical method used to identify the composition of the nonlinear response by the creation of harmonics and their respective phase may be cast into a linear algorithm. Evidently, such a methodology does not depend on advanced computational resources to work in real time, nevertheless with today's multi-thread processors opens a pathway for fast signal processing especially for simulations of complex configurations with amplifiers and effect pedals.

## 21.2 Preliminaries

A commonly used amplification model is the Wiener–Hammerstein model presented in Fig. 21.1. This block model is composed of a linear, a nonlinear followed by another linear component, where the input signal appears with a phase shift after the first filter followed by the generation of various harmonics as a consequence of the nonlinearity and last a frequency-dependent phase shift from the second filter that is the output signal. The latter shall simulate the characteristics of the nonlinear device in consideration. While filters with their influence on amplitude and phase are well understood, in approaches reported in the literature, the principal difficulty

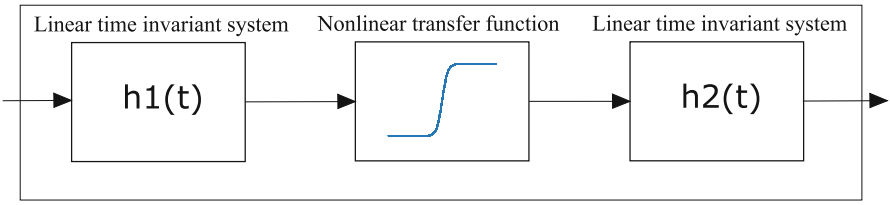



**Fig. 21.1** The Wiener–Hammerstein model diagram

lies in the task to identify the nonlinearity, which is sandwiched between the two linear blocks, i.e., the filters.

In the further, we present the idea of a new approach, which in principle should be able to overcome this difficulty, however so far considering only clean signals, in other words ignoring the presence of the always present noise in real signals. Note that this aspect becomes increasingly relevant for the higher part of the frequency spectrum. Additionally, we idealize the test case by considering an input signal with a single frequency only because input signals with multiple frequencies may be handled in close analogy to the presented case.

In order to get the response function of the system, the input–output analysis shall be performed for a sequence of input signals with different frequencies. For an input signal with frequency $\omega = \frac{2\pi}{T}$ ($T$ is the usual period), the frequency spectrum $\{k\omega\}_{k=1}^{K} = \{\omega, 2\omega, \ldots, K\omega\}$ generated by the nonlinearity is assumed to be discrete and compact, i.e., with a finite number ($K$) of generated harmonics. This assumption is reasonable since the filters are typically low-pass or band-pass filters and thus suppress the higher frequencies. Moreover, the data acquisition limits the highest detectable frequency by the sampling frequency $f_s$. Further, each frequency has an associated phase shift $\{\phi_k\}_{k=1}^{K} = \{\phi_1, \phi_2, \ldots, \phi_K\}$. The data for the analysis are then drawn from a digital signal acquired by sampling the analogue input and output signals with a digital oscilloscope.

For simplicity, but without restricting the method, the input signal shall have a zero phase $I = I(\omega|t)$, whereas the output signal is composed of a spectrum with the base frequency and its upper harmonics.

$$\Omega = \Omega(\{k\omega\}_{k=1}^{K}, \{\phi_k\}_{k=1}^{K}|t) = \Omega(\{k\omega\}_k, \{\phi_1, \phi_2, \ldots, \phi_K\}|t).$$

From the obtained experimental data to be analyzed, the initial transient is excluded, and only data from the total sample are used for the analysis that form a closed loop in the Lissajous figure $Z = I + \imath\Omega$. Thus one gets $N = f_s T$ as the number of samples (points that constitute the Lissajous figure) for a simply closed Lissajous curve, i.e., for exactly one period of the input signal.

The representation for the input and output signals is

$$I(\omega|t) = A_0 e^{\imath\omega t}\ ,\ \Omega(\{k\omega\}_k, \{\phi_k\}_k|t) = \sum_{k=1}^{K} A_k e^{\imath(k\omega t+\phi_k)}\ , \tag{21.1}$$

with the amplitudes ($A_k$) and phases ($\phi_k$) of the output signal to be determined from the experimental data. Here, $A_0$ is the amplitude and $\omega$ is the frequency of the input signal.

## 21.3 Data Analysis Using the Lissajous Curve

As already mentioned in the introduction, one of the challenges is to identify the nonlinearity, which is the principal characteristics of the response of an amplifier to an input signal. The use of real signals and efforts to disentangle the linear contributions from the nonlinear ones so far resulted in iterative algorithms only, where convergence is a crucial issue, so that we resorted to a method that allows to directly extract the nonlinearity from the experimental data. It is noteworthy that the usage of the input and output signals only characterizes the present analysis as a black-box approach, due to the fact that no detail of the electronic circuit was used to obtain a parameterized form of the amplification of the input signal, and in this sense, it is not limited to Wiener–Hammerstein block models but may in principle be applied to more complex signal processing architectures.

The following steps resemble the idea of the identification of the nonlinearity. First, the experimental data of the input and output signals compose the Lissajous figure, which may be analyzed in order to get the amplitudes and phase shifts of the output signal by a data fit procedure. By inspection of the Wiener–Hammerstein architecture, one observes that upon cancelling the phases, one obtains a representation of the nonlinearity since only the linear blocks affect the phase shifts. It is worth mentioning that the use of Lissajous curve is novel in the literature of amplification profiling, and reconstruction of the nonlinear curve separated from the linear block contributions was so far not obtained in analytical form.

An infinitesimal line element on the Lissajous curve is then given by

$$dZ = \left(\frac{dI}{dt} + \imath\frac{d\Omega}{dt}\right)\,dt \;= \left(\imath\omega A_0 e^{\imath\omega t} - \sum_k A_k k\omega e^{\imath\phi_k} e^{\imath k\omega t}\right)\,dt\;.$$

In order to separate specific output modes with its associated phase, we use an integral transform of $\frac{dZ}{dt}$.

$$\begin{aligned}\int_0^T \frac{dZ}{dt} e^{-\imath m\omega t}\,dt &= \imath\omega A_0 \int_0^T e^{\imath(1-m)\omega t}\,dt - \sum_k k\omega A_k e^{\imath\phi_k}\int_0^T e^{\imath(k-m)\omega t}\,dt \\ &= 2\pi\imath A_0\delta_{m1} - 2\pi m A_m e^{\imath\phi_m}. \end{aligned} \tag{21.2}$$

## 21.4 Amplitudes and Phases

From Eq. (21.2), one derives the equation that is then used to determine the amplitudes and phases,

$$A_m = \imath\left(A_0\delta_{1m} - \int_0^T Z e^{-\imath m\omega t}\,dt\right)e^{-\imath\phi_m}\;, \tag{21.3}$$

which allows to express the amplitudes and associated phases in terms of the Lissajous data and the known amplitude of the input signal.

$$\begin{aligned} A_m &= \left(\left(A_0\delta_{1m} - \int_0^T Z^\dagger e^{\imath m\omega t}\, dt\right)\left(A_0\delta_{1m} - \int_0^T Z e^{-\imath m\omega t}\, dt\right)\right)^{\frac{1}{2}} \\ &= \left(\left(A_0^2 - 2A_0\int_0^T I(t)cos(\omega t) + \Omega(t)\sin(\omega t)\, dt\right)\delta_{1m}\right. \\ &\quad \left. + \int_0^T\int_0^T Z^\dagger(t)Z(\tau)e^{\imath m\omega(t-\tau)}\, dt\, d\tau\right)^{\frac{1}{2}}. \end{aligned}$$

The integral with the $N$ Lissajous sampling points may be approximated making use of the sampling specification and the Riemann sum.

$$\frac{1}{T}\int_0^T Z e^{-\imath m\omega t}\, dt \approx \frac{1}{N}\sum_{n=0}^{N-1} Z(nf_s^{-1})e^{\imath 2\pi\frac{mn}{N}}, \tag{21.4}$$

where the time interval $\Delta t = f_s^{-1}$ is identified with the inverse sampling frequency and $N = Tf_s$. In the further and for convenience, we introduce the abbreviation $Z_n = Z(nf_s^{-1})$ for the Lissajous point at instant $t = nf_s^{-1}$.

$$\begin{aligned} A_m \approx \left(\left(A_0^2 - \frac{2}{N}A_0\sum_{n=0}^{N-1} I_n cos(2\pi n/N) + \Omega_n\sin(2\pi n/N)\right)\delta_{1m}\right. \\ \left. + \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{p=0}^{N-1} Z_n^\dagger Z_p e^{\imath 2\pi\frac{m(n-p)}{N}}\right)^{\frac{1}{2}}. \end{aligned} \tag{21.5}$$

From Eqs. (21.3) and (21.4), the phases are computed by

$$\phi_m^{(inv)} = \arctan\left(\frac{A_0\delta_{m1} + \mathrm{Im}\left\{\frac{\imath}{N}\sum_n Z_n e^{\imath 2\pi\frac{mn}{N}}\right\}}{\mathrm{Re}\left\{\frac{\imath}{N}\sum_n Z_n e^{\imath 2\pi\frac{mn}{N}}\right\}}\right). \tag{21.6}$$

Due to the fact that the arctan function in the function library returns only angles in the range $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, one has to identify the correct quadrant of the angle $\phi_m$. For completeness, we indicate all cases together with their correct phase value in Table 21.1 for the calculated value $\phi^{(inv)}$. From the last line in the table, one may understand that for very small amplitudes the calculation of the phases may fail because of instabilities caused by floating-point arithmetic quantization.

**Table 21.1** Correction to the phase shift depending on the nominator and denominator in the arctan function in Eq. (21.6)

| Nominator | Denominator | Corrected $\phi$ | | Nominator | Denominator | Corrected $\phi$ |
|---|---|---|---|---|---|---|
| $\gtreqless 0$ | $> 0$ | $\phi^{(inv)}$ | | $> 0$ | $= 0$ | $\phi^{(inv)} + \frac{\pi}{2}$ |
| $\geq 0$ | $< 0$ | $\phi^{(inv)} + \pi$ | | $< 0$ | $= 0$ | $\phi^{(inv)} - \frac{\pi}{2}$ |
| $< 0$ | $< 0$ | $\phi^{(inv)} - \pi$ | | $= 0$ | $= 0$ | Undefined |

## 21.5 Numerical Results

Our numerical examples for the Wiener–Hammerstein model presented in Fig. 21.1 were computed with equal band-pass filters $h_1$ and $h_2$. We used for each of these linear time invariant systems a high-frequency cut at 8000 Hz and a low-frequency cut at 50 Hz. Experiments were implemented with three different nonlinear transfer functions: a synthetic input–output signal, the hyperbolic tangent, and the hyperbolic sine function, respectively. The synthetic signals for the input are given by $I = sin(\omega t)$ and for the output in Eq. (21.7). In agreement with experimental findings for real tube amplifiers, the output signal contains besides the fundamental frequency also the second, the fourth, and the sixth upper harmonics. For data acquisition, a sampling frequency was set to $f_s = 200$ kHz, and the input frequency was $\omega = 2\pi$ kHz. The synthetic signal although non-physical was used to validate the profiling method by the usage of the analytical expressions (21.5) and (21.6).

$$\Omega = \sin(\omega t - \frac{\pi}{2}) + \frac{1}{2}\sin(3\omega t - \frac{\pi}{2}) + \frac{1}{4}\sin(5\omega t - \frac{\pi}{4}) + \frac{1}{8}\sin(7\omega t - \frac{\pi}{6}). \quad (21.7)$$

The input and output signals of the three transfer functions are presented in Fig. 21.2. For the proposed method, the signals shall be periodic so that the resulting Lissajous figure describes a closed curve apart from the initial transient. The curves from the input versus output signals of the synthetic signal and the two aforementioned nonlinear transfer functions are presented in Fig. 21.3, (a) Synthetic signal, (b) Hyperbolic tangent response, and (c) Hyperbolic sine response.

Upon inspecting the Lissajous curves, the nonlinear transfer functions are anything but obvious due to the presence of the phases. Thus, we removed the filters $h_1$ and $h_2$ to reduce the curve to the nonlinear transfer functions and show consistency of the proposed method. In the case of the synthetic signal, no specific transfer function was assumed so that we set all the phases of the Eq. (21.7) to zero in order to obtain the transfer function that generates the relation of the synthetic input and output signal. The Lissajous curves without phases and the filters influences are presented in Fig. 21.3 in parts (a) to (f).

From Eqs. (21.5) and (21.6), we computed the amplitudes and phases for each output signal. The spectrum of the output signals is presented in Fig. 21.4, where (a)

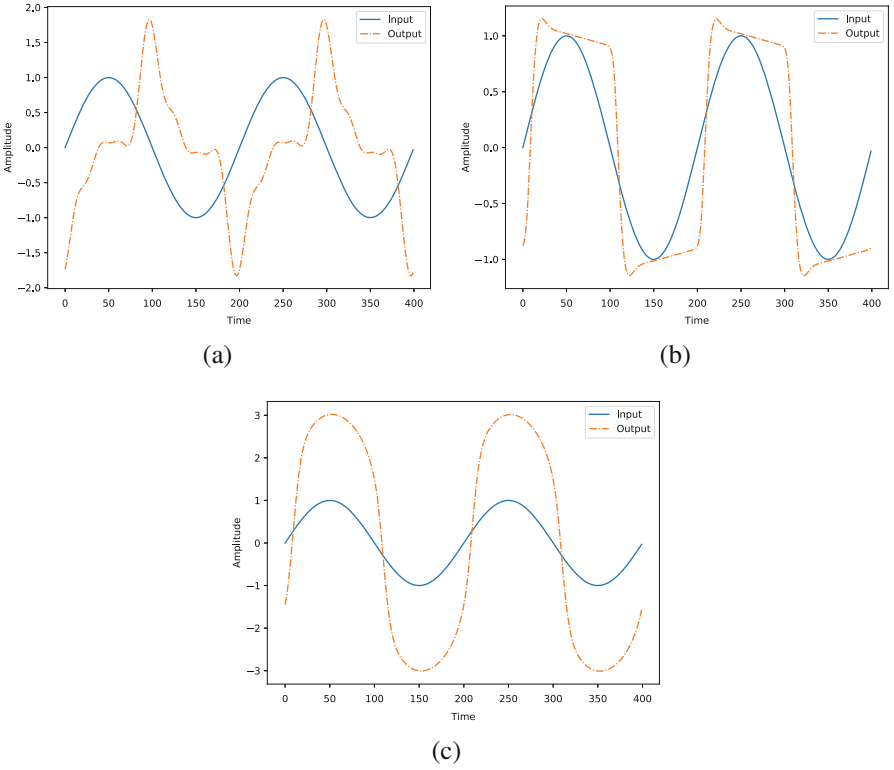



**Fig. 21.2** Input with $f = 1$ kHz and output signals for three different nonlinear transfer functions, (**a**) Synthetic signal, (**b**) Hyperbolic tangent response and (**c**) Hyperbolic sine response

shows the output of the synthetic signal simulation with the harmonics correctly computed according to Eq. (21.7). Figure 21.4b shows the response due to the hyperbolic tangent-shaped transfer function, and (c) shows the response due to the hyperbolic sine transfer function. As was to be expected, the hyperbolic tangent function produces a more strongly distorted output signal than the hyperbolic sine function, manifest in the smaller ratios between the basic frequency amplitude and the upper harmonics in comparison to the hyperbolic sine function.

Knowing the amplitudes and phases, it was possible to reconstruct the output signals based on Eq. (21.1). Figure 21.5 left column shows the input and reconstructed output for the case of canceled phases ($\phi_k = 0$). With the phases equal zero, one may easily verify the distortion behavior imposed by the nonlinear transfer functions. In the right column of the same figure, the output of the synthetic case and the outputs of the Wiener–Hammerstein model are shown together with the reconstructed output signals for the nonlinear transfer functions. By inspection, one may observe that the reconstruction has a fairly good similarity with the original output signal for the three simulations.

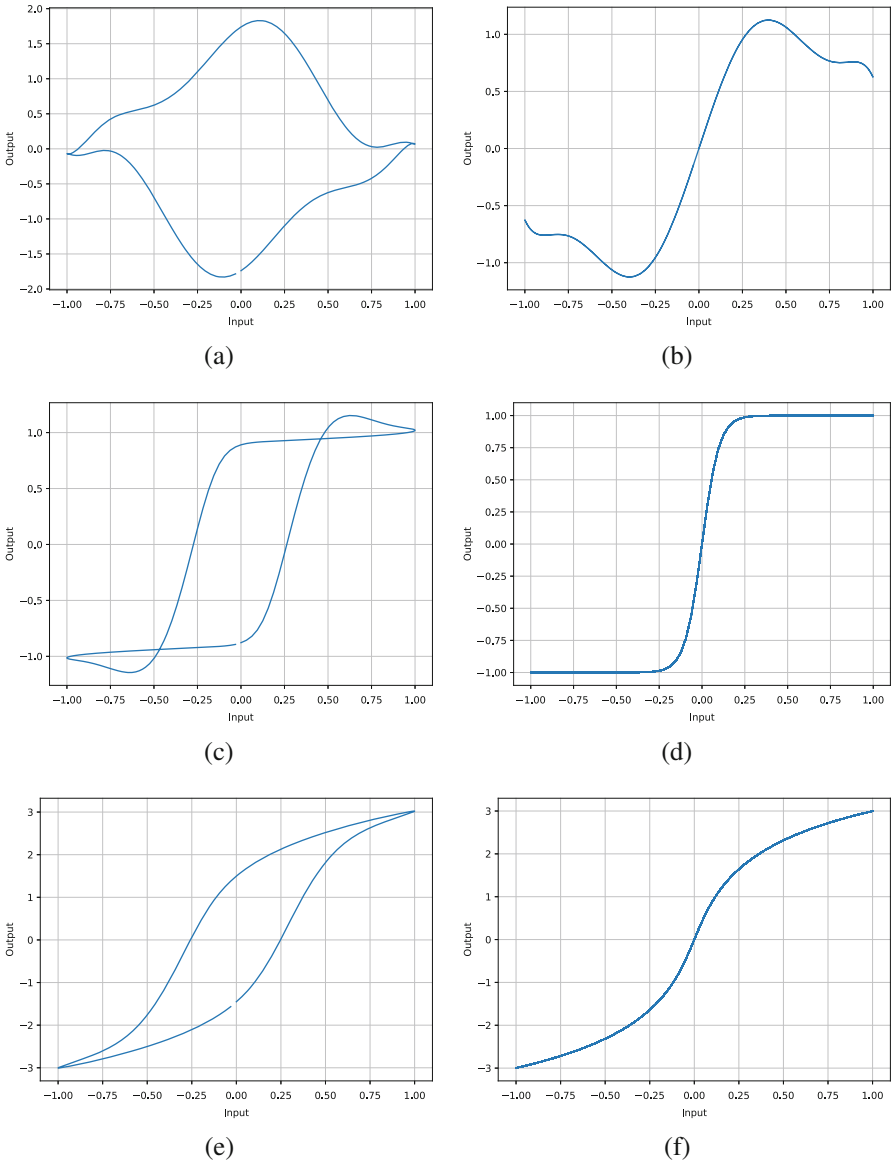



**Fig. 21.3** Lissajous curves of the nonlinear transfer functions with and without the filters $h_1$ and $h_2$ influence. In the case of the synthetic signal with and without phases. (**a**) Synthetic signal with phases. (**b**) Synthetic signal without phases. (**c**) Hyperbolic tangent with filters. (**d**) Hyperbolic tangent without filters. (**e**) Hyperbolic sine with filters. (**f**) Hyperbolic sine without filters

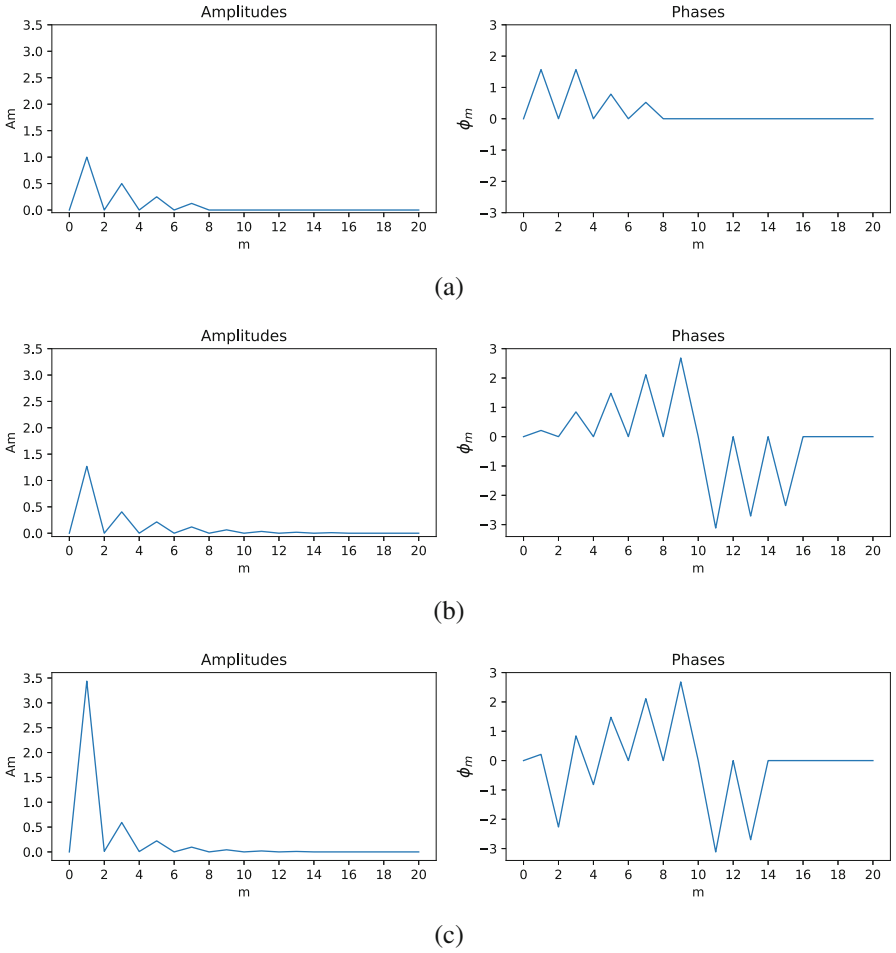
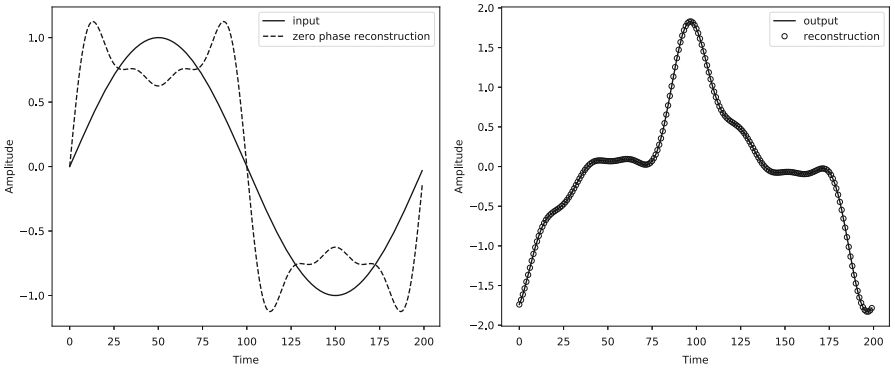
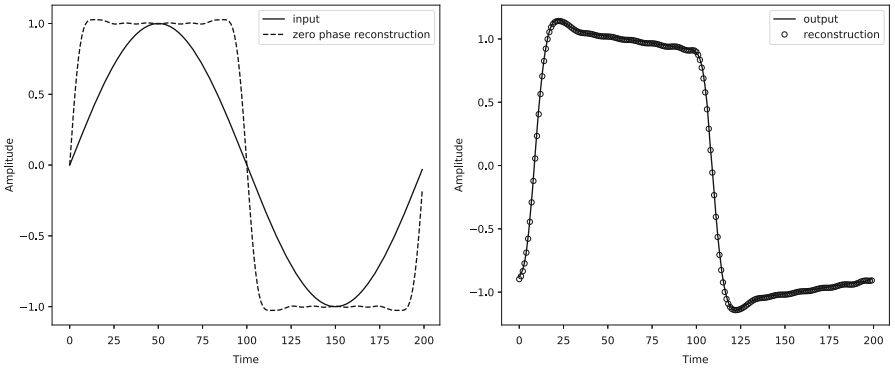



**Fig. 21.4** Spectrum of the output signal for the simulations with three different transfer functions. (**a**) Synthetic signal. (**b**) Hyperbolic tangent. (**c**) Hyperbolic sine

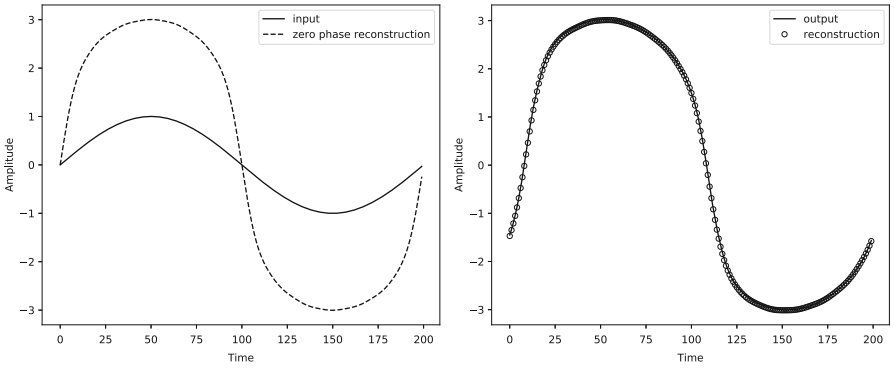
With the input and reconstructed output signals, it is possible to create the Lissajous curve for each nonlinear transfer function. The original input and output signals (Lissajous figure) are shown in the left column of Fig. 21.6, and in the right column, the nonlinear transfer functions from the experiment and with the reconstructed output signal are compared. Again, the reconstructed output signals (with $\phi_k = 0$) show for the three cases a fairly good similarity in comparison to the original output signals, which implies a reasonably good identification of the nonlinear transfer function. Note that in the hyperbolic tangent case, a small discrepancy is observable, and this behavior may be attributed to the $h_1$ and $h_2$ influence in the output signal imposing an imprecision in the phase calculation.




(a)

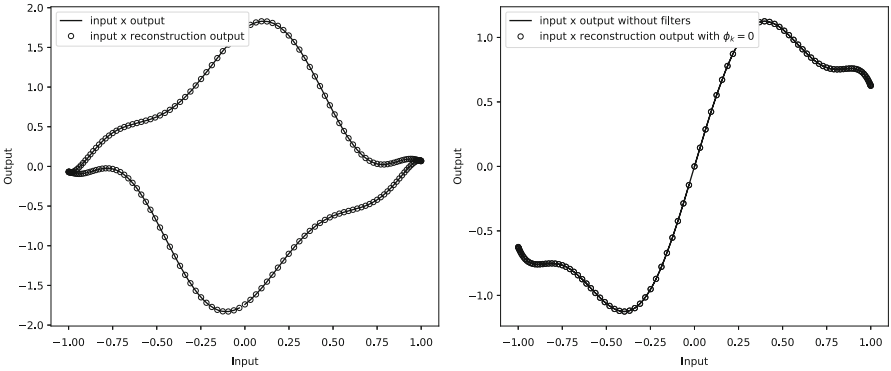



(b)

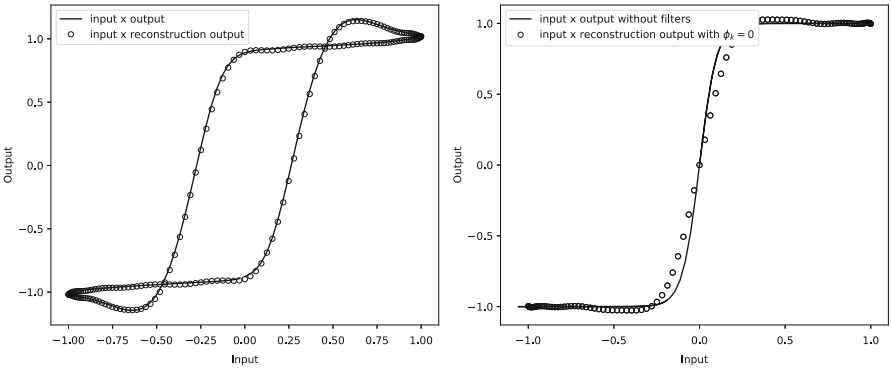



(c)

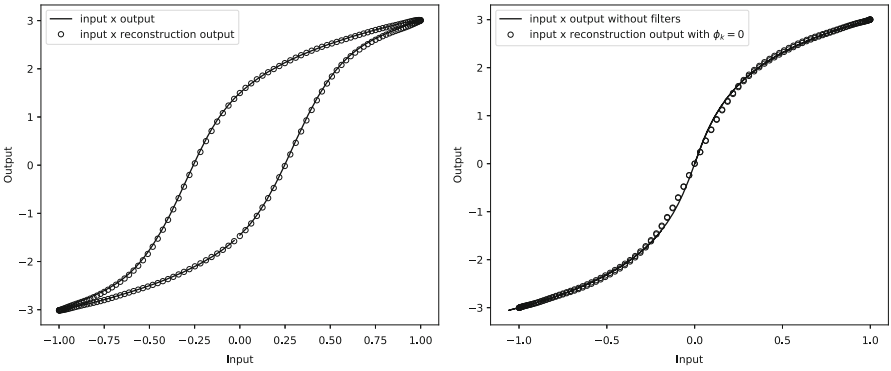
**Fig. 21.5** Comparison of the input and output signals as well as the reconstructed output signal for the three nonlinear transfer functions. In the left column, the input signals are compared to the reconstructed output signal with canceled phases ($\phi_k = 0$). The right column shows the output signal of the modeled Wiener–Hammerstein system and the synthetic signal compared to the reconstructed output signal. (**a**) Synthetic signal. (**b**) Hyperbolic tangent. (**c**) Hyperbolic sine




(a)




(b)




(c)

**Fig. 21.6** The left column shows the Lissajous curve of the original input versus output signals, and the right column shows the identification of the nonlinear transfer function, where for the reconstructed output signal the phases were canceled ($\phi_k = 0$). (**a**) Synthetic signal. (**b**) Hyperbolic tangent. (**c**) Hyperbolic sine

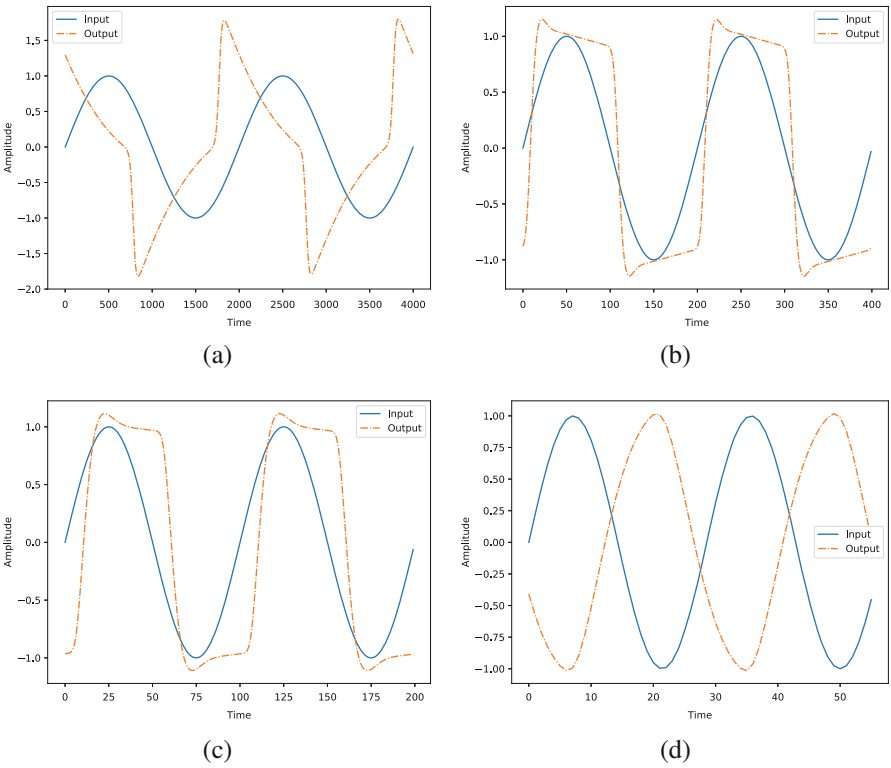



**Fig. 21.7** Input and output signals with different frequencies for the hyperbolic tangent as the nonlinear transfer function. (**a**) 100 Hz. (**b**) 1000 Hz. (**c**) 2000 Hz. (**d**) 7000 Hz

As already mentioned in the introduction in real situations, the signals to be amplified are typically a superposition of various frequencies so that one has to repeat the presented prescription for a set of frequencies where the transfer functions for values between the experimentally determined ones may be obtained by interpolation. In order to give an idea as to how the value of the fundamental frequency influences in the accuracy of the present implementation cases with $\omega = 2\pi 100$ Hz, $2\pi 2000$ Hz, and $2\pi 7000$ Hz were considered maintaining the two band-pass filters identical. The input signals for the cited frequencies and the respective outputs due to the nonlinear transfer functions are presented in Fig. 21.7. Note that as the frequency increases, the distortion tends to diminish so that the phase shifts are the principal effect on the output signal.

From the reconstructed output signals of 100, 2000, and 7000 Hz input frequencies, the spectrum was computed and presented in Fig. 21.8, while the spectrum for the input signal with 1000 Hz may be found in Fig. 21.4. As the distortions tend to be softer for higher frequencies, the number of harmonics tends to decrease. This phenomenon may be addressed to the influence of the $h_1$ and $h_2$ filters since

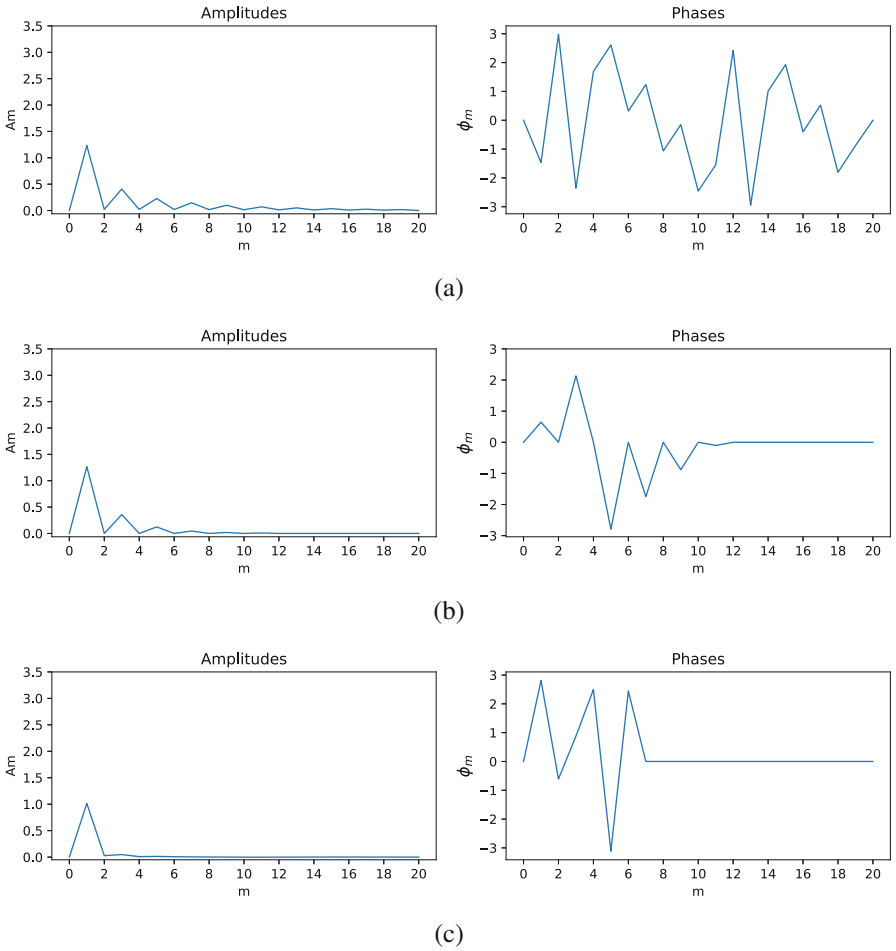



**Fig. 21.8** Spectrum of the output signals for different frequencies with the hyperbolic tangent as the nonlinear transfer function. (**a**) 100 Hz. (**b**) 2000 Hz. (**c**) 7000 Hz

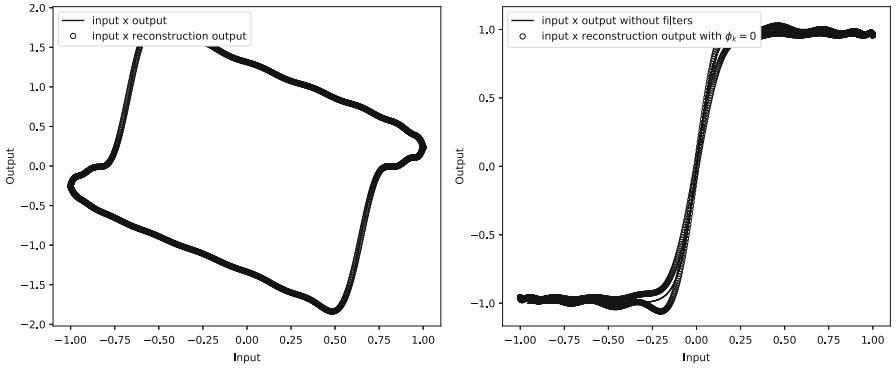
the fundamental frequency gets closer to the cutoff frequency of 8000 Hz, and thus suppression of the amplitudes becomes noticeable. For example, the second harmonics of the 7000 Hz input frequency is 14000 Hz, and it falls into the cutoff region of $h_1$ and $h_2$. Therefore, the more the frequency of the input signal tends to higher frequencies, the more the filters will influence the output signal, which turns it more difficult to profile the nonlinear transfer function. Further, the computed phases presented some instabilities at low frequencies, here for $f = 100$ Hz. These instabilities are due to the low-frequency cutoff for the combination of the filters $h_1$ and $h_2$ for the chosen frequency. Moreover, the topology of the digital filter used in the simulations was the so-called Infinite Impulse Response Filter, for which it is well known that unstable behaviors in the phase calculations occur.

Once the amplitudes and phases were computed from the output signals, the Lissajous curves were constructed together with the nonlinear transfer functions. The reconstructed output signals with input frequencies of 100, 2000, and 7000 Hz are shown in Fig. 21.9, and the result of the 1000 Hz input may be found in Fig. 21.6b. Upon analyzing the fidelity for the different fundamental frequencies, it is apparent that the lower the fundamental frequency is the more clipped the output signal is, turning the reconstruction of the nonlinear curve closer to the original one. This happens because the spectrum of the lower frequencies is richer in harmonics allowing for a reconstruction of the output signal with weak influences of the $h_1$ and $h_2$ filters. Nevertheless, some instabilities in the phase calculations occur, especially if the phase is close to $\pm\frac{\pi}{2}$ or $\pm\pi$, respectively.

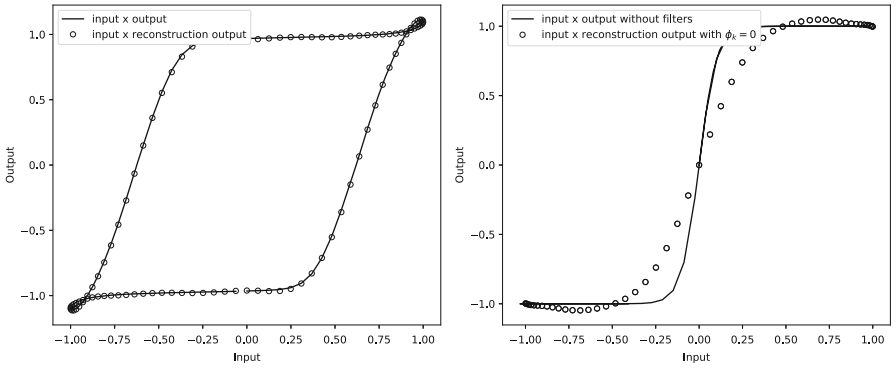
As a result, the model presented good reconstructions of the output signal; however, in some cases, the nonlinear transfer function reconstruction has imprecision. Obviously, the choice of the input frequencies has a great influence in the nonlinear functions determination due to the influence of the $h_1$ and $h_2$ filters. Therefore, a deeper study of the input frequencies selection is of crucial importance in the nonlinear transfer function determination.

## 21.6 Conclusions and Future Work

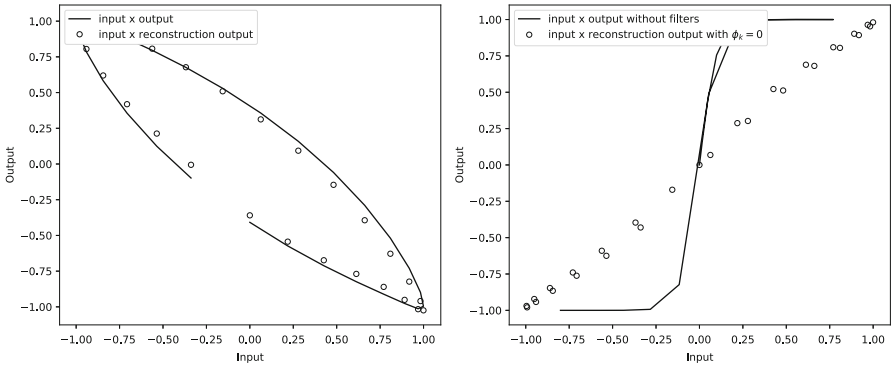
The novelty of the present work in the realm of amplification profiling is the use of Lissajous curves to separate the linear and nonlinear block contributions of a Wiener–Hammerstein model. It is the combination of an input signal and the experimental response of the linear–nonlinear–linear arrangement of components that allowed to establish an analytical expression for amplitudes and phases of the output signal based on the experimental values of the Lissajous data. There do exist several attempts in the literature for the same problem; however, these depend on iterative or other numerical schemes in order to attain the relevant parameters. In the contrary as outlined in Sect. 21.4, we could establish a direct relation between experimental data and parameterized signal descriptions, which work reliable for the amplitudes but still need some refinement for the phase determination. Especially, in the higher-frequency range where upper harmonics are damped by the band-pass filter, it may cause problems of arithmetic origin in the phase calculation due to small amplitudes and needs to be mitigated possibly by an expression containing the complex logarithm instead of the arctangent function. Ideas in this line are already in progress but will be a subject of a future work. Nevertheless, the comparisons between the input and reconstructed output signals show that the method is promising although some refinement is still in order. Various causes may affect the quality of results such as numerical precision but also precision of the experimental data due to the sampling process could be an issue, so that a more detailed analysis is in order to shed more light on the origin of the errors in the reconstructed nonlinear transfer function.




(a)




(b)




(c)

**Fig. 21.9** Reconstruction of the nonlinear transfer function for different input signal frequencies. (**a**) 100 Hz. (**b**) 2000 Hz. (**c**) 7000 Hz

There are immediate future challenges such as the pertinent question as to how to handle noisy signals. Further, the input signals of real sound scenarios have typically a rich spectrum so that a natural continuation of the present work is to consider more realistic input signals and extend the profiling analysis to such cases. An additional issue is that the present developments apply to input–output signals that allow to create closed Lissajous curves; however, every fragment of a musical piece is by virtue a transient phenomenon so that another integral transform shall be employed such as the Hilbert transform, which allows to determine instantaneous frequencies. In any case, all these aspects require an extension of the discussed method or even a modification by further developments to these types of problems.

## References

[DuEtAl20] Du, V.N., Reinhard, K., Anna, W., Weihe, P.: Confusingly similar: discerning between hardware guitar amplifier sounds and simulations with the kemper profiling amp. Music Sci. **3**, 1–16 (2020)

[EiEtAl17] Eixas, F., Möller, S., Zölzer, U.: Block-oriented gray box modeling of guitar amplifier. In: Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17), Edinburg, pp.184–191 (2016)

[EiZo16] Eixas, F., Zölzer, U.: Black-box modeling of distortion circuits with block-oriented models. In: Proceedings of the 19th International Conference on Digital Audio Effects (DAFx-16) Brno, Czech Republic, pp. 39–45 (2016)

[EiZo18] Eichas, F., Zölzer, U.: Virtual analog modeling of guitar amplifiers with Wiener-Hammerstein models. In: 44th Annual Convention on Acoustics (DAGA 2018) (2018)

[EiZo18a] Eichas, F., Zölzer, U.: Gray-box modeling of guitar amplifiers. J. Audio Eng. Soc. **66**(12), 1006–1015 (2018)

[Ke15] Kemper Profiler: Profiling guide (2015). Retrieved from http:/www.audioline.it/catalogo/allegati/2452_Kemper_Profiler_Rack.pdf

[MoEtAl15] Möller, S., Eichas, F., Zölzer, U.: Block-oriented modeling of distortion audio effects using iterative minimization. In: Proceedings of the 18th International Conference on Digital Audio Effects (DAFx-15), Trondheim, Norway, pp. 1–6 (2015)

[Og07] Ogunfunmi, T.: Adaptive Nonlinear System Identification: The Volterra and Wiener Model Approaches. Springer, Berlin (2007)

[OlSc13] De Oliveira, L.P.L., Schmith, J.: Ponderomotive force as a cause for chaos in loudspeakers. Appl. Math. Comput. **219**, 6449–6456 (2013)

[RoEtAl14] Rolain, Y., Schoukens, M., Vandersteen, G., Ferranti, F.: Fast identification of Wiener-Hammerstein systems using discrete optimisation. Electron. Lett. **50**(25), 1942–1944 (2014)

[ScOl21] Schmith, J., de Oliveira, L.P.L.: Dimensioning sealed enclosures for suppressing nonlinear distortions in woofers. Appl. Acoust. **178**, 107975 (2021)

[ScEtAl14] Schoukens, M., Zhang, E., Schoukens, J.: Structure detection of Wiener-Hammerstein systems with process noise. IEEE Trans. Instrum. Meas. **66**(3), 569–576 (2014)

[SjSc12] Sjöberg, J., Schoukens, J.: Initializing Wiener-Hammerstein models based on partitioning of the best linear approximation. Automatica **48**(1), 353–359 (2012)

[Zo11] Zölzer, U.: DAFX: Digital Audio Effects, vol. 2. Wiley Online Library (2011)

# Chapter 22
# Variation of Zero-Net Liquid Holdup in Gas–Liquid Cylindrical Cyclone (GLCC©)


**M. Shah, H. Zhao, R. Mohan, and O. Shoham**

## 22.1 Introduction

Separation of produced fluids is an important operation in the petroleum industry. For decades, the industry has relied mainly on conventional gravity-based separators to separate the production of gas, oil, and water. Conventional separators are large and heavy, retain large inventory of produced fluids, and require a considerable retention time for the separation process. Since the early 1990s, the industry has shown keen interest in developing an alternative to the conventional separator, namely compact multiphase cyclonic separators owing to their significant operational and economic merits. These include simplicity in construction, compactness, low weight, small footprint, and low capital and operational costs.

The Tulsa University Separation Technology Projects (TUSTP) university/industry research consortium has advanced state-of-the-art compact multiphase cyclonic separation technology for gas–oil–water–sand flow for the past 28 years. Individual compact separator devices were developed, such as Gas–Liquid Cylindrical Cyclone (GLCC©[1]), Liquid–Liquid Cylindrical Cyclone (LLCC©), Liquid–Liquid Hydro Cyclone (LLHC), Horizontal Pipe Separator (HPS©), and the integrated Compact Multiphase Separation System (CMSS©). Emphasis has been placed on measurement and understanding of the hydrodynamic flow behavior in compact separators. Mechanistic models were developed for each of the compact separators and on the development of design tools for the industry. To date, there are over 8,200 compact separators deployed in the field in the USA and around the world, including full separation or partial separation applications, as well as onshore, offshore, and subsea applications.

M. Shah · H. Zhao · R. Mohan · O. Shoham (✉)
The University of Tulsa, Tulsa, OK, USA
e-mail: mjs3302@utulsa.edu; haz7042@utulsa.edu; ram-mohan@utulsa.edu; ovadia-shoham@utulsa.edu



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_22

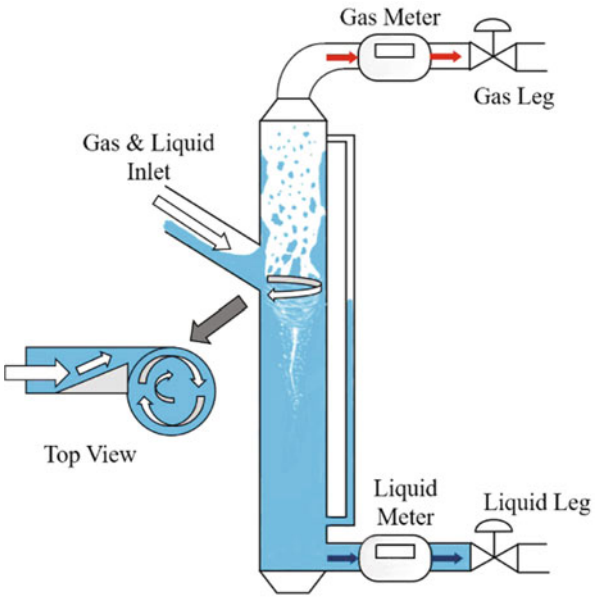



**Fig. 22.1** A schematic of the GLCC©

Figure 22.1 presents a schematic of the GLCC©. As can be seen, the GLCC© body is simple pipe section, mounted vertically with a downward inclined tangential inlet.

The top of the GLCC© is connected to the outlet gas leg, and the bottom part of the GLCC© is connected to the outlet liquid leg. The downward inclined inlet promotes pre-separation stratification of the gas–liquid flow, prior to flowing into the vertical GLCC© body. As demonstrated in the top view in the figure, a nozzle with a cross-sectional area of 25% of the full-bore inlet pipe is installed at the end of the inlet section tangentially to the GLCC© wall. Thus, the mixture flowing from the nozzle forms a swirling motion around the GLCC© inlet region. The swirling flow generates centrifugal forces that act on the gas and the liquid, whereby the heavier liquid phase is forced radially toward the wall and moves downward due to gravity. On the other hand, the lighter gas phase moves radially inward toward the center of the cyclone and flows upward to the GLCC© top. Finally, the liquid exits from the bottom of the GLCC© through the outlet liquid leg, and the gas flows out through the gas-leg outlet. For efficient separation, the liquid level is maintained at about 6" below the inlet.

GLCC© separators have a wide range of applications, as given by Gomez [Go98] and Shoham and Kouba [Sh98]. The most common field application of the GLCC© is the multiphase flow metering loop, as shown schematically in Fig. 22.2. In this application, a GLCC© separates the gas from the liquid, whereby single-phase meters, such as Coriolis mass flow meters, turbine meters, etc., are installed on the gas and liquid legs to measure the respective phase flow rates. The water cut can be measured utilizing a water cut meter that is installed on the liquid leg.

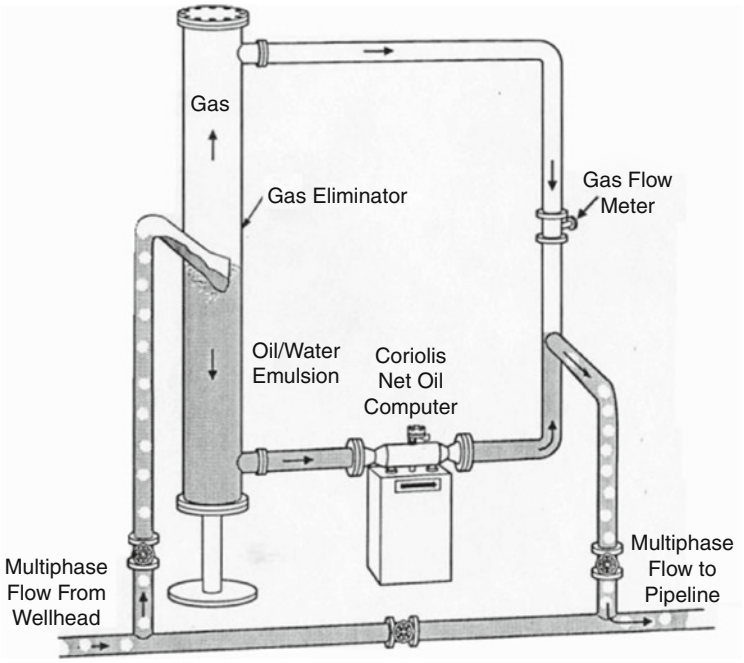



**Fig. 22.2** Schematic of GLCC© multiphase flow metering loop

As opposed to the available commercial multiphase flow meters, in the GLCC© multiphase flow metering loop configuration, actual measurements of the phases are carried out, which makes it a more reliable and more accurate system, in addition to its low cost. The gas and liquid legs are recombined downstream at a height of about 6 in. below the GLCC© inlet.

The GLCC© operation is limited by two physical phenomena, namely, liquid carry-over (LCO) and gas carry-under (GCU). In LCO, some liquid is carried over into the gas leg by the gas phase, while GCU occurs when some gas is carried by the liquid into the liquid leg. Figure 22.3 presents a schematic of the operational envelope (OPEN) for LCO. As can be seen, the OPEN for LCO separates the normal operating conditions (region below curve) from the LCO region (above the curve). This study is carried out for normal operating conditions below the OPEN.

During normal operating conditions below the OPEN for LCO, some liquid is held up in the upper part of the GLCC© in the form of churn flow, as shown in Fig. 22.1. The held-up liquid just moves up and down, but no liquid is produced into the gas leg. For these conditions, the gas phase velocity is sufficient to push some liquid into the GLCC© upper part but not high enough to carry the liquid into the gas leg. This phenomenon is termed zero-net liquid flow (ZNLF), and the associated liquid holdup in the upper part of the GLCC© is termed zero-net liquid holdup (ZNLH). Prediction of the ZNLH is important as it plays a significant role in the

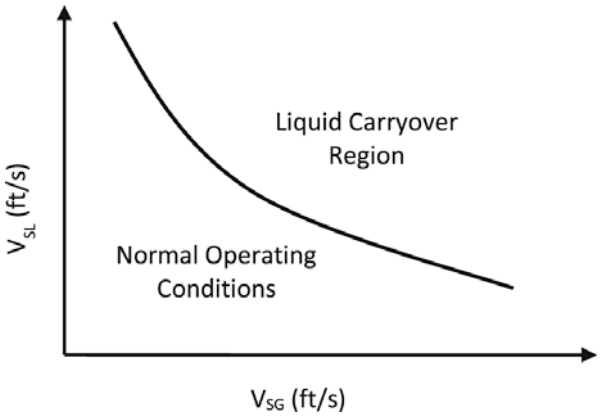



**Fig. 22.3** Schematic of GLCC© operational envelope for liquid carry-over

GLCC© pressure balance and in its design. In the past, measurements of the ZNLH were limited to either static liquid conditions or dynamic measurements at the OPEN for LCO flow conditions. No ZNLF data have been acquired on its variation in the upper part of the GLCC© under normal operating conditions, which is the gap that the present study aims to close. The main objectives of the study are as follows: (1) Acquire experimental data for the variation of the ZNLH and the associated churn flow height along the upper part of the GLCC© under normal operational conditions below the OPEN for LCO for both air–water and air–oil flow; (2) Develop a model for the prediction of the variation of ZNLH and churn region height in the upper part of the GLCC©; (3) Compare the developed model predictions with the acquired experimental data.

## 22.2 Literature Review

The Tulsa University Separation Technology Projects (TUSTP) has been a pioneer in the development of the GLCC©. Detailed experimental and theoretical studies on the GLCC© have been carried out since 1994, which are reviewed briefly. Also presented is a summary of published studies on ZNLF.

### *22.2.1 GLCC© Experimental Studies*

Arpandi et al. [Ar96] published the first TUSTP study on the hydrodynamic flow behavior in the GLCC©. The authors acquired data including the OPEN for LCO, equilibrium liquid level, static ZNLH, and pressure drop across the system. They also developed a rudimentary mechanistic model capable of predicting the GLCC© flow behavior, including the ZNLH and the OPEN for LCO.

The optimal design of the GLCC© was investigated by Kouba et al. [Ko95]. The authors presented laboratory and limited field data on the GLCC© flow behavior and separation efficiency. They noticed that the OPEN for LCO expands considerably utilizing an inclined inlet for the GLCC©. Later, Shoham and Kouba [Sh98] and Mohan and Shoham [Mo99] presented the development and design of the three-phase flow GLCC©.

Movafaghian et al. [Mo00] gathered experimental data on the effects of GLCC© geometry, fluid properties, and pressure on its performance. Acquired data were compared against the model developed by Arpandi et al.[Ar96] and showed good agreement. Wang [Wa00a], Wang et al. [Wa00b], and Wang et al. [Wa00c] studied control strategies for the GLCC©. In the following study, Wang et al. [Wa02] published a modified GLCC© for wet-gas separation and metering, which is equipped with an annular film extractor. The authors tested the wet-gas GLCC© in a field-scale flow loop at high pressure (up to 1,000 psia) with natural gas and Decane. The results show the expansion of the OPEN for LCO and improved performance of the modified GLCC©.

An extension study on the GLCC© performance under 3-phase gas–oil–water flow was presented by Kolla et al. [Ko19] using active control for liquid level. Experimental data were collected on the OPEN for LCO for water cuts between 0% and 100%. The authors reported that as the water cut reduces, the OPEN for LCO reduces, too. A rudimentary model was developed for the predictions of the OPEN for LCO for 3-phase flow, which predicts the experimental data fairly.

### *22.2.2 GLCC© Mechanistic Modeling*

An overall mechanistic model and design code for the GLCC were developed by Gomez [Go98] and Gomez et al. [Go99] and [Go00]. The model that enables the prediction of the hydrodynamic flow behavior in the GLCC©: the inlet gas and liquid tangential velocities, gas–liquid interface (vortex) profile, equilibrium liquid level, unified trajectory model for both bubbles and droplets, gas carry-under, and the OPEN for LCO. The model can also predict the aspect ratio of the GLCC©.

Mantilla et al. [Ma99] developed correlations, based on experimental data, for the prediction of the tangential and axial velocity profiles in the GLCC©. The correlations were utilized to improve the bubble trajectory model for the prediction of GCU.

In a later study, Gomez [Go01] developed a model for the prediction of GCU for normal operating conditions below the OPEN for LCO. Chirinos et al. [Ch00] acquired experimental data and proposed a model for the LCO at flow conditions higher than the OPEN for LCO.

### 22.2.3 GLCC© Zero-Net Liquid Holdup in Vertical Pipes

The effect of pressure on static ZNLH in vertical pipes was studied by Duncan and Scott [Du98] and Liu and Scott [Li01] utilizing a large diameter field-scale facility operated at high pressures. They found that the existing ZNLH models predict purely for pressures greater than 100 psig and proposed a methodology to predict ZNLH at high pressures.

Hyoung et al. [Hy00] investigated the effect of fluid properties on ZNLH. The study concluded that increasing the liquid density results in a decrease of the ZNLH, while as the liquid viscosity increases, the ZNLH is increased. The effects of foam and emulsion on ZNLH in a GLCC© were studied by Adebare [Ad06] in an attempt to optimize the GLCC© design. He noticed that the inlet slot configuration had no effect on ZNLH.

A more recent study, which preceded the current study and utilized the same facility, was published by Karpurapu [Ka18a]. Experimental data were acquired for GLCC© dynamic ZNLH at the OPEN for LCO flow conditions for air–water and air–oil flows. The air–oil flow exhibits higher ZNLH, as compared to air–water flow. The "Flooding and Flow Reversal" dimensionless model by Wallis [Wa61] for pipe flow was extended to enable the prediction of the dynamic ZNLH in a GLCC©, showing a good agreement with the acquired experimental data [Ka18b].

## 22.3 Experimental Program

This chapter presents the test facility and instrumentation, test matrix and testing procedure, as well as the acquired experimental results.

### 22.3.1 Experimental Facility

The experimental facility utilized is the TUSTP 4-phase flow loop, which enables conduct of experiments utilizing gas, water, oil, and solid particles. The flow loop components are described in the following sections.

#### 22.3.1.1 Experimental Flow Loop

A schematic of the TUSTP flow loop is presented in Fig. 22.4. Gas is supplied by a compressor with a flow capacity of 250 CFM at a pressure of 100 psig. Two 400-gallon tanks that are open to the atmosphere are utilized to store water and oil. The oil or the water can be pumped from their respective storage tanks into the flow loop by either a maximum of 10 HP (a maximum of 25 gpm) or a maximum of

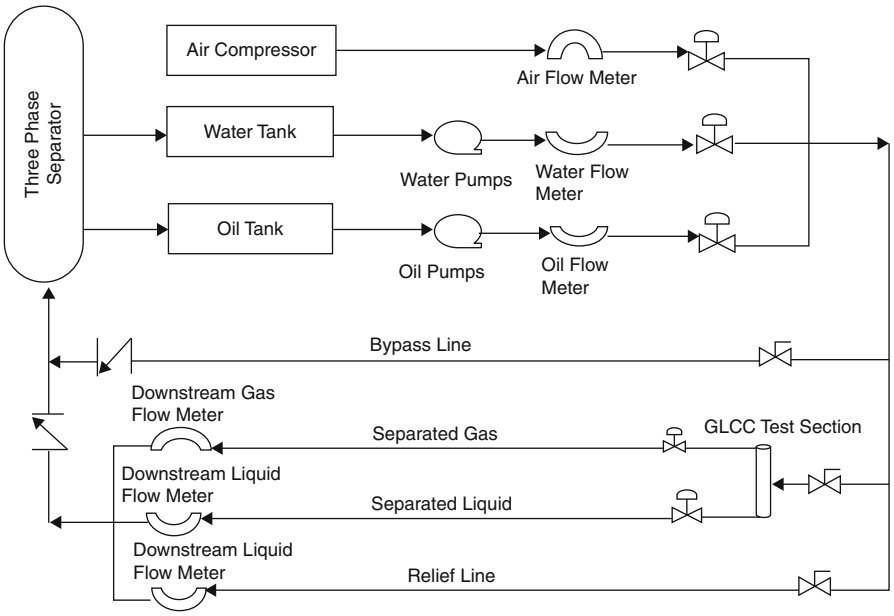



**Fig. 22.4** Schematic of experimental flow loop

25 HP (a maximum of 110 gpm) centrifugal pumps. The pumps are operated with variable frequency drive to control the flow rate. Note that in the current study, only the 25 HP pumps are used to pump the oil and water. The air, oil, and water are controlled and metered by the control valves and mass flow meters, respectively. The 3 phases can be mixed at the mixing tee, and the multiphase mixture is directed into the GLCC© test section. The separated gas and liquid phases are recombined downstream of the GLCC© before flowing into the 3-phase separator. The separator discharges the air to the atmosphere and recycles the clean water and clean oil into their respective storage tanks.

#### 22.3.1.2 GLCC©-Test Section

A schematic of the GLCC© test section is presented in Fig. 22.5. The test section is made of a schedule 80 acrylic pipe. The GLCC© body is a 3" ID 10-ft-tall vertical pipe, whereby the upper part of GLCC© above the inlet is 65" high and the lower part below the inlet is 55" high. The GLCC© inlet is 3" ID and is inclined downward at $-27°$ to the horizontal. A nozzle is installed at the end of the inlet section with a cross-sectional area of 25% of the inlet pipe cross-sectional area. The nozzle diverts the flow tangentially onto the GLCC© body wall.

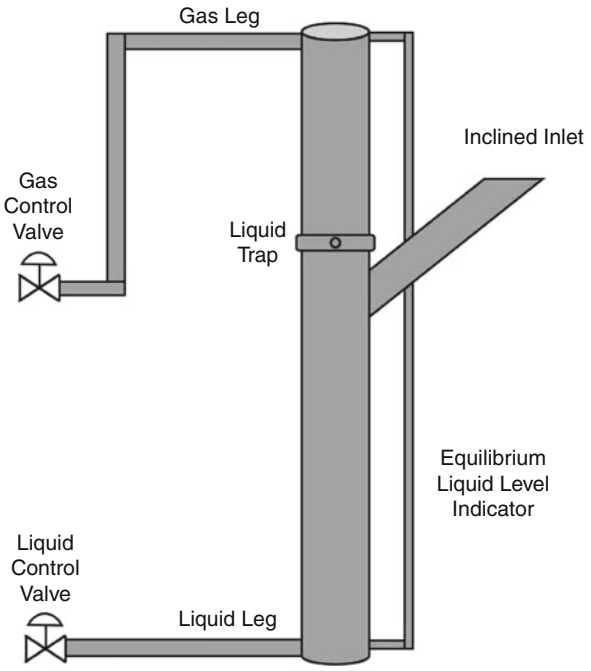



**Fig. 22.5** Schematic of GLCC© test section

**Table 22.1** Water properties

| | |
|---|---|
| Specific gravity | 0.998 |
| Viscosity @ 68 °F | 1.3cP |
| Surface tension @ 77 °F | 70 dyne/cm |

A 2" ID liquid leg and a 2" ID gas leg are installed, respectively, at the bottom and top of the GLCC© body. The GLCC© also includes a vertical liquid-level indicator made of a 0.75" ID transparent acrylic pipe, which is installed parallel to its body. The level in the GLCC© is monitored with the help of differential pressure transducer, which measures the pressure difference between the top and bottom of the GLCC©. A liquid trap is installed in the upper part of the GLCC© 15" above the inlet, which is utilized to trap the flow in the upper part of the GLCC© to measure the ZNLH. Two control valves are installed, one on each leg of the GLCC©, for controlling the pressure and liquid level in the GLCC©.

#### 22.3.1.3 Working Fluids

The working fluids utilized are air, tap water, and TULCO Tech 80 mineral oil. The main reasons for choosing these fluids are fast separation and low emulsification. The properties of water and oil are given in Tables 22.1 and 22.2, respectively.

**Table 22.2** TULCO tech 80 oil properties

| | |
|---|---|
| Specific gravity | 0.851 |
| Viscosity @ 68 °F | 31.7cP |
| Surface tension @ 77 °F | 25.5 dyne/cm |
| Flash point | 365 °F |
| Pour point | +10 °F |

#### 22.3.1.4 Instrumentation and Data Acquisition

The flow loop instrumentation and metering and control devices include the following: pressure and temperature transducers, Coriolis mass flow meters, pumps with variable frequency drives, gas control valves (GCV), and liquid control valves (LCV). The output–input signals of all the devices are connected to a computer via an analog-to-digital converter. The flow loop operation and data acquisition are carried out via the LabVIEW software. LabVIEW provides a graphical programming approach that helps visualize and operate all aspects of a system. These include hardware, operation and control, data acquisition, processing, and presentation. This software enables integration of complex algorithms on a simple visual diagram. In addition to the front panel interface, 2 more interfaces are created in LabVIEW, namely, the level control and pressure control interfaces. The acquired data during the experiments are converted into spreadsheet files, which are later processed to enable graphical and numerical (tables) representation of the experimental results.

### *22.3.2 Test Matrix*

Experiments are conducted for ZNLF at normal operating conditions below the OPEN for LCO. The collected data include the ZNLH and churn region height variations along the upper part of the GLCC© for both air–oil and air–water flows. The test matrix is presented in Figs. 22.6 and 22.7, respectively, for the air–oil and air–water experiments. Both figures depict the operational conditions, namely, the gas and liquid superficial velocity combinations, for which the data are acquired. As can be seen, as the superficial liquid velocity increases, the range of superficial gas velocities for a given superficial liquid velocity decreases. Note that the limiting conditions for the superficial velocities constitute the OPEN for LCO.

### *22.3.3 Testing Procedure*

The following procedure is implemented for acquiring the data, including the ZNLH and the churn flow region height:

1. Set the superficial liquid velocity to the desired value.
2. Slowly increase the gas flow rate to achieve the desired superficial gas velocity.

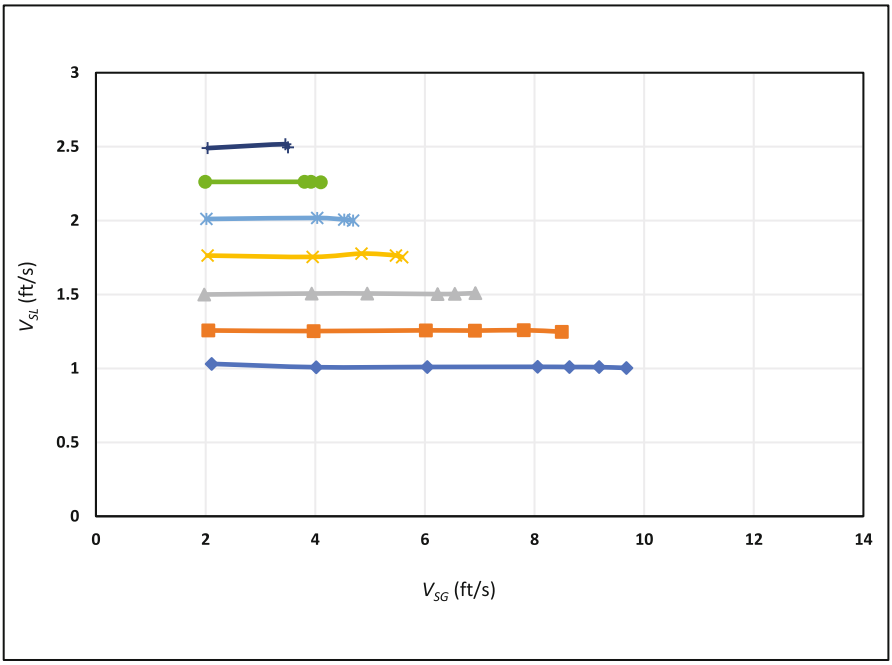



**Fig. 22.6** Test matrix for air–oil flow

3. Once the flow is stabilized, set the liquid level in the GLCC© to 6" below the inlet.
4. Record the churn flow region height.
5. Simultaneously shut the liquid trap and shut down the flow loop for safety reasons.
6. Allow the held-up liquid to settle and the trapped gas to separate and escape.
7. Measure the height of the liquid level above the liquid trap for calculating the ZNLH.
8. Repeat steps 1–7 for all liquid and gas superficial velocities for both air–oil and air–water flows.

## 22.4 Results and Discussion

This chapter presents the experimental results for the ZNLH variation and the associated churn region height along the GLCC© upper part for normal operational conditions below the OPEN for LCO. Also presented is the extension of the Karpurapu [Ka18a] correlation, which enables predictions of the ZNLH and the churn region height variations for normal operating conditions. Finally, a comparison is

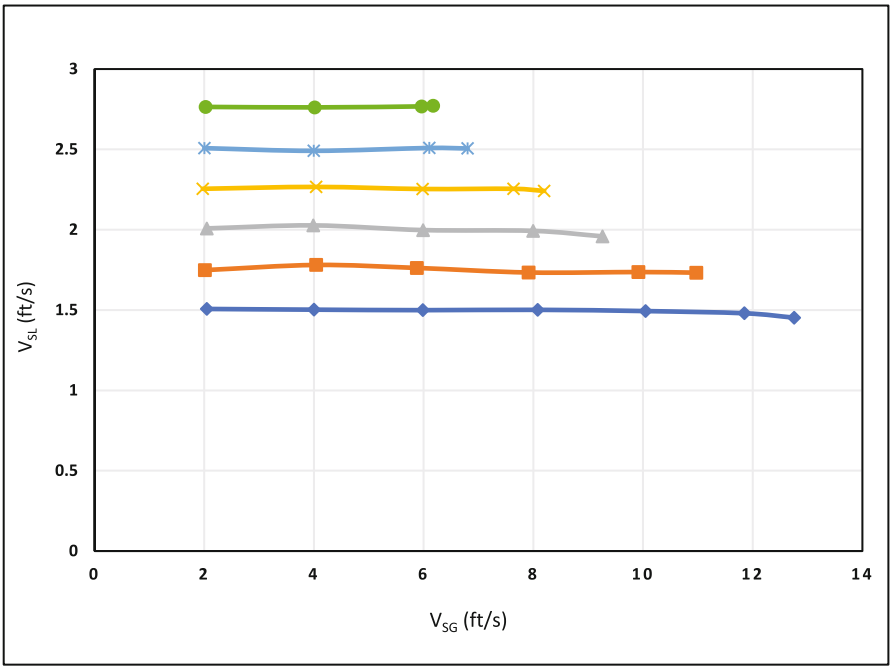



**Fig. 22.7** Test matrix for air–water flow

presented between the predictions of the developed extended correlation and the acquired experimental data.

### *22.4.1 Zero-Net Liquid Holdup Variation*

Experimental data are acquired for ZNLH variation along the GLCC© upper part for both air–oil and air–water flows, which are presented in Figs. 22.8 and 22.9, respectively. As can be seen, for superficial gas velocity below 2 ft/s, very small quantity of liquid may be carried into the GLCC© upper part, and the ZNLH is close to zero. However, as the superficial gas velocity is increased above 2 ft/s, liquid starts flowing into the GLCC© upper part in the form of churn flow, whereby the churn region height increases with increasing superficial gas velocity until it reaches the top of the GLCC© (the GLCC© gas leg). The growth of the ZNLH and churn region height is exponential for air–oil flow and linear for air–water flow. Note that the ZNLH increases with increasing superficial gas velocity, whereby the loci of the maximum ZNLH for each superficial liquid velocity constitute the OPEN for LCO, as depicted in the figures.

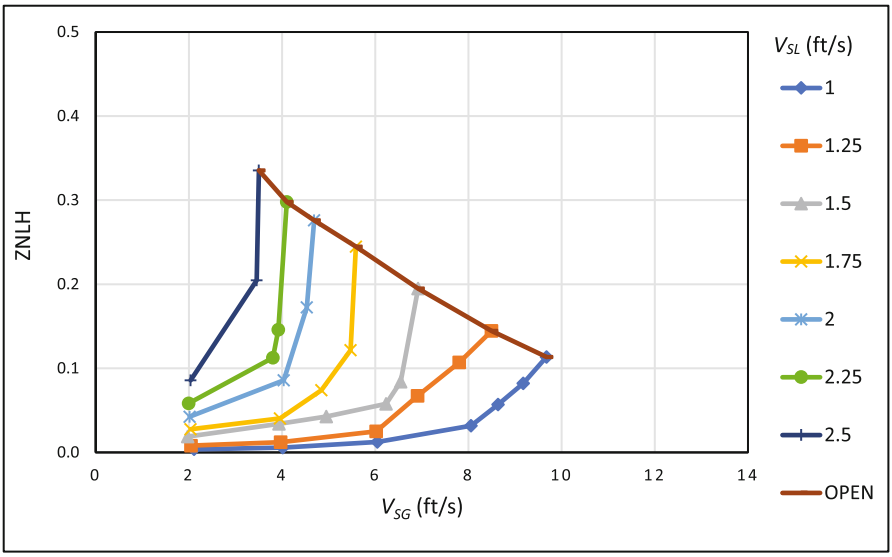



**Fig. 22.8** Variation of ZNLH for air–oil flow

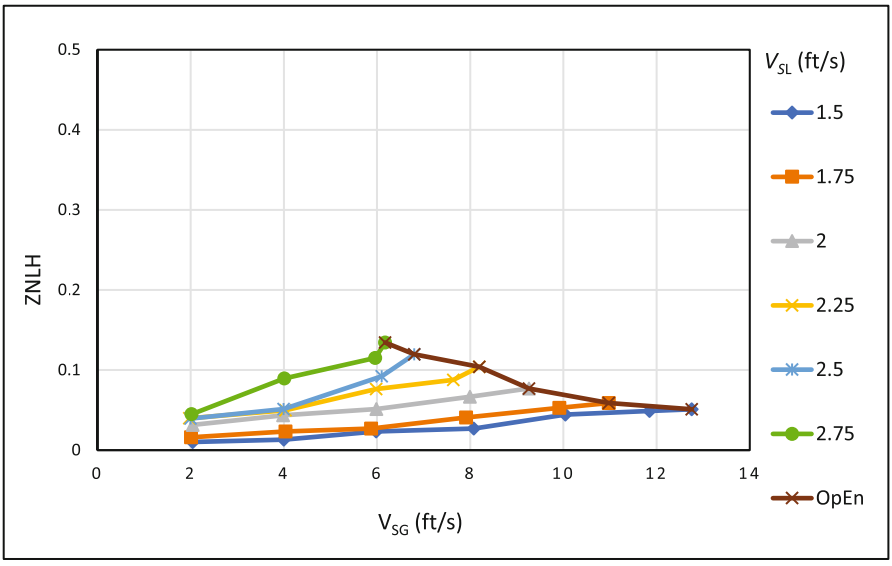



**Fig. 22.9** Variation of ZNLH for air–water flow

A comparison between the air–oil and air–water ZNLH experimental results demonstrates higher ZNLH for the air–oil flow, as compared to the air–water case, owing to the higher viscosity of the oil phase. The higher viscosity results in a higher drag force between the gas and the oil phase and higher friction between the oil and

the pipe wall. The higher frictional and drag forces lead to greater oil retention and higher ZNLH for air–oil flow. Also, for both cases, the ZNLH is very low at superficial gas velocity below 2 ft/s.

### 22.4.2 ZNLF Churn Region Height

The ZNLF churn region height in the GLCC© in upper part can vary in length from zero (at the GLCC© inlet) to its maximum when reaching the top of the GLCC©. Figures 22.10 and 22.11 give the acquired experimental data for the churn region height for air–oil and air–water flows, respectively. As can be seen, the churn region height curves are similar in shape to the ZNLH curves (see Figs. 22.8 and 22.9). As before, liquid is observed in the GLCC upper part for superficial gas velocities greater than 2 ft/s. Also, the churn region height is sensitive to the superficial gas velocity, whereby small increments of gas velocity result in an exponential growth of the churn region height. Note that the maximum churn region height of 60 in. corresponds to the height of the GLCC upper part. A comparison between Figs. 22.10 and 22.11 reveals that for the air–oil flow, owing to the higher oil viscosity, the churn region develops and reaches its maximum value at lower superficial gas velocities, as compared to the air–water case.

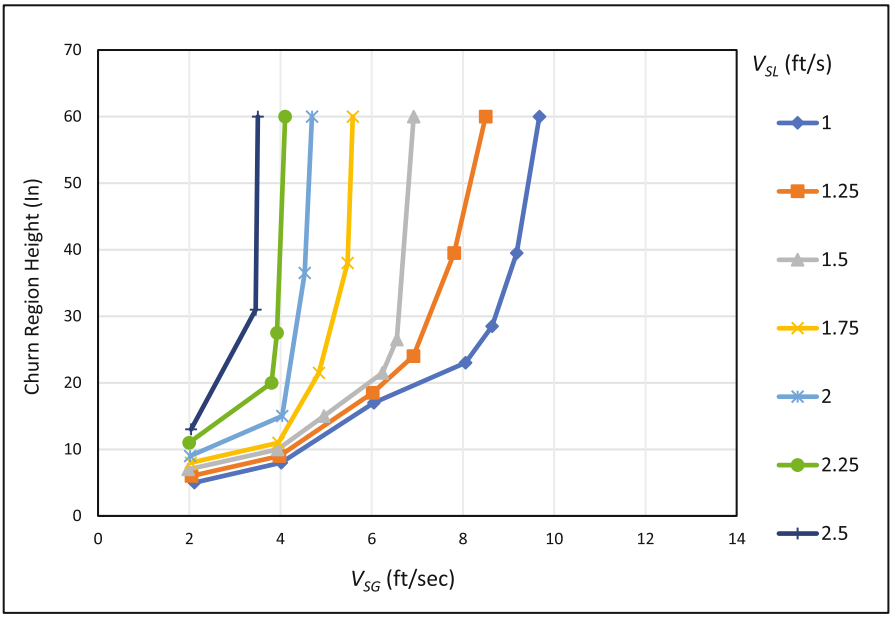



**Fig. 22.10** Churn region height for air–oil ZNLF

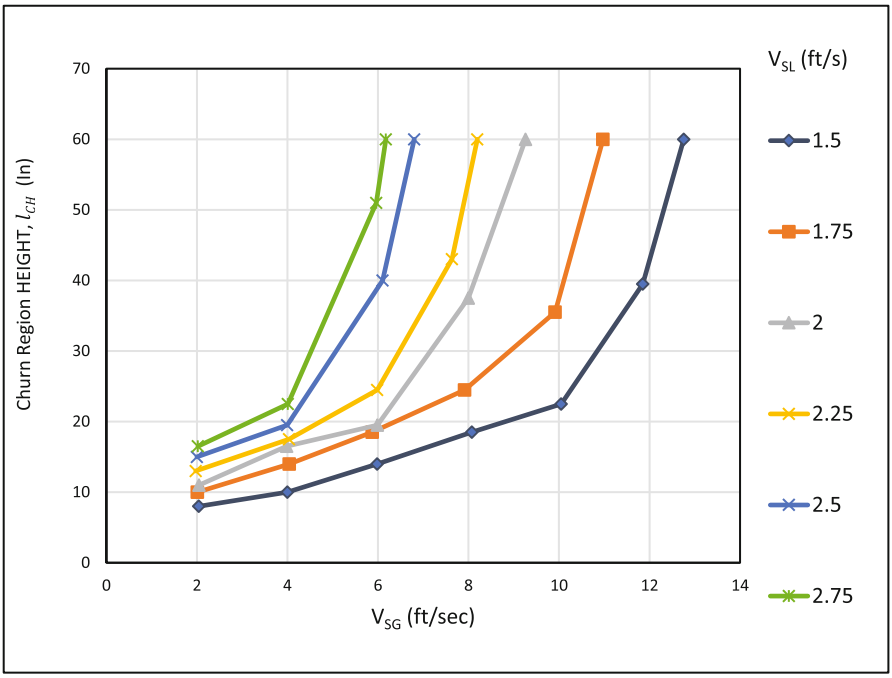



**Fig. 22.11** Churn region height for air–water ZNLF

### *22.4.3 ZNLH Variation Extended Correlation*

Karpurapu [Ka18a] developed a correlation for predicting the ZNLH at the OPEN for LCO flow conditions, $H_{ZNL0}$. For this case, the churn region height occupies the entire upper part of the GLCC©, $l_{TOP}$, but no liquid is carried out by the gas into the upper gas leg. This correlation is extended in this study to predict the variation ZNLH along the upper part of the GLCC for normal operating conditions below the OPEN for LCO, namely, $H_{ZNL}$, and the associated churn region height, $l_{CH}$.

The developed ZNLH variation extended model consists of 2 parts, as follows: The first part utilizes Karpurapu model [Ka18a] to determine $H_{ZNL0}$ for the operating superficial gas velocity. In the second part, the ZNLH for the operating conditions, namely, $H_{ZNL}$, is predicted, as well as the corresponding churn region height, i.e., $l_{CH}$. The developed extended correlation is presented next.

Part 1: The original Karpurapu [Ka18a] correlation is based on Wallis' [Wa61] flooding and flow reversal model, which is applied to the flow conditions at the OPEN for LCO, as follows:

$$\left(v^*_{SL0}\right)^{0.5} + \left(v^*_{SG0}\right)^{0.5} = 1. \tag{22.1}$$

In 22.1, $v^*_{SG0}$ and $v^*_{SL0}$ are dimensionless superficial gas and superficial liquid velocities, respectively, which correspond to the operational envelope condition. The superficial gas velocity at the operational envelope $v^*_{SG0}$ is given by

$$v^*_{SG0} = v^{0.5}_{SG0_{\rho_G}} \left[gd\left(\rho_L - \rho_G\right)\right]^{-0.5}. \tag{22.2}$$

The expression for $v^*_{SL0}$ is modified to incorporate the dynamic ZNLH phenomenon, namely,

$$v^*_{SL0} = v^{2n}_{SL0}\, \rho^n_L \left[gd\left(\rho_L - \rho_G\right)\right]^{-n}, \tag{22.3}$$

where $v_{SL0}$ is the associated superficial liquid velocity for ZNLF conditions at the OPEN for LCO. The exponent $n$ is given by

$$n = \left(\frac{25\mu\sqrt{gd}}{\sigma}\right)^{0.25}, \tag{22.4}$$

where $\mu$ is the liquid viscosity, $g$ is the acceleration of gravity, $d$ is the pipe diameter, and $\sigma$ is the surface tension. Finally, the $H_{ZNL0}$ is calculated by

$$H_{ZNL0} = \frac{v_{SL0}}{v_{SL0} + v_{SG0}}. \tag{22.5}$$

Note that the superficial velocities $v_{SG0}$ and $v_{SL0}$, as well as $H_{ZNL0}$, correspond to the flow conditions just below the OPEN for LCO.

Part 2: For a given superficial gas velocity,$v^*_{SG0}$ , the ZNLH predicted by Karpurapu [Ka18a], $H_{ZNL0}$, represents the maximum ZNLH for the given superficial gas velocity and the corresponding superficial liquid velocity at the OPEN for LCO, namely, $v^*_{SL0}$. For these conditions, the ZNLH churn flow region occupies the entire height of the upper part of the GLCC©, $l_{TOP}$. For the determination of the ZNLH below the OPEN for LCO, $H_{ZNL}$, the following ZNLH variation function is defined

$$f\left(v^*_{SG}, v^*_{SL}\right) = \left(v^*_{SG}\right)^3 \left(v^*_{SL}\right), \tag{22.6}$$

where $v^*_{SL}$ and $v^*_{SG}$ are the dimensionless forms of the operating conditions, $v_{SL}$ and $v_{SG}$, as given by

$$v^*_{SL} = v^{2n}_{SL} \rho^n_L \left[gd\left(\rho_L - \rho_G\right)\right]^{-n} \tag{22.7}$$

$$v^*_{SG} = v_{SG} \rho^{0.5}_G \left[gd\left(\rho_L - \rho_G\right)\right]^{-0.5}. \tag{22.8}$$

Finally, the ZNLH at the normal operating conditions below the OPEN for LCO is determined by

$$\frac{H_{ZNL}}{H_{ZNL0}} = \frac{f\left(v^*_{SG}, v^*_{SL}\right)}{f\left(v^*_{SG0}, v^*_{SL0}\right)}, \tag{22.9}$$

and the associated churn region height is calculated similarly by

$$\frac{l_{CH}}{l_{TOP}} = \frac{f\left(v_{SG}^*, v_{SL}^*\right)}{f\left(v_{SG0}^*, v_{SL0}^*\right)}. \tag{22.10}$$

Summarizing, Eq. 22.1 through 22.10 enable the prediction of the variations of the ZNLH and the corresponding churn region height in the upper part of the GLCC© for normal operating conditions. The solution procedure is as follows:

1. Determine $v_{SG0}^*$ from 22.2 for the given superficial gas velocity, $v_{SG0}$.
2. Substitute the calculated value of $v_{SG0}^*$(from step 1) into 22.1 and solve for $v_{SL0}^*$. Note that $v_{SG0}^*$ and $v_{SL0}^*$ correspond to operational envelope flow conditions.
3. Determine the exponent $n$ from 22.4 and solve for $v_{SL0}$ from 22.3.
4. Determine $H_{ZNL0}$ from 22.5.
5. Determine $v_{SL}^*$ and $v_{SG}^*$ from 22.7 and 22.8 for the superficial liquid velocity, $v_{SL}$, and superficial gas velocity, $v_{SG}$, corresponding to the operating conditions.
6. Determine the ZNLH variation function for the operating conditions, namely, $v_{SG}$ and $v_{SL}$, using 22.6, which is the numerator of 22.9.
7. Repeat Step 6 for the operational envelope conditions, i.e., $v_{SG0}^*$ and $v_{SL0}^*$, which is the denominator of 22.9.
8. Determine the zero-net liquid holdup, $H_{ZNL}$ for the operating conditions from 22.9.
9. Determine the churn region height, $l_{CH}$, for the operating conditions from 22.10.

### *22.4.4 Comparison Study*

This section presents comparisons between the predictions of the developed extended correlation of the ZNLH and churn region height (under normal operating conditions below the OPEN for LCO) with the acquired experimental data. Figures 22.12 and 22.13 present comparisons between ZNLH correlation predictions and experimental data, respectively, for air–oil flow and air–water flows. As can be seen, the developed correlation captures well the exponential trend of the ZNLH for air–oil flow, as well as the linear trend of the ZNLH for air–water flow.

Comparison between the correlation predictions and the acquired experimental data for the churn region height for air–oil flow and air–water flow is presented in Figs. 22.14 and 22.15, respectively. Again, the developed correlation captures the physical phenomena and follows the trend of the data well.

The discrepancies between the experimental data and the correlation predictions for the ZNLH are less than 12.7% for air–oil flow and less than 3% for air–water flow.

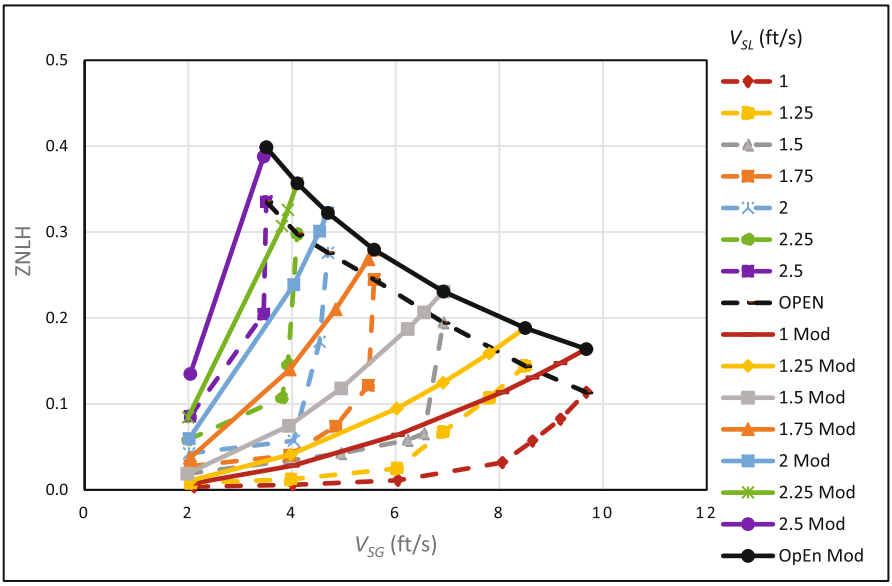



**Fig. 22.12** Comparison between ZNLH predictions and experimental data for air–oil flow

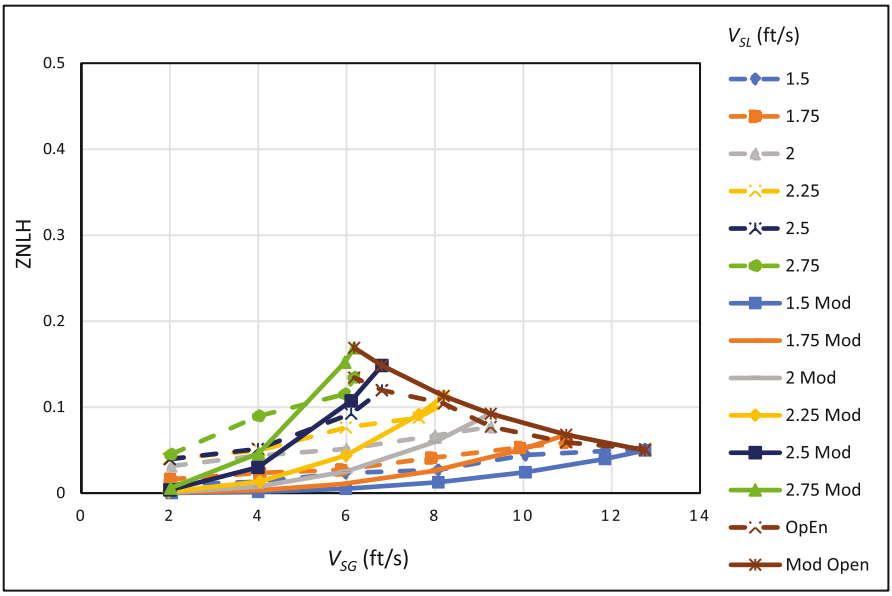



**Fig. 22.13** Comparison between ZNLH predictions and experimental data for air–water flow

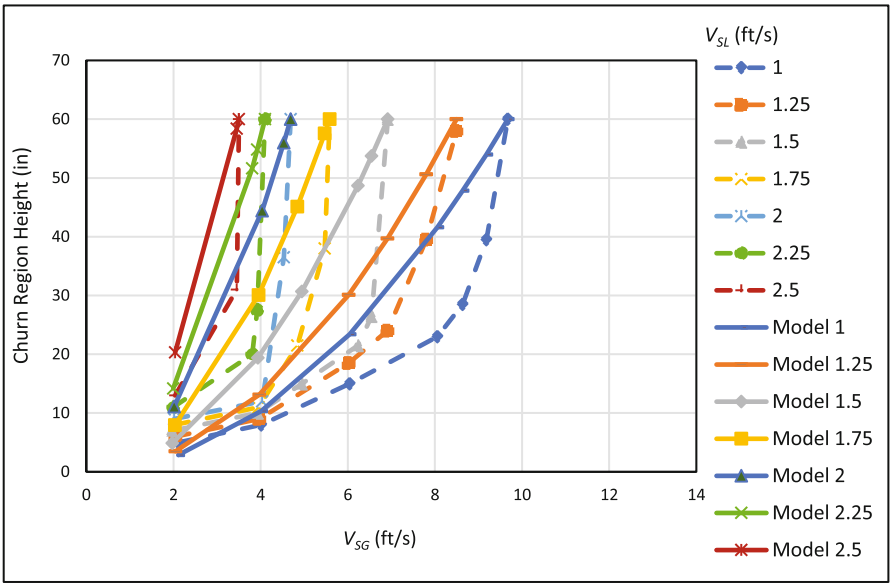



**Fig. 22.14** Comparison between churn region height predictions and experimental data for air–oil flow

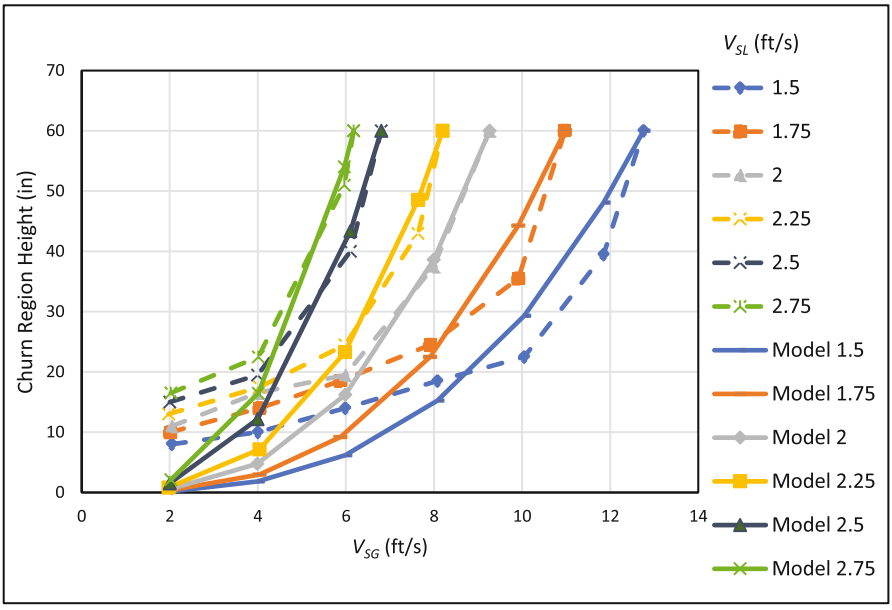



**Fig. 22.15** Comparison between churn region height predictions and experimental data for air–water flow

## References


[Ad06] Adebare, A.: Optimizing the efficiency of cylindrical cyclone gas/liquid separators for field applications. M.S. Thesis, Texas A&M University (2006)

[Ar96] Arpandi, I.A., Joshi A.R., Shoham, O., Shirazi, S., Kouba, G.E.: Hydrodynamics of two-phase flow in gas-liquid cylindrical cyclone separators. SPE J. **1 (4)**, 427–436 (1996)

[Ch00] Chirinos, W., Gomez, L., Wang, S., Mohan, R., Shoham, O., Kouba, G.: Liquid carry-over in gas-liquid cylindrical cyclone (GLCC©) compact separators. SPE J. **5**(3), 259–267 (2000)

[Du98] Duncan, R.W., Scott S.L.: Vertical zero net liquid flow: effects of high pressure on holdup. In: BHRG Multiphase Conference, vol. 31, pp. 43–60 (1998)

[Go98] Gomez, L.E.: A state-of-the-art simulator and field application design of gas-liquid cylindrical cyclone separators. M.S. Thesis, The University of Tulsa (1998)

[Go99] Gomez, L.E., Mohan, R.S., Shoham, O., Marrelli, J., Kouba, G.E.: Aspect ratio modeling and design procedure for GLCC© compact separators. J. Energy Resources Technol. **121**(1), 15–23 (1999)

[Go00] Gomez, L., Mohan, R., Shoham, O., Kouba, G.: Enhanced mechanistic model and field application design of gas-liquid cylindrical cyclone separator. SPE J. **5**(2), 190–198 (2000)

[Go01] Gomez, L.E.: Dispersed two-phase swirling flow characterization for predicting gas carry-under in gas-liquid cylindrical cyclone compact separators. Ph.D. Dissertation, The University of Tulsa (2001)

[Hy00] Hyoung, J.A., Langlinais J.P., Scott, S.L.: Effects of density and viscosity in vertical zero net liquid flow. J. Energy Resour. Technol. **122**, 49–55 (2000)

[Ka18a] Karpurapu, N.V.M.P.: Dynamic zero-net liquid holdup in a gas-liquid cylindrical cyclone (GLCC©) separator. M.S. Thesis, The University of Tulsa (2018)

[Ka18b] Karpurapu, N.V. M. P., Kolla, S.S., Mohan, R.S., Shoham, O.: Dynamic zero-net liquid holdup in gas-liquid cylindrical cyclones [conference presentation]. In: ASME 2018 5th Joint US-European Fluids Engineering Summer Conference, Montreal (2018)

[Ko19] Kolla, S.S., Mohan, R.S., Shoham, O.: A study on the effect of fluid properties and watercut on liquid carry-over in gas-liquid cylindrical cyclone compact separators. ASME J. Fluids Eng. **141**(9), 091303 (10pp.) (2019)

[Ko95] Kouba, G.E., Shoham, O., Shirazi, S.: Design and performance of gas-liquid cylindrical cyclone separators. In: BHR Group 7th International Meeting on Multiphase Flow, Cannes, France (1995)

[Li01] Liu, L., Scott, S.: Investigation on liquid holdup in vertical zero-net liquid flow. Chin. J. Chem. Eng. **9**(3), 284–290 (2001)

[Ma99] Mantilla, I., Shirazi, S., Shoham, O.: Flow field prediction and bubble trajectory model in GLCC© separators. ASME J. Energy Resour. Technol. **121**, 9–14 (1999)

[Mo99] Mohan, R., Shoham, O.: Technologies under development: design and development of gas-liquid cylindrical cyclone compact separators for three-phase flow. In: Oil and Gas Conference-Technology Options for Producers Survival, Co-Sponsored by DOE and PTTC, Dallas (1999)

[Mo00] Movafaghian, S., Jaua-Marturet, J., Mohan, R., Shoham, O., Kouba, K.: The effects of geometry, fluid properties and pressure on the hydrodynamics of gas-liquid cylindrical cyclone separators. Int. J. Multiphase Flow **26**(6), 999–1018 (2000)

[Sh98] Shoham, O., Kouba, G.: State-of-the-art of gas/liquid cylindrical-cyclone compact-separator technology. J. Pet. Technol. **50**(7), 58–65 (1998)

[Wa61] Wallis, G.B.: Flooding velocities for air and water in vertical tubes. The United Kingdom Atomic Energy Authority Report AEEW-R 123 (1961)

[Wa00a] Wang, S.: Dynamic simulation, experimental investigation and control system design of gas-liquid cylindrical cyclone separators. Ph.D. Dissertation, The University of Tulsa (2000)

[Wa00b] Wang, S., Mohan, R.S., Shoham, O., Marrelli, J.D., Kouba, G.E.: Performance improvement of gas-liquid cylindrical cyclone separators using integrated liquid level and pressure control systems. In: ASME Energy Sources Technology Conference and Exhibition, New Orleans (2000)

[Wa00c] Wang, S., Mohan, R., Shoham, O., Marrelli, J., Kouba, G.: Control system simulators for gas-liquid cylindrical cyclone separators. In: ASME Energy Sources Technology Conference and Exhibition, New Orleans (2000)

[Wa02] Wang, S., Gomez, L.E., Mohan, R.S., Shoham, O., Kouba, G.E.: High pressure testing of wet gas GLCC© separators. In: ASME Engineering Technology Conference on Energy, Houston (2002)

# Chapter 23
# On the Mono-Energetic Neutron Space Kinetics Equation in Cartesian Geometry: An Analytic Solution by a Spectral Method


**F. Tumelero, M. T. Vilhena, and B. E. J. Bodmann**

## 23.1 Introduction

The space kinetics equation for neutrons in multiplicative regimes is a crucial problem in nuclear reactor theory and control analysis. Since it is the scalar neutron flux that dominates the power response of a nuclear reactor, a model that is frequently used combines kinetic aspects with a diffusion approximation and provides as a solution the time evolution of the scalar neutron flux. A variety of control measures such as the control rod position or the chemical shim added to the moderating water represent effects that drive the spectral composition of the neutron population in a nuclear reactor. Hence, the accurate prediction of the flux behavior is an essential ingredient to attain operating efficiency and reactor safety.

Since the early beginnings of nuclear reactor operation, control in terms of kinetics was attributed to the presence of the delayed neutron precursors, usually taken into account by six groups with their characteristic time scales. Thus, from the numerical or arithmetic point of view, one has to provide algorithms that remain stable for time scales with differences of six orders in magnitude so that the problem is manifestly stiff. A review of the literature reveals that much effort has been made to solve the neutron kinetics equations [AbNa06, AbNa07, AbHa08, CoEtAl08, AbHa09, Qu10, NaEtAl12] to a numerical accuracy up to $10^{-14}$, which is even more accurate than the most precise experiment up to now (i.e., Mössbauer spectroscopy).

In numerical approaches for solving the equation system, the stiffness characteristic on the one hand imposes restrictions on the time step size and on the other hand sets limits for the maximum time interval that may be simulated by the solution. Recent years revealed in the literature a growing interest in determining

F. Tumelero (✉) · M. T. Vilhena · B. E. J. Bodmann
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: fernanda.tumelero@yahoo.com.br; marco.vilhena@ufrgs.br; bardo.bodmann@ufrgs.br



C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3_23

analytical representations of solutions since influences of the physical parameters on the solution may be directly determined, while numerical approaches need to compute the solution for each specific parameter choice.

In this line, the authors of reference [CeEtAl11] discussed an analytical solution of the multi-group neutron diffusion kinetics equation in a multilayered slab with six delayed neutron precursor groups, which they determine by the Generalized Integral Laplace Transform Technique (GILTT). In the previously cited work, the heterogeneous problem was cast into a set of recursive problems with constant parameters describing the physical properties of the domain and following the idea of Adomian's decomposition method [Ad94] for creating a solver. Based on a similar reasoning, the authors of reference [PeEtAl14] report on an analytical solution of the multi-group neutron kinetics diffusion equation in a homogeneous parallelepiped considering two-energy groups and six groups of delayed neutron precursors, by applying the Generalized Integral Transform Technique (GITT) in Cartesian coordinates. In another treatise [FeEtAl13a], the neutron diffusion equation for mono-energetic neutrons in cylinder geometry was solved analytically, assuming translational symmetry along the cylinder axis and using as an integral transform approach the Hankel transform in the radial coordinate $r$. A complementary work [FeEtAl13b] again using the Hankel transform applied to the diffusion equation for a homogeneous cylinder geometry considering two-energy groups together with one and six precursor concentrations. An approach to make progress with heterogeneous problems [CeEtAl15] presented the multi-region one-dimensional diffusion kinetics model with analytical expressions using a Taylor series, where the coefficients are found using the differential equation for a recursion relation together with the boundary and interface conditions, which establish the relations between the coefficients of the different regions.

## 23.2 What to Look For

More recently, a purely spectral method was applied to the spatial kinetics problem [OlEtAl17], where the solution was determined using the variable separation method for a case study in cylindrical geometry, mono-energetic neutrons, and one group of delayed neutron precursors. In a subsequent work [OlEtAl19], the same authors determined a general solution for two-energy groups and one delayed neutron precursor concentration for cylinder geometry. In a parallel approach [TuEtAl19], the space kinetics diffusion equation was solved in a one-dimensional geometry considering a homogeneous domain but for two-energy groups and six groups of delayed neutron precursors. In this chapter, a Taylor expansion was used in the space variable and allocating the time dependence to the expansion coefficients. Upon truncating the Taylor series at second order, a set of recursive systems of ordinary differential equations was obtained, where a modified decomposition method was applied to solve the equation system.

From comparison of the last decade's methods and their expediency with respect to obtaining analytical solutions, the Fourier method (also known as separation of variables) seemed promising for this type of problems in cylindrical geometry [OlEtAl17, OlEtAl19], where a heterogeneous problem along the cylinder axis could be solved exactly, considering the case of mono-energetic neutrons and one group of delayed neutron precursors. Here, heterogeneous problem means a composition of sectionally homogeneous domains each with varying nuclear parameters and their respective interfaces are located perpendicular to the symmetry axis. Due to the fact that the same method does not work for problems with inhomogeneities in concentric regions, i.e., the radial direction, the present study resorted to three-dimensional Cartesian geometry in order to review the inhomogeneous problem (sectionally homogeneous composition of the domain) using the method by separation of variables. Although such a solution has its relevance on its own right, it plays the role of an initialization in decomposition-method-based approaches [Ad94] to solve nonlinear reactor kinetics scenarios similar to those discussed in references [PeEtAl14, TuEtAl19]. Last but not least, the obtained results from the computational implementation of the algorithm associated with the obtained solution were compared against findings in the literature.

## 23.3 Model and Methodology

In this section, we present the methodology that leads to the analytical solution of the neutron space kinetics equation ((23.1) and (23.2)) assuming mono-energetic neutrons and one group of delayed neutron precursors in a three-dimensional Cartesian geometry. The found solution is then applied to a homogeneous as well as a nonhomogeneous domain setup, representing some properties that characterize some of the dynamics in nuclear reactor cores.

$$\frac{1}{u}\frac{\partial}{\partial t}\Phi(\mathbf{r},t) = D\nabla^2\Phi(\mathbf{r},t) - \Sigma_a\Phi(\mathbf{r},t) + (1-\beta)\nu\Sigma_f\Phi(\mathbf{r},t) + \lambda C(\mathbf{r},t), \quad (23.1)$$

$$\frac{\partial}{\partial t}C(\mathbf{r},t) = \beta\nu\Sigma_f\Phi(\mathbf{r},t) - \lambda C(\mathbf{r},t). \quad (23.2)$$

Here, $\Phi$ denotes the scalar neutron flux, $C$ is the delayed neutron precursor concentration, $u$ is the neutron velocity, $D$ is the diffusion coefficient, $\Sigma_a$ is the macroscopic absorption cross section, $\Sigma_f$ is the macroscopic fission cross section, $\nu$ is the average number of neutrons emitted by fission, $\beta$ is the delayed neutron fraction, and the $\lambda$ is the delayed neutron decay constant.

The system of Eqs. (23.1) and (23.2) is subject to the initial conditions given by the solution of the stationary problem (i.e., time-independent solution)

$$\Phi(\mathbf{r}, 0) = \Phi_0(\mathbf{r}) \,, \tag{23.3}$$

$$C(\mathbf{r}, 0) = \frac{\beta}{\lambda} \nu \Sigma_f \Phi_0(\mathbf{r}) \,, \tag{23.4}$$

and the boundary conditions are homogeneous current density and scalar flux conditions on the boundaries,

$$\frac{\partial}{\partial x}\Phi(0, y, z, t) = 0 \,, \qquad \Phi(L_x, y, z, t) = 0 \,,$$

$$\frac{\partial}{\partial y}\Phi(x, 0, z, t) = 0 \,, \qquad \Phi(x, L_y, z, t) = 0 \,,$$

$$\frac{\partial}{\partial z}\Phi(x, y, 0, t) = 0 \,, \qquad \Phi(x, y, L_z, t) = 0 \,.$$

A comment is in order here; from the formal point of view, the above conditions define the boundary conditions of the mathematical problem, whereas strictly speaking the homogeneous conditions of the derivative terms are physical symmetry conditions because the solution is constructed considering only one-eighth of the original domain.

For convenience and to end up with rather compact equations, we introduce in Eqs. (23.1) and (23.2) the following shorthand notations:

$$\mathscr{A} = uD \,,$$

$$\mathscr{B} = u\left((1-\beta)\nu\Sigma_f - \Sigma_a\right) \,,$$

$$\mathscr{D} = u\lambda \,,$$

$$\mathscr{E} = \beta\nu\Sigma_f \,,$$

$$\mathscr{H} = -\lambda \,,$$

$$\mathscr{P} = -\frac{\mathscr{E}}{\mathscr{H}} \,.$$

Then, the preparation for the application of variable separation is based on some propositions. We assume for every point $\mathbf{r}$ in the domain $\Omega \in \mathbb{R}^3$ that there exist functions $R_1 : \Omega \to \mathbb{R}$, $R_2 : \Omega \to \mathbb{R}$, $T_1 : [0, +\infty) \to \mathbb{R}$ and $T_2 : [0, +\infty) \to \mathbb{R}$ such that $\Phi(\mathbf{r}, t)$ and $C(\mathbf{r}, t)$ can be expressed as

$$\Phi(\mathbf{r}, t) = R_1(\mathbf{r})T_1(t) \,, \tag{23.5}$$

$$C(\mathbf{r}, t) = R_2(\mathbf{r})T_2(t) \,. \tag{23.6}$$

Upon replacing the expressions (23.5) and (23.6) into Eqs. (23.1) and (23.2), further making use of the shorthands, one gets

$$R_1(\mathbf{r})\frac{d}{dt}T_1(t) = T_1(t)\left(\mathscr{A}\nabla^2 R_1(\mathbf{r}) + \mathscr{B}R_1(\mathbf{r})\right) + \mathscr{D}R_2(\mathbf{r})T_2(t)\ ,$$
$$R_2(\mathbf{r})\frac{d}{dt}T_2(t) = \mathscr{E}R_1(\mathbf{r})T_1(t) + \mathscr{H}R_2(\mathbf{r})T_2(t)\ .$$

After some algebraic simplifications and arranging the space- and time-dependent variables on the respective left- and right-hand sides in the equations, one may proceed with the idea of separating the contributions from the different spatial and time dimensions. Moreover, assuming that there exist constants that physically make sense, the following equations arise:

$$R_2 = \frac{\mathscr{E}}{\sigma_1}R_1\ , \tag{23.7}$$
$$\frac{d}{dt}T_2 = \sigma_1 T_1 + \mathscr{H}T_2\ , \tag{23.8}$$
$$\nabla^2 R_1 + \frac{\mathscr{B}-\sigma_2}{\mathscr{A}}R_1 = 0\ , \tag{23.9}$$
$$\frac{d}{dt}T_1 = \sigma_2 T_1 + \frac{\mathscr{D}\mathscr{E}}{\sigma_1}T_2\ . \tag{23.10}$$

Here, $\sigma_1$ and $\sigma_2$ are the separation constants. The task is now to determine the solutions of the Eqs. (23.7), (23.8), (23.9), and (23.10) by studying the possible values of the new constants $\sigma_1$ and $\sigma_2$, i.e., the spectrum that contributes to the total solution. It is noteworthy that the spatial functions $R_1$ and $R2$ are related by physical parameters and the separation constant, whereas $T_1$ and $T_2$ set up a differential equation system to be solved. This was to be expected since the differential equation for the neutron precursors (23.2) contains only the time derivative, so that the spatial part is controlled by the partial differential equation for the scalar neutron flux (23.1).

The spatial part of the scalar neutron flux may be obtained by solving Eq. (23.9) upon substituting $R_1(x,y,z) = X(x)Y(y)Z(z)$ and successively applying the separation of variables method for the spatial dimensions.

$$Y(y)Z(z)\frac{\partial^2}{\partial x^2}X(x) + X(x)Z(z)\frac{\partial^2}{\partial y^2}Y(y) + X(x)Y(y)\frac{\partial^2}{\partial z^2}Z(z)$$
$$+\left(\frac{\mathscr{B}-\sigma_2}{\mathscr{A}}\right)X(x)Y(y)Z(z) = 0. \tag{23.11}$$

Upon dividing Eq. (23.11) by $X(x)Y(y)Z(z)$ and rearranging terms with the corresponding variable dependence on either side of the equation, so that they may

be separated by a constant one, arrive at three ordinary differential equations with known solution.

$$\frac{d^2}{dx^2}X(x) + \left[\sigma_3 + \frac{\mathscr{B} - \sigma_2}{\mathscr{A}} - \sigma_4\right] X(x) = 0, \tag{23.12}$$

$$\frac{d^2}{dy^2}Y(y) + \sigma_4 Y(y) = 0, \tag{23.13}$$

$$\frac{d^2}{dz^2}Z(z) - \sigma_3 Z(z) = 0. \tag{23.14}$$

Here, $\sigma_3$ and $\sigma_4$ are the separation constants. Using the original boundary conditions, one may separate the part of the conditions relevant for each spatial dimension.

$$\frac{d}{dx}X(0) = 0\,, \qquad X(L_x) = 0\,, \tag{23.15}$$

$$\frac{d}{dy}Y(0) = 0\,, \qquad Y(L_y) = 0\,, \tag{23.16}$$

$$\frac{d}{dz}Z(0) = 0\,, \qquad Z(L_z) = 0\,. \tag{23.17}$$

The next step is then to look for the eigenvalues for each of the Eqs. (23.12)–(23.14). To this end, these equations are solved for the set of boundary conditions (23.15)–(23.17), analyzing whether the constants that multiply the separable functions shall be less than, greater than, or equal to zero. Thus, the fundamental solutions for the spatial variables are generically given by

$$R_1^{(n,m,l)}(x,y,z) = b_{n,m,l}\cos\left(\frac{(2n-1)\pi}{2L_x}x\right)\cos\left(\frac{(2m-1)\pi}{2L_y}y\right) \times \cos\left(\frac{(2l-1)\pi}{2L_z}z\right), \tag{23.18}$$

and here $b_{n,m,l}$ represents the set of arbitrary constants to be determined from the initial condition. From the eigenvalue problem in the $z$ direction $\sigma_3 = -\xi_3^2 < 0$, where $\xi_3 = \frac{(2l-1)\pi}{2L_z}$, which are the eigenvalues associated with $l = 1, 2, \ldots$. Equivalently, one gets for the eigenvalue problem in the $y$ direction $\sigma_4 = \xi_4^2 > 0$, here with $\xi_4 = \frac{(2m-1)\pi}{2L_y}$, which are the eigenvalues associated with $m = 1, 2, \ldots$. The last spatial dimension has eigenvalues given by $\left(\sigma_3 + \frac{\mathscr{B}-\sigma_2}{\mathscr{A}} - \sigma_4\right) = \xi_2^2 > 0$, where $\xi_2 = \frac{(2n-1)\pi}{2L_x}$ are the eigenvalues associated with $n = 1, 2, \ldots$. Further, the

separation constant $\sigma_2$ depends on the previously determined spectra and is given by

$$\sigma_2 = \mathscr{B} - \mathscr{A}\left(\left(\frac{(2m-1)\pi}{2L_y}\right)^2 + \left(\frac{(2n-1)\pi}{2L_x}\right)^2 + \left(\frac{(2l-1)\pi}{2L_z}\right)^2\right) .$$

This determines completely the spatial amplitudes to the general solution and moreover may be used to also represent the solution of the stationary problem, i.e., the initial condition by setting $\sigma_1 = \lambda$ and $\sigma_2 = -u\beta\nu\Sigma_f$. It remains to solve the equation system for the time evolutions of the scalar neutron flux and the delayed neutron precursors. To this end, the ordinary differential equation system for $T_1^{(n,m,l)}(t)$ and $T_2^{(n,m,l)}(t)$ (Eqs. (23.8) and (23.10)) with the modes specified by *nml* was solved.

$$\frac{d}{dt}\begin{pmatrix} T_1(t) \\ T_2(t) \end{pmatrix} - \begin{pmatrix} \sigma_2 & \frac{\mathscr{D}\mathscr{E}}{\sigma_1} \\ \sigma_1 & \mathscr{H} \end{pmatrix}\begin{pmatrix} T_1(t) \\ T_2(t) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} . \tag{23.19}$$

Considering that the nuclear parameters are all positive, then the discriminant of the characteristic equation resulting from solving the problem (23.19) for both functions $T_1(t)$ and $T_2(t)$ is always positive, so that this characteristic equation has two distinct real roots. Then, applying the elimination method, one finds an ordinary differential equation for $T_2$,

$$\frac{d^2}{dt^2}T_2 - (\mathscr{H} + \sigma_2)\frac{d}{dt}T_2 + (\sigma_2\mathscr{H} - \mathscr{D}\mathscr{E})T_2 = 0,$$

and the following expression for the discriminant of the characteristic equation,

$$\Delta = (\mathscr{H} - \sigma_2)^2 + 4\mathscr{D}\mathscr{E} > 0 .$$

This guarantees the existence of the solution by exponential functions for the time-dependent amplitudes,

$$T_1(t) = c_1 \exp(\Lambda_1 t) + c_2 \exp(\Lambda_2 t)$$

and

$$T_2(t) = \sigma_1\left(\frac{c_1 \exp(\Lambda_1 t)}{\Lambda_1 - \mathscr{H}} + \frac{c_2 \exp(\Lambda_2 t)}{\Lambda_2 - \mathscr{H}}\right) .$$

Here, $c_1$ and $c_2$ are the arbitrary constants, and $\Lambda_1$ and $\Lambda_2$ are defined as

$$\Lambda_1 = \frac{1}{2}\left(\mathscr{H} + \sigma_2 + \sqrt{(\mathscr{H} + \sigma_2)^2 - 4\,(\sigma_2\mathscr{H} - \mathscr{D}\mathscr{E})}\right) ,$$

$$\Lambda_2 = \frac{1}{2}\left(\mathscr{H} + \sigma_2 - \sqrt{(\mathscr{H} + \sigma_2)^2 - 4\,(\sigma_2\mathscr{H} - \mathscr{D}\mathscr{E})}\right) .$$

Recalling that $R_1$ and $R_2$ are related by a scaling factor only, we can compose the general solution by the superposition of all the modes.

$$\Phi(\mathbf{r}, t) = \sum_{n=1}^{\infty}\sum_{l=1}^{\infty}\left(d_{n,m,l}\left(\exp(\Lambda_1 t) + \zeta_{n,m,l}\exp(\Lambda_2 t)\right)\right)$$
$$\cos\left(\frac{(2n-1)\pi}{2L_x}x\right)\cos\left(\frac{(2m-1)\pi}{2L_y}y\right)\cos\left(\frac{(2l-1)\pi}{2L_z}z\right) \tag{23.20}$$

$$C(\mathbf{r}, t) = \sum_{n=1}^{\infty}\sum_{l=1}^{\infty} E\left(d_{n,m,l}\left(\frac{\exp(\Lambda_1 t)}{\Lambda_1 - \mathscr{H}} + \zeta_{n,m,l}\frac{\exp(\Lambda_2 t)}{\Lambda_2 - \mathscr{H}}\right)\right)$$
$$\cos\left(\frac{(2n-1)\pi}{2L_x}x\right)\cos\left(\frac{(2m-1)\pi}{2L_y}y\right)\cos\left(\frac{(2l-1)\pi}{2L_z}z\right). \tag{23.21}$$

Here, the arbitrary constants $d_{n,m,l}$, related to the function $\Phi_0$, are determined by integrating over the spatial domain

$$d_{n,m,l} = \mathscr{D}_{n,m,l}\int_0^{L_z}\int_0^{L_y}\int_0^{L_x}\Phi_0(x, y)\cos\left(\frac{(2n-1)\pi}{2L_x}x\right)\cos\left(\frac{(2m-1)\pi}{2L_y}y\right)$$
$$\cos\left(\frac{(2l-1)\pi}{2L_z}z\right)\,dx\,dy\,dz ,$$

with

$$\mathscr{D}_{n,m,l} = \frac{8}{L_x L_y L_z(1 + \zeta_{n,m,l})} ,$$

and

$$\zeta_{n,m,l} = \frac{\mathscr{P}(\Lambda_2 - \mathscr{H})(\Lambda_1 - \mathscr{H}) - \mathscr{E}(\Lambda_2 - \mathscr{H})}{\mathscr{E}(\Lambda_1 - \mathscr{H}) - \mathscr{P}(\Lambda_2 - \mathscr{H})(\Lambda_1 - \mathscr{H})} ,$$

and where $\zeta_{n,m,l}$ is found from the initial condition of the delayed neutron precursor (Eq. (23.4)). Thus we end up with a unique and exact solution (Eqs. (23.20) and (23.21)), which is ready for applications except for the question of truncation for the purpose of computing numerical solution as presented next.

## 23.4 Results

In the further, we present two applications, the first one for a simple problem with a globally homogeneous domain; in other words, the nuclear parameters are time-independent and the same everywhere in the domain. The second application shows a case where the nuclear parameters for two subdomains are different and are changed stepwise in the vertical direction such as to mimic an action due to the movement of control rods. However, we do not consider an explicit time dependence in the interface between the two subdomains that is reasonable since the time scale of control actions is typically larger by roughly two orders of magnitude in comparison to measurable changes in the neutron flux.

### *23.4.1 Homogeneous Case*

This section presents the numerical results for the methodology implementation presented in the previous section. The homogeneous domain was defined in three-dimensional Cartesian geometry, where mono-energetic neutrons and a group of delayed neutron precursors were considered. According to the boundary conditions applied to find the unique solution presented in Eqs. (23.20) and (23.21), the solution for a quarter of a multiplicative medium with the shape of a cube is presented, considering the edge lengths of $L_x = L_y = L_z = 10\,cm$. For the numerical simulation, we used the nuclear parameter set provided in Table 23.1, further the fraction of delayed neutron precursors was as usual for uranium-235 $\beta = 0.0065$, and the mean decay constant of the precursors was $\lambda = 0.08\,\mathrm{s}^{-1}$. Figure 23.1 presents the scalar neutron flux as a function of $x$ for a time progression with $t = 1, 5, 15$ and $30\,\mathrm{s}$ with $y = 0\,\mathrm{cm}$ and $z = 4\,\mathrm{cm}$. The flux behavior tends to decrease with increasing time that characterizes the setup as sub-critical.

In Fig. 23.2a, the spatial profile of the neutron flux in the $x$-$y$ plane is shown for $t = 1\,\mathrm{s}$ and $z = 4\,\mathrm{cm}$. As was to be expected by the geometry specification, one observes a symmetry of the distribution under exchange of $x$ and $y$. Trivially, the same is true for exchanges of $x$ or $y$ with $z$, which is not shown here. As an example for the time evolution of the scalar neutron flux, Fig. 23.2b shows the flux along the vertical axis $z$ in the center of the multiplicative medium ($x = 0$ and $y = 0$). The time evolution of the flux shows the exponential attenuation behavior according to the solution for the found time function $T_1$.

**Table 23.1** Nuclear parameters

| $D$ (cm) | $u$ (cm/s) | $\Sigma_a$ (1/cm) | $\nu_g \Sigma_f$ (1/cm) |
|---|---|---|---|
| 0.96343 | $1.1035 \times 10^7$ | $1.5843 \times 10^{-2}$ | $3.3303 \times 10^{-2}$ |

Source: Adapted from [OlEtAl17]

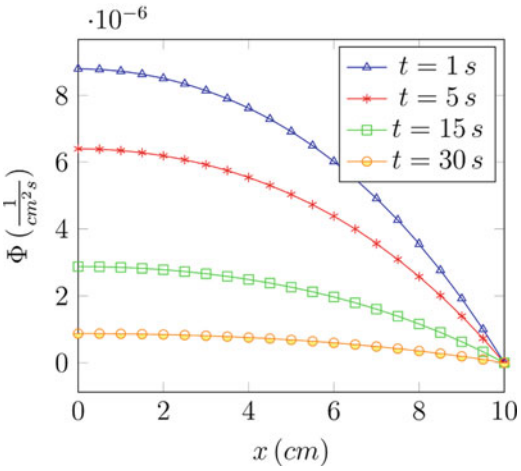



**Fig. 23.1** Neutron flux as a function of $x$ for several time instants

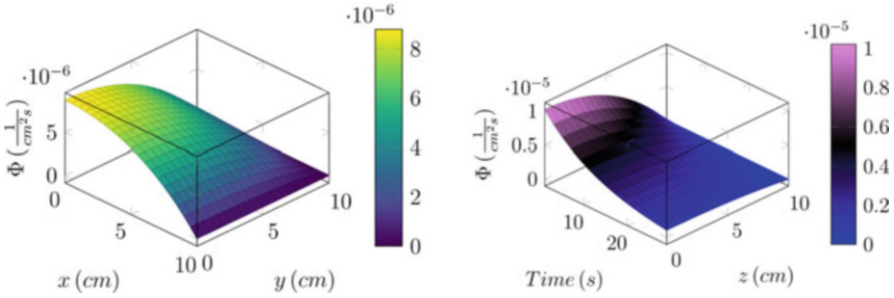



**Fig. 23.2** Neutron flux (**a**) in the $x$-$y$ plane for $t = 1\ s$ and $z = 4\ cm$, and (**b**) its time evolution in the center of the multiplicative medium ($x = y = 0$) along the vertical axes and $z$

A comparison of the findings of reference [OlEtAl17] shows compatibility of the two solutions. Recalling that the solution from the literature was obtained for a problem in a domain with cylinder geometry so that these have a curved boundary only, the time evolution in the center of the domain and along the $z$ axis coincides except for an arbitrary scale, which is due to scale invariance of both problems with homogeneous boundary/symmetry conditions.

### *23.4.2 Heterogeneous Case*

In this section, the previous problem was modified, and two subdomains with different nuclear parameters were considered maintaining the global geometry of the domain ($L_x = L_y = L_z = 10\,\text{cm}$) as well as the fraction of delayed neutron

precursors ($\beta = 0.0065$) and their characteristic decay constant $\lambda = 0.08\,s^{-1}$. Further, solutions with the interface located at positions $z = 2, 4, 6, 8\,\text{cm}$ are evaluated, which could mimic the instantaneous scalar neutron fluxes for a process where neutron absorbers are inserted into the multiplicative medium similar to the insertion of control rods in a nuclear reactor core. As already justified before, these results may be considered an acceptable approximation since the time scales for the insertion process ($\sim 10^1$ s) are typically considerably larger than the changes in the neutron flux ($\sim 10^{-2}$ s); however, more precise values depend on the reactor type and generation. The physical effect of such a change is a decrease of the scalar neutron flux in the region with the larger absorption cross section. Comparable results are reported in the study of [OlEtAl17] but for cylindrical geometry. In the further, numerical solutions are presented, where continuity of the flux and the current density at the interfaces were the matching criterion of the fluxes in the two subdomains. Note that our solution for Cartesian geometry works as a solution for more complex cases such as subdomain interfaces in $x$–$z$ or $y$–$z$ planes, whereas the solution of reference [OlEtAl17] is valid only in subdomain interfaces in planes perpendicular to the $z$ axis.

The idea to construct the global solution in the whole domain by solutions for each subdomain is based on the scale invariance property of the space kinetics equation and the homogeneous boundary/symmetry conditions. The fact that nonhomogeneous boundary conditions restrict scale invariance is then used at the interface and is where the fluxes and current densities are in general not vanishing. The formal steps to construct the solution are as follows. The structure of the obtained solution as obtained in Eq. (23.20) is valid for each of the subdomains except for the spectrum (separation constants) and consequently the integration constants to be determined, which is a direct consequence of the numerical differences in the nuclear parameters. As intermediate solutions, each subdomain is solved as an independent problem where the subdomain containing the origin imposes a restriction by a vanishing current density (symmetry condition), while at the boundary the vanishing flux restricts the solution in this subdomain. Local scale invariance is now explicitly broken in order to match the solutions and preserve the continuity equation, i.e., neutrons leaving one subdomain shall enter the second subdomain and vice versa. It is noteworthy that a global scale invariance still remains which is frequently fixed upon imposing a specific power for the considered nuclear medium and volume.

More specifically, the solution is valid for each segment, that is, in the upper cell $z \in [(1-\kappa)L_z, L_z]$ or the lower cell $z \in [0, (1-\kappa)L_z]$, where $\kappa \in [0, 1]$ specifies the location of the interface between the subdomains. The solutions in different cells (or regions of the reactor) have invariance under a global scale transformation so that for instance fixing the scale at the origin allows to match the solutions while preserving the continuity equation by adjusting the scale of the solution in the second domain. Thus, considering two distinct regions, the solution of the scalar neutron flux can be written as shown below.

**Table 23.2** Nuclear parameters for the heterogeneous case

| | $D$ (cm) | $u$ (cm/s) | $\Sigma_a$ (1/cm) | $\nu_g \Sigma_f$ (1/cm) |
|---|---|---|---|---|
| Region 1 | 0.96343 | $1.1035 \times 10^7$ | $1.5843 \times 10^{-2}$ | $3.3303 \times 10^{-2}$ |
| Region 2 | 0.96343 | $1.1035 \times 10^7$ | 0.115843 | $3.3303 \times 10^{-2}$ |

Source: Adapted from [OlEtAl17]

$$\Phi(x, y, z, t) = \begin{cases} \Phi^{[1]}(x, y, z, t), & \text{if } 0 \leq z \leq (1-\kappa)L_z \\ \Phi^{[2]}(x, y, z, t), & \text{if } (1-\kappa)L_z \leq z \leq L_z \end{cases}$$

The aforementioned scale degrees of freedom are exploited to satisfy the following interface conditions, i.e., continuity of the scalar neutron flux and continuity of the current density across the interface.

$$\Phi^{[1]}(x, y, (1-\kappa)L_z, t) = \Phi^{[2]}(x, y, (1-\kappa)L_z, t)$$
$$D^{[1]} \frac{\partial}{\partial z} \Phi^{[1]}(x, y, (1-\kappa)L_z, t) = D^{[2]} \frac{\partial}{\partial z} \Phi^{[2]}(x, y, (1-\kappa)L_z, t).$$

To perform the simulation, the nuclear parameters presented in Table 23.2 were used, where a larger macroscopic absorption cross section in region 2 was considered, representing the effect of neutron absorption by an operational control measure. The insertion of the absorber was evaluated for different positions from the top to the bottom in the considered geometry. The effect of the interface position on the scalar neutron flux is shown in Fig. 23.3a–d, where the time was frozen to $t = 1\,\text{s}$, and the projection into the $x$–$z$ plane is depicted for the positions $z = 8, 6, 4$ and $2\,\text{cm}$, respectively. Due to the symmetry under exchange of $x$ and $y$, the respective projection has the same appearance as the one displayed here. In the case of $z = 8\,\text{cm}$, the effect of the differences in the physical parameters for the subdomain above and below is hardly recognizable, which is partially due to the fact that at $z = 10\,\text{cm}$ we have homogeneous boundary conditions so that this region of the flux naturally has an accentuated attenuation. Nevertheless, for the interface positions at $z = 6\,\text{cm}$ down to $z = 2\,\text{cm}$, one observes a drop to lower fluxes with increasing $z$, which by numerical inspection shows that the following limits are complied.

$$\lim_{\epsilon \to 0} \Phi(x, y, z - \epsilon, t) = \lim_{\epsilon \to 0} \Phi(x, y, z + \epsilon, t).$$

The continuity of the current density was verified numerically and showed up to small differences from numerical imprecision a fairly well agreement for some instants, namely, $t = 1\,s, 5\,s, 10\,s$ and $15\,s$, respectively.

$$\mathbf{j}_1 = \lim_{\epsilon \to 0} D^{[1]} \nabla \Phi(x, y, z - \epsilon, t) = \lim_{\epsilon \to 0} D^{[2]} \nabla \Phi(x, y, z + \epsilon, t) = \mathbf{j}_2.$$

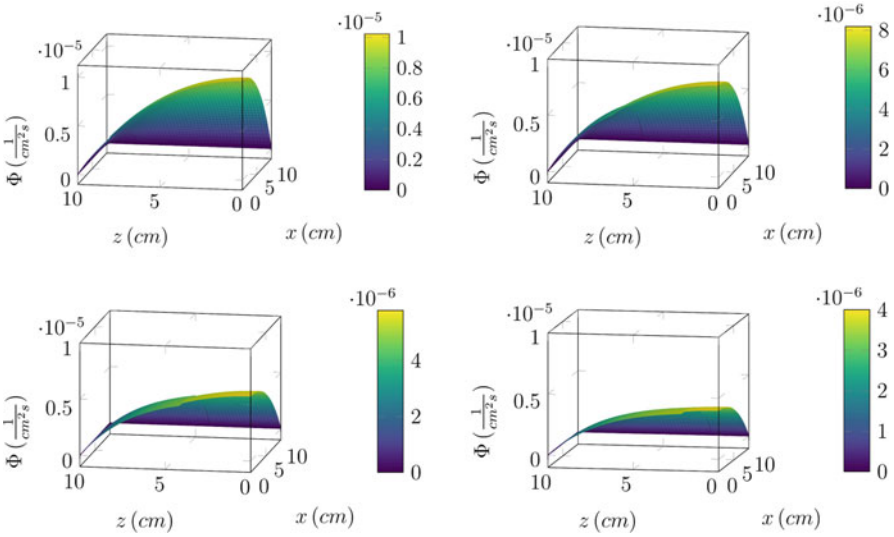



**Fig. 23.3** Scalar neutron flux as a function of $x$ and $z$ with interface between the two subdomains at (**a**) $z = 8\,\text{cm}$, (**b**) $z = 6\,\text{cm}$, (**c**) $z = 4\,\text{cm}$, and (**d**) $z = 2\,\text{cm}$

**Table 23.3** Numerical values for the current densities $\mathbf{j}_1$ and $\mathbf{j}_2$, for different times and with $x = 0$, $y = 0$ and considering the interface at $z = 8$

| Time | $\lvert\mathbf{j}_1\rvert$ | $\lvert\mathbf{j}_2\rvert$ |
|---|---|---|
| $1\,s$ | $1.6480 \times 10^{-6}$ | $1.6480 \times 10^{-6}$ |
| $5\,s$ | $1.1983 \times 10^{-6}$ | $1.1973 \times 10^{-6}$ |
| $10\,s$ | $8.0460 \times 10^{-7}$ | $8.0308 \times 10^{-7}$ |
| $15\,s$ | $5.4025 \times 10^{-7}$ | $5.3866 \times 10^{-7}$ |

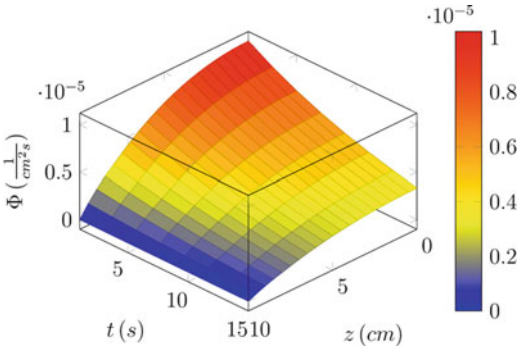



**Fig. 23.4** Time evolution of the scalar neutron flux along the $z$ axis with interface located at $z = 8\,\text{cm}$

One example is shown for the interface position $z = 8\,\text{cm}$ and with $x = y = 0\,\text{cm}$, where the numerical findings are presented in Table 23.3.

Figure 23.4 shows the time evolution of the scalar neutron flux along the $z$ axis and considering $x = y = 0$. Once again, one observes the same exponential atten-

uation with increasing time as in the homogeneous case, due to the characteristic of the nuclear parameters that characterize the problem as sub-critical, due to the increased absorption cross section in the upper subdomain.

## 23.5 Conclusion

In this chapter, we presented an analytical solution for the space kinetics equation for three-dimensional Cartesian geometry. The method of variable separation allowed to determine an in-principal exact solution except for the truncation of the spectrum associated to each spatial dimension, which is necessary for numerical applications. So far, we focused on the solution and left open the pertinent question of truncation and its implication on the precision of the solution. However, successive increase in the number of terms and their contribution to changes in comparison to the previous solution may give a hint as to how many terms will result close to the exact solution. Evidently, this is not a proof of convergence but indicates the stability of the solution and is a necessary (but not sufficient) condition for convergence.

The general solution was applied in two cases, one with a homogeneous domain and a second one where the domain was divided into two subdomains each with specific but different nuclear parameters. The latter problem is relevant for calculating the nuclear reactor core with its highly heterogeneous structure. If one considers each nuclear element as one region, one ends up with $\sim 10^2$ regions that define the domain analogue to the reactor core. For these cases, it is interesting to have an analytical solution that allows at least to simplify the optimization of the core assembly such that the resulting scalar neutron flux is as "homogeneous" as possible, which implies a "uniform" burn-up.

One of the features of the dynamical equation is scale invariance, which we exploited to determine in principle the solution in each individual subdomain. Moreover, initial and homogeneous boundary conditions maintain global scale invariance of the solution of the considered space kinetics problem. The scale degree of freedom opens up a pathway to construct the global solution fixing the scales of the solution in one subdomain with the neighboring ones using the same expressions but with differences in the physical parameter for each region. Once the appropriate spectrum is used from the nuclear parameters, the remaining integration constants are fixed by the initial and interface conditions. Thus, analytical procedures may be performed to investigate the influence of changes in the core assembly on the global profile of the solution.

Evidently, so far there are still several simplifications that have to be improved in future works, such as to increase the number of energy groups for neutrons as well as the number of groups of delayed neutron precursors. There is also a need to test the solution for critical and super-critical cases and compare these to findings in the literature. Another relevant issue is considering nuclear parameters with time

dependence using the idea presented in reference [TuEtAl19], which makes use of a modified decomposition method, where the solution presented in this chapter would serve as an initialization in a set of recursive equations. Solutions from this kind of approach could then substitute at least some of the benchmarks, which are so far exclusively generated by numerical approaches.

## References

[AbNa06] Aboanber, A.E., Nahla, A.A.: Solution of two-dimensional space-time multigroup reactor kinetics equations by generalized Padé and cut–product approximations. Ann. Nucl. Energy **33**, 209–222 (2006)

[AbNa07] Aboanber, A.E., Nahla, A.A.: Adaptive matrix formation AMF method of spacetime multigroup reactor kinetics equations in multidimensional model. Ann. Nucl. Energy **34**, 103–119 (2007)

[AbHa08] Aboanber, A.E., Hamada, Y.M.: Generalized Runge-Kutta method for two- and three-dimensional space-time diffusion equations with a variable time step. Ann. Nucl. Energy **35**, 1024–1040 (2008)

[AbHa09] Aboanber, A.E., Hamada, Y.M.: Computation accuracy and efficiency of a power series analytic method for two-and three- space-dependent transient problems. Prog. Nucl. Energy **51**, 451–464 (2009)

[Ad94] Adomian, G.: Solving Frontier Problems of Physics: The Decomposition Method. Kluwer Academic Publishers, Dordrecht (1994)

[CeEtAl11] Ceolin, C., Vilhena, M.T., Bodmann, B.E.J., Alvim, A.C.M.: On the analytical solution of the multi group neutron diffusion kinetic equation in a multilayered slab. In: Proceedings of the International Nuclear Atlantic Conference – INAC 2011, Belo Horizonte (2011)

[CeEtAl15] Ceolin, C., Schramm, M., Vilhena, M.T., Bodmann, B.E.J.: On the neutron multi-group kinetic diffusion equation in a heterogeneous slab: an exact solution on a finite set of discrete points. Ann. Nucl. Energy **76**, 271–282 (2015)

[CoEtAl08] Corno, S.E., Dulla, S., Picca, P., Ravetto, R.: Analytical approach to the neutron kinetics of the non-homogeneous reactor. Ann. Nucl. Energy **50**, 847–865 (2008)

[FeEtAl13a] Fernandes, J.C.L., Vilhena, M.T., Bodmann, B.E.J., Borges, V.: On the build-up factor from the multi-group neutron diffusion equation with cylindrical symmetry. World J. Nucl. Sci. Technol. **3**, 1–5 (2013).

[FeEtAl13b] Fernandes, J.C.L., Vilhena, M.T., Bodmann, B.E.J.: On a comparative analysis of the solutions of the kinetic neutron diffusion equation by the Hankel transform formalism and the spectral method. Prog. Nucl. Energy **69**, 71–76 (2013)

[NaEtAl12] Nahla, A.A., Al-Malki, F.A., Rokaya, M.: Numerical techniques for the neutron diffusion equations in the nuclear reactors. Adv. Stud. Theor. Phys. **6**, 640–664 (2012)

[OlEtAl17] Oliveira, F.R., Bodmann, B.E.J., Vilhena, M.T., Carvalho, F.: On an analytical formulation for the mono-energetic neutron space-kinetic equation in full cylinder symmetry. Ann. Nucl. Energy **99**, 253–257 (2017)

[OlEtAl19] Oliveira, F.R., Fernandes, J.C.L., Bodmann, B.E.J., Vilhena, M.T.: On an analytical solution for the two energy group neutron space-kinetic equation in heterogeneous cylindrical geometry. Ann. Nucl. Energy **133**, 216–220 (2019)

[PeEtAl14] Petersen, C.Z., Bodmann, B.E.J., Vilhena, M.T., Barros, R.C.: Recursive solutions for multi-group neutron kinetics diffusion equations in homogeneous three dimensional rectangular domains with time dependent perturbations. Kerntechnik **79**, 494–499 (2014)

[Qu10] Quintero-Leyva, B.: The multi-group integro-differential equations of the neutron diffusion kinetics. Solutions with the progressive polynomial approximation in multi-slab geometry. Ann. Nucl. Energy **37**, 766–770 (2010)

[TuEtAl19] Tumelero, F., Bodmann, B., Vilhena, M.T., Lapa, C.M.F.: On the solution of the neutron diffusion kinetic equation in planar geometry free of stiffness with convergence analysis. Ann. Nucl. Energy **125**, 272–282 (2019)

# Index





C. Constanda et al. (eds.), *Integral Methods in Science and Engineering*,
https://doi.org/10.1007/978-3-031-07171-3